UNIVERSITAT DE BARCELONA

Treball final de grau

Grau en Enginyeria Informàtica

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Aprenentatge automàtic aplicat a les dades d'apartaments turístics de Barcelona

---

Autor: Pau Bernabé Constans

Director: Santiago Seguí Mesquida
Realitzat a: Facultat de Matemàtiques i Informàtica

Barcelona, 22 de Gener 2022

Resum

La predicció de certes característiques de béns immobiliaris mitjançant aprenentatge automàtic és un tòpic que ja s'ha vist en diversos articles. No obstant això, cada problema és un desafiament que requereix ser analitzat i posat en context.

El que es busca en aquest projecte és predir el preu de pernoctació i l'ocupació del mes d'agost de 2021 dels apartaments anunciats a la plataforma Airbnb a la ciutat de Barcelona amb el seu respectiu conjunt de dades. La tasca consisteix a utilitzar mètodes de regressió, com el k-NN, el Random Forest, regressió lineal i xarxa neuronal, juntament amb altres tècniques i metodologies, com la selecció de característiques i la mineria de text entre d'altres, per tal d'aconseguir una predicció amb un error raonable.

Un altre aspecte important del projecte és l'anàlisi de sentiment. Es desenvolupa un model d'anàlisi de sentiment per a determinar l'opinió de les diferents ressenyes dels apartaments turístics, establint si una ressenya té un comentari positiu o negatiu. Aquesta informació també s'empra a l'hora d'acomplir la tasca principal.

Els resultats que s'han obtingut mostren la capacitat dels diversos mètodes i tècniques de resoldre problemes de diferents àmbits d'una manera notablement satisfactòria.

Abstract

Predicting certain aspects of real estate through machine learning is a topic that has already been seen in several articles. However, each problem is a challenge that needs to be analyzed and put into context.

This project aims to predict the price of overnight stays and occupancy for August 2021 of the apartments announced on the Airbnb platform in the city of Barcelona with their respective data set. The task is to use regression methods, such as k-NN, random forest, linear regression, and neural network, along with other techniques and methodologies, such as feature selection and text mining, among others, to get a prediction with a reasonable error.

Another important aspect of the project is sentiment analysis. A sentiment analysis model is developed to determine the opinion of the different reviews of the apartments, establishing whether a review has a positive or negative comment. This information is also used when performing the main task.

The results obtained show the ability of the various methods and techniques to solve problems in different areas in a remarkably satisfactory way.

Resumen

La predicción de ciertas características de bienes inmobiliarios mediante aprendizaje automático es un tópico ya visto en varios artículos. Sin embargo, cada problema es un desafío que requiere ser analizado y puesto en contexto.

Lo que se busca en este proyecto es predecir el precio de pernoctación y la ocupación de agosto de 2021 de los apartamentos anunciados en la plataforma Airbnb en la ciudad de Barcelona con su respectivo conjunto de datos. La tarea consiste en utilizar métodos de regresión, como el k-NN, el Random Forest, regresión lineal y red neuronal, junto con otras técnicas y metodologías, como la selección de características y la minería de texto entre otros, con el fin de conseguir una predicción con un error razonable.

Otro aspecto importante del proyecto es el análisis de sentimiento. Se desarrolla un modelo de análisis de sentimiento para determinar la opinión de las distintas reseñas de los apartamentos turísticos, estableciendo si una reseña tiene un comentario positivo o negativo. Esta información también se emplea a la hora de realizar la tarea principal.

Los resultados que se han obtenido muestran la capacidad de los diversos métodos y técnicas de resolver problemas de diferentes ámbitos de forma notablemente satisfactoria.

Agraïments

Vull agrair especialment a la meva mare tot el suport incondicional que m'ha donat al llarg de la meva vida i no ha sigut menys en aquest treball.

A en Pau, en Nico, la Flor, en Robert, l'Aniol i l'Oriol, els meus amics, on em puc recolzar sempre que em calgui.

A en Fran, en Manuel, l'Arnau i l'Aarón, primer companys i més tard amics, amb els que he compartit aquesta experiència.

I per últim, això no obstant, no menys important, al meu tutor del treball final de grau, en Santi, que m'ha ajudat a portar-lo a terme. També agrair a tots els professors del grau d'enginyeria informàtica que han fet que no perdi la passió pel que faig.

Índex

1 Introducció **1**

1.1 Context . . . 1

1.2 Motivació . . . 1

1.3 Objectiu principal . . . 2

1.4 Objectius específics . . . 2

1.5 Planificació . . . 3

2 Context **4**

2.1 Què és Airbnb? . . . 4

2.2 Barcelona i turisme . . . 4

3 Desenvolupament **5**

3.1 Programari emprat . . . 5

3.2 Conjunts de dades . . . 5

3.2.1 Dades Airbnb . . . 5

3.2.2 Dades de la ciutat de Barcelona . . . 6

3.2.3 Dades Tripadvisor . . . 6

3.3 Web scrapping . . . 6

3.3.1 Programari emprat en el Web Scrapping . . . 7

3.3.2 Reconeixement de la pàgina web . . . 7

3.3.3 Identificació dels elements i de les dades . . . 8

3.3.4 Obtenció de les dades . . . 8

3.3.5 Resultats i conclusions de l'us de web scrapping . . . 9

3.4 Anàlisi de les dades . . . 10

3.4.1 Dades InsideAirbnb . . . 10

3.4.2 Dades TripAdvisor . . . 18

3.4.3 Dades Barcelona Open Data . . . 19

3.4.4 Conclusions de l'anàlisi de les dades . . . 21

3.5 Anàlisi de sentiment . . . 22

3.5.1 Ús en el projecte . . . 22

3.5.2 Dades emprades . . . 22

3.5.3 Model d'anàlisi de sentiment . . . 23

3.5.4 Capa Embedding . . . 23

3.5.5 Word2Vec 24
3.5.6 Capa LSTM 26
3.5.7 Implementació i arquitectura del model 28
3.6 Predicció del preu i de l'ocupació 30
3.6.1 Mètodes emprats 30
3.6.2 Models Baseline 31
3.6.3 Feature engineering i selecció de característiques 32
3.6.4 Mineria de text: TF-IDF 37
3.6.5 Models finals 38

4 Resultats **39**
4.1 Entrenament model d'anàlisi de sentiment 39
4.2 Resultats del model d'anàlisi de sentiment 40
4.3 Resultats model de predicció preu i ocupació 42
4.3.1 Mesures d'avaluació 42
4.4 Resultats model Baseline 43
4.4.1 Base ocupació 43
4.4.2 Base preu 45
4.5 Resultats Feature engineering 46
4.5.1 Feature engineering ocupació 46
4.5.2 Feature engineering preu 50
4.6 Resultats mineria de text: TF-IDF 54
4.6.1 Resultats en el model d'ocupació 54
4.6.2 Resultats en el model de preu 54
4.7 Resultats Models finals 55
4.7.1 Resultat model final preu 55
4.7.2 Resultats model final ocupació 58

5 Conclusions i treball futur **61**
5.1 Conclusions 61
5.1.1 Model anàlisi de sentiment 61
5.1.2 Predicció del preu 61
5.1.3 Predicció de l'ocupació 62
5.2 Treball Futur 62

1 Introducció

El tòpic principal d'aquest projecte és l'aprenentatge automàtic. L'Aprenentatge automàtic és un camp de la intel·ligència artificial que està dedicat al disseny, l'anàlisi i el desenvolupament d'algorismes i tècniques que permeten que les màquines evolucionin. Es tracta de crear programes capaços de generalitzar comportaments a partir del reconeixement de patrons o classificació. L'aprenentatge automàtic està relacionat amb el camp de l'estadística, però també coincideix amb els mètodes de construcció de models, o l'aprenentatge estadístic. També hi ha punts de contacte amb la informàtica teòrica. Això és degut a la complexitat computacional dels problemes [23]. Cada cop estem més familiaritzats amb tecnologies basades en l'aprenentatge automàtic que ens fan la vida més fàcil de manera totalment transparent. Alguns camps on s'ha aplicat aquest tipus d'aprenentatge són les aplicacions dedicades al processament del llenguatge natural, als algorismes de cerca, la diagnosi mèdica, la bioinformàtica, la detecció de fraus i la classificació.

En aquest projecte se centra essencialment en l'aprenentatge automàtic supervisat. Aquesta variant consisteix a tenir un conjunt de dades d'entrenament que consta de les variables d'entrada i les variables de sortida corresponents. L'objectiu és crear una funció capaç de predir el valor corresponent a qualsevol objecte d'entrada després d'haver vist una sèrie d'exemples, les dades d'entrenament. Per això, ha de generalitzar a partir de les dades presentades a les situacions no vistes prèviament. Els models d'aprenentatge supervisat es poden agrupar en models de classificació; quan la variable a predir és categòrica (color, tipus d'animal, etc.) i models de regressió; quan la variable és contínua, per exemple, el preu d'un cotxe.

1.1 Context

Aquest projecte s'ha fet en el context d'un grau d'Enginyeria Informàtica i s'han aplicat coneixements i metodologies vistes al llarg de tot l'ensenyament desenvolupats en diferents escenaris.

Caldria destacar l'assignatura d'Aprenentatge Automàtic on es dona una base de les tècniques i coneixements del tòpic en qüestió.

1.2 Motivació

Actualment, l'ús de l'aprenentatge automàtic està present en múltiples àmbits de la nostra vida quotidiana; recomanacions de cançons segons el nostre gust musical, estimació de la durada trajecte en transport públic segons el lloc on ens trobem i un llarg etcètera. No obstant això, sempre hi ha aspectes que l'aprenentatge automàtic pot cobrir i encara no s'han desenvolupat o s'han desenvolupat de manera molt superficial.

La motivació del projecte rau principalment en l'oportunitat d'aprendre noves tècniques de l'àmbit de l'aprenentatge automàtic a més de dur a terme un projecte que resolgui un problema en concret de manera fiable.

El problema ha sigut definit de manera pròpia: esbrinar l'ocupació del mes d'agost de 2021, és a dir, el nombre de dies que un apartament està arrendat, i el preu de pernoctació dels apartaments anunciats a la pàgina web Airbnb a la ciutat de Barcelona. Principalment, aquesta tasca es desenvoluparà emprant dades dels mateixos apartaments, juntament amb mètodes d'aprenentatge automàtic supervisat.

1.3 Objectiu principal

L'objectiu principal del projecte es basa en desenvolupar un conjunt d'eines basades en l'aprenentatge automàtic que ens permeti 1) estimar l'ocupació dels apartaments i 2) predir el preu de la pernoctació de l'apartament. Per tal de resoldre aquest problema es desenvoluparan i validaran dos models de forma independent.

En ser un problema de regressió, ja que, tant l'ocupació de l'apartament i el seu preu són variables contínues, es buscarà reduir l'error entre la predicció dels models i la dada real tant com la qualitat de les dades i dels mètodes emprats ho permeti.

1.4 Objectius específics

Els objectius específics o parcials del projecte són els que s'han dut a terme per tal de completar l'objectiu principal:

- Analitzar els diferents conjunts de dades per saber la distribució de les diverses variables que es poden considerar rellevants a l'hora d'introduir als models.
- Creació de models simples per tal de veure el rendiment dels diferents mètodes amb les dades pertinents, per a tenir una primera fotografia de com enfocar el problema i detectar si hi ha marge de millora.
- Aplicar tècniques pròpies de Feature engineering i selecció de característiques per tal de treure el màxim profit a les dades emprades i aplicar-lo als diferents models.
- Aplicar tècniques de processament de llenguatge natural, com el TF-IDF, a les característiques dels apartaments que continguin text i incrementar el poder de predicció dels models amb una fiabilitat considerable.
- Creació d'un model d'anàlisi d'opinió que analitzi quin sentiment expressen les ressenyes dels apartaments.
- Desenvolupament d'un programari que permeti extreure dades d'una pàgina web en concret.
- Validació dels models creats.

1.5 Planificació

La planificació del projecte ha constat d'aproximadament vint setmanes, concretament 19, començant el mes de setembre de 2021 i acabant el projecte el gener de 2022, així ocupant un quadrimestre sencer del grau. A la figura 1, es pot veure en format de diagrama de Gantt la planificació inicial de les diverses tasques al llarg de cada setmana. És a la figura 2 on es mostra la dedicació real a cada tasca. Com es pot observar, la desviació de temps que hi ha hagut entre la planificació original i la planificació final ha sigut bastant raonable.

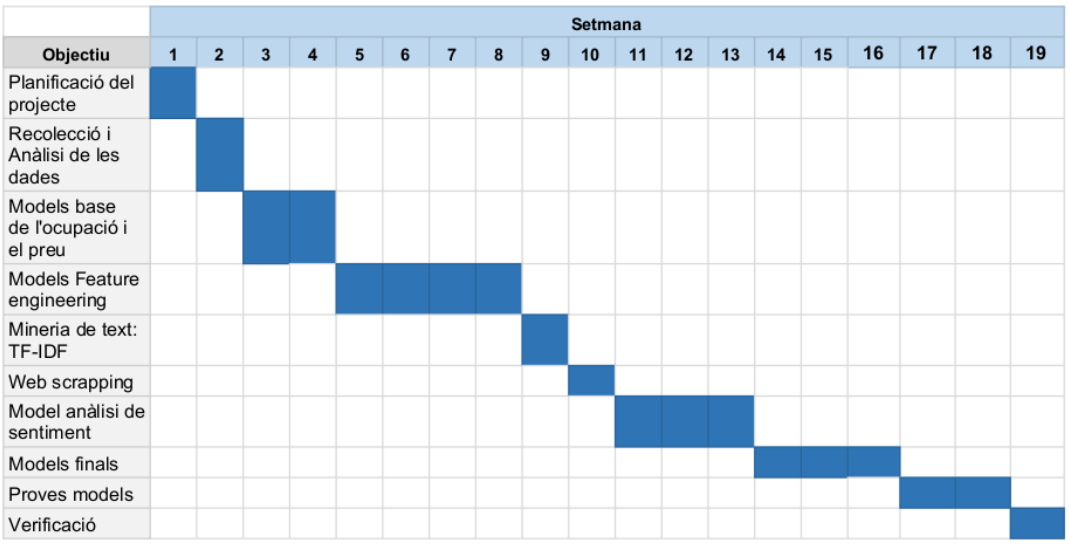

Figura 1: Diagrama de Gantt on s'exposa la planificació inicial del projecte

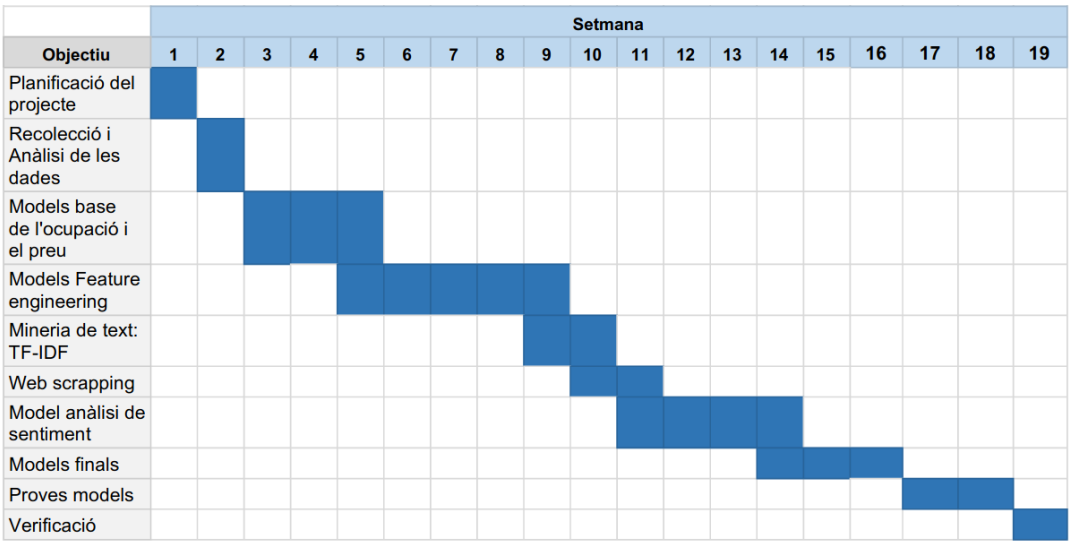

Figura 2: Diagrama de Gantt on s'exposa la inversió final de temps que ha suposat cada tasca del projecte.

2 Context

En aquesta secció s'explica l'idea bàsica darrere Airbnb i el lligam que té la ciutat de Barcelona amb el turisme.

2.1 Què és Airbnb?

Airbnb és una plataforma web que fa d'intermediària entre propietaris d'apartaments i arrendataris que volen llogar-ne un, generalment per fer-ne un ús turístic. Airbnb es pot considerar un exemple d'economia peer-to-peer, ja que, la mateixa plataforma no té en propietat els apartaments anunciats, sinó que en són els mateixos usuaris els que estableixen les relacions comercials per al lloguer dels habitatges.

Airbnb va començar el seu camí el 2008, quan dos dissenyadors que tenien espai lliure a casa van acollir tres viatgers que buscaven un lloc on allotjar-se. Actualment, milions d'amfitrions i viatgers tenen un compte gratuït a Airbnb per poder publicar els seus anuncis i reservar allotjaments únics a tot el món. Els amfitrions d'experiències a Airbnb comparteixen les seves passions i els seus interessos tant amb els viatgers com amb la gent del lloc [1].

2.2 Barcelona i turisme

La ciutat de Barcelona es caracteritza per ser un focus turístic durant tot l'any. Aquest fenomen germina a la celebració dels jocs olímpics del 1992 que van posar en el mapa la Ciutat Comtal després d'una gran remodelació. Va ser l'aparador perfecte perquè després de quasi tres dècades, Barcelona sigui una de les ciutats més visitades per turístes internacionals. Els turistes poden escollir quin tipus d'allotjament lloguen, tradicionalment sempre han sigut hotels, però ara opcions com Airbnb també entren al mercat.

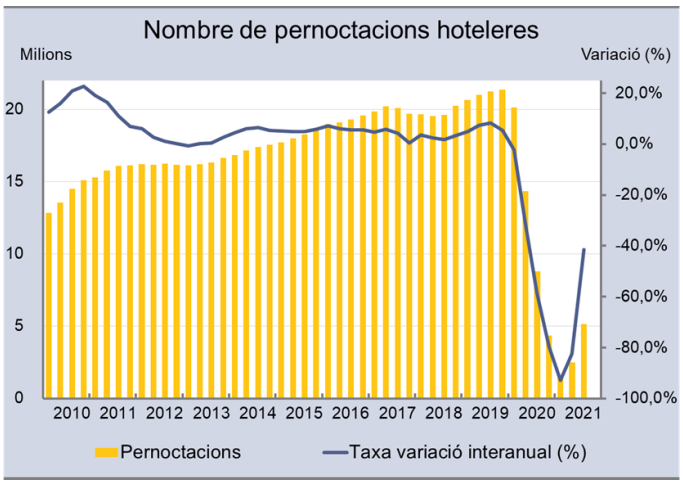

Figura 3: Es pot veure l'afluència de turisme i nombre de pernoctacions en hotels dels últims deu anys. La caiguda es deu als efectes de la pandèmia de Covid-19.

3 Desenvolupament

En l'etapa de desenvolupament es descriuen tots els procediments i tècniques que s'han fet servir per al compliment dels diversos objectius.

3.1 Programari emprat

El projecte s'ha desenvolupat en el llenguatge de programació Python, versió 3.7. S'han fet servir dos entorns de programació. D'una banda, Google Colab, que és un programari de Google, que es pot executar al navegador que consisteix en el format de Notebook de l'estil Jupyter. El seu ús té diversos avantatges, com no tenir dependències de llibreries i tenir la possibilitat d'utilitzar GPU o TPU per tal d'accelerar els entrenaments de xarxes neuronals. D'altra banda, en menor mesura, també s'ha utilitzat un editor de text especialitzat en Python, PyCharm, per a la tasca de *web scrapping*. Per a l'emmagatzematge dels conjunts de dades, es fa servir un repositori de Github públic.

La manipulació de les dades s'ha dut a terme mitjançant la llibreria Pandas (Versió 1.1.5), que és una extensió de la llibreria Numpy (Versió 1.19.5), que ofereix estructures de dades i operacions sobre taules. Per la visualització de les dades s'ha fet servir la llibreria Matplotlib (Versió 3.2.2) i també Seaborn (Versió 0.11.2).

Les llibreries més rellevants per a la creació dels models del projecte han sigut Sklearn [8] i Keras [2]. Sklearn és una llibreria de programari lliure d'aprenentatge automàtic que inclou mètodes de classificació, regressió entre molts altres. Keras és una llibreria que permet la creació de xarxes neuronals. La versió de Sklearn emprada en el projecte és la 1.0.1 i la versió de Keras és la 2.7.0.

3.2 Conjunts de dades

Una de les primeres etapes del projecte consisteix en la cerca i obtenció de dades per tal de resoldre el problema plantejat. A continuació es detallen totes les fonts de dades utilitzades per al desenvolupament del projecte.

3.2.1 Dades Airbnb

Les dades que pertanyen a Airbnb s'han obtingut a través del portal web InsideAirbnb[1], que no està associat a Airbnb. InsideAirbnb considera ciutats que tenen prou rellevància, les que tenen un nombre considerable d'apartaments, i n'extreu les dades de tots els que es troben anunciats a Airbnb.

Els conjunts de dades publicats contenen informació general sobre cada apartament:

- Informació general de l'apartament

[1] http://insideairbnb.com/

- Dades de l'amfitrió
- Ressenyes de l'apartament
- Disponibilitat de l'apartament a un any vista

Les dades s'actualitzen mensualment i es pot accedir a versions anteriors.

3.2.2 Dades de la ciutat de Barcelona

Es fa ús de conjunts de dades que descriuen aspectes de la ciutat de Barcelona que es troben al web de Dades Obertes de Barcelona [2].

- Transport públic
- Restauració

3.2.3 Dades Tripadvisor

Tripadvisor és una plataforma en línia on es recullen milions d'opinions al voltant de determinats negocis del sector turístic (hotels, restaurants, bars, discoteques, creuers, etc.). En aquest projecte es tenen en compte les opinions dels hotels. La raó de l'ús d'aquestes dades va lligat al model d'anàlisi de sentiment que s'explica més endavant.

Per a l'obtenció d'aquestes dades no es disposa de cap repositori on es trobin les dades requerides, per tant, s'hauran d'utilitzar tècniques de *web scrapping* al web de Tripadvisor i d'aquesta manera obtenir una quantitat considerable de dades.

3.3 Web scrapping

Com bé s'ha esmentat a la secció anterior, es disposen de la gran majoria de dades, ja que, estan disponibles en repositoris o webs on es pot descarregar l'arxiu que les conté. No obstant això, algunes dades que són necessàries per al desenvolupament d'alguna part del projecte no es troben disponibles i cal extreure-les directament de la seva font. Aquest és el cas de les ressenyes dels hotels de Barcelona del lloc web Tripadvisor, que s'emprarà *web scrapping* per tal d'extreure aquestes dades.

El *web scrapping* és una tècnica utilitzada mitjançant programari per extreure informació de llocs web. Usualment, aquests programes simulen la navegació d'un humà a una pàgina web ja sigui utilitzant manualment el protocol HTTP, o incrustant un navegador en una aplicació.

L'eina que s'ha desenvolupat és un programa que extreu les dades mitjançant aquesta tècnica i les emmagatzema perquè puguin ser usades com qualsevol conjunt de dades. El major avantatge és l'automatització d'aquest procés, que permet recol·lectar una gran quantitat de dades fàcilment.

