UNIVERSITAT AUTÒNOMA DE BARCELONA

Matemàtica Computacional i Analítica de Dades

Modelització de dades temporals sobre el comportament humà: impacte de variables exògenes en la predicció d'agressions sexuals

Carlota Loreto Castro Sánchez

Tutoritzat per: Jordi González Sabaté, Marc Borras Camarasa

(Àrea de Ciències de la Computació i Intel·ligència)

7 de juny de 2023

Índex

1 Introducció **1**
1.1 La importància de les dades temporals . . . 1
1.2 La Hipòtesi de Treball . . . 1
1.3 Objectius . . . 2
1.4 Organització d'aquest treball . . . 3

2 L'estat de l'Art en la modelització de dades temporals **3**

3 Metodologia **5**
3.1 XGBoost . . . 6
3.2 Random Forest . . . 8
3.3 Prophet . . . 9
3.4 DeepAR . . . 10
3.5 Mètriques . . . 12

4 Modelitzant el comportament humà en dades de consum **14**
4.1 Detalls d'implementació . . . 15
4.2 Avaluació dels mètodes . . . 16
4.3 Anàlisi i comparativa dels resultats . . . 16

5 Modelitzant el comportament humà en dades d'agressions de gènere **19**
5.1 Detalls d'implementació . . . 20
5.2 Avaluació dels mètodes . . . 22
5.3 Anàlisi i comparativa dels resultats . . . 25

6 Discussió **29**

7 Conclusions i Treball Futur **30**

Referències **32**

A Hiperparàmetres **37**

Resum

En l'actualitat, la modelització de dades temporals s'ha convertit en una de les gran fites de l'aprenentatge computacional, tant per a la predicció dels valors del Bitcoin al llarg del temps com per a la previsió del consum d'energia elèctrica en determinades regions.

Recentment, s'han aplicat eines computacionals de modelització de dades en el context de les agressions de gènere. Aquest treball se centra a aprendre a modelar el comportament humà mitjançant la modelització de dades, utilitzant exemples de consum humà en una cadena de supermercats i agressions sexuals en una ciutat.

Històricament, els mètodes més utilitzats en aquest àmbit han estat XGBoost i Random Forest, que han obtingut resultats excel·lents. No obstant això, s'han desenvolupat recentment una sèrie d'eines computacionals molt potents, com DeepAR i Prophet, que permeten representar l'evolució temporal de dades multidimensionals i obtenir resultats competitius, facilitant així la presa de decisions.

En aquest treball, analitzem i comparem aquestes eines des de perspectives teòriques i pràctiques en la predicció de del consum de vendes en supermercats. L'objectiu principal consisteix en comprendre els avantatges i les limitacions de cada eina, així com el seu comportament en funció de les particularitats del problema en qüestió, com per exemple l'influència dels articles promocionats en el volum de vendes dels supermercats. Com a resultat d'aquesta anàlisi, hem participat en una competició de predicció de consum de vendes entesa com un problema de regressió de vendes, obtenint un resultat situat entre els millors 5

Amb el coneixement teòric i pràctica assolits en aquesta competició, s'han aplicat aquestes eines en el problema de la modelització i predicció de les denúncies de violència masclista en el temps. En concret, hem utilitzat una base de dades que ens ha permès identificar les variables més influents per fer un seguiment de l'evolució d'aquest tipus de denúncies, com ara els mesos d'estiu, caps de setmana i cap d'any. Això ens ha proporcionat un coneixement més profund sobre la importància de les mètriques utilitzades per calcular l'error de predicció.

Com a resultat, hem arribat al límit de les possibilitats dels mètodes utilitzats amb les dades disponibles. Per tant, el proper pas consistirà en abordar tres aspectes fonamentals:: (i) obtenir dades més completes i precises sobre l'evolució de la violència masclista en els darrers anys, (ii) identificar possibles variables exògenes (com ara partits de futbol) que estiguin correlacionades amb l'increment d'aquestes denúncies; i (iii) predir com aquestes variables podrien facilitar la predicció de les denúncies.

Tot el codi desenvolupat durant aquest treball es pot trobar aquí: https://github.com/carlotacastro/TFG

Keywords— Modelització de sèries temporals, Agressions sexuals, Violència de Gènere, Comportament humà, Predicció, Estadística, Machine Learning...

1 Introducció

1.1 La importància de les dades temporals

La modelització i anàlisi de dades de sèries temporals estan adquirint cada vegada més rellevància, a causa de la producció massiva d'aquest tipus de dades a través de l'Internet de les Coses, la digitalització de l'assistència sanitària i l'auge de les ciutats intel·ligents. Aquesta disciplina s'ha convertit en un dels objectius fonamentals de l'aprenentatge computacional actual, ja sigui per per predir els valors del bitcoins al llarg del temps o el consum d'energia elèctrica en una certa regió. A més, un camp on l'anàlisi de dades temporals ha demostrat la seva importància, és en l'anàlisi de dades sociològiques, per exemple en la predicció del comportament compra en supermercats.

Un altre àmbit on recentment s'han aplicat eines de computacionals de modelització de dades és en la d'agressions de gènere, que és precisament el context d'aquest treball. Cal destacar l'ús del sistema VioGén pel Ministeri de l'Interior, el qual té com a finalitat el seguiment i protecció de les dones víctimes de violència de gènere i dels seus fills i filles en qualsevol part del territori[1]. Aquest sistema ha suscitat controvèrsia per l'ús de variables exògenes que expliquen possibles relacions entre les víctimes i determinats factors socioeconòmics.

Malgrat les limitacions de les dades públiques, la modelització de dades temporals pot contribuir a la prevenció d'aquestes agressions i a l'ajuda a les possibles víctimes, mitjançant l'assignació de recursos addicionals durant períodes de pic d'agressions. L'objectiu d'aquest treball és aprendre a modelar el comportament humà mitjançant la modelització de dades, utilitzant exemples del consum en una cadena de supermercats i agressions sexuals en una ciutat.

Es preveu que en un futur proper, la importància, quantitat i qualitat de les dades de sèries temporals augmentaran considerablement. A mesura que el seguiment continu i la recopilació de dades esdevinguin més habituals, sorgeix la necessitat d'anàlisi competent de les sèries temporals amb tècniques estadístiques i d'aprenentatge automàtic. Per aquest motiu, en aquest estudi examinarem i utilitzarem diverses tècniques de sèries temporals per a l'anàlisi i predicció en diferents àrees.

1.2 La Hipòtesi de Treball

A la modelització de dades sovint es troben problemes amb les variables. La obtenció d'aquestes, neteja, modificació, anàlisi i, fins i tot, creació no és una qüestió trivial. Primer es definim els dos tipus de variables que existeixen.

En primer lloc hi ha les variables exògenes. Una variable exògena és aquella que existeix fora del model que es tracta, és a dir, són similars a les variables independents. Factors aliens al model determinen el valor de les variables exògenes, llavors el model no pot predir el seu valor. Pel contrari, una variable endògena és una variable en un model el valor de la qual està determinat pel model, és a dir, és la variable dependent. Donat que la variable endògena existeix dins del model, el model pot predir el valor de la variable endògena.

L'inici d'aquest treball es dóna arran d'una sèrie de notícies que alerten d'una possible relació entre esdeveniments esportius diferents al voltant del món i l'augment d'agressions sexuals. En concret l'estudi *Can the FIFA World Cup Football (Soccer) Tournamnet Be Associated with an Increase in Domestic Abuse*[2] on es desatacava un augment en les agressions domèstiques del 26% en els dies que Anglaterra empatava o guanyava durant el Mundial de futbol i del 38% els dies que perdia. L'estudi concloïa que hi havia una tendència on les agressions domèstiques augmentaven de manera considerable en el temps després de fer un anàlisi quantitatiu, fent servir els models de regressió de Poisson i binomial negativa mirant els incidents d'agressions domèstiques diàries i mensuals comunicats a la policia de nord-oest d'Anglaterra en tres tornejos separats: La Copa Mundial de 2002, 2006 i 2010.

Altres estudis a destacar són *Effects of the 2010 World Cup football tournament on emergency department assault attendances in England*[3] on s'analitza la Copa Mundial de Futbol de 2010 a Sud-Àfrica mirant les assistències a 15 departaments d'emergència a Anglaterra i es conclou que hi ha un augment d'agressions del 37.5% els dies que juga Anglaterra i, també, *Effects of international football matches on ambulance call profiles and volumes during the 2006 World*

Cup[4], estudi anglès on es troba un augment de les trucades d'ambulàncies relacionades amb l'assalt immediatament després d'un partit de futbol de la Copa del Món (2006) en què va jugar Anglaterra i més tard al vespre.

Aquesta tendència no només té lloc a Anglaterra, sinó que altre estudis com *College Party Culture and Sexual Assault*[5] on es parla d'un augment del 28% en les violacions a campus universitaris dels Estats Units després de partits de futbol americà de primera divisió, suggereixen que l'augment en agressions es pot donar a diferents parts del món i tenir una relació amb esdeveniments concrets.

La idea inicial es tracta de modelitzar dades temporals de les agressions que es donin a Barcelona per a fer un anàlisi i predicció i trobar una relació entre les agressions sexuals i esdeveniments concrets, ja siguin relacionats amb el món esportiu o no.

Per al nostre problema la generació de variables és complicada. Sovint amb les variables de la base de dades no hi ha suficient informació i s'ha de recórrer a variables exògenes que ajudin al model a donar-li més graus de llibertat per a realitzar una predicció més precisa. L'anàlisi es farà a partir de diferents variables exògenes proporcionades per les bases de dades font dels problemes, conjuntament amb variables generades a partir de base de dades externes, així com la generació de noves variables endògenes per tal d'aprofundir en el model i aconseguir les prediccions més precises possibles.

La obtenció d'aquestes variables exògenes en les agressions sexuals donats els estudis que s'han esmentat sembla ser un factor clau per a poder arribar a unes bones prediccions. Gran part de la dificultat d'aquest treball ve donada per l'anàlisi i obtenció d'aquestes variables.

1.3 Objectius

L'objectiu principal d'aquest treball és la modelització del comportament humà mitjançant l'estudi exhaustiu de la modelització de dades temporals, que consisteix en una seqüència d'observacions d'una variable preses al llarg del temps. Mitjançant aquest estudi i l'ús de models predictius precisos, l'objectiu és predir diferents agressions sexuals que es produeixin al municipi de Barcelona i, si és possible, a tot el conjunt de Catalunya, utilitzant diferents conjunts de dades per tal de trobar correlacions i alertar possibles víctimes.

Entre els objectius específics es troba aconseguir un alt nivell de coneixement en la modelització de dades temporals i en l'aplicació dels mètodes d'aprenentatge computacional més complexes i avançats que es puguin utilitzar en aquestes dades, amb l'objectiu d'avaluar la seva precisió en la modelització. Per a aconseguir aquesta tasca, es participarà en la competició Kaggle *Store Sales - Time Series Forecasting*[6] on s'utilitza l'aprenentatge automàtic per predir les vendes d'una cadena de supermercats.

Un cop s'hagi realitzat aquest estudi, es pretén aplicar els coneixements adquirits per predir diferents agressions sexuals, analitzant els esdeveniments per identificar possibles patrons que expliquin quan i per què es produeixen aquests incidents, i corroborar la investigació existent fins a la data.

Per dur a terme aquesta anàlisi, es treballarà amb la base de dades *Chicago Crime*[7], centrant-se en la modelització de sèries temporals relacionades amb delictes, especialment les agressions sexuals. S'estudiaran i analitzaran els patrons identificats i es realitzarà una predicció de les dades utilitzant els mateixos models desenvolupats.

L'anàlisi de dades sobre agressions sexuals es vol realitzar utilitzant històrics de la regió de Catalunya, que es podrien obtenir a través dels Mossos d'Esquadra. Per aconseguir aquestes dades, es proposa presentar el projecte als Mossos d'Esquadra, juntament amb els dos anàlisis esmentats anteriorment, per explicar la necessitat d'obtenir dades específiques de la regió de Catalunya i, sobretot, demostrar la capacitat de proporcionar resultats fiables i beneficiosos per a totes les parts involucrades. Si no és possible obtenir aquestes dades, es considerarà l'ús de dades trimestrals públiques publicades pel Ministeri d'Igualtat [8], o bé, en cas d'estar disponibles, dades públiques de d'altres països que ofereixin informació més precisa sobre les agressions.

1.4 Organització d'aquest treball

Per a dur a terme aquesta investigació i trobar solucions adequades tant per a la predicció de vendes com per a la predicció del nombre total d'agressions, s'han seleccionat diversos models de regressió. L'anàlisi de regressió és una tècnica estadística utilitzada per a predir valors d'una variable dependent (resposta) a partir de valors d'una o més variables independents (predictors o característiques). Consisteix en trobar la millor funció que s'ajusti al conjunt de dades. La regressió es considera una forma d'aprenentatge automàtic supervisat, ja que els algoritmes construeixen un model matemàtic per a determinar la relació entre les entrades i les sortides desitjades.

Els models de regressió que es faran servir com a possibles solucions en aquest estudi seran Random Forest Regressor, XGBoost Regressor, Prophet i DeepAR. El funcionament i les característiques d'aquests models es detallaran més endavant.

La majoria de les bases de dades utilitzades en aquest anàlisi s'obtenen de la plataforma *Kaggle*. *Kaggle* és la comunitat científica de dades més gran del món, que proporciona eines i recursos per als usuaris per a trobar conjunts de dades, crear models d'intel·ligència artificial, publicar conjunts de dades, col·laborar amb altres científics de dades i enginyers d'aprenentatge automàtic, i participar en concursos per a resoldre reptes relacionats amb la ciència de dades[9].

Les bases de dades adquirides sovint contenen dades incompletes, errònies o que no s'ajusten perfectament al problema a resoldre. Per tant, és necessari aplicar modificacions per a obtenir els millors resultats amb els models utilitzats.

Aquest treball està estructurat de manera que es pugui modelar el comportament humà a través de dos exemples: el consum de vendes en un supermercat i el volum d'agressions sexuals en un lloc específic. En primer lloc, s'analitzaran els models de regressió utilitzant la base de dades *Store Sales*, i posteriorment es continuarà amb la segona temàtica del treball, que consisteix en l'anàlisi i la predicció d'agressions sexuals mitjançant la base de dades *Chicago Crimes*.

Tant per a les vendes com per a les agressions, l'objectiu és predir un valor concret, és a dir, el nombre total de vendes o d'agressions. Aquesta variable serà la variable dependent del model. Les variables independents, d'altra banda, corresponen a altres atributs que es troben en els conjunts de dades respectius, com ara els articles en promoció per a la predicció de vendes o el nivell de pobresa per a la predicció d'agressions. Donat que el conjunt de dades *Chicago Crimes* està dissenyat per a la classificació de delictes i no per a la regressió d'un tipus de crim concret, s'haurà de recórrer a la cerca, anàlisi i implementació de variables independents trobades a datasets externs així com la generació d'atributs temporals.

2 L'estat de l'Art en la modelització de dades temporals

La predicció de sèries temporals implica realitzar pronòstics científics basats en dades històriques que estan temporalment ordenades. Aquesta pràctica consisteix en construir models a partir de l'anàlisi de dades històriques i utilitzar-los per a realitzar observacions i impulsar la presa de decisions estratègiques futures. La modelització de dades de sèries temporals i la seva anàlisi són temes de gran importància en l'actualitat i, com bé ja hem esmentat amb anterioritat, es preveu que aquesta importància augmenti encara més en el futur.

No obstant això, aquesta necessitat no és una novetat. La modelització de dades temporals ha estat una necessitat al llarg de la història. L'anàlisi de sèries temporals depèn en gran mesura de la disponibilitat de dades, i no va ser fins a la dècada de 1900[10] que es van generar grans quantitats de dades coherents i de qualitat. A més, aquest camp ha estat influït pels avenços en la teoria de la probabilitat i les estadístiques. A continuació, es tractarà l'evolució dels models més rellevants en aquest àmbit.

Des de la suavització (smoothing) exponencial a la dècada de 1950 i ARIMA(X) a la dècada de 1970 fins a TBATS el 2011, els models estadístics tradicionals van dominar el camp de les sèries temporals, especialment per a la predicció univariada[11].

La suavització exponencial implica utilitzar el valor més recent observat com a predicció següent. També es pot calcular la mitjana simple de totes les dades observades per a fer la predicció[12]. La suavització exponencial lineal es troba en un punt intermig, donant més pes a les observacions

més recents[13]. Aquest concepte senzill es va ampliar posteriorment per incloure errors/nivells, tendències i components estacionaris, amb formulacions additives o multiplicatives.

El model ARIMA és una tècnica desenvolupada per George Box i Gwilym Jenkins que utilitza valors retardats de la variable objectiu com a predictors en una regressió. El model ARIMA es compon de tres components principals:

- Component autoregressiu (AR): utilitza valors passats per predir valors futurs i contribueix a explicar les tendències a llarg termini de les dades.
- Component integrat (I): substitueix els valors de les dades per les diferències entre valors successius i ajuda a eliminar les tendències i l'estacionalitat de les dades, assegurant així que la sèrie sigui estacionària, com requereix el model.
- Component de mitjana mòbil (MA): utilitza errors passats per predir valors futurs i ajuda a explicar les oscil·lacions brusques de les dades (s'utilitza per eliminar fluctuacions periòdiques de la sèrie temporal, com les degudes a l'estacionalitat). Es considera com a "soroll blanc" que s'assembla al terme de pertorbació en els models de regressió.

El model ARIMA s'utilitza principalment amb dades univariades i sense variables exògenes, tot i que hi ha variants que permeten una major flexibilitat, com l'ARIMA(X), que incorpora variables exògenes[14]. La sèrie temporal ha de ser estacionària per satisfer els supòsits subjacents del model. Normalment, prendre la primera diferència de la sèrie temporal aconsegueix convertir-la en estacionària. Aquest model sol funcionar bé amb quantitats moderades de dades i és particularment adequat per a pronòstics a curt termini, ja que s'adapta ràpidament a les tendències mitjanes de les dades.

La primera generalització dels models Box-Jenkins va ser acceptar models ARIMA multivariants[15], entre els quals es troben els models VAR (Vector AutoRegressive). No obstant això, aquestes tècniques només són aplicables a sèries temporals estacionàries.

Una altra línia de desenvolupament en l'àmbit de les sèries temporals, que es deriva dels models de Box-Jenkins, són les generalitzacions no lineals, especialment els models ARCH (Heteroscedasticitat condicional autoregressiva) i GARCH (G = Generalitzat). Aquests models permeten la parametrització i predicció de la variància no constant, i s'han mostrat especialment útils en el context de les sèries temporals financeres.

La regressió de la màquina vectorial de suport ($SVMR$), desenvolupada el 1963, és un algorisme que minimitza la funció de pèrdua per tal de maximitzar la distància entre els punts més llunyans de l'hiperplà[16]. Dins de l'àmbit de l'exploració automatitzada de les relacions entre els factors, trobem l'algorisme d'aprenentatge no supervisat de l'anàlisi de components principals (PCA). PCA és un mètode estadístic que simplifica la complexitat dels espais mostreig segons el principi de minimitzar la pèrdua d'informació de les dades, generant una relació clara dels paràmetres[17]. En un espai mostreig amb p dimensions, PCA permet trobar un nombre inferior de factors subjacents (z¡p) que expliquen aproximadament la mateixa quantitat d'informació que les variables originals. Aquesta metodologia serà utilitzada durant el treball.

