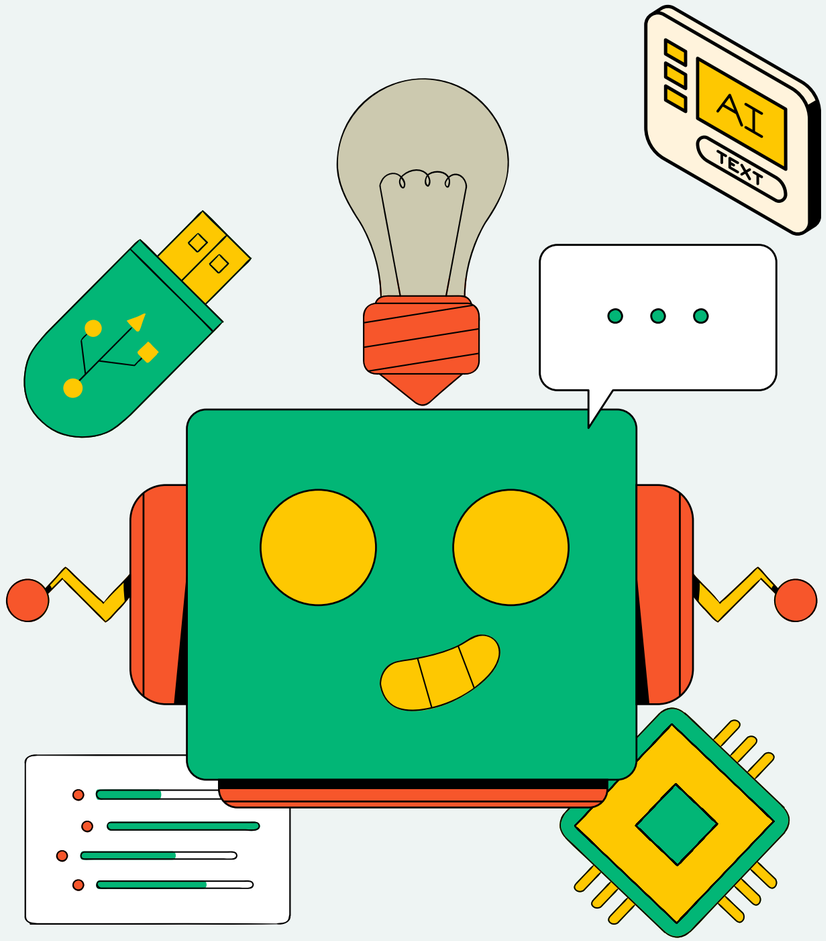


MODELOS DE PREDICCIÓN DE FRAUDE CREDITICIO

JUAN CARLOS AGUILAR ALFARO

MANFRED D. PORRAS ROJAS

HOLMAR RIVERA

DOMINICK RODRÍGUEZ TREJOS

INTRODUCCIÓN

En la era digital, con la mayoría de las entidades financieras ofreciendo servicios digitales y un incremento en las transacciones en línea, el fraude crediticio ha encontrado nuevas vías para expandirse.

- Detección: habilidad de reconocer fraude
- Prevención: medidas o acciones que se pueden tomar para evitar el fraude.

¿QUE ES MACHINE LEARNING?