[2] https://opendata-ajuntament.barcelona.cat/es/

3.3.1 Programari emprat en el Web Scrapping

Per al desenvolupament del *web scrapping*, s'ha escollit la llibreria de Selenium [13] que ofereix Python. Selenium és un entorn de proves de programari per a aplicacions web, que simula la interacció de l'usuari amb la seva interfície. D'aquesta manera, es poden programar i automatitzar les interaccions necessàries amb la pàgina web.

3.3.2 Reconeixement de la pàgina web

Abans de començar a desenvolupar el programa, cal tenir present quins tipus de dades es requereixen i com s'organitzen dins la pàgina web.

A la secció d'hotels de Tripadvisor es mostra una llista dels hotels que es troben a una ciutat o lloc en concret com es pot veure a la Figura 4.

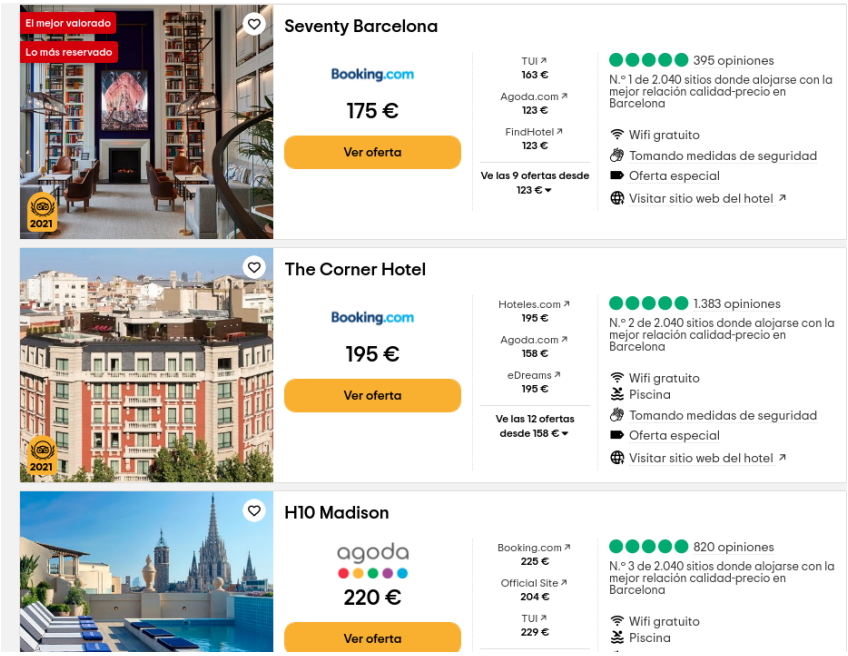

Figura 4: Secció d'hotels a tripadvisor

És quan s'entra a la pàgina d'un hotel on es troben les dades que es consideren rellevants pel projecte, les ressenyes i la puntuació dels hotels (veure Figura 5).

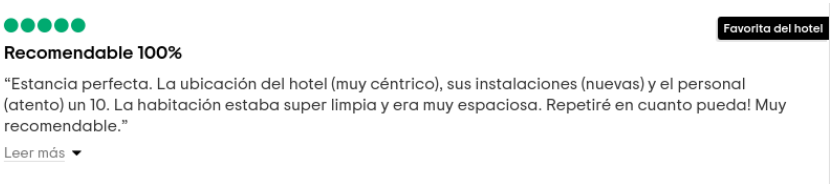

Figura 5: Ressenya d'un hotel a Tripadvisor

Per tant, la tasca del programa de web scrapping no només és la d'extreure dades de la pàgina web, sinó que també s'ha de navegar emprant les diverses funcions que

ofereix la llibreria Selenium.

3.3.3 Identificació dels elements i de les dades

Dins del paradigma de la programació web, la part que l'usuari pot veure és la interfície gràfica que ofereix la web. Aquesta interfície està formada per diversos elements, cadascun d'ells té certes característiques, com una amplada, color o tipus determinat.

Aquests elements es poden identificar mitjançant un identificador únic. Per a definir un grup d'elements amb unes característiques concretes es pot definir una classe. Un element pot tenir múltiples classes associades [6].

Selenium permet localitzar els elements d'una pàgina web mitjançant el seu identificador, nom de la classe o *xpath* entre d'altres. D'aquesta manera, es pot obtenir el contingut d'aquell element o bé realitzar algun tipus d'acció com és fer clic a l'element en qüestió.

Per tal d'indicar a Selenium sobre quins elements cal actuar i poder automatitzar el procés de recollida de dades, primer s'ha inspeccionat la pàgina web i cercat quins són els atributs de cada element que l'identifiquen de manera unívoca. Aquesta tasca es pot dur a terme amb un navegador web com pot ser Google Chrome, accedint a les opcions de desenvolupador (veure Figura 6). Un cop identificats tots els elements, es pot automatitzar la recolecció de les dades.

Figura 6: Inspecció d'elements a una pàgina web mitjançant les eines proporcionades pel navegador. A l'esquerra es troba el contingut renderitzat pel navegador que veu l'usuari, mentre a la dreta el codi HTML corresponent.

3.3.4 Obtenció de les dades

El fet de necessitar una gran quantitat de dades significa processar un nombre elevat d'hotels i les seves ressenyes corresponents.

Per tal de simplificar la navegació dins la pàgina web, és necessari obtenir i emmagatzemar els enllaços de tots els hotels que es processaran. Per realitzar aquesta primera acció cal recordar que en la pàgina web de Tripadvisor, on es troben els hotels (veure Figura 4), només es mostren al voltant de 30 entrades per pàgina, que són insuficients pel volum de dades que es necessita. Caldrà accionar, mitjançant una funció de Selenium, el botó de "Next" per a passar de pàgina i poder obtenir els enllaços dels següents 30 hotels i així successivament.

Un cop aconseguits tots els enllaços, es navega a cadascun dels hotels i es processen les ressenyes dels usuaris. Depenent de l'idioma en què estigui la pàgina web, només s'obtenen les opinions aquell idioma en concret. Per tenir dades en diferents llengües cal realizar clic, amb Selenium, al botó que permet obtindre totes les opinions en tots els idiomes disponibles. Veure figura 7.

Figura 7: Filtres de les opinions de Tripadvisor. Es pot filtrar per puntuació, per temporada de l'any on s'ha escrit la ressenya, tipus de viatger i l'idioma de l'opinió.

A banda de l'idioma, cal tenir en compte que el nombre d'opinions positives, amb una puntuació igual o major que 3 sobre 5, predomina en la gran majoria d'hotels. Per aquest motiu, si no s'apliquen filtres en funció de la puntuació, el nombre d'opinions negatives representarà un percentatge molt baix en el conjunt de dades. Perquè la distribució sigui equiparable, s'empren els filtres accionant els botons (veure Figura 7) de la mateixa manera que en el cas anterior. En la pàgina d'un hotel en concret també tenim la paginació que limita a 5 opinions per pàgina, aquest aspecte es resol de la mateixa manera que en el cas de la pàgina d'hotels.

A mesura que es van processant les dades de la pàgina, el text de l'opinió i la puntuació s'emmagatzema en un fitxer CSV.

3.3.5 Resultats i conclusions de l'us de web scrapping

L'ús de *web scrapping* en aquest projecte ha permès obtenir una gran quantitat de dades, al voltant d'unes 50000 ressenyes, de manera automatitzada. El conjunt de dades conté opinions en diversos idiomes i la distribució de les puntuacions de les ressenyes s'ha volgut equilibrar al màxim possible per a tenir un nombre d'opinions negatives equiparable al d'opinions positives en el conjunt de dades. Aquesta mateixa tècnica es pot aplicar fàcilment en l'extracció de dades d'altres pàgines web.

Cal esmentar que aquests tipus de programes estan subjectes als canvis que es produeixen a les pàgines web, cosa que provoca que el programa s'hagi de mantenir contínuament. En el cas de Tripadvisor, ha canviat la seva estructura mentre es produïa el desenvolupament d'aquest projecte, cosa que pot limitar algunes funcionalitats del *web scapper*, com l'ús del filtratge d'opinions.

3.4 Anàlisi de les dades

L'anàlisi de les dades és un pas essencial per a les diverses tasques que es porten a terme en el projecte. En aquesta secció es fa una inspecció general dels conjunts de dades que es disposen. En els conjunts de dades es tenen generalment dos tipus de variables: les variables numèriques i les variables categòriques. Les numèriques, com bé el seu nom indica, són nombres que comprenen un valor determinat com per exemple, el nombre de llits que disposa l'apartament. Les variables categòriques són tipus de variable que pot prendre un nombre limitat de valors diferents, per exemple, el districte on es troba l'apartament.

3.4.1 Dades InsideAirbnb

El conjunt de dades obtingut mitjançant el portal web InsideAirbnb consisteix de 17.079 registres, apartaments, i 74 columnes que representen cadascuna de les característiques dels apartament. Dins les 74 característiques en troben de númeriques, com el número d'habitacions, banys o preu i categòriques, com el barri o tipus d'apartament entre moltes altres.

A la Figura 8 es pot veure el nombre d'apartaments registrats a Airbnb en cadascun dels districtes de Barcelona. El gruix dels apartaments es concentra als districtes més cèntrics de la ciutat; l'Eixample i Ciutat Vella.

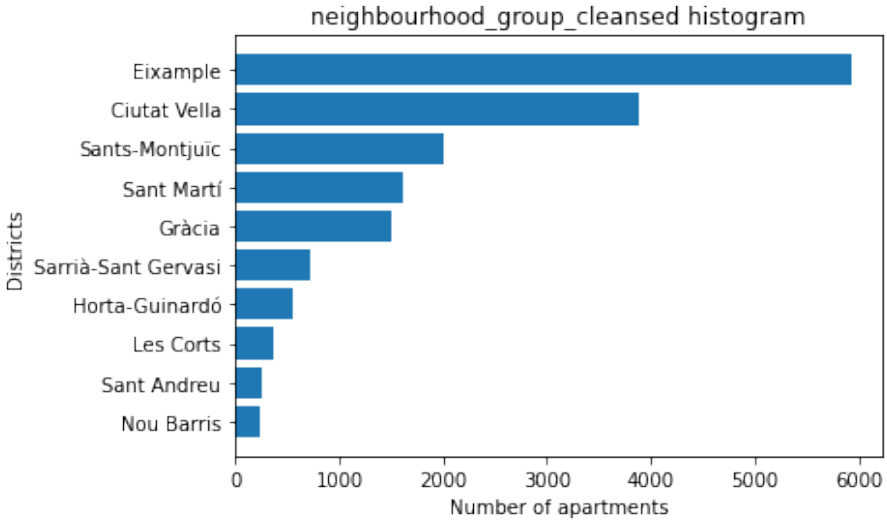

Figura 8: Nombre d'apartaments per districte

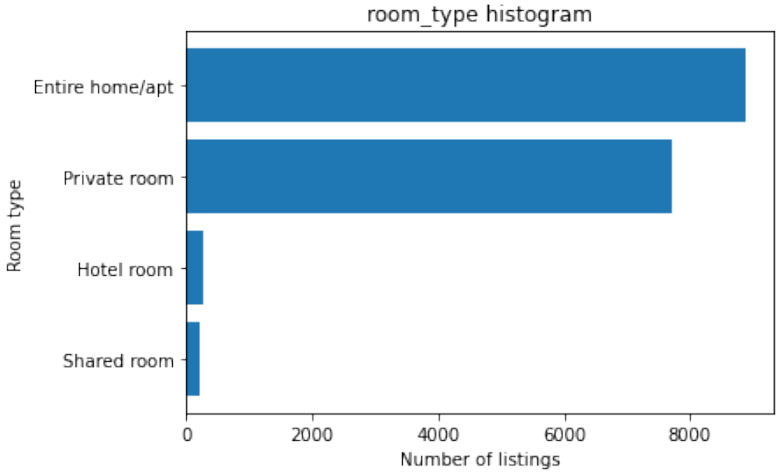

Figura 9: Freqüència del tipus d'habitació que s'anuncia a Airbnb

El tipus d'habitació que predomina a la ciutat és l'apartament complet, seguit de ben a prop per l'habitació privada dins d'un habitatge. Cal remarcar que alguns hotels o hostals també anuncien les seves habitacions a la pàgina web, però suposen un petit percentatge. En el cas de l'habitació compartida, és un cas poc freqüent. Veure la Figura 9.

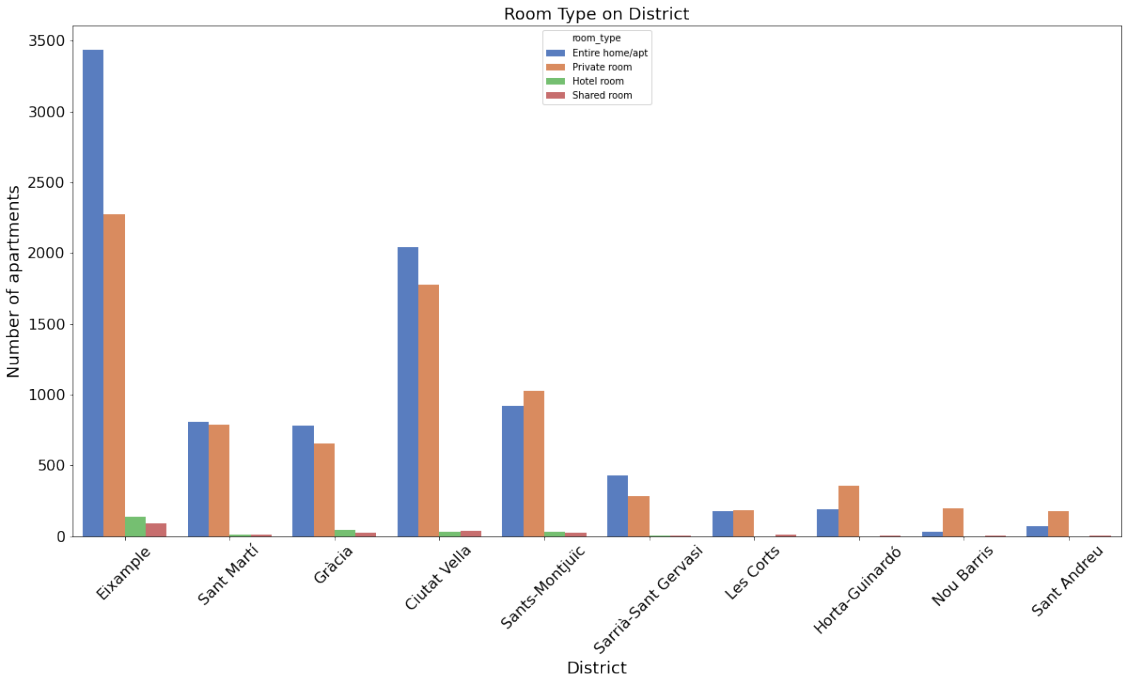

Figura 10: Tipus d'habitació més freqüent per districte

Donant una ullada al tipus d'habitació més freqüent per districte (veure Figura 10), es pot apreciar que als districtes cèntrics, com l'Eixample o Ciutat Vella, l'-habitació més comuna és l'apartament sencer. En canvi, en els districtes que són

menys cèntrics (Horta-Guinardó o Sant Andreu) és l'habitació privada el tipus d'habitació que predomina sobre l'apartament complet. L'habitació d'hotel i l'habitació compartida es concentren sobretot al districte de l'Eixample.

El preu dels apartaments que es troba al conjunt de dades té un rang de 0 a 9999 euros (veure Figura 11). Cal considerar que rarament un apartament es pugui llogar de manera gratuïta. També cal considerar els apartaments que superen els 1000 euros com casos més excepcionals, ja que, el gruix d'apartaments es concentra entre els 10 i 200 euros. Veure la figura 12.

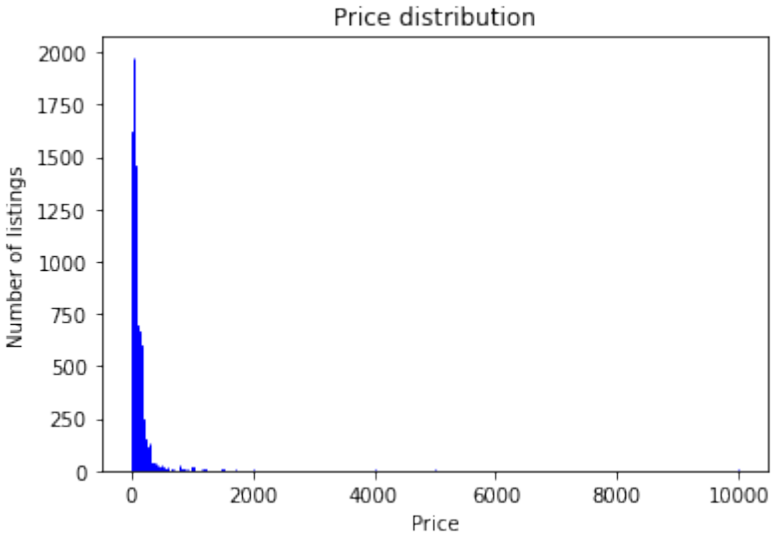

Figura 11: Distribució del preu dels apartaments

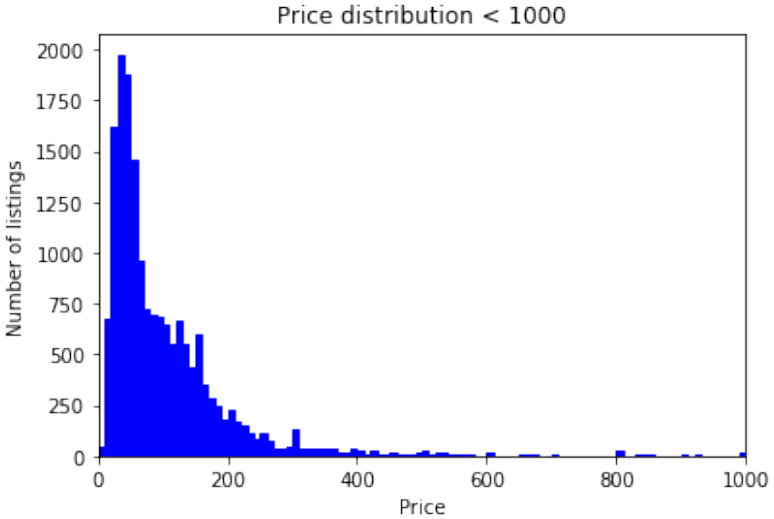

Figura 12: Distribució del preu dels apartaments per sota de 1000 euros

Si es relaciona la localització de l'apartament i el seu preu (veure Figura 13), es pot observar que generalment, el centre de la ciutat i també a alguns sectors del

litoral marítim és on es concentren els preus més elevats. A la figura 14 és pot veure com el preu mitjà dels districtes situats al centre, en aquest cas l'Eixample i Gràcia, és més elevat que el dels districtes perifèrics, com per exemple, Nou Barris i Sant Andreu.

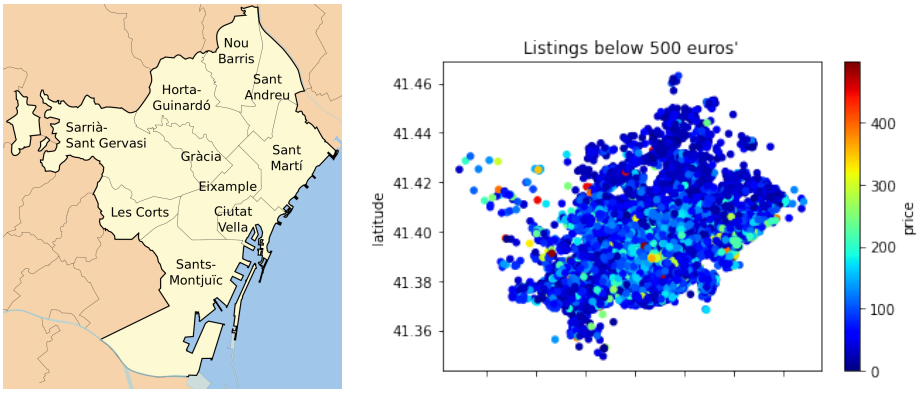

Figura 13: Relació localització de l'apartament i el seu preu. Només es mostren els apartaments amb un preu inferior a 500 euros per una millor visualització.

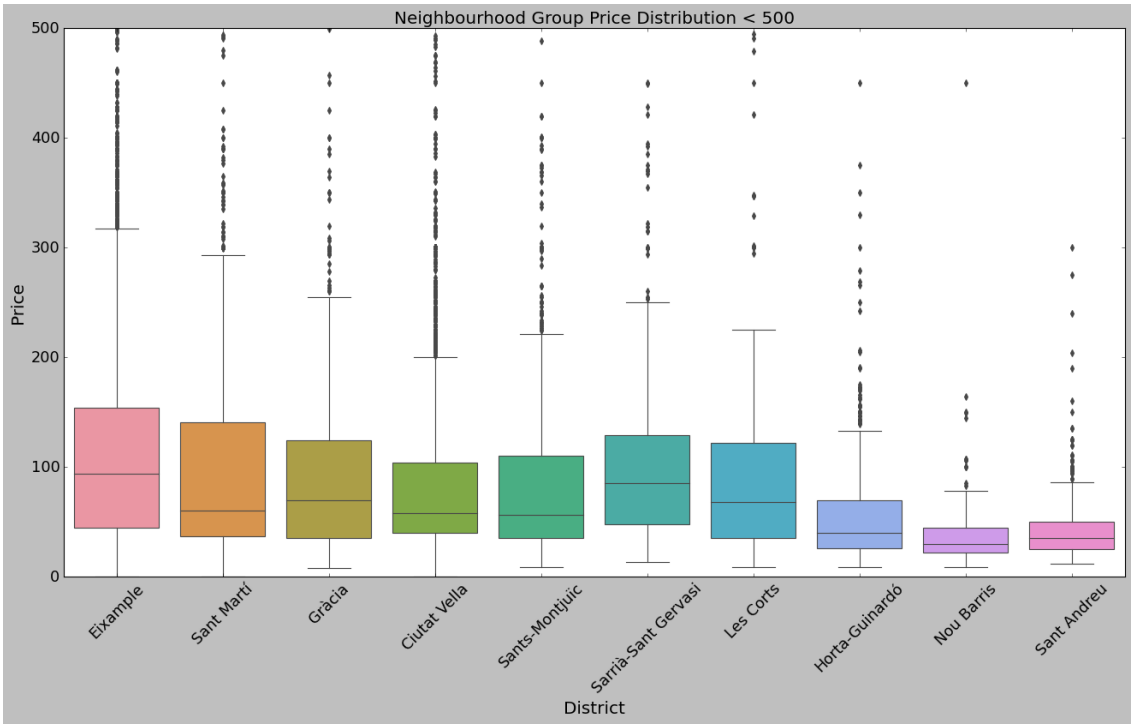

Figura 14: Preu de pernoctació dels apartaments de cada districte. Exceptuant els apartaments de més de 500 euros.

Els apartaments són anunciats pels amfitrions o hosts. La majoria són particulars que lloguen un apartament de la seva propietat, però és un cas comú el fet de trobar

empreses, com immobiliàries, que anuncien els apartaments. A la figura 15 es pot veure aquest fenomen.