Els algorismes d'augment del gradient d'aprenentatge supervisat, com XGBoost i LightGBM, van començar a ser introduïts al 2014/2016. Tot i no ser models de sèries temporals en sentit estricte, sovint es pot reduir un problema de sèries temporals a una problemàtica més general. En els enfocaments estadístics anteriors, com ARIMA, el sistema estava dissenyat específicament per a dades de sèries temporals, i només podíem alimentar les dades directament[13]. En canvi, en aquests enfocaments d'aprenentatge automàtic, hem de generar atributs que els algorismes puguin utilitzar per tenir en compte l'ordre temporal.

Prophet és un model presentat el 2017 [18] que millora les prediccions en treballar amb sèries temporals que presenten un fort component estacional amb un fort efecte de vacances, tot i no estar basat en aprenentatge profund. DeepAR, també presentat el 2017 [19], és un mètode de previsió basat en xarxes neuronals recurrents autoregressives que aprèn un model global a partir de les dades històriques de totes les sèries temporals del conjunt de dades. Aquest mètode es basa en treballs previs d'aprenentatge profund per a sèries temporals i adapta una memòria a curt termini

(LSTM), el que ajuda a mitigar el problema del gradient explosiu i desaparegut[20], aportant una arquitectura de xarxes neuronals recurrents al problema de la previsió probabilística[21].

Els models estadistics van deixar de dominar el camp en el moment que ESRNN va guanyar M4[1] el 2018. ESRNN va introduir una forma d'aprenentatge multitasca que supera la limitació de les quantitats massives de dades requerides per a l'aprenentatge profund. En lloc de construir un model independent per a cada sèrie temporal com ho faria un model estadístic tradicional, ESRNN utilitza un model complex per predir múltiples sèries temporals (multitasques). Aquest enfocament híbrid combina la suavització exponencial amb les xarxes neuronals recurrents (RNN). En primer lloc, cada sèrie temporal es descompon en components de nivell, tendència i estacionalitat utilitzant el mètode de suavització exponencial multiplicativa. Després, l'RNN s'encarrega d'aprendre les tendències no lineals sobre els valors desestacionalitzats i normalitzats. El model utilitza la funció de pèrdua quantil/pinball, que minimitza el quantil de la variable objectiu, i afegeix una penalització per a la variació o fluctuació de les prediccions.

En el món de les xarxes neuronals, el *Multilayer Perceptron (MLP)* és un model rellevant que cerca automàticament les relacions entre paràmetres des de la perspectiva del processament de la informació. Es tracta de capes denses totalment connectades, que transformen les dades d'entrada a la dimensió desitjada[22]. Gràcies a la seva capacitat d'autodetecció, els MLP tenen una precisió superior als models tradicionals. A partir del MLP, es pot construir una Xarxa Neuronal Bayesiana (BNN), que és una extensió que incorpora inferència posterior per controlar el sobreajustament[23]. Aquesta arquitectura combina la idea de probabilitat amb el MLP, on les variables aleatòries segueixen una distribució gaussiana.

Malgrat que el processament natural del llenguatge (PNL) treballa principalment amb seqüències, molts dels avenços en PNL no es poden aplicar eficaçment a les sèries temporals, ja que aquestes no tenen l'estructura profunda que es troba en les dades de text. Per tant, perquè l'aprenentatge profund funcioni bé amb les sèries temporals, es requereix una arquitectura de xarxa neuronal específica. Un exemple d'aquesta arquitectura és N-BEATS.

N-BEATS, presentat el 2019, introdueix un enfocament important d'empaquetar models amb diferents horitzons d'entrada, mètriques i inicialitzacions aleatòries. Es centra en "resoldre el problema de previsió de punts en sèries temporals univariades utilitzant l'aprenentatge profund". Posteriorment, el paquet DARTS adapta l'arquitectura original de N-BEATS a sèries temporals multivariades, convertint les dades d'origen en una sèrie unidimensional. D'aquesta manera, és possible incloure regressors addicionals com a característiques. Utilitza una arquitectura de xarxes de feed-forward agrupades amb una nova topologia jeràrquica doblement residual de previsions i "backcasts" [24]. A més, si la interpretabilitat és un factor important per a l'aplicació, aquest model ofereix una arquitectura ïnterpretable"que consta de dues piles: una pila per a les tendències i una pila per a l'estacionalitat.

Per concloure, al 2019, Facebook va adaptar AR-Net a Prophet i va crear NeuralProphet[25] al 2020, que és essencialment una extensió d'aprenentatge profund d'ARIMA de la mateixa manera que ESRNN és una extensió d'aprenentatge profund d'Exponential Smoothing.

3 Metodologia

Abans de centrar-nos específicament en els models, repassem els conceptes bàsics de l'aprenentatge supervisat. En l'aprenentatge supervisat, un model fa prediccions y_i basades en les dades d'entrada x_i. Un exemple comú és el model lineal, on la predicció es calcula com $\hat{y}i = \sum \theta j x_{ij}$, una combinació lineal de les característiques d'entrada ponderades. Els paràmetres θ són les incògnites que hem de determinar a partir de les dades.

L'entrenament del model consisteix a trobar els valors òptims dels paràmetres θ que s'ajustin millor a les dades d'entrenament. Per aconseguir això, hem de definir una funció objectiu que mesuri la qualitat de l'ajustament del model a les dades. Aquesta funció objectiu està composta per dues parts: la pèrdua (*loss*) L i el terme de regularització ω:

$$obj(\theta) = L(\theta) + \omega(\theta) \tag{1}$$

[1] M Competition és l'equivalència d'ImageNet amb la visió per ordinador per al model de sèrie temporal

La pèrdua (*loss*) mesura la qualitat predictiva del nostre model en relació a les dades d'entrenament. Representa la diferència entre els valors reals i els valors predits, és a dir, quant de lluny es troben els resultats del model dels valors reals. Una elecció comuna per a la pèrdua és l'error quadrat mitjà, que s'expressa com:

$$L(\theta) = \sum (y_i - \hat{y}_i)^2 \tag{2}$$

On y_i és el valor real i $\hat{y}_i$ és el valor predit pel model per a l'exemple i. L'error quadrat mitjà mesura la distància quadràtica promig entre els valors reals i els valors predits. Minimitzar aquesta pèrdua ens permet ajustar millor el model a les dades d'entrenament.

D'altra banda, el terme de regularització s'utilitza per controlar la complexitat del model i ajudar a prevenir el sobreajustament. La regularització introdueix una penalització addicional a la funció de pèrdua per evitar que els paràmetres del model prenguin valors molt grans. Això ajuda a obtenir un model més generalitzat que tingui una millor capacitat de fer prediccions en dades noves.

A continuació detallem la base teòrica dels models que hem fet servir en aquest treball.

3.1 XGBoost

És l'abreviatura d'*Extreme Gradient Boosting*. És un dels primers models dins del marc de l'algoritme de *gradient boosting*[2], utilitzat per a la construcció de models de regressió supervisada. Aquesta tècnica consisteix en entrenar i combinar models individuals per obtenir una única predicció. XGBoost té com a base un conjunt de models bàsics o "base learners".

L'*Ensemble leaning* implica la combinació de múltiples models individuals per aconseguir una predicció final. XGBoost està dissenyat de manera que els base learners siguin uniformement dolents entre ells, de manera que les prediccions dolentes es cancel·lin mútuament i només es conservin les millors prediccions per formar una predicció final de qualitat[26].

XGBoost utilitza conjunts d'arbres de decisió [27]com a base per construir el seu model. Aquests conjunts estan formats per arbres de classificació i regressió (CART). En un CART, a diferència dels arbres de decisió convencionals, cada fulla té associada una puntuació numèrica. Aquesta puntuació permet tenir interpretacions més riques i un enfocament unificat basat en principis d'optimització.

En pràctica, un sol arbre de decisió sovint no és suficientment fort per ser utilitzat de manera efectiva. En canvi, es recorre a l'ús de models de conjunt (*ensemble*),que combinen les prediccions de diversos arbres. Considerant un conjunt de dades $D = (x_i, y_i)$, on n és el nombre d'exemples i m és el nombre d'atributs, la predicció del conjunt es pot expressar com [28]:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{3}$$

on K és el número d'arbres, f_k és una funció en l'espai funcional F ($F = f(x) = w_q(q : \Re^m - > T, w \in \Re^T)$, i F és el conjunt de tots els CART possibles. Aquí q representa l'estructura de cada arbre i assigna un exemple a la fulla corresponent[29]. T és el nombre de fulles de l'arbre.

Cada f_k correspon a una estructura d'arbre independent q i pes de fulla w. A diferència dels arbres de decisió, cada arbre de regressió té una puntuació contínua a cada fulla. Utilitzem w_i per representar la puntuació a la fulla i. Les regles de decisió de l'arbre, donades per q, s'utilitzen per classificar els exemples a les fulles corresponents, i la predicció final es calcula sumant les puntuacions de les fulles corresponents, que estan donades per w.

La funció objectiu a optimitzar en XGBoost es defineix com:

$$obj(\theta) = \sum_{i}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \omega(f_k) \tag{4}$$

[2]El gradient boosting és una tècnica d'aprenentatge automàtic que s'utilitza en tasques de regressió i classificació, entre d'altres. Ofereix un model de predicció en forma d'un conjunt de models de predicció febles, que normalment són arbres de decisió

On la funció de regularització ω s'utilitza per controlar la complexitat de l'arbre i està definida com:

$$\omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2 \tag{5}$$

on γ i λ són hiperparàmetres que governen la importància de la complexitat de l'arbre en la funció objectiu. Aquesta formulació és similar a la utilitzada en Random Forest, un altre algoritme d'aprenentatge basat en arbres.

En XGBoost, les funcions f_i que defineixen l'estructura de l'arbre i les puntuacions de les fulles s'aprenen de forma additiva, corregint el que s'ha après fins a l'etapa anterior i afegint un arbre nou a cada iteració. La predicció final $\hat{y}_i^N$ per a la instància i a la iteració N es calcula com la suma de les prediccions de tots els arbres fins a la iteració t tenim:

$$\hat{y}_i^N = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{t-1} + f_t(x_i) \tag{6}$$

El model de conjunt d'arbres a l'eq. (5) inclou funcions com paràmetres i no es pot optimitzar mitjançant mètodes d'optimització tradicionals a l'espai euclidià. Cal afegir f_t per tal de minimitzar la següent funció objectiu:

$$obj^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \omega(f_t) + c \tag{7}$$

L'objectiu és trobar l'arbre f_t que millori el model segons aquesta funció objectiu. Per simplificar la formulació, s'utilitza una expansió de Taylor de segon ordre de la funció de pèrdua al voltant de $\hat{y}_i^{t-1}$. Això ens permet reescriure la funció objectiu com:

$$obj^t = \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \omega(f_t) + c \tag{8}$$

On g_i i h_i són les primeres i segones derivades parcials de la funció de pèrdua respecte a $\hat{y}_i^{t-1}$. Aquestes derivades són estadístiques de gradient de primer i segon ordre. Eliminant els termes constants, obtenim l'objectiu simplificat a l'iteració t:

$$\tilde{obj}^t = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \omega(f_t) \tag{9}$$

Un avantatge important d'aquesta formulació és que el valor de l'objectiu només depèn de les estadístiques g_i i h_i, el que permet utilitzar funcions de pèrdua personalitzades en XGBoost.

Podem reescriure la funció fent servir la definició de ω:

$$\tilde{obj}^t = \sum_{i=1}^{n} [g_i w_q(x_i) + \frac{1}{2} h_i w_q^2(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 = \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \tag{10}$$

on $I_j = i|q(x_i = j$ és el conjunt d'índexs de punts de dades assignats al j-èssima fulla. En la segona línia, s'ha reescrit la suma per agrupar els punts de dades de la mateixa fulla.

En aquesta equació, w_j són independents entre sí, i la forma $\sum_{i \in I_j} g_i) w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda) w_j^2$ és quadràtica. La millor puntuació w_j per a una estructura q donada és:

$$*w_j = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{11}$$

A partir d'aquí, podem obtenir una funció de puntuació per mesurar la qualitat d'una estructura d'arbre q utilitzant aquesta fórmula:

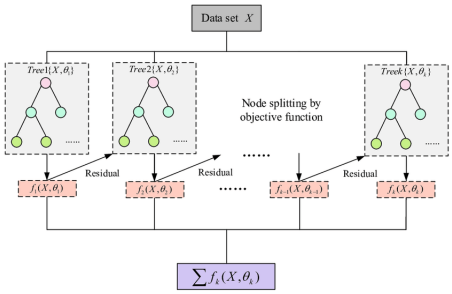

Figura 1: Diagrama XGBoost [30]

$$*obj^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{12}$$

Enumerar totes les combinacions possibles d'arbres per triar el millor és intractable en la pràctica. En lloc d'això, l'algorisme parteix d'una sola fulla i afegeix iterativament branques a l'arbre, optimitzant un nivell a la vegada. Aquest procés implica intentar dividir una fulla existent en dues fulles més petites. La reducció de pèrdua d'una divisió es calcula mitjançant la fórmula següent:

$$L_{divisio} = \frac{1}{2}[\frac{(\sum_{i\in I_E} g_i)^2}{\sum_{i\in I_E} h_i + \lambda} + \frac{(\sum_{i\in I_D} g_i)^2}{\sum_{i\in I_D} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}] - \gamma \tag{13}$$

La puntuació de la nova fulla esquerra, la puntuació de la nova fulla dreta, la puntuació de la fulla original i la regularització de la fulla addicional s'utilitzen per calcular aquesta reducció de pèrdua. Si la reducció de pèrdua és inferior a un llindar γ, es considera que afegir la branca no aporta suficient guany i pot ser millor no fer-ho. Aquesta és una tècnica de poda comú en els models basats en arbres.

3.2 Random Forest

L'algorisme de *bootstrapping*[3] [31]. Random Forest és una tècnica que combina mètodes d'*ensemble learning* amb arbres de decisió. Aquest algorisme genera múltiples arbres de decisió, dibuixats aleatòriament a partir de diverses submostres, i utilitza l'agregació mitjançant la mitjana per millorar la precisió predictiva i controlar el sobreajustament.

El procés de construcció d'un arbre de decisió [32] implica descompondre un conjunt de dades en subconjunts cada cop més petits mentre es desenvolupa un arbre de decisions associat. Aquest arbre consta de tres components principals: el node arrel, els nodes de decisió i els nodes fulla. El node arrel és el punt de partida on es realitzen les divisions del conjunt de dades. Els nodes de decisió són els punts de divisió resultants, mentre que els nodes fulla són les terminacions de l'arbre on no es realitzen més divisions.

Per aplicar l'algorisme Random Forest, es parteix del conjunt de dades original D i s'especifica el nombre d'arbres de decisió K a generar. Cada arbre es construeix fins que cada node conté com a màxim N mostres (per a tasques de regressió, N sovint és 5). A més, en cada node de l'arbre, es seleccionen aleatòriament F atributs per a la divisió. La característica a utilitzar per a la divisió és triada d'aquest conjunt aleatori d'atributs (F sol ser igual a la arrel quadrada del nombre total d'atributs en el conjunt de dades original D per a tasques de regressió). Així, Random Forest crea K subconjunts de dades a partir de D, i les mostres que no estan incloses en cap subconjunt es consideren mostres "fora de la bossa".

[3] Bootstrapping és el procés de mostreig aleatori de subconjunts d'un conjunt de dades durant un nombre determinat d'iteracions i un nombre determinat de variables. Aquests resultats es fan una mitjana conjuntament per obtenir un resultat més potent.

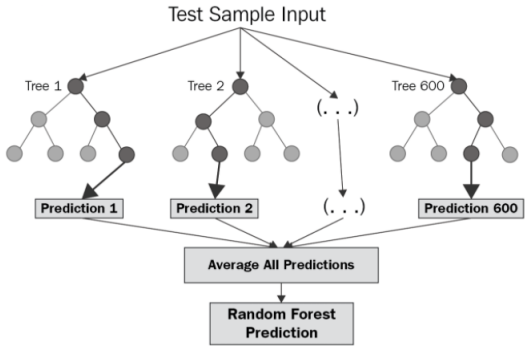

Figura 2: Diagrama Random Forest [33]

Els models entrenats de Random Forest formen un conjunt, i el resultat final d'una tasca de regressió s'obté fent la mitjana de les prediccions realitzades pels arbres individuals. Això permet obtenir una estimació més precisa i robusta, ja que els diferents arbres poden compensar els errors de cada un i proporcionar una predicció més fiable.

3.3 Prophet

És un software de codi obert publicat per l'equip principal de Data Science de Facebook. Es basa en un model estructural bayesià local[34] i utilitza un model additiu per a la predicció de sèries temporals. Aquest model s'ajusta a les tendències no lineals, incloent l'estacionalitat anual, setmanal i diària, així com els efectes de les vacances[35]. Per a modelar aquests efectes, Prophet utilitza sèries de Fourier o variables simulades. Aquesta eina és particularment eficaç en sèries temporals amb efectes estacionals forts i diverses temporades de dades històriques. En el seu nucli hi ha la suma de tres funcions de temps més un terme d'error per a fer la predicció, $y(t) = c(t) + e(t) + v(t) + e(t)$ que representen el **creixement c**, **estacionalitat e**, **vacances v** i **error e**.

La funció de **creixement** modela la tendència general de les dades, que pot ser present en tots els punts o canviar en els moments anomenats *changepoints* de Prophet, que són els punts on les dades experimenten un canvi de direcció.

Prophet ofereix dues opcions per al model de creixement, un logístic i un altre definit a trossos.

Per defecte, Prophet utilitza un model lineal definit a trossos. No obstant això, si les dades a predir segueixen un patró no lineal saturant, és a dir, un creixement no lineal que s'aproxima a una funció exponencial i després es manté estable o mostra canvis estacionals, llavors el model de creixement logístic és l'opció més adequada.

El model de creixement logístic s'ajusta mitjançant la següent equació estadística,

$$c(t) = \frac{C}{1 + e^{K(t-m)}} \tag{14}$$

on, C és la capacitat de càrrega (valor màxim de la corva), k és la taxa de creixement, i m és un paràmetre de compensació.

El model definit a trossos és una opció més adequada per a un creixement sense saturació i amb una taxa de creixement constant. Es representa mitjançant les següents equacions:

$$\begin{cases} \beta_0 + \beta_1 x & x \leq c \\ \beta_0 - \beta_2 c + (\beta_1 + \beta_2)x & x > c \end{cases}$$

on c és el punt de canvi de tendència i β és el paràmetre de tendència que es pot ajustar segons les necessitats de la predicció.

