UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

MODELITZACIÓ DEL RISC DE CRÈDIT A TRAVÉS DEL MACHINE LEARNING

Autor: Joan Llano Carcasona

Director: Dr. Josep Fortiana
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 21 de juny de 2020

Abstract

EBA (European Banking Authority) is requiring the banking institutions to show the procedures they are carrying out, alongside the models that they are using, to estimate risk metrics, particularly, the credit risk.

Against this background, the financial institutions are finding it difficult to justify how the Machine Learning Methods work. This gives rise to the following question: Is it possible to justify the Machine Learning Methods as applied to the credit risk estimation through Mathematics? Aiming at answering this question, we study several Machine Learning methods, both from a mathematical perspective and their reals applications.

The structure of the work is divided into four parts.
In the first part, we describe in detail the main issue to be studied, that is, in the event of a customer applying for a loan, the model should be able to predict whether the loan should be granted or not. We illustrate our models with a real database from a financial institution.

The second part of this research is devoted to the theory of the following Machine Learning Methods: Logistic regression, Classification Trees (including Bagging and Random Forest), and Boosting (Adaboost).

The third part is a practical application of the above models, using the statistical software "R". We train models on a subset form our database, and assess the discriminatory capacity against the new observations.

Finally, we analyze the results obtained, propose future research areas and draw final conclusions.

Resum

La EBA exigeix a les entitats bancàries a justificar els procediments aplicats i els models utilitzats per estimar mètriques de risc, en particular, el risc de crèdit.

Davant d'aquesta situació, les entitats financeres tenen dificultats per mostrar com funcionen els mètodes de *machine learning*. Això dóna lloc a la següent pregunta. És possible justificar els models de *machine learning* aplicats al risc de crèdit a través de les matemàtiques? Per respondre aquesta pregunta estudiem, diferents metodologies de *machine learning* tant des d'una perspectiva matemàtica com de la seva aplicació real.

L'estructura del treball es divideix en quatre parts.
En la primera part descrivim amb detall el problema a estudiar, és a dir, en el cas que un client sol·liciti un préstec, el model hauria de predir si concedir-li o no. Per construir els models utilitzem una base de dades real d'una entitat financera.

La segona part del treball la dediquem a la teoria dels següents models de *machine learning*: Regressió Logística, Arbres de Classificació (incloent Bagging i Random Forest) i el mètode Boosting (Adaboost).

La tercera part és una aplicació real dels models anteriors mitjançant el programa estadístic "R". Entrenem els models en un subconjunt de la nostra base de dades i avaluem la capacitat discriminatòria del model final enfront noves observacions.

La quarta i última part, analitzem els resultats obtinguts, proposem línies d'investigació futures i arribem a unes conclusions finals.

2010 Mathematics Subject Classification. 62J05, 62J99, 62P05

Agraïments

En primer lloc, agrair al tutor del treball, en Dr. Josep Fortiana ja que, a pesar de les complicacions que hem tingut a causa del confinament, ha estat disposat en tot moment a atendre'm i ajuda'm amb els dubtes que he tingut al llarg del treball.

Agraïr també a la meva família, en especial a la meva mare, per tots els esforços que ha fet durant la meva etapa universitària.

I no em puc oblidar dels FELLININI's, que sense saber molt bé com, sempre acaben fent que les coses siguin mes fàcils.

Índex

1 Introducció

1.1 Problema a Estudiar

Avui en dia, cada cop més se sent parlar de l'ús de tècniques d'Intel·ligència Artificial en camps com la investigació, biomedicina, màrqueting i altres àmbits. En particular en el món de les finances. L'ús d'aquestes eines al sistema financer i en concret al món bancari, està generant molta controvèrsia ja que, hi ha regulacions, com l'Autoritat Bancaria Europea que obliguen a les entitats a plasmar i traçar tots els procediments aplicats i els models utilitzats per l'obtenció de resultats com, probabilitats d'incompliment, ràtios de transició i altres paràmetres i magnituds financeres.

La controvèrsia neix perquè ben cert és que els mètodes d'Intel·ligència Artificial funcionen molt bé, sovint millor que els mètodes convencionals i simplifiquen procediments costosos a nivell de recursos i d'implementació. Tanmateix, justificar el perquè funcionen no és una tasca senzilla, ja que es coneixen els *inputs* i els *outputs*, però els passos intermedis, que és on intervé el *Machine Learning* a través de la Intel·ligència Artificial, són dificultosos de comprendre i explicar com funcionen. Per tant, en l'àmbit regulador, és complicat justificar l'aplicació d'aquests mètodes.

Per tant, la motivació i la finalitat d'aquest treball és fer ús de les matemàtiques per ser capaços d'entendre mètodes de *Machine Learning* que es poden utilitzar en el l'àmbit bancari, amb l'objectiu d'explicar aquests passos intermedis entre els *inputs* i el *output* per posteriorment, implementar-los i justificar-los.

Un cop estudiats els mètodes, es procedeix a la posada en pràctica dels mateixos des de la següent perspectiva.

Ens situem des de la posició d'un banc financer, i volem saber si concedir finançament a través d'un préstec a un nou client. Per a fer-ho hem de definir una sèrie de criteris i preguntes, aquests, són els nostres inputs, i l'output conceptualment és si se li coincideix el préstec o se li denega. En cas que compleixi amb els requisits fixats per l'entitat o el regulador, se li concedeix el préstec, en cas contrari no se li concedeix.

És a dir, el problema a estudiar és donat uns inputs, estimar la probabilitat que un determinat client no compleixi amb les seves obligacions creditícies. Aquesta probabilitat es coneix com a Probabilitat d'Incompliment, ($Probability\ of\ Default$) d'ara endavant PD i serà el criteri a seguir per decidir la concessió del préstec, per tant, la PD és l'output o variable dependent a predir. Dir que aquesta PD també es pot utilitzar per fer una qualificació creditícia dels clients de l'entitat financera per predir si en un futur poden tenir dificultats per fer front a les seves obligacions de pagament.

Els models de probabilitat d'incompliments, utilitzen una combinació de dades de comportament determinat, com la tendència en la seva distribució, informació financera, ràtios financeres, situació econòmica i informació subjectiva, és a dir, altra informació que l'entitat consideri important.

Per tant, el problema a estudiar és la modelització de la PD fent ús d'aquesta informació i aplicant mètodes de *Machine Learning*. La construcció del model, de forma

resumida consisteix en els següents passos:

1. **Anàlisi estadístic de les variables explicatives:** Estudiarem la raonabilitat econòmica de cada una de les variables explicatives i analitzarem la capacitat discriminatòria respecte a la variable de resposta. Un cop seleccionades les variables predictives, si s'escau, les transformarem per deixar-les en magnituds comparables i després amb l'objectiu d'identificar les variables predictives candidates al model, analitzarem la correlació entre elles.

2. **Estimació del model** En aquest apartat, aplicarem la teoria que hi ha darrere els mètodes de *Machine Learning* estudiats en la tesi. És on descriurem la metodologia i els mètodes utilitzats per l'obtenció del model per cadascun dels mètodes estudiats.

3. **Validació** En aquest apartat descriurem l'anàlisi de la capacitat predictiva del model sobre la base de les observacions no utilitzades en el procés d'entrenament del model. Per tant, compararem les capacitats discriminatòries dels models entrenats amb els models entrenats aplicats en la mostra de validació.

1.2 Anàlisi de dades

Per dur a terme el problema explicat anteriorment, utilitzarem informació històrica relativa als clients d'una entitat financera i els seus contractes. Aquests clients suposen persones no físiques (Empreses) les quals la mateixa entitat ha segmentat en funció del volum de facturació per fer l'estudi. És a dir, totes les empreses tenen un mínim de facturació. La informació que es presenta és a nivell client i no a nivell contracte. (Un client pot tenir més d'un contracte).

A cada client se l'hi fa una marca de *bo* o *dolent* en funció de si compleix alguna de les següents característiques, si en compleix alguna, es marca com a *dolent*, altrament *bo*:

- Clients amb algun contracte amb impagaments de 30 o més dies i import mig total impagat igual o superior a 500 euros.
- Clients amb algun contracte refinançat superior al 20% del finançament total.
- Clients amb contractes marcats com dubtosos subjectius.
- Clients en situació de concurs creditor.

Per tant la variable resposta es defineix com

$$Y = BM\_12M = \begin{cases} 0 \quad si \quad \text{Client marcat com } bo \\ \\ 1 \quad si \quad \text{Client marcat com } dolent \end{cases}$$

Les dades recullen informació en set finestres temporals, corresponents als mesos de març i setembre entre març de 2015 i març de 2018. En aquestes set finestres temporals, considerant els criteris prèviament definits, el nombre de clients i la distribució de *bons* i *dolents* en cada mes de referència és:

Finestra Temporal	bons	dolents	Total	% bons	% dolents
201503	4.776	277	5.053	94,5%	5,5%
201509	5.335	332	5.667	94,1%	5,9%
201603	5.575	502	6.077	91,7%	8,3%
201609	6.177	552	6.699	92,2%	7,8%
201703	6.698	370	7.068	94,8%	5,2%
201709	7.532	344	7.876	95,6%	4,4%
201803	8.085	365	8.414	95,7%	4,3%
Total	**44.143**	**2.711**	**46.854**	**94,2%**	**5,8%**

Taula 1: Registres d'observacions.

En les dades trobem la següent informació

Codi	Informació mes i Client	Tipologia
ID MESES	Mes de referència (AAAAMM)	Discreta
ID CLIEN	Identificador del Client	Discreta

Taula 2: Registres de mesos i clients

També se'ns ha facilitat un llistat d'inputs, que seran subjectes d'estudi per veure quines d'elles acaben formant part del model final. A les variables explicatives, les anomenen Factors de Risc.

FR	Descripció del Factor de Risc	Tipologia
FR_1	Despeses financeres sobre passiu exigible	Contínua
FR_2	Fons propis sobre actiu total	Contínua
FR_3	Despeses financeres sobre BAIT	Contínua
FR_4	Ràtio de cobertura al servei del deute	Contínua
FR_5	Mitja últims 365D del saldo mitjà passiu vista sobre facturació	Contínua
FR_6	Disponible mitjà sobre el límit U60D	Contínua
FR_7	Termini mitjà de finançament circulant U365D	Discreta
FR_8	Nombre d'impagaments 10D en aquesta o altres entitats	Discreta
FR_9	Nombre d'excedits o descoberts ($\geq$3D) U180D en aquesta o altres entitats	Discreta
FR_{10}	Variació del benefici net sobre recursos propis	Contínua
FR_{11}	Rotació dies d'existències sobre compres	Contínua
FR_{12}	Variació del resultat de l'exercici	Contínua
FR_{13}	Resultat de l'exercici sobre facturació	Contínua
FR_{14}	Variació del cash-flow	Contínua
FR_{15}	Variació EBITDA sobre facturació	Contínua
FR_{16}	Variació dels fons propis	Contínua
FR_{17}	Fons de maniobra relatiu	Contínua
FR_{18}	Període de dies de cobrament	Contínua
FR_{19}	Variació període de dies de cobrament	Contínua
FR_{20}	Variació fons de maniobra sobre actiu total	Contínua

FR_{21}	Variació fons de maniobra	Contínua
FR_{22}	Variació del ràtio de fluxos de caixa sobre deute exigible a ll/t	Contínua
FR_{23}	Ràtio entre l'import d'efectes reclamats i prorrogats sobre descomptats	Contínua
FR_{24}	Nombre de mesos que l'empresa porta operant	Discreta
FR_{25}	Indicador d'increment en el termini de venciment d'algun dels efectes descomptats U60D respecte el termini de venciment dels efectes descomptats en els U365D	Discreta
FR_{26}	Percentatge d'impagaments línies comercials U180D	Contínua
FR_{27}	Morosos o fallits en altres entitats	Discreta

Taula 3: Factors de Risc

Per tant, tenim 27 factors de risc a analitzar. Generalment són factors de tipus comportamental, d'impagament i d'estats financers de l'empresa. Estudiem l'estructura de les dades.

Veiem que a la taula 3, tenim un total de 46.854. Dels quals, 44.143 corresponents a observacions marcades com *bons* i 2.711 com *dolents*. Percentualment tenim que el $94,2\%$ de la mostra són *bons* i el $5,8\%$ són *dolents*.

Missings i valors atípics
S'estudia el nombre de *missings* (són valors marcats com NA el que significa que no estan informats). El total de *missings* en les dades ascendeix a set i provenen tots del mateix factor de risc *Nombre de mesos que l'empresa porta operant*. Donat que les observacions amb *missings* no són prou significatives per a tractar-les com un grup a part i provenen del mateix factor, decidim omplir el camp amb el valor zero ja que sembla coherent pensat que si encara no estan informats, és que fa poc que operen al mercat.

Per altra banda, hi ha factors de risc que presenten el valor $10E + 13$. Aquest valor està marcat per la pròpia entitat quan el valor presenta valors atípics o bé valors anormals com per exemple, quan en una ràtio s'esta dividint entre zero, o quan algun input de construcció del factor no té sentit, com ara un actiu o passiu negatiu.

Segon el Pla General Comptable espanyol definim actiu i passiu com:
- Actiu: béns, drets i altres recursos controlats econòmicament per l'empresa, dels que s'espera obtenir beneficis o rendiments econòmics en un futur.
- Passiu: Obligacions actuals sorgides com a conseqüència de successos passats.

Desbalancejament de les dades
Ens trobem un una situació on la població de *dolents* és minoritària, ja que percentualment tenim poques mostres d'aquesta classe. Donat el context que estem, no és d'estranyar que hi hagi molta més població de bons que de dolents, ja que si no l'entitat financera tindria problemes de viabilitat econòmica.

Aquesta característica, de cara a l'anàlisi de dades, presenta un problema conegut com a *dades desbalancejades*. Aquest problema genèricament afecta els algorismes en el seu procés de generalització de la informació perjudicant la classe minoritària, ja que el més

probable és que el model no aconsegueixi diferenciar entre una classe i una altra marcant com a *bons* clients que podrien ser dolents. Per tant l'objectiu ara és trobar una solució a aquest problema.

Algunes de les estratègies per contrarestar aquest problema són les següents.

1. *Oversampling*: Consisteix a duplicar observacions de la classe minoritària. Però no és recomanable, ja que el model pot caure en problemes de $overfitting$. És a dir, pot fer bones prediccions sobre el conjunt d'entrenament però el model no s'ajusta a noves observacions.

2. *Undersampling*: És l'exercici contrari de l'$oversampling$. Es tracta d'eliminar observacions de la classe majoritària per equilibrar les dues poblacions de classes. Però és perillós perquè podem estar prescindint d'observacions importants.

3. *Artificial sampling*: Es pot intentar crear mostres sintètiques utilitzant diversos algorismes que intenten seguir la tendència de grup minoritari. El risc que presenta aquesta estratègia és que podem estar alterant la distribució original de la classe minoritària i confondre al model en la seva classificació.

4. *Splitting data*: Aquesta tècnica es basa a dividir el conjunt de dades. És a dir, s'agafa un percentatge superior al 50% de les observacions per entrenar el model i amb el percentatge restant es fa el test d'error del model. És a dir, es valida el model amb les observacions que no s'han utilitzat per entrenar-lo.

   Per més informació veure Guo *et al.* (2008).

Les tres primeres estratègies són recomanables si es tenen poblacions de 99% i 1%. Com que no és el nostre cas, triem el punt 4. Aplicant-lo ens podem trobar el problema que estem tenint, és a dir, després de fer la divisió de dades, podem seguir tenint una població majoritària en el conjunt d'entrenament. No obstant, no cal fer cap canvi, utilitzarem estadístics que avaluen la capacitat predictiva i discriminatòria del model, com ara la corba *ROC*, sobre la que entrarem en més detall a la secció Modelització, diu que el rendiment global de la prova és correcta si el *ROC* és superior al 70%.

En la nostra mostra, el conjunt d'entrenament conté el 80% de les observacions i el conjunt de validació el conté el 20% restant. Dir que aquest 80% s'ha aplicat sobre les dues poblacions perquè tinguin distribucions semblants.
En la taula següent veiem com han quedat les distribucions de les mostres.

Classe	**Mostra**	**Mostra d'entrenament**	**Mostra de validació**
bons	44.143	35.321	8.822
dolents	2.711	2.175	536
Total	**46.854**	**37.496**	**9.358**

Taula 4: Splitting Data.

Un cop tenim el conjunt d'entrenament, passem a analitzar les variables. Amb l'objectiu d'analitzar les distribucions i la raonabilitat econòmica de cada una de les variables explicatives s'analitza la seva distribució verificant els percentils:

0, 1, 10, 25, 50, 75, 90, 95, 95, 99, 100. Posteriorment, amb l'objectiu d'analitzar el poder discriminant de cada variable predictiva respecte a la variable depenent, s'analitzen els següents estadístics:

- **Weight of Evidence**: És una mesura estadística que determina el poder predictiu d'una variable predictiva en relació amb la variable resposta. És a dir, descriu com un factor de risc es capaç de separar clients bons i dolents i per tant, és un indicador de la correlació entre la variable predictiva i la variable resposta, en el capítol 4 s'entra més en detall amb un exemple. Aquesta tècnica i les primeres aplicacions es remunten al treball d'Alan Turing en la Segona Guerra Mundial (Smith *et al.* (2002)). Per al càlcul del $Weight\ of\ Evidence$ (WOE) és necessari discretitzar les variables. Per a fer-ho, es fan reunions de clients amb característiques similars, i es defineixen trams per a cada reunió buscant un comportament diferenciat de cada tram respecte a la variable dependent.

  El WOE per a cada factor es calcula com:

  $$WOE_i = \log\left(\frac{\%bons_i}{\%dolents_i}\right) \quad 1 \leq i \leq N$$

  On N és el nombre de trams que té el factor (pot ser diferent per a cada factor de risc). $\%Bons$ i $\%dolents$ correspon al percentatge de bons i dolents al tram i respecte el total d'observacions de *bons* i *dolents* en el tram.

  El WOE pot ser positiu, negatiu o zero en funció de quin percentatge de població de bons i dolents és més gran o igual. Un tram tindrà major poder discriminant com més gran sigui el valor del WOE en valor absolut. És a dir, quan les poblacions entre bons i dolents siguin el més distants possible.

- **Information Value**: Mitjançant el $Information\ Value$ (IV) s'avalua la capacitat discriminant o predictiva del factor de risc respecte a la definició de *dolent*.
  El IV es calcula com una combinació lineal dels WOE_i per cadascun dels i trams definits prèviament per a cada factor de risc.

  $$IV = \sum_{i=1}^{N} (\%bons_i - \%dolents_i) \cdot WOE_i$$

  Un $IV = 0$ significa que en tots els N trams del factor de risc, el nombre de bons i dolents és el mateix i per tant el factor no pot discriminar entre bons i dolents. Per tant, a mesura que augmenta el valor del IV significa que major és la capacitat discriminant del factor. Com a punt de referència, es considera que un factor de risc té una predictibilitat acceptable per entrar al model a partir de $IV \geq 0,1$. Veure Majer, I. (2006) per a més informació.

