ITBA
Instituto Tecnológico de Buenos Aires

82.05 Análisis Predictivo

Trabajo Práctico N° 2 - Modelo Predictivo

Manuel Dominguez

Caso de negocio: desarrollo de estrategias y campañas de marketing para mantener clientes

Objetivos

- Predecir con precisión si una persona tiene más probabilidad de dejar de ser cliente de un banco
- Establecer estrategias de marketing personalizadas para aquellas personas que posean una mayor probabilidad de dejar de ser clientes

Variable respuesta

Buscamos predecir si una persona se irá o no del banco

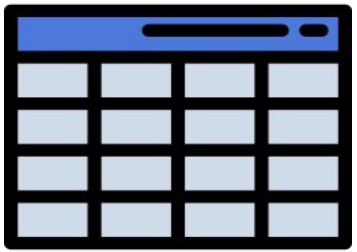
Dataset: descripción de los datos

Descripción del dataset utilizado

Se poseía un dataset de entrenamiento y un dataset de testeo con datos personales de aquellos clientes que se quería predecir si iban a dejar de ser clientes del banco o no.

A partir de los datos, se creó un dataframe de entrenamiento con las siguientes características:

- 8001 filas
- 15 columnas
- Contiene 7 variables numéricas y 8 categóricas



Descripción de las variables

Id	Id del cliente
Surname	el apellido del cliente o su nombre de familia
Credit Score	un valor numérico que representa el puntaje crediticio del cliente
Geography	el país donde reside el cliente (Francia, España o Alemania)
Gender	el género del cliente (Masculino o Femenino)
Age	la edad del cliente
Tenure	el número de años que el cliente ha estado con el banco
Balance	el saldo de la cuenta del cliente
NumOfProducts	el número de productos bancarios que utiliza el cliente (por ejemplo, cuenta de ahorros, tarjeta de crédito)
HasCrCard	si el cliente tiene una tarjeta de crédito (1 = sí, 0 = no)
IsActiveMember	si el cliente es un miembro activo (1 = sí, 0 = no)

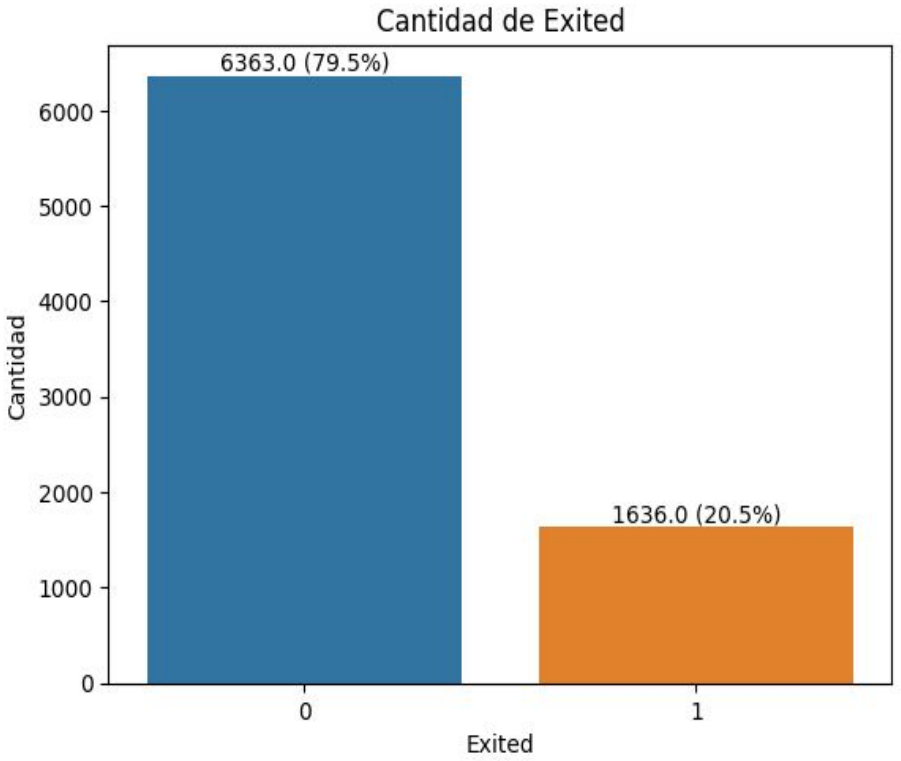
EstimatedSalary	el salario estimado del cliente
Passport	número de pasaporte
EducationYears	años de educación
Exited	si el cliente ha abandonado (1 = sí, 0 = no)