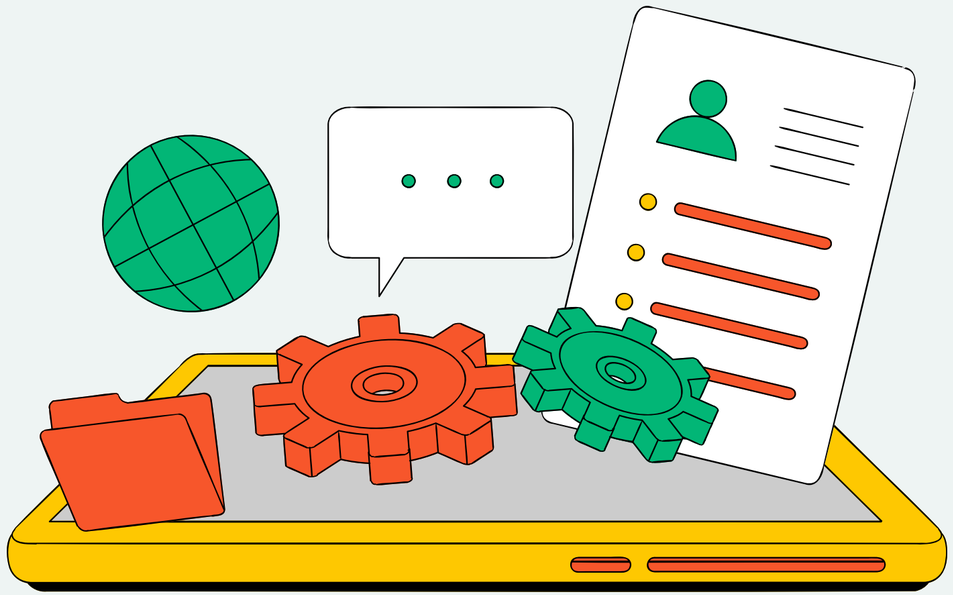
El objetivo principal del machine learning es entender cómo se estructuran los datos y asociarlos a modelos que puedan ser interpretados y utilizados por humanos. Esto implica crear algoritmos que puedan identificar patrones dentro de grandes conjuntos de datos, permitiendo así a las máquinas aprender y tomar decisiones basadas en la información obtenida.

TIPOS DE MACHINE LEARNING

Aprendizaje supervisado

Aprendizaje no supervisado

Aprendizaje de reforzamiento



PROCESO DE CREACION DE UN MODELO

PREPARACIÓN DE LOS DATOS

- Recolección, limpieza y preprocesamiento de datos
- Asegurar la calidad y confiabilidad de los datos

DESARROLLO DEL MODELO

- Selección de algoritmos de aprendizaje automático
- Proceso de entrenamiento

EVALUACIÓN DEL RENDIMIENTO

- Validación cruzada y pruebas en conjuntos de datos de validación
- Evaluación del rendimiento del modelo

PREPARACION DE LA DATA

01

STANDARDSCALER

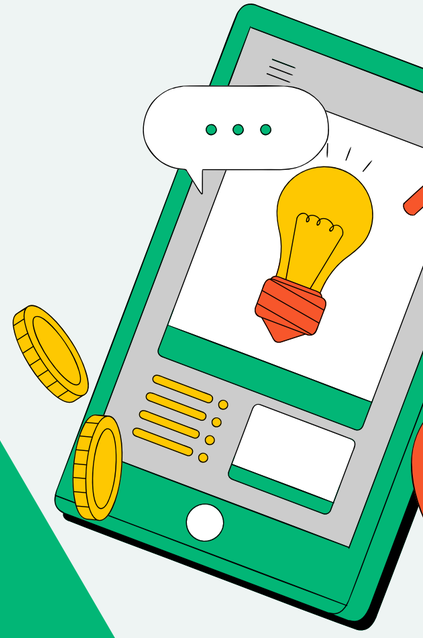
02

ONE-HOT ENCODER

03

PIPELINE

MODELOS USADOS



- Random forest
- Regresión Logística
- Support Vector Classification (SVC)
- XGBoost
- LightGBM