Per últim, es transformen els factors que no s'hagin descartat en l'anàlisi de la capacitat predictiva amb l'objectiu que totes les variables siguin comparables entre elles. Un mètode bastant utilitzat és la transformació a partir del WOE. El valor del factor per a cada client "perd"el seu valor original i passa a ser una variable discreta amb el valor el valor del WOE calculat per al tram corresponent. En el capítol de la Modelització es troba una explicació i un exemple detallat de com portar-ho a terme.

2 Teoria dels mètodes del Machine Learning

2.1 Introducció

La informació d'aquesta introducció, s'extreu de Bourel, M. (2012) i Vapnik, V. (1998). El principi de l'aprenentatge automàtic o *Machine Learning*, en el context supervisat, a partir d'una mostra d'aprenentatge $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ constituïda per n observacions d'un parell de variables aleatòries (x, y) on $x \in X$ són vectors independents i idènticament distribuïts (i.i.d.) que segueixen una funció de distribució desconeguda (però fixa) $F(X)$, consisteix en construir una funció $f\colon X \longrightarrow Y$, amb la qual, donat un vector d'entrada x, es pugui predir amb cert grau de certesa la variable $y = f(x)$. Per cada variable d'entrada $x_i \in X$ se li diu variable explicativa o predictiva (normalment és un vector) i a $y_i \in Y$ variable dependent. Quan la variable dependent és discreta parlem d'un problema de classificació i quan és contínua d'un problema de regressió.

Matemàticament, per aconseguir-ho, suposem que la funció f pertany a una certa classe de funcions $\mathcal{H}$ i es busca f^* que minimitza el risc R_L sobre una funció de pèrdua L, és a dir:

$$f^* = Arg\min_{f\in\mathcal{H}} R_L(f) = Arg\min_{f\in\mathcal{H}} E_{x,y}(L(f, x, y))$$

En un problema de classificació, si $y \in \{1, \ldots, K\}$, la funció de pèrdua sovint és

$$L(f, x, y) = \mathbb{1}_{\{y\neq f(x)\}}$$

i el risc i la funció que minimitza el risc:

$$R_L(f) = P(y \neq f(x)) \quad \text{i} \quad f^* = Arg\max_{k\in\{1,\ldots,K\}} P(y = k|x)$$

És a dir, f^* prediu la classe k que fa màxima la probabilitat a posteriori de y coneixent x. Aquest classificador és coneix com *classificador de Bayes*.

A la pràctica, com es desconeix la funció de distribució conjunta de (x, y), es busca trobar una funció f_n^*, basada en la distribució empírica, que minimitzi el risc empíric $R_{n,L}(f)$ sobre la mostra $\mathcal{L}$, és a dir:

$$f_n^* = Arg\min_{f\in\mathcal{H}} R_{n,L}(f) = Arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L(f(x_i), x_i, y_i)$$

En problemes de classificació binaris, dos mètodes d'aplicació molt utilitzats en aprenentatge automàtic són les estimacions a través de regressions logístiques i arbres de classificació (CART).

Per tant, en els capítols següents veurem una en detall com fer un model de regressió logística i veurem diferents mètodes de classificació aplicats als arbres de decisió.

2.2 de Regressió Logística

La informació d'aquest capítol s'ha extret, en gran part de Peña, D. (2002) i s'ha complementat amb Hastie, T., Tibshirani, R., Friedman, J. (2009) i James, G., Witten, D.,

Hastie, T., Tibshirani, R. (2013).
La regressió logística forma part dels mètodes de discriminació o classificació com els arbres de decisió que explicarem més endavant. Aquests mètodes s'utilitzen per categoritzar dades de resposta en funció d'un conjunt de variables explicatives o predictives. Un dels avantatges que presenta el mètode de regressió logística respecte a altres mètodes de classificació és que és un model estadístic, per oposició a un algorisme de classificació. Una conseqüència d'aquest fet és que els coeficients de la regressió són interpretables d'una manera semblant als de la regressió ordinària, en aquesta un coeficient $\hat{\beta}$ comporta que un increment d'una unitat de la variable predictora produeix un increment de $\hat{\beta}$ unitats de la resposta. En la regressió logística la relació és més complicada, ho veurem més endavant, però segueix essent directa. En altres mètodes és possible determinar la importància relativa de cada predictor, però no hi ha aquesta connexió.

Aquest mètode s'utilitza cada vegada més en una àmplia varietat d'aplicacions. Els primers usos van ser en estudis biomèdics, però els últims vint anys també s'han utilitzat molt en investigació i màrqueting de ciències socials. Recentment, la regressió logística s'ha convertit en una eina popular en aplicacions comercials. Algunes aplicacions de qualificació creditícia utilitzen la regressió logística per modelar la probabilitat que un subjecte sigui solvent, és a dir, el nostre cas.

2.2.1 Principis bàsics

Considerem el problema de discriminar entre dues poblacions P_0 i P_1.
Una forma d'abordar el problema és definir una variable de classificació Y, que pren el valor 0 quan l'element pertany a P_0 i 1 quan l'element pertany a P_1.

Per tant la mostra consisteix en n observacions del tipus $(x_1, y_1), \ldots, (x_n, y_n)$ on x_i és un vector de k variables predictives i y_i ens indica a quina població pertany la i-èssima observació.

Matricialment la mostra ens queda organitzada de la següent manera:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

L'objectiu actual és fer un model que predigui el valor de la variable binaria y_i quan coneixem noves dades de les variables predictives, de manera que puguem analitzar estadísticament quines variables influeixen en la resposta i puguem valorar la capacitat predictiva del model.

Formulem el model de regressió com:

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon_i \text{ amb } \epsilon_i \sim \ N(0, \sigma^2) \tag{2.1}$$

Es podrien estimar els β_i mitjançant el mètode de mínims quadrats ordinaris obtenint així una regressió lineal, però veurem que aquest model presenta problemes.

Comencem calculant l'esperança del model anterior un cop hem estimat els paràmetres β_i:

$$E[y_i|x_1,\ldots,x_k] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{2.2}$$

Definició 2.1. *Definim p_i com la probabilitat que y_i prengui el valor 1, és a dir, que pertanyi a la població P_1 , per tant*

$$p_i = P(y_i = 1|x_1,\ldots,x_k)$$

Donat que la variable y_i és binomial, y_i segueix una distribució de Bernoulli de paràmetre p_i. Per tant $y_i \sim \ Be(p_i)$ i la seva esperança és:

$$E[y_i|x_1,\ldots,x_k] = p_i \times 1 + (1-p_i) \times 0 = p_i$$

I de 2.2 obtenim,

$$p_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Arran d'aquest resultat ens sorgeixen dos problemes:

1. Si estimem el model 2.1, la predicció $\hat{y_i} = \hat{p_i}$ estima que un cert individu amb característiques x_i pertanyi a la població P_1. No obstant això, no tenim cap garantia que p_i estigui entre 0 i 1 i per tant, podem obtenir probabilitats negatives o majors que 1 presentant un problema d'interpretació.

2. Els únics valors de y_i són 1 o 0 per tant la pertorbació ϵ_i pot prendre els valors:

$$\epsilon_i = \begin{cases} 1-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k) & si \quad y_i = 1 \\ \\ -(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k) & si \quad y_i = 0 \end{cases} = \begin{cases} 1-p_i & si \quad y_i = 1 \\ \\ -p_i & si \quad y_i = 0 \end{cases}$$

   Passem a calcular la seva esperança i variància:

$$E[\epsilon_i] = p_i \times (1-p_i) + (1-p_i) \times (-p_i) = 0$$

$$Var[\epsilon_i] = E[{\epsilon_i}^2] - E[\epsilon_i]^2 = E[{\epsilon_i}^2] = p_i \times (1-p_i)^2 + (1-p_i) \times (-p_i)^2 = (1-p_i)p_i$$

   La variància de la pertorbació no és constant, ja que la probabilitat pot variar per a cada individu i en conseqüència no segueix una distribució normal.

2.2.2 Construcció de la regressió logística

En l'apartat anterior hem vist per què no podem construir el nostre model com una regressió lineal múltiple. Si volem que el nostre model construït per discriminar ens proporcioni directament la probabilitat de pertànyer a cada població, hem de transformar la variable resposta per assegurar que estigui entre 0 i 1. Escrivim:

$$p_i = F(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

Definició 2.2. *Sigui F: $\mathbb{R} \longrightarrow (0,1)$ una funció continua, definim la funció Sigmoide com $F(x) = \frac{1}{1+e^{-x}}$*

Per tant, la funció Sigmoide F, és la funció de distribució logística buscada donada per:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$

Passem a calcular $1 - p_i$:

$$1 - p_i = \frac{e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}} = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$

Fent ara el quocionet de p_i i $1 - p_i$ tenim:

$$O = \frac{p_i}{1 - p_i} = \frac{\frac{1}{1+e^{-(\beta_0 + \beta_1^t x)}}}{\frac{1}{1+e^{(\beta_0 + \beta_1^t x)}}} = \frac{1 + e^{(\beta_0 + \beta_1^t x)}}{1 + e^{-(\beta_0 + \beta_1^t x)}} = e^{(\beta_0 + \beta_1^t x)} \tag{2.3}$$

on $\beta_1^t = (\beta_1, \ldots, \beta_k)$ *i* $x^t = (x_1, \ldots, x_k)$

I aplicant logaritmes al quocient anterior arribem a:

$$g(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1^t x = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{2.4}$$

que és un model lineal que es denomina *lògit*. La variable *lògit* g, representa en una escala logarítmica la diferència entre les probabilitats de pertànyer a les dues poblacions, i en ser una funció lineal de les variables predictives permet ens facilita l'estimació i la interpretació del model.

Podem observar el gràfic de la funció $F(x)$ en la Figura 1.

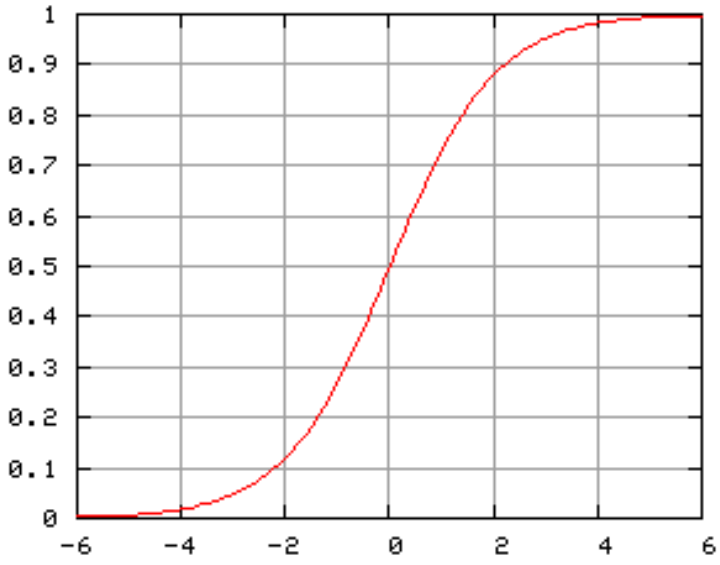


Figura 1: Funció Sigmoide

2.2.3 Estimació del model lògit

Suposem que tenim una mostra aleatòria de dades independents entre si i idènticament distribuïdes $(x_1, y_1), \ldots, (x_n, y_n)$ i k variables predictives. Com que la variable resposta y_i es distribueix com una Bernoulli, la funció de probabilitats per una resposta y_i qualsevol és:

$$f_\beta(y_i) = p_i^{y_i}(1-p_i)^{1-y_i} \text{ amb } p_i = \frac{1}{1+e^{-(\beta_0+\beta_1^t x)}} \text{ i } y_i \in \{0,1\}$$

Donat que les variables aleatòries són i.i.d. Tenen la mateixa funció de probabilitat, la funció de probabilitats conjunta de la mostra és:

$$f_\beta(y_i, \ldots, y_n) = f_\beta(y_i) \cdot \ldots \cdot f_\beta(y_n) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{2.5}$$

Aquesta funció, com a funció del vector de paràmetres $\beta = (\beta_0, \ldots, \beta_k)$ es coneix com a funció de versemblança i és la funció a maximitzar. Prenent logaritmes a l'expressió 2.5 obtenim la següent transformació:

$$\log(f_\beta(y_i, \ldots, y_n)) = \log(\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}) = \sum_{i=1}^{n} \log(p_i^{y_i}(1-p_i)^{1-y_i}) =$$

$$= \sum_{i=1}^{n} (\log(p_i^{y_i}) + \log((1-p_i)^{1-y_i})) = \sum_{i=1}^{n} (y_i \log(p_i) + (1-y_i)\log(1-p_i)) =$$

$$= \sum_{i=1}^{n} (y_i \log(p_i) - y_i \log(1-p_i) + \log(1-p_i)) = \sum_{i=1}^{n} (y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i))$$

Tenint en compte que $1 - p_i = \frac{1}{1+e^{\beta^t x_i}}$ on x_i és el vector de variables predictives per l'observació i-èssima. És a dir, $x_i^t = (1, x_{i1}, \ldots, x_{ik})$ i utilitzant la funció lògit 2.4 arribem a la següent funció:

$$L(\beta) = \sum_{i=1}^{n} (y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)) = \sum_{i=1}^{n} \left(y_i \beta^t x_i + \log\left(\frac{1}{1+e^{\beta^t x_i}}\right)\right) =$$

$$= \sum_{i=1}^{n} (y_i \beta^t x_i + \log(1+e^{\beta^t x_i})) = \sum_{i=1}^{n} (y_i \beta^t x_i) - \sum_{i=1}^{n} log(1+e^{\beta^t x_i})$$

Per tant, la funció a maximitzar és $L(\beta) = \sum_{i=1}^{n} (y_i \beta^t x_i) - \sum_{i=1}^{n} log(1+e^{\beta^t x_i})$ que és continua i diferenciable en tot $\mathbb{R}^{k+1}$. La condició necessària per trobar un extrem relatiu és veure en quins punts s'anul·la la derivada, per tant, per obtenir els estimadors de màxima versemblança (MV), derivem $L(\beta)$ com a vector columna.

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \left(\frac{e^{\beta^t x_i}}{1+e^{\beta^t x_i}}\right) = \sum_{i=1}^{n} x_i \left(y_i - \frac{1}{1+e^{-\beta^t x_i}}\right) = 0$$

Anomenem $\hat{\beta}$ al vector de paràmetres que satisfan el sistema d'equacions anterior, ens queden les següents igualtats:

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} x_i \left(\frac{1}{1+e^{-\hat{\beta}^t x_i}}\right) = \sum_{i=1}^{n} x_i \hat{p}_i \Rightarrow \sum_{i=1}^{n} x_i (y_i - \hat{p}_i) = \sum_{i=1}^{n} x_i (y_i - \hat{y}_i) = 0 \tag{2.6}$$

Aquestes equacions estableixen que el producte escalar dels valors observats per les variables explicatives ha de ser igual al producte escalar dels valors previstos per les variables explicatives. També ens diu que el vector de variables explicatives ha de ser ortogonal als residus del model $e_i = y_i - \hat{y}_i$.

Per obtenir el vector $\hat{\beta}_{MV}$ de paràmetres que maximitza la versemblança utilitzarem el mètode de Newton - Raphson. Desenvolupant $\frac{\partial L(\beta)}{\partial \beta}$ al voltant d'una condició inicial β_a es té:

$$\frac{\partial L(\beta)}{\partial \beta} = \frac{\partial L(\beta_a)}{\partial \beta} + \frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^t}(\beta - \beta_a) \tag{2.7}$$

Perquè el punt β_a sigui el màxim de versemblança, s'ha de complir $\frac{\partial L(\beta_a)}{\partial \beta} = 0$, imposant aquesta condició a 2.7 tenim:

$$\frac{\partial L(\beta)}{\partial \beta} = \frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^t}(\beta - \beta_a)$$

L'objectiu ara, és trobar analíticament β_a. Comencem multiplicant per la inversa de $\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^t}$ obtenint

$$\beta_a = \hat{\beta} + \left(-\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^t}\right)^{-1} \cdot \frac{\partial L(\beta)}{\partial \beta} \tag{2.8}$$

Aquesta expressió depèn de la matriu inversa de segones derivades. Per conèixer-la derivem $-\frac{\partial L(\beta)}{\partial \beta}$

$$-\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^t} = \sum_{i=1}^{n} x_i \left(\frac{x_i e^{\beta^t x_i}(1 + e^{\beta t x_i}) - x_i e^{\beta^t x_i} e^{\beta^t x_i}}{(1 + e^{\beta t x_i})^2}\right) = \sum_{i=1}^{k} x_i \left(\frac{x_i e^{\beta^t x_i} + \cancel{x_i e^{2\beta^t x_i}} - \cancel{x_i e^{2\beta^t x_i}}}{(1 + e^{\beta t x_i})^2}\right) =$$

$$= \sum_{i=1}^{n} x_i \left(\frac{x_i e^{\beta^t x_i}}{(1 + e^{\beta t x_i})^2}\right) = \sum_{i=1}^{n} x_i^t x_i \omega_i$$

on els coeficients w_i venen donats per:

$$w_i = \left(\frac{e^{\beta^t x_i}}{(1 + e^{\beta t x_i})^2}\right) = \left(\frac{1}{1 + e^{-\beta t x_i}}\right)\left(\frac{1}{1 + e^{\beta t x_i}}\right) = p_i(1 - p_i)$$

I les variàncies i covariàncies dels paràmetres estimats s'obtenen a partir de la inversa de la matriu de les segones derivades.

$$I^{-1}(\beta) = \frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^t}$$

Substituint ara les expressions avaluades en $\hat{\beta}$

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i(y_i - \hat{p_i}) \qquad -\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^t} = \sum_{i=1}^{n} x_i^t x_i \omega_i$$

a l'equació 2.8 tenim:

$$\beta_a = \hat{\beta} + \left(\sum_{i=1}^{n} x_i^t x_i \hat{w}_i\right)^{-1} \cdot \left(\sum_{i=1}^{n} x_i(y_i - \hat{p_i})\right)$$

Matricialment, podem escriure aquest algorisme com:

$$\beta_a = \hat{\beta} + (X^t \hat{W} X)^{-1} X^t (Y - \hat{Y})$$

$I(\hat{\beta}) = X^t \hat{W} X$ és una matriu on X té dimensió $n \times (k+1)$, $\hat{W}$ és una matriu diagonal $n \times n$ amb termes $\hat{p}_i(1-\hat{p}_i)$ i $\hat{Y}$ és el vector de valors esperats de Y amb dimensió $n \times 1$.

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \hat{W} = \begin{bmatrix} \hat{p}_1(1-\hat{p}_1) & 0 & \cdots & 0 \\ 0 & \hat{p}_2(1-\hat{p}_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \hat{p}_n(1-\hat{p}_n) \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Aquesta equació ens permet definir el següent mètode iteratiu:

$$\hat{\beta}_{h+1} = \hat{\beta}_h + (X^t \hat{W} X)^{-1} X^t (Y - \hat{Y})$$

que convergirà quan $\hat{\beta}_{h+1} \approx \hat{\beta}_h$, és a dir, quan la matriu de variables explicatives sigui ortogonal al vector residus ($X^t(Y - \hat{Y}) \approx 0$).

No obstant a tot l'explicat anteriorment, a l'estimació de màxima versemblança se li pot donar un altre enfocament. Maximitzar la versemblança es pot expressar com minimitzar una funció que mesura la desviància entre les dades i el model.