Análisis Exploratorio

Análisis Descriptivo

Distribución de Exited

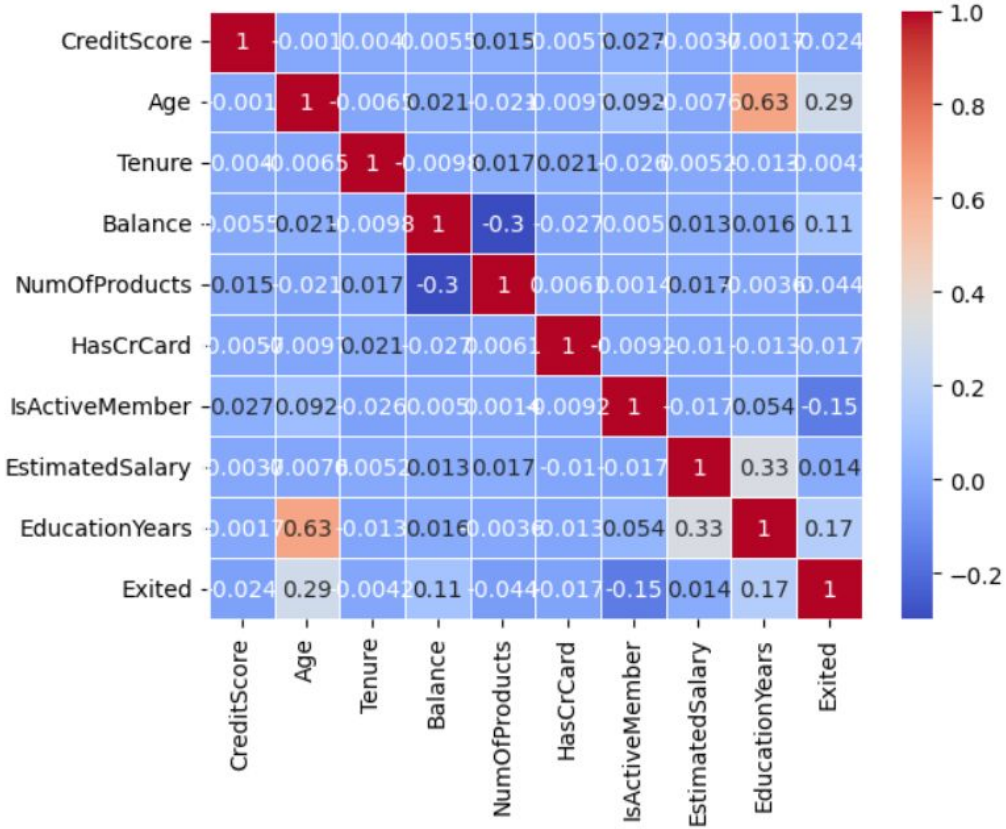
- De los 8001 registros, 6.363 se terminan quedando
- Casi el 20% de los clientes del banco terminan yéndose



