UNIVERSITY OF EDINBURGH
Business School

2023-24

Applied Machine Learning

Individual Coursework

Credit Risk Modelling

Lahiru Alahakoon

Table of Contents

List of Tables

List of Figures

Executive Summary

- Financial institutions face significant challenges in managing credit risk. This study investigated the use of machine learning to predict customer defaulting on credit card loans.

- Two machine learning models were developed, Random Forest and XG Boost, and both were affected from poor data quality and class imbalance.

- While both models were under performed due to underfitting XG Boost model achieved a slightly better accuracy and f-1 score demonstrating its potential to detect at-risk borrowers.

- Future efforts should be focused on acquiring qualitative and alternative data sources and to incorporate domain expertise to improve models' overall accuracy and generalizability.

1. Introduction

1.1 Background

Financial Institutions like Banks face significant challenges in managing credit risk where the possibility of a borrower failing to repay a loan. This can potentially cause significant financial strain and impacting overall capital health. Traditionally, credit risk assessment has relied heavily on manual analysis, financial statements and credit history, a process inherently subjective and data limited. Machine learning with its powerful tools offers a revolutionary approach to credit risk assessment, measuring customer creditworthiness, and allowing credit card providers to make informed decisions regarding approvals.

This project sets out to develop a machine learning model to carry out credit risk assessments, using the provided dataset of credit card customer information over a one-year period. By analysing historical data and associated customer characteristics, the model will aim to estimate the likelihood of customers defaulting on their credit card payments. This crucial information then can be used by banks to optimize their risk management strategies, improve lending decisions, and eventually safeguard their financial stability. A comprehensive exploration will be embarked to address these challenges effectively and build a highly accurate credit risk assessment model through series of investigative steps. This in-depth analysis will serve as the basis for selecting the most appropriate techniques for data preparation, model training and, ultimately, achieving optimal model performance.

1.2 Problem Statement

This project develops and implement two machine learning models for credit risk assessment using synthetic customer demographic data related to credit cards to estimate the likelihood of a borrower defaulting on future loans. Later two models will be compared and discussed on their performance and accuracy. Eventually, the aim is to identify the most effective machine learning model for predicting customer default risk and provide the bank with valuable tool for making informed lending decisions.

2. Methodology

2.1 Exploratory Data Analysis

The given data set consists of 50,000 data points and 54 features. From the data set, ‘CLIENT_ID’ can be removed, leaving 52 features to examine along with the target variable ‘TARGET_LABEL_BAD.1’. There are nominal, discrete, continuous, and binary data present and, some of the categorical data have been already encoded. (e.g. FLAG_VISA, etc.)

2.1.1 Data Description

Following table consists of variable titles and their descriptions.

Table 1 - Variable Description

Id	Var_Title	Var_Description
1	ID_CLIENT	Sequential number for the applicant (to be used as a key)
2	CLERK_TYPE	Not informed
3	PAYMENT_DAY	Day of the month for bill payment, chosen by the applicant
4	APPLICATION_SUBMISSION_TYPE	Indicates if the application was submitted via the internet or in person/posted
5	QUANT_ADDITIONAL_CARDS	Quantity of additional cards asked for in the same application form
6	POSTAL_ADDRESS_TYPE	Indicates if the address for posting is the home address or other. Encoding not informed.
7	SEX	
8	MARITAL_STATUS	Encoding not informed
9	QUANT_DEPENDANTS	
10	EDUCATION_LEVEL	Educational level in gradual order not informed
11	STATE_OF_BIRTH	
12	CITY_OF_BIRTH	
13	NACIONALITY	Country of birth. Encoding not informed but Brazil is likely to be equal 1.
14	RESIDENCIAL_STATE	State of residence
15	RESIDENCIAL_CITY	City of residence
16	RESIDENCIAL_BOROUGH	Borough of residence
17	FLAG_RESIDENCIAL_PHONE	Indicates if the applicant possesses a home phone
18	RESIDENCIAL_PHONE_AREA_CODE	Three-digit pseudo-code

19	RESIDENCE_TYPE	Encoding not informed. In general, there are the types: owned, mortgage, rented, parents, family etc.
20	MONTHS_IN_RESIDENCE	Time in the current residence in months
21	FLAG_MOBILE_PHONE	Indicates if the applicant possesses a mobile phone
22	FLAG_EMAIL	Indicates if the applicant possesses an e-mail address
23	PERSONAL_MONTHLY_INCOME	Applicant's personal regular monthly income in Brazilian currency (R$)
24	OTHER_INCOMES	Applicant's other incomes monthly averaged in Brazilian currency (R$)
25	FLAG_VISA	Flag indicating if the applicant is a VISA credit card holder
26	FLAG_MASTERCARD	Flag indicating if the applicant is a MASTERCARD credit card holder
27	FLAG_DINERS	Flag indicating if the applicant is a SINERS credit card holder
28	FLAG_AMERICAN_EXPRESS	Flag indicating if the applicant is an AMERICAN EXPRESS credit card holder
29	FLAG_OTHER_CARDS	Despite being label "FLAG", this field presents three values not explained
30	QUANT_BANKING_ACCOUNTS	
31	QUANT_SPECIAL_BANKING_ACCOUNTS	
32	PERSONAL_ASSETS_VALUE	Total value of the personal possessions such as houses, cars etc. in Brazilian currency (R$).
33	QUANT_CARS	Quantity of cars the applicant possesses
34	COMPANY	If the applicant has supplied the name of the company where he/she formally works
35	PROFESSIONAL_STATE	State where the applicant works
36	PROFESSIONAL_CITY	City where the applicant works
37	PROFESSIONAL_BOROUGH	Borough where the applicant works
38	FLAG_PROFESSIONAL_PHONE	Indicates if the professional phone number was supplied
39	PROFESSIONAL_PHONE_AREA_CODE	Three-digit pseudo-code
40	MONTHS_IN_THE_JOB	Time in the current job in months
41	PROFESSION_CODE	Applicant's profession code. Encoding not informed
42	OCCUPATION_TYPE	Encoding not informed
43	MATE_PROFESSION_CODE	Mate's profession code. Encoding not informed
44	EDUCATION_LEVEL	Mate's educational level in gradual order not informed
45	FLAG_HOME_ADDRESS_DOCUMENT	Flag indicating documental confirmation of home address
46	FLAG_RG	Flag indicating documental confirmation of citizen card number
47	FLAG_CPF	Flag indicating documental confirmation of taxpayer status

48	FLAG_INCOME_PROOF	Flag indicating documental confirmation of income
49	PRODUCT	Type of credit product applied. Encoding not informed
50	FLAG_ACSP_RECORD	Flag indicating if the applicant has any previous credit delinquency
51	AGE	Applicant's age at the moment of submission
52	RESIDENCIAL_ZIP_3	Three most significant digits of the actual home zip code
53	PROFESSIONAL_ZIP_3	Three most significant digits of the actual job zip code
54	TARGET_LABEL_BAD=1	Target Variable: BAD=1, GOOD=0

2.1.2 Missing Values

It is crucial to handle missing values because they might lead to bias results, reduce the accuracy, and hinders model performance. The figure below shows the presence of missing values in the data set.