Definició 2.3. *Definim la desviància global del model com* $D(\beta) = -2L(\beta)$

Tenim

$$L(\beta) = \sum_{i=1}^{n} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

per tant,

$$D(\beta) = -2 \sum_{i=1}^{n} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \tag{2.9}$$

i com més gran sigui la funció $L(\beta)$, més gran és la concordança entre el valor dels paràmetres i les dades i menor és la desviància $D(\beta)$.

Definició 2.4. *Per a cada dada es defineix la seva desviància com* $d_i = -2(y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$.

i mesura l'ajust a l'observació (y_i, x_i).

Com que $0 < p_i < 1$ tenim que $log(p_i) < 0$ i $log(1 - p_i) < 0$. Per tant la desviància sempre és positiva. A més donat que $y_i \in \{0, 1\}$ la desviància es veu explicada únicament per un dels dos termes de l'expressió. Estudiem les casuístiques:

1. Si $y_i = 1$ el segon terme de la desviància és nul pel que queda $d_i = -2log(p_i)$. L'observació tindrà una desviància gran en cas que la probabilitat estimada p_i sigui petita, el que indica que aquesta observació està mal explicada pel model.

2. Si $y_i = 0$ el primer terme de la desviància és nul pel que queda $d_i = -2log(1 - p_i)$. La desviància serà gran en cas que p_i sigui gran. El vol dir que la probabilitat de pertànyer a la població P_0 és petita i per tant, el model ajusta malament l'observació.

Per tant la desviància serà 0 quan $y_i = p_i$.

2.2.4 Interpretació dels paràmetres

Un aspecte important és saber com interpretar els paràmetres β_i del model Lògit.
El paràmetre β_0 suposa l'ordenada en l'origen i el vector $\beta_1 = (\beta_1, \ldots, \beta_k)$ és el vector de pendents.

Definició 2.5. *Denominem odds ratio o ràtio de probabilitats als paràmetres e^{β_0} i e^{β_i}. Indiquen en quant es modifiquen les probabilitats per una unitat de canvi en les variables predictives.*

En efecte, per una observació i qualsevol tenim de l'equació 2.3

$$O_i = \frac{p_i}{1 - p_i} = e^{\beta_0} \prod_{j=1}^{k} e^{\beta_j x_j}$$

Suposem una altra observació $j \neq i$, amb tots els valors de les variables predictives iguals excepte per la variable h on $x_{ih} = x_{jh} + 1$.

El coeficient de la ràtio de probabilitats per aquestes dues observacions ve determinat per:

$$\frac{O_i}{O_j} = e^{\beta_h}$$

i ens indica com varia la ràtio de probabilitats per un increment en la variable x_h.

2.2.5 Contrasts d'hipòtsis

Si es vol contrastar si una variable o un conjunt de variables del model són estadísticament significatives podem construir un contrast de la raó de versemblança comparant el màxim de la funció de versemblança $L(\beta)$ pel model amb aquestes variables o sense. Suposem que $\beta = (\beta_1, \beta_2)$ on β_1 té dimensió $k - s$ i β_2 té dimensió s. Es desitja contrastar, per exemple:

$$\begin{cases} H_0 : \beta_2 & = 0 \\ H_1 : \beta_2 & \neq 0 \end{cases}$$

La ràtio de versemblança és:

$$\Lambda = \frac{\sup_{\beta_1} f_{\beta_1}(y_1, \ldots, y_n)}{\sup_{\beta} f_{\beta}(y_1, \ldots, y_n)}$$

Sigui $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ el vector de paràmetres estimats, prenent logaritmes a la ràtio de versemblança es té:

$$\log(\Lambda) = \log\left(\frac{f_{\hat{\beta}_1}(y_1,\ldots,y_n)}{f_{\hat{\beta}}(y_1,\ldots,y_n)}\right) = \log(f_{\hat{\beta}_1}(y_1,\ldots,y_n)) - \log(f_{\hat{\beta}}(y_1,\ldots,y_n)) = L(\hat{\beta}_1) - L(\hat{\beta})$$

Sota la hipòtesis nul·la H_0, la variable $-2\log(\Lambda)$ es distribueix asimptòticament com una χ^2_s amb s graus de llibertat quan $k \to \infty$.

Una manera equivalent d'analitzar el contrast és utilitzant les desviàncies.

1. Suposem $H_0 : \beta_2 = 0$ és certa, llavors la desviància és $D_{H_0}(\hat{\beta}) = -2L(\hat{\beta}_1)$.
2. Suposem ara que $H_1 : \beta_2 \neq 0$ és certa, llavors la desviància és $D_{H_1}(\hat{\beta}) = -2L(\hat{\beta}_1, \hat{\beta}_2)$.

Tenim que si la hipòtesi nul·la, la diferència de desviàncies $D_{H_o}(\hat{\beta}_1) - D_{H_1}(\hat{\beta}) = 2L(\hat{\beta}_1, \hat{\beta}_2) - 2L(\hat{\beta}_1)$ es distribueix asimptòticament com una χ^2_s amb s graus de llibertat.

Considerem vector de paràmetres $\beta = (\beta_0, \ldots, \beta_k)$, per estudiar si un β_i de forma individual és estadísticament significatiu, contrastem el següent mitjançant l'estadístic de Wald.

$$\begin{cases} H_0 : \beta_i & = 0 \\ H_1 : \beta_i & \neq 0 \end{cases}$$

Definició 2.6. *Definim l'estadístic de Wald com* $\frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$ *on* $s(\hat{\beta}_i)$ *és la desviació típica del paràmetre* $\hat{\beta}_i$.

Sota la hipòtesi nul·la, $\frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \sim N(0,1)$ quan $k \to \infty$.

2.2.6 Mesura global de l'ajust

Podem definir una mesura global de l'ajust a través de la desviància del model amb tots els paràmetres respecte a la desviància del model que sol inclou el paràmetre constant β_0.

Definició 2.7. *Definim la mesura global de l'ajust com* $R^2 = 1 - \frac{D(\hat{\beta})}{D(\hat{\beta}_0)} = 1 - \frac{L(\hat{\beta})}{L(\hat{\beta}_0)}$

Tenim que, en el model on sol s'inclou la constant, la probabilitat p és constant en tota la mostra.
Si $card(P_1) = m$ $(y_i = 1)$ llavors $p = \frac{m}{n}$ i per l'expressió 2.9 es té:

$$D(\hat{\beta}_0) = -2\sum_{i=1}^{n}\left(\frac{m}{n}\log\left(\frac{m}{n}\right) + (1-\frac{m}{n})\log\left(1-\frac{m}{n}\right)\right) =$$

$$= -2\sum_{i=1}^{n}\left(\frac{m}{n}\log\left(\frac{m}{n}\right) + (1-\frac{m}{n})\log\left(\frac{n-m}{n}\right)\right) =$$

$$= -2\sum_{i=1}^{n}\left(\frac{m}{n}\log(m) - \cancel{\frac{m}{n}\log(n)} + \log(n-m) - \log(n) - \frac{m}{n}\log(n-m) + \cancel{\frac{m}{n}\log(n)}\right) =$$

$= -2m\log(m) - 2n(1-\frac{m}{n})\log(n-m) - 2n\log(n) = -2m\log(m) - 2(n-m)\log(n-m) - 2n\log(n)$

En el cas que el model s'ajusti perfectament a les observacions, és a dir, per a cada $y_i = 1$ tenim $\hat{p_i} = 1$ llavors $D(\hat{\beta}) = 0$ i $R^2 = 1$.
En cas que les variables explicatives siguin estadísticament no significatives, el model sol depèn de la constant β_0 i $D(\hat{\beta}) = D(\hat{\beta_0})$ i $R^2 = 0$.

2.3 Mètode CART

Un procediment alternatiu al mètode de classificació a través de la regressió logística és utilitzar els arbres de classificació. (Clasification and Regression Trees).
La informació de continuació, s'ha extret, en gran part de Breiman, L., Friedman, J., Olshen R., i Stone, C. (1984), James, G., Witten, D., Hastie, T., Tibshirani, R. (2013) i Peña, D. (2002).

2.3.1 Teòria d'arbres de classificació

Suposem que tenim una mostra d'entrenament amb n observacions $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ n variables explicatives, x_i i un vector Y indicant la classe de cada observació y_i.

Matricialment la mostra ens queda organitzada de la següent manera:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Per construir l'arbre, es realitzen una sèrie de divisions binàries de les dades en subconjunts els més homogenis possibles, segons diverses regles de decisió, fins a arribar a un arbre maximal on es reparteixen totes les observacions i que conté en cada fulla (node terminal) un nombre reduït de dades. El procediment és el següent:

Partim d'un node inicial, on s'ha escollit una variable explicativa, per exemple x_1 i un punt de tall c. Llavors es divideixen les dades segons si $x_1 \leq c$ d'aquells que $x_1 > c$. Del node inicial neixen dues branques amb un altre node per branca, a un hi arribem les observacions que verifiquen $x_1 \leq c$ i a l'altre node hi arriben les que verifiquen $x_1 > c$. En aquests dos nous nodes, es repeteix el procediment anterior, és a dir, s'escull una nova variable explicativa i un nou punt de tall i es torna a dividir la mostra en dues parts més homogènies. El procés s'acaba quan s'han classificat totes les observacions (o la majoria) correctament al seu grup.

Per tant, un arbre de decisió divideix recursivament l'espai d'observacions amb plans perpendiculars als eixos de coordenades. La construcció de l'arbre requereix les decisions següents:

1. La selecció de variables i els seus punts de tall.
2. Quan un node es considera terminal o quan es continua dividint.
3. L'assignació del grup als nodes terminals.

A continuació es mostra un exemple fictici d'arbre de classificació.

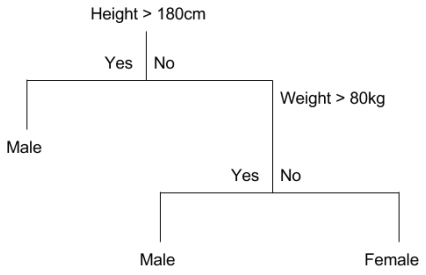


Figura 2: Exemple d'arbre de classificació

Tenim dues variables explicatives $x = (alçada, pes)$ i volem predir el sexe, és a dir, $y_i \in \{Masculí, Femení\}$. El procediment de classificació s'inicia en el node inicial, que hi trobem la variable explicativa alçada i el seu punt de tall són 180cm. En cas que la persona tingui una alçada superior a 180 cm se'l classifica al grup masculí. En cas contrari arribem a un nou node on la variable explicativa és pes i el punt de tall són 80 kg. Per tant si la persona mesura menys de 180cm i pesa més de 80 kg pertany al grup masculí. Si pesa menys de 80 kg, se'l classifica al grup femení.

Per decidir la variable a utilitzar per fer cada partició, en primer lloc es calcula la proporció d'observacions que passen pel node per a cada un dels grups.
Un criteri per decidir si partir un node o no, és utilitzar l'Índex d'impuresa de Gini o Entropia creuada que es defineix com:

Definició 2.8. *Suposem que tenim $t = 1, \ldots, T$ nodes i G grups, es defineix l'Índex de Gini com:*

$$I(t) = -\sum_{g=1}^{G} p(g|t) \log(p(g|t))$$

$p(g|t)$ representa la probabilitat que les observacions que arriben al node t pertanyin, a cadascuna de les classes.
l'Índex d'impuresa Gini és una mesura del la variància total entre els G grups. Aquestes mesures són positives i mesuren la desviància. Adquirixen un valor petit si tots els $p(g|t)$ estan a prop de 0 o de 1.
És a dir, per exemple, si $p(s|t) = 1$ i $p(k|t) = 0$ amb $s \neq k$ (totes les observacions que passen pel node t pertanyen al grup s), llavors la desviància del node t és $I(t) = 0$.

La variable utilitzada per realitzar una divisió de les observacions en un node, se selecciona minimitzant la impuresa resultant de la divisió. Considerem un conjunt de q preguntes del tipus: és $x_i < a$? $\forall i = 1, \ldots, n$ i $a \in \mathbb{R}$. Siguin p_S i p_K les proporcions de les observacions del node t que aniran als nodes resultants de respondre SI (node t_S) o NO (node t_K) a la pregunta. El canvi en l'entropia després de la pregunta q, serà la diferència entre l'entropia $I(t)$ i l'entropia després del node, que serà definida per $p_S I(t_S) + p_K I(t_K)$.

Per tant el canvi en entropia en la pregunta q és:

$$\Delta I(t, q) = I(t) - (p_S I(t_S) + p_K I(t_K)) \tag{2.10}$$

I es desitja escollir q per maximitzar el canvi d'entropia en cada node. És a dir, s'escull per a cada node, de totes les preguntes aquella que maximitza 2.10, ja que és la pregunta que proporciona els dos grups més homogenis possibles com a resultat de la divisió.

A efectes d'evitar el sobre ajustament ($overfitting$) sobre les dades en els quals s'entrena el model, s'utilitza un algorisme d'aglomeració ($pruning$) que reuneix diverses fulles i branques de l'arbre, obtenint un arbre que tingui major poder de predicció.

Un problema important dels arbres és la seva alta variància. Sovint, un canvi petit en les dades dóna lloc a un resultat molt diferent de les divisions. La raó principal d'aquesta inestabilitat és la naturalesa jeràrquica del procés: l'efecte d'un error en la divisió superior es propaga en totes les seves posteriors divisions.

2.4 Mètode Agregació de models

La següent informació es segueix de Breiman, L. (1996) i Bourel, M. (2012).
Una manera de pal·liar el problema anterior és fer una agregació de models. Els mètodes d'agregació de models consisteixen en, a partir d'un conjunt de dades $\mathcal{L}$, construir varis predictors i posteriorment, combina'ls d'alguna forma per obtenir un predictor més estable que disminueixi la variància.

En un problema de classificació de diverses classes, i seguint el conjunt d'observacions $\mathcal{L}$, podem construir M classificadors $g_1, g_2, \ldots, g_M$ per predir la variable dependent y_i i combinar-los amb la finalitat d'aconseguir un model més consistent que si utilitzem únicament model predictiu.
Es pot definir un classificador agregat, a través del vot per majoria dels M classificadors intermedis, és a dir:

$$f_A(x) = Arg \max_{k\in\{1,\ldots,K\}} (\#\colon \{m\colon g_m(x) = k\}) = Arg \max_{k\in\{1,\ldots,K\}} \left(\sum_{m=1}^{M} \mathbb{1}_{\{g_m(x)=k\}} \right)$$

A més, podem combinar aquests classificadors intermedis mitjançant un vot ponderat. Si $\{\alpha_1, \ldots, \alpha_M\}$ tals que $\alpha_i \in \mathbb{R}\ \forall i \in \{1, \ldots, M\}$, podem definir el classificador agregat com:

$$f_A(x) = Arg \max_{k\in\{1,\ldots,K\}} \left(\sum_{m=1}^{M} \alpha_m \cdot \mathbb{1}_{\{g_m(x)=k\}} \right)$$

És a dir, construïm un model agregat que prediu la classe que fa màxima la suma ponderada dels classificadors $g_1, g_2, \ldots, g_M$.

Si la variable resposta és contínua numèrica, un predictor agregat es podria definir com la mitja de les prediccions fetes per $g_1, g_2, \ldots, g_M$ sobre la mostra $\mathcal{L}$. També es pot considerar una ponderació. En aquest cas, el predictor agregat seria una combinació lineal de la forma

$$f_A(x) = \sum_{m=1}^{M} \alpha_m g_m(x)$$

Comprarem l'error mitjà del mètode CART i l'error agregat. El fonament teòric es basa que l'error mitjà del mètode CART que s'obté sobre el conjunt d'entrament $\mathcal{L}$ és més gran o igual que l'error obtingut per l'estimador Agregat. En efecte, siguin (x, y) de $\mathcal{L}$ independents que segueixen la mateixa probabilitat de distribució P. Per exemple, si y és numèrica i $f_{\mathcal{L}}(x)$ el seu predictor. Llavors el predictor agregat és:

$$f_A(x) = E_{\mathcal{L}} f_{\mathcal{L}}(x)$$

l'error mitjà de predicció de $f_{\mathcal{L}}(x)$ com:

$$e = E_{\mathcal{L}} E_{x,y}(y - f_{\mathcal{L}}(x))^2$$

i l'error de predicció agregat del predictor $f_A(x)$ com:

$$e_A = E_{x,y}(y - f_A(x))^2$$

Ara provem que la desigualtat $(Ez)^2 \leq Ez^2$ és certa:

$$Ez = \mu \quad \text{i} \quad E[z - \mu]^2 = \sigma^2$$

$$Ez^2 = E[z - \mu + \mu]^2 = E[z - \mu]^2 - 2\mu \cancel{E(z - \mu)} + E\mu^2 = \sigma^2 + \mu^2$$

I aplicant el resultat anterior és té:

$$e = E_{x,y} y^2 - 2E_{x,y} y E_{\mathcal{L}} f_{\mathcal{L}} + E_{x,y} E_{\mathcal{L}} f_{\mathcal{L}}(x)^2 \geq E_{x,y}(y - f_A(x))^2 = e_A$$

ja que:

$$E_{x,y}[E_{\mathcal{L}} f_{\mathcal{L}}(x)^2] \geq E_{x,y}[E_{\mathcal{L}} f_{\mathcal{L}}(x)]^2$$

2.4.1 Bagging

La major part del següent contingut s'ha seguit de Breiman, L. (1996) i Hastie, T., Tibshirani, R., Friedman, J. (2009). La tècnica Bagging, és un mètode d'agregació de models que es basa en el vot majoritari o la mitja. És un procediment per reduir la variància d'un mètode d'aprenentatge estadístic.

Seguim amb la mostra d'entrenament amb n observacions $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ on els y_i són la variable respostes de classificació o numèriques. El predictor d'aquest conjunt d'entrenament és $f_{\mathcal{L}}(x)$ on els inputs x prediuen y a través de $f_{\mathcal{L}}(x)$.

Suposem ara que donats els conjunts d'entrenament $\{\mathcal{L}_k\}_{k \in I}$ on $I = \{1, \ldots, K\}$. Cada conjunt té n observacions independents i segueixen la mateixa funció de distribució dels elements de $\mathcal{L}$. L'objectiu és utilitzar $\{\mathcal{L}_k\}_{k \in I}$ per obtenir un millor predictor que $f_{\mathcal{L}}(x)$. Amb la restricció d'utilitzar el predictor $f_{\mathcal{L}}(x)$ en els altres conjunts d'obvservació $\{f_{\mathcal{L}_k}(x)\}_{k \in I}$.