Análisis Descriptivo

Matriz de correlaciones

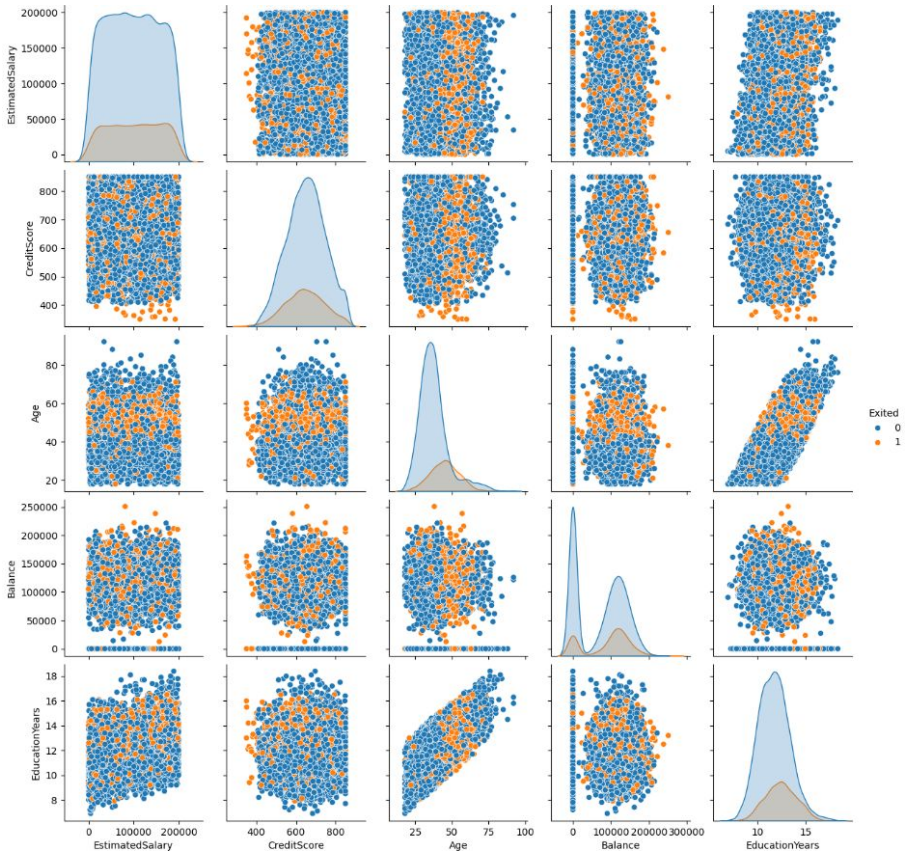
- Se utilizó el método de “Pearson” debido a su robustez
- No se observan correlaciones fuertes entre las variables
- La correlación más fuerte es entre la edad y los años de educación



Análisis Descriptivo

Scatter-plots

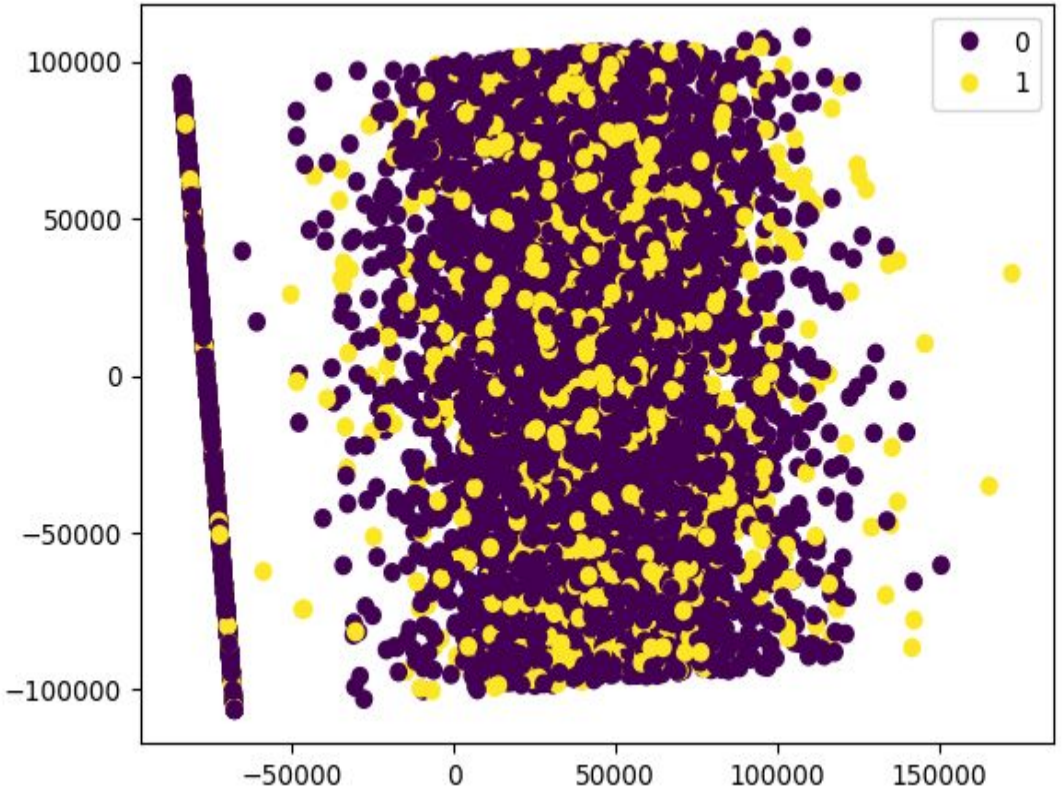
- No hay relaciones lineales entre las variables.
- Las únicas variables con relación son la edad y los años de educación