La funció d'**estacionalitat** consisteix simplement en una sèrie de Fourier en funció del temps que proporciona flexibilitat al model per a capturar components d'estacionalitat setmanal, anual i diària [36]. S'expressa com:

$$e(t) = \sum (a_n cos(\frac{2\pi nt}{P}) + b_n sin(\frac{2\pi nt}{P})) = X(t)\beta \tag{15}$$

on,

- $X(t) = [cos(\frac{2\pi 1t}{P}), sin(\frac{2\pi 1t}{P}), ..., cos(\frac{2\pi Nt}{P}), sin(\frac{2\pi Nt}{P})$
- $\beta = [a_1, b_1, ..., a_N, b_N]$ (vector de pesos)
- N és l'ordre de les sèries de Fourier
- P és el període (any, mes, setmana, etc.)

L'estacionalitat es calcula mitjançant una suma de Fourier parcial, i l'ordre de Fourier determina la velocitat a la qual pot canviar l'estacionalitat. Augmentar aquest ordre permet adaptar-se a cicles d'estacionalitat més ràpids.

La funció de **vacances** permet a Prophet ajustar la predicció quan hi ha vacances o esdeveniments importants que poden afectar-la. Es proporciona una llista de dates i, quan cada data està present a la predicció, s'afegeix o s'estalvia valor als termes de creixement i estacionalitat basats en les dades històriques[37]. Aquest efecte es descriu com:

$$Z(t) = [1(t \in D_1, ..., 1(t \in D_L] \tag{16}$$

$$v(t) = Z(t)k \tag{17}$$

on $Z(t)$ modela l'efecte linear de les vacances mitjançant un vector de variables indicadores (*dummy*), on el valor 1 indica la presència de vacances i el valor 0 indica la seva absència.El terme k representa l'efecte de les vacances. Aquest segueix una distribució normal de mitjana zero i variància 0.5 $k \sim N(0, var^2) \sim N(0, 0.5)$ [38]

La predicció final de Prophet contindrà una predicció per a cada valor històric present al conjunt de dades, així com previsions addicionals per al nombre de períodes futurs determinat pel mètode. A més, aquesta predicció estarà acompanyada de les parts inferiors i superiors de l'interval de confiança, que proporcionen una estimació de la incertesa associada a la previsió.

3.4 DeepAR

DeepAR és un algorisme de previsió de sèries temporals escalars mitjançant una xarxa neuronal recurrent amb neurones LSTM, desenvolupat per Amazon. Aquest algorisme utilitza l'aprenentatge supervisat per entrenar el model amb les dades històriques de les sèries temporals incloses en el conjunt de dades. Mitjançant l'entrenament simultani en diverses sèries temporals, el model DeepAR és capaç d'aprendre els comportaments complexos i les dependències entre les sèries, el que sovint resulta en un millor rendiment en comparació amb els mètodes estàndard com ARIMA i ETS[39].

Durant l'entrenament, s'utilitza una o més sèries temporals com a entrada per a la construcció del model, el qual aprèn una representació d'aquest procés o processos i l'utilitza per predir l'evolució de la sèrie temporal objectiu. Aquest procés d'entrenament es realitza mitjançant la presa de mostres aleatòries de diversos exemples d'entrenament, seleccionats de les diferents sèries temporals del conjunt de dades d'entrenament [40]. Per controlar fins a quin punt es poden fer prediccions en el futur, s'utilitza l'hiperparàmetre l·longitud de predicció", mentre que l'hiperparàmetre l·longitud de context"determina fins a quin punt en el passat la xarxa té accés per realitzar les prediccions.

Passem ara a descriure aquest procés en termes matemàtics. Si denotem el valor de la sèrie temporal i en el temps t com $z_{i,t}$, el nostre objectiu és modelar la distribució condicional del futur de cada sèrie temporal, donat el seu passat, representada per $P(z_i, t_0 : T | z_i, 1 : t_0 - 1, x_i, 1 : T)$

[21]. Aquí t_0 indica el punt de temps a partir del qual considerem que $z_{i,t}$ és desconegut en el moment de la predicció, i $x_{i,1:T}$ són covariables que se suposa que són conegudes en tot moment.

El model es basa en una arquitectura de xarxa recurrent autoregressiva. Suposem que el model de distribució $Q_\Theta(z_{i,t_0:T}|z_{i,1:t_0-1}, x_{i,1:T})$ consta d'un producte de factors de versemblança

$$Q_\Theta(z_{i,t_0:T}|z_{i,1:t_0-1}, x_{i,1:T}) = \prod_{t=t_0}^{T} Q_\Theta(z_{i,t}|z_{i,1:t-1}, x_{i,1:T}) = \prod_{t=t_0}^{T} L(z_{i,t}|\theta(h_{i,t}, \Theta)) \tag{18}$$

parametritzat per la sortida $h_{i,t}$ d'una xarxa recurrent autoregressiva

$$h_{i,t} = h(h_{i,t-1}, z_{i,t-1}, x_{i,t}, \Theta) \tag{19}$$

on h és una funció implementada per una xarxa neuronal recurrent multicapa amb cèl·lules LSTM. Aquest model és autoregressiu ja que utilitza l'observació de l'últim pas de temps $z_{i,t-1}$ com a entrada, i és recurrent ja que la sortida anterior de la xarxa $h_{i,t-1}$ es retroalimenta com a entrada en el següent pas de temps. La versemblança $L(z_{i,t}|\theta(h_{i,t}))$ és una distribució fixa els paràmetres de la qual es donen mitjançant una funció $\theta(h_{i,t}, \Theta)$ de la sortida de la xarxa $h_{i,t}$. La selecció de la versemblança $L(z|\theta)$ és important perquè ha de correspondre amb les propietats estadístiques de les dades i permet determinar el model de soroll en el procés de previsió.

La informació relativa a les observacions en el rang de condicionament $z_{i,1:t_0-1}$ és transferida al rang de predicció mitjançant l'estat inicial h_{i,t_0-1}. En la configuració seqüència a seqüència, aquest estat inicial és la sortida d'una xarxa de codificadors.

Donats els paràmetres del model Θ, podem obtenir mostres conjuntes

$\hat{z}_{i,t_0:T} \sim Q\Theta(z_{i,t_0:T}|z_{i,1:t_0-1}, x_{i,1:T})$ directament mitjançant mostreig ancestral: en primer lloc, calculem h_{i,t_0-1} utilitzant l'equació (20) per a $t = 1, ..., t_0$. Per a $t = t_0, t_0 + 1, ..., T$ mostrem $\hat{z}_{i,t} \sim L(\cdot|\theta(\hat{h}_{i,t}, \Theta))$ on $\hat{h}_{i,t} = h(h_{i,t-1}, \hat{z}_{i,t-1}, x_{i,t}, \Theta)$ amb inicialització $\hat{h}_{i,t0-1} = h_{i,t_0-1}$ i $\hat{z}_{i,t_0-1} = z_{i,t_0-1}$.

Els paràmetres Θ del model, que inclouen els paràmetres de la xarxa neuronal recurrent $h(\cdot)$, així com els paràmetres de $\theta(\cdot)$, es poden aprendre maximitzant la log-versemblança

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{t=t_0}^{T} logL(z_{i,t}|\theta(h_{i,t})) \tag{20}$$

Com que $h_{i,t}$ és una funció determinista de l'entrada, s'observen totes les quantitats necessàries per calcular la log-versemblança, de manera que no és necessari realitzar cap inferència i la log-versemblança es pot optimitzar directament mitjançant descens de gradient estocàstic, calculant gradients respecte a Θ.

Per facilitar l'aprenentatge de patrons que varien en el temps, com ara pics durant els caps de setmana, DeepAR crea automàticament atributs de temps basts en la freqüència de la sèrie temporal objectiu. Aquests atributs s'utilitzen juntament amb els atributs personalitzats proporcionats durant l'entrenament. Entre aquests atributs derivats es troben minut de la hora, hora del dia, el dia de la setmana, dia del mes, mes de l'any, dia de l'any i setmana de l'any.

La imatge següent mostra el procés d'entrenament a l'esquerra i el de predicció a la dreta:

Per capturar patrons d'estacionalitat, DeepAR incorpora automàticament valors retardats (*lagged*) de la sèrie temporal objectiu.

Durant la inferència, el model entrenat pren com a entrada sèries temporals objectiu, que poden o no haver estat utilitzades durant l'entrenament, i realitza una predicció de la distribució de probabilitat pels següents valors del paràmetre de llargada de predicció. Com que DeepAR s'entrena amb totes les dades del conjunt, la predicció té en compte els patrons apresos de sèries temporals similars.

L'algorisme avalua la precisió de la distribució de la previsió mitjançant la pèrdua de quantil ponderada. Per a un quantil en el rang [0, 1], la pèrdua quantil ponderada es defineix de la següent manera [41]:

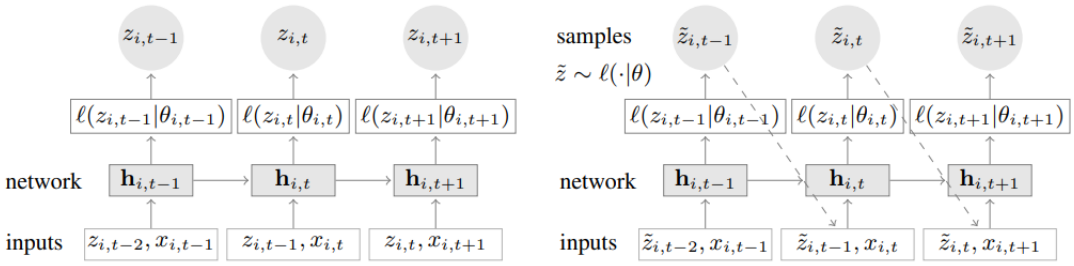

Figura 3: Diagram schem of DeepAR [21]

$$\text{wQuantileLoss}[\tau] = 2\frac{\sum_{i,t} Q_{i,t}^{(\tau)}}{\sum_{i,t} |y_{i,t}|}, \qquad Q_{i,t}^{(\tau)} = \begin{cases} (1-\tau)|q_{i,t}^{(\tau)} - y_{i,t}| & \text{if } q_{i,t}^{(\tau)} > y_{i,t} \\ \tau|q_{i,t}^{(\tau)} - y_{i,t}| & \text{otherwise} \end{cases}$$

on $q_{i,t}^{(\tau)}$ representa el τquantil de la distribució predita pel model[42].

3.5 Mètriques

En l'anàlisi de múltiples sèries temporals, se presenta un desafiament significatiu: la selecció adequada de la mètrica per avaluar l'error de la predicció. L'avaluació del rendiment de qualsevol model d'aprenentatge automàtic és crucial, tant des d'un punt de vista tècnic com empresarial. Especialment quan les decisions empresarials depenen dels coneixements generats a partir dels models de previsió, és vital conèixer amb precisió la seva rendiment.

Moltes mètriques populars es coneixen com a dependents de l'escala [43]. D'això significa que les mètriques d'error s'expressen en les mateixes unitats (per exemple, dòlars, polzades, etc.) que les dades. El principal avantatge d'aquestes mètriques és que solen ser fàcils de calcular i interpretar [44]. No obstant això, no es poden utilitzar per comparar sèries diferents a causa de la seva dependència de l'escala.

Error Absolut Mitjà (MAE)

L'Error Absolut Mitjà (MAE) es calcula mitjançant la mitjana de les diferències absolutes entre els valors reals (anomenats y) i els valors predits (y_hat).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{21}$$

Aquesta mètrica és fàcil de comprendre i calcular, i es recomana per avaluar la precisió en una sola sèrie [43]. No obstant això, no és adequada per comparar sèries diferents amb unitats diferents. A més, no s'ha de utilitzar si es vol penalitzar els valors atípics.

Error Quadràtic Mitjà (MSE)

Si es desitja donar més importància als valors atípics, es pot considerar l'Error Quadràtic Mitjà (MSE). Com indica el seu nom, es calcula mitjançant la mitjana dels errors al quadrat (diferències entre y i y_hat). Degut a la seva elevació al quadrat, els errors grans tenen un pes més significatiu que els errors petits, cosa que pot ser un inconvenient en certes situacions. Així doncs, el MSE és apropiat en situacions en què es vol donar un major èmfasi als errors importants. Cal tenir en compte que, a causa de l'elevació al quadrat, la mètrica perdrà la seva unitat original.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{22}$$

Root Mean Squared Error (RMSE)

El Root Mean Squared Error (RMSE) evita la pèrdua de la unitat que ocorre amb el MSE al treure'n l'arrel quadrada. De manera similar al MSE, l'RMSE també és sensible als valors atípics.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{23}$$

Algunes fonts, com ara [45], argumenten que l'RMSE és una mesura inadequada i mal interpretada d'un error mitjà, i recomanen utilitzar el MAE en lloc de l'RMSE. No obstant això, [46] suggereixen utilitzar l'RMSE per a l'optimització del model i per a l'avaluació de diferents models en situacions on s'espera que la distribució d'errors sigui gaussiana.

Les mètriques dependents de l'escala no són adequades per a comparar diferents sèries temporals. En canvi, les mètriques d'error percentual resolen aquest problema. Aquestes mètriques són independents de l'escala i s'utilitzen per comparar el rendiment de les prediccions entre diferents sèries temporals. No obstant això, presenten una limitació quan es tracta de valors zero en una sèrie temporal, ja que aleshores les mètriques es converteixen en valors infinits o indefinits, la qual cosa els fa poc interpretables [43].

Error de Percentatge Absolut Mitjà (MAPE)

L'Error de Percentatge Absolut Mitjà (MAPE) es calcula mitjançant la mitjana de les diferències absolutes entre els valors reals i els valors predits, dividida pels valors reals i multiplicada per 100.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i} * 100\right| \tag{24}$$

Els avantatges del MAPE són la seva independència de l'escala i la seva fàcil interpretació. No obstant això, aquesta mètrica genera valors infinits o indefinits quan els valors reals són zero o propers a zero [47]. A més, el MAPE penalitza més els errors negatius que els positius, la qual cosa pot conduir a una asimetria en la seva aplicació [43].

Error percentual absolut mitjà simètric (SMAPE)

Per abordar l'asimetria del MAPE, es va proposar l'error percentual absolut mitjà simètric (SMAPE). El SMAPE és una de les mètriques d'error més controvertides, ja que hi ha diferents definicions i fórmules disponibles, i alguns crítics afirmen que aquesta mètrica no és veritablement simètrica, malgrat el seu nom [48].

El SMAPE es calcula mitjançant la mitjana de totes les previsions realitzades per a un horitzó de temps determinat. Els seus avantatges radiquen en la capacitat d'evitar el problema del MAPE amb errors grans quan els valors reals (y) són propers a zero, i en la reducció de les diferències significatives entre els errors percentuals absoluts quan el valor real (y) és més gran o més petit que el valor predit (y_hat).A diferència del MAPE, el SMAPE oscil·la entre el 0% i el 200% i té límits ben definits [49]. No obstant això, [48] assenyala que el SMAPE penalitza més les subestimacions que les sobreestimacions.

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{25}$$

La Cobertura

La mètrica de cobertura (Coverage) calcula el percentatge de prediccions que superen el valor real. Aquesta és una tècnica utilitzada per avaluar la qualitat del model de predicció.

$$Coverage = \frac{100}{n} \sum_{i=1}^{n} (\hat{y}_i < y_i) \tag{26}$$

Tot i que aquesta mètrica ens permet interpretar el funcionament del nostre model, no proporciona un error numèric concret. En canvi, ens dóna una indicació de la proporció de prediccions que són més altes que els valors reals, cosa que pot ser útil per a determinar la qualitat del model en termes de sobreestimació o subestimació.

En resum, quan totes les sèries temporals estan en la mateixa escala i l'objectiu és avaluar el rendiment de la predicció, es pot preferir l'ús del MAE, ja que és més senzill d'explicar [43]. No obstant això, en cas de tenir dades amb valors extrems, el MAE resulta ineficient, igual que el MAPE. Per exemple, si tenim un districte on es registren una mitjana de 40 agressions mensuals i un altre on la mitjana és de 0, el MAPE sobrevalorarà les sèries amb baix volum, mentre que el SPMAE farà el contrari.

El MSE és beneficiós quan la variabilitat de les prediccions és significativa i es vol penalitzar més els errors amb valors més grans [50]. No obstant això, com que és un valor al quadrat, aquesta mètrica sovint és difícil de comprendre. Per abordar aquesta dificultat i tenir una mètrica més interpretable, es pot utilitzar el RMSE [46], especialment quan s'espera una distribució gaussiana dels errors i per corregir l'error de les unitats de mesura.

No existeix una única mètrica que pugui combinar adequadament totes les sèries, ja siguin de baix o alt volum. Quan es vol fer una mitjana entre sèries amb volums molt diferents, sorgeixen problemes. En aquest cas, s'ha de proporcionar una mètrica global per a cada sèrie, tenint en compte que cada una pot tenir una escala completament diferent.

Una opció seria normalitzar les dades, però en el nostre anàlisi aquesta opció no té sentit. Per exemple, en el cas de les vendes de supermercats, on hi ha una distribució exponencial de les vendes i un grup reduït de sèries representa la major part de les vendes, la normalització convertiria aquesta distribució en lineal. Això implicaria que el 90% de les sèries, que només representen el 10% de les vendes, tindrien un pes desproporcionat. En altres paraules, s'estaria ignorant el 10% de les sèries que representen el 90% de les vendes.

En definitiva, no existeix una única mètrica d'error. Cada mètrica té els seus avantatges i debilitats. Per tant, l'elecció de la mètrica depèn sempre del cas d'ús o l'objectiu específic, així com de les dades subjacents. És important no limitar-se a analitzar una sola mètrica d'error quan s'avalua el rendiment del model.

4 Modelitzant el comportament humà en dades de consum

En primer lloc, es va iniciar un estudi per a avaluar fins a quin punt aquests models poden arribar en la predicció del comportament humà. Es va triar com a exemple un comportament bastant típic, però amb un grau elevat de treball i complexitat: el comportament de compres i vendes, participant en la competició de Kaggle *Store Sales Time Series Forecasting*. L'objectiu d'aquesta competició consisteix a crear diversos models que prediguessin amb major precisió les vendes per unitats de milers d'articles venuts a diferents botigues de Corporación Favorita, una gran botiga d'aliments.

En el context del comerç minorista, les previsions probabilístiques de l'oferta i la demanda de productes són d'utilitat per a una gestió òptima de l'inventari o la planificació del personal, i són considerades una tecnologia crucial per a la majoria dels aspectes de l'optimització de la cadena de subministrament. Es va partir de la idea que el millor model seria presentat a la competició de Kaggle.

Les dades d'entrenament inclouen dates, informació sobre la botiga i el producte, com el número de la botiga, família del producte, nombre total d'articles d'una família de productes que s'estaven

promocionant, així com les xifres de vendes. Les dates es comprenen en el període del 01/01/2013 al 31/08/2017. Per altra banda, les dades de prova tenen les mateixes característiques que les dades d'entrenament. Serveixen per predir les vendes objectiu per a les dates corresponents i corresponen als 15 dies posteriors a l'última data de les dades d'entrenament (del 16/08/2017 al 31/08/2017).

A nivell de variables endògenes i exògenes, s'inclouen altres dades relacionades amb la localització de la botiga i del preu diari del petroli, donat que l'Equador és un país que depèn del petroli i la seva salut econòmica és molt vulnerable als xocs dels preus del petroli.

