PRA2 - Projecte de mineria de dades

Mineria de dades

Autor: Marc Nosàs Pomares

Data Realització: Barcelona, 10 de Gener 2024

Introducció

Abans de començar amb la realització d'aquesta segona part de la pràctica repasso el que vaig fer en la pràctica anterior i del *dataset*.

El *dataset* consisteix de les reserves de 3 anys d'un motor de reserves. Per conservar la privacitat dels clients, vaig anonimitzar alguns elements del *dataset*. Originalment contàvem amb 391.579 registres amb un total de 45 columnes.

També vam assolir alguns objectius analítics. El primer dels objectius és la predicció de cancel·lació de les reserves. La intenció és millorar el *reporting* que donem als clients sobre les reserves adquirides i el *revenue* que s'espera del mes següent, així com trobar *insights* que ens ajudin a reduir el número de cancel·lacions en el futur.

Després de fer un anàlisis exploratori del dataset i esudiar les correlacions entre columnes vam establir un procés de neteja de dades així com la normalització de camps numèrics, la discretització d'alguns camps i l'aplicació d'un estudi SVD.

En aquesta pràctica l'objectiu serà aplicar diferents tipus de models analítics, tant supervisats com no supervisats, sobre el conjunt de dades per resoldre l'objectiu plantejat en el punt anterior.

Igual que en la pràctica anterior he utilitzat python. La documentació referent al codi l'he posat en forma de comentaris en el mateix codi (bàsicament les llibreries) i no estan referenciades a la bibliografia d'aquest informe.

1 Exercici 1

El model no supervisat que he utilitzat és el k-means amb 2 clusters. A continuació, mostraré les mesures de qualitat del model generat i analitzem la qualitat del model generat.

Per a avaluar el model el que ens resultarà més útil és el valor de l'exactitud i la matriu de confusió (que ens permetrà calcular la **Precisió**, la **Sensibilitat** o la ***F-measure***). Totes aquestes mesures les hem fetes gràcies a la funció `metriques_qualitat`.

En un problema binari no equilibrat, com en el cas de les reserves hoteleres on només un 21% de les reserves es cancel·len, l'exactitud pot ser enganyosa. Si el model simplement classifica totes les mostres com a reserves no cancel·lades, encara pot aconseguir una alta exactitud, ja que la gran majoria de les mostres pertanyen a aquesta classe. No obstant això, aquest model no aporta cap valor en termes de predicció de reserves cancel·lades. Això ens passarà en tota la pràctica, per això serà important fixar-nos en els altres indicatius.

Així doncs, aquest és el resultat per al nostre K-means:

- Exactitud: 40.51%
- Precisió: 62.13%
- F-measure: 49.04%
- TRUE-TRUE: 27.52% — FALSE-TRUE: 7.92%
- TRUE-FALSE: 51.57% — FALSE-FALSE: 12.99%

D'aquests resultats en podem treure algunes conclusions:

- L'exactitud de 40.51% indica que el model té un rendiment força baix en la predicció si la reserva es cancelarà o no. Això implica que, s'espera que el 40.51% de les prediccions siguin correctes.
- La precisió de 62.13% ens indica que el model té una molt major capacitat per predir correctament les reserves cancel·lades. Això significa que el 62.13% de les reserves identificades com a cancel·lades són realment cancel·lades. En altres paraules, el model té una taxa de falsos positius del 37.87% en la predicció de cancel·lacions. Aquesta és una dada prou bona tenint en compte el resultat de l'exactitud i que la intenció és detectar la potencialitat de que la reserva sigui cancelada. Podria ser que de les que no es van cancelar siguessin potencialment çancelables".
- El valor de F-measure de 49.04% indica que el model té un rendiment baix en termes de precisió i recuperació combinades. Aquesta mesura té en compte tant la precisió com la recuperació i ens proporciona una mitjana harmònica dels dos.