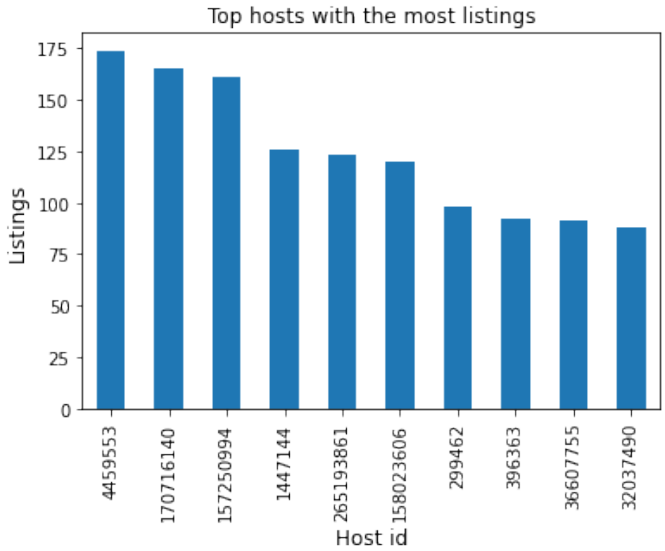

Figura 15: Els 10 amfitrions amb més apartaments anunciats a Airbnb (a la ciutat de Barcelona)

Normalment en tots els conjunts de dades sempre s'hi troben alguns valors que no tenen cap valor, també anomenats Nan o valors nuls. InsideAirbnb empra *web scrapping* per obtenir totes les dades i freqüentment hi ha absència d'algunes dades, per exemple, un apartament que no ha sigut llogat en cap ocasió, no tindrà cap mena de puntuació ni ressenya i el procés de recol·lecció de dades ho deixa com a nul. En el cas del conjunt de dades d'apartaments, la data d'actualització del calendari i el nombre de lavabos pràcticament tenen tots els valors nuls. Veure figura 16.

Com s'ha esmentat anteriorment, Barcelona és una ciutat turística que rep turistes durant tot l'any, però el gruix més gran de visitants s'allotja a la ciutat a la temporada d'estiu. La majoria dels apartaments de Barcelona tenen una ocupació elevada a la temporada estival, com és el cas de l'agost. Com es pot veure a la figura 17, una gran part dels apartaments no es pot reservar, ja que està totalment reservat o bé, l'amfitrió el no permet llogar-lo durant un cert període. Aquest últim cas és minoritari i més probable de trobar-se en temporades amb menys afluència turística.

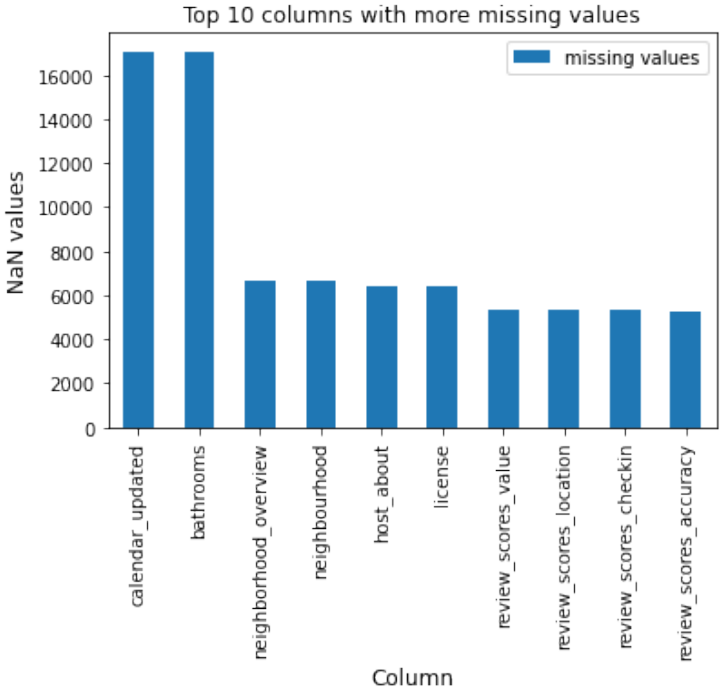

Figura 16: Les columnes del conjunt de dades que tenen més valors nuls o NaN.

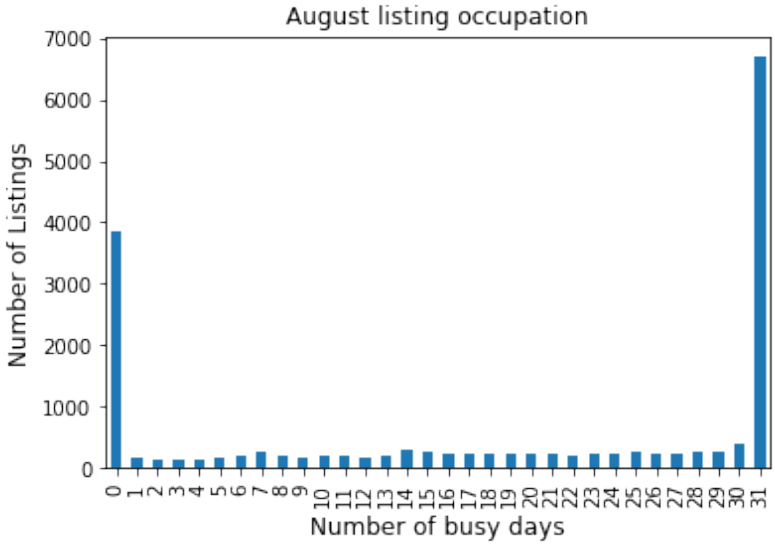

Figura 17: Disponibilitat dels apartaments de Barcelona al mes d'agost.

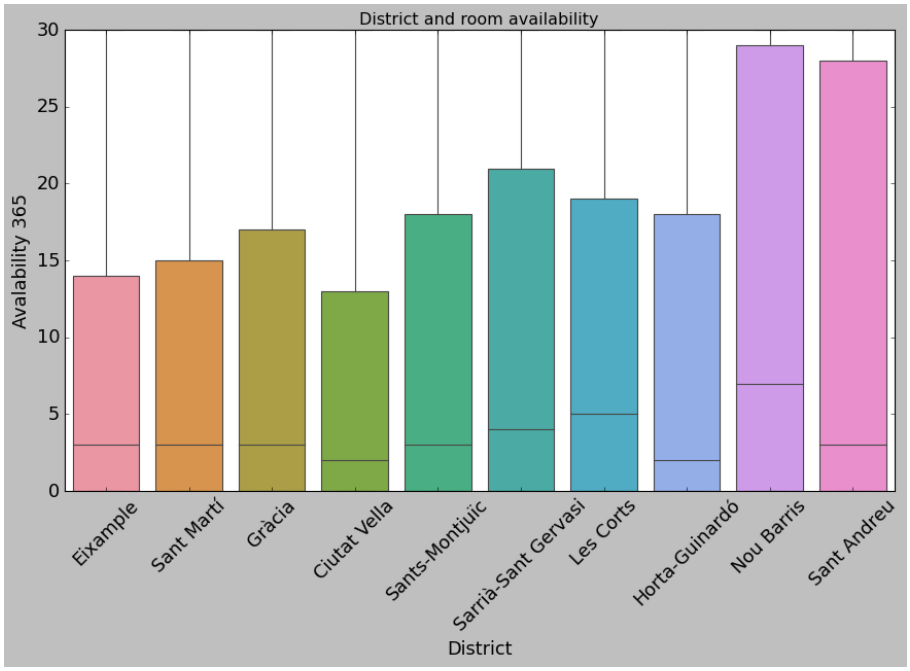

Figura 18: Boxplot on es mostra la mitjana de disponibilitat dels apartaments per cadascun dels districtes de la ciutat a un any vista.

Depenent de la zona de la ciutat, un apartament pot tenir una demanda més alta o més baixa segons la qualitat dels habitatges, atractius de la zona on es trobi, etcètera. Com es pot veure a la figura 18, els districtes més cèntrics són els que tenen una major ocupació com l'Eixample i Ciutat Vella, en canvi, districtes més perifèrics com Nou Barris, tenen una major disponibilitat.

InsideAirbnb també proporciona un conjunt de dades de les ressenyes que fan els usuaris quan s'allotgen a un dels apartaments que hi ha al web. En el cas de Barcelona, s'hi troben al voltant de 500000 opinions. Alguns apartaments tenen un major nombre d'opinions pel fet d'haver estat més temps anunciats a Airbnb i també perquè s'han llogat amb una freqüència superior als altres. A la figura 19 es pot veure que els apartaments amb més ressenyes superen el mig miler d'opinions.

Cada opinió té una puntuació general, que és una mitjana ponderada de diversos ítems, com la neteja de l'apartament, com de ben localitzat es troba, com ha sigut la comunicació amb l'amfitrió, etc. La puntuació mitjana de la majoria dels apartaments és elevada, 4.5 sobre 5, cosa que significa que la gran majoria de clients tenen una experiència satisfactòria.

A banda de la puntuació, trobem la ressenya en qüestió. Amb l'ajuda d'un *wordcloud* o núvol de paraules, es mostren les paraules més utilitzades en les ressenyes dels usuaris. Com bé es pot veure a la figura 20, les ressenyes majoritàriament parlen d'aspectes positius, fent referència a la localització, a alguns monuments emblemàtics de la ciutat i a opinions personals entre d'altres.

Pel que fa a la longitud de les ressenyes (veure Figura 21), el gruix principal té una longitud d'entre 20 i 70 paraules.

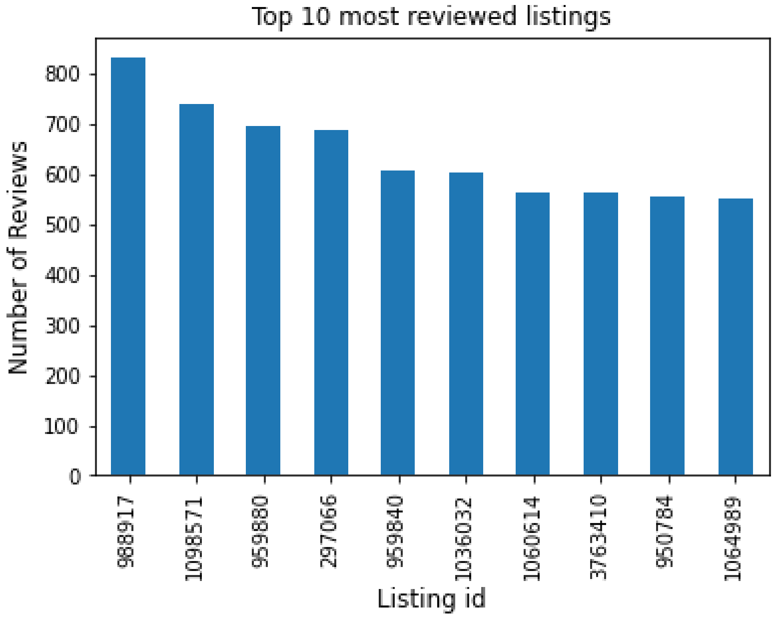

Figura 19: Els 10 apartaments amb més opinions d'Airbnb a la ciutat de Barcelona

Figura 20: Núvol de paraules. Les paraules més emprades en les ressenyes dels usuaris d'Airbnb

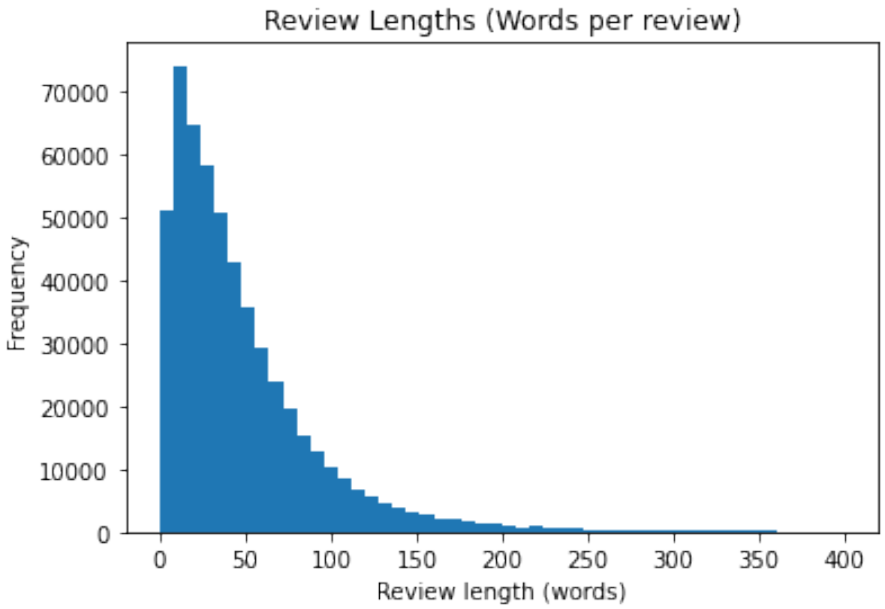

Figura 21: Longitud de les ressenyes dels apartaments d'Airbnb

3.4.2 Dades TripAdvisor

Les dades obtingudes de la pàgina web TripAdvisor mitjançant web scrapping simplement tenen dos dades: la puntuació i la ressenya.

La puntuació mínima és 1 i la màxima 5. Com es pot veure a la figura 22, el nombre de ressenyes positives, considerant com a positives 3, 4 i 5 és major al nombre de ressenyes negatives.

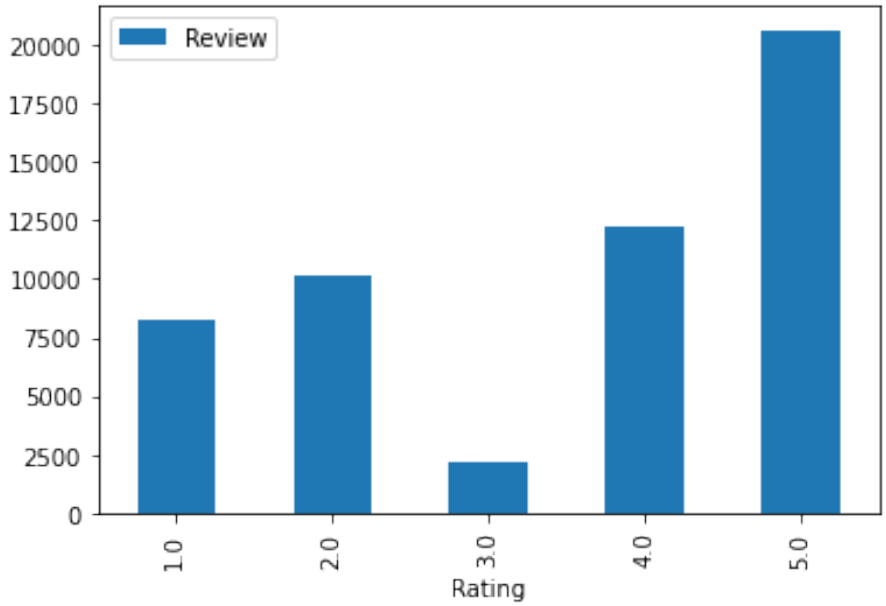

Figura 22: Distribució de la puntuació de les ressenyes de TripAdvisor

El núvol de paraules de les ressenyes de TripAdvisor (veure Figura 23), les paraules més emprades a les opinions dels usuaris de TripAdvisor, parlen essencialment d'aspectes relacionats amb l'hotel, però també s'hi poden trobar paraules referents al personal de l'hotel, la localització, etc. En general, tòpics que són bastant semblants al núvol de paraules de les ressenyes d'Airbnb.

Figura 23: Núvol de paraules del conjunt de dades de TripAdvisor. Les paraules més emprades en les ressenyes dels usuaris.

3.4.3 Dades Barcelona Open Data

Les dades de Barcelona es divideixen en dos conjunts de dades: establiments de restauració i estacions de transport públic. En el cas de transport públic es dona informació sobre les coordenades geogràfiques de cada estació, el tipus de transport, en quin districte i en quin barri es localitza entre d'altres, està format aproximadament per 650 registres i 16 característiques. Al conjunt de dades de la restauració de Barcelona s'hi troba informació sobre tots els establiments d'hostaleria de Barcelona, en aquest conjunt hi han més dades que al conjunt de transport amb aproximadament 2500 registres i 36 característiques. Dels dos conjunts, es pot destacar la informació sobre el tipus de servei que s'hi dona, les seves coordenades geogràfiques i el districte.

A les figures 24 i 25 es pot observar la distribució d'equipaments de transport públic i la localització dels diferents establiments de restauració respectivament.

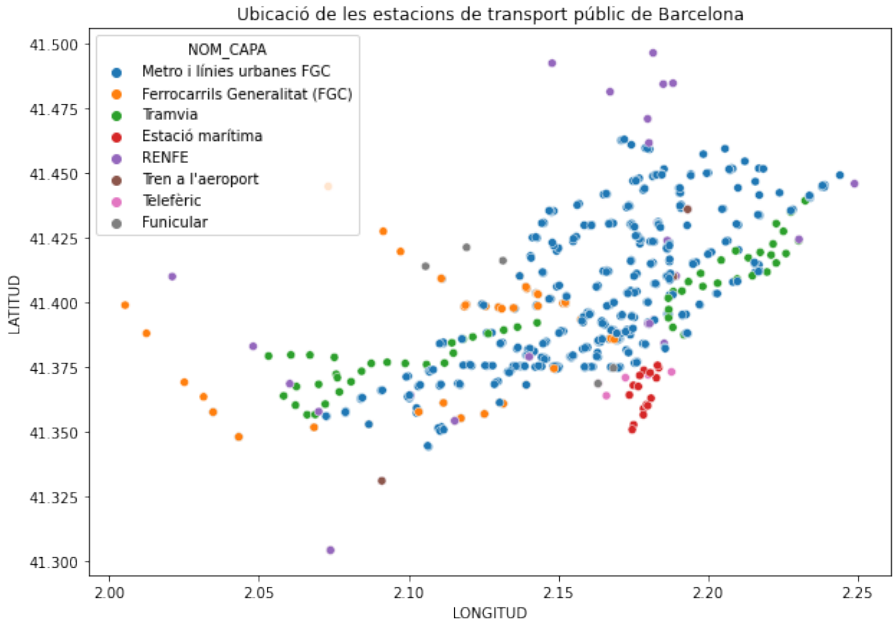

Figura 24: Distribució dels diferents tipus de transport públic a la ciutat de Barcelona

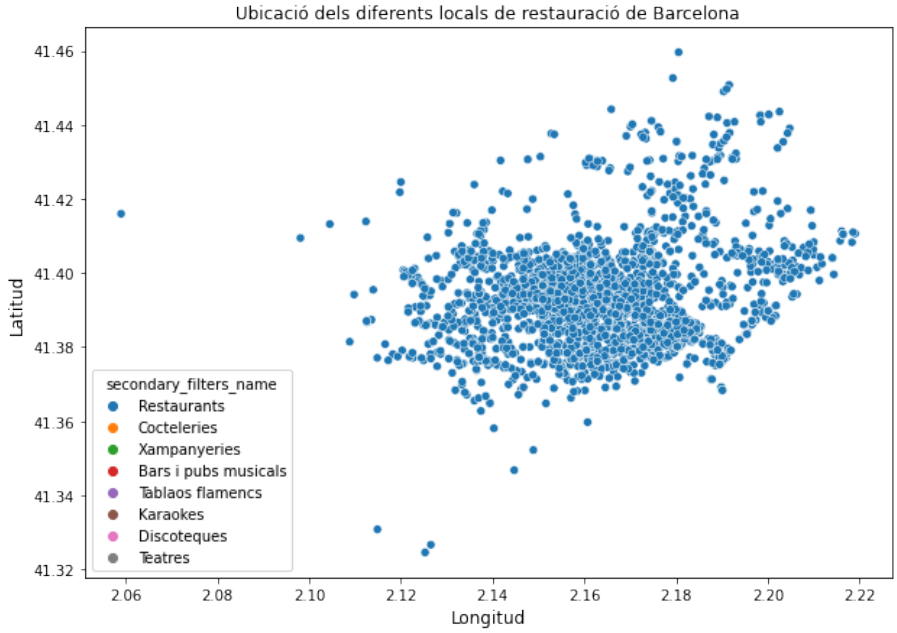

Figura 25: Distribució dels diferents establiments de restauració la ciutat de Barcelona

3.4.4 Conclusions de l'anàlisi de les dades

Un cop fet l'anàlisi de les dades que es disposen en el projecte, es poden treure una sèrie de conclusions sobre la qualitat de cadascun dels conjunts de dades.

En el cas de les dades referents a Airbnb, que són extretes per l'organització InsideAirbnb, en general aporten una gran quantitat d'informació, ja que, tenim dades de com és l'apartament de manera detallada, característiques de l'amfitrió, disponibilitat de l'apartament durant l'any i les ressenyes que hi deixen els usuaris. En alguns casos aïllats es troben errors o incongruències en les dades. Un exemple clar és que alguns apartaments tenen un preu molt elevat, en el rang de 5000 a 9000 euros, i si es revisa l'apartament en qüestió al web d'Airbnb es pot verificar que el preu donat al conjunt de dades és erroni. També es descobreixen casos d'apartaments a zero euros. Més endavant a l'etapa de desenvolupament s'explicarà com es tracten aquests casos. Sobre els valors nuls o NaN d'aquest conjunt de dades, es pot dir que generalment se'n descobreixen de forma freqüent, ja que, es disposen de moltes característiques de l'habitatge i en la majoria de casos és informació que no s'ha afegit encara o bé no existeix, com és el cas de la puntuació d'un apartament quan aquest no ha rebut cap hoste.

Les dades de Tripadvisor extretes mitjançant el *web scrapping* desenvolupat en el projecte és menys complex verificar la seva validesa, ja que, només consten de les columnes on s'hi troben la ressenya i la puntuació respectivament. No hi ha cap valor absent cosa, que permet aprofitar tot el conjunt de dades sencer. Finalment, s'aconsegueix una distribució on s'equilibra el nombre de ressenyes positives amb el nombre de ressenyes negatives, a més de tenir ressenyes de múltiples idiomes.

Les dades referents al transport públic i la restauració de Barcelona poden representar d'alguna manera la centralitat de les diferents zones de la ciutat. S'hi troben moltes característiques amb valors nuls, cosa que les úniques dades útils poden ser les coordenades i el tipus de transport o de local en cas de la restauració.

3.5 Anàlisi de sentiment

L'anàlisi de sentiment fa referència a l'ús del processament de llenguatge natural, anàlisi de text i lingüística computacional per identificar i extreure informació subjectiva de certa informació, com pot ser text o veu entre d'altres. L'anàlisi de sentiment es fa servir àmpliament en el camp de les opinions i xarxes socials amb una gran gamma d'aplicacions, com per exemple, analitzar l'opinió d'un client que ha comprat un producte, la satisfacció dels votants envers un determinat partit polític, etc.

En termes generals, l'anàlisi de sentiment intenta determinar l'actitud un interlocutor, parlant o escrit, respecte a algun tema o a la polaritat contextual global d'un document. L'actitud pot ser el seu judici o avaluació, estat afectiu (és a dir, l'estat emocional de l'autor en el moment de parlar o escriure), o la intenció comunicativa emocional (és a dir, l'efecte emocional que l'autor pretén causar en el lector)[18].

3.5.1 Ús en el projecte

En aquest projecte s'ha desenvolupat un model d'anàlisi de sentiment per tal d'esbrinar el sentiment, positiu o negatiu, de les opinions que escriuen els usuaris que s'han allotjat als diferents apartaments anunciats a Airbnb.

3.5.2 Dades emprades

Els conjunts de dades emprats per desenvolupar aquesta tasca són les opinions dels hotels anunciats a Tripadvisor obtingudes mitjançant *web scrapping* i el conjunt de dades de les opinions dels apartaments d'Airbnb que ofereix InsideAirbnb.

