

# PREDICCIÓN DE TRANSACCIONES DE CLIENTES DE SANTANDER

---

Proyecto de Data Analytics

Junio 2024



# CONTEXTO

---

El proyecto se centra en predecir si un cliente realizará una transacción específica utilizando un conjunto de datos proporcionado por Santander. Esta predicción es crucial para identificar patrones de comportamiento y mejorar las estrategias de marketing y fidelización.





## INFORMACION DE DATOS

El conjunto de datos proviene de Kaggle y contiene:

- Variables anónimas numeradas de var\_0 a var\_199 (características).
- Variable target (0 o 1), donde 1 indica que el cliente realizó la transacción y 0 que no la realizó (objetivo).

Se utilizó el conjunto de datos train.csv para entrenar y validar el modelo y test.csv para evaluar el rendimiento.

# VISUALIZACIONES

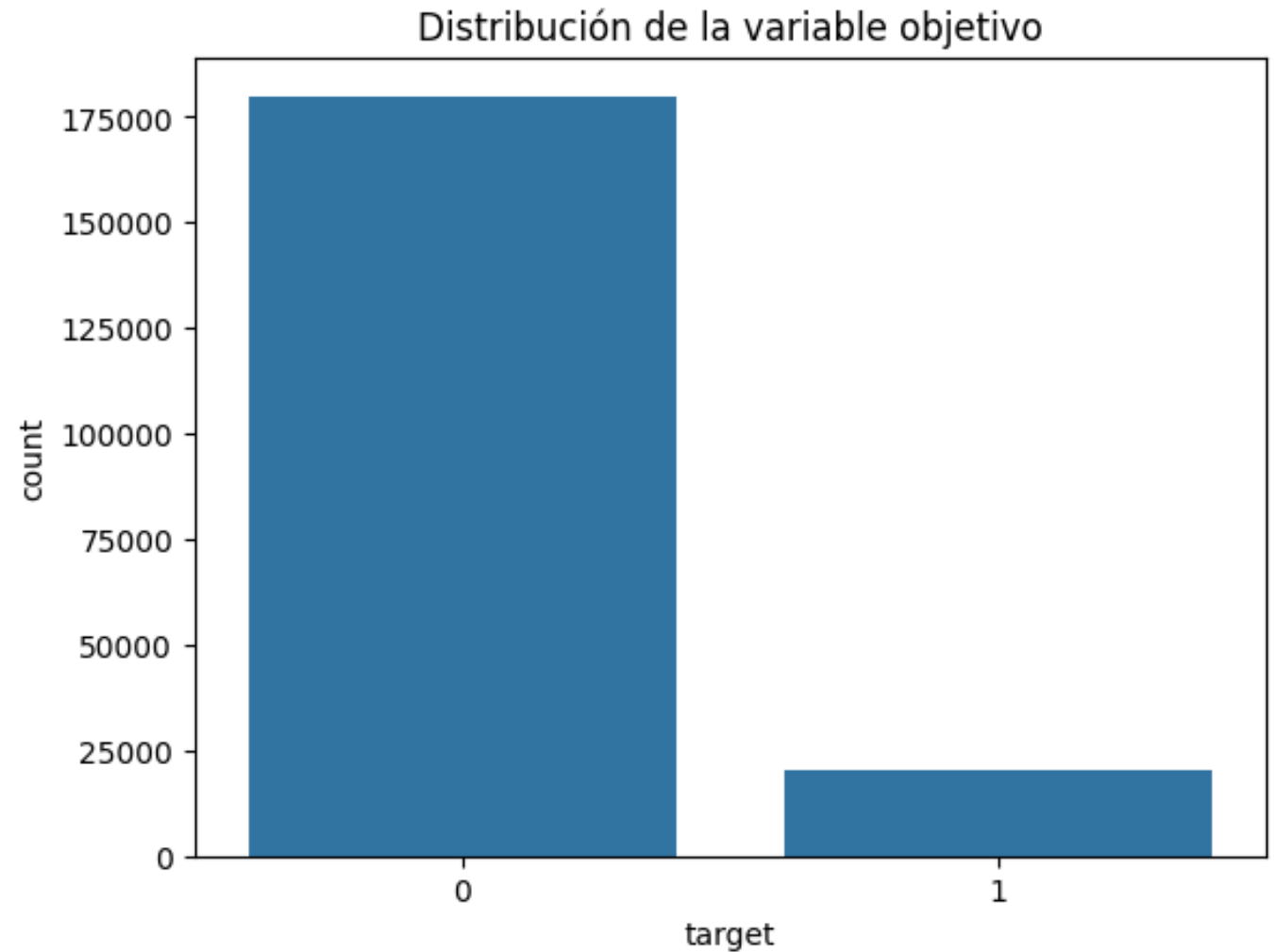
A continuación, se presentan algunas visualizaciones clave y sus interpretaciones:





## DISTRIBUCIÓN DE LA VARIABLE OBJETIVO

La variable objetivo está desbalanceada, con más casos de no transacción (90%) que de transacción (10%).



