TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Anàlisi de les diferents eines de modelització en el camp del risc de crèdit

Autor: Albert Martos Ramírez

Director: Hector Rufino Alcalde

Departament: Econometria, Estadística i Economia Aplicada

Convocatòria: Juny 2021

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

AGRAÏMENTS

Abans d'iniciar l'informe, volia agrair en primera persona l'ajuda, les reunions telemàtiques i el suport que m'ha brindat l'Hector Rufino Alcalde des de l'inici del treball. Tot i les ensopegades que han sorgit durant aquests cinc mesos, el projecte s'ha fet molt amè.
Moltes gràcies Hector.

RESUM

La importància de gestionar òptimament el risc de crèdit ha evolucionat a l'alça durant l'última dècada i per això bancs i institucions demanden una solució al respecte. Per donar resposta al col·lectiu financer, aquesta investigació es centra en una comparació exhaustiva de les capacitats predictives entre els mètodes clàssics utilitzats per la majoria d'institucions financeres i els mètodes alternatius que proporciona l'aprenentatge automàtic. Tot això amb la finalitat de trobar l'algoritme que millor respon a les necessitats de les entitats tenint en un compte els beneficis i costos de cada model utilitzant una base de dades realista que podria tenir perfectament un banc qualsevol.

Paraules clau: entitats financeres, risc de crèdit, probabilitat d'impagament, models clàssics, Machine Learning, Interpretative Machine Learning.

Classificació del treball segons la Mathematics Subject Classification (MSC):

- 62J12 – Models lineals generalitzats.
- 62J20 – Diagnòstics.
- 62P05 – Aplicacions a les ciències actuarials i matemàtiques financeres.
- 65C20 – Models, mètodes numèrics.
- 68T05 – Aprenentatge i sistemes adaptatius.
- 91G40 – Risc de crèdit.
- 91G70 – Mètodes estadístics, econometria.
- 97M30 – Matemàtiques financeres i actuarials.

ABSTRACT

The importance of optimally managing credit risk has increased over the last decade, which is why banks and other institutions are demanding a solution in this regard. To respond to the financial community, this research focuses on a thorough comparison of the predictive capabilities between the classical methods used by most financial institutions and the alternative methods provided by Machine Learning methodology. All this to find the algorithm that best meets the needs of banking institutions considering the benefits and costs of each model using a realistic database that could have perfectly any bank.

Key words: banks, credit risk, credit risk modelling, probability of default, classical models, Machine Learning, Interpretative Machine Learning.

Project classification according to the Mathematics Subject Classification (MSC):

- 62J12 – Generalized linear models.
- 62J20 – Diagnostics.
- 62P05 – Applications to actuarial sciences and financial mathematics.
- 65C20 – Models, numerical methods.
- 68T05 – Learning and adaptive systems.
- 91G40 – Credit risk.
- 91G70 – Statistical methods, econometrics.
- 97M30 – Financial and insurance mathematics.

ÍNDEX GENERAL

1. INTRODUCCIÓ....11
2. METODOLOGIA....14
3. CONCEPTES BÀSICS....17
   3.1 Risc financer....18
   3.2 Risc de crèdit....19
   3.3 Probabilitat d'impagament (PD)....21
   3.4 Exposure at Default (EAD)....24
   3.5 Loss Given Default (LGD)....24
4. ALTERNATIVES ALS MODELS TRADICIONALS EN EL CAMP DEL RISC DE CRÈDIT....28
   4.1 Situació actual del ML en les entitats creditícies....28
   4.2 Quantificació dels beneficis....29
      4.2.1 Corba ROC i àrea sota la corba (AUC)....29
      4.2.2 Índex de gini....33
   4.3 Factors de cost....36
   4.4 Funció de cost....39
5. EXPOSICIÓ DEL CAS PRÀCTIC....43
   5.1 Base de dades....43
   5.2 Anàlisi descriptiu avançat....50
      5.2.1 Variables numèriques....52
      5.2.2 Variables categòriques....60
   5.3 Construcció dels models I: Modelització tradicional....71
      5.3.1 Model logístic....71
      5.3.2 Model logístic amb penalització Lasso....77
      5.3.3 Model logístic amb penalització Ridge....78
   5.4 Construcció dels models II: Modelització alternativa....79
      5.4.1 Arbres de decisió....80
      5.4.2 Random Forest....84
      5.4.3 Support Vector Machine....87
      5.4.4 XGBoost....88
   5.5 Quantificació dels beneficis....91
   5.6 Quantificació dels costos i resultats....97
   5.7 Aplicació de Interpretative Machine Learning (IML)....99
6. CONCLUSIONS....105
7. BIBLIOGRAFIA....109
8. ANNEX....113
   8.1 Descripció de les variables....113

8.2 Mapes de calor de les correlacions 119
8.3 Resultats dels models clàssics 124
8.3.1 Primer model logístic amb totes les variables 124
8.3.2 Segon model logístic amb totes les variables 127
8.3.3 Tercer model logístic amb totes les variables 130
8.3.4 Model logístic amb menys variables 133
8.3.5 Model Lasso amb totes les variables 136
8.3.6 C.V Lasso amb menys variables 139
8.3.7 Model Lasso amb menys variables 140
8.3.8 Model Ridge amb totes les variables 142
8.3.9 C.V Ridge amb menys variables 145
8.3.10 Model Ridge amb menys variables 146
8.4 Interpretative Machine Learning 148
8.4.1 IML d'un client amb una baixa probabilitat d'impagament 148
8.4.2 IML d'un client amb una elevada probabilitat d'impagament 149
8.4.3 IML d'un client amb una probabilitat d'impagament mitjana 150
8.5 Codi emprat 151
8.5.1 Codi SQL per a la creació dels fitxers d'entrenament i test 151
8.5.2 Anàlisi descriptiu 167
8.5.3 Construcció dels models clàssics amb totes les variables 176
8.5.4 Construcció dels models clàssics amb menys variables 181
8.5.5 Construcció dels models alternatius 186
8.5.6 Script per al càlcul de l'índex de gini i AUC. Gràfics de probabilitat . 194
8.5.7 Interpretative Machine Learning per al model XGBoost 198

1. INTRODUCCIÓ

L'auge en l'interès per quantificar el risc creditici es deu principalment a les conseqüències que va comportar la forta crisi financera internacional del 2008 a moltes empreses i particulars. Una quantitat nombrosa de negocis no va aguantar l'impacte econòmic de la crisi i van haver de tancar o declarar-se en fallida. Això provocà grans pèrdues monetàries en totes aquelles entitats financeres que havien concedit crèdits anteriors a la recessió a empreses que van ser incapaces de mantenir els pagaments. Per evitar una pèrdua de capital significativa, els bancs i, conjuntament amb altres institucions i organismes, estan invertint temps per a trobar nous algoritmes que optimitzin la presa de decisió per cada sol·licitud de crèdit dels prestataris. Així doncs, es busca un conjunt d'eines estadístiques que permeti preveure futurs incompliments.

El projecte s'endinsa en l'anàlisi del rendiment de múltiples metodologies per a la predicció de morosos emfatitzant en les probabilitats d'impagament estimades. En especial, una comparació exhaustiva entre els mètodes clàssics utilitzats per la majoria d'entitats financeres i la metodologia alternativa de Machine Learning. Una de les principals motivacions serà, per tant, trobar un model amb una elevada capacitat predictiva tenint en compte els seus beneficis i els seus inconvenients. Per això s'utilitzaran diferents algoritmes derivats de la metodologia clàssica i alternativa com podrien ser la inserció de penalitzacions en el model logístic clàssic o els Random Forest. Paral·lelament amb aquesta idea, és també destacable saber expressar adequadament, a partir del model, els motius pels quals s'accepta o es denega un préstec qualsevol i és especialment en l'últim cas quan pot afectar la vida personal del subjecte. Altres objectius marcats en la realització del treball són entendre les diverses mesures que caracteritzen el risc de crèdit com també conèixer el procediment que podria seguir un banc per a dur a terme la modelització dels impagaments a partir de les dades històriques dels clients amb les quals disposa.

Es va escollir l'anàlisi dels models per al risc de crèdit com a tema per donar resposta a una de les inquietuds personals referents a la modelització estadística en el camp de les finances. Alhora, alguns dels temes més apassionants que s'han fet a la carrera són la creació de models i la codificació, aspectes que es tracten durant tota la pràctica.

Aquest treball s'estructura en dues parts diferents: una de teoria dividida en dos grans blocs diferenciats i la part pràctica. La primera part de teoria detalla tots els aspectes necessaris per posar en context el risc de crèdit i la segona explica els

aspectes imprescindibles a considerar quan es modelitza una situació del risc creditici. Per últim, l'exposició de la part pràctica tracta la construcció del fitxer de dades mitjançant l'ús de dades reals trobades d'internet d'una entitat financera com també de l'elaboració pas a pas dels diferents algoritmes i els respectius resultats.

2. METODOLOGIA

El projecte es basa en les idees principals de l'estudi *"Machine Learning in credit risk: Measuring the dillemma between prediction and supervisory cost"* realitzat pels autors Andrés Alonso i José Manuel Carbó i publicat pel Banc d'Espanya l'any 2020. En l'article es tracta el marge de millora que poden aportar els models de Machine Learning respecte els models clàssics utilitzats en el sector financer des de fa dècades utilitzant l'índex de gini i l'AUC com a mètriques de bondat d'ajust. No només es quantifiquen els beneficis sinó que també ho fan per als costos a partir de l'algoritme de la caixa negra. D'aquesta relació benefici – cost se'n extreu el model definitiu.

Per a donar resposta a tots els objectius marcats s'ha agafat una base de dades procedent de la pàgina web *Kaggle*. Aquesta proporciona les característiques financeres i socials de més de dos-cents mil clients que han demanat un préstec en una entitat financera. Inclou també si l'operació financera va acabar en un impagament, el pilar on es sustenta l'estudi. Les dades estan formades per múltiples taules o arxius que contenen informació diferents entre elles. A partir d'aquestes s'han construït dos fitxers d'entrenament i de testatge per avaluar el rendiment de cadascuna de les metodologies utilitzades.

Els mètodes emprats per a la modelització de les dades i per tant, per a la estimació de les probabilitats d'impagament, formen part d'un gran ventall d'opcions disponibles per als estadístics. El model logístic, amb penalització Lasso o Ridge, models titllats de clàssics, seran comparats amb algoritmes procedents del Machine Learning popularment coneguts per oferir bons resultats. Els arbres de decisió, el Random Forest, el Support Vector Machine i el XGBoost són els algoritmes que s'estudien i es comparen en fases posteriors del projecte d'investigació.

Per al correcte desenvolupament del projecte, s'ha optat per utilitzar tres llenguatges de programació diferents: R, Python i el SQL implementat també amb Python. En la primera fase del projecte s'han construït els fitxers de dades definitiu (entrenament i testatge) a partir de la unió de múltiples taules mitjançant codi SQL. L'R, en canvi, s'ha fet servir únicament per a la construcció dels models clàssics i per la avaluació o comparació entre els diferents models considerant especialment les estimacions realitzades al conjunt de dades de testatge. Les dues llibreries més importants que s'han fet servir són el *openxlsx* (per la lectura dels fitxers) i el *glmnet* (per la construcció dels models clàssics amb penalització). Per una altra

banda, s'ha utilitzat Python (a partir de Spyder (Anaconda)) per a la transformació i construcció de taules, la descriptiva de les dades i la creació dels diferents algoritmes de Machine Learning. El *pandasql* (càrrega del llenguatge SQL), *pandas* (construcció i ús de dataframes), *matplotlib* (elaboració de gràfics avançats) i el *sklearn* (creació dels algoritmes de Machine Learning) són les llibreries seleccionades per a la implantació dels models alternatius. La implementació d'Interpretative Machine Learning també s'ha fet amb el Python, però utilitzant Jupyter Notebook en comptes de Spyder amb la llibreria *epi5*.

3. CONCEPTES BÀSICS

Imagina una dona amb el seu fill esperant a que el semàfor es posi en verd per tal que puguin passar pel pas de vianants. Al mateix instant, un home travessa la carretera tot i tenir el semàfor en vermell. Es desconeixen els motius pels quals l'home ha decidit no esperar-se i creuar, però ell ha assumit internament que podria ser envestit per un vehicle. En canvi, la dona i el seu fill esperen tranquil·lament sabent que no els hi passarà res en cap de les circumstàncies. D'aquesta manera l'home està assumint un risc al travessar mentre que la família no presenta cap risc. Un altre situació semblant podria ser la dels habitants japonesos que viuen en un arxipèlag propens a terratrèmols i tsunamis. En el dia a dia assumeixen aquest risc intrínsec de viure episodis puntuals de moviments tectònics.

Així doncs, què és exactament el risc? Segons l'enciclopèdia catalana és la "*contingència desfavorable a la qual està exposat algú o alguna cosa, perill incert*" (Enciclopedia Catalana, 2021). En altres paraules, és la exposició a una situació on hi ha una certa possibilitat d'estar en perill o de que succeeixi un esdeveniment no desitjat per a la persona. D'aquesta manera, una situació de risc seria aquella en la que podria ocórrer quelcom dolent, però que no té per què passar. Per tant, podria definir-se també com una mesura quantitativa per a quantificar aquesta possibilitat com a probabilitat.

Al tractar-se d'una mesura quantitativa i deguda a l'existència de factors externs, anomenats factors de risc, es poden trobar solucions o alternatives que permetin disminuir la probabilitat del succés no favorable. I és que les persones conviuen diàriament amb moltes situacions on l'exposició al perill és present, però que en la majoria d'ocasions s'aconsegueix disminuir el risc gairebé en la seva totalitat, com ara la incorporació d'airbags als automòbils.

L'exposició no és sempre la mateixa i succeeix sempre sota unes circumstàncies o característiques concretes, com per exemple no portar casc al conduir una motocicleta o bé viure a prop d'una central nuclear. Això comporta a que existeixi un gran ventall de tipus de risc, entre els quals es podrien trobar els biològics, físics i financers. Aquest últim és el de més interès per entendre el plantejament del risc de crèdit i les seves motivacions per a prevenir-lo tant com es pugui.

3.1 Risc financer

L'anàlisi del risc financer és una de les variables més importants per a totes les empreses, inversors o qualsevol altra persona o organització que pugui patir algun esdeveniment amb conseqüències financeres, especialment negatives. Fa referència a la incertesa en el rendiment d'una inversió deguda a canvis en els sectors on s'opera, a la impossibilitat de la devolució del capital per alguna de les parts involucrades o també a causa de la inestabilitat del mercat.

Normalment el context del risc financer es troba en la possible pèrdua monetària o de capital que es podria assolir si les circumstàncies són desfavorables. És per aquest motiu que és d'interès trobar solucions per a reduir l'exposició.

Es poden distingir quatre tipus de riscs financers bàsics:

1. Risc de mercat: Es dona quan es compren actius financers en el mercat, donat que no es pot assegurar el retorn de la inversió. Es un dels més importants que existeixen degut a que és el més comú i perquè el seus efectes són molt amplis, ja que es deu, sobretot, a la dinàmica entre l'oferta i la demanda del mercat. Aquesta dinàmica respon, en gran part, a les incerteses econòmiques del moment que afecten en gran part al rendiment de totes les empreses. Existeixen tres factors claus que permeten explicar aquest tipus de risc:
    - Els tipus d'interès: Es tracta del risc associat als moviments en contra dels tipus d'interès. En funció de la situació de l'empresa o individu pot interessar més o menys la pujada o baixada d'aquests valors.
    - El tipus de canvi: Fa referència a les possibles variacions en la taxa de canvi en el mercat de divises.
    - Demanda i oferta: Afecta en gran part als accionistes i inversors en que poden patir una davallada del valor de les accions de les empreses que posseeixen, disminuint així el valor de la cartera de la persona.

2. Risc de crèdit: És aquell que ocorre quan una de les dues parts d'un contracte no assumeix les seves obligacions de pagament. S'ha dedicat un apartat complet per a explicar-lo.
3. Risc de liquiditat: Aquest tipus es refereix a la possibilitat de ser incapaç de complir les obligacions que té una empresa i això comporta que la seva situació financera i la seva existència es vegi amenaçada.

4. Risc operatiu: Es tracta d'aquella exposició al risc no externa. Succeeixen especialment per la falta de controls interns dins de l'empresa, fallides tecnològiques, mala administració, errors humans, fallides en els processos, la falta de personal, entre d'altres.

És important saber identificar quins són els riscs potencials de la situació concreta i especialment avaluar els possibles impactes que podrien tenir. En cas contrari no es podrà reaccionar ni tampoc prevenir qualsevol esdeveniment que podria malmetre la imatge de l'empresa, perdre competitivitat o fins i tot arribar a declarar la fallida.

3.2 Risc de crèdit

Aquest tipus de risc sorgeix en la possibilitat de que una de les parts del contracte de l'instrument financer, normalment els clients, empreses o fins i tot els governs dels països, incompleix les seves obligacions de pagament per motius d'insolvència o per la incapacitat de pagament i produeixi, d'aquesta manera, pèrdues financeres a l'altra part. La pèrdua de capital, coneguda com pèrdua esperada (PE) pot calcular-se a partir de la fórmula següent:

$$PE = PD * EAD * LGD$$

On la probabilitat d'incompliment (PD) i en anglès *probability of default* recull la probabilitat de que el prestatari no compleixi amb les seves obligacions contractuals. L'EAD (*exposure at default*) fa referència a l'exposició a l'impagament que es veu definida com l'import del deute restant en el moment que el client deixa de pagar. La LGD (*loss given default*) és el rati de pèrdua en cas d'incompliment, un percentatge d'un préstec que, una vegada impagat i efectuades totes les gestions per a poder tornar-lo a cobra, resulta finalment incobrable.

Un dels objectius principals és quantificar la probabilitat d'impagament existent amb els prestataris. Es tracta d'una mesura molt important per a totes aquelles entitats que ofereixen préstecs, béns o qualsevol altre producte a canvi d'una retribució major a l'ofert a l'inici del contracte. D'aquesta manera, i especialment per als bancs i altres entitats financeres dedicades a la concessió de crèdits com podrien ésser *Cofidis* o *Creditea*, és important saber gestionar adequadament les incerteses que es generen i per tant quantificar quant de perillós pot arribar a ser un client per tal d'evitar les màximes pèrdues possibles. És important per als bancs tenir una mesura de les pèrdues que poden derivar dels impagaments per poder generar provisions o reserves de capital. Les provisions són els diners que els bancs tenen

guardats per fer front als impagaments que esperen tenir durant un temps establert, normalment d'un any, de forma que no causin danys en els seus comptes.

Ara bé, no existeix una única forma de risc, sinó que n'hi ha vàries. Primer, el propi risc d'impagament i que s'ha descrit prèviament. Segon, la incertesa del pagament va associada a l'actitud o característiques del deutor i les variables de mercat. Seguidament el risc colateral que depèn de les garanties del deute existents. Per últim, el risc de concentració que és present quan s'entreguen quantitats copioses o excessives a pocs prestataris que provoca en un augment de les pèrdues en el supòsit que acabin essent morosos.

Va ser a partir de la greu crisi financera del 2008 – 2009 a escala mundial que multituds de bancs i empreses van declarar-se en fallida degut a la gran quantitat d'hipoteques amb un alt risc creditici en els seves respectives carteres de balanç. El gran nombre d'empreses declarades en ruïna va comportar a una necessitat urgent d'entendre i saber com gestionar el risc creditici. És el que s'anomena popularment com a *Credit Risk Management*. Sense una avaluació íntegra d'aquest, els bancs no tenen cap forma de saber si les reserves de capital reflecteixen amb tota seguretat els riscos presents, o i si les reserves destinades a les pèrdues dels préstecs cobreixen correctament les pèrdues potencials de crèdit a curt termini. Si un banc presenta una reserva de capital insuficient no serà capaç de fer front a les despeses causades pels impagaments mentre que si és massa elevada sí podrà pagar les insuficiències de capital, però perdrà competitivitat per no haver invertit tant com la resta d'entitats bancàries. Per aquest motiu el millor que es pot fer es trobar un punt mig en el cúmul de provisions, sent conservadors per tenir un cert marge per sobre.

Perquè aquests esdeveniments no succeeixin, i garantint que les reserves de capital reflecteixin apropiadament el perfil del risc, s'ha d'implementar una solució integrada i quantitativa que no afecti en el rendiment del banc. Aquesta hauria d'incloure:

- Una gestió de l'etapa de modelització perfecta.
- Avaluació i monitorització en temps real.
- Comprovar amb altres dades que el model funciona segons els criteris o necessitats establertes.
- Disposar d'eines de visualització de dades i d'intel·ligència de negoci que permetin extreure la informació imprescindible per aquelles persones que la necessitin.

Tot i poder implementar aquesta solució, s'ha de tenir en compte que el procés serà diferent si s'està analitzant el risc per a grans empreses i corporacions o si bé són particulars o pymes. En el cas del primer grup es fa un anàlisi individualitzat que fan un conjunt d'experts que mesuren el risc d'impagament, mentre que per l'altre agrupació es construeixen models que avaluen les operacions massivament a partir dels impagaments històrics que ha tingut l'entitat financera. Aquest punt es detallarà més en el següent apartat amb la distinció entre puntuació y la classificació de crèdit.

3.3 Probabilitat d'impagament (PD)

Com s'ha descrit en l'apartat anterior, la probabilitat d'incompliment és una mesura de qualificació creditícia que se li dona a cada client segons les seves característiques i de la situació econòmica que es viu en aquell precís moment amb la finalitat d'estimar, durant un cert període de temps que normalment és d'un any, quant de probable és que acabi no pagant. Quan arriba un client que desitja un préstec a l'entitat financera, el banc estudia el seu cas i determina la seva probabilitat. En el supòsit que acabi presentant un valor elevat, s'aplicaran interessos importants en el préstec per assegurar-se'n de la seva viabilitat.

Per a orientar les PD s'utilitzen dues eines anomenades *Scoring* i *Rating* amb unes característiques concretes per a cada metodologia.

Respecte el primer, es basa en l'ús d'instruments estadístiques focalitzats en l'estimació de la probabilitat d'impagament per a després decidir si la persona o qualsevol altra institució interessada en el crèdit l'acaba rebent o no. En l'*scoring* s'inclouen particulars i les pymes. La puntuació de crèdit impacta en una multitud de transaccions financeres que inclouen les hipoteques, les targetes de crèdit o els préstecs privats.

Existeixen popularment dues maneres per a obtenir les puntuacions. Es pot utilitzar la FICO, que és la més utilitzada amb aproximadament un 90% del sector financer o la puntuació Vantage. Aquestes dues metodologies són molt semblants, si bé poden prendre valors entre 300 i 850, i la única diferència destacable és la classificació de l'individu dins d'aquests valors. Tot i això la puntuació màxima que es pot obtenir és la mateixa en ambdós casos.

La puntuació de crèdit individual s'obté tenint en compte cinc factors als quals se'ls incorpora un pes per a cadascun d'ells:

- L'historial dels pagaments. (35%)
- El capital que deu el client. (30%)
- La llargada de l'historial dels crèdits. (15%)
- El nou crèdit. (10%)
- Altres informacions. (10%)

Rating	Escala de FICO	Escala del Vantage
Excel·lent	800 – 850	781 - 850
Molt bona / Bona	740 – 799	661 – 780
Bona / Justa	670 – 739	601 – 660
Justa / Pobre	580 – 669	500 – 600
Pobre / Molt pobre	300 - 579	300 – 499

Taula 1. Classificació de les puntuacions de risc segons el sistema.

Aquest valor obtingut és només una mesura del risc que presenta el client i no proporciona cap estimació de la probabilitat d'impagament. La principal funció de l'*scoring* és classificar els nivells dels riscos des dels més elevats fins als més petits. Existeixen altres metodologies més avançades per a modelitzar el risc de crèdit millors la pròpia puntuació creditícia per a avaluar la probabilitat d'impagament, però requereixen de més coneixements tecnològics i de programari que pocs experts tenen.

La valoració dels crèdits i l'avaluació per a grans empreses, corporacions i governs es fa normalment a partir d'una empresa dedicada a la qualificació de crèdit. El que fan aquestes empreses és mesurar la solvència de les entitats. Si la qualificació és elevada voldrà dir que hi ha una alta probabilitat que l'empresa torni el préstec en la seva totalitat sense presentar problemes, mentre que si és baixa fa referència, normalment, a que ha tingut problemes per pagar en el passat i que pot presentar la mateixa dinàmica en el futur. Així doncs, la qualificació del crèdit *rating* que obté una empresa acaba tenint un pes important per a l'obtenció del préstec.

La determinació del crèdit *rating* es fa mitjançant un sistema alfabètic elaborat per les agències de qualificació de risc Moody's, Standard & Poor's (S&P) i Fitch tot i que no existeix un codi homogeni sinó que cadascuna presenta algunes diferències respecte les altres. Per exemple, i tal com es pot observar a la taula que es mostra a continuació, Moody's utilitza Aaa per a distingir aquelles empreses que presenten un risc mínim, mentre que S&P i Fitch utilitzen AAA com a excel·lència. Per contra, el Caa3 seria el grup que presentaria major risc creditici. A mesura que s'avança numèricament, amb l'abecedari i es redueixen les *a* augmenta progressivament el risc.

Taula 2. Sistema alfabètic que utilitza l'agència Moody's. (Pelz, 2019)

Al cap i a la fi, el *rating* o la puntuació serveix per classificar el client i un cop està classificat, en funció de la puntuació, se li assigna una probabilitat d'entrar en impagament, que és la PD.

Aquestes dues tècniques són molt utilitzades en el món del risc de crèdit, però no s'utilitzaran en aquest projecte. Existeix una altra manera de calcular-les a partir de les dades històriques d'un banc en que es coneixen les característiques dels clients com també si van incomplir amb les seves obligacions de pagament o no. Aquesta és la metodologia que es seguirà i que s'explicarà detalladament en els capítols posteriors.

3.4 Exposure at Default (EAD)

L'exposició en el moment d'incompliment és un factor necessari a tenir en compte per al càlcul de la pèrdua esperada o capital, definida com l'import del deute pendent de pagar en el moment que es produeix l'esdeveniment. Aquesta pèrdua depèn de la quantitat amb la qual el banc està exposat respecte el prestatari al moment de l'impagament ja que l'incompliment té lloc en una data futura desconeguda.

Normalment els bancs calculen l'EAD per cadascun dels crèdits disponibles per després determinar el risc conjunt d'impagament. Es tracta d'un nombre dinàmic que varia cada vegada que el prestatari torna els diners corresponents. El que s'acaba mesurant és la extensió en la qual el banc es veu exposat respecte la contrapart en el cas que, passat un temps, aquest no pagui.

L'EAD pot calcular-se de la següent manera:

$$EAD = Disposat + Disponible \times CCF$$

En la fórmula, el disposat fa referència a la quantitat que s'ha fet servir del límit del crèdit mentre que el disponible és aquell import restant que el prestatari encara té pendent per a consumir. El CCF, dit *crèdit conversor factor* és el rati percentual entre la quantitat monetària que el client demana d'un nou crèdit i el capital restant del préstec actual. És a dir, si una persona posseeix un crèdit al que li resten 20.000€ disponibles i demana un altre de 5.000€. Llavors el CCF resultant seria del 25%.

3.5 Loss Given Default (LGD)

Es pot definir la Loss Given Default (LGD) o severitat com una mètrica imprescindible en el risc de crèdit definida com la quantitat de diners que el banc o entitat financera perd en el moment que el prestatari incompleix amb el pagament, expressada normalment en forma de percentatge sobre el total de l'exposició.

Les entitats creditícies solen determinar les pèrdues analitzant altres impagaments. La quantificació de les pèrdues és complexa i és necessari analitzar un conjunt ampli de variables que, després d'un procés exhaustiu, permet determinar la LGD. Es podria considerar un cas en que un banc qualsevol ofereix un préstec de 10.000.000€ d'euros a una empresa i aquesta, per causes desconegudes, acaba no pagant. La pèrdua per al banc no és necessàriament de deu milions, sinó que existeixen altres factors, com podrien ser la quantitat de béns que disposa l'entitat en forma de

garantia del contracte o si ja s'havien pagat quotes prèviament que reduirien substancialment la pèrdua inicial del capital atorgat a l'empresa. La gestió que realitza l'entitat davant els impagaments, ja sigui amb l'enviament de cartes, costos judicials, advocats o tenir empleats destinats a la gestió de la morositat fa augmentar-ne els costos i acaben repercutint en el càlcul de la LGD.

Una altra possible situació, com a circumstància puntual, és que el client no pagui però que en un termini raonable retorni els rebuts vençuts i es posi al corrent del contracte. En aquest cas la LGD seria un percentatge proper al 0%, donat que es considerarien com a pèrdua les gestions realitzades pel banc.

La LGD i l'EAD són dos factors molt semblants. La diferència principal entre aquests dos és que la LGD inclou la possibilitat de recuperar capital monetari de l'incompliment. En el cas que el prestatari incompleix en una hipoteca després de varis anys pagant mensualment, l'EAD seria la quantitat que li quedava per a liquidar-la. Si el banc vengués l'habitatge recuperaria una certa quantitat de l'EAD, que sí estaria inclosa en el càlcul de la LGD.

Vegem un cas que té en compte tots els factors vists fins ara. Imagina una parella que demana una quantitat de 225.000€ per un préstec d'un condomini. Després de pagar múltiples quotes durant varis anys, la parella s'estanca financerament parlant i tenen problemes per a seguir amb els pagaments rutinaris fins que acaben incomplint amb el pagament acordat. En aquest moment, l'EAD és de 186.500€. Després de tot això, el banc es fa amb el control del condomini i pot vendre'l per 160.000€ obtenint així unes pèrdues netes de 26.500€ i un *loss given default* del 14,21%.

$$Pèrdues\ netes = 186.500€ - 160.000€ = 26.500€$$

$$LGD = \frac{(186.500€ - 160.000€)}{186.500€} = 0{,}14209 = 14{,}209\%$$

En aquesta mateixa situació, la pèrdua esperada es calcula tenint en compte els següents factors: un LGD del 14,21%, una probabilitat d'incompliment (PD) del 100% ja que es coneix que van incomplir i una EAD de 186.500€, obtenint la següent:

$$Pèrdua\ esperada\ (PE) = 0{,}14209 \times 1 \times 186.500€ = 26.500€$$

Aquesta pèrdua esperada s'ha calculat a partir d'un cas conegut, però podria ser que l'entitat financera estigui interessada en conèixer les pèrdues potencials que podrien patir, sense que sigui segur que acabi succeint. Per això, el banc assumeix que la

probabilitat d'impagament de la parella no és del 100% sinó del 60%. D'aquesta manera la nova pèrdua esperada seria de 15.900€.

$$Pèrdua\ esperada\ (PE) = 0{,}14209 \times 0.6 \times 186.500€ = 15.900€$$

4. ALTERNATIVES ALS MODELS TRADICIONALS EN EL CAMP DEL RISC DE CRÈDIT

Els mètodes emprats fins al dia d'avui pels bancs per a la predicció i previsió d'impagaments han estat sempre els més simples i fàcils d'entendre: els models lineals generalitzats, concretament el model lògit i també els que inclouen un factor de penalització, com podria ser el model lògit amb penalització Lasso. Però tot fa pensar que podria canviar en els propers anys.

Els coneixements matemàtics requerits per a l'ús de les metodologies derivades de la intel·ligència artificial, que inclou també el Machine Learning (ML), existeixen des de la dècada dels 50, però la seva implantació i utilització ha pogut donar a llum gràcies als avenços tecnològics produïts en les últimes dècades. El ML podria definir-se com un conjunt de tècniques estadístiques que permeten als ordinadors aprendre i millorar únicament a partir de la inserció d'una gran col·lecció de dades.

Les eines derivades del ML poden usar-se per a la predicció i classificació segons quin sigui el problema i que per tant encaixarien amb el concepte del risc de crèdit. Segons múltiples enquestes, les entitats financeres estan adoptant aquestes tècniques en moltes àrees de la gestió del risc creditici com podrien ser en el càlcul de les provisions o el *credit scoring*. És conegut que els algoritmes de Machine Learning són capaços de millorar significativament la capacitat predictiva en la majoria de situacions respecte els models tradicionals i que farien, al final, decantar-se per aquests models. Malgrat aquesta millora, també és capaç de produir mals de caps a causa de la seva complexitat i dificultat, fet que ha comportat les reticències dels supervisors bancaris en que les entitats utilitzessin aquestes metodologies. Per aquests motius fa que sigui necessària l'existència d'una mena de manual per a dominar i entendre els nous algoritmes.

4.1 Situació actual del ML en les entitats creditícies

Segons l'Institut Financer Internacional (IIF) les tècniques de Machine Learning s'utilitzen en el sector del crèdit *scoring*. S'han fet múltiples enquestes per saber l'estat d'aquests models en les entitats financeres. Una d'elles, que consistia en demanar-ho en 60 empreses a nivell internacional, acabà concloent que el 37% d'elles utilitzaven un procés íntegre de ML en aquest sector. Una altra enquesta desenvolupada per les autoritats bancàries europees, en l'àrea del capital regulatori, mostra una davallada en l'ús d'aquesta metodologia al passar d'un 20% l'any 2018 al

10% el següent any. Les empreses constaten que necessiten models simples i fàcils d'interpretar, mentre que els de ML són més complexos d'explicar, però no impossible.

De totes maneres, en els darrers anys ha augmentat el nombre d'entitats que utilitzen aquests models en producció o en proves pilot. No obstant, el creixement no ha estat homogeni arreu dels continents, sinó que aquelles que disposen d'un cert grau de tecnologia avançada són les que acaben desenvolupant i utilitzant més el ML i, les que presenten un nivell menor acaben utilitzant-les esporàdicament.

Aquí es presenta el problema: és necessiten coneixements i certa tecnologia per poder construir i supervisar aquests algoritmes. Moltes empreses no són capaces d'assumir aquest cost per aconseguir una possible millora en les seves capacitats predictives. Andrés Alonso i José Manuel Carbó proposen una solució en l'article "*Machine Learning in Credit Risk: Measuring the dilema between prediction and supervisory cost*" publicat pel Banc d'Espanya per a comparar els models tradicionals amb els innovadors per a l'estimació de les probabilitats d'impagament a partir de la quantificació d'unes mètriques que reflecteixen els beneficis i costos de cada model. Malgrat que el paper es basa únicament en la predicció de les PD, la metodologia de ML també és aplicable per a la resta de paràmetres vistos en l'apartat 3.

4.2 Quantificació dels beneficis

La poca disponibilitat de bases de dades de llargues series temporals d'alta qualitat dificulta la recerca en aquest camp del risc de crèdit, però és necessari establir formes per a mesurar els beneficis i els costos dels models. Quan es fa referència als beneficis es parlarà de la capacitat de predicció, del nombre d'impagaments predits adequadament o del nombre de prediccions totals correctes.

La mètrica emprada en el *paper* i que resulta en ser la més utilitzada per a establir els guanys és l'àrea sota la corba perquè permet contenir amb un únic valor tota la informació d'interès. Paral·lelament s'ha cregut convenient afegir una altra mètrica molt utilitzada en el sector creditici per a quantificar la desigualtat o diferències entre grups, l'índex de gini.

4.2.1 Corba ROC i àrea sota la corba (AUC)

La corba ROC és una de les metodologies més utilitzades per a observar el comportament dels models a l'hora de classificar els individus. Pot definir-se com

una corba de probabilitat que comença des del 0 i acaba a l'1. Tot i això, es sol utilitzar un derivat de la corba ROC, l'AUC. Aquest mesura l'àrea que deixa per sota la corba de probabilitat, d'aquí les seves sigles AUC (*Area under the curve*), i ambdós serveixen per informar de la capacitat que té el model per a distingir entre les classes de la variable resposta. L'àrea sota la corba pot prendre els mateixos valors que la corba ROC i que, quant més s'apropi a 1 millor capacitat predictiva tindrà el model.

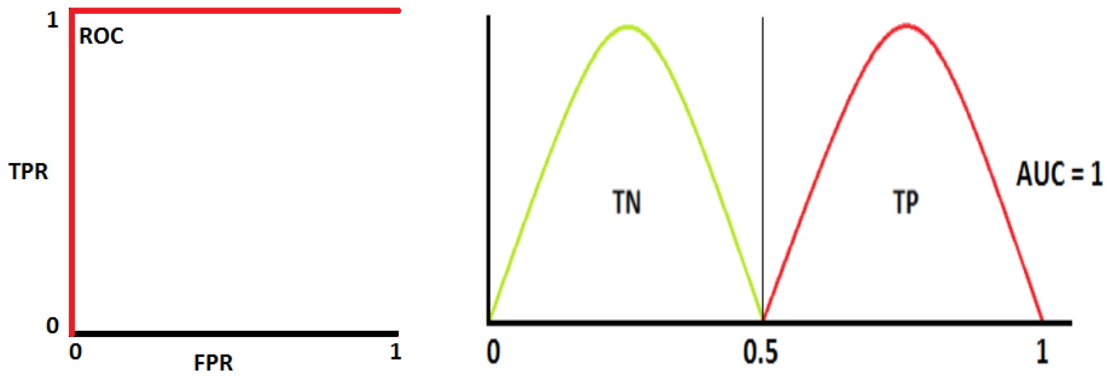
És important conèixer com es realitza el gràfic de la corba ROC. Es representa gràficament la taxa de vertaders positius contra la taxa de falsos positius. La taxa de vertaders positius (TPR), també anomenada sensibilitat, és la proporció de casos positius que han estat ben classificats pel model. En canvi, la taxa de falsos positius (FPR) és la proporció de casos negatius que el model detecta com a positius. Aquest últim es pot calcular fent el complementari de la taxa de vertaders negatius (especificitat). Llavors es representa la FPR a l'eix de les abscisses i la TPR al de coordenades.

$$Sensibilitat = \frac{Vertaders\ positius}{Vertaders\ positius + Falsos\ negatius}$$

$$Especificitat = \frac{Vertaders\ negatius}{Vertaders\ negatius + Falsos\ positius}$$

$$Taxa\ de\ falsos\ positius = 1 - Especificitat$$

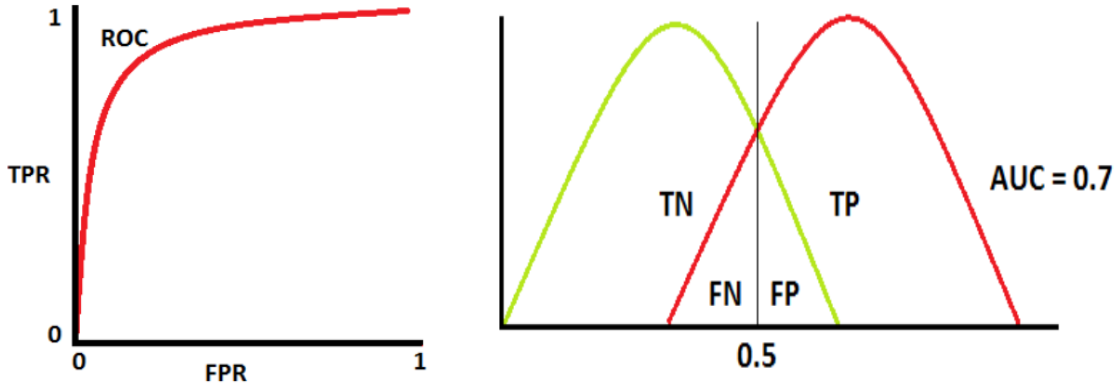
L'AUC, tal i com s'ha dit anteriorment, és l'àrea que hi ha per sota de la corba ROC. Es poden trobar quatre casos principals dels resultats de la corba ROC i de l'AUC. Primer, que el model predigui perfectament. Segon, en que les classificacions estiguin força bé, però amb errors. Un altre en que no sàpiga classificar les observacions i un últim que seria una predicció completament errònia.



Figures 1 i 2. Gràfics corba ROC perfectes. (Narkhede, 2018)

El primer cas es mostra en les figures 1 i 2. Es tracta de la situació ideal, però que a la pràctica és molt difícil que succeeixi. Ara bé, si podrien trobar-se valors pròxims a 1. En aquesta situació l'àrea sota la corba és 1 degut a que totes les classificacions realitzades han estat correctes. Això significa que la taxa de vertaders positius és d'1 i, per tant, la de falsos negatius és de 0 ja que cap negatiu ha estat classificat com a positiu. D'aquesta manera s'obté el gràfic de la figura 1. La figura 2 mostra la classificació realitzada. Per exemple, en el cas de la modelització de les probabilitats d'impagament, es pot considerar un valor crític de 0.5 per a distingir si una probabilitat estimada acaba essent classificada com a impagament (major que 0.5, com a TP) o no ho és (menor que 0.5, com a TN). En aquesta situació el model seria capaç d'identificar-ho tot de forma correcta tal i com ensenya la figura 2.

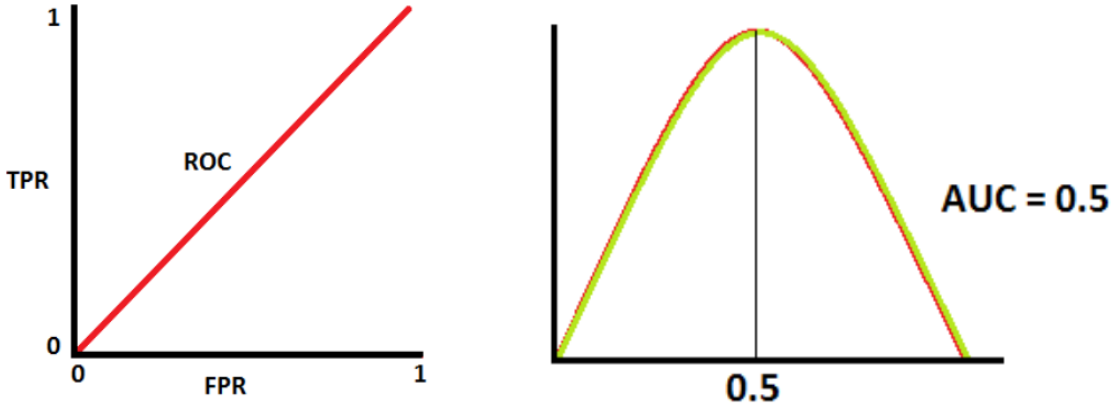
Una situació més quotidiana és quan el model és capaç de distingir les diferents classes sense arribar a la perfecció cometent una sèrie d'errors obtenint AUC inferiors a 1 i superiors a 0,5. Acaba classificant individus a un grup quan realment no pertanyen a aquest. Els següents gràfics mostren clarament la situació:



Figures 3 i 4. Gràfics corba ROC bons. (Narkhede, 2018)

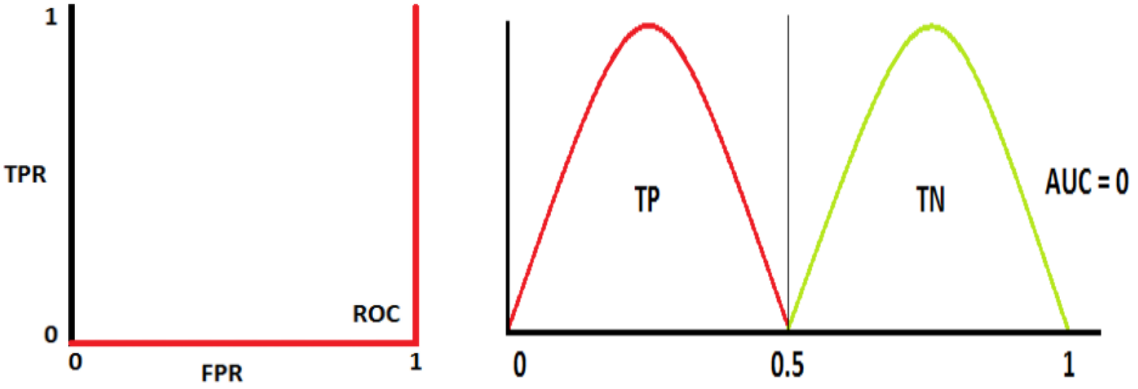
Amb el mateix exemple anterior, obtenir un AUC de 0,7 indicaria que hi hauria individus que el model prediria que complirien amb les seves obligacions, però al final no és així i acaba fallant. El model no és perfecte, però és capaç de detectar alguns patrons en les dades.

La pitjor situació que un es podria trobar és quan l'AUC és del 0,5. Això succeeix quan el model no té capacitat per a discriminar entre els prestataris que acaben complint o incomplint, és a dir, no té idea de com classificar els individus. La corba ROC resulta ser, per aquests casos, diagonal.



Figures 5 i 6. Gràfics corba ROC dolents. (Narkhede, 2018)

Per últim existeix el cas contrari al primer en que el model prediu incorrectament totes les observacions. En altres paraules, tots els que presentaven impagaments els acaba classificant com a clients sans i viceversa. Es tracta d'un cas generalment atípic que fa acte de presència quan no s'han treballat adequadament les dades. Normalment es deu a una mala codificació de la variable resposta. Com que totes les prediccions acaben essent errònies l'AUC que s'obté és de 0.



Figures 7 i 8. Gràfics corba ROC contraris. (Narkhede, 2018)

Tornant a l'article, aquest compara les millores obtingudes en termes d'AUC respecte un dels models tradicionals (el logístic) amb la resta de models de classificació. La comparació la realitzen tenint en compte altres resultats obtinguts per altres investigadors també enfocats en l'àrea de la modelització del risc creditici.

Per mostrar les millores percentuals de l'AUC respecte el lògit, s'efectua un gràfic on l'eix x es veu representat per les diferents metodologies de ML ordenades de menor a major complexitat algorítmica i, posteriorment, l'eix de coordenades en que apareix el percentatge de millora de l'AUC.

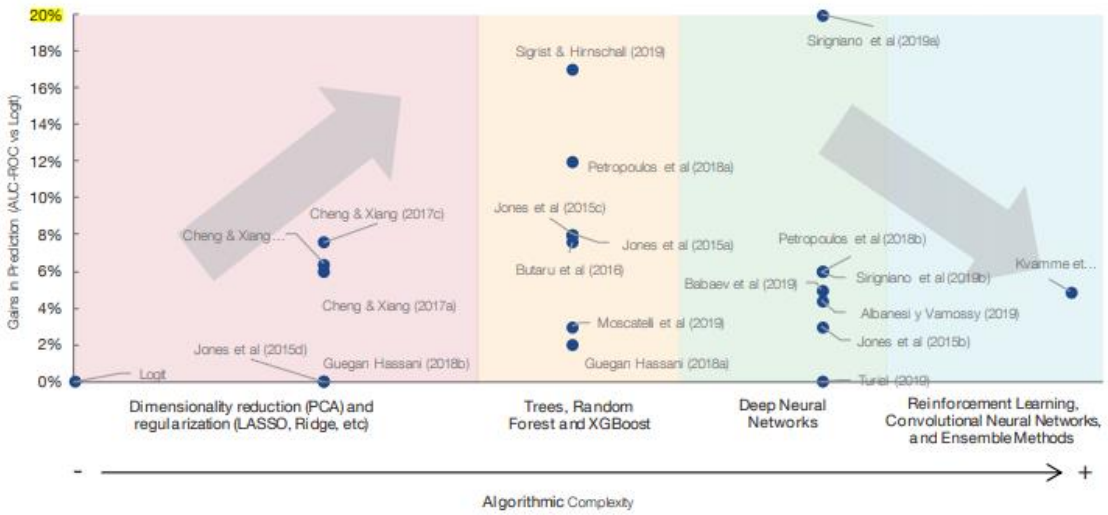


Figura 10. Millores percentuals dels diferents models respecte el Logit. (Alonso i Carbó, 2020)

Els models de Machine Learning són popularment coneguts per aconseguir, generalment, una molt bona capacitat predictiva del problema. Aquesta imatge permet afirmar aquesta creença. D'altra banda, també es pot extreure que a mesura que s'utilitzen eines de ML més avançades, com podrien ser el *random forest* o les *xarxes neuronals,* milloren la capacitat de predicció respecte els models estadístics més simples. Malgrat això, les millores que s'obtenen són heterogènies arribant fins a una millora màxima percentual en l'AUC del 20% i trencant la tendència de millora quan la complexitat és ja molt elevada. Seria necessari realitzar més estudis utilitzant algoritmes d'aprenentatge per reforç o xarxes neuronals convolucionals per acabar d'afirmar-ho.

4.2.2 Índex de gini

Una mesura molt utilitzada en el món economètric per a quantificar les desigualtats d'una distribució és el coeficient de gini. Pot prendre valors compresos entre el 0 i l'1, on el 0 fa referència a la perfecta igualtat i quan l'índex és 1 es parlaria d'una situació perfectament desigual. Existeix la possibilitat d'expressar aquests valors en forma de percentatge tenint així registres entre el 0 i el 100. Les aplicacions més comunes que es poden trobar són en la mesura de les diferències salarials en un país o entre gèneres.

El coeficient de gini es calcula, teòricament, a partir del diagrama de la corba de Lorenz. El diagrama representa gràficament la distribució relativa d'una variable en una població. Cadascun dels punts de la corba descriu un percentatge acumulatiu dels valors del domini. D'aquesta manera, el punt (0,0) representaria el 0% mentre

que el (100,100) representaria el 100% del gràfic. Posteriorment s'estableix una recta entre aquests dos ponts que correspondria a una distribució perfecta o, dita d'una altra forma, igualitària.

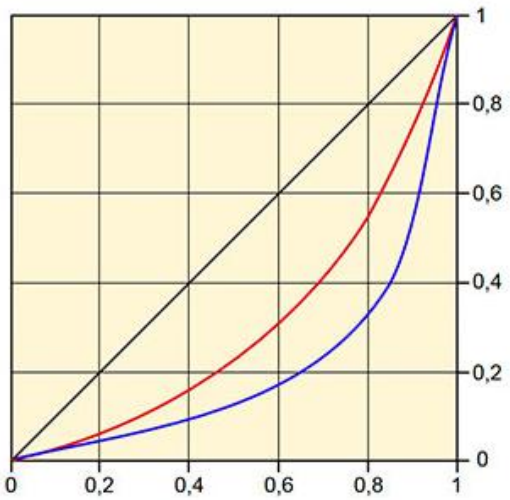


Figura 10. Corba de Lorenz. (Peña, 2019)

Amb aquesta línia com a referència, es pot determinar el grau de llunyania que es mostra la corba de Lorenz respecte la diagonal. Com més s'apropi a la recta més igualitària serà la distribució i viceversa. Aquesta corba té dues propietats importants: és sempre positiva i convexa. La figura de la pàgina 34 es mostra de color negre la recta, i en vermell o blau possibles corbes de Lorenz.

Suposa que s'estan analitzant les desigualtats en els salaris anuals de dos països G i H i es decideix fer els gràfics de Lorenz per ambdós casos. El primer es veu representat per la corba vermella i el segon per la blava. Com que la corba blava és la més allunyada respecte la diagonal es dirà que els salaris de la població del país H són més desiguals que els del país G. Quan es parla de desigualtat, en aquest cas, es fa referència a que existeix un petit percentatge de la població que acumula més riquesa. En el supòsit que fos igualitari, tothom tindria el mateix salari.

La fórmula matemàtica del coeficient de gini deriva del propi diagrama de Lorenz. Són necessàries l'àrea entre la corba i la recta, que anomenarem com a A i l'àrea per sota de la corba de Lorenz, que serà B. D'aquesta manera es pot definir gini com a resultat de l'operació matemàtica següent:

$$Índex\ de\ gini = \frac{A}{A+B}$$

On A i B poden veure's en la figura que es mostra a continuació:

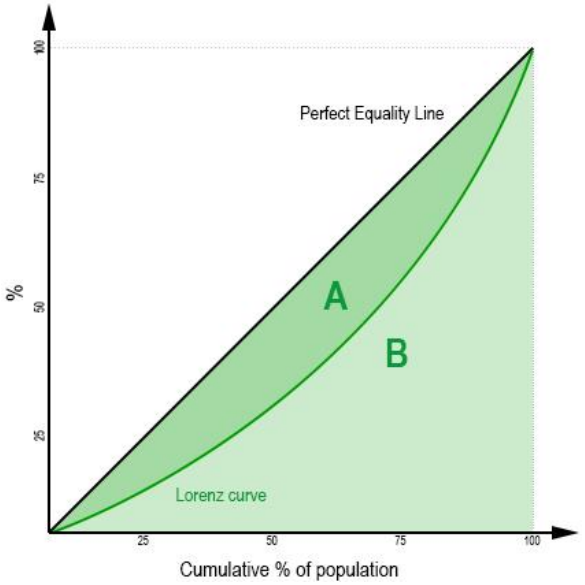


Figura 11. Àrees necessàries per al càlcul del coeficient de gini. (Zeder, 2018)

Tot i que es pot calcular gini fent el quocient de les àrees A i B, no és la metodologia més comunament utilitzada. L'altre opció és construir una taula que contingui informació sobre el problema en concret. Una taula que hauria de tractar les distribucions de la variable d'interès, per exemple, si s'està analitzant la distribució dels salaris d'una empresa es podria construir una taula que contingui els diferents salaris que es cobrin, quantes persones el cobren i a partir d'aquí tota la informació probabilística que es pugui obtenir.

Salari	Treballadors	Treballadors Acumulats	Proporció (p_i)	Salari x Treballadors	Salari total Acumulat	Proporció (q_i)	$p_i - q_i$
1.000€	100	100	0,2564	100.000€	100.000€	0,1504	0,106
1.500€	200	300	0,7692	300.000€	400.000€	0,6015	0,1677
2.000€	50	350	0,8974	100.000€	500.000€	0,7519	0,1455
3.000€	25	375	0,9615	75.000€	575.000€	0,8647	0,0968
5.000€	10	385	0,9872	50.000€	625.000€	0,9399	0,0473
8.000€	5	390	1	40.000€	665.000€	1	0
SUMA	390		3,8717	665.000€			0,5633

Taula 3. Distribució dels salaris dels treballadors d'una empresa desconeguda.

En aquests casos, gini es podria calcular de la següent forma:

$$Índex\ de\ gini = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i}$$

On ara la p_i fa referència a la proporció acumulada de la població, és a dir, quant representen els treballadors d'un salari concret de forma acumulada, mentre que la q_i és la proporció acumulada dels ingressos. Així doncs, el coeficient de gini per a una empresa amb les característiques salarials de la taula 3 seria de 0,1455 indicant que és bastant igualitària en termes de salari per la seva proximitat al 0.

$$Índex\ de\ gini = \frac{0{,}5633}{3{,}8717} = 0{,}1455$$

Aquest coeficient té certa aplicabilitat en el risc creditici. Serà d'interès poder quantificar les desigualtats de les probabilitats estimades pels models segons si l'individu havia incomplit els pagaments o no. En qualsevol cas, es buscaran models que facin augmentar l'índex perquè, com més gran sigui, més desigualtats hi haurà en les estimacions de les probabilitats i per tant, el model identificaria millor les dues classes d'interès.

4.3 Factors de cost

Des d'un punt de vista conservador, és necessari establir unes bases per a poder controlar què és el que passa quan s'està modelitzant. Aquestes bases poden tenir en compte un gran número de criteris necessaris a tenir presents a l'hora de veure si els processos que s'estan fent són els millors que hi ha o bé, existeix algun aspecte en que el model falla. Pel motiu anterior és necessari establir un sistema que permeti quantificar aquests problemes.

El sistema que s'utilitzarà estarà basat en les validacions dels sistemes IRB (*Internal Ratings-Based*) per a identificar i classificar tots els factors que poden acabar essent un cost per als desenvolupadors. Els sistemes IRB són models estadístics per avaluar internament el perfil de risc, que han de complir una sèrie de requisits previs definits per un expert. Es tracta d'un dels més utilitzats i especialment per a l'estimació de les probabilitats d'impagaments.

Les institucions creditícies són responsables en l'avaluació dels resultats dels sistemes IRB. Quan s'utilitza metodologia l'objectiu principal és acabar estimant el paràmetre d'interès, que podria ser la PD com també les altres components del risc

com la LGD i la EAD. Un cop el disseny del model ha estat aprovat i compleixi amb els requisits que demana el supervisor, serà acceptat i podrà ser utilitzat per a la predicció del capital regulatori, per exemple. La part de la validació és principalment quantitativa, tot i que existeixen altres factors que recull el sistema IRB com podrien ser la privacitat de les dades i la seva qualitat, la dificultat per a resoldre problemes o la governança o procés per a controlar accessos, activitat dels models.

D'aquesta manera és important estudiar la compatibilitat dels models de Machine Learning amb el sistema de validació IRB. Així es poden establir costos potencials per al supervisor. Es definiran cinc grans grups de costos que permetran analitzar els inconvenients de la nova metodologia:

- L'estadística.
- La tecnologia.
- La conducta del mercat.
- La interpretabilitat.
- Els biaixos.

Els factors estadístics que poden afectar en la creació del model i que estan fortament lligats a l'ús dels algoritmes de ML són la presència d'hiperparàmetres, la necessitat de processar les dades (*feature engineering*) o la complexitat de comprovar el comportament d'aquest amb altres dades de testatge. Un altre problema que pateixen aquests algoritmes és que existeix la possibilitat de sobreestimar les dades ja que ofereixen molta flexibilitat i si no es controla es poden arribar a situacions no desitjades i que el model no estimi adequadament.

Una capacitat tecnològica suficient és necessària per al desenvolupament de models ja que s'han d'implementar, pujar i mantenir-los en producció per a la seva operació. Una variable que pot explicar l'anterior és el temps computacional que triga el model en executar i el seu impacte mediambiental, especialment a partir de les petjades de carboni que tenen en compte el consum elèctric. També existeixen altres factors derivats de la tecnologia com podrien ser la dependència dels models als servidors externs o, fins i tot, l'existència al risc cibernètic.

Igual d'importants és la qualitat de les dades amb les quals es treballa i tot el que envolta a la privacitat. Moltes de les institucions financeres utilitzen, principalment, les seves pròpies dades històriques amb l'objectiu de desenvolupar models

predictius. Es poden incloure altres aspectes importants com la transparència de tot el procés de modelització o la capacitat de l'equip o dels analistes per a poder replicar els models, coneguda amb el nom d'*auditability*.

Tenir la capacitat d'interpretar el model i els resultats és la tasca més important per a poder explicar al client les raons per les quals se l'ha acceptat o denegat el crèdit. Igual d'important és no deixar aquesta decisió en mans d'un model; ha d'existir una persona humana que jutgi i validi les estimacions dels factors de risc. Addicionalment, la Comissió Europea va establir un comunicat en que els resultats dels models de ML han de ser interpretables per a totes les persones que participen en el procés, tant equip tècnic de l'entitat com el propi client, ja que la decisió de donar o no el préstec pot afectar significativament l'impacte econòmic de la persona. Per tant, el banc ha de ser capaç d'explicar quins han estat els elements de risc, variables o qualsevol altre aspecte que han determinat la presa de decisió. Aquí pot entrar en joc una branca dels ML que permet identificar aquelles característiques que han definit la resolució. És el que s'anomena com a Interpretative Machine Learning, amb sigles IML, branca del ML en la qual s'estan produint avenços importants en els últims anys.

L'últim factor a incloure i per això no menys important, és el biaix en el sentit que s'han de complir els principis d'igualtat: individus i grups lliures d'injustícies, discriminacions o estigmatitzacions.

Existeixen múltiples fonts potencials de biaixos. Una d'elles podria ser treballar amb mostres esbiaixades, que sorgeixen quan l'algoritme de ML s'entrena a partir de dades històriques amb soroll. Llavors les decisions automàtiques que es prendrien augmentarien les dimensions d'aquests biaixos. En segon lloc les decisions no poden estar basades en les característiques personals com la raça o la ètnia exceptuant aquelles situacions en les que s'han consentit l'ús d'aquestes per a la predicció o per a altres fins. Finalment podria existir també el biaix algorítmic. Fa acte de presència quan el model utilitza únicament un petit conjunt de variables per a realitzar la predicció o classificació.

Es pot construir una taula resum amb el conjunt de factors de costos més importants:

Grups	Factors de cost
Estadístics	Estabilitat, sobreestimació i hiperparàmetres
Tecnològics	Transparència, petjades de carboni, dependència d'altres servidors i risc cibernètic
Conducta	Privacitat i replicació dels models
Interpretabilitat	Interpretabilitat
Biaixos	Biaix

Taula 4. Factors de risc segons el seu grup de referència.

Al final la relació entre costos i beneficis dependrà de la finalitat marcada. Serà diferent si el propòsit és construir un model per a l'estimació del capital regulatori on la precisió serà fonamental o si bé l'objectiu és estimar les probabilitats d'impagament perquè, en aquest cas, la capacitat de classificació serà més important.

4.4 Funció de cost

Amb aquests factors anteriors es pot construir una mesura del cost de supervisió per cadascun dels models que es vulguin analitzar en funció de la tolerància al risc establerta i el propòsit de la construcció del model. Per dur a terme aquesta mesura, es planteja un procés de dues etapes. En la primera etapa es realitzarà una avaluació estructural quantificant les dificultats tècniques intrínseques per a cada model basada en els factors anteriors. És el que s'anomena com a algoritme de caixa negra (*black-box algorithm*) La taula posterior mostra un exemple:

		Lasso	Tree	Random Forest	XGBoost	Deep Learning	RL & Ensamble Methods
Statistics	Stability	1.0	3.0	2.0	2.0	4.0	4.0
	Nº (Hyper) parameters	1.0	2.0	3.0	4.0	5.0	6.0
	Over-fitting	1.0	3.0	2.0	3.0	5.0	5.0
	Feature engineering	1.0	—	—	—	3.0	3.0
	Dynamic calibration	—	—	—		—	1.0
Technology	Transparency	1.0	1.0	1.0	1.0	1.0	2.0
	Carbon Footprint	1.0	1.0	3.0	2.0	5.0	6.0
	Third-party providers dependencies	1.0	1.0	1.0	1.0	3.0	4.0
	Cyber-attacks	1.0	1.0	1.0	1.0	2.0	2.0
Conduct	Privacy	1.0	1.0	3.0	3.0	3.0	3.0
	Auditability	1.0	1.0	3.0	4.0	5.0	6.0
	Interpretability	1.0	1.0	2.0	2.0	3.0	4.0
	Biases	1.0	3.0	4.0	4.0	5.0	5.0

Taula 5. Algoritme de la caixa negra dels models. (Alonso i Carbó, 2020)

Aquesta quantificació pot deixar-se fixe perquè és en la fase 2 quan el supervisor pot modificar segons li convingui. S'afegeix un paràmetre nou, uns pesos entre el 0% i el 100%, que permeten donar cert pes a cadascún dels factors de cost. Finalment, amb la suma del producte entre el pes i la quantificació definida en la fase 1 es trobarà la mètrica conjunta del cost de supervisió per cada model. Es pot definir aquest cost supervisat per cada model *i-èssim*, per cada factor *f*, els pesos *W* i el valor establert per als factor de risc *X*:

$$CdS_i = \sum_{f}^{F} W_f \times X_f$$

El supervisor pot modificar els pesos d'acord amb les seves preferències per tal de complir amb la regulació. Per a fer-ho i per supervisar, conjuntament amb la fase 1, es necessiten experts amb alts coneixements estadístics, tecnològics i financers. A continuació es mostra la taula 5 modificada amb l'addició d'uns pesos *W* d'exemple:

		Weigth=f (model use)	Lasso	Tree	Random Forest	XGBoost	Deep Learning	RL & Ensamble Methods
Statistics	Stability	10.0%	1.0	3.0	2.0	2.0	4.0	4.0
	Nº (Hyper) parameters	10.0%	1.0	2.0	3.0	4.0	5.0	6.0
	Over-fitting	10.0%	1.0	3.0	2.0	3.0	5.0	5.0
	Feature engineering	10.0%	1.0	—	—	—	3.0	3.0
	Dynamic calibration	10.0%	—	—	—		—	1.0
Technology	Transparency	5.0%	1.0	1.0	1.0	1.0	1.0	2.0
	Carbon Footprint	5.0%	1.0	1.0	3.0	2.0	5.0	6.0
	Third-party providers dependencies	10.0%	1.0	1.0	1.0	1.0	3.0	4.0
	Cyber-attacks	10.0%	1.0	1.0	1.0	1.0	2.0	2.0
Conduct	Privacy	0.0%	1.0	1.0	3.0	3.0	3.0	3.0
	Auditability	10.0%	1.0	1.0	3.0	4.0	5.0	6.0
	Interpretability	10.0%	1.0	1.0	2.0	2.0	3.0	4.0
	Biases	0.0%	1.0	3.0	4.0	4.0	5.0	5.0
	Supervisory cost of the model	100%	0.90	1.30	1.60	1.85	3.30	3.90

Taula 6. Algoritme de la caixa negra dels models. (Alonso i Carbó, 2020)

Segons aquesta taula, i que s'utilitzarà parcialment en aquest projecte, els models d'aprenentatge reforçat i la metodologia *ensemble* són els que presenten major cost de supervisió mentre que el lasso presenta el valor més petit.

No s'utilitzaran tots els models que apareixen en la taula 5. Es veuran els models logístics amb i sense penalització (lògit, lasso i Ridge) essent els tradicionals i per una altra part els moderns o innovadors de Machine Learning (arbres de decisió, random forest, support vector Machine i el XGBoost). S'analitzarà quin d'ells acaba proporcionant millors resultats a partir de l'anàlisi dels beneficis – costs.

5. EXPOSICIÓ DEL CAS PRÀCTIC

Després de donar èmfasi en la teoria essencial i al plantejament elaborat pels analistes Andrés Alonso i José Manuel Carbó sobre els beneficis i costos en l'aplicació de les noves metodologies de Machine Learning en la modelització de les probabilitats d'impagament, és moment de recapitular momentàniament als objectius marcats a l'inici del projecte. Primer, poder decidir quin model dels que es construiran funciona millor com també quantificar l'eficàcia i la utilitat dels models revolucionaris de *Data Mining* respecte els mètodes clàssics emprats per les institucions financeres.

Trobar el model no és tasca fàcil. S'han de tenir coneixements avançats en programació i estadística per a que l'etapa de modelització sigui un èxit. Poden sorgir impediments, problemes o errors que dificultin aquest procés i sense ajut o experiència serà bastant improbable que el procediment surti bé.

Per donar resposta al problema en qüestió i, per a obtenir resultats fiables, s'ha procurat ésser realista a partir de l'ús d'una base de dades que podria tenir un banc o institució creditícia qualsevol amb les dades històriques, financeres i personals de clients en un període de temps concret.

Posteriorment, s'elaborarà una descriptiva d'algunes de les variables més considerades més representatives. Això ens donarà una idea de com són les dades. Seguidament s'analitzaran un per un cadascun dels models emprats a partir de la qualitat de les precisions i anàlisi dels beneficis i costos per acabar escollint el millor model. Finalment, s'aplicaran eines d'Interpretative Machine Learning per a analitzar els factors amb major pes per a la classificació dels clients.

5.1 Base de dades

La informació que s'ha fet servir per analitzar els models per a la predicció de les probabilitats d'impagament prové de *Home Credit* (HC) que és una institució financera no bancària internacional fundada l'any 1997 a la República Txeca. Opera en 10 països i les seves tasques es basen en proveir préstecs a particulars amb un registre històric creditici petit o nul d'una forma segura i positiva per al client. Més de 135 milions de persones han rebut algun crèdit per part d'aquesta entitat.

És difícil trobar bases de dades públiques relacionades amb el risc creditici perquè normalment s'inclou informació personal i privada dels clients, tant empreses com

particulars, i aquestes dades podrien afectar a moltes persones. *Home Credit* va proporcionar l'any 2018 una base de dades molt extensa amb la informació dels préstecs dels seus clients amb l'objectiu d'incentivar als experts i interessats en la modelització per intentar construir el millor model que expliqui el comportament de les dades dels seus clients, és a dir, trobar l'algoritme que faci les millors prediccions. Els participants tenien un marge de tres mesos per a presentar els seus resultats. A canvi de l'ajuda, l'entitat financera va oferir tres premis amb un valor total de 70.000 dòlars per aquells tres millors models.

La base de dades que han fet pública conté un total de deu arxius. Cadascun d'ells emmagatzema un gran ventall d'informació històrica que és de molt interès analitzar-la. Per tenir en compte totes les variables d'interès és important agrupar-les d'alguna manera en un únic fitxer i, posteriorment, dividir les observacions en el grup d'entrenament del model i un altre de testeig per comprovar que el model és adequat. Aquest procés s'ha fet mitjançant l'ús del llenguatge SQL que permetrà incloure modificacions, transformacions i filtratges a les variables quan sigui d'interès.

D'aquests deu fitxers, n'hi ha dos que fan referència a les dades del préstec i del client en estudi. Aquests dos són els més importants pel seu contingut. Un d'ells (*train*) inclou la variable resposta de si el client va incomplir aquell crèdit o no, mentre que l'altre (*test*) no la inclou. L'estructura d'aquests dos arxius és així perquè els aspirants havien de crear un model a partir de les dades de *train* i, un cop establert el model final, fer la predicció amb *test*. Segurament, els creadors van pensar en que era millor no incorporar la variable dels impagaments en la de test per a que els participants fossin incapaços de crear un model ajustat únicament a les dades de *test*. El fitxer *train* és el que es farà servir en el projecte i test, en canvi, no s'utilitzarà perquè no es coneix la variable resposta d'interès.

Hi ha uns altres dos fitxers addicionals que són merament informatius en els quals es dona una explicació i definició exhaustiva de cadascuna de les variables que apareix en cada fitxer i, per una altra part, també queda detallat el funcionament de la competició.

Una vegada eliminats els que no es faran servir, queden comptabilitzats set fitxers de dades. Dins d'aquests es poden classificar segons la informació que ofereixen en tres tipologies diferents: actual, interna i externa. El primer tipus fa referència a les dades de *train* esmentades anteriorment. Les internes són aquelles que proporcionen informació històrica dels préstecs anteriors demanats a *Home Credit* pel client. Per

últim, les externes serien exactament el mateix que les internes exceptuant que les dades històriques de diferents préstecs són d'altres bancs. És bastant comú que una entitat bancària disposi de l'historial dels crèdits demanats en altres entitats del client.

És hora de posar nom als diferents fitxers. "*Application_train*" és la taula més important ja que disposa, com s'ha dit, de la variable resposta de si va incomplir o no amb els pagaments. Cada fila representa un préstec. Inclou, a més a més, informació del crèdit com podria ser la quantitat demanada, però també apareixen aspectes personals del propi client.

Respecte al grup intern, es poden trobar totes les aplicacions prèvies realitzades pel client per a aconseguir un crèdit a l'entitat amb el nom de "*previous_application*". Posteriorment hi ha la taula dita "*POS_CASH_balance*" que mostra l'estat mensual dels préstecs donats anteriorment per *Home Credit* on molts d'ells segueixen actius. També hi ha informació sobre cada pagament i impagament efectuat pel client a "*installments_payments*". D'aquesta manera és aquí on es guarden les transferències monetàries. Per últim, hi ha "*credit_card_balance*" que mostra l'historial mensual de les targetes de crèdit que tenen els sol·licitants a l'entitat.

Només hi ha dues taules que tracten del grup extern. Reben el nom de "*bureau*" i "*bureau_balance*" i detallen tots els crèdits atorgats per altres entitats financeres com també l'estat mensual d'aquests respectivament. A la taula següent queda la tipologia de dades que tracta cada fitxer:

Tipologia d'informació	**Fitxers**
Actual	"application_train"
Interna	"previous_application", "POS_CASH_balance", "installments_payments", "credit_card_balance"
Externa	"bureau", "bureau_balance"

Taula 5. Tipus d'informació que tracta cadascuna de les taules.

La volumetria dels fitxers amb la que es treballa és molt gran, trobant múltiples arxius amb més de 10 milions de registres. I la majoria d'ells tenen també un nombre elevat d'atributs, amb una suma total de 218 columnes, aspectes pels quals indiquen que s'està treballant amb una base de dades complexa. Per a evitar aquesta complexitat, s'agafaran únicament els factors que es creguin, a priori, que puguin

explicar d'alguna manera el comportament dels impagaments produïts a *Home Credit*.

