Tasca 7.2: Projecte de ML per predir si un client d'una entitat bancària contractarà un nou dipòsit a termini.

Introducció

Es tracta de una entitat bancària que ha fet una campanya de màrqueting per incentivar la contractació de un nou producte (dipòsit a termini). Disposa tant de dades dels clients (financeres, administratives i personals) i també de la eficàcia de la campanya portada a terme. Per futures campanyes de màrqueting volen conèixer quin clients i amb quines característiques estan més predisposats a contractar un nou producte i per tant poder focalitzar la nova campanya en aquests clients, tot millorant la eficiència dels recursos disponibles per a la campanya de màrqueting.

Objectius del Projecte

1. ***Quins són els objectius del negoci?***

Augmentar la contractació de nous productes focalitzant la campanya de màrqueting en els clients mes predisposats a contractar un nou depòsit a termini. Conèixer els motius que fan que un client estigui més predisposat a contractar un nou producte.

2. ***Quines decisions o processos específics voleu millorar o automatitzar amb ML?***

Es cerca optimitzar els recursos de màrqueting (persones, trucades, publicitat,...), evitant contactar els clients amb baixa probabilitat de que contractin el nou producte i centrant els esforços en els clients amb més predisposició a la contractació del nou producte.

3. ***Es podria resoldre el problema de manera no automatitzada?***

Si, es podria resoldre el problema sense utilitzar un model de Machine Learning fent servir estadístiques, la pròpia experiència del comercials de la entitat bancaria o un anàlisi amb fulls de càlcul. Tot i així, es tractaria de una tasca força manual i no automatitzada. Addicionalment, donat el volum de dades requeriria un temps considerable per l'anàlisi.

En canvi, utilitzant Machine Learning el procés serà mes automàtic, amb menys intervenció de persones i es podrà anar millorant en el futur amb les noves dades que es recullin.

Metodologia Proposta

4. ***Quin és l'algorisme de Machine Learning més adequat per resoldre aquest problema? Com justifica l'elecció d'aquest algorisme? Que mètriques d'avaluació s'utilitzaran per a mesurar el rendiment del model?***

Els algorismes mes adients són l'**Arbre de Decisió** (*Decision Tree Classifier*), la **Regressió Logistica** (*Logistic Regression*) i el **Clasificador Bosc Aleatori** (*Random Forest Classifier*) , per les següents raons:

- Donat que tenim un conjunt de dades etiquetades i volem predir una d'aquestes dades ('*deposit*'), estem davant d'un aprenentatge supervisat. Tots dos algorismes proposats son de tipus ML supervisat.
- Volem una predicció binaria entre '*si*' o '*no*' un client contractarà el nou dipòsit ofert durant la nova campanya de màrqueting (la predicció 'y') en funció de tots o una part del paràmetres restants (les característiques 'X').

- L'objectiu fixat es filtrar els clients que amb més alta probabilitat contractaran el nou producte, es a dir, prioritzar els clients pels que la predicció del model de la característica '*deposit*' sigui '*yes*'.
- El model també ens donarà informació de quins son les característiques mes importants a l'hora de que un client contracti un nou producte. Amb aquesta informació podrem millorar en el futur els processos i la manera de adreçar-se al client. Si per exemple, el model determina que la característica mes rellevant per la nova contractació es que el client ja tingui un depòsit contractat, el banc podria decidir fer una campanya de dipòsits amb molt bones condicions per clients que no en tinguin, tot esperant que en el futur aquest clients contractin més nous dipòsits.

Addicionalment afegirem un classificador aleatori (*Dummy Classifier*) amb distribució de probabilitat igual a la distribució de la variable objectiu (*deposit)* com model de referència, de manera que cap model hauria de tenir mètriques pitjors que aquest model de referència.

Per mesurar el rendiment del model les mètriques candidates son les habituals de un model de classificació binaria: *Accuracy*, *Precision*, *Recall* i *F1-Score*, ja que cada mètrica ens dona informació diferent i complementaria. Amb els responsables de l'entitat s'ha de decidir quins seran els valors mínims a considerar per validar el model i també el impacte que puguin tenir els errors de predicció (clients no contactats que si que haguessin contractat el depòsit o clients contactats que finalment rebutgen l'oferta del nou producte). El càlcul de les mètriques es farà tant amb el data set d'entrenament com amb el de test (que ja hem separat prèviament). Les mètriques sobre el data set de test ens permetrà conèixer el rendiment sobre dades que el model no ha vist mai i la comparativa amb les mètriques sobre el data set d'entrenament ens permetrà conèixer si estem en situació de sub-entrenament (*underfitting)* o sobre-entrenament *(overfitting)*.

Tot i que no coneixem els criteris desitjats per l'entitat bancaria per decidir quina seria la mètrica mes adient podem predir que:

- L'entitat bancaria estarà interessada en conèixer el rendiment global del model, es a dir, encertar tant amb els clients que contractaran el depòsit (sobre els que prioritzarà les trucades de la campanya) i els que no el contractaran (que no trucarà o trucarà amb menys prioritat). Per mesurar el rendiment global del model farem servir la mètrica ***accuracy***.
- El cost econòmic d'un fals positiu (el model prediu que el client contractarà el dipòsit però finalment no el contracta) es baix (cost d'una trucada telefònica i temps de la persona que truca). Pel contrari, el cost de un fals negatiu (el model prediu que no contractarà el dipòsit però realment l'hagués contractat si l'haguessin trucat) pot ser superior ja que es perd la contractació de un servei. En definitiva, es important mantenir baix el número de falsos negatius i per tant utilitzarem la mètrica ***recall***.

Per tal de comparar els diferents models i millorar el rendiment del model finalment triat, calcularem la corba ROC que ens permetrà ajustar el llindar de decisió dels models per tal de minimitzar el número de falsos negatius como hem comentat abans.

Dades Disponibles

5. *Quines dades estan disponibles per abordar aquest problema?*

Tenim disponibles mes d'onze mil registres de clients, convenientment etiquetats. Estan disponibles dades que a priori tenen rellevància de cara a predir si contractaran o no el nou

dipòsit (balanç actual, si ja te un dipòsit, si la campanya anterior va tenir èxit, temps des de l'últim contacte...).

Mètrica d'Èxit

6. ***Quina és la mètrica d'èxit per a aquest projecte?***

Increment, comparat amb campanyes anteriors, en el percentatge de nous depòsits contractats respecte del número de contactes realitzats.

Responsabilitats Ètiques i Socials

7. **Quines responsabilitats ètiques i socials és *important* tenir en compte?**

Tindrem en compte les següents responsabilitats ètiques i socials:

- Benefici humà, tant per el client (oferiment de un producte desitjat) com per la entitat bancaria (increment del nombre de nous depòsits contractats).
- Transparència i explicabilitat, gràcies a treballar amb models coneguts (arbre de decisió, regressió logística) que es poden avaluar amb mètriques i aporten informació de com estan calculant les prediccions (característiques mes importants).
- Justícia i equitat, donat que no utilitzarem dades que puguin esbiaixar els resultats, como podrien ser el gènere, rasa, religió o altres de similars. Així mateix, garantirem que tenim prou dades i amb elevada diversitat.
- Privacitat i seguretat, anonimitzant (o fins i tot eliminant) dades de caràcter personal (estat civil, feina, edat, educació..) o sensible (balanç, morositat,...).
- Responsabilitat per part dels desenvolupadors del model de predicció i redempció de comptes en el cas de danys o conseqüències negatives derivades del model aplicat.