## MATRIZ DE CORRELACIÓN

No hay correlaciones fuertes entre las características, lo que sugiere que cada variable contribuye de manera única al modelo.

target	1	0.052	0.05	0.056	0.011	0.011	0.031	0.067	-0.003	0.02	-0.043	-0.0022
var_0	0.052	1	-0.00054	0.0066	0.0038	0.0013	0.003	0.007	0.0024	0.005	-0.0026	0.00035
var_1	0.05	-0.00054	1	0.004	1e-05	0.0003	-0.0009	0.0033	0.0015	0.0041	-0.00083	0.0029
var_2	0.056	0.0066	0.004	1	0.001	0.00072	0.0016	0.00088	-0.00099	0.0026	-0.0019	-0.00047
var_3	0.011	0.0038	1e-05	0.001	1	-0.00032	0.0033	-0.00077	0.0025	0.0036	-0.00083	-0.0009
var_4	0.011	0.0013	0.0003	0.00072	-0.00032	1	-0.0014	4.9e-05	0.0045	0.0012	-0.00092	-0.0034
var_5	0.031	0.003	-0.0009	0.0016	0.0033	-0.0014	1	0.0026	-0.00099	0.00015	-0.0053	0.00033
var_6	0.067	0.007	0.0033	0.00088	-0.00077	4.9e-05	0.0026	1	-0.0025	-0.0012	-0.0057	0.0015
var_7	-0.003	0.0024	0.0015	-0.00099	0.0025	0.0045	-0.00099	-0.0025	1	0.00081	0.0029	0.00036
var_8	0.02	0.005	0.0041	0.0026	0.0036	0.0012	0.00015	-0.0012	0.00081	1	-0.0011	-0.00075
var_9	-0.043	-0.0026	-0.00083	-0.0019	-0.00083	-0.00092	-0.0053	-0.0057	0.0029	-0.0011	1	0.0016
var_10	-0.0022	0.00035	0.0029	-0.00047	-0.0009	-0.0034	0.00033	0.0015	0.00036	-0.00075	0.0016	1

# TÉCNICAS DE MODELADO

Se probaron varios algoritmos de clasificación, incluyendo Regresión Logística, Random Forest y Máquina de Vectores de Soporte (SVM).



## Preprocesamiento

- Imputación de datos faltantes con 'SimpleImputer'
- Escalado de características con 'StandardScaler'

## Desbalanceo

- Balancear la clase minoritaria con 'SMOTE'

## Evaluación

- División de datos (70% train y 30% test)
- Métricas (Precisión, recall y matriz de confusión)

# RESULTADOS

Se probaron varios algoritmos de clasificación, incluyendo Regresión Logística, Random Forest y Máquina de Vectores de Soporte (SVM).

## Regresión Logística

```
Resultados de Logistic Regression:
[[42759 11239]
 [10564 43380]]
      precision    recall  f1-score   support

     0       0.80      0.79      0.80     53998
     1       0.79      0.80      0.80     53944

 accuracy          0.80     107942
 macro avg       0.80      0.80      0.80     107942
 weighted avg    0.80      0.80      0.80     107942

Accuracy: 0.7980118952770933
```

## Random Forest

```
Resultados de Random Forest:
[[53152   846]
 [ 3261 50683]]
      precision    recall  f1-score   support

     0       0.94      0.98      0.96     53998
     1       0.98      0.94      0.96     53944

 accuracy          0.96     107942
 macro avg       0.96      0.96      0.96     107942
 weighted avg    0.96      0.96      0.96     107942

Accuracy: 0.9619517889236813
```

## Support Vector Machine (SVM)

```
Resultados de Support Vector Machine:
[[51315  2683]
 [ 2890 51054]]
      precision    recall  f1-score   support

     0       0.95      0.95      0.95     53998
     1       0.95      0.95      0.95     53944

 accuracy          0.95     107942
 macro avg       0.95      0.95      0.95     107942
 weighted avg    0.95      0.95      0.95     107942

Accuracy: 0.9483704211521002
```



# CONCLUSIONES Y RECOMENDACIONES

---

- ❑ La predicción de transacciones es factible con un alto grado de precisión.
  - ❑ Random Forest ofrece la mejor precisión (96.2%) y balance entre precisión y recall.
  - ❑ El desbalanceo de clases es un reto significativo que se puede mitigar con técnicas como SMOTE.
  - ❑ Los modelos de machine learning pueden ayudar a identificar patrones complejos que no son evidentes a simple vista.
- 
- ❖ Implementar el modelo en producción para monitorear y ajustar continuamente su rendimiento.
  - ❖ Realizar un análisis más profundo de las variables más influyentes para diseñar estrategias de marketing personalizadas.
  - ❖ Continuar explorando técnicas avanzadas como el ensamble de modelos para mejorar la precisión y robustez del modelo.

# CONTACTAME

---

**Nombre:** Martín Sotelo

**Correo:** [martinsotelo2569@gmail.com](mailto:martinsotelo2569@gmail.com)

**LinkedIn:**

<https://www.linkedin.com/in/mart%C3%A1nSoteloloarte/>