2 Exercici 2

En aquest punt se'ns demana canviar la mètrica de distància. He decidit utilitzar la similitud dels cosinus. Mentre que la distància euclidiana es calcula com l'arrel quadrada de la suma dels quadrats de les diferències entre dos punts, la similitud del cosinus mesura el cosinus de l'angle entre aquests dos punts. Això pot ser útil per a dades disperses i d'alta dimensionalitat (com és el cas). També normalitza el resultat, de manera que no es veu afectat per la magnitud de les dades, el que pot ser beneficiós en molts escenaris.

Aixi doncs, com que només volem conèixer l'angle entre vectors per canviar la mètrica només ens caldrà normalitzar les dades abans d'aplicar Kmeans. Per la resta podem utilitzar les mateixes condicions.[1].

Aquest ha sigut el nou resultat:

- Exactitud: 40.53%
- Precisió: 62.11%
- F-measure: 49.05%
- TRUE-TRUE: 27.54% — FALSE-TRUE: 7.92%

[1] Aquest concepte està ben explicat al següent article de Medium: "Exploring Cosine Similarity and Cosine Distance"

- TRUE-FALSE: 51.55% — FALSE-FALSE: 12.99%

Veiem que els resultats han sigut gairebé idèntics. Per tant el canvi de mètrica de distància no ha ajudat en absolut.

3 Exercici 3

En aquest exercici treballarem amb els algorismes de DBSCAN i OPTICS.

En ambdós casos he reduit molt el nombre de registres per a fer l'avaluació dels valors òptims i després he corregut els valors òptims amb una part més gran del *dataset*[2].

1. **DBSCAN**:
   - Per a l'avaluació del model he provat els valors per a diferents valors de ϵ:

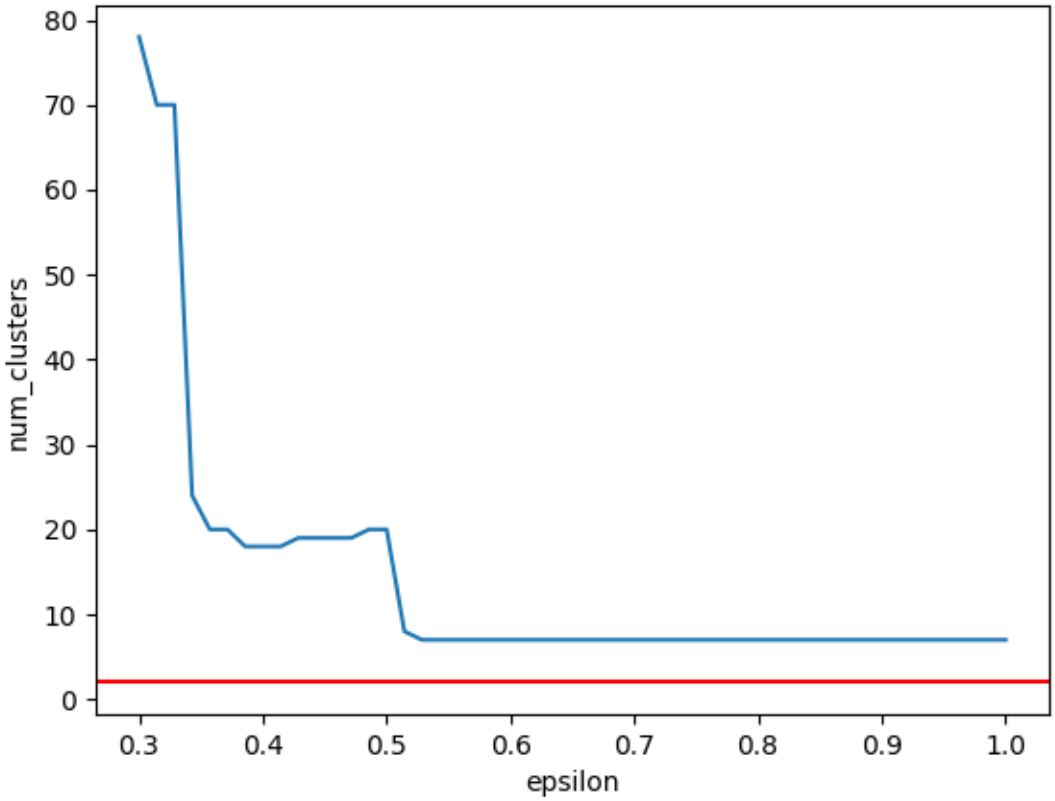


Figura 1: Relació entre el valor de ϵ i el número de clústers generats. La linia vermella marca n=2.

