Predicció de la Subscricpió de Dipòsits a Termini en Clients Bancaris

Introducció

En aquest projecte, l'objectiu és millorar l'eficiència de les campanyes de màrqueting d'una entitat bancària mitjançant la predicció de quins clients estan més inclinats a subscriure un dipòsit a termini.
Això permetrà a l'entitat optimitzar els seus recursos, dirigir millor els esforços de venda i augmentar la taxa de conversió en les seves campanyes.
El projecte és crucial per a l'entitat bancària, ja que la millora en les estratègies de màrqueting ajudarà a reduir costos i a incrementar els ingressos.

Objectius del Projecte

1. Quins són els objectius del negoci?
L'objectiu principal és augmentar la taxa de subscripcions de dipòsits a termini, maximitzant el retorn de les campanyes de màrqueting.
El banc busca identificar els clients amb més probabilitat de subscriure aquest producte, per així enfocar millor les seves estratègies de comunicació i venda.
***2. Quines decisions o processos específics voleu millorar o automatitzar amb* ML?**
El projecte busca automatitzar la classificació de clients en funció de la seva probabilitat de subscriure un dipòsit a termini.
També analitzar, dins del grup dels que tenen poca probabilitat, quins són els factors que hi incideixen i com transformar-los perquè tinguin més probabilitat de subscriure's.
Això permetrà millorar les decisions sobre quins esforços dirigir als grups que tenen probabilitat però també quins recursos dirigir als que no la tenen per canviar l'índex de probabilitat a alta.
Podem oferir condicions específiques o avantatges exclusius per certs perfils per tal de revertir la baixa probabilitat.
Tot aquest anàlisi permetrà dirigir els esforços de màrqueting, reduint costos associats amb campanyes massives i millorant la precisió dels missatges promocionals.
3. Es podria resoldre el problema de manera no automatitzada?
Tot i que és possible analitzar manualment dades de clients per identificar patrons de comportament, aquest procés seria lent i ineficient per a una gran quantitat de dades.
El Machine Learning permet automatitzar aquest procés, identificant patrons complexos que serien difícils de detectar manualment, i actualitzant les prediccions en temps real amb noves dades.

Metodologia Proposada
Es proposa utilitzar un algorisme de classificació binària i supervisada, ja que l'objectiu és predir si un client subscriurà o no un dipòsit a termini. Ja comptem amb la variable objectiu "dipòsit" amb resultats "sí" o "no".

Els algorismes més adequats per aquest tipus de predicció, que poden treballar tant amb variables categòriques com numèriques, són:

1. **Regressió Logística**
    - Modelarà la probabilitat que un esdeveniment ocorri, en aquest cas, si el client subscriurà o no el dipòsit. És fàcil d'interpretar i ofereix bons resultats en problemes de classificació binària amb dades estructurades.
    - **Inconvenient**: Pot no funcionar bé si hi ha relacions no lineals complexes entre les variables.
2. **Random Forest**
    - S'utilitzarà un conjunt d'arbres de decisió que combinarà els resultats de múltiples arbres per fer una predicció. Utilitzarà el principi de "bagging" per millorar la precisió i reduir el sobreajustament. Podrà gestionar tant dades numèriques com categòriques i detectar relacions complexes entre variables. És molt fiable i precís, especialment amb dades desequilibrades.
3. **Support Vector Machines (SVM)**

- Buscarà un hiperplà que separi les dues classes amb el marge més gran possible. Serà capaç de trobar solucions òptimes per a la classificació binària en casos de separació no lineal mitjançant el "kernel trick". És ideal per a relacions no lineals.
- **Inconvenients**: Serà difícil d'interpretar i requerirà una configuració acurada de paràmetres com el tipus de kernel utilitzat.

4. **Gradient Boosting (XGBoost)**
    - XGBoost serà un algoritme de gradient boosting que construirà models seqüencialment, corregint els errors dels models anteriors. És molt potent en problemes de classificació binària, excel·lent per treballar amb dades desequilibrades i amb moltes variables. Sol oferir resultats molt precisos.
    - **Inconvenients**: Podrà requerir un entrenament més llarg i una optimització acurada dels paràmetres.
5. **Arbres de Decisió**
    - Seran fàcils d'interpretar i visualitzar, cosa que facilitarà entendre com es prenen les decisions. No requeriran gaire preparació de les dades i podran gestionar tant variables categòriques com numèriques.
    - **Inconvenients**: Es podran sobreajustar fàcilment si no es poden adequadament. Seran sensibles a petites variacions en les dades, cosa que podrà generar arbres molt diferents i menys estables.