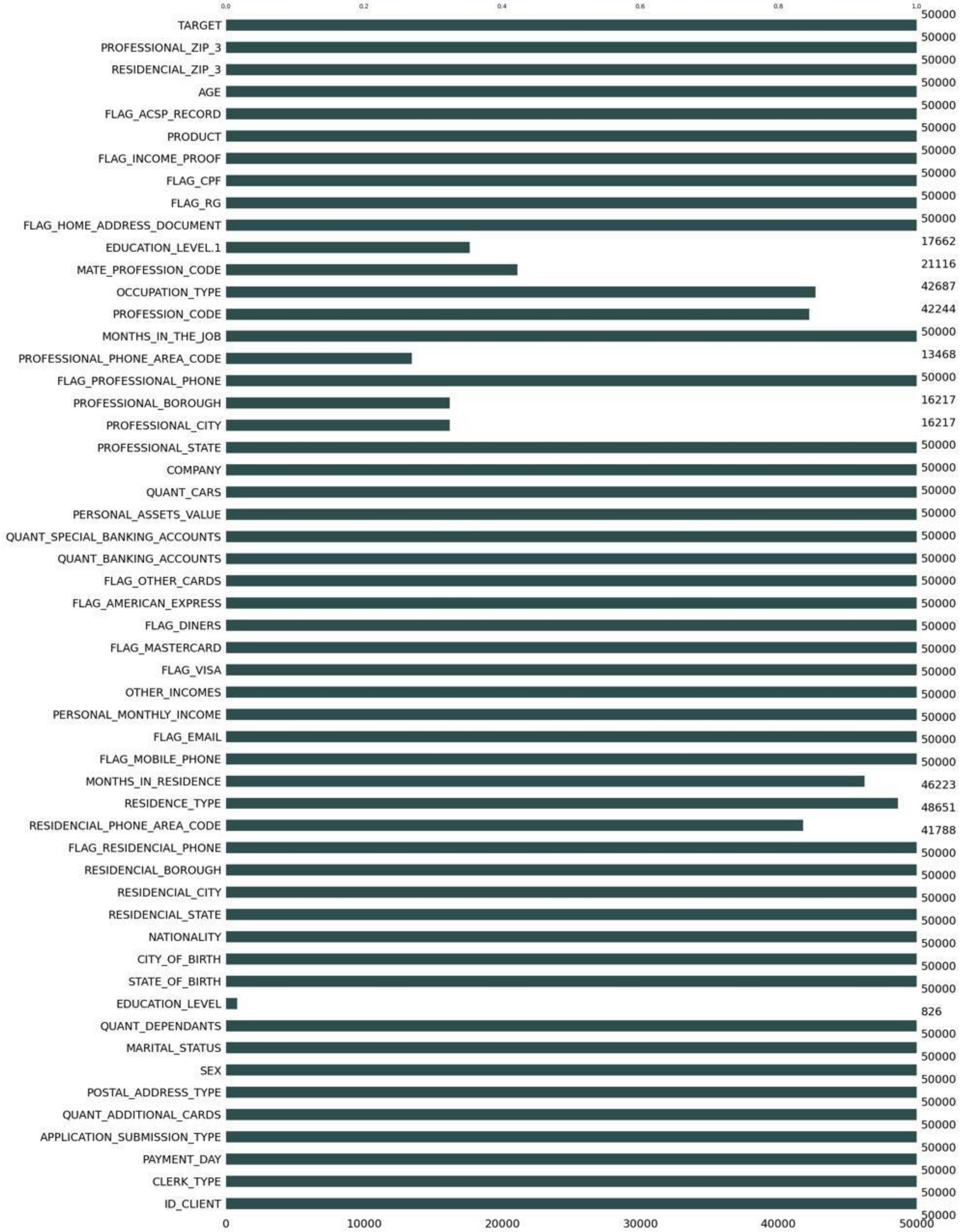


Figure 1 - Missing Values

There are 242,901 missing values present across 11 features. Their percentage distribution is as follows.

Table 2 - Missing Value Table

Variable	Missing %
EDUCATION_LEVEL	98.35%
PROFESSIONAL_PHONE_AREA_CODE	73.06%
PROFESSIONAL_CITY	67.57%
PROFESSIONAL_BOROUGH	67.57%
EDUCATION_LEVEL.1	64.68%
MATE_PROFESSION_CODE	57.77%
RESIDENCIAL_PHONE_AREA_CODE	16.42%
PROFESSION_CODE	15.51%
OCCUPATION_TYPE	14.63%
MONTHS_IN_RESIDENCE	7.55%
RESIDENCE_TYPE	2.70%

EDUCATION_LEVEL and RESIDENCE_TYPE has the highest and lowest missing value count in the data set, which is 98.35% and 2.7% respectively. Appropriate imputation techniques will be adapted for variables that has a missing percentage less than 20%. The rest will be removed because it will be ineffective to impute that many missing values.

2.1.3 Exploring Target Variable

Target column represents the credit default status of customers (0= GOOD, 1= BAD). The distribution of classes is as follows.

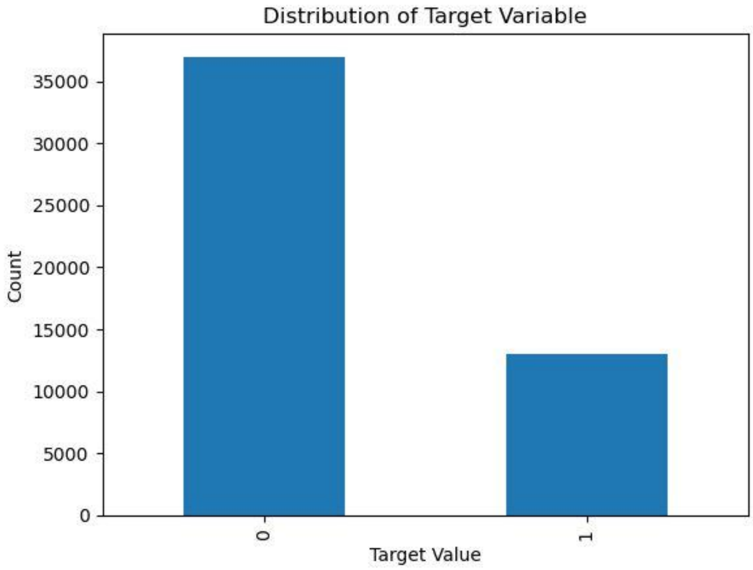


Figure 2 - Distribution of Target Variable

There is a significant class imbalance that needs to be addressed in further analysis.

2.1.4 Correlation Heat Map

Below shows a correlation heat map generated for the original data set. It visualizes the relationships between multiple numerical variables in the data set.