En la Fig.1, es recomanable buscar un valor de ϵ on es produeix un canvi significatiu en la pendent de la corba. Aquest canvi significatiu indica una transició en la densitat de les dades i, per tant, pot ser utilitzat com una bona estimació per al valor óptim. En la figura n'observem

[2]Ho he fet amb les 100.000 primeres files del dataset i ha trigat unes hores però els resultats han sigut força dolents per tant he decidit no augmentar el número de files

2, però és rellevant que a partir de $\epsilon = 0.5$ es perd el pendent. Per això he utilitzat aquell valor (on tindrem un número reduït de clústers).

- Els resultats per a aquest estudi han sigut:
  - Exactitud: 74.37 %
  - Precisió: 3.53 %
  - F-measure: 6.74 %
  - TRUE-TRUE: 73.52 % — FALSE-TRUE:23.30%
  - TRUE-FALSE: 2.33% — FALSE-FALSE: 0.85%

  A continuació, analitzarem els indicadors de qualitat i compararem els resultats amb els anteriors:

  - **Exactitud**: En aquest cas, la exactitud és del 74.37%. Això indica que el model té un millor rendiment en la predicció de les reserves cancel·lades que el model anterior. Esperem que el 74.37% de les prediccions siguin correctes. Atenció, això no vol dir que sigui més pertinent.
  - **Precisió**: La precisió és del 3.53%. Aquesta és una taxa **molt** baixa, el que significa que el model té una alta proporció de falsos positius en la predicció de les reserves cancel·lades. Només el 3.53% de les reserves identificades com a cancel·lades són realment cancel·lades.
  - **F-measure**: L'F-measure és del 6.74%, que és una mesura molt baixa de la precisió i la recuperació combinades. Això indica que el model no té un bon rendiment en termes de precisió i recuperació de les reserves cancel·lades.

  En resum, si bé el model DBSCAN té una millor exactitud en comparació amb el model k-means, la seva precisió i F-measure són molt baixes. Això indica que el model té una alta proporció de falsos positius i un rendiment global pobre en la predicció de les reserves cancel·lades. Una manera de millorar això seria intentar balancejar el volum de positius i negatius. Tot i així ho he provat i els resultats milloren molt poc.

2. **OPTICS**:

- En aquest cas també hem d'optimitzar una variable que entrem al algoritme, el número mínim de punts que ha de tenir un clúster. Igualment, busquem observar un canvi de tendència en el número de clústers en funció del mínim de punts. Si tenim molts punts l'algoritme s'equivocarà poc, però simplement perquè gairebé tindrem 1 punt per clúster.

  Doncs bé hem avaluat diferents valors entre 1 i 100 per obtenir el següent resultat:

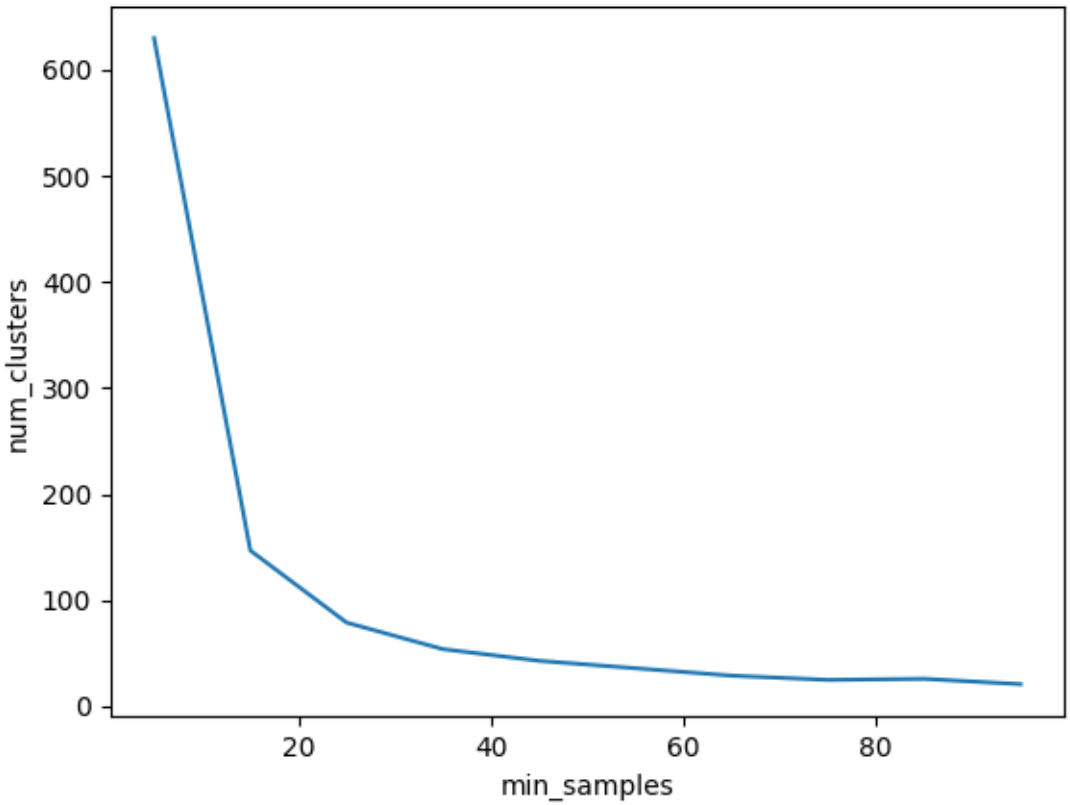


Figura 2: Relació entre el valor del mínim de punts i el número de clústers generats.

Veient els resultas he decidit obtar per el valor `min_samples` = 30. Amb clústers més petits segurament obtindrem millors resultats però serien poc útils de cara a dades de les que no coneixem la resposta.

- Els resultats han sigut similars (tot i que pitjors) als obtinguts amb DBSCAN:
  - Exactitud: 76.01%
  - Precisió: 0.02%
  - F-measure: 0.05%
  - TRUE-TRUE: 76.01% — FALSE-TRUE: 23.93%
  - TRUE-FALSE: 0.05% — FALSE-FALSE: 0.01%

Una vegada més ens trobem que la precisió és molt baixa.Veiem que hi ha un biaix molt gran a les dades. En aquest cas, com que la majoria de les reserves no es cancel·len, el model podria tenir una tendència inherent a predir que les reserves no es cancel·laran, independentment de les característiques específiques de cada reserva. A més el fet de que hi hagi molta dimensionalitat no ajuda al model. Per tal d'intentar solucionar aquest fet he agafat només algunes columnes que (per experiència) sé que són més rellevants de cara a la cancel·lació, tot i així els resultats han millorat

de forma insignificant.

Per tant, hem vist que DBSCAN i OPTICS no són el mètode ideal per detectar la probabilitat de cancel·lació de les reserves. Estem tractant un problema de classificació binària, aplicar tècniques de *clustering* a un problema de classificació pot no donar resultats òptims per diverses raons.

En primer lloc, tant DBSCAN com OPTICS busquen regions d'alta densitat separades per regions de baixa densitat. Si les cancel·lacions estan distribuïdes en molts tipus diferents de reserves i no es limiten a una àrea densament agrupada, aquests algoritmes poden tenir un rendiment baix. Això es pot veure afectat per al gran número de dimensions que tenim. Per exemple, les reserves estan repartides de forma equitativa en el temps per tant la densitat en aquest sentit serà força uniforme. Això passarà amb la gran majoria de les dimensions.