Definim $z_i = (x_i, y_i)$ amb mitja $\bar{z}$ i variància σ^2. Per tant la variància de $\bar{z}$ és:

$$Var[\bar{z}] = \frac{1}{n^2} \sum_{i=1}^{n} Var[z_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Si y_i és numèrica, una manera natural de reduir la variància és substituir $f_{\mathcal{L}}(x)$ per la mitjana dels $f_{\mathcal{L}_k}(x)$ sobre k, és a dir per:

$$f_A(x) = \frac{1}{K}\sum_{k=1}^{K} f_{\mathcal{L}_k}(x)$$

Si $f_{\mathcal{L}}(x)$ prediu una classe $j \in \{1, \ldots, J\}$, un mètode d'agregar els $\{f_{\mathcal{L}_k}(x)\}_{k\in I}$ és per votació:

$$f_A(x) = Arg \max_{j\in\{1,\ldots,J\}} (\#\{k\colon f_{\mathcal{L}_k}(x) = j\}$$

Però aquest procediment no és pràctic perquè, generalment no disposem de múltiples conjunts d'entrenament. Per tant s'utilitza, la tècnica de Bagging que segueix els següents passos:

1. Dividir el conjunt d'entrenament en diferents subconjunts obtenint com a resultat mostres aleatòries amb les següents característiques:
   - Mostra uniforme (mostra *bootstrap*) (mateixa quantitat d'observacions en cada subconjunt).
   - Mostres amb reposició (les observacions poden repetir-se en el mateix subconjunt de dades).
2. Es crea un model predictiu per a cada subconjunt, obtenint models diferents. És a dir, es crea un arbre amb el mètode CART on el qual és maximal (no s'aplica *pruning*).
3. Per últim, es construeix un únic model que és resultat de fer la mitjana de tots els models de cada subconjunt.

Estudiem del nombre d'observacions que tindrà cada mostra *bootstrap*. Per a fer-ho comencem definint el nostre espai de probabilitats com $(\mathcal{L}, \mathcal{A}, P)$. Com $\mathcal{L}$ és un conjunt finit de n observacions independents tenim que la $\sigma - àlgebra$ és $\mathcal{A} = 2^n$. L'esdeveniment d'interès E ocorre quan és selecciona alguna observació $z_i = (x_i, y_i)$ per una mostra B.

Per tant, definim una mesura de probabilitat com $P(E) = P(z_i \in B)$. Podem pensar construir la nostra mostra com un experiment de n assajos. Cada prova selecciona una de les nostres observacions de manera uniforme a l'atzar amb reemplaçament, pel que inclourà z_i amb probabilitat $P(E) = \frac{|E|}{|\mathcal{L}|} = \frac{1}{n}$ o excloure z_i amb probabilitat $P(E^c) = 1 - \frac{1}{n}$.

Ara l'espai de probabilitats $(\mathcal{L}, \mathcal{A}, P)$ està completament definit. L'experiment que estem realitzant consta de n intents amb reposició, pel que la probabilitat que z_i s'ometi de tots ells és:

$$\left(\bigcup_{i=1}^{n} P(E)\right)^c = \bigcap_{i=1}^{n} P(E^c) = \left(1 - \frac{1}{n}\right)^n$$

I per tant, la probabilitat que z_i s'inclogui almenys una vegada és $1 - \left(1 - \frac{1}{n}\right)^n$.
Utilitzant la definició del nombre e, $e = \lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n$ tenim:

$$\lim_{n\to\infty} 1 - \left(1 - \frac{1}{n}\right)^n = 1 - \lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^n = 1 - \lim_{n\to\infty}\left\{\left(1 + \frac{1}{-n}\right)^{-n}\right\}^{\frac{1}{-n}\cdot n} = 1 - \frac{1}{e}$$

Cada element $z_i \in \mathcal{L}$ té una probabilitat de sortir en la mostra, si n és gran, de $1 - \frac{1}{e} \approx \frac{2}{3}$. Per tant, cada mostra conté aproximadament $n' = \frac{2}{3}n$ observacions. El $\frac{1}{3}$ restant d'observacions s'utilitzen per al test d'error de model. (*Out of Bag Error*).

Siguin $\{\mathcal{L}^b\}$ amb $b \in \{1, \ldots, B\}$ les mostres de $\{\mathcal{L}\}$. Entrenem els B models $f_{\mathcal{L}^b}(x)$ i per obtenir el model predictiu final, es fa la mitjana de totes les B prediccions obtenint:

$$f_{Bag}(x) = \frac{1}{B} \sum_{b=1}^{B} f_{\mathcal{L}^b}(x)$$

No obstant, aquest mètode s'utilitza per predir un resultat quantiatiu.

Si la variable de resposta és qualitativa amb J classes, i estimem els B models $\{f_{\mathcal{L}^b}(x)\}$ entrenats en $\{\mathcal{L}^b\}$ llavors el model final és:

$$f_{Bag}(x) = Arg \max_{j \in \{1,\ldots,J\}} (\#\{j \colon f_{\mathcal{L}^b}(x) = j\}$$

Visualment, podem trobar a la figura 21 de l'Annex com funciona el mètode Bagging.

Out-of-Bag Error

La mostra *Out of Bag Error* (*OOB*) és una mesura de l'error aplicada als mètodes que utilitzen la tècnica *bootstrapping*. Donada la naturalesa de procés Bagging, resulta possible estimar de forma directa el test d'error sense necessitat de recórrer a validació creuada ($cross - validation$).

Hem vist que s'utilitzen aproximadament $\frac{2}{3}$ d'observacions de $\mathcal{L}$ per les mostres *bootstrap*. Per tant, el OOB error, representa l'error de predicció en el conjunt d'observacions que no formen part de les mostres *bootstrap*, és a dir, que han quedat "fora de la bossa".

Si per cada arbre ajustat en el procés Bagging es registren les observacions utilitzades, es pot predir la resposta de l'observació j-èssima fent ús d'aquells arbres en què aquesta observació ha sigut exclosa (OOB). Aquest procediment produirà al voltat de $\frac{B}{3}$ prediccions per la j-èssima observació. Per obtenir una predicció única, podem fer la mitjana de les respostes en cas que sigui quantitativa o sotmetre-la a votació si és qualitativa.

Seguint aquest procés, es poden obtenir les prediccions per les n observacions i calcular el $OOB - mean\ square\ error$ per regressió o el $OOB - classification\ error$ per als arbres de classificació.

2.4.2 Random Forest

La informació referent a *RandomForest* s'ha extret principalment de Breiman, L. (1996) i Hastie, T., Tibshirani, R., Friedman, J. (2009). El mètode Ranom forest, és una modificació més complexa del mètode Bagging que aconsegueix millors resultats gràcies al fet que descorrelaciona els arbres generats duran el procés. Recordant el procés del mètode Bagging, es basa en el fet que fent la mitjana o votació d'un conjunt de prediccions, s'aconsegueix reduir la variància. Això és cert sempre que les prediccions de cada mostra

bootstrap no estiguin correlacionades. Si la correlació és alta, la reducció de la variància que es pot aconseguir serà petita.

Suposem que en el conjunt d'observacions $\mathcal{L}$ hi ha un predictor molt influent respecte la resta. En aquest context, la majoria d'arbres creats al fer Bagging estaran dominats pel mateix predictor i seran semblants entre ells. A conseqüència de l'alta correlació, el mètode Bagging no disminuirà la variància com s'esperava i per tant no hi haurà una millora en el model agregat.

Per tant, l'algorisme *Random Forest* combina les tècniques *CART* i *Bagging*. Igual que a *Bagging*, se sortegen B mostres *bootstrap* $\{\mathcal{L}^b\}$ del conjunt de dades $\mathcal{L}$, i sobre les mateixes es construeixen B arbres pels quals, en cada node, s'escull la millor subdivisió feta per un conjunt de variables predictives seleccionades aleatòriament. En el cas de regressió es sol escollir $m = \frac{k}{3}$ variables explicatives a l'atzar, en el cas de classificació, $m = \sqrt{k}$, en cas que $m = k$ no hi ha diferència amb *Bagging*. Els arbres són maximals, és a dir, no s'aplica *pruning* i com a *Bagging*, en el cas de classificació, la predicció d'una observació feta per *Random Forest* és la classe més votada dels B arbres i en cas de regressió es fa la mitjana dels valors assignats.

Afegir que, ni en *Bagging* ni *Ranfom Forest* el fet d'augmentar el nombre de mostres *bootstrap* no provoca *overfitting* per tant, a partir d'un cert nombre d'arbres, hi ha una convergència de l'error generalitzat.

En aquest context, passem a veure què ocorre amb la variància quan augmentem el nombre d'arbres. Suposem que cada arbre té variància igual σ^2 i correlació ρ.
La variància de l'estimador de *Random Forest* serà:

$$Var\left(\frac{1}{B}\sum_{b=1}^{B} f_{\mathcal{L}^b}(x)\right) = \frac{1}{B^2}\sum_{b=1}^{B}\sum_{l=1}^{B} Cov(f_{\mathcal{L}^b}(x), f_{\mathcal{L}^l}(x)) =$$

$$= \frac{1}{B^2}\sum_{b=1}^{B}\sum_{l\neq b}^{B}\Big(Cov(f_{\mathcal{L}^b}(x), f_{\mathcal{L}^l}(x)) + Var(\mathcal{L}^b(x)\Big) =$$

$$= \frac{1}{B^2}\sum_{b=1} B((B-1)\rho\sigma^2 + \sigma^2) = \frac{B\rho\sigma^2(B-1) + B\sigma^2}{B^2} = \rho\sigma^2 + \frac{\sigma^2(1-\rho)}{B}$$

Per tant, si $B \to \infty$, la variància depèn únicament de $\rho\sigma^2$. I per tant, fent divisions aleatòries dels predictors en cada arbre podem disminuir la correlació entre ells i conseqüentment, es reduirà la variància del model.

Importància de les variables

En aplicacions de modelització, les variables de predicció quasi mai són igual de rellevants. Sovint, sol poques variables acaben tenint una influència en la resposta tot i haver fet anteriorment una anàlisi univariant analitzant la correlació del predictor amb la variable resposta.
El mecanisme de construcció de *Random Forest* i *Bagging* ens permet establir un barem de la importància de cada variable en la predicció final. El procediment és el següent.

Es cultiva l'arbre b-èssim $\mathcal{L}^b(x)$ i per aquest, s'agafa la mostra OOB i es registra la precisió de la predicció. A continuació s'agafa una variable de l'arbre $\mathcal{L}^b(x)$, es permuta el seu valor aleatòriament en la mostra OOB i es calcula la precisió de nou. La precisió de cada permutació hauria de ser pitjor que per la mostra OOB original, ja que el mateix arbre s'encarrega de fer créixer l'arbre de la forma més òptima. Aquest procés es realitza per a totes les variables i es calcula la mitja. Així les variables menys importants haurien d'alterar menys la diferència entre l'error de la mostra OOB i l'error de la mostra OOB permutada, que les variables més importants. Podem trobar més detall visitant Louppe *et al*. (2014).

2.4.3 Boosting

La informació d'aquest capítol s'ha seguit, majoritàriament de Alfaro *et al* (2013), Yoav F., Schapire, R.E. (1997) i Hastie, T., Tibshirani, R., Friedman, J. (2009). La tècnica *Boosting* és una de les idees d'aprenentatge més potents en els últims vint anys. Originalment va ser dissenyat per problemes de classificació però és possible estendre'l a problemes de regressió.

La idea que hi ha darrere el *Boosting* és ajustar, de forma seqüencial, múltiples prediccions dèbils (*weak learner*), en el nostre cas arbres senzills amb una o poques divisions que prediuen sol lleugerament millor que l'esperat de forma aleatòria). Cada nou model utilitza informació del model anterior per aprendre dels seus errors, millorant iteració a iteració.

A diferència del mètode *Bagging*, el *boosting* no fa ús del *bootstrapping*, pel que cada arbre depèn dels arbres previs.

Boosting i el Model Agregat

El mètode *Boosting* s'ajusta a un model agregat a través d'un conjunt de funcions base. En el nostre cas, les funcions base són els classificadors (arbres) de cada iteració. Aquest model agregat és de la forma

$$f(x) = \sum_{i=1}^{M} \beta_m b(x; \gamma_m)$$

On els β_m són els coeficients del desenvolupament i $b(x; \gamma_m) \in \mathbb{R}$ normalment són funcions senzilles que depenen de x i un conjunt de paràmetres γ. Aquests paràmetres, en el cas dels arbres s'utilitzen per determinar les variables explicatives i els punts de tall de cada node.

Generalment, aquests models s'ajusten minimitzant una funció de pèrdua mitjana sobre les dades d'entrenament, com l'error quadràtic o una funció de pèrdua basada en versemblança.

$$\min_{\beta_m, \gamma_m} \sum_{i=1}^{n} L\left(y_i, f(x_i)\right) = \min_{\beta_m, \gamma_m} \sum_{i=1}^{n} L\left(y_i, \sum_{m=1}^{M} \beta_m b(x_i; \gamma_m)\right) \tag{2.11}$$

No obstant, per moltes funcions de pèrdua o funcions bàsiques, minimitzar-les requereix tècniques d'optimització numèrica computacionalment intensives. Sovint es pot trobar

una alternativa més senzilla quan és factible resoldre el subproblema d'ajustar una funció bàsica, és a dir

$$\min_{\beta,\gamma} \sum_{i=1}^{n} L\left(y_i, \beta b(x_i;\gamma)\right)$$

Mitjançant aquest mètode es pot aproximar la solució de 2.11 afegint, progressivament noves funcions base a l'expansió sense ajustar els paràmetres i els coeficients dels models que ja s'han agregat. L'algorisme segueix el següent procediment:

1. S'inicialitza $f_0(x) = 0$
2. Desde $m = 1$ fins M

   a) Es calcula

   $$(\beta_m, \gamma_m) = Arg \min_{\beta,\gamma} \sum_{i=1}^{n} L\left(y_i, f_{m-1}(x_i) + \beta b(x_i;\gamma)\right) \tag{2.12}$$

   b) Es defineix

   $$f_m(x) = f_{m-1}(x) + \beta_m b(x;\gamma_m)$$

Comencem presentant l'algorisme més popular basat en el mètode *Boosting*.

AdaBoost

Suposem el conjunt d'entrenament amb n observacions $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ que segueixen una funció de distribució F desconeguda però fixa en $X \times Y$. Suposem el cas en què la variable resposta té dues classes. $Y = \{-1, 1\}$ i com sempre, l'objectiu és predir la classe y en funció dels inputs x. Utilitzem *boosting* per buscar una predicció la qual sigui consistent amb la major part de la mostra. És a dir $h(x_i) = y_i$ per la majoria $1 \leq i \leq n$. La ràtio d'error de la predicció en el conjunt d'entrenament és definit per

$$err = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i \neq h(x_i)\}}$$

I la ràtio de l'error esperat en futures prediccions és

$$E_{x,y} \mathbb{1}_{\{y \neq h(x)\}}$$

Com ja hem comentat, el propòsit de *Boosting* és aplicar progressivament l'algorisme dels classificadors dèbils per versions modificades de les dades, obtenint així una seqüència de classificadors dèbils $h_m(x)$ amb $m \in \{1, \ldots, M\}$.

La hipòtesi final ve donada per una combinació lineal ponderada de tots els classificadors dèbils que s'han produït anteriorment

$$h(x) = sign\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right)$$

on els $\alpha_1, \ldots, \alpha_n$ són els coeficients de cada classificador dèbil. Posteriorment s'explicaran com es calculen. La funció d'aquests pesos és donar més importància als classificadors

amb menys taxa d'error en la seqüència.

Anteriorment, s'ha dit "versions modificades de les dades". Això significa que en cada pas *boosting* s'apliquen els pesos $\omega_1, \ldots, \omega_n$ en cada observació d'aprenentatge (x_i, y_i) $0 \leq i \leq n$. Inicialment s'aplica una distribució uniforme en cada observació, és a dir,

$$\omega_i = \frac{1}{n} \qquad 0 \leq i \leq n$$

En cada iteració posterior $(m = 2, 3, \ldots, M)$ els pesos son modificats individualment en cada observació (x_i, y_i) i s'aplica l'algoritme de classificació a aquestes noves observacions ponderades. La lògica d'aquest procediment és la següent: en la iteració m, les observacions que hagin estat mal classificades pel classificador $h_{m-1}(x)$ en el pas $m-1$, es veuran afectades per un increment de les seves ponderacions. En canvi, les observacions ben classificades en la iteració anterior, la seva ponderació disminuirà.

És a dir, en la iteració m tindran més pes aquelles observacions que hagin estat mal classificades pel classificador de l'etapa $m-1$. D'aquesta manera cada classificador és forçat a tenir-les "més en compte". El predictor que s'obté al final resulta d'un vot ponderat majoritari o d'una mitja ponderada dels predictors en cada etapa. Al focalitzar-se sobre les observacions mal classificades, l'error empíric disminueix ràpidament. Es prova que l'error generalitzat també disminueix, per tant *Adaboost* és un algoritme amb molt bon rendiment.

Presentem l'agorisme *Adaboost* resumidament:

1. Inicialitzem el pes de les observacions: $\omega_i = \frac{1}{n} \qquad 0 \leq i \leq n$
2. Desde $m = 1$ fins M
   a) Ajustem el classificador $h_m(x)$ al conjunt d'entrenament $\mathcal{L}$ utilitzant els pesos ω_i
   b) Calculem l'error err_m en funció dels pesos ω_i i del classificador $h_m(x)$

$$err_m = \frac{\sum_{i=1}^{n} \omega_i \cdot \mathbb{1}_{\{y_i \neq h_m(x_i)\}}}{\sum_{i=1}^{n} \omega_i}$$

   c) Calculem el coeficient corresponent al classificador $h_m(x)$

$$\alpha_m = \log\left(\frac{1 - err_m}{err_m}\right)$$

   d) Actualitzem els pesos ω_i

$$\omega_i \leftarrow \omega_i \cdot e^{\alpha_m \cdot \mathbb{1}_{\{y_i \neq h_m(x_i)\}}}, \qquad 0 \leq i \leq n$$

3. Classificador final

$$h(x) = sign\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right)$$

Com ja s'ha dit al principi del capítol, el *weak learner* prediu lleugerament millor que la predicció esperada de forma aleatòria. Com que la variable resposta té dues classes, podem concloure que cada classificador dèbil té un error màxim lleugerament inferior a 0.5. Per tant,

$$1 - err_m > err_m \quad \Rightarrow \quad \frac{1 - err_m}{err_m} > 1 \quad \Rightarrow \quad \alpha_m = \log\left(\frac{1 - err_m}{err_m}\right) > 0$$

I tenim que α_m és positiu. Les observacions mal classificades queden multiplicades per $e^{\alpha_m} > 1$ augmentant el seu pes en la següent iteració.

Importància de les variables

Com ja hem comentat en el mètode *Random Forest* és interessant veure com influeix cada variable en l'entrenament de model. La metodologia que segueix la tècnica *Adaboost* per determinar la importància de les variables, tal i com s'explica en Natekin A. i Knoll A. (2013), es defineix la influència de la variable j-èssima en un sol arbre b i considerem que l'arbre té L nodes. Aleshores la mesura de la importància de la variable p ve definida per:

$$Imp_j(b) = \sum_{i=1}^{L} I_i^2 \cdot \mathbf{1}_{\{V_i = j\}}$$

V_i suposa la variable que divideix el i-èssim node. Per tant, aquesta mesura es basa en el nombre de vegades que es selecciona una variable per dividir un node. El terme I_i^2 suposa la millora empírica al quadrat de l'error de classificació en cada node. Finalment, la influència de cada variable en tots els arbres la podem calcular com:

$$Imp_p = \frac{1}{B} \sum_{b=1}^{B} Imp_j(b_i)$$

Funció de pèrdua exponencial

Veiem que l'algorisme *Adaboost* és equivalent a un model agregat amb funció de pèrdua

$$L(y, f(x)) = e^{-yf(x)}$$

En l'algorisme Adaboost, les funcions base són els classificadors dèbils individuals $h_m(x) \in \{-1, 1\}$. Utilitzant la funció de pèrdua presentada i aplicant la recurrència 2.12 s'ha de resoldre

$$(\beta_m, h_m) = Arg \min_{\beta, h} \sum_{i=1}^{n} e^{-y_i(f_{m-1}(x_i) + \beta h(x_i))} = Arg \min_{\beta, h} \sum_{i=1}^{n} \omega_i^{(m)} e^{-y_i(\beta h(x_i))} \qquad (2.13)$$

on $\omega_i^{(m)} = e^{-y_i f_{m-1}(x_i)}$. Com que els $\omega_i^{(m)}$ no depenen ni de β ni de $h(x)$, els podem considerar com els pesos aplicats a les observacions. Aquest pes sol depèn de $f_{m-1}(x_i)$ per tant, s'anirà modificant en cada iterada.