- PROFESSIONAL_PHONE_AREA_CODE has a high correlation with RESIDENCIAL PHONE_AREA_CODE which is explainable.
- FLAG_EMAIL is highly correlated with QUANT_BANKING_ACCOUNTS, QUANT_SPECIAL_BANKING_ACCOUNTS and QUANT_CARS
- None of the variables have a high correlation with the target variable.

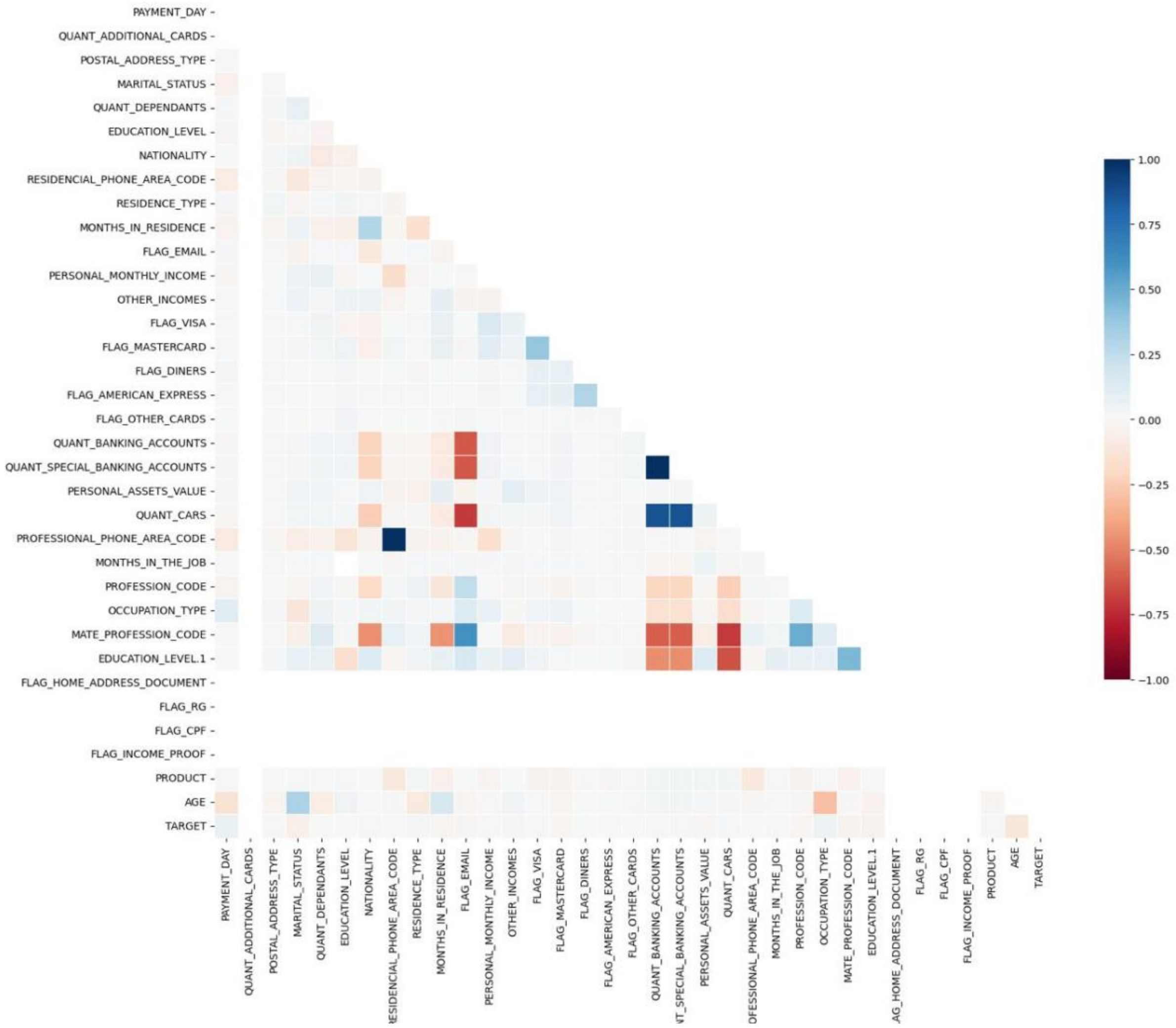


Figure 3 - Correlation Heat Map

2.1.5 Bivariate Analysis

The below figure provides a comprehensive overview of bivariate relationships of few selected features. A Count plot of each feature is presented in the diagonal of the matrix. The rest are bubble plots, used to explore the distribution and magnitude of the relationship.