En segon lloc, com que aquests algoritmes no utilitzen etiquetes de classe per formar clústers, poden fallar en detectar els patrons complexos de les dades que separen les cancel·lacions de les no-cancel·lacions, cosa que ha conduit a una baixa precisió.

4 Exercici 4

Sempre que he fet particions per a test-train he utilitzat proporcions d'entre 70-30 i 80-20. En funció de la complexitat i el volumn de dades. Donat que el tamany del conjunt de dades és d'aproximadament 325.437 registres, una divisió 80/20 tindria sentit, proporcionant 260k registres per a l'entrenament i 65k per al test.

La raó d'aquesta divisió és tenir un conjunt de dades prou gran per a l'entrenament per permetre que l'algoritme de Random Forest aprengui i creï un model precís, alhora que es disposa d'una quantitat suficient de dades per provar el poder predictiu del model. El conjunt de prova es manté separat del procés d'entrenament i s'utilitza per donar una estimació imparcial de l'adequació final del model als dades d'entrenament.

També és important tenir en compte la distribució de les cancel·lacions dins del conjunt de dades. Donada la distribució de les cancel·lacions (el 20% de les dades), és crucial assegurar-se que això es reflecteixi adequadament tant en els conjunts d'entrenament com en els de prova, per això he comprobat aquestes dades en el meu codi. Aquestes són les distribucions:

	Train		Test	
	False	True	False	True
Count	205.769	54.580	51.613	13.475
%	79	21	79	21

5 Exercici 5

En aquest punt he fet dos models diferents. Un per poder entendre i generar una imatge de l'arbre amb les normes (amb una complexitat molt baixa) i un altre més complex, amb diferents arbres i més difícil d'interpretar.

- El primer cas el resultat ha sigut molt dolent (com es podria esperar):
  - Exactitud: 79.45%

- Precisió: 1.07%
- F-measure: 2.11%
- TRUE-TRUE: 79.23% — FALSE-TRUE: 20.48%
- TRUE-FALSE: 0.07% — FALSE-FALSE: 0.22%

Els resultats segurament es deuen a la poca profunditat de l'arbre i el nombre d'arbres. Però ens permet crear una imatge llegible d'un dels arbres.