Cal destacar diverses observacions sobre els conjunts de dades. Pel que fa a les festivitats, un festiu que està oficialment marcat com a transferit cau en aquest dia natural, però el govern decideix traslladar-lo a una altra data. Un dia transferit es semblarà més a un dia normal que a un dia festiu. Per exemple, la festa de la Independència de Guayaquil es va traslladar del 2012-10-09 al 2012-10-12, el que significa que es va celebrar el 2012-10-12. Els dies del tipus *Bridge* són dies addicionals que s'afegeixen a un festiu (p. ex., per allargar el descans durant un cap de setmana llarg). Els dies festius addicionals són dies afegits a un dia festiu del calendari normal, com sol succeir al voltant de Nadal (convertint la Nit de Nadal en un dia festiu).

En relació amb la particularitat de la zona on es processen les dades, és essencial tenir en compte que en el sector públic els salaris es paguen quinzenalment, els dies 15 i el darrer dia de cada mes. Aquesta circumstància pot afectar les vendes dels supermercats. A més a més, cal tenir present que l'Equador va patir un terratrèmol de magnitud 7,8 el 16 d'abril de 2016, el qual va ocasionar una resposta massiva d'ajuda humanitària. La població es va mobilitzar per participar en els esforços de socors, proporcionant aigua i altres productes bàsics, fet que va tenir un impacte significatiu en les vendes dels supermercats durant diverses setmanes posteriors al sisme.

4.1 Detalls d'implementació

En relació amb el preprocessament de les dades, el primer pas ha consistit en realitzar la neteja individual de cadascun dels conjunts de dades, tenint en compte la seva gran magnitud. Després de cada neteja, s'han fusionat els diferents conjunts de dades aplicant diversos mètodes específics.

Pel que fa a la neteja de les dades, el primer pas consisteix en eliminar les entrades duplicades presents en tots els conjunts de dades. A més, s'ha seleccionat únicament aquells esdeveniments que no han estat transferits, les festes com a tipus nacionals (no es categoritzen entre locals o regionals ja que sembla impactar de forma negativa el rendiment) i es fan correccions en esdeveniments amb dates mal assignades per tal de tenir una concordança en les dades. A continuació, en el procés de neteja de les dades, s'han omplert els valors buits de les transaccions amb el promig diari de transaccions per botiga.

Posteriorment, s'han generat diversos atributs addicionals, com ara la categoria *unique store* per identificar aquelles botigues que són exclusives d'una localitat, la categoria *new store* per a les botigues que han obert després de la data d'inici del conjunt de dades. També s'ha creat l'atribut *wd* (work day) per a definir els dies laborables (True) i festius (False), així com l'atribut *isclosed* per a les dates importants com Pasqua o el primer dia de l'any. Finalment, s'han afegit noves característiques relacionades amb el temps, incloent els períodes en què comencen les escoles (abril-maig i agost-setembre, ja que és important captar la seva estacionalitat en les vendes de productes escolars), i les dates del terratrèmol d'abril de 2016.

Un altre atribut que ha tingut un efecte molt positiu en la predicció de les dades és la inclusió d'atributs relacionats amb els productes en promoció (*onpromotion*). S'han inclòs atributs que indiquen la mitjana de productes en promoció setmanals, bisetmanals, mensuals, etc., així com les mitjanes de productes en promoció per a cada botiga i per a diferents desfasaments en el temps (lags). Aquests desfasaments han estat especialment importants per a la predicció, ja que han permès que les sèries temporals retardades siguin útils com a característiques per modelar la dependència en sèrie. També s'han afegit atributs que indiquen la mitjana de l'oli en períodes setmanals, bisetmanals, mensuals, etc

Finalment, s'ha pres la decisió de començar l'anàlisi de les dades a partir del 30 d'abril de 2017, tenint en compte que l'última botiga va obrir el 24 d'abril de 2017. Aquesta elecció ens permet disposar d'un període de temps suficient per a estabilitzar les vendes abans d'iniciar l'avaluació. Cal destacar que tant l'entrenament com la predicció es realitzen mitjançant el càlcul de l'error per

a cada botiga, tenint en compte les famílies de vendes. Aquest enfocament ens ajuda a comprendre millor la correlació entre les diferents categories de vendes.

En el que fa a la modelització de les dades, s'ha aplicat una transformació logarítmica a les dades de vendes. Aquesta transformació s'ha utilitzat per abordar un problema de linealitat detectat en el diagrama de residuals, que probablement estigui relacionat amb el creixement compost de les dades. A més, s'ha adoptat un procés determinista per garantir que es produeixi sempre la mateixa sortida a partir de les mateixes condicions inicials o estat. Això ens permet obtenir resultats coherents i repetibles en l'anàlisi de les dades.

4.2 Avaluació dels mètodes

A continuació, es va desenvolupar l'entrenament i les prediccions. En primer lloc, s'han seleccionat els atributs que s'han determinat, mitjançant la selecció de característiques (Feature Selection), que funcionen de manera òptima per a cada un dels nostres models. La realització de l'entrenament i la prova per a cada botiga del nostre conjunt de dades (54 botigues) ha donat lloc a una millora significativa de les prediccions. Cal destacar que l'afegiment d'una funció de cost amb pesos exponencials, que emfatitza el mes últim de vendes abans de la predicció, ha contribuït a una reducció addicional de l'error.

Per avaluar les prediccions, s'utilitza l'RMSLE (Root Mean Squared Logarithmic Error):

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(log(1+\hat{y}_i) - log(1+y_i))^2} \tag{27}$$

On N és el nombre total d'observacions al conjunt de dades (públic/privat), $\hat{y}_i$ és la predicció de l'objectiu, i y_i és l'objectiu real per a i.

És important tenir en compte que, a diferència del RMSE, el RMSLE és asimètric, penalitzant més les prediccions de valors més baixos que les de valors més alts de vendes. Per exemple, si suposem que el valor correcte és yi=1000, subestimar en 600 és gairebé el doble que sobreestimar en 600.

Aquesta asimetria es deu al següent raonament:

$$(log(1+\hat{y}_i) - log(1+y_i))^2 = log(\frac{1+\hat{y}_i}{1+y_i}) \tag{28}$$

Per tant, bàsicament estem considerant ràtios en lloc de diferències, com succeeix amb el RMSE. Per calcular l'RMSLE, en lloc de prendre directament la mitjana, primer s'aplica el logaritme a les vendes i, després, es calcula la mitjana dels valors transformats. Posteriorment, s'inverteixen els valors, és a dir, $y = np.expm1(mean(np.log1p(df.sales)))$. Aquesta es la y que es presenta com a submissió a Kaggle per a calcular l'RMSLE.

Així doncs, durant l'entrenament, els valors de vendes (les prediccions) són transformats utilitzant el logaritme i, posteriorment, els valors predits són invertits mitjançant l'exponencial. A més a més, s'ha considerat imprescindible dur a terme un procés d'ajustament dels hiperparàmetres per a cadascun dels models, amb l'objectiu de seleccionar els millors valors per a aquests paràmetres.

4.3 Anàlisi i comparativa dels resultats

El rendiment obtingut amb els diferents mètodes explicats en les seccions anteriors es pot observar en la Taula 1.

El *Weighted Model* correspon a un model on les prediccions finals es calculen a partir de 30% del XGBoost Regressor i la resta del Random Forest Regressor, percentatge escollit per validació creuada.

La fila *Top Kaggle* fa referència al millor valor dels notebooks públics de la competició en la submissió, consultat a data 1 de juny del 2023. Aquest resultat s'obté mitjançant DARTS

Model	**Error**
DeepAr	3.428
Prophet	3.247
XGBoost	0.424
Random Forest	0.405
Weighted Model	0.401
Top Kaggle[51]	**0.380**

Taula 1: Resultats competicio *Store Sales*

(Dropouts meet Multiple Additive Regression Trees), que és una llibreria de Python per a la previsió fàcil d'utilitzar i la detecció d'anomalies en sèries temporals[52].

Per a arribar als valors mostrats en aquesta taula, s'han ajustat els hiperparàmetres mitjançant una *Randomized Search*, algoritme que prova un conjunt d'hiperparàmetres de manera aleatòria i calcula el resultat, retornant el conjunt que generi els millors resultats després d'haver provat tots els hiperparàmetres necessaris. Els paràmetres de l'estimador utilitzats per a aquests mètodes s'han optimitzat mitjançant una cerca validada creuada sobre la configuració dels paràmetres.

A diferència del *GridSearchCV*, no es proven tots els valors dels paràmetres, sinó que es mostren un nombre fix de paràmetres de les distribucions especificades[53]. El nombre de paràmetres que s'han provat en el nostre cas ha estat el valor per defecte de *n_iter*, 10. Com que s'especifica almenys un paràmetre com a distribució, s'utilitza el mostreig amb reemplaçament.

Per a **XGBoost**, s'ha triat el nombre més elevat d'arbres que generi una diferència significativa. Aquest model es construeix seqüencialment, on cada nou arbre intenta corregir els errors comesos pels arbres anteriors. Ràpidament, el model arriba a un punt on el rendiment decreixent, per tant s'ha triat el valor d'arbres més alt que generava una diferència notable, 500 en aquest cas determinat. És important trobar un equilibri entre arbres poc profunds, que tenen un rendiment baix a causa de la seva limitada capacitat per tenir en compte els detalls del problema, i arbres massa profunds, que sobreajusten el conjunt de dades d'entrenament. En el nostre cas, s'ha determinat que un valor òptim per a la mida de l'arbre és 3, tot i que no s'ha observat una diferència notable entre els valors de 3 i 6.

A XGBoost, s'implementen nous arbres per corregir els errors residuals de la seqüència d'arbres existents. La taxa d'aprenentatge és un paràmetre clau que modula l'aprenentatge del model aplicant un factor de ponderació a les correccions dels arbres nous quan s'incorporen al model. L'establiment de valors inferiors a 1,0 té com a efecte una reducció en les correccions aportades per cada arbre afegit, requerint així la inclusió d'un major nombre d'arbres al model. És habitual utilitzar valors reduïts en el rang de 0,1 a 0,3, o fins i tot inferiors a 0,1. En aquest cas, s'ha determinat que el valor òptim és de 0,01, atès que el nombre d'arbres generats és elevat (500).

Pel que fa a **Random Forest**, s'ha assignat el nombre màxim possible d'arbres a generar. Random Forest obté millors prediccions a mesura que s'incrementa el nombre d'arbres abans de calcular la predicció mitjana. Per tant, s'ha triat el valor màxim que el processador pot suportar per tal d'obtenir prediccions més fortes i estables. Quant a la profunditat màxima dels arbres, s'ha seleccionat el valor màxim que no provoca sobreajustament. Un arbre amb major profunditat conté més informació de les dades, per la qual cosa s'ha optat pel valor de 50 en lloc de valors com 100 o superiors.

En l'anàlisi dels hiperparàmetres per a **Prophet** s'ha utilitzat el conjunt de dades *event_holidays*, que inclou totes les dates festives. La flexibilitat de la tendència, en particular, en quant canvia als punts de canvi, és un factor rellevant. Si el valor és massa petit, la tendència serà insuficient i la variància associada als canvis de tendència s'interpretarà com a soroll. D'altra banda, si el valor és massa gran, la tendència s'ajustarà en excés. Per controlar l'adaptabilitat als efectes de les vacances, s'ha establert un valor de 0,01, que indica una flexibilitat molt limitada. Pel que fa a la flexibilitat de l'estacionalitat, un valor elevat permet ajustar-se a grans fluctuacions, mentre que un valor reduït redueix la magnitud de l'estacionalitat. S'ha determinat que un valor petit és el més adequat en aquest cas.

El model additiu incorpora tendència, estacionalitat i altres efectes en les prediccions. Aquest model és adequat per al conjunt de dades en qüestió, ja que presenta una variació estacional relativament constant al llarg del temps. Això s'identifica mirant la sèrie temporal i si la magnitud de les fluctuacions estacionals creix amb la magnitud de la sèrie temporals. Altres paràmetres activen l'estacionalitat anual, setmanal i diària si hi ha un any, una setmana o un dia de dades, i desactivarà en cas contrari. Com que s'ha decidit tractar únicament les dades de 2017, l'estacionalitat anual és falsa i tampoc es troba cap estacionalitat diària donat que només tenim una dada de vendes per el total del dia.

A **DeepAR** destaca l'ús de el nombre de punts de temps que el model arriba a veure abans de fer la predicció. També rep entrades retardades de l'objectiu, de manera que el nombre de punts de temps que el model arriba a veure abans de fer la predicció pot ser molt més petit que les estacionalitats habituals. Per exemple, una sèrie temporal diària pot tenir estacionalitat anual. Així mateix, El nombre màxim de passades sobre les dades d'entrenament ajuda a determina la precisió de la predicció donat que determina l'aprenentatge de l'algoritme.

Per avaluar com de bones eren les prediccions dels models s'han comparat els resultats de l'execució amb un *mean model*. El *mean model* es tracta d'un model que prediu de manera constant el mateix valor, la mitjana, és a dir, la suma de les vendes dividit entre el número de dies. Tots els models anteriors han obtingut unes prediccions amb un error més petit que el del *mean model*, per tant, es pot dir que els models fan prediccions més precises i que tenen en compte les altres variables.

El model *Prophet* no ha semblat donar els resultats que s'esperaven. En una primera instancia es va suposar que els paràmetres que modulen la flexibilitat d'adaptació per als efectes de les vacances i l'estacionalitat estarien valors elevats donat que el s'ha vist que moltes vendes es veuen influenciades per festius que es puguin donar a les regions o l'estació de l'any com per exemple per a la venta de material escolar, però el model ha obtingut les millors prediccions quan aquests paràmetres se'ls assignava uns valors més baixos indicant una flexibilitat menor a aquests efectes. Aquest fet s'explica ja que s'ha limitat a només 4 mesos l'entrenament i dues setmanes per a les prediccions, per tant, els efectes de les vacances i la estacionalitat no són tan forts. Prophet no busca relacions casuals entre el passat i el futur. Simplement troba la millor corba per adaptar-se a les dades mitjançant un component de corba logística lineal per al regressor extern.

DeepAR és un model d'interès, ja que permet fer prediccions recurrents de sèries temporals per a cada botiga. Aquesta capacitat és crucial per comprendre la correlació de les vendes entre dues botigues, com ara aquelles situades a la mateixa zona geogràfica. No obstant això, l'aplicació de DeepAR presenta un desafiament significatiu en la transformació de les dades al format requerit pel model. La tasca de classificar cada variable com a dinàmica o estàtica, i com a categòrica o no categòrica, no és trivial i difícilment automatitzable. Malgrat això, s'han desenvolupat algunes funcions per facilitar aquest procés.

Per distingir entre variables estàtiques i dinàmiques, es compila una llista amb tots els elements únics de cada variable per a cada família i botiga. Si el nombre d'elements únics en relació amb el total de combinacions de famílies i botigues supera el 5%, es considera una variable dinàmica; en cas contrari, es considera estàtica. Pel que fa a la determinació de si una variable és categòrica o no, es fa servir un sistema de votació. Si una variable en una agrupació té més d'un element únic i el nombre d'elements únics és inferior al 10% de la longitud de la sèrie completa, o només hi ha un únic valor únic, es resta 1; sinó, s'afegeix 1,5. Si el total de vots al finalitzar l'avaluació per a totes les agrupacions és igual o superior a zero, la variable es considera numèrica; en cas contrari, es considera categòrica.

L'aplicació de DeepAR no ha obtingut els resultats esperats, en gran part per la limitació de temps per implementar aquest model, donada la seva complexitat. S'han enfrontat diversos problemes durant la categorització de les variables, la qual cosa ha retardat la seva implementació completa per obtenir prediccions òptimes. A més, les restriccions de recursos han limitat la precisió de l'algorisme, ja que el seu temps d'execució era elevat i presentava dificultats per a ser traslladat a la GPU.

En realitat, s'ha observat que altres models que incorporen Deep Learning per a altres participants de la competició han obtingut resultats lleugerament superiors. DeepAR és un algorisme

que ha demostrat ser altament eficaç en la predicció de sèries temporals. Amb més dedicació i recursos, seria possible obtenir resultats més satisfactoris.

El model *Weighted Model* ha demostrat ser el més efectiu en termes de resultats. Aquest model consisteix en l'aplicació conjunta de Random Forest i XGBoost a les dades de vendes, generant prediccions noves basades en un 70% de les prediccions de Random Forest i un 30% de les prediccions de XGBoost. Aquests percentatges van ser determinats mitjançant diverses proves, modificant les proporcions i els paràmetres i atributs seleccionats. No obstant això, sempre s'ha mantingut una major proporció per a Random Forest, ja que ha estat el model individual amb els millors resultats.

Malauradament, l'addició d'una regressió lineal a l'Stacking amb els models de XGBoost i Random Forest ha donat lloc a resultats lleugerament inferiors. Aquesta regressió lineal no ha funcionat adequadament per a aquest model, ja que els atributs tenen una alta correlació i no existeix una relació lineal clara entre aquests atributs i les vendes finals. Això limita la seva capacitat per capturar estacionalitats o picades que el model pugui exhibir.

Random Forest demostra un millor rendiment quan hi ha una alta proporció de dades mancants i quan el conjunt de dades és ampli. D'altra banda, XGBoost supera Random Forest en casos on les dades estan desbalancejades i les diferències en les vendes no són notables. És important destacar que Random Forest és més fàcil de sintonitzar que XGBoost, i s'ha observat una millora significativa en la fase final de l'execució gràcies als canvis realitzats als paràmetres. Per aquesta raó, sembla ser més adequat per a aquest conjunt de dades, tot i que la diferència en el rendiment no és molt significativa. La combinació d'aquests dos models ha donat lloc a resultats molt satisfactoris.

En l'última submissió del model *Weighted Model* amb l'error més baix, es va situar en la posició 29 de 660 equips participants, representant el 4.3% de les millors submissions. Amb les altres submissions, amb excepció de Prophet i DeepAR, es va trobar en el 8% de les millors prediccions.

5 Modelitzant el comportament humà en dades d'agressions de gènere

Amb l'objectiu d'aprofundir en la predicció d'agressions sexuals, s'analitzarà la base de dades de Kaggle denominada *Chicago Crimes*, la qual registra incidents delictius denunciats (excloent els homicidis amb dades per a cada víctima) ocorreguts a la ciutat de Chicago des del 2001 fins a la present, excepte els set dies més recents.

Un obstacle freqüent en la modelització i predicció d'agressions sexuals és la manca de detall en les bases de dades disponibles. En una sèrie temporal de baixa freqüència (és a dir, amb dades mensuals, trimestrals, etc.), les característiques més comunes són: (i) una tendència (el moviment a llarg termini d'una sèrie); (ii) estacionalitat (oscil·lacions a curt termini que es repeteixen en anys successius) i (iii) una variància (dispersió al voltant de la mitjana) que creix amb la mitjana[54].

D'altra banda, en una sèrie temporal d'alta freqüència (és a dir, amb dades diàries, horàries, etc.), s'observa: (i) una mitjana estable al llarg del temps, (ii) absència d'estacionalitat i (iii) una variància que canvia amb el temps, alternant períodes d'alta volatilitat (alta variància) amb períodes de baixa volatilitat (baixa variància). La variància canvia de forma no sistemàtica.