La solució de l'expressió 2.13 pot ser obtinguda en dues etapes:

1. $\forall \beta > 0$ la solució de 2.13 per $h_m(x)$ ve donada per:

$$h_m(x) = Arg \min_{h} \sum_{i=1}^{n} \omega_i^{(m)} \cdot \mathbb{1}_{\{y_i \neq h(x_i)\}}$$

que és el classificador que minimitza la ràtio de l'error ponderat.

2. Per trobar la solució del paràmetre β escrivim l'expressió 2.13 com segueix:

$$\sum_{i=1}^{n} \omega_i^{(m)} e^{-y_i(\beta h(x_i))} = e^{-\beta} \cdot \sum_{y_i = h(x_i)} \omega_i^{(m)} + e^{\beta} \cdot \sum_{y_i \neq h(x_i)} \omega_i^{(m)}$$

Que a la vegada pot ser escrit com

$$(e^{\beta} - e^{-\beta}) \sum_{i=1}^{n} \omega_i^{(m)} \cdot \mathbb{1}_{\{y_i \neq h(x_i)\}} + e^{-\beta} \sum_{i=1}^{n} \omega_i^{(m)}$$

Ara se substitueix $h(x)$ per $h_m(x)$ que per 1) és el classificador que ens dóna la mínima taxa d'error. Suposem els pesos normalitzats,

$$(e^{\beta} - e^{-\beta}) \sum_{i=1}^{n} \omega_i^{(m)} \cdot \mathbb{1}_{\{y_i \neq h_m(x_i)\}} + e^{-\beta} \sum_{i=1}^{n} \omega_i^{(m)} = (e^{\beta} - e^{-\beta}) err_m + e^{-\beta} =$$

$$= e^{\beta} err_m + e^{-\beta}(1 - err_m)$$

Per tant, l'objectiu ara és trobar el valor òptim de β. Per a fer-ho considerem la funció

$$g(\beta) = \log(e^{\beta} err_m + e^{-\beta}(1 - err_m))$$

Fent la seva derivada i igualant a zero tenim:

$$g'(\beta) = \frac{e^{\beta} err_m - e^{-\beta}(1 - err_m)}{e^{\beta} err_m + e^{-\beta}(1 - err_m)} = 0 \quad \Leftrightarrow \quad e^{\beta} err_m - e^{-\beta}(1 - err_m) = 0 \quad \Leftrightarrow$$

$$\Leftrightarrow \quad e^{\beta} err_m = e^{-\beta}(1 - err_m) \quad \Leftrightarrow \quad \beta + \log(err_m) = -\beta + \log(1 - err_m) \quad \Leftrightarrow$$

$$\Leftrightarrow \quad 2\beta = \log(1 - err_m) - \log(err_m)$$

$$\Leftrightarrow \quad \beta^* = \frac{1}{2} \log\left(\frac{1 - err_m}{err_m}\right)$$

Un cop trobats la solució dels dos paràmetres, s'actualitza l'aproximació

$$f_m(x) = f_{m-1} + \beta_m h_m(x)$$

El que fa que els pesos per la pròxima iteració siguin:

$$\omega_i^{(m+1)} = e^{-y_i f_m(x_i)} = \omega_i^{(m)} \cdot e^{-\beta_m y_i h_m(x_i)}$$

I utilitzant que $-y_i h_m(x_i) = 2 \cdot \mathbb{1}_{\{y_i \neq h_m(x_i)\}} - 1$ tenim que

$$\omega_i^{(m+1)} = \omega_i^{(m)} \cdot e^{\alpha_m \cdot \mathbb{1}_{\{y_i \neq h_m(x_i)\}}} \cdot e^{-2\beta_m}$$

on $\alpha_m = 2\beta_m$. El factor $e^{-2\beta_m}$ és constant en cada iterada per tant, el podem obviar i aconseguim l'expressió definida en l'algorisme *Adaboost*. Amb això podem concloure que *Adaboost* minimitza una funció de pèrdua exponencial mitjançant un model agregat.

3 Modelització

En aquest capítol, aplicarem alguns dels mètodes vists anteriorment i els enfrontarem per extreure'n conclusions. Per a fer-ho, tal com hem explicat al capítol introductori, els passos a seguir seran els següents:

1. **Anàlisi estadístic de les variables explicatives**
2. **Estimació del model**
3. **Validació**

Totes les taules i figures són d'elaboració pròpia i tota la modelització s'ha fet amb el llenguatge de programació R. Podem trobar el detall del codi amb comentaris visitant: *https://gist.github.com/joanllano*

3.1 Anàlisi Estadístic

Per començar, se segmenta el conjunt d'observacions en la mostra d'entrenament i la mostra de validació.

Classe	Mostra	Mostra d'entrenament	Mostra de validació
bons	44.143	35.321	8.822
dolents	2.711	2.175	536
Total	**46.854**	**37.496**	**9.358**

Amb l'objectiu de veure com es distribueixen tots els factors, s'analitzen els percentils

$$0,\ 1\%,\ 10\%,\ 25\%,\ 50\%,\ 75\%,\ 90\%,\ 95\%,\ 100\%$$

Un cop vist com es distribueixen els factors, prosseguim a categoritzar tots els factors, és a dir, la funció *scorecard* :: *woebin*() ens dóna trams òptims (heterogenis entre ells) d'un mínim de 5% de població (*bins*) per a cada factor.
Podem trobar el detall de les distribucions a la figura 19 i del detall dels camps que es calculen mitjançant la funció *woebin*() en la taula 20 de l'Annex del treball.

En la secció introducció, trobem una explicació conceptual del WOE. La majoria de la informació prové de Majer, I. (2006), analitzem amb detall quins avantatges presenta aquest estadístic. Si el considerem com la contribució independent de cada variable al model final, el WOE ens permet:

- Detectar relacions lineals i no lineals.
- Classificar les variables en termes de poder predictiu univariant.
- Visualitzar les correlacions entre les variables predictives i la variable resposta.
- Comparar el poder predictiu de les variables contínues i discretes sense necessitar crear variables fictícies.

Algunes premisses que s'han de complir per fer ús del WOE són:

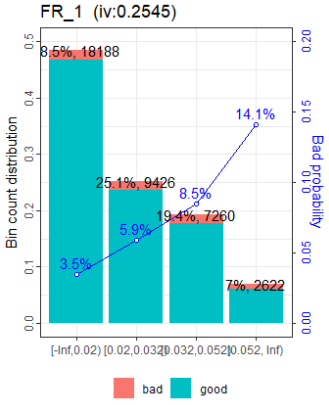
1. Cada categoria (bin) ha de tenir mínim un 5% de població.
2. Cada categoria ha de contenir observacions de les dues poblacions.
3. El WOE ha de ser diferent en cada categoria.

Per tant, les nostres variables "perdran"el seu valor real i els hi assignem el WOE de cada tram convertint-les totes en variables categòriques. No obstant, abans de realitzar la transformació, considerem els valors de l'estadístic IV:

IV	**Capacitat Predictiva de la variable**
< 0.02	No predictiva
$[0.02, 0.1)$	Dèbil
$[0.1, 0.3)$	Mitja
≥ 0.3	Forta

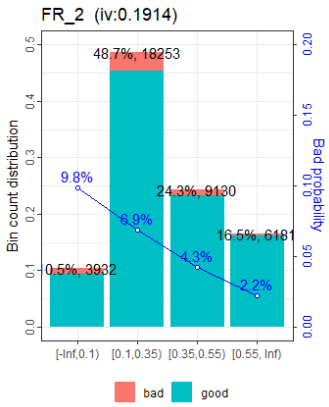
Taula 5: Capacitat predictiva IV

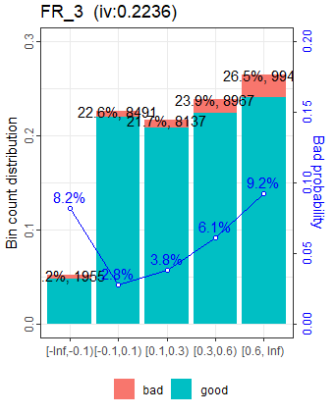
En el nostre estudi, seleccionem aquells factors de risc que presenten un IV més gran o igual que 0.1. A continuació mostrarem la distribució de bons i dolents en cada tram i la taxa de mora de cada factor amb $IV \geq 0.1$:



FR_1: Despeses financeres sobre passiu exigible: Per increments del valor del factor, l'empresa presenta una pitjor situació financera i per tant, la probabilitat de fer $Default$ augmenta, és a dir, la relació entre aquest factor i la variable dependent és positiva i per tant, té sentit que la taxa de mora sigui creixent.

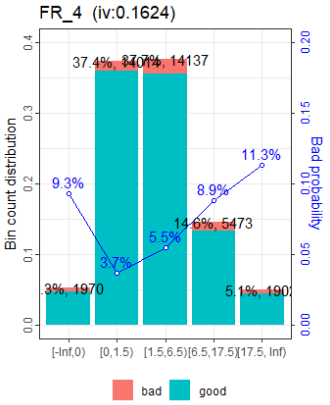
FR_2 Fons propis sobre actiu total: Per increments del valor del factor, significa que l'empresa presenta una millor situació financera i per tant, la probabilitat de fer $Default$ disminueix, és a dir, la relació entre aquest factor i la variable dependent és negativa i per tant, té sentit que la taxa de mora sigui decreixent.

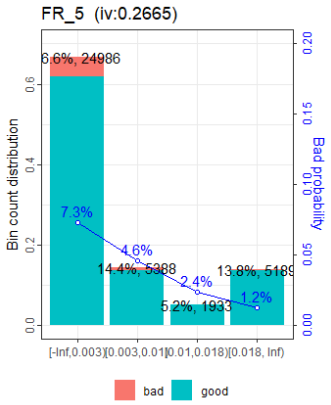




FR_3 Despeses financeres sobre BAIT (Benefici brut): Veiem que la TM té forma de U. No és d'estranyar, ja que en el primer tram, la ràtio pren valors negatius el que significa que el BAIT és negatiu. Un cop el BAIT és positiu, augments de la ràtio vol dir que les despeses financeres estan augmentant o el BAIT està disminuint per tant, el que s'espera és que la taxa de mora creixi.

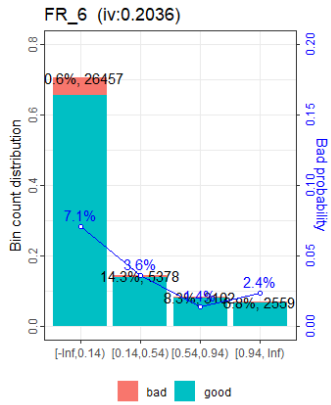
FR_4 Ràtio de cobertura al servei del deute: Es calcula com el passiu no exigible sobre els fons generats. La taxa de mora veiem que no és monòtona. Si els fons generats són negatius, la taxa de mora és més alta que si són positius. En canvi, si augmenta la ràtio és perquè el passiu no corrent augmenta o els fons generats disminueixen i per tant, la taxa de mora també creix.

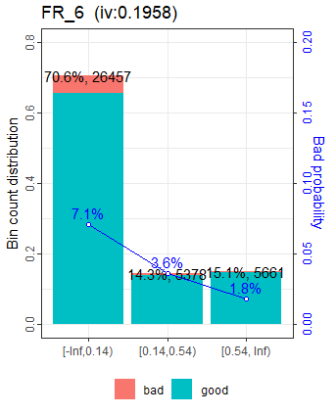




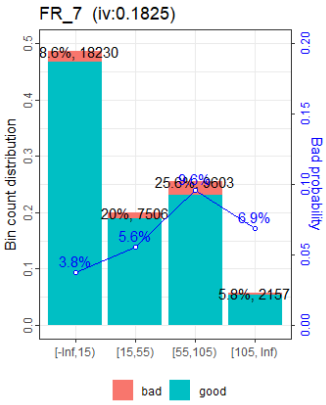
FR_5 Mitja últims 365D del saldo mitjà passiu vista sobre facturació: Per augments dels valors del factor vol dir o bé un augment de facturació o una disminució del saldo mitjà passiu, la situació financera de l'empresa millora i per tant la TM decreix.

FR_6 Disponible mitjà sobre el límit U60D: Segons la definició del factor, a més disponible, menys disposat i per tant menys TM. No obstant, veiem que l'últim tram presenta una TM que l'anterior.

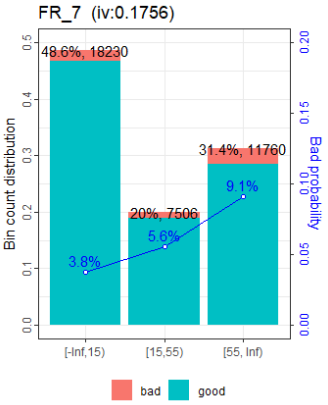




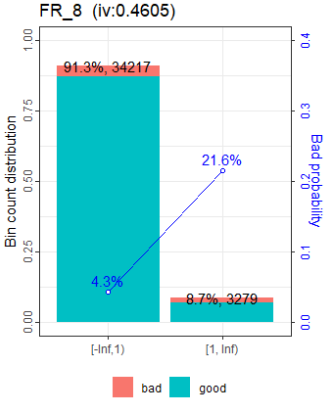
Per solucionar el problema anterior, el que fem és ajuntar els dos últims trams, el IV disminueix però segueix sent superior a 0.1 i aconseguim que la taxa de mora sigui monòtona i tingui sentit amb la definició del factor.



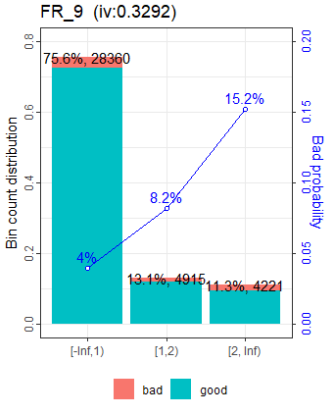
FR_7 Termini mitjà de finançament circulant U365D: Segons l'anàlisi financer, augments del termini mitjà de finançament pot provocar un desequilibri entre els passius i actius de llarg i a curt termini. Per tant, per augments del factor la taxa de mora ha d'augmentar. Veiem que l'últim tram presenta una TM superior al tram anterior.



Per solucionar el problema anterior, utilitzem el criteri seguit en el factor de risc anterior i el que fem és ajuntar els dos últims trams, el IV disminueix però segueix sent superior a 0.1 i aconseguim que la taxa de mora sigui monòtona i tingui sentit amb la definició del factor.



FR_8 Nombre d'impagaments 10D: Per la pròpia definició del factor, a més vegades que un client no abona la quota de capital pendent un cop passats deu dies del venciment de la data contractual, la probabilitat que acabi fent *default* augmenta i en conseqüència, la taxa de mora ha d'augmentar.



FR_9 Nombre d'excedits o descoberts: Diem que un client s'excedeix o es queda en descobert quan el saldo en la targeta de crèdit no és suficient per a fer front a un determinat pagament i l'entitat banca-

ria avança el capital necessari per fer front al deute. Per tant, per la pròpia definició, a més vegades que el client es queda en descobert o s'excedeix del límit de crèdit, la taxa de mora augmenta.

Un cop analitats i ajustat els trams dels factors, es transforma el factor. És a dir, a cada *Bin* del factor, li assignem el valor del WOE del tram en qüestió. Vegem-ho amb un exemple.

FR_1 Despeses financeres sobre passiu exigible

Factor	Bin	WOE	IV	TM
FR_1	[-Inf, 0.02)	0.5385	0.2545	3.47%
FR_1	[0.02, 0.032)	-0.0235	0.2545	5.93%
FR_1	[0.032, 0.052)	-0.4075	0.2545	8.47%
FR_1	[0.052, Inf)	-0.9814	0.2545	14.11%

Taula 6: WOE_FR_1

Pel factor de risc FR_1, El procés de transformació és el següent, en primer lloc mirem en quin *Bin* cau el valor del factor i un cop identificat, el factor perd el seu valor real i se li assigna el WOE del tram en què ha caigut el valor real. En la figura 20 de l'Annex podem trobar el detall de tots els factors.

A continuació mostrem les matrius de correlacions dels factors i dels factors transformats. Al factor transformat l'anomenem WOE_FR_i, les correlacions s'han obtingut mitjançant el *Coeficient de correlació de Spearman*.

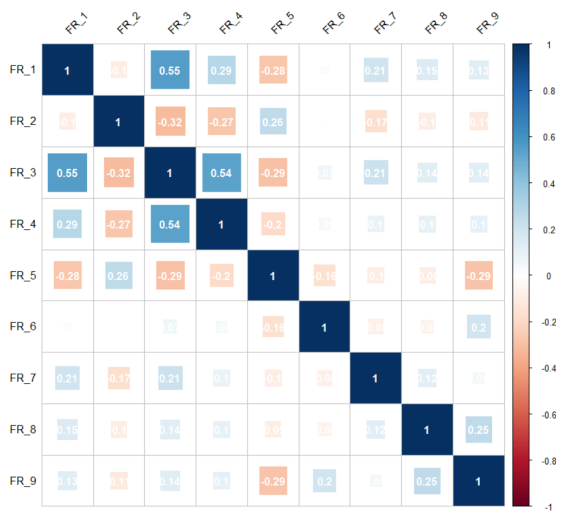


Figura 3: Matriu de Correlacions FR

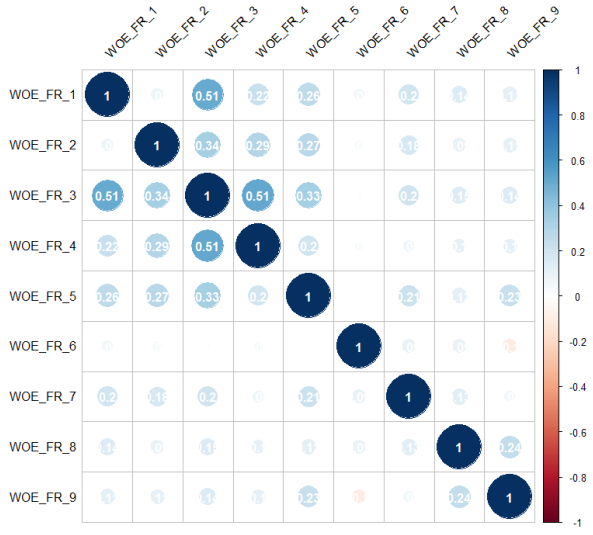


Figura 4: Matriu de Correlacions WOE

La matriu de l'esquerra representa les correlacions dels factors reals i la matriu de la dreta, és la matriu de correlacions dels factors transformats. Veiem que la intensitat del color i la mida del quadrat i la bola són proporcionals a la correlació entre els factors.