La raó de l'ús de les dades de Tripadvisor com a conjunt d'entrenament del model rau en el fet que les ressenyes d'Airbnb disposen del text de l'opinió, però no de la puntuació atorgada, de la mateixa manera que al web d'Airbnb no apareix la puntuació individual de cada ressenya, sinó que es mostra una puntuació mitjana general de l'apartament de totes les ressenyes que ha rebut.

No s'han utilitzat altres conjunts de dades més coneguts com el famós conjunt de dades de Twitter, Sentiment140, o el conjunt de dades d'opinions de pel·lícules la pàgina web IMDB perquè són dades majoritàriament en anglès, mentre que al conjunt d'opinions d'Airbnb ens trobem amb múltiples idiomes, fet que pot dificultar l'aprenentatge del model.

L'ús de les dades de Tripadvisor és apropiat perquè tenim dades en diferents idiomes i les opinions són d'un àmbit similar a les ressenyes dels apartaments d'Airbnb, és a dir, parlen de l'experiència que ha tingut l'usuari en un hotel, en quin estat ha trobat l'habitació, com l'ha tractat el personal de l'hotel, etc.

3.5.3 Model d'anàlisi de sentiment

Per desenvolupar la tasca de classificació de sentiment es construeix una xarxa neuronal. Les xarxes neuronals són una classe d'algorismes d'aprenentatge automàtic que s'utilitzen per modelar patrons complexos en conjunts de dades mitjançant múltiples capes ocultes i funcions d'activació no lineals. Una xarxa neuronal pren una entrada, la passa a través de múltiples capes de neurones amagades i emet una predicció que representa l'entrada combinada de totes les neurones. Les xarxes neuronals s'entrenen de manera iterativa fent servir tècniques d'optimització com el descens del gradient. Després de cada cicle d'entrenament, es calcula l'error entre la predicció i l'objectiu. Les derivades d'aquesta mètrica d'error es calculen i els errors es propaguen a través de la xarxa mitjançant una tècnica anomenada *back-propagation*. Aleshores, els coeficients (pesos) de cada neurona s'ajusten en relació amb quant van contribuir a l'error total. Aquest procés es repeteix iterativament fins que l'error de xarxa cau per sota d'un llindar acceptable o bé l'entrenament arriba al seu fi. [7].

La gran majoria de xarxes neuronals dedicades a àmbit de l'anàlisi d'opinió tenen una arquitectura com la que es pot veure a la Figura 26.

Figura 26: Arquitectura d'aprenentatge profund comú en els problemes d'anàlisi de sentiment.

3.5.4 Capa Embedding

El que s'anomena *Word Embedding* és un conjunt de tècniques pròpies del tractament de llenguatge natural i dels models de llenguatge on les paraules es representen mitjançant vectors de nombres reals. Per tant, dues paraules que tenen el mateix significat, o semblant, tenen una representació similar. En altres paraules, representa paraules en un sistema de coordenades on les paraules relacionades són més properes [11]. Veure figura 27.

La capa Embedding [16] s'utilitza principalment en aplicacions relacionades amb el processament de llenguatge natural, com ara el modelatge del llenguatge, però també es pot utilitzar amb altres tasques que involucren xarxes neuronals. Mentre es tracten problemes de NLP, es poden emprar *word embeddings* prèviament entrenats amb mètodes com ara GloVe o Word2Vec. L'entrada d'aquesta capa requereix que cadascuna de les paraules sigui codificada amb un enter, aquest pas es pot realitzar mitjançant el Tokenizer que també ofereix la llibreria Keras.

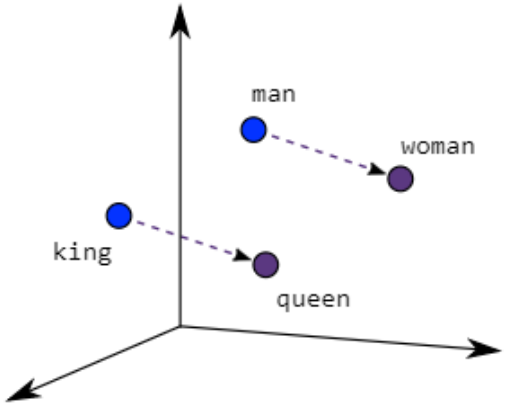

Figura 27: Exemple de *word embedding*, on la paraula 'man' és propera a la paraula 'woman', ja que, aquestes dues paraules es poden trobar en contextos gairebé idèntics. El mateix passa amb els mots 'king' i 'queen'.

3.5.5 Word2Vec

Aquest model va ser creat per Google l'any 2013 i és un model predictiu basat en l'aprenentatge profund per calcular i generar representacions vectorials denses, distribuïdes i contínues d'alta qualitat de paraules, que capturen la similitud contextual i semàntica [4]. El Word2Vec té la capacitat d'aprendre associacions entre paraules a partir d'un gran corpus. Un cop entrenat, aquest model pot detectar paraules sinònimes o suggerir paraules addicionals per a una frase parcial.

L'arquitectura de la xarxa neuronal del Word2Vec és simple. Consisteix en una capa d'entrada, una sola capa oculta i una capa de sortida (veure figura 28) i com totes les xarxes neuronals, té pesos, i durant l'entrenament, el seu objectiu és ajustar aquests pesos per reduir una funció de pèrdua. Però en el cas de Word2Vec només s'agafen els pesos ocults i s'empren com a word embeddings. En altres àmbits de l'aprenentatge automàtic es troben arquitectures similars a aquesta, un clar exemple en són els autoencoders.

La capa d'entrada no pot emprar les paraules en el seu format com a cadena de caràcters, per tant, cal que mitjançant One Hot Encoding s'identifiqui cada paraula amb un vector de la mateixa longitud que el vocabulari, ple de zeros excepte a l'índex que representa la paraula que volem representar, al qual se li assigna "1". La capa de sortida també tindrà tantes unitats com el vocabulari, que representarà les probabilitats de cadascuna de les paraules que siguin veïnes de la paraula representada a l'entrada. La capa oculta és una capa Dense, totalment connectada, on es troben els *word embeddings*.

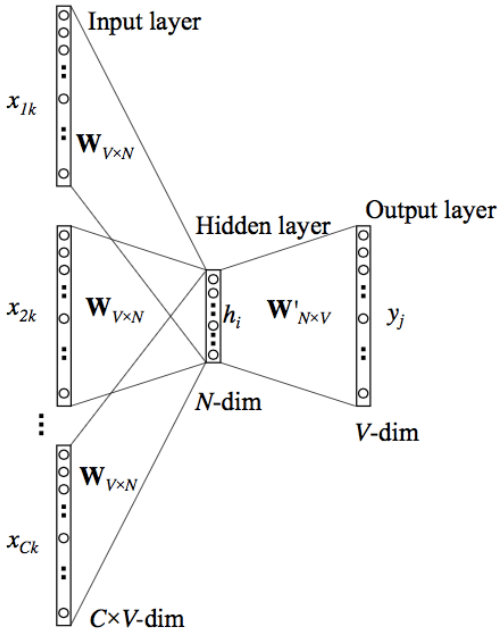

Figura 28: Arquitectura de la xarxa neuronal del Word2Vec, en aquest cas, el model CBOW.

Word2Vec té dos arquitectures diferents a l'hora de crear els *word embeddings*: el model CBOW (Continuous Bag of words) i el model Skip-Gram.

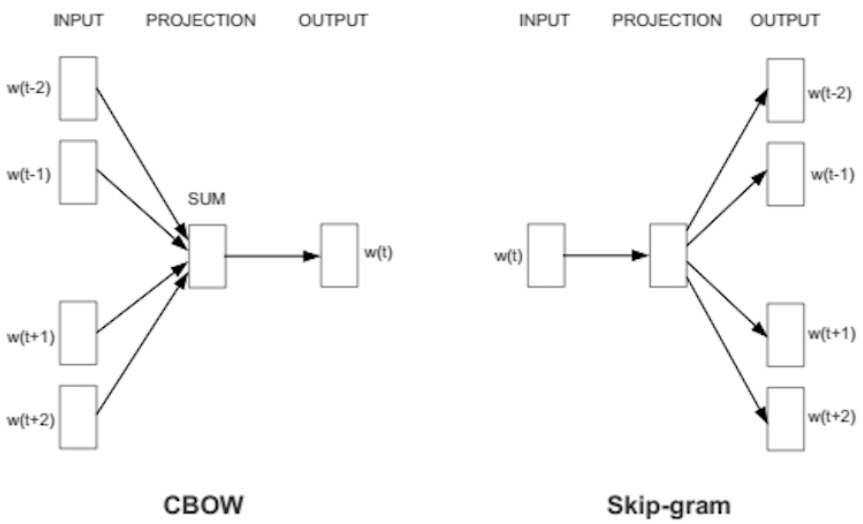

Figura 29: Arquitectures dels dos models que es troben a Word2Vec.

El model CBOW és un model d'aprenentatge profund que agafa un conjunt de paraules X, també anomenat context, com a entrada per predir una paraula y, l'objectiu. El context és definit per una finestra, és a dir, si el valor de la finestra = 2, el context tindrà parells de dues paraules. Donada la següent frase -"La ràpida guineu marró salta per sobre del gos mandrós", s'obtindran parells de dues paraules

amb l'estructura (finestra de context, objectiu), els parells serien ([ràpida, marró], guineu), ([del, mandrós], gos), etc.

El model Skip-Gram, a diferència del CBOW, utilitza com entrada la paraula objectiu per predir el context, en model també es té en compte la finestra. Considerant la mateixa frase -"La ràpida guineu marró salta per sobre del gos mandrós" i una finestra = 2, els parells que es generarien serien de l'estructura (objectiu, finestra de context). En aquest cas els parells obtinguts serien (guineu, [ràpida, marró]), etc.

En el projecte s'ha optat per emprar els pesos generats pel model Word2Vec i transferir-los a la capa Embedding.

3.5.6 Capa LSTM

Les xarxes neuronals convencionals no tenen la capacitat de mantenir informació prèvia i això suposa una limitació en alguns casos. Les xarxes neuronals recurrents (també conegudes com a RNN) són especialment útils a l'hora de modelar dades seqüencials, que empren informació prèvia.

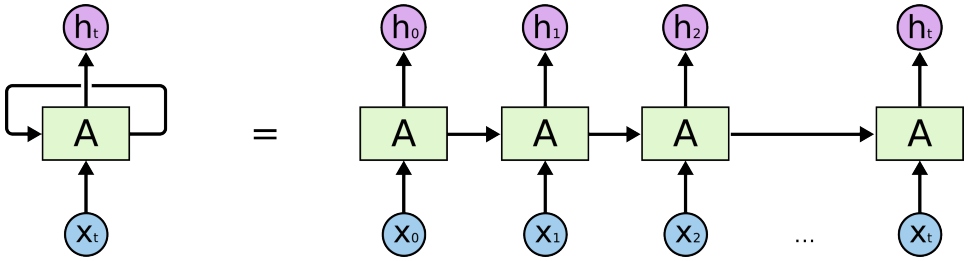

Figura 30: Esquema RNN

A la figura 30, la xarxa neuronal A rep una entrada x_t i emet un valor h_t. Un bucle permet passar informació d'un pas de la xarxa al següent. Aquest bucle és el que s'anomena *hidden state* o estat ocult. Aquesta naturalesa semblant a un bucle revela que les xarxes neuronals recurrents estan íntimament relacionades amb seqüències i llistes.

Encara que a la teoria les RNNs tinguin la capacitat de connectar informació prèvia a la tasca actual, a la pràctica és més complex depenent del problema que s'hagi de resoldre. A vegades, només es necessita mirar la informació recent per acomplir la tasca actual. Per exemple, si es considera un model de llenguatge que intenta predir la paraula següent a partir de les anteriors. Si s'intenta predir l'última paraula a "els núvols són al cel", no es necessita més context; és bastant obvi que la següent paraula serà çel". En aquests casos, on la bretxa entre la informació rellevant i el lloc que es necessita és petita, els RNN poden aprendre a utilitzar la informació prèvia. Veure figura 31.

Però també hi ha casos en què es necessita més context. Si s'intenta predir l'última paraula del text "Vaig créixer a França... Parlo francès amb fluïdesa". La informació recent suggereix que la paraula següent és probablement el nom d'una

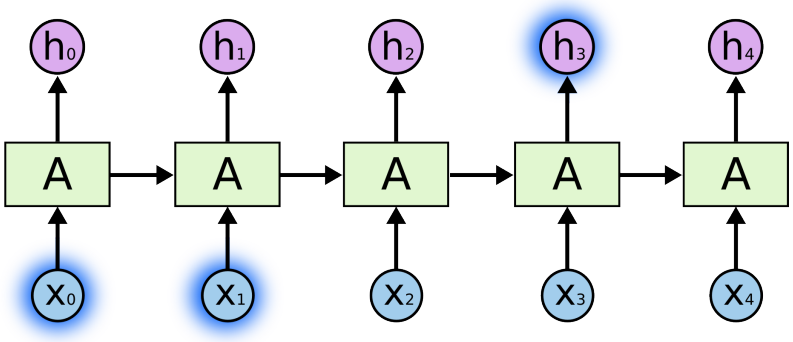

Figura 31: Short term dependency

llengua, però si es vol esbrinar quina llengua, es necessita el context de França, més enrere en la seqüència. És totalment possible que la bretxa entre la informació rellevant i el punt on es necessita es faci molt gran. Veure figura 32.

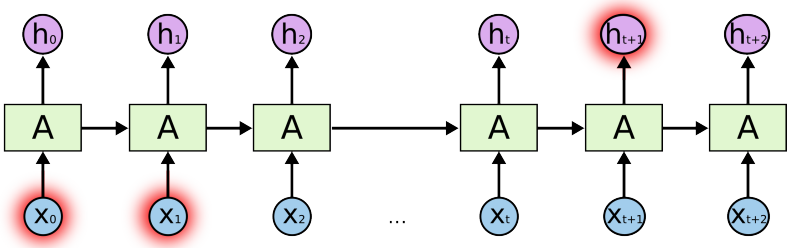

Figura 32: Long term dependency

Aquesta limitació és el problema de les dependències a llarg termini, que les capes LSTM el resolen [15].

3.5.7 Implementació i arquitectura del model

Les dades que s'empren per a l'entrenament, el conjunt de dades de Tripadvisor, conté els diferents comentaris i les puntuacions corresponents com bé s'ha mencionat anteriorment. En voler predir el sentiment d'un comentari, cal definir que es considera una opinió positiva, que s'identifiquen amb un 1, i que es considera una opinió negativa, identificades amb un 0; es defineix com una opinió positiva les ressenyes que tenen una puntuació igual o major que 3 sobre 5. Les opinions negatives tenen una puntuació inferior a 3 sobre 5. El text de les opinions cal que sigui processat per tal de només treballar amb paraules, els caràcters de puntuació, i caràcters especials com emoticones són eliminats.

El model Word2Vec es pot usar utilitzant la implementació que es proporciona a la llibreria gensim [10]. Aquesta classe té una sèrie de paràmetres de configuració que poden ser modificats per tal d'escollir quin model es fa servir, CBOW o Skip-Gram, la finestra de paraules, dimensió del vector, etc. L'entrenament del Word2Vec es fa sobre totes les ressenyes, definint un nombre d'epochs. D'aquest model, com bé s'ha esmentat anteriorment, només s'utilitzen els *word embeddings* resultants (pesos de la xarxa neuronal del W2V) amb els quals es fa una matriu.

La xarxa neuronal es construeix mitjançant la llibreria Keras. La primera capa (entrada) és una de tipus Embedding, on s'especifiquen les dimensions de l'entrada (longitud del vocabulari i la dimensió del vector) i també s'introdueixen els pesos del model Word2Vec previament entrenat, que és la matriu de *word embeddings*.

Segueix una capa LSTM Bidireccional amb 64 unitats, la diferència d'aquesta capa envers la LSTM unidireccional rau en l'estat ocult. La LSTM bidireccional connecta dues capes ocultes en direccions oposades a la mateixa sortida. D'aquesta manera, la capa de sortida pot obtenir informació d'estats passats i futurs simultàniament.

La capa prèvia a la capa de sortida, és una capa Dense o totalment connectada de 32 unitats amb una activació relu. La sortida és una capa Dense amb dues unitats de sortida amb una activació softmax. L'activació softmax dona una distribució de probabilitats on en aquest cas, es determinarà quin és el sentiment de la ressenya en qüestió.

Per evitar el sobreajustament de la xarxa neuronal, és convenient utilitzar tècniques com *dropout* i regularització dels pesos que ajuden a evitar aquest problema. El *dropout* consisteix a ometre de manera aleatòria un percentatge de neurones de la xarxa durant l'entrenament [14].

Com més temps entrenem la xarxa, més especialitzats seran els pesos de les dades d'entrenament, sobreajustant les dades d'entrenament. Els pesos augmentaran de mida per tal de gestionar les particularitats dels exemples que es veuen a les dades d'entrenament, fent que la xarxa sigui inestable. Tot i que els pesos estaran especialitzats en el conjunt de dades d'entrenament, una variació menor o un soroll estadístic a les entrades esperades donarà lloc a grans diferències en la sortida. La regularització dels pesos es basa a afegir un nou terme a la funció de pèrdua que penalitza els pesos grans per tal de controlar el seu creixement. Hi ha diversos tipus

de regularització, en aquest cas s'ha utilitzat l'L2, que és la suma dels quadrats dels pesos (1).

$$x = (\sum_{j=i}^{n} ||w^{[j]}||^2)\frac{\lambda}{2m} \tag{1}$$

On n és el nombre de capes de la xarxa, $w^{[j]}$ és la matriu de pesos per la j capa, m és el nombre d'entrades i λ és el paràmetre de regularització. Aquest paràmetre de regularització λ es pot modificar mitjançant l'API de Keras. Si el valor de lambda és gran, més pesos decauran propers a zero.

Per tal d'avaluar el rendiment del model, es presta atenció al comportament de la funció de pèrdua i l'exactitud de la predicció com mètrica d'avaluació. La funció de pèrdua emprada en el model és la *sparse categorical cross entropy*, ja que, les etiquetes del tipus de sentiment s'etiqueten amb un enter, segons el tipus de sentiment.

Abans que el model faci la predicció del sentiment sobre les ressenyes d'Airbnb, s'avalua el rendiment del model sobre un subconjunt de ressenyes que han sigut etiquetades i estan en diferents idiomes. Un cop validat, es realitza la predicció per tot el conjunt de ressenyes d'Airbnb i s'emmagatzema tota la informació rellevant (identificador de l'apartament, ressenya en qüestió i sentiment) en un fitxer csv.

3.6 Predicció del preu i de l'ocupació

En aquesta secció s'explica el desenvolupament dels models de predicció del preu de pernoctació i la seva ocupació dels apartaments anunciats a la pàgina web Airbnb a la ciutat de Barcelona.

3.6.1 Mètodes emprats

- **La regressió lineal** intenta modelar la relació entre dues o més variables ajustant una equació lineal a les dades observades. Una variable es considera una variable explicativa i l'altra es considera una variable dependent. Una recta de regressió lineal té una equació de la forma $Y = a + bX$, on X és la variable explicativa i Y és la variable dependent. El pendent de la recta és b, i a és la intercepció (el valor de Y quan $X = 0$). Veure figura 33. El mètode més comú per ajustar una recta de regressió és el mètode dels mínims quadrats. Aquest mètode calcula la línia més adequada per a les dades observades minimitzant la suma dels quadrats de les desviacions verticals de cada punt de dades a la línia (si un punt es troba exactament a la línia ajustada, aleshores la seva desviació vertical és 0). Com que les desviacions primer es s'eleven al quadrat i després se sumen, no hi ha cancel·lacions entre els valors positius i negatius.[5].

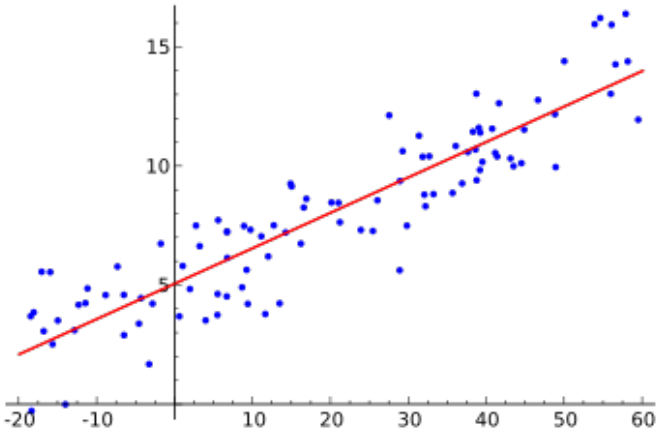

Figura 33: Exemple d'una regressió lineal amb una variable dependent i una variable independent

- **El k-NN** és un mètode que s'utilitza per tasques de regressió i de classificació. En la regressió k-NN, la sortida és el valor de la propietat de l'objecte. Aquest valor és la mitjana dels valors dels k veïns més propers. Els exemples d'entrenament són vectors en un espai de característiques multidimensionals, cadascun amb una etiqueta de classe o valor de l'objecte en qüestió. La fase d'entrenament de l'algorisme consisteix només a emmagatzemar els vectors de característiques i les variables a predir de les mostres d'entrenament. En la fase de predicció, k és una constant definida per l'usuari i un vector sense

etiquetar (una consulta o un punt de prova) es classifica assignant l'etiqueta que és més freqüent entre les k mostres d'entrenament més properes a aquest punt de consulta. Veure figura 34. La mètrica de distància que s'utilitza normalment és la distància Euclidiana [22].

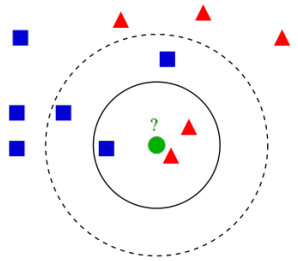

Figura 34: Exemple on es proposa a quina classe pertany una mostra de test (punt verd). Si el valor de k = 3, el valor de la mostra serà un triangle vermell, ja que, la majoria dels seus veins més propers són dos triangles vermells. En canvi, si el valor de k = 5, se li assignarà a la mostra el quadrat blau.

- **El random forest** està composat per múltiples arbres de decisió, motiu pel qual opera com un ensemble [20]. Cadascun dels arbres del bosc fa una predicció i en el cas d'una tasca de regressió, es retorna la mitjana de les prediccions dels diferents arbres. Cada arbre de decisió, en principi, no utilitza totes les dades d'entrenament, sinó que cada arbre agafa una porció. Una mostra d'entrenament pot ser emprada per més d'un arbre, d'això se'n diu Boostrapping. A escala de node no s'usen totes les característiques, sinó que només s'empren subconjunts, d'aquesta manera s'evita incloure característiques que tenen un potencial predictiu més elevat. Com més gran sigui la diversitat d'arbres, millor: reduïm la variància i obtenim un model de millor rendiment [9].

- En el model final, també s'usa una **xarxa neuronal**. El concepte s'ha explicat anteriorment al model d'anàlisi de sentiment. L'arquitectura de la xarxa neuronal per resoldre aquest problema s'explica més endavant.

3.6.2 Models Baseline

La primera iteració del desenvolupament de cada model per predir el preu i l'ocupació es basa en un model experimental que és simple i interpretable, d'ara endavant, anomenat *baseline* o base. El *baseline* dona una visió general de com són d'útils les dades que s'estan utilitzant i en quins àmbits es pot millorar el desenvolupament del model.

En el model base de l'ocupació s'empra el conjunt de dades que té tota la informació general dels apartaments (també anomenat *listings*) i el conjunt de dades referent a la disponibilitat de l'apartament a un any vista (també anomenat *calendar*). L'ocupació del mes d'agost de 2021, que és la variable a predir, es troba filtrant en el conjunt de dades de *calendar*. En el model base del preu, només es fan servir les dades de *listings*.

