Santander Customer Satisfaction

Which customers are happy customers?

Presenter:
Nishitha BANGALORE NAGENDRA
Julien SUZAT
Lou GODARD
Yue Wu

07/12/2023

CONTENT

Problem Explanation

1. Introduction

Banco Santander is a multinational financial services company. Customer satisfaction is a key to success in banking. Unsatisfied customers don't stick for long and they also don't say anything about their dissatisfaction. Santander group needs to identify unsatisfied customers to remain competitive in the banking sector.

2. Explanation of Business Problem

Based on the customer's information, the machine learning model should be able to predict satisfied and unsatisfied customers.

Problem Explanation

3. Machine Learning Formulation

- Binary classification problem
- Dataset : 370 anonymous features
- TARGET feature : 1 --- unsatisfied customer; 0 --- satisfied customer

4. Dataset Description

- train.csv : 370 features and the TARGET feature.
- test.csv : 370 features(identical to train.csv) and no TARGET feature.
- No information about the features.
- Only the TARGET feature gives information about customer satisfaction.

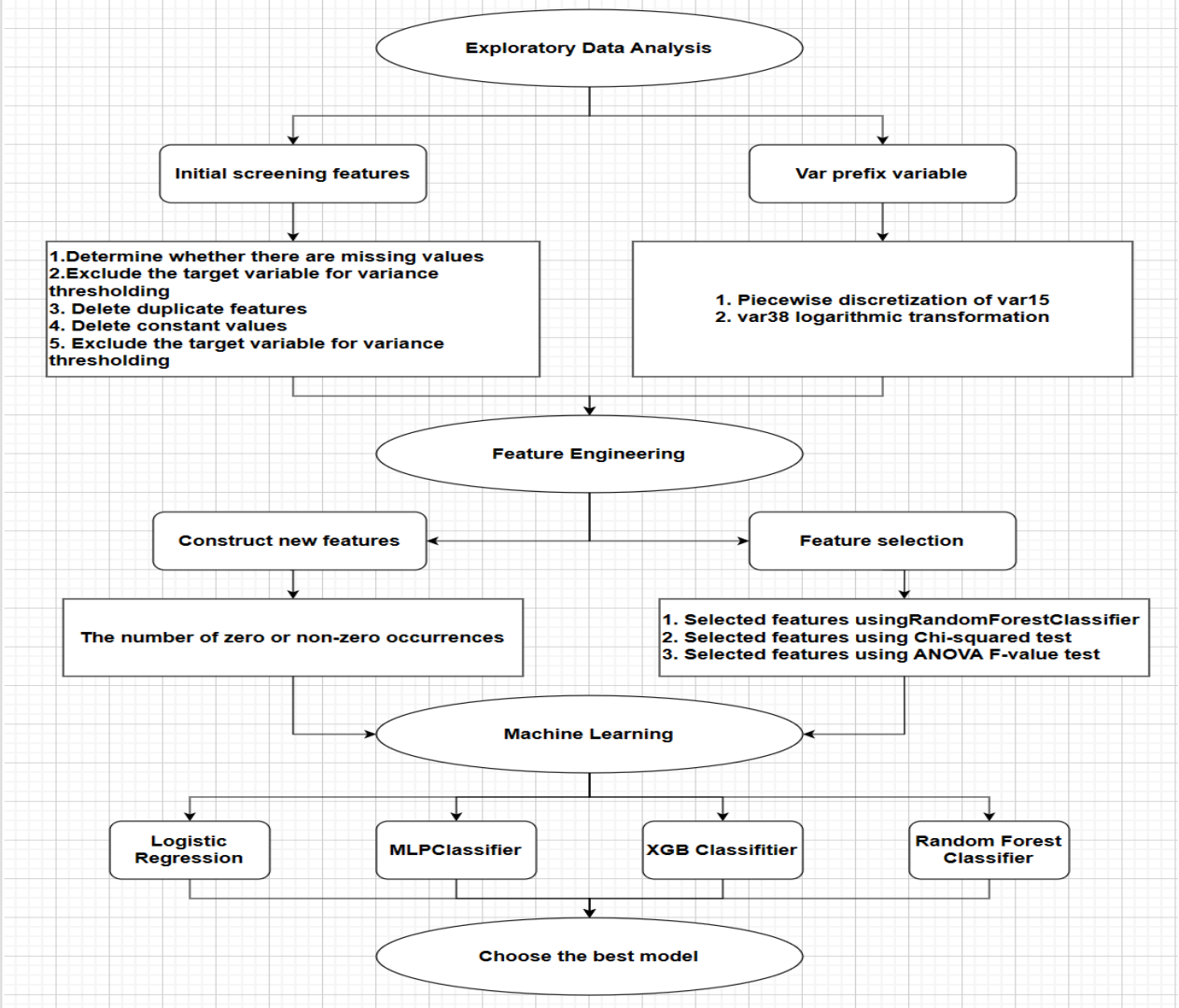
Problem Explanation

5. Performance Metric

The performance metrics that can be used for this problem are AUC. The Santander group in the Kaggle competition used AUC as the metric.

6. Link

https://www.kaggle.com/competitions/santander-customer-satisfaction/overview



Implementation steps

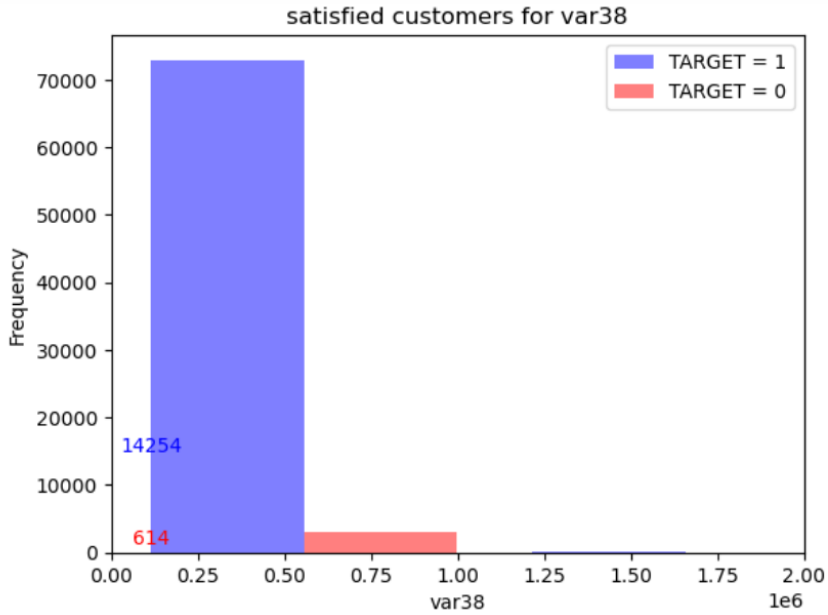
Exploratory Data Analysis

Instead of analyzing all the 369 features, we selected few assumed important features and analyzed a few of them.

The top 3 important features are Var 38, Var 15, Saldo_medio_var5_ult3.

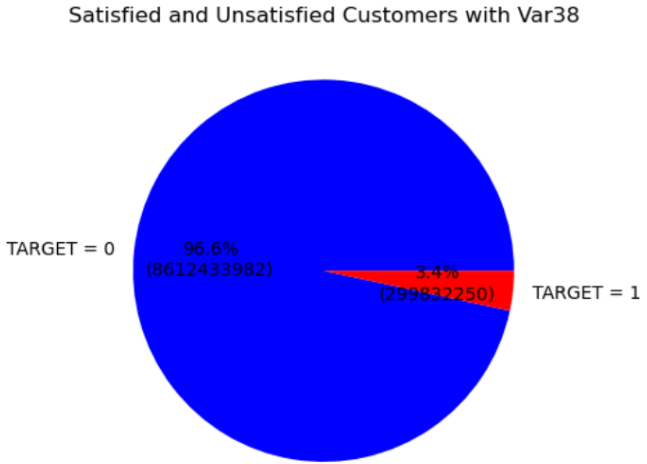
- Var38: We assume that this feature might be the mortgage value of the customers.
- Var15: We assume that this feature might be the age of the customers.
- Saldo_medio_var5_ult3 : The feature is a numerical feature and the value 0 has the highest frequency among all other values.

