# PREDICTIVE MODELING OF LOAN DEFAULT: A DATA MINING APPROACH

João Figueiredo, 1230194, MEI
João Araújo, 1200584, MEI
Data Mining, December 2023

Instituto Superior de Engenharia do Porto

# TABLE OF CONTENTS

# INTRODUCTION

## Context

- Bank loans
- Loan Defaulters

## Motivation

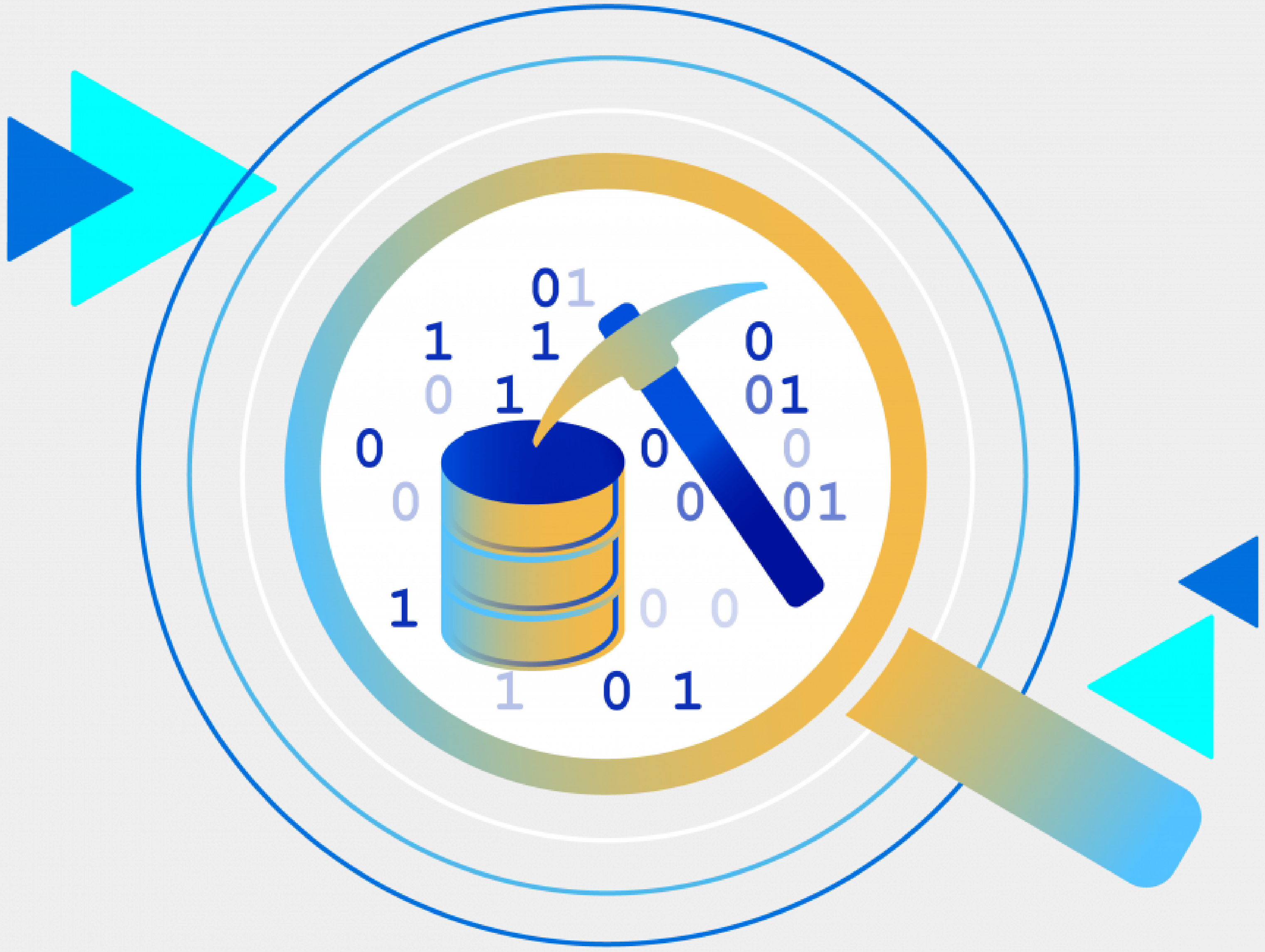
- Financial Consequences of Loan Defaults
- Importance of Analysis

## Goals

- Development of predictive models to determine if bank clients will default on loans.

