Customer Classification with XGBoost

Joan Fernández Navarro

STATISTICAL LEARNING FOR DATA SCIENCE

Contents

1. Introduction
2. Data Preprocessing
3. Train-Test Split
4. Feature Importance Analysis
5. Final Model & Results
6. Conclusion

Introduction

- **Problem**: Customer classification for business optimization
- **Dataset**: 116,934 customers with 196 features
- **Target**: Binary classification based on KPI "BMA_corregido"
- **Objective**: Identify profitable customers to maximize business benefit

Business Goal

Maximize business profit

by selecting customers with positive KPI values

Dataset Overview

Characteristic	**Value**
Total customers	116,934
Features	196
After preprocessing	51,618
Target variable	"true_class"
KPI threshold	0.0

- **Target definition**: 0 (selected) if KPI > 0, otherwise 1 (excluded)
- **Missing data**: No missing values after preprocessing

Data Preprocessing

Data Preprocessing

- **Initial cleaning**: Removed irrelevant columns (PrimaTotalPoliza, ComisionTotalPoliza, etc.)
- **SINCO filtering**: Kept only customers with SINCO data (51,618 rows)
- **Missing values**: No missing data after preprocessing
- **Feature selection**: Numerical features only, excluding target and KPI

Preprocessing Results

- Clean dataset: 51,618 customers × 195 features
- Target variable: Binary classification based on KPI threshold
- Ready for model training and evaluation

KPI Analysis

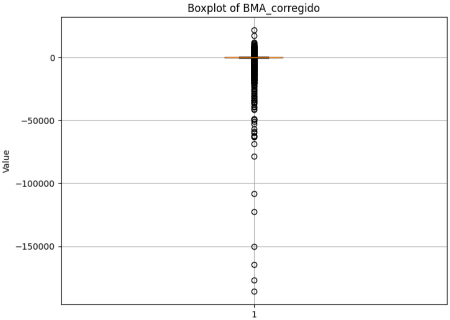

Figure 1: Distribution of BMA_corregido

Statistic	Value
Mean	17.63
Std	2,277.25
Min	-185,990.40
Max	21,746.52

- **High variance** in KPI values
- **Right-skewed** distribution
- Some customers with **significant losses**

Data Splitting Strategy

- **Test size**: 30% of original data (15,486 customers)
- **Validation size**: 10% of training data (3,614 customers)
- **Training size**: 32,518 customers
- **Random state**: 42 for reproducibility

Dataset	Rows	Features
X_train	32,518	194
X_val	3,614	194
X_test	15,486	194
Total	51,618	194

Feature Importance Analysis

Initial Model

- **Model**: XGBoost Classifier with 100 estimators
- **Evaluation metric**: ROC-AUC score
- **Feature selection**: Top 5% most important features
- **Training time**: 2.10 seconds

Initial Model Performance

- **Train AUC**: 0.958 (potential overfitting)
- **Validation AUC**: 0.715
- **Test AUC**: 0.708

Most Important Features

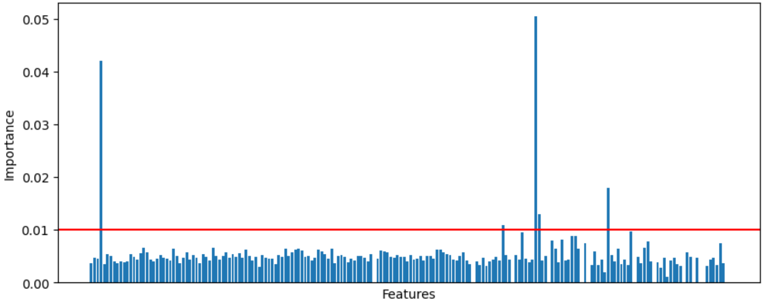

Figure 2: Feature importance distribution

Top 9 Most Important Features:

1. Result_siniestros_SINCO
2. id55
3. Cliente_Diverso
4. AnyoPoliza
5. PerteneceSINCO
6. id13_H
7. FrecuenciaSiniestroSINCO
8. NumeroDanyosMaterialesSINC
9. id72_2

Final Model & Results

Final Model Configuration

- **Features**: Only top 9 most important features
- **Model**: XGBoost Classifier (same hyperparameters)
- **Training time**: $<$1 second
- **Performance**: Maintained with 95% fewer features

Final Model Performance

- **Train AUC**: 0.796 (reduced overfitting)
- **Validation AUC**: 0.717
- **Test AUC**: 0.709

Business Impact Analysis

Metric	Before Model	After Model
Total customers	15,486	15,486
Max possible benefit	3,674,846€	3,674,846€
Actual benefit	406,006€	1,227,273€
Benefit percentage	11%	33%
Avg. benefit per customer	26.22€	79.25€

Key Improvements

- **202% increase** in total benefit
- **3x improvement** in average benefit per customer
- **22 percentage point increase** in benefit capture

Customer Selection Analysis

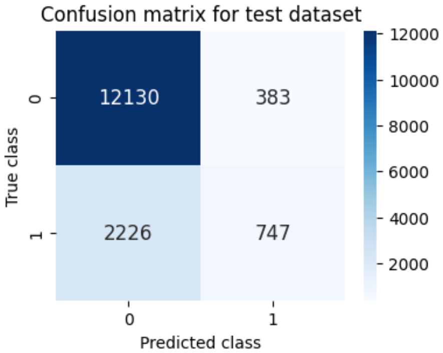

Figure 3: Confusion matrix for test dataset

Selection Results:

- **Selected customers**: 14,356
- **Excluded customers**: 1,130
- **Correct selections**: 12,130 (84.5%)
- **Incorrect selections**: 2,226 (15.5%)
- **Missed opportunities**: 383 customers

Efficiency

Model successfully identifies 97% of profitable customers while excluding most unprofitable ones.

Customer Segmentation Statistics

Statistic	Selected	Excluded
Count	14,356	1,130
Mean KPI	85€	-727€
Std KPI	1,990€	4,379€
Min KPI	-176,973€	-122,770€
Q1 KPI	93€	-664€
Median KPI	183€	-184€
Q3 KPI	267€	129€
Max KPI	11,774€	2,947€

Segmentation Insights

- Model effectively separates profitable vs unprofitable customers
- Excluded group shows significantly negative average KPI
- Selected group maintains positive KPI across all quartiles

Conclusion

- **Feature Reduction**: 95% feature reduction maintained model performance
- **Business Impact**: 3x improvement in average benefit per customer
- **Model Efficiency**: Fast training (<1s) with only 9 features
- **Selection Accuracy**: 84.5% correct selection rate
- **Scalability**: Lightweight model suitable for production deployment

Thank You!

Questions?