Pel que fa a les característiques de cada apartament, com s'ha vist a la secció d'anàlisi de dades, es disposen de variables numèriques i de variables categòriques. Les numèriques es poden utilitzar directament en el model, a diferència de les categòriques, ja que, poden ser cadenes de caràcters, booleans, etc. Per a poder usar les variables categòriques cal crear variables anomenades *dummy* mitjançant One hot encoding. Aquest tipus de variable només compren dos valors, 0 o 1, i se'n creen tantes com valors diferents comprèn la variable categòrica a la qual fan referència.

La selecció de característiques és el procés de reduir el nombre de variables d'entrada quan es desenvolupa un model predictiu. Per això, és necessari reduir el nombre de variables d'entrada tant per reduir el cost computacional del modelatge com, en alguns casos, per millorar el rendiment del model. La selecció de les característiques que es realitza en un model base rarament esdevé el conjunt de característiques que dona els millors resultats, no obstant això, és útil per saber quines característiques tenen i quines no un pes considerable a l'hora de predir la variable en qüestió.

Per tal de definir el conjunt de característiques per predir l'ocupació o el preu cal entendre bé i esbrinar quins factors diferencials fan que un apartament tingui una ocupació major que un altre o tingui un preu determinat. En ser un model base, el mètode més àgil i relativament eficaç és plantejar una primera hipòtesi, que es basi en el sentit comú, de quines variables independents poden tenir una influència en la predicció de la variable dependent.

En el cas del preu de pernoctació d'un apartament, es pot pensar d'entrada que un apartament sencer amb 5 habitacions serà més car que un que en tingui únicament 1, però aquesta afirmació no és del tot certa, ja que, altres factors com el districte on se situa l'apartament o els altres equipaments que ofereix l'habitatge poden influir en el valor d'aquesta variable. De l'ocupació es pot dir el mateix, tenint en compte també aspectes propis de l'afluència turística com la localització preferida dels turistes, el preu, etc.

3.6.3 Feature engineering i selecció de característiques

Els models *baseline* donen una primera versió de com enfocar la selecció de característiques i el model. Això no obstant, cal treballar d'una manera més profunda les dades que es disposen i esbrinar quines característiques i conjunts de característiques tenen un major efecte sobre l'eficàcia del model, a més de la creació de noves variables.

Tal com s'ha esmentat en la secció de l'anàlisi de les dades, rarament es dona el cas que un conjunt de dades no tingui dades nul·les. En el conjunt de listings s'hi troben múltiples característiques amb valor absent o NaN (*Not a number*). La qüestió rau en com gestionar els valors absents, ja que, no es poden emprar en els models. Generalment, hi ha diverses opcions: imputar els valors mitjançant un valor arbitrari (com podria ser 0) o fer servir el valor mitjà o la mediana dels valors no absents per assignar-lo als registres nuls. L'opció de descartar-los és igual

de vàlida si no representen un percentatge elevat de les dades, perquè si es vol fer ús en el model una característica amb una gran quantitat de dades absents, es descartarà un gruix molt gran de registres, cosa que pot provocar que hi hagi una quantitat insuficient de dades que pot afectar negativament l'entrenament del model. El que s'ha optat per fer en aquest projecte és eliminar del conjunt de dades les característiques que superen els 1000 valors absents.

Un altre aspecte a considerar són les mostres de dades que són numèricament distants a la resta o també el que s'anomenen *outliers*. Els *outliers* es detecten a l'hora de predir el preu i no tant en l'ocupació, bàsicament perquè l'ocupació és un valor acotat entre el nombre de dies que té l'agost. En canvi, el preu és un valor que té un rang més ampli que va de 0 a 9999 euros. Com s'ha esmentat a la secció d'anàlisi de dades, és poc comú el fet de llogar de franc un apartament, també ho és el fet de llogar un apartament a un preu tan elevat com 9000 euros. Aquestes mostres són propies d'errors en el procés de recol·lecció de les dades i, per tant, cal que siguin descartades.

Per detectar els *outliers* de manera senzilla, es pot emprar la visualització de les dades i interpretar-les. Un boxplot és un mètode per il·lustrar gràficament els grups de dades (numèriques) a través dels seus quartils. Com es pot apreciar a la figura 35, la distribució del preu es concentra a prop de 10 i aproximadament fins als 300 euros. Fora d'aquest rang, les mostres restants es poden considerar *outliers*.

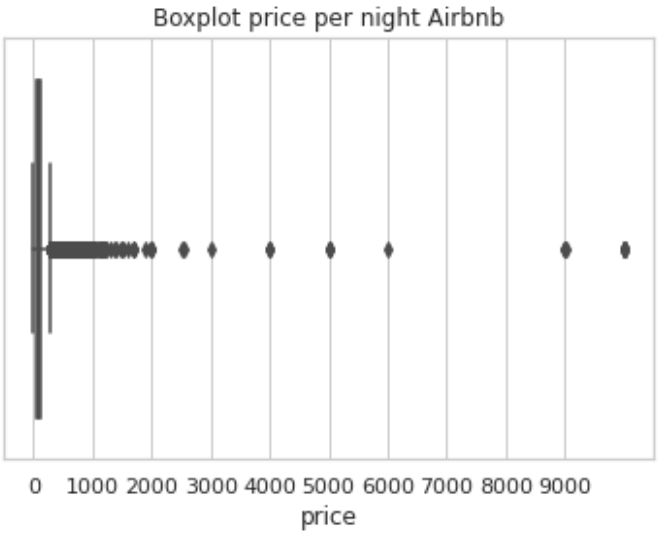

Figura 35: Diagrama de caixa del preu per nit als apartaments d'Airbnb (Barcelona).

A banda de gestionar les dades existents, en l'àmbit de la *feature engineering* també es contempla la creació de noves característiques amb l'objectiu de donar un millor rendiment al model.

En el conjunt de dades de *listings*, es disposa de les variables latitud i longitud, que són les coordenades geogràfiques dels apartaments. A partir d'aquestes dues característiques, es genera una nova dada anomenada distància o distància al centre.

La distància al centre es basa a determinar una localització geogràfica com el centre de la ciutat i calcular la distància que hi ha des de cada apartament fins al centre utilitzant les coordenades geogràfiques. Es determina com el centre de la ciutat la plaça de Catalunya perquè té una comunicació molt amplia i actua com a node de diverses línies de metro, autobusos i línies directes a l'aeroport entre d'altres. A més del seu atractiu turístic i la proximitat a llocs d'interès.

La distància al centre es calcula mitjançant la fórmula de Haversine. La fórmula del Haversine és una important equació per a la navegació astronòmica, pel que fa al càlcul de la distància de cercle màxim entre dos punts d'un globus sabent la seva longitud i la seva latitud (veure Figura 36) . És un cas especial d'una fórmula més general de trigonometria esfèrica, la llei del haversine, sobre els costats i angles d'un "triangle esfèric"[21].

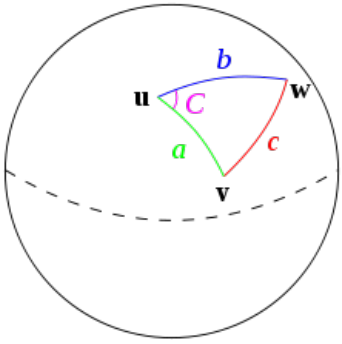

Figura 36: Il·lustració de la distància Haversine

Al conjunt de dades de *listings* també s'hi poden trobar quines comoditats té cada apartament. Les comoditats acostumen a ser elements complementaris de l'apartament, com un assecador de cabell, una piscina, un detector de monòxid de carboni, etc. Aquesta dada es proporciona en format cru, és a dir, és una cadena de caràcters que conté una llista amb totes les comoditats de l'apartament. Veure figura 37.

	id	amenities
0	18674	["Washer", "Host greets you", "Iron", "Hair dr...
1	23197	["Washer", "Iron", "Hair dryer", "Security cam...
2	32711	["Washer", "Iron", "Hair dryer", "Security cam...
3	34241	["Oven", "Microwave", "Kitchen", "Elevator", "...
4	34981	["Washer", "Extra pillows and blankets", "Iron...

Figura 37: Taula on es mostra l'identificador de l'apartament i les seves comoditats

Per a veure si el nombre de comoditats té una influència en la predicció de l'ocupació i en la del preu, bàsicament s'agafa la longitud de la llista en qüestió i se'n fa una nova variable per cadascun dels apartaments. També es fan proves creant variables*dummy* de cadascuna de les comoditats mitjançant One Hot Encoding.

Per resoldre l'absència de dades del nombre de banys, s'empra una característica ja existent anomenada *bathroom text*. Com bé el seu nom indica, descriu com són els lavabos de l'apartament, ja que, existeixen dos tipus de banys: privat i compartit. D'aquesta manera, es creen dues noves característiques que fan referència a cada tipus de bany i s'especifiquen el nombre de banys de cada tipus que disposa l'apartament. El procediment que se segueix es basa en l'extracció dels dígits de la cadena de caràcters per tal de determinar el tipus.

	id	bathrooms_text
0	18674	2 baths
1	23197	2 baths
2	32711	1.5 baths
3	34241	1 private bath
4	34981	3 baths
5	35379	2 shared baths
6	35388	2 shared baths
7	35390	2 shared baths
8	35392	2 shared baths
9	36763	1 shared bath

→

	id	private_bath	shared_bath
0	18674	2.0	0.0
1	23197	2.0	0.0
2	32711	1.5	0.0
3	34241	1.0	0.0
4	34981	3.0	0.0
5	35379	0.0	2.0
6	35388	0.0	2.0
7	35390	0.0	2.0
8	35392	0.0	2.0
9	36763	0.0	1.0

Figura 38: Taula on es mostra el processament de la dada *bathroom text* per tal de ser utilitzada per la creació de noves característiques.

A banda de les dades que es proporcionen en el conjunt dels listings, s'opta per emprar conjunts de dades externes, ja vistos a la secció d'anàlisi de les dades, com són les dades de l'hostaleria i el transport públic a la ciutat de Barcelona. L'ús d'aquestes dades rau esbrinar si utilitzades d'una manera determinada, el nombre d'equipaments públics com estacions de metro o bus i el nombre de locals de restauració propers tenen una influència en el preu de l'apartament i/o en la seva ocupació.

Tant en el cas dels transports com el de la restauració, es generen dues noves variables on se suma el nombre d'estacions de transport públic i locals de restauració que té cada districte. També es creen dues noves característiques considerant la localització de cada apartament. Fent ús la distància de Haversine, es calcula la distància de cada apartament a cadascun dels equipaments de transport públic i locals de restauració. S'estableix un llindar de distància per considerar si un servei és proper i si és menor o igual s'incrementa el comptador.

En aquesta secció, la selecció de característiques es gestiona utilitzant mètodes que impliquen avaluar la relació entre cada variable d'entrada i la variable objectiu mitjançant estadístiques i seleccionar aquelles variables d'entrada que tenen més relació amb la variable objectiu. El problema que es tracta és una regressió amb variables numèriques d'entrada, per tant, s'usa el coeficient de correlació de Pearson. El coeficient de correlació de Pearson és la relació entre la covariància de dues variables i el producte de les seves desviacions estàndard; per tant, és essencialment una mesura normalitzada de la covariància, de manera que el resultat sempre té un valor entre -1 i 1 [19]. Si dues variables tenen una correlació de per sota de 0,

significa que estan associades en el sentit invers i si tenen una correlació per sobre de 0, estan associades en el mateix sentit. A l'hora de cercar les característiques que poden tenir un millor funcionament pel model, les variables independents amb una major correlació amb la variable dependent (sigui negativa o positiva) són les que teòricament poden millorar el rendiment del model.

La llibreria sklearn, proporciona un mètode de selecció de característiques anomenat SelectKBest que proporciona les k millors variables d'acord amb el coeficient de correlació obtingut [12]. Per a contrastar els resultats que dona el SelectKBest, s'opta per utilitzar un altre mètode per a la selecció de variables, el DecisionTreeRegressor [3]. Aquest mètode també està disponible a la llibreria sklearn i després de ser entrenat amb les dades, es poden obtenir el grau d'importància de les característiques.

3.6.4 Mineria de text: TF-IDF

En el conjunt de dades de *listings*, a banda de les variables numèriques i categòriques, s'hi troben característiques que contenen text que és informació referent a l'apartament, com pot ser el nom, la descripció que l'amfitrió dona, entre d'altres. L'objectiu de l'ús d'aquest mètode en el projecte és descobrir si hi ha alguna relació entre l'ús de certes paraules en les característiques que són essencialment text (nom de l'apartament, descripció, comoditats de l'apartament, etc.) amb el preu i l'ocupació.

El TF-IDF (Term Frequency - Inverse document frequency) és una mètrica estadística per quantificar el pes que tenen les paraules en un document. *Term frequency*, com bé el seu nom indica, quantifica la freqüència que apareix una paraula en un document. Suposant que en el conjunt de descripcions dels apartaments hi ha 1000 paraules i es troba 5 cops la paraula "Güell", el tf de Güell serà: tf = 5/1000 = 0.005.

Només utilitzant el *term frequency*, el pes associat a cada paraula ve donat per la freqüència que apareix en el conjunt de documents. Per aquest motiu les paraules buides, que s'empren de forma recurrent i no tenen cap significat, tindran un pes major. Aquí és on juga un paper important l'altra part d'aquest mètode. El *inverse document frequency* (2) assigna un pes més baix a les paraules freqüents i assigna més pes a les paraules que són poc freqüents.

$$idf(t, D) = log\frac{|D|}{|d \in D : t \in d|} \tag{2}$$

On $|D|$ és el nombre de documents, $|d \in D : t \in d|$ és el nombre de documents on apareix el terme t. S'acostuma a ajustar aquesta fòrmula a $1 + |d \in D : t \in d|$ en els casos que el terme no es troba a la colecció de documents. Per tant, el TF-IDF és el producte del TF i el IDF.

En el projecte s'ha emprat aquest mètode a diverses característiques, algunes ja esmentades com bé són el nom, la descripció i el conjunt de comoditats que ofereix l'apartament. Per tal d'un correcte processament de les dades, es neteja el text dels caràcters que no aporten informació, per només quedar-se amb les paraules. No obstant això, hi han paraules que tampoc aporten informació rellevant, com bé s'han mencionat anteriorment, les paraules buides. Aquests mots són preposicions, articles, pronoms, articles, etc.

Per a construir la matriu TF-IDF de les característiques, s'empra una classe de la llibreria sklearn anomenada TfidfVectorizer que facilita molt la tasca. Aquesta classe permet modificar la configuració i excloure termes que tinguin una certa freqüència (màxima i mínima) en el conjunt de documents. D'aquesta manera es poden descartar els termes que apareixen a gairebé tots els documents i també obviar termes molt específics com per exemple, noms propis de persona o fins i tot algunes errades ortogràfiques.

Un cop creada la matriu de característiques tf-idf, s'empren les paraules que tenen una major importància (valor tf-idf) dins del conjunt per a realitzar la predicció

del preu o l'ocupació. S'utilitzen juntament amb les característiques emprades a la secció de feature engineering per tal de seguir millorant els models.

3.6.5 Models finals

Els models finals per a la predicció de l'ocupació i el preu es construeixen a partir dels resultats que han donat els models de les iteracions anteriors, utilitzant les característiques i metodologies que han funcionat positivament en la tasca proposada.

En definitiva, aquest últim model introdueix l'ús de les dades generades a partir del model d'anàlisi de sentiment (cada ressenya de l'apartament amb el sentiment predit) com una sèrie de noves característiques: nombre total de ressenyes, nombre total de ressenyes positives, nombre total de ressenyes negatives, ràtio de ressenyes positives i ràtio de ressenyes negatives.

En el cas dels mètodes, es busca millorar la configuració dels diferents mètodes mitjançant els seus paràmetres. En el cas del Random forest, s'empra la classe GridSearchCV, de la llibreria sklearn, la qual permet trobar els paràmetres òptims per al model.

Es fa ús d'un nou mètode per a la predicció del preu i l'ocupació, una xarxa neuronal. L'estructura és similar a la que s'ha vist per al model d'anàlisi de sentiment, la diferència rau en l'arquitectura de la mateixa xarxa, ja que, es treballa amb diferents tipus de dades i tracta una tasca de regressió. Es desenvolupa una xarxa neuronal per cada tasca (ocupació i preu), on després de diverses etapes de desenvolupament, les dues xarxes són idèntiques. Les capes que s'utilitzen essencialment són les Totalment connectades (Dense) i la sortida és una capa amb una única sortida amb una activació linear, que dona la predicció del preu o l'ocupació de l'apartament corresponent.

La funció de pèrdua que s'usa és el MSE (Mean squared error), ampliament feta servir en les tasques de regressió, que és la mitjana de dades supervisades de les diferències al quadrat entre els valors reals i predits. La mètrica d'avaluació que s'empra en el model durant l'entrenament és l'error mitjà absolut (MAE) dels valors reals i els predits.

Per tal que els models generalitzin correctament, en aquest cas, tal com s'ha aplicat al model d'anàlisi de sentiment, també es fa ús de l'eliminació de neurones i de tècniques de regularització de pesos, en concret l'L2. Addicionalment, es fa ús de l'aturada prematura o *early stop* de l'entrenament per tal de no consumir recursos innecessaris quan la funció de pèrdua no varia i també és útil per evitar sobreajustament del model a les dades d'entrenament.

4 Resultats

En aquesta secció es descriuen els resultats obtinguts durant l'etapa de desenvolupament. Es divideix en els resultats que dona el model d'anàlisi de sentiment i els resultats que donen els models de predicció del preu i els models de predicció de l'ocupació.

4.1 Entrenament model d'anàlisi de sentiment

El model d'anàlisi de sentiment, que com anteriorment s'ha explicat, consta d'una xarxa neuronal. L'entrenament és de 50 èpoques, una època és una passada completa del conjunt de dades d'entrenament a través de l'algorisme. En aquests resultats, s'analitza mitjançant diverses mètriques com és l'etapa d'entrenament del model i el rendiment dels resultats que obtenim.

Per tal d'avaluar l'entrenament de la xarxa neuronal, s'observa la funció de pèrdua del conjunt d'entrenament, que indica com de bé el model ajusta les dades d'entrenament, i del conjunt de validació, que indica com s'ajusta el model a dades noves. D'aquesta manera, observant les dues funcions de pèrdua, la del conjunt d'entrenament i la del conjunt de validació, es pot veure si el model generalitza correctament o només té un bon rendiment en el mateix conjunt d'entrenament.

Com es pot veure a la figura 39, a les primeres èpoques de l'entrenament (0 a 10), les dues funcions de pèrdua tenen tendència a disminuir notablement, fet que és positiu perquè vol dir que el model esta aprenent. A partir d'aquell moment, ambdues funcions de pèrdua continuen disminuint però ja no amb el mateix ritme.

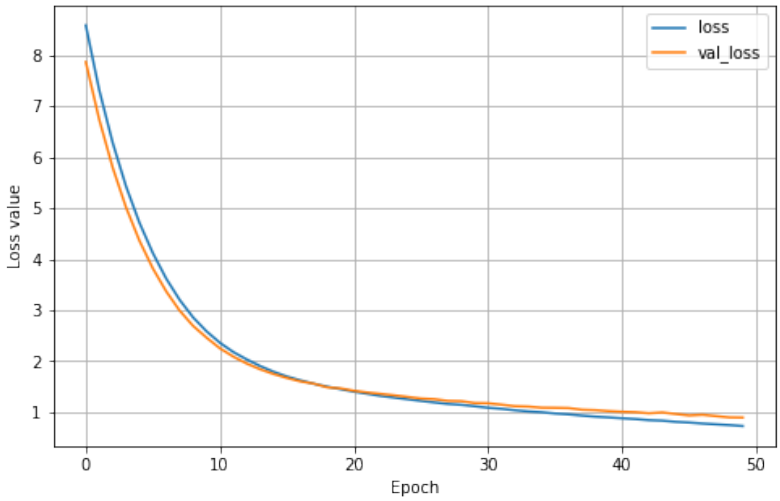

Figura 39: Funcions de pèrdua del conjunt d'entrenament i del conjunt de validació del model d'anàlisi de sentiment amb les dades etiquetades de Tripadvisor.

L'altre aspecte a observar és l'exactitud de la predicció. Cal recordar que el model fa una predicció de quina classe pertany la ressenya en qüestió; positiva o negativa. Com es pot veure a la figura 40, l'exactitud del conjunt d'entrenament i validació

augmenta ràpidament i a partir de la vintena època, ambdues s'estabilitzen. Els valors que prenen són notablement bons, ja que el conjunt de validació obté una exactitud propera a 0.91.

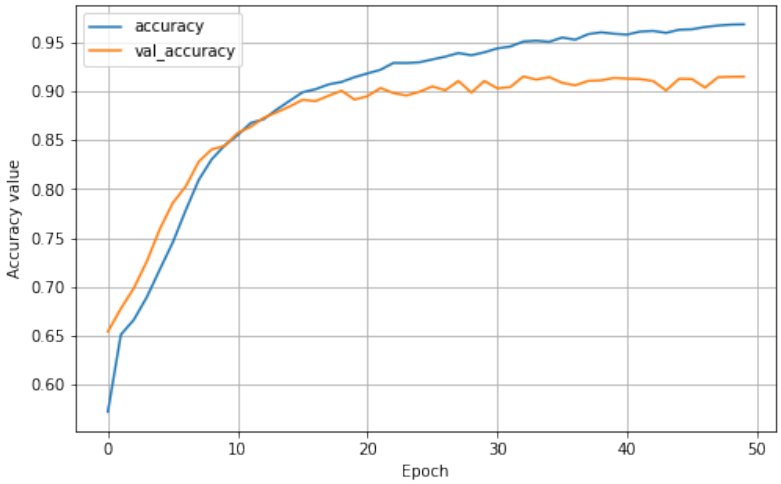

Figura 40: Corbes de l'*accuracy* del conjunt d'entrenament i del conjunt de validació del model d'anàlisi de sentiment amb les dades etiquetades de Tripadvisor.

Pel que fa al rendiment del model, es considera que generalitza correctament. Cal considerar les funcions de pèrdua dels dos conjunts; tal com s'ha vist a la figura 39, ambdues funcions disminueixen a la vegada i no hi ha una diferència important entre elles. La funció de pèrdua dona aquests resultats gràcies al fet d'afegir tècniques de regularització a les diverses capes de la xarxa neuronal.

L'exactitud també dona informació sobre un possible sobreajustament de les dades. En aquest cas, podria ser un clar indici d'overfitting que el conjunt d'entrenament tingui una accuracy propera a 0.96, no obstant això, el conjunt de validació també té una exactitud elevada, la qual no dista substancialment de la del conjunt d'entrenament.

El nombre d'èpoques s'ha definit observant les oscil·lacions de les funcions de cost i els valors de l'exactitud. Com bé ja s'ha descrit en aquesta secció, els valors arriben a un punt que són gairebé estàtics, per tant, es decideix acabar l'entrenament del model quan es detecta aquest fenomen, d'aquesta manera, en alguns casos s'evita l'aparició de sobreajustament. No obstant això, l'entrenament realitza les 50 èpoques que s'han especificat, sense aturar-se abans.

4.2 Resultats del model d'anàlisi de sentiment

Cal recordar que el model d'anàlisi de sentiment s'entrena amb les dades etiquetades de Tripadvisor i fa la predicció sobre les dades no etiquetades de les ressenyes d'Airbnb. Per a verificar que el model funciona correctament amb el conjunt de dades no etiquetades, se n'etiqueta un subconjunt, de diferents llengües, on la predicció té una exactitud al voltant del 91%. D'aquesta manera, es pot extrapolar

aquest rendiment al conjunt no etiquetat.