# CRISP-DM METHODOLGY

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation

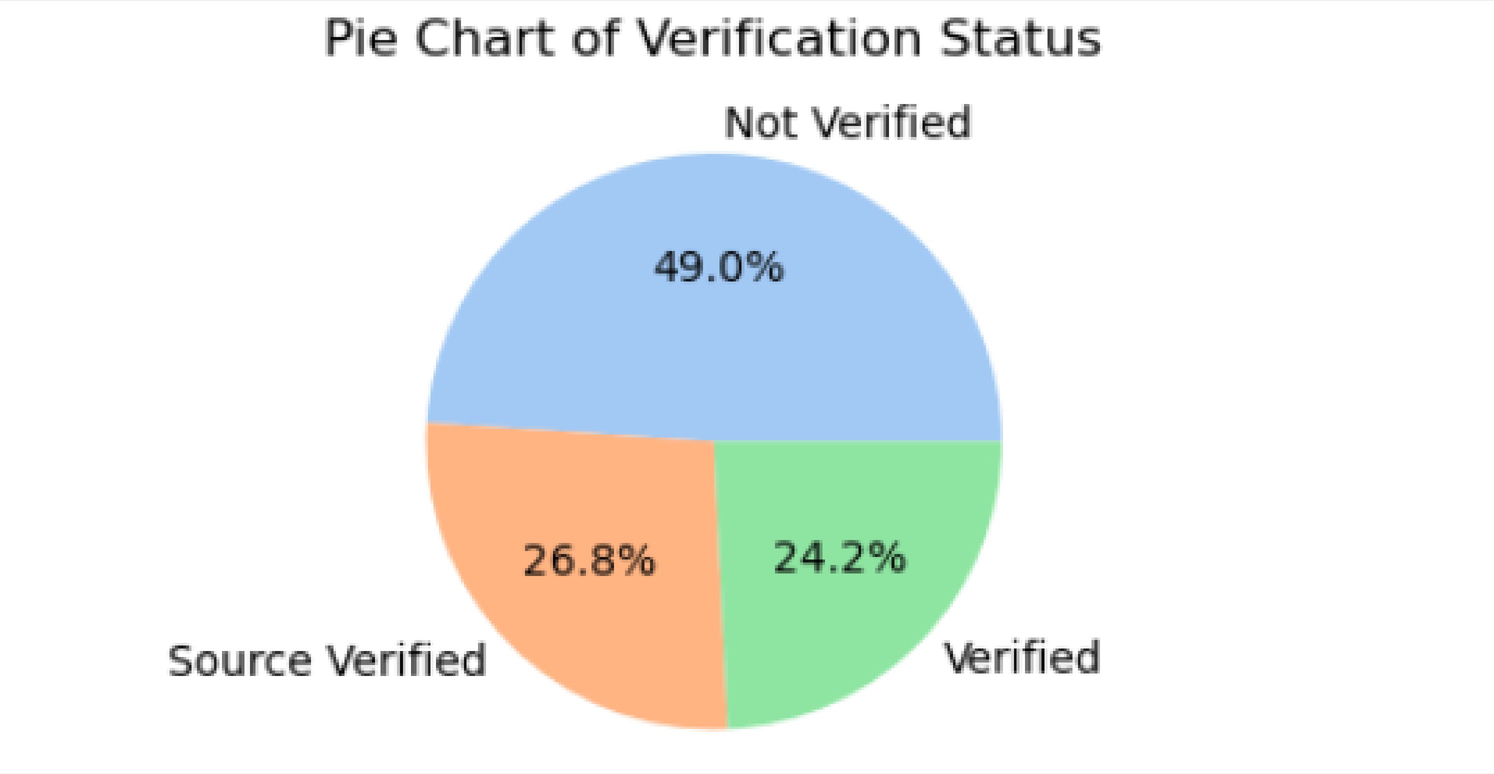


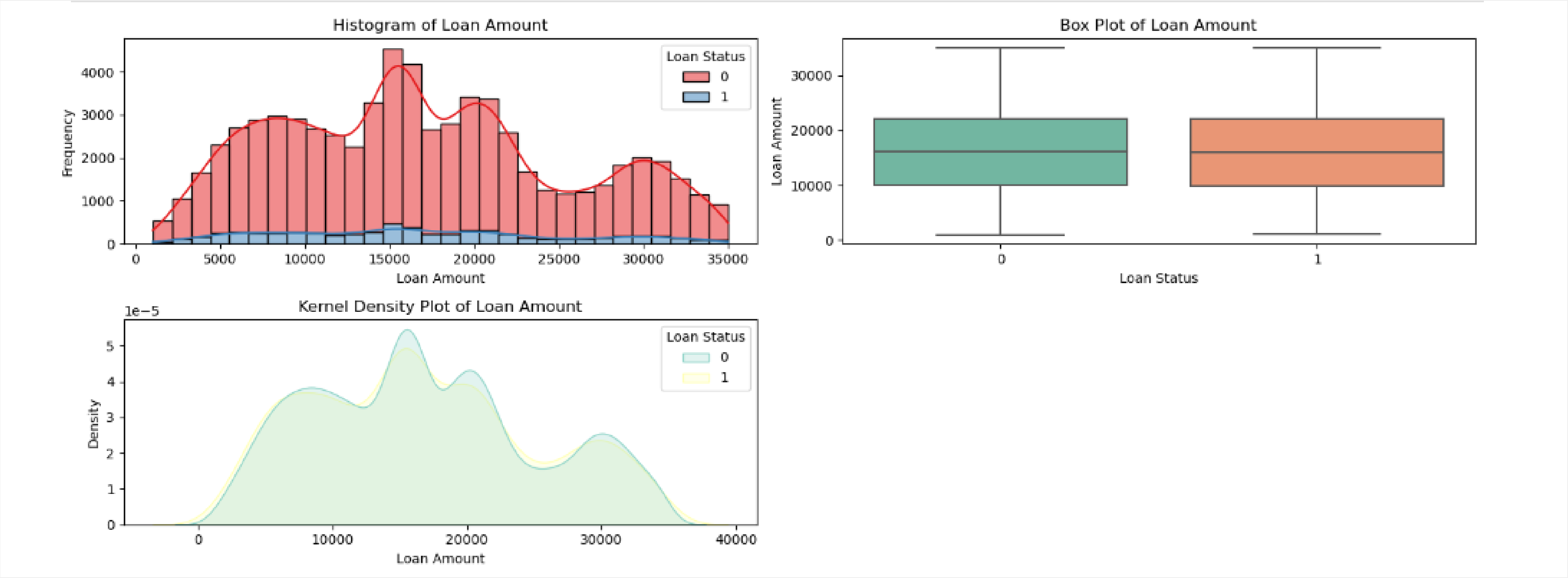
# DATA EXPLORATION

## PLOTS

FOR VISUAL ANALYSIS, DIVERSE PLOTS WERE CREATED BASED ON VARIABLE TYPES (NUMERICAL OR CATEGORICAL)

- **Numerical variables**
  - Histograms
  - Box Plot
  - Kernel Density Plot
- **Categorical variables**
  - Bar plots
  - heatmap
  - TreeMap
  - Pie Chart
  - Word Cloud

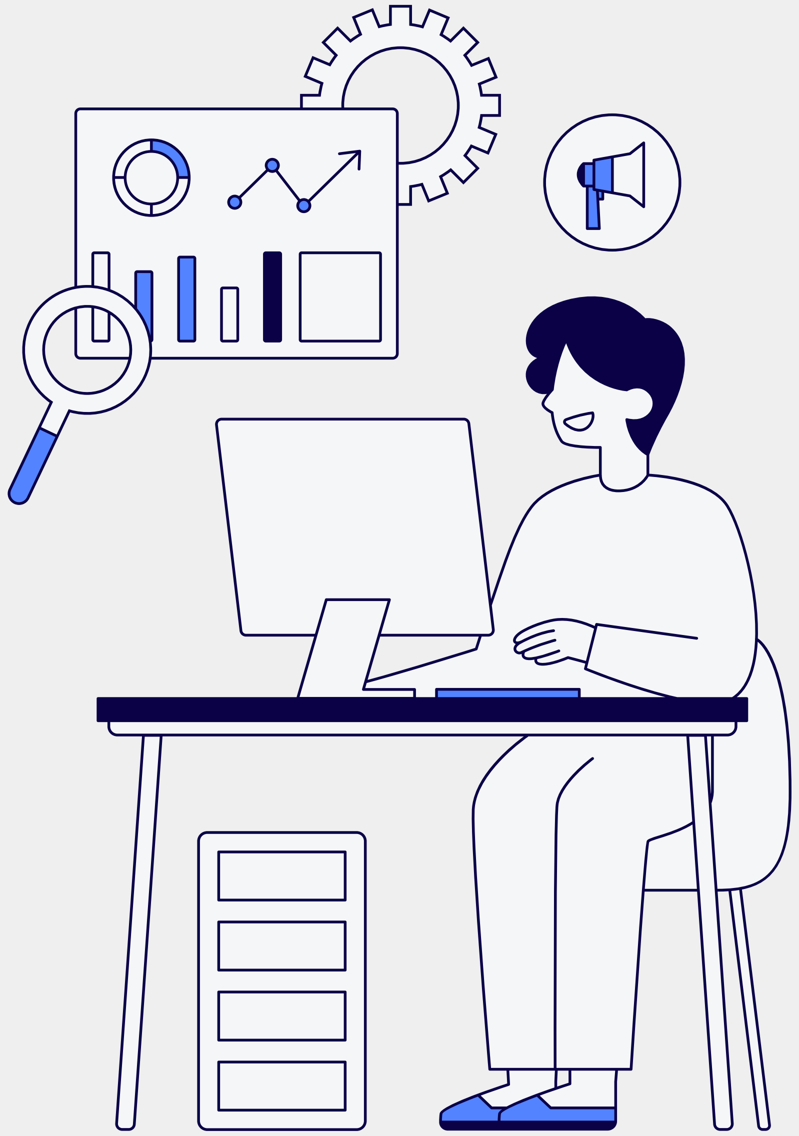


# DATA EXPLORATION

## DATAFRAME INFORMATION

- **Number of rows**
  - 67463 rows
- **Number of columns**
  - 35 columns
- **Size**
- **Missing values and duplicates**
  - None were found
- **Columns with only 1 possible value**
  - Columns found were dropped immediately