Es separaran les dades en proporció 70%-30% per entrenament i test, es crearà un **pipeline** i s'utilitzarà **GridSearchCV** per trobar els millors hiperparàmetres (**best estimators**) per a cada model. Després, es realitzarà un nou entrenament utilitzant aquests paràmetres optimitzats.

Posteriorment, es visualitzarà la **matriu de confusió** per avaluar la classificació de falsos positius, falsos negatius, veritables positius i veritables negatius. Els models es valoraran utilitzant les següents mètriques: **accuracy**, **precision**, **recall** i **F1 score**.

Després, es seleccionarà el model que es consideri més adequat, tenint en compte que no es contempla un gran risc en cas de falsos positius i es comprovarà si hi ha **overfitting** o **underfitting** mitjançant la **corba ROC**.

També es visualitzarà la **corba d'aprenentatge** per determinar si és necessària una reducció de dimensionalitat mitjançant **PCA**. Si es requereix reduir la dimensionalitat, s'utilitzaran dues components principals. A continuació, s'entrenarà el model amb aquesta nova transformació i es compararan els resultats amb les mateixes mètriques esmentades anteriorment.

Dades Disponibles

El conjunt de dades del dataset disponible "banc_dataset.CSV" inclou informació demogràfica i financera dels clients, així com dades relacionades amb campanyes de màrqueting anteriors. Les variables disponibles inclouen:

- **Dades del client**: Edat, ocupació, estat civil, nivell educatiu, balanç mitjà anual, préstecs (habitatge i personal), etc.
- **Dades de contacte**: Tipus de comunicació (cel·lular o telefònica), dia i mes del contacte, durada de la darrera trucada, etc.
- **Historial de campanyes**: Nombre de contactes en campanyes anteriors, resultat de les campanyes prèvies, etc.
- **Variable objectiu**: Subscriurà el client un dipòsit a termini? (Sí o no).

Mètrica d'èxit del projecte

La mètrica d'èxit serà l'impacte real en els resultats, o sigui, l'augment de subscripcions de dipòsits a termini. Medirem la taxa de conversió entre l'inversió en recursos i màrketing amb els resultats obtinguts.

Responsabilitats Ètiques i Socials

En la implementació d'aquest projecte, és crucial tenir en compte els aspectes ètics següents:

1. **Privadesa de les dades**: Les dades dels clients han de ser tractades amb la màxima confidencialitat, seguint totes les normatives legals com el GDPR. És important que el banc obtingui el consentiment dels clients abans d'utilitzar les seves dades per a la construcció de models de ML.
2. **Evitar discriminacions en les segmentacions**: Tot i que el model farà una segmentació sociodemogràfica, haurem d'aplicar posibles restriccions que assegurin que les decisions comercials siguin equitatives per a tots els grups d'edat, tots els estats civils i gènere, per exemple, establint

llindars mínims d'incentius per a tots els grups demogràfics identificats e intentar igualar els incentius per tots els gèneres i estats civils.
A nivell tècnic, si el model detecta que les persones solteres o casades reben tractaments diferents sense cap justificació objectiva, s'ajustarien els pesos d'aquestes variables perquè el model deixi de considerar-les com a factors determinants, desensibilitzatirem les variables.
Després d'entrenar el model, realitzarem una auditoria per assegurar-nos que no hi hagi resultats discriminatoris segons la segmentació realitzada en relació a estat civil, gènere o classe social.
3. **Transparència i explicabilitat**: Les decisions preses pel model han de ser explicables i transparents, per tal de generar confiança en el sistema, tant dins del banc com entre els clients. Tenint en compte tots els clients, edats, gèneres i llindars de renta.
4. **Impacte en els clients**: El model ha d'assegurar-se que els esforços de màrqueting no siguin invasius o molestos per als clients, promovent una relació equilibrada, responsable i consentida.
Thaïs Rocafull / 17 setembre 2024 / ML Tasca 3.2