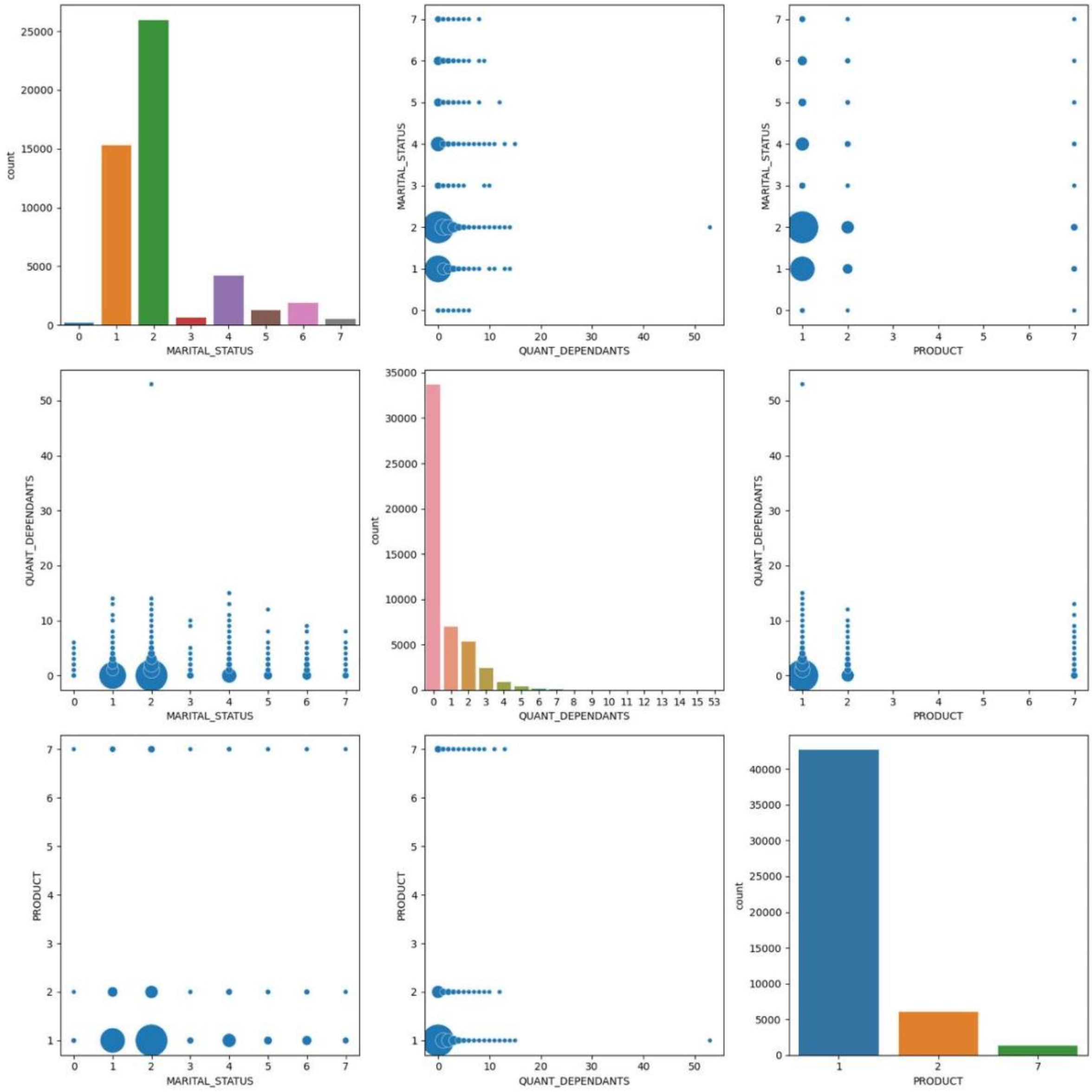


Figure 4 - Distribution of Customer Demographics and Product Requested

The following figure shows the Bivariate analysis concerning the target variable. The dataset exhibits a class imbalance.

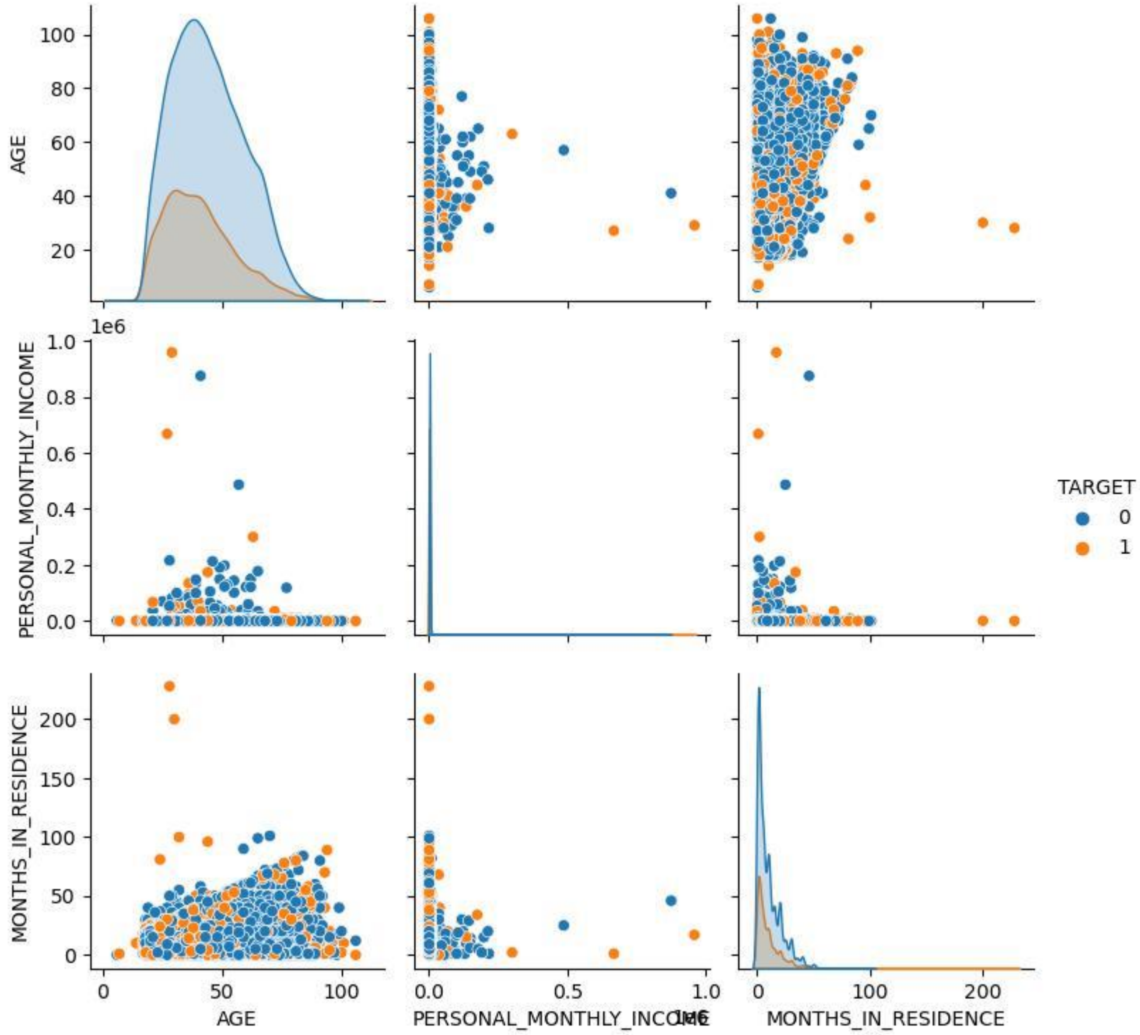


Figure 5 - Bivariate Analysis of Numerical Variables Against Target Variable

The following boxplot represents the numerical features and a heavy set of outliers have been identified. This will be addressed in future steps.

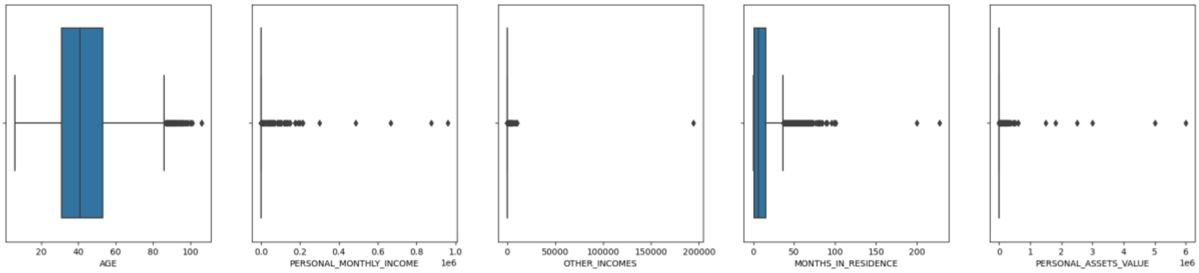


Figure 6 - Box Plot s of Numerical Features

2.2 Data Preprocessing

Data preprocessing involves Manipulation, Filtration and Augmentation of data before it is analysed. The following data preprocessing steps are considered in this study.

