Churn Prediction

Machine Learning Technical Presentation

Nathan Jones

Contents

Problema de *Clasificación Binario*: Churn vs No Churn

Datos

- Customer Churn Dataset: 500,000 filas
- 10 Variables
- 350,000 Train
- 150,000 Test
- 55% Churn

Datos Limpios, Predicciones Confiables

Variables Categoricas

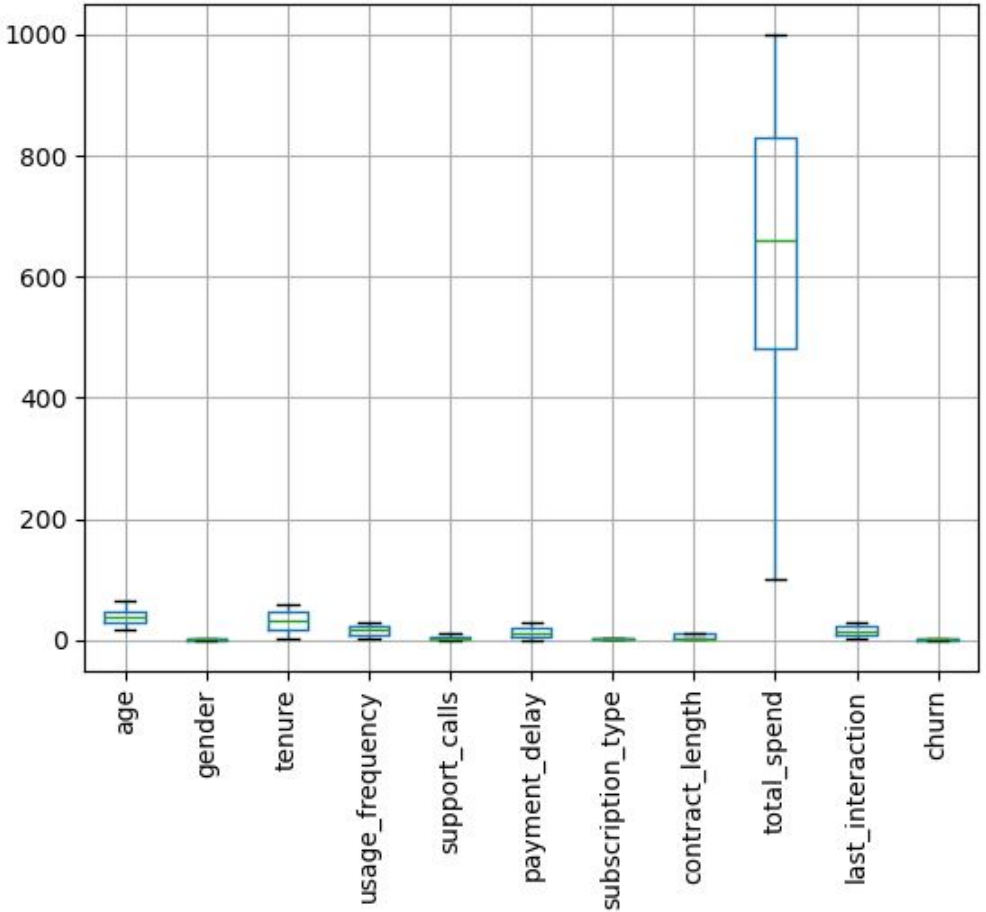
- Gender
- Subscription Type: Basic, Standard, Premium
- Contract Length: Monthly, Quarterly, Annual

Variables Numericas

- **Persona**: age, gender,
- **Tiempo**: tenure, usage_frequency, last_interaction
- **Problemas**: support_calls,
- **Contrato**: subscription_type [*encoded*], contract_length [*encoding*].
- **Pago**: payment_delay, total_spend,
- **Clasificador**: churn