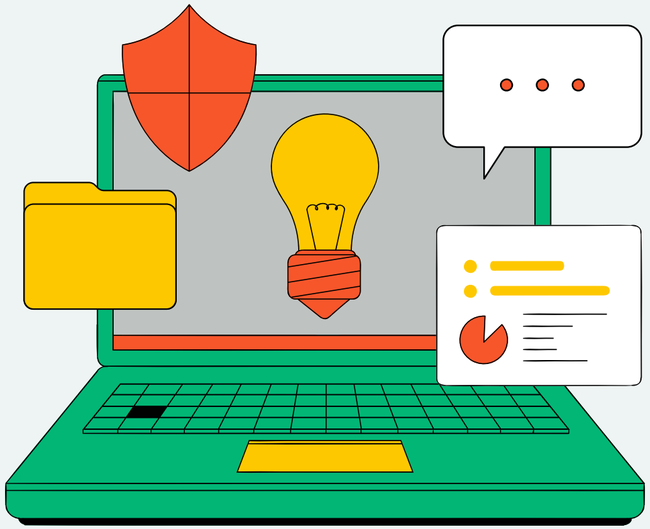
MÉTODOS DE ENSAMBLAJE

Stacking

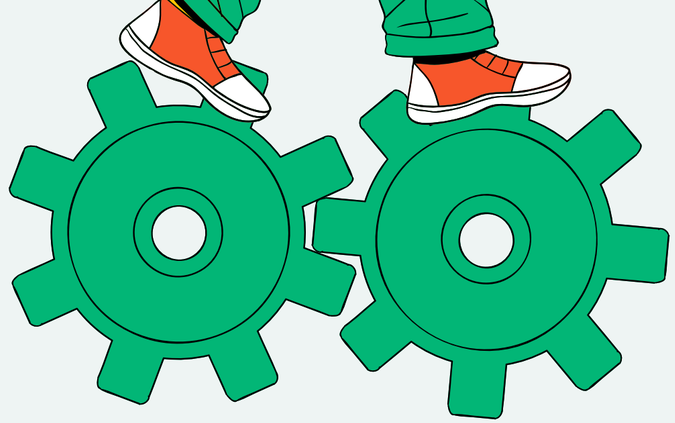
Funciona en dos capas:
Modelos base
Modelo meta

Voting

Funciona en dos formas:
Hard Voting
Soft Voting



RESULTADOS



HGBOOST

VOTING CLASSIFIER

STACKING CLASSIFIER

¿QUE ES LA VALIDACION CRUZADA ESTRATIFICADA?

- Se divide el conjunto de datos en k folds
- Mantiene la proporción de las clases
- El modelo se entrena con k-1 pliegues y se valida en el pliegue restante
- Se promedian los resultados

ROC AUC

Métrica para evaluar el redimiendo de un algoritmo de clasificación binario

True Positive Rate (**TPR**) is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (**FPR**) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

Modelo XGBoost:

- Conjunto de prueba: 0.999124
- Validación cruzada estratificada: 0.99739

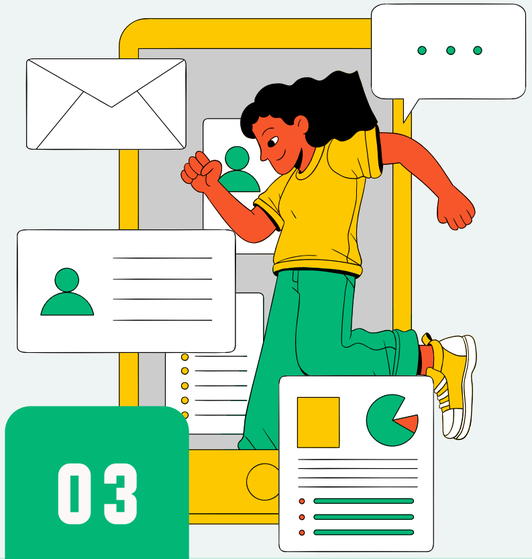
Modelo VotingClassifier:

- Conjunto de prueba: 0.99867
- Validación cruzada estratificada: 0.99629

Modelo Stacking:

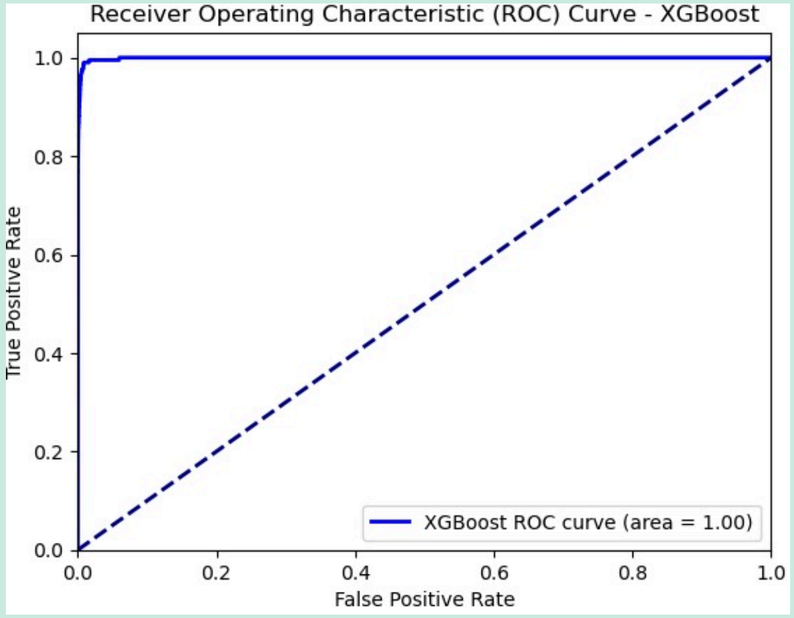
- Conjunto de prueba: 0.99
- Validación cruzada estratificada:

GRAFICOS DE RESULTADOS



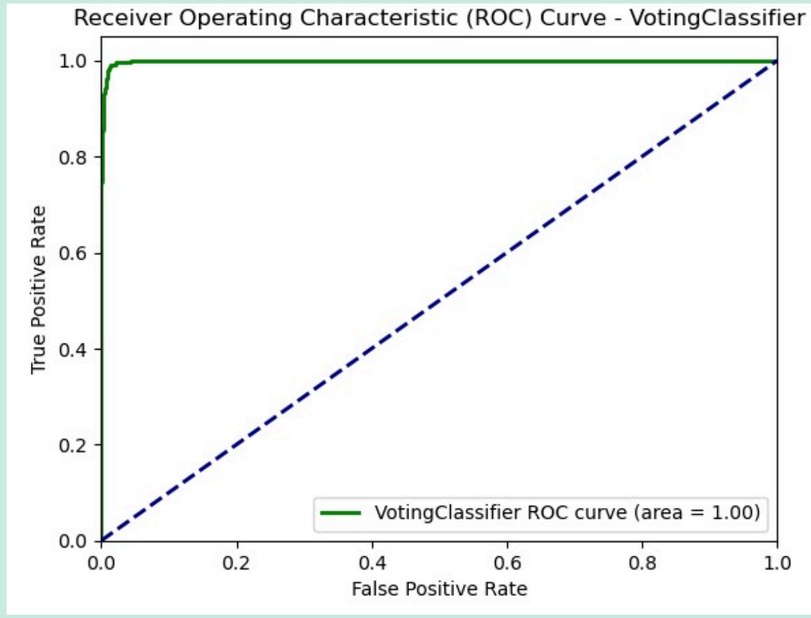
01

HGBOOST



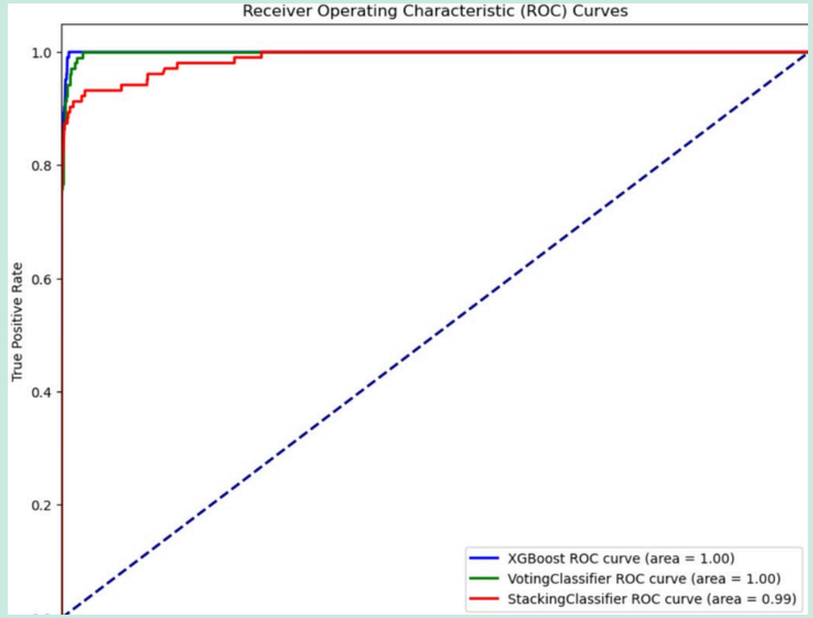
02

VOTING-CLASSIFIER



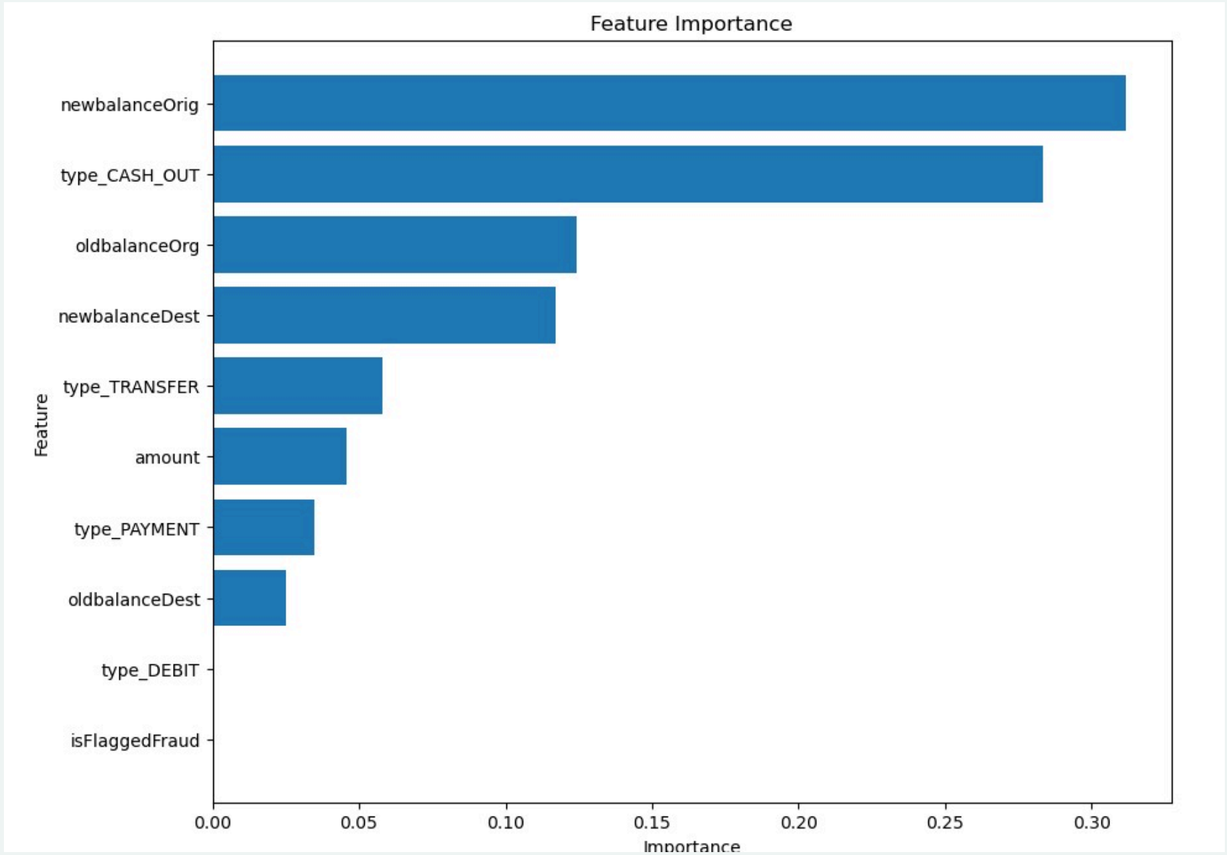
03

STACKING



XGBOOST

- Fue el algoritmo con mejores resultados
- Algoritmo de boosting
- Rápido y eficiente



MEJORES PARAMETROS:

- TASA DE APRENDIZAJE: 0.1
- MAXIMA PROFUNDIDAD: 5
- NUMERO DE ARBOLES: 100
- MUESTRA: 100%

CONCLUSIONES

El modelo XGBoost mostro los mejores resultados

A pesar de implementar modelos de ensamblaje, estos no dieron el rendimiento esperado

La variable oldbalanceOrg es la que brinda mayor informacion. Seguida de Amount

Los tres modelos dieron resultados excelentes

RECOMENDACIONES

- Manejar la información sensible y personal de manera cuidadosa
- Optimizar los hiperparametros de todos los modelos
- Tomar en cuenta la potencia de calculo
- Se podrían considerar otros modelos para los ensamblajes

¿PREGUNTAS?