Exploratory Data Analysis



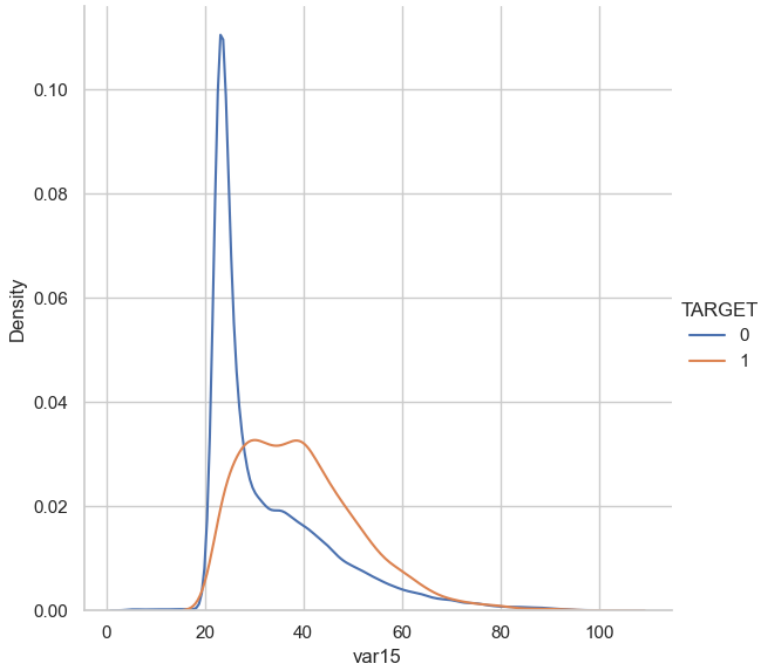
The pie chart suggests that a significant portion of satisfied customers has a 'var38' value around 117310.979016494, contributing to almost the entire sum of 'var38' for this category. This could imply that this particular value is a common characteristic among satisfied customers in the dataset.

Examining the histogram, an intriguing pattern emerges. Initially, there is a higher concentration of satisfied customers within the var38 value range of approximately 12,500 to 55,000. However, as we progress to the var38 values ranging from around 55,000 to 100,000, the scenario shifts, revealing a higher prevalence of dissatisfied customers. Beyond the 100,000 mark, the trend reverts, showcasing a greater number of satisfied customers once again. This implies that, specifically for mortgage values between 55,000 and 100,000, customers tend to express dissatisfaction, while beyond the 100,000 threshold, the proportion of satisfied customers decreases compared to their dissatisfied counterparts.




```
The mode for 'var38' in 'TARGET = 0' is: 117310.979016494
```

Exploratory Data Analysis



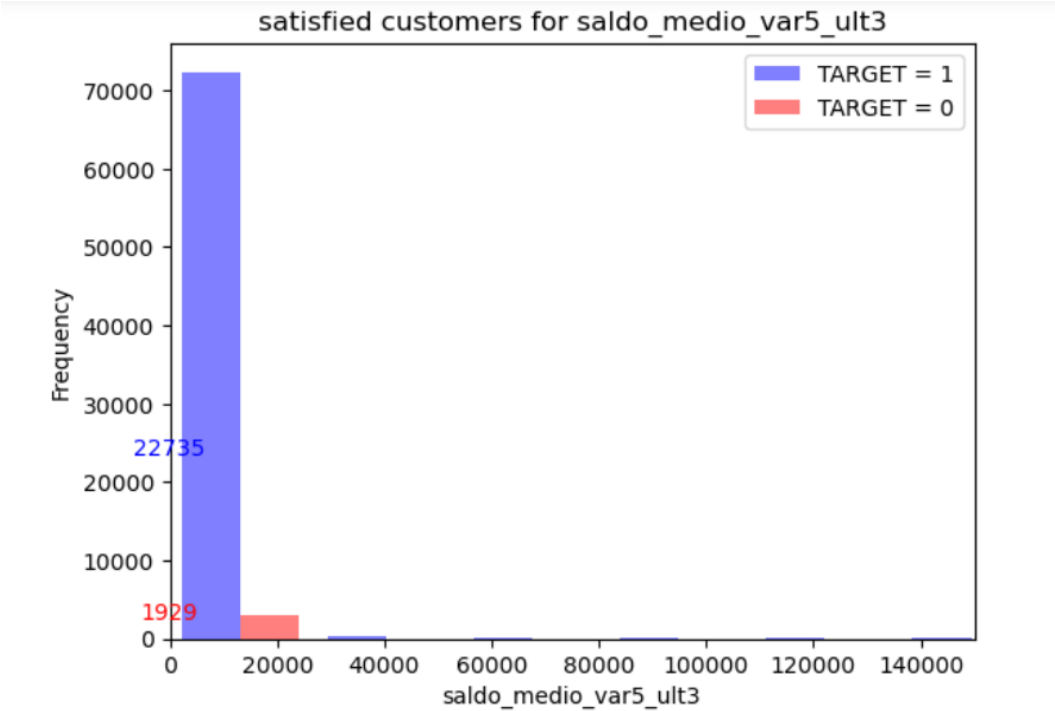
The var15 feature encompasses values within the range of 5 to 105, strongly suggesting that it represents the age of the customers.