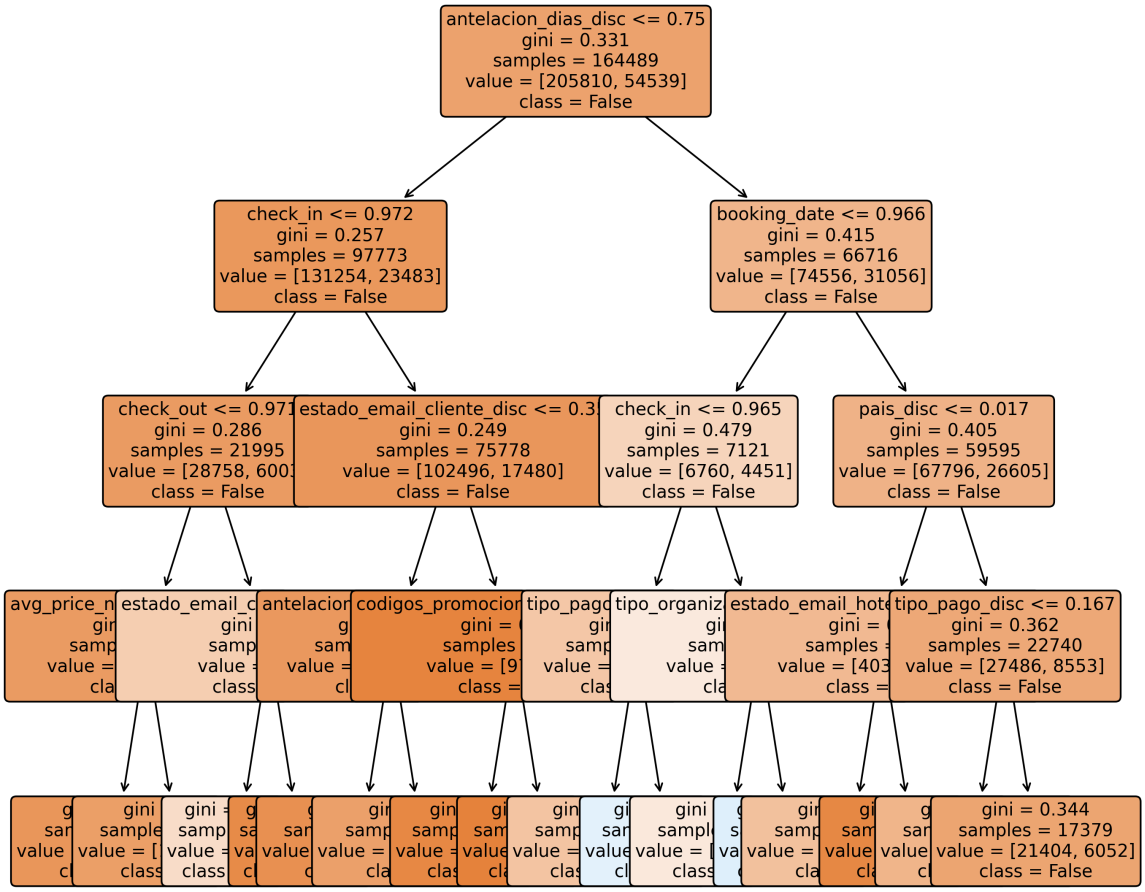
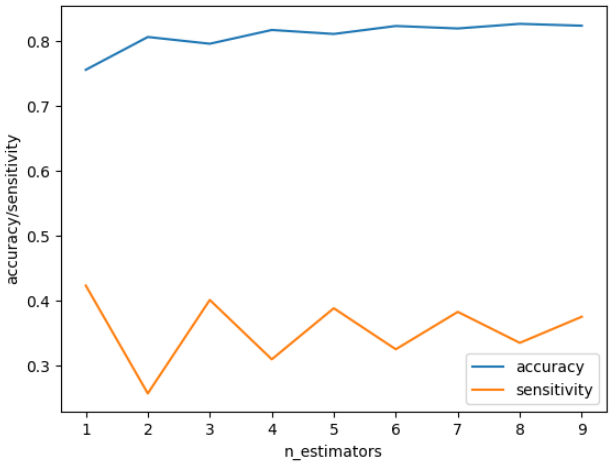


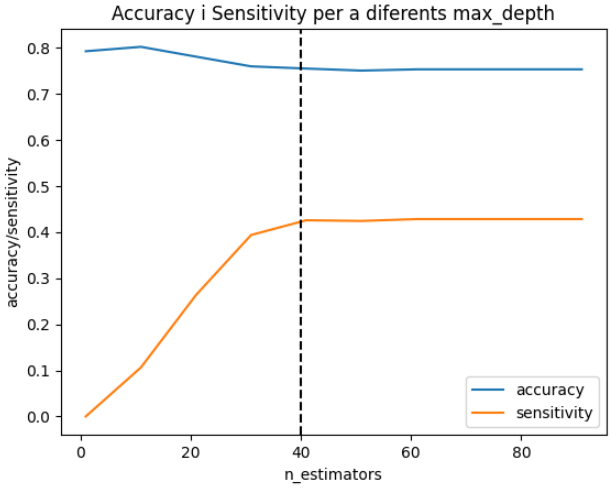
Figura 3: Imatge d'un dels arbres del Random Forest Classification, amb una profunditat de 4 capes.

A més amb el codi de python també es genera el document de text amb totes les regles, tant per

aquest cas com per al cas de més profunditat.