Es pot utilitzar el llenguatge SQL com a mètode d'unió entre les taules perquè estan connectades ja que existeixen variables identificadores per a cada préstec. Hi ha la variable identificadora del préstec de *train* que permet la relació entre la majoria d'elles. També hi ha altres ID com podria ser la que connecta els préstecs externs de *Home Credit* amb el seu estat mensual. En la figura que es mostra a continuació s'observen les relacions existents i també el nombre de files i variables:

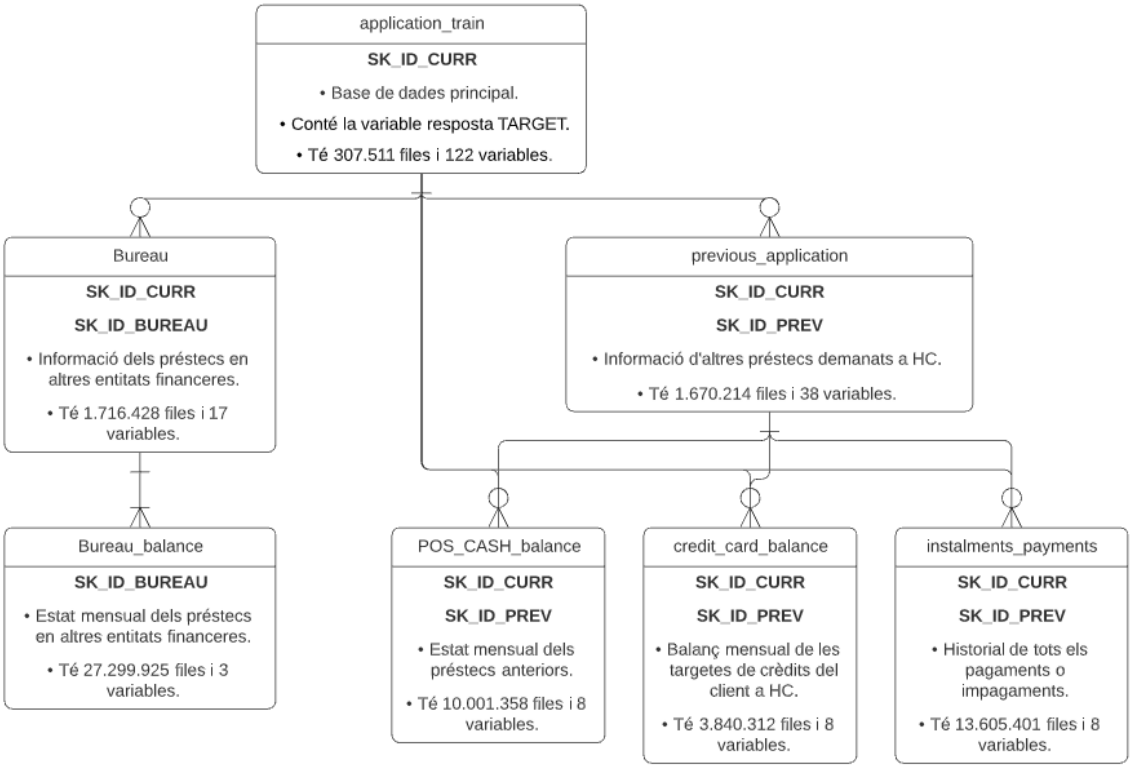


Figura 12. Diagrama relacional de la base de dades. En negreta es mostren els identificadors que permeten relacionar-se amb la resta de taules.

El procés de la construcció dels fitxers de dades finals d'entrenament i testatge ha consistit en tres fases. La primera, consisteix en agafar les variables d'interès de cadascuna de les taules i guardar-les en *dataframes* diferents per identificar-ne ràpidament el seu origen. Per a facilitar la feina, els nous *dataframes* han estat creats tenint en compte els préstecs d'"*application_train*" (on hi ha la variable resposta), és a dir, la informació que hi ha en una fila dels diferents conjunts de dades fa referència al crèdit de la mateixa línia. En aquesta fase s'ha aplicat un preprocessament per a les dades mancants. A l'unir les taules, a partir dels *joins* agrupats pels identificadors corresponents, sorgien una quantitat considerable de NA's. Això era degut, en gran part, perquè el client i-èssim no estava en una de les

taules provocant així la falta d'informació de la variable en concret per aquest individu. En la majoria de casos es pot solucionar incorporant un 0 en els llocs on faltava la dada. Per exemple, *IS_NEW_CLIENT*, factor que informa si el client és nou a *Home Credit,* ha estat creada a partir del nivell *New* de la variable *NAME_CLIENT_TYPE* expressat com *NAME_CLIENT_TYPE = "New"*. Les observacions amb aquest nivell eren codificades com a 1 (sí presenta la característica) mentre que la resta, clients ja registrats, eren codificades automàticament com a NA's. Per a evitar-ho, s'han transformat els NA's per zeros. També s'han modificat algunes categories de múltiples variables perquè s'ha cregut convenient efectuant canvis de noms, unions, etc. En una segona fase s'han incorporat totes les variables distribuïdes en els diferents *dataframes* en un únic conjunt de dades. Posteriorment es troba la fase final: dividir les dades de forma completament aleatòria en entrenament i test. S'ha considerat que el primer contingui el 80% de les dades totals (246.008 registres) i el 20% restant li correspon a test (61.503 registres). En paral·lel, el nombre de factors en cadascuna de les dues taules és de 88.

Considerant el funcionament de les llibreries que s'han utilitzat per a la realització dels models de ML al Python, s'ha necessitat aplicar certs canvis en els factors qualitatius: la creació de variables fictícies. Aquestes, també anomenades variables *dummies*, són atributs que prenen 0 o 1 en funció de si presenten la característica d'una variable categòrica en concret i se'n crearan tants *dummies* com nivells tingui aquesta variable. Així doncs, si existeix un factor (una sola columna) amb tres nivells diferents A,B i C es passaria a tenir tres noves columnes i cadascuna d'elles codificades amb 0 i 1. Si aquest concepte es traspassa a les dades de *train* i test s'obtenen un total de 186 columnes. Finalment, s'emmagatzemen un total de tres fitxers: les dades d'entrenament i de testatge i un altre fitxer amb totes les dades sense variables *dummies*. Aquest últim és el que servirà per a entendre com són les dades a partir d'un curós anàlisi descriptiu.

De les 218 columnes o factors totals que existeixen tenint en compte totes les taules, s'utilitzaran 88. La reducció considerable en el nombre de variables d'interès és deguda a que l'objectiu principal no és construir el millor model per a les dades de Home Credit, sinó que el es vol és veure els diferents comportaments dels models i identificar quins podrien ser beneficiosos per a l'estimació del risc. Alhora, la disminució de variables permetrà agilitzar els processos. Les variables seleccionades formen part, en la gran majoria, de factors que podrien utilitzar perfectament els bancs en aquestes situacions per a l'estimació de les PD. Les dades finals aconseguides segueixen presentant unes dimensions molt grans i això fa difícil

explicar les variables una per una. Per aquest motiu es llista, seguidament, d'un conjunt reduït de variables que poden ésser de gran interès.

Variable	Descripció
TARGET	Variable resposta: el client va incomplir o no.
AMT_INCOME_TOTAL	Salari total del client.
AMT_CREDIT	Crèdit que ha demanat.
AMT_ANNUITY	Anualitat del crèdit.
MEAN_EXT_SOURCE	Mitjana de la valoració externa del client.
AMT_ANNUITY_TOT	Suma de les anualitats dels crèdits actius.
PERCENTATGE_ANNUITY	Divisió entre l'anualitat total i el salari.
RATI_DEUTE_GARANTIA	Divisió entre el crèdit i la garantia.
NUM_ACTIVE_CREDITS	Nombre de crèdits actius del client.
NUM_CREDITS_PREVIS_TANCATS	Nombre de crèdits finalitzats del client.
DID_OVERDUE_Altres_EF	Va tenir algun pagament vençut.
DID_PROLONG_Altres_EF	Si va allargar algun pagament.
AGE_EXPECTED	Edat del client.
OCCUPATION_TYPE	Professió del client.
ORGANIZATION_TYPE	Sector de l'empresa on treballa el client.

Taula 6. Definició d'algunes de les variables més importants considerades a priori.

El llistat complet de variables que s'utilitzaran en la modelització es troba a l'apartat 8.1 de l'annex (pàgina 113). S'inclou, addicionalment, d'informació que pot ser d'interès per al lector, com per exemple, una definició clara i breu del que tracta cadascuna d'elles, si l'atribut és numèric o categòric, els diferents nivells o categories si escau i de quina taula provenia en l'origen. També apareixen aquelles variables que s'han creat a partir d'unes altres.

A grans trets, les variables que existeixen tant a entrenament com a *test*, poden classificar-se en tres grups diferents:

- Variables relacionades amb el capital monetari del client i les seves afectacions amb el préstec. El salari o l'anualitat corresponent al crèdit formen part d'aquesta tipologia.
- Variables que expliquen com ha estat l'historial del client independentment de quina entitat financera ha estat. El nombre de crèdits actius o els tancats serien bons exemples.
- Variables que defineixen les característiques socials del client. S'estaria parlant, per exemple, del nivell educatiu, tipus d'habitatge en el que viu, amb qui viu o de què treballa.

En general, es tracta d'informació que un banc podria disposar al seu servei perfectament. Malgrat això, s'han inclòs també altres factors que poden descriure's com a discriminatius, especialment de caràcter personal com podria ser el gènere del sol·licitant, una característica que no s'utilitza en la modelització del risc creditici en els bancs espanyols perquè podria operar en contra de la persona.

Un cop vists tot el procediment emprat per a la construcció de les taules definitives i la definició d'alguna de les variables principals, és important remarcar que la fase d'aconseguir les dades és de les més importants de qualsevol estudi. Unes dades sense estar treballades, sense control pot comportar a l'obtenció de resultats inesperats i fins i tot dolents. Això no succeeix únicament en el sector del risc de crèdit, sinó que tothom que treballa amb bases de dades de qualsevol àmbit ha de ser conscient dels perills que comporta no dedicar-li temps a les dades. Tot i això, és necessari conèixer sobre el tema del qual s'està tractant.

El codi SQL que correspon al tractament de les taules i de la construcció de les dades definitives es mostra a l'annex al punt 8.5.1 (pàgina 151).

5.2 Anàlisi descriptiu avançat

Qualsevol projecte que tracta amb dades, independentment de la seva volumetria, necessita un estudi i anàlisi de cadascuna de les variables que presenten les taules. Estudiar-les permet obtenir una idea general de com és la base de dades, però fonamentalment s'utilitza per a identificar tendències, problemes o anomalies. En el supòsit que s'haguessin detectat errors en aquesta etapa del projecte, s'estaria encara a temps per a ajustar les dades per a solucionar els problemes. Si l'error s'hagués detectat en les fases finals del projecte comportaria a refer-ho tot des del principi. D'aquí la importància d'estudiar l'entorn en el que es treballa.

Idealment s'ha d'anar variable per variable. Malauradament, el nombre d'atributs que es tracten és molt elevat i si s'anés una per una s'estaria dedicant molt espai a un aspecte que no es troba en els objectius del treball. Tot i això, al tractar-se d'una etapa important en la verificació de les dades, es centraran únicament en algunes variables que s'han cregut imprescindibles per seguir i entendre les fases posteriors.

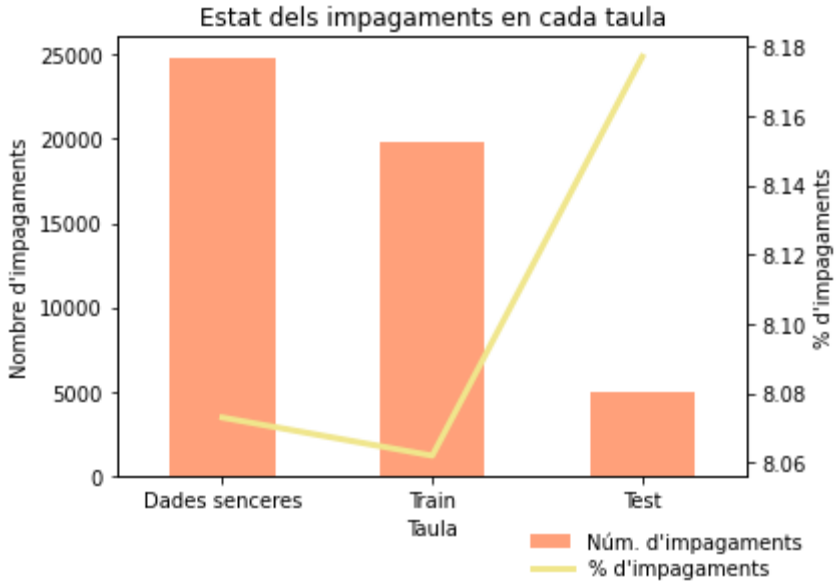
Una de les primeres tasques imprescindibles és comprovar l'existència de dades faltants o també anomenades *missings*, és a dir, observacions que els hi falta informació sobre una o varis factors. En cas afirmatiu s'haurà de decidir què fer amb aquestes observacions perquè podrien comportar múltiples problemes o en errors computacionals a l'hora de la modelització. En la taula següent es mostra la situació d'aquests en la base de dades amb totes les observacions.

Variable	Nombre de dades faltants	Representació sobre el total
OCCUPATION_TYPE	41.034	13,35%
NAME_TYPE_SUITE	1.292	0,42%
AMT_GOODS_PRICE	278	0,09%
RATI_DEUTE_GARANTIA	278	0,09%
MEAN_EXT_SOURCE	172	0,06%
MAX_EXT_SOURCE	172	0,06%
MIN_EXT_SOURCE	172	0,06%
AMT_ANNUITY_TOT	12	0,004%
PERCENTATGE_ANNUITY	12	0,004%

Taula 6. Comptatge i percentatge dels missings en la taula completa

La presència de dades mancants és, en general, molt poc notòria. Únicament és *OCCUPATION_TYPE* la que presenta un nombre de *missings* molt diferenciat respecte a la resta. Tot i que pugui suposar un problema no ho serà. Al tractar-se d'una variable categòrica va patir un canvi estructural; es van crear tantes variables com nivells tenia la variable. A l'aplicar-ho es va considerar que si presentava la característica valdria 1 i sinó seria un 0. Com que els *missings* no és res s'ha considerat com que no tenia la característica i, per tant, totes les observacions amb dades mancants en *OCCUPATION_TYPE* valen 0 en totes les variables *dummies*.

Aquest projecte es troba immers en la creació de diferents models amb la finalitat de classificar els individus segons si l'operació acabarà en un impagament o no el màxim de bé possible. Per això, el que serà d'interès és veure les relacions que existeixen entre les variables, agafant grups nombrosos i representatius, i la resposta. D'aquesta manera es podrien determinar patrons o grups de major risc respecte d'altres que presentaran un risc menor. Primer de tot es podria examinar els casos d'incompliments, i els percentatges respecte el total, de cadascuna de les taules que s'utilitzaran en l'estudi.



Dades	Registres totals	Nombre d'impagaments	% d'impagaments
Senceres	307.511	24.825	8,073%
Train	246.008	19.804	8,062%
Test	61.503	5.021	8,177%

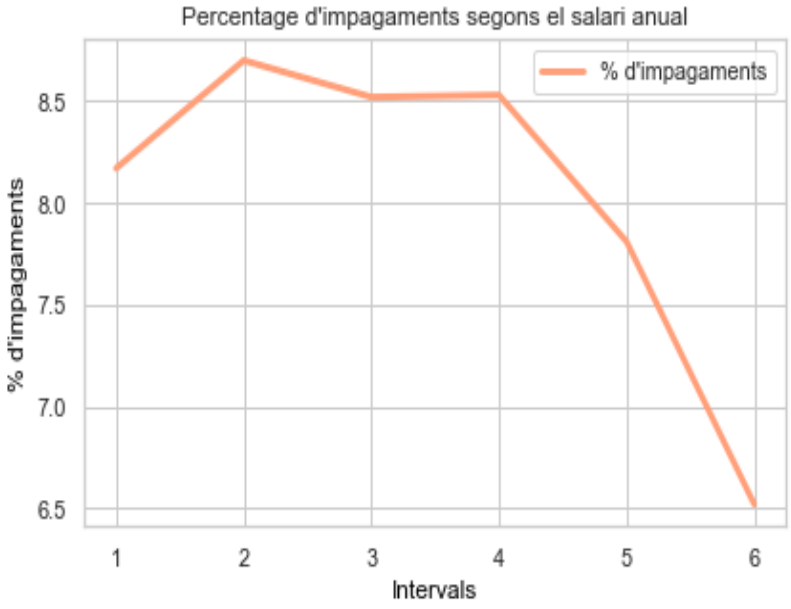
Figura 13. Gràfica de l'estat dels impagaments (TARGET) a les diferents taules
Taula 6. Nombre i percentatge d'impagaments (TARGET) a les diverses taules.

Hi ha un total de 24.825 impagaments que representen el 8,073% del total dels resultats de les operacions repartides en 19.804 a la taula d'entrenament i 5.021 a la de testatge. Com era d'esperar, els impagaments no tenen una alta freqüència en les taules. La gran majoria, aproximadament del 92%, compleix amb les seves obligacions de pagament. Si no fos així a les entitats bancàries no els hi sortiria a compte oferir aquest tipus de servei.

És també interessant veure la possible presència d'efectes en la producció d'impagaments que poden tenir els diferents valors o nivells de les variables. Es dividiran les dades en funció de si han presentat impagament o no per a poder avaluar les diferents conductes d'ambdós grups. No obstant, per a limitar l'extensió d'aquesta part, només s'estudiaran algunes de les variables del conjunt.

5.2.1 Variables numèriques

Són aquelles que es poden representar amb números i acaben permetent quantificar algun element a partir de comptatges, mesures, entre d'altres. En el conjunt de dades existeixen nombroses variables numèriques, sobretot en la part de variables centrades en la quantificació del capital i de l'historial creditici. S'analitzaran, únicament, com afecten els salaris, el rati deute garantia, el nombre de crèdits actius i els anys del vehicles dels clients d'aquells que sí tenen cotxe aplicant una discretització dels valors numèrics amb intervals per veure quins són els grups més exposats a desenvolupar un impagament. La volumetria de cada grup serà equitativa entre tots ells. Es mostrarà, primer, el gràfic resultant i seguidament una taula amb la informació dels intervals respectius començant inicialment pels salaris.



Grup	Intervals reals	% d'impagaments
1	[25.650;90.000]	8,17%
2	(90.000;117.000]	8,70%
3	(117.000;147.150]	8,52%
4	(147.150;180.000]	8,53%
5	(180.000;225.000]	7,81%
6	(225.000;117.000.000]	6,52%

Figura 14. Evolució del percentatge d'impagament segons grup de salari.
Taula 7. Intervals reals amb el corresponent percentatge d'impagament.

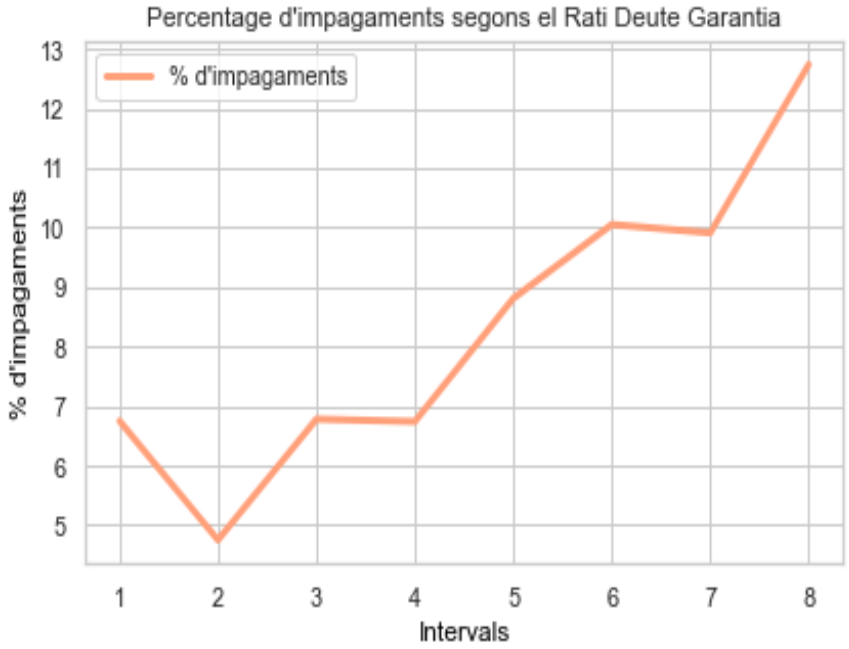
És ben conegut que la població amb pocs recursos econòmics els hi és més dificil poder afrontar les anualitats. Aquest plantejament faria pensar que el grup de risc amb salaris baixos seria més propens a la producció d'impagaments.

Paral·lelament, existeix popularment una creença que les persones amb altes rendes no tenen per què incomplir amb els seus pagaments ja que, al disposar de més capital que la resta, els hi és més fàcil efectuar els pagaments. Això podria ser cert sempre i quan els rics demanessin préstecs de la mateixa quantitat que els de classe mitjana o baixa. Una altra opció que és bastant viable és que demanin crèdits amb quantitat superior a la del seu salari anual. En aquests casos la situació podria ser molt similar a la de la classe obrera, amb salaris baixos, quan demanen una hipoteca, generalment molt més elevada que els salaris de la classe baixa o mitjana.

Els percentatges d'impagament obtinguts per a cada grup de salari semblen recolzar les creences anteriors. Els quatre primers grups presenten un percentatge d'impagament major a la mitjana del conjunt d'observacions i els dos últims, i especialment el sisè amb major riquesa salarial, tenen menys probabilitats d'incomplir. Tot i això, el primer grup presenta una menor probabilitat respecte els tres següents indicant que possiblement són més previnguts en quant als pagaments.

Una de les altres mesures molt utilitzades pels bancs és el quocient entre la quantitat de crèdit demanat i la garantia del préstec establert codificada com a *Rati Deute Garantia*. La garantia fa referència a l'acreditació monetària del particular o empresa que ofereix a l'entitat financera per fer front als pagaments. Així doncs, la garantia va vinculada al nombre d'actius que té l'empresa o sol·licitant i com més gran sigui, majors facilitats tindrà per afrontar futures despeses.

Així doncs, seria d'esperar que els ratis més petits presentessin un menor risc d'impagament i, per tant, una menor probabilitat. A continuació es mostren els resultats a partir de la generació de vuit intervals.



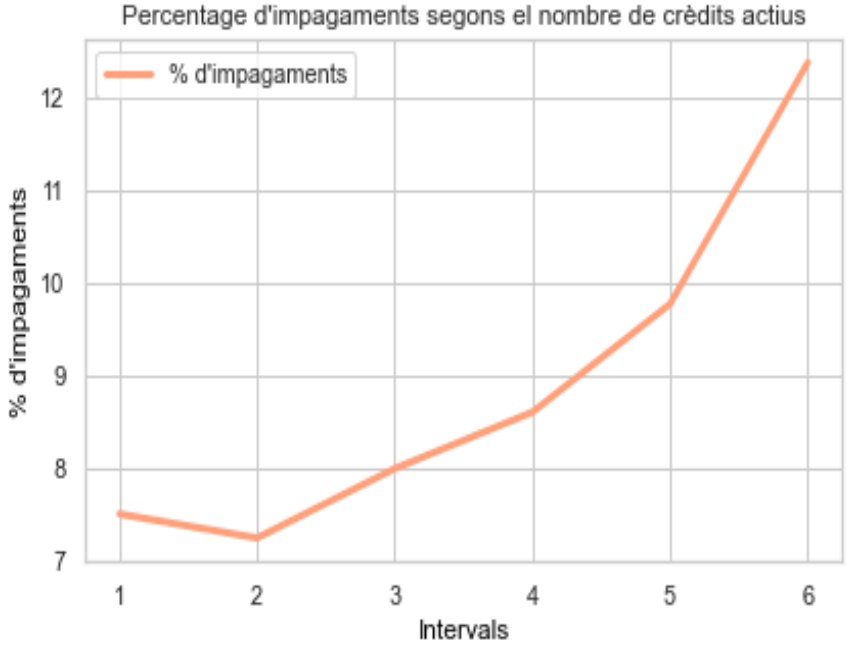
Grup	Intervals reals	% d'impagaments
1	(0,149;1]	6,75%
2	(1;1,079]	4,75%
3	(1,079;1,119]	6,78%
4	(1,119;1,145]	6,74%
5	(1,145;1,168]	8,81%
6	(1,168;1,211]	10,05%
7	(1,211;1,277]	9,91%
8	(1,277;6]	12,74%

Figura 15. Evolució del percentatge d'impagaments segons el quocient deute garantia.
Taula 8. Intervals reals amb el corresponent percentatge d'impagament.

En efecte és així. Quan el quocient és més petit (els quatre primers grups) resulta en una proporció d'impagaments més baixa que la mitjana global. Sembla que les situacions més segures per al banc és quan el prestatari ofereix una garantia molt pròxima a la quantitat creditícia demanada. D'altra banda, a mesura que el quocient augmenta per sobre d'1,145 les probabilitats d'impagament també augmenten

significativament, assolint un màxim d'un 12,74% per als ratis més grans (grup 8) compresos entre 1,277 i 6.

El nombre de crèdits actius pot ésser una altra variable que pugui ajudar al model a classificar un individu. Un crèdit en estat vigent comporta a que el client segueix pagant quotes alienes al préstec d'estudi. Això significa que s'està destinant part del capital del prestatari a pagar una altra operació financera i que, per tant, presentar un nombre elevat de crèdits actius podrà suposar majors dificultats a efectuar els pagaments obligatoris. *NUM_ACTIVE_CREDITS* és un atribut de les taules creat a partir de la suma del nombre de crèdits actius tant fora com dins de *Home Credit*.

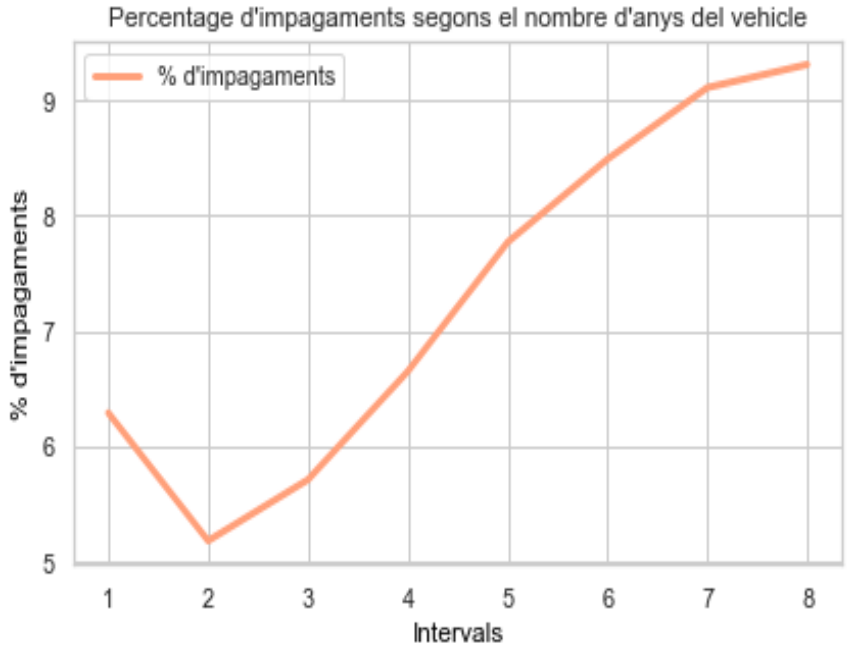


Grup	Intervals reals	% d'impagaments
1	[0,1]	7,50%
2	(1,2]	7,24%
3	(2,3]	7,99%
4	(3,4]	8,60%
5	(4,5]	9,77%
6	(5,34]	12,38%

Figura 16. Evolució del percentatge d'impagaments segons el nombre de crèdits actius.
Taula 9. Intervals reals amb el corresponent percentatge d'impagament.

El plantejament anterior aparenta ser cert. Sembla que presentar entre cap i tres crèdits en estat vigent no té tan risc com presentar-ne més de tres. S'observa un risc d'impagament creixent a mesura que augmenten els crèdits actius.

Una altra variable que no fa referència al capital i que també podria tenir un cert pes en el model és els anys del vehicle del client. Per fer-ho s'han tingut en compte únicament aquells clients que sí presentaven cotxe. Es podria pensar que aquells clients que posseeixen un vehicle antic són més propensos a incomplir degut a que no tenen el capital necessari per a renovar-lo.



Grup	Intervals reals	% d'impagaments
1	[0,5;2]	6,29%
2	(2;5]	5,18%
3	(5;7]	5,71%
4	(7;9]	6,65%
5	(9;12]	7,77%
6	(12;15]	8,49%
7	(15;20]	9,11%
8	(20;91]	9,31%

Figura 17. Evolució del percentatge d'impagaments segons el nombre d'anys del vehicle.
Taula 10. Intervals reals amb el corresponent percentatge d'impagament.

La informació que es pot extreure són els següents: a mesura que el vehicle es fa vell la probabilitat d'impagament augmenta considerablement amb un màxim del 9,31% quan l'automòbil té més de 20 anys. Tot i això hi ha dos factors a destacar:

- Quan el vehicle té menys de 12 anys, la probabilitat d'incomplir es troba per sota de la mitjana del conjunt i és entre dos i cinc anys després de l'adquisició del vehicle el moment amb menor probabilitat.
- Hi ha una dada curiosa: si fa menys o igual a 2 anys des de l'adquisició de l'automòbil pot suposar un lleuger augment en la probabilitat d'impagament. Potser és degut a la pèrdua de capital que ha suposat la compra del vehicle.

Més variables a destacar podrien ser l'edat i la puntuació financera externa que se li atorga al client. Per a representar els gràfics d'aquestes dues variables s'ha utilitzat la densitat de *kernel* ja que permet "balancejar" el nombre d'observacions per als nivells de *TARGET* facilitant la comparació entre ambdós grups.

Respecte el primer, l'edat, es mostra com la distribució d'aquesta dels prestataris es mou entre els vint-i-pocs fins pocs menys dels setanta. El més important, però, és que el grup d'impagaments marcat amb color salmó presenta una asimetria positiva bastant marcada acumulant més impagaments entre la població més jove. A mesura que augmenta l'edat, a partir de la trentena, es redueixen progressivament el nombre d'impagaments. D'altra banda, la proporció de no impagaments es mostra constant en qualsevol edat trobant-se únicament per sobre dels impagaments a partir dels quaranta anys.

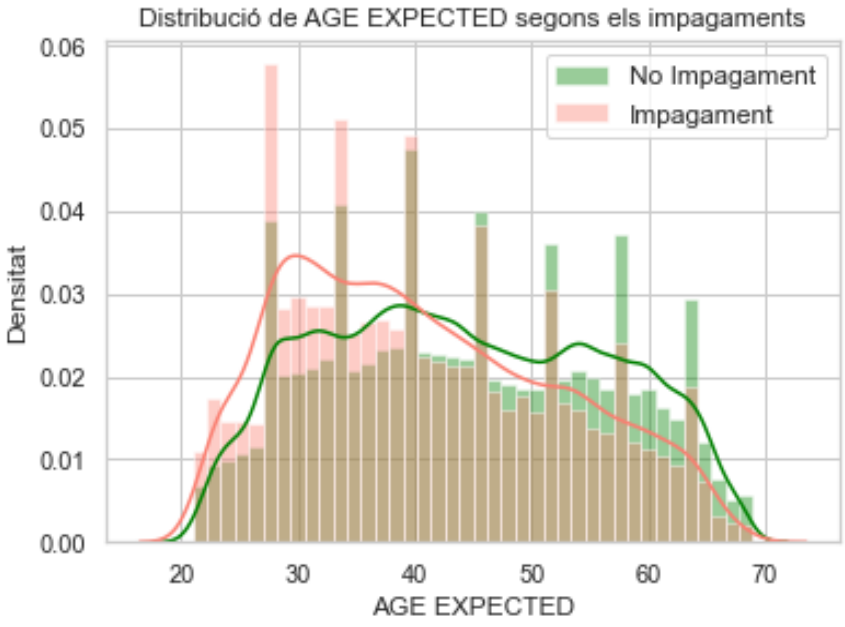


Figura 18. Distribució de les edats segons els grups de TARGET.

La puntuació mitjana (*MEAN_EXT_SOURCE*) és calcula mitjançant la mitjana calculada a partir de les tres puntuacions donades per altres entitats financeres al client. Aquesta valoració pot prendre valors entre 0 i 1. A continuació es mostra la gràfica amb les densitats:

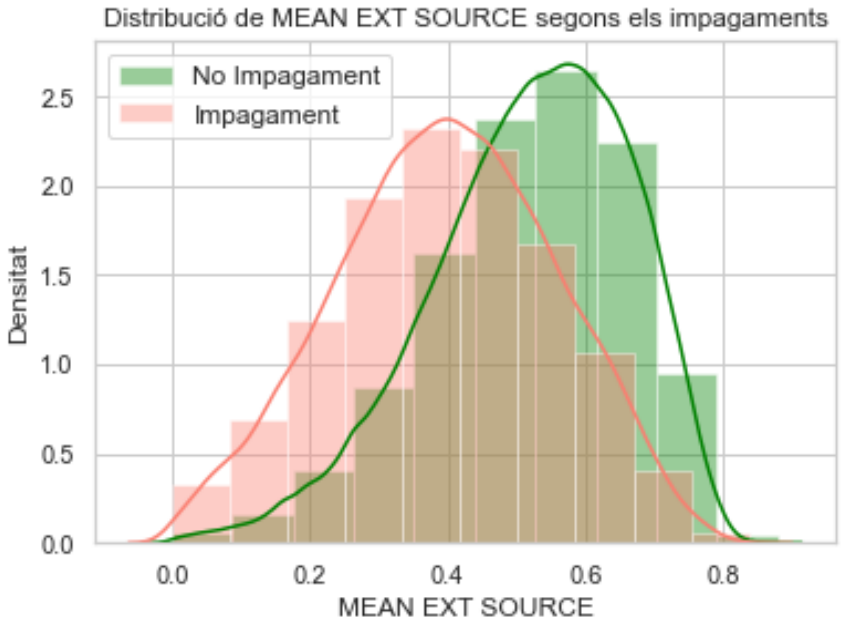


Figura 19. Distribució de les puntuacions mitjanes segons els grups de TARGET.

La distribució d'aquestes puntuacions acords amb els diferents nivells de la variable resposta podrien seguir perfectament una normal. Mentre que el grup salmó és simètrica i mesocúrtica, els no impagaments presenten una asimetria negativa notable. De totes maneres, presentar una mitjana petita pot comportar a presentar una major probabilitat d'impagament mentre que, si la mitjana és elevada, el client tindria bastants números en que acabaria complint fins al final de l'operació financera.

Per últim, es podria analitzar el conjunt de variables numèriques a partir d'estudiar-ne les relacions entre elles. Per a efectuar-ho es pot fer ús de la correlació de Pearson, que quantifica la dependència lineal entre dues variables sense veure's afectada per les diferents escales. Malauradament, l'existència de moltes variables numèriques fa difícil que es mostrin les relacions una per una, però es pot utilitzar una altra tècnica visual apte en aquestes situacions: un mapa de calor. Aquesta metodologia permetrà pintar la relació entre dues variables tenint en compte dos colors i la intensitat d'aquests. Si la correlació és positiva és pintarà de color blau i a mesura que augmenti el blau es tornarà més intens. Pel contrari, si és negativa, serà vermella. A continuació es mostra el mapa de calor.

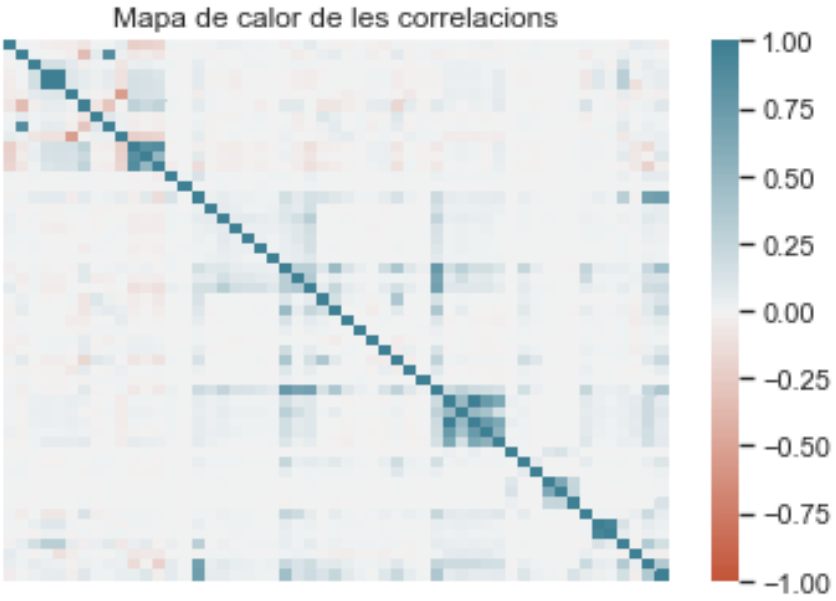


Figura 20. Mapa de calor de les correlacions de totes les variables.

Excloent la diagonal, que és tenint en compte la mateixa variable dues vegades, s'observa com la majoria de relacions són positives. Això significa que quan una variable augmenta, l'altre també ho fa. També es pot extreure que les correlacions, a grans trets, no són fortes (majors a 0,6 o menors a -0,6). S'haurien d'estudiar quines són aquestes parelles de variables amb altes correlacions per a tenir identificar-les en el supòsit que aparegui algun problema de dependència lineal entre variables. En la següent taula es mostren aquelles correlacions fortes amb la mateixa descripció dita anteriorment:

Variable 1	Variable 2	Correlació
AMT_GOODS_PRICE	*AMT_CREDIT*	0,9870
CNT_FAM_MEMBERS	*CNT_CHILDREN*	0,8792
MAX_EXT_SOURCE	*MEAN_EXT_SOURCE*	0,8430
MIN_EXT_SOURCE	*MEAN_EXT_SOURCE*	0,8799
NUM_Repeater_Client_HC	*Approved_STATUS_HC*	0,7696
TOT_RECEIVABLE_CC	*SUM_BALANCE_CC*	0,9999
TOT_DRAWINGS_CC	*SUM_BALANCE_CC*	0,7552
TOT_INSTALMENTS_CC	*SUM_BALANCE_CC*	0,6468
DIES_NOMES_GRAN_DEUTE_PCB	*TOT_CO_ALT_PCB_MENSUAL*	0,6353

PERCENTATGE_ANNUITY	*AMT_ANNUITY_TOT*	0,9516
NUM_ACTIVE_CREDITS	*CRE_TOTAL_Altres_EF*	0,6983
NUM_CREDITS_PREVIS_TANCATS	*CRE_TOTAL_Altres_EF*	0,7199

Taula 11. Parelles de variables amb correlació forta.

Es destaca la relació entre *AMT_GOODS_PRICE* (garantia) i *AMT_CREDIT* (quantitat de crèdit demanada) amb una correlació de 0,9870 i *TOT_RECEIVABLE_CC* (quantitat rebuda en la targeta de crèdit del préstec anterior) i *SUM_BALANCE_CC* (balanç mensual de la targeta de crèdit) amb un valor pràcticament d'1. També *PERCENTATGE_ANNUITY* i *AMT_ANNUITY_TOT* destaquen degut a que aquesta última s'ha utilitzat per a calcular el percentatge.

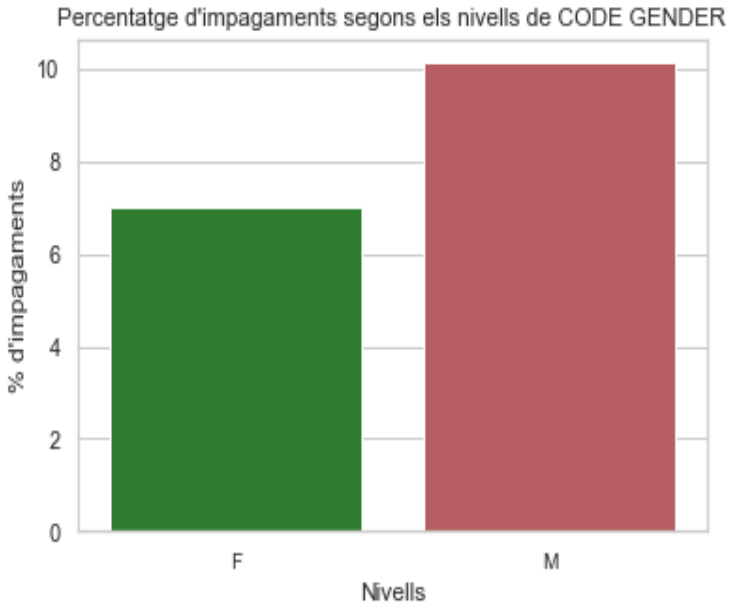
Addicionalment, per a clarificar la *figura 20*, s'han afegit mapes de calor a l'annex (pàgina 118) que permeten identificar la correlació entre totes les parelles existents.

5.2.2 Variables categòriques

Igual d'importants són les variables categòriques. Aquests factors poden prendre un nombre finit de valors, dit comunament com a nivell/s, que normalment estan establerts i són fixes. S'inclouen, també, les variables que són binàries (poden prendre únicament dos valors). La majoria de variables qualitatives en la base de dades expliquen o donen a entendre les característiques del client a nivell personal, social o laboral i es posaran en manifest algunes d'elles, començant per les binàries, amb la mateixa metodologia emprada en l'apartat anterior.

Per a fer-ho s'utilitzaran gràfics de barres amb una escala de color constant. L'escala construïda per aquests gràfics consta de cinc colors ordenats de menys a major risc: verd intens, seguit d'un verd més clar, groc, taronja i finalment el color granat. L'ús d'aquests cinc colors és per a identificar visualment i fàcilment aquells grups amb menor i major proporció d'impagament (verd intens i granat respectivament), especialment útil quan es tinguin múltiples nivells.

El gènere de la persona és un dels aspectes que, en l'actualitat, més apareixen quan es sol·licita informació sobre el client. Malgrat la seva presència abundant, forma part d'un conjunt de variables perilloses perquè podria considerar-se com a factor discriminatiu. De totes maneres, s'ha decidit incloure'l en la taula definitiva.



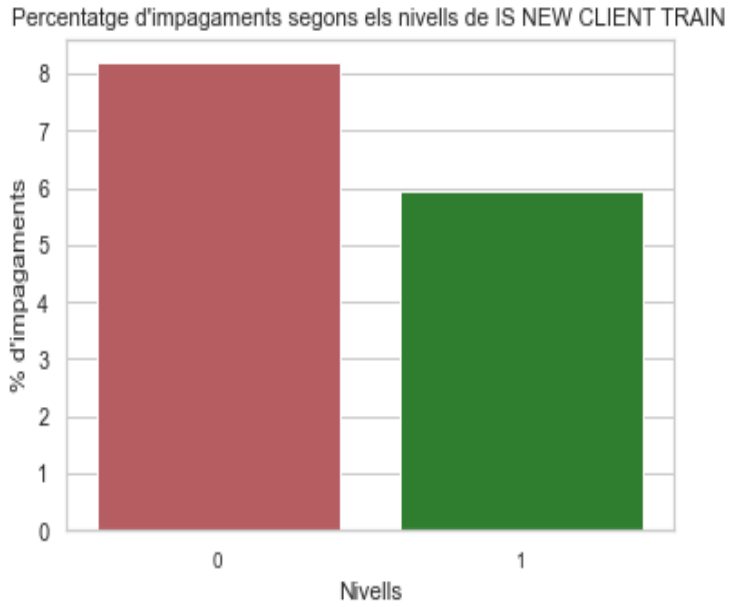
Nivells	Registres totals	Nombre d'impagaments	% d'impagaments
***Female* (F)**	202.452	14.170	7,00%
***Male* (M)**	105.059	10.655	10,14%

Figura 21. Percentatge d'impagaments segons el gènere.
Taula 12. Descriptiva genèrica del nombre d'impagaments per sexe.

Amb gairebé dues terceres parts dels registres, les dones són el col·lectiu amb menor risc d'acabar amb un impagament. Els homes semblen presentar una major probabilitat d'incomplir, presentant un 44% més de risc respecte les dones.

Una altra variable a considerar és el fet de si la persona ja havia contractat o demanat prèviament algun dels productes que ofereix *Home Credit*, com podrien ser préstecs, assegurances, garanties per a mòbils, entre d'altres. Per a recollir aquesta informació, s'han creat dues variables: una numèrica i una altra categòrica. La primera fa referència al nombre de crèdits demanats i la categòrica explica si es tracta d'un nou client o no. Per codificar-la s'han utilitzat els valors 0 i 1 on el 0 representa els clients ja coneguts per a HC mentre que l'1 seria pels que sol·liciten un crèdit a HC per primera vegada.

S'observa que la gran majoria dels prestataris ja estaven registrats a les bases de dades de l'entitat. És destacable, però, la menor probabilitat d'impagament que presenten els nous. Ser un nou client disminueix en un 27% el risc d'incomplir respecte aquells que ja havien operat amb HC.



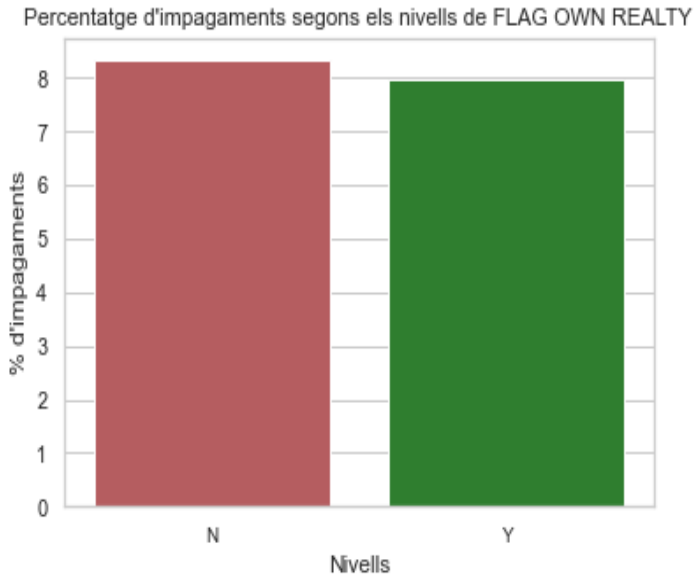
Nivells	Registres totals	Nombre d'impagaments	% d'impagaments
No (0)	291.057	23.845	8,19%
Sí (1)	16.454	980	5,96%

Figura 22. Percentatge d'impagaments segons si és un client nou a HC.
Taula 13. Descriptiva genèrica del nombre d'impagaments segons si és un client nou.

Molts cops els bancs disposen d'informació sobre les propietats del prestatari. En la taula definitiva es poden trobar aquest tipus de dades, més concretament sobre la possessió d'habitatge i de vehicles. Normalment, el que té moltes propietats és perquè disposa de capital i s'ho pot permetre. Seguint aquest plantejament no seria d'estranyar si el fet de tenir un habitatge propi i/o un vehicle reduís considerablement les probabilitats d'impagaments.

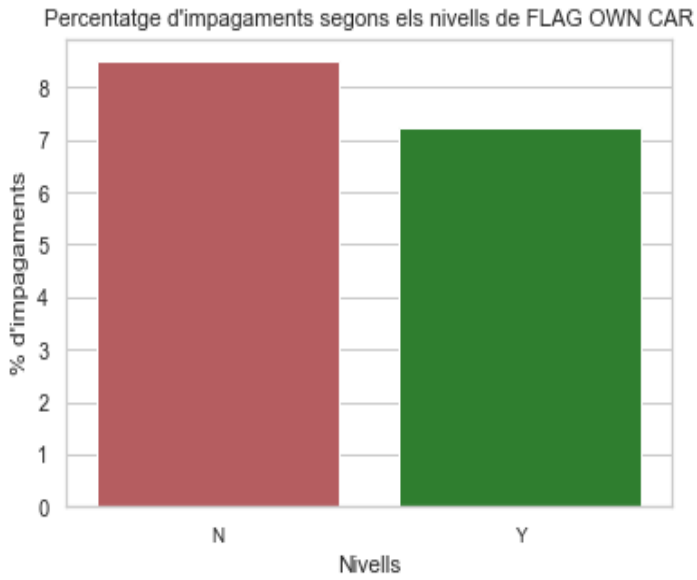
A la pàgina següent es mostren els resultats segons la possessió d'habitatge o d'un cotxe. Respecte el primer, es pot dir que més de la meitat dels clients tenen adquirit un habitatge i que són aquells que no en tenen els que presenten una major probabilitat d'impagament. Tot i això, la diferència entre ambdós grups no és tan marcada com els dos casos previs.

La possessió d'un vehicle segueix la mateixa línia que el de tenir un immoble. Aquells que el disposen presenten una probabilitat inferior a la mitjana mentre els que no en tenen és major. Malgrat tot, les probabilitats per als grups és similar a la mitjana, tot i que el fet de tenir un cotxe és un factor més important a l'hora de reduir-ne la probabilitat respecte l'habitatge.



Nivells	Registres totals	Nombre d'impagaments	% d'impagaments
No (N)	94.199	7.842	8,32%
Sí (Y)	213.312	16.983	7,96%

Figura 23. Percentatge d'impagaments segons si té un habitatge propi.
Taula 14. Descriptiva genèrica del nombre d'impagaments segons si té un habitatge propi.



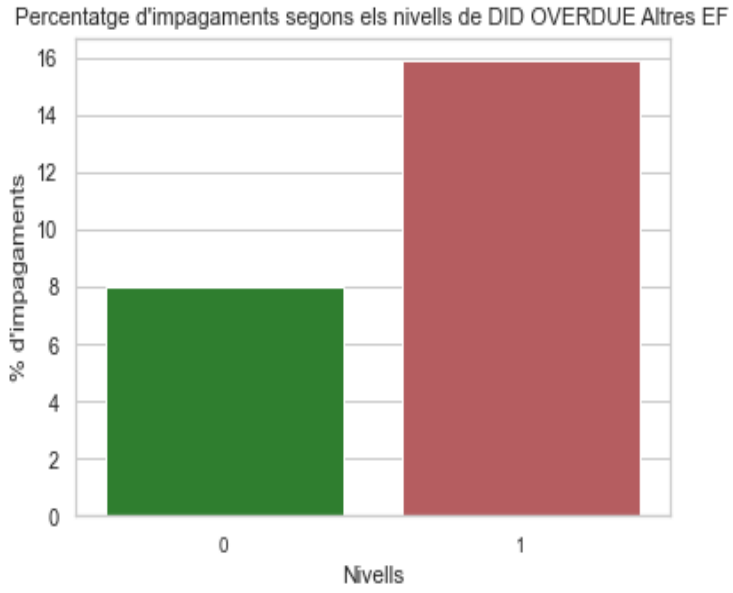
Nivells	Registres totals	Nombre d'impagaments	% d'impagaments
No (N)	202.924	17.249	8,50%
Sí (Y)	104.587	7.576	7,24%

Figura 24. Percentatge d'impagaments segons si disposa d'un vehicle.
Taula 15. Descriptiva genèrica del nombre d'impagaments segons si disposa d'un vehicle.

Aquí es podria lligar amb el vist en la secció anterior. Al final, tenir cotxe disminueix el risc, però dintre dels que en tenen s'observen diferències segons els anys d'antiguitat del mateix. De fet, és interessant veure que els que no tenen cotxe presenten un percentatge d'impagament del 8,5%. Dels que presenten un vehicle el percentatge és igual o més gran del 8,5% quan el vehicle passa dels dotze anys d'antiguitat. És a dir, tenir cotxe beneficia si el cotxe té menys de dotze anys. En cas contrari, posseir un cotxe no aporta cap benefici en la quantificació del risc.

Existeixen múltiples variables binàries més, però es tancarà aquest grup de variables amb una que fa referència a l'historial creditici del client en una altra entitat aliena a HC. La variable *DID_OVERDUE_Altres_EF*, creada a partir d'altres variables, posa en manifest si el client va incomplir en un préstec anterior. En cas afirmatiu pren 1 com a valor i 0 altrament.

Quan es produeix un impagament per part del client és molt possible que es torni a passar perquè significa que el particular o empresa es troba en una situació financera delicada i que podria seguir incorrent a més impagaments en les següents quotes. També podria ser que, en el supòsit que fos el client que hagués d'anar personalment a l'entitat a pagar, s'oblidés d'acudir-hi de manera puntual.



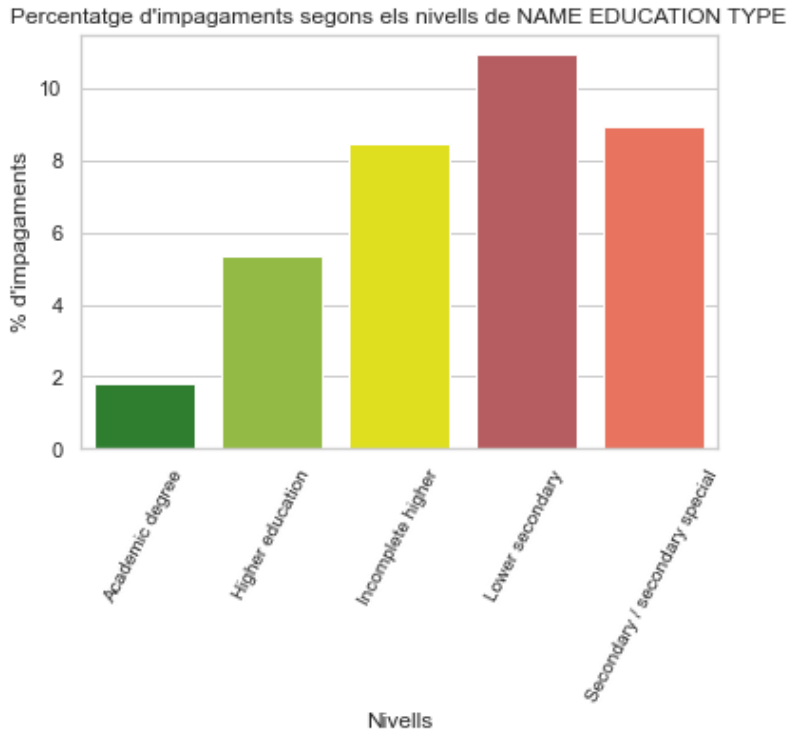
Nivells	Registres totals	Nombre d'impagaments	% d'impagaments
No (0)	304.114	24.285	7,99%
Sí (1)	3.397	540	15,90%

Figura 25. Percentatge d'impagaments segons si el client va retardar algun pagament.
Taula 16. Descriptiva genèrica del nombre d'impagaments segons si el client va retardar algun pagament.

Independentment del motiu de no pagar, sembla ser que presentar-ne augmenta significativament la probabilitat d'impagament. Això afirma la teoria de que si has impagat una vegada, és molt possible que ho torni a fer. També és destacable la poca volumetria per aquest grup en concret: respecte les dades senceres, només una mica més d'un 1% dels individus registrats va incomplir en una altra entitat.

Un cop vistes algunes de les variables binàries d'interès, es pot passar a analitzar aquelles que presenten més d'un nivell. Com s'ha dit a l'inici de l'apartat, s'usarà una escala de colors *verd-granat* per a la visualització dels percentatges. Els tons verds indicaran grups amb menor proporció i els granats amb un major valor.

Una de les variables que més podria destacar és el nivell educatiu. Es pot analitzar si el grau d'estudis té algun efecte en acabar l'operació amb impagaments o no. Normalment, aquells que tenen una educació superior que permetria aspirar a oficis amb una millor retribució respecte a la dels altres grups.



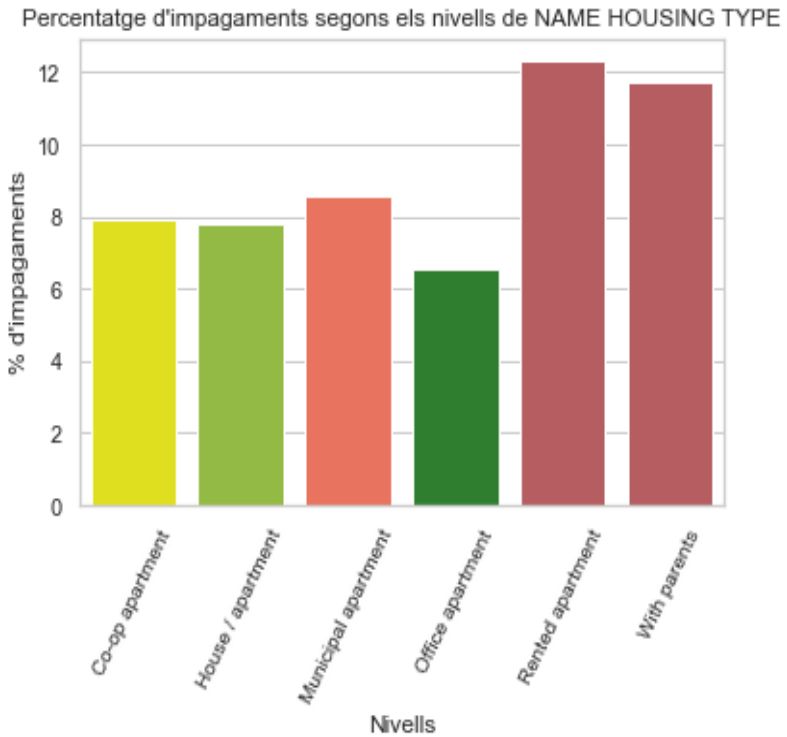
Dades	Registres totals	Nombre d'impagaments	% d'impagaments
Academic degree	164	3	1,83%
Higher education	74863	4009	5,36%
Incomplete higher	10277	872	8,48%

Lower secondary	3816	417	10,93%
Secondary / secondary special	218391	19524	8,94%

Figura 26. Percentatge d'impagaments segons el nivell educatiu.
Taula 17. Descriptiva genèrica del nombre d'impagaments segons el nivell educatiu.

Tot i les diferències en la volumetria de cada grup, es confirma l'existència d'un patró: a mesura que augmenta el grau d'estudis de la persona es redueix la probabilitat d'impagament. Així doncs, els que ostenten un títol de grau universitari presenten una probabilitat d'1,83%, mentre aquells que tenen el grau més baix, la *Lower Secondary* (equivalent als tres primers anys de l'Educació Secundària Obligatòria espanyola), tenen un risc molt més elevat que els anteriors, amb un valor del 10,93%. Sorprèn com més de dues terceres parts de la població d'estudi disposa únicament d'un nivell educatiu semblant al d'un alumne espanyol de 4t d'ESO.

Un aspecte completament diferent a analitzar podria ser la tipologia d'habitatge on viu el client, és a dir, si viu en una casa, en un habitatge de lloguer o social, entre molts altres. La informació sobre l'habitatge la recull *NAME_HOUSING_TYPE*.



Nivells	Registres totals	Nombre d'impagaments	% d'impagaments
Co-op	1.122	89	7,93%

apartament			
House / apartament	272.868	21.272	7,80%
Municipal apartament	11.183	955	8,54%
Office apartament	2.617	172	6,57%
Rented apartament	4.881	601	12,31%
With parents	14.840	1.736	11,70%

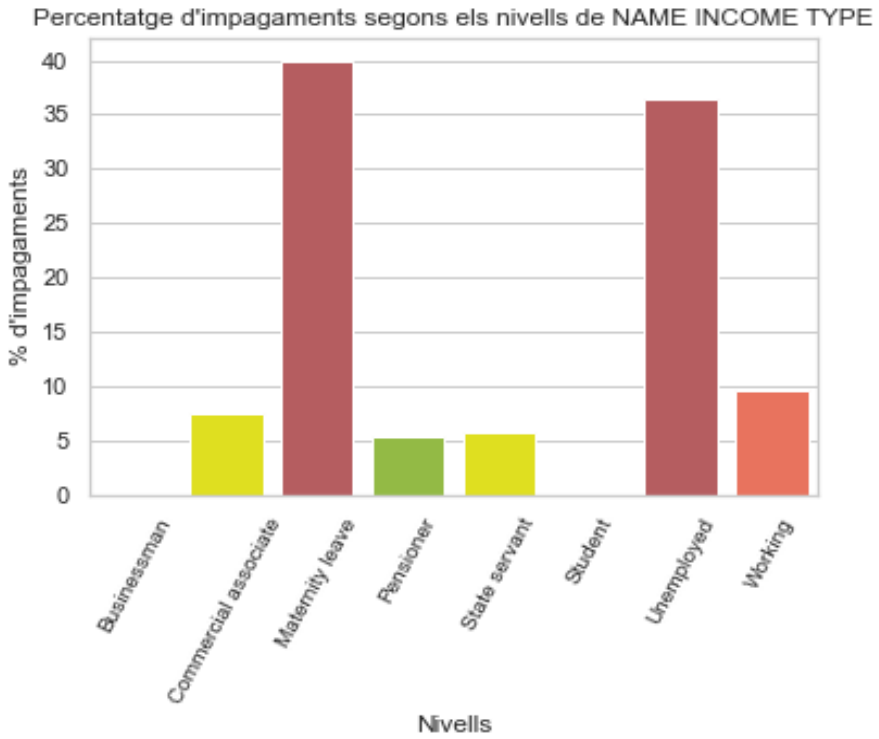
Figura 27. Percentatge d'impagaments segons el tipus d'habitatge on viu.
Taula 18. Descriptiva genèrica del nombre d'impagaments segons el tipus d'habitatge on viu.

Succeeix el mateix que per al cas del nivell educatiu: les volumetries són dispars entre els diferents grups, però no suposa cap problema. La gran majoria dels clients de la taula completa viuen en una casa o apartament propi (89%) mentre que la resta viuen en altres tipus d'habitatges. Destaquen per tenir una probabilitat d'impagament significativament per sobre de la mitjana aquells que viuen de lloguer o amb els pares. Aquests tenen valors del 12,31% i 11,70% respectivament. És possible que els que estiguin de lloguer ja paguin una part important del seu salari per mantenir el pis mentre que els que viuen amb els pares no deuen rebre una aportació econòmica suficient. D'altra banda, els que habiten en un apartament-oficines són els que presenten un menor grau d'incompliment.

Ja s'ha vist en la part numèrica com evolucionava la probabilitat d'impagament acord als intervals de salari però ara es pot fer des d'una perspectiva més qualitativa: el tipus de salari i el seu ofici. El tipus de salari correspon al per què o quina és l'activitat que efectua per a rebre una retribució econòmica. Al gràfic i taula següent es mostren les probabilitats d'impagaments segons el tipus.

El primer aspecte a destacar és l'alta probabilitat d'impagament que tenen aquelles persones que reben ingressos a partir d'una baixa de maternitat o bé una prestació d'atur. Aquests grups presenten una probabilitat major al 35%. Tot i això, i conjuntament als homes de negocis i estudiants, presenten molt poques observacions. Això podria explicar el per què d'un valor tant alt, però també tindria sentit que fos així perquè normalment les prestacions que atorguen els governs a aquests col·lectius és molt petita. D'altra banda, els homes de negoci i estudiants presenten un risc nul d'impagament, seguits dels pensionistes i dels funcionaris amb

probabilitats que superen el 5%, notablement per sota de la mitjana establerta al voltant del 8%.



Dades	Registres totals	Nombre d'impagaments	% d'impagaments
Businessman	10	0	0%
Commercial associate	71.617	5.360	7,48%
Maternity leave	5	2	40%
Pensioner	55.362	2.982	5,39%
State servant	21.703	1.249	5,75%
Student	18	0	0%
Unemployed	22	8	36,36%
Working	158.774	15.224	9,59%

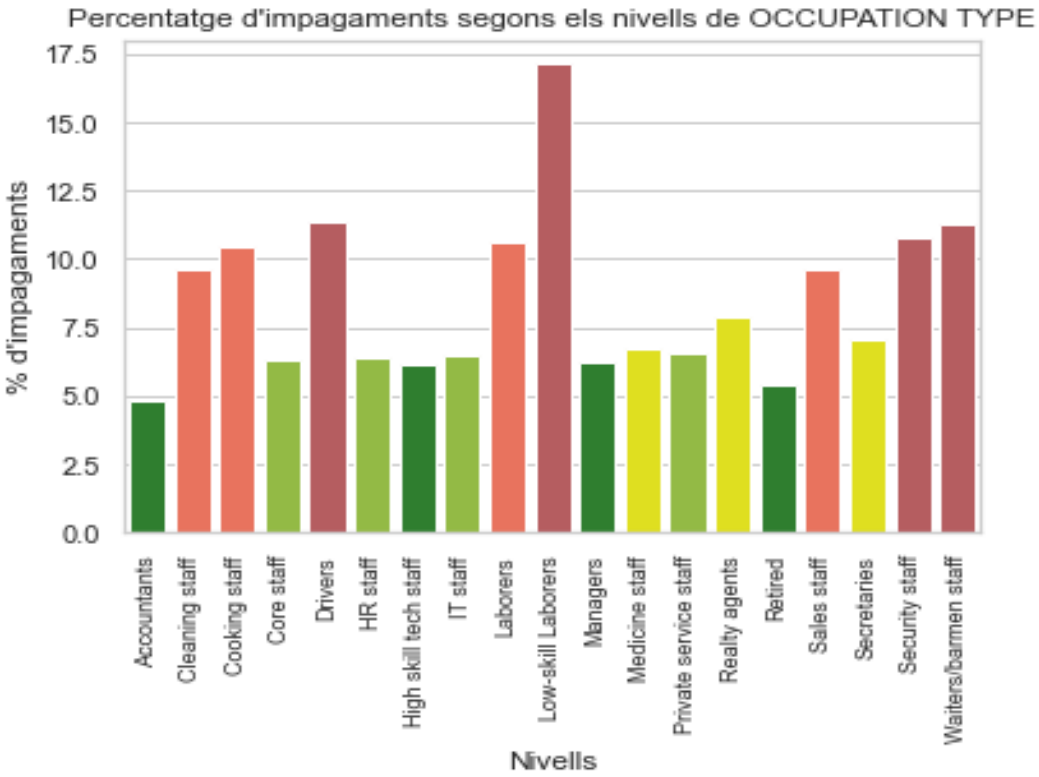
Figura 28. Percentatge d'impagaments segons el tipus d'ingrés.
Taula 19. Descriptiva genèrica del nombre d'impagaments segons el tipus d'ingrés.

El primer aspecte a destacar és l'alta probabilitat d'impagament que tenen aquelles persones que reben ingressos a partir d'una baixa de maternitat o bé una prestació d'atur. Aquests grups presenten una probabilitat major al 35%. Tot i això, i

conjuntament als homes de negocis i estudiants, presenten molt poques observacions. Això podria explicar el per què d'un valor tant alt, però també tindria sentit que fos així perquè normalment les prestacions que atorguen els governs a aquests col·lectius és molt petita. D'altra banda, els homes de negoci i estudiants presenten un risc nul d'impagament, seguits dels pensionistes i dels funcionaris amb probabilitats que superen el 5%, notablement per sota de la mitjana establerta al voltant del 8%.

Finalment, com a última variable que es veurà, s'estudiaran les probabilitats d'impagament segons els oficis. La base de dades incorpora informació sobre l'ocupació professional del client tot i que té present una gran quantitat de dades mancants.

Serà d'interès veure si els oficis poden jugar un paper important a l'hora de pràcticar la classificació. D'entrada s'esperaria que fossin els oficis amb menys professionalització, i per tant menor salari, els que presentesin major probabilitat i, per contra, els que tinguin una major especialització que tindrien una retribució força més elevada que la resta.



Dades	Registres totals	Nombre d'impagaments	% d'impagaments
Accountants	9.813	474	4,83%

Cleaning staff	4.651	447	9,61%
Cooking staff	5.946	621	10,44%
Core staff	27.570	1.738	6,30%
Drivers	18.603	2.107	11,33%
HR staff	563	36	6,39%
High skill tech staff	11.380	701	6,16%
IT staff	526	34	6,46%
Laborers	55.185	5.838	10,58%
Low-skill laborers	2.093	359	17,15%
Managers	21.371	1.328	6,21%
Medicine staff	8.536	572	6,70%
Private service staff	2.652	175	6,60%
Realty agents	751	59	7,86%
Retired	55.362	2.982	5,39%
Sales staff	32.101	3.092	9,63%
Secretaries	1.305	92	7,05%
Security staff	6.721	722	10,74%
Waiters / barmen staff	1.348	152	11,28%

Figura 29. Percentatge d'impagaments segons l'ofici del client.
Taula 20. Descriptiva genèrica del nombre d'impagaments segons l'ofici del client.

El col·lectiu que destaca amb diferència és el *Low-skill laborers* que formarien part les persones amb unes habilitats laborals deficients o notablement millorables. Es tracta doncs d'un grup amb molt baixa professionalització. El segueixen oficis que no necessiten d'una acreditació concreta com podrien ser els guàrdies de seguretat, xefs o cambrers d'un restaurant.

Per una altra banda es trobarien els comptables, pensionistes o treballadors amb molt bones habilitats tecnològiques. Es tracten de professions, menys els pensionistes, que sí necessiten d'un títol o curs específic per a poder exercir. Al

final, totes les variables vistes fins ara poden relacionar-se d'alguna forma o una altra, com seria el nivell educatiu i l'ofici.

En tot aquest apartat d'anàlisi descriptiu s'ha intentat descriure com són en general tot el conjunt de variables en les taules. En les fase següent, el procés de modelització, acabaran apareixent la resta de variables, però no es detallaran tots els seus valors o nivells que la corresponen. De totes maneres i per a més informació, es pot consultar en qualsevol moment l'annex en cas de dubtes.

5.3 Construcció dels models I: Modelització tradicional

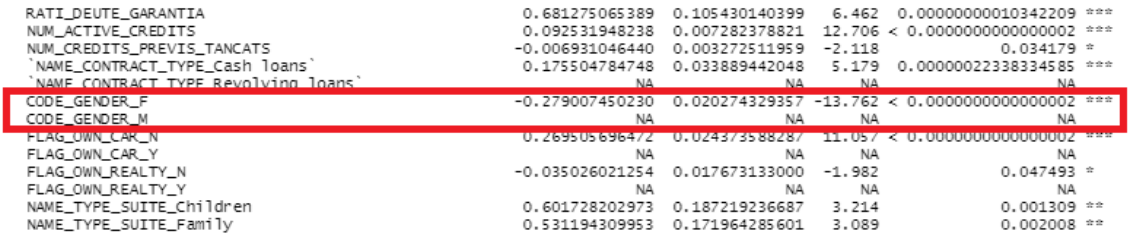
L'etapa de modelar és sens dubte la que requereix de més coneixements i aptituds per portar-la a terme. S'han de tenir en compte tots els problemes potencials que poden sorgir en la construcció i desenvolupament del model per poder utilitzar-se en la fase de producció. En aquesta part es veuran totes les etapes posteriors al muntatge de la base de dades dels models que corresponen a la metodologia clàssica: els models binaris. Hi ha múltiples tècniques que fan referència als models d'elecció binària, però únicament es veuran el Logístic i l'anterior amb penalització Lasso o Ridge per a la modelització de la variable resposta *TARGET* a partir del fitxer d'entrenament. No es farà referència a la capacitat predictiva del model en si mateixa.

5.3.1 Model logístic

El model logístic, el qual rep el seu nom per utilitzar el link *logit* com a forma d'estimar els coeficients, és el model més accessible per al supervisor que es tractarà en el conjunt del treball. Permet classificar als individus en un conjunt finit de possibilitats a partir d'introduir informació a l'algoritme per a que identifiqui certes característiques de les observacions que permetin al model atorgar-li una categoria en concret.