Upon closer inspection of the visual representation, a compelling observation arises: nearly all dissatisfied customers possess a var15 (age) value greater than or equal to 23.

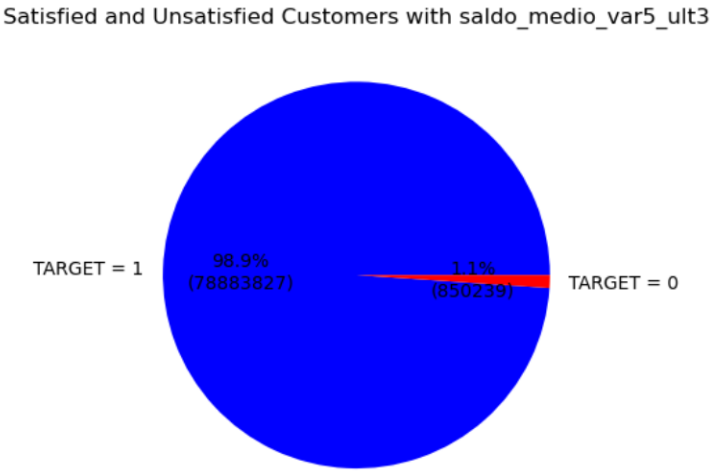
In contrast, the age distribution of satisfied customers spans the entire range from 5 to 105 years, making it challenging to derive any distinct insights.

Remarkably, there are no instances of dissatisfied customers below the age of 23 or above the age of 102. This intriguing pattern prompts the consideration of a new feature – one that identifies customers with an age less than 23, effectively serving as a distinguishing factor between satisfied and dissatisfied customers.

Exploratory Data Analysis



The distributions largely overlap, and a notable observation is that the variable "saldo_medio_var5_ult3" is predominantly 0 for unsatisfied customers. Further analysis is required to compare this with the scenario for satisfied customers where "saldo_medio_var5_ult3" is also 0.



Result (0.0): The most common value for 'saldo_medio_var5_ult3' among unsatisfied customers is 0.0. This could imply that a significant proportion of unsatisfied customers have a zero balance in this particular variable.

```
The mode for 'saldo_medio_var5_ult3' in 'TARGET = 0' is: 0.0
```

Data Preprocessing Overview

```
Entrée [23]: train_df.shape

Out[23]: (76020, 370)
```

In this section of the code, we perform preprocessing operations on our dataset, represented by the train_df DataFrame

- Data Preprocessing Steps:
- Cleaning
- Transformation
- Reduction
- Exploration
- Objective:
- Prepare data for advanced analysis and modeling.

Data Removal (Line 23)

First, we removed a specific column called 'ID' from the DataFrame. This column is often removed because the ID generally doesn't provide useful information for our predictive model

5 rows × 371 columns

```
train_df = train_df.drop ( 'ID' ,  axis  =  1 )
```

```
train_df.shape
```

```
(76020, 370)
```

Data Shape Check (Line 24)

5 rows × 371 columns

```
train_df = train_df.drop ( 'ID' ,  axis  =  1 )
```

```
train_df.shape
```

```
(76020, 370)
```

Next, we checked the current shape of the DataFrame to understand how many rows and columns we have after deleting 'ID'. This gives us an overview of the structure of our data

Data Types Examination

```
# Checking the data type of the dataset variables.

train_df . dtypes . value_counts ()
# All variables in the training data set were classified as being of the numeric data type .
```

```
int64      259
float64    111
dtype: int64
```

We also examined the data types of our variables after deleting 'ID'. This step is crucial as it helps us understand the nature of our variables. Here, we found that we have 259 variables of integer type (int64) and 111 variables of float type (float64). This suggests that all the remaining variables are numeric."

