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.
Això permetrà millorar les decisions sobre a qui 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 Proposta
Es proposa utilitzar un algorisme de **classificació binària**, ja que l'objectiu és predir si un client subscriurà o no un dipòsit a termini (resultat "sí" o "no").

Ens trobem amb: numèriques senceres com l'edat o el saldo mitjà, variables categòriques com la professió, estat civil, nivell educatiu... i algunes variables booleanes que indiquen si el client té crèdit de mora, un préstec habitatge o altre préstec personal. També varies variables dedicades al seguiment del client y markeging.

Els algorismes més adequats per aquest tipus de predicció i que poguin treballar amb variables categòriques i numèriques són:

1. **Regressió Logística**
   - Modela la probabilitat que un esdeveniment ocorri, en aquest cas, si el client subscriurà o no el dipòsit. És senzill 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
   - Conjunt d'arbres de decisió que combina els resultats de múltiples arbres per fer una predicció. Utilitza el principi de "bagging" per millorar l'exactitud i reduir el sobreajustament. Pot gestionar tant dades numèriques com categòriques i detectar relacions complexes entre les variables. Molt precís i capaç de gestionar dades desequilibrades.

- **Inconvenient**: Difícil d'interpretar, pot ser més lent a l'hora de fer prediccions.

3. Support Vector Machines (SVM)

- Busca un hiperplà que separa les dues classes amb el marge més gran possible. Capaç de trobar solucions òptimes per a la classificació binària en casos de separació no lineal mitjançant el "kernel trick". Ideal per relacions no lineals.
- **Inconvenients**: Difícil d'interpretar, i requereix una bona configuració de paràmetres com el tipus de kernel utilitzat.

4. Gradient Boosting (XGBoost)

- XGBoost és un algoritme de gradient boosting que construeix models seqüencialment, corregint els errors dels models anteriors. És un dels algoritmes més potents en problemes de classificació binària. Excel·lent per treballar amb conjunts de dades desequilibrats i amb moltes variables, i sol oferir resultats molt precisos.
- **Inconvenients**: Pot requerir un entrenament més llarg i una optimització acurada dels paràmetres.

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

Cada algoritme serà avaluat utilitzant mètriques com la **precisió**, **recall**, **F1-score** i **l'àrea sota la corba ROC (AUC-ROC)** per determinar quin model s'ajusta millor als objectius del projecte.

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 segmentacions**: El model ha d'evitar segmentacions relacionades amb característiques demogràfiques com l'edat, el gènere o l'estat civil, per no discriminar certs grups de clients.
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.
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.