Varianza

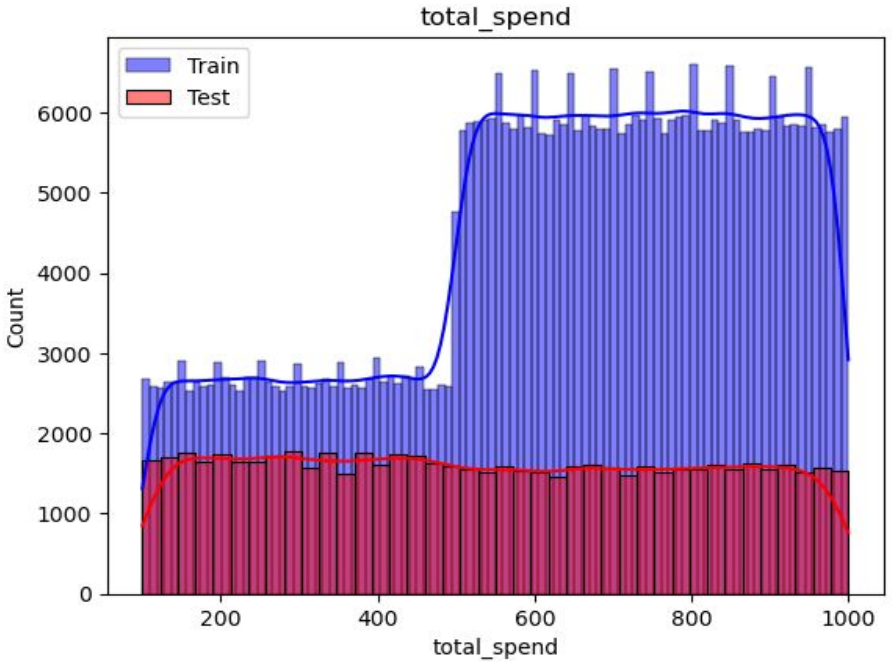
Escalar datos en fase de preprocesamiento para modelos excepto Decision Classifiers.



Problema de Representatividad

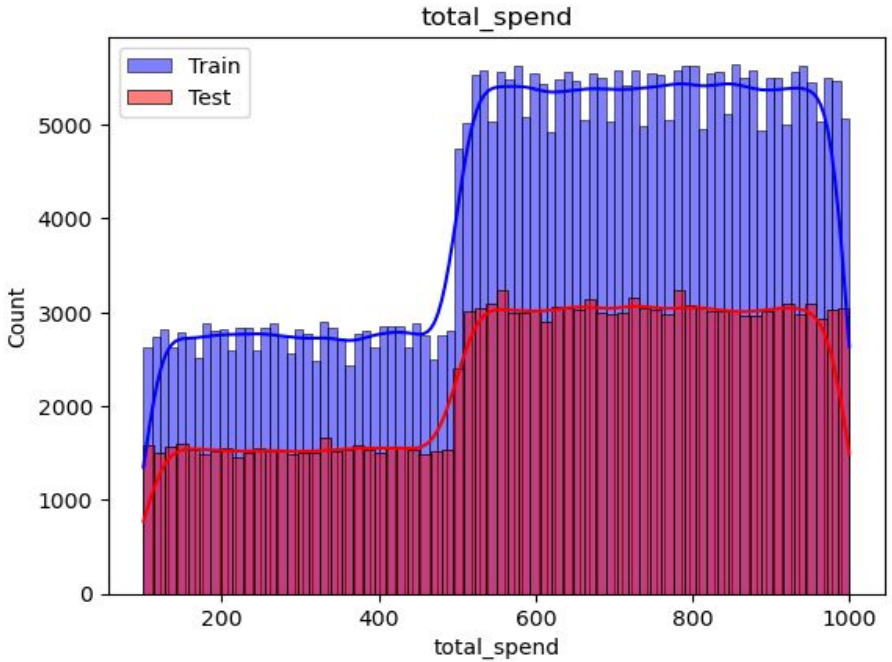
Sesgos presentes en train y no en test:

- total_spend
- payment_delay
- support_calls
- last_interaction



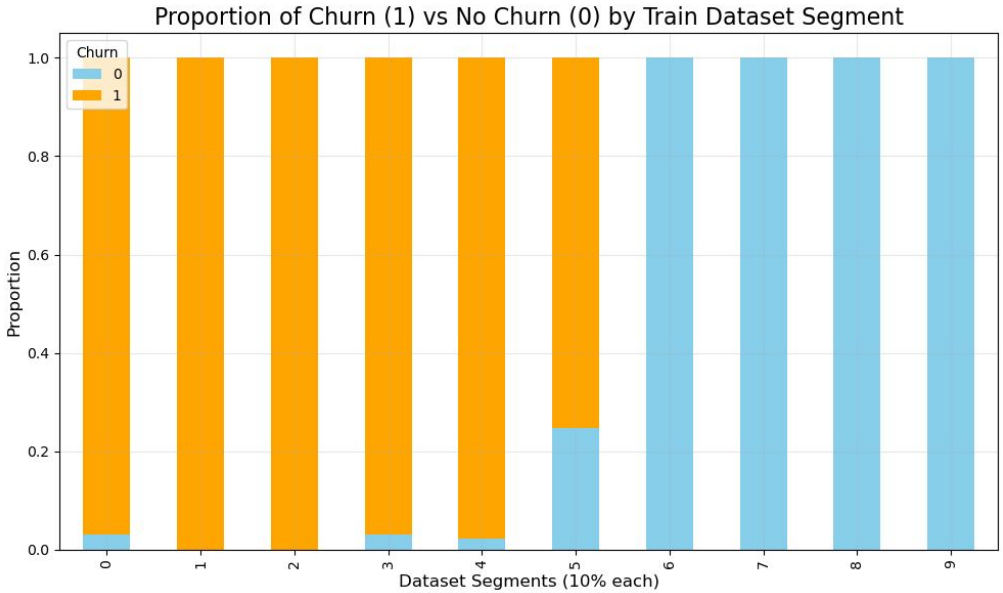
Solución de Representatividad

Concatenación y Estratificación



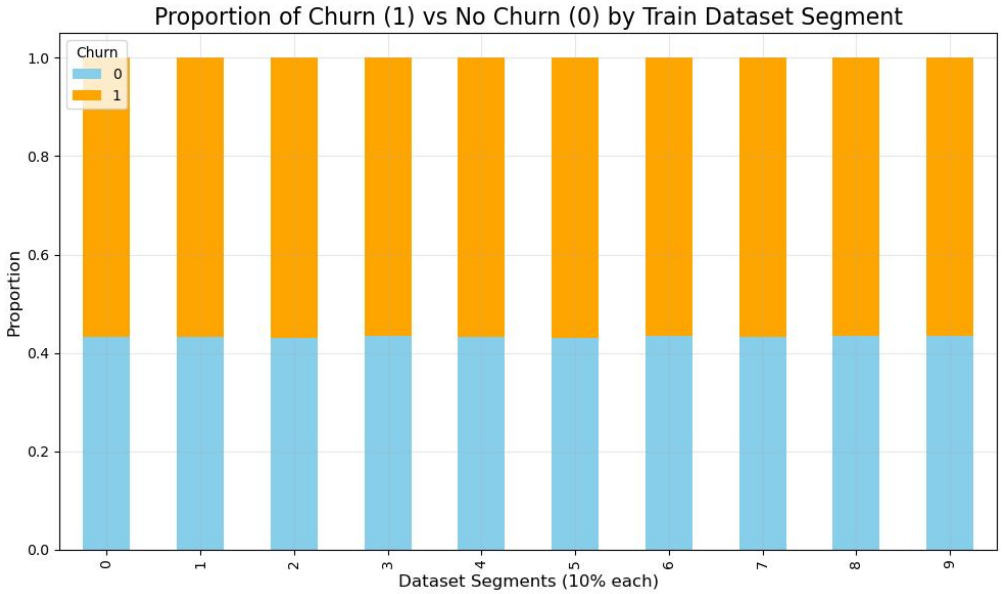
Problema: Distribución Desequilibrado

Churn (1) se ve desproporcionadamente representado al principio de dataset, introduciendo sesgos al futuro modelo.



Solución para Distribución Desequilibrado

Aplicando shuffle (sklearn.utils), para equilibrar distribución.