- Per a obtenir millors resultats el que he fet és optimitzar el número d'arbres i la profunditat. El que he fet és intentar optimitzar la sensitivitat modificant la profunditat i el nombre d'arbres. Els resultats han sigut interessants, amb un sol arbre la sensitivitat és la més alta. En quan a la profunditat al voltant de les 40 capes es satura i no millora gaire. Aquests són els resultats:





Així doncs ens hem quedat amb: `max_depth` = 40 i `n_estimators` = 1. I els resultats han sigut els següents:

  - Exactitud: 75.65%
  - Precisió: 42.39%
  - F-measure: 54.33%

Per tant, podem concloure que el Random Forest Classifier és clarament el mètode d'aprenentatge automàtic més eficient per predir si la reserva es cancel·larà, en comparació amb els altres mètodes

utilitzats.

Si mirem les xifres, la precisió del Random Forest Classifier és considerablement més alta que la dels altres dos mètodes. Aquest té una precisió del 42,39%, amb una exactitud de més d'un 75% (en el cas de K-means la precisió era bona però moltes que no es cancel·len es veien com a cancel·lades i l'exactitud era molt baixa).

A més, l'exactitud del Random Forest Classifier també és la més elevada (75,65%), el que significa que aquest mètode té una capacitat general més gran per predir correctament si les reserves es cancel·laran o no.

Una altra cosa que he fet és intentar veure quines són les columnes més importants per l'algoritme. La que agafa com a més important és la columna de "booking_date". És molt rellevant ja que es podria deure a factors dependents de les dates, (per exemple, el Covid) que són imprevisibles.

6 Exercici 6

com que ja he aprofundit força en l'algoritme de `RandomForestClassifier` he decidit provar-ne un de diferent. A les guies d'Scikit-learn he trobat GradientBoostingClassifier que sembla una bona opció per a classificacions com la que ens ocupa. Bàsicament, és una combinació de diferents algoritmes senzills.

Així doncs anem a veure els resultats obtinguts:

- Exactitud: 80.04%
- Precisió: 6.25%
- F-measure: 11.59%

Veiem que els resultats d'exactitud és el millor en tots els casos anteriors, però en canvi a nivell de Precisió els resultats són pitjors que en el cas del Random Forest Classifier.

7 Exercici 7: Conclusions

En aquesta secció, abordarem les limitacions inherents al nostre conjunt de dades. Prèviament hem detallat els nostres resultats, ara ens centrarem en aquesta discussió sobre les restriccions del dataset utilitzat.

Recordem l'objectiu de la nostra pràctica: utilitzem un conjunt de dades sobre reserves hoteleres per desenvolupar models d'aprenentatge automàtic, amb l'esperança de predir la probabilitat de cancel·lació d'una reserva.

El nostre dataset inclou un total de 391.579 registres, però presenta un desequilibri significatiu. La majoria de les reserves, concretament el 79%, no són cancel·lades, enfront d'un 21% que sí ho són. Aquesta disparitat pot afectar l'aprenentatge dels nostres models, especialment en relació a l'extracció de les característiques associades a les reserves que es cancel·len. Sense una supervisionat adequada, els algoritmes podrien inclinar-se a predir que la majoria de reserves no seran cancel·lades, la qual cosa contradiu la finalitat de la nostra pràctica.

El nostre objectiu principal és l'optimització de la precisió dels algoritmes. Busquem determinar la

probabilitat de cancel·lació d'una reserva de la forma més accurada possible.

- **Limitacions**:

  Una de les principals limitacions del *dataset* radica en la temporalitat de les dades, que va de 2019 fins a 2023. Aquest període inclou la pandèmia del Covid-19, un esdeveniment d'impacte enorme que causà que els índexs de cancel·lació augmentessin dràsticament. A més, la variabilitat en les raons de cancel·lació fa que la identificació de patrons sigui particularment complexa, ja que aquestes raons poden ser altament aleatòries i no sempre estan clarament relacionades amb les dades disponibles.