Tot i ser simple o accessible per al supervisor, hi ha un gran nombre d'aspectes destacables que s'han de tenir en compte per vetllar pel bon funcionament del model. Per a simplificar-ho tot, es considerarà imprescindible la no presència i existència de combinacions lineals entre variables i de dades atípiques. Aquests dos factors a observar poden alterar el model i estimar valors que no s'adeqüen a la realitat. De totes maneres, es farà una descripció per cada model construït fins arribar a l'últim a partir de les millores aplicades en cadascun dels models.

El primer model a considerar és el que conté com a variable resposta si l'operació financera va acabar amb un impagament o no i tota la resta de variables seran utilitzades per a intentar explicar el comportament d'aquesta. En l'apartat *5.1 Base de dades* es s'ha detallat el procés de creació dels fitxers finals, incloent-hi la incorporació de variables fictícies. Les *dummies* generen tantes variables com nivells tingui l'atribut de referència i, si no es fa cap tractament d'eliminació d'alguna d'elles, s'estaria en una situació amb múltiples combinacions lineals. L'R arregla parcialment aquest problema al no incloure un dels nivells en el model expressant-ho amb un *NA* (dada mancant). En la figura 30 s'emmarca amb un rectangle vermell una de les situacions amb combinació lineal entre les dues variables fictícies del gènere del prestatari. Es tracta, igualment, d'uns inconvenients amb fàcil remei.



RATI_DEUTE_GARANTIA	0.681275065389	0.105430140399	6.462	0.00000000010342209	***
NUM_ACTIVE_CREDITS	0.092531948238	0.007282378821	12.706	< 0.0000000000000002	***
NUM_CREDITS_PREVIS_TANCATS	-0.006931046440	0.003272511959	-2.118	0.034179	*
`NAME_CONTRACT_TYPE_Cash loans`	0.175504784748	0.033889442048	5.179	0.00000022338334585	***
`NAME_CONTRACT_TYPE_Revolving loans`	NA	NA	NA	NA	
CODE_GENDER_F	-0.279007450230	0.020274329357	-13.762	< 0.0000000000000002	***
CODE_GENDER_M	NA	NA	NA	NA	
FLAG_OWN_CAR_N	0.269505696472	0.024373588287	11.057	< 0.0000000000000002	***
FLAG_OWN_CAR_Y	NA	NA	NA	NA	
FLAG_OWN_REALTY_N	-0.035026021254	0.017673133000	-1.982	0.047493	*
FLAG_OWN_REALTY_Y	NA	NA	NA	NA	
NAME_TYPE_SUITE_Children	0.601728202973	0.187219236687	3.214	0.001309	**
NAME_TYPE_SUITE_Family	0.531194309953	0.171964285601	3.089	0.002008	**

Figura 30. Alguns coeficients del primer model lògit amb combinacions lineals.

És important recordar quines són les variables fictícies que s'acaben eliminant perquè afecten a l'estimació dels coeficients. Aquest canvi es deu a que les altres variables avaluen les seves diferències respecte a la *dummie* que s'ha eliminat prèviament. És a dir, si existeixen cinc variables fictícies per a explicar el nivell educatiu del client i es decideix eliminar-ne *Higher Education* per a prevenir la combinació lineal, les estimacions dels coeficients de les altres fictícies seran comparades amb *Higher Education* en funció si són un col·lectiu amb major o menor risc d'impagament. Al següent llistat es mostren les variables que s'han eliminat:

Variables fictícies eliminades
"NAME_CONTRACT_TYPE_Cash loans"
"CODE_GENDER_F"
"FLAG_OWN_CAR_N"
"FLAG_OWN_REALTY_N"
"NAME_TYPE_SUITE_Uaccompanied"
"NAME_INCOME_TYPE_Commercial associate"

“NAME_EDUCATION_TYPE_Secondary / secondary special”
“NAME_FAMILY_STATUS_Married”
“NAME_HOUSNG_TYPE_House / apartament”
“LAST_YEARS_EMPLOYED_0-3”
“OCCUPATION_TYPE_Accountants”
“ORGANIZATION_TYPE_Business”
“DIES_MAX_IMPAGATS_Altres_EF_PERFECT”
“Credits_Assegurats_HC_Mateixos credits”
“SUM_Diferencia_AMT_HC_Less credit than ATB”
“LAST_DUE_MONTH_Less_6Months”

Taula 21. Variables fictícies eliminades per a la construcció del model logístic.

Després de l’eliminació de les variables anteriors, és hora de plantejar un segon model. S’ha cregut convenient afegir dues interaccions (factors que tenen en compte dues variables) al segon model: el fet de tenir casa i automòbil i haver allargat els pagaments i impagat els préstecs en altres entitats financeres. L’addició d’interaccions permet donar més informació al model que podria ser important per a la classificació.

Les combinacions lineals ocorregudes en el primer model eren degudes a la transformació en variables *dummies* dels factors categòrics. Això no exclou que no pugui succeir també per a les variables numèriques. És el que s’anomena com a multicol·linealitat (també per a les categòriques). Es tracta d’un problema que s’ha d’evitar ja que, en cas contrari, pot incrementar la variància dels coeficients de la regressió, fent-los inestables.

Hi ha múltiples maneres d’analitzar-la, però s’utilitzarà el VIF (*Variance Inflation Factor*) com a mètode per a quantificar la dependència lineal entre les variables. Es tracta d’una mètrica senzilla d’interpretar: quan el VIF és major a 5 es considera una dependència lineal gran i, si sobrepassa els 10, la dependència serà molt gran. Per a reduir-ne l’espai es detallaran únicament aquelles variables que presenten un VIF major a 5.

Variable	VIF
AMT_CREDIT	103,1696
AMT_GOODS_PRICE	105,0233
MEAN_EXT_SOURCE	64,4157
MAX_EXT_SOURCE	23,5853
MIN_EXT_SOURCE	21,5141
CRE_TOTAL_Altres_EF	6,7182
Approved_STATUS_HC	11,4043

Variable	VIF
Canceled_STATUS_HC	5,9700
Refused_STATUS_HC	9,1830
NUM_Repeater_Client_HC	31,1275
SUM_BALANCE_CC	12.888,5249
TOT_REICEVABLE_CC	12.848,7265
AMT_ANNUITY_TOT	26,8860
PERCENTATGE_ANNUITY	26,7669

Taula 22. Variables amb un VIF major a 5.

Hi ha un total de 14 variables que presenten valors majors a 5. El que es fa en situacions similars és eliminar una o múltiples variables de la dependència lineal per aconseguir que el VIF sigui inferior o proper a 5. La majoria d'elles ja s'havien vist en les correlacions de la descriptiva avançada i per això es sap quines són les variables que formen les diferents combinacions lineals. Per exemple, *SUM_BALANCE_CC* i *TOT_RECEIVABLE_CC* presenten un VIF extremadament gran i també es coneix que la seva correlació era de pràcticament 1. Serà suficient eliminar-ne una d'elles. Altres dependències lineals són *AMT_CREDIT* amb *AMT_GOODS_PRICE*, *MEAN_EXT_SOURCE* amb *MAX_EXT_SOURCE* i *MIN_EXT_SOURCE*, entre d'altres. A la taula 22 es mostren, en un color salmó, les variables que s'eliminaran per a reduir els VIF.

Un últim aspecte a comprovar són les observacions influents. Aquestes poden ser molt perilloses perquè es tracten de registres amb valors poc comuns que acaben tenint un impacte desproporcionat en un model. En altres paraules, un client pot presentar informació molt diferent a la resta de clients i que el model doni més pes al primer fent, per exemple, que un coeficient realment significatiu no ho sigui. Aquesta influència es quantificarà a partir de la distància de Cook i s'eliminaran les observacions que presentin una distància extremadament alta en comparació de la resta.

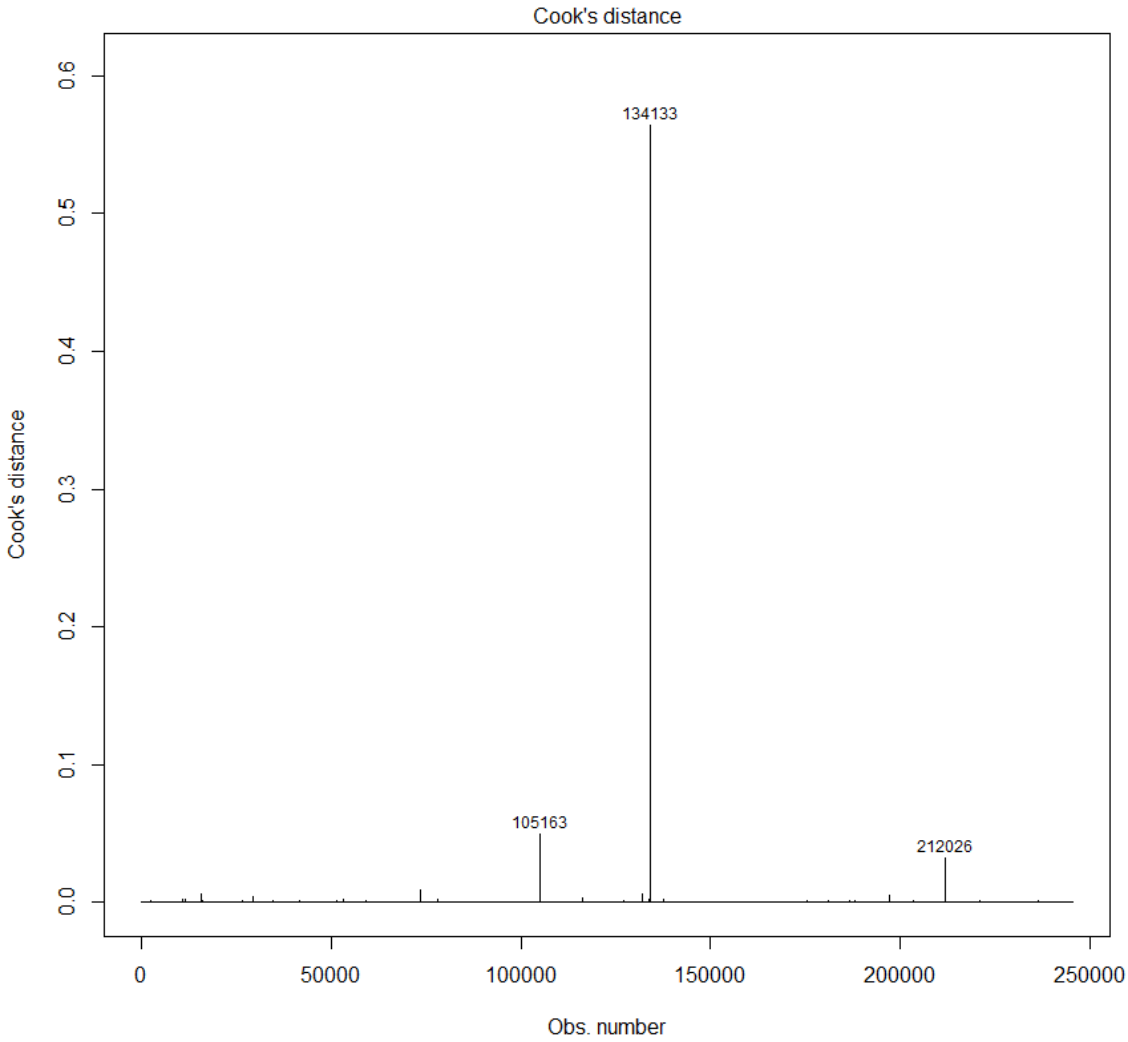


Figura 31. Distància de Cook per a cada observació.

La majoria de distàncies es mouen lleugerament per sobre de zero, però n'hi ha tres que destaquen per sobre de la resta que resulten ser la 105.163, 134.133 i 212.026. Tot i que les tres presentin valors per sobre de la resta, només la 134.133 serà eliminada per ser extremadament influent. S'haurien pogut eliminar també els altres registres, però s'ha de tenir molta cura. A vegades, eliminar dades atípiques pot no ser un encert. Podria ésser que s'estigui deixant de banda col·lectius que realment existeixen i que a l'eliminar-les, el model no classifica adequadament. Per aquest motiu es mantindran, excepte aquelles que verdaderament puguin portar problemes.

El tercer model és el que resol tots els problemes plantejats anteriorment. S'han eliminat variables dependentment lineals i també l'observació influent. D'aquesta manera s'ha obtingut un model tenint en compte totes les variables possibles del conjunt de dades. Es tracta d'un model apte per ser utilitzat i que podria ser emprat per a entendre com d'importants considera cadascuna de les variables i quins efectes

tenen cada variable en la classificació entre impagament i no impagament. A l'annex es troba l'estimació dels coeficients feta pel model logístic (pàgina 130).

Interpretar els models logístics és una tasca que no suposa de gaire dificultat per al supervisor. Tot i això és d'importància saber que són els coeficients amb un p-valor més petit de 0,05 els que permeten diferenciar d'alguna manera les dues categories diferents de la variable resposta. Els que resulten ser majors són variables o nivells que no ajuden a discriminar. Del model tres es poden extreure informació sobre els següents factors:

- Els coeficients de la majoria de les variables que fan referència al capital del client són positius. Això indica que valors elevats d'aquestes fan augmentar la probabilitat d'impagament. La quantitat demanada del crèdit (AMT_CREDIT), el nombre de quotes pendents a pagar (NUM_QUOTES_PENDENTS), el rati deute garantia (RATI_DEUTE_GARANTIA) i el fet de no pagar la quantitat monetària exacta de la quota (NOT_ENOUGH_MONEY_PAID_IP), entre d'altres, que el model detecta com a importants.
- Per una altra banda, el coeficient de la variable salari del client (AMT_INCOME_TOTAL) és significatiu i negatiu i per tant redueix la probabilitat d'impagament a mesura que augmenta el salari.
- Respecte les variables que fan referència a l'historial creditici del client, es troben, per exemple, l'aprovació prèvia d'algun crèdit a HC (Approved_STATUS_HC), haver tingut múltiples crèdits tancats tenint en compte HC com altres entitats financeres (NUM_CREDITS_PREVIS_TANCATS) i la puntuació financera mitjana atorgada al client (MEAN_EXT_SOURCE) redueixen la probabilitat d'impagament. En canvi, el nombre de crèdits actius (NUM_ACTIVE_CREDITS) i retirar molts diners de la targeta de crèdit (TOT_DRAWINGS_CC) augmenten considerablement el risc d'incomplir.
- Per últim i considerant les variables socials, es pot extreure que els clients amb major nivell educatiu, amb vehicle (FLAG_OWN_CAR_Y), que tenen un préstec rotatori (NAME_CONTRACT_TYPE_Revolving Loans) i que porten treballant més de 10 anys (LAST_YEARS_EMPLOYED_10-80) són característiques que detecta el model per a reduir la probabilitat d'impagament. Si es miren els que provoquen l'efecte contrari es trobarien el nombre d'infants (CNT_CHILDREN), si es viu divorciat (NAME_FAMILY_STATUS_Separated) o en un matrimoni establert a partir del registre civil (NAME_FAMILY_STATUS_Civil Marriage), entre d'altres.

La complexitat d'aquest model, deguda al gran nombre de variables inserides, comporta a la necessitat de reduir-ne la complexitat sempre que sigui possible. Una opció seria construir un algoritme que permetés construir el model que optimitzés l'AIC, una mètrica de qualitat del model. Es tracta d'un mètode fàcil d'aplicar amb els programes d'avui en dia, però per a aquestes dades té un problema: comporta a un alt cost computacional. Trobar aquest model seria una tasca que comportaria més d'un dia de feina per a l'ordinador. Per aquesta raó s'ha optat a considerar únicament les variables significatives (incloent tots els nivells d'una variable categòrica) del model. Amb això s'aconseguirà reduir-ne la complexitat i facilitar-ne la interpretació. Es tractarà d'un nou model i que en fases posteriors es compararà entre aquest i el que utilitza totes les variables existents.

Recordar que a l'annex es poden consultar totes les sortides del programari R que permeten explicar cadascun dels coeficients de tots els models vistos.

5.3.2 Model logístic amb penalització Lasso

Existeixen vàries alternatives per a convertir un model complex en un de més senzill. Hi ha uns mètodes que permeten trobar un subconjunt de variables resultant en un model més òptim que l'inicial. Utilitzant-se per a aquest cas de classificació binària, la metodologia en qüestió rep el nom de regressió logística penalitzada. Per a trobar el millor model s'incorpora un paràmetre de penalització ja que conté moltes variables. El que s'aconsegueix és que tots aquells coeficients dels factors que resulten ser no significatius prenguin valors al voltant del zero. Un dels mètodes relacionats amb la inserció del factor de penalització és el Lasso.

La penalització Lasso consisteix en establir un coeficient de zero per a totes aquelles variables no significatives o bé, que no són suficientment importants com per a explicar la resposta. Llavors, només les que ha considerat importants es mantenen en el model final. S'elaboraran dos models: un amb totes les variables existents i un altre que inclogui únicament aquelles significatives del model logístic anterior.

L'R disposa d'un conjunt de funcions molt hàbils per a modelitzar aquest tipus de models. Permet dissenyar la rigorositat de la reducció de coeficients del model. Et dona la possibilitat de trobar el valor del paràmetre referent a la reducció òptima a partir de la tècnica de cross-validació. L'R el codifica com a lambda λ. En el gràfic següent es mostra l'evolució del logaritme de lambda (eix de les abscisses) a mesura que s'afegeixen variables (barra superior) i es redueix l'error (eix d'ordenades). La millor lambda resulta ser la més petita (0,0000982) i inclou la majoria de les

variables explicatives, és a dir, no s'elimina gairebé cap coeficient tal i com considera el principi del Lasso.

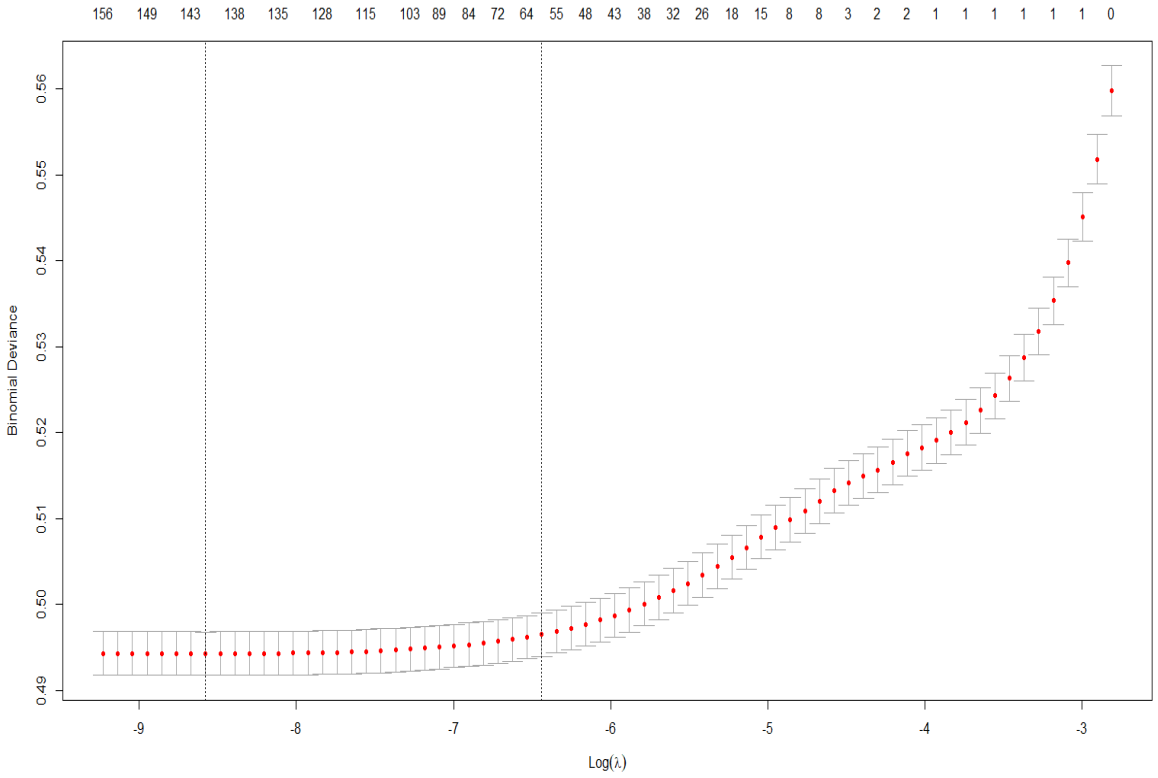


Figura 32. Criteri de selecció per a la lambda òptima que redueix l'error de predicció a partir de la cross-validació. Gràfic amb el model Lasso amb totes les variables. El logaritme de la lambda òptima (0,0000982) és -9,23.

Hi ha una altra forma de trobar la millor lambda, però d'una manera més complexa i amb un major cost computacional: elaborar tants models com lambdes diferents existents hi hagi i llavors comparar l'índex de gini i l'AUC de les estimacions realitzades per a les dades d'entrenament agafant aquella que maximitzi les mètriques anteriors. De totes maneres, s'aconsegueix la mateixa lambda utilitzant aquest mètode.

A partir de l'obtenció de la lambda anterior es procedeix a estimar els coeficients del model. El mateix procediment es segueix per a construir el model considerant únicament les variables significatives trobades al model logístic.

5.3.3 Model logístic amb penalització Ridge

L'última opció aplicable que s'ha utilitzat és la regressió logística amb la penalització Ridge. Aquesta, en comptes de considerar amb zeros totes les variables no

importants com considerava el Lasso, té en compte cadascuna d'elles però les que resulten ser no significatives el coeficient serà molt proper a zero.

Les característiques per a la construcció del model Ridge són les mateixes que les del Lasso: és necessari trobar la lambda que minimitzi l'error de predicció. En la següent gràfica es mostra l'evolució de les lambdes i el seu respectiu error.

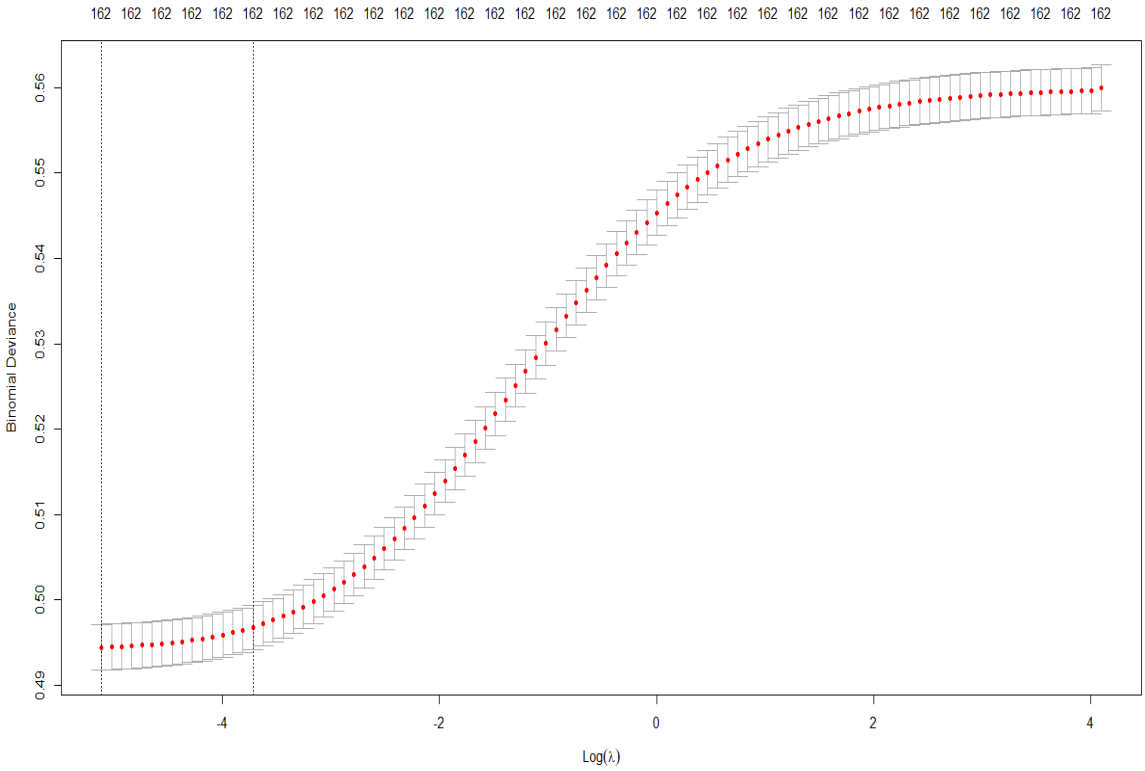


Figura 33. Criteri de selecció per a la lambda òptima que redueix l'error de predicció a partir de la cross-validació. Gràfic amb el model Ridge amb totes les variables. El logaritme de la lambda òptima (0,006024) és -5,11.

La lambda òptima, és de nou, la més petita i és la que es troba més a l'esquerra. És a partir d'aquesta amb la que s'elaborarà el model Ridge considerant totes les variables. El mateix procediment s'aplica amb el model que conté les variables significatives.

5.4 Construcció dels models II: Modelització alternativa

La capacitat d'utilitzar els algoritmes de ML en qualsevol ordinador i moment, és sens dubte, un dels avenços més importants d'aquest segle en el camp de l'estadística. Es tracten d'algoritmes amb els quals s'obté, normalment, una capacitat predictiva major que la desenvolupada pels models lineals o generalitzats. De totes maneres, és important tenir cura de com es creen perquè hi ha factors, com

ara el sobreentrenament, que poden reduir considerablement la capacitat predictiva del model.

El ventall d'algoritmes disponibles per utilitzar és molt ampli, però en el projecte es veuran els que s'estudien al *paper*, exceptuant les xarxes neuronals i els *Ensemble Methods*. S'ha afegit com a model d'estudi el *Support Vector Machine* (SVM). Així doncs, els models de Machine Learning que es construiran són els arbres de decisió, el *Random Forest*, el SVM i el *XGBoost*. Dins del SVM se'n veuran dos tipus: el lineal i el no lineal i el mateix s'aplica per al *XGBoost*, on s'utilitzaran dos mètodes d'estimació dels coeficients.

Per a aconseguir el model per cadascun dels anteriors, s'ha seguit un procediment on s'ha tingut en compte l'índex de gini i l'AUC. En molts casos, quan es tracten models de ML, es fan proves fins trobar els paràmetres òptims que permeten aconseguir el millor model i aquesta és la metodologia que s'ha seguit. Abans de començar a modelitzar, s'han estandarditzat totes les variables per a vetllar pel bon funcionament dels algoritmes. Llavors, s'han construït diversos models amb diferents hiperparàmetres i s'han calculat, amb les dades de test, les mètriques esmentades anteriorment i s'ha seleccionat el que maximitzi gini o l'AUC. Si s'agafessin les dades d'entrenament per a calcular les mètriques, es podria ocórrer a resultats incorrectes. Això seria degut a que els models de Machine Learning s'aprenen les dades de *train* i queden ajustats sota unes característiques molt concretes reflectint en una alta capacitat predictiva aparent, però no real. Llavors, al realitzar la classificació de noves dades diferents a les d'entrenament, es troba en una situació molt diferent i pitjor al que s'havia dit inicialment degut a que la capacitat predictiva del model és, realment, dolenta.

5.4.1 Arbres de decisió

Els arbres de decisió són models predictius formats per regles binàries (Sí/No o Compleix / No compleix) que permet repartir les observacions en funció dels seus atributs i d'aquesta manera predir la categoria de la variable resposta. Es tracta d'un mètode que sorgeix a partir de l'idea d'un arbre que es ramifica. Les branques de l'arbre permeten distingir individus en funció de les variables més importants que detecta el model per la classificació.

L'algoritme disposa d'una gran flexibilitat en termes de paràmetres. La més important, però, és el nombre de branques que es volen perquè permet establir un nombre màxim de regles binàries. En el supòsit que no s'inclogués un màxim de

branques i s'utilitzessin les dades d'entrenament, l'estructura i condicions de l'arbre funcionaria únicament per a aquestes dades i no unes altres. Això és el que provocaria el sobreentrenament: l'algoritme no té capacitat predictiva per a altres conjunts de dades. Per exemple, i avançant-se a les proves de l'algoritme, es troba una situació amb 30 branques en que ja es detecta un sobreentrenament elevat. S'observa com al primer gràfic, d'entrenament, classifica amb impagament gairebé a tots els individus que veritablement van presentar un incompliment (groc) mentre que per a test, el segon gràfic, l'algoritme no és capaç de classificar els individus.

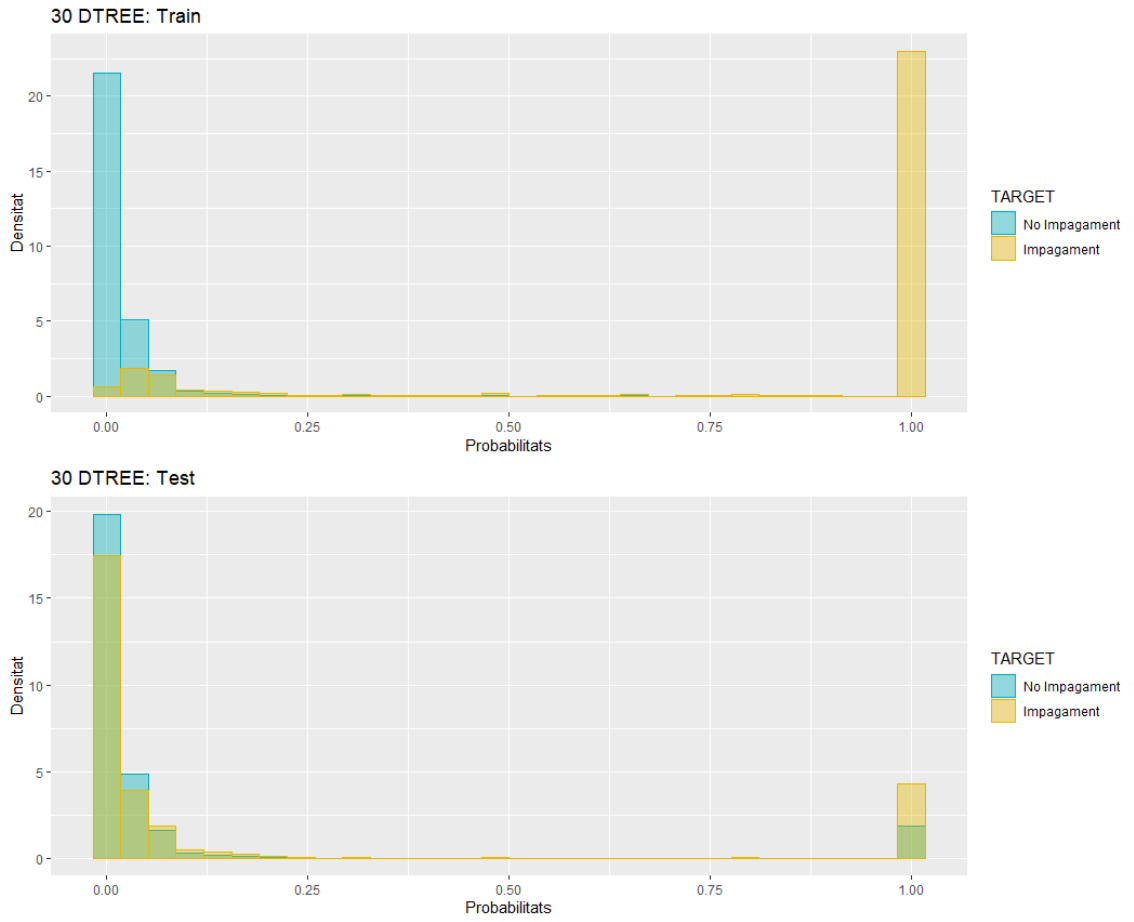


Figura 34. Exemple de sobreentrenament de l'algoritme amb 30 branques màximes.

El gràfic anterior mostra clarament els perills del sobreentrenament. Per això es proposa estimar múltiples models amb un nombre de branques molt dispars entre ells i calcular-ne les mètriques corresponents a partir de l'estimació de les probabilitats del model. Exemplificant-ho, i per aquest conjunt de dades, es podria començar amb 5, llavors, 15, 30 i 70 branques i analitzar el seu rendiment. Posteriorment es creen models amb un nombre de branques similars als models amb millor rendiment per a trobar el que més s'adeqüi correctament a les dades sempre tenint en compte gini i l'AUC. Addicionalment, es pot veure el grau de diferenciació que aconsegueix cada

model a partir de la gràfica de la seva distribució de probabilitat amb les dades de test, tal i com es mostra en la figura 35.

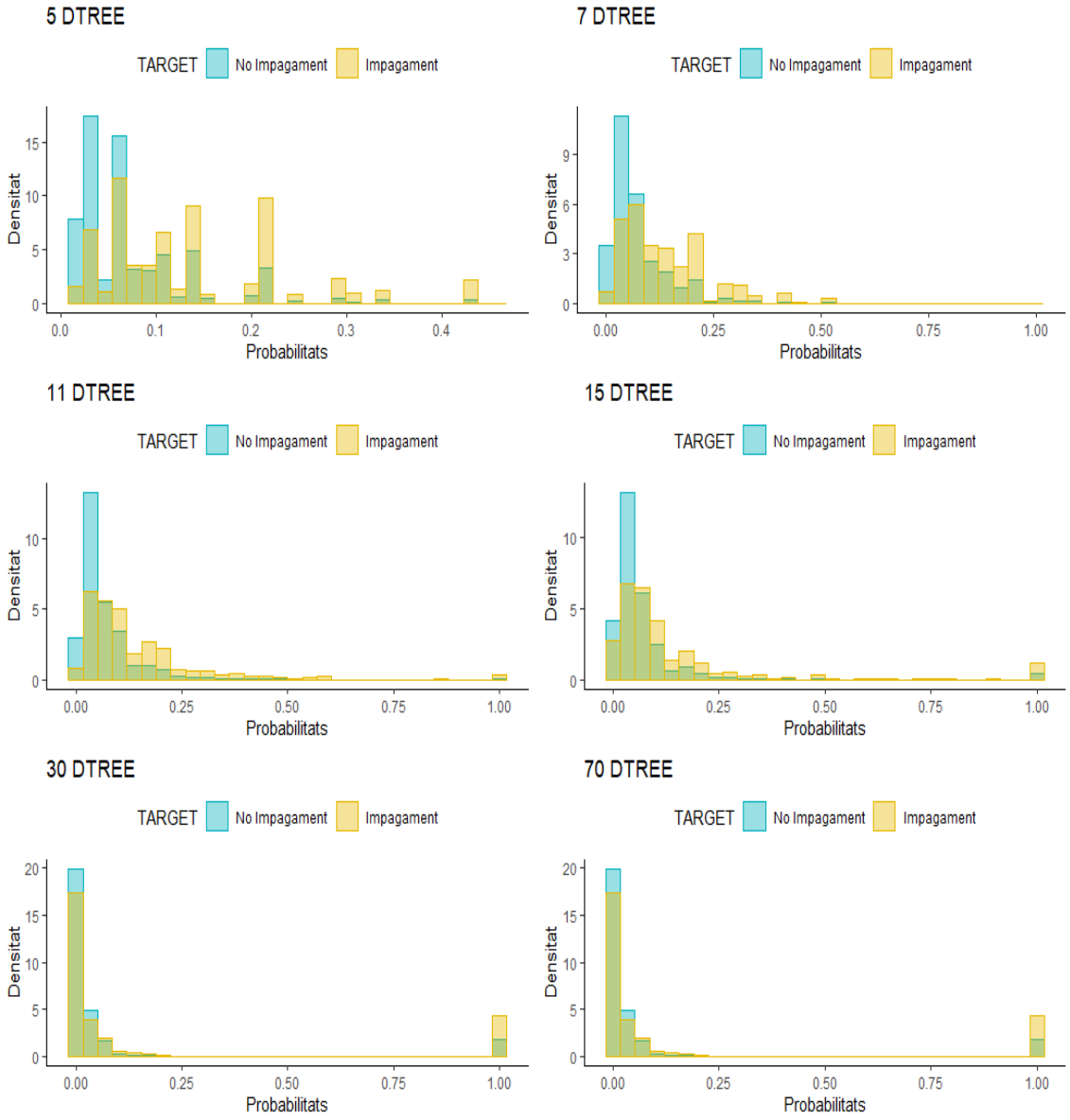


Figura 35. Distribució de probabilitat segmentant entre impagament i no impagament per a múltiples arbres amb diferent nombre de branques.

S'ha vist que el model amb 30 branques presenta sobreestimació, però tot indica que aquest problema comença a generar-se a partir de les 15 branques. Alhora, tot fa indicar que el nombre de ramificacions òptim es troba entre 5 i 11. Es pot considerar aquest interval perquè a l'arbre amb cinc branques predomina el color groc a la part dreta del gràfic i, per al model amb onze branques, comencen a estimar-se probabilitats molt elevades per a individus que realment no van incomplir.

Per a entrar en més detall, s'han calculat les mètriques per als models que apareixen en la figura 35 a partir de les dades de testatge. També s'han inclòs altres models que podrien ser d'interès. D'aquesta manera es podrà determinar de forma precisa el nombre de branques a considerar.

Nombre de branques	Gini	AUC
5	0,4459	0,7230
6	0,4484	0,7242
7	0,4547	0,7274
8	0,4536	0,7268
11	0,4314	0,7157
13	0,3932	0,6966
15	0,3345	0,6672
20	0,1951	0,5975
30	0,046	0,5231
50	0,0869	0,5434
70	0,0908	0,5454

Taula 23. Índex de gini i AUC per als diferents arbres de decisió.
En verd es troba el millor model amb nombre de branques òptim.

Segons els resultats de la taula s'han de considerar 7 branques per a treballar amb el millor arbre. L'estimació de les probabilitats per a les dades de test presenten un índex de gini de 0,4547 i al comparar-les amb el valor real s'obtindria un AUC de 0,7274. De fet, a la taula s'observa com tant el gini i l'AUC augmenten lentament fins a les 7 ramificacions per, posteriorment, canviar de tendència i desplomar-se assolint un mínim en les 30 branques.

5.4.2 Random Forest

Es tracta d'un algoritme que segueix amb la mateixa línia dels arbres de decisió però amb un pas més enllà. Com bé indica el seu nom, els Random Forest podrien definir-se com un conjunt d'arbres de decisió, així definint-se com un bosc. S'estimen múltiples arbres i es classifiquen els individus per cadascun dels arbres. D'aquesta manera, la categoria més freqüent dintre de les estimades per l'observació i-èssima serà la que realment estimi l'algoritme de Random Forest. Amb altres paraules, es fa una consulta a diferents arbres per a que diguin la seva opinió sobre quina possible categoria li pertocaria al client i, un cop recol·lectades, s'escull la més repetida. Així doncs, es necessiten dos paràmetres: el nombre d'arbres a estimar i l'extensió de l'arbre. Al poder generar un bosc d'arbres, i així comparar entre els diferents opinions permet, generalment, l'obtenció de millors resultats que agafant-ne un de sol.

De totes maneres, al basar-se plenament amb els arbres, ocorre als mateixos inconvenients i problemes que aquests. Malgrat el desenvolupament de varis arbres i el contrast d'opinions, la sobreestimació segueix essent un risc per a la construcció del Random Forest. Un nombre de branques elevat comportarà, de nou, a un model amb una baixa capacitat predictiva tal i com mostra la figura 36. En aquests gràfics, extrets a partir de les dades de *Home Credit*, es mostren diferents realitats que es poden trobar quan es modelitzen algoritmes de Machine Learning. Per a saber amb quin algoritme es tracta, els paràmetres esmentats anteriorment apareixen al títol de cada gràfica: primer el nombre d'arbres estimats seguits del nombre de branques. S'observa com a mesura que augmenten les ramificacions, augmenta també la capacitat predictiva en el conjunt d'entrenament, especialment quan hi ha 60 branques, mentre que no sembla que existeixi una evolució significativa en la de testatge. Si es para atenció als primers valors de l'eix de les abscisses de les distribucions de probabilitat en el subgrup groc d'impagament, es veurà com aquestes augmenten al mateix temps que ho fan les ramificacions. És a dir, a mesura que s'incorporen noves branques, augmenta la quantitat de clients amb una probabilitat baixa d'impagament quan en realitat aquesta hauria de ser elevada. En definitiva, s'està sobreestimant.

Pel motiu anterior i tal com s'ha fet pels arbres, és necessari trobar el nombre de ramificacions òptimes. Per aconseguir-ho es seguirà la mateixa metodologia emprada.

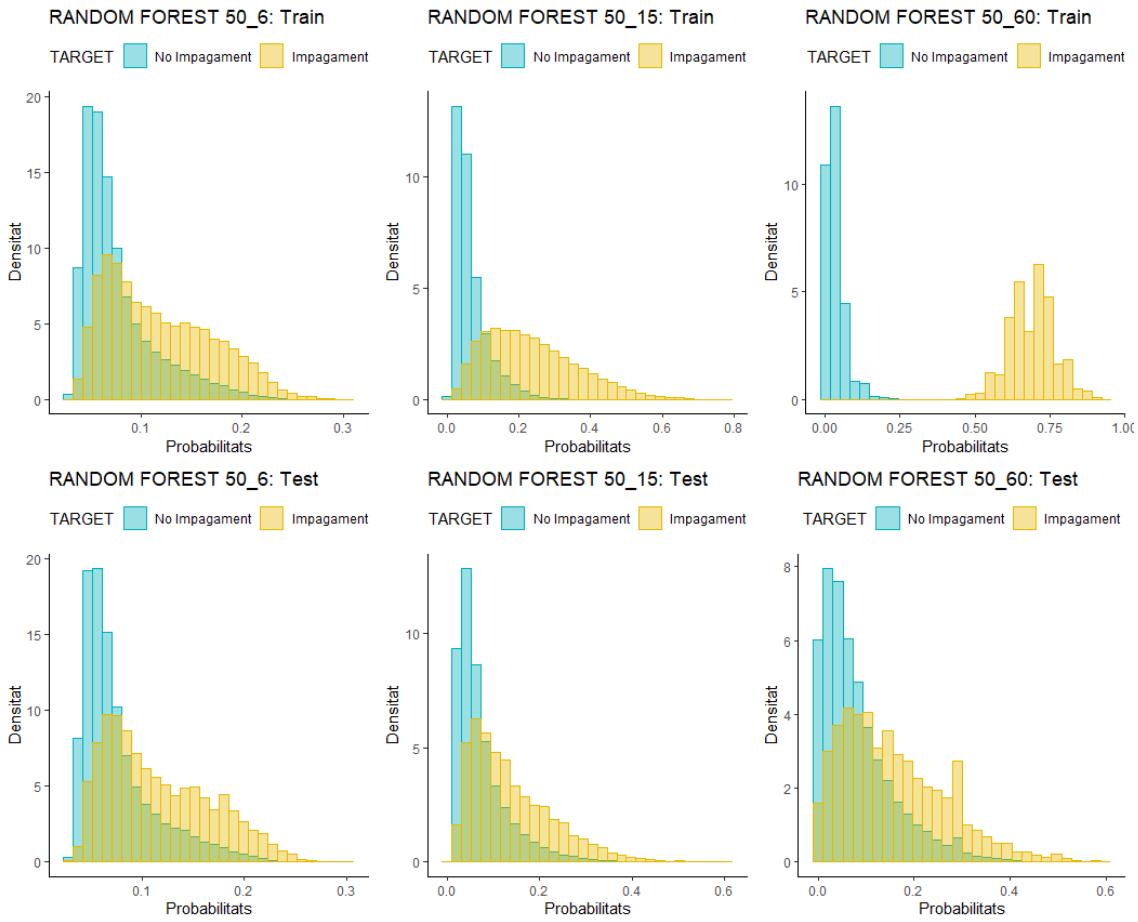


Figura 36. Evolució de les probabilitats d'impagament per a les dades d'entrenament i testatge en funció dels diferents hiperparàmetres del Random Forest.

L'existència de dos hiperparàmetres pot comportar un retard en la troballa del millor model degut a les nombres combinacions possibles. S'ha de trobar el nombre de ramificacions i el d'arbres a estimar. El més important, però, és el determinar el primer ja que pot comportar la tant esmentada sobreestimació si presenta moltes branques o que el model acabi disposant d'informació insuficient degut a una ramificació insuficient. Per determinar el nombre de branques òptimes s'han construir múltiples models estimant 50 arbres canviant l'hiperparàmetre d'interès. A continuació es mostra la taula amb els resultats obtinguts:

Nombre d'arbres estimats	Nombre de branques	Gini	AUC
50	6	0,4926	0,7463
50	10	0,5053	0,7527
50	11	0,5079	0,7540
50	12	0,5100	0,7550
50	13	0,5087	0,7544

50	15	0,5027	0,7513
50	20	0,4910	0,7455
50	25	0,4648	0,7324
50	30	0,4520	0,7260
50	35	0,4562	0,7281
50	40	0,4515	0,7257
50	60	0,4430	0,7213

Taula 24. Índex de gini i AUC per als diferents Random Forest.
En verd es troba el millor model amb el nombre de branques òptim.

L'últim pas és determinar el nombre d'arbres a estimar. Com s'ha argumentat anteriorment, si s'escull un nombre elevat permetrà contrastar diferents opinions d'una forma més correcta. Imagina una situació en que has de decidir entre dues peces de roba totalment diferents. En molts cops demanaràs consell a les persones del teu entorn per a donar resposta a la indecisió, i quantes més siguin, més segur estaràs que has comprat la peça correcta. Amb el nombre d'arbres succeeix el mateix. Considerar un nombre elevat pot ajudar a millorar la capacitat predictiva dels Random Forest.

Nombre d'arbres estimats	Nombre de branques	GINI	AUC
35	12	0,5075	0,7538
100	12	0,5128	0,7564
500	12	0,5159	0,7579
1000	12	0,5168	0,7585
2000	12	0,5169	0,7585

Taula 25. Índex de gini i AUC per als diferents Random Forest.
En verd es troba el millor model amb el nombre d'arbres estimats òptims; en blau el model escollit.

En efecte, l'augment en el nombre dels diferents arbres millora lleugerament la capacitat predictiva. De totes maneres, la millora reflectida en les mètriques es va reduint progressivament fins que gairebé és inexistent. És el que passa entre els 1.000 i 2.000 arbres: la millora és ínfima. Per això i acord al principi de parsimònia s'escollirà el model amb 1.000 arbres i 12 branques per ser el més simple.

5.4.3 Support Vector Machine

L'algoritme de Support Vector Machine, o conegut per les seves sigles SVM, és un mètode de llenguatge supervisat utilitzat tant en problemes de classificació com de regressió. En aquest cas, es basarà amb la idea de trobar el millor hiperplà que millor separi les dues categories referents a la variable resposta *TARGET*. De forma més senzilla i per a la situació dels impagaments seria trobar la millor recta, si s'utilitza el SVM lineal, o bé una corba que permeti distingir de forma precisa les dues classes.

Els SVM és una de les tècniques de ML més emprades degut, principalment, a les funcions kernel. Aquestes tenen un rol molt important en la classificació i són utilitzades per analitzar patrons en el conjunt de dades. Resulten ésser de molta ajuda quan es tracta d'un problema no lineal. En una fase avançada de l'algoritme es transforma l'espai del núvol de punts original a un altre amb major dimensionalitat. D'aquesta manera permet crear la frontera o recta de decisió òptima. El procediment anterior rep el nom de truc del kernel. Malgrat la seva popularitat, es tracta d'un algoritme poc eficient per a mostres amb una gran quantitat de registres que necessita de molt temps d'execució.

La gran varietat de kernels ofereix moltes possibilitats de modelització dels SVM. No obstant, Es plantegen un total de tres models diferents: dos amb el kernel lineal, però un amb totes les variables i l'altre amb només les significatives i un últim model que presenti un kernel no lineal. Hi ha múltiples kernels no lineals que podrien utilitzar-se però en aquest projecte s'estudiarà només el RBF (*Radial Basis Function*),un dels més emprats en el SVM, amb les variables significatives per a reduir el temps de computació. Els resultats d'aquest model apareixen explícits en el punt 5.5. Respecte als models que incorporen el kernel lineal, s'obtenen els següents resultats.

Model	Dades	Gini	AUC
Lineal_TV	Train	0.5145	0.7573
Lineal_TV	Test	0.5159	0.7580
Lineal_MV	Train	0.5171	0.7585
Lineal_MV	Test	0.5162	0.7581

Taula 26. Índex de gini i AUC per als Support Vector Machine lineals. Lineal_TV fa referència al model amb totes les variables i MV amb les significatives. En verd es troba el millor models segons les mètriques d'interès.

Els valors de les mètriques de l'índex de gini i AUC suggereixen que el millor SVM amb kernel lineal és el que conté únicament les variables significatives. D'aquesta manera, i conjuntament amb el SVM amb kernel RBF, són els models que avancen a la fase comparativa de models.

5.4.4 XGBoost

XGBoost són les sigles de "eXtreme Gradient Boosting" i està revolucionant el sector de Machine Learning degut a les millores substancials en la capacitat predictiva respecte altres algoritmes. Com el Random Forest, el XGBoost implementa arbres de decisió amb la metodologia de *Gradient boosting* per a minimitzar el temps d'execució i maximitzar-ne el rendiment. Tot i que aquests dos algoritmes puguin semblar molt semblants, en veritat, no ho són. El primer utilitza arbres complets (biaix petit, variància elevada) formats per conjunts de variables independents. D'altra banda, el XGBoost construeix arbres de manera seqüencial (biaix elevat i variància petita), on cada arbre es creat tenint en compte el marge d'error que deixen les variables pitjor classificades per l'anterior amb l'objectiu de minimitzar el biaix. Amb altres paraules, el *Boosting* es basa en entrenar els classificadors dèbils múltiples vegades utilitzant diferents conjunts d'entrenament per a efectuar proves amb diferents distribucions o pesos sobre les dades. Una altra diferència destacable és que el XGBoost no és tant vulnerable al sobreentrenament com el Random Forest.

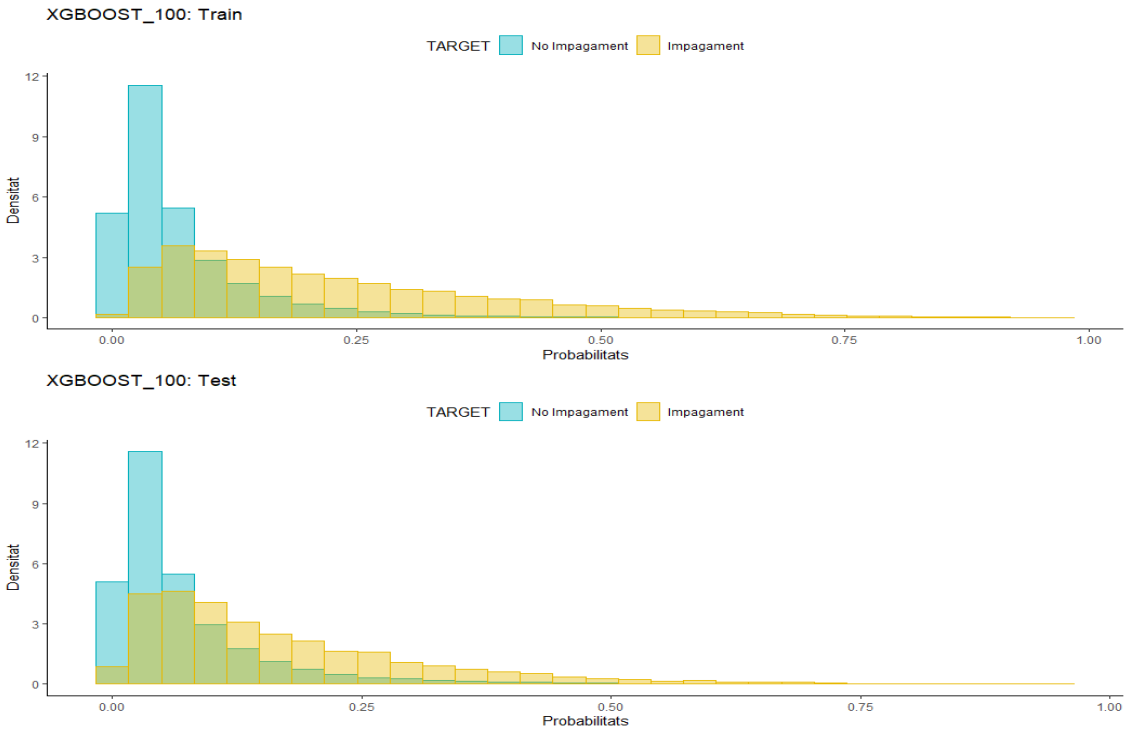


Figura 37. Distribució de les probabilitats estimades per un XGBoost amb 100 estimacions.

Això implica que no s'hagi de determinar el nombre de branques per aconseguir una bona capacitat predictiva. Llavors, el nombre d'hiperparàmetres a considerar és només un: la quantitat d'arbres del model XGBoost. Al trobar-se en un context diferent al *Random Forest*, es podria proposar un model amb un gran nombre d'arbres tal i com s'ha fet en el seu apartat. Aquest plantejament seria erroni degut a que els arbres que es creen al *Random Forest* són independents entre ells i es comparen al final del procés, però els arbres del XGBoost es construeixen a partir de l'anterior. Al ser dependents un de l'altre pot comportar a que, si s'estimen molts arbres, puguin aportar més soroll que beneficis.

Una de les maneres per trobar el nombre d'arbres a estimar, i la que es farà servir, és calcular la pèrdua logarítmica dels models amb ramificacions diferents. Es tracta d'una mesura de rendiment sobre el model de classificació. L'ideal seria escollir aquell nombre d'arbres que presenti la pèrdua més propera a 0. Si la mesura valgués 0 s'estaria parlant d'un model perfecte.

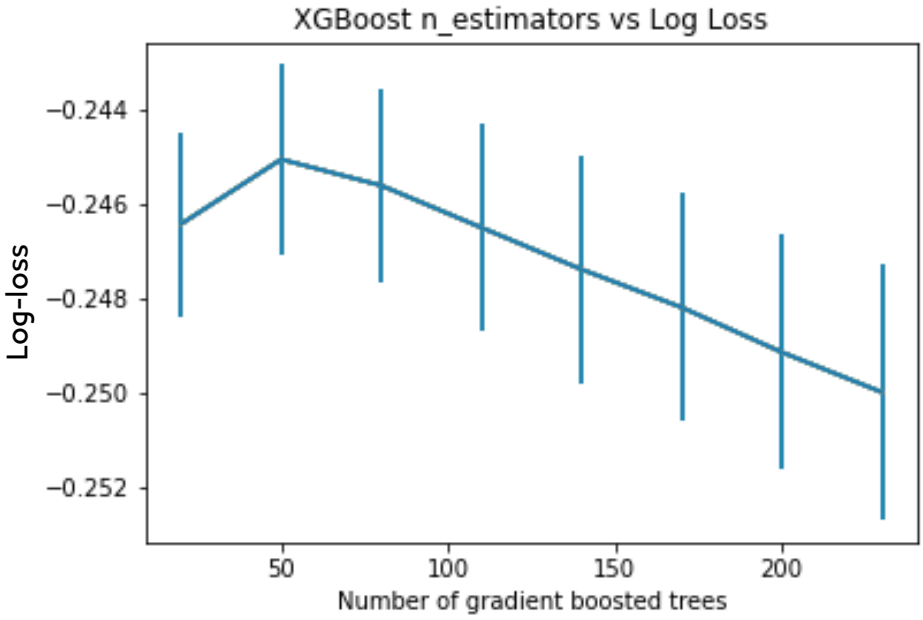


Figura 38. Evolució de les pèrdues logarítmiques segons el nombre d'arbres estimats. Inclou també els intervals de confiança de la pèrdua per cada model

El model que presenta la pèrdua logarítmica més pròxima a 0 és el de cinquanta arbres estimats. Un cop superat aquest model, disminueix la qualitat del model a mesura que augmenten els arbres a estimar. La figura 38 s'ha fet considerant totes les variables i, per verificar la gràfica, s'han quantificat també l'índex de gini i l'AUC de múltiples models en la taula 27. Com s'esperava, el model amb arbres propers a 50 presenten bons resultats.

Arbres estimats	Gini	AUC
35	0,5327	0,7663
45	0,5224	0,7612
50	0,5342	0,7671
55	0,5335	0,7668
60	0,5340	0,7670
100	0,5306	0,7653
500	0,4897	0,7448

Taula 27. Índex de gini i AUC per models XGBoost amb diferents arbres.

Es podria plantejar la mateixa situació però considerant les variables significatives del model logístic. A la taula 28 es mostra el rendiment de varis models amb menys variables on, en aquest cas, el model amb un total de 55 presenta millor capacitat predictiva.

Arbres estimats	Gini	AUC
35	0,5299	0,7650
50	0,5308	0,7654
55	0,5311	0,7655
60	0,5310	0,7655

Taula 28. Índex de gini i AUC per models XGBoost amb diferents arbres i només amb les variables significatives.

Si es comparen els dos millors models de les taules 27 i 28 s'observa com el model amb més variables té millor capacitat predictiva que considerant únicament les significatives. D'aquesta manera, el XGBoost és capaç d'extreure més informació de les variables no rellevants diagnosticades pel model logístic. Així doncs, el model XGBoost que es tindrà en compte per a la comparació dels set diferents tipus de models serà el que estima 50 arbres amb totes les variables.

5.5 Quantificació dels beneficis

En els dos apartats previs s'han construït una gran quantitat de models amb diferents condicions i característiques per verificar que s'ha estat escollint el millor. En aquest moment, es tenen sis models derivats de la metodologia clàssica i cinc del Machine Learning. Un dels objectius principals és trobar el model que ofereixi una major capacitat predictiva i llavors comparar-ne el cost tal i com s'exposava en el *paper* per a establir un millor model global. Per fer-ho es poden analitzar l'índex de gini i l'AUC dels models com també els seus intervals de confiança respectius. Alhora, també es poden representar les distribucions de les probabilitats estimades tal i com s'ha fet anteriorment.

No obstant, i per a reduir el nombre de models a treball, es pot analitzar primer quin és el model tradicional amb millor capacitat predictiva. Per poder determinar-lo és necessari estudiar els resultats de les mètriques per al model logístic, amb penalització Lasso i Ridge en funció del nombre de variables incorporades. S'agafarà únicament el que presenti el valor més gran en les mètriques per a les dades de testatge i per tant resulti en millors prediccions per a dades alienes a les d'entrenament. En la taula següent es mostren els resultats on *TV* significa totes les variables i *MV* només les significatives.

Dades	Model	Gini	AUC
Train_TV	Logístic	0,5151	0,7576
	Lasso	0,5138	0,7569
	Ridge	0,5128	0,7564
Test_TV	Logístic	0,5147	0,7574
	Lasso	0,5156	0,7578
	Ridge	0,5148	0,7574
Train_MV	Logístic	0,5141	0,7571
	Lasso	0,5146	0,7573
	Ridge	0,5133	0,7567
Test_MV	Logístic	0,5153	0,7577
	Lasso	0,5166	0,7583
	Ridge	0,5156	0,7578

Taula 29. Índex de gini i AUC per als diferents models tradicionals. En verd es mostra el millor model.

La diferència entre els models és pràcticament nul·la ja que les mètriques prenen valors molt similars. Així doncs, qualsevol dels sis podria ser un bon candidat per a representar el grup tradicional però s'agafarà el model Lasso amb menys variables com el millor perquè és el que maximitza tant gini com l'àrea sota la corba considerant tots els casos i, especialment, per ser el millor amb les dades de test. Per aquestes dades i amb menys variables, el Lasso presenta una millora del 0,25% respecte el logístic.

Dels onze models inicials queden finalment sis per determinar quin és el que funciona millor. A partir d'aquí només es parlarà del model com a tal sense importar-ne els seus hiperparàmetres.

Tipus de model	Model
Tradicional	Lasso_MV
Alternatiu	Arbres de decisió
Alternatiu	Random Forest
Alternatiu	SVM lineal
Alternatiu	SVM RBF
Alternatiu	XGBoost

Taula 30. Llistat dels millors models.

La manera més senzilla de trobar el model que permeti classificar millor entre els individus que incompliran o no és a partir de l'índex de gini i l'AUC obtinguts per a les dades de test. L'algoritme amb millor capacitat predictiva segons els indicadors és el XGBoost, seguit del Random Forest i el model Lasso.

Model	Gini	AUC
Logístic	0,5153	0,7577
Lasso	0,5166	0,7583
Arbres	0,4547	0,7274
Random Forest	0,5169	0,7585
SVM Lineal	0,5162	0,7581
SVM RBF	0,4704	0,7352
XGBoost	0,5342	0,7671

Taula 31. Bondat d'ajust per a test segons els diferents models.

Quatre dels algoritmes anteriors proporcionen resultats lleugerament millor que el logístic. Si es compara amb aquest, utilitzant el model XGBoost s'obté una millora del 3,67% i del 0,31% si es tracta del Random Forest en termes de gini. Si fos l'AUC, els guanys serien de l'1,24% i del 0,11%. En canvi, els arbres i el Support Vector Machine amb kernel no lineal són els que presenten un rendiment notablement inferior al logístic si bé presenten una pèrdua de l'índex de gini de l'11,76% i del 8,71% respectivament.

Addicionalment, també es poden comparar els intervals de confiança de les mètriques. Utilitzar intervals ens permetria quantificar la franja de capacitat predictiva que s'esperaria si s'incorporessin noves dades amb una probabilitat del 95%.

Model		Gini		AUC	
		$IC_{0.025}$	$IC_{0.975}$	$IC_{0.025}$	$IC_{0.975}$
Logístic	Train	0,5044	0,5226	0,7522	0,7613
Logístic	Test	0,5040	0,5236	0,7520	0,7618
Lasso	Train	0,5030	0,5260	0,7515	0,7630
Lasso	Test	0,5062	0,5246	0,7531	0,7623
Arbres	Train	0,4633	0,4871	0,7316	0,7436
Arbres	Test	0,4333	0,4744	0,7166	0,7372
RF	Train	0,7010	0,7189	0,8505	0,8595
RF	Test	0,4953	0,5379	0,7476	0,7690
SVM_L	Train	0,4904	0,5418	0,7452	0,7709
SVM_L	Test	0,4974	0,5351	0,7487	0,7675
SVM_RBF	Train	0,7846	0,8152	0,8923	0,9076
SVM_RBF	Test	0,4457	0,4898	0,7230	0,7450
XGBoost	Train	0,6323	0,6551	0,8161	0,8275
XGBoost	Test	0,5121	0,5490	0,7561	0,7745

Taula 32. Intervals de confiança per a les mètriques de cada model.

Com era d'esperar, l'algoritme XGBoost és el que més destaca respecte les dades de testatge. Tot i això, és interessant veure com el SVM amb el kernel RBF presenta una franja molt bona amb les dades d'entrenament. La no linealitat ha estat construïda per explicar el conjunt de *train* però no per la de testatge que obté valors molt inferiors. Per ordre i, respecte l'AUC, s'obtindrien millores potencials de l'1,67%, 0,95% i 0,70% per al XGBoost, Random Forest i SVM lineal respecte el logístic.

Les probabilitats estimades són també un altre aspecte clau a comparar. Un bon model seria capaç de donar una probabilitat per sobre la mitjana a aquells individus que proporcionessin un risc considerable a incomplir i viceversa. És important, doncs, que puguin discernir entre qualsevol tipus de prestataris. Així són les distribucions de les probabilitats estimades a partir de les dades de test dels models amb els quals es treballa.

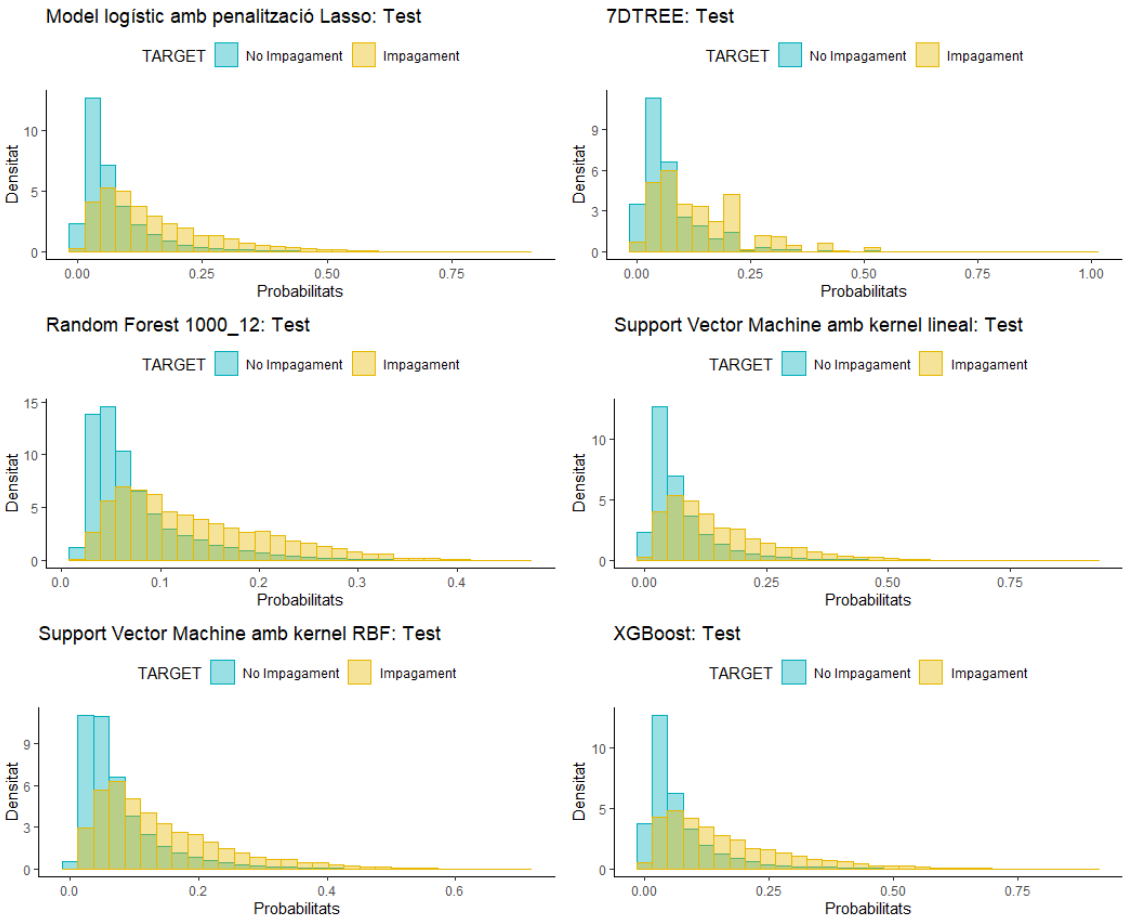


Figura 39. Distribució de les probabilitats estimades per cada model.

Tots els models segueixen el mateix patró: al principi, s'acumulen un gran nombre de clients sans i, a mesura que augmenta la probabilitat, aquests clients disminueixen ràpidament. Per altra banda, una vegada la proporció de morosos és superior a la de clients sans, el nombre de morosos comença a disminuir però sempre mantenint-se la mateixa estructura en les proporcions.

També, però, tenen una diferència destacable: el rang de les probabilitats no és el mateix, si bé per al Lasso, SVM lineal i XGBoost es mouen entre 0 i un nombre inferior a 1 i el valor màxim del Random Forest no assoleix el 0,5.

Una altra forma de veure la capacitat predictiva de cada model és calcular el percentatge d'impagaments reals en múltiples intervals creats a partir de les probabilitats estimades. És a dir, es poden dividir les estimacions en *n* grups on cadascun incorpora totes aquells clients amb una probabilitat estimada compresa entre dos valors *x* i *y*. Llavors, si els models s'han construït correctament, com més gran fossin els valors estimats que comprenen l'interval major seria el percentatge d'impagaments reals.

S'ha plantejat la creació de nou intervals diferents: [0; 0,04], (0,04; 0,08], (0,08; 0,12], (0,12; 0.16], (0,16; 0,20], (0,20; 0,30], (0,30; 0,50], (0,50; 0,75], (0,75; 1]. A la figura 40 es mostra l'evolució del percentatge de morosos detectats segons la mida de cada interval de les probabilitats estimades per cada model. En l'eix de les abscisses es mostra únicament el primer valor de l'interval; si al 0,3 un model pren per exemple 25%, vol dir que el 25% de les observacions compreses en l'interval (0,3; 0,5] són morosos.

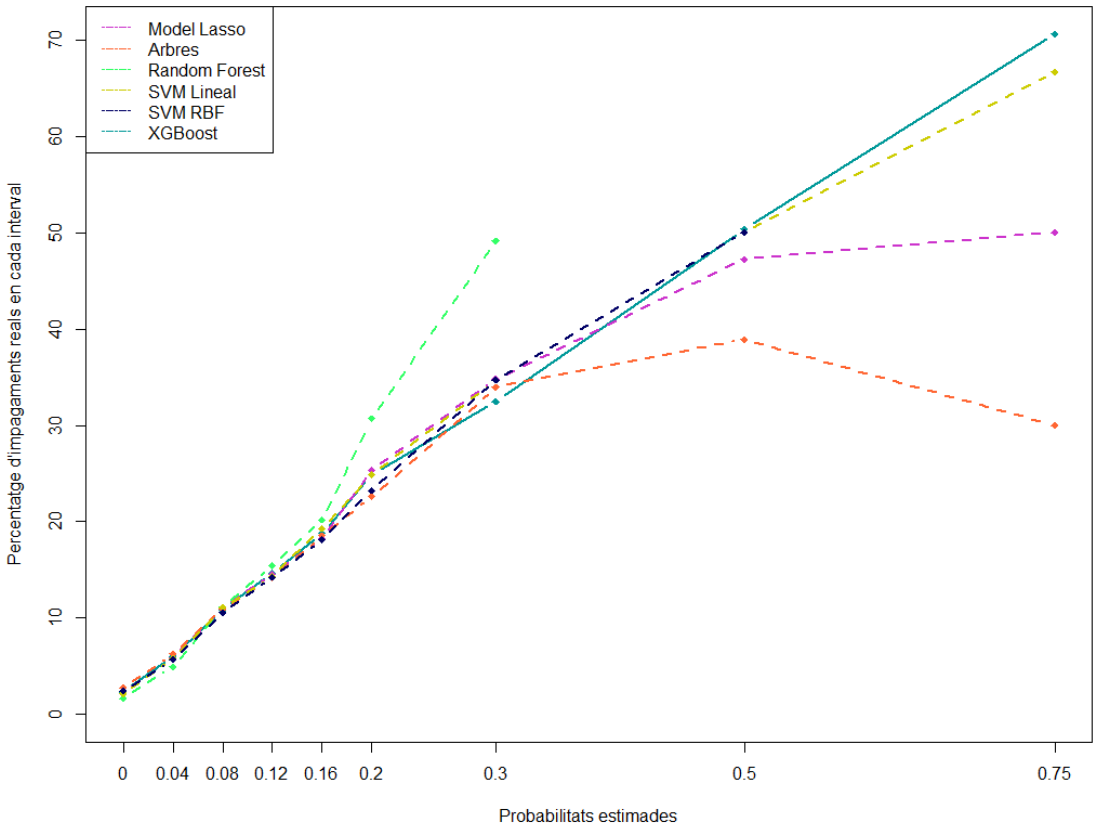


Figura 40. Evolució del percentatge de morosos reals segons els intervals de les probabilitats estimades.

Es mostra com al principi tots els models tenen un comportament molt similar. Tot i això, el Random Forest és el primer en desmarcar-se del grup ja que a partir de l'interval (0,16; 0,20] pren una pendent totalment diferent a la resta però és també

el que el que presenta una probabilitat estimada màxima més petita. Si s'observen els últims dos intervals que fan referència a individus que el model considera d'alt risc d'incompliment, el model XGBoost és el que té un encert major. El 70% de les observacions dins de l'interval (0,75; 1] són realment morosos. En canvi, l'arbre de decisió i el model Lasso només detecten aproximadament un 30% i 50% en aquest mateix interval respectivament.