# DATA EXPLORATION

## 03

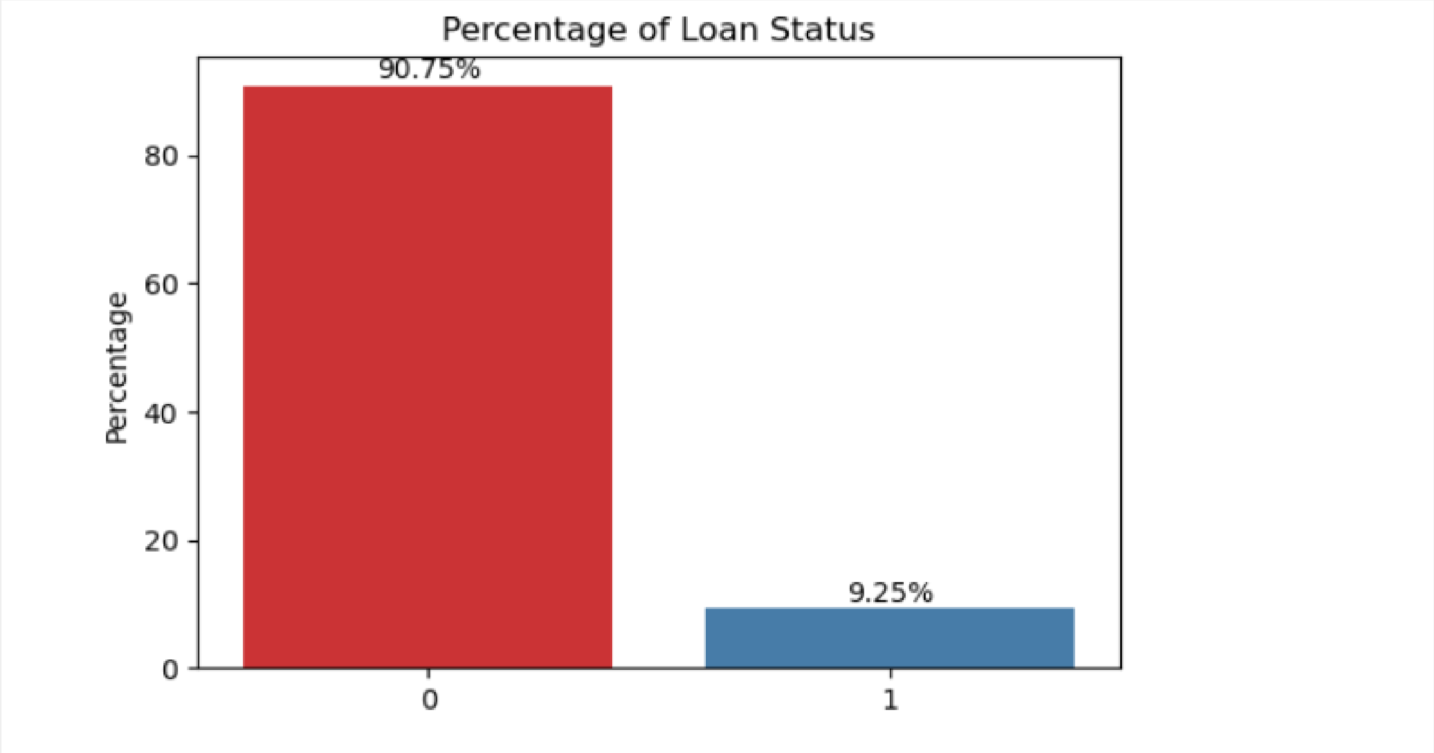
### CORRELATION ANALYSIS WITH TARGET VARIABLE

- ANOVA and CHI-Squared functions were used
- Columns not correlated were dropped