Análisis Descriptivo

PCA

- Se aplicó Análisis de componentes principales con el objetivo de intentar identificar posibles clusters
- No hay grupos claros entre las personas que dejan el banco y las que no



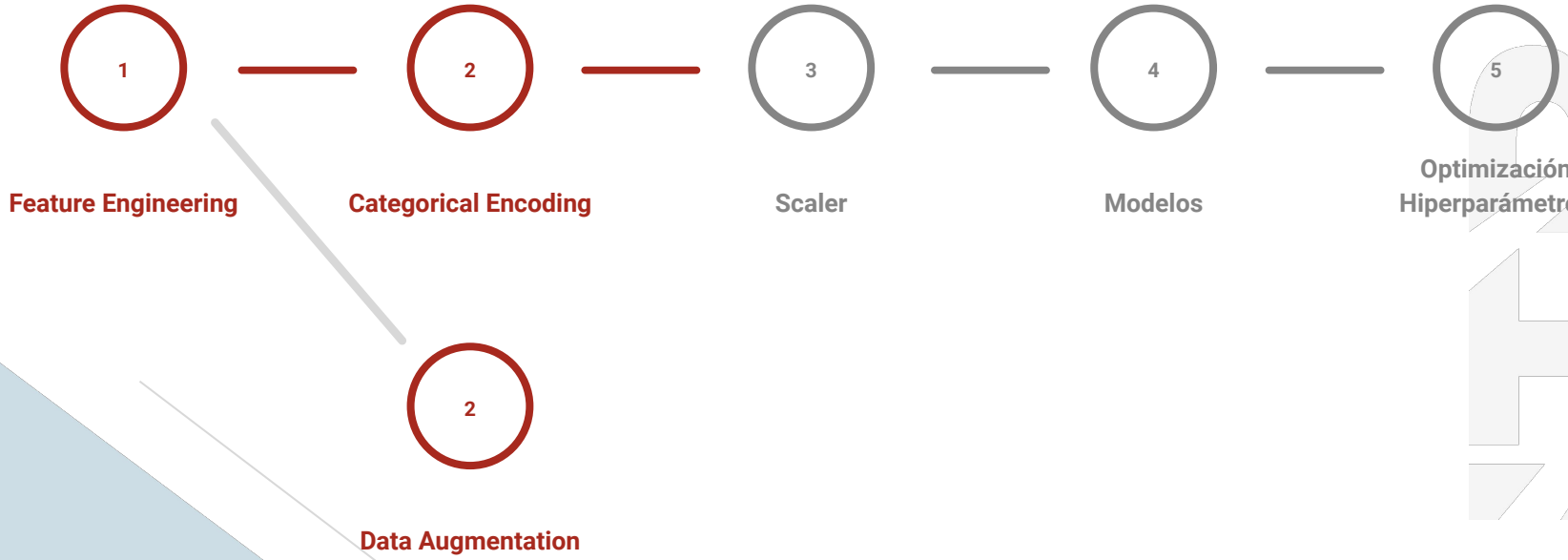
Modelos y Feature Engineering

Modelo Baseline

- Se eliminaron los datos nulos
- Se fitteo un XGBoost sin usar hiperparámetros ni usando Feature Engineering
- Se aplicó “One Hot Encoding” a las variable Geografía
- Se eliminaron las variables Género, Apellido, Pasaporte y ID