- Handling Missing Values, Errors, and Outliers
- Free Text Elimination
- Dominant Category Removal
- Correlation Analysis
- Domain Expertise Integration

2.2.1 Handling Missing Values, Errors, and Outliers

2.2.1.1 Missing Value Imputation

Imputation of missing values is considered the best option, as removing data will reduce the number of records which could affect the model accuracy. But features that has a missing value percentage less than 20% will only be considered for imputation and others will be neglected.

- Mode Imputation is carried out for Categorical variables.
- Median Imputation is carried out for Numerical variables.

2.2.1.2 Handling Errors

In 'PAYMENT _DAY' column there are 90 records which had an error value of -99999. These will be Mode Imputed.

2.2.1.3 Winsorizing Method

Winsorization replaces outliers with values at pre-defined percentiles (i.e., 5th and 95th percentile). This reduces the impact of outliers on metrics like mean or standard deviation, providing more robust estimates.

Extreme outliers are present in following features; hence they are winsorized.

- PERSONAL_MONTHLY_INCOME
- OTHER_INCOMES
- AGE
- MONTHS_IN_RESIDENCE
- QUANT_DEPENDANTS

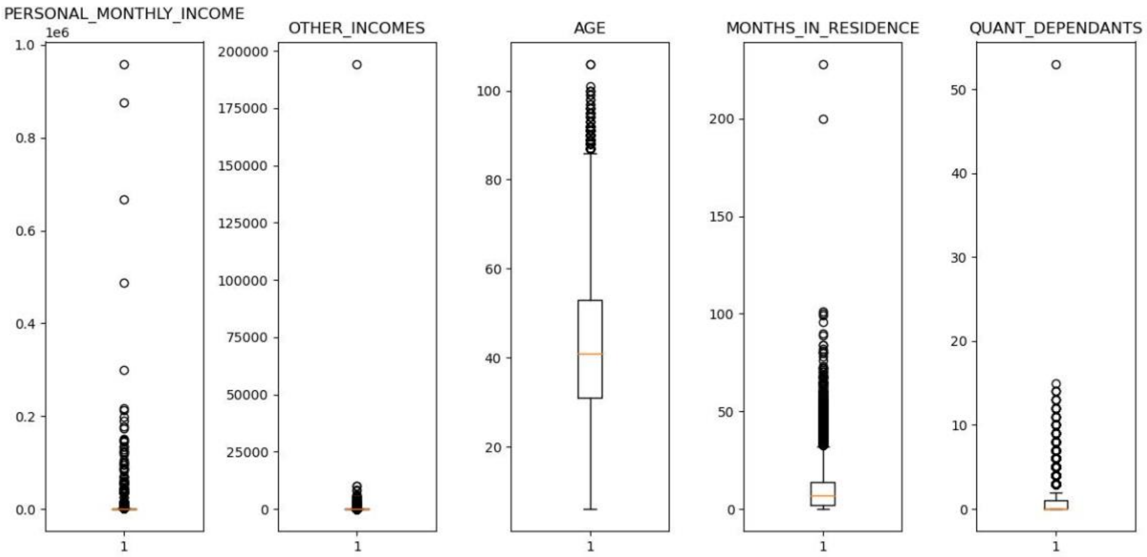


Figure 7 - Before Winsorization

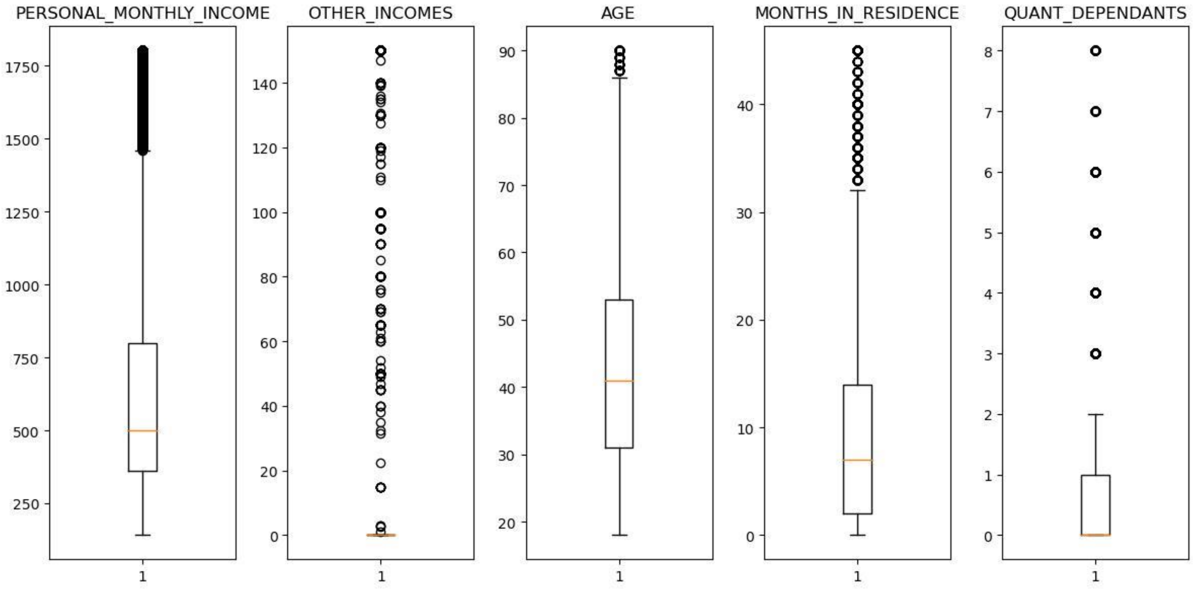


Figure 8 - After winsorization

2.2.2 Free Text Examination

Features containing free text entries are individually assessed. It should be determined if these features require conversion into a numerical format suitable for modelling, or if they should be excluded altogether.

- CITY_OF_BIRTH
- RESIDENCIAL_CITY
- RESIDENCIAL_BOROUGH
- STATE_OF_BIRTH

Above features include geographical data of the clients, which do not hold a great significance. Therefore, they will not be considered for modelling.

2.2.3 Dominant Category Removal

Categorical features where single class hold a disproportionately large share of the dataset are identified and removed. These features offer minimal predictive power. But in some cases, the dominant category might be relevant. Below table shows the occurrences where each feature had a single value that accounts for at least 95% of the entire column.

Table 3 - Dominant Features

Feature	**Dominant Percentage**
CLERK_TYPE	100%
QUANT_ADDITIONAL_CARDS	100%
FLAG_MOBILE_PHONE	100%
FLAG_DINERS	100%
FLAG_AMERICAN_EXPRESS	100%
FLAG_OTHER_CARDS	100%
MONTHS_IN_THE_JOB	100%
FLAG_HOME_ADDRESS_DOCUMENT	100%
FLAG_RG	100%
FLAG_CPF	100%
FLAG_INCOME_PROOF	100%
FLAG_ACSP_RECORD	100%
POSTAL_ADDRESS_TYPE	99%
NATIONALITY	96%
PERSONAL_ASSETS_VALUE	95%

Let’s look at few important features in the above table.