La freqüència de la sèrie temporal, ja sigui baixa o alta, pot afectar significativament la precisió de les prediccions. En el cas de la base de dades del consum en supermercats, que és d'alta freqüència, s'han obtingut resultats molt precisos. No obstant això, en el context de les agressions sexuals, no es troben bases de dades a nivell estatal que presentin les mateixes característiques de freqüència que les dades dels supermercats. Per exemple, en el cas de les dades públiques disponibles a molts països, inclosa Espanya, aquestes només es reporten de forma trimestral i els anàlisis se solen presentar anualment, tal i com es visualitza en la Fig. 4.

Per aquesta raó, s'utilitza la base de dades pública de Chicago. Aquestes dades s'extreuen del sistema CLEAR (Citizen Law Enforcement Analysis and Reporting) del Departament de Policia de Chicago, i la darrera data disponible és el 8 de febrer del 2023, que correspon al moment en què es va començar a processar la base de dades. Cal tenir en compte que aquestes dades inclouen informes no verificats enviats al Departament de Policia, i les classificacions preliminars dels delictes poden canviar posteriorment com a resultat d'una investigació addicional. A més, sempre hi ha la possibilitat d'errors mecànics o humans[55]. Aquestes dades inclouen la localització del

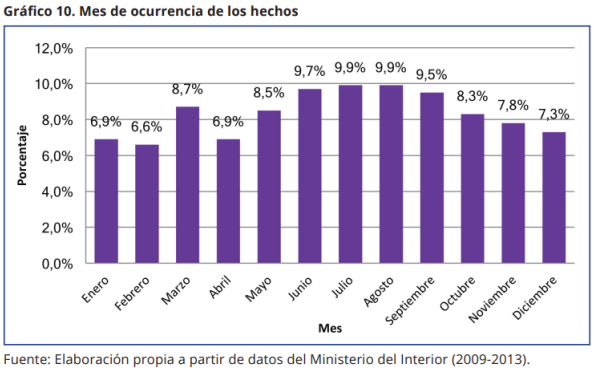

Figura 4: Total Agressions sexuals mensuals proporcionades pel Ministeri d'Interior [56]

crim i les coordenades, el codi de l'FBI, el número de la comunitat, el barri, el districte, si es va realitzar l'arrest, si el crim va ser domèstic i el tipus de crims entre d'altres. Per fer l'anàlisi de les agressions sexuals seleccionarem de la base de dades únicament els crims classificats com *Sex Offense* i *Crime Sexual Assault*.

5.1 Detalls d'implementació

Una característica rellevant de la base de dades és la tendència a acumular un gran nombre d'agressions el primer dia de cada mes. Aquesta acumulació es pot explicar pel fet que les dades provenen del sistema CLEAR, on s'agrupen totes les agressions conegudes dins d'un mes determinat, però sense una data específica. El mateix fenomen s'observa en relació a l'any, on les agressions es registren el primer dia de l'any, i pel que fa a les hores, on les agressions es registren a les 12:01 a. m. o a les 12:00 a. m. Aquestes suposicions es fonamenten en l'existència d'un nombre extraordinari d'agressions a aquestes hores, que no es produeixen amb la mateixa freqüència en altres dies del mes.

S'ha dut a terme un estudi exhaustiu de les dades que revelen una alta incidència d'agressions en dies que no són el primer del mes. En primer lloc, s'han analitzat totes les notícies corresponents als dies amb un nombre total d'agressions superior a 15, així com la notícia del dia anterior, amb l'objectiu de determinar si un esdeveniment concret afectava el nombre d'agressions del dia següent. Les notícies s'han consultat a l'arxiu històric del diari *Chicago Sun Times* [57], un dels diaris més rellevants de la ciutat de Chicago.

S'ha elaborat una llista de les 10 notícies més rellevants de cada dia, agrupant-les per temàtiques. En una primera instància, cap esdeveniment concret ha semblat ser la causa de les agressions, ja que en un període de 22 anys no s'ha observat un patró específic relacionat amb un ambient polític més convuls, esdeveniments nacionals o regionals. No obstant això, s'ha observat que els dies amb notícies més rellevants (aquelles que han rebut un major nombre de visites) coincideixen amb esdeveniments esportius, especialment relacionats amb els equips de beisbol de Chicago, com ara *The Chicago Cubs* i *The Chicago White Sox*, tot i que també s'inclouen en menor proporció l'equip de futbol americà *The Chicago Bears* i l'equip de bàsquet *The Chicago Bulls*. A continuació, s'ha realitzat una comparativa entre els resultats dels dies anteriors als partits d'aquests equips i els resultats dels dies amb un nombre elevat d'agressions, amb l'objectiu de determinar una possible correlació mitjançant una anàlisi exhaustiva de l'historial dels resultats d'aquests equips en els últims 22 anys.

A més, s'han elaborat diferents atributs temporals per a la predicció del model (com l'any, el mes, el dia de la setmana, el dia, etc.). Cal destacar els següents atributs: *New Years*, que indica amb un valor de 1 els dies que corresponen al primer dia de l'any i amb 0 la resta de dies; *firstday*, que identifica el primer dia de cada mes; i *isweekend*, que assigna el valor de 1 als dies dissabte i diumenge i 0 als altres dies. Altres atributs que han contribuït a l'optimització del model inclouen l'ús de retardades en les agressions entre 1 i 5 dies, així com atributs de mitjanes de

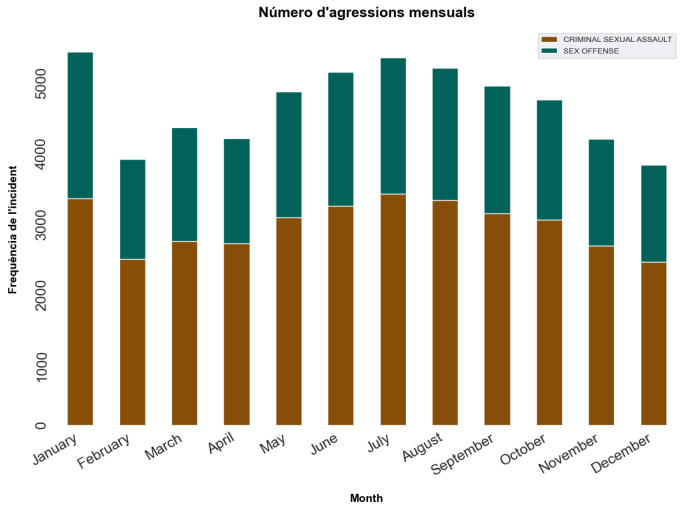

Figura 5: Total Agressions sexuals mensuals base de dades Chicago Crimes

les agressions setmanals, bisetmanals, mensuals i bimensuals. Finalment, s'ha adoptat un procés determinista per a assegurar que el model produeixi sempre la mateixa sortida en funció de les mateixes condicions inicials o l'estat inicial.

És evident que les agressions sexuals augmenten durant les estacions més càlides, particularment a l'estiu, destacant els mesos de juliol i agost, i disminueixen durant les estacions més fredes com es pot veure a la Fig. 5. A més, la majoria de les agressions tenen lloc durant els caps de setmana. Aquests resultats concorden amb l'estudi realitzat pel Ministeri de l'Interior sobre les agressions sexuals[58], on es constata una tendència similar, tenint en compte que les dades de Chicago acumulen, al gener, les denúncies en les quals no s'especifica el mes ni el dia en què van tenir lloc les agressions.

D'altra banda, mitjançant un estudi sobre les hores mitjanes de llum mensuals a la ciutat de Chicago durant tot l'any, s'ha constatat, com es pot veure a la Fig. 7 que la majoria de les agressions tenen lloc durant la franja nocturna, representant un 55% del total (cal destacar que s'han exclòs les agressions que es produeixen entre les 12 a.m. i les 12:01 a.m. per a normalitzar les dades de Chicago). Aquesta observació és concorde amb l'estudi realitzat a Catalunya sobre la prevenció d'agressions facilitades per drogues [59], on s'evidencia un increment d'agressions en horaris nocturns, especialment en entorns d'oci nocturn i durant els caps de setmana. A més, els autors d'aquest estudi coincideixen en destacar el consum d'alcohol com a factor de risc facilitador d'aquestes situacions.

A nivell de distribució geogràfica es veu a la Fig. 8 com el 43% de les agressions tenen lloc a la zona sud de Chicago (South Side), seguit per 29% a la zona oest, 22% a la zona nord i només un 6% a la zona est i que la majoria de les agressions tenen lloc a domicilis o apartaments (el 55.56%). A la Fig. 9 es pot veure la distribució de les agressions per comunitat a Chicago.

En l'estudi realitzat per investigar els patrons de pic d'agressions, utilitzant les notícies del diari Chicago Sun Times, no es van trobar correlacions clares. Tant els esdeveniments destacats dels dies analitzats com els resultats dels equips de futbol americà, bàsquet i beisbol no van demostrar ser factors causals. Així doncs, aquestes variables extrínseques no semblen ser conclusives en aquest context. Per a obtenir una comprensió més profunda de la situació de la ciutat, seria necessari disposar de dades més detallades que, lamentablement, no van ser accessibles durant el període d'investigació, tant per raons de temps com de recursos.

D'altra banda, altres variables es van formular basant-se en diferents estudis, com l'auditoria del sistema VioGén [60] i el mapa nacional de solucions per a posar fi a la violència contra les dones [61]. Aquests estudis han revelat que els col·lectius més afectats per l'agressió sexual i la violència de gènere són les dones immigrants, les desocupades, les persones grans, les amb discapacitat,

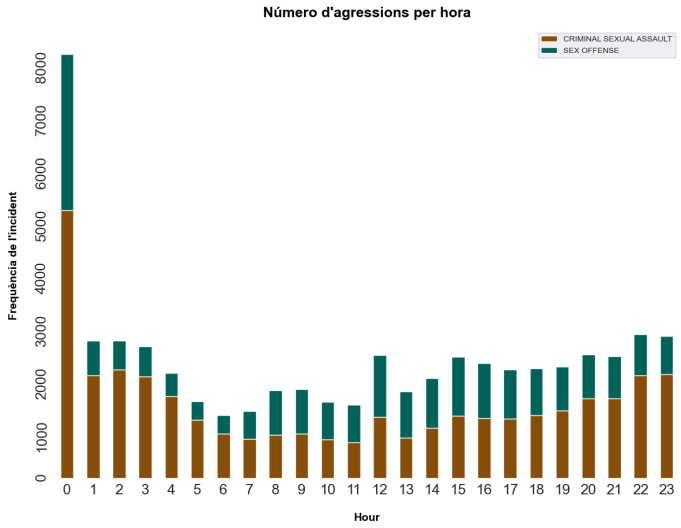

Figura 6: Número d'agressions per hora base de dades Chicago Crimes

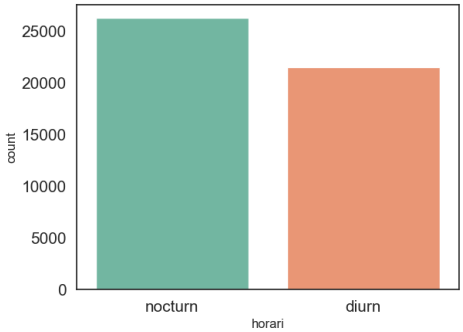

Figura 7: Total d'agressions en horari nocturn i diurn

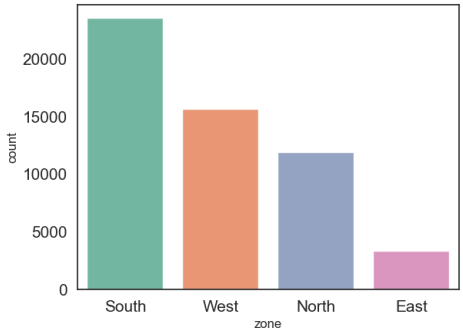

Figura 8: Total d'agressions per zona a Chicago

les amb persones dependents a càrrec i les amb un baix nivell de competències digitals. Aquests col·lectius es troben especialment vulnerables, ja que tenen un coneixement limitat sobre la igualtat de gènere i es troben en situacions de desprotecció. Per aquest motiu, s'ha accedit a les dades de Chicago Data [62] i s'ha extret informació sobre el percentatge de persones amb estudis secundaris i superiors, el nivell de pobresa i el nombre de persones nascudes a l'estranger, ja que aquestes dades han estat les més detallades disponibles.

5.2 Avaluació dels mètodes

L'objectiu principal d'aquest treball es poder predir el comportament humà per tal preveure les agressions amb la màxima precisió possible en una localitat determinada. Per aconseguir aquest objectiu, s'ha abordat l'anàlisi i la predicció de les agressions a Chicago de dues maneres. En primer lloc, s'ha realitzat una predicció de les dades mensuals per a tota la ciutat, així com per a cada comunitat, utilitzant adequadament les variables relacionades amb els col·lectius més vulnerables, adaptades en forma de percentatges. Amb l'objectiu d'obtenir dades encara més precises, s'ha elaborat també un anàlisi i una predicció de les dades diàries, tant per a la ciutat en conjunt com per a les comunitats específiques.

Per a l'anàlisi i la predicció de les dades per a tota la ciutat, s'han utilitzat variables relacionades amb els col·lectius anuals de tot l'àrea de Chicago, que es trobaven disponibles a la base de dades de *Rob Paral and Associates*. En primer lloc, s'ha realitzat una predicció d'aquestes variables,

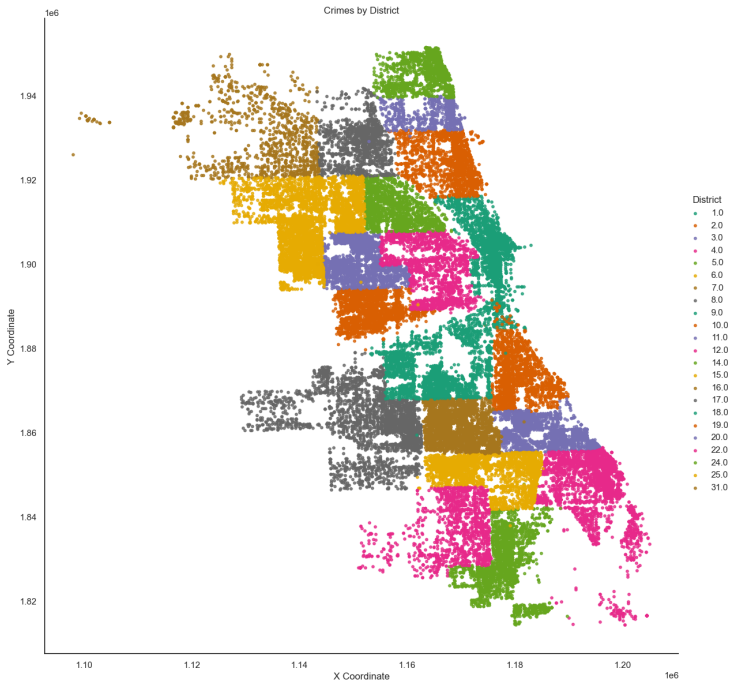

Figura 9: Total d'agressions per comunitat a Chicago

una vegada s'han modificat en el seu format, per a l'any 2022 (data en què es realitzaran la majoria de les prediccions), utilitzant el model ARIMA. Aquest model s'ha seleccionat ja que, segons la investigació realitzada, és el model amb millors resultats i més utilitzat per aquest tipus de prediccions en casos on les dades són limitades. Durant l'anàlisi, s'han avaluat diversos factors com, per exemple, la verificació de si una sèrie de dades és estacionària o l'ordre de diferenciació [63]. També s'han eliminat les variables *day, dayofweek, weekofyear, isweekend i firstday* per a l'anàlisi mensual i s'ha adaptat la variable *New Years* al primer dia del mes.

Cal destacar que, en afegir una funció de cost de pesos exponencial com es va fer per a la predicció de vendes, no es va observar cap millora significativa. Aquesta funció es va dissenyar per destacar els últims 14 dies, l'últim mes i el mateix període de temps de l'any anterior abans de la predicció. Malauradament, cap de les proves realitzades va mostrar una millora en els resultats.

Durant les previsions per a totes les regions durant els darrers 6 mesos, es va observar un fort sobreajust. La predicció en l'entrenament seguia perfectament la sèrie de dades, mentre que en la prova les agressions estaven significativament subestimades. Es van realitzar intents per reduir aquest sobreajust, com utilitzar PCA (anàlisi de components principals) i mantenir només 2 components, però no es va observar cap canvi en els resultats. Això suggereix que la correlació entre la sèrie i les variables exògenes ha canviat sobtadament. Una manera de provar-ho és canviar 6 mesos el problema, predint de gener a juny, de manera que el nostre període de prova pertany ara a la part de la sèrie temporal on pensem que la correlació es manté, com es veu a la Fig. 10.

Com s'ha vist, ara les prediccions per al període de prova ja no subestimen les agressions sinó que segueixen perfectament la tendència. Per donar suport a la hipòtesi que la correlació amb l'exogen s'ha mogut, podem intentar veure la correlació per als diferents conjunts a la Fig. 11.

Això ens proporciona dues conclusions importants:

- Hem de garantir que els conjunts d'entrenament i de prova estiguin lliures de qualsevol biaix. Una forma d'abordar això és entrenar un model que intenti classificar si una fila pertany al conjunt d'entrenament o al conjunt de prova.

- És crucial reciclar els models quan es despleguen en producció. Hem de tenir sempre present que la relació entre la variable objectiu i les variables exògenes pot canviar amb el temps.

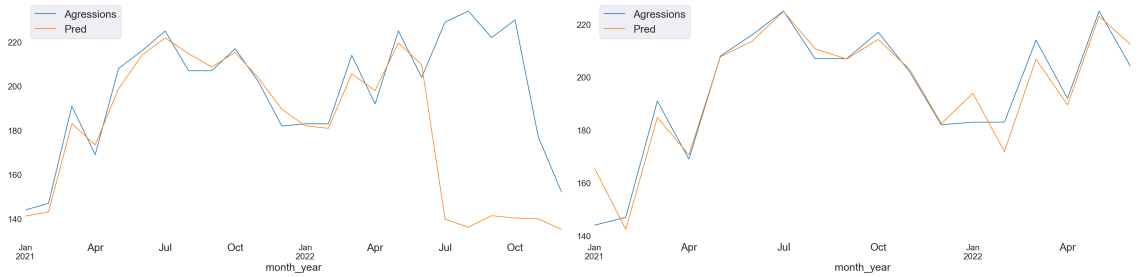

Figura 10: Predicció 6 últims i primers mesos 2022 respectivament

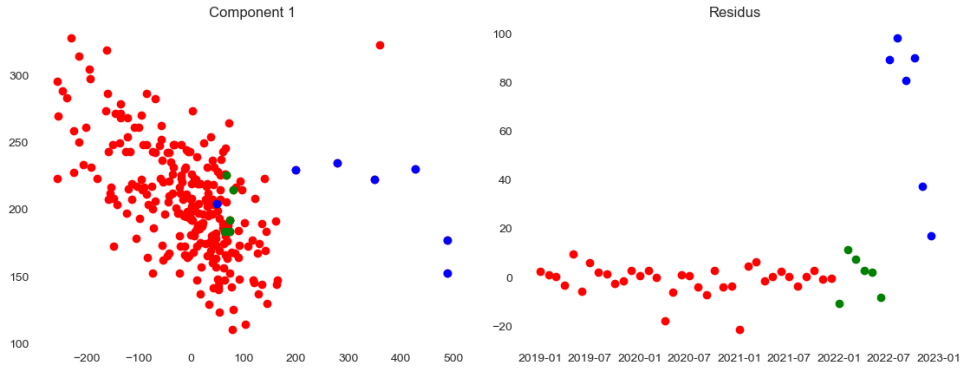

Figura 11: A la imatge de la dreta la primera component de PCA sobre el total d'agressions. En vermell les mostres del training, en verd el període de gener a juny de 2022 i en blau el període de juliol a desembre de 2022. A la imatge de l'esquerra els residus.