Com es pot veure a la la figura 41, es mostra un petit conjunt dels resultats obtinguts amb algunes de les ressenyes etiquetades on es vol mostrar quin és el rendiment que té el model i les seves particularitats. En definitiva, el model funciona notablement bé, com es pot veure, pot classificar bé el sentiment en diferents idiomes que era un dels objectius que es requeria.

Ressenya	Sentiment real	Predicció
very cozy apartment with many large rooms very close to the stunning sagrada familia and eating places we love it hope to return one day again	1	1
all perfect just like in the pictures im sure ill come back	1	1
un apartamento muy agradable y adaptado	1	1
all perfect just like in the pictures im sure ill come back	1	1
host is very helpful and easy to communicate with gave great suggestions on where to go what to see and where to eat	1	1
mala localizacion bastante descuidado	0	0
our stay was a bad experience the pictures are very different from what they sell the terrace was not at all what the pictures offered it is dirty with dead and plastic plants has construction materials scattered in different areas the living room had a broken table the central ceiling lamp was a black stick hanging without screen or bulb it didn´t work the white couch was really dirty and the carpets where disgusting the noise from the gran via is uncomfortable and the whole flat trembles when the metro passes by i definitely don't recommend it especially when travelling with kids the communication was good on the first night my phone fell and broke the apt opens with an app so we couldn´t get in and they were helpful to come and open at midnight	0	0
es war eine schlechte erfahrung wir werden nicht in diese wohnung zurückkehren	0	0
alles perfekt genau wie auf den bildern ich komme sicher wieder	1	1
je ne recommanderais certainement pas cette chambre à qui que ce soit	0	0
hautement recommandé	1	1
cétait une mauvaise expérience nous ne reviendrons pas dans cet appartement	0	0
séjour très agréable	1	1
muy desconforme con la limpieza y el mantenimiento ya que nos encontramos el apartamento muy sucio y sin productos para poderlo limpiar además demasiado ruido con respecto al resto de huéspedes del mismo bloque asimismo la puerta principal del apartamento deja mucho que desear aún así el apartamento está muy bien ubicado si mejoran estas cosas es totalmente recomendable	0	0
buen sitio para estancia cómoda	1	1
apartamento sucio no volveremos	0	1
appartement sale nous ny retournerons pas	0	1
would definitely not suggest this room for anyone	0	1

Figura 41: Mostra de resultats de la predicció d'anàlisi de sentiment del conjunt etiquetat de ressenyes.

Els errors que es poden observar a la figura 41 són falsos positius, que poden ser originats per la complexitat del llenguatge emprat o bé expressions contradictòries que el model no determina bé de quin sentiment es tracta quan no té prou informació. Aquests tipus d'errors en la predicció, les dues classes de sentiment obtenen una probabilitat bastant semblant, propera al 50% cadascuna, que en aquests casos es decanta cap al sentiment positiu. Es pot considerar que el model té un nivell d'exactitud notable i que compleix els requisits que s'han plantejat.

4.3 Resultats model de predicció preu i ocupació

En aquesta secció, es discutiran els resultats que ha donat cada model amb les seves millores corresponents en les diferents iteracions dels models fets per la predicció del preu i l'ocupació. S'avalua amb diverses mesures la qualitat de la regressió, és a dir, quin és el rendiment que dona el model fent la predicció d'aquestes variables.

4.3.1 Mesures d'avaluació

Les mesures d'error avaluen la diferència que hi ha entre dos conjunts de dades. En el projecte s'empren la RMSE i la MAE.

L'arrel de l'error quadratic mitjà, també anomenat RMSE (3) és una mesura freqüent en l'avaluació dels models d'aprenentatge automàtic.

$$rmse = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2} \tag{3}$$

El RMSE mai pot ser negatiu i assolir un valor de 0 significaria un ajustament perfecte dels dos conjunts de dades (predicció i reals), per tant, es busca que aquesta mètrica tingui valors baixos per considerar que el model té un bon rendiment. No obstant això, s'ha de tenir en compte l'escala de les dades, no seria correcte comparar la RMSE de diferents conjunts de dades amb variables dependents de diferent magnitud.

L'efecte que té el RMSE és proporcional a l'error quadrat. Per aquest motiu, un error gran penalitza de forma desproporcionada i fa que l'error sigui molt més gran. Una mostra que sigui un *outlier*, podria fer que l'error sigui més gran.

El MAE (4), o error absolut mitjà, és una mesura conceptualment senzilla que consisteix a calcular la diferència mitjana entre dos conjunts de dades.

$$mae = (\frac{1}{n})\sum_{i=1}^{n}|y_i - x_i| \tag{4}$$

Aquesta mesura té en comú amb el RMSE el fet que no es pot comparar amb diferents conjunts de dades i que un MAE = 0 significaria un ajustament perfecte de les dues observacions.

La R^2 o coeficient de determinació, és el "percentatge de variància explicat"pel model. És a dir, R^2 és la fracció per la qual la variància dels errors és menor que la variància de la variable dependent [17]. La R^2 és una mesura estadística que permet mesurar la qualitat de la predicció del model, no obstant això, no hi ha un valor de R^2 fix que determini com ha sigut de bo el rendiment del model, sinó que depèn del problema que s'està resolent.

En el cas del Random Forest s'utilitza el *out of bag score* per validar el rendiment. Aquesta mesura és l'exactitud de predicció de cada mostra d'entrenament x_i, fent servir només els arbres que no tenien x_i a la seva mostra d'entrenament.

Normalment, l'OOB score dona uns resultats "pitjors" d'exactitud en la predicció que els que dona el coeficient de determinació.

4.4 Resultats model Baseline

En els models baseline, s'empren tres dels mètodes esmentats anteriorment: k-NN, la regressió lineal i el Random Forest.

4.4.1 Base ocupació

En el model base de l'ocupació s'empren com a característiques la latitud i la longitud on es troba l'apartament, la disponibilitat del mes actual (cal recordar que el conjunt de dades utilitzat és el de juliol de 2021), una variable *dummy* que indica si la propietat es tracta d'un apartament sencer, el preu i una variable *dummy* que indica si l'amfitrió és superhost.

L'heurística que s'usa per a la selecció de les característiques, com bé s'ha esmentat a la secció de desenvolupament, és hipotetitzar quines variables poden funcionar millor i provar diversos conjunts de característiques. La latitud i la longitud són dues característiques que donen la localització de l'apartament, per tant, hi haurà punts de Barcelona on els apartaments tinguin una demanda superior. La disponibilitat del mes de juliol pot tenir una relació prou forta amb l'ocupació del mes a predir, ja que, indica la demanda que té l'apartament en un mes amb una forta afluència turística. La variable *dummy* que indica si la propietat és un apartament complet o no pot tenir una influència si algun tipus de propietat en concret és més sol·licitada que alguna altra. La característica de superhost indica que l'amfitrió dona un tracte de més qualitat als seus hostes.

Perquè el mètode k-NN pugui generalitzar correctament, cal especificar una k que defineixi una regió amb un nombre de veïns prou gran d'acord amb el nombre de mostres que hi ha. El mètode k-NN amb una $k = 10$ i amb les característiques anteriors, té un rendiment que es pot veure a la figura 42 amb el conjunt d'entrenament i de test respectivament. El RMSE del conjunt d'entrenament és 6.35, mentre que al conjunt de test és 6.96.

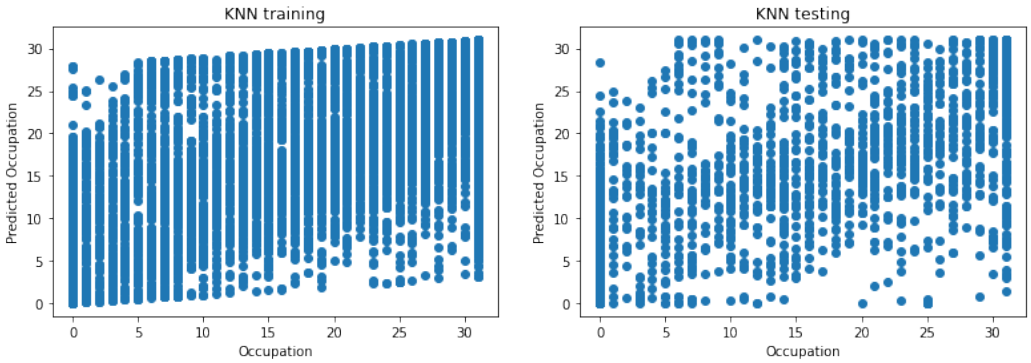

Figura 42: Figures on es mostren els valors de la predicció del model que utilitza el k-NN en comparativa amb els valors reals de l'ocupació. El coeficient de determinació en el conjunt d'entrenament és $R^2 = 0.758$ i al conjunt de test és $R^2 = 0.711$. (Etapa Baseline)

La regressió lineal té un RMSE molt similar tant a l'entrenament (7.45), com al conjunt de test (7.36). L'error és lleugerament major que al k-NN. Veure figura 43.

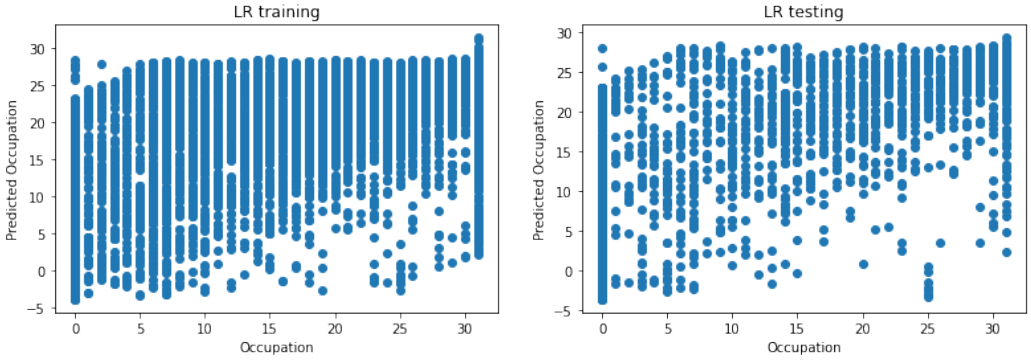

Figura 43: Figures on es mostren els valors de la predicció del model que utilitza la regressió lineal en comparativa amb els valors reals de l'ocupació. El coeficient de determinació del model és al conjunt d'entrenament $R^2 = 0.667$ i al conjunt de test és $R^2 = 0.677$. (Etapa Baseline)

El Random Forest amb una configuració de 100 predictors (arbres de decisió) té un OOB score de 0.723 i $R^2 = 0.733$ al conjunt de test. El RMSE que té aquest model al conjunt de test és de 6.7.

Els tres models base de l'ocupació del mes d'agost, donen uns resultats notablement bons sense fer una anàlisi exhaustiva de les característiques que poden donar un millor rendiment ni fent una modificació important de la configuració dels paràmetres dels diferents mètodes. Cal considerar que l'escala de la variable a predir pot prendre un rang de 32 valors diferents (0 a 31), per tant, cal ser conscient que la magnitud de l'error que donen els tres models és bastant raonable i també dona marge de millora. Pel que fa a *overfitting*, els tres models generalitzen correctament, ja que, donant una ullada a l'error del conjunt d'entrenament i a l'error del conjunt de test, tenen una diferència poc significant. Si en lloc de l'error observem la R^2,

es pot veure que el model que utilitza el k-NN és lleugerament major al conjunt d'entrenament, fet que poc indicar un sobreajustament menor. Es pot considerar que el model que usa el Random Forest és el que dona el millor rendiment.

4.4.2 Base preu

En el model base per predir el preu s'utilitza la mateixa metodologia que al model base de l'ocupació per a escollir les característiques. En aquest cas també s'empra la latitud i la longitud de l'apartament, la variable *dummy* que indica si la propietat és un apartament sencer, el nombre de persones que poden allotjar-s'hi i la variable que indica si l'amfitrió és un superhost. El conjunt de característiques és molt semblant a la del model d'ocupació, la diferència rau a afegir el nombre de persones que s'hi poden allotjar, ja que, es creu que té una relació directament proporcional amb el preu.

Com ja s'ha vist a la secció d'anàlisi de dades, el preu té un rang de valors molt més ampli que l'ocupació, per aquest motiu s'hi troben *outliers* que dificulten el rendiment del model. En aquest model base s'exclouen els apartaments que tenen un preu major de 400 euros.

El model utilitzat amb el mètode k-NN s'empra una $k = 13$ i el RMSE d'entrenament és 49.92, mentre que al test és 52.49.

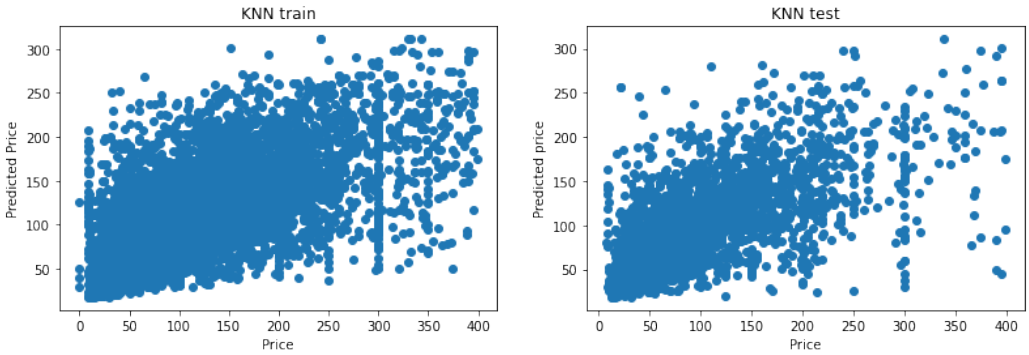

Figura 44: Figures on es mostren els valors de la predicció del model que utilitza el k-NN en comparativa amb els valors reals del preu. En el conjunt d'entrenament el coeficient de determinació és $R^2 = 0.53$ i al conjunt de test és $R^2 = 0.47$. (Etapa Baseline)

A la regressió lineal el RMSE al conjunt d'entrenament és 55.34 i al conjunt de test 56.22. Veure 45.

En el cas del Random forest, la out of bag score és 0.451 i la $R^2 = 0.48$ al conjunt de test. El RMSE d'aquest model al conjunt de prova arriba a 51.98.

Fent un balanç general dels tres models de la predicció del preu, el rendiment és millorable i hi ha marge de millora a l'hora d'escollir les característiques i la configuració dels paràmetres dels mètodes. Eliminar del model una part dels apartaments amb un preu anormal a la resta, ha fet que els models tinguessin un millor

rendiment, encara que aquesta mesura no ha sigut suficient i cal acotar millor els límits dels *outliers*. El sobreajustament es veu lleugerament en el cas del model que utilitza el mètode de k-NN on el conjunt d'entrenament té un millor rendiment que el conjunt de test. Veure figura 44.

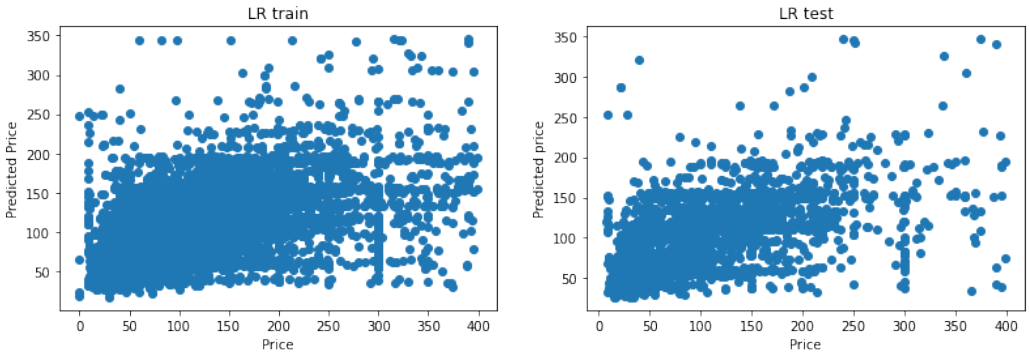

Figura 45: Figures on es mostren els valors de la predicció del model que utilitza la regressió lineal en comparativa amb els valors reals del preu. En el conjunt d'entrenament el coeficient de determinació és $R^2 = 0.40$ i al conjunt de test és $R^2 = 0.39$. (Etapa Baseline)

4.5 Resultats Feature engineering

En la nova iteració dels models, s'analitzen i es busquen característiques que tinguin un major rendiment. L'ús de la classe SelectKBest permet veure quines característiques poden ser útils a l'hora de predir l'ocupació i el preu, es segueix en certa mesura el criteri de les variables amb una major puntuació obtinguda. Addicionalment, s'utilitza un arbre de decisió per contrastar els resultats obtinguts amb el mètode SelectKBest.

4.5.1 Feature engineering ocupació

Els resultats del mètode SelectKBest (veure figura 46), amb una configuració per obtenir les 7 variables amb la puntuació més gran per a predir el preu de l'ocupació, determina que la variable més rellevant és la disponibilitat del mes de juliol. La segueix la disponibilitat dels 365 dies vinents i el nombre mínim de nits que cal arrendar l'apartament.

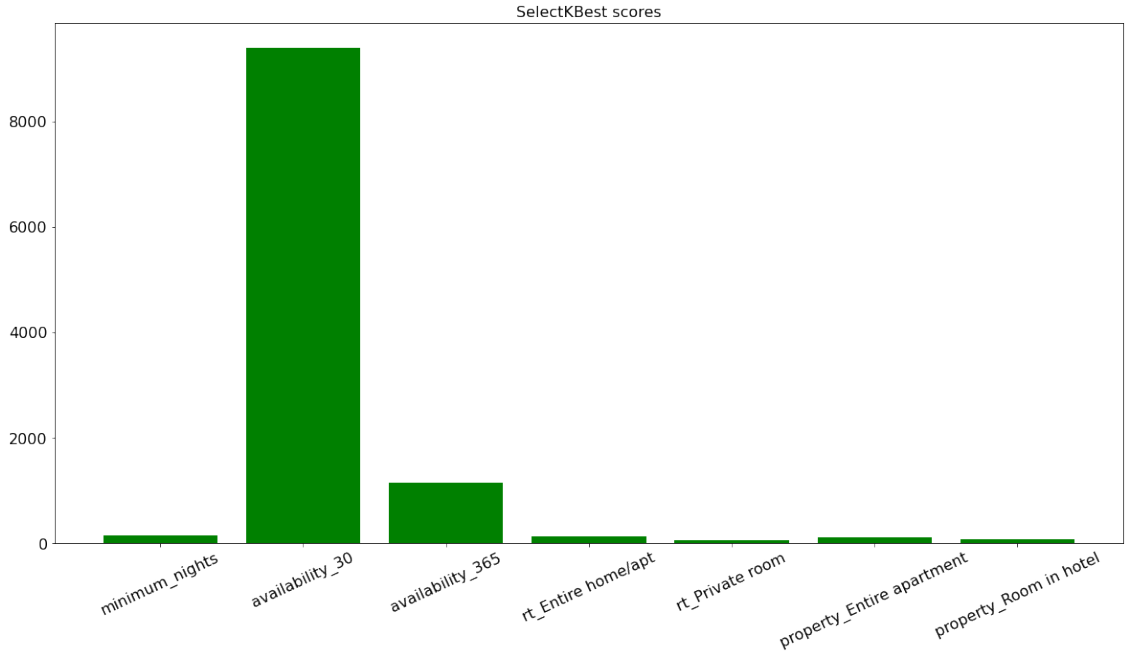

Figura 46: Gràfica on s'indiquen les k característiques amb una influència considerable per predir la variable dependent (ocupació).

En aquest model s'empra una característica que indica el nombre mínim de nits que s'ha de llogar l'habitatge, la disponibilitat (nombre de dies) del mes de juliol, la variable que indica la disponibilitat a un any vista, una variable *dummy* que indica si l'apartament és complet, la distància de l'apartament al centre de la ciutat (Plaça Catalunya), el nombre de persones que s'hi poden allotjar i el nombre de banys privats que disposa.

Utilitzant el mètode de k-NN amb $k = 10$, s'assoleix un RMSE de 4.56 al conjunt d'entrenament, mentre que al conjunt de test 4.73.

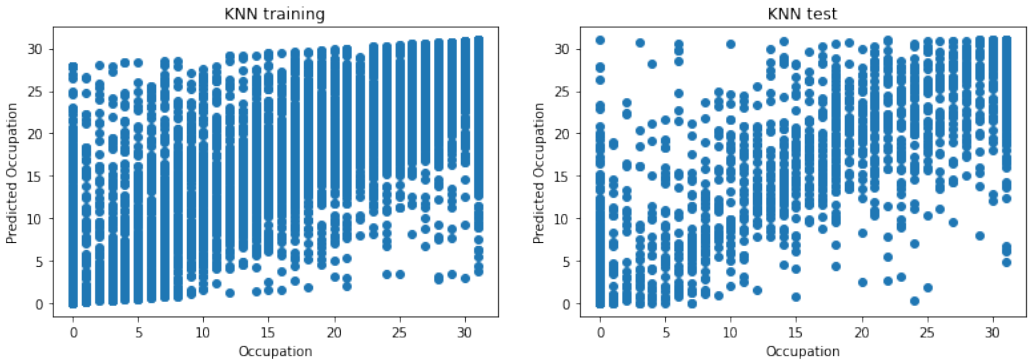

Figura 47: Figures on es mostren els valors de la predicció del model que utilitza el k-NN en comparativa amb els valors reals de l'ocupació. El coeficient de determinació al conjunt d'entrenament és $R^2 = 0.875$ i al conjunt de test és $R^2 = 0.865$. (Etapa Feature engineering)

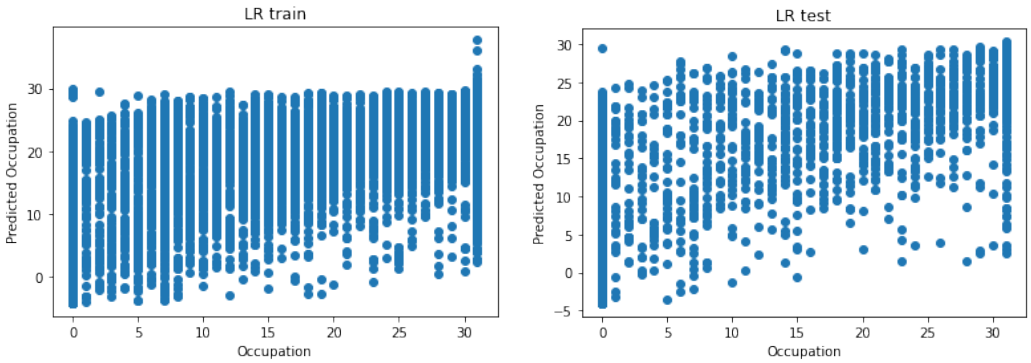

Figura 48: Figures on es mostren els valors de la predicció del model que utilitza la regressió lineal en comparativa amb els valors reals de l'ocupació. El coeficient de determinació al conjunt d'entrenament és $R^2 = 0.69$ i al conjunt de test és $R^2 = 0.70$. (Etapa Feature engineering)

El model que utilitza la regressió lineal, en el conjunt d'entrenament el RMSE és 7.17 mentre al conjunt de test és 6.97. Veure figura 48. En el cas del random forest, la out of bag score és de 0.856 i el conjunt de test dona una R^2 de 0.87. El RMSE d'aquest model és de 4.6. Veure figura 49.