Alternativament, es pot comparar directament amb les estimacions realitzades amb el model logístic. A la figura 41 s'observen els canvis percentuals en el percentatge d'impagaments reals respecte el logístic. Si el percentatge per un interval és positiu, comporta a que el model predigui més impagaments en aquesta franja i, si és negatiu, en predigui menys.

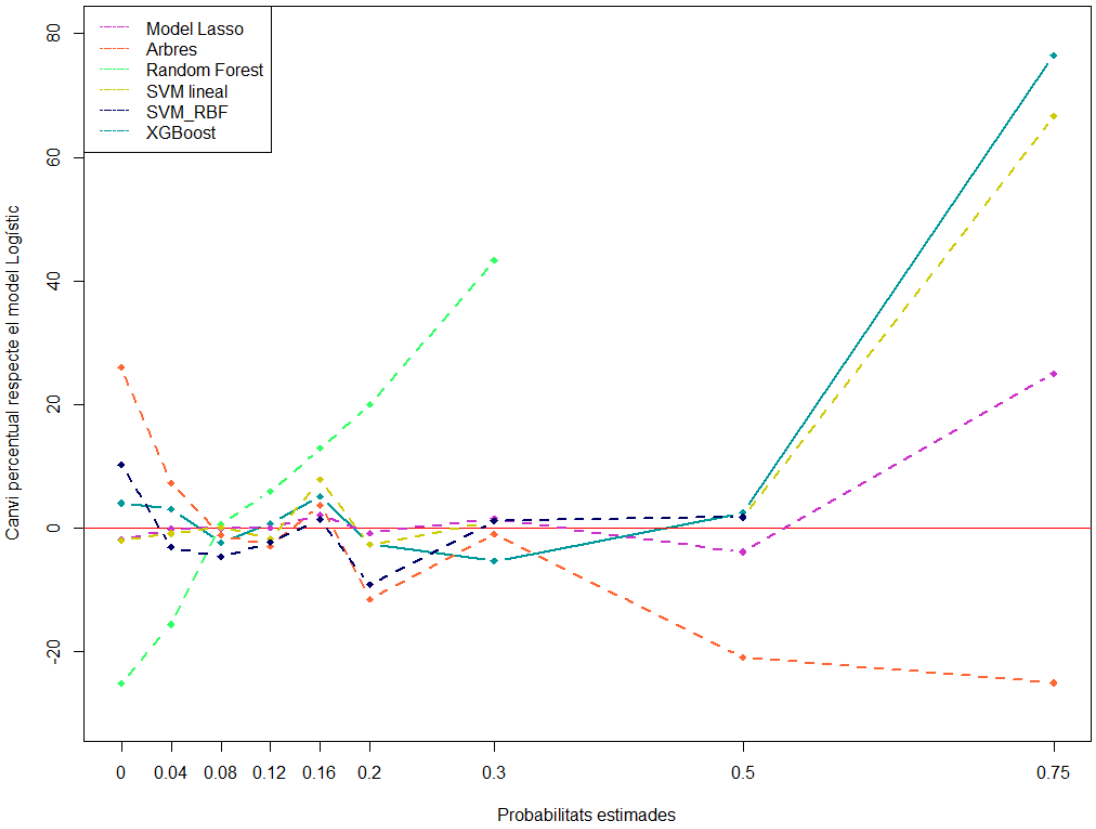


Figura 41. Diferències percentuals en el percentatge d'impagaments reals respecte el model logístic.

A simple vista l'algoritme que més destaca és el Random Forest degut a que prediu molts menys morosos quan la probabilitat estimada és menor a la mitjana i, un cop passat aquest valor, en detecta molt més. La resta es mouen al voltant del 0, si bé els arbres de decisió presenten una pendent negativa que el permeten definir, d'aquesta manera, com el pitjor model en comparació amb el logístic. El XGBoost,

SVM lineal i el model logístic amb penalització Lasso acaben detectant molts més impagaments en l'interval final que el logístic.

5.6 Quantificació dels costos i resultats

Un anàlisi exhaustiu no només analitza els beneficis dels models sinó que també s'han d'esmentar els seus inconvenients. La tria del millor s'ha de fer basant-se amb els avantatges que aporten, però també els seus problemes. Ara bé, un és capaç d'avaluar-ne els inconvenients i solucionar-los a partir del coneixement i experiència prèvia en l'àmbit de modelització especialment en els de ML.

En la secció *4.4 Funció de cost* s'havia visualitzat una taula que quantificava el cost per cadascun dels models segons un conjunt de factors de risc. Aquella taula permet al supervisor actualitzar els percentatges acord a les funcionalitats que es desitgen per al model en qüestió. Aquí, a l'estar analitzant els models en funció dels resultats presos, s'han modificat lleugerament alguns valors dels factors com també els pesos d'aquests. S'ha donat màxima importància al sobreentrenament i la interpretabilitat del model que és tal i com es mostra en la taula 33.

	%	*Logístic*	*Lasso*	*Arbres*	*RF*	*SVM*	*SVM_RBF*	*XGBoost*
Estabilitat	10	1	1	3	2	1	1	2
Nombre de paràmetres	5	1	1	2	3	2	2	4
Sobreentrenament	15	1	1	3	2	3	3	3
Enginyeria de dades	0	1	-	-	-	-	-	-
Calibratge dinàmic	5	-	-	-	-	-	-	-
Transparència	5	1	1	1	1	1	1	1
Empremta de carboni	10	1	1,5	1	2	3	5	2
Dependència externa	5	1	1	1	1	1	1	1
Atacs cibernètics	0	1	1	1	1	1	1	1
Privacitat	5	1	1	1	3	2	2	3
Replicació	10	1	1	1	3	2	3	4
Interpretabilitat	20	1	1	1	2	2	2	2
Biaixos	10	1	1	3	4	3	3	4
Suma / Cost	100	0,95	1	1,7	2,2	2,05	2,35	2,5

Taula 33. Cost del model per cadascun dels inconvenients.

S'ha donat major pes a aquests factors perquè es tracten de dificultats que el supervisor haurà de controlar. Especialment la interpretabilitat ja que és en el sector del risc creditici on és més important quan el prestatari demana per què se l'hi ha denegat o acceptat el crèdit. El que no es podria fer és dir-li: "*L'algoritme amb el qual es treballa ha denegat la seva sol·licitud del préstec degut a les seves característiques X. Ho sentim molt*". És necessari explicitar de forma clara quines són les característiques X. Hi ha models on la interpretabilitat és senzilla com és el cas dels models tradicionals però es complica quan augmenta la complexitat algorítmica.

Els models amb menor cost resulten ser els tradicionals mentre que, a mesura que augmenta la complexitat algorítmica, també ho fa el cost. De totes maneres, un augment del cost no comporta automàticament que sigui descartable sinó que fa més difícil la construcció del model. Per això es necessiten experts que dominin tots els aspectes relacionats amb els algoritmes de modelització.

Es poden representar la relació entre els beneficis i els costos a partir de les seves estimacions. Això és el que apareix a la figura 42. Cada punt fa referència a un model i per ordre són el Lasso, arbres de decisió, SVM amb kernel lineal, Random Forest, SVM amb kernel RBF i el XGBoost.

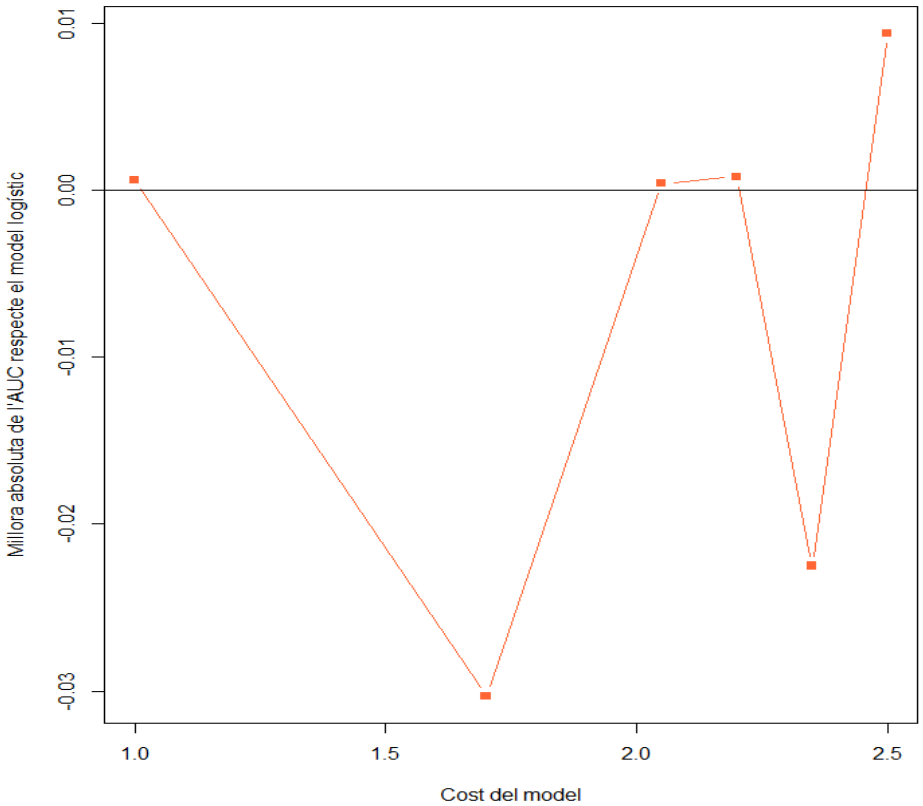


Figura 42. Relació beneficis i costos respecte el model Logístic.

Tal i com s'havia demostrat anteriorment, els arbres i el SVM amb kernel no lineal proporcionen resultats pitjors i amb un cost superior al model logístic per aquestes dades. Llavors, el Lasso, SVM lineal i Random Forest no ofereixen una millora significativa i que, a més a més, aquests dos últims presenten un cost molt més elevat. Per una altra banda, el XGBoost proporciona el cost més elevat de tots però també és el que aporta uns majors beneficis a nivell de capacitat predictiva. D'aquesta manera i assumint que es tenen els coneixements necessaris per solucionar els inconvenients que planteja l'algoritme XGBoost és, sens dubte, el millor model amb el qual es podria treballar. Això és el que pot observar-se en la taula següent on es mostra el rati entre benefici de l'índex de gini respecte el model logístic i cost de cada model.

Model	Rati benefici i cost
Logístic	0
Lasso	+0,25
Arbres	-6,92
Random Forest	+0,14
SVM Lineal	+0,09
SVM RBF	-3,71
XGBoost	+1,47

Taula 34. Rati entre beneficis tenint en compte gini i el cost de cada model.

5.7 Aplicació de Interpretative Machine Learning (IML)

La tria de l'algoritme definitiu és un aspecte fonamental per a determinar quina metodologia emprar per a interpretar els factors de risc del prestatari. Si s'escollís el logístic, per exemple, es podrien utilitzar les *Odds Ratio* (OR) com a mesura per quantificar el risc de presentar una característica respecte d'una altra. En canvi, els models que formen part del col·lectiu de ML no permeten el càlcul de mesures semblants a les OR. Per aquest motiu s'estan produint avenços en el camp de la interpretació per a tots els models de Machine Learning. És el que ja s'ha anomenat com a IML, Interpretative Machine Learning.

Amb aquesta nova metodologia, aplicada per a l'algoritme XGBoost resultant, es pot extreure els pesos de les variables que considera més rellevants per a la classificació.

És el que apareix a la figura 43. També permet veure quins factors fan augmentar o disminuir la probabilitat estimada d'impagament per a l'individu.

Weight	Feature
0.0848	MEAN_EXT_SOURCE
0.0261	FLAG_OWN_CAR_N
0.0180	NAME_EDUCATION_TYPE_Higher education
0.0178	LAST_YEARS_EMPLOYED_10-80
0.0157	LAST_YEARS_EMPLOYED_0-3
0.0124	MORTGAGE_LOAN_Altres_EF
0.0120	UrgentN_PURP_HC
0.0116	NAME_INCOME_TYPE_Working
0.0104	NOT_ENOUGH_MONEY_PAID_IP
0.0101	NAME_EDUCATION_TYPE_Secondary / secondary special
0.0101	Home_CAT_HC
0.0100	Refused_STATUS_HC
0.0099	CODE_GENDER_F
0.0098	SUM_Diferencia_AMT_HC_Less credit than ATB
0.0097	IS_Refreshed_Client_HC

Figura 43. Les 15 variables amb major pes global del XGBoost.

La variable que més destaca per sobre de la resta és la mitjana de les valoracions externes del client. La segueix el valor de no tenir cotxe, tenir una educació elevada o el nombre d'anys treballats prèviament. Altrament, hi ha variables que el seu pes és 0 degut a que aporten gairebé o directament res a la classificació. A continuació es mostren alguna d'elles.

0	BUSINESS_LOAN_Altres_EF
0	ORGANIZATION_TYPE_Mobile
0	ORGANIZATION_TYPE_Services
0	ORGANIZATION_TYPE_Religion
0	CRE_Another_Type_LOAN_Altres_EF
0	TE_CREDIT_Altres_EF
0	CRE_BDE_Altres_EF
0	ORGANIZATION_TYPE_University
0	DIES_MAX_IMPAGATS_Altres_EF__61-90_
0	UNKNOWN_LOAN_Altres_EF

Figura 44. Algunes variables del model XGBoost amb pes nul.

Aquests pesos fan referència al model en si, però també es poden analitzar els pesos de les variables per cadascun dels clients, és a dir, els algoritmes de IML permeten indicar quins són els factors que fan possible prendre una decisió o una altra. Com que s'està en un cas no balancejat on hi ha més presència de clients sans que de morosos, la probabilitat d'impagament estimada es mourà al voltant del percentatge real de la base de dades del 8%. En les sortides del paquet que s'ha fet servir al Python, surt per defecte la categoria estimada (si ha impagat o no), la seva probabilitat, els factors amb major pes per determinar aquesta categoria (en verd) i alguns d'altres que augmentarien la probabilitat de l'altre categoria (en vermell). En la gran majoria de casos, la probabilitat d'impagament estimada és menor a 0,5. Això comporta a que el model classifiqui aquestes observacions com a no impagament amb una probabilitat de *1 - p*. Únicament seria impagament si la probabilitat d'incompliment fos major de 0,5. Per això, es pot considerar persona de

risc que hagi estat estimada com a sana aquelles que tinguin un valor estimat proper o menor a 0,92.

Es veuran tres casos diferents. Primer, un client amb una molt baixa probabilitat d'incompliment. Llavors, un altre en que aquesta probabilitat es troba al voltant del 8% i per últim un prestatari que presenta un alt risc segons el model.

y=0 (probability **0.992**, score **-4.877**) top features

Contribution[?]	Feature
+2.541	<BIAS>
+1.040	MEAN_EXT_SOURCE
+0.399	TOT_INSTALMENTS_CC
+0.356	MAX_EXT_SOURCE
+0.304	AMT_INCOME_TOTAL
+0.220	MAX_CREDIT_LIM_CC
+0.171	AMT_CREDIT_ACT_TOT
+0.127	RATI_DEUTE_GARANTIA
-0.087	TOT_RECEIVABLE_CC
-0.089	NAME_TYPE_SUITE_Family
-0.095	LAST_YEARS_EMPLOYED_0-3
-0.127	Approved_STATUS_HC
-0.206	NOT_ENOUGH_MONEY_PAID_IP

Figura 45. Factors amb major pes per a un client amb probabilitat d'impagament del 0,08%.

Aquesta persona té un risc d'impagament estimat del 0,008 % degut a que presenta molt bones característiques en múltiples variables i pocs aspectes de risc segons el model. Alguna d'aquestes variables es troben explicades a la descriptiva amb l'evolució dels percentatges d'impagaments. L'individu forma part de col·lectius en que la probabilitat d'impagament és de les més baixes que existeixen. Per exemple, la seva mitjana de valoració externa és de 0,69 situada a la part dreta de la figura 19 diferenciada del grup on s'acumulen més morosos. Altres variables que afavoririen individualment o conjuntament amb d'altres la classificació de no impagament serien el nombre de quotes pagades amb una targeta de crèdit prèvia, els ingressos totals i el límit màxim de la targeta de crèdit. El subjecte en qüestió va acabar no impagant.

Un altre individu completament diferent al primer podria ser un que tingui una probabilitat estimada d'impagament del 70%. El model denegaria la sol·licitud de préstec per part del client degut a vuit factors diferents, especialment la mitjana de les valoracions externes. Presenta una mitjana de 0,12, l'hi han rebutjat el crèdit fins un total de quatre vegades (Refused_STATUS_HC), la valoració mínima és de 0,03 (MIN_EXT_SOURCE) és home (CODE_GENDER_F = 0), té un rati entre deute i garantia d'1,26, no està casat i no ha tancat cap crèdit prèviament. Aquests vuit valors dels respectius valors són, després d'efectuar la descriptiva i analitzar per sobre el model logístic, realment dolents. La persona no té cap característica significativa que pugui salvar-lo de la denegació del préstec. I en efecte, si es compara la predicció amb la dada real es veurà que el client sí va acabar incomplint.

y=1 (probability 0.700, score 0.850) top features

Contribution	Feature
+1.407	MEAN_EXT_SOURCE
+0.538	Refused_STATUS_HC
+0.410	MIN_EXT_SOURCE
+0.255	CODE_GENDER_F
+0.219	RATI_DEUTE_GARANTIA
+0.205	NAME_FAMILY_STATUS_Married
+0.197	NUM_CREDITS_PREVIS_TANCATS
+0.121	Approved_STATUS_HC
-0.058	NUM_ACTIVE_CREDITS
-0.069	NOT_ENOUGH_MONEY_PAID_IP
-0.091	Home_CAT_HC
-0.116	PERCENTATGE_ANNUITY
-2.541	<BIAS>

Figura 46. Factors amb major pes per a un client amb probabilitat d'impagament del 70%.

Dels casos més clars es pot passar a un de més complex. Què passa quan la probabilitat estimada és propera al percentatge real d'impagaments. Recordar que s'està treballant amb dades no balancejades. Una opció per solucionar aquest problema seria mitjançant la inserció de pesos al model XGBoost però que és un aspecte que no és d'interès. Es presenta un cas amb una probabilitat del 10,5%, dos punts per sobre de la mitjana real. Una possibilitat seria considerar com a impagaments totes aquelles estimacions que fossin més grans que la mitjana. Així doncs, si es seguís el plantejament anterior el prestatari acabaria incomplint. Presenta un rati entre deute i garantia de 1 i no es troba en el col·lectiu que porta treballant des de fa menys de tres anys. Són característiques bones per reduir-ne la probabilitat. No obstant, la valoració mitjana de 0,37, ser home (CODE_GENDER_F = 0) i treballar de conductor permet a l'algoritme augmentar la probabilitat d'impagament. Tot i que la probabilitat estimada sigui més gran que la proporció de morosos reals i que per tant pugui ser un client amb un alt risc, va acabar complint amb tots els pagaments que pertocaven convertint-se així en una predicció errònia elaborada pel model.

y=0 (probability 0.895, score -2.139) top features

Contribution	Feature
+2.541	<BIAS>
+0.179	RATI_DEUTE_GARANTIA
+0.151	LAST_YEARS_EMPLOYED_0-3
+0.146	REGION_POPULATION_RELATIVE
+0.141	LAST_YEARS_EMPLOYED_10-80
+0.112	Refused_STATUS_HC
+0.099	TOT_DRAWINGS_CC
-0.079	Approved_STATUS_HC
-0.082	Mobile_CAT_HC
-0.099	MAX_EXT_SOURCE
-0.104	NAME_FAMILY_STATUS_Married
-0.142	OCCUPATION_TYPE_Drivers
-0.155	CODE_GENDER_F
-0.572	MEAN_EXT_SOURCE

Figura 47. Factors amb major pes per a un client amb probabilitat d'impagament del 10,5%.

Com s'ha vist amb els tres casos pràctics anteriors, la metodologia derivada de l'IML resulta molt útil per poder explicar les decisions que pren el model. No és exclusiu del XGBoost, sinó que el seu ús es pot ampliar a tota la resta d'algoritmes de ML i els models tradicionals. De totes maneres, és necessari verificar amb l'equip que els resultats i interpretacions tinguin sentit.

6. CONCLUSIONS

Després d'arribar fins aquest punt és moment de recapitular al principi i recordar quins eren els objectius marcats. El propòsit principal era trobar un algoritme estadístic que permetés classificar adequadament els individus entre aquells que acabarien incomplint i els que pagarien totes les anualitats sense donar problemes a l'entitat financera. Tot plegat, posant èmfasi en la millora que poden aportar els models de Machine Learning respecte els tradicionals utilitzats per l'àmplia majoria de bancs. L'altre aspecte destacable és entendre i realitzar els mateixos procediments que podria seguir una empresa financera interessada en la construcció de models a partir de dades pròpies.

Començant per aquest últim, és sense cap mena de dubtes un procés llarg i costós. La construcció dels fitxers d'entrenament i de testatge és l'etapa més important i que se l'hi ha de dedicar més temps per a obtenir resultats correctes en les fases posteriors. S'ha de tenir cura de com es tracten les variables i s'uneixen les diferents taules existents. Posteriorment, és imprescindible analitzar una per una totes les variables per comprovar que contenen la informació que s'espera d'elles i que no siguin conseqüència d'una recollida de dades errònia o esbiaixada, d'alguna mala interpretació de les variables o de codificació. La construcció dels models i els respectius anàlisis depenen en gran part del tractament aplicat a les dades. Així doncs, és important que l'encarregat de les dades disposi tant d'aptituds tècniques, teòriques i de coneixement del sector i entitat.

Les nombroses alternatives als models clàssics i les diferents capacitats predictives que s'obtenen comporta a que no hi hagi un mètode concret establert per a la modelització d'impagaments. Entre els estudiats, l'algoritme XGBoost és el que presenta millors resultats respecte la resta. Proporciona una millora del 3,67% en l'índex de gini i de l'1,24% en l'AUC si es compara amb el model logístic. El segueixen el Random Forest, el Lasso i el SVM lineal amb uns guanys molt parells i pràcticament insignificatius. D'altra banda estimant el SVM amb kernel no lineal i els arbres de decisió s'obtenen resultats definitivament pitjors.

Rànquing	Model	Gini	AUC	Cost
1	XGBoost	+3,67%	+1,24%	Alt
2	Random Forest	+0,31%	+0,11%	Alt
3	Lasso	+0,25%	+0,08%	Baix
4	SVM lineal	+0,18%	+0,05%	Mitjà

5	SVM RBF	-8,71%	-2,97%	Alt
6	Arbres de decisió	-11,76%	-4,00%	Mitjà

Taula 35. Benefici i cost dels models definitius respecte el logístic amb test.

Considerant també els costos i, sabent que l'algoritme XGBoost és el que presenta major cost, aquest segueix essent un model viable per ser utilitzat. La relació entre beneficis i costos respecte la resta de models és assumible si bé, s'obtenen alts guanys elevats a partir d'un cost elevat. Presentar un cost elevat no comporta necessàriament a eliminar-lo de la llista de possibles models a posar en producció sinó que de nou són obligatoris coneixements i capacitats tècniques per a mantenir el funcionament d'aquest. El següent grup compost pel Random Forest, Lasso i el SVM lineal presenten beneficis similars però diferenciats pel cost. L'opció més viable i senzilla seria escollir d'aquest grup el model Lasso al ésser de menor cost dels tres.

Malgrat que el Lasso presenta un cost més fàcil d'assumir que el del XGBoost, existeixen formes de solucionar els problemes que presenta aquest últim. S'ha vist com l'aplicació de les eines de l'IML donen llum a l'inconvenient de la interpretabilitat en els algoritmes de Machine Learning. Com aquestes mesures n'hi ha moltes més, i no tenen per què ser altres algoritmes, que es poden implementar per disminuir els riscos que caracteritzen cadascun dels models.

Per concloure es pot garantir que amb aquestes dades l'algoritme que ha destacat més per la seva capacitat predictiva ha estat el XGBoost. Un dels problemes principals que podria ocórrer al tractar algoritmes de Machine Learning i que podria preocupar l'entitat financera és la seva interpretabilitat. Això, però no és cap mena de problema tal i com s'ha vist en les eines de l'Interpretative Machine Learning en el punt anterior. Addicionalment, sembla que en general existeixen algoritmes que funcionen de forma similar o lleugerament millor que els models clàssics com són el Random Forest i el SVM lineal mentre que n'hi ha d'altres que el seu rendiment és significativament inferior als tradicionals com serien els arbres de decisió i el SVM amb kernel no lineal. De totes maneres, es necessiten de més projectes d'investigació amb dades completament diferents i amb la incorporació d'algoritmes addicionals com podrien ser les xarxes neuronals per a poder afirmar quin model és el més adequat per a la predicció d'impagaments. Aquí s'ha vist com el XGBoost és una molt bona opció per a la modelització d'aquests casos però això no extreu que no hi hagi un altre algoritme semblant o molt diferent que pugui millorar lleugerament o significativament la capacitat predictiva.

Aquesta publicació obre les portes a les entitats financeres a utilitzar noves eines de modelització, sempre tenint en compte els costos que comporten tant per les persones que creen i utilitzen els models, com pels inspectors que han de vetllar per la replicabilitat dels mateixos. A partir de la feina realitzada amb les dades de Home Credit s'ha pogut concloure que el millor model resultant és el XGBoost, demostrant que els mètodes de modelització alternativa poden proporcionar beneficis sobre els models tradicionals. Però serà cada entitat és l'encarregada de valorar sobre les dades i experiència els mètodes que millor se'ls adaptin.

7. BIBLIOGRAFIA

Alonso, A., & Carbó, J. (2020). MACHINE LEARNING IN CREDIT RISK:. *Banco de España*, 34.

ANDBANK. (04 de Juny de 2014). *ANDBANK*. Obtenido de https://www.andbank.es/observatoriodelinversor/que-es-el-coeficiente-de-gini/

Avato. (21 / Desembre / 2019). *Avato*. Recollit de Avato: https://www.avato-consulting.com/?p=28903&lang=en

Awasthi, S. (17 / Desembre / 2020). *Dataaspirant*. Recollit de Dataaspirant: https://dataaspirant.com/svm-kernels/

Bhalla, D. (2019). *Listen Data*. Recollit de Listen Data: https://www.listendata.com/2019/09/gini-cumulative-accuracy-profile-auc.html#Gini-Coefficient

Bluhm, C., Overbeck, L., & Wagner, C. (2010). *Introduction to Credit Risk Modeling*. Boca Raton: Chapman & Hall/CRC.

Castellanos, Y. M. (31 de Gener de 2014). *Economipedia*. Obtenido de https://economipedia.com/definiciones/indice-de-gini.html

Enciclopedia Catalana. (sense data). *Enciclopèdia Catalana*. Recollit de Enciclopèdia Catalana: https://www.enciclopedia.cat/ec-gdlc-e00119378.xml

Finanzas para Mortales. (sense data). *Finanzas para Mortales*. Recollit de https://www.wiki-finanzas.com/index.php?seccion=Contenido&id=2015C526032728

Hale, J. (4 / Març / 2019). *towards data science*. Recollit de towards data science: https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02

Home Credit. (18 / Maig / 2018). *Kaggle*. Recollit de https://www.kaggle.com/c/home-credit-default-risk/data?select=bureau_balance.csv

kassambara. (11 / Març / 2018). *Statistical tools for high-throughput data analysis*. Recollit de Statistical tools for high-throughput data analysis: http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/

Koehrsen, W. (27 / Desembre / 2017). *towards data science*. Recollit de towards data science: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

Minaie, N. (20 / Juliol / 2019). *towards data science*. Recollit de towards data science: https://towardsdatascience.com/sql-in-python-for-beginners-b9a4f9293ecf

Moreno, M. A. (21 / Octubre / 2011). *El Blog Salmón*. Recollit de https://www.elblogsalmon.com/conceptos-de-economia/que-es-el-coeficiente-de-gini

Narkhede, S. (26 / Juny / 2018). *towards data science*. Recollit de https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Narkhede, S. (26 / Juny / 2018). *towards data science*. Recollit de https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Navlani, A. (27 / Desembre / 2019). *datacamp*. Recollit de datacamp: https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

Nguyen, A. (10 / Agost / 2020). *towards data science*. Recollit de towards data science: https://towardsdatascience.com/credit-risk-management-classification-models-hyperparameter-tuning-d3785edd8371

Pelz, D. (12 / Març / 2019). *DW*. Recollit de DW: https://www.dw.com/en/are-ratings-agencies-hurting-africas-economies/a-47870146

Peña, C. S. (30 / Octubre / 2019). *Revistadigital*. Recollit de Revistadigital: https://revistadigital.inesem.es/educacion-sociedad/indicadores-de-desigualdad-social/

Ray, S. (13 / Setembre / 2017). *Analytics Vidhya*. Recollit de Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

Rodrigo, J. A. (Desembre / 2020). *Ciencia de datos*. Recollit de Ciencia de datos: https://www.cienciadedatos.net/documentos/py24-svm-python.html

Rodríguez, A. T. (03 / Setembre / 2020). *BBVA*. Recollit de https://www.bbva.com/es/coeficiente-gini-detector-la-desigualdad-salarial/

Solanki, S. (23 / Octubre / 2020). *CoderzColumn*. Recollit de CoderzColumn: https://coderzcolumn.com/tutorials/machine-learning/how-to-use-eli5-to-understand-sklearn-models-their-performance-and-their-predictions

Tanner, G. (13 / Maig / 2019). *towards data science*. Recollit de towards data science: https://towardsdatascience.com/introduction-to-machine-learning-model-interpretation-55036186eeab

Urooj, W. (27 / Desembre / 2019). *Medium*. Recollit de Medium: https://medium.com/edureka/support-vector-machine-in-python-539dca55c26a

Vickery, R. (6 / Agost / 2019). *towards data science*. Recollit de towards data science: https://towardsdatascience.com/python-libraries-for-interpretable-machine-learning-c476a08ed2c7

Wikipedia. (2021). *Wikipedia, the free encyclopedia*. Recollit de https://en.wikipedia.org/wiki/Home_Credit

Wikipedia. (2021). *Wikipedia, the free encyclopedia*. Obtenido de https://es.wikipedia.org/wiki/Coeficiente_de_Gini

Wikipedia. (2021). *Wikipedia, the free encyclopedia*. Recollit de https://es.wikipedia.org/wiki/Curva_de_Lorenz

XGBoost developers. (2020). *XGBoost*. Recollit de XGBoost: https://xgboost.readthedocs.io/en/latest/python/python_api.html#xgboost.XGBClassifier

Zeder, R. (31 / Juliol / 2018). *Quickonomics*. Recollit de Quickonomics: https://quickonomics.com/what-is-the-gini-index/

8. ANNEX

En aquest apartat d'annex es troben altres taules, gràfics, sortides de programari, entre d'altres, diferents a les vistes prèviament i que poden ser d'interès pel lector. No apareixen en el cos del treball per fer-lo més amè i ràpid de llegir.

8.1 Descripció de les variables

A continuació es detalla cadascuna de les variables utilitzades per a la construcció dels models. També es mostra, per a les categòriques, els nivells i el nombre d'observacions d'aquests.

Variable	Origen	Valors que pren	Descripció breu
Target	application_train	0: Client sa - 282.686 1: Impagament - 24.825	Variable resposta d'interès.
NAME_CONTRACT_TYPE	application_train	Cash loans: 278.232 Revolving loans: 29.279	Identificació si és un préstec en efectiu o giratori.
FLAG_OWN_CAR	application_train	N: No - 202.924 Y: Yes - 104.587	El client posseeix d'un vehicle o no.
ANY_COTXE_INTERACTION	application_train	Numèric	Aquells que tenen cotxe, indica els anys que té el vehicle. Si no en té, la variable pren 0.
FLAG_OWN_REALTY	application_train	N: No - 94.199 Y: Yes - 213.312	El client posseeix un pis / casa o no.
CODE_GENDER	application_train	F: Female - 202.452 M: Male - 105.059	Sexe del client.
AMT_INCOME_TOTAL	application_train	Numèric	Ingressos del client.
AMT_CREDIT	application_train	Numèric	Crèdit del préstec.
AMT_ANNUITY	application_train	Numèric	Anualitat del préstec.
AMT_GOODS_PRICE	application_train	Numèric	Preu dels béns pels quals s'ha demanat el crèdit. (hipoteques, préstec cotxes, targetes de crèdits, crèdits estudiants o préstecs personals).
CNT_CHILDREN	application_train	Numèric	Nombre de nens que el client té a càrrec seu.
CNT_FAM_MEMBERS	application_train	Numèric	Nombre de membres familiars totals.
NAME_TYPE_SUITE	application_train	Children: 3.267 Family: 40.149 Group of people: 271 Other:2.636 Spouse, partner: 11.370 Unaccompanied: 248.526	Amb qui viu el client.

NAME_INCOME_TYPE	application_train	Businessman: 10 Commercial associate: 71.617 Maternity leave: 5 Pensioner: 55.362 State servant: 21.703 Student: 18 Unemployed: 22 Working: 158.774	Tipus d'ingrés del client. És a dir, què fa per a rebre ingressos.
NAME_EDUCATION_TYPE	application_train	Academic degree: 164 Higher education: 74.863 Incomplete higher: 10.277 Lower secondary: 3.816 Secondary / secondary special: 218.391	Nivell màxim d'estudis del client.
NAME_FAMILY_STATUS	application_train	Civil marriage: 29.775 Married: 196432 Separated: 19770 Single / not married: 45444 Unknown: 2 Widow: 16088	Estat civil del client.
NAME_HOUSING_TYPE	application_train	Co-op apartament: 1.122 House / apartament: 272.868 Municipal apartament: 11.183 Office apartament: 2.617 Rented apartament: 4.881 With parents: 14.840	Tipus d'habitatge on viu el client.
REGION_POPULATION_RELATIVE	application_train	Numèric	Valor normalitzat que informa si el client viu en una zona molt o poc poblada (nombres elevats comporten a zones més poblades i viceversa).
REGION_RATING_CLIENT	application_train	1: 32.197 2: 226.984 3: 48.330	Puntuació de l'empresa en funció del lloc de residència del client.
AGE_EXPECTED	application_train	Numèric [21,69]	Edat del client.
LAST_YEARS_EMPLOYED	application_train	[0-3]: 115980 [4-9]: 79294 [10-80]: 47111 Retired: 55374	Nombre d'anys que porta treballant el client.
OCCUPATION_TYPE	application_train	Accountants: 9813 Cleaning staff: 4.651 Cooking staff: 5.946 Core staff: 27.570 Drivers: 18.603 HR staff: 563 High skill tech staff: 11.380	Ofici laboral del client.

		IT staff: 526 Laborers: 55.185 Low-skill Laborers: 2.093 Managers: 21.371 Medicine staff: 8536 Private service staff: 2.652 Realty agents: 751 Retired: 55.362 Sales staff: 32.101 Secretaries: 1.305 Security staff: 6.721 Waiters / barmen staff: 1.348	
ORGANIZATION_TYPE	application_train	Advertising: 429 Agriculture: 2.454 Bank: 2.507 Business: 84.525 Cleaning: 260 Construction: 6.721 Culture: 379 Electricity: 950 Emergency: 560 Government: 10.404 Hotel: 966 Housing: 2.958 Industry: 14.310 Insurance: 597 Kindergarten: 6.880 Legal Services: 305 Medicine: 11.192 Military: 2.633 Mobile: 317 Other: 16.683 Police: 2.341 Postal: 2.157 Realtor: 396 Religion: 85 Restaurant: 1.811 Retired: 55.362 School: 8.892 Security: 3.247 Security Ministries: 1.974 Self-employed: 38.411 Services: 1.575 Telecom: 577 Trade: 14.314 Transport – 8.990 University: 1.327 XNA - 22	Tipus d'empresa en la que el client treballa.
MEAN_EXT_SOURCE	application_train	Numèric	Mitjana de la puntuació del client segons les 3 variables de valoració.
MAX_EXT_SOURCE	application_train	Numèric	Màxim de les puntuacions del client segons les 3 variables de valoració.
MIN_EXT_SOURCE	application_train	Numèric	Mínim de les puntuacions del client segons les 3 variables de valoració.
CRE_SOL_Altres_EF	bureau	Numèric	Nombre total de préstecs en

			estat "*Sold*" en altres EF.
CRE_BDE_Altres_EF	bureau	Numèric	Nombre total de préstecs en estat "*Bad Debt*" en altres EF.
CRE_TOTAL_Altres_EF	bureau	Numèric	Número total de crèdits totals en altres entitats financeres.
TE_CREDIT_Altres_EF	bureau	0: 44.020 1: 263.491	Si el client té crèdit en una altra entitat financera [1] o no [0].
CRE_Another_Type_LOAN_Altres_EF	bureau	0: 306.698 1: 813	Té altres tipus de crèdit en una altra entitat financera.
CAR_LOAN_Altres_EF	bureau	0: 287.936 1: 19.575	Té crèdit/s per a pagar un vehicle en una altra entitat financera.
CONSUMER_CREDIT_Altres_EF	bureau	0: 60.662 1: 246.849	Té crèdit/s per a pagar consumicions en una altra entitat financera.
CREDIT_CARD_Altres_EF	bureau	0: 135.469 1: 172.042	Té crèdit/s per a targetes de crèdit en una altra entitat financera.
BUSINESS_LOAN_Altres_EF	bureau	0: 306.120 1: 1.391	Té crèdit/s d'empresa en una altra entitat financera.
MICROLOAN_Altres_EF	bureau	0: 304.012 1: 3.499	Té un micropréstec en una altra entitat financera.
MORTGAGE_LOAN_Altres_EF	bureau	0: 293.206 1: 14.305	Té un crèdit per a pagar una hipoteca en una altra entitat financera.
UNKNOWN_LOAN_Altres_EF	bureau	0: 307.087 1: 424	Té crèdit/s sense saber ben bé per a quin ús es necessitava en una altra entitat financera.
DID_OVERDUE_Altres_EF	bureau	0: 304.114 1: 3.397	Si el client ha impagat en alguna altra entitat, es codifica com a 1. Sinó, pren 0.
DID_PROLONG_Altres_EF	bureau	0: 299.003 1: 8.508	Si el client ha allargat el pagament en alguna altra entitat, es codifica com a 1. Sinó, pren 0.
DIES_MAX_IMPAGATS_Altres_EF	bureau	PERFECT: 276.459 [1-30]: 24.686 [31-60]: 2.961 [61-90]: 2.206 [90-120]: 794 +120: 405	Màxim entre els dies impagats en una altra entitat codificat amb intervals.
CASH_LOANS_TY_HC	previous_application.csv	0: 136.129 1: 171.382	Identificació si ha demanat prèviament un crèdit en efectiu a HC o no.
CONS_LOANS_TY_HC	previous_application.csv	0: 38.488 1: 269.023	Identificació si ha demanat prèviament un crèdit de consum a HC o no.
REV_LOANS_TY_HC	previous_application.csv	0: 203.452 1: 104.059	Identificació si ha demanat prèviament un crèdit rotatiu a HC o no.
NC_Goods_PURP_HC	previous_application.csv	Numèric	Nombre total d'aplicacions prèvies a HC per si el propòsit del crèdit era per a béns de necessitat.
Repairs_PURP_HC	previous_application.csv	Numèric	Nombre total d'aplicacions prèvies a HC per si el propòsit del crèdit era per a la realització de reparacions.
UrgentN_PURP_HC	previous_application.csv	Numèric	Nombre total d'aplicacions prèvies a HC per si el propòsit del crèdit era per a un cas urgent.
GET_HOME_PURP_HC	previous_application.csv	Numèric	Nombre total d'aplicacions prèvies a HC per si el propòsit era per a obtenir una propietat /

			casa.
CARS_PURP_HC	previous_application.csv	Numèric	Nombre total d'aplicacions prèvies a HC per si el propòsit era per a adquirir un vehicle.
FUN_PURP_HC	previous_application.csv	Numèric	Nombre total d'aplicacions prèvies a HC per si el propòsit era per a divertir-se.
Approved_STATUS_HC	previous_application.csv	Numèric	Nombre total de crèdits a HC en estat aprovats
Canceled_STATUS_HC	previous_application.csv	Numèric	Nombre total de crèdits a HC en estat cancel·lat.
Refused_STATUS_HC	previous_application.csv	Numèric	Nombre total de crèdits a HC en estat rebutjat
UnusedOF_STATUS_HC	previous_application.csv	Numèric	Nombre total de crèdits a HC sense fer servir.
ElectD_CAT_HC	previous_application.csv	Numèric	Nombre total d'aplicacions del client a HC per a béns relacionats amb l'electrònica.
Clothes_CAT_HC	previous_application.csv	Numèric	Nombre total d'aplicacions del client a HC per a béns o relacionats amb la roba / accessoris.
FreeT_CAT_HC	previous_application.csv	Numèric	Nombre total d'aplicacions del client a HC per a béns o relacionats amb el temps lliure.
Health_CAT_HC	previous_application.csv	Numèric	Nombre total d'aplicacions del client a HC per a béns o relacionats amb la salut.
Home_CAT_HC	previous_application.csv	Numèric	Nombre total d'aplicacions del client a HC per a béns o relacionats amb la casa / llar.
Mobile_CAT_HC	previous_application.csv	Numèric	Nombre total d'aplicacions del client a HC per a béns o relacionats amb el mòbil.
ConstructionM_CAT_HC	previous_application.csv	Numèric	Nombre total d'aplicacions del client a HC per a béns o relacionats amb la construcció.
Vehicles_CAT_HC	previous_application.csv	Numèric	Nombre total d'aplicacions del client a HC per a béns o relacionats amb el vehicle.
Credits_Assegurats_HC	previous_application.csv	Menys crèdits assegurats: 197.822 Mateixos crèdits assegurats: 57.983 Més crèdits assegurats: 51.706	Si la persona tenia més, menys o igual nombre de crèdits assegurats.
IS_NEW_CLIENT_TRAIN	previous_application.csv	0: 291.057 1: 16.454	Si el client és nou a HC [1] o no [0].
IS_Refreshed_Client_HC	previous_application.csv	0: 230.806 1: 76.705	Si és un client renovat a HC [1] o no [0].
NUM_Repeater_Client_HC	previous_application.csv	Numèric	Quants cops ha repetit el clienta HC? (Sense comptar aquesta vegada)
SUM_Diferencia_AMT_HC	previous_application.csv	More credit than ATB: 194.795 Less credit than ATB: 77.454 Same credit than ATB: 35.262	Resultat de restar: Crèdit total final demanat a HC – Crèdit demanat inicialment a HC.
LAST_DUE_MONTH	previous_application.csv	Less_6Months: 75.394 Less_Year: 62.097	Quant fa de l'últim mes que el client va incomplir a HC.

		Between_1Year_2Years: 60.971 Between_2Years_5Years: 48.178 No_DPD_PA: 23.079 Retired: 20.648 More_5Years: 17.144	
SUM_BALANCE_CC	credit_card_balance.csv	Numèric	Balanç total de les targetes de crèdit mes per mes.
MAX_CREDIT_LIM_CC	credit_card_balance.csv	Numèric	Límit de la targeta de crèdit màxim del client durant tots els mesos corresponents.
TOT_RECEIVABLE_CC	credit_card_balance.csv	Numèric	Quantitat pendent total a pagar de la targeta de crèdit a HC.
TOT_DRAWINGS_CC	credit_card_balance.csv	Numèric	Nombre de vegades que el client ha tret diners de la targeta de crèdit de HC.
TOT_INSTALMENTS_CC	credit_card_balance.csv	Numèric	Nombre de quotes pagades en els préstecs anteriors.
MEAN_DPD_CC	credit_card_balance.csv	Numèric	Mitjana dels dies vençuts del crèdit anterior tenint en compte tots els mesos.
TE_CREDIT_CARD_CC	credit_card_balance.csv	0: 220.606 1: 86.905	Té targeta de crèdit a HC [1] o no [0].
NUM_QUOTES_PENDENTS	POS_CASH_balance	Numèric	Total de quotes pendents a pagar dels préstecs vigents a HC.
TOT_CO_SIG_PCB_MENSUAL	POS_CASH_balance	Numèric	Nombre de contractes signats a HC.
TOT_CO_RTS_PCB_MENSUAL	POS_CASH_balance	Numèric	Nombre de contractes retornats a HC.
TOT_CO_ALT_PCB_MENSUAL	POS_CASH_balance	Numèric	Nombre de contractes amb un altre estat a HC.
DIES_NOMES_GRAN_DEUTE_PCB	POS_CASH_balance	Numèric	Total de dies amb deute gran del client a HC.
DIES_TOT_DEUTE_PCB	POS_CASH_balance	Numèric	Total dies amb deute: inclou qualsevol impagament.
LATE_PAYMENTS_IP	installments_payments.csv	0: 299.767 1: 7.744	Va pagar tard almenys una vegada a HC.
NOT_ENOUGH_MONEY_PAID_IP	installments_payments.csv	0: 196.530 1: 110.981	No va pagar la quantitat demanada almenys una vegada.
AMT_ANNUITY_TOT	Nova variable	Numèric	Suma d'anualitats de tots els crèdits actius del client.
PERCENTATGE_ANNUITY	Nova variable	Numèric	Resultat de dividir les anualitats totals i el salari total. Llavors, es multiplica per 100.
AMT_CREDIT_ACT_TOT	Nova variable	Numèric	Quantitat total dels crèdits actius del client.
RATI_DEUTE_GARANTIA	Nova variable	Numèric	Resultat de la divisió entre AMT_CREDIT/AMT_GOODS_PRICE. És a dir, la quantitat de crèdit sol·licitada dividida entre la garantia.
NUM_ACTIVE_CREDITS	Nova variable	Numèric	Nombre de crèdits actius que té el client en qualsevol entitat financera, HC inclosa.
NUM_CREDITS_PREVIS_TANCATS	Nova variable	Numèric	Nombre de crèdits tancats que té el client en qualsevol entitat financera, HC inclosa.

Taula 35. Descripció completa de la base de dades definitiva.

8.2 Mapes de calor de les correlacions

En aquest apartat es mostren tots els mapes de calor possibles a partir de les diferents combinacions. Cada grup de variables de l'esquerra es troben tres vegades seguides mentre que les variables de baix varien.

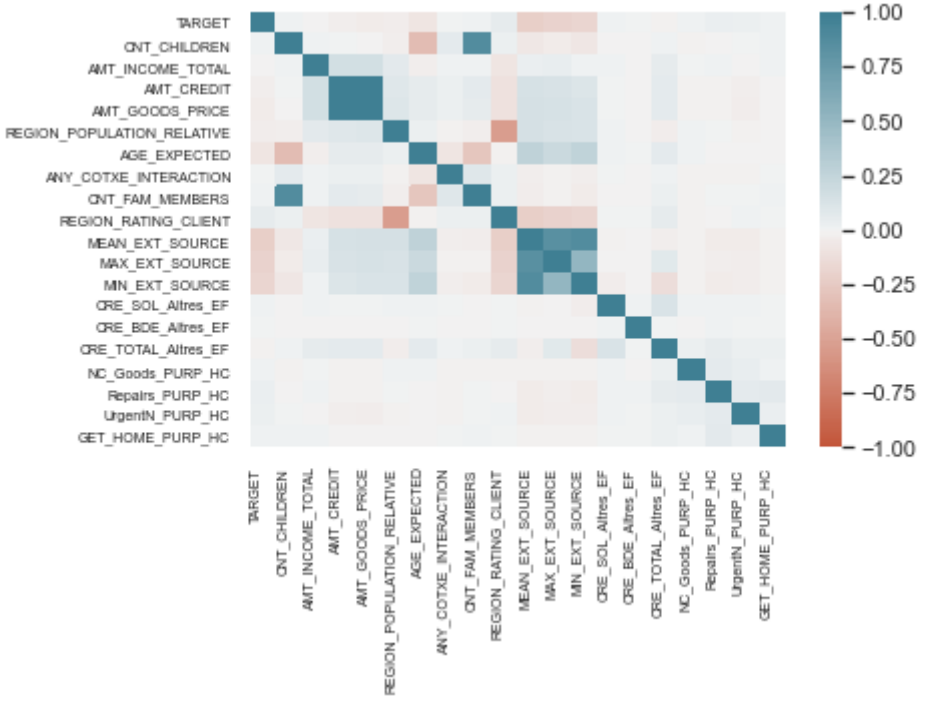


Figura 47. Mapa de calor de les correlacions I.

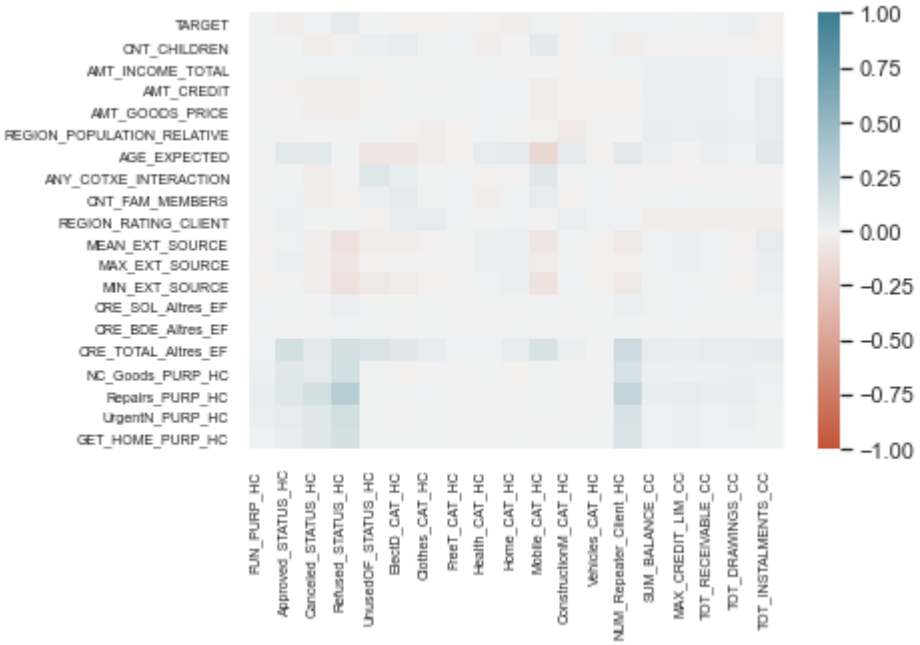


Figura 48. Mapa de calor de les correlacions II.

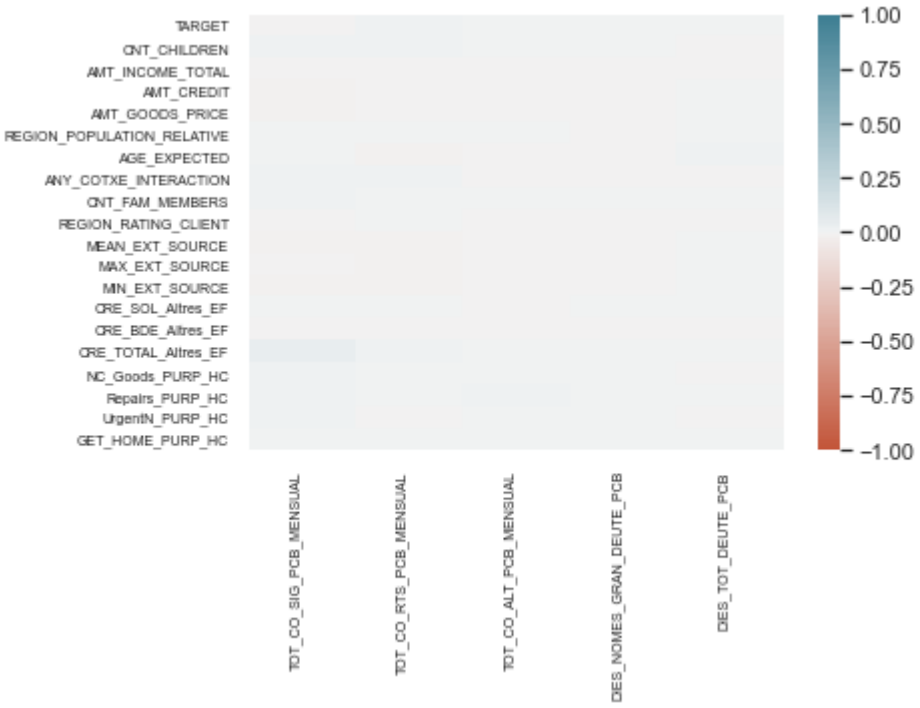


Figura 49. Mapa de calor de les correlacions III.

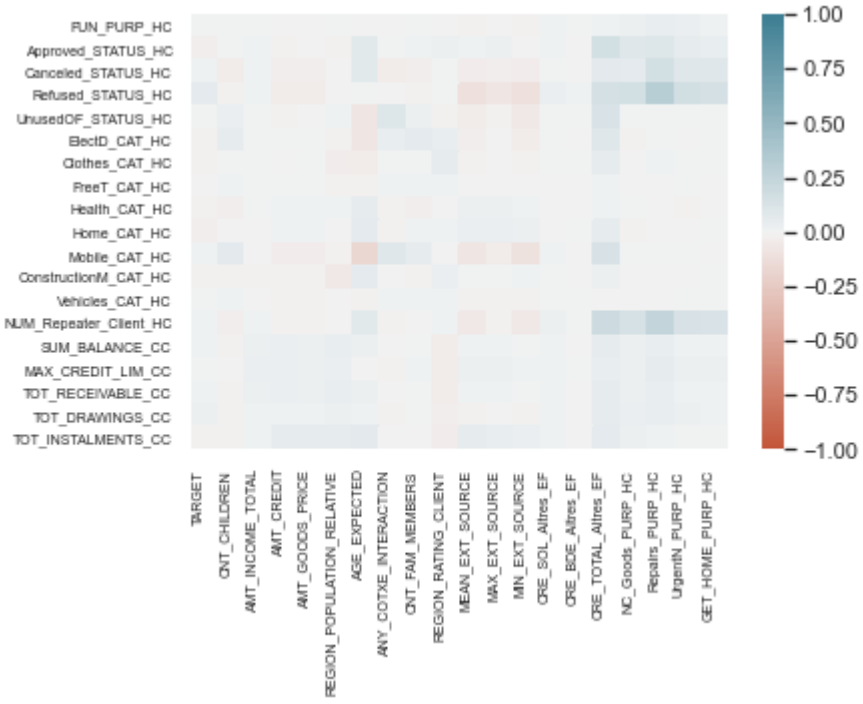


Figura 50. Mapa de calor de les correlacions IV.

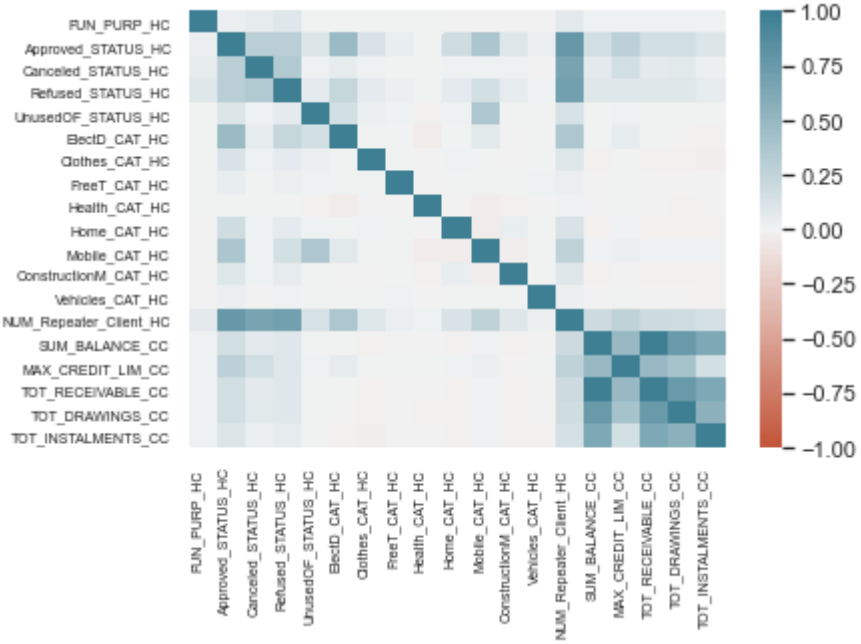


Figura 51. Mapa de calor de les correlacions V.

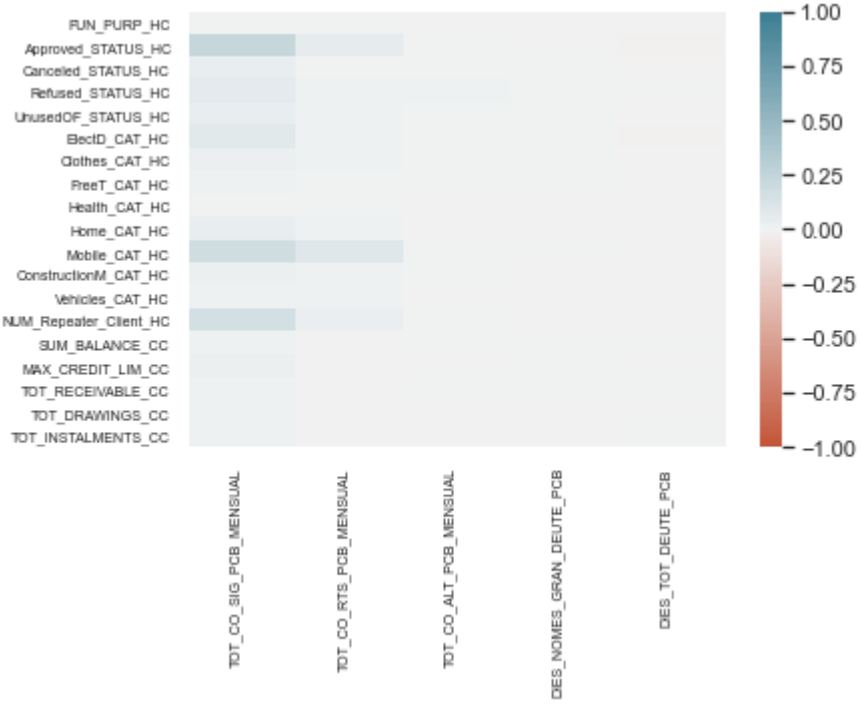


Figura 52. Mapa de calor de les correlacions VI.

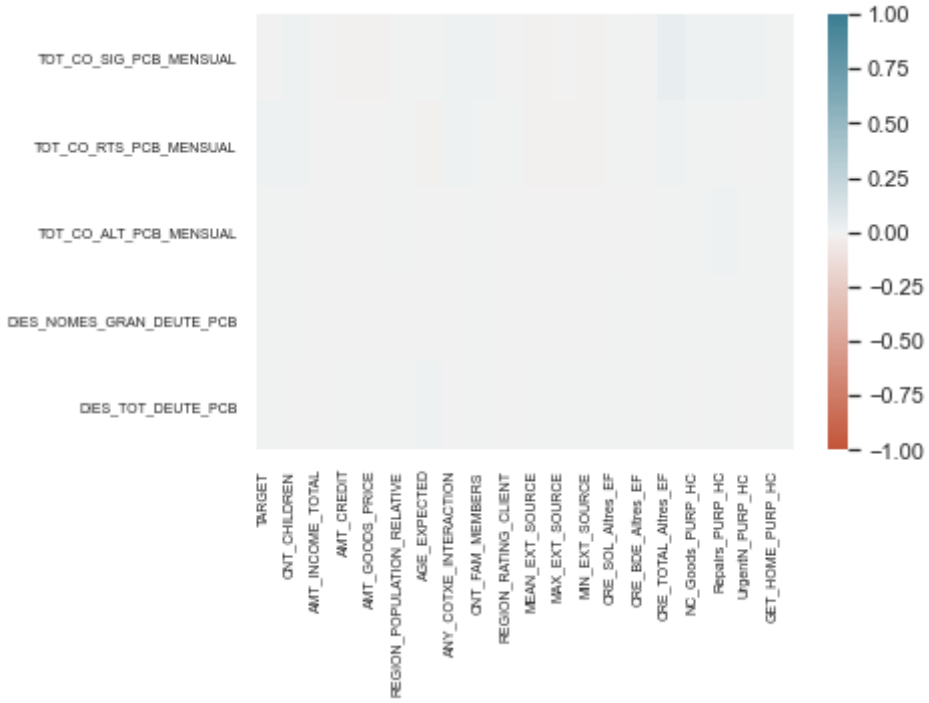


Figura 53. Mapa de calor de les correlacions VII.

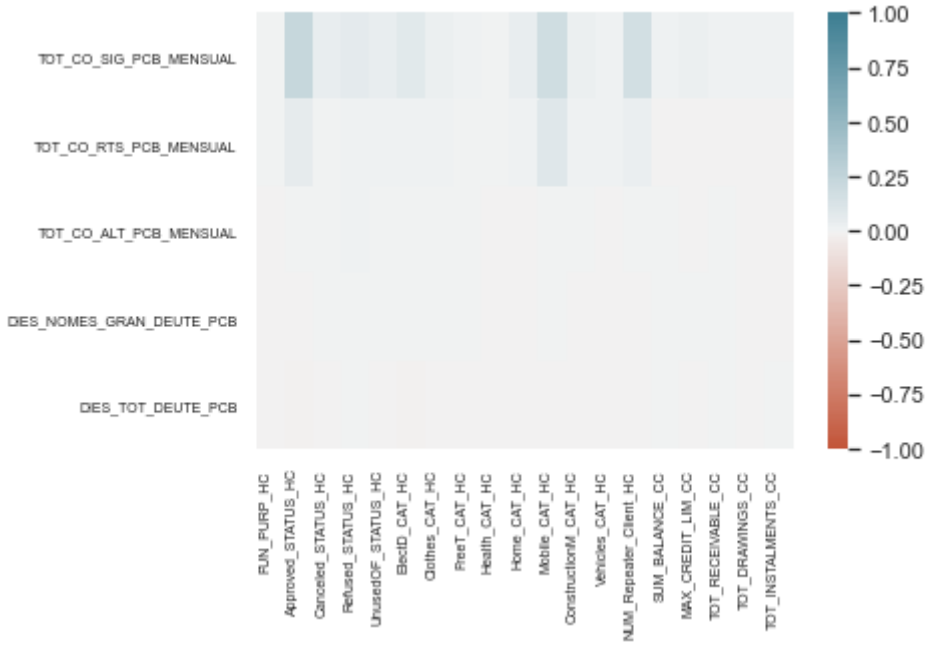


Figura 54. Mapa de calor de les correlacions VIII.

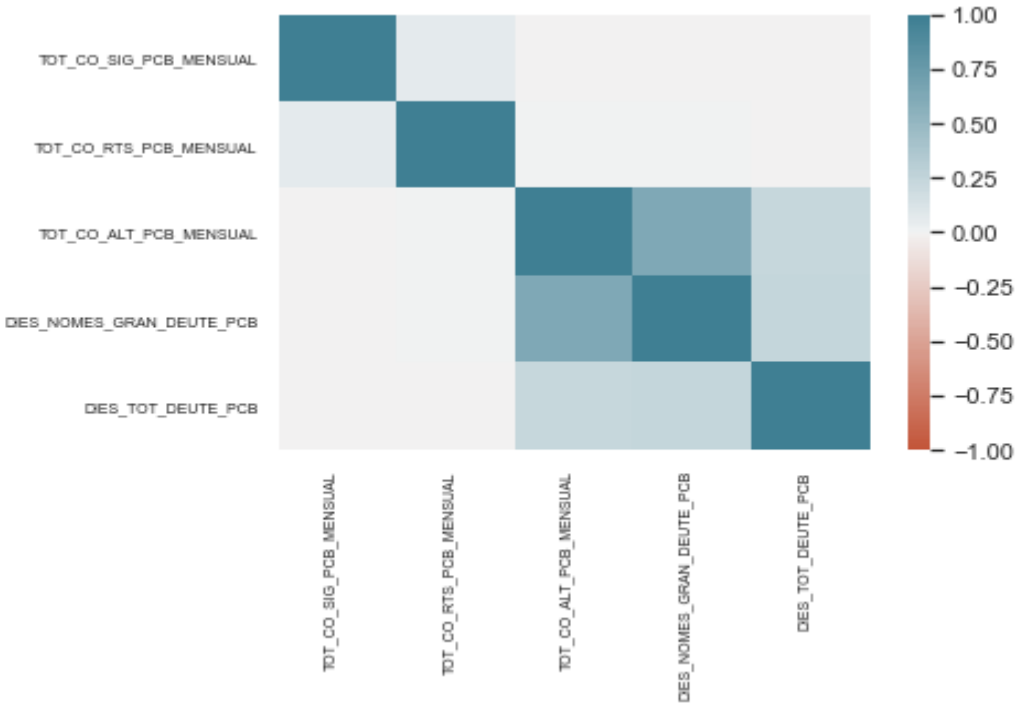


Figura 55. Mapa de calor de les correlacions IX.

8.3 Resultats dels models clàssics

Aquí es troben totes les sortides del programari R per a cada model clàssic creat.