Desarrollo Iterativo Incremental

Misma Metología, Distintos Modelos

Sklearn, Keras Libraries

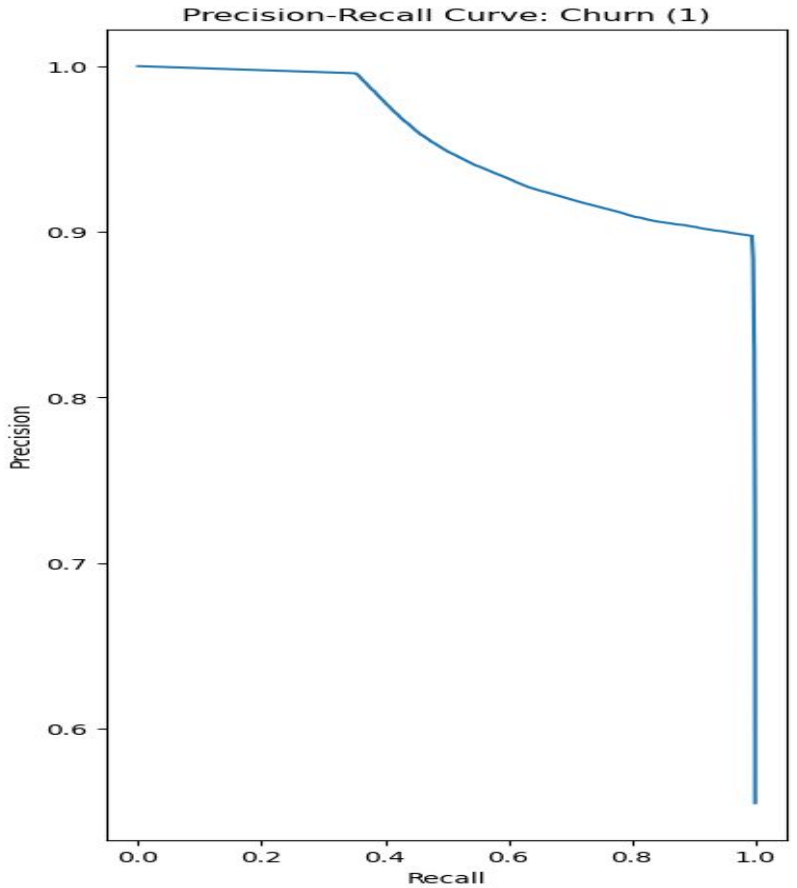
- Logistic Regression Classifier
- **Decision Tree Classifier**
- Random Forest Classifier
- K Neighbours Classifier
- Support Vector Classifier con Principal Component Analysis
- Neural Network Sequential Model

Decision Tree Precisión/Recall

Modelo Optimizado: 99% recall

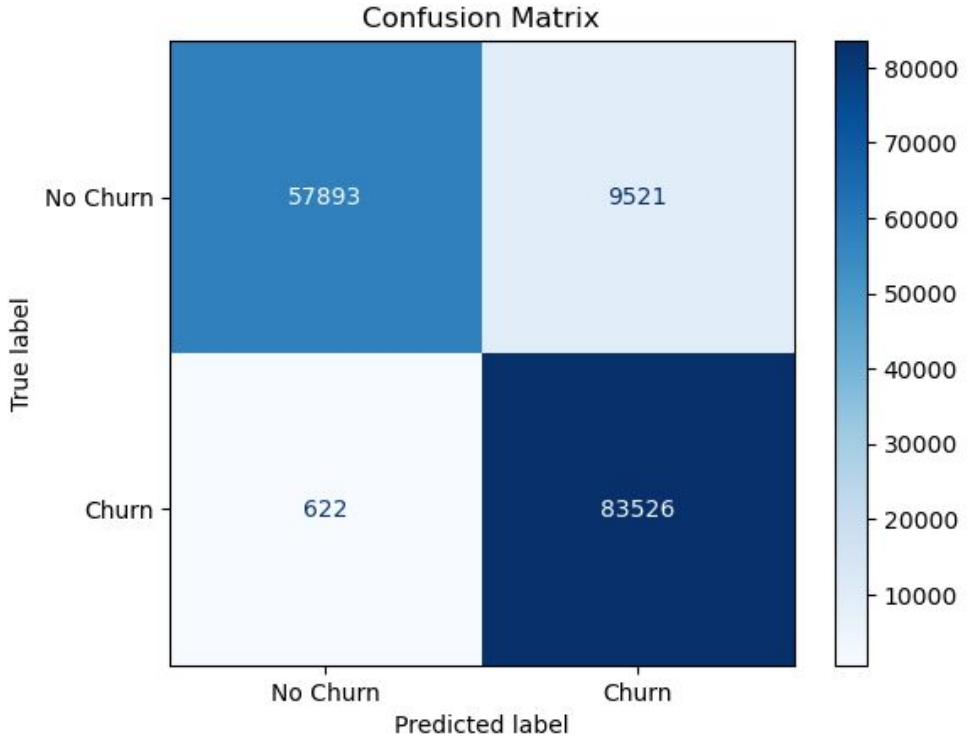
13% mejora en recall (class 1, churn).