## 04

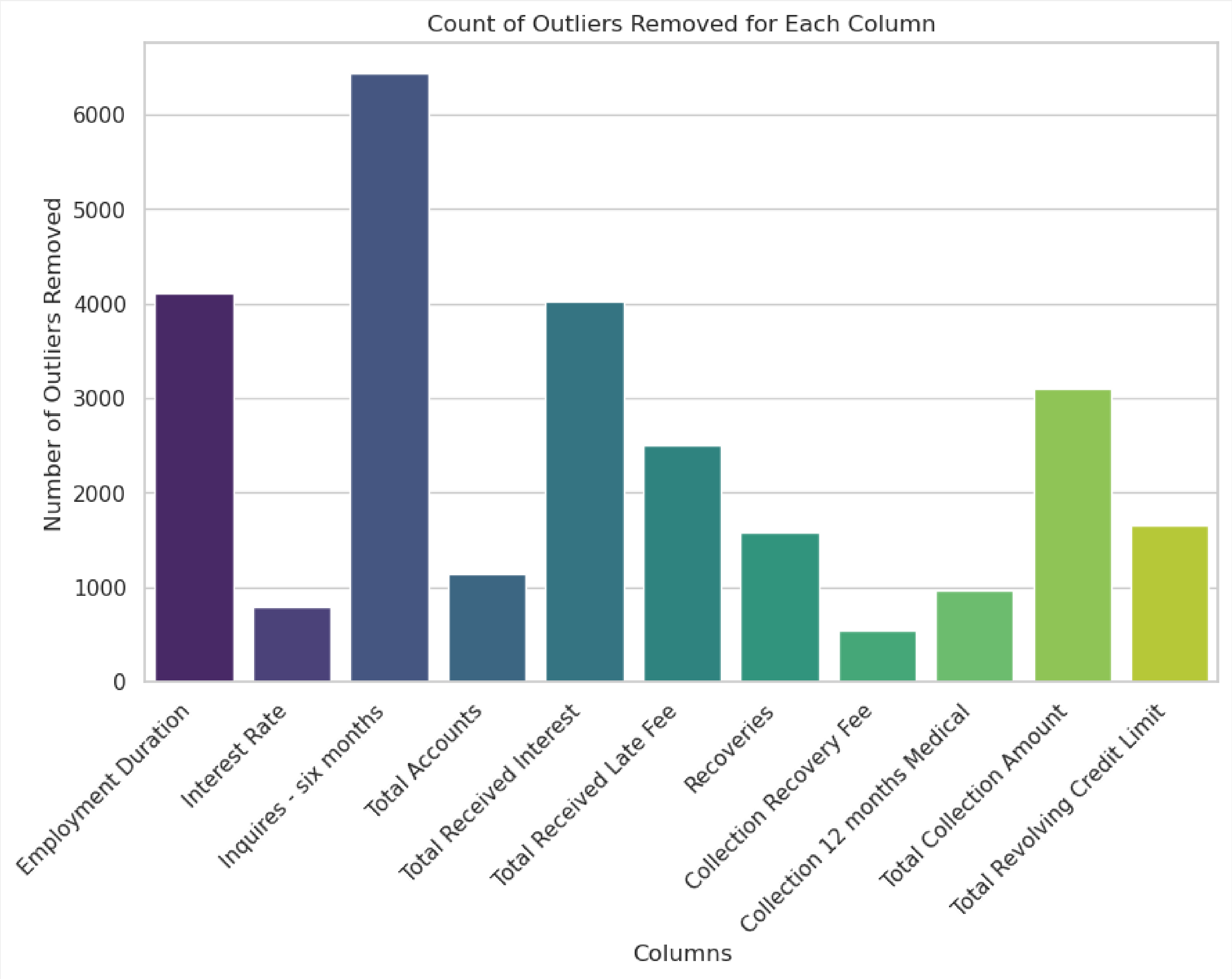
### DATA IMBALANCEMENT

- Data regarding target variable is extremely imbalanced



# PREPROCESSING
## OUTLIER HANDLING

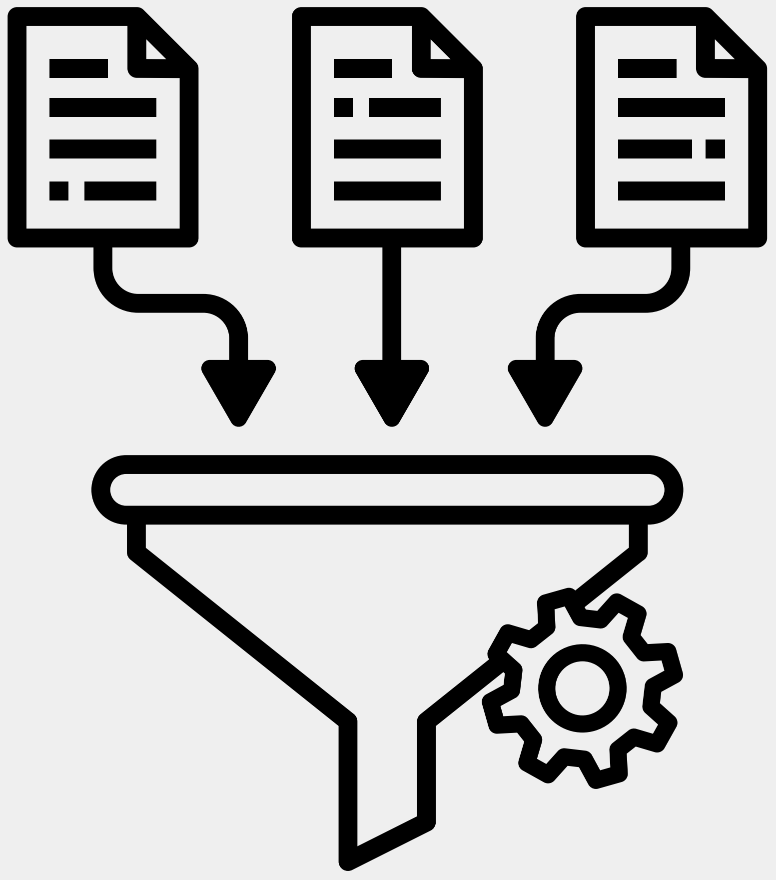
- Before dropping non-correlated columns, outliers must be dealt with.
- Visual analysis pinpointed variables with outliers in the dataset.
- The IQR measure was employed to assess which values should be considered outliers and removed.



# PREPROCESSING

## COLUMN REMOVAL AND CATEGORICAL ENCODING

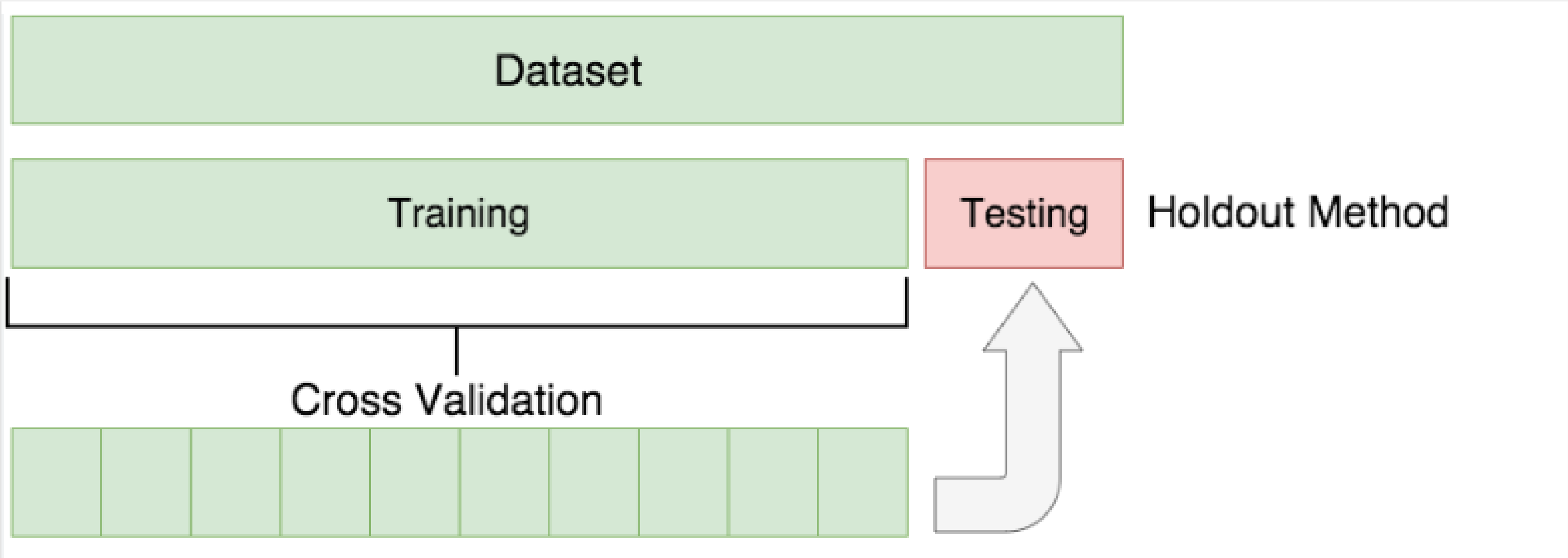
- Outliers addressed, non-correlated columns removed.
- Binary category columns were label encoded
- For Non-binary categorical columns (> 2 possible values), one-hot encoding was applied



# CREATION OF MODELS

## DATA PREPARATION

- Train/Test split
  - Training Data -> 80%
  - Testing Data -> 20%
- Undersampling techniques
  - Near-miss and random were used
- Data was scaled
  - StandardScaler