8.3.1 Primer model logístic amb totes les variables

```
> model1 <-  glm(TARGET ~ ., data = train_x, family = binomial)
> summary(model1)

Call:
glm(formula = TARGET ~ ., family = binomial, data = train_x)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6844  -0.4289  -0.3048  -0.2157   3.2954

Coefficients: (14 not defined because of singularities)
                                              Estimate      Std. Error z value             Pr(>|z|)
(Intercept)                                5.875310817332 67.634591305527   0.087              0.930776
CNT_CHILDREN                               0.006277985821  0.044114308753   0.142              0.886834
AMT_INCOME_TOTAL                           0.000000023920  0.000000021946   1.090              0.275725
AMT_CREDIT                                 0.000001148245  0.000000212991   5.391  0.00000007004915831 ***
AMT_GOODS_PRICE                           -0.000001080423  0.000000238722  -4.526  0.00000601511380991 ***
REGION_POPULATION_RELATIVE                 0.349178206883  0.713093785170   0.490              0.624370
AGE_EXPECTED                              -0.004942566771  0.001012057669  -4.884  0.00000104123500526 ***
ANY_COTXE_INTERACTION                      0.004062484085  0.001120543443   3.625              0.000288 ***
CNT_FAM_MEMBERS                            0.023964145668  0.042253201066   0.567              0.570608
REGION_RATING_CLIENT                       0.155903707007  0.017649991939   8.833 < 0.0000000000000002 ***
MEAN_EXT_SOURCE                           -5.130002648624  0.397762186372 -12.897 < 0.0000000000000002 ***
MAX_EXT_SOURCE                             0.808103440018  0.204833845223   3.945  0.00007974484807233 ***
MIN_EXT_SOURCE                             0.339535119740  0.204740198256   1.658              0.097243 .
CRE_SOL_Altres_EF                          0.183926258636  0.048275741275   3.810              0.000139 ***
CRE_BDE_Altres_EF                          0.544187120627  0.693078896453   0.785              0.432352
CRE_TOTAL_Altres_EF                       -0.014518958927  0.004386590164  -3.310              0.000933 ***
TE_CREDIT_Altres_EF                       -0.118019916306  0.041921567637  -2.815              0.004874 **
CRE_Another_Type_LOAN_Altres_EF           -0.202205604528  0.157891736191  -1.281              0.200313
CAR_LOAN_Altres_EF                        -0.155250882327  0.038725980740  -4.009  0.00006098691431717 ***
CONSUMER_CREDIT_Altres_EF                 -0.028539118778  0.036017971483  -0.792              0.428152
CREDIT_CARD_Altres_EF                     -0.050760956080  0.021630519475  -2.347              0.018939 *
BUSINESS_LOAN_Altres_EF                   -0.232139297622  0.132732606281  -1.749              0.080304 .
MICROLOAN_Altres_EF                        0.574541784970  0.052541461496  10.935 < 0.0000000000000002 ***
MORTGAGE_LOAN_Altres_EF                   -0.395232258692  0.047817701738  -8.265 < 0.0000000000000002 ***
UNKNOWN_LOAN_Altres_EF                     0.055161108627  0.221585529900   0.249              0.803409
DID_OVERDUE_Altres_EF                      0.223440257744  0.057805839377   3.865              0.000111 ***
DID_PROLONG_Altres_EF                      0.001492254653  0.046957277070   0.032              0.974648
CASH_LOANS_TY_HC                           0.050434090407  0.021622507108   2.332              0.019675 *
CONS_LOANS_TY_HC                          -0.028529065103  0.032672364908  -0.873              0.382561
REV_LOANS_TY_HC                            0.098585862777  0.026326908177   3.745              0.000181 ***
NC_Goods_PURP_HC                           0.071441378450  0.036254108367   1.971              0.048773 *
Repairs_PURP_HC                            0.034672657368  0.017351427353   1.998              0.045689 *
UrgentN_PURP_HC                            0.078322438157  0.029095938042   2.692              0.007105 **
GET_HOME_PURP_HC                           0.007289625767  0.036523744658   0.200              0.841804
CARS_PURP_HC                              -0.041957033367  0.034604016167  -1.212              0.225325
FUN_PURP_HC                               -0.000648624784  0.066569410656  -0.010              0.992226
Approved_STATUS_HC                        -0.057978048868  0.013230153366  -4.382  0.00001174513833964 ***
Canceled_STATUS_HC                         0.018076069738  0.010700191678   1.689              0.091158 .
Refused_STATUS_HC                          0.083459441665  0.010485585167   7.959  0.00000000000000173 ***
UnusedOF_STATUS_HC                         0.027969313732  0.028493180299   0.982              0.326290
ElectD_CAT_HC                             -0.061653873319  0.009148075879  -6.740  0.00000000001588829 ***
Clothes_CAT_HC                            -0.109628827432  0.023429864903  -4.679  0.00000288248070698 ***
FreeT_CAT_HC                              -0.281275608838  0.072514372932  -3.879              0.000105 ***
Health_CAT_HC                             -0.033968196816  0.064222422439  -0.529              0.596864
Home_CAT_HC                               -0.110478235342  0.020276258763  -5.449  0.00000005075360489 ***
Mobile_CAT_HC                              0.004590687464  0.010758410055   0.427              0.669593
ConstructionM_CAT_HC                      -0.032979933490  0.025502111151  -1.293              0.195934
Vehicles_CAT_HC                           -0.040181220879  0.065811054441  -0.611              0.541495
IS_NEW_CLIENT_TRAIN                        0.133142849823  0.083651192224   1.592              0.111465
IS_Refreshed_Client_HC                    -0.148106097414  0.023638925641  -6.265  0.00000000037199351 ***
NUM_Repeater_Client_HC                    -0.019021098602  0.009777448724  -1.945              0.051726 .
SUM_BALANCE_CC                             0.000000332908  0.000000407833   0.816              0.414337
MAX_CREDIT_LIM_CC                         -0.000000256563  0.000000088983  -2.883              0.003935 **
TOT_RECEIVABLE_CC                         -0.000000310247  0.000000408203  -0.760              0.447237
TOT_DRAWINGS_CC                            0.008694758834  0.001147057752   7.580  0.00000000000003454 ***
TOT_INSTALMENTS_CC                        -0.000197695378  0.000018150299 -10.892 < 0.0000000000000002 ***
MEAN_DPD_CC                                0.001255154999  0.001181594086   1.062              0.288120
TE_CREDIT_CARD_CC                         -0.060528326188  0.036078189413  -1.678              0.093406 .
TOT_CO_SIG_PCB_MENSUAL                     0.003304594146  0.011959821612   0.276              0.782312
TOT_CO_RTS_PCB_MENSUAL                     0.157063850536  0.053781946162   2.920              0.003496 **
TOT_CO_ALT_PCB_MENSUAL                    -0.001763203417  0.008587572954  -0.205              0.837322
DIES_NOMES_GRAN_DEUTE_PCB                  0.000001048090  0.000005842574   0.179              0.857633
DIES_TOT_DEUTE_PCB                         0.000002702188  0.000001374698   1.966              0.049338 *
```

DIES_TOT_DEUTE_PCB	0.000002702188	0.000001374698	1.966	0.049338	*
NUM_QUOTES_PENDENTS	0.008221426869	0.000885055591	9.289	< 0.0000000000000002	***
LATE_PAYMENTS_IP	0.257406974942	0.043883046822	5.866	0.00000000447107463	***
NOT_ENOUGH_MONEY_PAID_IP	0.305273443671	0.017147313890	17.803	< 0.0000000000000002	***
AMT_ANNUITY_TOT	-0.000000422283	0.000000179252	-2.356	0.018483	*
PERCENTATGE_ANNUITY	0.000788231668	0.000270362552	2.915	0.003552	**
AMT_CREDIT_ACT_TOT	0.000000013483	0.000000003572	3.775	0.000160	***
RATI_DEUTE_GARANTIA	0.681275065389	0.105430140399	6.462	0.00000000010342209	***
NUM_ACTIVE_CREDITS	0.092531948238	0.007282378821	12.706	< 0.0000000000000002	***
NUM_CREDITS_PREVIS_TANCATS	-0.006931046440	0.003272511959	-2.118	0.034179	*
`NAME_CONTRACT_TYPE_Cash loans`	0.175504784748	0.033889442048	5.179	0.00000022338334585	***
`NAME_CONTRACT_TYPE_Revolving loans`	NA	NA	NA	NA	
CODE_GENDER_F	-0.279007450230	0.020274329357	-13.762	< 0.0000000000000002	***
CODE_GENDER_M	NA	NA	NA	NA	
FLAG_OWN_CAR_N	0.269505696472	0.024373588287	11.057	< 0.0000000000000002	***
FLAG_OWN_CAR_Y	NA	NA	NA	NA	
FLAG_OWN_REALTY_N	-0.035026021254	0.017673133000	-1.982	0.047493	*
FLAG_OWN_REALTY_Y	NA	NA	NA	NA	
NAME_TYPE_SUITE_Children	0.601728202973	0.187219236687	3.214	0.001309	**
NAME_TYPE_SUITE_Family	0.531194309953	0.171964285601	3.089	0.002008	**
`NAME_TYPE_SUITE_Group of people`	0.586706888184	0.300660439073	1.951	0.051010	.
NAME_TYPE_SUITE_Other	0.576834911326	0.187726562279	3.073	0.002121	**
`NAME_TYPE_SUITE_Spouse, partner`	0.462673720139	0.175393725119	2.638	0.008342	**
NAME_TYPE_SUITE_Unaccompanied	0.523919016161	0.170707682427	3.069	0.002147	**
NAME_INCOME_TYPE_Businessman	-7.913208485622	69.066264587061	-0.115	0.908783	
`NAME_INCOME_TYPE_Commercial associate`	-0.114429814874	0.020107189902	-5.691	0.00000001263049715	***
`NAME_INCOME_TYPE_Maternity leave`	2.548139474871	1.208112960230	2.109	0.034928	*
NAME_INCOME_TYPE_Pensioner	-8.504068920389	67.633798595430	-0.126	0.899940	
`NAME_INCOME_TYPE_State servant`	-0.099524315347	0.039247892633	-2.536	0.011219	*
NAME_INCOME_TYPE_Student	-9.455714730134	43.581138945843	-0.217	0.828233	
NAME_INCOME_TYPE_Unemployed	-7.060915363440	67.636636510868	-0.104	0.916856	
NAME_INCOME_TYPE_Working	NA	NA	NA	NA	
`NAME_EDUCATION_TYPE_Academic degree`	-1.303939953096	0.598147314769	-2.180	0.029260	*
`NAME_EDUCATION_TYPE_Higher education`	-0.290492027849	0.022708896629	-12.792	< 0.0000000000000002	***
`NAME_EDUCATION_TYPE_Incomplete higher`	-0.224821877540	0.043210139611	-5.203	0.00000019610856812	***
`NAME_EDUCATION_TYPE_Lower secondary`	0.076093984706	0.062201776366	1.223	0.221201	
`NAME_EDUCATION_TYPE_Secondary / secondary special`	NA	NA	NA	NA	
`NAME_FAMILY_STATUS_Civil marriage`	0.122593199569	0.025503103713	4.807	0.00000153218598601	***
NAME_FAMILY_STATUS_Married	NA	NA	NA	NA	
NAME_FAMILY_STATUS_Separated	0.156503959923	0.050521327781	3.098	0.001950	**
`NAME_FAMILY_STATUS_Single / not married`	0.095702333852	0.046018609559	2.080	0.037558	*
NAME_FAMILY_STATUS_Widow	NA	NA	NA	NA	
`NAME_HOUSING_TYPE_Co-op apartment`	-0.041242884903	0.134763793308	-0.306	0.759575	
`NAME_HOUSING_TYPE_House / apartment`	-0.008477820760	0.032929251671	-0.257	0.796827	
`NAME_HOUSING_TYPE_Municipal apartment`	0.069148182665	0.050912495127	1.358	0.174408	
`NAME_HOUSING_TYPE_Office apartment`	-0.264198038582	0.099598837644	-2.653	0.007987	**
`NAME_HOUSING_TYPE_Rented apartment`	0.129639446060	0.059201209405	2.190	0.028538	*
`NAME_HOUSING_TYPE_With parents`	NA	NA	NA	NA	
`LAST_YEARS_EMPLOYED_0-3`	-8.377287560884	67.633633425922	-0.124	0.901424	
`LAST_YEARS_EMPLOYED_4-9`	-8.532772996089	67.633633853586	-0.126	0.899604	
`LAST_YEARS_EMPLOYED_10-80`	-8.721976301119	67.633635416557	-0.129	0.897390	
OCCUPATION_TYPE_Accountants	-0.170983884116	0.060098560380	-2.845	0.004440	**
`OCCUPATION_TYPE_Cleaning staff`	0.095985556578	0.062880066219	1.526	0.126889	
`OCCUPATION_TYPE_Cooking staff`	0.100888455320	0.055996510308	1.802	0.071594	.
`OCCUPATION_TYPE_Core staff`	-0.052700319513	0.039415467521	-1.337	0.181207	
OCCUPATION_TYPE_Drivers	0.166737240300	0.037041226033	4.501	0.00000675084787502	***
`OCCUPATION_TYPE_HR staff`	-0.199775259996	0.219041120098	-0.912	0.361745	
`OCCUPATION_TYPE_High skill tech staff`	-0.099849509954	0.049900599181	-2.001	0.045396	*
`OCCUPATION_TYPE_IT staff`	0.060355265755	0.203460613425	0.297	0.766739	
OCCUPATION_TYPE_Laborers	0.072633123452	0.027869912385	2.606	0.009157	**
`OCCUPATION_TYPE_Low-skill Laborers`	0.227402617817	0.072807206963	3.123	0.001788	**
OCCUPATION_TYPE_Managers	0.027845833837	0.040277608859	0.691	0.489347	
`OCCUPATION_TYPE_Medicine staff`	-0.029362425947	0.068274346790	-0.430	0.667148	
`OCCUPATION_TYPE_Private service staff`	-0.029615767078	0.095961357556	-0.309	0.757609	
`OCCUPATION_TYPE_Realty agents`	-0.081897388879	0.164827622742	-0.497	0.619283	
`OCCUPATION_TYPE_Sales staff`	0.026175936929	0.033244368754	0.787	0.431060	
OCCUPATION_TYPE_Secretaries	0.185355635459	0.124912184616	1.484	0.137839	
`OCCUPATION_TYPE_Security staff`	0.108584379249	0.059822748901	1.815	0.069508	.
`OCCUPATION_TYPE_Waiters/barmen staff`	0.181117801257	0.104421276945	1.734	0.082831	.
ORGANIZATION_TYPE_Advertising	0.434957881344	0.248789077458	1.748	0.080412	.
ORGANIZATION_TYPE_Agriculture	0.309478664493	0.168724025337	1.834	0.066620	.
ORGANIZATION_TYPE_Bank	-0.064366700421	0.184562570964	-0.349	0.727275	
ORGANIZATION_TYPE_Business	0.279140570147	0.150621210663	1.853	0.063845	.

```
ORGANIZATION_TYPE_Business                          0.279140570147  0.150621210663   1.853            0.063845 .
ORGANIZATION_TYPE_Cleaning                          0.379238184508  0.280511410717   1.352            0.176390
ORGANIZATION_TYPE_Construction                      0.483274692692  0.156780175294   3.082            0.002053 **
ORGANIZATION_TYPE_Culture                           0.337358120637  0.291398341300   1.158            0.246978
ORGANIZATION_TYPE_Electricity                       0.199915608340  0.210397014996   0.950            0.342019
ORGANIZATION_TYPE_Emergency                         0.125849576299  0.236237669941   0.533            0.594224
ORGANIZATION_TYPE_Government                        0.143940899691  0.156282651353   0.921            0.357035
ORGANIZATION_TYPE_Hotel                            -0.154269969347  0.219366685595  -0.703            0.481899
ORGANIZATION_TYPE_Housing                           0.259912637739  0.168934503762   1.539            0.123916
ORGANIZATION_TYPE_Industry                          0.207580829952  0.153930879081   1.349            0.177487
ORGANIZATION_TYPE_Insurance                        -0.085093646746  0.273153814674  -0.312            0.755403
ORGANIZATION_TYPE_Kindergarten                      0.175388858878  0.159917286562   1.097            0.272752
`ORGANIZATION_TYPE_Legal Services`                  0.855883860755  0.280694880090   3.049            0.002295 **
ORGANIZATION_TYPE_Medicine                          0.142297731905  0.159471026360   0.892            0.372226
ORGANIZATION_TYPE_Military                         -0.187161073666  0.181156289595  -1.033            0.301535
ORGANIZATION_TYPE_Mobile                            0.200125880883  0.276207304225   0.725            0.468728
ORGANIZATION_TYPE_Other                             0.219516102802  0.153412488453   1.431            0.152462
ORGANIZATION_TYPE_Police                           -0.085502773598  0.186856739147  -0.458            0.647251
ORGANIZATION_TYPE_Postal                            0.333000735705  0.174617105681   1.907            0.056516 .
ORGANIZATION_TYPE_Realtor                           0.963903809507  0.245168922158   3.932  0.00008438571169679 ***
ORGANIZATION_TYPE_Religion                          0.134684743806  0.619984663393   0.217            0.828022
ORGANIZATION_TYPE_Restaurant                        0.310778105092  0.175350485090   1.772            0.076341 .
ORGANIZATION_TYPE_School                            0.084167047306  0.158367548815   0.531            0.595096
ORGANIZATION_TYPE_Security                          0.120842240508  0.170355686413   0.709            0.478106
`ORGANIZATION_TYPE_Security Ministries`            -0.067329655543  0.189967104736  -0.354            0.723018
`ORGANIZATION_TYPE_Self-employed`                   0.375396650065  0.151603415146   2.476            0.013280 *
ORGANIZATION_TYPE_Services                          0.176252285868  0.195414499581   0.902            0.367088
ORGANIZATION_TYPE_Telecom                           0.351671160252  0.238252452307   1.476            0.139932
ORGANIZATION_TYPE_Trade                             0.250419738135  0.154330631528   1.623            0.104671
ORGANIZATION_TYPE_Transport                         0.305971717692  0.155831130853   1.963            0.049590 *
ORGANIZATION_TYPE_University                                   NA              NA      NA                  NA
`DIES_MAX_IMPAGATS_Altres_EF_+120`                 -0.113476344362  0.201880282468  -0.562            0.574050
DIES_MAX_IMPAGATS_Altres_EF_PERFECT                -0.215132705860  0.186207347772  -1.155            0.247951
`DIES_MAX_IMPAGATS_Altres_EF_[1-30]`               -0.134616053489  0.187590181083  -0.718            0.473000
`DIES_MAX_IMPAGATS_Altres_EF_[31-60]`              -0.142467142364  0.200172207892  -0.712            0.476636
`DIES_MAX_IMPAGATS_Altres_EF_[61-90]`              -0.408344597828  0.241248354469  -1.693            0.090526 .
`DIES_MAX_IMPAGATS_Altres_EF_[91-120]`                         NA              NA      NA                  NA
`Credits_Assegurats_HC_Mateixos crèdits assegurats` -0.041443607089  0.029428569599  -1.408            0.159049
`Credits_Assegurats_HC_Menys crèdits assegurats`    -0.022662462319  0.025405225017  -0.892            0.372372
`Credits_Assegurats_HC_Més crèdits assegurats`                 NA              NA      NA                  NA
`SUM_Diferencia_AMT_HC_Less credit than ATB`       -0.055458807075  0.038135038393  -1.454            0.145870
`SUM_Diferencia_AMT_HC_More credit than ATB`        0.051972805990  0.036096755920   1.440            0.149919
`SUM_Diferencia_AMT_HC_Same credit than ATB`                   NA              NA      NA                  NA
LAST_DUE_MONTH_Less_6Months                         0.204363887661  0.060324744887   3.388            0.000705 ***
LAST_DUE_MONTH_Less_Year                            0.342009178338  0.060481637175   5.655  0.00000001560636522 ***
LAST_DUE_MONTH_Between_1Year_2Years                 0.301026613320  0.060729105472   4.957  0.00000071635857728 ***
LAST_DUE_MONTH_Between_2Years_5Years                0.305738010063  0.061821391460   4.946  0.00000075946661145 ***
LAST_DUE_MONTH_More_5Years                          0.432922861938  0.067826103930   6.383  0.00000000017383865 ***
LAST_DUE_MONTH_Retired                              0.422638017747  0.064441789409   6.558  0.00000000005437127 ***
LAST_DUE_MONTH_No_DPD_PA                                       NA              NA      NA                  NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 137558  on 245640  degrees of freedom
Residual deviance: 120991  on 245471  degrees of freedom
  (367 observations deleted due to missingness)
AIC: 121331

Number of Fisher Scoring iterations: 10
```

Figura 56. Resum del primer model logístic creat amb totes les variables.

8.3.2 Segon model logístic amb totes les variables

```
> model_totes_var <- (glm(TARGET ~ . - CNT_FAM_MEMBERS + DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF
 +  FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y, data = train_x2new, family = binomial))
> summary(model_totes_var)

Call:
glm(formula = TARGET ~ . - CNT_FAM_MEMBERS + DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
    FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y, family = binomial, data = train_x2new)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6867  -0.4285  -0.3050  -0.2159   3.3021

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -1.763e+00  1.483e-01 -11.891  < 2e-16 ***
CNT_CHILDREN                             3.016e-02  1.122e-02   2.687 0.007201 **
AMT_INCOME_TOTAL                         2.375e-08  2.175e-08   1.092 0.274828
AMT_CREDIT                               1.144e-06  2.130e-07   5.374 7.71e-08 ***
AMT_GOODS_PRICE                         -1.078e-06  2.387e-07  -4.518 6.25e-06 ***
REGION_POPULATION_RELATIVE               3.516e-01  7.132e-01   0.493 0.621996
AGE_EXPECTED                            -4.882e-03  1.012e-03  -4.825 1.40e-06 ***
ANY_COTXE_INTERACTION                    4.093e-03  1.120e-03   3.655 0.000257 ***
REGION_RATING_CLIENT                     1.557e-01  1.765e-02   8.822  < 2e-16 ***
MEAN_EXT_SOURCE                         -5.142e+00  3.978e-01 -12.928  < 2e-16 ***
MAX_EXT_SOURCE                           8.102e-01  2.048e-01   3.955 7.64e-05 ***
MIN_EXT_SOURCE                           3.465e-01  2.047e-01   1.692 0.090616 .
CRE_SOL_Altres_EF                        1.841e-01  4.829e-02   3.813 0.000138 ***
CRE_BDE_Altres_EF                        5.086e-01  6.943e-01   0.733 0.463841
CRE_TOTAL_Altres_EF                     -1.449e-02  4.387e-03  -3.304 0.000954 ***
TE_CREDIT_Altres_EF                     -1.177e-01  4.192e-02  -2.807 0.004994 **
CRE_Another_Type_LOAN_Altres_EF         -2.046e-01  1.578e-01  -1.297 0.194745
CAR_LOAN_Altres_EF                      -1.569e-01  3.872e-02  -4.052 5.08e-05 ***
CONSUMER_CREDIT_Altres_EF               -2.896e-02  3.602e-02  -0.804 0.421373
CREDIT_CARD_Altres_EF                   -5.041e-02  2.163e-02  -2.330 0.019780 *
BUSINESS_LOAN_Altres_EF                 -2.320e-01  1.327e-01  -1.748 0.080469 .
MICROLOAN_Altres_EF                      5.756e-01  5.255e-02  10.953  < 2e-16 ***
MORTGAGE_LOAN_Altres_EF                 -3.975e-01  4.780e-02  -8.316  < 2e-16 ***
UNKNOWN_LOAN_Altres_EF                   5.884e-02  2.216e-01   0.266 0.790566
DID_OVERDUE_Altres_EF                    1.970e-01  5.968e-02   3.301 0.000964 ***
DID_PROLONG_Altres_EF                   -1.620e-02  4.800e-02  -0.338 0.735720
CASH_LOANS_TY_HC                         5.088e-02  2.162e-02   2.353 0.018614 *
CONS_LOANS_TY_HC                        -2.806e-02  3.267e-02  -0.859 0.390318
REV_LOANS_TY_HC                          9.839e-02  2.633e-02   3.737 0.000186 ***
NC_Goods_PURP_HC                         7.169e-02  3.626e-02   1.977 0.048000 *
Repairs_PURP_HC                          3.402e-02  1.736e-02   1.960 0.049965 *
UrgentN_PURP_HC                          7.892e-02  2.909e-02   2.713 0.006675 **
GET_HOME_PURP_HC                         6.403e-03  3.653e-02   0.175 0.860879
CARS_PURP_HC                            -4.306e-02  3.465e-02  -1.243 0.213867
FUN_PURP_HC                             -3.779e-03  6.657e-02  -0.057 0.954727
Approved_STATUS_HC                      -5.802e-02  1.323e-02  -4.385 1.16e-05 ***
Canceled_STATUS_HC                       1.791e-02  1.070e-02   1.673 0.094243 .
Refused_STATUS_HC                        8.334e-02  1.049e-02   7.946 1.93e-15 ***
UnusedOF_STATUS_HC                       2.748e-02  2.849e-02   0.964 0.334805
ElectD_CAT_HC                           -6.170e-02  9.150e-03  -6.743 1.55e-11 ***
Clothes_CAT_HC                          -1.099e-01  2.343e-02  -4.688 2.76e-06 ***
FreeT_CAT_HC                            -2.818e-01  7.252e-02  -3.885 0.000102 ***
Health_CAT_HC                           -3.475e-02  6.427e-02  -0.541 0.588650
Home_CAT_HC                             -1.107e-01  2.027e-02  -5.460 4.76e-08 ***
Mobile_CAT_HC                            4.443e-03  1.076e-02   0.413 0.679630
ConstructionM_CAT_HC                    -3.307e-02  2.549e-02  -1.297 0.194469
Vehicles_CAT_HC                         -4.068e-02  6.585e-02  -0.618 0.536691
IS_NEW_CLIENT_TRAIN                      1.248e-01  8.359e-02   1.493 0.135553
IS_Refreshed_Client_HC                  -1.476e-01  2.364e-02  -6.242 4.31e-10 ***
NUM_Repeater_Client_HC                  -1.887e-02  9.780e-03  -1.930 0.053665 .
SUM_BALANCE_CC                           3.349e-07  4.079e-07   0.821 0.411708
MAX_CREDIT_LIM_CC                       -2.581e-07  8.898e-08  -2.900 0.003728 **
TOT_RECEIVABLE_CC                       -3.125e-07  4.083e-07  -0.765 0.444115
TOT_DRAWINGS_CC                          8.732e-03  1.147e-03   7.612 2.69e-14 ***
TOT_INSTALMENTS_CC                      -1.974e-04  1.815e-05 -10.879  < 2e-16 ***
MEAN_DPD_CC                              1.260e-03  1.181e-03   1.067 0.285837
TE_CREDIT_CARD_CC                       -5.966e-02  3.608e-02  -1.654 0.098211 .
TOT_CO_SIG_PCB_MENSUAL                   3.269e-03  1.197e-02   0.273 0.784701
TOT_CO_RTS_PCB_MENSUAL                   1.574e-01  5.379e-02   2.926 0.003428 **
```

TOT_CO_RTS_PCB_MENSUAL	1.574e-01	5.379e-02	2.926	0.003428	**
TOT_CO_ALT_PCB_MENSUAL	-1.615e-03	8.575e-03	-0.188	0.850634	
DIES_NOMES_GRAN_DEUTE_PCB	1.011e-06	5.836e-06	0.173	0.862436	
DIES_TOT_DEUTE_PCB	2.704e-06	1.375e-06	1.966	0.049300	*
NUM_QUOTES_PENDENTS	8.243e-03	8.851e-04	9.314	< 2e-16	***
LATE_PAYMENTS_IP	2.578e-01	4.389e-02	5.874	4.26e-09	***
NOT_ENOUGH_MONEY_PAID_IP	3.051e-01	1.715e-02	17.793	< 2e-16	***
AMT_ANNUITY_TOT	-4.050e-07	1.783e-07	-2.271	0.023147	*
PERCENTATGE_ANNUITY	7.580e-04	2.701e-04	2.806	0.005016	**
AMT_CREDIT_ACT_TOT	1.342e-08	3.563e-09	3.766	0.000166	***
RATI_DEUTE_GARANTIA	6.833e-01	1.054e-01	6.482	9.05e-11	***
NUM_ACTIVE_CREDITS	9.238e-02	7.282e-03	12.686	< 2e-16	***
NUM_CREDITS_PREVIS_TANCATS	-6.918e-03	3.273e-03	-2.114	0.034547	*
`NAME_CONTRACT_TYPE_Revolving loans`	-1.743e-01	3.389e-02	-5.145	2.67e-07	***
CODE_GENDER_M	2.856e-01	2.016e-02	14.169	< 2e-16	***
FLAG_OWN_CAR_Y	-3.214e-01	3.492e-02	-9.202	< 2e-16	***
FLAG_OWN_REALTY_Y	1.400e-02	2.097e-02	0.668	0.504345	
NAME_TYPE_SUITE_Children	8.027e-02	7.782e-02	1.031	0.302322	
NAME_TYPE_SUITE_Family	9.312e-03	2.380e-02	0.391	0.695596	
`NAME_TYPE_SUITE_Group of people`	6.179e-02	2.479e-01	0.249	0.803143	
NAME_TYPE_SUITE_Other	5.515e-02	7.914e-02	0.697	0.485879	
`NAME_TYPE_SUITE_Spouse, partner`	-6.005e-02	4.218e-02	-1.424	0.154575	
NAME_INCOME_TYPE_Businessman	-7.791e+00	6.906e+01	-0.113	0.910179	
`NAME_INCOME_TYPE_Maternity leave`	2.638e+00	1.209e+00	2.182	0.029137	*
NAME_INCOME_TYPE_Pensioner	-2.689e-01	4.100e-02	-6.559	5.42e-11	***
`NAME_INCOME_TYPE_State servant`	1.601e-02	4.216e-02	0.380	0.704126	
NAME_INCOME_TYPE_Student	-9.340e+00	4.378e+01	-0.213	0.831072	
NAME_INCOME_TYPE_Unemployed	1.182e+00	6.204e-01	1.905	0.056756	.
NAME_INCOME_TYPE_Working	1.160e-01	2.010e-02	5.772	7.81e-09	***
`NAME_EDUCATION_TYPE_Academic degree`	-1.308e+00	5.979e-01	-2.187	0.028754	*
`NAME_EDUCATION_TYPE_Higher education`	-2.973e-01	2.262e-02	-13.146	< 2e-16	***
`NAME_EDUCATION_TYPE_Incomplete higher`	-2.267e-01	4.320e-02	-5.248	1.54e-07	***
`NAME_EDUCATION_TYPE_Lower secondary`	7.779e-02	6.220e-02	1.251	0.211084	
`NAME_FAMILY_STATUS_Civil marriage`	1.233e-01	2.550e-02	4.836	1.32e-06	***
NAME_FAMILY_STATUS_Separated	1.333e-01	3.249e-02	4.103	4.07e-05	***
`NAME_FAMILY_STATUS_Single / not married`	7.205e-02	2.301e-02	3.131	0.001741	**
NAME_FAMILY_STATUS_Widow	-2.221e-02	4.225e-02	-0.526	0.599144	
`NAME_HOUSING_TYPE_Co-op apartment`	-3.032e-02	1.316e-01	-0.230	0.817773	
`NAME_HOUSING_TYPE_Municipal apartment`	7.666e-02	4.142e-02	1.851	0.064181	.
`NAME_HOUSING_TYPE_Office apartment`	-2.556e-01	9.509e-02	-2.688	0.007178	**
`NAME_HOUSING_TYPE_Rented apartment`	1.383e-01	5.220e-02	2.649	0.008065	**
`NAME_HOUSING_TYPE_With parents`	8.832e-03	3.293e-02	0.268	0.788545	
`LAST_YEARS_EMPLOYED_4-9`	-1.556e-01	1.961e-02	-7.937	2.07e-15	***
`LAST_YEARS_EMPLOYED_10-80`	-3.450e-01	2.848e-02	-12.112	< 2e-16	***
`OCCUPATION_TYPE_Cleaning staff`	1.189e-01	6.245e-02	1.903	0.057009	.
`OCCUPATION_TYPE_Cooking staff`	1.245e-01	5.545e-02	2.246	0.024710	*
`OCCUPATION_TYPE_Core staff`	-2.612e-02	3.845e-02	-0.679	0.496988	
OCCUPATION_TYPE_Drivers	1.859e-01	3.654e-02	5.087	3.63e-07	***
`OCCUPATION_TYPE_HR staff`	-1.711e-01	2.188e-01	-0.782	0.434298	
`OCCUPATION_TYPE_High skill tech staff`	-7.585e-02	4.927e-02	-1.540	0.123666	
`OCCUPATION_TYPE_IT staff`	8.226e-02	2.035e-01	0.404	0.686043	
OCCUPATION_TYPE_Laborers	9.334e-02	2.707e-02	3.448	0.000565	***
`OCCUPATION_TYPE_Low-skill Laborers`	2.440e-01	7.258e-02	3.361	0.000775	***
OCCUPATION_TYPE_Managers	5.244e-02	3.942e-02	1.330	0.183418	
`OCCUPATION_TYPE_Medicine staff`	-4.893e-03	6.782e-02	-0.072	0.942490	
`OCCUPATION_TYPE_Private service staff`	-4.208e-03	9.562e-02	-0.044	0.964903	
`OCCUPATION_TYPE_Realty agents`	-5.973e-02	1.646e-01	-0.363	0.716781	
`OCCUPATION_TYPE_Sales staff`	5.096e-02	3.221e-02	1.582	0.113690	
OCCUPATION_TYPE_Secretaries	2.113e-01	1.246e-01	1.696	0.089936	.
`OCCUPATION_TYPE_Security staff`	1.281e-01	5.951e-02	2.153	0.031298	*
`OCCUPATION_TYPE_Waiters/barmen staff`	2.018e-01	1.042e-01	1.938	0.052675	.
ORGANIZATION_TYPE_Advertising	1.574e-01	1.993e-01	0.790	0.429738	
ORGANIZATION_TYPE_Agriculture	2.433e-02	7.913e-02	0.308	0.758450	
ORGANIZATION_TYPE_Bank	-3.631e-01	1.092e-01	-3.325	0.000885	***
ORGANIZATION_TYPE_Cleaning	1.033e-01	2.378e-01	0.434	0.664027	
ORGANIZATION_TYPE_Construction	1.997e-01	4.730e-02	4.221	2.43e-05	***
ORGANIZATION_TYPE_Culture	5.531e-02	2.508e-01	0.221	0.825440	
ORGANIZATION_TYPE_Electricity	-7.926e-02	1.486e-01	-0.533	0.593808	
ORGANIZATION_TYPE_Emergency	-1.573e-01	1.839e-01	-0.855	0.392425	
ORGANIZATION_TYPE_Government	-1.368e-01	4.939e-02	-2.769	0.005615	**
ORGANIZATION_TYPE_Hotel	-4.305e-01	1.613e-01	-2.669	0.007613	**
ORGANIZATION_TYPE_Housing	-2.391e-02	7.973e-02	-0.300	0.764276	
ORGANIZATION_TYPE_Industry	-7.116e-02	3.761e-02	-1.892	0.058497	.
ORGANIZATION_TYPE_Insurance	-3.591e-01	2.292e-01	-1.567	0.117138	

```
ORGANIZATION_TYPE_Insurance                        -3.591e-01  2.292e-01  -1.567 0.117138
ORGANIZATION_TYPE_Kindergarten                     -1.050e-01  6.072e-02  -1.729 0.083895 .
`ORGANIZATION_TYPE_Legal Services`                  5.770e-01  2.384e-01   2.420 0.015503 *
ORGANIZATION_TYPE_Medicine                         -1.389e-01  6.043e-02  -2.299 0.021526 *
ORGANIZATION_TYPE_Military                         -4.636e-01  1.048e-01  -4.423 9.75e-06 ***
ORGANIZATION_TYPE_Mobile                           -7.958e-02  2.324e-01  -0.342 0.732047
ORGANIZATION_TYPE_Other                            -5.874e-02  3.768e-02  -1.559 0.119023
ORGANIZATION_TYPE_Police                           -3.656e-01  1.149e-01  -3.181 0.001468 **
ORGANIZATION_TYPE_Postal                            5.174e-02  9.201e-02   0.562 0.573875
ORGANIZATION_TYPE_Realtor                           6.869e-01  1.946e-01   3.529 0.000417 ***
ORGANIZATION_TYPE_Religion                         -1.427e-01  6.020e-01  -0.237 0.812621
ORGANIZATION_TYPE_Restaurant                        2.435e-02  9.222e-02   0.264 0.791722
ORGANIZATION_TYPE_School                           -1.929e-01  5.682e-02  -3.396 0.000684 ***
ORGANIZATION_TYPE_Security                         -1.600e-01  8.292e-02  -1.930 0.053585 .
`ORGANIZATION_TYPE_Security Ministries`            -3.423e-01  1.197e-01  -2.859 0.004246 **
`ORGANIZATION_TYPE_Self-employed`                   9.337e-02  2.556e-02   3.652 0.000260 ***
ORGANIZATION_TYPE_Services                         -1.024e-01  1.261e-01  -0.812 0.416792
ORGANIZATION_TYPE_Telecom                           6.387e-02  1.861e-01   0.343 0.731469
ORGANIZATION_TYPE_Trade                            -3.068e-02  3.873e-02  -0.792 0.428348
ORGANIZATION_TYPE_Transport                         2.506e-02  4.481e-02   0.559 0.575989
ORGANIZATION_TYPE_University                       -2.753e-01  1.506e-01  -1.828 0.067532 .
`DIES_MAX_IMPAGATS_Altres_EF_+120`                  1.036e-01  8.010e-02   1.293 0.195893
`DIES_MAX_IMPAGATS_Altres_EF_[1-30]`                8.128e-02  2.723e-02   2.985 0.002835 **
`DIES_MAX_IMPAGATS_Altres_EF_[31-60]`               7.176e-02  7.510e-02   0.956 0.339317
`DIES_MAX_IMPAGATS_Altres_EF_[61-90]`              -1.952e-01  1.542e-01  -1.266 0.205450
`DIES_MAX_IMPAGATS_Altres_EF_[91-120]`              2.155e-01  1.862e-01   1.158 0.246960
`Credits_Assegurats_HC_Menys crèdits assegurats`    1.896e-02  2.433e-02   0.779 0.435800
`Credits_Assegurats_HC_Més crèdits assegurats`      4.163e-02  2.943e-02   1.415 0.157147
`SUM_Diferencia_AMT_HC_More credit than ATB`        1.076e-01  2.209e-02   4.871 1.11e-06 ***
`SUM_Diferencia_AMT_HC_Same credit than ATB`        5.474e-02  3.814e-02   1.435 0.151147
LAST_DUE_MONTH_Less_Year                            1.380e-01  2.331e-02   5.921 3.20e-09 ***
LAST_DUE_MONTH_Between_1Year_2Years                 9.675e-02  2.402e-02   4.029 5.61e-05 ***
LAST_DUE_MONTH_Between_2Years_5Years                1.015e-01  2.704e-02   3.753 0.000175 ***
LAST_DUE_MONTH_More_5Years                          2.277e-01  3.939e-02   5.780 7.46e-09 ***
LAST_DUE_MONTH_Retired                              2.183e-01  3.365e-02   6.486 8.79e-11 ***
LAST_DUE_MONTH_No_DPD_PA                           -2.058e-01  6.032e-02  -3.412 0.000646 ***
DID_OVERDUE_Altres_EF:DID_PROLONG_Altres_EF         4.942e-01  2.390e-01   2.068 0.038639 *
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y                    7.370e-02  3.628e-02   2.032 0.042193 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 137558  on 245640  degrees of freedom
Residual deviance: 121002  on 245472  degrees of freedom
AIC: 121340

Number of Fisher Scoring iterations: 10
```

Figura 57. Resum del segon model logístic creat amb totes les variables.

8.3.3 Tercer model logístic amb totes les variables

```
> model_bo3 <-  glm(TARGET ~ . - NUM_Repeater_Client_HC - MAX_CREDIT_LIM_CC - CNT_FAM_MEMBERS - MAX
_EXT_SOURCE -
+                     MIN_EXT_SOURCE - SUM_BALANCE_CC - AMT_GOODS_PRICE - AMT_ANNUITY_TOT +
+                     DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +  FLAG_OWN_CAR_Y:FLAG_OWN_REALTY
_Y +
+                     TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC , data = train_x3, family = binomial)
> summary(model_bo3)

Call:
glm(formula = TARGET ~ . - NUM_Repeater_Client_HC - MAX_CREDIT_LIM_CC -
    CNT_FAM_MEMBERS - MAX_EXT_SOURCE - MIN_EXT_SOURCE - SUM_BALANCE_CC -
    AMT_GOODS_PRICE - AMT_ANNUITY_TOT + DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
    FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC,
    family = binomial, data = train_x3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7569  -0.4271  -0.3058  -0.2178   3.2873

Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                -2.098e+00  1.116e-01 -18.794  < 2e-16 ***
CNT_CHILDREN                                3.073e-02  1.122e-02   2.739 0.006164 **
AMT_INCOME_TOTAL                           -2.909e-07  1.097e-07  -2.653 0.007972 **
AMT_CREDIT                                  2.003e-07  2.443e-08   8.199 2.42e-16 ***
REGION_POPULATION_RELATIVE                  5.051e-01  7.144e-01   0.707 0.479581
AGE_EXPECTED                               -5.268e-03  1.010e-03  -5.217 1.82e-07 ***
ANY_COTXE_INTERACTION                       4.067e-03  1.120e-03   3.630 0.000283 ***
REGION_RATING_CLIENT                        1.509e-01  1.774e-02   8.504  < 2e-16 ***
MEAN_EXT_SOURCE                            -3.998e+00  5.511e-02 -72.554  < 2e-16 ***
CRE_SOL_Altres_EF                           1.872e-01  4.823e-02   3.881 0.000104 ***
CRE_BDE_Altres_EF                           5.194e-01  6.943e-01   0.748 0.454402
CRE_TOTAL_Altres_EF                        -1.419e-02  4.395e-03  -3.229 0.001242 **
TE_CREDIT_Altres_EF                        -8.734e-02  4.152e-02  -2.103 0.035430 *
CRE_Another_Type_LOAN_Altres_EF            -1.976e-01  1.579e-01  -1.252 0.210581
CAR_LOAN_Altres_EF                         -1.527e-01  3.880e-02  -3.936 8.29e-05 ***
CONSUMER_CREDIT_Altres_EF                  -2.298e-02  3.600e-02  -0.638 0.523359
CREDIT_CARD_Altres_EF                      -4.926e-02  2.163e-02  -2.278 0.022725 *
BUSINESS_LOAN_Altres_EF                    -2.221e-01  1.327e-01  -1.674 0.094065 .
MICROLOAN_Altres_EF                         5.772e-01  5.257e-02  10.979  < 2e-16 ***
MORTGAGE_LOAN_Altres_EF                    -3.998e-01  4.778e-02  -8.368  < 2e-16 ***
UNKNOWN_LOAN_Altres_EF                      5.815e-02  2.218e-01   0.262 0.793212
DID_OVERDUE_Altres_EF                       2.049e-01  5.968e-02   3.434 0.000596 ***
DID_PROLONG_Altres_EF                      -1.413e-02  4.799e-02  -0.295 0.768355
CASH_LOANS_TY_HC                            5.345e-02  2.160e-02   2.475 0.013321 *
CONS_LOANS_TY_HC                           -2.790e-02  3.266e-02  -0.854 0.393003
REV_LOANS_TY_HC                             1.024e-01  2.622e-02   3.907 9.36e-05 ***
NC_Goods_PURP_HC                            7.158e-02  3.630e-02   1.972 0.048647 *
Repairs_PURP_HC                             3.419e-02  1.736e-02   1.970 0.048879 *
UrgentN_PURP_HC                             8.012e-02  2.910e-02   2.753 0.005899 **
GET_HOME_PURP_HC                            6.159e-03  3.642e-02   0.169 0.865718
CARS_PURP_HC                               -4.199e-02  3.459e-02  -1.214 0.224749
FUN_PURP_HC                                -1.595e-04  6.661e-02  -0.002 0.998089
Approved_STATUS_HC                         -7.885e-02  7.762e-03 -10.159  < 2e-16 ***
Canceled_STATUS_HC                          7.458e-05  5.217e-03   0.014 0.988594
Refused_STATUS_HC                           6.552e-02  4.622e-03  14.176  < 2e-16 ***
UnusedOF_STATUS_HC                          1.129e-02  2.716e-02   0.416 0.677670
ElectD_CAT_HC                              -6.048e-02  9.118e-03  -6.633 3.28e-11 ***
Clothes_CAT_HC                             -1.083e-01  2.344e-02  -4.620 3.83e-06 ***
FreeT_CAT_HC                               -2.833e-01  7.265e-02  -3.899 9.65e-05 ***
Health_CAT_HC                              -3.385e-02  6.431e-02  -0.526 0.598627
Home_CAT_HC                                -1.096e-01  2.028e-02  -5.403 6.56e-08 ***
Mobile_CAT_HC                               5.902e-03  1.073e-02   0.550 0.582106
ConstructionM_CAT_HC                       -3.253e-02  2.548e-02  -1.277 0.201607
Vehicles_CAT_HC                            -3.823e-02  6.591e-02  -0.580 0.561940
IS_NEW_CLIENT_TRAIN                         1.031e-01  8.297e-02   1.242 0.214123
IS_Refreshed_Client_HC                     -1.262e-01  2.042e-02  -6.181 6.37e-10 ***
TOT_RECEIVABLE_CC                           2.365e-08  6.702e-09   3.528 0.000419 ***
TOT_DRAWINGS_CC                             8.986e-03  1.108e-03   8.111 5.03e-16 ***
TOT_INSTALMENTS_CC                         -2.062e-04  1.768e-05 -11.659  < 2e-16 ***
MEAN_DPD_CC                                 5.178e-04  9.165e-04   0.565 0.572087
TE_CREDIT_CARD_CC                          -5.921e-02  3.603e-02  -1.643 0.100379
TOT_CO_SIG_PCB_MENSUAL                      3.162e-03  1.198e-02   0.264 0.791809
TOT_CO_RTS_PCB_MENSUAL                      1.571e-01  5.381e-02   2.920 0.003500 **
```

TOT_CO_RTS_PCB_MENSUAL	1.571e-01	5.381e-02	2.920	0.003500	**
TOT_CO_ALT_PCB_MENSUAL	-1.114e-03	8.494e-03	-0.131	0.895685	
DIES_NOMES_GRAN_DEUTE_PCB	1.530e-06	5.754e-06	0.266	0.790327	
DIES_TOT_DEUTE_PCB	2.626e-06	1.377e-06	1.908	0.056455	.
NUM_QUOTES_PENDENTS	8.116e-03	8.851e-04	9.170	< 2e-16	***
LATE_PAYMENTS_IP	2.568e-01	4.390e-02	5.850	4.91e-09	***
NOT_ENOUGH_MONEY_PAID_IP	3.058e-01	1.715e-02	17.833	< 2e-16	***
PERCENTATGE_ANNUITY	1.173e-04	4.804e-05	2.442	0.014625	*
AMT_CREDIT_ACT_TOT	1.388e-08	3.573e-09	3.885	0.000102	***
RATI_DEUTE_GARANTIA	1.071e+00	6.140e-02	17.443	< 2e-16	***
NUM_ACTIVE_CREDITS	9.442e-02	7.293e-03	12.947	< 2e-16	***
NUM_CREDITS_PREVIS_TANCATS	-7.220e-03	3.285e-03	-2.198	0.027959	*
`NAME_CONTRACT_TYPE_Revolving loans`	-1.570e-01	3.371e-02	-4.656	3.23e-06	***
CODE_GENDER_M	2.912e-01	2.032e-02	14.333	< 2e-16	***
FLAG_OWN_CAR_Y	-3.195e-01	3.496e-02	-9.141	< 2e-16	***
FLAG_OWN_REALTY_Y	1.634e-02	2.098e-02	0.779	0.436113	
NAME_TYPE_SUITE_Children	8.145e-02	7.782e-02	1.047	0.295237	
NAME_TYPE_SUITE_Family	8.312e-03	2.380e-02	0.349	0.726916	
`NAME_TYPE_SUITE_Group of people`	5.737e-02	2.480e-01	0.231	0.817057	
NAME_TYPE_SUITE_Other	5.362e-02	7.922e-02	0.677	0.498538	
`NAME_TYPE_SUITE_Spouse, partner`	-6.282e-02	4.221e-02	-1.488	0.136693	
NAME_INCOME_TYPE_Businessman	-7.846e+00	6.911e+01	-0.114	0.909612	
`NAME_INCOME_TYPE_Maternity leave`	2.638e+00	1.204e+00	2.191	0.028478	*
NAME_INCOME_TYPE_Pensioner	-2.735e-01	4.110e-02	-6.654	2.84e-11	***
`NAME_INCOME_TYPE_State servant`	1.345e-02	4.217e-02	0.319	0.749857	
NAME_INCOME_TYPE_Student	-9.354e+00	4.373e+01	-0.214	0.830634	
NAME_INCOME_TYPE_Unemployed	1.168e+00	6.124e-01	1.907	0.056456	.
NAME_INCOME_TYPE_Working	1.112e-01	2.014e-02	5.520	3.38e-08	***
`NAME_EDUCATION_TYPE_Academic degree`	-1.294e+00	5.980e-01	-2.164	0.030430	*
`NAME_EDUCATION_TYPE_Higher education`	-2.928e-01	2.270e-02	-12.899	< 2e-16	***
`NAME_EDUCATION_TYPE_Incomplete higher`	-2.220e-01	4.322e-02	-5.137	2.78e-07	***
`NAME_EDUCATION_TYPE_Lower secondary`	7.686e-02	6.226e-02	1.235	0.217007	
`NAME_FAMILY_STATUS_Civil marriage`	1.247e-01	2.554e-02	4.882	1.05e-06	***
NAME_FAMILY_STATUS_Separated	1.372e-01	3.253e-02	4.218	2.47e-05	***
`NAME_FAMILY_STATUS_Single / not married`	7.397e-02	2.304e-02	3.210	0.001325	**
NAME_FAMILY_STATUS_Widow	-1.978e-02	4.228e-02	-0.468	0.639839	
`NAME_HOUSING_TYPE_Co-op apartment`	-3.161e-02	1.316e-01	-0.240	0.810090	
`NAME_HOUSING_TYPE_Municipal apartment`	8.120e-02	4.141e-02	1.961	0.049906	*
`NAME_HOUSING_TYPE_Office apartment`	-2.528e-01	9.512e-02	-2.657	0.007876	**
`NAME_HOUSING_TYPE_Rented apartment`	1.420e-01	5.222e-02	2.720	0.006532	**
`NAME_HOUSING_TYPE_With parents`	9.135e-03	3.295e-02	0.277	0.781626	
`LAST_YEARS_EMPLOYED_4-9`	-1.558e-01	1.961e-02	-7.944	1.96e-15	***
`LAST_YEARS_EMPLOYED_10-80`	-3.474e-01	2.847e-02	-12.201	< 2e-16	***
`OCCUPATION_TYPE_Cleaning staff`	1.136e-01	6.250e-02	1.818	0.069087	.
`OCCUPATION_TYPE_Cooking staff`	1.233e-01	5.549e-02	2.222	0.026269	*
`OCCUPATION_TYPE_Core staff`	-2.584e-02	3.846e-02	-0.672	0.501668	
OCCUPATION_TYPE_Drivers	1.861e-01	3.654e-02	5.094	3.50e-07	***
`OCCUPATION_TYPE_HR staff`	-1.708e-01	2.190e-01	-0.780	0.435464	
`OCCUPATION_TYPE_High skill tech staff`	-7.568e-02	4.927e-02	-1.536	0.124523	
`OCCUPATION_TYPE_IT staff`	8.521e-02	2.034e-01	0.419	0.675256	
OCCUPATION_TYPE_Laborers	9.327e-02	2.708e-02	3.445	0.000572	***
`OCCUPATION_TYPE_Low-skill Laborers`	2.376e-01	7.267e-02	3.270	0.001074	**
OCCUPATION_TYPE_Managers	6.251e-02	3.968e-02	1.575	0.115203	
`OCCUPATION_TYPE_Medicine staff`	-5.877e-03	6.785e-02	-0.087	0.930981	
`OCCUPATION_TYPE_Private service staff`	5.775e-04	9.565e-02	0.006	0.995183	
`OCCUPATION_TYPE_Realty agents`	-5.056e-02	1.646e-01	-0.307	0.758654	
`OCCUPATION_TYPE_Sales staff`	5.141e-02	3.222e-02	1.596	0.110587	
OCCUPATION_TYPE_Secretaries	2.070e-01	1.246e-01	1.660	0.096859	.
`OCCUPATION_TYPE_Security staff`	1.250e-01	5.956e-02	2.098	0.035880	*
`OCCUPATION_TYPE_Waiters/barmen staff`	1.955e-01	1.042e-01	1.877	0.060578	.
ORGANIZATION_TYPE_Advertising	1.584e-01	1.993e-01	0.795	0.426824	
ORGANIZATION_TYPE_Agriculture	2.004e-02	7.920e-02	0.253	0.800264	
ORGANIZATION_TYPE_Bank	-3.665e-01	1.092e-01	-3.355	0.000794	***
ORGANIZATION_TYPE_Cleaning	1.004e-01	2.379e-01	0.422	0.673085	
ORGANIZATION_TYPE_Construction	2.015e-01	4.732e-02	4.259	2.05e-05	***
ORGANIZATION_TYPE_Culture	5.068e-02	2.508e-01	0.202	0.839855	
ORGANIZATION_TYPE_Electricity	-7.986e-02	1.486e-01	-0.537	0.590923	
ORGANIZATION_TYPE_Emergency	-1.601e-01	1.840e-01	-0.870	0.384267	
ORGANIZATION_TYPE_Government	-1.409e-01	4.943e-02	-2.851	0.004364	**
ORGANIZATION_TYPE_Hotel	-4.316e-01	1.614e-01	-2.674	0.007499	**
ORGANIZATION_TYPE_Housing	-2.266e-02	7.974e-02	-0.284	0.776313	
ORGANIZATION_TYPE_Industry	-7.046e-02	3.761e-02	-1.873	0.061030	.
ORGANIZATION_TYPE_Insurance	-3.570e-01	2.289e-01	-1.560	0.118826	
ORGANIZATION_TYPE_Kindergarten	-1.073e-01	6.077e-02	-1.766	0.077390	.

```
ORGANIZATION_TYPE_Kindergarten                  -1.073e-01  6.077e-02  -1.766 0.077390 .
`ORGANIZATION_TYPE_Legal Services`               5.781e-01  2.382e-01   2.427 0.015210 *
ORGANIZATION_TYPE_Medicine                      -1.400e-01  6.045e-02  -2.315 0.020592 *
ORGANIZATION_TYPE_Military                      -4.578e-01  1.048e-01  -4.366 1.26e-05 ***
ORGANIZATION_TYPE_Mobile                        -8.345e-02  2.324e-01  -0.359 0.719510
ORGANIZATION_TYPE_Other                         -6.101e-02  3.769e-02  -1.619 0.105486
ORGANIZATION_TYPE_Police                        -3.651e-01  1.149e-01  -3.177 0.001489 **
ORGANIZATION_TYPE_Postal                         4.638e-02  9.206e-02   0.504 0.614429
ORGANIZATION_TYPE_Realtor                        6.877e-01  1.943e-01   3.539 0.000402 ***
ORGANIZATION_TYPE_Religion                      -1.447e-01  6.021e-01  -0.240 0.810119
ORGANIZATION_TYPE_Restaurant                     2.579e-02  9.225e-02   0.280 0.779791
ORGANIZATION_TYPE_School                        -1.981e-01  5.687e-02  -3.483 0.000495 ***
ORGANIZATION_TYPE_Security                      -1.584e-01  8.293e-02  -1.911 0.056055 .
`ORGANIZATION_TYPE_Security Ministries`         -3.357e-01  1.197e-01  -2.805 0.005039 **
`ORGANIZATION_TYPE_Self-employed`                9.276e-02  2.557e-02   3.628 0.000286 ***
ORGANIZATION_TYPE_Services                      -1.036e-01  1.261e-01  -0.822 0.411264
ORGANIZATION_TYPE_Telecom                        6.964e-02  1.860e-01   0.374 0.708048
ORGANIZATION_TYPE_Trade                         -3.360e-02  3.874e-02  -0.867 0.385762
ORGANIZATION_TYPE_Transport                      2.625e-02  4.481e-02   0.586 0.557987
ORGANIZATION_TYPE_University                    -2.844e-01  1.510e-01  -1.884 0.059557 .
`DIES_MAX_IMPAGATS_Altres_EF_+120`               1.091e-01  8.008e-02   1.362 0.173090
`DIES_MAX_IMPAGATS_Altres_EF_[1-30]`             8.450e-02  2.723e-02   3.104 0.001911 **
`DIES_MAX_IMPAGATS_Altres_EF_[31-60]`            7.600e-02  7.508e-02   1.012 0.311399
`DIES_MAX_IMPAGATS_Altres_EF_[61-90]`           -1.837e-01  1.540e-01  -1.193 0.232763
`DIES_MAX_IMPAGATS_Altres_EF_[91-120]`           2.173e-01  1.862e-01   1.167 0.243200
`Credits_Assegurats_HC_Menys crèdits assegurats` 1.741e-02  2.432e-02   0.716 0.474084
`Credits_Assegurats_HC_Més crèdits assegurats`   4.150e-02  2.943e-02   1.410 0.158533
`SUM_Diferencia_AMT_HC_More credit than ATB`     1.074e-01  2.209e-02   4.860 1.17e-06 ***
`SUM_Diferencia_AMT_HC_Same credit than ATB`     5.184e-02  3.814e-02   1.359 0.174067
LAST_DUE_MONTH_Less_Year                         1.378e-01  2.330e-02   5.913 3.36e-09 ***
LAST_DUE_MONTH_Between_1Year_2Years              9.593e-02  2.400e-02   3.996 6.43e-05 ***
LAST_DUE_MONTH_Between_2Years_5Years             1.037e-01  2.702e-02   3.838 0.000124 ***
LAST_DUE_MONTH_More_5Years                       2.297e-01  3.938e-02   5.831 5.50e-09 ***
LAST_DUE_MONTH_Retired                           2.205e-01  3.365e-02   6.553 5.65e-11 ***
LAST_DUE_MONTH_No_DPD_PA                        -2.067e-01  6.034e-02  -3.426 0.000613 ***
DID_OVERDUE_Altres_EF:DID_PROLONG_Altres_EF      5.003e-01  2.393e-01   2.091 0.036538 *
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y                 7.382e-02  3.627e-02   2.035 0.041846 *
MAX_CREDIT_LIM_CC:TE_CREDIT_CARD_CC             -2.599e-07  8.876e-08  -2.929 0.003405 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 137553  on 245639  degrees of freedom
Residual deviance: 121057  on 245477  degrees of freedom
AIC: 121383

Number of Fisher Scoring iterations: 10
```

Figura 58. Resum del tercer model logístic creat amb totes les variables.

8.3.4 Model logístic amb menys variables

```
> model_bo2 <-  glm(TARGET ~ .  -  FLAG_OWN_REALTY_Y - NUM_Repeater_Client_HC - MAX_CREDIT_
LIM_CC - CNT_FAM_MEMBERS - MAX_EXT_SOURCE - MIN_EXT_SOURCE - SUM_BALANCE_CC - AMT_GOODS_PRI
CE - AMT_ANNUITY_TOT + DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +  FLAG_OWN_CAR_Y:FLAG_O
WN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC , data = train_x5, family = binomial)
> summary(model_bo2)

Call:
glm(formula = TARGET ~ . - FLAG_OWN_REALTY_Y - NUM_Repeater_Client_HC -
    MAX_CREDIT_LIM_CC - CNT_FAM_MEMBERS - MAX_EXT_SOURCE - MIN_EXT_SOURCE -
    SUM_BALANCE_CC - AMT_GOODS_PRICE - AMT_ANNUITY_TOT + DID_PROLONG_Altres_EF:DID_OVERDUE_
Altres_EF +
    FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC,
    family = binomial, data = train_x5)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7686  -0.4273  -0.3059  -0.2180   3.2941

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -2.073e+00  9.965e-02 -20.805  < 2e-16 ***
CNT_CHILDREN                             3.046e-02  1.121e-02   2.718 0.006563 **
AMT_INCOME_TOTAL                        -2.747e-07  1.087e-07  -2.527 0.011513 *
AMT_CREDIT                               1.991e-07  2.438e-08   8.165 3.21e-16 ***
AGE_EXPECTED                            -5.014e-03  9.901e-04  -5.064 4.10e-07 ***
ANY_COTXE_INTERACTION                    4.085e-03  1.116e-03   3.661 0.000251 ***
REGION_RATING_CLIENT                     1.429e-01  1.607e-02   8.894  < 2e-16 ***
MEAN_EXT_SOURCE                         -4.001e+00  5.502e-02 -72.726  < 2e-16 ***
CRE_SOL_Altres_EF                        1.878e-01  4.819e-02   3.897 9.73e-05 ***
CRE_TOTAL_Altres_EF                     -1.446e-02  4.326e-03  -3.342 0.000832 ***
TE_CREDIT_Altres_EF                     -1.079e-01  2.547e-02  -4.238 2.25e-05 ***
CAR_LOAN_Altres_EF                      -1.503e-01  3.867e-02  -3.886 0.000102 ***
CREDIT_CARD_Altres_EF                   -4.411e-02  2.097e-02  -2.103 0.035434 *
MICROLOAN_Altres_EF                      5.804e-01  5.237e-02  11.083  < 2e-16 ***
MORTGAGE_LOAN_Altres_EF                 -3.963e-01  4.761e-02  -8.324  < 2e-16 ***
DID_OVERDUE_Altres_EF                    2.072e-01  5.958e-02   3.478 0.000505 ***
DID_PROLONG_Altres_EF                   -1.242e-02  4.796e-02  -0.259 0.795618
REV_LOANS_TY_HC                          1.050e-01  2.603e-02   4.034 5.49e-05 ***
NC_Goods_PURP_HC                         7.819e-02  3.607e-02   2.168 0.030174 *
Repairs_PURP_HC                          3.783e-02  1.718e-02   2.201 0.027708 *
UrgentN_PURP_HC                          8.555e-02  2.896e-02   2.954 0.003133 **
Approved_STATUS_HC                      -7.308e-02  6.331e-03 -11.544  < 2e-16 ***
Refused_STATUS_HC                        6.659e-02  4.232e-03  15.736  < 2e-16 ***
ElectD_CAT_HC                           -6.690e-02  7.848e-03  -8.525  < 2e-16 ***
Clothes_CAT_HC                          -1.160e-01  2.298e-02  -5.046 4.50e-07 ***
FreeT_CAT_HC                            -2.913e-01  7.255e-02  -4.015 5.94e-05 ***
Home_CAT_HC                             -1.190e-01  1.956e-02  -6.081 1.19e-09 ***
IS_Refreshed_Client_HC                  -1.185e-01  2.001e-02  -5.922 3.18e-09 ***
TOT_RECEIVABLE_CC                        2.363e-08  6.675e-09   3.541 0.000399 ***
TOT_DRAWINGS_CC                          8.994e-03  1.107e-03   8.125 4.47e-16 ***
TOT_INSTALMENTS_CC                      -2.029e-04  1.759e-05 -11.533  < 2e-16 ***
TE_CREDIT_CARD_CC                       -6.323e-02  3.561e-02  -1.775 0.075836 .
TOT_CO_RTS_PCB_MENSUAL                   1.539e-01  5.354e-02   2.875 0.004035 **
DIES_TOT_DEUTE_PCB                       2.781e-06  1.284e-06   2.166 0.030333 *
NUM_QUOTES_PENDENTS                      8.272e-03  8.768e-04   9.435  < 2e-16 ***
LATE_PAYMENTS_IP                         2.555e-01  4.383e-02   5.829 5.57e-09 ***
NOT_ENOUGH_MONEY_PAID_IP                 3.076e-01  1.705e-02  18.040  < 2e-16 ***
PERCENTATGE_ANNUITY                      1.176e-04  4.776e-05   2.462 0.013800 *
AMT_CREDIT_ACT_TOT                       1.367e-08  3.543e-09   3.857 0.000115 ***
RATI_DEUTE_GARANTIA                      1.075e+00  6.123e-02  17.562  < 2e-16 ***
NUM_ACTIVE_CREDITS                       9.388e-02  7.271e-03  12.911  < 2e-16 ***
NUM_CREDITS_PREVIS_TANCATS              -7.289e-03  3.273e-03  -2.227 0.025953 *
`NAME_CONTRACT_TYPE_Revolving loans`    -1.573e-01  3.357e-02  -4.687 2.77e-06 ***
CODE_GENDER_M                            2.891e-01  2.024e-02  14.281  < 2e-16 ***
FLAG_OWN_CAR_Y                          -3.291e-01  3.237e-02 -10.165  < 2e-16 ***
NAME_INCOME_TYPE_Businessman            -7.820e+00  6.911e+01  -0.113 0.909901
`NAME_INCOME_TYPE_Maternity leave`       2.632e+00  1.206e+00   2.183 0.029000 *
NAME_INCOME_TYPE_Pensioner              -2.787e-01  4.102e-02  -6.793 1.10e-11 ***
`NAME_INCOME_TYPE_State servant`         9.575e-03  4.212e-02   0.227 0.820187
NAME_INCOME_TYPE_Student                -9.350e+00  4.366e+01  -0.214 0.830426
NAME_INCOME_TYPE_Unemployed              1.179e+00  6.136e-01   1.922 0.054660 .
NAME_INCOME_TYPE_Working                 1.076e-01  2.007e-02   5.361 8.27e-08 ***
`NAME_EDUCATION_TYPE_Academic degree`   -1.301e+00  5.983e-01  -2.175 0.029646 *
`NAME_EDUCATION_TYPE_Higher education`  -2.916e-01  2.266e-02 -12.868  < 2e-16 ***
```

`NAME_EDUCATION_TYPE_Higher education`	-2.916e-01	2.266e-02	-12.868	< 2e-16	***
`NAME_EDUCATION_TYPE_Incomplete higher`	-2.198e-01	4.320e-02	-5.088	3.62e-07	***
`NAME_EDUCATION_TYPE_Lower secondary`	7.533e-02	6.223e-02	1.211	0.226086	
`NAME_FAMILY_STATUS_Civil marriage`	1.250e-01	2.552e-02	4.897	9.73e-07	***
NAME_FAMILY_STATUS_Separated	1.401e-01	3.243e-02	4.321	1.55e-05	***
`NAME_FAMILY_STATUS_Single / not married`	7.613e-02	2.291e-02	3.323	0.000890	***
NAME_FAMILY_STATUS_Widow	-1.718e-02	4.220e-02	-0.407	0.683885	
`NAME_HOUSING_TYPE_Co-op apartment`	-3.114e-02	1.315e-01	-0.237	0.812809	
`NAME_HOUSING_TYPE_Municipal apartment`	7.833e-02	4.096e-02	1.912	0.055868	.
`NAME_HOUSING_TYPE_Office apartment`	-2.583e-01	9.507e-02	-2.717	0.006588	**
`NAME_HOUSING_TYPE_Rented apartment`	1.391e-01	5.199e-02	2.675	0.007463	**
`NAME_HOUSING_TYPE_With parents`	5.924e-03	3.260e-02	0.182	0.855801	
`LAST_YEARS_EMPLOYED_4-9`	-1.560e-01	1.960e-02	-7.957	1.77e-15	***
`LAST_YEARS_EMPLOYED_10-80`	-3.477e-01	2.846e-02	-12.217	< 2e-16	***
`OCCUPATION_TYPE_Cleaning staff`	1.162e-01	6.247e-02	1.859	0.062982	.
`OCCUPATION_TYPE_Cooking staff`	1.248e-01	5.547e-02	2.251	0.024395	*
`OCCUPATION_TYPE_Core staff`	-2.509e-02	3.845e-02	-0.653	0.514042	
OCCUPATION_TYPE_Drivers	1.855e-01	3.652e-02	5.079	3.79e-07	***
`OCCUPATION_TYPE_HR staff`	-1.616e-01	2.188e-01	-0.739	0.460107	
`OCCUPATION_TYPE_High skill tech staff`	-7.507e-02	4.926e-02	-1.524	0.127492	
`OCCUPATION_TYPE_IT staff`	8.861e-02	2.033e-01	0.436	0.662859	
OCCUPATION_TYPE_Laborers	9.316e-02	2.707e-02	3.441	0.000579	***
`OCCUPATION_TYPE_Low-skill Laborers`	2.370e-01	7.263e-02	3.263	0.001103	**
OCCUPATION_TYPE_Managers	6.142e-02	3.966e-02	1.549	0.121456	
`OCCUPATION_TYPE_Medicine staff`	-5.152e-03	6.783e-02	-0.076	0.939454	
`OCCUPATION_TYPE_Private service staff`	1.520e-03	9.561e-02	0.016	0.987313	
`OCCUPATION_TYPE_Realty agents`	-4.627e-02	1.645e-01	-0.281	0.778457	
`OCCUPATION_TYPE_Sales staff`	5.215e-02	3.221e-02	1.619	0.105435	
OCCUPATION_TYPE_Secretaries	2.087e-01	1.246e-01	1.675	0.093898	.
`OCCUPATION_TYPE_Security staff`	1.245e-01	5.955e-02	2.090	0.036599	*
`OCCUPATION_TYPE_Waiters/barmen staff`	1.957e-01	1.042e-01	1.879	0.060245	.
ORGANIZATION_TYPE_Advertising	1.606e-01	1.993e-01	0.806	0.420297	
ORGANIZATION_TYPE_Agriculture	1.538e-02	7.917e-02	0.194	0.845986	
ORGANIZATION_TYPE_Bank	-3.673e-01	1.092e-01	-3.363	0.000772	***
ORGANIZATION_TYPE_Cleaning	1.019e-01	2.378e-01	0.429	0.668202	
ORGANIZATION_TYPE_Construction	2.012e-01	4.730e-02	4.253	2.11e-05	***
ORGANIZATION_TYPE_Culture	4.881e-02	2.507e-01	0.195	0.845630	
ORGANIZATION_TYPE_Electricity	-8.051e-02	1.485e-01	-0.542	0.587772	
ORGANIZATION_TYPE_Emergency	-1.614e-01	1.839e-01	-0.878	0.380211	
ORGANIZATION_TYPE_Government	-1.435e-01	4.942e-02	-2.903	0.003697	**
ORGANIZATION_TYPE_Hotel	-4.285e-01	1.613e-01	-2.656	0.007898	**
ORGANIZATION_TYPE_Housing	-2.442e-02	7.972e-02	-0.306	0.759384	
ORGANIZATION_TYPE_Industry	-7.124e-02	3.760e-02	-1.895	0.058134	.
ORGANIZATION_TYPE_Insurance	-3.577e-01	2.289e-01	-1.563	0.118075	
ORGANIZATION_TYPE_Kindergarten	-1.079e-01	6.075e-02	-1.776	0.075798	.
`ORGANIZATION_TYPE_Legal Services`	5.774e-01	2.380e-01	2.426	0.015273	*
ORGANIZATION_TYPE_Medicine	-1.400e-01	6.043e-02	-2.316	0.020557	*
ORGANIZATION_TYPE_Military	-4.603e-01	1.048e-01	-4.392	1.12e-05	***
ORGANIZATION_TYPE_Mobile	-8.994e-02	2.323e-01	-0.387	0.698660	
ORGANIZATION_TYPE_Other	-6.248e-02	3.767e-02	-1.658	0.097243	.
ORGANIZATION_TYPE_Police	-3.653e-01	1.149e-01	-3.179	0.001477	**
ORGANIZATION_TYPE_Postal	4.626e-02	9.204e-02	0.503	0.615255	
ORGANIZATION_TYPE_Realtor	6.904e-01	1.942e-01	3.555	0.000378	***
ORGANIZATION_TYPE_Religion	-1.461e-01	6.020e-01	-0.243	0.808261	
ORGANIZATION_TYPE_Restaurant	2.346e-02	9.223e-02	0.254	0.799212	
ORGANIZATION_TYPE_School	-2.004e-01	5.685e-02	-3.525	0.000423	***
ORGANIZATION_TYPE_Security	-1.575e-01	8.291e-02	-1.900	0.057393	.
`ORGANIZATION_TYPE_Security Ministries`	-3.355e-01	1.196e-01	-2.804	0.005040	**
`ORGANIZATION_TYPE_Self-employed`	9.158e-02	2.553e-02	3.587	0.000334	***
ORGANIZATION_TYPE_Services	-1.067e-01	1.261e-01	-0.846	0.397677	
ORGANIZATION_TYPE_Telecom	6.684e-02	1.859e-01	0.360	0.719208	
ORGANIZATION_TYPE_Trade	-3.285e-02	3.872e-02	-0.848	0.396259	
ORGANIZATION_TYPE_Transport	2.657e-02	4.480e-02	0.593	0.553045	
ORGANIZATION_TYPE_University	-2.860e-01	1.510e-01	-1.894	0.058198	.
`DIES_MAX_IMPAGATS_Altres_EF_+120`	1.012e-01	8.004e-02	1.265	0.205993	
`DIES_MAX_IMPAGATS_Altres_EF_[1-30]`	8.478e-02	2.720e-02	3.117	0.001828	**
`DIES_MAX_IMPAGATS_Altres_EF_[31-60]`	7.540e-02	7.506e-02	1.005	0.315078	
`DIES_MAX_IMPAGATS_Altres_EF_[61-90]`	-1.857e-01	1.539e-01	-1.207	0.227579	
`DIES_MAX_IMPAGATS_Altres_EF_[91-120]`	2.051e-01	1.859e-01	1.103	0.270090	
`SUM_Diferencia_AMT_HC_More credit than ATB`	1.261e-01	2.086e-02	6.044	1.51e-09	***
`SUM_Diferencia_AMT_HC_Same credit than ATB`	7.114e-02	3.521e-02	2.021	0.043312	*
LAST_DUE_MONTH_Less_Year	1.385e-01	2.328e-02	5.947	2.74e-09	***
LAST_DUE_MONTH_Between_1Year_2Years	9.601e-02	2.392e-02	4.013	5.99e-05	***
LAST_DUE_MONTH_Between_2Years_5Years	9.871e-02	2.680e-02	3.683	0.000231	***

```
LAST_DUE_MONTH_More_5Years                    2.305e-01  3.855e-02   5.980 2.23e-09 ***
LAST_DUE_MONTH_Retired                        2.175e-01  3.315e-02   6.563 5.27e-11 ***
LAST_DUE_MONTH_No_DPD_PA                     -1.522e-01  4.211e-02  -3.615 0.000300 ***
DID_OVERDUE_Altres_EF:DID_PROLONG_Altres_EF   4.959e-01  2.393e-01   2.072 0.038228 *
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y              8.669e-02  3.050e-02   2.843 0.004471 **
MAX_CREDIT_LIM_CC:TE_CREDIT_CARD_CC          -2.593e-07  8.862e-08  -2.927 0.003428 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 137553  on 245639  degrees of freedom
Residual deviance: 121086  on 245507  degrees of freedom
AIC: 121352

Number of Fisher Scoring iterations: 10
```

Figura 59. Resum del model logístic amb les variables significatives

8.3.5 Model Lasso amb totes les variables

```
coef(lasso.model2)[,70]
```

(Intercept)	-2.045899e+00
AMT_INCOME_TOTAL	-2.398569e-07
REGION_POPULATION_RELATIVE	2.453981e-01
ANY_COTXE_INTERACTION	3.637512e-03
MEAN_EXT_SOURCE	-4.006899e+00
CRE_BDE_Altres_EF	4.196901e-01
TE_CREDIT_Altres_EF	-8.845651e-02
CAR_LOAN_Altres_EF	-1.479758e-01
CREDIT_CARD_Altres_EF	-4.231088e-02
MICROLOAN_Altres_EF	5.726408e-01
UNKNOWN_LOAN_Altres_EF	4.952549e-03
DID_PROLONG_Altres_EF	-2.917162e-03
CONS_LOANS_TY_HC	-3.168588e-02
NC_Goods_PURP_HC	6.738017e-02
UrgentN_PURP_HC	7.742929e-02
CARS_PURP_HC	-3.575749e-02
Approved_STATUS_HC	-7.699119e-02
Refused_STATUS_HC	6.559219e-02
ElectD_CAT_HC	-5.933861e-02
FreeT_CAT_HC	-2.666874e-01
Home_CAT_HC	-1.069032e-01
ConstructionM_CAT_HC	-2.914754e-02
IS_NEW_CLIENT_TRAIN	2.808420e-02
TOT_RECEIVABLE_CC	2.111360e-08
TOT_INSTALMENTS_CC	-1.980685e-04
TE_CREDIT_CARD_CC	-4.531924e-02
TOT_CO_RTS_PCB_MENSUAL	1.486877e-01
DIES_NOMES_GRAN_DEUTE_PCB	6.540590e-07
NUM_QUOTES_PENDENTS	8.054780e-03
NOT_ENOUGH_MONEY_PAID_IP	3.034060e-01
AMT_CREDIT_ACT_TOT	1.278324e-08
NUM_ACTIVE_CREDITS	8.882608e-02
`NAME_CONTRACT_TYPE_Revolving loans`	-1.522645e-01
FLAG_OWN_CAR_Y	-2.946554e-01
NAME_TYPE_SUITE_Children	6.480066e-02
`NAME_TYPE_SUITE_Group of people`	1.673568e-02
CNT_CHILDREN	2.792943e-02
AMT_CREDIT	1.880831e-07
AGE_EXPECTED	-5.219835e-03
REGION_RATING_CLIENT	1.447308e-01
CRE_SOL_Altres_EF	1.755344e-01
CRE_TOTAL_Altres_EF	-1.181481e-02
CRE_Another_Type_LOAN_Altres_EF	-1.648615e-01
CONSUMER_CREDIT_Altres_EF	-1.874401e-02
BUSINESS_LOAN_Altres_EF	-1.923881e-01
MORTGAGE_LOAN_Altres_EF	-3.861167e-01
DID_OVERDUE_Altres_EF	2.019302e-01
CASH_LOANS_TY_HC	4.701647e-02
REV_LOANS_TY_HC	9.070996e-02
Repairs_PURP_HC	3.286735e-02
GET_HOME_PURP_HC	1.438796e-03
FUN_PURP_HC	0.000000e+00
Canceled_STATUS_HC	0.000000e+00
UnusedOF_STATUS_HC	7.008999e-03
Clothes_CAT_HC	-1.055072e-01
Health_CAT_HC	-2.112317e-02
Mobile_CAT_HC	4.819566e-03
Vehicles_CAT_HC	-2.649844e-02
IS_Refreshed_Client_HC	-1.219195e-01
TOT_DRAWINGS_CC	8.829482e-03
MEAN_DPD_CC	4.423992e-04
TOT_CO_SIG_PCB_MENSUAL	6.902207e-04
TOT_CO_ALT_PCB_MENSUAL	0.000000e+00
DIES_TOT_DEUTE_PCB	2.547830e-06
LATE_PAYMENTS_IP	2.521423e-01
PERCENTATGE_ANNUITY	1.121322e-04
RATI_DEUTE_GARANTIA	1.073469e+00
NUM_CREDITS_PREVIS_TANCATS	-8.379764e-03
CODE_GENDER_M	2.877867e-01
FLAG_OWN_REALTY_Y	1.682644e-02
NAME_TYPE_SUITE_Family	2.626143e-03
NAME_TYPE_SUITE_Other	4.020144e-02

`NAME_TYPE_SUITE_Group of people`	1.673568e-02
NAME_TYPE_SUITE_Other	4.020144e-02
`NAME_TYPE_SUITE_Spouse, partner`	-5.441977e-02
NAME_INCOME_TYPE_Businessman	-3.293724e-01
`NAME_INCOME_TYPE_Maternity leave`	2.472344e+00
NAME_INCOME_TYPE_Pensioner	-2.687088e-01
`NAME_INCOME_TYPE_State servant`	0.000000e+00
NAME_INCOME_TYPE_Student	-2.384964e+00
NAME_INCOME_TYPE_Unemployed	1.111300e+00
NAME_INCOME_TYPE_Working	1.089547e-01
`NAME_EDUCATION_TYPE_Academic degree`	-1.114670e+00
`NAME_EDUCATION_TYPE_Higher education`	-2.908191e-01
`NAME_EDUCATION_TYPE_Incomplete higher`	-2.131658e-01
`NAME_EDUCATION_TYPE_Lower secondary`	6.874767e-02
`NAME_FAMILY_STATUS_Civil marriage`	1.183869e-01
NAME_FAMILY_STATUS_Separated	1.283591e-01
`NAME_FAMILY_STATUS_Single / not married`	6.782322e-02
NAME_FAMILY_STATUS_Widow	-1.642619e-02
`NAME_HOUSING_TYPE_Co-op apartment`	-2.963514e-03
`NAME_HOUSING_TYPE_Municipal apartment`	7.317232e-02
`NAME_HOUSING_TYPE_Office apartment`	-2.383211e-01
`NAME_HOUSING_TYPE_Rented apartment`	1.345975e-01
`NAME_HOUSING_TYPE_With parents`	4.921237e-03
`LAST_YEARS_EMPLOYED_4-9`	-1.510060e-01
`LAST_YEARS_EMPLOYED_10-80`	-3.424915e-01
`OCCUPATION_TYPE_Cleaning staff`	9.552088e-02
`OCCUPATION_TYPE_Cooking staff`	1.080705e-01
`OCCUPATION_TYPE_Core staff`	-3.581644e-02
OCCUPATION_TYPE_Drivers	1.748814e-01
`OCCUPATION_TYPE_HR staff`	-1.290287e-01
`OCCUPATION_TYPE_High skill tech staff`	-7.530877e-02
`OCCUPATION_TYPE_IT staff`	3.465276e-02
OCCUPATION_TYPE_Laborers	8.629988e-02
`OCCUPATION_TYPE_Low-skill Laborers`	2.271847e-01
OCCUPATION_TYPE_Managers	4.445172e-02
`OCCUPATION_TYPE_Medicine staff`	-1.176495e-02
`OCCUPATION_TYPE_Private service staff`	0.000000e+00
`OCCUPATION_TYPE_Realty agents`	-2.199143e-02
`OCCUPATION_TYPE_Sales staff`	4.159227e-02
OCCUPATION_TYPE_Secretaries	1.721378e-01
`OCCUPATION_TYPE_Security staff`	1.007458e-01
`OCCUPATION_TYPE_Waiters/barmen staff`	1.754857e-01
ORGANIZATION_TYPE_Advertising	1.284917e-01
ORGANIZATION_TYPE_Agriculture	1.717524e-02
ORGANIZATION_TYPE_Bank	-3.346439e-01
ORGANIZATION_TYPE_Cleaning	7.299306e-02
ORGANIZATION_TYPE_Construction	2.022421e-01
ORGANIZATION_TYPE_Culture	4.682191e-03
ORGANIZATION_TYPE_Electricity	-4.496546e-02
ORGANIZATION_TYPE_Emergency	-1.167923e-01
ORGANIZATION_TYPE_Government	-1.196119e-01
ORGANIZATION_TYPE_Hotel	-3.866249e-01
ORGANIZATION_TYPE_Housing	0.000000e+00
ORGANIZATION_TYPE_Industry	-5.699533e-02
ORGANIZATION_TYPE_Insurance	-2.934122e-01
ORGANIZATION_TYPE_Kindergarten	-8.158060e-02
`ORGANIZATION_TYPE_Legal Services`	5.438191e-01
ORGANIZATION_TYPE_Medicine	-1.191476e-01
ORGANIZATION_TYPE_Military	-4.244346e-01
ORGANIZATION_TYPE_Mobile	-3.537404e-02
ORGANIZATION_TYPE_Other	-4.627114e-02
ORGANIZATION_TYPE_Police	-3.257039e-01
ORGANIZATION_TYPE_Postal	4.523275e-02
ORGANIZATION_TYPE_Realtor	6.525932e-01
ORGANIZATION_TYPE_Religion	-2.277195e-03
ORGANIZATION_TYPE_Restaurant	2.439652e-02
ORGANIZATION_TYPE_School	-1.731132e-01
ORGANIZATION_TYPE_Security	-1.245983e-01
`ORGANIZATION_TYPE_Security Ministries`	-2.969287e-01
`ORGANIZATION_TYPE_Self-employed`	9.812759e-02
ORGANIZATION_TYPE_Services	-7.810279e-02
ORGANIZATION_TYPE_Telecom	4.238544e-02

ORGANIZATION_TYPE_Services	-7.810279e-02
ORGANIZATION_TYPE_Telecom	4.238544e-02
ORGANIZATION_TYPE_Trade	-1.695296e-02
ORGANIZATION_TYPE_Transport	2.596422e-02
ORGANIZATION_TYPE_University	-2.391943e-01
`DIES_MAX_IMPAGATS_Altres_EF_+120`	9.516420e-02
`DIES_MAX_IMPAGATS_Altres_EF_[1-30]`	7.816628e-02
`DIES_MAX_IMPAGATS_Altres_EF_[31-60]`	6.019086e-02
`DIES_MAX_IMPAGATS_Altres_EF_[61-90]`	-1.585913e-01
`DIES_MAX_IMPAGATS_Altres_EF_[91-120]`	1.864437e-01
`Credits_Assegurats_HC_Menys crèdits assegurats`	2.412494e-03
`Credits_Assegurats_HC_Més crèdits assegurats`	2.762900e-02
`SUM_Diferencia_AMT_HC_More credit than ATB`	1.006034e-01
`SUM_Diferencia_AMT_HC_Same credit than ATB`	4.208651e-02
LAST_DUE_MONTH_Less_Year	1.264609e-01
LAST_DUE_MONTH_Between_1Year_2Years	8.451476e-02
LAST_DUE_MONTH_Between_2Years_5Years	8.834831e-02
LAST_DUE_MONTH_More_5Years	2.090992e-01
LAST_DUE_MONTH_Retired	2.089582e-01
LAST_DUE_MONTH_No_DPD_PA	-1.698743e-01
DID_OVERDUE_Altres_EF:DID_PROLONG_Altres_EF	4.649309e-01
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y	5.348507e-02
MAX_CREDIT_LIM_CC:TE_CREDIT_CARD_CC	-2.315470e-07

Figura 60. Coeficients dels nivells de totes les variables del model Lasso.