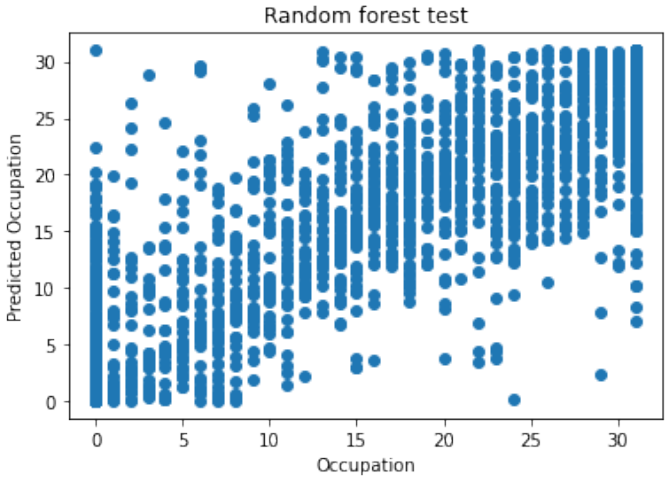

Figura 49: Comparació de la predicció de l'ocupació amb l'ocupació real dels apartaments del conjunt de prova amb el Random Forest. (Etapa Feature engineering)

En els resultats obtinguts es veu una millora en els tres mètodes utilitzats en comparació al model base. Si es revisa la taula 1, El model que utilitza el k-NN i el que utilitza el Random Forest tenen una millora significativa, disminuint l'error i amb un coeficient de determinació major. En el cas de la regressió lineal, el seu rendiment millora tímidament. La raó d'aquesta millora general rau en la selecció de les característiques, que amb l'ajuda dels mètodes de selecció ha sigut una tasca més senzilla, no obstant això, s'han fet les proves pertinents amb diferents conjunts per a verificar el que donava els millors resultats. Es pot afirmar que tots els models

generalitzen correctament, en comparar els errors del conjunt d'entrenament i els del conjunt de test, hi ha una diferència menor.

De les noves dades introduïdes, les dades referents al transport i la restauració de Barcelona, no s'han trobat útils i, per tant, no s'han utilitzat, ja que, els resultats dels tres models es veia afectat negativament.

Mètode	Etapa	RMSE	R^2
k-NN	Baseline	6.96	0.711
k-NN	FE	4.73	0.865
Linear Reg.	Baseline	7.36	0.677
Linear Reg.	FE	6.97	0.7
Random Forest	Baseline	6.7	0.733
Random Forest	FE	4.6	0.87

Taula 1: Comparativa dels resultats de la predicció de l'ocupació a les etapes base i feature engineering amb dels diferents mètodes amb el conjunt de test a l'hora de predir l'ocupació

4.5.2 Feature engineering preu

En la configuració del mètode de selecció de característiques SelectKBest, s'especifica que retorni les 10 característiques amb la puntuació més gran que atorga a les variables independents, per a predir el preu de pernoctació dels apartaments. Com es pot veure a la figura 50, la característica que obté la puntuació més gran és el nombre de persones que es poden allotjar a l'habitatge. La resta de característiques són referents a l'apartament i una de l'amfitrió. Alguns parells de característiques tenen similitud, com per exemple la variable *dummy* que indica el tipus d'apartament sencer i la que indica el tipus de propietat d'apartament sencer. En aquests casos, en ser variables que tenen una correlació considerable, se n'escull només una per tal de reduir la complexitat del model (menys característiques) i obtenir un millor rendiment.

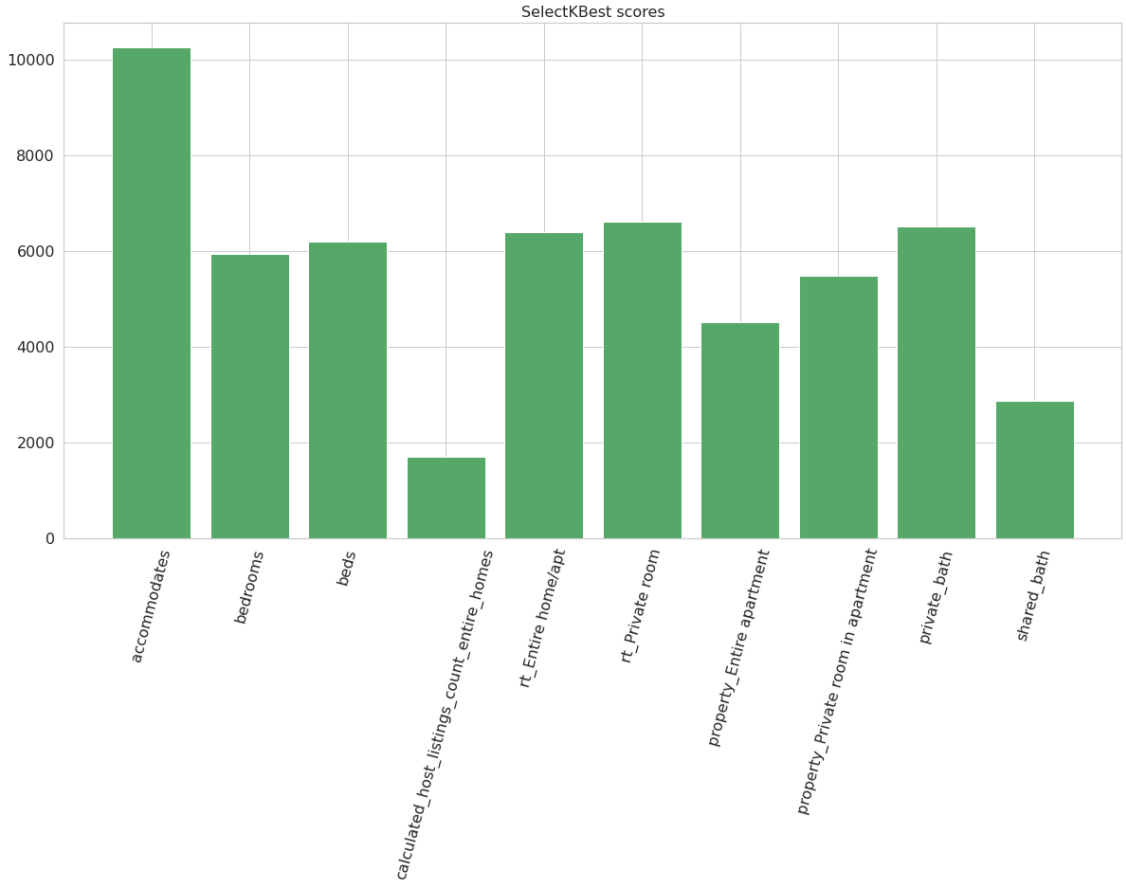

Figura 50: Gràfica on s'indiquen les k característiques amb una influència considerable per predir la variable dependent (preu).

Un cop fet l'anàlisi de les característiques, se n'agafen 9. El nombre de persones que es poden allotjar a l'apartament, el nombre de dormitoris, el nombre de lavabos privats, el nombre mínim de nits per a efectuar la reserva, la distància al centre de la ciutat, la variable *dummy* que indica si el lloguer es de l'apartament complet, el

nombre de propietats anunciades que té l'amfitrió a Airbnb, la variable *dummy* que indica si l'apartament es troba al districte de l'Eixample i una altra variable *dummy* que indica si l'apartament pot ser llogat a l'instant sense cap mena d'acceptació per part de l'amfitrió.

En el model que utilitza el mètode de k-NN, amb una $k = 13$, tal com al model base, el RMSE al conjunt d'entrenament és de 37.16, mentre que al conjunt de test és de 40.15.

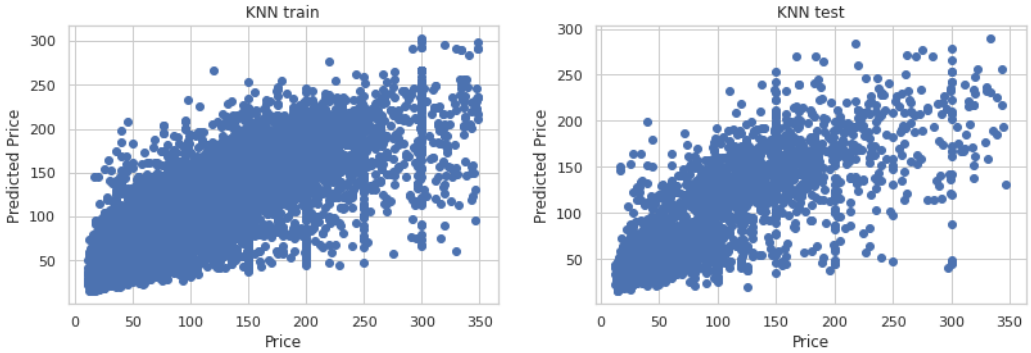

Figura 51: Coeficient de determinació al model amb el mètode de k-NN: En el conjunt d'entrenament la $R^2 = 0.69$ i al conjunt de test és $R^2 = 0.64$. (Etapa Feature engineering)

En el cas de la regressió lineal el RMSE és de 48.33 i al conjunt de test, 48.79. El model que utilitza el random forest té un OOB score de 0.658 i el conjunt de test dona un coeficient de determinació $R^2 = 0.66$. El RMSE d'aquest model al conjunt de prova és 39.09.

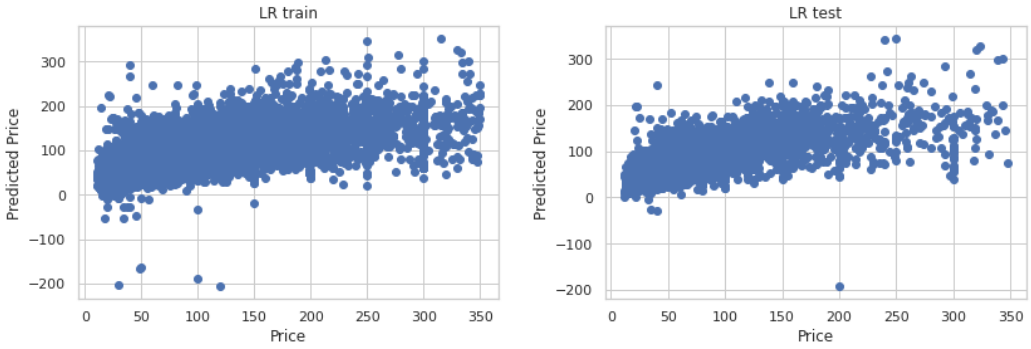

Figura 52: Coeficient de determinació al model amb el mètode de regressió lineal: En el conjunt d'entrenament la $R^2 = 0.48$ i al conjunt de test és $R^2 = 0.47$. (Etapa Feature engineering)

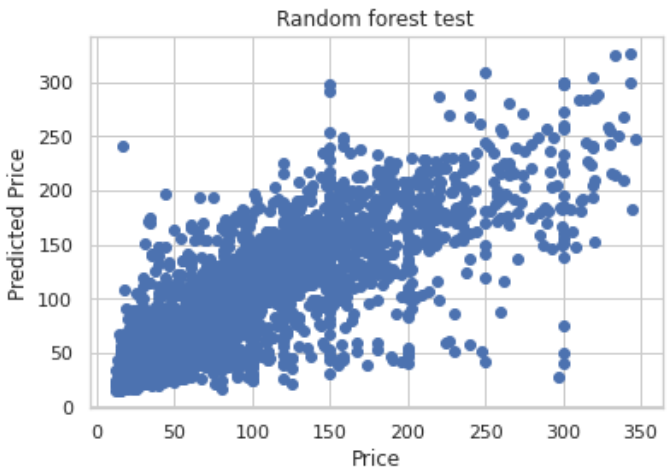

Figura 53: Comparació de la predicció del preu amb el preu real dels apartaments del conjunt de prova amb el Random Forest. (Etapa Feature engineering)

Els resultats obtinguts en els diferents models per predir el preu són, en comparació amb els models base, molt millors. A la taula 2 es poden veure els diversos resultats en el conjunt de test on els mètodes de k-NN i Random forest fan una millora significativa i la regressió lineal ho fa lleugerament. En quant a *overfitting*, el mètode k-NN és el que lleugerament té l'error amb una diferència major, tal com passava al model base. Per corregir aquest error s'han fet diverses proves amb el paràmetre k de la configuració del k-NN i amb $k = 13$ és un valor que dona un equilibri entre resultat acceptable i diferència del RMSE dels dos conjunts menor.

La selecció de les característiques ha tingut una influència en el rendiment dels diferents models amb l'ajuda del SelectKBest. La detecció més acurada els *outliers*; els apartaments amb un preu superior a 350 euros i inferiors a 10 euros es descarten del model, fet que també influeix en una millora dels resultats.

Mètode	Model	RMSE	R^2
k-NN	Baseline	52.49	0.47
k-NN	FE	40.41	0.64
Linear Reg.	Baseline	56.22	0.39
Linear Reg.	FE	48.7	0.47
Random Forest	Baseline	51.9	0.48
Random Forest	FE	39.09	0.66

Taula 2: Comparativa dels resultats de la predicció del preu a les etapes base i feature engineering al conjunt de prova dels diferents mètodes utilitzats

Un aspecte que també està present al model d'ocupació, però afecta molt més al model de la predicció del preu és l'existència d'elements duplicats als conjunts d'entrenament i test. Aquest fenomen és comú en els problemes d'aprenentatge

automàtic i és causat per la selecció de característiques del model, on es redueix la informació i es dona el cas que dos apartaments tenen les mateixes característiques, així i tot, amb un preu diferent. Per a pal·liar aquesta qüestió s'han cercat característiques que mantinguin o millorin els resultats i alhora puguin diferenciar cada apartament tant com sigui possible, un exemple és la variable *dummy* que determina si l'apartament es troba al districte de l'Eixample.

Les noves característiques creades mitjançant les dades de transport públic i restauració no donen uns resultats significativament bons en ser emprades, per tant, igual que al model d'ocupació, no se'n farà ús.

4.6 Resultats mineria de text: TF-IDF

Com bé s'ha esmentat a la secció de desenvolupament, s'empren les tres principals característiques que contenen text (equipaments de l'apartament, descripció i nom) i per cadascuna d'elles es genera una matriu tf-idf. Per escollir les característiques que es faran servir al model, s'ha buscat reduir la dimensionalitat del subconjunt, descartant les que no tenien una influència en la millora i l'introducció d'aquestes noves característiques no té efecte.

4.6.1 Resultats en el model d'ocupació

En el model d'ocupació, l'ús de les característiques obtingudes a partir de la mineria de text no té influència sobre el rendiment del model. Cal recordar que els resultats obtinguts en el model previ de feature engineering els errors de la predicció i els coeficients de determinació dels diferents mètodes donen poc marge de millora.

4.6.2 Resultats en el model de preu

En el cas del model de predicció del preu, les paraules "workspace", "tv" i "dedicated" s'afegeixen al conjunt.

Mètode	Etapa	RMSE	R^2
k-NN	TF-IDF	40.01	0.647
k-NN	FE	40.41	0.64
Linear Reg.	TF-IDF	48.29	0.48
Linear Reg.	FE	48.7	0.47
Random Forest	TF-IDF	38.33	0.676
Random Forest	FE	39.09	0.66

Taula 3: Comparativa dels resultats de la predicció del preu de les etapes de mineria de text i feature engineering amb diferents mètodes amb el conjunt de test.

Com es pot observar a la taula 3, els resultats dels tres mètodes presenten una millora, tot i això, és una millora poc significativa en comparació a la que s'ha donat anteriorment del model base al model on s'ha fet ús de feature engineering.

En definitiva, tant a la predicció de l'ocupació, com a la predicció del preu, la mineria de text no funciona tal com s'esperava. Un dels principals motius és la presència de múltiples idiomes, que incapaciten la tasca de lematitzar les paraules, a banda d'això, una paraula existeix en el corpus en diversos idiomes motiu pel qual té una dificultat afegida.

4.7 Resultats Models finals

En el model finals s'introdueixen les dades generades mitjançant l'anàlisi de sentiment. En ambdós models es parteix amb el subconjunt de característiques i configuracions dels mètodes que ha tingut bons resultats en cada cas.

4.7.1 Resultat model final preu

Després d'un anàlisi de les noves característiques creades i comparant el rendiment que ofereixen, es decideix afegir la característica que dona el ratio de ressenyes positives que té l'apartament juntament amb les característiques que s'han utilitzat a la seccio de feature engineering. Els apartaments que no tenen cap ressenya es descarten, cosa que fa reduir el nombre de dades de les que es disposa. L'opció d'imputar els valors absents no es contempla pel possible fet de tenir un augment de dades duplicades que perjudicarien el rendiment del model.

El model que utilitza el mètode k-NN, el nombre de veïns es canvia de 13 a 14 per tal d'aconseguir que el model generalitzi millor, ja que, amb el valor de $k = 13$ juntament amb el nou conjunt de característiques, els resultats de l'entrenament mostren una lleugera tendència a sobreajustar-se. En aquest cas, el model amb el mètode de k-NN s'acosegueix un RMSE = 37.22 al conjunt d'entrenament i un RMSE = 39.71 al conjunt de prova.

El model que utilitza la regressió lineal, obté un RMSE = 45.96 al conjunt d'entrenament i un RMSE = 46.40 al conjunt de prova, millorant els resultats dels models que utilitzen aquest mètode en els desenvolupaments anteriors.

El random forest es configura amb el conjunt de paràmetres que la classe GridSearchCV determina que són els millors per aquest problema. En aquest cas, el nombre de predictors que s'utilitzen són 300, el nombre màxim de característiques en un node és 3 i la màxima profunditat de cadascun dels predictors passa a ser 40. Obté un OOB Score = 0.71 i un coeficient de determinació = 0.72 al conjunt de prova. Veure figura 54.

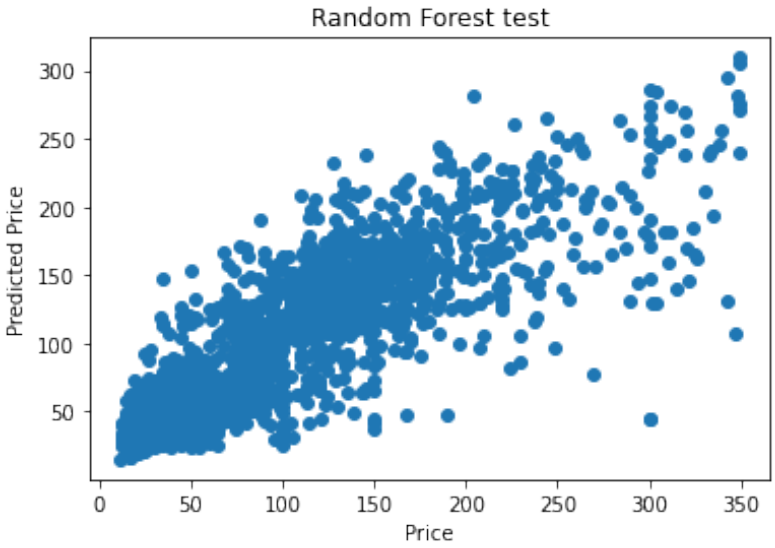

Figura 54: Model final de la predicció del preu amb el preu real dels apartaments del conjunt de prova amb el Random Forest. Es pot apreciar una distribució més semblant a una línia recta, cosa que significa un millor rendiment del model.

Pel que fa a la xarxa neuronal, es pot observar a la figura 55 que la funció de pèrdua en el conjunt d'entrenament (loss) i al conjunt de validació (val loss) a les primeres èpoques descendeixen de manera notable i els valors de la funció és van estabilitzant al llarg de les següents èpoques.

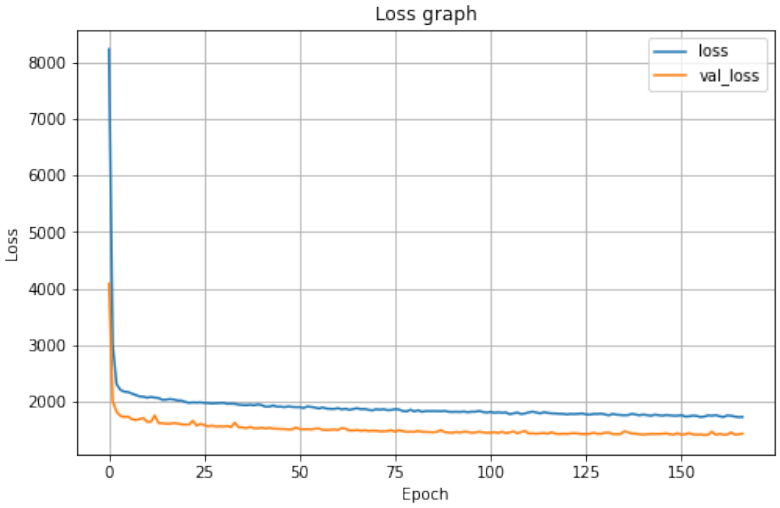

Figura 55: Funcions de pèrdua del conjunt d'entrenament i del conjunt de validació del model que empra la xarxa neuronal per a la predicció del preu.

L'error mitjà absolut es comporta de manera similar a la funció de pèrdua tal com es pot veure a la figura 56. Analitzant el comportament de la funció de pèrdua i del MAE, es pot confirmar que la xarxa neuronal és capaç de generalitzar correctament,

ja que, el conjunt de validació i el d'entrenament tenen una similitud força notable a les dues gràfiques esmentades.

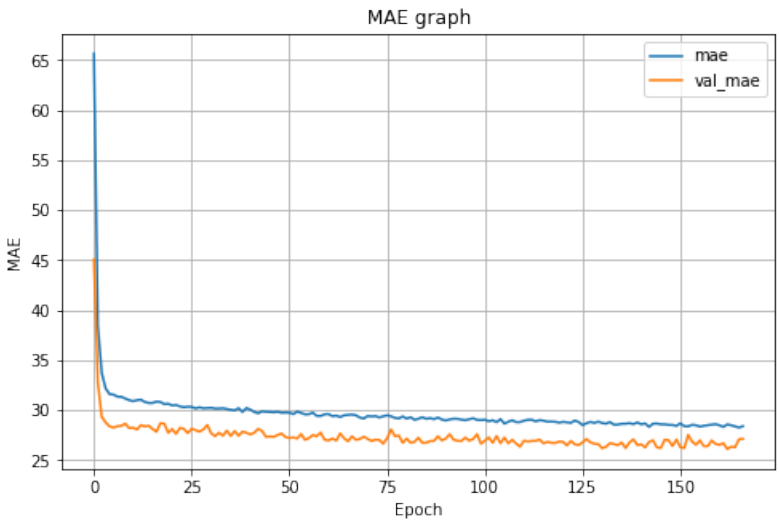

Figura 56: Corbes de l'error absolut mitjà dels conjunts d'entrenament i validació del model que empra la xarxa neuronal per a la predicció del preu.

El resultat de la predicció que s'obté és un RMSE = 40.4 en el conjunt de test. També es disposa de l'error absolut mitjà on al conjunt d'entrenament és 28.4 i al conjunt de test és 27.9. Veure figura 57.