Accuracy Validation	Accuracy Kaggle
0,85125	0,84405

Proceso hacia modelo final



Feature Engineering

- Categorización de variables numéricas (Edad, CreditScore,...)
- Polynomial Features
- Uso de apellidos y pasaportes
- Transformaciones logarítmicas y potencias
- Creación de nuevas variables a partir de las originales
- Uso de técnicas de FeatureSelection (Forward Feature Selection)

Data Augmentation

- Concatenación de datasets agregando ruido (incremento a 100k registros)
- Uso de GANs para generar data tabular sintética (CTGAN)

Nuevas variables

CreditScore_x_Age	CreditScore según la edad del cliente
CreditScore_x_Balance	CreditScore según el balance de la cuenta
NumOfProducts_x_Age	Cantidad de Productos por edad
Tenure_x_Age	Tiempo en la empresa según la edad
%SalaryInBank	Porcentaje del salario que puede estar en el banco
Balance _x_ Estimated Salary	Balance según el salario
AgeofEntry	Edad cuando se entró al banco
CustomerEngagement	Nivel de actividad en el banco según edad, CreditScore y cantidad de productos
EducationProduct	Interacción entre la edad y los años del de educación del cliente y la cantidad de productos

Categorical Encoding

- One-Hot Encoding
- Label Encoder
- Count Frequency Encoder
- OrdinalEncoder
- Target Encoder
- CatBoost Encoder

Escalamiento de datos

- Análisis de **MinMaxScaler**, **StandardScaler**, **Normalizer** y **MaxAbs**
- Se evaluaron en distintos modelos teniendo en cuenta todas las métricas presentadas y sacando una ponderación
- **Standard Scaler ganador**

$$\frac{(Accuracy \times 1.5) + (Precision \times 0.75) + Recall + (F1\text{-}Score \times 1.5) + (AUC\text{-}Score \times 1.25) + (PRAUC \times 0.75)}{6}$$

Modelo	Scaler	Accuracy	Precision	Recall	F1-Score	AUC-Score	PRAUC	Ponderacion
Random Forest	MinMax	0,8395	0,606	0,631	0,618	0,762	0,656	0,8680417
Random Forest	StandardScaler	0,8415	0,611	0,631	0,621	0,763	0,659	0,870875
Random Forest	MaxAbs	0,8405	0,608	0,631	0,619	0,762	0,658	0,8692917
Random Forest	Normalizer	0,782	0,25	0,029	0,05	0,5	0,239	0,408
LGBM	MinMax	0,8625	0,696	0,589	0,638	0,761	0,685	0,8900833
LGBM	MinMax	0,8605	0,685	0,597	0,638	0,762	0,682	0,889
LGBM	StandardScaler	0,8605	0,693	0,577	0,6304	0,755	0,679	0,8825583
LGBM	StandardScaler	0,862	0,688	0,601	0,642	0,765	0,686	0,8930417
LGBM	MaxAbs	0,855	0,666	0,597	0,629	0,759	0,673	0,880125
LGBM	Normalizer	0,541	0,246	0,597	0,348	0,561	0,45	0,581875
XGB	MinMax	0,847	0,669	0,507	0,577	0,721	0,639	0,8340833
XGB	StandardScaler	0,847	0,669	0,507	0,577	0,721	0,639	0,8340833
XGB	MaxAbs	0,847	0,669	0,507	0,577	0,721	0,639	0,8340833
XGB	Normalizer	0,794	0	0	0	0,5	0,603	0,4534167
CatBoost	MinMax	0,843	0,638	0,604	0,621	0,756	0,663	0,8696667
CatBoost	MinMax	0,84	0,624	0,613	0,619	0,757	0,66	0,867625
CatBoost	StandardScaler	0,845	0,642	0,604	0,623	0,757	0,665	0,871875
CatBoost	StandardScaler	0,839	0,623	0,61	0,616	0,755	0,658	0,8650833
CatBoost	MaxAbs	0,845	0,642	0,604	0,623	0,757	0,665	0,871875
CatBoost	MaxAbs	0,839	0,623	0,61	0,616	0,755	0,658	0,8650833
CatBoost	Normalizer	0,211	0,211	1	0,349	0,5	0,605	0,5884583
CatBoost	StandardScaler	0,858	0,694	0,589	0,637	0,76	0,685	0,88825
CatBoost	StandardScaler	0,86	0,706	0,581	0,637	0,758	0,688	0,88925
CatBoost	StandardScaler	0,860625	0,707	0,584	0,639	0,759	0,689	0,8909896
CatBoost	StandardScaler	0,859	0,692	0,604	0,645	0,766	0,69	0,89525