8.3.6 C.V Lasso amb menys variables

Elecció de la lambda per al model Lasso amb menys variables.

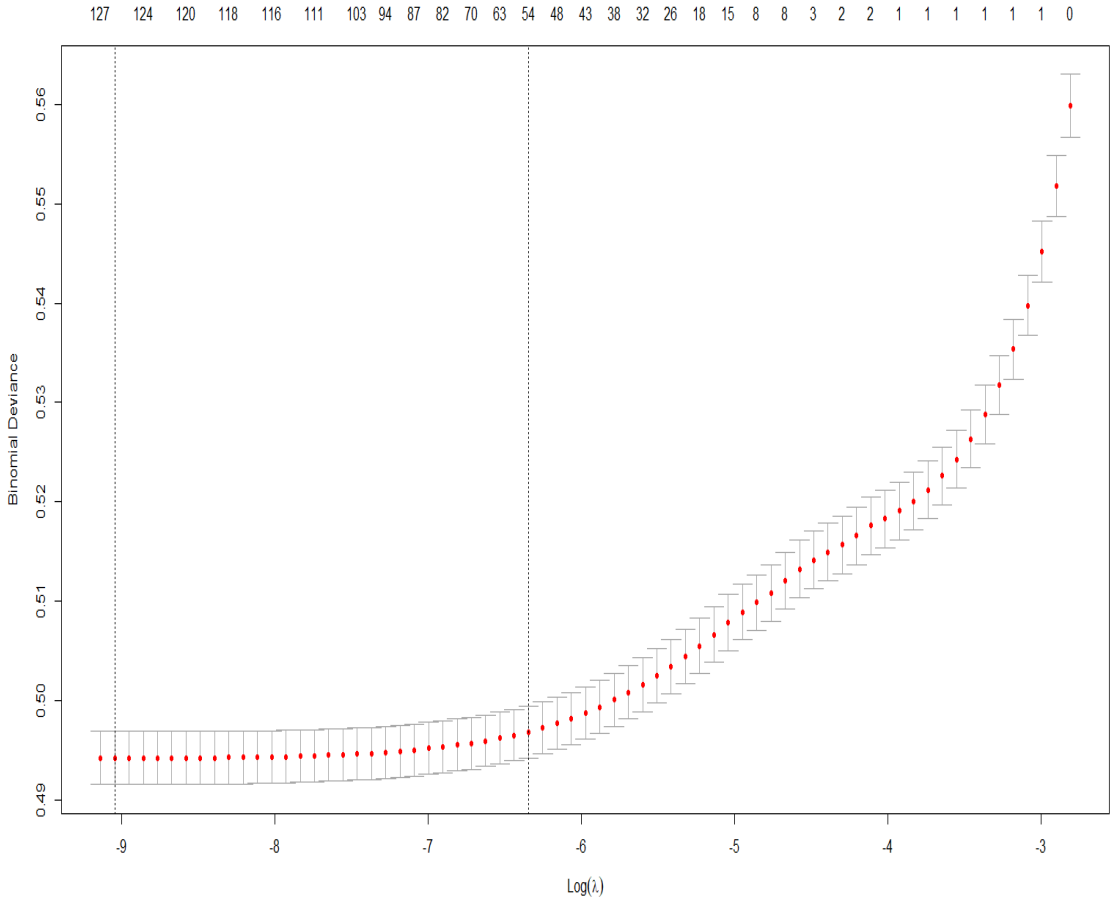


Figura 61. Criteri de selecció per a la lambda òptima que redueix l'error de predicció a partir de la cross-validació. Gràfic amb el model Lasso amb les variables significatives.

8.3.7 Model Lasso amb menys variables

```
coef(lasso.model)[,69]
```

(Intercept)	CNT_CHILDREN
-2.051147e+00	2.745474e-02
AMT_INCOME_TOTAL	AMT_CREDIT
-2.204792e-07	1.848077e-07
AGE_EXPECTED	ANY_COTXE_INTERACTION
-4.932149e-03	3.595268e-03
REGION_RATING_CLIENT	MEAN_EXT_SOURCE
1.397617e-01	-4.009655e+00
CRE_SOL_Altres_EF	CRE_TOTAL_Altres_EF
1.748848e-01	-1.189743e-02
TE_CREDIT_Altres_EF	CAR_LOAN_Altres_EF
-1.051008e-01	-1.457367e-01
CREDIT_CARD_Altres_EF	MICROLOAN_Altres_EF
-3.759511e-02	5.754691e-01
MORTGAGE_LOAN_Altres_EF	DID_OVERDUE_Altres_EF
-3.819131e-01	2.029470e-01
DID_PROLONG_Altres_EF	REV_LOANS_TY_HC
-7.506523e-04	9.337926e-02
NC_Goods_PURP_HC	Repairs_PURP_HC
7.369229e-02	3.652160e-02
UrgentN_PURP_HC	Approved_STATUS_HC
8.297115e-02	-7.197667e-02
Refused_STATUS_HC	ElectD_CAT_HC
6.642253e-02	-6.555287e-02
Clothes_CAT_HC	FreeT_CAT_HC
-1.124338e-01	-2.725993e-01
Home_CAT_HC	IS_Refreshed_Client_HC
-1.155074e-01	-1.137406e-01
TOT_RECEIVABLE_CC	TOT_DRAWINGS_CC
2.076530e-08	8.822196e-03
TOT_INSTALMENTS_CC	TE_CREDIT_CARD_CC
-1.937801e-04	-4.832853e-02
TOT_CO_RTS_PCB_MENSUAL	DIES_TOT_DEUTE_PCB
1.434448e-01	2.684808e-06
NUM_QUOTES_PENDENTS	LATE_PAYMENTS_IP
8.209499e-03	2.512097e-01
NOT_ENOUGH_MONEY_PAID_IP	PERCENTATGE_ANNUITY
3.052792e-01	1.122717e-04
AMT_CREDIT_ACT_TOT	RATI_DEUTE_GARANTIA
1.251272e-08	1.077972e+00
NUM_ACTIVE_CREDITS	NUM_CREDITS_PREVIS_TANCATS
8.804466e-02	-8.549899e-03
`NAME_CONTRACT_TYPE_Revolving loans`	CODE_GENDER_M
-1.528567e-01	2.855598e-01
FLAG_OWN_CAR_Y	NAME_INCOME_TYPE_Businessman
-3.025940e-01	-2.221908e-01
`NAME_INCOME_TYPE_Maternity leave`	NAME_INCOME_TYPE_Pensioner
2.446098e+00	-2.719049e-01
`NAME_INCOME_TYPE_State servant`	NAME_INCOME_TYPE_Student
0.000000e+00	-2.287812e+00
NAME_INCOME_TYPE_Unemployed	NAME_INCOME_TYPE_Working
1.116468e+00	1.066596e-01
`NAME_EDUCATION_TYPE_Academic degree`	`NAME_EDUCATION_TYPE_Higher education`
-1.104683e+00	-2.897949e-01
`NAME_EDUCATION_TYPE_Incomplete higher`	`NAME_EDUCATION_TYPE_Lower secondary`
-2.101915e-01	6.637614e-02
`NAME_FAMILY_STATUS_Civil marriage`	NAME_FAMILY_STATUS_Separated
1.180859e-01	1.299921e-01
`NAME_FAMILY_STATUS_Single / not married`	NAME_FAMILY_STATUS_Widow
6.894568e-02	-1.402842e-02
`NAME_HOUSING_TYPE_Co-op apartment`	`NAME_HOUSING_TYPE_Municipal apartment`
-1.583280e-03	6.958175e-02
`NAME_HOUSING_TYPE_Office apartment`	`NAME_HOUSING_TYPE_Rented apartment`
-2.417568e-01	1.312240e-01
`NAME_HOUSING_TYPE_With parents`	`LAST_YEARS_EMPLOYED_4-9`
1.280254e-03	-1.507168e-01
`LAST_YEARS_EMPLOYED_10-80`	`OCCUPATION_TYPE_Cleaning staff`
-3.423878e-01	9.641568e-02
`OCCUPATION_TYPE_Cooking staff`	`OCCUPATION_TYPE_Core staff`
1.078075e-01	-3.633727e-02
OCCUPATION_TYPE_Drivers	`OCCUPATION_TYPE_HR staff`
1.736017e-01	-1.176820e-01

OCCUPATION_TYPE_Drivers	1.736017e-01
`OCCUPATION_TYPE_HR staff`	-1.176820e-01
`OCCUPATION_TYPE_High skill tech staff`	-7.467500e-02
`OCCUPATION_TYPE_IT staff`	3.201284e-02
OCCUPATION_TYPE_Laborers	8.555204e-02
`OCCUPATION_TYPE_Low-skill Laborers`	2.256069e-01
OCCUPATION_TYPE_Managers	4.205052e-02
`OCCUPATION_TYPE_Medicine staff`	-1.212758e-02
`OCCUPATION_TYPE_Private service staff`	0.000000e+00
`OCCUPATION_TYPE_Realty agents`	-1.564581e-02
`OCCUPATION_TYPE_Sales staff`	4.148762e-02
OCCUPATION_TYPE_Secretaries	1.696066e-01
`OCCUPATION_TYPE_Security staff`	9.822031e-02
`OCCUPATION_TYPE_Waiters/barmen staff`	1.734617e-01
ORGANIZATION_TYPE_Advertising	1.277066e-01
ORGANIZATION_TYPE_Agriculture	1.282820e-02
ORGANIZATION_TYPE_Bank	-3.314164e-01
ORGANIZATION_TYPE_Cleaning	7.197631e-02
ORGANIZATION_TYPE_Construction	2.021653e-01
ORGANIZATION_TYPE_Culture	0.000000e+00
ORGANIZATION_TYPE_Electricity	-4.182709e-02
ORGANIZATION_TYPE_Emergency	-1.146065e-01
ORGANIZATION_TYPE_Government	-1.207256e-01
ORGANIZATION_TYPE_Hotel	-3.801340e-01
ORGANIZATION_TYPE_Housing	0.000000e+00
ORGANIZATION_TYPE_Industry	-5.642568e-02
ORGANIZATION_TYPE_Insurance	-2.874491e-01
ORGANIZATION_TYPE_Kindergarten	-8.042324e-02
`ORGANIZATION_TYPE_Legal Services`	5.404831e-01
ORGANIZATION_TYPE_Medicine	-1.178449e-01
ORGANIZATION_TYPE_Military	-4.247303e-01
ORGANIZATION_TYPE_Mobile	-3.577735e-02
ORGANIZATION_TYPE_Other	-4.629396e-02
ORGANIZATION_TYPE_Police	-3.242250e-01
ORGANIZATION_TYPE_Postal	4.400375e-02
ORGANIZATION_TYPE_Realtor	6.516068e-01
ORGANIZATION_TYPE_Religion	0.000000e+00
ORGANIZATION_TYPE_Restaurant	2.270217e-02
ORGANIZATION_TYPE_School	-1.735227e-01
ORGANIZATION_TYPE_Security	-1.205393e-01
`ORGANIZATION_TYPE_Security Ministries`	-2.945161e-01
`ORGANIZATION_TYPE_Self-employed`	9.813498e-02
ORGANIZATION_TYPE_Services	-7.778764e-02
ORGANIZATION_TYPE_Telecom	3.693385e-02
ORGANIZATION_TYPE_Trade	-1.481398e-02
ORGANIZATION_TYPE_Transport	2.615335e-02
ORGANIZATION_TYPE_University	-2.365685e-01
`DIES_MAX_IMPAGATS_Altres_EF_+120`	8.688451e-02
`DIES_MAX_IMPAGATS_Altres_EF_[1-30]`	7.795644e-02
`DIES_MAX_IMPAGATS_Altres_EF_[31-60]`	5.813818e-02
`DIES_MAX_IMPAGATS_Altres_EF_[61-90]`	-1.580301e-01
`DIES_MAX_IMPAGATS_Altres_EF_[91-120]`	1.719440e-01
`SUM_Diferencia_AMT_HC_More credit than ATB`	1.180921e-01
`SUM_Diferencia_AMT_HC_Same credit than ATB`	4.992376e-02
LAST_DUE_MONTH_Less_Year	1.255059e-01
LAST_DUE_MONTH_Between_1Year_2Years	8.293159e-02
LAST_DUE_MONTH_Between_2Years_5Years	8.105541e-02
LAST_DUE_MONTH_More_5Years	2.079997e-01
LAST_DUE_MONTH_Retired	2.075430e-01
LAST_DUE_MONTH_No_DPD_PA	-1.467169e-01
DID_OVERDUE_Altres_EF:DID_PROLONG_Altres_EF	4.593700e-01
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y	6.560040e-02
MAX_CREDIT_LIM_CC:TE_CREDIT_CARD_CC	-2.293844e-07

Figura 62. Coeficient dels nivells de les variables significatives del model Lasso.

8.3.8 Model Ridge amb totes les variables

```
coef(ridge.model2)[,100]
```

(Intercept)	CNT_CHILDREN
-2.087448e+00	2.679179e-02
AMT_INCOME_TOTAL	AMT_CREDIT
-2.069104e-07	1.479901e-07
REGION_POPULATION_RELATIVE	AGE_EXPECTED
-1.507629e-02	-6.174217e-03
ANY_COTXE_INTERACTION	REGION_RATING_CLIENT
2.651167e-03	1.474506e-01
MEAN_EXT_SOURCE	CRE_SOL_Altres_EF
-3.712244e+00	1.692784e-01
CRE_BDE_Altres_EF	CRE_TOTAL_Altres_EF
4.771966e-01	-5.554772e-03
TE_CREDIT_Altres_EF	CRE_Another_Type_LOAN_Altres_EF
-9.058155e-02	-1.726183e-01
CAR_LOAN_Altres_EF	CONSUMER_CREDIT_Altres_EF
-1.518967e-01	-3.491208e-02
CREDIT_CARD_Altres_EF	BUSINESS_LOAN_Altres_EF
-2.970619e-02	-1.998879e-01
MICROLOAN_Altres_EF	MORTGAGE_LOAN_Altres_EF
5.775215e-01	-3.555748e-01
UNKNOWN_LOAN_Altres_EF	DID_OVERDUE_Altres_EF
3.937224e-02	2.304291e-01
DID_PROLONG_Altres_EF	CASH_LOANS_TY_HC
-8.818374e-03	4.167840e-02
CONS_LOANS_TY_HC	REV_LOANS_TY_HC
-3.943316e-02	9.375475e-02
NC_Goods_PURP_HC	Repairs_PURP_HC
7.079809e-02	3.919123e-02
UrgentN_PURP_HC	GET_HOME_PURP_HC
8.669629e-02	1.251776e-02
CARS_PURP_HC	FUN_PURP_HC
-3.287722e-02	7.919997e-03
Approved_STATUS_HC	Canceled_STATUS_HC
-6.197617e-02	8.573773e-04
Refused_STATUS_HC	UnusedOF_STATUS_HC
6.180281e-02	1.276626e-02
ElectD_CAT_HC	Clothes_CAT_HC
-5.890749e-02	-1.045646e-01
FreeT_CAT_HC	Health_CAT_HC
-2.530441e-01	-4.145962e-02
Home_CAT_HC	Mobile_CAT_HC
-1.069661e-01	2.583155e-03
ConstructionM_CAT_HC	Vehicles_CAT_HC
-3.520898e-02	-3.182349e-02
IS_NEW_CLIENT_TRAIN	IS_Refreshed_Client_HC
2.251133e-02	-1.168756e-01
TOT_RECEIVABLE_CC	TOT_DRAWINGS_CC
1.669852e-08	8.095840e-03
TOT_INSTALMENTS_CC	MEAN_DPD_CC
-1.632272e-04	5.369256e-04
TE_CREDIT_CARD_CC	TOT_CO_SIG_PCB_MENSUAL
-6.070670e-02	-6.962590e-04
TOT_CO_RTS_PCB_MENSUAL	TOT_CO_ALT_PCB_MENSUAL
1.448436e-01	-8.257732e-04
DIES_NOMES_GRAN_DEUTE_PCB	DIES_TOT_DEUTE_PCB
1.367359e-06	2.954092e-06
NUM_QUOTES_PENDENTS	LATE_PAYMENTS_IP
7.250786e-03	2.549149e-01
NOT_ENOUGH_MONEY_PAID_IP	PERCENTATGE_ANNUITY
2.779332e-01	1.121382e-04
AMT_CREDIT_ACT_TOT	RATI_DEUTE_GARANTIA
1.191070e-08	1.037044e+00
NUM_ACTIVE_CREDITS	NUM_CREDITS_PREVIS_TANCATS
7.608326e-02	-1.283722e-02
`NAME_CONTRACT_TYPE_Revolving loans`	CODE_GENDER_M
-1.618155e-01	2.616463e-01
FLAG_OWN_CAR_Y	FLAG_OWN_REALTY_Y
-2.272354e-01	2.988492e-02
NAME_TYPE_SUITE_Children	NAME_TYPE_SUITE_Family
6.934521e-02	3.632113e-03
`NAME_TYPE_SUITE_Group of people`	NAME_TYPE_SUITE_Other
6.437083e-02	5.332202e-02

`NAME_TYPE_SUITE_Group of people`	6.437083e-02
`NAME_TYPE_SUITE_Spouse, partner`	-5.459103e-02
`NAME_INCOME_TYPE_Maternity leave`	2.469129e+00
`NAME_INCOME_TYPE_State servant`	-6.533048e-03
NAME_INCOME_TYPE_Unemployed	1.179317e+00
`NAME_EDUCATION_TYPE_Academic degree`	-1.024537e+00
`NAME_EDUCATION_TYPE_Incomplete higher`	-1.946778e-01
`NAME_FAMILY_STATUS_Civil marriage`	1.157149e-01
`NAME_FAMILY_STATUS_Single / not married`	6.943574e-02
`NAME_HOUSING_TYPE_Co-op apartment`	-2.154548e-02
`NAME_HOUSING_TYPE_Office apartment`	-2.292315e-01
`NAME_HOUSING_TYPE_With parents`	2.199980e-02
`LAST_YEARS_EMPLOYED_10-80`	-3.081361e-01
`OCCUPATION_TYPE_Cooking staff`	1.226045e-01
OCCUPATION_TYPE_Drivers	1.868280e-01
`OCCUPATION_TYPE_High skill tech staff`	-7.878546e-02
OCCUPATION_TYPE_Laborers	1.022624e-01
OCCUPATION_TYPE_Managers	3.951958e-02
`OCCUPATION_TYPE_Private service staff`	-1.127773e-02
`OCCUPATION_TYPE_Sales staff`	5.096440e-02
`OCCUPATION_TYPE_Security staff`	1.282234e-01
ORGANIZATION_TYPE_Advertising	1.494331e-01
ORGANIZATION_TYPE_Bank	-3.184697e-01
ORGANIZATION_TYPE_Construction	2.065465e-01
ORGANIZATION_TYPE_Electricity	-7.423503e-02
ORGANIZATION_TYPE_Government	-1.127595e-01
ORGANIZATION_TYPE_Housing	-9.394405e-03
ORGANIZATION_TYPE_Insurance	-3.109295e-01
`ORGANIZATION_TYPE_Legal Services`	5.257895e-01
ORGANIZATION_TYPE_Military	-3.876753e-01
ORGANIZATION_TYPE_Other	-4.444460e-02
ORGANIZATION_TYPE_Postal	6.218450e-02
ORGANIZATION_TYPE_Religion	-1.375170e-01
ORGANIZATION_TYPE_School	-1.637109e-01
`ORGANIZATION_TYPE_Security Ministries`	-2.853518e-01
NAME_TYPE_SUITE_Other	5.332202e-02
NAME_INCOME_TYPE_Businessman	-1.229804e+00
NAME_INCOME_TYPE_Pensioner	-2.130722e-01
NAME_INCOME_TYPE_Student	-2.243571e+00
NAME_INCOME_TYPE_Working	1.116215e-01
`NAME_EDUCATION_TYPE_Higher education`	-2.679857e-01
`NAME_EDUCATION_TYPE_Lower secondary`	8.418308e-02
NAME_FAMILY_STATUS_Separated	1.186889e-01
NAME_FAMILY_STATUS_Widow	-3.238624e-02
`NAME_HOUSING_TYPE_Municipal apartment`	7.794341e-02
`NAME_HOUSING_TYPE_Rented apartment`	1.467067e-01
`LAST_YEARS_EMPLOYED_4-9`	-1.377240e-01
`OCCUPATION_TYPE_Cleaning staff`	1.166911e-01
`OCCUPATION_TYPE_Core staff`	-4.203423e-02
`OCCUPATION_TYPE_HR staff`	-1.635197e-01
`OCCUPATION_TYPE_IT staff`	5.753023e-02
`OCCUPATION_TYPE_Low-skill Laborers`	2.648983e-01
`OCCUPATION_TYPE_Medicine staff`	-2.157828e-02
`OCCUPATION_TYPE_Realty agents`	-4.385610e-02
OCCUPATION_TYPE_Secretaries	1.666121e-01
`OCCUPATION_TYPE_Waiters/barmen staff`	1.879115e-01
ORGANIZATION_TYPE_Agriculture	4.159892e-02
ORGANIZATION_TYPE_Cleaning	1.177577e-01
ORGANIZATION_TYPE_Culture	3.921587e-02
ORGANIZATION_TYPE_Emergency	-1.274958e-01
ORGANIZATION_TYPE_Hotel	-3.619287e-01
ORGANIZATION_TYPE_Industry	-5.470874e-02
ORGANIZATION_TYPE_Kindergarten	-7.745657e-02
ORGANIZATION_TYPE_Medicine	-1.081891e-01
ORGANIZATION_TYPE_Mobile	-6.457377e-02
ORGANIZATION_TYPE_Police	-3.073590e-01
ORGANIZATION_TYPE_Realtor	6.403757e-01
ORGANIZATION_TYPE_Restaurant	4.802817e-02
ORGANIZATION_TYPE_Security	-1.191724e-01
`ORGANIZATION_TYPE_Self-employed`	9.981447e-02

`ORGANIZATION_TYPE_Security Ministries`	-2.853518e-01
`ORGANIZATION_TYPE_Self-employed`	9.981447e-02
ORGANIZATION_TYPE_Services	-9.185231e-02
ORGANIZATION_TYPE_Telecom	6.458383e-02
ORGANIZATION_TYPE_Trade	-1.372355e-02
ORGANIZATION_TYPE_Transport	3.694447e-02
ORGANIZATION_TYPE_University	-2.442211e-01
`DIES_MAX_IMPAGATS_Altres_EF_+120`	1.171418e-01
`DIES_MAX_IMPAGATS_Altres_EF_[1-30]`	8.559680e-02
`DIES_MAX_IMPAGATS_Altres_EF_[31-60]`	7.335438e-02
`DIES_MAX_IMPAGATS_Altres_EF_[61-90]`	-1.618950e-01
`DIES_MAX_IMPAGATS_Altres_EF_[91-120]`	2.155069e-01
`Credits_Assegurats_HC_Menys crèdits assegurats`	1.186525e-02
`Credits_Assegurats_HC_Més crèdits assegurats`	2.773089e-02
`SUM_Diferencia_AMT_HC_More credit than ATB`	9.684720e-02
`SUM_Diferencia_AMT_HC_Same credit than ATB`	3.837487e-02
LAST_DUE_MONTH_Less_Year	1.129363e-01
LAST_DUE_MONTH_Between_1Year_2Years	7.282078e-02
LAST_DUE_MONTH_Between_2Years_5Years	7.247215e-02
LAST_DUE_MONTH_More_5Years	1.881901e-01
LAST_DUE_MONTH_Retired	2.066046e-01
LAST_DUE_MONTH_No_DPD_PA	-1.567544e-01
DID_OVERDUE_Altres_EF:DID_PROLONG_Altres_EF	4.864237e-01
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y	8.683785e-03
MAX_CREDIT_LIM_CC:TE_CREDIT_CARD_CC	-1.814210e-07

Figura 63. Coeficients dels nivells de totes les variables del model Ridge.

8.3.9 C.V Ridge amb menys variables

Elecció de la lambda per al model Ridge amb menys variables.

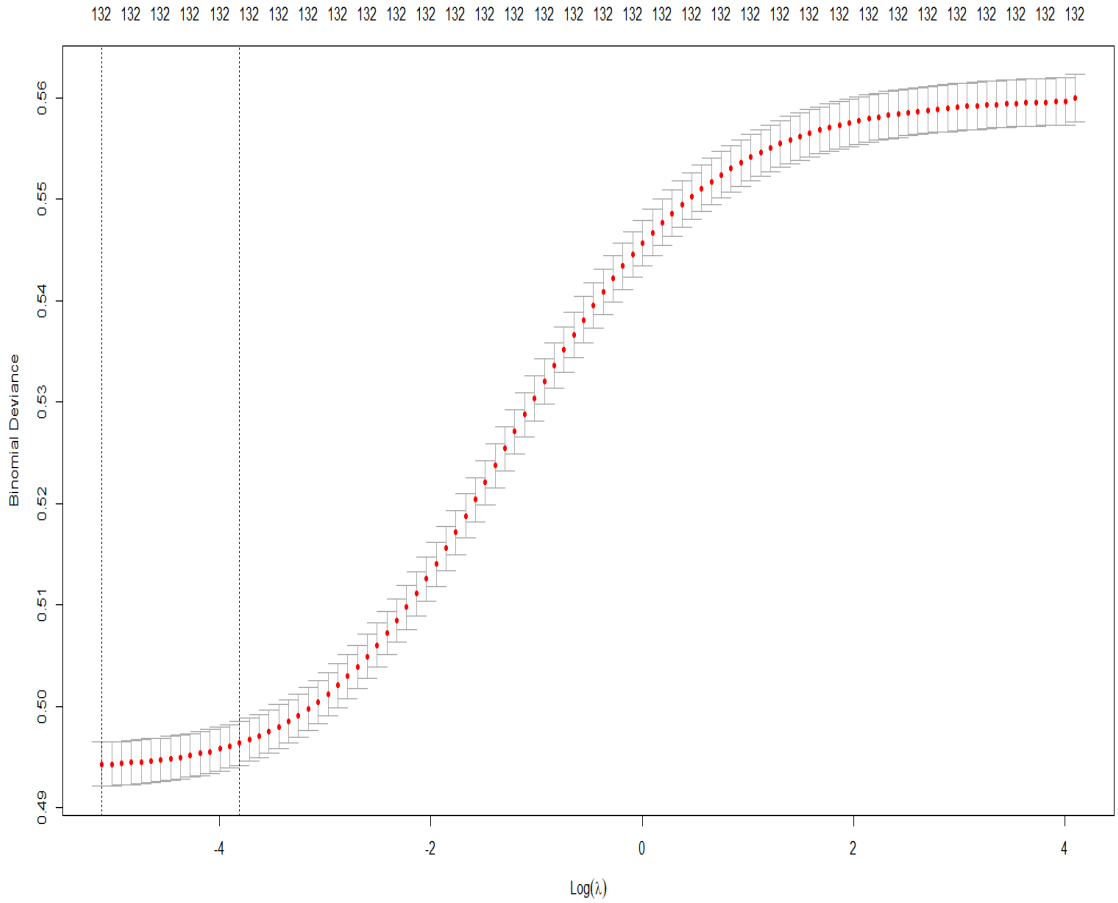


Figura 64. Criteri de selecció per a la lambda òptima que redueix l'error de predicció a partir de la cross-validació. Gràfic amb el model Ridge amb les variables significatives.

8.3.10 Model Ridge amb menys variables

```
coef(ridge.model)[,100]
```

(Intercept)	CNT_CHILDREN
-2.096864e+00	2.657863e-02
AMT_INCOME_TOTAL	AMT_CREDIT
-1.950681e-07	1.460010e-07
AGE_EXPECTED	ANY_COTXE_INTERACTION
-5.924073e-03	2.595212e-03
REGION_RATING_CLIENT	MEAN_EXT_SOURCE
1.450240e-01	-3.715652e+00
CRE_SOL_Altres_EF	CRE_TOTAL_Altres_EF
1.689823e-01	-6.075498e-03
TE_CREDIT_Altres_EF	CAR_LOAN_Altres_EF
-1.193428e-01	-1.500041e-01
CREDIT_CARD_Altres_EF	MICROLOAN_Altres_EF
-2.529894e-02	5.821683e-01
MORTGAGE_LOAN_Altres_EF	DID_OVERDUE_Altres_EF
-3.513993e-01	2.314259e-01
DID_PROLONG_Altres_EF	REV_LOANS_TY_HC
-7.383180e-03	9.748949e-02
NC_Goods_PURP_HC	Repairs_PURP_HC
7.809374e-02	4.385467e-02
UrgentN_PURP_HC	Approved_STATUS_HC
9.324445e-02	-5.943259e-02
Refused_STATUS_HC	ElectD_CAT_HC
6.267435e-02	-6.303285e-02
Clothes_CAT_HC	FreeT_CAT_HC
-1.092038e-01	-2.582003e-01
Home_CAT_HC	IS_Refreshed_Client_HC
-1.130573e-01	-1.118282e-01
TOT_RECEIVABLE_CC	TOT_DRAWINGS_CC
1.676737e-08	8.107456e-03
TOT_INSTALMENTS_CC	TE_CREDIT_CARD_CC
-1.611189e-04	-5.998566e-02
TOT_CO_RTS_PCB_MENSUAL	DIES_TOT_DEUTE_PCB
1.407221e-01	3.101140e-06
NUM_QUOTES_PENDENTS	LATE_PAYMENTS_IP
7.454774e-03	2.539228e-01
NOT_ENOUGH_MONEY_PAID_IP	PERCENTATGE_ANNUITY
2.796260e-01	1.125352e-04
AMT_CREDIT_ACT_TOT	RATI_DEUTE_GARANTIA
1.171856e-08	1.043614e+00
NUM_ACTIVE_CREDITS	NUM_CREDITS_PREVIS_TANCATS
7.567343e-02	-1.286645e-02
`NAME_CONTRACT_TYPE_Revolving loans`	CODE_GENDER_M
-1.610987e-01	2.595188e-01
FLAG_OWN_CAR_Y	NAME_INCOME_TYPE_Businessman
-2.393523e-01	-1.222749e+00
`NAME_INCOME_TYPE_Maternity leave`	NAME_INCOME_TYPE_Pensioner
2.460344e+00	-2.152708e-01
`NAME_INCOME_TYPE_State servant`	NAME_INCOME_TYPE_Student
-8.871105e-03	-2.244608e+00
NAME_INCOME_TYPE_Unemployed	NAME_INCOME_TYPE_Working
1.187396e+00	1.092550e-01
`NAME_EDUCATION_TYPE_Academic degree`	`NAME_EDUCATION_TYPE_Higher education`
-1.029627e+00	-2.672020e-01
`NAME_EDUCATION_TYPE_Incomplete higher`	`NAME_EDUCATION_TYPE_Lower secondary`
-1.928237e-01	8.212256e-02
`NAME_FAMILY_STATUS_Civil marriage`	NAME_FAMILY_STATUS_Separated
1.160765e-01	1.210427e-01
`NAME_FAMILY_STATUS_Single / not married`	NAME_FAMILY_STATUS_Widow
7.131247e-02	-3.003864e-02
`NAME_HOUSING_TYPE_Co-op apartment`	`NAME_HOUSING_TYPE_Municipal apartment`
-2.328703e-02	7.199727e-02
`NAME_HOUSING_TYPE_Office apartment`	`NAME_HOUSING_TYPE_Rented apartment`
-2.342292e-01	1.413910e-01
`NAME_HOUSING_TYPE_With parents`	`LAST_YEARS_EMPLOYED_4-9`
1.553036e-02	-1.379815e-01
`LAST_YEARS_EMPLOYED_10-80`	`OCCUPATION_TYPE_Cleaning staff`
-3.080252e-01	1.196218e-01
`OCCUPATION_TYPE_Cooking staff`	`OCCUPATION_TYPE_Core staff`
1.236435e-01	-4.150285e-02
OCCUPATION_TYPE_Drivers	`OCCUPATION_TYPE_HR staff`
1.866002e-01	-1.554630e-01

OCCUPATION_TYPE_Drivers	1.866002e-01
`OCCUPATION_TYPE_High skill tech staff`	-7.799724e-02
OCCUPATION_TYPE_Laborers	1.025145e-01
OCCUPATION_TYPE_Managers	3.928099e-02
`OCCUPATION_TYPE_Private service staff`	-1.046329e-02
`OCCUPATION_TYPE_Sales staff`	5.191352e-02
`OCCUPATION_TYPE_Security staff`	1.287828e-01
ORGANIZATION_TYPE_Advertising	1.525987e-01
ORGANIZATION_TYPE_Bank	-3.178512e-01
ORGANIZATION_TYPE_Construction	2.068456e-01
ORGANIZATION_TYPE_Electricity	-7.263269e-02
ORGANIZATION_TYPE_Government	-1.148003e-01
ORGANIZATION_TYPE_Housing	-1.041772e-02
ORGANIZATION_TYPE_Insurance	-3.081035e-01
`ORGANIZATION_TYPE_Legal Services`	5.260412e-01
ORGANIZATION_TYPE_Military	-3.894151e-01
ORGANIZATION_TYPE_Other	-4.497903e-02
ORGANIZATION_TYPE_Postal	6.194709e-02
ORGANIZATION_TYPE_Religion	-1.398940e-01
ORGANIZATION_TYPE_School	-1.649541e-01
`ORGANIZATION_TYPE_Security Ministries`	-2.853916e-01
ORGANIZATION_TYPE_Services	-9.399018e-02
ORGANIZATION_TYPE_Trade	-1.289992e-02
ORGANIZATION_TYPE_University	-2.449489e-01
`DIES_MAX_IMPAGATS_Altres_EF_[1-30]`	8.550463e-02
`DIES_MAX_IMPAGATS_Altres_EF_[61-90]`	-1.642698e-01
`SUM_Diferencia_AMT_HC_More credit than ATB`	1.119096e-01
LAST_DUE_MONTH_Less_Year	1.134163e-01
LAST_DUE_MONTH_Between_2Years_5Years	6.653328e-02
LAST_DUE_MONTH_Retired	2.073518e-01
DID_OVERDUE_Altres_EF:DID_PROLONG_Altres_EF	4.842229e-01
MAX_CREDIT_LIM_CC:TE_CREDIT_CARD_CC	-1.820764e-07
`OCCUPATION_TYPE_HR staff`	-1.554630e-01
`OCCUPATION_TYPE_IT staff`	5.990790e-02
`OCCUPATION_TYPE_Low-skill Laborers`	2.645370e-01
`OCCUPATION_TYPE_Medicine staff`	-2.107282e-02
`OCCUPATION_TYPE_Realty agents`	-4.030018e-02
OCCUPATION_TYPE_Secretaries	1.675640e-01
`OCCUPATION_TYPE_Waiters/barmen staff`	1.877350e-01
ORGANIZATION_TYPE_Agriculture	3.832433e-02
ORGANIZATION_TYPE_Cleaning	1.192501e-01
ORGANIZATION_TYPE_Culture	3.834649e-02
ORGANIZATION_TYPE_Emergency	-1.293333e-01
ORGANIZATION_TYPE_Hotel	-3.601548e-01
ORGANIZATION_TYPE_Industry	-5.512604e-02
ORGANIZATION_TYPE_Kindergarten	-7.745034e-02
ORGANIZATION_TYPE_Medicine	-1.077889e-01
ORGANIZATION_TYPE_Mobile	-6.872324e-02
ORGANIZATION_TYPE_Police	-3.079171e-01
ORGANIZATION_TYPE_Realtor	6.432366e-01
ORGANIZATION_TYPE_Restaurant	4.685686e-02
ORGANIZATION_TYPE_Security	-1.178802e-01
`ORGANIZATION_TYPE_Self-employed`	9.985036e-02
ORGANIZATION_TYPE_Telecom	6.259772e-02
ORGANIZATION_TYPE_Transport	3.729475e-02
`DIES_MAX_IMPAGATS_Altres_EF_+120`	1.093316e-01
`DIES_MAX_IMPAGATS_Altres_EF_[31-60]`	7.220833e-02
`DIES_MAX_IMPAGATS_Altres_EF_[91-120]`	2.028449e-01
`SUM_Diferencia_AMT_HC_Same credit than ATB`	4.606529e-02
LAST_DUE_MONTH_Between_1Year_2Years	7.246329e-02
LAST_DUE_MONTH_More_5Years	1.853427e-01
LAST_DUE_MONTH_No_DPD_PA	-1.410547e-01
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y	2.824669e-02

Figura 65. Coeficient dels nivells de les variables significatives del model Lasso.

8.4 Interpretative Machine Learning

8.4.1 IML d'un client amb una baixa probabilitat d'impagament

y=0 (probability **0.992**, score **-4.877**) top features

Contribution	Feature
+2.541	<BIAS>
+1.040	MEAN_EXT_SOURCE
+0.399	TOT_INSTALMENTS_CC
+0.356	MAX_EXT_SOURCE
+0.304	AMT_INCOME_TOTAL
+0.220	MAX_CREDIT_LIM_CC
+0.171	AMT_CREDIT_ACT_TOT
+0.127	RATI_DEUTE_GARANTIA
+0.073	CAR_LOAN_Altres_EF
+0.069	ANY_COTXE_INTERACTION
+0.055	CODE_GENDER_F
+0.055	NUM_CREDITS_PREVIS_TANCATS
+0.053	DIES_NOMES_GRAN_DEUTE_PCB
+0.049	OCCUPATION_TYPE_Core staff
+0.045	MIN_EXT_SOURCE
+0.034	SUM_BALANCE_CC
+0.032	DIES_MAX_IMPAGATS_Altres_EF__1-30_
+0.030	SUM_Diferencia_AMT_HC_More credit than ATB
+0.028	PERCENTATGE_ANNUITY
+0.024	IS_NEW_CLIENT_TRAIN
+0.021	FLAG_OWN_CAR_N
+0.019	CRE_SOL_Altres_EF
+0.019	CONSUMER_CREDIT_Altres_EF
+0.019	NAME_FAMILY_STATUS_Married
+0.016	REGION_RATING_CLIENT
+0.014	NUM_QUOTES_PENDENTS
+0.014	AMT_CREDIT
+0.012	NAME_FAMILY_STATUS_Single / not married
+0.012	NUM_ACTIVE_CREDITS
+0.012	AGE_EXPECTED
+0.011	ORGANIZATION_TYPE_Transport
+0.009	UnusedOF_STATUS_HC
+0.009	OCCUPATION_TYPE_Drivers
+0.008	OCCUPATION_TYPE_Laborers
+0.008	LAST_DUE_MONTH_Less_Year
+0.007	OCCUPATION_TYPE_Private service staff
+0.007	ORGANIZATION_TYPE_Postal
+0.005	TOT_DRAWINGS_CC
+0.005	MICROLOAN_Altres_EF
+0.004	DIES_TOT_DEUTE_PCB
+0.004	ConstructionM_CAT_HC
+0.002	LAST_DUE_MONTH_Retired
+0.002	ORGANIZATION_TYPE_Telecom
+0.001	TOT_CO_ALT_PCB_MENSUAL
+0.001	ORGANIZATION_TYPE_Construction
+0.001	TOT_CO_RTS_PCB_MENSUAL
+0.000	ORGANIZATION_TYPE_Realtor
+0.000	GET_HOME_PURP_HC
-0.000	NAME_EDUCATION_TYPE_Academic degree
-0.000	CARS_PURP_HC
-0.000	FUN_PURP_HC
-0.000	ORGANIZATION_TYPE_Hotel
-0.000	ORGANIZATION_TYPE_Security Ministries
-0.001	ORGANIZATION_TYPE_Military
-0.001	NAME_HOUSING_TYPE_Office apartment
-0.001	ORGANIZATION_TYPE_Bank
-0.001	ORGANIZATION_TYPE_Medicine
-0.001	FreeT_CAT_HC
-0.001	ORGANIZATION_TYPE_School
-0.002	ORGANIZATION_TYPE_Government
-0.002	NAME_INCOME_TYPE_State servant
-0.002	NAME_CONTRACT_TYPE_Cash loans
-0.003	OCCUPATION_TYPE_Accountants
-0.003	LAST_DUE_MONTH_No_DPD_PA
-0.005	FLAG_OWN_REALTY_N
-0.006	Clothes_CAT_HC
-0.006	LAST_DUE_MONTH_Less_6Months
-0.008	AMT_ANNUITY_TOT
-0.009	AMT_GOODS_PRICE
-0.009	IS_Refreshed_Client_HC
-0.009	MORTGAGE_LOAN_Altres_EF
-0.009	CRE_TOTAL_Altres_EF
-0.012	Home_CAT_HC
-0.013	LAST_YEARS_EMPLOYED_4-9
-0.019	NAME_INCOME_TYPE_Working
-0.024	NAME_EDUCATION_TYPE_Higher education
-0.028	LAST_YEARS_EMPLOYED_10-80
-0.028	REGION_POPULATION_RELATIVE
-0.037	CASH_LOANS_TY_HC
-0.039	ORGANIZATION_TYPE_Self-employed
-0.044	Mobile_CAT_HC
-0.046	NAME_EDUCATION_TYPE_Secondary / secondary special
-0.050	ElectD_CAT_HC
-0.050	Refused_STATUS_HC
-0.087	TOT_RECEIVABLE_CC
-0.089	NAME_TYPE_SUITE_Family
-0.095	LAST_YEARS_EMPLOYED_0-3
-0.127	Approved_STATUS_HC
-0.206	NOT_ENOUGH_MONEY_PAID_IP

Figura 66. Pes per variable per a un individu amb baixa probabilitat d'impagament.

8.4.2 IML d'un client amb una elevada probabilitat d'impagament

y=1 (probability 0.700, score 0.850) top features

Contribution	Feature
+1.407	MEAN_EXT_SOURCE
+0.538	Refused_STATUS_HC
+0.410	MIN_EXT_SOURCE
+0.255	CODE_GENDER_F
+0.219	RATI_DEUTE_GARANTIA
+0.205	NAME_FAMILY_STATUS_Married
+0.197	NUM_CREDITS_PREVIS_TANCATS
+0.121	Approved_STATUS_HC
+0.066	FLAG_OWN_CAR_N
+0.058	NAME_FAMILY_STATUS_Civil marriage
+0.047	AMT_CREDIT_ACT_TOT
+0.045	REGION_RATING_CLIENT
+0.040	AMT_GOODS_PRICE
+0.040	LAST_DUE_MONTH_Less_6Months
+0.038	CONSUMER_CREDIT_Altres_EF
+0.038	LAST_YEARS_EMPLOYED_0-3
+0.037	LAST_DUE_MONTH_Retired
+0.030	TOT_INSTALMENTS_CC
+0.028	OCCUPATION_TYPE_Laborers
+0.028	MAX_EXT_SOURCE
+0.021	ANY_COTXE_INTERACTION
+0.020	NAME_EDUCATION_TYPE_Higher education
+0.019	NAME_INCOME_TYPE_Working
+0.019	LAST_YEARS_EMPLOYED_10-80
+0.017	OCCUPATION_TYPE_Core staff
+0.017	ElectD_CAT_HC
+0.017	SUM_Diferencia_AMT_HC_More credit than ATB
+0.015	MAX_CREDIT_LIM_CC
+0.013	MORTGAGE_LOAN_Altres_EF
+0.010	ORGANIZATION_TYPE_Business
+0.009	NAME_INCOME_TYPE_State servant
+0.009	IS_Refreshed_Client_HC
+0.006	Clothes_CAT_HC
+0.005	NAME_EDUCATION_TYPE_Secondary / secondary special
+0.004	CRE_TOTAL_Altres_EF
+0.004	LAST_DUE_MONTH_No_DPD_PA
+0.004	CAR_LOAN_Altres_EF
+0.003	OCCUPATION_TYPE_Accountants
+0.002	NAME_HOUSING_TYPE_Office apartment
+0.002	OCCUPATION_TYPE_High skill tech staff
+0.002	NAME_CONTRACT_TYPE_Cash loans
+0.002	ORGANIZATION_TYPE_Military
+0.001	ORGANIZATION_TYPE_School
+0.001	FreeT_CAT_HC
+0.001	ORGANIZATION_TYPE_Medicine
+0.001	ORGANIZATION_TYPE_Bank
+0.001	Canceled_STATUS_HC
+0.001	ORGANIZATION_TYPE_Police
+0.000	ORGANIZATION_TYPE_Security Ministries
+0.000	ORGANIZATION_TYPE_Hotel
+0.000	CARS_PURP_HC
+0.000	NAME_EDUCATION_TYPE_Academic degree
-0.000	GET_HOME_PURP_HC
-0.000	DIES_TOT_DEUTE_PCB
-0.000	ORGANIZATION_TYPE_Legal Services
-0.000	ORGANIZATION_TYPE_Realtor
-0.001	DIES_MAX_IMPAGATS_Altres_EF__91-120_
-0.001	CRE_SOL_Altres_EF
-0.001	SUM_BALANCE_CC
-0.001	TOT_CO_RTS_PCB_MENSUAL
-0.001	DID_OVERDUE_Altres_EF
-0.002	NAME_HOUSING_TYPE_Rented apartment
-0.002	OCCUPATION_TYPE_Secretaries
-0.002	NUM_QUOTES_PENDENTS
-0.002	NAME_FAMILY_STATUS_Separated
-0.003	UrgentN_PURP_HC
-0.004	MICROLOAN_Altres_EF
-0.004	ORGANIZATION_TYPE_Construction
-0.006	LAST_DUE_MONTH_More_5Years
-0.007	ORGANIZATION_TYPE_Self-employed
-0.007	MEAN_DPD_CC
-0.009	Credits_Assegurats_HC_Mateixos crèdits assegurats
-0.013	Repairs_PURP_HC
-0.013	REGION_POPULATION_RELATIVE
-0.015	TOT_DRAWINGS_CC
-0.018	OCCUPATION_TYPE_Drivers
-0.023	DIES_NOMES_GRAN_DEUTE_PCB
-0.028	Mobile_CAT_HC
-0.030	TOT_RECEIVABLE_CC
-0.031	AMT_ANNUITY_TOT
-0.040	AMT_CREDIT
-0.040	REV_LOANS_TY_HC
-0.048	AGE_EXPECTED
-0.058	NUM_ACTIVE_CREDITS
-0.069	NOT_ENOUGH_MONEY_PAID_IP
-0.091	Home_CAT_HC
-0.116	PERCENTATGE_ANNUITY
-2.541	<BIAS>

Figura 67. Pes per variable per a un individu amb una alta probabilitat d'impagament.

8.4.3 IML d'un client amb una probabilitat d'impagament mitjana

y=0 (probability **0.895**, score **-2.139**) top features

Contribution	Feature
+2.541	<BIAS>
+0.179	RATI_DEUTE_GARANTIA
+0.151	LAST_YEARS_EMPLOYED_0-3
+0.146	REGION_POPULATION_RELATIVE
+0.141	LAST_YEARS_EMPLOYED_10-80
+0.112	Refused_STATUS_HC
+0.099	TOT_DRAWINGS_CC
+0.081	MIN_EXT_SOURCE
+0.073	CAR_LOAN_Altres_EF
+0.066	NOT_ENOUGH_MONEY_PAID_IP
+0.062	PERCENTATGE_ANNUITY
+0.047	NUM_ACTIVE_CREDITS
+0.032	OCCUPATION_TYPE_Laborers
+0.030	SUM_Diferencia_AMT_HC_More credit than ATB
+0.025	DIES_NOMES_GRAN_DEUTE_PCB
+0.024	ORGANIZATION_TYPE_Self-employed
+0.019	CONSUMER_CREDIT_Altres_EF
+0.016	TOT_RECEIVABLE_CC
+0.016	AMT_INCOME_TOTAL
+0.013	Canceled_STATUS_HC
+0.012	NAME_FAMILY_STATUS_Single / not married
+0.010	AGE_EXPECTED
+0.010	MICROLOAN_Altres_EF
+0.009	LAST_DUE_MONTH_Retired
+0.006	ORGANIZATION_TYPE_Medicine
+0.006	LAST_DUE_MONTH_More_5Years
+0.004	NAME_HOUSING_TYPE_With parents
+0.004	ORGANIZATION_TYPE_Transport
+0.004	REV_LOANS_TY_HC
+0.003	UrgentN_PURP_HC
+0.002	SUM_BALANCE_CC
+0.002	NAME_FAMILY_STATUS_Separated
+0.002	NAME_HOUSING_TYPE_Municipal apartment
+0.002	NAME_HOUSING_TYPE_Rented apartment
+0.001	ANY_COTXE_INTERACTION
+0.001	DID_OVERDUE_Altres_EF
+0.001	DIES_MAX_IMPAGATS_Altres_EF__31-60_
+0.001	ORGANIZATION_TYPE_Construction
+0.001	CRE_SOL_Altres_EF
+0.000	ORGANIZATION_TYPE_Realtor
+0.000	DIES_TOT_DEUTE_PCB
+0.000	GET_HOME_PURP_HC
-0.000	NAME_EDUCATION_TYPE_Academic degree
-0.000	CARS_PURP_HC
-0.000	FUN_PURP_HC
-0.000	ORGANIZATION_TYPE_Hotel
-0.000	ORGANIZATION_TYPE_Security Ministries
-0.001	ORGANIZATION_TYPE_Military
-0.001	ORGANIZATION_TYPE_Bank
-0.001	OCCUPATION_TYPE_Accountants
-0.001	REGION_RATING_CLIENT
-0.001	FreeT_CAT_HC
-0.001	ORGANIZATION_TYPE_School
-0.002	Clothes_CAT_HC
-0.002	NAME_INCOME_TYPE_State servant
-0.002	OCCUPATION_TYPE_High skill tech staff
-0.002	OCCUPATION_TYPE_Medicine staff
-0.002	NAME_HOUSING_TYPE_Office apartment
-0.003	LAST_DUE_MONTH_No_DPD_PA
-0.005	NAME_EDUCATION_TYPE_Secondary / secondary special
-0.005	NUM_QUOTES_PENDENTS
-0.005	FLAG_OWN_REALTY_N
-0.009	IS_Refreshed_Client_HC
-0.009	NAME_CONTRACT_TYPE_Cash loans
-0.010	ORGANIZATION_TYPE_Business
-0.012	Home_CAT_HC
-0.013	AMT_CREDIT_ACT_TOT
-0.016	TOT_INSTALMENTS_CC
-0.017	MORTGAGE_LOAN_Altres_EF
-0.020	NAME_EDUCATION_TYPE_Higher education
-0.020	NAME_INCOME_TYPE_Working
-0.020	LAST_DUE_MONTH_Less_6Months
-0.021	NUM_Repeater_Client_HC
-0.025	CASH_LOANS_TY_HC
-0.027	AMT_CREDIT
-0.031	ElectD_CAT_HC
-0.033	DIES_MAX_IMPAGATS_Altres_EF__1-30_
-0.034	MAX_CREDIT_LIM_CC
-0.054	AMT_ANNUITY_TOT
-0.057	FLAG_OWN_CAR_N
-0.058	AMT_GOODS_PRICE
-0.060	NUM_CREDITS_PREVIS_TANCATS
-0.079	Approved_STATUS_HC
-0.082	Mobile_CAT_HC
-0.099	MAX_EXT_SOURCE
-0.104	NAME_FAMILY_STATUS_Married
-0.142	OCCUPATION_TYPE_Drivers
-0.155	CODE_GENDER_F
-0.572	MEAN_EXT_SOURCE

Figura 68. Pes per variable per a un individu amb una probabilitat d'impagament mitjana.

8.5 Codi emprat

A continuació es mostra tot el codi que s'ha fet servir per a la construcció de les dades, la descriptiva, gràfics i construcció dels models.