Cal destacar també que l'ús de lags (retards) i mitjanes de mesos com a variables sovint perjudicava la predicció, ja que el conjunt de proves contenia molts valors zero, el que provocava prediccions més baixes. Aquesta situació es va observar especialment en la predicció mensual per comunitats, i es va optar per eliminar definitivament aquestes variables.

Durant l'anàlisi de les agressions diàries per comunitat, també es van presentar diversos problemes. En primer lloc, el volum de dades era massa gran per assumir tot el conjunt de dades, per la qual cosa es va decidir seleccionar les dades a partir del 2017 en endavant (un procediment que es va replicar posteriorment per a altres prediccions).

La segona problemàtica que es va enfrontar és que la predicció tendia cap a zeros. Donat que s'estava analitzant les agressions diàries per a cada comunitat de Chicago, moltes d'aquestes comunitats no tenien agressions cada dia. Fet que fa que en molts casos els models tendeixin a predir zeros sempre.

S'identifiquen dues formes de modelar aquest problema de manera més eficient i precisa:

- **Renewal model**: Aquest model proposa un canvi de paradigma en l'abordatge de la variable a predir. En lloc de predir la sèrie temporal en si mateixa, s'estima el temps que transcorrerà entre dos esdeveniments successius, així com la magnitud del proper esdeveniment. Aquests models són particularment útils en sèries temporals on predominen els valors nuls (zeros), ja que permeten capturar aquests intervals i esdeveniments significatius. No obstant això, aquests models tenen restriccions en sèries temporals on la majoria dels valors són zeros.

- **Models Jeràrquics** (Hierarchical models): Aquests models no només aborden el problema dels valors nuls (zeros), sinó que també permeten modelar altres efectes i treballar amb diversos tipus de sèries temporals. Mitjançant aquest enfocament, es prenen en consideració les característiques jeràrquiques de les dades i s'integren múltiples nivells d'informació. Això permet capturar les relacions entre les diferents unitats o agrupacions de dades, i millora la capacitat de predicció en casos amb valors nuls i altres complexitats.

Amb aquests dos enfocaments, s'espera superar les limitacions inherents a la predicció de valors nuls en les sèries temporals d'agressions sexuals. Així, es pot millorar la precisió i l'eficàcia dels models, proporcionant resultats més fiables i significatius per a l'anàlisi i la prevenció d'aquest tipus d'agressions.

El treball previ en previsió jeràrquica segueix un enfocament en dues etapes: Primerament, s'obtenen les previsions base de manera independent per a cada sèrie temporal de la jerarquia, i després es combinen i revisen en un pas de postprocessament per garantir la coherència. Aquest procediment en dues etapes planteja dos problemes principals: (i) els paràmetres del model per a cada sèrie temporal s'aprenen de manera independent i (ii) les previsions base es revisen sense tenir en compte els paràmetres del model apresos. A més, la majoria dels mètodes existents només poden produir pronòstics puntuals en lloc de pronòstics probabilístics.

El model jeràrquic incorpora tant l'aprenentatge com la reconciliació en un únic model, on els paràmetres del model s'aprenen simultàniament per totes les sèries temporals de la jerarquia. Això garanteix que les previsions probabilístiques del model siguin coherents sense requerir cap pas de postprocessament. Les idees clau darrere del mètode proposat inclouen la diferenciabilitat de l'operació de mostreig i la implementació de la reconciliació de les mostres com un problema d'optimització convexa [64]. Això permet combinar components que típicament són independents, com la generació de previsions base, el mostreig i la reconciliació, en un únic model entrenable. No obstant això, en la pràctica, aquest model requereix un grau molt elevat de processament de càlcul i és computacionalment costós, el que dificulta la seva implementació en aquest treball. Per això, s'ha decidit continuar amb les altres tres agrupacions per a la predicció.

En el nostre cas particular a Chicago, l'ús del model jeràrquic implicaria entrenar un model per a cada comunitat específica, així com un model global. Aquest enfocament permetria abordar el fenomen conegut com a çanibalisme entre sèries", un terme utilitzat en el context de les vendes, que fa referència a la transferència de volum d'una sèrie temporal a una altra.

En el cas de les vendes, aquesta transferència de demanda és més clara i desitgem tenir un model que pugui comprendre i modelar aquesta dinàmica. No obstant això, en el context de les agressions sexuals, pot ser més complex establir una relació causal directa per a aquesta transferència de volum d'agressions. Tot i això, els models jeràrquics permeten abordar aquests efectes i modelar-los de manera adequada, tenint en compte les interaccions entre les diferents comunitats i els factors que poden influir en la propagació o concentració d'agressions.

Pel que fa a les prediccions, s'utilitzaran els models Weighted Model i Random Forest, ja que han mostrat els millors resultats per a les Vendes de la Botiga, juntament amb Prophet, XGBoost i DeepAR per a comparació. Les prediccions es realitzaran durant un període de 14 dies per al total d'agressions diàries a Chicago, amb avaluacions que abasten des del 16 de desembre fins al 31 de desembre de 2022 i del 24 de gener al 6 de febrer de 2023. Per a les prediccions mensuals del total de la ciutat i per comunitat, es considerarà un període de sis mesos, concretament els últims sis mesos de l'any 2022, ja que és l'any més recent amb dades completes. També es duran a terme avaluacions en altres períodes, com tot l'any 2021 i 2022, com a referències per a l'avaluació.

Per a avaluar els models s'han fet servir les mètriques esmentades a la metodologia, comparant-les amb les mateixes mètriques per al model mitjà. Com s'ha parlat anteriorment és complicat trobar una mètrica que s'ajusti a les necessitats exactes de la base de dades de manera global, per tant, es tenen en compte diverses mètriques per a veure l'evolució en conjunt. Tot i així, es tindrà en compte els valors residuals de cadascuna de les prediccions com a referència de la qualitat de la predicció.

5.3 Anàlisi i comparativa dels resultats

Com s'ha comentat anteriorment, la predicció de les agressions diàries per comunitat no ha estat satisfactòria. No obstant això, s'han obtingut resultats bastant satisfactoris per a les agressions diàries totals de la ciutat de Chicago i les agressions mensuals per comunitat i totals de la ciutat. En general, el model que ha mostrat millors resultats per a totes les prediccions ha estat el Weighted Model, seguit de Random Forest, XGBoost, Prophet i, finalment, DeepAR. En totes aquestes implementacions, s'han eliminat les variables de mitjanes mensuals i gairebé tots els lags mensuals. Com s'ha explicat anteriorment, aquestes variables semblen mostrar un fort sobreajustament i

subestimen les prediccions a partir de juliol. Això també es pot atribuir al fet que s'ha demostrat que durant l'estiu hi ha un pic d'agressions, de manera que tenir en compte aquests lags i agressions dels últims mesos no contribueix a aquest període.

Per al fitxer amb les vacances que se li passa com a argument a Prophet, s'ha creat un csv de nou a partir de les dades de les vacances anuals a Chicago[65].

A les següents gràfiques es pot veure el rendiment dels models als diferents períodes de predicció agafant les quatre mètriques destacades a la metodologia del treball. Cada columna representa un període on *Q1* és la predicció del 24 de gener de 2023 al 6 de febrer del 2023 (mitja de les 14 prediccions temporals), *Q2* és la predicció del 18 de desembre de 2022 al 31 de desembre del 2022 (mitja de les 14 prediccions), *Q3* és la predicció dels sis últims mesos de 2022 per al total de la ciutat (6 prediccions, una per mes), *Q4* és la predicció de tot l'any 2022 per al total de la ciutat (mitja de les 12 prediccions, una per mes), *Q5* és la predicció dels sis últims mesos de 2022 per comunitat (mitja de les prediccions de 6 mesos per les 77 comunitats), *Q6* és la predicció de tot l'any 2022 per comunitat (mitja de les prediccions de 12 mesos per les 77 comunitats).

Model	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**
DeepAr	1.88	1.57	12.70	44.2	2.13	1.98
Prophet	4.04	1.44	14.59	14.86	2.10	2.05
Random Forest	**1.58**	1.39	13.99	11.5	**1.36**	**1.26**
XGBoost	2.15	**1.26**	7.80	**9.81**	1.43	1.38
Weighted Model	1.85	1.37	**7.27**	9.96	1.37	**1.26**

Taula 2: Mètrica MAE per períodes

Model	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**
DeepAr	2.84	1.76	16.77	53.34	3.04	2.83
Prophet	4.8	**1.64**	17.71	17.99	3.05	2.95
Random Forest	**2.60**	1.78	14.87	15.96	1.89	1.75
XGBoost	3.24	1.77	8.81	**12.89**	1.92	1.82
Weighted Model	2.72	1.77	**8.21**	13.15	**1.87**	**1.71**

Taula 3: Mètrica RMSE per períodes

Model	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**
DeepAr	27.81	58.25	6.22	19.98	93.03	88.88
Prophet	103.05	54.19	7.38	7.44	90.4	89.44
Random Forest	**25.01**	50.36	6.72	5.94	82.79	80.10
XGBoost	39.95	**45.04**	3.82	**4.97**	**80.88**	78.55
Weighted Model	28.50	49.26	**3.64**	5.06	81.18	**77.8**

Taula 4: Mètrica SMAPE en percentatge per tots els períodes

La selecció dels períodes s'ha realitzat amb l'objectiu de realitzar prediccions més precises per a les dates disponibles. S'ha optat per predir els últims 14 dies del conjunt de dades, fins al 6 de febrer, i també els últims 14 dies de febrer de 2022, amb l'objectiu d'evitar el pic d'agressions del primer dia de l'any. Pel que fa a les prediccions mensuals, s'ha considerat pertinent utilitzar les dades més recents disponibles, és a dir, els últims 6 mesos de 2022. Aquest mateix raonament s'aplica a les dades anuals, utilitzant l'any complet més recent, que és el 2022.

Les taules han estat organitzades en ordre alfabètic, resaltant els millors resultats per període. Es va pensar una ordenació per mitjana, però s'hauria de ponderar per al número de prediccions

Model	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**
DeepAr	**50.00**	21.43	**50.00**	41.67	59.96	58.12
Prophet	0.00	45.86	33.33	33.33	**57.14**	57.25
Random Forest	57.14	**50.00**	66.67	58.33	58.22	**55.30**
XGBoost	14.29	57.14	33.33	**50.00**	66.23	68.72
Weighted Model	64.29	**50.00**	33.33	58.33	61.26	59.41

Taula 5: Mètrica Coverage en percentatge per tots els períodes

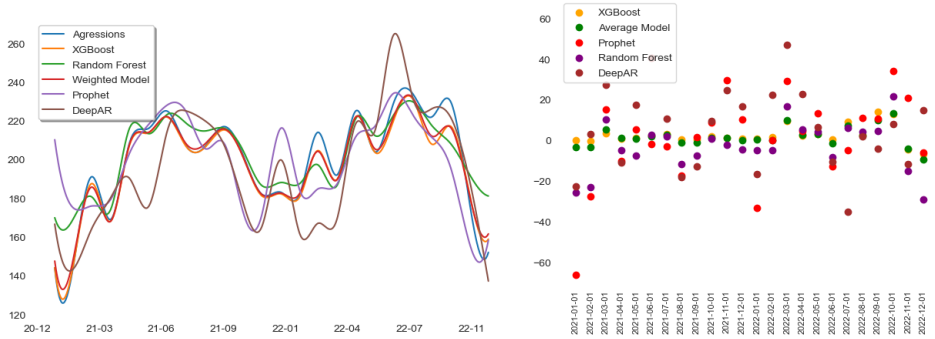

Figura 12: Predicció amb tots els models dels 6 últims mesos de 2022

que es realitzen per període. Pel que fa a la mètrica de Cobertura (Coverage), el millor resultat per període s'ha establert mitjançant el càlcul d'una distància lineal al valor del 50 (la meitat de les prediccions es troben per sobre del valor real) de l'error. En el futur, seria interessant considerar una diferència asimètrica, com la que s'utilitza amb la mètrica RMSLE en el consum de supermercats, on, en termes de prevenció de víctimes, és millor predir per sobre dels valors reals.

Com era d'esperar, en les prediccions on hi ha dies o comunitats sense cap agressió, el MAPE tendeix a infinit. Per aquest motiu, no s'ha tingut en compte en la comparativa entre mètriques. Amb l'SMAPE, ens trobem amb el problema de les penalitzacions per subestimació. Les prediccions dels models mostren una tendència a subestimar les agressions, la qual cosa es reflecteix en errors molt alts. La mètrica de Cobertura ens indica que sovint les prediccions es troben a la meitat o per sota del valor real. En general, es pot observar que a mesura que el valor de Cobertura és més baix (indicant més subestimació), el valor de l'SMAPE és més alt. També cal assenyalar que aquesta mètrica és controvertida, ja que no té una definició única i no està totalment acotada.

L'MSE és una mesura més precisa de l'error de predicció que el MAE, ja que té en compte la magnitud dels errors, mentre que el MAE és ineficient per a valors extrems. L'MSE penalitza les grans desviacions entre els valors reals i els previstos. És per això que es pot observar que l'error de l'MSE i el RMSE són més grans que el del MAE en totes les prediccions, sent el RMSE l'arrel quadrada de l'MSE.

A les Figs. 12 i 13 es realitza una representació gràfica de les prediccions de dos períodes, així com de les prediccions de tots els models utilitzats, acompanyades d'una representació dels residus d'aquestes prediccions. És observable que els pics assenyalats anteriorment al llarg de tot el treball coincideixen amb una major concentració de residus, com ara del 25 al 29 de desembre o bé els pics del juliol i l'octubre per a la predicció mensual. Aquests fenòmens requereixen un estudi en profunditat per determinar un conjunt de variables exògenes que els expliquin i, d'aquesta manera, reduir l'error i, per tant, els residus.

El rendiment dels models és fàcil de visualitzar tant en els gràfics com en les taules, on en general es veu un millor rendiment per al Weighted Model i XGBoost, seguit de Random Forest amb un rendiment lleugerament inferior però encara acceptable. Random Forest mostra robustesa davant de dades sorolloses i atípiques, a més de proporcionar importància a les característiques,

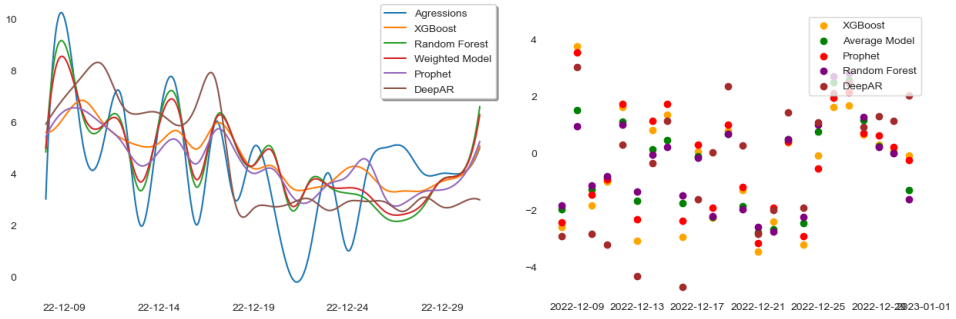

Figura 13: Predicció amb tots els models del 18 de desembre al 31 de desembre de 2022

la qual cosa ajuda a identificar les variables més influents en la predicció. Donat que tenim un conjunt de dades voluminós amb moltes característiques, és comprensible que el seu rendiment sigui bo. Random Forest té molts arbres amb fulles d'igual pes de manera que es pot obtenir una gran precisió fàcilment amb les dades disponibles. Això fa que afegir més característiques a les dades sigui una bona idea per veure el seu rendiment.No obstant això, és important tenir en compte que Random Forest té tendència al sobreajustament i pot no arribar a ser tan precís en les prediccions per aquesta raó.

Com a norma general, ha estat poc comú generar prediccions per sobre del valor real als períodes de prova. Una possible raó és el fet que hi ha pics sobtats en les dades que són més difícils de predir, i per tant, hi ha una tendència a predir valors més baixos.

En general, XGBoost sol tenir un rendiment destacat a l'hora de detectar patrons i relacions no lineals en les dades. Sembla ser el model que millor captura la tendència de les prediccions. De la mateixa manera, XGBoost té una gran capacitat per gestionar grans quantitats de dades i característiques complexes. S'explica, igual que en la predicció de vendes, que la combinació d'ambdós algoritmes ens doni els millors resultats amb el Weighted Model.

Prophet té un rendiment més baix en comparació amb els altres models. Aquest model, com ja es va observar en la predicció del consum de supermercats, no és tan adequat per a problemes amb característiques complexes o no lineals, i requereix un pre-processament addicional per gestionar-les. La precisió d'aquest model es pot veure afectada si les dades d'entrada no segueixen patrons estacionals o no són prou estables, com és el nostre cas, on tenim una gran sèrie de pics d'agressions en determinats dies pels quals necessitem més investigació i recórrer a variables exògenes per a explicar-los.

Per a DeepAR, tot i que no ha tingut un rendiment molt superior als altres algoritmes, ha aconseguit millors prediccions que bastants models, en especial per a les prediccions diaries. DeepAR és capaç de capturar patrons a curt termini a les dades. Amb dades diàries, és més probable que hi hagi patrons i variacions intradiàries més pronunciades, com ara estacionalitat diària o efectes de dies de la setmana. Aquests patrons a curt termini poden ser més desafiadors de capturar amb dades mensuals, on la variabilitat dins de cada mes pot ser menor. Un fet que distingeix a DeepAR de la resta d'algoritmes en el rendiment és que ha estat capaç de predir valors per sobre dels reals, fet que és interessant a tenir en compte quan es tracta amb dades d'agressions sexuals. No obstant, és un algoritme que requereix d'una gran sintonització dels hiperparàmetres i és computacionalment molt costós, en especial quan s'intenten predir grans volums de dades com ara a les comunitats.

El rendiment de DeepAR es veu empitjorat per el rendiment del període 4 (predicció mensual de tot l'any 2022). Si es treies aquest període, DeepAR es mostra entre els tres millors models per a totes les mètriques, destacant el seu rendiment en les prediccions diàries.

Com a comentari final, és important destacar l'ús de les variables exògenes en aquesta predicció. S'ha demostrat que hi ha pics d'agressions que no es poden explicar únicament amb les dades disponibles, sinó que cal recórrer a les variables exògenes. Encara que no s'han pogut predir aquests pics amb les variables de pobresa, educació i immigració, s'han obtingut resultats lleugerament millors en la predicció de la tendència, especialment en les prediccions per comunitat. A la Fig.