- estimator__min_samples_split: 2
- estimator__min_samples_leaf: 5
- estimator__max_features: None
- estimator__max_depth: 10
- (class_weights='balanced')

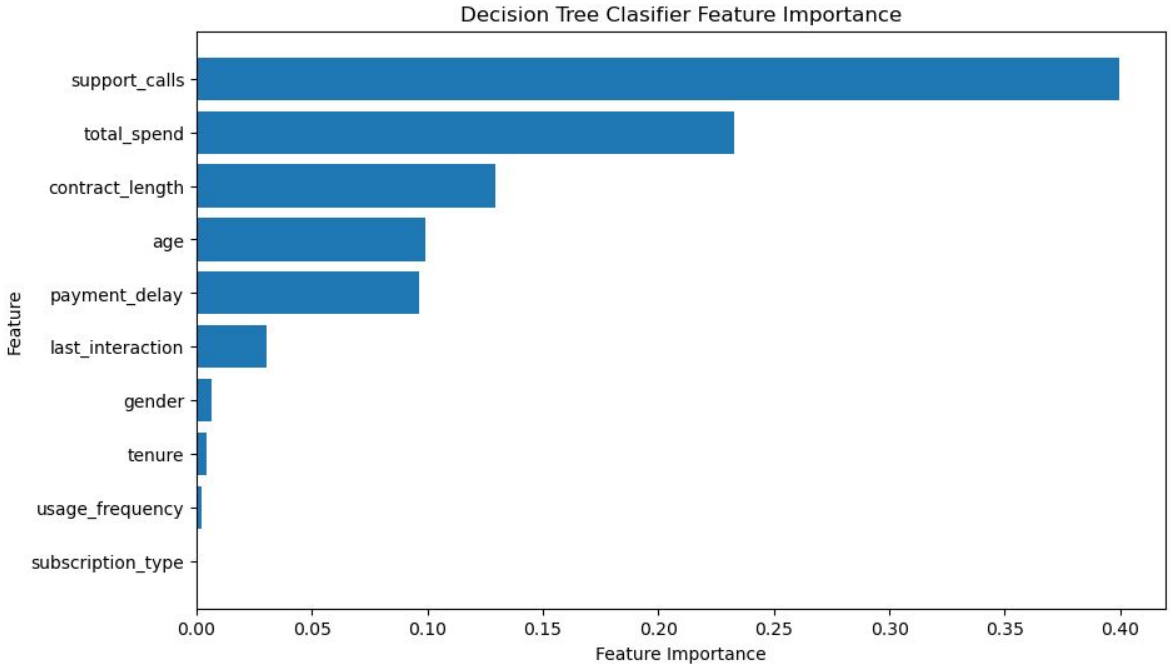


Decision Tree Confusion Matrix

Valores reales para contextualizar el recall alto 99% y precisión 90% respeto al Churn (clase 1).



Feature Importance



Optimizar hoy, liderar mañana

Asunciónes Claves

1. Es más barato y rentable *retener* clientes existentes que atraer nuevos.
2. Es manageable asumir los costes adicionales de mantener los 10% de clientes erróneamente predichos ‘Churn’.

A lo contrario, posible alteración:

- Mayor umbral para favorecer precisión y reducir falsos positivos.