  Les nostres prediccions actuals mostren una precisió baixa (un màxim de menys d'un 45% en el cas del RFC) i anticipo que això podria ser encara més evident al treballar amb dades futures. Principalment, això és a causa de l'existència d'un biaix temporal en les nostres dades i dels possibles canvis de paradigma de cara al futur.

  Un clar exemple d'aquesta problemàtica és l'ocurrència d'incidents imprevistos que pot causar la cancel·lació en bloc de nombroses reserves (com l'exemple ja esmentat o d'altres, imaginem una cancel·lació d'un esdeveniment a una ciutat o un desastre natural, per exemple). Aquest tipus d'esdeveniments segueix un patró clar a la nostra base de dades, però la seva naturalesa és inherentment imprevisible. Això planteja una qüestió sobre l'eficàcia dels nostres models per afrontar aquestes situacions en les dades del futur.

  En resum, tot i que els nostres models d'aprenentatge automàtic es plantejaven amb l'esperança d'identificar les tendències de cancel·lació, he pogut constatar que aquests objectius no es poden assolir de manera satisfactòria. La majoria de patrons identificats es veuen fortament distorsionats pel biaix temporal i l'aleatorietat en les cancel·lacions, limitant significativament la utilitat de l'estudi.

  Dit això, el model que ha funcionat millor és el Random Forest Classifier i és el model que crec que hauríem d'utilitzar si volem seguir amb l'estudi.

- **Riscos d'ús d'aquest model**:

  En primer lloc, malgrat que la taxa d'exactitud del model es manté a un 75.65%, aquesta xifra pot ser enganyosa. Aquest percentatge d'exactitud indica que una tres quarts parts de les prediccions del model són correctes, però no distingeix entre les prediccions positives i negatives. El nostre principal interès aquí rau en les prediccions de cancel·lacions de reserva, que és un resultat menys freqüent, com ja he explicat.

  En aquesta línia, la precisió del model es manté a un 42.39%, una xifra relativament baixa. Això significa que gairebé la meitat de les vegades que el model prediu una cancel·lació, aquesta no es produeix. Des d'una perspectiva pràctica, això podria dur a una mala gestió de les reserves, ja que l'hotel podria preparar-se per a una cancel·lació que realment no ocorrerà.

  Addicionalment, la F-measure del model és de 54.33%. Aquesta mesura combina la precisió i la recordació en un sol valor, proporcionant-nos una visió més completa del rendiment del model quan les classes estan desequilibrades, com és el nostre cas. Llevat que estiguem disposats a acceptar falsos positius amb freqüència (cancel·lacions que el model prediu però que no es produeixen en realitat), un valor de F-measure per sota del 60% no és adequat.

Així doncs les conclusions no són massa esperançadores, el treball fet pot servir per intentar observar tendències i, potser, veure quines són les variables que més afavoreixen la cancel·lació. Per millorar el rendiment del model proposo algunes millores:

- En primer lloc, una estratègia podria ser començar l'entrenament del model amb dades de l'any 2019 i anar progressivament incorporant dades fins a 2024. D'aquesta manera, donaríem més importància a aquelles dades més recents que, presumiblement, es semblaran més a les situacions futures que volem preveure.
- Un altre canvi podria ser la inclusió de noves variables en l'anàlisi, com el dia de la setmana o el mes de la reserva. No obstant això, suggeriria eliminar la data exacta de reserva, així com les dates de check-in i check-out, ja que aquestes poden produïr un overfitting per les raons que ja he comentat.
- Finalment, crec que també ajudaria la reducció de la dimensionalitat del nostre conjunt de dades. Aquest procés, que implica l'eliminació de columnes que no contribueixen a la potència predictiva del model, o que inclús la poden perjudicar.

Referències

Colás, Raúl Montoliu (2021a). *Avaluació de models*.
— (2021b). *Models supervisats*.
Gironés, J., J. Casas i J. Minguillón (2017). *Minería de datos: Modelos y algoritmos*. Editorial UOC.
Roig, Jordi Gironés (2021). *Models no supervisat*.