Modelos

- Análisis de **RandomForest**, **LGBM**, **XGB**, **GradientBoosting, Neural Networks y Voting Classifier**

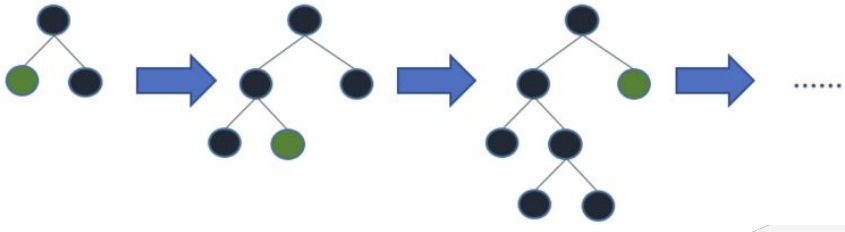
- Se utilizó GridSearchCV con StratifiedKFold

- **LGBM ganador**

Modelo	Accuracy	Precision	Recall	F1-Score	AUC-Score	PRAUC	Ponderacion
Random Forest	0,8704	0,7527	0,5535	0,6379	0,7531	0,6992	0,807708333
LGBM	0,8788	0,7914	0,5595	0,6556	0,7606	0,7209	0,824345833
XGB	0,8521	0,6768	0,5414	0,6016	0,7371	0,6564	0,773870833
GB	0,8563	0,6984	0,5333	0,6048	0,7367	0,664	0,7779375
CatBoost	0,8788	0,8018	0,5475	0,6507	0,7562	0,7213	0,821554167
NN	0,8554	0,7164	0,4949	0,5854	0,722	0,6577	0,7648625
VC (LGBM-RF)	0,8554	0,6468	0,6586	0,6527	0,7826	0,6879	0,816670833