8.5.1 Codi SQL per a la creació dels fitxers d'entrenament i test

```
# Llibreries a utilitzar
import os
import pandas as pd

# Establir el directori de treball: si treballo al meu pc (- potent)
os.getcwd() # Directori actual
os.chdir("E:/Grau d'Estadística/4t/TFG/Bases de dades")

# Directori si treballo al pc d'en Carles: (+ potent)


# Lectura de les bases de dades
# Base de dades train: Cada fila representa un préstec.
# Per aplicar les comandes de SQL més ràpid i ocupar menys només es llegeix
l'identificador.
train = pd.read_csv("application_train.csv", sep=",", decimal=".", usecols =
["SK_ID_CURR"])

# Extracció de la base de dades definitiva mitjançant codi SQL a partir del
package pandasql
from pandasql import sqldf

###############################################################################
# %%

## Altres Entitats Financeres: bureau i bureau_balance
#### BUREAU

# Base de dades bureau: Altres crèdits proporcionats per altres institucions
financeres.

bureau = pd.read_csv("bureau.csv", sep=",", decimal=".", usecols =
["SK_ID_CURR", "SK_ID_BUREAU", "CREDIT_ACTIVE",
                     "CREDIT_TYPE",
"CREDIT_DAY_OVERDUE","CNT_CREDIT_PROLONG","AMT_ANNUITY","AMT_CREDIT_SUM"])

# Quants ID diferents de train hi ha a bureau
len(set(train['SK_ID_CURR'].unique()) & set(bureau['SK_ID_CURR'].unique()))

# %%

bureau_SQL = """ SELECT train.SK_ID_CURR,
    CASE WHEN CREDITS_ACTIUS_TOT IS NULL THEN 0 ELSE CREDITS_ACTIUS_TOT END AS
CRE_ACT_Altres_EF,
    CASE WHEN CREDITS_TANCATS_TOT IS NULL THEN 0 ELSE CREDITS_TANCATS_TOT END
AS CRE_CLO_Altres_EF,
    CASE WHEN CREDITS_VENUTS_TOT IS NULL THEN 0 ELSE CREDITS_VENUTS_TOT END AS
CRE_SOL_Altres_EF,
    CASE WHEN CREDITS_DEUTEDOLENT_TOT IS NULL THEN 0 ELSE
CREDITS_DEUTEDOLENT_TOT END AS CRE_BDE_Altres_EF,
    CASE WHEN CRE_TOT_Altres_EF IS NULL THEN 0 ELSE CRE_TOT_Altres_EF END AS
CRE_TOTAL_Altres_EF,
    CASE WHEN CRE_TOT_Altres_EF IS NULL THEN 0 ELSE 1 END AS
TE_CREDIT_Altres_EF,
    CASE WHEN Another_Type_LOAN_EF > 0 THEN 1 ELSE 0 END as
CRE_Another_Type_LOAN_Altres_EF,
    CASE WHEN CAR_LOAN_EF > 0 THEN 1 ELSE 0 END as CAR_LOAN_Altres_EF,
```

```
    CASE WHEN CONSUMER_CREDIT_EF > 0 THEN 1 ELSE 0 END as
CONSUMER_CREDIT_Altres_EF,
    CASE WHEN CREDIT_CARD_EF > 0 THEN 1 ELSE 0 END as CREDIT_CARD_Altres_EF,
    CASE WHEN BUSINESS_LOAN_EF > 0 THEN 1 ELSE 0 END as
BUSINESS_LOAN_Altres_EF,
    CASE WHEN MICROLOAN_EF > 0 THEN 1 ELSE 0 END as MICROLOAN_Altres_EF,
    CASE WHEN MORTGAGE_LOAN_EF > 0 THEN 1 ELSE 0 END as
MORTGAGE_LOAN_Altres_EF,
    CASE WHEN UNKNOWN_LOAN_EF > 0 THEN 1 ELSE 0 END as UNKNOWN_LOAN_Altres_EF,
    SUM(CASE WHEN AMOUNT_ANNUITY_Active_EF > 0 THEN AMOUNT_ANNUITY_Active_EF
ELSE 0 END) AS AMOUNT_ANNUITY_Active_EF,
    CASE WHEN AMOUNT_CURRENT_CREDIT_Active_EF IS NULL THEN 0 ELSE
AMOUNT_CURRENT_CREDIT_Active_EF END AS AMOUNT_CURRENT_CREDIT_Active_EF,
    CASE WHEN SUM(CASE WHEN bureau.CREDIT_DAY_OVERDUE IS NULL THEN 0 ELSE
bureau.CREDIT_DAY_OVERDUE END) > 0 THEN 1 ELSE 0 END AS DID_OVERDUE_Altres_EF,
    CASE WHEN SUM(CASE WHEN bureau.CNT_CREDIT_PROLONG IS NULL THEN 0 ELSE
bureau.CNT_CREDIT_PROLONG END) > 0 THEN 1 ELSE 0 END AS DID_PROLONG_Altres_EF

    FROM train LEFT JOIN (SELECT bureau.SK_ID_CURR,
SUM(bureau.CREDIT_DAY_OVERDUE) AS CREDIT_DAY_OVERDUE, SUM(CNT_CREDIT_PROLONG)
AS CNT_CREDIT_PROLONG,
            SUM(CASE WHEN bureau.CREDIT_TYPE= "Another type of loan" THEN 1
ELSE 0 END) as Another_Type_LOAN_EF,
            SUM(CASE WHEN bureau.CREDIT_TYPE= "Car loan" THEN 1 ELSE 0 END) as
CAR_LOAN_EF,
            SUM(CASE WHEN bureau.CREDIT_TYPE= "Consumer credit" THEN 1 ELSE 0
END) as CONSUMER_CREDIT_EF,
            SUM(CASE WHEN bureau.CREDIT_TYPE= "Credit card" THEN 1 ELSE 0 END)
as CREDIT_CARD_EF,
            SUM(CASE WHEN bureau.CREDIT_TYPE= "Loan for business development"
THEN 1 ELSE 0 END) as BUSINESS_LOAN_EF,
            SUM(CASE WHEN bureau.CREDIT_TYPE= "Microloan" THEN 1 ELSE 0 END)
as MICROLOAN_EF,
            SUM(CASE WHEN bureau.CREDIT_TYPE= "Mortgage" THEN 1 ELSE 0 END) as
MORTGAGE_LOAN_EF,
            SUM(CASE WHEN bureau.CREDIT_TYPE= "Unknown type of loan" THEN 1
ELSE 0 END) as UNKNOWN_LOAN_EF,
            COUNT(bureau.SK_ID_CURR) as CRE_TOT_Altres_EF,
            SUM(CASE WHEN bureau.CREDIT_ACTIVE= "Active" THEN 1 ELSE 0 END) as
CREDITS_ACTIUS_TOT,
            SUM(CASE WHEN bureau.CREDIT_ACTIVE= "Closed" THEN 1 ELSE 0 END) as
CREDITS_TANCATS_TOT,
            SUM(CASE WHEN bureau.CREDIT_ACTIVE= "Sold" THEN 1 ELSE 0 END) as
CREDITS_VENUTS_TOT,
            SUM(CASE WHEN bureau.CREDIT_ACTIVE= "Bad debt" THEN 1 ELSE 0 END)
as CREDITS_DEUTEDOLENT_TOT,
            SUM(CASE WHEN bureau.CREDIT_ACTIVE= "Active" THEN
bureau.AMT_ANNUITY ELSE 0 END) AS AMOUNT_ANNUITY_Active_EF,
            SUM(CASE WHEN bureau.CREDIT_ACTIVE= "Active" THEN
bureau.AMT_CREDIT_SUM ELSE 0 END) AS AMOUNT_CURRENT_CREDIT_Active_EF
            FROM bureau
            GROUP BY bureau.SK_ID_CURR) AS bureau ON train.SK_ID_CURR =
bureau.SK_ID_CURR
  GROUP BY train.SK_ID_CURR;"""

bureau_data = sqldf(bureau_SQL, locals())
del(bureau_SQL)

#%%
##############################################################################
#### BUREAU_BALANCE
    #Veure si l'individu (ID) va pagar, almenys amb x dies de retard:

# Base de dades bureau_balance
bureau_balance = pd.read_csv("bureau_balance.csv", sep=",", decimal=".",
usecols = ["SK_ID_BUREAU","MONTHS_BALANCE", "STATUS"])

# Reducció de les dimensions de la taula
```

```
new_bureau_balance = """ SELECT bureau_balance.SK_ID_BUREAU,
bureau_balance.STATUS, bureau.SK_ID_CURR
                         FROM bureau_balance INNER JOIN bureau ON
bureau_balance.SK_ID_BUREAU = bureau.SK_ID_BUREAU
                         ;"""

new_bureau_balance = sqldf(new_bureau_balance, locals())
del(bureau_balance)

#Eliminem aquells nivells que no ens interessen:
new_bureau_balance =
new_bureau_balance.drop(new_bureau_balance[new_bureau_balance.STATUS ==
"C"].index)
new_bureau_balance =
new_bureau_balance.drop(new_bureau_balance[new_bureau_balance.STATUS ==
"X"].index)
new_bureau_balance =
new_bureau_balance.drop(new_bureau_balance[new_bureau_balance.STATUS ==
"0"].index)
new_bureau_balance = new_bureau_balance.drop(["SK_ID_BUREAU"],1)

# Màxim dies impagats: Veure el màxim de dies que l'individu (ID) va tardar en
pagar. Si el màxim són 90 dies, els de 1-30, 30-60 i 60-90 no es tenen en
compte.

dies_retard_EF = """SELECT train.SK_ID_CURR,
                    CASE WHEN COMPTATS IS NULL THEN "PERFECT" ELSE COMPTATS
END AS DIES_MAX_IMPAGATS_Altres_EF
                    FROM train LEFT JOIN (SELECT
new_bureau_balance.SK_ID_CURR,
                        CASE WHEN MAX(new_bureau_balance.STATUS) == '1' THEN
"[1-30]"
                        WHEN MAX(new_bureau_balance.STATUS) == '2' THEN "[31-
60]"
                        WHEN MAX(new_bureau_balance.STATUS) == '3' THEN "[61-
90]"
                        WHEN MAX(new_bureau_balance.STATUS) == '4' THEN "[91-
120]"
                        WHEN MAX(new_bureau_balance.STATUS) == '5' THEN "+120"
ELSE "PERFECT" END AS COMPTATS
                        FROM new_bureau_balance
                        GROUP BY SK_ID_CURR) AS UNIO ON train.SK_ID_CURR =
UNIO.SK_ID_CURR;"""

dies_retard_Altres_EF = sqldf(dies_retard_EF,locals())

del(bureau)
del(new_bureau_balance)
del(dies_retard_EF)

# Unió del bureau amb el bureau_balance
information_Altres_EF = pd.merge(bureau_data, dies_retard_Altres_EF , left_on
= "SK_ID_CURR", right_on = "SK_ID_CURR")
information_Altres_EF.dtypes

del(bureau_data)
del(dies_retard_Altres_EF)

#%%
##############################################################################

# Les següents quatre bases de dades estan connectades:
# PREVIOUS_APPLICATION: (és la que connecta les altres tres)
#XNA: Information not available.
#XPA: No aplicable.

previous_application = pd.read_csv("previous_application.csv", sep=",",
decimal=".", usecols = ["SK_ID_PREV","SK_ID_CURR",
```

```
                    "NAME_CASH_LOAN_PURPOSE", "NAME_CONTRACT_TYPE",
"AMT_ANNUITY", "AMT_APPLICATION","AMT_CREDIT","NAME_CONTRACT_STATUS",

"NAME_PAYMENT_TYPE","CODE_REJECT_REASON","NAME_CLIENT_TYPE",
"NAME_GOODS_CATEGORY",
                     "CNT_PAYMENT", "NAME_YIELD_GROUP",
"NFLAG_INSURED_ON_APPROVAL", "DAYS_LAST_DUE"])

len(set(train['SK_ID_CURR'].unique()) &
set(previous_application['SK_ID_CURR'].unique()))

#eee = previous_application.AMT_ANNUITY[previous_application["AMT_ANNUITY"] ==
0]

previous_application[["SK_ID_CURR","AMT_ANNUITY"]].groupby(["SK_ID_CURR"]).agg
(['mean'])


# Les variables categòriques que CREC que poden ser d'interès:
previous_application[["SK_ID_PREV","NAME_CONTRACT_TYPE"]].groupby(["NAME_CONTR
ACT_TYPE"]).agg(['count'])

##
previous_application[["SK_ID_PREV","NAME_CASH_LOAN_PURPOSE"]].groupby(["NAME_C
ASH_LOAN_PURPOSE"]).agg(['count'])

# Ajuntem alguns nivells:
    # No Information available: XAP i XNA
    # Desconegut (poden ser que siguin un d'aquests següents) + Refusal to
name the goal?
    # Urgent: Urgent needs.
    # Reparacions: Repairs.
    # Altres (teoricament es saben quins són): Others
    # Pagament per altres préstecs: Payments on other loans.
    # Comprar / Adquirir una vivenda o terreny: Building a house or an annex,
buying a garage, buying a holiday home / land, Buying a home
    # Tot el que tingui a veure amb un cotxe: Buying a new / used car, Car
repairs (aquesta podria anar a reparacions)
    # Serveis bàsics: Education, Medicine, Gasification / Water supply,
Everyday expenses
    # Serveis secundaris / ocasions esporàdiques: Furniture, Hobby, Journey,
Money for a third person, Purchase of electronic equipment, Wedding / Gift /
Holiday

# Recodificació (No sé com ajuntar aquestes d'aqui) (PROPOSTA)
previous_application.loc[(previous_application['NAME_CASH_LOAN_PURPOSE'] ==
'XAP')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'XNA'),
'NAME_CASH_LOAN_PURPOSE'] = "No information available"

previous_application.loc[(previous_application['NAME_CASH_LOAN_PURPOSE'] ==
'Building a house or an annex')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'Buying a
garage')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'Buying a
holiday home / land')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'Furniture')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'Buying a
home'), 'NAME_CASH_LOAN_PURPOSE'] = "Getting a property (a house mainly)"

previous_application.loc[(previous_application['NAME_CASH_LOAN_PURPOSE'] ==
'Buying a new car')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'Buying a used
car')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'Car repairs'),
'NAME_CASH_LOAN_PURPOSE'] = "Related to cars (purchase / repairs)"
```

```
previous_application.loc[(previous_application['NAME_CASH_LOAN_PURPOSE'] ==
'Education')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'Medicine')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'Gasification /
water supply')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'Purchase of
electronic equipment')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'Everyday
expenses'), 'NAME_CASH_LOAN_PURPOSE'] = "Necessity Goods"

previous_application.loc[(previous_application['NAME_CASH_LOAN_PURPOSE'] ==
'Hobby')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'Journey')|
           (previous_application['NAME_CASH_LOAN_PURPOSE'] == 'Wedding / gift
/ holiday'), 'NAME_CASH_LOAN_PURPOSE'] = "Related to have fun"

previous_application.loc[(previous_application['NAME_CASH_LOAN_PURPOSE'] ==
'Money for a third person')|
                         (previous_application['NAME_CASH_LOAN_PURPOSE'] ==
'Refusal to name the goal'), 'NAME_CASH_LOAN_PURPOSE'] = "Other"

previous_application[["SK_ID_PREV","NAME_CASH_LOAN_PURPOSE"]].groupby(["NAME_C
ASH_LOAN_PURPOSE"]).agg(['count'])

###

previous_application[["SK_ID_PREV","NAME_CONTRACT_STATUS"]].groupby(["NAME_CON
TRACT_STATUS"]).agg(['count'])
#
previous_application[["SK_ID_PREV","NAME_PAYMENT_TYPE"]].groupby(["NAME_PAYMEN
T_TYPE"]).agg(['count'])

previous_application[["SK_ID_PREV","NAME_CLIENT_TYPE"]].groupby(["NAME_CLIENT_
TYPE"]).agg(['count'])
previous_application.loc[(previous_application['NAME_CLIENT_TYPE'] == 'XNA'),
'NAME_CLIENT_TYPE'] = "No information available"
previous_application[["SK_ID_PREV","NAME_CLIENT_TYPE"]].groupby(["NAME_CLIENT_
TYPE"]).agg(['count'])

### RECODIFICACIÓ De la variable NAME_GOODS_CATEGORY
previous_application[["SK_ID_PREV","NAME_GOODS_CATEGORY"]].groupby(["NAME_GOOD
S_CATEGORY"]).agg(['count'])

previous_application.loc[(previous_application['NAME_GOODS_CATEGORY'] ==
'XNA'), 'NAME_GOODS_CATEGORY'] = "No information given"

previous_application.loc[(previous_application['NAME_GOODS_CATEGORY'] ==
'Audio/Video')|
           (previous_application['NAME_GOODS_CATEGORY'] == 'Computers')|
           (previous_application['NAME_GOODS_CATEGORY'] == 'Consumer
Electronics')|
           (previous_application['NAME_GOODS_CATEGORY'] == 'Office
Appliances')|  #appliance: electrodomèstic
           (previous_application['NAME_GOODS_CATEGORY'] == 'Photo / Cinema
Equipment'), 'NAME_GOODS_CATEGORY'] = "Related to electronic devices"

previous_application.loc[(previous_application['NAME_GOODS_CATEGORY'] == 'Auto
Accessories')|
           (previous_application['NAME_GOODS_CATEGORY'] == 'Clothing and
Accessories')|
           (previous_application['NAME_GOODS_CATEGORY'] == 'Jewelry'),
'NAME_GOODS_CATEGORY'] = "Related to clothes and accessories"

previous_application.loc[(previous_application['NAME_GOODS_CATEGORY'] ==
'Insurance')|
           (previous_application['NAME_GOODS_CATEGORY'] == 'Medical
Supplies')|
           (previous_application['NAME_GOODS_CATEGORY'] == 'Education')|
```

```
          (previous_application['NAME_GOODS_CATEGORY'] == 'Medicine'),
'NAME_GOODS_CATEGORY'] = "Related to health and goods"

previous_application.loc[(previous_application['NAME_GOODS_CATEGORY'] ==
'Furniture')|
          (previous_application['NAME_GOODS_CATEGORY'] == 'Homewares')|
          (previous_application['NAME_GOODS_CATEGORY'] == 'House
Construction')|
          (previous_application['NAME_GOODS_CATEGORY'] == 'Gardening'),
'NAME_GOODS_CATEGORY'] = "Related to home"

previous_application.loc[(previous_application['NAME_GOODS_CATEGORY'] ==
'Fitness')|
          (previous_application['NAME_GOODS_CATEGORY'] == 'Sport and
Leisure')|
          (previous_application['NAME_GOODS_CATEGORY'] == 'Tourism'),
'NAME_GOODS_CATEGORY'] = "Related to free time"

previous_application.loc[(previous_application['NAME_GOODS_CATEGORY'] ==
'Additional Service')|
          (previous_application['NAME_GOODS_CATEGORY'] == 'Animals')|
          (previous_application['NAME_GOODS_CATEGORY'] == 'Direct Sales')|
          (previous_application['NAME_GOODS_CATEGORY'] == 'Weapon'),
'NAME_GOODS_CATEGORY'] = "Other"

previous_application[["SK_ID_PREV","NAME_GOODS_CATEGORY"]].groupby(["NAME_GOOD
S_CATEGORY"]).agg(['count'])

##

previous_application[["SK_ID_PREV","CNT_PAYMENT"]].groupby(["CNT_PAYMENT"]).ag
g(['count'])
previous_application[["SK_ID_PREV","NFLAG_INSURED_ON_APPROVAL"]].groupby(["NFL
AG_INSURED_ON_APPROVAL"]).agg(['count'])
previous_application.loc[(previous_application['NFLAG_INSURED_ON_APPROVAL'] ==
1), 'NFLAG_INSURED_ON_APPROVAL'] = "Insured"
previous_application.loc[(previous_application['NFLAG_INSURED_ON_APPROVAL'] ==
0), 'NFLAG_INSURED_ON_APPROVAL'] = "Uninsured"
previous_application[["SK_ID_PREV","NFLAG_INSURED_ON_APPROVAL"]].groupby(["NFL
AG_INSURED_ON_APPROVAL"]).agg(['count'])

# Consulta SQL
previous_app_SQL = """ SELECT train.SK_ID_CURR,
                       CASE WHEN CASH_LOANS_TY_HC > 0 THEN 1 ELSE 0 END AS
CASH_LOANS_TY_HC,
                       CASE WHEN CONS_LOANS_TY_HC > 0 THEN 1 ELSE 0 END AS
CONS_LOANS_TY_HC,
                       CASE WHEN REV_LOANS_TY_HC > 0 THEN 1 ELSE 0 END AS
REV_LOANS_TY_HC,
                       CASE WHEN NC_Goods_PURP_HC IS NULL THEN 0 ELSE
NC_Goods_PURP_HC END AS NC_Goods_PURP_HC,
                       CASE WHEN Repairs_PURP_HC IS NULL THEN 0 ELSE
Repairs_PURP_HC END AS Repairs_PURP_HC,
                       CASE WHEN UrgentN_PURP_HC IS NULL THEN 0 ELSE
UrgentN_PURP_HC END AS UrgentN_PURP_HC,
                       CASE WHEN GET_HOME_PURP_HC IS NULL THEN 0 ELSE
GET_HOME_PURP_HC END AS GET_HOME_PURP_HC,
                       CASE WHEN CARS_PURP_HC IS NULL THEN 0 ELSE CARS_PURP_HC
END AS CARS_PURP_HC,
                       CASE WHEN FUN_PURP_HC IS NULL THEN 0 ELSE FUN_PURP_HC
END AS FUN_PURP_HC,
                       CASE WHEN Approved_STATUS_HC IS NULL THEN 0 ELSE
Approved_STATUS_HC END AS Approved_STATUS_HC,
                       CASE WHEN Canceled_STATUS_HC IS NULL THEN 0 ELSE
Canceled_STATUS_HC END AS Canceled_STATUS_HC,
                       CASE WHEN Refused_STATUS_HC IS NULL THEN 0 ELSE
Refused_STATUS_HC END AS Refused_STATUS_HC,
```

```
                        CASE WHEN UnusedOF_STATUS_HC IS NULL THEN 0 ELSE
UnusedOF_STATUS_HC END AS UnusedOF_STATUS_HC,
                        CASE WHEN ElectD_CAT_HC IS NULL THEN 0 ELSE
ElectD_CAT_HC END AS ElectD_CAT_HC,
                        CASE WHEN Clothes_CAT_HC IS NULL THEN 0 ELSE
Clothes_CAT_HC END AS Clothes_CAT_HC,
                        CASE WHEN FreeT_CAT_HC IS NULL THEN 0 ELSE FreeT_CAT_HC
END AS FreeT_CAT_HC,
                        CASE WHEN Health_CAT_HC IS NULL THEN 0 ELSE
Health_CAT_HC END AS Health_CAT_HC,
                        CASE WHEN Home_CAT_HC IS NULL THEN 0 ELSE Home_CAT_HC
END AS Home_CAT_HC,
                        CASE WHEN Mobile_CAT_HC IS NULL THEN 0 ELSE
Mobile_CAT_HC END AS Mobile_CAT_HC,
                        CASE WHEN ConstructionM_CAT_HC IS NULL THEN 0 ELSE
ConstructionM_CAT_HC END AS ConstructionM_CAT_HC,
                        CASE WHEN Vehicles_CAT_HC IS NULL THEN 0 ELSE
Vehicles_CAT_HC END AS Vehicles_CAT_HC,
                        CASE WHEN Uninsured_HC < Insured_HC THEN "Més crèdits
assegurats"
                             WHEN Uninsured_HC > Insured_HC THEN "Menys crèdits
assegurats" ELSE "Mateixos crèdits assegurats" END AS Credits_Assegurats_HC,
                        CASE WHEN Recent_Client_HC IS NULL THEN 1 ELSE
Recent_Client_HC END AS IS_NEW_CLIENT_TRAIN,
                        CASE WHEN Refreshed_Client_HC > 0 THEN 1 ELSE 0 END AS
IS_Refreshed_Client_HC,
                        CASE WHEN Repeater_Client_HC > 0 THEN
Repeater_Client_HC + Recent_Client_HC ELSE 0 END AS NUM_Repeater_Client_HC,
                        CASE WHEN DIFERENCIA > 0 THEN 'More credit than ATB'
                             WHEN DIFERENCIA < 0 THEN 'Less credit than ATB'
                             ELSE 'Same credit than ATB' END AS
SUM_Diferencia_AMT_HC,
                        LAST_DUE_MONTH
                        FROM train LEFT JOIN (SELECT SK_ID_CURR,
                                                     SUM(CASE WHEN
NAME_CONTRACT_TYPE = "Cash loans" THEN 1 ELSE 0 END) AS CASH_LOANS_TY_HC,
                                                     SUM(CASE WHEN
NAME_CONTRACT_TYPE = "Consumer loans" THEN 1 ELSE 0 END) AS CONS_LOANS_TY_HC,
                                                     SUM(CASE WHEN
NAME_CONTRACT_TYPE = "Revolving loans" THEN 1 ELSE 0 END) AS REV_LOANS_TY_HC,
                                                     SUM(CASE WHEN
NAME_CASH_LOAN_PURPOSE = "Repairs" THEN 1 ELSE 0 END) AS Repairs_PURP_HC,
                                                     SUM(CASE WHEN
NAME_CASH_LOAN_PURPOSE = "Necessity Goods" THEN 1 ELSE 0 END) AS
NC_Goods_PURP_HC,
                                                     SUM(CASE WHEN
NAME_CASH_LOAN_PURPOSE = "Urgent needs" THEN 1 ELSE 0 END) AS UrgentN_PURP_HC,
                                                     SUM(CASE WHEN
NAME_CASH_LOAN_PURPOSE = "Getting a property (a house mainly)" THEN 1 ELSE 0
END) AS GET_HOME_PURP_HC,
                                                     SUM(CASE WHEN
NAME_CASH_LOAN_PURPOSE = "Related to cars (purchase / repairs)" THEN 1 ELSE 0
END) AS CARS_PURP_HC,
                                                     SUM(CASE WHEN
NAME_CASH_LOAN_PURPOSE = "Related to have fun" THEN 1 ELSE 0 END) AS
FUN_PURP_HC,
                                                     SUM(CASE WHEN
NAME_CONTRACT_STATUS = "Approved" THEN 1 ELSE 0 END) AS Approved_STATUS_HC,
                                                     SUM(CASE WHEN
NAME_CONTRACT_STATUS = "Canceled" THEN 1 ELSE 0 END) AS Canceled_STATUS_HC,
                                                     SUM(CASE WHEN
NAME_CONTRACT_STATUS = "Refused" THEN 1 ELSE 0 END) AS Refused_STATUS_HC,
                                                     SUM(CASE WHEN
NAME_CONTRACT_STATUS = "Unused offer" THEN 1 ELSE 0 END) AS
UnusedOF_STATUS_HC,
                                                     SUM(CASE WHEN
NAME_CLIENT_TYPE = "New" THEN 0 ELSE 0 END) AS Recent_Client_HC,
```

```
                                                    SUM(CASE WHEN
NAME_CLIENT_TYPE = "Refreshed" THEN 1 ELSE 0 END) AS Refreshed_Client_HC,
                                                    SUM(CASE WHEN
NAME_CLIENT_TYPE = "Repeater" THEN 1 ELSE 0 END) AS Repeater_Client_HC,
                                                    SUM(CASE WHEN
NAME_GOODS_CATEGORY = "Related to electronic devices" THEN 1 ELSE 0 END) AS
ElectD_CAT_HC,
                                                    SUM(CASE WHEN
NAME_GOODS_CATEGORY = "Related to clothes and accessories" THEN 1 ELSE 0 END)
AS Clothes_CAT_HC,
                                                    SUM(CASE WHEN
NAME_GOODS_CATEGORY = "Related to free time" THEN 1 ELSE 0 END) AS
FreeT_CAT_HC,
                                                    SUM(CASE WHEN
NAME_GOODS_CATEGORY = "Related to health and goods" THEN 1 ELSE 0 END) AS
Health_CAT_HC,
                                                    SUM(CASE WHEN
NAME_GOODS_CATEGORY = "Related to home" THEN 1 ELSE 0 END) AS Home_CAT_HC,
                                                    SUM(CASE WHEN
NAME_GOODS_CATEGORY = "Mobile" THEN 1 ELSE 0 END) AS Mobile_CAT_HC,
                                                    SUM(CASE WHEN
NAME_GOODS_CATEGORY = "Construction Materials" THEN 1 ELSE 0 END) AS
ConstructionM_CAT_HC,
                                                    SUM(CASE WHEN
NAME_GOODS_CATEGORY = "Vehicles" THEN 1 ELSE 0 END) AS Vehicles_CAT_HC,
                                                    SUM(CASE WHEN
NFLAG_INSURED_ON_APPROVAL = "Insured" THEN 1 ELSE 0 END) AS Insured_HC,
                                                    SUM(CASE WHEN
NFLAG_INSURED_ON_APPROVAL = "Uninsured" THEN 1 ELSE 0 END) AS Uninsured_HC,
                                                    SUM(AMT_CREDIT -
AMT_APPLICATION) AS DIFERENCIA,

ROUND(MIN(ABS(DAYS_LAST_DUE))/30,0) AS LAST_DUE_MONTH
                                                FROM previous_application
                                                GROUP BY
previous_application.SK_ID_CURR) AS UNIO ON train.SK_ID_CURR =
UNIO.SK_ID_CURR;"""

previous_application_SQL = sqldf(previous_app_SQL, locals())

# Modifiquem la variable LAST_DUE_MONTH.
import numpy as np
bins = [0, 6, 12, 24,60,12000,50000, np.inf]
names = ['Less_6Months', 'Less_Year', 'Between_1Year_2Years',
'Between_2Years_5Years','More_5Years','Retired','No_DPD_PA']
previous_application_SQL['LAST_DUE_MONTH'] =
pd.cut(previous_application_SQL['LAST_DUE_MONTH'], bins, labels=names)

del(bins)
del(names)
del(previous_app_SQL)

##############################

# CREDIT_CARD_BALANCE:
credit_card_balance = pd.read_csv("credit_card_balance.csv", sep=",",
decimal=".", usecols = ["SK_ID_PREV", "SK_ID_CURR",

"MONTHS_BALANCE","AMT_BALANCE","AMT_CREDIT_LIMIT_ACTUAL","AMT_PAYMENT_TOTAL_CU
RRENT","AMT_TOTAL_RECEIVABLE",

"CNT_DRAWINGS_ATM_CURRENT","CNT_INSTALMENT_MATURE_CUM","NAME_CONTRACT_STATUS",
"SK_DPD","SK_DPD_DEF"])

len(set(train['SK_ID_CURR'].unique()) &
set(credit_card_balance['SK_ID_CURR'].unique()))

# Consulta SQL
```

```
credit_card = """ SELECT train.SK_ID_CURR,
                  CASE WHEN SUM_BALANCE_CC IS NULL THEN 0 ELSE SUM_BALANCE_CC
END AS SUM_BALANCE_CC,
                  MAX_CREDIT_LIM_CC,
                  CASE WHEN TOT_RECEIVABLE_CC IS NULL THEN 0 ELSE
TOT_RECEIVABLE_CC END AS TOT_RECEIVABLE_CC,
                  CASE WHEN TOT_DRAWINGS_CC IS NULL THEN 0 ELSE
TOT_DRAWINGS_CC END AS TOT_DRAWINGS_CC,
                  CASE WHEN TOT_INSTALMENTS_CC IS NULL THEN 0 ELSE
TOT_INSTALMENTS_CC END AS TOT_INSTALMENTS_CC,
                  CASE WHEN MEAN_DPD_CC IS NULL THEN 0 ELSE MEAN_DPD_CC END AS
MEAN_DPD_CC,
                  SUM(CASE WHEN TE_CONTR_ACT_CC IS NULL THEN 0 ELSE
TE_CONTR_ACT_CC END) AS TE_CONTR_ACT_CC,
                  CASE WHEN TE_CREDIT_CARD_CC IS NULL THEN 0 ELSE 1 END AS
TE_CREDIT_CARD_CC

                  FROM train LEFT JOIN (SELECT SK_ID_CURR AS SK_ID_CURR_2,
                                        SUM(AMT_BALANCE) AS
SUM_BALANCE_CC,
                                        MAX(AMT_CREDIT_LIMIT_ACTUAL) AS
MAX_CREDIT_LIM_CC,
                                        SUM(AMT_TOTAL_RECEIVABLE) AS
TOT_RECEIVABLE_CC,
                                        SUM(CNT_DRAWINGS_ATM_CURRENT) AS
TOT_DRAWINGS_CC,
                                        SUM(CNT_INSTALMENT_MATURE_CUM) AS
TOT_INSTALMENTS_CC,
                                        AVG(SK_DPD_DEF) AS MEAN_DPD_CC,
                                        CASE WHEN NAME_CONTRACT_STATUS =
"Active" AND MONTHS_BALANCE = MAX(MONTHS_BALANCE) THEN 1 ELSE 0 END AS
TE_CONTR_ACT_CC,
                                        COUNT(SK_ID_CURR) AS
TE_CREDIT_CARD_CC

                                        FROM credit_card_balance
                                        GROUP BY
credit_card_balance.SK_ID_PREV) AS UNIO ON train.SK_ID_CURR =
UNIO.SK_ID_CURR_2
                  GROUP BY train.SK_ID_CURR;"""

credit_card_SQL = sqldf(credit_card, locals())
del(credit_card)

############################
# POS_CASH_balance
POS_CASH_balance = pd.read_csv("POS_CASH_balance.csv", sep=",", decimal=".",
usecols = ["SK_ID_PREV","SK_ID_CURR", "MONTHS_BALANCE",
                   "CNT_INSTALMENT", "CNT_INSTALMENT_FUTURE",
"NAME_CONTRACT_STATUS", "SK_DPD", "SK_DPD_DEF"])

hola = POS_CASH_balance[POS_CASH_balance.MONTHS_BALANCE == -1]
hola.NAME_CONTRACT_STATUS.head(100)

len(set(train['SK_ID_CURR'].unique()) &
set(POS_CASH_balance['SK_ID_CURR'].unique()))

hola[['SK_ID_PREV','SK_ID_CURR', 'CNT_INSTALMENT_FUTURE']].head(30)

#Reducció de les dimensions per a que no doni Memory Error:
reduccio1 = """ SELECT POS_CASH_balance.SK_ID_PREV,
POS_CASH_balance.SK_ID_CURR, POS_CASH_balance.CNT_INSTALMENT,
POS_CASH_balance.CNT_INSTALMENT_FUTURE,
                POS_CASH_balance.NAME_CONTRACT_STATUS,
POS_CASH_balance.SK_DPD, POS_CASH_balance.SK_DPD_DEF,
POS_CASH_balance.MONTHS_BALANCE
                FROM POS_CASH_balance INNER JOIN train ON
POS_CASH_balance.SK_ID_CURR = train.SK_ID_CURR;"""
```

```
POS_CASH_balance_new = sqldf(reduccio1, locals())
del(reduccio1)

# Agafar l'últim mes per a no duplicar. Pas a pas per evitar el Memory Error.
POS_CASH_balance_1 = """ SELECT train.SK_ID_CURR,
            CASE WHEN SUM(Cont_Active) IS NULL THEN 0 ELSE SUM(Cont_Active)
END AS TOT_CO_ACT_PCB,
            CASE WHEN SUM(Cont_Completed) >=0 THEN SUM(Cont_Completed) ELSE 0
END AS TOT_CO_COM_PCB_MENSUAL,
            CASE WHEN SUM(Cont_Signed) >= 0 THEN SUM(Cont_Signed) ELSE 0 END
AS TOT_CO_SIG_PCB_MENSUAL,
            CASE WHEN SUM(Cont_RTTS) >=0 THEN SUM(Cont_RTTS) ELSE 0 END AS
TOT_CO_RTS_PCB_MENSUAL,
            CASE WHEN SUM(Cont_Others) >= 0 THEN SUM(Cont_Others) ELSE 0 END
AS TOT_CO_ALT_PCB_MENSUAL
            FROM train LEFT JOIN (SELECT SK_ID_CURR, CNT_INSTALMENT_FUTURE,
MONTHS_BALANCE,
                                  CASE WHEN NAME_CONTRACT_STATUS = "Active"
AND MONTHS_BALANCE = -1 THEN 1 ELSE 0 END AS Cont_Active,
                                  CASE WHEN NAME_CONTRACT_STATUS =
"Completed" THEN 1 ELSE 0 END AS Cont_Completed,
                                  CASE WHEN NAME_CONTRACT_STATUS = "Signed"
THEN 1 ELSE 0 END AS Cont_Signed,
                                  CASE WHEN NAME_CONTRACT_STATUS = "Returned
to the store" THEN 1 ELSE 0 END AS Cont_RTTS,
                                  CASE WHEN NAME_CONTRACT_STATUS = "Amortized
debt" OR NAME_CONTRACT_STATUS = "Approved"
                                  OR NAME_CONTRACT_STATUS = "Canceled" OR
NAME_CONTRACT_STATUS = "Demand" OR
                                  NAME_CONTRACT_STATUS = "XNA" THEN 1 ELSE 0
END AS Cont_Others
                                  FROM POS_CASH_balance_new) AS UNIO ON
train.SK_ID_CURR = UNIO.SK_ID_CURR
            GROUP BY train.SK_ID_CURR;"""

POS_1 = sqldf(POS_CASH_balance_1, locals())
POS_CASH_balance_2 = """ SELECT *,
                         CASE WHEN DIES_NOMES_GRAN_DEUTE_PCB IS NULL THEN 0
ELSE DIES_NOMES_GRAN_DEUTE_PCB END AS DAYS_BIG_DUE_ONLY_PCB,
                         CASE WHEN DIES_TOT_DEUTE_PCB IS NULL THEN 0 ELSE
DIES_TOT_DEUTE_PCB END AS DAYS_ALL_DUE_PCB
                         FROM POS_1 LEFT JOIN (SELECT SK_ID_CURR AS
SK_ID_CURR2,
                                  SUM(SK_DPD_DEF) AS
DIES_NOMES_GRAN_DEUTE_PCB,
                                  SUM(SK_DPD) AS DIES_TOT_DEUTE_PCB
                                  FROM POS_CASH_balance_new
                                  GROUP BY SK_ID_CURR) AS UNIO ON
POS_1.SK_ID_CURR = UNIO.SK_ID_CURR2;"""

POS_2 = sqldf(POS_CASH_balance_2, locals())
POS_CASH_data = POS_2
POS_CASH_data.drop(["DIES_NOMES_GRAN_DEUTE_PCB", "DIES_TOT_DEUTE_PCB"],1)

POS_CASH_balance_3 = """ SELECT train.SK_ID_CURR, CASE WHEN
NUM_QUOTES_PENDENTS IS NULL THEN 0 ELSE NUM_QUOTES_PENDENTS END AS
NUM_QUOTES_PENDENTS
                         FROM train LEFT JOIN (SELECT SK_ID_CURR AS
SK_ID_CURR2,
                                               SUM(CASE WHEN MONTHS_BALANCE =
-1 THEN CNT_INSTALMENT_FUTURE ELSE 0 END) AS NUM_QUOTES_PENDENTS
                                  FROM POS_CASH_balance_new
                                  GROUP BY SK_ID_CURR) AS UNIO ON
train.SK_ID_CURR = UNIO.SK_ID_CURR2;"""

POS_3 = sqldf(POS_CASH_balance_3, locals())
```

```
POS_CASH_data = pd.merge(POS_CASH_data, POS_3 , left_on = "SK_ID_CURR",
right_on = "SK_ID_CURR")

del(POS_1)
del(POS_2)
del(POS_3)
del(POS_CASH_balance_1)
del(POS_CASH_balance_2)
del(POS_CASH_balance_3)

# Prova de merge
POS_CASH_data.drop(POS_CASH_data.columns[6], axis=1, inplace=True) #Elimino un
SK_ID_CURR

##### POS_CASH, CREDIT_CARD I PREVIOUS_APPLICATION
len(set(POS_CASH_balance_new['SK_ID_PREV'].unique()) &
set(credit_card_balance['SK_ID_PREV'].unique()))

AMOUNT_ACTIVE_SQL = """ SELECT previous_application.SK_ID_CURR,
previous_application.SK_ID_PREV,
                        CASE WHEN Cont_Active == 1 THEN SUM(AMT_ANNUITY) ELSE
0 END AS AMT_ANNUITY_ACTIVE_HC,
                        CASE WHEN Cont_Active == 1 THEN SUM(AMT_CREDIT) ELSE 0
END AS AMT_CREDIT_ACTIVE_HC
                        FROM previous_application LEFT JOIN (SELECT
SK_ID_PREV, CASE WHEN MONTHS_BALANCE = MAX(MONTHS_BALANCE) AND
NAME_CONTRACT_STATUS = "Active" THEN 1 ELSE 0 END AS Cont_Active
                                                              FROM
POS_CASH_balance
                                                              GROUP BY
SK_ID_PREV) AS UNIO ON previous_application.SK_ID_PREV = UNIO.SK_ID_PREV
                        GROUP BY previous_application.SK_ID_CURR; """

AMT_1 = sqldf(AMOUNT_ACTIVE_SQL, locals())

AMOUNT_ACTIVE_SQL2 = """SELECT previous_application.SK_ID_CURR,
previous_application.SK_ID_PREV,
                        CASE WHEN Cont_Active == 1 THEN SUM(AMT_ANNUITY) ELSE
0 END AS AMT_ANNUITY_ACTIVE_HC_2
                        FROM previous_application LEFT JOIN (SELECT
SK_ID_PREV, CASE WHEN MONTHS_BALANCE = MAX(MONTHS_BALANCE) AND
NAME_CONTRACT_STATUS = "Active" THEN 1 ELSE 0 END AS Cont_Active
                                                              FROM
credit_card_balance
                                                              GROUP BY
SK_ID_PREV) AS UNIO ON previous_application.SK_ID_PREV = UNIO.SK_ID_PREV
                        GROUP BY previous_application.SK_ID_CURR; """

AMT_2 = sqldf(AMOUNT_ACTIVE_SQL2, locals())
AMT = pd.merge(AMT_1, AMT_2, left_on = "SK_ID_PREV", right_on = "SK_ID_PREV")
AMT['AMT_ANNUITY_HC'] = AMT.AMT_ANNUITY_ACTIVE_HC +
AMT.AMT_ANNUITY_ACTIVE_HC_2

AMOUNT_FINAL = """ SELECT train.SK_ID_CURR, CASE WHEN AMT_ANNUITY_HC IS NULL
THEN 0 ELSE AMT_ANNUITY_HC END AS AMT_ANNUITY_HC,
                    CASE WHEN AMT_CREDIT_ACTIVE_HC IS NULL THEN 0 ELSE
AMT_CREDIT_ACTIVE_HC END AS AMT_CREDIT_ACTIVE_HC
                    FROM train LEFT JOIN AMT ON
train.SK_ID_CURR=AMT.SK_ID_CURR_x
                    GROUP BY train.SK_ID_CURR; """

AMT_FINAL = sqldf(AMOUNT_FINAL, locals())
amt_final = AMT_FINAL.copy()

del(AMT_FINAL)
del(AMT_1)
del(AMT_2)
del(AMOUNT_ACTIVE_SQL)
```

```
del(AMOUNT_ACTIVE_SQL2)
del(AMT)
del(AMOUNT_FINAL)
del(credit_card_balance)
del(POS_CASH_balance)

##############################
# INSTALLMENTS_PAYMENTS:
installments_payments = pd.read_csv("installments_payments.csv", sep=",",
decimal=".")
installments_payments =
installments_payments.drop(["NUM_INSTALMENT_VERSION","NUM_INSTALMENT_NUMBER"],
1)

len(set(train['SK_ID_CURR'].unique()) &
set(installments_payments['SK_ID_CURR'].unique()))

installments_payments["Diferencia_Dies"] =
installments_payments.DAYS_ENTRY_PAYMENT -
installments_payments.DAYS_INSTALMENT
installments_payments["Diferencia_Amt"] = installments_payments.AMT_PAYMENT -
installments_payments.AMT_INSTALMENT

installments_payments.Diferencia_Amt.head(20)
installments_payments_SQL = """SELECT train.SK_ID_CURR,
           CASE WHEN Dif_Dies > 0 THEN 1 ELSE 0 END AS LATE_PAYMENTS_IP,
           CASE WHEN Dif_Amt < 0 THEN 1 ELSE 0 END AS NOT_ENOUGH_MONEY_PAID_IP
           FROM train LEFT JOIN (SELECT SK_ID_CURR AS SK_ID_CURR2,
SUM(Diferencia_Dies) AS Dif_Dies, SUM(Diferencia_Amt) AS Dif_Amt
                                 FROM installments_payments
                                 GROUP BY SK_ID_CURR) AS UNIO ON
UNIO.SK_ID_CURR2 = train.SK_ID_CURR;"""

install_payments = sqldf(installments_payments_SQL, locals())
install_payments.dtypes

del(installments_payments)

#################

# Ajuntar les bases de dades de la part de crèdits previs:
#Unir train amb previous
train = pd.merge(train, previous_application_SQL, left_on = "SK_ID_CURR",
right_on = "SK_ID_CURR")

# Unir l'anterior amb crèdit_card
train = pd.merge(train, credit_card_SQL, left_on = "SK_ID_CURR", right_on =
"SK_ID_CURR")

# Unir l'anterior amb POS_CASH_balance
train = pd.merge(train, POS_CASH_data, left_on = "SK_ID_CURR", right_on =
"SK_ID_CURR")
train = pd.merge(train, amt_final, left_on = "SK_ID_CURR", right_on =
"SK_ID_CURR")

# Unir l'anterior amb install_payments:
train = pd.merge(train, install_payments, left_on = "SK_ID_CURR", right_on =
"SK_ID_CURR")

# %%

##########
# Hem de retocar la bbdd de application_train (recodificar el que sigui
necessari) i després ajuntar-ho

# Llegim les variables d'interés de la base de dades train:
application_train = pd.read_csv("application_train.csv", sep=",", decimal=".",
usecols = ["SK_ID_CURR",
```

```
                    "TARGET","NAME_CONTRACT_TYPE","FLAG_OWN_CAR",
"FLAG_OWN_REALTY","AMT_INCOME_TOTAL","AMT_CREDIT","AMT_ANNUITY","AMT_GOODS_PRI
CE",

"NAME_TYPE_SUITE","NAME_INCOME_TYPE","NAME_EDUCATION_TYPE","NAME_FAMILY_STATUS
","NAME_HOUSING_TYPE",

"DAYS_BIRTH","DAYS_EMPLOYED","OCCUPATION_TYPE","ORGANIZATION_TYPE","REGION_POP
ULATION_RELATIVE","CODE_GENDER",

"CNT_CHILDREN","CNT_FAM_MEMBERS","REGION_RATING_CLIENT","OWN_CAR_AGE",
"EXT_SOURCE_1", "EXT_SOURCE_2","EXT_SOURCE_3"])

application_train.isnull().sum(axis = 0)
sum(application_train.DAYS_EMPLOYED==0)

##
valoracio_externa = application_train.iloc[:,24:27]
mitjana_valoracio_externa = valoracio_externa.mean(axis=1)
max_valoracio_externa = valoracio_externa.max(axis=1)
min_valoracio_externa = valoracio_externa.min(axis=1)

dat2 = pd.DataFrame({'MEAN_EXT_SOURCE':
mitjana_valoracio_externa,'MAX_EXT_SOURCE':
max_valoracio_externa,'MIN_EXT_SOURCE': min_valoracio_externa   })

application_train = application_train.join(dat2)
application_train = application_train.drop(["EXT_SOURCE_1",
"EXT_SOURCE_2","EXT_SOURCE_3"],1)

##
application_train[["SK_ID_CURR","CODE_GENDER"]].groupby(["CODE_GENDER"]).agg([
'count'])
application_train.loc[(application_train['CODE_GENDER'] == 'XNA'),
'CODE_GENDER'] = "F" #EL més comú
application_train[["SK_ID_CURR","CODE_GENDER"]].groupby(["CODE_GENDER"]).agg([
'count'])

##
application_train[["SK_ID_CURR","OWN_CAR_AGE"]].groupby(["OWN_CAR_AGE"]).agg([
'count'])

application_train.loc[(application_train['OWN_CAR_AGE'] == 0), 'OWN_CAR_AGE']
= 0.5 #EL més comú
application_train.loc[(application_train['OWN_CAR_AGE'] == 'NaN'),
'OWN_CAR_AGE'] = 0 #EL més comú
application_train.OWN_CAR_AGE = application_train.OWN_CAR_AGE.fillna(0)

application_train.OWN_CAR_AGE.head(10)

te_cotxe = application_train.FLAG_OWN_CAR
te_cotxe = te_cotxe.replace("N", 0)
te_cotxe = te_cotxe.replace("Y", 1)
te_cotxe.sum()

application_train.OWN_CAR_AGE = application_train.OWN_CAR_AGE * te_cotxe


##
application_train[["SK_ID_CURR","NAME_CONTRACT_TYPE"]].groupby(["NAME_CONTRACT
_TYPE"]).agg(['count'])
application_train[["SK_ID_CURR","FLAG_OWN_CAR"]].groupby(["FLAG_OWN_CAR"]).agg
(['count'])  # Té cotxe.
application_train[["SK_ID_CURR","FLAG_OWN_REALTY"]].groupby(["FLAG_OWN_REALTY"
]).agg(['count']) # Té casa / pis, etc.

##
```

```
application_train[["SK_ID_CURR","NAME_TYPE_SUITE"]].groupby(["NAME_TYPE_SUITE"
]).agg(['count'])

application_train.loc[(application_train['NAME_TYPE_SUITE'] == 'Other_A')|
           (application_train['NAME_TYPE_SUITE'] == 'Other_B'),
'NAME_TYPE_SUITE'] = "Other"

application_train[["SK_ID_CURR","NAME_TYPE_SUITE"]].groupby(["NAME_TYPE_SUITE"
]).agg(['count']) #Aquesta s'haurà d'eliminar 110% segur.

##
application_train[["SK_ID_CURR","NAME_INCOME_TYPE"]].groupby(["NAME_INCOME_TYP
E"]).agg(['count'])
application_train[["SK_ID_CURR","NAME_EDUCATION_TYPE"]].groupby(["NAME_EDUCATI
ON_TYPE"]).agg(['count'])

application_train[["SK_ID_CURR","NAME_FAMILY_STATUS"]].groupby(["NAME_FAMILY_S
TATUS"]).agg(['count'])
application_train.loc[(application_train['NAME_FAMILY_STATUS'] == 'Unknown') ,
'NAME_FAMILY_STATUS'] = "Married"
application_train[["SK_ID_CURR","NAME_FAMILY_STATUS"]].groupby(["NAME_FAMILY_S
TATUS"]).agg(['count'])


application_train[["SK_ID_CURR","NAME_HOUSING_TYPE"]].groupby(["NAME_HOUSING_T
YPE"]).agg(['count'])

#OCCUPATION_TYPE més important que la organization_type?
application_train[["SK_ID_CURR","OCCUPATION_TYPE"]].groupby(["OCCUPATION_TYPE"
]).agg(['count'])

application_train.loc[(application_train['NAME_INCOME_TYPE'] == 'Pensioner'),
'OCCUPATION_TYPE'] = "Retired"
application_train.loc[(application_train['OCCUPATION_TYPE'] == 'XNA'),
'OCCUPATION_TYPE'] = "Other"

application_train[["SK_ID_CURR","OCCUPATION_TYPE"]].groupby(["OCCUPATION_TYPE"
]).agg(['count'])


application_train[["SK_ID_CURR","ORGANIZATION_TYPE"]].groupby(["ORGANIZATION_T
YPE"]).agg(['count'])

application_train.loc[(application_train['NAME_INCOME_TYPE'] == 'Pensioner'),
'ORGANIZATION_TYPE'] = "Retired"

application_train.loc[(application_train['ORGANIZATION_TYPE'] == 'Business
Entity Type 1')|
           (application_train['ORGANIZATION_TYPE'] == 'Business Entity Type
2')|
           (application_train['ORGANIZATION_TYPE'] == 'Business Entity Type
3'), 'ORGANIZATION_TYPE'] = "Business"

application_train.loc[(application_train['ORGANIZATION_TYPE'] == 'Industry:
type 1')|
           (application_train['ORGANIZATION_TYPE'] == 'Industry: type 2')|
           (application_train['ORGANIZATION_TYPE'] == 'Industry: type 3')|
           (application_train['ORGANIZATION_TYPE'] == 'Industry: type 4')|
           (application_train['ORGANIZATION_TYPE'] == 'Industry: type 5')|
           (application_train['ORGANIZATION_TYPE'] == 'Industry: type 6')|
           (application_train['ORGANIZATION_TYPE'] == 'Industry: type 7')|
           (application_train['ORGANIZATION_TYPE'] == 'Industry: type 8')|
           (application_train['ORGANIZATION_TYPE'] == 'Industry: type 9')|
           (application_train['ORGANIZATION_TYPE'] == 'Industry: type 10')|
           (application_train['ORGANIZATION_TYPE'] == 'Industry: type 11')|
           (application_train['ORGANIZATION_TYPE'] == 'Industry: type 12')|
           (application_train['ORGANIZATION_TYPE'] == 'Industry: type 13'),
'ORGANIZATION_TYPE'] = "Industry"
```

```
application_train.loc[(application_train['ORGANIZATION_TYPE'] == 'Trade: type
1')|
           (application_train['ORGANIZATION_TYPE'] == 'Trade: type 2')|
           (application_train['ORGANIZATION_TYPE'] == 'Trade: type 3')|
           (application_train['ORGANIZATION_TYPE'] == 'Trade: type 4')|
           (application_train['ORGANIZATION_TYPE'] == 'Trade: type 5')|
           (application_train['ORGANIZATION_TYPE'] == 'Trade: type 6')|
           (application_train['ORGANIZATION_TYPE'] == 'Trade: type 7'),
'ORGANIZATION_TYPE'] = "Trade"

application_train.loc[(application_train['ORGANIZATION_TYPE'] == 'Transport:
type 1')|
           (application_train['ORGANIZATION_TYPE'] == 'Transport: type 2')|
           (application_train['ORGANIZATION_TYPE'] == 'Transport: type 3')|
           (application_train['ORGANIZATION_TYPE'] == 'Transport: type 4'),
'ORGANIZATION_TYPE'] = "Transport"

application_train[["SK_ID_CURR","ORGANIZATION_TYPE"]].groupby(["ORGANIZATION_T
YPE"]).agg(['count'])


# Edat en el moment de l'extracció
application_train["DAYS_BIRTH"] =
pd.to_numeric(application_train["DAYS_BIRTH"])
application_train["DAYS_BIRTH"] = round((abs(application_train.DAYS_BIRTH) /
365),0)

application_train["DAYS_EMPLOYED"] =
pd.to_numeric(application_train["DAYS_EMPLOYED"])
application_train["DAYS_EMPLOYED"] =
round((abs(application_train.DAYS_EMPLOYED) / 365),0)
application_train.isnull().sum(axis = 0)

import numpy as np
bins = [-0.1, 4, 10, 81, np.inf]
names = ['0-3', '4-9', '10-80', 'Retired']
application_train['DAYS_EMPLOYED'] =
pd.cut(application_train['DAYS_EMPLOYED'], bins, labels=names)

# Rename columns:
application_train = application_train.rename(columns={'DAYS_BIRTH':
'AGE_EXPECTED', 'DAYS_EMPLOYED': 'LAST_YEARS_EMPLOYED', 'OWN_CAR_AGE':
'ANY_COTXE_INTERACTION'})
application_train.dtypes

application_train['TARGET'] = application_train['TARGET'].astype(object)

#%%
##############################################################################
# Juntar bureau information:
train = pd.merge(information_Altres_EF, train , left_on = "SK_ID_CURR",
right_on = "SK_ID_CURR")

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', None)
pd.set_option('display.max_colwidth', None)

train.dtypes

# Juntar les bases de dades de train amb l'anterior:

train2 = pd.merge(application_train, train , left_on = "SK_ID_CURR", right_on
= "SK_ID_CURR")

train2.dtypes
```

```
# Modificacions i creacions de variables: (Ajuntar informació de les altres
bbdd)
#Anualitats actives
train2['AMT_ANNUITY_TOT'] = train2.AMT_ANNUITY + train2.AMT_ANNUITY_HC +
train2.AMOUNT_ANNUITY_Active_EF
train2['PERCENTATGE_ANNUITY'] = round((train2.AMT_ANNUITY_TOT /
train2.AMT_INCOME_TOTAL) * 100,5) # Amt_Annuity (Active) / Income_Total

#Crèdits actius quantitat
train2['AMT_CREDIT_ACT_TOT'] = train2.AMT_CREDIT +
train2.AMOUNT_CURRENT_CREDIT_Active_EF + train2.AMT_CREDIT_ACTIVE_HC
train2['RATI_DEUTE_GARANTIA'] = train2.AMT_CREDIT / train2.AMT_GOODS_PRICE

# Nombre de crèdits actius totals
train2['NUM_ACTIVE_CREDITS'] = train2.TOT_CO_ACT_PCB + train2.TE_CONTR_ACT_CC
+ train2.CRE_ACT_Altres_EF # Crèdits actius totals
train2['NUM_ACTIVE_CREDITS_HC'] = train2.TOT_CO_ACT_PCB +
train2.TE_CONTR_ACT_CC # Crèdits actius al propi banc
train2['NUM_CREDITS_PREVIS_TANCATS'] =  train2.CRE_CLO_Altres_EF +
train2.TOT_CO_COM_PCB_MENSUAL # Crèdits previs completats.

#Comprovació
train2.AMT_CREDIT_ACT_TOT.head(10)
train2.AMT_INCOME_TOTAL.head(10)

train2.PERCENTATGE_ANNUITY.head(40)
train2 =
train2.drop(["AMT_ANNUITY","AMT_ANNUITY_HC","AMOUNT_ANNUITY_Active_EF","AMOUNT
_CURRENT_CREDIT_Active_EF","AMT_CREDIT_ACTIVE_HC"],1)
train2 = train2.drop(["TOT_CO_ACT_PCB","TE_CONTR_ACT_CC", "CRE_CLO_Altres_EF",
"TOT_CO_COM_PCB_MENSUAL", "SK_ID_CURR"],1)
train2 = train2.drop(["DAYS_BIG_DUE_ONLY_PCB","DAYS_ALL_DUE_PCB",
"NUM_ACTIVE_CREDITS_HC", "CRE_ACT_Altres_EF"],1)

# Breus modificacions:
train2.DIES_TOT_DEUTE_PCB.isna().sum()
train2.DIES_TOT_DEUTE_PCB = train2.DIES_TOT_DEUTE_PCB.fillna(0)
train2.DIES_NOMES_GRAN_DEUTE_PCB = train2.DIES_NOMES_GRAN_DEUTE_PCB.fillna(0)
train2.MAX_CREDIT_LIM_CC = train2.MAX_CREDIT_LIM_CC.fillna(0)
train2.LAST_DUE_MONTH = train2.LAST_DUE_MONTH.fillna('No_DPD_PA')

train2.isnull().sum(axis = 0)


# Fitxer amb el que es guarda la base de dades final:
train2.to_excel('F:/Tot TFG/TFG/Data/AllData.xlsx', sheet_name='Dades')

# Elaboració de la bbdd de train i de test:

from sklearn.model_selection import train_test_split
target = train2.TARGET
train2 = train2.drop(["TARGET"],1)
train2.dtypes

# Fem la divisió:
X_train, X_test, y_train, y_test = train_test_split(train2, target,
test_size=0.2, random_state=2021)

# All_Dummies
X_train2 = pd.get_dummies(X_train)
X_train2 = X_train2.drop(['ORGANIZATION_TYPE_Retired',
'OCCUPATION_TYPE_Retired','LAST_YEARS_EMPLOYED_Retired'],1)
X_train2['TARGET'] = y_train
X_train2.to_excel('F:/Tot TFG/TFG/Data/TrainData_AllDummiesNew.xlsx')

X_test2 = pd.get_dummies(X_test)
X_test2 = X_test2.drop(['ORGANIZATION_TYPE_Retired',
'OCCUPATION_TYPE_Retired','LAST_YEARS_EMPLOYED_Retired'],1)
```

```
X_test2['TARGET'] = y_test
X_test2.to_excel('F:/Tot TFG/TFG/Data/TestData_AllDummiesNew.xlsx')
```

8.5.2 Anàlisi descriptiu

```
import os
import sys
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import describe
from scipy.stats import chi2_contingency
from scipy.stats import mannwhitneyu
import random
from scipy.stats import ttest_ind
import seaborn as sns
import matplotlib.pyplot as plt
from time import time

pd.options.display.float_format = '{:.2f}'.format

## Dades categòriques:

# Descriptiva i gràfics per agrupacions d'una variable:
def descriptiva_agrup (df, agrupacio):
    # Mirar com és la variable: categòrica: TRUE, FALSE -> sys.exit()
    tipus = tipus_var(df,agrupacio)

    if tipus: #si és una variable categòrica, llavors que faci un diagrama de
barres.
        results = pd.DataFrame()
        array_sum = df.groupby(agrupacio)['TARGET'].sum()
        array_n = df.groupby(agrupacio)['TARGET'].count()
        array_perc = (array_sum.values / array_n.values)*100
        index = list(array_sum.index)

        results['Nivells'] = index
        results['Nombre de registres'] = array_n.values
        results["Nombre d'impagaments"] = array_sum.values
        results["% d'impagaments"] = array_perc
        return(results)
    else: # Si no és categòrica, que surti de la funció
        print('La variable introduïda no és categòrica!')
        sys.exit()

def barplot_desc_agrup (df, agrupacio):
    if tipus_var(df,agrupacio) == True:
        pd_intervals = pd.DataFrame(columns=['Interval', 'Color'])

        if any(df[agrupacio].unique() == 'XNA'):
            df = df[df[agrupacio] != 'XNA']

        descriptiva = descriptiva_agrup(df, agrupacio)
        nom_columna = remove_unwanted(agrupacio)

        col = []
        for val in descriptiva["% d'impagaments"]:
            if val < descriptiva["% d'impagaments"].quantile(0.2):
                col.append('forestgreen')
            elif (val >= descriptiva["% d'impagaments"].quantile(0.2)) & (val
< descriptiva["% d'impagaments"].quantile(0.4)):
                col.append('yellowgreen')
            elif (val >= descriptiva["% d'impagaments"].quantile(0.4)) & (val
< descriptiva["% d'impagaments"].quantile(0.6)):
                col.append('yellow')
            elif (val >= descriptiva["% d'impagaments"].quantile(0.6)) & (val
< descriptiva["% d'impagaments"].quantile(0.8)):
```

```
                col.append('tomato')
            else:
                col.append('r')

        values = [min(descriptiva["% d'impagaments"]),descriptiva["%
d'impagaments"].quantile(0.2),descriptiva["%
d'impagaments"].quantile(0.4),descriptiva["%
d'impagaments"].quantile(0.6),descriptiva["%
d'impagaments"].quantile(0.8),descriptiva["% d'impagaments"].quantile(1)]
        int_0 = "[" + str(round(values[0],2)) + " , " +
str(round(values[1],2)) + ')'
        int_1 = "[" + str(round(values[1],2)) + " , " +
str(round(values[2],2)) + ')'
        int_2 = "[" + str(round(values[2],2)) + " , " +
str(round(values[3],2)) + ')'
        int_3 = "[" + str(round(values[3],2)) + " , " +
str(round(values[4],2)) + ')'
        int_4 = "[" + str(round(values[4],2)) + " , " +
str(round(values[5],2)) + ')'

        listOfSeries = [pd.Series([int_0, "forestgreen"],
index=pd_intervals.columns ) ,
                        pd.Series([int_1, "yellowgreen"],
index=pd_intervals.columns ) ,
                        pd.Series([int_2, "yellow"],
index=pd_intervals.columns ) ,
                        pd.Series([int_3, "tomato"],
index=pd_intervals.columns ) ,
                        pd.Series([int_4, "r"], index=pd_intervals.columns)]
        pd_intervals = pd_intervals.append(listOfSeries, ignore_index=True)

        sns.set_theme(style="whitegrid")
        tips = sns.load_dataset("tips")

        ax = sns.barplot(x="Nivells", y="% d'impagaments", data=descriptiva,
palette=col)
        if len(descriptiva.Nivells.unique()) <= 2:
            plt.xticks(rotation=0,fontsize=10)
        elif len(descriptiva.Nivells.unique()) <= 10 and
len(descriptiva.Nivells.unique()) >= 3:
            plt.xticks(rotation=62,fontsize=10)
        else:
            plt.xticks(rotation=90,fontsize=9)

        plt.title("Percentatge d'impagaments segons els nivells de" + ' ' +
nom_columna)

        return ax, pd_intervals

    else:
        print("La variable introduïda no és categòrica!")
        sys.exit()


## Dades numèriques:

# Correlacions I: Gràfiques
def correlacions_imp (df, num):
    # Amb el num agafem les variables d'interés. (perquè no carrega tots els
noms quan num = 0)
    # num = 0: mapa de calor sencer.
    # num = 1: 20:20
    # num = 2: 20:(21:40)
    # num = 3: 41:46
    # num = 4: (21:40):20
    # num = 5: (21:40):(21:40)
    # num = 6: (21:40):(41:46)
    # num = 7: (41:46):20
```

```
    # num = 8: (41:46):(21:40)
    # num = 9: (41:46):(41:46)

    dades_num = df.select_dtypes(include=[np.number])
    corr = dades_num.corr()

    if num == 1:
        corr = corr.iloc[0:20,0:20]
    if num == 2:
        corr = corr.iloc[0:20,21:40]
    if num == 3:
        corr = corr.iloc[0:20,41:46]
    if num == 4:
        corr = corr.iloc[21:40,0:20]
    if num == 5:
        corr = corr.iloc[21:40,21:40]
    if num == 6:
        corr = corr.iloc[21:40,41:46]
    if num == 7:
        corr = corr.iloc[41:46,0:20]
    if num == 8:
        corr = corr.iloc[41:46,21:40]
    if num == 9:
        corr = corr.iloc[41:46,41:46]

    ax = sns.heatmap(corr, vmin=-1, vmax=1,
center=0,cmap=sns.diverging_palette(20, 220,
n=200),square=False,xticklabels=True, yticklabels=True)

ax.set_xticklabels(ax.get_xticklabels(),rotation=90,horizontalalignment='right
', Fontsize = 7)
    ax.set_yticklabels(ax.get_ymajorticklabels(), fontsize = 7);
    #ax.set_title('Mapa de calor de les correlacions')

    return ax

## Gràfics:

def boxplot_TARGET (df, variable):

    df2 = df[df.TARGET == 0]
    df3 = df[df.TARGET == 1]
    DF = pd.DataFrame({'No Impagament': df2[variable], 'Impagament':
df3[variable]})

    ax = DF[['No Impagament', 'Impagament']].plot(kind='box', title='boxplot',
showmeans=True)

    return plt.show()

def barres_TARGET (df, variable):
    tipus = tipus_var(df,variable)
    nom_columna = remove_unwanted(variable)

    if tipus: #si és una variable categòrica, llavors que em faci un diagrama
de barres.
        array = df[['TARGET',variable]].groupby(variable).agg(['sum'])
        index = list(array.index)
        values = array.values[:,0]

        plt.bar(index,values, color='green')
        plt.xlabel("Diferents nivells de la variable" + ' ' + nom_columna)
        plt.ylabel("Nombre d'impagaments")
        plt.title("Impagaments segons" + ' ' + nom_columna)
        if len(df[variable].unique()) > 2:
            plt.xticks(rotation = 75,fontsize=8)
        else:
            plt.xticks(fontsize=9)
```

```
        return(plt.show())

    else: #llavors la variable és numèrica -> histograma
        if df[variable].max() > 800000:
            a = 2000
        elif (df[variable].max() > 40) & (df[variable].max() < 800000):
            a = 40
        else:
            a = 10

        df2 = df[df.TARGET == 0]
        sns.distplot(df2[[variable]], kde=False, label='No Impagament', color
= 'seagreen', bins = a)

        df2 = df[df.TARGET == 1]
        sns.distplot(df2[[variable]], kde=False,label='Impagaments', color =
'salmon', bins = a)
        sns.color_palette("Paired")
        # Plot formatting
        plt.legend(prop={'size': 12})
        plt.title("Impagaments segons" + ' ' + nom_columna)
        plt.xlabel(nom_columna)
        plt.ylabel('Nombre de registres')

        if df[variable].max() > 800000:
            plt.xlim(left = 0,right = df2[variable].quantile(0.9975))

        return(plt.show())


def histograma_relatiu_TARGET (df, variable):
    tipus = tipus_var(df,variable)
    nom_columna = remove_unwanted(variable)

    if tipus: #si és una variable categòrica, llavors que em faci un diagrama
de barres.
        print('La variable introduida no és numèrica!')
        sys.exit()

    else: #llavors la variable és numèrica
        if df[variable].max() > 800000:
            a = 2000
        elif (df[variable].max() > 40) & (df[variable].max() < 800000):
            a = 40
        else:
            a = 10

        df2 = df[df.TARGET == 0]
        sns.distplot(df2[[variable]], kde=True, label='No Impagament', color =
'green', bins = a)

        df3 = df[df.TARGET == 1]
        sns.distplot(df3[[variable]], kde=True, label='Impagament', color =
'salmon', bins = a)

        # Plot formatting
        plt.legend(prop={'size': 12})
        plt.title("Distribució de" + ' ' + nom_columna + ' ' + 'segons els
impagaments')
        plt.xlabel(nom_columna)
        plt.ylabel('Densitat')


        return(plt.show())

def relatiu_TARGET (df, variable):
    tipus = tipus_var(df,variable)
```

```
    nom_columna = remove_unwanted(variable)

    if tipus: #si és una variable categòrica, llavors que em faci un diagrama
de barres.
        print('La variable introduida no és numèrica!')
        sys.exit()

    else: #llavors la variable és numèrica

        df2 = df[df.TARGET == 0]
        sns.distplot(df2[[variable]], kde=True, label='No Impagament', kde_kws
= {'shade': True, 'linewidth': 3}, color = 'green', hist = False)


        df3 = df[df.TARGET == 1]
        sns.distplot(df3[[variable]], kde=True, label='Impagament', kde_kws =
{'shade': True, 'linewidth': 3}, color = 'salmon', hist = False)

        # Plot formatting
        plt.legend(prop={'size': 12})
        plt.title("Impagaments segons" + ' ' + nom_columna)
        plt.xlabel(nom_columna)
        plt.ylabel('Proporció de registres')

        plt.xlim(left = min(df[variable])-5, right =
df[variable].quantile(0.9999999))
        return(plt.show())


## Altres:

def descriptiva_general_num (df): #Ha de tenir el target!
    #Dividir entre categòriques i numèriques:
    dades_num = df.select_dtypes(include=[np.number])
    dades_num['TARGET'] = df.TARGET
    resultats_sign = pd.DataFrame(columns=['Variable', 'Variable 2', 'P-
valor'])
    if dades_num.shape[1] == 0:
        print('No hi ha cap variable numèrica en la base de dades!')
        sys.exit()

    else:
        dades_num1 = dades_num[dades_num['TARGET'] == 1]
        dades_num0 = dades_num[dades_num['TARGET'] == 0]


        stats = dades_num.describe()
        stats.loc['missings'] = dades_num.shape[0] - dades_num.count()
        stats.loc['var'] = dades_num.var()
        stats.loc['skew'] = dades_num.skew()
        stats.loc['kurt'] = dades_num.kurtosis()

        stats1 = dades_num1.describe()
        stats1.loc['missings'] = dades_num1.shape[0] - dades_num1.count()
        stats1.loc['var'] = dades_num1.var()
        stats1.loc['skew'] = dades_num1.skew()
        stats1.loc['kurt'] = dades_num1.kurtosis()

        stats0 = dades_num0.describe()
        stats0.loc['missings'] = dades_num0.shape[0] - dades_num0.count()
        stats0.loc['var'] = dades_num0.var()
        stats0.loc['skew'] = dades_num0.skew()
        stats0.loc['kurt'] = dades_num0.kurtosis()

        # stats: resum numèric de totes les observacions.
        # stats0: resum numèric dels no impagaments.
        # stats1: resum numèric dels impagaments.
    return stats, stats0, stats1
```

```
# Mitjana d'impagaments per interval!
def descriptiva_num_intervals (df, variable, num_talls):
    df["TARGET"] = pd.to_numeric(df["TARGET"])
    tipus = tipus_var(df,variable)
    nom_columna = remove_unwanted(variable)


    if tipus: #si és una variable categòrica, llavors que em faci un diagrama
de barres.
        print('La variable introduida no és numèrica!')
        sys.exit()

    else:
        dades = df[['TARGET', variable]]
        dades[[variable]] = pd.qcut(dades[variable], q = num_talls,
duplicates='drop')
        return dades.groupby([variable]).mean()*100

    # Agrupar pels intervals, i que tregui el número d'impagaments, número
d'obs, i % d'impagaments (mitjana).


def tipus_var (df, variable):
    boolea = (df[[variable]].dtypes == 'object').bool()
    if boolea: #És categòrica
        tipus = True
    else: #És numèrica
        tipus = False
    return(tipus)


def remove_unwanted (name):
# initializing bad_chars_list
    bad_chars = [';', ':', '!', '*', '_', '.', '-', '/', '\ ']
# using replace() to
    for i in bad_chars :
       name = name.replace(i, ' ')

    name = str(name)
    return name
```

```
import os
os.chdir("F:/TOT TFG/TFG")
from Scripts.Descriptiva.Descriptiva_functions import *
dades = pd.read_excel("Data/AllData.xlsx", index_col=0)

# Convertim les variables que són binàries en categòriques:
for column in dades.select_dtypes(include=np.number):
    if len(dades[column].unique()) == 2 and column!= "CRE_BDE_Altres_EF":
            dades[column] = dades[column].astype(object)

# Determinar Missings
NAS = dades.isnull().sum()
PERC = NAS / 307511 * 100
dataset = pd.DataFrame({'Noms': dades.columns,
                  "Nombre de NA's": NAS,
                  "Percentatge NA's": PERC})

#### CATEGÒRIQUES ####

# Anàlisi segons impagaments
descriptiva_agrup(dades,'NAME_TYPE_SUITE')
descriptiva_agrup(dades,'CODE_GENDER')
descriptiva_agrup(dades,'NAME_INCOME_TYPE') # IMPORTANT!
descriptiva_agrup(dades,'NAME_EDUCATION_TYPE') # IMPORTANT!
descriptiva_agrup(dades,'NAME_FAMILY_STATUS') # IMPORTANT!
```

```
descriptiva_agrup(dades,'FLAG_OWN_CAR') # IMPORTANT!
descriptiva_agrup(dades,'NAME_HOUSING_TYPE') # IMPORTANT!
descriptiva_agrup(dades,'LAST_YEARS_EMPLOYED') # IMPORTANT!
descriptiva_agrup(dades,'OCCUPATION_TYPE') # IMPORTANT
descriptiva_agrup(dades,'ORGANIZATION_TYPE') # IMPORTANT
descriptiva_agrup(dades,'NAME_CONTRACT_TYPE') # IMPORTANT
descriptiva_agrup(dades,'FLAG_OWN_REALTY')
descriptiva_agrup(dades,'LATE_PAYMENTS_IP') # IMPORTANT
descriptiva_agrup(dades,'NOT_ENOUGH_MONEY_PAID_IP') # IMPORTANT
descriptiva_agrup(dades,'TE_CREDIT_CARD_CC')
descriptiva_agrup(dades,'LAST_DUE_MONTH')
descriptiva_agrup(dades,'SUM_Diferencia_AMT_HC')
descriptiva_agrup(dades,'IS_Refreshed_Client_HC') # IMPORTANT
descriptiva_agrup(dades,'IS_NEW_CLIENT_TRAIN') # IMPORTANT
descriptiva_agrup(dades,'Credits_Assegurats_HC')
descriptiva_agrup(dades,'DIES_MAX_IMPAGATS_Altres_EF') # IMPORTANT
descriptiva_agrup(dades,'DID_PROLONG_Altres_EF')
descriptiva_agrup(dades,'DID_OVERDUE_Altres_EF')
descriptiva_agrup(dades,'TE_CREDIT_Altres_EF') # IMPORTANT
descriptiva_agrup(dades,'CAR_LOAN_Altres_EF')
descriptiva_agrup(dades,'MORTGAGE_LOAN_Altres_EF')
descriptiva_agrup(dades,'CONSUMER_CREDIT_Altres_EF')


## Gràfic de barres ## S'haurien d'eliminar aquelles categories que tinguin
menys de 3 observacions (per posar algo)
barplot_desc_agrup(dades, 'NAME_TYPE_SUITE') # 'important'
barplot_desc_agrup(dades, 'CODE_GENDER') # 'important'
barplot_desc_agrup(dades, 'NAME_INCOME_TYPE') # Important
barplot_desc_agrup(dades, 'NAME_EDUCATION_TYPE') # Important
barplot_desc_agrup(dades, 'NAME_FAMILY_STATUS') # S'hauria d'eliminar el
Unknown perquè esbiaixa el gràfic.
barplot_desc_agrup(dades, 'FLAG_OWN_CAR')
barplot_desc_agrup(dades, 'NAME_HOUSING_TYPE') # Important
barplot_desc_agrup(dades, 'LAST_YEARS_EMPLOYED') # Important
barplot_desc_agrup(dades, 'OCCUPATION_TYPE') # Important
barplot_desc_agrup(dades, 'ORGANIZATION_TYPE') # Important
barplot_desc_agrup(dades, 'NAME_CONTRACT_TYPE') # Important
barplot_desc_agrup(dades, 'FLAG_OWN_REALTY')
barplot_desc_agrup(dades, 'LATE_PAYMENTS_IP') # Important
barplot_desc_agrup(dades, 'NOT_ENOUGH_MONEY_PAID_IP') # Important
barplot_desc_agrup(dades, 'TE_CREDIT_CARD_CC')
barplot_desc_agrup(dades, 'LAST_DUE_MONTH')
barplot_desc_agrup(dades, 'SUM_Diferencia_AMT_HC')
barplot_desc_agrup(dades, 'IS_Refreshed_Client_HC')
barplot_desc_agrup(dades, 'IS_NEW_CLIENT_TRAIN') # Important
barplot_desc_agrup(dades, 'Credits_Assegurats_HC')
barplot_desc_agrup(dades, 'DIES_MAX_IMPAGATS_Altres_EF')
barplot_desc_agrup(dades, 'TE_CREDIT_Altres_EF')
barplot_desc_agrup(dades, 'DID_PROLONG_Altres_EF')
barplot_desc_agrup(dades, 'DID_OVERDUE_Altres_EF')
barplot_desc_agrup(dades, 'MORTGAGE_LOAN_Altres_EF')
barplot_desc_agrup(dades, 'CONSUMER_CREDIT_Altres_EF')

#### CATEGÒRIQUES ####

# Anàlisi segons impagaments
descriptiva_agrup(dades,'NAME_TYPE_SUITE')
descriptiva_agrup(dades,'CODE_GENDER')
descriptiva_agrup(dades,'NAME_INCOME_TYPE') # IMPORTANT!
descriptiva_agrup(dades,'NAME_EDUCATION_TYPE') # IMPORTANT!
descriptiva_agrup(dades,'NAME_FAMILY_STATUS') # IMPORTANT!
descriptiva_agrup(dades,'FLAG_OWN_CAR') # IMPORTANT!
descriptiva_agrup(dades,'NAME_HOUSING_TYPE') # IMPORTANT!
descriptiva_agrup(dades,'LAST_YEARS_EMPLOYED') # IMPORTANT!
descriptiva_agrup(dades,'OCCUPATION_TYPE') # IMPORTANT
descriptiva_agrup(dades,'ORGANIZATION_TYPE') # IMPORTANT
descriptiva_agrup(dades,'NAME_CONTRACT_TYPE') # IMPORTANT
```

```
descriptiva_agrup(dades,'FLAG_OWN_REALTY')
descriptiva_agrup(dades,'LATE_PAYMENTS_IP') # IMPORTANT
descriptiva_agrup(dades,'NOT_ENOUGH_MONEY_PAID_IP') # IMPORTANT
descriptiva_agrup(dades,'TE_CREDIT_CARD_CC')
descriptiva_agrup(dades,'LAST_DUE_MONTH')
descriptiva_agrup(dades,'SUM_Diferencia_AMT_HC')
descriptiva_agrup(dades,'IS_Refreshed_Client_HC') # IMPORTANT
descriptiva_agrup(dades,'IS_NEW_CLIENT_TRAIN') # IMPORTANT
descriptiva_agrup(dades,'Credits_Assegurats_HC')
descriptiva_agrup(dades,'DIES_MAX_IMPAGATS_Altres_EF') # IMPORTANT
descriptiva_agrup(dades,'DID_PROLONG_Altres_EF')
descriptiva_agrup(dades,'DID_OVERDUE_Altres_EF')
descriptiva_agrup(dades,'TE_CREDIT_Altres_EF') # IMPORTANT
descriptiva_agrup(dades,'CAR_LOAN_Altres_EF')
descriptiva_agrup(dades,'MORTGAGE_LOAN_Altres_EF')
descriptiva_agrup(dades,'CONSUMER_CREDIT_Altres_EF')


## Gràfic de barres ## S'haurien d'eliminar aquelles categories que tinguin
menys de 3 observacions (per posar algo)
barplot_desc_agrup(dades, 'NAME_TYPE_SUITE') # 'important'
barplot_desc_agrup(dades, 'CODE_GENDER') # 'important'
barplot_desc_agrup(dades, 'NAME_INCOME_TYPE') # Important
barplot_desc_agrup(dades, 'NAME_EDUCATION_TYPE') # Important
barplot_desc_agrup(dades, 'NAME_FAMILY_STATUS') # S'hauria d'eliminar el
Unknown perquè esbiaixa el gràfic.
barplot_desc_agrup(dades, 'FLAG_OWN_CAR')
barplot_desc_agrup(dades, 'NAME_HOUSING_TYPE') # Important
barplot_desc_agrup(dades, 'LAST_YEARS_EMPLOYED') # Important
barplot_desc_agrup(dades, 'OCCUPATION_TYPE') # Important
barplot_desc_agrup(dades, 'ORGANIZATION_TYPE') # Important
barplot_desc_agrup(dades, 'NAME_CONTRACT_TYPE') # Important
barplot_desc_agrup(dades, 'FLAG_OWN_REALTY')
barplot_desc_agrup(dades, 'LATE_PAYMENTS_IP') # Important
barplot_desc_agrup(dades, 'NOT_ENOUGH_MONEY_PAID_IP') # Important
barplot_desc_agrup(dades, 'TE_CREDIT_CARD_CC')
barplot_desc_agrup(dades, 'LAST_DUE_MONTH')
barplot_desc_agrup(dades, 'SUM_Diferencia_AMT_HC')
barplot_desc_agrup(dades, 'IS_Refreshed_Client_HC')
barplot_desc_agrup(dades, 'IS_NEW_CLIENT_TRAIN') # Important
barplot_desc_agrup(dades, 'Credits_Assegurats_HC')
barplot_desc_agrup(dades, 'DIES_MAX_IMPAGATS_Altres_EF')
barplot_desc_agrup(dades, 'TE_CREDIT_Altres_EF')
barplot_desc_agrup(dades, 'DID_PROLONG_Altres_EF')
barplot_desc_agrup(dades, 'DID_OVERDUE_Altres_EF')
barplot_desc_agrup(dades, 'MORTGAGE_LOAN_Altres_EF')
barplot_desc_agrup(dades, 'CONSUMER_CREDIT_Altres_EF')



# dataframes amb les relacions del test de chi-quadrat
sign, no_sign = chi_independencia(dades)

sign_mespetites = sign[sign['P-valor']<(0.01/(10**300))]
sign_mespetites = sign[sign['P-valor']==0]
sign_mespetites = sign_mespetites.dropna(axis=0)

barres_TARGET(dades, 'NAME_EDUCATION_TYPE')
dades.TARGET

# Boxplot:
boxplot_TARGET(dades, 'AGE_EXPECTED')

#########################
## Variables numèriques:

total, total0, total1 = descriptiva_general_num(dades)
```

```
# Descriptiva per intervals
descriptiva_num_intervals(dades, "AGE_EXPECTED",5)
descriptiva_num_intervals(dades, "AMT_ANNUITY_TOT",30)
descriptiva_num_intervals(dades, "PERCENTATGE_ANNUITY",30)
descriptiva_num_intervals(dades, "AMT_CREDIT_ACT_TOT",18)
descriptiva_num_intervals(dades, "RATI_DEUTE_GARANTIA",10)
descriptiva_num_intervals(dades, "NUM_ACTIVE_CREDITS",10)
descriptiva_num_intervals(dades, "AMT_INCOME_TOTAL",6)
descriptiva_num_intervals(dades, "MEAN_EXT_SOURCE",10)
descriptiva_num_intervals(dades, "REGION_POPULATION_RELATIVE",10)
descriptiva_num_intervals(dades, "ANY_COTXE_INTERACTION",20)

dades2 = dades[dades.ANY_COTXE_INTERACTION > 0]
descriptiva_num_intervals(dades2, "ANY_COTXE_INTERACTION",8)


barres_TARGET(dades, 'AMT_ANNUITY_TOT')
barres_TARGET(dades, 'PERCENTATGE_ANNUITY')
barres_TARGET(dades, 'AMT_CREDIT_ACT_TOT')
barres_TARGET(dades, 'RATI_DEUTE_GARANTIA')
barres_TARGET(dades, 'NUM_ACTIVE_CREDITS')
barres_TARGET(dades, 'NUM_QUOTES_PENDENTS')
barres_TARGET(dades, 'DIES_NOMES_GRAN_DEUTE_PCB')
barres_TARGET(dades, 'NUM_Repeater_Client_HC')
barres_TARGET(dades, 'AGE_EXPECTED')
barres_TARGET(dades, 'AMT_INCOME_TOTAL')
barres_TARGET(dades, 'AMT_CREDIT')
barres_TARGET(dades, 'AMT_GOODS_PRICE')
barres_TARGET(dades, 'Refused_STATUS_HC')
barres_TARGET(dades, 'NC_Goods_PURP_HC')
barres_TARGET(dades, 'Repairs_PURP_HC')
barres_TARGET(dades, 'UrgentN_PURP_HC')
barres_TARGET(dades, 'GET_HOME_PURP_HC')
barres_TARGET(dades, 'CARS_PURP_HC')
barres_TARGET(dades, 'FUN_PURP_HC')
barres_TARGET(dades, 'REGION_POPULATION_RELATIVE')
barres_TARGET(dades, 'ANY_COTXE_INTERACTION')

histograma_relatiu_TARGET(dades, 'AMT_ANNUITY_TOT')
histograma_relatiu_TARGET(dades, 'PERCENTATGE_ANNUITY')
histograma_relatiu_TARGET(dades, 'AMT_CREDIT_ACT_TOT')
histograma_relatiu_TARGET(dades, 'RATI_DEUTE_GARANTIA')
histograma_relatiu_TARGET(dades, 'NUM_ACTIVE_CREDITS')
histograma_relatiu_TARGET(dades, 'NUM_QUOTES_PENDENTS')
histograma_relatiu_TARGET(dades, 'DIES_NOMES_GRAN_DEUTE_PCB')
histograma_relatiu_TARGET(dades, 'NUM_Repeater_Client_HC')
histograma_relatiu_TARGET(dades, 'AGE_EXPECTED')
histograma_relatiu_TARGET(dades, 'AMT_INCOME_TOTAL')
histograma_relatiu_TARGET(dades, 'AMT_CREDIT')
histograma_relatiu_TARGET(dades, 'AMT_GOODS_PRICE')
histograma_relatiu_TARGET(dades, 'Refused_STATUS_HC')
histograma_relatiu_TARGET(dades, 'NC_Goods_PURP_HC')
histograma_relatiu_TARGET(dades, 'Repairs_PURP_HC')
histograma_relatiu_TARGET(dades, 'UrgentN_PURP_HC')
histograma_relatiu_TARGET(dades, 'GET_HOME_PURP_HC')
histograma_relatiu_TARGET(dades, 'CARS_PURP_HC')
histograma_relatiu_TARGET(dades, 'FUN_PURP_HC')
histograma_relatiu_TARGET(dades, 'MEAN_EXT_SOURCE')
histograma_relatiu_TARGET(dades, 'ElectD_CAT_HC')
histograma_relatiu_TARGET(dades, 'Home_CAT_HC')
histograma_relatiu_TARGET(dades, 'CRE_CLO_Altres_EF')
histograma_relatiu_TARGET(dades, 'CRE_SOL_Altres_EF')
histograma_relatiu_TARGET(dades, 'TARGET')

relatiu_TARGET(dades, 'AMT_CREDIT')

# Correlacions entre totes les variables:
correlacions_imp(dades,9) #0,1,2,...,9.
```

```
# Gràfic impagaments.
import matplotlib.pyplot as plt

# Creem un dataframe per fer el gràfic
dataset = pd.DataFrame({'Taula':['Dades senceres', 'Train', 'Test'],
                   "Núm. d'impagaments": [24825,19804,5021],
                   "% d'impagaments": [8.073,8.062,8.177]})

# creating axes object and defining plot
ax = dataset.plot(kind = 'bar', x = 'Taula',
                  y = "Núm. d'impagaments", color = 'lightsalmon',
                  linewidth = 3)

ax2 = dataset.plot(kind = 'line', x = 'Taula',
                   y = "% d'impagaments", secondary_y = True,
                   color = 'khaki',  linewidth = 3,
                   ax = ax)
plt.title("Estat dels impagaments en cada taula")

# Etiquetant els eixos
ax.set_xlabel('Taula', color = 'black')
ax.set_ylabel("Nombre d'impagaments", color = "black")
ax2.set_ylabel("% d'impagaments", color = 'black')
ax.legend(bbox_to_anchor=(0.77, 1),fancybox=True,prop={'size': 10}, frameon =
False)
ax2.legend(bbox_to_anchor=(0.717, 0.935), fancybox=True, frameon = False)
plt.tight_layout()

# Es mostra el gràfic
plt.show()


# Altres gràfics:

# Creem un dataframe per fer el gràfic
dataset = pd.DataFrame({'Intervals':['1','2','3','4','5','6','7','8'],
                   "% d'impagaments":
[6.29,5.18,5.71,6.65,7.77,8.49,9.11,9.31]})

# Construim els eixos i el gràfic
ax = dataset.plot(kind = 'line', x = 'Intervals',
                  y = "% d'impagaments", color = 'lightsalmon',
                  linewidth = 3)
  plt.title("Percentage d'impagaments segons el nombre d'anys del vehicle")

# Etiquetem els eixos
ax.set_xlabel('Intervals', color = 'black')
ax.set_ylabel("% d'impagaments", color = 'black')
#defining display layout
plt.tight_layout()

# Es mostra el gràfic
plt.show()
```