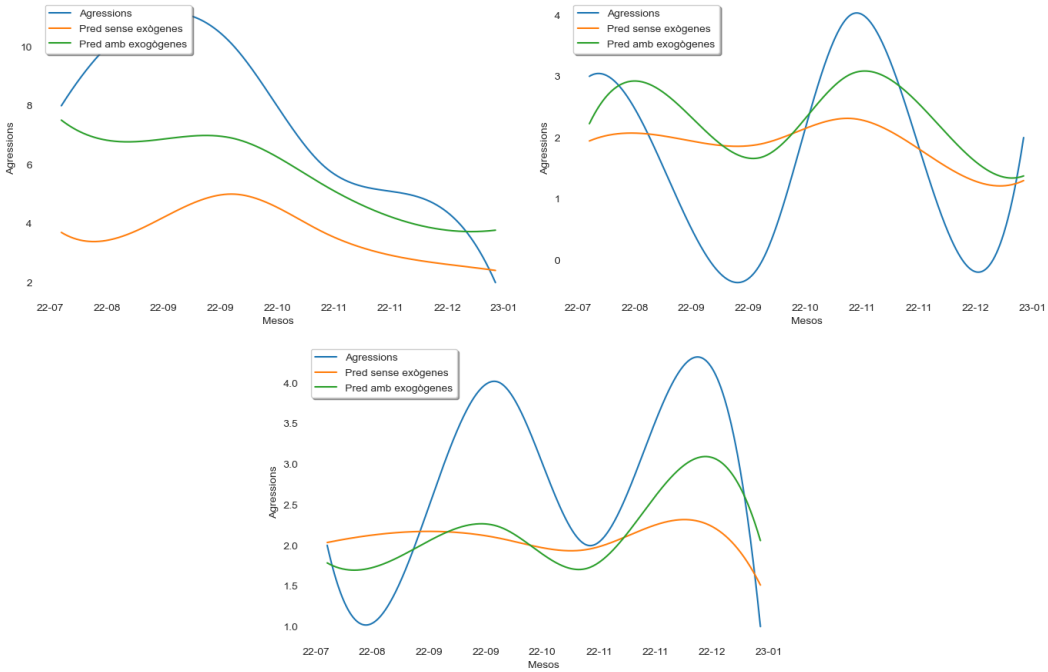

Figura 14: Predicció amb variable exògenes i sense varibales exògenes per als 6 últims mesos de 2022

14 es pot observar la diferència en aquestes prediccions per a tres comunitats diferents.

6 Discussió

Els resultats d'ambdós exemples ens demostren que la modelització del comportament humà és una tasca complexa i multifactorial. Aquest comportament està determinat per un gran nombre de variables tant endògenes com exògenes, les quals pot ser complicat trobar o adaptar amb les dades amb les quals es treballa.

És evident com les dades temporals tenen un impacte significatiu en la predicció, tant en les vendes als supermercats com en les agressions sexuals. D'una banda, s'ha observat que variables associades a períodes concrets, com la temporada escolar o la setmana de cobrament de salaris, provoquen un augment de les vendes. D'altra banda, la temporada d'estiu genera un increment considerable en les agressions sexuals.

Les festivitats també semblen exercir una influència notable en el comportament humà. Tant les festivitats pròpies de la regió on es troba la botiga com les festivitats nacionals, com el Nadal, Pasqua o Cap d'Any, provoquen un augment en el consum de productes, i en el cas de les agressions, es segueix un patró similar. Els dies festius, com els caps de setmana, i les festivitats nacionals com el Cap d'Any, es correlacionen amb un augment de les agressions sexuals.

A més, hi ha processos externs que sovint no es poden predir, com el terratrèmol del 2016, que va comportar un increment en la venda de productes de primera necessitat per ajudar les víctimes. També es troben dates amb pics en el nombre d'agressions que no tenen una correlació immediata amb les dades tractades.

Una tècnica que sembla contribuir a obtenir resultats més precisos en les prediccions és la segmentació de les dades. En el cas de les vendes, s'ha realitzat una predicció per a cada botiga o família de productes, mentre que en el cas de les agressions, s'ha fet una predicció per a cada comunitat de l'àrea de Chicago. Aquest enfocament permet comprendre millor la correlació entre les vendes/agressions i, en definitiva, el comportament humà. Es pot observar com evoluciona una família o comunitat concreta al llarg del temps i com això afecta les altres.

El recórrer a informació de variables exògenes és un mètode que resulta molt efectiu a l'hora de

realitzar les prediccions. Per a les vendes s'ha fet servir el preu diari del petroli donat que l'Equador és un país que depèn del petroli i la seva salut econòmica és molt vulnerable als xocs dels preus del petroli. Així mateix, com s'ha vist als estudis tant de VioGén com d'altres elaborats pel ministeri, en el moment en que es van afegir les variables exògenes que explicaven els percentatges de pobresa, estudis i immigració per comunitat i a l'àrea de Chicago, es va poder obtenir uns resultats més precisos del nombre d'agressions.

Pel que fa a les variables, l'enginyeria de característiques va jugar un paper crucial en ambdues prediccions. Amb l'objectiu de reduir l'overfitting en la majoria dels models, va ser necessari adaptar i eliminar diverses variables, la qual cosa va resultar en una millora de les prediccions. Cal destacar que no sempre més variables porten a resultats més precisos.

Per a concloure, en relació a les variables, s'ha demostrat que la generació de lags i mitjanes amb certs atributs resulta molt efectiva en l'aprenentatge de la majoria dels models. En el cas de les vendes, tant l'oli com el total d'articles de cada família que estaven en promoció tenen un fort impacte en les vendes, i generar mitjanes setmanals i mensuals ajuda a comprendre l'evolució de les vendes. A més, l'ús de lags fa que els valors de les vendes passades semblin contemporanis amb els valors que s'intenta predir, la qual cosa resulta útil per modelar la dependència en sèrie. El mateix s'aplica a la generació de lags i mitjanes per a les agressions sexuals.

Un altre factor destacable és la similitud en el comportament dels models en ambdós tipus de prediccions. Així com s'ha observat en les prediccions de vendes a Store Sales, on la mètrica ha penalitzat la divisió per famílies de productes, el mateix ha ocorregut amb la divisió per comunitats en les agressions. A causa de la competició a Kaggle, on l'objectiu és aconseguir el menor error segons una mètrica específica, s'ha treballat en funció d'aquesta mètrica més que en la predicció que s'ajusti millor a l'evolució de la vida real. Amb les agressions, també s'ha enfrontat el mateix problema de triar una mètrica de referència per avaluar el rendiment dels models.

No obstant això, sembla que en ambdós casos, l'ordre de rendiment dels models és bastant similar per al conjunt de mètriques. El millor model és el Weighted Model, amb els percentatges corresponents per a Random Forest i XGBoost segons el seu rendiment en les prediccions. A continuació, es troben Random Forest i XGBoost amb errors molt similars, i després, en ordre, Prophet i DeepAR en ambdues prediccions. DeepAR és un model que caldria explorar més en el moment en que es disposessin de dades més detallades on DeepAR pot fer ús de la seva capacitat d'anàlisi de patrons a curt termini.

Les variables creades amb termes deterministes de la sèrie de Fourier basats en el temps del calendari són considerades entre els components més importants per a ambdós tipus de prediccions. Els termes de Fourier són periòdics, i això els converteix en eines útils per descriure un patró periòdic. Una avantatge d'aquests termes és que descriuen l'estacionalitat com una funció relativament suau, i la capacitat de modificar el nombre de termes permet ajustar el grau de suavitat de la representació.

Finalment, cal ressaltar la importància d'ajustar els hiperparàmetres dels models. Després d'executar aquest procés, tots els models van experimentar millores significatives. En aquest sentit, és rellevant destacar la influència de certs paràmetres i les seves similituds i diferències en les prediccions.

7 Conclusions i Treball Futur

Es pot concloure que, en gran mesura, els models de Deep Learning disponibles en l'actualitat permeten predir el comportament humà de manera efectiva. Aquest treball ha posat de manifest la gran importància de les variables en les prediccions. Les variables endògenes de temps generades per a la predicció del comportament humà, tant en les vendes com en les agressions, han aportat millores significatives en la qualitat de les prediccions.

D'altra banda, les variables exògenes ens ajuden a comprendre altres factors externs a les nostres dades que justifiquen canvis bruscos en el comportament humà. La inclusió de dades com el preu diari del petroli o els terratrèmols en les vendes, o l'índex de pobresa, educació i immigració en les dades d'agressions, ha mostrat una millora, especialment en la identificació de pics en dades específiques o comunitats concretes.

Per tant, tant les variables endògenes com les exògenes afecten la modelització de les dades. L'enfocament de modelitzar les dades tal com es presenten sovint no produeix prediccions precises i pot resultar en resultats molt distants de la realitat. Cal recórrer a aquestes dades externes per obtenir una visió més completa del problema.

Aquest és el camí a seguir en treballs futurs. Per a predir dades més localitzades, s'hauria de centrar en l'objectiu principal de predir les agressions a Barcelona. Quan sigui possible obtenir dades més detallades dels Mossos d'Esquadra, més enllà de les dades trimestrals disponibles públicament, caldrà realitzar un estudi de variables exògenes. Com ha demostrat l'algoritme VioGén, aquestes variables exògenes com la nacionalitat, la pobresa, l'edat, la comunitat i la situació familiar de la víctima expliquen el nivell de perill al qual estan exposades. L'ús de variables exògenes que fa servir VioGén, tot i la controvèrsia generada per destacar grups específics on les agressions són més freqüents exclusivament per factors socioeconòmics, és un enfocament correcte i que realment ajuda en la detecció i prevenció de futures agressions sexuals.

Com s'ha observat durant la competició de Kaggle, la llibreria DARTS amb LightGBM s'ha demostrat com el millor model, i seria interessant considerar la seva aplicació per a la predicció d'agressions. DARTS ofereix una àmplia gamma de models, des de clàssics com ARIMA fins a xarxes neuronals profundes. Encara que en el moment de l'elaboració dels models aquesta llibreria no era coneguda i no es va tenir en compte, molts dels millors resultats s'han obtingut amb la seva aplicació.

La idea principal darrere de DARTS és utilitzar la tècnica de regularització anomenada *dropout*, habitual en xarxes neuronals profundes, i incorporar-la a l'algoritme d'augment del gradient. En DARTS, el dropout implica ignorar una part dels arbres durant el càlcul dels pseudo-residuals en cada iteració. Aquesta tècnica fa que l'efecte relatiu dels arbres inicials disminueixi dràsticament i els arbres posteriors, fins i tot després de centenars d'iteracions, continuïn contribuint a les prediccions, ajudant a combatre l'overfitting [66]. La capacitat de generar nous aprenents que no es especialitzen excessivament en mostres específiques és la raó per la qual DARTS és un model tan generalitzat i potent.

Per altra banda, LightGBM és un marc de *gradient boosting* que utilitza algorismes d'aprenentatge basats en arbres de decisió. Aquesta llibreria està dissenyada per a la distribució i destaca per la seva velocitat d'entrenament més ràpida, eficiència i menor requeriment de memòria. LightGBM és compatible amb l'aprenentatge paral·lel, distribuït i GPU, i és capaç de gestionar dades a gran escala [67]. Aquesta eina utilitza un mètode basat en histograma, on les dades s'agrupen en bins utilitzant un histograma de la seva distribució. En lloc de treballar amb punts de dades individuals, LightGBM itera, calcula les millores i divideix les dades basant-se en els bins. Aquest mètode també pot ser optimitzat per a conjunts de dades dispersos. LightGBM ofereix la característica d'empaquetament exclusiu d'entitats, on l'algorisme combina entitats exclusives per reduir la dimensionalitat, augmentant així la seva velocitat i eficiència [68].

La combinació de la llibreria DARTS amb LightGBM sembla com una idea prometedora. LightGBM destaca per la seva alta velocitat d'entrenament, un avantatge significatiu quan es treballa amb conjunts de dades massius. A més, si a això se li afegeix l'ús de la tècnica de regularització dropout proporcionada per la llibreria DARTS, la qual ajuda a combatre l'overfitting, es pot millorar els resultats de manera significativa. Cal remarcar que els millors resultats obtinguts fins ara han estat amb models d'augment de gradient.

Un altre model que ofereix prediccions acurades és el model ARIMA, el qual ha guanyat popularitat en els últims anys. Aquest model destaca en la predicció de dades no estacionàries i permet una gran flexibilitat en la descomposició de sèries temporals en diferents components. ARIMA utilitza mitjanes mòbils retardades per suavitzar les dades de les sèries temporals. Cal tenir en compte que els models autoregressius assumeixen implícitament que el futur seguirà el mateix patró que el passat, però aquesta suposició pot resultar inexacta segons les dades amb les quals es treballi. Una evolució d'ARIMA és SARIMAX, que permet tenir en compte l'estacionalitat i les variables exògenes, factors importants per a la predicció del comportament humà.

No obstant, no s'ha de deixar de banda els models que hem fet servir per a les prediccions actuals. El Weighted Model mostra un molt bon rendiment i estaria entre els models ha repetir per a futures prediccions fent servir les ponderacions per a models diferents. També el model

DeepAR, com s'ha comentat a la discussió, és un model que té molta possibilitat de millora amb recursos òptims a nivell de dades i computació.

A futur, l'objectiu principal consistiria en analitzar dades diàries i identificar les variables exògenes que expliquen els factors socioeconòmics determinants. Aquesta tasca implicaria realitzar una divisió de Barcelona en barris, similar a com s'ha dut a terme a Chicago amb les seves comunitats, i utilitzar els nous models esmentats, juntament amb aquells amb els quals s'han obtingut els millors resultats, per a predir les agressions.

Mitjançant aquest enfocament, es podria obtenir una visió més detallada i localitzada de la dinàmica de les agressions a Barcelona. L'anàlisi de dades diàries permetria capturar les fluctuacions i els canvis a curt termini en el comportament humà. A més, la inclusió de les dades exògenes relacionades amb factors socioeconòmics clau permetria entendre millor els determinants subjacents de les agressions.

Aquesta aproximació més granular i amb models més avançats té com a finalitat millorar la precisió i la capacitat de pronòstic de les prediccions d'agressions. S'estaria treballant amb dades més actualitzades i es tindria en compte una gamma més àmplia de factors que influeixen en el comportament humà, enfortint així la comprensió i la prevenció d'aquests incidents en la ciutat.

Agraïments

M'agradaria agrair aquesta feina als meravellosos professors que he tingut al llarg de la meva vida que han fet que m'enamori de les matemàtiques i han fet de mi una persona curiosa. Gràcies a tots els docents que m'han format com a persona, mai no es valorarà suficient la vostra feina.

Agrair en especial la feina dels meus tutors durant aquest treball, en Jordi González i en Marc Borras, que han estat uns guies fantàstics ajudant-me en tot moment i ensenyant-me coses noves cada dia.

També m'agradaria agrair als meus pares per ser la millor educació que mai hom pugui rebre. Per ser pacients, per escoltar-me sempre i dedicar-me tot el seu temps, però sobretot per empènyer-me sempre a fer el que jo vulgui, desenvolupar el meu pensament crític i fer de mi una persona millor cada dia.

Finalment, agrair a María Pallejà i Clara Sorolla per ensenyar-me com són de meravelloses les dones, fer-me més forta en un entorn tan crític, acompanyar-me i aprendre'n d'elles cada dia.

Referències

[1] El Sistema viogén supera Los Seis millones de valoraciones de riesgo a víctimas de violencia de género. La Moncloa. 19/05/2023. El Sistema VioGén supera los seis millones de valoraciones de riesgo a víctimas de violencia de género [Prensa/Actualidad/Interior]. (Sense data). https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/interior/Paginas/2023/190523-sistema-viogen-victimas-violencia-genero.aspx#:~:text=Es%20un%20sistema%20policial%20centralizado,cualquier%20parte%20del%20territorio%20nacional

[2] Kirby, S., Francis, B., & O'Flaherty, R. (2013). Can the FIFA World Cup football (soccer) tournament be associated with an increase in domestic abuse? Journal of Research in Crime and Delinquency, 51(3), 259-276. doi:10.1177/0022427813494843

[3] Quigg, Z., Hughes, K., & Bellis, M. A. (2012). Effects of the 2010 World Cup Football Tournament on emergency department assault attendances in England. The European Journal of Public Health, 23(3), 383–385. https://doi.org/10.1093/eurpub/cks098

[4] Deakin, C. D., Thompson, F., Gibson, C., & Green, M. (2007). Effects of international football matches on ambulance call profiles and volumes during the 2006 World Cup. Emergency Medicine Journal, 24(6), 405–407. https://doi.org/10.1136/emj.2007.046920

[5] Lindo, J., Siminski, P., & Swensen, I. (2015). College Party culture and sexual assault. NATIONAL BUREAU OF ECONOMIC RESEARCH. https://doi.org/10.3386/w21828

[6] Store sales - time series forecasting (sense data) Kaggle. https://www.kaggle.com/competitions/store-sales-time-series-forecasting

[7] Chicago, C.of (2018) Chicago crime, Kaggle. https://www.kaggle.com/datasets/chicago/chicago-crime.

[8] Ministerio de Igualdad (sense data) Portal estadístico, Portal Estadístico Violencia de Género. http://estadisticasviolenciagenero.igualdad.mpr.gob.es/

[9] Kaggle - Introduction. https://www.kaggle.com.