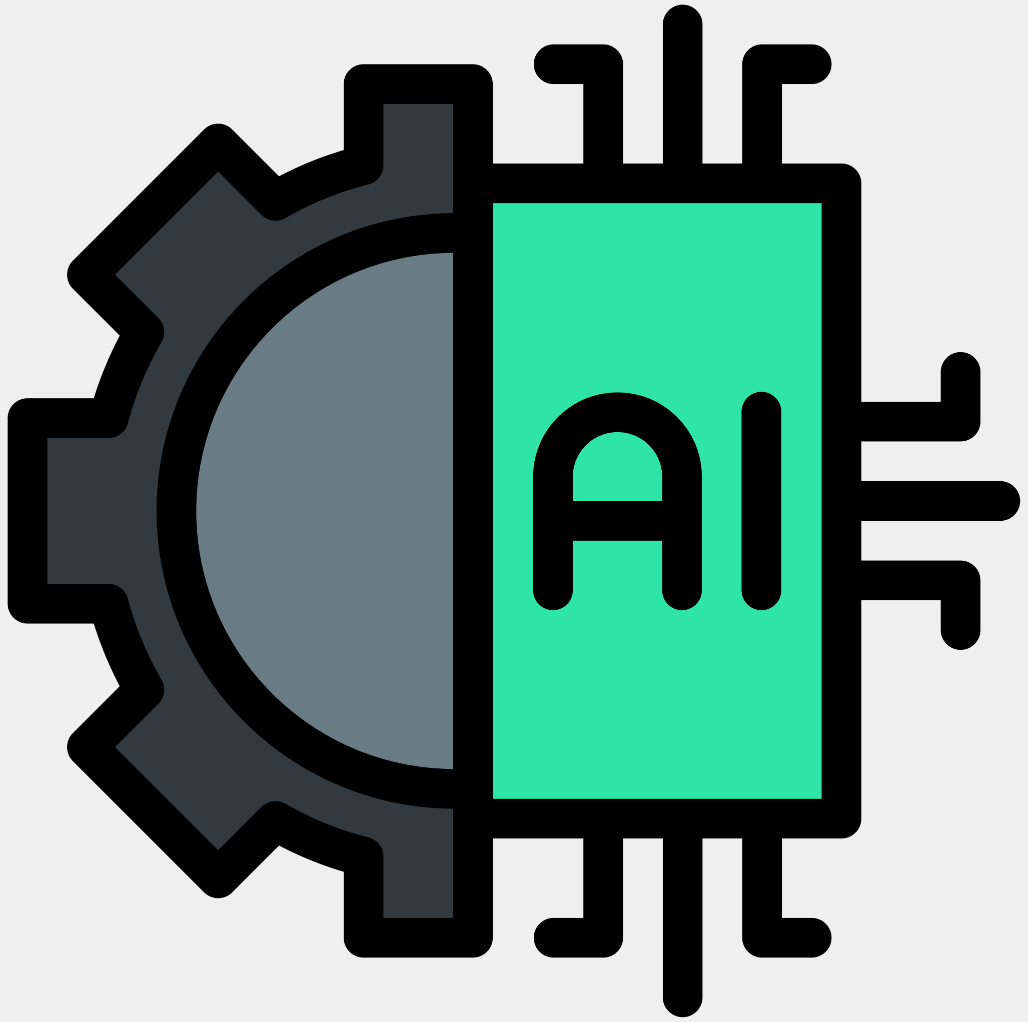


# CREATION OF MODELS

## ALGORTHMS

A wide range of algorithms/models were used:

- **Boosting Algorithms**
- **Linear Algorithms**
- **Decision Tree Algorithms**
- **Instance-Based Learning**
- **Neural Networks**
- **Probabilistic Methods**
- **Support Vector Machines**
- **Ensemble methods**