- In MONTHS_IN_THE_JOB, 99% of the entries are 0.
- In PERSONAL_ASSETS_VALUE, 95.2% of the entries are 0.
- In QUANT_ADDITIONAL_CARDS, 100% of the entries are 0.

Since these features contain a dominant category, they will be eliminated.

2.2.4 Correlation Analysis

Based on the EDA, following features show a greater correlation with each other.

- QUANT_BANKING_ACCOUNTS
- QUANT_SPECIAL_BANKING_ACCOUNTS
- QUANT_CARS
- FLAG_EMAIL

Therefore, we can keep one (QUANT_BANKING_ACCOUNTS) and eliminate others as they all have a linear relationship.

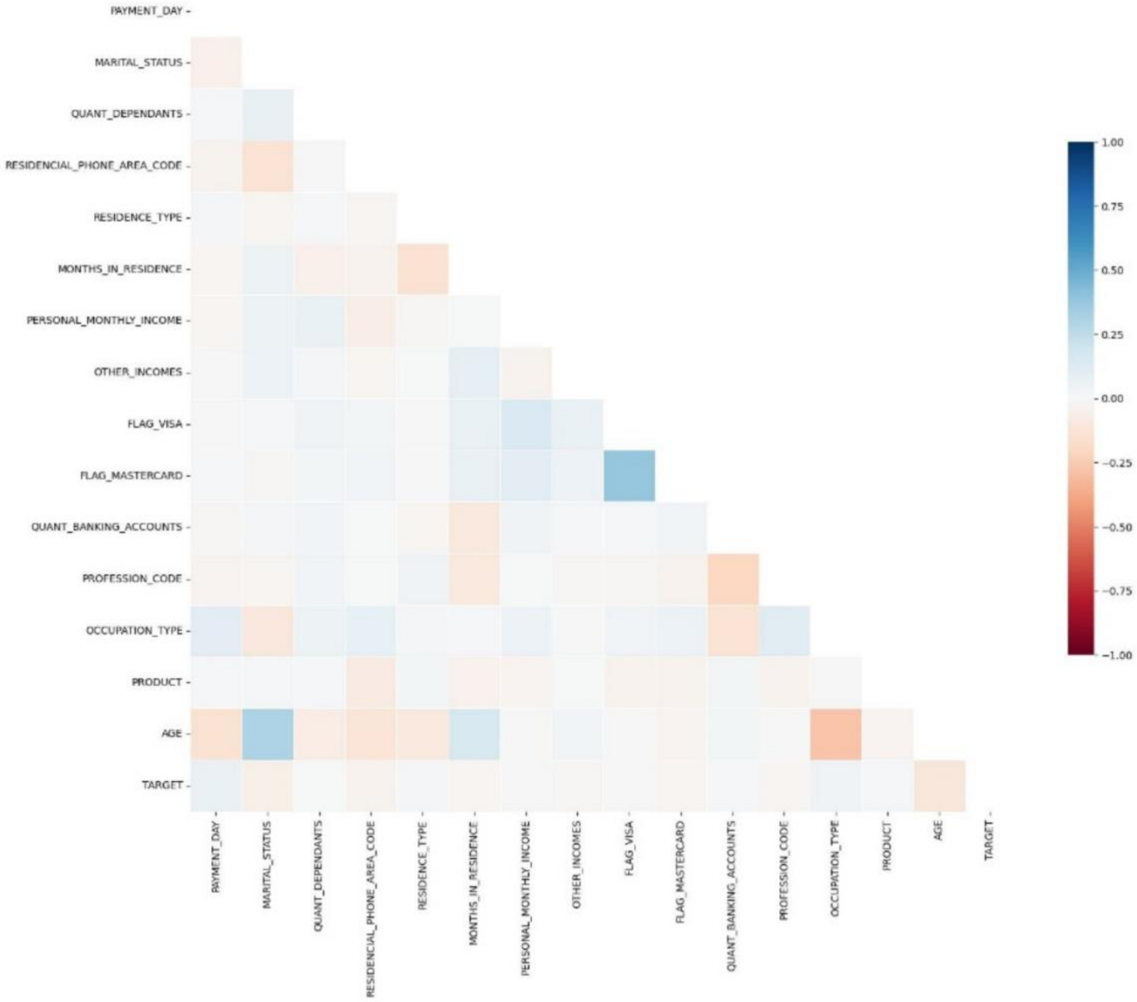


Figure 9 - Heat Map after Correlation Analysis

2.2.5 Domain Expertise Integration

Domain knowledge is utilized to identify irrelevant features in the dataset that was not flagged by the automated selection process.

Following features will be removed based on their nature and relevancy to the model. (DeNicola, 2023) (Langager)

- ID_CLIENT
- SEX
- RESIDENCIAL_PHONE_AREA_CODE
- PROFESSIONAL_STATE
- RESIDENCIAL_ZIP_3
- PROFESSIONAL_ZIP_3
- APPLICATION_SUBMISSION_TYPE

2.3 Feature Engineering and Feature Selection

2.3.1 Feature Engineering

A new feature 'TOTAL_INCOME' can be engineered by summing 'PERSONAL_MONTHLY_INCOME' and 'OTHER_INCOME'. Both features represent monthly income in Brazilian currency (R$) and share the same unit.

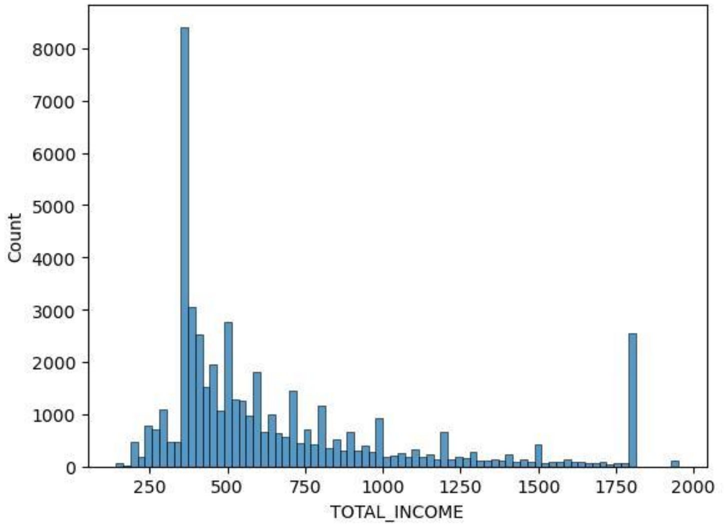


Figure 10 - Distribution of new Column: Total Income

2.3.2 Feature Selection

Feature selection is the process of identifying and selecting a subset of relevant and important features to the model. In this study two methods are carried out.

1) Tree Based Feature Selection

   A tree-based model can be used to compute impurity-based feature importances which in turn can be used to discard irrelevant features. Below graph showcases the feature importances generated using an Extra Tree Classifier. Label encoder and standard scaler were used to encode and scale categorical and numerical data respectively.