Considerem que dos factors estan correlacionats si presenten un coeficient de correlació igual o superior a 0.6. Veiem que, sense tenir en compte la diagonal principal de la matriu (lògicament cada factor està totalment correlacionat amb si mateix), la correlació màxima es dóna amb els factors FR_1 amb FR_3 i els factors FR_3 i FR_4 on el màxim d'aquestes correlacions és 0.55. Per altra banda els factors transformats que es correlacionen són els mateixos però amb els seus respectius WOE's, és a dir, WOE_FR_1 amb el factor WOE_FR_3 i els factors WOE_FR_3 i WOE_FR_4 on les dues correlacions són d'un 0.51. Per tant, sota el nostre criteri cap factor està correlacionat.

3.2 Estimació de models

Un cop analitzades les capacitats predictives i les correlacions dels factors, apliquem la teoria vista al capítol [2]. S'han estimat i comparat tres models en el conjunt d'entrenament de la taula 14 mitjançant els següents mètodes:

1. **Regressió Logística**

   Per entrenar el model, donada la dificultat d'interpretar la importància de les variables, s'opta per estimar el model utilitzant els factors transformats, ja que com hem explicat a l'inici del capítol tenen alguns avantatges. Per tant s'ha estimat el model:

   $$P(BM\_12M = 1) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{9} \beta_i \cdot WOE\_FR_i)}}$$

   L'estimació del model es realitza amb el mètode *stepwise*, aquest mètode consisteix a fer el model sense predictors i va agregant seqüencialment els predictors més influents i després d'agregar cada variable, elimina qualsevol variable que ja no proporcioni una millora en l'ajust del model. Seguint aquest mètode, s'han obtingut els següents coeficients:

WOE_Factor	**Coeficient**	**p-value**
Intercept	-2.7877	$<$ 2e-16
WOE_FR$_1$	-0.5404	$<$ 2e-16
WOE_FR$_2$	-0.5191	7.06e-16
WOE_FR$_4$	-0.4674	9.68e-15
WOE_FR$_5$	-0.3758	9.98e-10
WOE_FR$_6$	-0.8277	$<$ 2e-16
WOE_FR$_7$	-0.4944	$<$ 2e-16
WOE_FR$_8$	-0.5978	$<$ 2e-16
WOE_FR$_9$	-0.6446	$<$ 2e-16

Taula 7: Model Logistic WOE

Veiem que, és el factor WOE_FR_3 ha sigut descartat, per tant la informació que aportava es trobava recollida en altres predictors. Veient la matriu de correlació,

probablement aquests siguin els factors WOE_FR_1 i WOE_FR_4.

Observem que treballant amb un nivell de significació del 5% tots els factors són significatius. Donat que totes les variables predictives estan la mateixa magnitud (mitjançant els WOE's), la importància que pren cada variable la podem saber mirant el valor més gran en valor absolut del coeficient de cada factor de risc. Segons aquest model, la variable que presenta més importància és WOE_FR_6 amb un coeficient $|\beta_6| = 0.8277$ i la que menys, és el factor WOE_FR_5 amb un coeficient $|\beta_5| = 0.3758$. Sense tenir en compte el terme independent, podem estudiar quin és el pes de cada factor al model final. Per a fer-ho mirarem quina proporció representa el valor del coeficient sobre la resta. D'aquesta manera tenim:

β_1	β_2	β_4	β_5	β_6	β_7	β_8	β_9
12.1%	11.6%	10.5%	8.4%	18.5%	11.1%	13.4%	14.4%

Taula 8: Pes relatiu WOE_FR

Per tant, podem veure que tots els factors aporten al model aproximadament el mateix, amb l'excepció dels factors WOE_FR_6 i WOE_FR_5 que aporten més i menys, respectivament, que la resta.

Principalment, per avaluar el rendiment d'un model predictiu es valora la seva capacitat discriminatòria. Ens referim a capacitat discriminatòria a la capacitat que té el model per distingir entre individus de diferents classes. En el nostre cas, ve a ser la capacitat que té el model per distingir entre *bons* i *dolents*. Aquesta capacitat pot ser descrita mitjançant la corba ROC (*Receiver Operating Characteristic*), denotant per AUC (Area Under Curve) a l'àrea que queda per sota la corba mitjançant una representació gràfica relacionant les següents magnituds (Ens referirem al valor de l'AUC directament com la ROC del model): Considerem la següent matriu:

		Actual Value	
		0	**1**
Prediction outcome	**0**	True Negative	False Positive
	1	False Negative	True Positive

Aquesta matriu s'anomena matriu de confusió. Podem representar la corba ROC considerant la raó de verdaders positius (VPR) que es coneix com a sensitivitat enfront de la raó de falsos positius (FPR) que es coneix 1 - Especificitat. La capacitat discriminatòria ens ve reflectida pel valor de l'AUC.

$$Sensitivitat = \frac{TP}{TP + FP} \qquad Especificitat = \frac{TN}{TN + FN}$$

On l'Especificitat és la raó dels negatius verdaders. Els conceptes explicats sobre la corba ROC i la matriu de confusió s'han seguit de Narkhede, S. (2018).

En el context del risc de crèdit si un client resulta ser *bo* i el model el marca com a dolent i li deneguem la concessió d'un préstec, el més probable és que la pèrdua és molt menor en comparació amb la concessió d'un préstec a un client *dolent*, classificat com a *bo*. Per tant, ens interessa que la ràtio de verdaders positius sigui alta tot i que una situació equilibrada de les dues ràtios és l'ideal, però no sempre és possible ja que, si volem incrementar la raó VPR, aleshores la raó FPR es veurà afectada i disminuirà. És a dir, existeix un $tradeoff$ entre les dues raons. Es recomana a les entitats que la sensitivitat estigui al voltant del 70%.

Per defecte, en les funcions de R que tornem la matriu de confusió, utilitzen com a tall 0.5. En el nostre cas, interpretaria que si $p > 0.5$, aleshores el client és dolent. Però, en el cas en què una de les poblacions és minoritària, aquest tall no sempre és el millor.

Per establir el criteri per determinar si un client és bo o és dolent s'ha analitzat la distribució de les prediccions i la seva mitja. Les prediccions es distribueixen de la següent manera:

0%	1%	10%	25%	50%	75%	90%	95%	100%
0.0023	0.0037	0.0086	0.018	0.0464	0.0667	0.1241	0.1974	0.6172

Taula 9: Distribució de les prediccions

La mitja de les prediccions és 0.0580 i l'hem establit com a tall (és a dir, si $p > 0.058$ és dolent), ja que com el 94% de la mostra són bons, la mitja queda predominada per aquests, i per tant, si el model rendeix correctament, ens hauria de sortir una sensitivitat i especificitat equilibrades i elevades.

Analitzant la matriu de confusió hem obtingut els següents resultats:

	0	**1**
0	25407	645
1	9914	1530

Taula 10: Matriu de confusió WOE_logistic

Per tant,

$$Sensitivitat = \frac{1530}{1530 + 640} = 0.70 \qquad Especificitat = \frac{25407}{25407 + 9914} = 0.71$$

Utilitzant aquest tall sembla que el model és bastant capaç de discriminar entre bons i dolents.

Afegir que el tall per classificar un client com a bo o dolent depèn de l'entitat i de l'aversió al risc que tingui, per tant, és de caràcter subjectiu.

Tal com hem explicat, el que ens determinarà la capacitat discriminatòria del model és la ROC. Abans, establim els llindars veure el comportament de poder discriminatori a través de la ROC.

ROC	**Capacitat discriminatòria**
< 0.7	Dèbil
$[0.7, 0.8)$	Acceptable
$[0.8, 0.9)$	Forta
≥ 0.9	Molt Forta

Taula 11: Capacitat Discriminatòra ROC.

I en el nostre model, en el conjunt d'entrenament s'obté una ROC de:

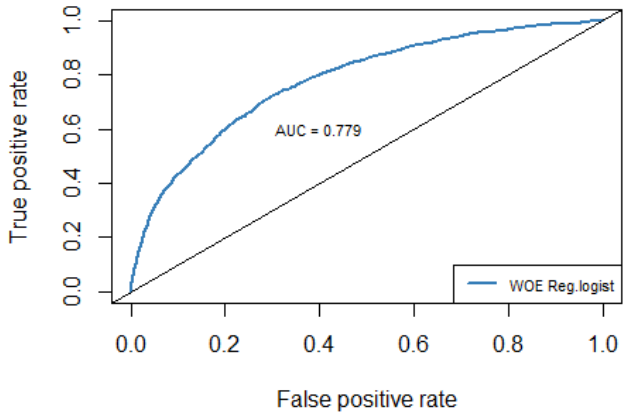


Figura 5: ROC WOE Logistic

Veiem que la diagonal de 45 graus $y = x$ és el model que classifica aleatòriament quan un client és bo o dolent. El que interessa maximitzar és l'àrea entre la corba ROC i la recta de 45º. Aquesta àrea com hem explicat, es coneix com a AUC i veiem que és d'un 77.9%. Per tant, aquest model és candidat a ser el model final que l'entitat necessita.

2. **Random Forest**

A continuació, estimarem un altre model però utilitzant una metodologia diferent

del vist fins ara. Els factors de risc, continuaran sent els 9 descrits en l'apartat [1] d'aquesta secció, no obstant, utilitzarem el valor real del factor i no el transformat. Tal i com hem dit, el valor transformat, entre altres característiques, ens permet comparar cada variable per a veure la seva importància. En el cas del Random Forest, com veurem, el R ja ens proporciona la importància de les variables predictores.

Donada la naturalesa dels arbres de decisió, a diferència de la regressió logística, aquests no tenen una estructura predeterminada sinó que s'han de generar a partir de les pròpies dades a partir d'uns hípers-paràmetres que explicarem posteriorment. En altres paraules, el mètode Random Forest genera arbres independents, per la construcció dels arbres s'utilitzen un subconjunt total de variables disponibles i d'observacions d'entrenament, aconseguint així que tinguin una estructura diferent, posterioment es passa cada client en el conjunt de arbres i es conta el número de vots de cada arbre, es mira quin és el vot majoritari i es classifica al client.
Un exemple d'arbre perdictiu aleatòri és:

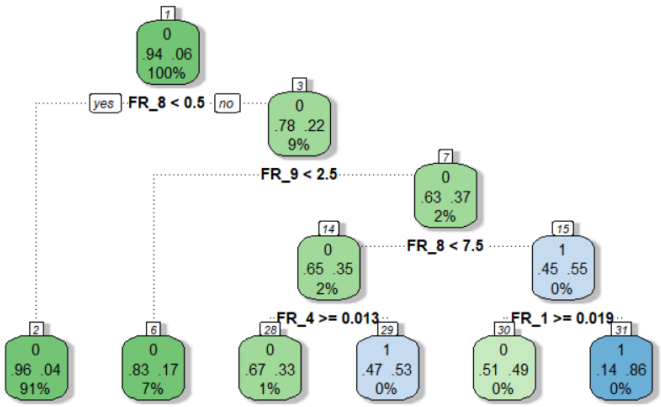


Figura 6: Arbre de decisió

Veiem que, per exemple aquest arbre consta de quatre variables i sis nodes terminals. Analitzem que conté cada node. Per exemple, situem-nos al node terminal 1, és a dir, el de baix a l'esquerra. El primer que veiem és la categoria majoritària del model, després veiem el percentatge d'individus que pertanyen en aquesta categoria i el percentatge d'individus que pertanyen a l'altra, que en aquest cas és un 96% de les observacions d'aquest node són de classe 0 i el 4% són de classe 1 i en aquest node terminal, hi ha un 91% de les observacions del conjunt d'entrenament.

El rendiment d'un model estimat mitjançant el mètode Random Forest depèn en gran manera dels hípers-paràmetres que hem mencionat anteriorment. Aquests hípers-paràmetres són:

(a) **Nº de variables predictives**: hem explicat a la teòria del Random Forest que, per descorrelacionar els arbres dels conjunts *bootstrap* és selecciona un nombre a l'atzar de variables predictives per a cada arbre. Normalment, en el cas de classificació, s'utilitza $m = \sqrt{n}$ variables. (sent n el nombre total de variables, en el nostre cas 9), no obstant això, no sempre és la millor decisió.

Per saber el nombre de variables òptim que ha de tenir cada arbre, estudiem l'evolució de la ràtio d'error de OOB (conjunt d'observacions que no s'han

utilitzat en les mostres *bootstrap*. Per a fer-ho em considerat un total de 500 arbres on cada node terminal conté mínim una observació. El resultat és:

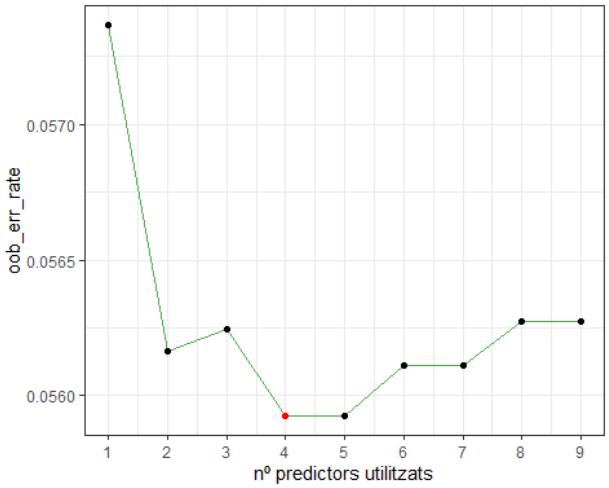


Figura 7: Evolució del out-of-bag-error vs mtry

Veiem doncs, que el nombre de variables que minimitza l'error del OOB és $m = 4$ (tenint en compte que hem fixat 500 arbres *bootstrap* i que cada node conté mínim una observació).

(b) **Nº mínim d'observacions**: el procediment és similar al que acabem d'explicar però, en aquest cas, volem saber el nombre òptim d'observacions mínimes ha de tenir cada node terminal, considerant 4 variables predictores i 500 arbres.

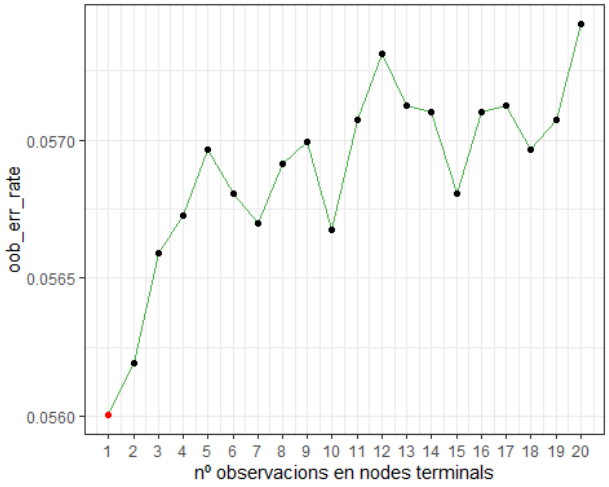


Figura 8: Evolució del out-of-bag-error vs nodesize

En aquest cas, veiem que el mínim d'observacions per cada node terminal que minimitza l'error en OOB és 1.

(c) **Nº d'arbres** *bootstrap*: per acabar, anem a veure, tal com hem dit a la part teòrica del treball, quan l'error de OOB s'estabilitza considerant, en aquest cas,

quatre variables predictives i una observació mínima en cada node terminal.

Figura 9: Evolució del out-of-bag-error vs ntrees

Veiem que a partir de més de 100 arbres. L'error de *OOB* efectivament, convergeix.

Per tant, els hípers-paràmetres que utilizarem per estimar el model són:

- Nº Variables: 4
- Nº Observacions: 1 per node terminal
- Nº arbres *bootstrap*: 500 arbres

Com ja hem explicat, per la construcció dels arbres s'utilitzen un subconjunt total de variables disponibles i d'observacions d'entrenament. Per tant, resulta interessant estudiar l'impacte que té la presència o no de cada variable en l'obtenció del model final. Per tant, els hípers-paràmetres que utilitzàrem per estimar el model són:

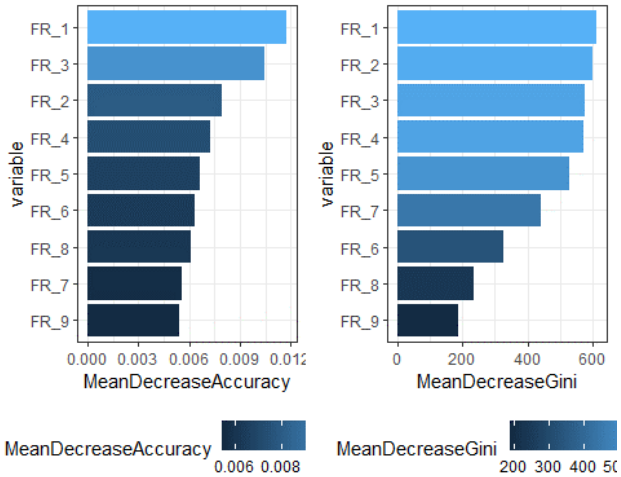


Figura 10: Importància Variables Predictives

La mètrica *Mean Decrease Accuracy* (Reducció de l'Accuracy) ens permet visualitzar l'impacte relatiu que té no incloure una variable concreta en el rendiment del

model. Aquesta mètrica ens permet saber quina és la variable amb major poder predictiu en el model. Veiem que totes les variables presenten un impacte superior a 0 i la variable que presenta més importància és el factor FR_1. Per tant tots els factors son rellevants i el model els té en compte.

Per altra banda, la mètrica *Mean Decrease Gini* (Reducció de la impuresa (Gini)) està relacionada amb la forma en què es construeixen els arbres de decisió aleatoris i quina variable escollir en cada node. La mètrica mesura la capacitat que té una variable per dividir les dades en particions més pures en cada node, és a dir, que la majoria de les dades que es classifiquen en una categoria, realment pertanyin a ella. A diferència de la mètrica anterior, tenim que els factors de risc FR_8 i FR_9 mostren una puresa a l'hora de partir les dades inferior a la resta, no obstant i donat que la primera mètrica les considera rellevants, és convenient considerar-les i com a símil tenim que el factor FR_1 és la variable que millor divideix les dades.

La ràtio d'error en la mostra OOB es calcula com:

$$OOB = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{\{y_i \neq V_i\}}$$

Que en el nostre cas és d'un 5.56%.