Modelo Final

LightGBM



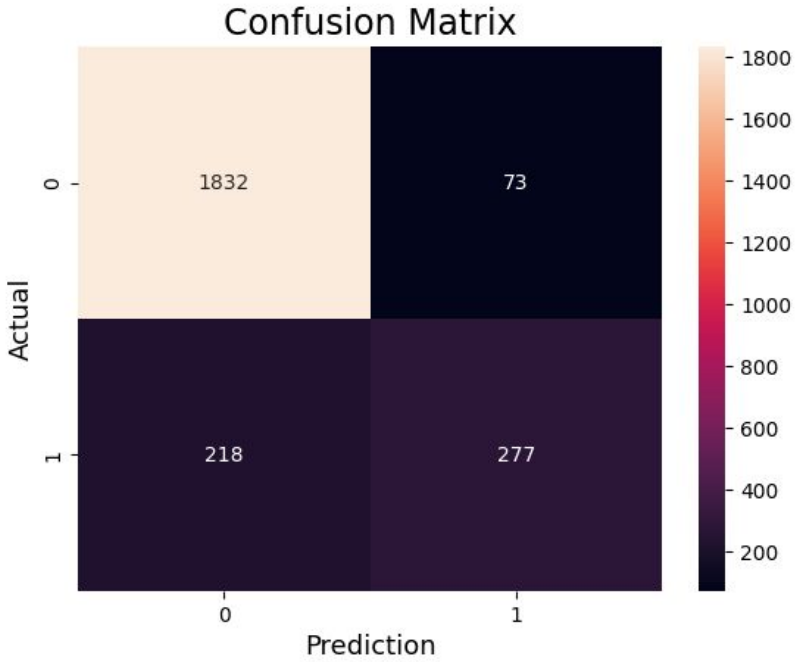
Resultados

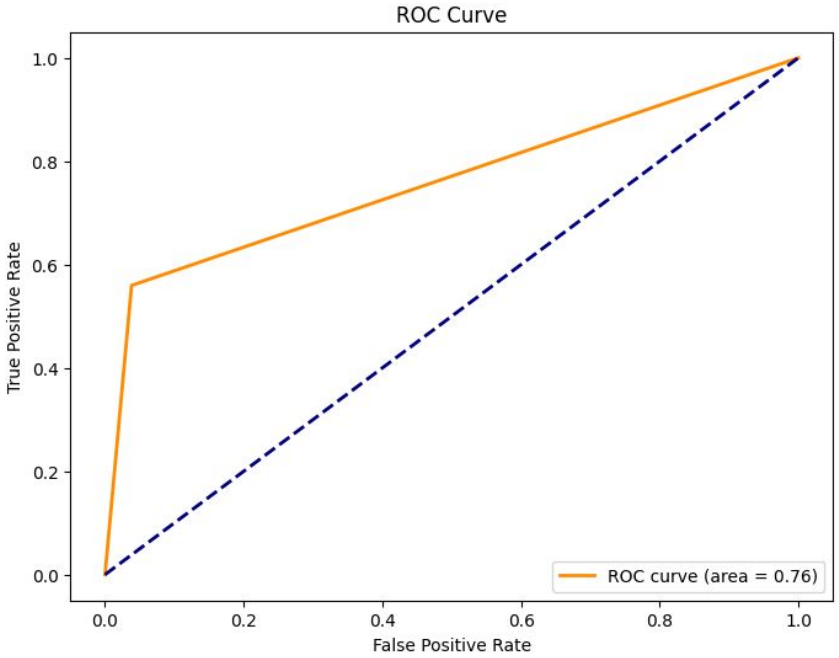
Accuracy	0,8788
Precision	0,7914
Recall	0,5595
F1	0,6556
AUC	0,7606
PRAUC	0,7209