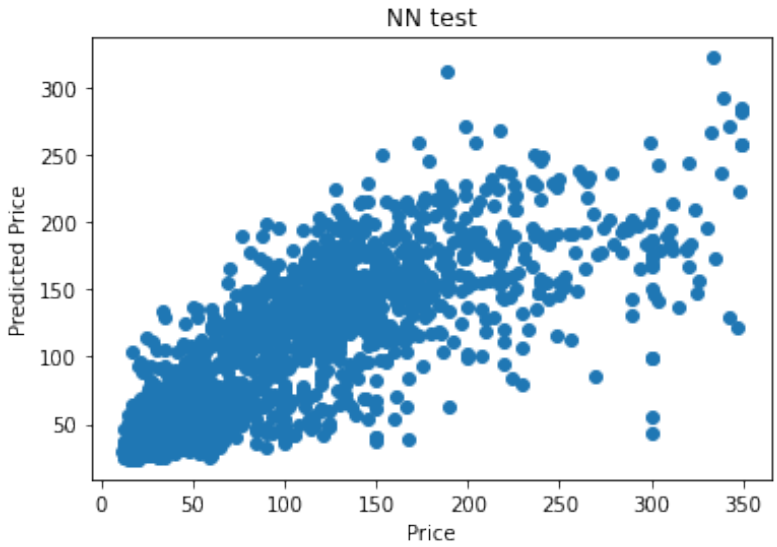

Figura 57: Model final de la predicció del preu amb el preu real dels apartaments del conjunt de prova amb una xarxa neuronal. El coeficient de determinació és 0.64, cosa que indica que s'ajusta notablement a les dades, no obstant això, el model que utilitza el random forest té un millor rendiment.

Tal com es pot veure a la taula 4, els mètodes emprats en l'etapa de feature

engineering (k-NN, regressió lineal i Random Forest) milloren en aquest model final, les millores més significatives les tenen el model que utilitza la regressió lineal i el model que utilitza el random forest. El model que fa servir la xarxa neuronal dona uns resultats en la línia dels altres mètodes. Per tant, observant els resultats, es pot considerar el random forest com el mètode que millor prediu el preu de pernoctació d'un apartament, ja que, té el RMSE més baix i el coeficient de determinació més alt.

Mètode	Etapa	RMSE	R^2
k-NN	Final	39.71	0.67
k-NN	FE	40.41	0.64
Linear Reg.	Final	46.40	0.55
Linear Reg.	FE	48.7	0.47
Random Forest	Final	36.75	0.72
Random Forest	FE	39.09	0.66
Neural Network	Final	40.41	0.66

Taula 4: Comparativa dels resultats de la predicció del preu a l'etapa final i a l'etapa de feature engineering amb diferents mètodes amb el conjunt de test.

4.7.2 Resultats model final ocupació

El model final de l'ocupació ja té uns resultats notablement bons des de l'etapa de feature engineering. En aquest últim model s'afegeixen les dades d'anàlisi de sentiment i es fan diverses proves per veure si poden aportar alguna millora. Les diverses proves conclouen que afegir característiques relacionades amb aquest conjunt de dades no millora la predicció dels models, sinó que l'empitjora considerablement. Essencialment, la millora del rendiment dels diversos models en aquesta etapa és causada per ajustar la millor configuració dels paràmetres dels diferents mètodes.

En el cas del k-NN, amb un nombre de veïns $k = 15$, té un RMSE = 4.71 al conjunt d'entrenament i un RMSE = 4.75 al conjunt de prova. Per tant, es pot afirmar que el model generalitza notablement bé a l'haver-hi molt poca diferència entre l'error d'entrenament i l'error de prova. No obstant això, no s'aconsegueix una millora siginificativa.

El model que fa servir el mètode de la regresió lineal no pateix canvis en relació al model que utilitza el mateix mètode en l'etapa de feature engineering.

El random forest canvia lleugerament la configuració mitjançant GridSearchCV on només com a canvi el nombre de predictors, els altres paràmetres queden per defecte. La millora del rendiment que pateix és poc rellevant.

La xarxa neuronal té un procés d'entrenament molt similar al que s'ha vist amb la tasca de la predicció del preu. Com es pot apreciar a la figura 58, la funció de pèrdua dels cadascun dels conjunts (entrenament i validació) descendeix ràpidament i s'estabilitza amb ambdues funcions amb valors molt similars. El mateix cas es dona

amb el MAE. Veure figura 59. El rendiment del model que utilitza la xarxa neuronal és acceptable, obtenint un RMSE de 5.6 i amb un coeficient de determinació de 0.81.

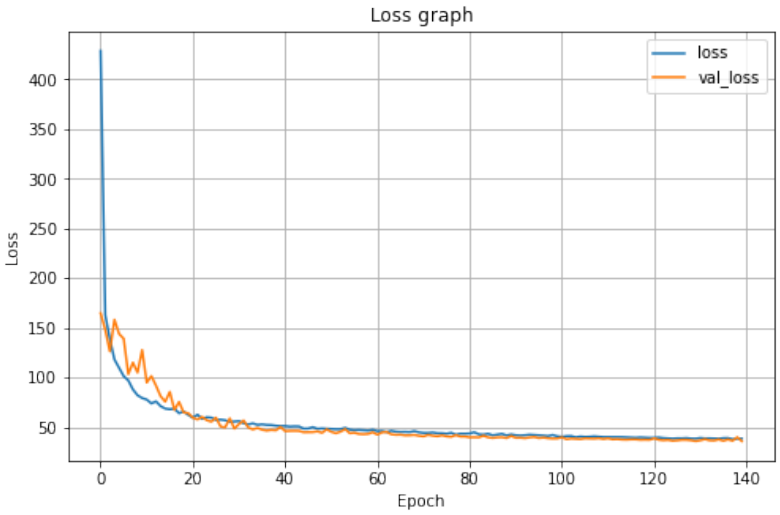

Figura 58: Funcions de pèrdua del conjunt d'entrenament i del conjunt de validació del model que empra la xarxa neuronal per a la predicció de l'ocupació.

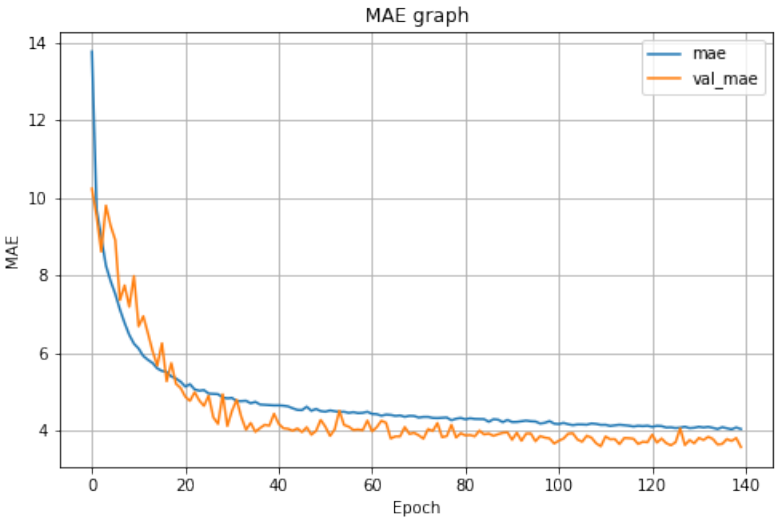

Figura 59: Corbes de l'error absolut mitjà dels conjunts d'entrenament i validació del model que empra la xarxa neuronal per a la predicció de l'ocupació.

Mètode	Etapa	RMSE	R^2
k-NN	Final	4.71	0.864
k-NN	FE	4.73	0.865
Linear Reg.	Final	6.97	0.7
Linear Reg.	FE	6.97	0.7
Random Forest	Final	4.57	0.874
Random Forest	FE	4.6	0.87
Neural Network	Final	5.57	0.81

Taula 5: Comparativa dels resultats de la predicció de l'ocupació a l'etapa final i a l'etapa de feature engineering amb dels diferents mètodes amb el conjunt de test a l'hora de predir l'ocupació

En definitiva, el desenvolupament aplicat en aquesta etapa no obté els resultats esperats, ja que, les millores han sigut molt discretes. Com s'ha esmentat anteriorment, els resultats del model d'ocupació en etapes prèvies han sigut molt satisfactoris (veure taula 5) i difícilment s'han pogut millorar de manera significativa. La realització d'aquest model final ha sigut útil per acabar de configurar correctament els diferents mètodes que s'utilitzen.

El model que es pot considerar que dona el millor rendiment a la tasca de predir l'ocupació és el model que fa servir el Random Forest.

5 Conclusions i treball futur

5.1 Conclusions

Una de les motivacions d'aquest projecte ha sigut poder aprendre noves tècniques i metodologies relacionades amb l'aprenentatge automàtic i que poguessin ser aplicades en un problema amb casos reals, cosa que s'ha pogut assolir satisfactòriament.

A escala general, es pot considerar que els objectius plantejats a l'inici del projecte s'han aconseguit correctament, no obstant això, algunes de les tècniques i metodologies aplicades, no han tingut la utilitat en l'objectiu principal que s'esperava.

5.1.1 Model anàlisi de sentiment

El model d'anàlisi de sentiment requeria uns resultats prou fiables sobre un conjunt etiquetat de ressenyes de múltiples idiomes perquè es pogués considerar que era una bona solució. Després d'entrenar-lo amb les dades obtingudes a través del *web scrapping* del web de Tripadvisor, aquest model té la capacitat de predir bastant acuradament el sentiment d'una ressenya, en idiomes comuns, com pot ser l'anglès, castellà, francès, etc.

5.1.2 Predicció del preu

La predicció del preu de pernoctació es considera que ha sigut exitosa, ja que, mitjançant diverses tècniques i metodologies, s'ha reduït considerablement l'error de predicció en comparació al model base, que era el model a batre. Els mètodes que han funcionat millor han sigut els no paramètrics, el k-NN i el Random forest, aquest últim ha donat els millors resultats per aquesta tasca.

L'etapa on s'han pogut millorar els resultats del model ha sigut en la de feature engineering i selecció de característiques. La detecció d'alguns errors en el preu i l'eliminació d'*outliers* també han tingut una influència positiva en els resultats. L'etapa posterior, on s'han introduït les dades obtingudes mitjançant l'anàlisi de sentiment i s'ha millorat la configuració dels diversos mètodes ha tingut una influència.

Les dades emprades en el projecte han tingut una influència clara en el rendiment del model. InsideAirbnb proporciona totes les dades dels apartaments que es mostren al web de manera correcta, no obstant això, caldrien característiques més específiques sobre l'habitatge en qüestió per poder tenir més marge de millora en la predicció del preu. Les dades dels apartaments que es proporcionen, sobretot les característiques referents a l'apartament, són superficials perquè se centren en el nombre d'equipaments que té l'habitatge (a banda d'altres aspectes) com el nombre de llits, d'habitacions, etc. Aquest fenomen provoca que apartaments que són diferents en termes de qualitat dels equipaments que ofereix, se'ls tracti com a habitatges similars.

5.1.3 Predicció de l'ocupació

La predicció de l'ocupació de l'agost ha resultat una tasca relativament senzilla, ja que, al disposar de les dades del mes anterior (juliol) el qual també pertany a la temporada de major afluència turística.

En l'etapa de feature engineering i selecció de característiques el model en qüestió ha millorat considerablement deixant un marge molt petit de millora que s'ha pogut ampliar mitjançant la millora de la configuració dels diferents mètodes a l'última etapa del desenvolupament.

5.2 Treball Futur

El projecte té la capacitat de continuar desenvolupant-se i obtenir millors resultats en les dues tasques principals: predir el preu i l'ocupació d'un apartament.

- **Anàlisi de les imatges dels apartaments per a la predicció del preu.** Les imatges dels apartaments anunciats a Airbnb aporten informació de com és l'apartament de manera més detallada, fet que pot fer que el rendiment del model per a la predicció del preu es vegi incrementat. L'objectiu seria que mitjançant tècniques d'aprenentatge automàtic es poguessin extreure les característiques de la imatge per tal de predir el preu de manera directa o bé crear noves variables perquè s'utilitzin dins del model.

- **Predir l'ocupació d'un apartament en qualsevol període de temps.** És un aspecte important dins del projecte, no obstant això, la metodologia hauria de ser diferent per tal de poder obtenir bons resultats.

- **Aprofundir en l'anàlisi de les ressenyes.** Mitjançant un model semblant al d'anàlisi de sentiment, poder determinar de què parla la ressenya de l'usuari. Per exemple detectar si posa èmfasi al tracte que ha rebut amb l'amfitrió o bé si la localització de l'apartament és molt bona. D'aquesta manera es podrien crear noves dades que es podrien aplicar a la predicció del preu o la predicció de l'ocupació o en altres àmbits.

Referències

[1] Airbnb. *What is Airbnb and how does it work*. URL: https://www.airbnb.co.uk/help/article/2503/what-is-airbnb-and-how-does-it-work?_set_bev_on_new_domain=1627040147_NmZkMzc3MzdhNDc0.

[2] Francois Chollet et al. *Keras*. 2015. URL: https://github.com/fchollet/keras.

[3] *Decision tree regressor*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html.

[4] *Implementing Deep Learning Methods and Feature Engineering for Text Data: The Continuous Bag of Words*. URL: https://www.kdnuggets.com/2018/04/implementing-deep-learning-methods-feature-engineering-text-data-cbow.html.

[5] *Linear Regression*. URL: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm.

[6] MDN Web Docs. *Introduction to Web development*. URL: https://developer.mozilla.org/en-US/docs/Web/Guide/Introduction_to_Web_development.

[7] *Neural Network concepts*. URL: https://ml-cheatsheet.readthedocs.io/en/latest/nn_concepts.html#neural-network.

[8] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". A: *Journal of Machine Learning Research* 12 (2011), pàg. 2825 - 2830.

[9] *Random Forests*. URL: http://machinelearning202.pbworks.com/w/file/fetch/60606349/breiman_randomforests.pdf.

[10] Radim Řehůřek i Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. A: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, maig de 2010, pàg. 45 - 50.

[11] Towards Data Science. *Anàlisi de sentiment*. URL: https://towardsdatascience.com/machine-learning-word-embedding-sentiment-classification-using-keras-b83c28087456.

[12] *SelectKBest*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html.

[13] *Selenium*. URL: https://selenium-python.readthedocs.io/.

[14] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". A: *Journal of Machine Learning Research* 15.56 (2014), pàg. 1929 - 1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[15] *Understanding LSTM*. URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[16] Analytics Vidhya. *Understanding Embedding Layer in Keras*. URL: https://medium.com/analytics-vidhya/understanding-embedding-layer-in-keras-bbe3ff1327ce.

[17] *What's a good value for R-squared?* URL: https://people.duke.edu/~rnau/rsquared.htm.

[18] Wikipedia contributors. *Anàlisi de sentiment*. URL: https://en.wikipedia.org/wiki/Sentiment_analysis.

[19] Wikipedia contributors. *Correlació de Pearson*. URL: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.

[20] Wikipedia contributors. *Ensemble Learning*. URL: https://en.wikipedia.org/wiki/Ensemble_learning.

[21] Wikipedia contributors. *Haversine Distance*. URL: https://ca.wikipedia.org/wiki/F%C3%B3rmula_del_Haversine.

[22] Wikipedia contributors. *K nearest neighbors algorithm*. URL: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm.

[23] Wikipedia contributors. *Machine learning — Wikipedia, The Free Encyclopedia*. URL: https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=1065233729.

Índex de figures

1 Diagrama de Gantt on s'exposa la planificació inicial del projecte . 3

2 Diagrama de Gantt on s'exposa la inversió final de temps que ha suposat cada tasca del projecte. 3

3 Es pot veure l'afluència de turisme i nombre de pernoctacions en hotels dels últims deu anys. La caiguda es deu als efectes de la pandèmia de Covid-19. 4

4 Secció d'hotels a tripadvisor 7

5 Ressenya d'un hotel a Tripadvisor 7

6 Inspecció d'elements a una pàgina web mitjançant les eines proporcionades pel navegador. A l'esquerra es troba el contingut renderitzat pel navegador que veu l'usuari, mentre a la dreta el codi HTML corresponent. 8

7 Filtres de les opinions de Tripadvisor. Es pot filtrar per puntuació, per temporada de l'any on s'ha escrit la ressenya, tipus de viatger i l'idioma de l'opinió. 9

8 Nombre d'apartaments per districte 10

9 Freqüència del tipus d'habitació que s'anuncia a Airbnb 11

10 Tipus d'habitació més freqüent per districte 11

11 Distribució del preu dels apartaments 12

12 Distribució del preu dels apartaments per sota de 1000 euros . . . 12

13 Relació localització de l'apartament i el seu preu. Només es mostren els apartaments amb un preu inferior a 500 euros per una millor visualització. 13

14 Preu de pernoctació dels apartaments de cada districte. Exceptuant els apartaments de més de 500 euros. 13

15 Els 10 amfitrions amb més apartaments anunciats a Airbnb (a la ciutat de Barcelona) 14

16 Les columnes del conjunt de dades que tenen més valors nuls o NaN. 15

17 Disponibilitat dels apartaments de Barcelona al mes d'agost. 15

18 Boxplot on es mostra la mitjana de disponibilitat dels apartaments per cadascun dels districtes de la ciutat a un any vista. 16

19 Els 10 apartaments amb més opinions d'Airbnb a la ciutat de Barcelona 17

20 Núvol de paraules. Les paraules més emprades en les ressenyes dels usuaris d'Airbnb 17

21 Longitud de les ressenyes dels apartaments d'Airbnb 18

22 Distribució de la puntuació de les ressenyes de TripAdvisor 18

23 Núvol de paraules del conjunt de dades de TripAdvisor. Les paraules més emprades en les ressenyes dels usuaris. 19
24 Distribució dels diferents tipus de transport públic a la ciutat de Barcelona 20
25 Distribució dels diferents establiments de restauració la ciutat de Barcelona 20
26 Arquitectura d'aprenentatge profund comú en els problemes d'anàlisi de sentiment. 23
27 Exemple de *word embedding*, on la paraula 'man' és propera a la paraula 'woman', ja que, aquestes dues paraules es poden trobar en contextos gairebé idèntics. El mateix passa amb els mots 'king' i 'queen'. 24
28 Arquitectura de la xarxa neuronal del Word2Vec, en aquest cas, el model CBOW. 25
29 Arquitectures dels dos models que es troben a Word2Vec. 25
30 Esquema RNN 26
31 Short term dependency 27
32 Long term dependency 27
33 Exemple d'una regressió lineal amb una variable dependent i una variable independent 30
34 Exemple on es proposa a quina classe pertany una mostra de test (punt verd). Si el valor de k = 3, el valor de la mostra serà un triangle vermell, ja que, la majoria dels seus veins més propers són dos triangles vermells. En canvi, si el valor de k = 5, se li assignarà a la mostra el quadrat blau. 31
35 Diagrama de caixa del preu per nit als apartaments d'Airbnb (Barcelona). 33
36 Il·lustració de la distància Haversine 34
37 Taula on es mostra l'identificador de l'apartament i les seves comoditats 34
38 Taula on es mostra el processament de la dada *bathroom text* per tal de ser utilitzada per la creació de noves característiques. 35
39 Funcions de pèrdua del conjunt d'entrenament i del conjunt de validació del model d'anàlisi de sentiment amb les dades etiquetades de Tripadvisor. 39
40 Corbes de l'*accuracy* del conjunt d'entrenament i del conjunt de validació del model d'anàlisi de sentiment amb les dades etiquetades de Tripadvisor. 40
41 Mostra de resultats de la predicció d'anàlisi de sentiment del conjunt etiquetat de ressenyes. 41

42 Figures on es mostren els valors de la predicció del model que utilitza el k-NN en comparativa amb els valors reals de l'ocupació. El coeficient de determinació en el conjunt d'entrenament és $R^2 = 0.758$ i al conjunt de test és $R^2 = 0.711$. (Etapa Baseline) 44

43 Figures on es mostren els valors de la predicció del model que utilitza la regressió lineal en comparativa amb els valors reals de l'ocupació. El coeficient de determinació del model és al conjunt d'entrenament $R^2 = 0.667$ i al conjunt de test és $R^2 = 0.677$. (Etapa Baseline) . . 44

44 Figures on es mostren els valors de la predicció del model que utilitza el k-NN en comparativa amb els valors reals del preu. En el conjunt d'entrenament el coeficient de determinació és $R^2 = 0.53$ i al conjunt de test és $R^2 = 0.47$. (Etapa Baseline) 45

45 Figures on es mostren els valors de la predicció del model que utilitza la regressió lineal en comparativa amb els valors reals del preu. En el conjunt d'entrenament el coeficient de determinació és $R^2 = 0.40$ i al conjunt de test és $R^2 = 0.39$. (Etapa Baseline) 46

46 Gràfica on s'indiquen les k característiques amb una influència considerable per predir la variable dependent (ocupació). 47

47 Figures on es mostren els valors de la predicció del model que utilitza el k-NN en comparativa amb els valors reals de l'ocupació. El coeficient de determinació al conjunt d'entrenament és $R^2 = 0.875$ i al conjunt de test és $R^2 = 0.865$. (Etapa Feature engineering) . . . 47

48 Figures on es mostren els valors de la predicció del model que utilitza la regressió lineal en comparativa amb els valors reals de l'ocupació. El coeficient de determinació al conjunt d'entrenament és $R^2 = 0.69$ i al conjunt de test és $R^2 = 0.70$. (Etapa Feature engineering) . . . 48

49 Comparació de la predicció de l'ocupació amb l'ocupació real dels apartaments del conjunt de prova amb el Random Forest. (Etapa Feature engineering) 48

50 Gràfica on s'indiquen les k característiques amb una influència considerable per predir la variable dependent (preu). 50

51 Coeficient de determinació al model amb el mètode de k-NN: En el conjunt d'entrenament la $R^2 = 0.69$ i al conjunt de test és $R^2 = 0.64$. (Etapa Feature engineering) 51

52 Coeficient de determinació al model amb el mètode de regressió lineal: En el conjunt d'entrenament la $R^2 = 0.48$ i al conjunt de test és $R^2 = 0.47$. (Etapa Feature engineering) 51

53 Comparació de la predicció del preu amb el preu real dels apartaments del conjunt de prova amb el Random Forest. (Etapa Feature engineering) 52

54 Model final de la predicció del preu amb el preu real dels apartaments del conjunt de prova amb el Random Forest. Es pot apreciar una distribució més semblant a una línia recta, cosa que significa un millor rendiment del model. 56

55 Funcions de pèrdua del conjunt d'entrenament i del conjunt de validació del model que empra la xarxa neuronal per a la predicció del preu. 56

56 Corbes de l'error absolut mitjà dels conjunts d'entrenament i validació del model que empra la xarxa neuronal per a la predicció del preu. 57

57 Model final de la predicció del preu amb el preu real dels apartaments del conjunt de prova amb una xarxa neuronal. El coeficient de determinació és 0.64, cosa que indica que s'ajusta notablement a les dades, no obstant això, el model que utilitza el random forest té un millor rendiment. 57

58 Funcions de pèrdua del conjunt d'entrenament i del conjunt de validació del model que empra la xarxa neuronal per a la predicció de l'ocupació. 59

59 Corbes de l'error absolut mitjà dels conjunts d'entrenament i validació del model que empra la xarxa neuronal per a la predicció de l'ocupació. 59

Índex de taules

1 Comparativa dels resultats de la predicció de l'ocupació a les etapes base i feature engineering amb dels diferents mètodes amb el conjunt de test a l'hora de predir l'ocupació 49

2 Comparativa dels resultats de la predicció del preu a les etapes base i feature engineering al conjunt de prova dels diferents mètodes utilitzats 52

3 Comparativa dels resultats de la predicció del preu de les etapes de mineria de text i feature engineering amb diferents mètodes amb el conjunt de test. 54

4 Comparativa dels resultats de la predicció del preu a l'etapa final i a l'etapa de feature engineering amb diferents mètodes amb el conjunt de test. 58

5 Comparativa dels resultats de la predicció de l'ocupació a l'etapa final i a l'etapa de feature engineering amb dels diferents mètodes amb el conjunt de test a l'hora de predir l'ocupació 60