Missing Values Check

```
train_df . isna () . sum () .sum() #There are no null values within the dataset.
```

```
0
```

Finally, we checked our dataset for missing values. The good news is that after all these manipulations, there are no missing values.

This reinforces the quality of our data and allows us to move on to the next stage of the analysis process.

Preprocessing Conclusion

5 rows × 371 columns

```
train_df = train_df.drop ( 'ID' ,  axis  =  1 )
```

```
train_df.shape
```

```
(76020, 370)
```

```
# Checking the number of duplicate records.

train_df . duplicated () . sum ()
```

```
4807
```

```
# Checking the data type of the dataset variables.

train_df . dtypes . value_counts ()
# All variables in the training data set were classified as being of the numeric data type .
```

```
int64      259
float64    111
dtype: int64
```

```
train_df . isna () . sum () .sum() #There are no null values within the dataset.
```

```
0
```

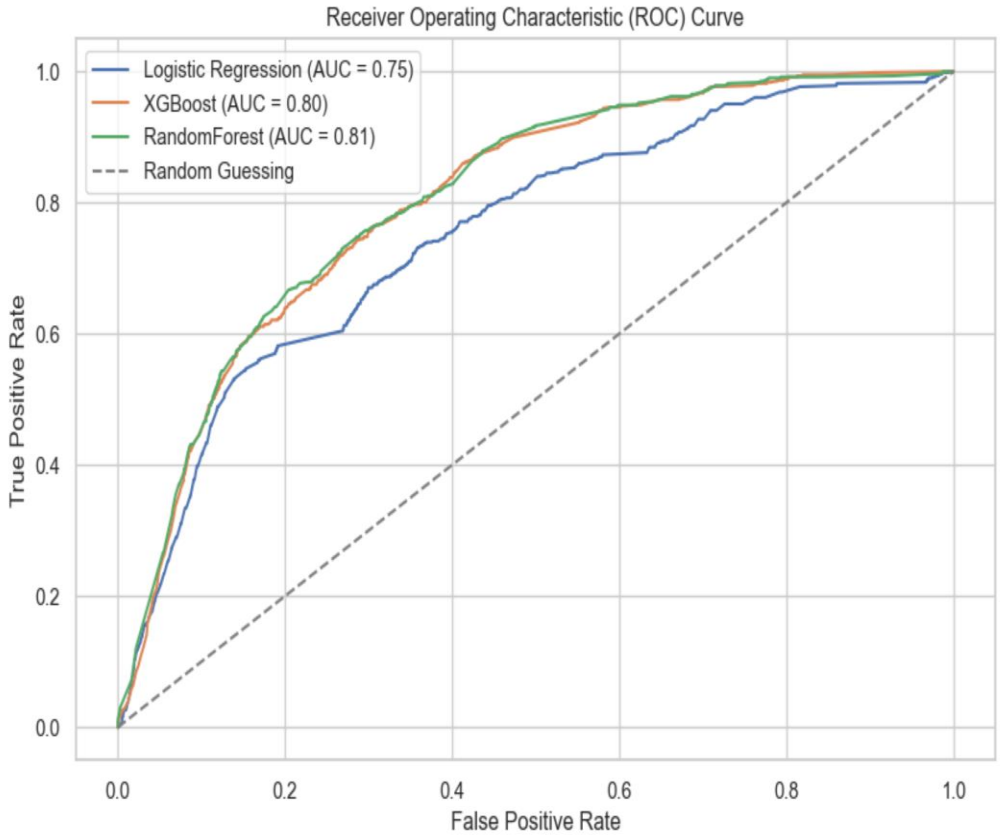
In summary, these lines of code represent essential data preprocessing steps, putting us in a strong position to further explore our variables and build predictive models.

Conclusion

Aurruacy score of four methods

Methods	Train	Test
Logistic Regression	0.772	0.752
MLPC	0.816	0.799
XGBoost	0.816	0.802
Random Forest	0.817	0.806

Conclusion



The use of three integrated models has brought great to AUC, among which the Random Forest performed the best. Followed by XGBoost model, among which the Logistic Regression model performed the worst.

THANK YOU