Pel que fa a la matriu de confusió, anteriorment hem vist un criteri per assignar el tall per la classificació, però aquest no és l'únic.
Un altre criteri per determinar el tall, és calcular en punt que queda més a prop del punt $(0, 1)$, és a dir, el punt tal que:

$$\min((1 - Sensitivitat)^2 + (1 - Especificitat)^2)$$

I el punt que minimitza l'expressió anterior és $p = 0.0718$. Amb aquest tall, la matriu de confusió és:

	0	**1**
0	27628	566
1	7693	1609

Taula 12: Matriu de confusió FR_RandomForest

Per tant,

$$Sensitivitat = \frac{1609}{1609 + 566} = 0.74 \qquad Especificitat = \frac{27628}{27628 + 7693} = 0.78$$

I la corba ROC del model en el conjunt d'entrenament és:

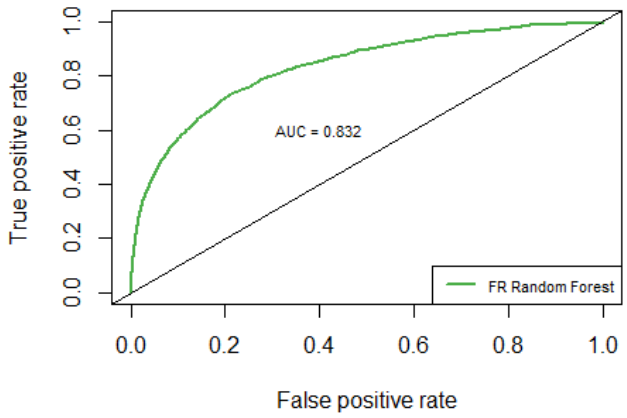


Figura 11: ROC FR RandomForest

Veiem que amb aquesta metodologia, la ROC augmenta fins a un 83,2%. Per tant, el rendiment del model és superior a l'entrenat mitjançant la regressió logística.

3. **Adaboost**
   Per últim, anem a entrenar un model mitjançant la tècnica *Adaboost*. El mètode *Adaboost* proposa entrenar una sèrie de classificadors dèbils de forma interactiva, de manera que cada un nou classificador *weak learner* s'enfoca en les dades que han estat classificades erròniament pel seu predecessor, d'aquesta manera cada cop s'obtenen classificadors que prediuen millor les dades fins a obtenir el classificador final, que és capaç de predir la majoria de dades correctament.
   En un primer moment, assignem el mateix pes a totes les observacions i en cada iteració, en funció de les observacions mal classificades, es va actualitzant donant més pes a aquestes. D'aquesta manera es busca minimitzar l'error esperat i s'enfoca a classificar correctament les dades que ara tenen un major pes.

   El classificador final ve donat per una combinació lineal ponderada de tots els classificadors dèbils que s'han produït anteriorment. Per tant, el rendiment del classificador final només del nombre de classificadors dèbils que hi ha. En altres paraules, de les iteracions que es fan sobre els pesos de les observacions mal classificades.

   D'aquesta forma, en el mètode *Adaboost* sol hem d'escollir l'híper-paràmetre, donat el cost computacional que té la funció *adabag* :: *boosting*() per utilitzar el mètode *Adaboost* en unes dades com les que tenim, hem fixat l'híper-paràmetre a 100 iteracions. Estudiem els resultats obtinguts.

   Per construir la matriu de confusió, utilitzem un altre criteri diferent dels vist fins ara. Utilitzem el criteri de *Youden* (Schisterman, E. F., Perkins, N. J., Liu, A., Bondell, H. (2005)), que és el punt de tall que fa que la corba ROC estigui a distància màxima del model aleatori. Aquest punt és el 2.305752 i la matriu de confusió resultant és

	0	**1**
0	26431	612
1	8890	1563

Taula 13: Matriu de confusió FR_Adaboost

Per tant,

$$Sensitivitat = \frac{1563}{1563 + 612} = 0.72 \qquad Especificitat = \frac{26431}{26431 + 8890} = 0.74$$

I la corba ROC del model és

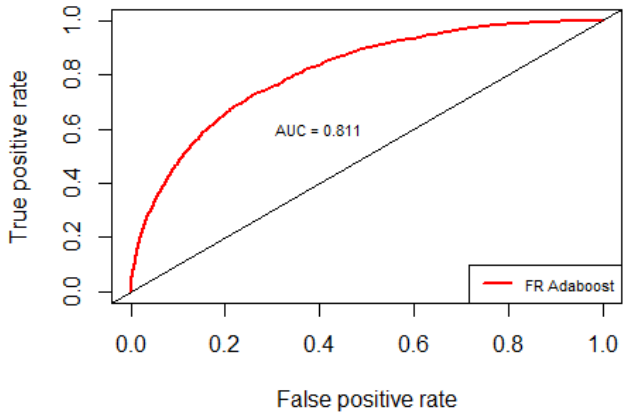


Figura 12: ROC FR Adaboost

Veiem que el ROC és 81.1%, és a dir, el model s'ajusta bé a les dades d'entrenament. Haurem de valorar el rendiment i el possible *overfitting* del model a la mostra de validació.
Passem a veure la influència de els variables en el mètode Adaboost

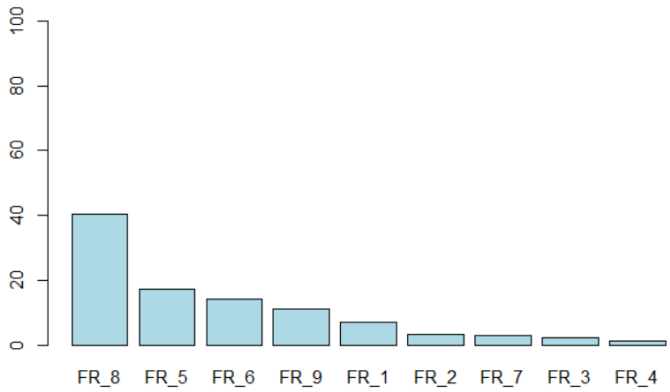


Figura 13: Influència Factors

Veiem que, segons aquest model, la variable que més pesa en el model és FR_8 amb un 40% mentres que la la resta de variables, influeixen menys d'un 20%.

Ara ja tenim 3 models entrenats amb diferents mètodes. Analitzem les corbes ROC de cada model.

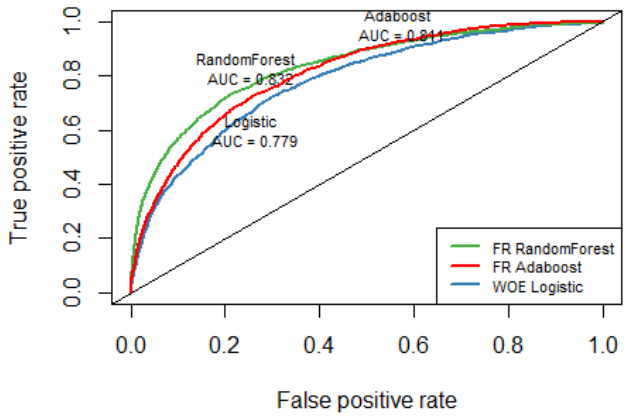


Figura 14: ROCs Models Entrenats

Si ens quedéssim aquí i haguéssim d'escollir un model, escolliríem el model entrenat pel mètode *Adaboost*, però seria un error, ja que s'ha d'avaluar la capacitat discriminatòria amb noves dades. És a dir, amb la mostra de validació.

3.3 Validació de models

Com acabem de comentar, no és suficient veure la ROC del model en el conjunt d'entrenament. La validació d'un model és el procés d'avaluar model generat amb un bon rendiment en la mostra d'entrenament sobre un conjunt de noves dades. Aquest procés ens proporciona la capacitat de generalització d'un model entrenat.
Per a fer-ho utilitzem la mostra de validació, que representa un 20% de les observacions totals.

Classe	Mostra de validació
bons	8.822
dolents	536
Total	**9.358**

Taula 14: Mostra de Validació.

Seguim amb la mateixa metodologia anterior, és a dir, analitzem per separat cada model sobre les noves dades i seguint el criteri utilitzat en el conjunt d'entrenament, veurem la matriu de confusió, la sensitivitat i especificitat del model i la corba ROC.

1. **Model logístic**

Per aplicar el model estimat en les dades d'entrenament

$$P(BM\_12M = 1) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \sum_{i=1}^{9} \hat{\beta}_i \cdot WOE\_FR_i)}}$$

en les noves dades, el primer que hem de fer és transformar-les. Per a fer-ho utilitzem, per a cada factor, els mateixos *bins* definits en l'apartat [1] d'aquesta secció. A cada *bin* del factor li assignem el WOE que li toca. Un cop tenim els factors transformats, utilitzem els coeficients anteriors i els resultats que obtenim són els següents.

Respecte a la matriu de confusió, utilitzem el mateix criteri que en el procés de desenvolupament de model. És a dir, un client el classifiquem com a bo si té una probabilitat inferior a la mitja de prediccions. En cas contrari és dolent.
La mitja de prediccions en el conjunt d'entrenament és $\bar{p} = 0.0580$ i la matriu de confusió és

	0	**1**
0	6365	177
1	2457	359

Taula 15: Matriu de confusió Validació Logística

Per tant,

$$Sensitivitat = \frac{359}{359 + 177} = 0.67 \qquad Especificitat = \frac{6365}{6365 + 2457} = 0.72$$

I la corba ROC del model de validació és

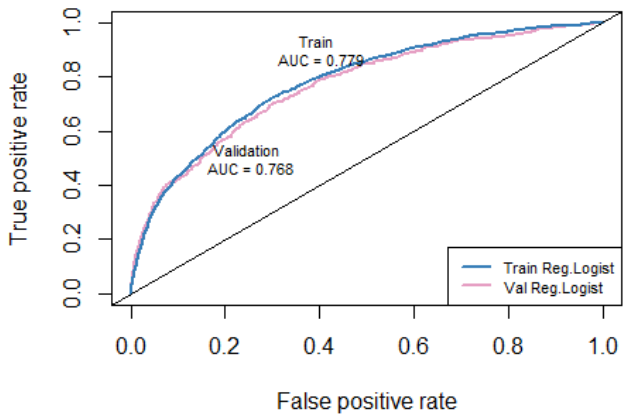


Figura 15: ROC Logistic Train i Validation

Veiem doncs, que el ROC del model en la mostra de validació és d'un 76.8%, mentre que el ROC de la mostra d'entrenament és 77.9%. EL ROC en noves dades ha

disminuït lleugerament, però podem concloure que el model no presenta problemes de *overfitting* i té una capacitat discriminatòria bastant bona.

2. **Random Forest**

Sovint, els arbres de classificació es solen sobre ajustar a les dades. Analitzem el rendiment de model entrenat mitjançant la tècnica *Random Forest* en la mostra de validació.

Per construir la matriu de confusió, seguim el mateix criteri que en la mostra d'entrenament. És a dir, el punt de tall per classificar un individu com a bo o dolent és aquell que fa que la corba ROC estigui més a prop del punt (0,1).

En aquest cas, el punt obtingut en la mostra d'entrenament és $p = 0.0718$, és a dir, els clients que tinguin una predicció superior a aquest punt, el model els classificarà com a dolents. La matriu de confusió resultant és

	0	**1**
0	6930	146
1	1892	390

Taula 16: Matriu de confusió Validació Random Forest

Amb una sensitivitat i especificitat:

$$Sensitivitat = \frac{390}{390 + 146} = 0.73 \qquad Especificitat = \frac{6930}{6930 + 1892} = 0.78$$

I la corba ROC del model en la mostra de validació és

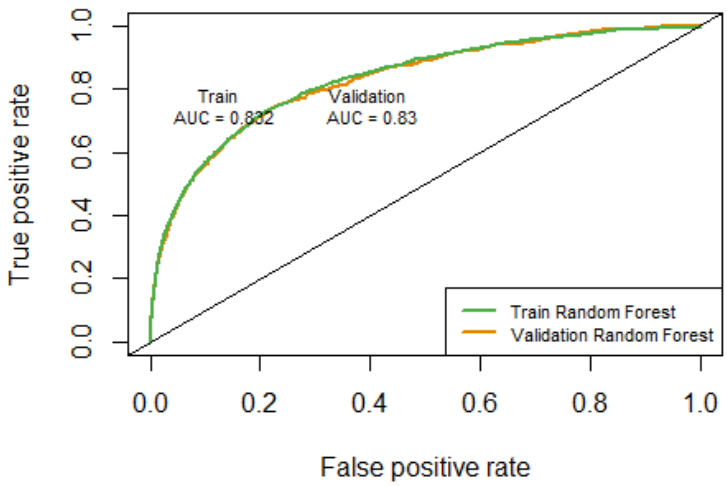


Figura 16: ROC Random Forest Train i Validation

Com podem veure, les dues corbes ROC són quasi iguals. La ROC en l'entrenament és d'un 83.2% mentre que en la validació és d'un 83.0%. Anteriorment hem comentat que els arbres de classificació solen sobre ajustar-se a les dades. En el nostre cas no és així, el rendiment del model és igual de bo en noves dades que en el conjunt d'entrenament, per tant podem concloure que el model té una capacitat discriminatòria forta.

3. **Adaboost**

Anem a estudiar l'últim model. Com ja s'ha comentat, aquest model és candidat a presentar *overfitting*.

El criteri a seguir en aquest model per classificar un client bo o dolent, és assignar el punt tal que la corba ROC està a distància màxima del model aleatori en el model d'entrenament. Aquest punt és $p = 2.305752$ i la matriu de confusió resultant en aquest cas és:

	0	**1**
0	6218	156
1	2604	380

Taula 17: Matriu de confusió Validació Adaboost

Amb una sensitivitat i especificitat:

$$Sensitivitat = \frac{380}{380 + 156} = 0.71 \qquad Especificitat = \frac{6218}{6218 + 2604} = 0.70$$

i la corba ROC del model en la mostra del validació

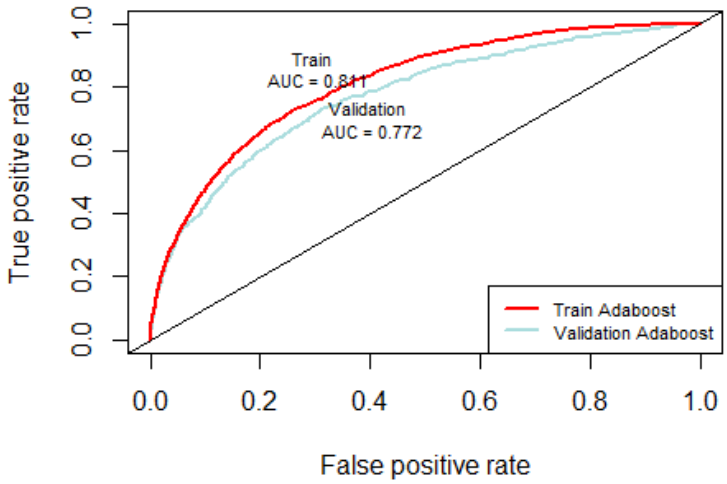


Figura 17: ROC Adaboost Train i Validation

Mirant la matriu de confusió veiem que tant la sensibilitat com l'especificitat són

d'un 70% i la corba ROC de model és d'un 77.2%, per tant, sembla que no el model no presenta problemes d'overfitting davant de noves dades.

De forma resumida, veiem com han quedat les corbes ROCs dels models en la mostra de validació.

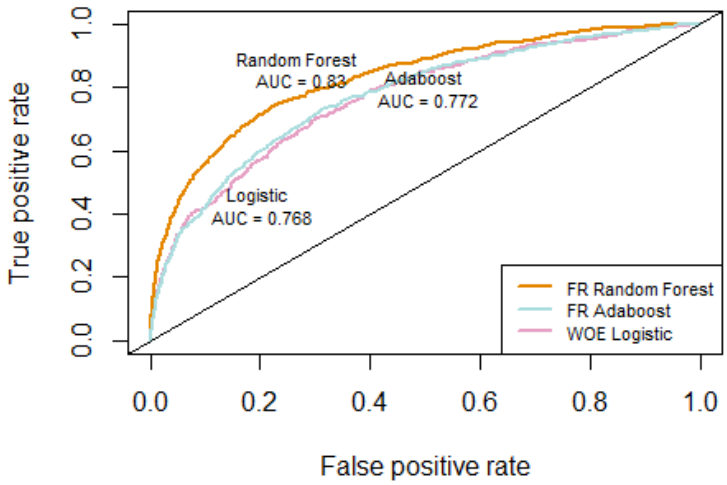


Figura 18: ROCs Models Validació

Hem vist que el models en la mostra d'entrenament presenten bones capacitats predictives. Respecte dels models de validació, és natural que la sensibilitat, especificitat i AUC dels models en noves dades empitjori, no obstant, cap dels tres models presenten diferències significatives entre la mostra d'entrenament i la mostra de validació.

4 Conclusions

En aquest capítol detallarem els aspectes més importants del treball i analitzarem els resultats obtinguts per extreure'n conclusions. A més proposarem línies d'investigació futures que no s'han fet a causa de la limitació d'extensió del treball.

La gestió del risc de crèdit és un aspecte decisiu que afronten les entitats bancàries avui en dia. La regulació bancaria, exigeix que les entitats financeres tinguin un mínim de capital i reserves en els seus balanços. La finalitat és que donada una crisi econòmica i la taxa de mora augmenti considerablement, l'entitat no es quedi sense liquiditat i pugui suportar-ho amb el capital i les reserves de la pròpia entitat. Per tant, els bancs han d'aprovisionar una part dels crèdits/préstecs que emeten.

Davant d'aquesta situació, una bona gestió del risc de crèdit és fonamental, ja que com millor decideixi l'entitat a qui concedeix finançament, menys risc haurà de sofrir davant d'un escenari de crisi.

Arran d'això, la demanda de models predictius (siguin desenvolupats per la mateixa entitat o per consultors) s'ha vist incrementada i és precisament en aquest punt on recau la major part d'aquesta tesi.

En la tesi hem estudiat diferents punts. Hem començat explicant una descripció del problema a estudiar i el tractament de dades, seguidament hem analitzat la teoria que hi ha darrere d'alguns mètodes de *Machine Learning* i hem acabat per la modelització dels mateixos mètodes estudiats.

Pel que fa a la regressió logística, hem sigut capaços de veure tot el procés des d'un punt de vista matemàtic per arribar als estimadors del model, per tant, aquests models sovint són acceptats per les entitats bancàries ja que es possible traçar tot el procediment intermedi pel desenvolupament del model.

Referent als arbres de predicció aleatòria, hem descrit dos mètodes que els utilitzen. Aquests són el mètode *Random Forest* i el mètode *Adaboost*, i també hem pogut explicar en detall com funcionen aquests models.

Responent a la pregunta que ens fèiem inicialment de si és possible descriure i detallar tot el procés que hi ha darrere d'un model de *Machine Learning*, podem concloure que sí, i de cara al regulador és possible poder justificar perquè un model predictiu construït amb regressió lineal o bé amb arbres de classificació, classifica a un client com a bo o dolent.

Analitzem els resultats obtinguts, la taula següent ens resumeix els resultats dels models obtinguts en la mostra d'entrenament i en la mostra de validació.

1. **Rendiment dels models en la mostra d'entrenament**

	Regressió Logística	Random Forest	Adaboost
Sensitivitat	0.70	0.74	0.72
Especificitat	0.71	0.78	0.74
ROC	77.9%	83.2%	81.1%

Taula 18: Comparació models mostra entrenament

La primera observació és que els tres mètodes en les dades d'entrenament presenten fortes capacitats de discriminació. El model entrenat mitjançant la regressió logística és qui presenta "pitjor" ROC d'un 77.9%, mentres que els models *Random Forest* i *Adaboost* presenten ROC's superior als 80%. Respecte a la matriu de confusió, tots els models com a mínim tenen una sensitivitat del 70%. L'aspecte positiu a part de presentar bones sensibilitats, és que les especificitats en tots els models són superiors al 70%. Per tant, una primera conclusió és que els models en les mostres d'entrenament mostren rendiments molt positius.