8.5.3 Construcció dels models clàssics amb totes les variables

```
setwd("G:/Tot TFG/TFG/Data")
options(scipen=2)

library(readxl)
library(tidyverse)
library(caret)
library(car)
```

```
# Base de dades per defecte del Python

train_x <- read_xlsx("TrainData_AllDummiesNew.xlsx")
test_x <- read_xlsx("TestData_AllDummiesNew.xlsx")

# Eliminar variables que apareixen:
train_x[["...1"]] <- NULL
train_x$ORGANIZATION_TYPE_XNA <- NULL
test_x[["...1"]] <- NULL
test_x$ORGANIZATION_TYPE_XNA <- NULL

if (!any(names(train_x) == 'TARGET')){
  train_x$TARGET <- train_y$TARGET
}

# Anàlisi dels NA's:
apply(train_x, MARGIN = 2, function(x) sum(is.na(x))) #Comptar el número de
NA's
NAS <- apply(train_x, MARGIN = 2, function(x) sum(is.na(x))) #Comptar el
número de NA's

# Hem vist que el model presenta VIF's elevats:
suma <- apply(train_x, MARGIN = 2, sum,na.rm = TRUE)
data_suma <- data.frame( Missings = NAS)
data_suma$Percentatge <- data_suma$Missings/nrow(train_x)*100
#data_suma$Nom_Columna <- names(train_x2)
data_suma$Num_columna <- 1:nrow(data_suma)

# Model
model1 <-  glm(TARGET ~ ., data = train_x, family = binomial)
summary(model1)

# No hem eliminat cap variable, així que s'estan creant variables linealment
dependents.
vif(model1) # Error

## Intentem corregir el model anterior!
train_x2 <- train_x
train_x2new <- train_x2

# Eliminar variables que apareixen:
train_x2[["...1"]] <- NULL
train_x2$ORGANIZATION_TYPE_XNA <- NULL

#if (!any(names(train_x) == 'TARGET')){
#  train_x$TARGET <- train_y$TARGET
#}

train_x2new <- train_x2

# MODEL 2 amb millores:

# Si base de dades antiga:
#train_x2 <- train_x2[,-
c(64,66,68,75,77,88,89,95,100,103,124,156,161,164,167)]
# Si base de dades nova:
#100: Name_Family_Status_Married (el que he eliminat)
train_x2new <- train_x2new[,-
c(72,74,76,78,85,87,98,100,105,110,113,134,166,171,174,177)]
train_x2new <- na.omit(train_x2new)

model_totes_var <- (glm(TARGET ~ . - CNT_FAM_MEMBERS +
DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC, data =
train_x2new, family = binomial))
summary(model_totes_var)

vif(model_totes_var)[which(vif(model_totes_var)>5)]
```

```
plot(model_totes_var)
plot(model_totes_var, which=4)

# MODEL 3: Amb menys multicol·linealitat:
# Solucionar VIF
model_totes_var_menysVIF <- (glm(TARGET ~ . - NUM_Repeater_Client_HC -
MAX_CREDIT_LIM_CC - CNT_FAM_MEMBERS - MAX_EXT_SOURCE - MIN_EXT_SOURCE -
SUM_BALANCE_CC - AMT_GOODS_PRICE - AMT_ANNUITY_TOT +
DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC , data
= train_x2new, family = binomial))
summary(model_totes_var_menysVIF)
vif(model_totes_var_menysVIF)[which(vif(model_totes_var_menysVIF)>5)]
plot(model_totes_var_menysVIF, which=4)


# Treiem l'outlier: observació 134133.
# MODEL 4: El model bo
train_x3 <- train_x2new[-134133,]
train_x3 <- na.omit(train_x3)
model_bo3 <-  glm(TARGET ~ . - NUM_Repeater_Client_HC - MAX_CREDIT_LIM_CC -
CNT_FAM_MEMBERS - MAX_EXT_SOURCE -
              MIN_EXT_SOURCE - SUM_BALANCE_CC - AMT_GOODS_PRICE -
AMT_ANNUITY_TOT +
              DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y +
                TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC , data = train_x3, family
= binomial)
summary(model_bo3)
position_vif5 <- which(vif(model_bo3)>5)
vif(model_bo3)[position_vif5]
plot(model_bo3, which = 4)

# Obtenir les probabilitats de TRAIN
prediccions <- as.data.frame(model_bo3$fitted.values)
prediccions$RESPOSTA <- train_x3$TARGET
colnames(prediccions)[1] <- "PROBABILITATS"

library(openxlsx)
write.xlsx(prediccions, "G:/Tot
TFG/TFG/Scripts/03_Models/GLM/Probabilitats_Logit_TOT_TRAIN.xlsx")


# Obtenir les probabilitats de TEST
test_x2 <- test_x[,-
c(72,74,76,78,85,87,98,100,105,110,113,134,166,171,174,177)]

NAS <- apply(test_x2, MARGIN = 2, function(x) sum(is.na(x))) #Comptar el
número de NA's
suma <- apply(test_x2, MARGIN = 2, sum,na.rm = TRUE)
data_suma <- data.frame( Missings = NAS)
data_suma$Percentatge <- data_suma$Missings/nrow(test_x2)*100
#data_suma$Nom_Columna <- names(train_x2)
data_suma$Num_columna <- 1:nrow(data_suma)

test_x2 <- na.omit(test_x2)

prediccions_test <- as.data.frame(predict(model_bo3, newdata = test_x2, type =
"response"))
prediccions_test$RESPOSTA <- test_x2$TARGET
colnames(prediccions_test)[1] <- "PROBABILITATS"

write.xlsx(prediccions_test, "G:/Tot
TFG/TFG/Scripts/03_Models/GLM/Probabilitats_Logit_TOT_TEST.xlsx")

train_x3 <- train_x2new[-134133,]
train_x3 <- na.omit(train_x3)
y <- train_x3$TARGET
```

```
train_x3$TARGET <- NULL

matriu_model2 <- model.matrix( ~ .-1 - NUM_Repeater_Client_HC -
MAX_CREDIT_LIM_CC - CNT_FAM_MEMBERS - MAX_EXT_SOURCE - MIN_EXT_SOURCE -
SUM_BALANCE_CC - AMT_GOODS_PRICE - AMT_ANNUITY_TOT +
DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC,
train_x3)

# MODEL LASSO
cv.lasso2 <- cv.glmnet(matriu_model2, y, alpha = 1, family = "binomial",
standardize = TRUE)
plot(cv.lasso2)
lasso.model2 <- glmnet(matriu_model2, y, alpha = 1, family = "binomial",
lambda = cv.lasso2$lambda)
coef(lasso.model2)

set.seed(2021)
files <- sample(1:nrow(train_x3), size = 50000, replace = FALSE)

millor_lambda <- function(model, lambdes, y_train, files, matriu_model){
  gini <- c()
  auc <- c()
  lambda <- c()
  for (i in lambdes){
    prediccions <- as.data.frame(predict(model, matriu_model, s = i,
type="response"))
    colnames(prediccions)[1] <- "PROBABILITAT"
    prediccions$RESPOSTA <- y_train
    prediccions <- prediccions[files,]
    valoracio <- ModelPerformance(prediccions$RESPOSTA,
prediccions$PROBABILITAT)

    gini <- c(gini, valoracio$Gini)
    print(gini)
    auc <- c(auc, valoracio$AUC)
    print(auc)
    lambda <- c(lambda, i)
  }
  df <- data.frame("Lambda" = lambda,"GINI" = gini, "AUC" = auc)
  return(df)
}

ModelPerformance <- function (actuals, predictedScores){
  fitted <- data.frame (Actuals=actuals, PredictedScores=predictedScores) #
actuals and fitted
  colnames(fitted) <- c('Actuals','PredictedScores') # rename columns

  ones <- fitted[fitted$Actuals==1, ] # Subset ones
  zeros <- fitted[fitted$Actuals==0, ] # Subsetzeros

  totalPairs <- nrow(ones) * nrow(zeros) # calculate total number of pairs to
check

  # A pair is concordant if 1 (event) has a higher predicted probability than
0
  conc <- sum (c(vapply(ones$PredictedScores, function(x) {((x >
zeros$PredictedScores))}, FUN.VALUE=logical(nrow(zeros)))), na.rm=T)

  # A pair is disconcordant if 1 (event) has a lower predicted probability
than 0
  disc <- sum(c(vapply(ones$PredictedScores, function(x) {((x <
zeros$PredictedScores))}, FUN.VALUE = logical(nrow(zeros)))), na.rm = T)

  # Calculate concordance, discordance, ties and AUC
  concordance <- conc/totalPairs
  discordance <- disc/totalPairs
  tiesPercent <- (1-concordance-discordance)
```

```
  Gini = (conc-disc)/totalPairs
  AUC  = concordance + 0.5*tiesPercent
  return(list("Concordance"=round(concordance,6),
"Discordance"=round(discordance,6),
             "Tied"= round(tiesPercent,6), "Gini"= round(Gini,6),"AUC"=
round(AUC,6)))
}

df <- millor_lambda(lasso.model2, cv.lasso2$lambda, y, files, matriu_model2)

which(df$GINI == max(df$GINI))
which(df$AUC == max(df$AUC))
coef(lasso.model2)[,70]

# La millor lambda és:

# Probabilitats per a TRAIN
dataframe <- as.data.frame(matriu_model2)
prediccions <- as.data.frame(predict(lasso.model2, matriu_model2, s =
cv.lasso2$lambda[70], type="response"))
colnames(prediccions)[1] <- "PROBABILITATS"
prediccions$RESPOSTA <- y

head(prediccions)
tapply(prediccions$PROBABILITAT, prediccions$RESPOSTA, mean)

write.xlsx(prediccions, "G:/Tot TFG/TFG/Scripts/03_Models/LASSO i
RIDGE/Probabilitats_LASSO_TOT_Train.xlsx")


# Probabilitats per a TEST
test_x2 <- test_x[,-
c(72,74,76,78,85,87,98,100,105,110,113,134,166,171,174,177)]
test_x2 <- na.omit(test_x2)

nums <- unlist(lapply(test_x2, is.numeric))
results_test <- test_x2[,nums]
results_test <- na.omit(results_test)
y_test <- results_test$TARGET
results_test$TARGET <- NULL

matriu_model_test <- model.matrix( ~ .-1- NUM_Repeater_Client_HC -
MAX_CREDIT_LIM_CC - CNT_FAM_MEMBERS - MAX_EXT_SOURCE - MIN_EXT_SOURCE -
SUM_BALANCE_CC - AMT_GOODS_PRICE - AMT_ANNUITY_TOT +
DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC
,results_test)

prediccions_test <- as.data.frame(predict(lasso.model2, s =
cv.lasso2$lambda[70], matriu_model_test, type="response"))
colnames(prediccions_test)[1] <- "PROBABILITATS"
prediccions_test$RESPOSTA <- y_test
head(prediccions_test)
tapply(prediccions_test$PROBABILITAT, prediccions_test$RESPOSTA, mean)

write.xlsx(prediccions_test, "G:/Tot TFG/TFG/Scripts/03_Models/LASSO i
RIDGE/Probabilitats_LASSO_TOT_Test.xlsx")

rm(lasso.model2, dataframe, df, model_bo, model1, prediccions,
prediccions_test)

##########
# MODEL RIDGE

cv.ridge2 <- cv.glmnet(matriu_model2, y, alpha = 0, family = "binomial",
standardize = TRUE)
plot(cv.ridge2)
```

```
ridge.model2 <- glmnet(matriu_model2, y, alpha = 0, family = "binomial",
lambda = cv.ridge2$lambda)
coef(ridge.model2)

df <- millor_lambda(ridge.model2, cv.ridge2$lambda, y, files, matriu_model2)
which(df$GINI == max(df$GINI))
which(df$AUC == max(df$AUC))
coef(ridge.model2)[,100]

# Probabilitats per a TRAIN
prediccions <- as.data.frame(predict(ridge.model2, matriu_model2, s =
cv.ridge2$lambda[100], type="response"))
colnames(prediccions)[1] <- "PROBABILITATS"
prediccions$RESPOSTA <- y

head(prediccions)
tapply(prediccions$PROBABILITAT, prediccions$RESPOSTA, mean)

write.xlsx(prediccions, "G:/Tot TFG/TFG/Scripts/03_Models/LASSO i
RIDGE/Probabilitats_RIDGE_TOT_Train.xlsx")


# Probabilitats per a TEST
test_x2 <- test_x[,-
c(72,74,76,78,85,87,98,100,105,110,113,134,166,171,174,177)]
test_x2 <- na.omit(test_x2)

nums <- unlist(lapply(test_x2, is.numeric))
results_test <- test_x2[,nums]
results_test <- na.omit(results_test)
y_test <- results_test$TARGET
results_test$TARGET <- NULL

matriu_model_test <- model.matrix( ~ .-1- NUM_Repeater_Client_HC -
MAX_CREDIT_LIM_CC - CNT_FAM_MEMBERS - MAX_EXT_SOURCE - MIN_EXT_SOURCE -
SUM_BALANCE_CC - AMT_GOODS_PRICE - AMT_ANNUITY_TOT +
DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC
,results_test)

prediccions_test <- as.data.frame(predict(ridge.model2, s =
cv.ridge2$lambda[100], matriu_model_test, type="response"))
colnames(prediccions_test)[1] <- "PROBABILITATS"
prediccions_test$RESPOSTA <- y_test
head(prediccions_test)
tapply(prediccions_test$PROBABILITAT, prediccions_test$RESPOSTA, mean)

write.xlsx(prediccions_test, "G:/Tot TFG/TFG/Scripts/03_Models/LASSO i
RIDGE/Probabilitats_RIDGE_TOT_Test.xlsx")
```

8.5.4 Construcció dels models clàssics amb menys variables

```
setwd("G:/Tot TFG/TFG/Data")
options(scipen=2)

library(readxl)
library(tidyverse)
library(caret)
library(car)
library(MASS)
library(openxlsx)

# Model sense outliers i amb variables significatives

train_x2 <- read_xlsx("TrainData_AllDummiesNew.xlsx")
test_x2 <- read_xlsx("TestData_AllDummiesNew.xlsx")
```

```
# Eliminar variables que apareixen:
train_x2[["...1"]] <- NULL
train_x2$ORGANIZATION_TYPE_XNA <- NULL
test_x2[["...1"]] <- NULL
test_x2$ORGANIZATION_TYPE_XNA <- NULL

train_x2new <- train_x2
test_x2new <- test_x2

# An?lisi dels NA's:
apply(train_x2, MARGIN = 2, function(x) sum(is.na(x))) #Comptar el número de
NA's

# Hem vist que el model presenta VIF's elevats:
suma <- apply(train_x2, MARGIN = 2, sum)
data_suma <- data.frame(Suma = suma)
data_suma$Nom_Columna <- names(train_x2)
data_suma$Num_columna <- 1:nrow(data_suma)

#103: Name_Family_Status_Widow
#100: Name_Family_Status_Married
train_x2new <- train_x2new[,-
c(72,74,76,78,85,87,98,100,105,110,113,134,166,171,174,177)]
train_x2new <- na.omit(train_x2new)

test_x2new <- test_x2new[,-
c(72,74,76,78,85,87,98,100,105,110,113,134,166,171,174,177)]
test_x2new <- na.omit(test_x2new)


##### MODEL LOGISTIC #####
train_x3 <- train_x2new[-134133,]
train_x3 <- na.omit(train_x3)
model_bo3 <-  glm(TARGET ~ . - NUM_Repeater_Client_HC - MAX_CREDIT_LIM_CC -
CNT_FAM_MEMBERS - MAX_EXT_SOURCE -
                    MIN_EXT_SOURCE - SUM_BALANCE_CC - AMT_GOODS_PRICE -
AMT_ANNUITY_TOT +
                    DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y +
                    TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC , data = train_x3,
family = binomial)
summary(model_bo3)

names(train_x3)

train_x5 <- train_x3[,-
c(5,14,17,19,21,24,27,28,33,34,35,37,39,43,45,46,47,48,56,58,60,61,76:80,158,1
59)] # Treiem les variables no significatives del model.
test_x5 <- test_x2new[,-
c(5,14,17,19,21,24,27,28,33,34,35,37,39,43,45,46,47,48,56,58,60,61,76:80,158,1
59)] #Treiem les variables no significatives del model.

model_bo2 <-  glm(TARGET ~ .  -  FLAG_OWN_REALTY_Y - NUM_Repeater_Client_HC -
MAX_CREDIT_LIM_CC - CNT_FAM_MEMBERS - MAX_EXT_SOURCE - MIN_EXT_SOURCE -
SUM_BALANCE_CC - AMT_GOODS_PRICE - AMT_ANNUITY_TOT +
DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC , data
= train_x5, family = binomial)
summary(model_bo2)

prediccions <- as.data.frame(model_bo2$fitted.values)
prediccions$RESPOSTA <- train_x5$TARGET
colnames(prediccions)[1] <- "PROBABILITATS"

write.xlsx(prediccions, "G:/Tot TFG/TFG/Scripts/03_Models/GLM, LASSO, RIDGE
amb variables no significatives/Probabilitats/Probabilitats_GLM_Train.xlsx")
```

```
# Predicci? GLM test:
y_test <- test_x5$TARGET
test_x5$TARGET <- NULL

prediccions_test <- as.data.frame(predict(model_bo2, newdata = test_x5,
type="response"))
prediccions_test$RESPOSTA <- y_test
colnames(prediccions_test)[1] <- "PROBABILITATS"

write.xlsx(prediccions_test, "G:/Tot TFG/TFG/Scripts/03_Models/GLM, LASSO,
RIDGE amb variables no
significatives/Probabilitats/Probabilitats_GLM_Test.xlsx")

##### MODEL LASSO #####
library(glmnet)

nums <- unlist(lapply(train_x5, is.numeric))
results <- train_x5[,nums]
results <- na.omit(results)

y <- results$TARGET
results$TARGET <- NULL

matriu_model <- model.matrix( ~ .-1 -  FLAG_OWN_REALTY_Y -
NUM_Repeater_Client_HC - MAX_CREDIT_LIM_CC - CNT_FAM_MEMBERS - MAX_EXT_SOURCE
- MIN_EXT_SOURCE - SUM_BALANCE_CC - AMT_GOODS_PRICE - AMT_ANNUITY_TOT +
DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC,
results)

# Model LASSO
cv.lasso <- cv.glmnet(matriu_model, y, alpha = 1, family = "binomial",
standardize = TRUE)
plot(cv.lasso)
lasso.model <- glmnet(matriu_model, y, alpha = 1, family = "binomial", lambda
= cv.lasso$lambda)

prediccions <- as.data.frame(predict(lasso.model, matriu_model, s =
cv.lasso$lambda, type="response"))
colnames(prediccions)[1] <- "PROBABILITAT"
prediccions$RESPOSTA <- y

# QUINA ES LA MILLOR LAMBDA
# Trobar la millor lambda acord amb el màxim GINI i AUC d'una mostra de train.
set.seed(2021)
files <- sample(1:nrow(train_x3), size = 65000, replace = FALSE)
millor_lambda <- function(model, lambdes, y_train, files, matriu_model){
  gini <- c()
  auc <- c()
  lambda <- c()
  for (i in lambdes){
    prediccions <- as.data.frame(predict(model, matriu_model, s = i,
type="response"))
    colnames(prediccions)[1] <- "PROBABILITAT"
    prediccions$RESPOSTA <- y_train
    prediccions <- prediccions[files,]
    valoracio <- ModelPerformance(prediccions$RESPOSTA,
prediccions$PROBABILITAT)

    gini <- c(gini, valoracio$Gini)
    print(gini)
    auc <- c(auc, valoracio$AUC)
    print(auc)
    lambda <- c(lambda, i)
  }
  df <- data.frame("Lambda" = lambda,"GINI" = gini, "AUC" = auc)
  return(df)
```

```
}

ModelPerformance <- function (actuals, predictedScores){
  fitted <- data.frame (Actuals=actuals, PredictedScores=predictedScores) #
actuals and fitted
  colnames(fitted) <- c('Actuals','PredictedScores') # rename columns

  ones <- fitted[fitted$Actuals==1, ] # Subset ones
  zeros <- fitted[fitted$Actuals==0, ] # Subsetzeros

  totalPairs <- nrow(ones) * nrow(zeros) # calculate total number of pairs to
check

  # A pair is concordant if 1 (event) has a higher predicted probability than
0
  conc <- sum (c(vapply(ones$PredictedScores, function(x) {((x >
zeros$PredictedScores))}, FUN.VALUE=logical(nrow(zeros)))), na.rm=T)

  # A pair is disconcordant if 1 (event) has a lower predicted probability
than 0
  disc <- sum(c(vapply(ones$PredictedScores, function(x) {((x <
zeros$PredictedScores))}, FUN.VALUE = logical(nrow(zeros)))), na.rm = T)

  # Calculate concordance, discordance, ties and AUC
  concordance <- conc/totalPairs
  discordance <- disc/totalPairs
  tiesPercent <- (1-concordance-discordance)
  Gini = (conc-disc)/totalPairs
  AUC  = concordance + 0.5*tiesPercent
  return(list("Concordance"=round(concordance,6),
"Discordance"=round(discordance,6),
              "Tied"= round(tiesPercent,6), "Gini"= round(Gini,6),"AUC"=
round(AUC,6)))
}

df <- millor_lambda(lasso.model, cv.lasso$lambda, y, files, matriu_model )
which(df$GINI == max(df$GINI))
which(df$AUC == max(df$AUC))
df$Lambda[69]

#Coeficients del model
coef(lasso.model)[,69]
mitjana <- comparar_lambdes(prediccions)
which(mitjana$IMP == max(mitjana$IMP))

# PREDICCIÓ LASSO TRAIN
prediccions <- as.data.frame(predict(lasso.model, matriu_model, s =
cv.lasso$lambda[69], type="response"))
colnames(prediccions)[1] <- "PROBABILITATS"
prediccions$RESPOSTA <- y

head(prediccions)
tapply(prediccions$PROBABILITAT, prediccions$RESPOSTA, mean)

write.xlsx(prediccions, "G:/Tot TFG/TFG/Scripts/03_Models/GLM, LASSO, RIDGE
amb variables no significatives/Probabilitats/Probabilitats_LASSO_Train.xlsx")

# PREDICCIÓ LASSO TEST
test_x5 <- test_x2new[,-
c(5,14,17,19,21,24,27,28,33,34,35,37,39,43,45,46,47,48,56,58,60,61,76:80,158,1
59)] #Treiem les variables no significatives del model.
nums <- unlist(lapply(test_x5, is.numeric))
results_test <- test_x5[,nums]
results_test <- na.omit(results_test)

y_test <- results_test$TARGET
results_test$TARGET <- NULL
```

```
# matriu_model_test <- model.matrix( ~ .-1 - CARS_PURP_HC -
DID_PROLONG_Altres_EF -  FLAG_OWN_REALTY_Y - NUM_Repeater_Client_HCC -
MAX_CREDIT_LIM_CC - CNT_FAM_MEMBERS - MAX_EXT_SOURCE - MIN_EXT_SOURCE -
SUM_BALANCE_CC - AMT_GOODS_PRICE - AMT_ANNUITY_TOT +
DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC,
results_test)
matriu_model_test <- model.matrix( ~ .-1 -  FLAG_OWN_REALTY_Y -
NUM_Repeater_Client_HC - MAX_CREDIT_LIM_CC - CNT_FAM_MEMBERS - MAX_EXT_SOURCE
- MIN_EXT_SOURCE - SUM_BALANCE_CC - AMT_GOODS_PRICE - AMT_ANNUITY_TOT +
DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC,
results_test)

prediccions_test <- as.data.frame(predict(lasso.model, s =
cv.lasso$lambda[69], matriu_model_test, type="response"))
colnames(prediccions_test)[1] <- "PROBABILITATS"
prediccions_test$RESPOSTA <- y_test
head(prediccions_test)
tapply(prediccions_test$PROBABILITAT, prediccions_test$RESPOSTA, mean)

write.xlsx(prediccions_test, "G:/Tot TFG/TFG/Scripts/03_Models/GLM, LASSO,
RIDGE amb variables no
significatives/Probabilitats/Probabilitats_LASSO_Test.xlsx")

##### MODEL RIDGE #####

cv.ridge <- cv.glmnet(matriu_model, y, alpha = 0, family = "binomial",
standardize = TRUE)
plot(cv.ridge)

ridge.model <- glmnet(matriu_model, y, alpha = 0, family = "binomial",
                      lambda = cv.ridge$lambda)
coef(ridge.model)

df_ridge <- millor_lambda(ridge.model, cv.ridge$lambda, y, files, matriu_model
)
which(df_ridge$GINI == max(df_ridge$GINI))
which(df_ridge$AUC == max(df_ridge$AUC))
df_ridge$Lambda[100]
coef(ridge.model)[,100]

# PREDICCIÓ RIDGE TRAIN
prediccions <- as.data.frame(predict(ridge.model, matriu_model, s =
cv.ridge$lambda[100], type="response"))
colnames(prediccions)[1] <- "PROBABILITATS"
prediccions$RESPOSTA <- y
tapply(prediccions$PROBABILITAT, prediccions$RESPOSTA, mean)

write.xlsx(prediccions, "G:/Tot TFG/TFG/Scripts/03_Models/GLM, LASSO, RIDGE
amb variables no significatives/Probabilitats/Probabilitats_Ridge_Train.xlsx")

# PREDICCIÓ RIDGE TEST
test_x5 <- test_x2new[,-
c(5,14,17,19,21,24,27,28,33,34,35,37,39,43,45,46,47,48,56,58,60,61,76:80,158,1
59)] #Treiem les variables no significatives del model.
names(test_x5)
nums <- unlist(lapply(test_x5, is.numeric))
results_test <- test_x5[,nums]
results_test <- na.omit(results_test)

y_test <- results_test$TARGET
results_test$TARGET <- NULL
matriu_model_test <- model.matrix( ~ .-1 -  FLAG_OWN_REALTY_Y -
NUM_Repeater_Client_HC - MAX_CREDIT_LIM_CC - CNT_FAM_MEMBERS - MAX_EXT_SOURCE
- MIN_EXT_SOURCE - SUM_BALANCE_CC - AMT_GOODS_PRICE - AMT_ANNUITY_TOT +
DID_PROLONG_Altres_EF:DID_OVERDUE_Altres_EF +
```

```
FLAG_OWN_CAR_Y:FLAG_OWN_REALTY_Y + TE_CREDIT_CARD_CC:MAX_CREDIT_LIM_CC,
results_test)

prediccions_test <- as.data.frame(predict(ridge.model, s =
cv.ridge$lambda[100], matriu_model_test, type="response"))
colnames(prediccions_test)[1] <- "PROBABILITATS"
prediccions_test$RESPOSTA <- y_test
tapply(prediccions_test$PROBABILITAT, prediccions_test$RESPOSTA, mean)

write.xlsx(prediccions_test, "G:/Tot TFG/TFG/Scripts/03_Models/GLM, LASSO,
RIDGE amb variables no
significatives/Probabilitats/Probabilitats_Ridge_Test.xlsx")
```

8.5.5 Construcció dels models alternatius

```
import pandas as pd
from sklearn.tree import DecisionTreeClassifier # Import Decision Tree
Classifier
from sklearn import metrics #Import scikit-learn metrics module for accuracy
calculation
import os

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', None)
pd.set_option('display.max_colwidth', None)

os.getcwd() # Directori actual
os.chdir("G:\Tot TFG\TFG\Data")
os.chdir("F:\Tot TFG\TFG\Data")

#Càrrega TRAIN
train_x = pd.read_excel("TrainData_AllDummiesNew.xlsx")
train_x2 = train_x

train_x2 = train_x2.drop(["Unnamed: 0", "ORGANIZATION_TYPE_XNA"],1)
train_x2 = train_x2.dropna()
train_x2 = train_x2.reset_index(drop=True)
# train_x2 = train_x2.drop(134133)

y_train = train_x2.TARGET
X_train = train_x2.drop(["TARGET"],1)

# Càrrega TEST
test_x = pd.read_excel("TestData_AllDummiesNew.xlsx")
test_x2 = test_x

test_x2 = test_x2.drop(["Unnamed: 0", "ORGANIZATION_TYPE_XNA"],1)
test_x2 = test_x2.dropna()
test_x2 = test_x2.reset_index(drop=True)

y_test = test_x2.TARGET
X_test = test_x2.drop(["TARGET"],1)

###############################################################################

#Estandaritzar train i test
from sklearn.preprocessing import StandardScaler

sc = StandardScaler()

columnes = train_x2.columns

X_train = train_x2
target = X_train.TARGET
X_train = X_train.drop(["TARGET"], axis=1)
```

```
# No sé si la fila es correspon amb el TARGET...
X_train = sc.fit_transform(X_train)
df = pd.DataFrame(X_train, columns = columnes[0:183])
df['TARGET'] = target

df.TARGET.isna().sum()
X_train = df.dropna()
y_train = X_train.TARGET
X_train = X_train.drop(["TARGET"], axis=1)

#Test
target = test_x2.TARGET
test_x2 = test_x2.drop(["TARGET"], axis=1)

X_test = sc.transform(test_x2)

df = pd.DataFrame(X_test, columns = columnes[0:183])
df['TARGET'] = target

df.TARGET.isna().sum()
X_test = df.dropna()
y_test = X_test.TARGET
X_test = X_test.drop(["TARGET"], axis=1)

########################################################################

# Construim l'arbre de decisió:

clf = DecisionTreeClassifier(max_depth=30, random_state= 2021)

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = clf.predict(X_test)

y_pred_prob = clf.predict_proba(X_test)


prob_arbres = pd.DataFrame(data = y_pred_prob[:,1], columns=["PROBABILITATS"])
prob_arbres['RESPOSTA'] = y_test
print(round(prob_arbres.groupby('RESPOSTA')['PROBABILITATS'].mean(),4))
#prob_arbres.groupby('RESPOSTA')['PROBABILITATS'].max()
#prob_arbres.groupby('RESPOSTA')['PROBABILITATS'].min()
#prob_arbres.groupby('RESPOSTA')['PROBABILITATS'].median()

print("Nombre d'impagaments")
print(sum(y_pred))
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score, f1_score
print("Matriu de Confusió")
print(confusion_matrix(y_test,y_pred))
# print(classification_report(y_test,y_pred))
#cl = classification_report(y_test,y_pred2,output_dict=True)
#cl = pd.DataFrame(cl).transpose()
#cl.to_excel("CLASSREPORT300RF.xlsx")
print("Accuracy")
print(round(accuracy_score(y_test, y_pred),4))
print("F1-Score")
print(round(f1_score(y_test, y_pred),4))


prob_arbres.to_excel("G:/Tot TFG/TFG/Scripts/03_Models/Arbres de
decisió/Probabilitats/7DTREE.xlsx")

# Guardar TRAIN
y_pred_prob = clf.predict_proba(X_train)
```

```
y_pred = clf.predict(X_train)

prob_arbres = pd.DataFrame(data = y_pred_prob[:,1], columns=["PROBABILITATS"])
prob_arbres['RESPOSTA'] = y_train
print(round(prob_arbres.groupby('RESPOSTA')['PROBABILITATS'].mean(),4))
print("Accuracy")
print(round(accuracy_score(y_train, y_pred),4))
print("F1-Score")
print(round(f1_score(y_train, y_pred),4))
prob_arbres.to_excel("G:/Tot TFG/TFG/Scripts/03_Models/Arbres de
decisió/Probabilitats/30DTREE_TRAIN.xlsx")

import pickle
filename = 'C:/Users/Albert/Desktop/Models TFG/Arbres/6DTree.sav'
pickle.dump(model, open(filename, 'wb'))

#######################################################################

from sklearn.ensemble import RandomForestClassifier

regressor = RandomForestClassifier(n_estimators= 1000 ,
max_depth=12,random_state=2021) # Amb n_estimators = 20 funcionava bé
regressor.fit(X_train, y_train)

y_pred = regressor.predict(X_test)
y_pred_forest = regressor.predict_proba(X_test)


prob_arbres = pd.DataFrame(data = y_pred_forest[:,1],
columns=["PROBABILITATS"])
prob_arbres['RESPOSTA'] = y_test
print(round(prob_arbres.groupby('RESPOSTA')['PROBABILITATS'].mean(),4))

print("Nombre d'impagaments")
print(sum(y_pred))
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score, f1_score
print("Matriu de Confusió")
print(confusion_matrix(y_test,y_pred))
cl = classification_report(y_test,y_pred,output_dict=True)
cl = pd.DataFrame(cl).transpose()
# cl.to_excel("F:/Tot TFG/TFG/Scripts/03_Models/Random Forest/TAULA
GINI/CLASSREPORT10000RF.xlsx")
print("Accuracy")
print(round(accuracy_score(y_test, y_pred),4))
print("F1-Score")
print(round(f1_score(y_test, y_pred),4))

prob_arbres.to_excel("G:/Tot TFG/TFG/Scripts/03_Models/Random
Forest/Probabilitats_proves/1000_12_RANDOMFOREST_TEST.xlsx")


y_pred = regressor.predict(X_train)
y_pred_forest = regressor.predict_proba(X_train)

prob_arbres = pd.DataFrame(data = y_pred_forest[:,1],
columns=["PROBABILITATS"])
prob_arbres['RESPOSTA'] = y_train
print(round(prob_arbres.groupby('RESPOSTA')['PROBABILITATS'].mean(),4))

prob_arbres.to_excel("G:/Tot TFG/TFG/Scripts/03_Models/Random
Forest/Probabilitats_proves/1000_12_RANDOMFOREST_TRAIN.xlsx")

        #prob_arbres.groupby('RESPOSTA')['PROBABILITATS'].max()
        #prob_arbres.groupby('RESPOSTA')['PROBABILITATS'].min()
        #prob_arbres.groupby('RESPOSTA')['PROBABILITATS'].median()

print("Nombre d'impagaments")
```

```
print(sum(y_pred))
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score, f1_score
print("Matriu de Confusió")
print(confusion_matrix(y_test,y_pred))
cl = classification_report(y_test,y_pred,output_dict=True)
cl = pd.DataFrame(cl).transpose()
# cl.to_excel("F:/Tot TFG/TFG/Scripts/03_Models/Random Forest/TAULA
GINI/CLASSREPORT10000RF.xlsx")
print("Accuracy")
print(round(accuracy_score(y_test, y_pred),4))
print("F1-Score")
print(round(f1_score(y_test, y_pred),4))

import pickle
filename = 'G:/Tot TFG/TFG/Scripts/03_Models/Random Forest/RF_1000_12_BO.sav'
pickle.dump(regressor, open(filename, 'wb'))


########################################################################


import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt


pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', None)
pd.set_option('display.max_colwidth', None)

os.getcwd() # Directori actual
os.chdir("G:\Tot TFG\TFG\Data")
os.chdir("F:\Tot TFG\TFG\Data")


#Càrrega TRAIN
train_x = pd.read_excel("TrainData_AllDummiesNew.xlsx")
train_x2 = train_x

train_x2 = train_x2.drop(["Unnamed: 0", "ORGANIZATION_TYPE_XNA"],1)
train_x2 = train_x2.dropna()
train_x2 = train_x2.reset_index(drop=True)
# train_x2 = train_x2.drop(134133)

y_train = train_x2.TARGET

# Càrrega TEST
test_x = pd.read_excel("TestData_AllDummiesNew.xlsx")
test_x2 = test_x

test_x2 = test_x2.drop(["Unnamed: 0", "ORGANIZATION_TYPE_XNA"],1)
test_x2 = test_x2.dropna()
test_x2 = test_x2.reset_index(drop=True)

y_test = test_x2.TARGET

# Treiem variables no importants segons GLM
train_x2 = train_x2.drop(["Credits_Assegurats_HC_Mateixos crèdits assegurats",
"Credits_Assegurats_HC_Menys crèdits assegurats",
                          "Credits_Assegurats_HC_Més crèdits assegurats",
"NAME_TYPE_SUITE_Children", "NAME_TYPE_SUITE_Family",
                          "NAME_TYPE_SUITE_Group of people",
"NAME_TYPE_SUITE_Other", "NAME_TYPE_SUITE_Spouse, partner",
"NAME_TYPE_SUITE_Unaccompanied",
```

```
                           "CNT_FAM_MEMBERS", "MAX_EXT_SOURCE",
"MIN_EXT_SOURCE", "CRE_BDE_Altres_EF", "CRE_Another_Type_LOAN_Altres_EF",
"CAR_LOAN_Altres_EF",
                           "BUSINESS_LOAN_Altres_EF", "UNKNOWN_LOAN_Altres_EF",
"CASH_LOANS_TY_HC", "CONS_LOANS_TY_HC", "GET_HOME_PURP_HC", "CARS_PURP_HC",
                           "FUN_PURP_HC", "Health_CAT_HC", "Mobile_CAT_HC",
"Vehicles_CAT_HC", "SUM_BALANCE_CC"],1)
train_x2 = train_x2.reset_index(drop=True)


test_x2 = test_x2.drop(["Credits_Assegurats_HC_Mateixos crèdits assegurats",
"Credits_Assegurats_HC_Menys crèdits assegurats",
                           "Credits_Assegurats_HC_Més crèdits assegurats",
"NAME_TYPE_SUITE_Children", "NAME_TYPE_SUITE_Family",
                           "NAME_TYPE_SUITE_Group of people",
"NAME_TYPE_SUITE_Other", "NAME_TYPE_SUITE_Spouse, partner",
"NAME_TYPE_SUITE_Unaccompanied",
                           "CNT_FAM_MEMBERS", "MAX_EXT_SOURCE",
"MIN_EXT_SOURCE", "CRE_BDE_Altres_EF", "CRE_Another_Type_LOAN_Altres_EF",
"CAR_LOAN_Altres_EF",
                           "BUSINESS_LOAN_Altres_EF", "UNKNOWN_LOAN_Altres_EF",
"CASH_LOANS_TY_HC", "CONS_LOANS_TY_HC", "GET_HOME_PURP_HC", "CARS_PURP_HC",
                           "FUN_PURP_HC", "Health_CAT_HC", "Mobile_CAT_HC",
"Vehicles_CAT_HC", "SUM_BALANCE_CC"],1)
test_x2 = test_x2.reset_index(drop=True)

#############################################################################


#Estandaritzar train i test
from sklearn.preprocessing import StandardScaler

sc = StandardScaler()

columnes = train_x2.columns

X_train = train_x2
target = X_train.TARGET
X_train = X_train.drop(["TARGET"], axis=1)

# No sé si la fila es correspon amb el TARGET...
X_train = sc.fit_transform(X_train)
df = pd.DataFrame(X_train, columns = columnes[0:(len(columnes)-1)])
df['TARGET'] = target

df.TARGET.isna().sum()
X_train = df.dropna()
y_train = X_train.TARGET
X_train = X_train.drop(["TARGET"], axis=1)

#Test
target = test_x2.TARGET
test_x2 = test_x2.drop(["TARGET"], axis=1)

X_test = sc.transform(test_x2)

df = pd.DataFrame(X_test, columns = columnes[0:(len(columnes)-1)])
df['TARGET'] = target

df.TARGET.isna().sum()
X_test = df.dropna()
y_test = X_test.TARGET
X_test = X_test.drop(["TARGET"], axis=1)

#############################################################################

# SVM
```

```
X_train2 = X_train
y_train2 = y_train
# Lineal d'una altra forma

from sklearn.svm import LinearSVC, SVC
from sklearn.calibration import CalibratedClassifierCV


svc_classifier = SVC(kernel='linear', C=1).fit(X_train2, y_train2)

svm = LinearSVC(random_state=2021, class_weight= "balanced", dual = False,
max_iter = 1000)
clf = CalibratedClassifierCV(svm)
clf.fit(X_train2, y_train2)

y_pred_lineal = clf.predict(X_test)
 = clf.predict_proba(X_test)

import pickle
# save the model to disk
filename = 'G:/Tot
TFG/TFG/Scripts/03_Models/SVM/SVC_linear_kernel_totesvariables.sav'
pickle.dump(clf, open(filename, 'wb'))

# load the model from disk
filename = 'C:/Users/Albert/Desktop/Models
TFG/SVM/SVC_linearkernel_menysvariables.sav'
loaded_model = pickle.load(open(filename, 'rb'))

prob_lineal_tot = loaded_model.predict_proba(X_test)

prob_svm_lineal = pd.DataFrame(data = prob_lineal_tot[:,1],
columns=["PROBABILITATS"])
prob_svm_lineal['RESPOSTA'] = y_test
print(round(prob_svm_lineal.groupby('RESPOSTA')['PROBABILITATS'].mean(),4))

prob_svm_lineal.to_excel("G:/Tot
TFG/TFG/Scripts/03_Models/SVM/Probabilitat_Default_Lineal_MenysVAR_Test.xlsx")

from sklearn import svm

# Temps d'execució per a la bbdd sencera per al kernel rbf: 1847.47 minuts.

start_time = time()
wclf = svm.SVC(kernel='rbf', class_weight= "balanced", random_state = 2021,
probability= True)

wclf.fit(X_train2, y_train2)

elapsed_time = (time() - start_time)/60
print("Ha acabat de fit el model")
print("Elapsed time: %.10f minutes." % elapsed_time)

y_pred_rbf_75000 = wclf.predict(X_test)

elapsed_time = (time() - start_time)/60
print("Ha acabat de fer les prediccions")
print("Elapsed time: %.10f minutes." % elapsed_time)

y_prob_svm_rbf_75000 = wclf.predict_proba(X_test)

elapsed_time = (time() - start_time)/60
print("Elapsed time: %.10f minutes." % elapsed_time)

import pickle
# save the model to disk
filename = 'C:/Users/Albert/Desktop/Models TFG\SVM/SVC_RBFkernel_245641.sav'
pickle.dump(wclf, open(filename, 'wb'))
```

```
sum(y_pred_rbf)
sum(y_pred_rbf_75000)

y_prob_svm_rbf_75000 = wclf.predict_proba(X_train2)

prob_svm_rbf = pd.DataFrame(data = y_prob_svm_rbf_75000[:,1],
columns=["PROBABILITATS"])
prob_svm_rbf['RESPOSTA'] = y_train2
print(round(prob_svm_rbf.groupby('RESPOSTA')['PROBABILITATS'].median(),4))

prob_svm_rbf.to_excel("G:/Tot
TFG/TFG/Scripts/03_Models/SVM/Probabilitat_Default_RBF_All_Train.xlsx")

#########################################################################

import pandas as pd
# Use numpy to convert to arrays
import numpy as np
import os
import matplotlib.pyplot as plt


pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', None)
pd.set_option('display.max_colwidth', None)

os.getcwd() # Directori actual
os.chdir("G:\Tot TFG\TFG\Data")
os.chdir("F:\Tot TFG\TFG\Data")


#Càrrega TRAIN
train_x = pd.read_excel("TrainData_AllDummiesNew.xlsx")
train_x2 = train_x

train_x2 = train_x2.drop(["Unnamed: 0", "ORGANIZATION_TYPE_XNA"],1)
train_x2 = train_x2.dropna()
train_x2 = train_x2.reset_index(drop=True)
# train_x2 = train_x2.drop(134133)

y_train = train_x2.TARGET

# Càrrega TEST
test_x = pd.read_excel("TestData_AllDummiesNew.xlsx")
test_x2 = test_x

test_x2 = test_x2.drop(["Unnamed: 0", "ORGANIZATION_TYPE_XNA"],1)
test_x2 = test_x2.dropna()
test_x2 = test_x2.reset_index(drop=True)

y_test = test_x2.TARGET

from sklearn.preprocessing import StandardScaler

sc = StandardScaler()

columnes = train_x2.columns

X_train = train_x2
target = X_train.TARGET
X_train = X_train.drop(["TARGET"], axis=1)

# No sé si la fila es correspon amb el TARGET...
X_train = sc.fit_transform(X_train)
df = pd.DataFrame(X_train, columns = columnes[0:(len(columnes)-1)])
df['TARGET'] = target
```

```
df.TARGET.isna().sum()
X_train = df.dropna()
y_train = X_train.TARGET
X_train = X_train.drop(["TARGET"], axis=1)

#Test
target = test_x2.TARGET
test_x2 = test_x2.drop(["TARGET"], axis=1)

X_test = sc.transform(test_x2)

df = pd.DataFrame(X_test, columns = columnes[0:(len(columnes)-1)])
df['TARGET'] = target

df.TARGET.isna().sum()
X_test = df.dropna()
y_test = X_test.TARGET
X_test = X_test.drop(["TARGET"], axis=1)

X_train = X_train.drop([134133], axis = 0)
y_train = y_train.drop([134133])
# XGBoost

import xgboost as xgb
from xgboost import XGBClassifier
from sklearn.metrics import mean_squared_error
import pandas as pd
import numpy as np
from sklearn.metrics import roc_auc_score

from time import time
start_time = time()

import re
regex = re.compile(r"\[|\]|<", re.IGNORECASE)
X_train.columns = [regex.sub("_", col) if any(x in str(col) for x in set(('[',
']', '<'))) else col for col in X_train.columns.values]
X_test.columns = [regex.sub("_", col) if any(x in str(col) for x in set(('[',
']', '<'))) else col for col in X_test.columns.values]

# Quan és no balancejat es posa un scale_pos_weight: num_obs_negatives /
num_obs_positives

model = XGBClassifier(n_estimators = 50, random_state = 2021, booster =
"gbtree")
model.fit(X_train, y_train)

import pickle
filename = 'G:/Tot TFG/TFG/Scripts/03_Models/XGBoost/XGBoost_NOOUTLIER_50.sav'
pickle.dump(model, open(filename, 'wb'))

probabilitat_XGB = model.predict_proba(X_test)
class_test = model.predict(X_test)

prob_xgb = pd.DataFrame(data = probabilitat_XGB[:,1],
columns=["PROBABILITATS"])
prob_xgb['RESPOSTA'] = y_test
print(round(prob_xgb.groupby('RESPOSTA')['PROBABILITATS'].mean(),4))

prob_xgb.to_excel("G:/Tot
TFG/TFG/Scripts/03_Models/XGBoost/Probabilitats/Probabilitats_50TREE_NOOUTLIER
_Test.xlsx")

# Per a train
probabilitat_XGB = model.predict_proba(X_train)
```

```
prob_xgb = pd.DataFrame(data = probabilitat_XGB[:,1],
columns=["PROBABILITATS"])
prob_xgb['RESPOSTA'] = y_train
print(round(prob_xgb.groupby('RESPOSTA')['PROBABILITATS'].mean(),4))

prob_xgb.to_excel("G:/Tot
TFG/TFG/Scripts/03_Models/XGBoost/Probabilitats/Probabilitats_50TREE_NOOUTLIER
_Train.xlsx")
```

8.5.6 Script per al càlcul de l'índex de gini i AUC. Gràfics de probabilitat

```
ModelPerformance <- function (actuals, predictedScores){

  fitted <- data.frame (Actuals=actuals, PredictedScores=predictedScores) #
actuals and fitted
  colnames(fitted) <- c('Actuals','PredictedScores') # rename columns

  ones <- fitted[fitted$Actuals==1, ] # Subset ones
  zeros <- fitted[fitted$Actuals==0, ] # Subsetzeros

  totalPairs <- nrow(ones) * nrow(zeros) # calculate total number of pairs to
check

  # A pair is concordant if 1 (event) has a higher predicted probability than
0
  conc <- sum (c(vapply(ones$PredictedScores, function(x) {((x >
zeros$PredictedScores))}, FUN.VALUE=logical(nrow(zeros)))), na.rm=T)

  # A pair is disconcordant if 1 (event) has a lower predicted probability
than 0
  disc <- sum(c(vapply(ones$PredictedScores, function(x) {((x <
zeros$PredictedScores))}, FUN.VALUE = logical(nrow(zeros)))), na.rm = T)

  # Calculate concordance, discordance, ties and AUC
  concordance <- conc/totalPairs
  discordance <- disc/totalPairs
  tiesPercent <- (1-concordance-discordance)
  Gini = (conc-disc)/totalPairs
  AUC  = concordance + 0.5*tiesPercent
  return(list("Concordance"=round(concordance,6),
"Discordance"=round(discordance,6),
              "Tied"= round(tiesPercent,6), "Gini"= round(Gini,6),"AUC"=
round(AUC,6)))
}


setwd("G:/Tot TFG/TFG/Scripts/03_Models/Random Forest")

prediccions <- read.xlsx("Menys variables/Probabilitats_GLM_Train.xlsx")
resposta = prediccions$RESPOSTA
probabilitats = prediccions$PROBABILITATS
tapply(prediccions$PROBABILITAT, prediccions$RESPOSTA, mean)


prediccions <- read.xlsx("Totes les
variables/Probabilitats_RIDGE_TOT_Test.xlsx")
resposta = prediccions$RESPOSTA
probabilitats = prediccions$PROBABILITATS
tapply(prediccions$PROBABILITAT, prediccions$RESPOSTA, mean)


ModelPerformance(resposta, probabilitats)

# Obtenir GINI i AUC per als models de TRAIN (utilitzar la funció anterior en
la seva totalitat comporta a Errors de Memòria)
```

```
i <- 1
conco <- c()
disco <- c()
tied <- c()
gini <- c()
auc <- c()

while (i <= 50){
  indexs <- sample(1:nrow(prediccions), 45000, replace = FALSE)
  sample_i <- ModelPerformance(resposta[indexs], probabilitats[indexs])
  conco <- c(conco, sample_i$Concordance)
  disco <- c(disco, sample_i$Discordance)
  tied <- c(tied, sample_i$Tied)
  gini <- c(gini, sample_i$Gini)
  auc <- c(auc, sample_i$AUC)
  i <- i + 1
}

mean(conco)
mean(disco)
mean(tied)
mean(gini)#; max(gini); min(gini)

mean(auc); max(auc); min(auc)

#########################################################################

# Gràfics: Lasso, Arbres, RF, els dos SVM, XGBoost.
library(openxlsx)
library(ggplot2)
library(dplyr)
library(Epi)
theme_set(theme_classic() +  theme(legend.position = "top"))

# GLM, LASSO o RIDGE TEST

prediccions_test <- read.xlsx("G:/Tot TFG/TFG/Scripts/03_Models/GLM, LASSO,
RIDGE amb variables no
significatives/Probabilitats/Probabilitats_Lasso_Test.xlsx")
titol1 = "Model logístic amb penalització Lasso"
num_imp_grup <- prediccions_test %>% mutate(Cuts = ncut(PROBABILITATS, breaks
= c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01))) %>% group_by(Cuts) %>%
summarise(Resposta = sum(RESPOSTA))
valors_Lasso <- (num_imp_grup$Resposta /
table(ncut(prediccions_test$PROBABILITATS, breaks =
c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01)))) * 100
prediccions_test$RESPOSTA <- as.factor(prediccions_test$RESPOSTA)
levels(prediccions_test$RESPOSTA) <- c("No Impagament", "Impagament")
colnames(prediccions_test)[2] <- "TARGET"

plot1 <- ggplot(prediccions_test, aes(x = PROBABILITATS)) +
  geom_histogram(aes(color = TARGET, y=..density.., fill = TARGET),
                 position = "identity", bins = 30, alpha = 0.4) +
  scale_color_manual(values = c("#00AFBB", "#E7B800")) +
  scale_fill_manual(values = c("#00AFBB", "#E7B800")) +
  ggtitle(titol1) +
  xlab("Probabilitats") + ylab("Densitat")

prediccions_test <- read.xlsx("G:/Tot TFG/TFG/Scripts/03_Models/Arbres de
decisió/Probabilitats/7DTREE.xlsx")
titol1 = "7DTREE"
num_imp_grup <- prediccions_test %>% mutate(Cuts = ncut(PROBABILITATS, breaks
= c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01))) %>% group_by(Cuts) %>%
summarise(Resposta = sum(RESPOSTA))
valors_Arbres <- (num_imp_grup$Resposta /
table(ncut(prediccions_test$PROBABILITATS, breaks =
c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01)))) * 100
prediccions_test$RESPOSTA <- as.factor(prediccions_test$RESPOSTA)
```

```
levels(prediccions_test$RESPOSTA) <- c("No Impagament", "Impagament")
colnames(prediccions_test)[3] <- "TARGET"

plot2 <- ggplot(prediccions_test, aes(x = PROBABILITATS)) +
  geom_histogram(aes(color = TARGET, y=..density.., fill = TARGET),
                 position = "identity", bins = 30, alpha = 0.4) +
  scale_color_manual(values = c("#00AFBB", "#E7B800")) +
  scale_fill_manual(values = c("#00AFBB", "#E7B800")) +
  ggtitle(titol1) +
  xlab("Probabilitats") + ylab("Densitat")


prediccions_test <- read.xlsx("G:/Tot TFG/TFG/Scripts/03_Models/Random
Forest/Probabilitats/1000_12_RANDOMFOREST_TEST.xlsx")
titol1 = "Random Forest 1000_12"
num_imp_grup <- prediccions_test %>% mutate(Cuts = ncut(PROBABILITATS, breaks
= c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01))) %>% group_by(Cuts) %>%
summarise(Resposta = sum(RESPOSTA))
valors_RF <- (num_imp_grup$Resposta /
table(ncut(prediccions_test$PROBABILITATS, breaks =
c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01)))) * 100
prediccions_test$RESPOSTA <- as.factor(prediccions_test$RESPOSTA)
levels(prediccions_test$RESPOSTA) <- c("No Impagament", "Impagament")
colnames(prediccions_test)[3] <- "TARGET"

plot3 <- ggplot(prediccions_test, aes(x = PROBABILITATS)) +
  geom_histogram(aes(color = TARGET, y=..density.., fill = TARGET),
                 position = "identity", bins = 30, alpha = 0.4) +
  scale_color_manual(values = c("#00AFBB", "#E7B800")) +
  scale_fill_manual(values = c("#00AFBB", "#E7B800")) +
  ggtitle(titol1) +
  xlab("Probabilitats") + ylab("Densitat")

prediccions_test <- read.xlsx("G:/Tot
TFG/TFG/Scripts/03_Models/SVM/Probabilitats/Probabilitat_Default_Lineal_MenysV
AR_Test.xlsx")
titol1 = "Support Vector Machine amb kernel lineal"
num_imp_grup <- prediccions_test %>% mutate(Cuts = ncut(PROBABILITATS, breaks
= c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01))) %>% group_by(Cuts) %>%
summarise(Resposta = sum(RESPOSTA))
valors_SVM_Lineal <- (num_imp_grup$Resposta /
table(ncut(prediccions_test$PROBABILITATS, breaks =
c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01)))) * 100
prediccions_test$RESPOSTA <- as.factor(prediccions_test$RESPOSTA)
levels(prediccions_test$RESPOSTA) <- c("No Impagament", "Impagament")
colnames(prediccions_test)[3] <- "TARGET"

plot4 <- ggplot(prediccions_test, aes(x = PROBABILITATS)) +
  geom_histogram(aes(color = TARGET, y=..density.., fill = TARGET),
                 position = "identity", bins = 30, alpha = 0.4) +
  scale_color_manual(values = c("#00AFBB", "#E7B800")) +
  scale_fill_manual(values = c("#00AFBB", "#E7B800")) +
  ggtitle(titol1) +
  xlab("Probabilitats") + ylab("Densitat")

prediccions_test <- read.xlsx("G:/Tot
TFG/TFG/Scripts/03_Models/SVM/Probabilitats/Probabilitat_Default_RBF_Test.xlsx
")
titol1 = "Support Vector Machine amb kernel RBF"
num_imp_grup <- prediccions_test %>% mutate(Cuts = ncut(PROBABILITATS, breaks
= c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01))) %>% group_by(Cuts) %>%
summarise(Resposta = sum(RESPOSTA))
valors_SVM_RBF <- (num_imp_grup$Resposta /
table(ncut(prediccions_test$PROBABILITATS, breaks =
c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01)))) * 100
prediccions_test$RESPOSTA <- as.factor(prediccions_test$RESPOSTA)
levels(prediccions_test$RESPOSTA) <- c("No Impagament", "Impagament")
colnames(prediccions_test)[3] <- "TARGET"
```

```
plot5 <- ggplot(prediccions_test, aes(x = PROBABILITATS)) +
  geom_histogram(aes(color = TARGET, y=..density.., fill = TARGET),
                 position = "identity", bins = 30, alpha = 0.4) +
  scale_color_manual(values = c("#00AFBB", "#E7B800")) +
  scale_fill_manual(values = c("#00AFBB", "#E7B800")) +
  ggtitle(titol1) +
  xlab("Probabilitats") + ylab("Densitat")

prediccions_test <- read.xlsx("G:/Tot
TFG/TFG/Scripts/03_Models/XGBoost/Probabilitats/Probabilitats_50TREE_Test.xlsx
")
titol1 = "XGBoost"
num_imp_grup <- prediccions_test %>% mutate(Cuts = ncut(PROBABILITATS, breaks
= c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01))) %>% group_by(Cuts) %>%
summarise(Resposta = sum(RESPOSTA))
valors_XGBOOST <- (num_imp_grup$Resposta /
table(ncut(prediccions_test$PROBABILITATS, breaks =
c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01)))) * 100
prediccions_test$RESPOSTA <- as.factor(prediccions_test$RESPOSTA)
levels(prediccions_test$RESPOSTA) <- c("No Impagament", "Impagament")
colnames(prediccions_test)[3] <- "TARGET"

plot6 <- ggplot(prediccions_test, aes(x = PROBABILITATS)) +
  geom_histogram(aes(color = TARGET, y=..density.., fill = TARGET),
                 position = "identity", bins = 30, alpha = 0.4) +
  scale_color_manual(values = c("#00AFBB", "#E7B800")) +
  scale_fill_manual(values = c("#00AFBB", "#E7B800")) +
  ggtitle(titol1) +
  xlab("Probabilitats") + ylab("Densitat")

require(gridExtra)
grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6, ncol=2, nrow =3)

# Gràfic:

plot(valors_XGBOOST, type = "b", frame = TRUE, pch = 18,
     col = "#009999", xlab = "Probabilitats estimades", ylab = "Percentatge
d'impagaments reals en cada interval")
# Add a second line
lines(valors_Lasso, pch = 18, col = "#CC33CC", type = "b", lty = 2)
lines(valors_Arbres, pch = 18, col = "#FF6633", type = "b", lty = 2)
lines(valors_RF, pch = 18, col = "#33FF66", type = "b", lty = 2)
lines(valors_SVM_Lineal, pch = 18, col = "#CCCC00", type = "b", lty = 2)
lines(valors_SVM_RBF, pch = 18, col = "#000066", type = "b", lty = 2)

# Add a legend to the plot
legend("topleft", legend=c("Model Lasso", "Arbres", "Random Forest", "SVM
Lineal", "SVM RBF", "XGBoost"),
       col=c("#CC33CC", "#FF6633", "#33FF66", "#CCCC00", "#000066",
"#009999"), lty = 18, cex=1)


# Comparació amb el logístic:

prediccions_test <- read.xlsx("G:/Tot TFG/TFG/Scripts/03_Models/GLM, LASSO,
RIDGE amb variables no
significatives/Probabilitats/Probabilitats_GLM_Test.xlsx")

num_imp_grup <- prediccions_test %>% mutate(Cuts = ncut(PROBABILITATS, breaks
= c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01))) %>% group_by(Cuts) %>%
summarise(Resposta = sum(RESPOSTA))
valors_Logistic <- (num_imp_grup$Resposta /
table(ncut(prediccions_test$PROBABILITATS, breaks =
c(0,0.04,0.08,0.12,0.16,0.20,0.30,0.5,0.75,1.01)))) * 100

valors_XGBOOST2 <- (valors_XGBOOST - valors_Logistic)/valors_Logistic*100
valors_Lasso2 <- (valors_Lasso - valors_Logistic)/valors_Logistic*100
```

```
valors_Arbres2 <- (valors_Arbres - valors_Logistic)/valors_Logistic*100
valors_RF2 <- (valors_RF - valors_Logistic[1:7])/valors_Logistic[1:7]*100
valors_SVM_Lineal2 <- (valors_SVM_Lineal -
valors_Logistic)/valors_Logistic*100
valors_SVM_RBF2 <- (valors_SVM_RBF -
valors_Logistic[1:8])/valors_Logistic[1:8]*100


plot(valors_XGBOOST2, type = "b", frame = TRUE, pch = 18,
     col = "#009999", xlab = "Probabilitats estimades", ylab = "Canvi
percentual respecte el model Logístic", ylim = c(-30,80))
abline (h = 0, col= "red")
# Add a second line
lines(valors_Lasso2, pch = 18, col = "#CC33CC", type = "b", lty = 2)
lines(valors_Arbres2, pch = 18, col = "#FF6633", type = "b", lty = 2)
lines(valors_RF2, pch = 18, col = "#33FF66", type = "b", lty = 2)
lines(valors_SVM_Lineal2, pch = 18, col = "#CCCC00", type = "b", lty = 2)
lines(valors_SVM_RBF2, pch = 18, col = "#000066", type = "b", lty = 2)

# Add a legend to the plot
legend("topleft", legend=c("Model Lasso", "Arbres", "Random Forest", "SVM
lineal", "SVM_RBF", "XGBoost"),
       col=c("#CC33CC", "#FF6633", "#33FF66", "#CCCC00", "#000066",
"#009999"), lty = 18, cex=1)
```

8.5.7 Interpretative Machine Learning per al model XGBoost

```
import os
import pickle
import pandas as pd
filename = 'XGBoost_50.sav'
xgboost = pickle.load(open(filename, 'rb'))

#Càrrega TRAIN
train_x = pd.read_excel("TrainData_AllDummiesNew.xlsx")
train_x2 = train_x

train_x2 = train_x2.drop(["Unnamed: 0", "ORGANIZATION_TYPE_XNA"],1)
train_x2 = train_x2.dropna()
train_x2 = train_x2.reset_index(drop=True)
# train_x2 = train_x2.drop(134133)

y_train = train_x2.TARGET

# Càrrega TEST
test_x = pd.read_excel("TestData_AllDummiesNew.xlsx")
test_x2 = test_x

test_x2 = test_x2.drop(["Unnamed: 0", "ORGANIZATION_TYPE_XNA"],1)
test_x2 = test_x2.dropna()
test_x2 = test_x2.reset_index(drop=True)

y_test = test_x2.TARGET

# Si base de datos entera.
train_x2 = train_x

train_x2 = train_x2.drop(["Unnamed: 0", "ORGANIZATION_TYPE_XNA"],1)
train_x2 = train_x2.dropna()
train_x2 = train_x2.reset_index(drop=True)
# train_x2 = train_x2.drop(134133)

y_train = train_x2.TARGET

test_x2 = test_x
```

```
test_x2 = test_x2.drop(["Unnamed: 0", "ORGANIZATION_TYPE_XNA"],1)
test_x2 = test_x2.dropna()
test_x2 = test_x2.reset_index(drop=True)

y_test = test_x2.TARGET


#Estandaritzar train i test
from sklearn.preprocessing import StandardScaler

sc = StandardScaler()

columnes = train_x2.columns

X_train = train_x2
target = X_train.TARGET
X_train = X_train.drop(["TARGET"], axis=1)

# No sé si la fila es correspon amb el TARGET...
X_train = sc.fit_transform(X_train)
df = pd.DataFrame(X_train, columns = columnes[0:(len(columnes)-1)])
df['TARGET'] = target

df.TARGET.isna().sum()
X_train = df.dropna()
y_train = X_train.TARGET
X_train = X_train.drop(["TARGET"], axis=1)

#Test
target = test_x2.TARGET
test_x2 = test_x2.drop(["TARGET"], axis=1)

X_test = sc.transform(test_x2)

df = pd.DataFrame(X_test, columns = columnes[0:(len(columnes)-1)])
df['TARGET'] = target

df.TARGET.isna().sum()
X_test = df.dropna()
y_test = X_test.TARGET
X_test = X_test.drop(["TARGET"], axis=1)

# XGBOOST ELI5
filename = 'XGBoost_50.sav'
xgboost = pickle.load(open(filename, 'rb'))
# Les variables més importants en conjunt:
eli5.explain_weights(xgboost, top = 100)

import re
regex = re.compile(r"\[|\]|<", re.IGNORECASE)
X_test.columns = [regex.sub("_", col) if any(x in str(col) for x in set(('[',
']', '<'))) else col for col in X_test.columns.values]

eli5.show_prediction(xgboost, X_test.iloc[3], feature_names =
list(X_train.columns))
[test_x2.iloc[3].MEAN_EXT_SOURCE, test_x2.iloc[3].TOT_INSTALMENTS_CC,
test_x2.iloc[3].MAX_EXT_SOURCE, test_x2.iloc[3].AMT_INCOME_TOTAL, y_test[3]]

eli5.show_prediction(xgboost, X_test.iloc[11], feature_names =
list(X_train.columns))
[test_x2.iloc[11].MEAN_EXT_SOURCE, test_x2.iloc[11].Refused_STATUS_HC,
test_x2.iloc[11].MIN_EXT_SOURCE,
test_x2.iloc[11].CODE_GENDER_F,test_x2.iloc[11].RATI_DEUTE_GARANTIA,
y_test.iloc[11]]
[test_x2.iloc[11].RATI_DEUTE_GARANTIA,
test_x2.iloc[11].NAME_FAMILY_STATUS_Married,
test_x2.iloc[11].NUM_CREDITS_PREVIS_TANCATS]
```

```
eli5.show_prediction(xgboost, X_test.iloc[43], feature_names =
list(X_train.columns))
[test_x2.iloc[43].RATI_DEUTE_GARANTIA, test_x2['LAST_YEARS_EMPLOYED_0-
3'].iloc[43],
test_x2.iloc[43].MEAN_EXT_SOURCE,test_x2.iloc[43].CODE_GENDER_F,test_x2.iloc[4
3].OCCUPATION_TYPE_Drivers, y_test.iloc[43]]
```