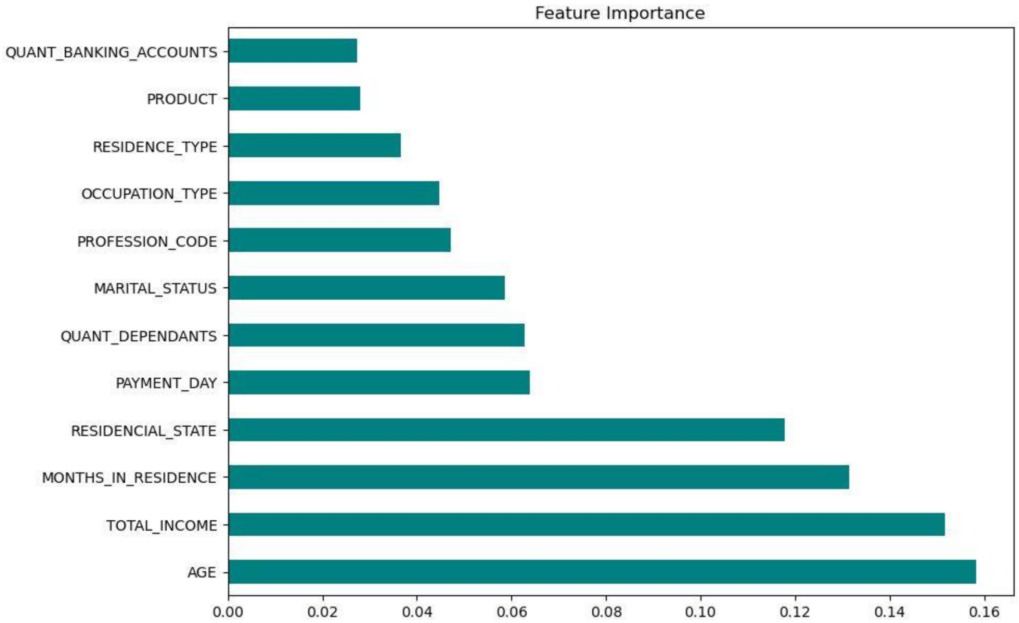


Figure 11 - Feature importances based on an Extra Tree Classifier

2) Recursive Feature Elimination

   The recursive feature elimination is the process of selecting the feature by recursively considering smaller and smaller sets of features, given an external estimator. Here the same estimator (Extra Tree Classifier) was used to verify the results from tree-based selection.

The selected features are,

- AGE
- TOTAL_INCOME
- MONTHS_IN_RESIDENCE
- RESIDENCIAL_STATE
- PAYMENT_DAY
- QUANT_DEPENDANTS
- MARITAL_STATUS

2.3.3 Data Scaling and Encoding

Encoding is the process of converting categorical data into numerical form for algorithms to process effectively. Feature scaling transforms numerical data into a common scale. This is important because numerical variables may have different units, scales and ranges which affect the model training time. In this study, One-Hot Encoding was used for categorical variables and Standard Scaler were used for numerical variables.

2.3.4 Data Resampling

Data resampling is a strategy that is used to address the Class Imbalance of classification models. There is a significant Class Imbalance present in the target variable. SMOTE (Synthetic Minority Oversampling Technique) resampling technique is used in this study to overcome imbalance. Below figure shows the distribution of target variable before and after resampling.

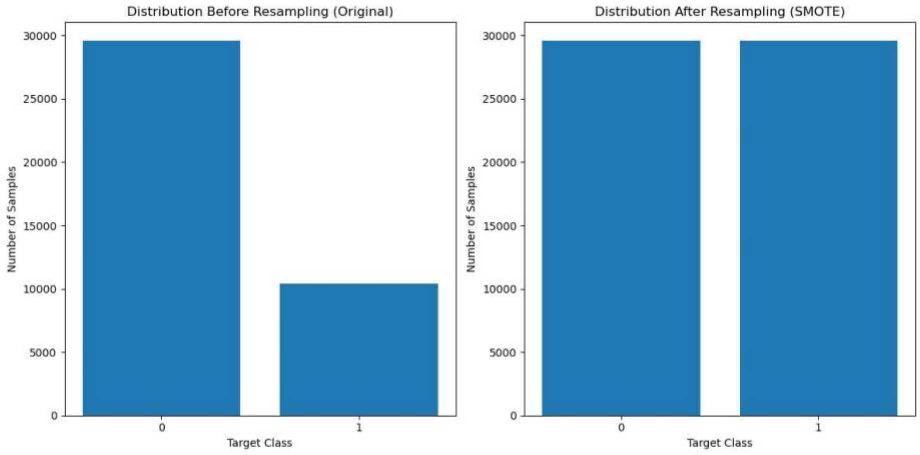


Figure 12 - Effect of SMOTE Resampling

2.4 Model Selection and Development

This study will employ a comprehensive approach to model selection based on accuracy. From a pool of potential algorithms including Random Forest, Logistic Regression, Decision Tree, Gradient Boosting, K-Nearest Neighbours, and XG Boost, two models will be chosen for further evaluation.

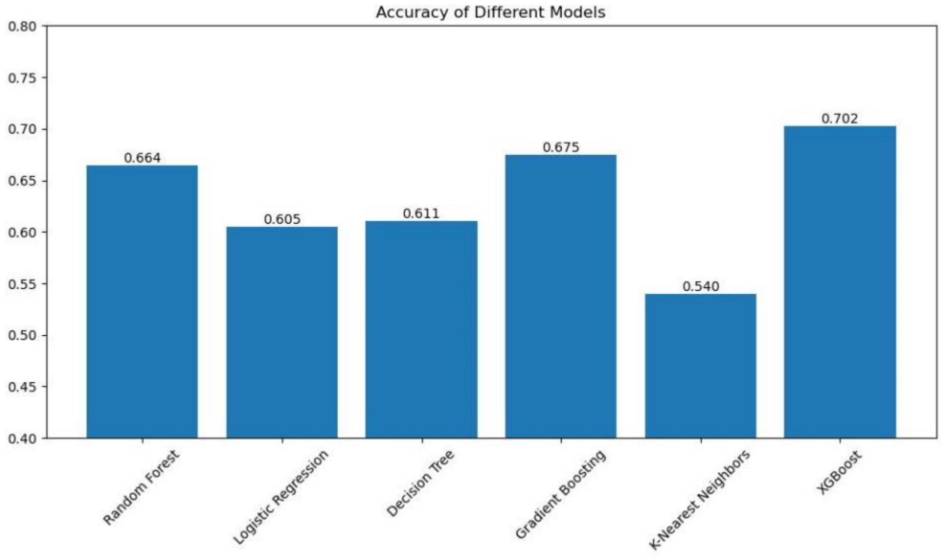


Figure 13 - Accuracy of Different Models

Above graph shows the accuracy for 6 different models on their default parameters. Based on the accuracy score, Random Forest and XG Boost are selected.