2. **Rendiment dels models en la mostra de validació**

	Regressió Logística	Random Forest	Adaboost
Sensitivitat	0.67	0.73	0.71
Especificitat	0.72	0.78	0.70
ROC	76.8%	83.0%	77.2%

Taula 19: Comparació models mostra validació

Com hem dit, el models entrenats amb arbres de decisió, com són *Random Forest* i, en especial *Adaboost* es solen sobre-ajustar. En comparació el model d'entrenament, el que mostra menys variablitat és el model entrenat per *Random Forest*, que segueix presentat una ROC superior al 80% i tant l'especificitat com la sensibilitat són superiors al 70%. Similar succeeix amb el model *Adaboost*, encara que la seva ROC hagi disminuït, segueix estan a prop del 80% i tan la sensibilitat com la especificitat són superiors o iguals al 70%. Per últim, la regressió logística segueix tenint una bona capacitat predictiva, però presenta una sensibilitat inferior al 70% no obstant, davant d'aquesta situació, si l'entitat tingués alguna política interna que imposes una sensitivitat mínima d'un 70% (que sol ser l'habitual) i tingués un especial interès a utilitzar aquest model, podria recalibrar aquest punt de tall penalitzant l'especificitat ja que és un model amb bona capacitat discriminatòria.

Un cop vist els models tan en la mostra d'entrenament i en la mostra de validació, concloem que, tant el model logístic com el *Random Forest* i *l'Adaboost*, tenen una capacitat discriminatòria per classificar clients bons i dolents satisfactòria. En particular el *Random Forest* perquè, a part de presentar una ROC superior al 80%, és el model que mostra menys variabilitat de sensitivitat i especificitat respecte a la mostra d'entrenament.

En conclusió, tal i com hem vist, els tres models en termes de discriminació serien adequats i el banc els podria instaurar en els seus sistemes interns, però no sempre és suficient analitzar la capacitat discriminatòria.
La influència de les variables en cada model també és un aspecte a tenir en compte. Pel que fa a la regressió logística, de nou variables transformades, al model n'entren vuit, de

les quals, ha grans trets, tenien un pes relatiu en el model. Pel que fa al *Random Forest* el resultat és similar a la regressió logística, però en aquest cas, el model utilitza tots els factors risc. En canvi, al model *Adaboost*, si que és cert que presenta una bona capacitat discriminatòria però, d'un total de 9 factors de risc. Un factor representa el 40% de model, mentre que la resta no arriben ni al 20%. Per tant, de cara a l'entitat bancària, no és atractiu que una única variable predomini tant i tota la resta influeixi tan poc ja que el model s'acaba resumint en una única variable i en cas que no hi hagi informació d'aquesta variable per a un client, el model pot fer classificacions errònies.

No obstant a tot el vist anteriorment, l'aplicació del model podria tenir més aplicacions que el fet de decidir si concedir o denegar un préstec a un client.Si el banc vol únicament saber si concedir un préstec a un client, hauria d'escollir el model de classificació que millor discrimina, en el nostre cas, és el model *Random Forest*. Ara bé, si el banc li vol donar un altre enfocament al model, i aquí entren les línies d'investigació futures, com per exemple, fer un estudi intern sobre els clients que ja els hi ha concedit finançament per classificar-los en diferents grups de risc i així poder preveure quins presenten més risc d'incompliment.

Davant d'aquesta situació seria interessant utilitzar la regressió logística, ja que amb els valors dels coeficients podríem establir un score a cada client que quantifiques la seva qualitat creditícia mitjançant la següent expressió:

$$Score = c_1 \cdot \ln\left(\frac{P(BM\_12M = 1)}{1 - P(BM\_12M = 1)}\right) + c_2$$

És a dir, utilitzant la *odds* ràtio, podem establir una puntuació a cada client i després fer grups homogenis entre els clients d'un mateix grup i heterogenis entre els diferents grups i establir protocols de seguiment per als clients que presenten més risc. (c_1 i c_2 són dues constants que serveixen per definir l'escala a la puntuació).

Una altra línia d'investigació seria provar models predictius utilitzant *deep learning*, és a dir, xarxes neuronals. És probable que encara milloressin més els resultats d'aquests models, però com hem dit, la traçabilitat del mètode és un aspecte necessari per a l'acceptació del model de cara al regulador, i el major inconvenient de les xarxes neuronals, és que el que passa per a dins de les xarxes no es té la certesa de poder-ho explicar.

Com a última conclusió, comentar la complexitat general de l'elaboració de la tesi i el temps dedicat. Des del tractament de dades, fins a la mateixa validació dels models. També remarcar tots els conceptes de *Machine Learning* i del llenguatge R que hem après fent aquest treball i sobretot, de la importància que tenen les matemàtiques ja no sol en la descripció dels models, sinó en la pròpia realitat econòmica i financera.

Per últim, dir que hem assolit l'objectiu de poder explicar, des d'un punt de vista matemàtic, com funcionen alguns models de *MachineLearning* i haver pogut comprovar empíricament com funcionen aquests models i l'aplicabilitat directa que tenen.

Referències

Alfaro, E., Gámez, M., García N. (2013), *Adabag: An R Package for Classification with Boosting and Bagging*, Journal of Statistical Software, 54(2), pp. 1-35.

Agresti A. (2002), *Categorical Data Analysis*, Vol. 482, John Wiley Sons, Inc., Hoboken, New Jersey.

Breiman, L. (1996), *Bagging Predictors*, Machine Learning 24(2): pp. 123-140.

Breiman, L. (1996), *Random Forests*, Machine Learning 45(1): pp. 5-32.

Breiman, L., Friedman, J., Olshen R., i Stone, C. (1984), *Classification and Regression Trees*, Belmont, CA: Chapman & Hall.

Bourel, M. (2012), *Métodos de agregación de modelos y aplicaciones. Memoria Investigaciones en Ingeniería* 10, pp. 19-32.

Freund, Y., i Schapire, R. E. (1995), *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*,In European conference on computational learning theory. Springer, Berlin, Heidelberg. pp. 23-37.

Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G. (2008). En Guo, M., Zhao, L., Wang, L. (Eds.), *On the Class Imbalance Problem*, In 2008 Fourth international conference on natural computation, (Vol. 4, pp. 192-201). IEEE.

Hastie, T., Tibshirani, R., Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science Business Media, New York.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), *An Introduction to Statistical Learning with applications in R* Vol 112, Springer, New York.

Louppe, G., Wehenkel, L., Sutera, A., Geurts, P. (2013), *Understanding variable importances in forests of randomized trees. In Advances in neural information processing systems*, Vol 1, pp. 431-439.

Majer, I. (2006), *Application scoring: logit model approach and the divergence method compared*, Department of Applied Econometrics Working Papers, (10), Warsaw.

Narkhede, S. (2018), *Understanding Confusion Matrix.*
`https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62`

Narkhede, S. (2018), *Understanding AUC - ROC Curve.*
`https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5`

Natekin, A., Knoll A. (2012), *Gradient boosting machines, a tutorial*, Frontiers in Neurorobotics.
`https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021/full`

Parmar, R. (2018), *Common Loss functions in machine learning.*
`https://towardsdatascience.com/common-loss-functions-in-machine-learning-46af0ffc4d23`

Peña, D. (2002), *Análisis de datos multivariantes*, McGraw-Hill, Madrid.

Sanz-Solé, M. (1999), *Probabilitats*, Vol 28, Publicacions i Edicions de la UB, Barcelona.

Schisterman, E. F., Perkins, N. J., Liu, A., Bondell, H. (2005). *Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples*, Epidemiology, pp. 73-81.

Smith, E, P., Lipkonch, I., i Ye, K (2002), *Weight-of-Evidence (WOE): Quantitative Estimation of Probability of Impairment for Individual and Multiple Lines of Evidence*, Human and Ecological Risk Assessment: 8(7), pp. 1585-1596.

Vapnik, V. (1998), *Statistical Learning Theory*, Jhon Wiley & Sons INC, New York.

Visa, S., Ramsay, B., Ralescu,A., Van der Knaap, E. (2011), *Confusion Matrix-based Feature Selection*, MAICS, 710, pp. 120-127.

5 Annex

1. Distribució dels factors de risc

Factor	min	p1	p10	p25	p50	p75	p90	p95	p99	max
FR_1	0.00	0.00	0.00	0.01	0.02	0.03	0.05	0.06	0.10	9,999,999,999,999.00
FR_2	-3.84	0.00	0.10	0.18	0.30	0.46	0.64	0.74	0.89	9,999,999,999,999.00
FR_3	-636.13	-2.13	0.00	0.08	0.30	0.62	0.88	1.03	3.11	9,999,999,999,999.00
FR_4	-1,397,894.00	-18.29	0.00	0.41	2.08	5.20	11.05	17.68	61.92	9,999,999,999,999.00
FR_5	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.06	0.19	8.12
FR_6	0.00	0.00	0.00	0.00	0.00	0.23	0.79	0.99	1.00	9,999,999,999,999.00
FR_7	0.00	0.00	0.00	0.00	20.07	64.20	90.00	109.70	156.00	365.00
FR_8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	5.00	39.00
FR_9	0.00	0.00	0.00	0.00	0.00	0.00	2.00	2.00	5.00	27.00
FR_10	-3,272.18	-14.09	-0.76	-0.33	0.04	0.75	2.23	4.86	62.09	9,999,999,999,999.00
FR_11	0.00	0.00	0.02	13.99	60.67	137.70	274.05	486.53	9,999,999,999,999.00	9,999,999,999,999.00
FR_12	-360,229.00	-11.82	-0.72	-0.24	0.18	1.01	2.86	6.31	70.98	9,999,999,999,999.00
FR_13	-10.37	-0.13	0.00	0.00	0.01	0.03	0.07	0.11	0.25	17.18
FR_14	-473,497.00	-3.74	-0.51	-0.17	0.13	0.66	1.99	4.36	352.02	9,999,999,999,999.00
FR_15	-1,089.80	-2.85	-0.46	-0.20	0.01	0.32	1.15	2.41	428.00	9,999,999,999,999.00
FR_16	-310,229.00	-0.62	-0.04	0.01	0.07	0.19	0.51	1.04	11.84	9,999,999,999,999.00
FR_17	0.00	0.31	0.84	1.05	1.29	1.75	2.63	3.81	8.52	9,999,999,999,999.00
FR_18	-30.28	1.17	22.78	49.00	81.07	117.31	166.04	207.43	338.40	4,899.24
FR_19	-2,544.25	-0.87	-0.37	-0.17	0.00	0.20	0.58	1.07	19.25	1,000,000,000.00
FR_20	-247.07	-0.73	-0.23	-0.09	0.02	0.13	0.36	0.72	6.26	9,999,999,999,999.00
FR_21	-7,222.83	-8.58	-0.69	-0.16	0.07	0.40	1.38	3.13	50.68	9,999,999,999,999.00
FR_22	-9,761.13	-3.32	-0.69	-0.26	0.27	1.60	9,999,999,999,999.00	9,999,999,999,999.00	9,999,999,999,999.00	9,999,999,999,999.00
FR_23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.07	14.91
FR_24	3.00	36.00	94.00	163.00	259.00	356.00	448.00	531.00	812.47	1,767.00
FR_25	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
FR_26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
FR_27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figura 19: Distribució FR

2. Análisi dels factors

Els camps que calculem mitjançant la funció woebin() en R són:

Camp	Descripció
Factor	Factor que estem analitzant
Bin	Tram del factor
Frec	Nombre de clients que hi ha en el tram corresponent
Good	Nombre de clients marcats com *bons*
Bad	Nombre de clients marcats com *dolents*
p_Pob	Percentatge de clients respecte el total de clients
p_Good	Percentatge de *bons* respecte el total de clients *bons*
p_Bad	Percentatge de *dolents* respecte el total de clients *dolents*
WOE	Es calcula com $\ln(\frac{p\_good}{p\_Bad})$ en cada tram
IV	Es calcula com $\sum(p\_good - p\_Bad)\ln(\frac{p\_good}{p\_Bad})$
TM	És la ràtio $\frac{Bad}{\sum Bad}$
TMR	Es calcula com $\frac{TM}{\frac{\sum Bad}{\sum Frec}}$

Taula 20: Detall Camps.

I el detall per a cada factor és:

Factor	Bin	Break	Frec	p_Pob	Good	Bad	p_Good	p_Bad	WOE	Bin_IV	IV	TM	TMR
FR_1	[-Inf,0.02)	0.02	18188	48.51%	17557	631	49.71%	29.01%	0.5385	0.111435	0.2545	3.47%	0.598095
FR_1	[0.02,0.03	0.032	9426	25.14%	8867	559	25.10%	25.70%	-0.0235	0.00014	0.2545	5.93%	1.022375
FR_1	[0.032,0.0	0.052	7260	19.36%	6645	615	18.81%	28.28%	-0.4075	0.038556	0.2545	8.47%	1.460374
FR_1	[0.052, Inf	Inf	2622	6.99%	2252	370	6.38%	17.01%	-0.9814	0.104376	0.2545	14.11%	2.432735
FR_2	[-Inf,0.1)	0.1	3932	10.49%	3545	387	10.04%	17.79%	-0.5726	0.044413	0.1914	9.84%	1.696771
FR_2	[0.1,0.35)	0.35	18253	48.68%	16993	1260	48.11%	57.93%	-0.1858	0.018243	0.1914	6.90%	1.190041
FR_2	[0.35,0.55]	0.55	9130	24.35%	8740	390	24.74%	17.93%	0.3221	0.021944	0.1914	4.27%	0.73641
FR_2	[0.55, Inf)	Inf	6181	16.48%	6043	138	17.11%	6.34%	0.992	0.106774	0.1914	2.23%	0.384898
FR_3	[-Inf,-0.1)	-0.1	1955	5.21%	1795	160	5.08%	7.36%	-0.3699	0.008412	0.2236	8.18%	1.410909
FR_3	[-0.1,0.1)	0.1	8491	22.65%	8256	235	23.37%	10.80%	0.7717	0.096995	0.2236	2.77%	0.477128
FR_3	[0.1,0.3)	0.3	8137	21.70%	7824	313	22.15%	14.39%	0.4313	0.03347	0.2236	3.85%	0.663141
FR_3	[0.3,0.6)	0.6	8967	23.91%	8419	548	23.84%	25.20%	-0.0555	0.000754	0.2236	6.11%	1.05356
FR_3	[0.6, Inf)	Inf	9946	26.53%	9027	919	25.56%	42.25%	-0.5028	0.08394	0.2236	9.24%	1.592915
FR_4	[-Inf,0)	0	1970	5.25%	1787	183	5.06%	8.41%	-0.5086	0.017062	0.1624	9.29%	1.60144
FR_4	[0,1.5)	1.5	14014	37.37%	13499	515	38.22%	23.68%	0.4788	0.06961	0.1624	3.67%	0.633535
FR_4	[1.5,6.5)	6.5	14137	37.70%	13360	777	37.82%	35.72%	0.0571	0.0012	0.1624	5.50%	0.947522
FR_4	[6.5,17.5)	17.5	5473	14.60%	4988	485	14.12%	22.30%	-0.4568	0.037353	0.1624	8.86%	1.527714
FR_4	[17.5, Inf)	Inf	1902	5.07%	1687	215	4.78%	9.89%	-0.7274	0.037161	0.1624	11.30%	1.948739
FR_5	[-Inf,0.003	0.003	24986	66.64%	23167	1819	65.59%	83.63%	-0.243	0.043844	0.2665	7.28%	1.255052
FR_5	[0.003,0.0	0.01	5388	14.37%	5142	246	14.56%	11.31%	0.2524	0.008197	0.2665	4.57%	0.787106
FR_5	[0.01,0.018	0.018	1933	5.16%	1887	46	5.34%	2.11%	0.9267	0.029908	0.2665	2.38%	0.410253
FR_5	[0.018, Inf	Inf	5189	13.84%	5125	64	14.51%	2.94%	1.5956	0.184562	0.2665	1.23%	0.212629
FR_6	[-Inf,0.14)	0.14	26457	70.56%	24579	1878	69.59%	86.34%	-0.2158	0.036156	0.1958	7.10%	1.223716
FR_6	[0.14,0.54]	0.54	5378	14.34%	5185	193	14.68%	8.87%	0.5034	0.029227	0.1958	3.59%	0.618674
FR_6	[0.54, Inf)	Inf	5661	15.10%	5557	104	15.73%	4.78%	1.191	0.130426	0.1958	1.84%	0.316713
FR_7	[-Inf,15)	15	18230	48.62%	17544	686	49.67%	31.54%	0.4541	0.082335	0.1756	3.76%	0.648729
FR_7	[15,55)	55	7506	20.02%	7089	417	20.07%	19.17%	0.0458	0.000411	0.1756	5.56%	0.957752
FR_7	[55, Inf)	Inf	11760	31.36%	10688	1072	30.26%	49.29%	-0.4879	0.092827	0.1756	9.12%	1.571496
FR_8	[-Inf,1)	1	34217	91.26%	32751	1466	92.72%	67.40%	0.3189	0.080762	0.4605	4.28%	0.738614
FR_8	[1, Inf)	Inf	3279	8.74%	2570	709	7.28%	32.60%	-1.4996	0.379733	0.4605	21.62%	3.72761
FR_9	[-Inf,1)	1	28360	75.63%	27231	1129	77.10%	51.91%	0.3956	0.099636	0.3292	3.98%	0.686299
FR_9	[1,2)	2	4915	13.11%	4512	403	12.77%	18.53%	-0.3719	0.0214	0.3292	8.20%	1.413537
FR_9	[2, Inf)	Inf	4221	11.26%	3578	643	10.13%	29.56%	-1.071	0.208137	0.3292	15.23%	2.626161

Figura 20: Anàlisis estadístic FR

3. Mètode Bagging

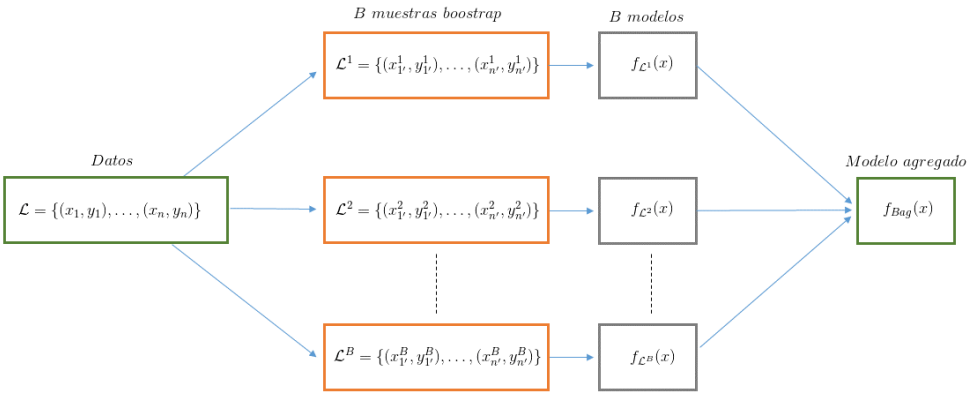


Figura 21: Exemple del mètode Bagging

Índex de taules

Índex de figures