# MODEL'S EVALUATION
## CROSS-VALIDATION

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample

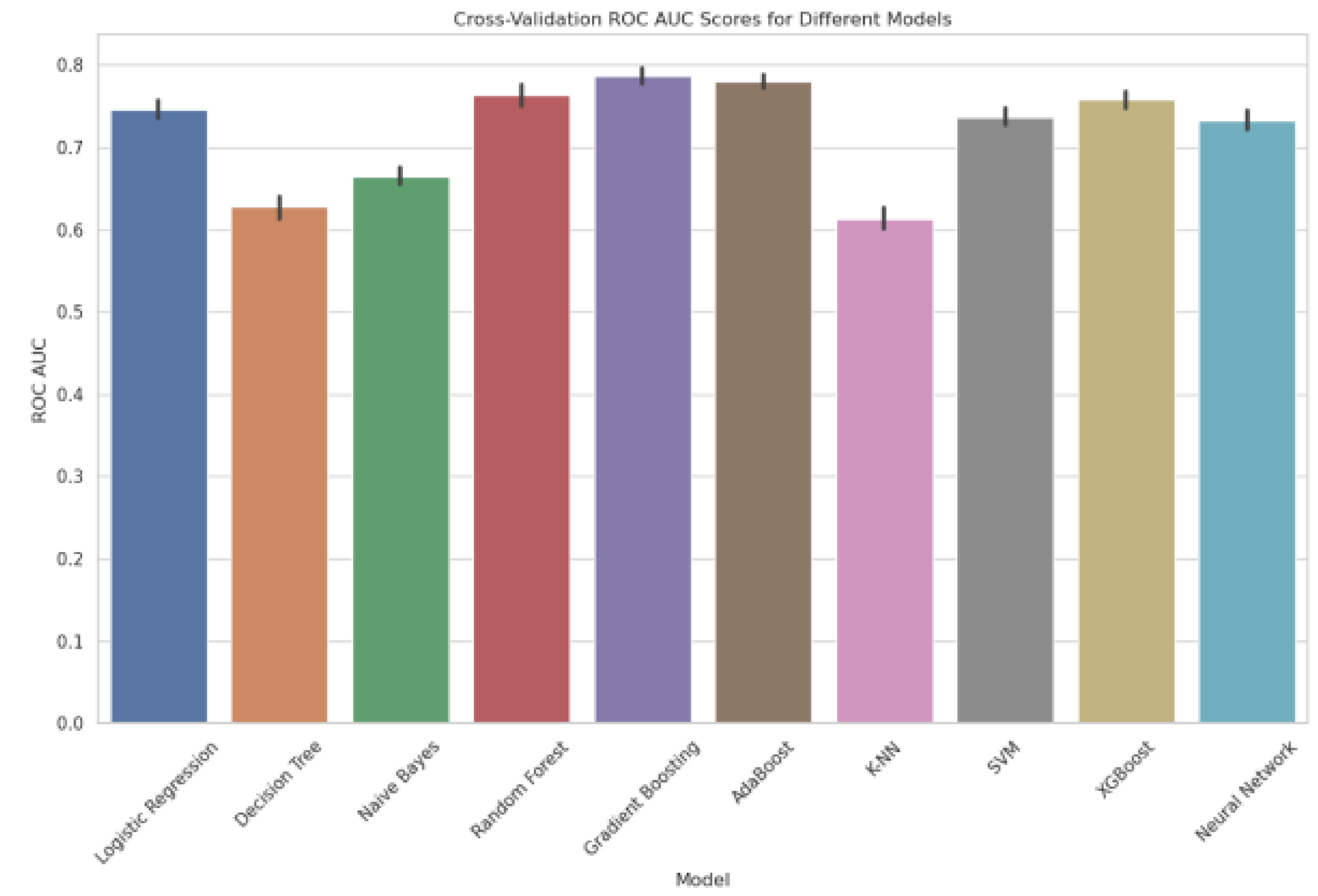


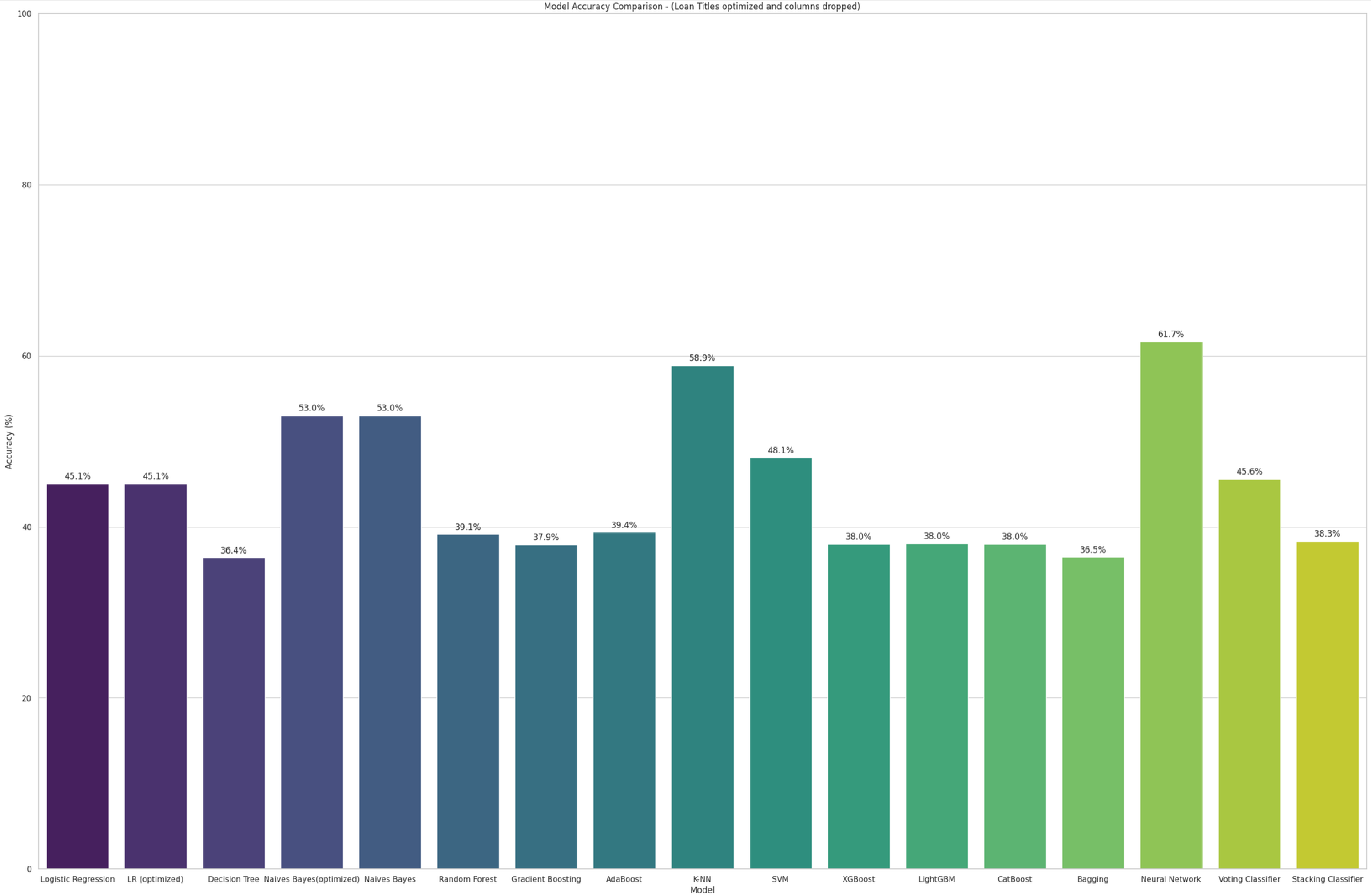

Figure 4.2: Cross-Validation NearMiss Undersampler

Cross-validation details:

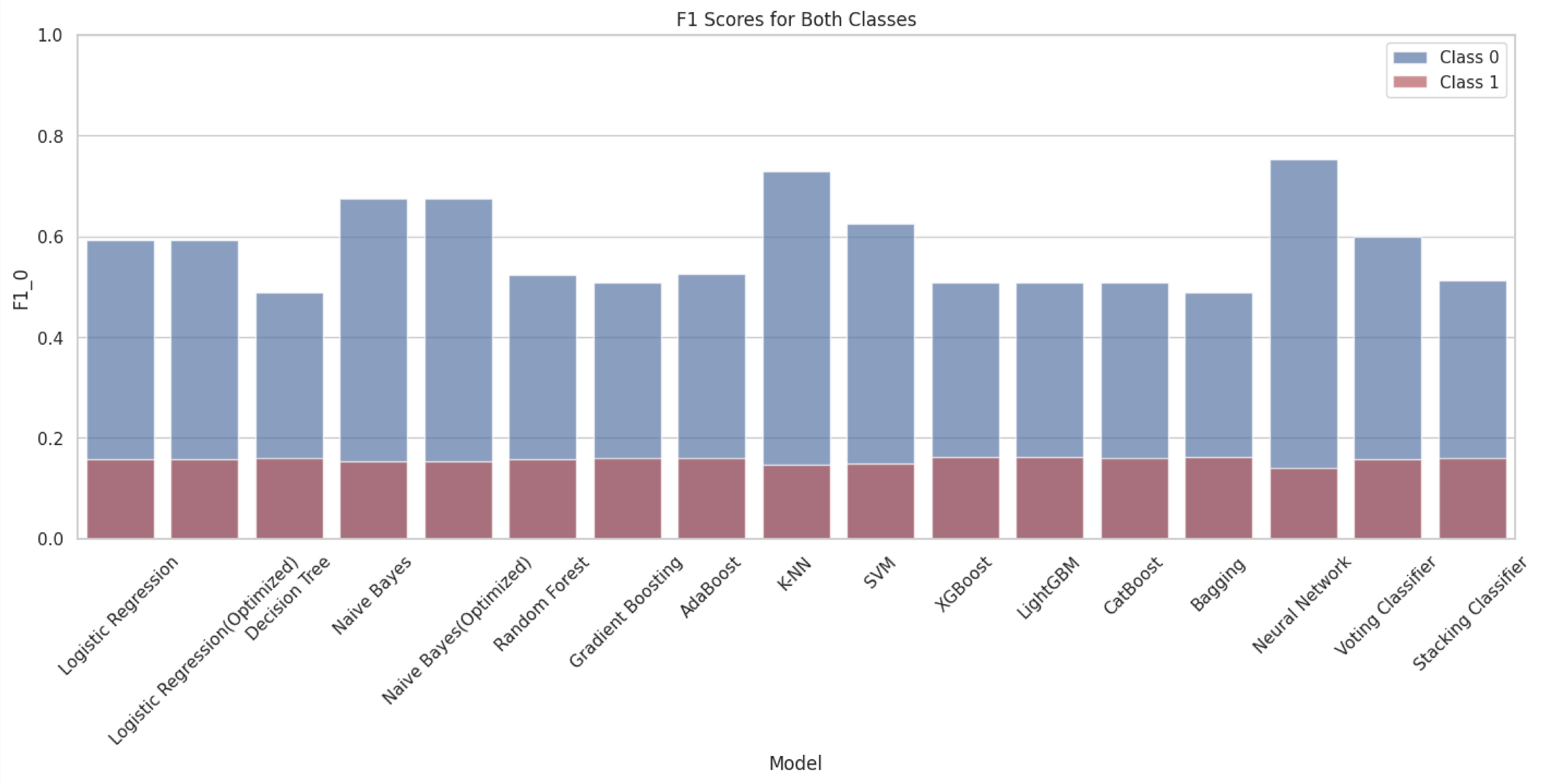
- 10 kfolds
- Scoring Metric: Roc-AUC

# MODEL'S EVALUATION

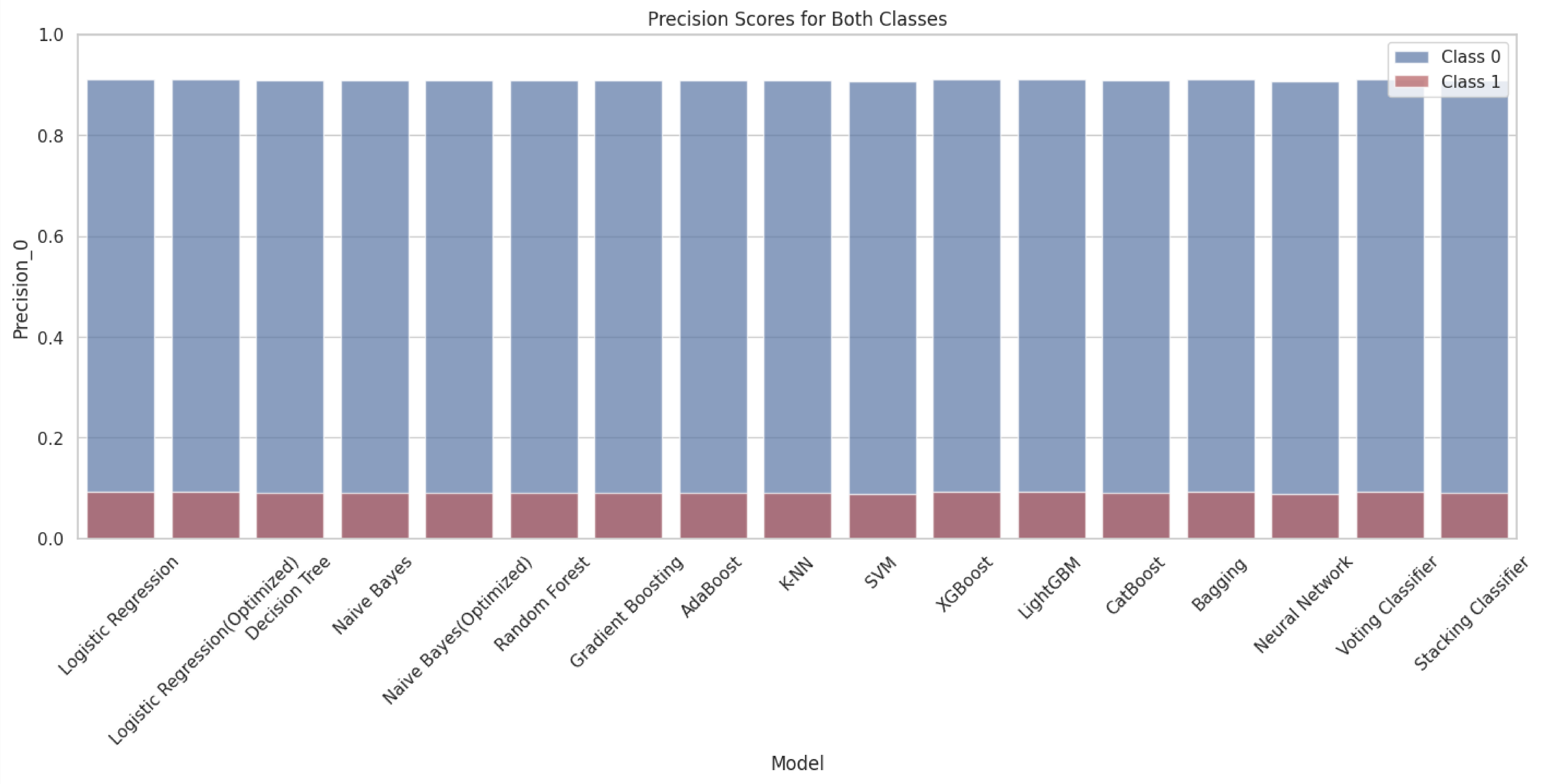
## ACCURACY



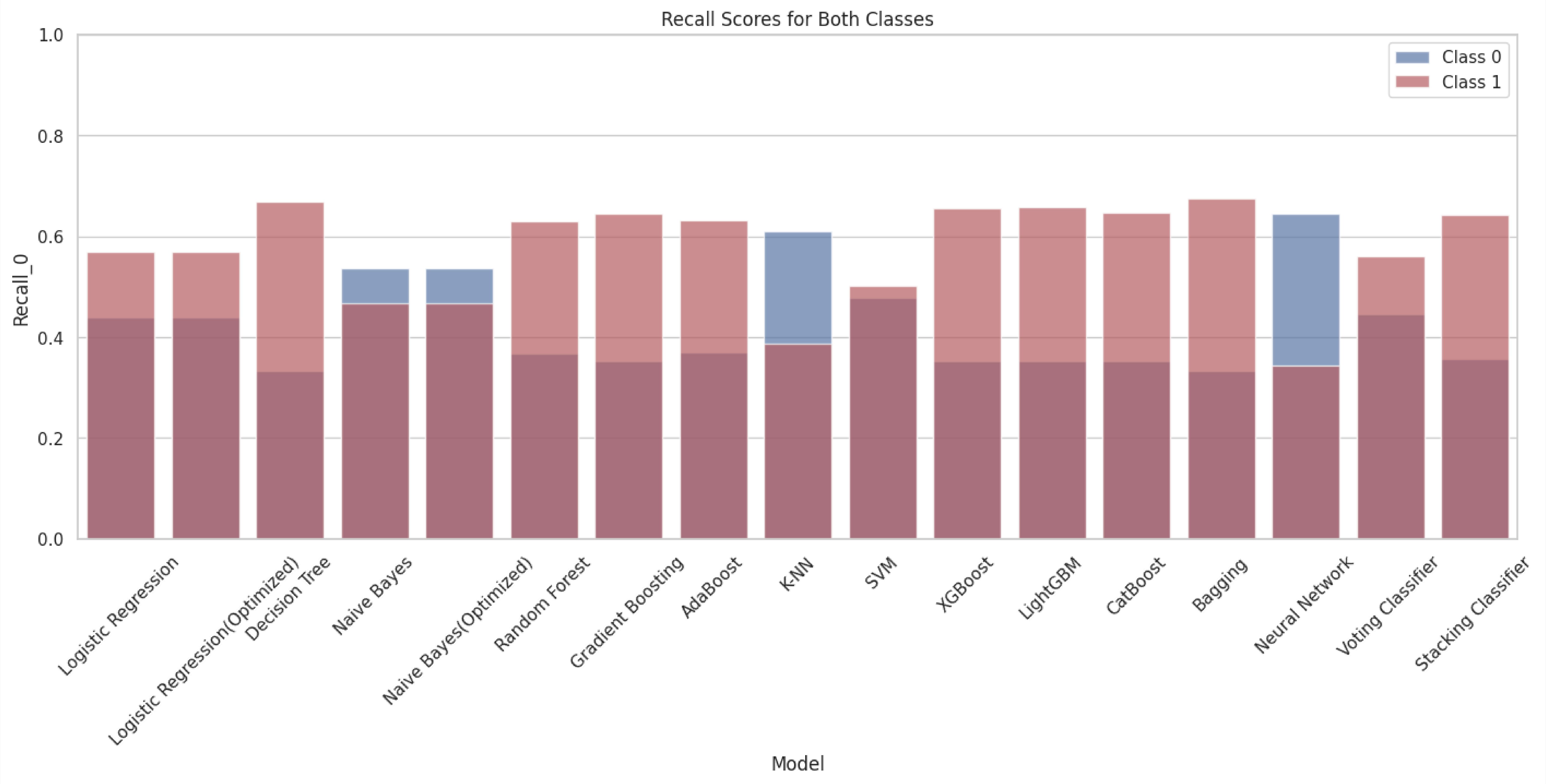
# MODEL'S EVALUATION
## F1



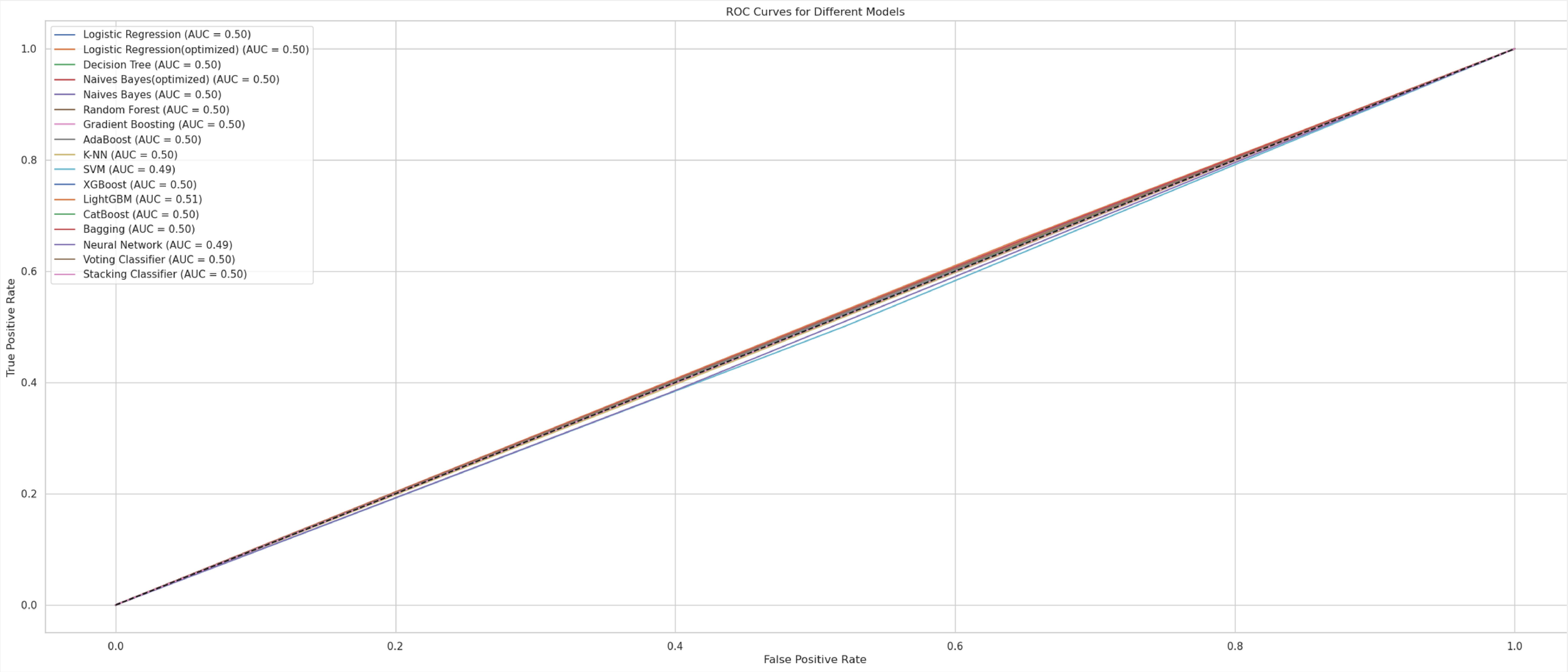
# MODEL'S EVALUATION
## PRECISION



# MODEL'S EVALUATION
## RECALL



# MODEL'S EVALUATION
## ROC-AUC



# CONSTRAINTS

Data imbalance poses a constraint as it may lead to biased model predictions, where the algorithm may favor the majority class and struggle to accurately predict the minority class.

## DATASET IMBALANCE CHALLENGE

To address the imbalance, undersampling techniques like NearMiss were employed. While these techniques help balance class distribution, they come with the drawback of reducing the amount of training data available to the models, potentially limiting their ability to learn intricate patterns.

## UNDERSAMPLING TECHNIQUES EMPLOYED

The reduction in training data due to undersampling may have affected the models' capacity to learn subtle patterns and nuances within the data, potentially impacting their performance on the test data.

## IMPACT ON MODEL PERFORMANCE

# CONCLUSION

- **Model Performance Overview:**
  - Diverse machine learning models showed varied performances in predicting loan defaults.
  - Notably, k-Nearest Neighbors (KNN) achieved the highest accuracy at 58.1%, closely followed by Neural Networks at 58.6%, and Naive Bayes at 53.0%.
- **Consistent ROC AUC Trends:**
  - Most models exhibited ROC AUC scores around 0.5, indicating limited discrimination ability.
  - The models struggled to effectively distinguish between loan default and non-default instances.
- **Call for Further Investigation:**
  - Despite commendable accuracy in some models, consistent 0.5 ROC AUC scores suggest a need for deeper investigation.
  - Exploring alternative models and refining features could enhance discrimination capabilities.
- **Insights and Future Considerations:**
  - Deeper analysis of dataset characteristics may uncover challenges faced by models.
  - The study provides valuable insights for refining predictive modeling approaches in predicting loan defaults.