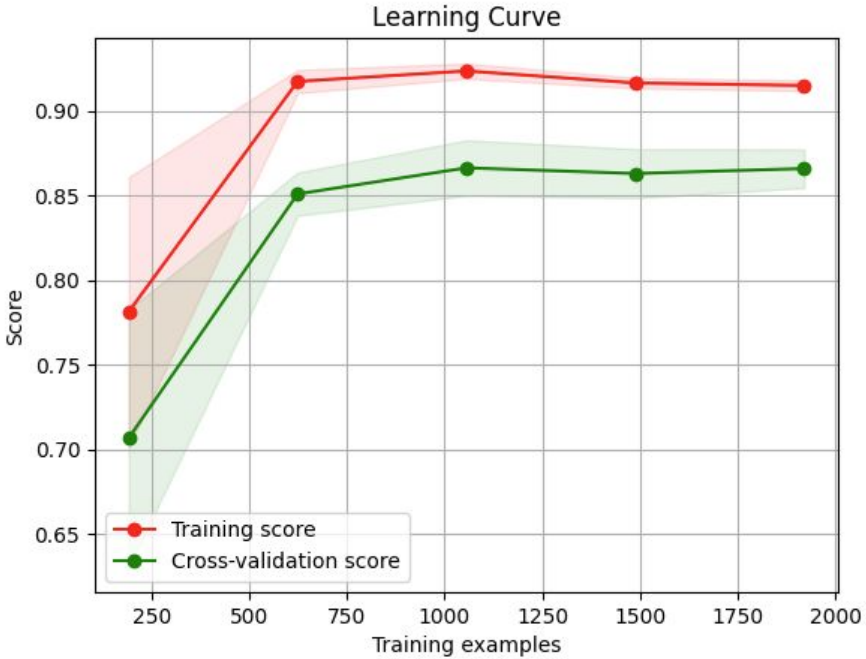
Hiperparámetros

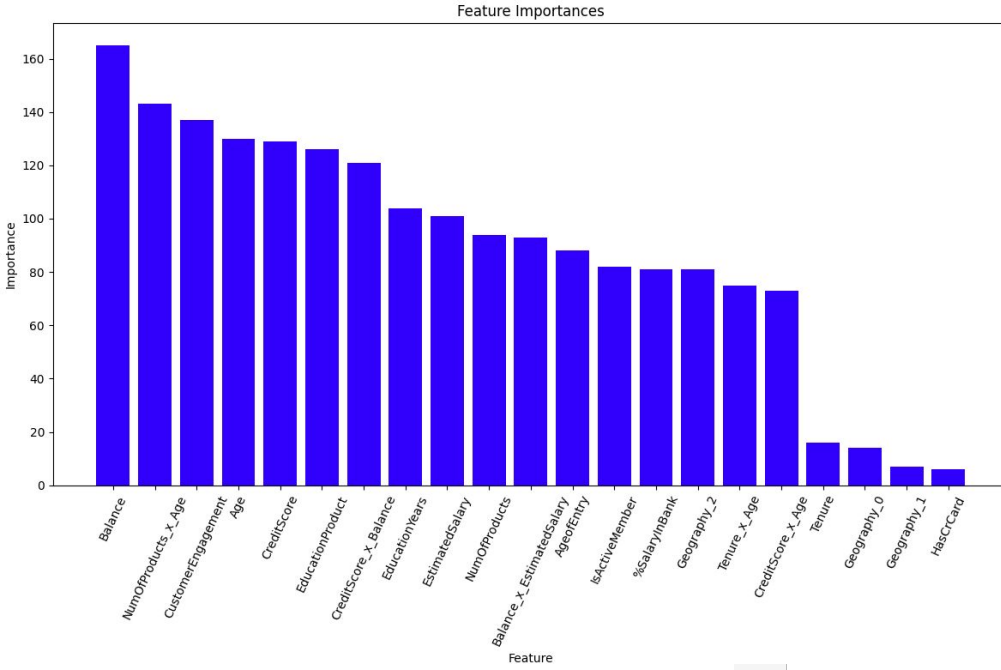
Nombre del parámetro	Valor del parámetro
boosting_type	dart
colsample_bytree	1.0
learning_rate	0.1
max_depth	10
min_child_samples	20
min_child_weight	0.001
min_split_gain	0.0
n_estimators	100
num_leaves	20
objective	binary
reg_alpha	0.0
reg_lambda	0.0
metric	binary_logloss
feature_freaction	0.8
bagging_fraction	0.6
bagging_freq	10
lambda_l1	1
lambda_l2	1
min_data_in_leaf	20
min_gain_to_split	0.1

Análisis


Confusion Matrix
1832
73
218
277
Actual
Prediction
0
1
1800
1600
1400
1200
1000
800
600
400
200


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.76)


Learning Curve
Score
0.90
0.85
0.80
0.75
0.70
0.65
Training score
Cross-validation score
250
500
750
1000
1250
1500
1750
2000
Training examples


Feature Importances
Importance
160
140
120
100
80
60
40
20
0
Balance
NumOfProducts_x_Age
CustomerEngagement
Age
CreditScore
EducationProduct
CreditScore_x_Balance
EducationYears
EstimatedSalary
NumOfProducts
Balance_x_EstimatedSalary
AgeofEntry
IsActiveMember
%SalaryInBank
Geography_2
Tenure_x_Age
CreditScore_x_Age
Tenure
Geography_0
Geography_1
HasCrCard
Feature

Espacios de mejora y limitaciones

Espacios de mejora y limitaciones

Limitaciones

- Muy pocos registros
- Poca cantidad de variables

Mejoras

- Mejor organización y registro de cambios en el código
- Usar diferentes encoders
- Usar técnicas de reducción de dimensionalidad para vectorización de apellidos y PolynomialFeatures
- Mejor optimización de hiperparámetros
- Mejor forma de imputación de nulos
- Usar otras técnicas de Data Augmentation

ITBA
Instituto Tecnológico de Buenos Aires

¡MUCHAS GRACIAS!

MÁS INFORMACIÓN > **www.itba.edu.ar**