[10] Macias, D. A. (2022). A brief history on time series analysis; forecasting. Medium. Disponoble a: https://medium.com/@deniseamacias1/a-brief-history-on-time-series-analysis-forecasting-f5a22bbd0641

[11] Wong, J. (2023). https://medium.com/@ycwong.joe/a-brief-history-of-time-series-models-38455c2cd78d

[12] Hyndman, R.J., Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3

[13] Tum, P. (2021). Overview of Time Series Forecasting from Statistical to Recent ML Approaches. https://medium.com/codex/introduction-to-prophet-algorithm-a59e463a6c72

[14] Berk, R. A. (Sense data). TIME SERIES ANALYSIS. Disponible a https://www.encyclopedia.com/social-sciences/encyclopedias-almanacs-transcripts-and-maps/time-series-analysis

[15] Hager, R. (2022). https://www.su.se/english/research/research-subjects/statistics/statistical-models-in-the-social-sciences/a-brief-history-of-time-series-analysis-1.612367

[16] Liu, G., Zhong, K., Li, H., Chen, T.,; Wang, Y. (2022). A state of Art Review on time series forecasting with machine learning for environmental parameters in agricultural greenhouses. Information Processing in Agriculture. https://doi.org/10.1016/j.inpa.2022.10.005

[17] Rodrigo, J. A. (2017). Análisis de Componentes Principales. https://www.cienciadedatos.net/documentos/35_principal_component_analysis

[18] Nadeem. (2021). Introduction to prophet algorithm. https://medium.com/codex/introduction-to-prophet-algorithm-a59e463a6c72

[19] (Sense data). Algoritmo de previsión DeepAR. https://docs.aws.amazon.com/es_es/sagemaker/latest/dg/deepar.html

[20] Khurana, P. (2019). How LSTMs solve the problem of Vanishing Gradients? https://prvnk10.medium.com/how-lstms-solve-the-problem-of-vanishing-gradients-ea88f08c78ca

[21] Salinas, D.; Flunkert, V. (2019). https://www.sciencedirect.com/science/article/pii/S0169207019301888

[22] swarnimrai. (2021). Multi-Layer Perceptron Learning in Tensorflow. Disponible a https://www.geeksforgeeks.org/multi-layer-perceptron-learning-in-tensorflow/ (Data d'accès: 19 maig, 2023)

[23] What is a Bayesian neural network? (Sense data). https://www.databricks.com/glossary/bayesian-neural-network

[24] Oreshkin, B. N., Carpov, D., Chapados, N.,; Bengio, Y. (2020). https://arxiv.org/abs/1905.10437

[25] Triebe, O., Hewamalage, H., Pilyugina, P., Laptev, N., Bergmeir, C., amp; Rajagopal, R. (2021). Neuralprophet: Explainable forecasting at scale. arXiv.org. https://arxiv.org/abs/2111.15397

[26] What is XGBoost? (sense data) NVIDIA Data Science Glossary. https://www.nvidia.com/en-us/glossary/data-science/xgboost/

[27] Hudgeon, D.,; Nichol, R. (2020). Machine Learning for Business: Using amazon sagemaker and Jupyter. https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

[28] Introduction to Boosted Trees (Sense data). https://xgboost.readthedocs.io/en/stable/tutorials/model.html

[29] Tianqi Chen; Carlos Guestrin. (2016, January 10). XGBoost: A scalable tree boosting system arxiv. https://arxiv.org/pdf/1603.02754.pdf

[30] Degradation state recognition of piston pump based on ICEEMDAN and XGBoost. (2020). https://www.researchgate.net/publication/345327934_Degradation_state_recognition_of_piston_pump_based_on_ICEEMDAN_and_XGBoost

[31] Bootstrap aggregation, random forests and boosted trees. (Sense data). https://www.quantstart.com/articles/bootstrap-aggregation-random-forests-and-boosted-trees/

[32] Decision Tree - Regression. (sense data). https://www.saedsayad.com/decisiontreereg.htm

[33] Nadeem. (2021). Introduction to prophet algorithm. Medium. Disponible a https://medium.com/codex/introduction-to-prophet-algorithm-a59e463a6c72

[34] Nadeem, N. (2021) Introduction to prophet algorithm, Medium. CodeX. https://medium.com/codex/introduction-to-prophet-algorithm-a59e463a6c72

[35] Forecasting at scale. (sense data) Prophet. https://facebook.github.io/prophet/

[36] Goyal, D. (2020, Abril 06). How does prophet work? part-2. https://medium.com/analytics-vidhya/how-does-prophet-work-part-2-c47a6ceac511

[37] Time series analysis using Facebook Prophet. (2023, Març 10). from https://www.geeksforgeeks.org/time-series-analysis-using-facebook-prophet/

[38] Yang, S. (2020, Decembre 26). Time series analysis using Prophet in pythonpart 1: Math explained. https://medium.com/analytics-vidhya/time-series-analysis-using-prophet-in-python-part-1-math-explained-5936509c175c

[39] Maklin, C. (2022, Juliol 15). Deepar forecasting algorithm. Medium. https://medium.com/@corymaklin/deepar-forecasting-algorithm-6555efa63444

[40] Hudgeon, D. and Nichol, R. (2020) How DeepAR Works, Amazon. Manning Publications Co. https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html

[41] Hudgeon, D.; Nichol, R. (2020a). DeepAR Forecasting Algorithm. Amazon. https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html

[42] YouTube. (2020). Probabilistic Forecasting with DeepAR and AWS SageMaker. YouTube. https://www.youtube.com/watch?v=zQvHESgqFcU.

[43] Hyndman, R. J. (2006). [PDF] another look at forecast accuracy metrics for intermittent demand: Semantic scholar. Foresight: The International Journal of Applied Forecasting. https://www.semanticscholar.org/paper/Another-Look-at-Forecast-Accuracy-Metrics-for-Hyndman/6fdc9d43ad105d3e68a319430cc3e7d60264d7df

[44] Rink, K. (2021). Time Series Forecast Error Metrics you should know. Medium. https://towardsdatascience.com/time-series-forecast-error-metrics-you-should-know-cc88b8c67f27

[45] Willmott, C. J.; Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (RMSE) in assessing average model performance. Climate Research. https://www.int-res.com/abstracts/cr/v30/n1/p79-82/

[46] Chai, T. and Draxler, R. R.: (2014) Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, Geosci. Model Dev., 7, 1247–1250, https://doi.org/10.5194/gmd-7-1247-2014

[47] Kim, S.; Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. International Journal of Forecasting, 32(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

[48] Goodwin, P.; Lawton, R. (1999). On the asymmetry of the symmetric Mape. International Journal of Forecasting, 15(4), 405–408. Dispnible a: https://doi.org/10.1016/s0169-2070(99)00007-2

[49] Makridakis, S.; Hibon, M. (2000). The M3-competition: Results, conclusions and implications. International Journal of Forecasting, 16(4), 451–476. https://doi.org/10.1016/s0169-2070(00)00057-1

[50] Lendave, V. (2021). A guide to different evaluation metrics for time series forecasting models. Analytics India Magazine. https://analyticsindiamag.com/a-guide-to-different-evaluation-metrics-for-time-series-forecasting-models/

[51] https://www.kaggle.com/code/dmitryshekhov/darts-ensemble-with-lightgbm-stacking

[52] CodeX. (2021). Introduction to prophet algorithm. Medium. https://medium.com/codex/introduction-to-prophet-algorithm-a59e463a6c72

[53] RANDOMIZEDSEARCHCV. scikit. (sense data). Disponible a https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

[54] (2013). Tema 6: Modelización con datos de series temporales. https://www.ucm.es/data/cont/docs/518-2013-10-25-Tema_6_EctrGrado.pdf

[55] Chicago, C. of. (2018, 17 abril). Chicago crime. Kaggle. https://www.kaggle.com/datasets/chicago/chicago-crime

[56] https://www.interior.gob.es/opencms/pdf/archivos-y-documentacion/documentacion-y-publicaciones/publicaciones-descargables/seguridad-ciudadana/Agresores_sexuales_con_victima_desconocida_126180061_web.pdf

[57] Chicago Sun-Times Archives. (Sense Dara). https://chicagosuntimes.newsbank.com/

[58] Gimenez Salinas, A.; Pérez Ramirez, M. (2018, Diciembre). Agresores sexuales con victima desconocida implicaciones para la ... Research Gate., https://www.researchgate.net/publication/329844005_AGRESORES_SEXUALES_CON_VICTIMA_DESCONOCIDA_Implicaciones_para_la_investigacion_criminal

[59] Panyella-Carbó, M. N., Martin-Fumadó, C.,; Gómez-Durán, E. L. (2021, 1 gener). Prevention of drug-facilitated sexual assault. Spanish Journal of Legal Medicine. https://www.elsevier.es/en-revista-spanish-journal-legal-medicine-446-articulo-prevention-drug-facilitated-sexual-assault-S2445424921000078

[60] ETICAS. (3 març 2022). (rep.). Auditoria Externa del Sistema VioGen. https://eticasfoundation.org/wp-content/uploads/2022/04/ETICAS-_-Auditoria-Externa-del-sistema-VioGen-_-20220308.docx.pdf.

[61] 00Mapa StopVioGen. (2022). (rep.). MAPA NACIONAL DE SOLUCIONES PARA EL FIN DE LAS VIOLENCIAS CONTRA LAS MUJERES. https://www.mapastopviogen.es/

[62] Chicago data. Rob Paral Associates. (Sense Data). https://robparal.com/chicago-data/

[63] CodeX. (2021). Introduction to prophet algorithm. Medium. https://medium.com/codex/introduction-to-prophet-algorithm-a59e463a6c72

[64] Rangapuram, S. S., Werner, L. D., Benidis, K., Mercado, P., Gasthaus, J.,; Januschowski, T. (2021). End-to-end learning of coherent probabilistic forecasts for hierarchical time series. PMLR. http://proceedings.mlr.press/v139/rangapuram21a.html

[65] City holidays (offices closed). City of Chicago:: City Holidays (Offices Closed). (Sense data). https://www.chicago.gov/city/en/narr/misc/city-holidays.html

[66] Meir, U. (2022). D.A.R.T-your new weapon against overfitting in boosting models. Medium. https://medium.com/@meir412_37692/d-a-r-t-your-new-weapon-against-overfitting-in-boosting-models-9ea4e6aa435b

[67] Welcome to LightGBM's documentation! LightGBM 3.3.2 documentation. (Sense data). https://lightgbm.readthedocs.io/en/v3.3.2/

[68] Data Science Team. (Sense data). Cómo funciona el algoritmo lightgbm. Cómo funciona el algoritmo LightGBM-ArcGIS Pro — Documentación. https://pro.arcgis.com/es/pro-app/latest/tool-reference/geoai/how-lightgbm-works.htm

A Hiperparàmetres

Model	**Hiperparametres**
Random Forest	n_estimators=1200, max_depth=50, max_features='auto', bootstrap=True, min_samples_leaf=2, min_samples_split=2, random_state=0
XGBoost	n_estimators=500, learning_rate=0.01, max_depth=3, subsample=0.5, colsample_bytree=0.4, colsample_bylevel=1, random_state=0
Weighted Model	RF: n_estimators=1200, max_depth = 50, max_features = 'auto', bootstrap = True, min_samples_leaf=2, min_samples_split=2, random_state=0, XG: n_estimators=500, learning_rate = 0.01, max_depth= 3, subsample = 0.5, colsample_bytree = 0.6, colsample_bylevel = 1, random_state=0
Stacking	Mateixos paràmetres que a Random Forest i a XGBoost
Prophet	holidays = event_holiday, changepoint_prior_scale = 0.05, holidays_prior_scale = 0.01, seasonality_prior_scale = 0.01, seasonality_mode = 'additive', yearly_seasonality = False, weekly_seasonality = True, daily_seasonality = False
DeepAr	freq = 'd',context_length = 30, prediction_length = 16,num_layers = 2, num_cells = 40,cell_type = l·lstm",epochs = 50,cardinality = cardinality.tolist()

Taula 6: Resultats competicio *Store Sales*

Model	**Hiperparametres**
Random Forest	n_estimators=1200, max_depth=100, min_samples_split=2, random_state=0
Weighted Model	Mateixos valors que a RF i XGBoost
XGBoost	n_estimators=500, learning_rate=0.01, max_depth=3, subsample=0.5, colsample_bytree=0.4, colsample_bylevel=1, random_state=0
Prophet	holidays = hol, changepoint_prior_scale = 0.05, holidays_prior_scale = 0.01, seasonality_prior_scale = 0.01, seasonality_mode = 'multiplicative', yearly_seasonality = False, weekly_seasonality = True, daily_seasonality = False
DeepAr	freq = 'd',context_length = 45, prediction_length = 14, num_layers = 2, num_cells = 40, cell_type = "gru", epochs = 50, cardinality = cardinality.tolist()

Taula 7: Q1

Model	**Hiperparametres**
Random Forest	n_estimators=1200, max_depth=100, min_samples_leaf=4, random_state=0
Weighted Model	Mateixos valors que a RF i XGBoost
XGBoost	n_estimators=500, learning_rate=0.01, max_depth=3, subsample=0.5, colsample_bytree=0.4, colsample_bylevel=1, random_state=0
Prophet	holidays = hol, changepoint_prior_scale = 0.05, holidays_prior_scale = 5, seasonality_prior_scale = 5, seasonality_mode = 'multiplicative', yearly_seasonality = True, weekly_seasonality = True, daily_seasonality = False
DeepAr	freq = 'd',context_length = 100, prediction_length = 10, num_layers = 2, num_cells = 40, cell_type = "gru", epochs = 50, cardinality = cardinality.tolist()

Taula 8: Q2

Model	**Hiperparametres**
Random Forest	n_estimators=1200, max_depth=100, min_samples_split=4, random_state=0
Weighted Model	Mateixos valors que a RF i XGBoost
XGBoost	n_estimators=500, learning_rate=0.01, max_depth=20, subsample=0.8999999999999999, colsample_bytree=0.7, colsample_bylevel=0.8999999999999999, random_state=0
Prophet	holidays = hol, changepoint_prior_scale = 0.05, holidays_prior_scale = 0.01, seasonality_prior_scale = 0.01, seasonality_mode = 'additive', yearly_seasonality = False, weekly_seasonality = True, daily_seasonality = False
DeepAr	freq = 'd',context_length = 100, prediction_length = 18, num_layers = 2, num_cells = 40, cell_type = "gru", epochs = 50, cardinality = cardinality.tolist()

Taula 9: Q3

Model	**Hiperparametres**
Random Forest	n_estimators=1200, max_depth=100, min_samples_split=10, random_state=0
Weighted Model	Mateixos valors que a RF i XGBoost
XGBoost	n_estimators=500, learning_rate=0.01, max_depth=15, subsample=0.6, colsample_bytree=0.6, colsample_bylevel=1, random_state=0
Prophet	holidays = hol, changepoint_prior_scale = 0.08, holidays_prior_scale = 0.01, seasonality_prior_scale = 0.01, seasonality_mode = 'multiplicative', yearly_seasonality = False, weekly_seasonality = True, daily_seasonality = False
DeepAr	freq = 'd',context_length = 100, prediction_length = 12, num_layers = 2, num_cells = 40, cell_type = "gru", epochs = 50, cardinality = cardinality.tolist()

Taula 10: Q4

Model	**Hiperparametres**
Random Forest	n_estimators=1200, max_depth=50, min_samples_leaf=4, random_state=0
Weighted Model	Mateixos valors que a RF i XGBoost
XGBoost	n_estimators=500, learning_rate=0.1, max_depth=3, subsample=0.8999999999999999, colsample_bytree=0.5, colsample_bylevel=1, random_state=0
Prophet	holidays = hol, changepoint_prior_scale = 0.05, holidays_prior_scale = 0.01, seasonality_prior_scale = 0.01, seasonality_mode = 'additive', yearly_seasonality = False, weekly_seasonality = True, daily_seasonality = False
DeepAr	freq = 'd', context_length = 100, prediction_length = 462, num_layers = 2, num_cells = 40, cell_type = l·lstm", epochs = 50, cardinality = cardinality.tolist()

Taula 11: Q5

Model	**Hiperparametres**
Random Forest	n_estimators=1200, max_depth=50, min_samples_leaf=4, random_state=0
Weighted Model	Mateixos valors que a RF i XGBoost
XGBoost	n_estimators=500, learning_rate=0.1, max_depth=3, subsample=0.8999999999999999, colsample_bytree=0.5, colsample_bylevel=1, random_state=0
Prophet	holidays = hol, changepoint_prior_scale = 0.05, holidays_prior_scale = 0.01, seasonality_prior_scale = 0.01, seasonality_mode = 'additive', yearly_seasonality = False, weekly_seasonality = True, daily_seasonality = False
DeepAr	freq = 'd',context_length = 100, prediction_length = 924, num_layers = 2, num_cells = 40, cell_type = "gru", epochs = 50, cardinality = cardinality.tolist()

Taula 12: Q6

Universitat Autònoma de Barcelona

Departament de Ciències de la Computació

Bellaterra, 28 d'Agost, 2023

A qui correspongui,

Escric aquesta carta per recomanar el projecte de final del grau Matemàtica Computacional i Analítica de Dades de la UAB, realitzat per l'estudiant Carlota Castro Sánchez titulat "*Modelització de dades temporals sobre el comportament humà: impacte de variables exògenes en la predicció d'agressions sexuals*". Com a professor de la UAB des de fa més de 20 anys he tingut el privilegi de dirigir nombrosos projectes, molts d'ells excepcionals, i crec fermament que aquest que es presenta en la competició és un exemple d'excel·lència aplicant metodologia científica a una lacra de la societat.

Aquest projecte destaca pel seu profund enfocament interdisciplinari, combinant a la perfecció les tècniques avançades d'aprenentatge automàtic per a la modelització i predicció de l'evolució temporal de la violència domèstica. La capacitat de l'estudiant d'aplicar mètodes d'avantguarda per predir dates específiques d'augment potencial de la violència domèstica és innovadora i socialment amb un impacte prometedor. El projecte aborda amb rigor la necessitat d'un millor anàlisi en les dades sobre la violència domèstica que es recullen diàriament per part de la forces i cossos de seguretat.

El rigor tècnic que l'estudiant Carlota Castro ha demostrat en aquest projecte és destacable, sobretot per l'aplicació de mètodes avançats de "deep learning" en sèries temporals de dades de violència domèstica en els darrers anys de la ciutat de Chicago, que són públicament disponibles. L'habilitat de l'estudiant en el preprocessament de dades, la programació de les solucions, i la selecció dels millors models demostra la seva comprensió profunda dels principis de l'aprenentatge automàtic. La utilització de diversos conjunts de dades sobre comportament humà, incloses dades de comportament de compra en supermercats de l'Equador, mostra l'enfocament integral del projecte per intentar capturar aspectes holístics en el comportament humà.

El rendiment predictiu del model desenvolupat per la Carlota ha estat molt bo, amb resultats estadísticament significatius i fiables. Aquest fet mostra el potencial del projecte per contribuir al benestar social, amb la capacitat de proporcionar informació oportuna per a la intervenció i l'assignació de recursos per a la lluita de la violència de gènere. De fet, aquest tipus de projectes subratllen el potencial transformador de la investigació científica en Intel·ligència Artificial per abordar desafiaments socials

Edifici Q, Campus de la UAB, 08193 Bellaterra (Cerdanyola del Vallès)
Tel: +34935813015 Fax: +34935814740 Email: d.ciencies.computacio@uab.cat

complexos. De fet, aquest projecte ha permès estar en converses amb l'Àrea Central de Violències Sexuals dels Mossos d'Esquadra, creada recentment, per tal d'adaptar els models apresos amb les dades recopilades en els darrers anys a Catalunya.

A més dels mèrits tècnics del projecte, destaca la comunicació hàbil de l'estudiant de la seva recerca. El document és exhaustiu i comprensible, fent que el projecte sigui accessible també al públic no tècnic.

Així doncs, destaco en aquest projecte la combinació d'innovació, excel·lència tècnica, importància del problema i potencial d'impacte en el món real, que el converteixen en una representació exemplar de recerca multidisciplinar La dedicació de l'estudiant Carlota Castro a superar els límits del coneixement i aplicar la seva experiència per crear un canvi positiu és realment inspiradora.

Estic segur que reconèixer els mèrits d'aquest projecte no només demostra el treball excepcional que ha fet l'estudiant, sinó que també posa l'accent en la importància d'aprofitar les tecnologies avançades en Intel·ligència Artificial per tenir impacte en la societat.

Gràcies per tenir en compte aquesta carta de recomanació, i organitzar aquests premis per reconèixer i celebrar l'excel·lència en la investigació científica en Intel·ligència Artificial que es fa a nivell de Graus i Màsters universitaris.

Sincerament,

Jordi Gonzalez Sabaté - DNI 52606264C (TCAT)

2023.08.28 17:27:08 +02'00'

Dr. Jordi Gonzàlez
Departament de Ciències de la Computació
Universitat Autònoma de Barcelona
Bellaterra, Barcelona, Spain 08193
Telf: +34 935 81 1828