2.5 Model Training and Evaluation

2.5.1 Model 1 – Random Forest

Random forest is chosen because it is a powerful ensemble method operates by constructing a number of decision trees, which combine predictions of individual trees to ensure robustness and accuracy. Initially the model is trained on resampled data, employing their default parameters. Next, the hyperparameters of the model are tuned using GridSearchCV. The optimized parameters for the models are,

- Max depth
- Number of estimators

The below figure shows the fluctuation of training and testing the accuracy of the model against maximum depth of the Random Forest.

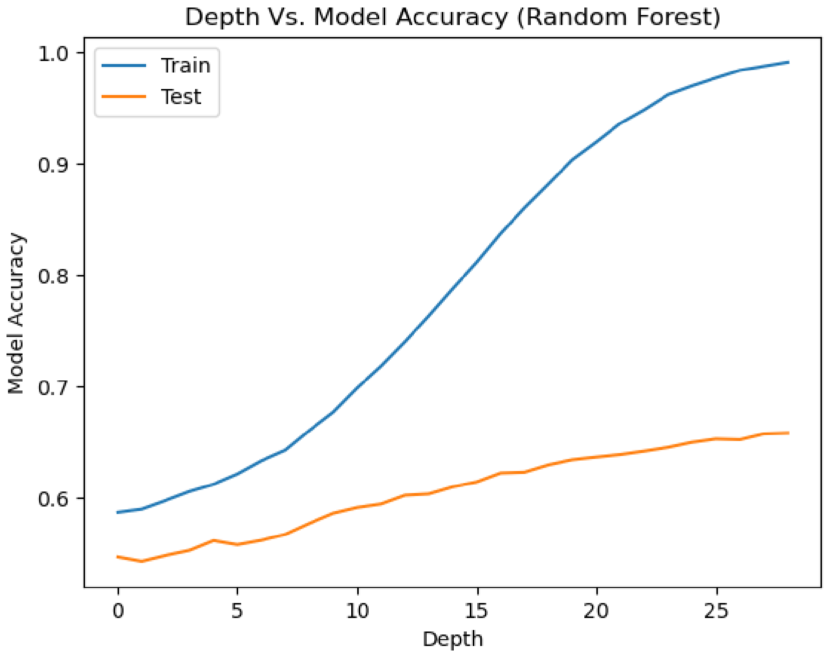


Figure 14 - Depth Vs. Model Accuracy of Random Forest

Based on the above figure the model is underfitting. (Brownlee, 2019)

The below figure shows the change of accuracy against the number of estimators.

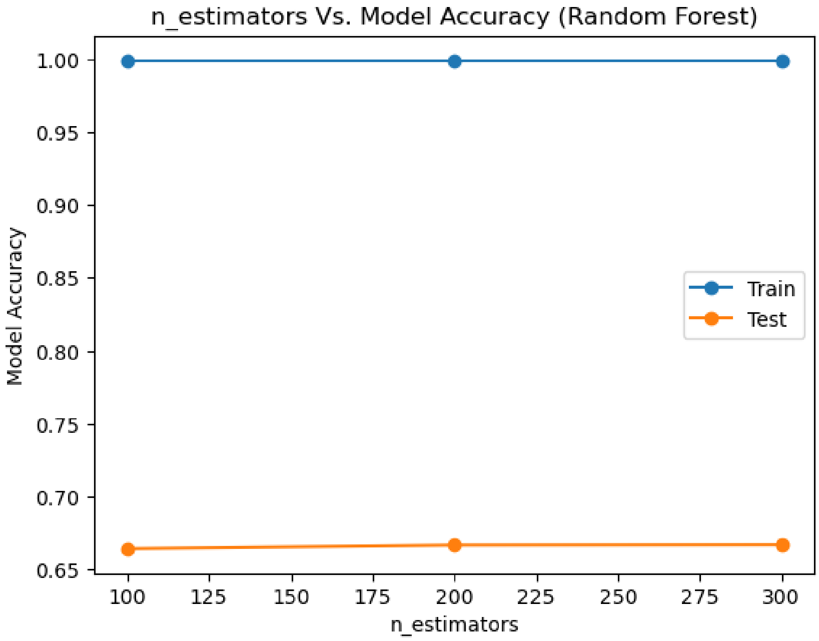


Figure 15 - n_estimators Vs. Model Accuracy of Random Forest

The above figure confirms the under-fitting nature of the model.

2.5.2 Model 2 – XG Boost

XG Boost method is built upon the concept of gradient boosting. It trains multiple decision trees sequentially, with each tree focusing on improving the predictions of the previous tree by addressing its errors.

Same as the previous model, after the training step hyperparameters were optimized using GridSearchCV. Best parameters for the model are,

- Maximum Depth
- Learning Rate
- Number of Estimators

As shown in the below figure, the model is under fitting. (Brownlee, 2019)

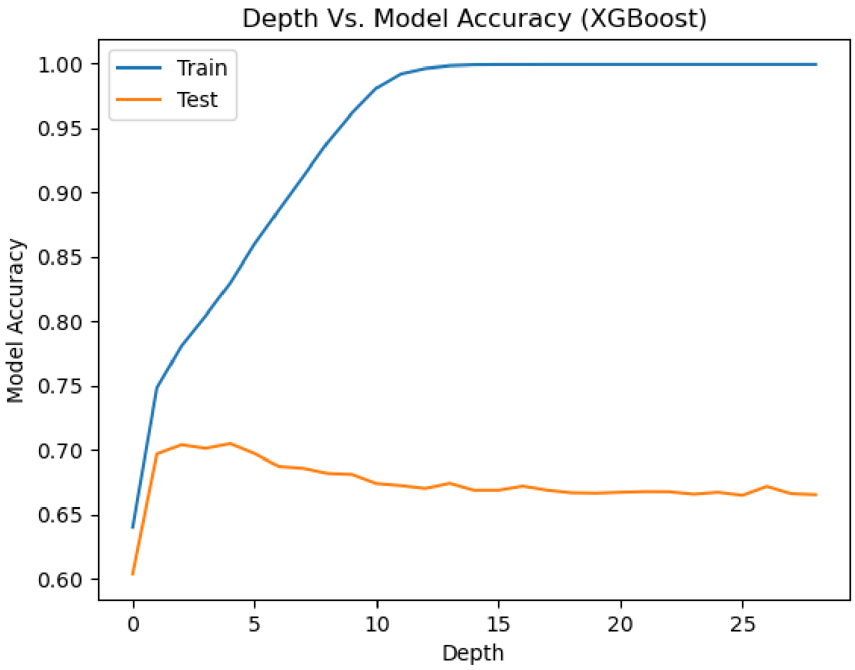


Figure 16 - Depth Vs. Model Accuracy of XG Boost

The below figure shows the change of accuracy against the number of estimators.

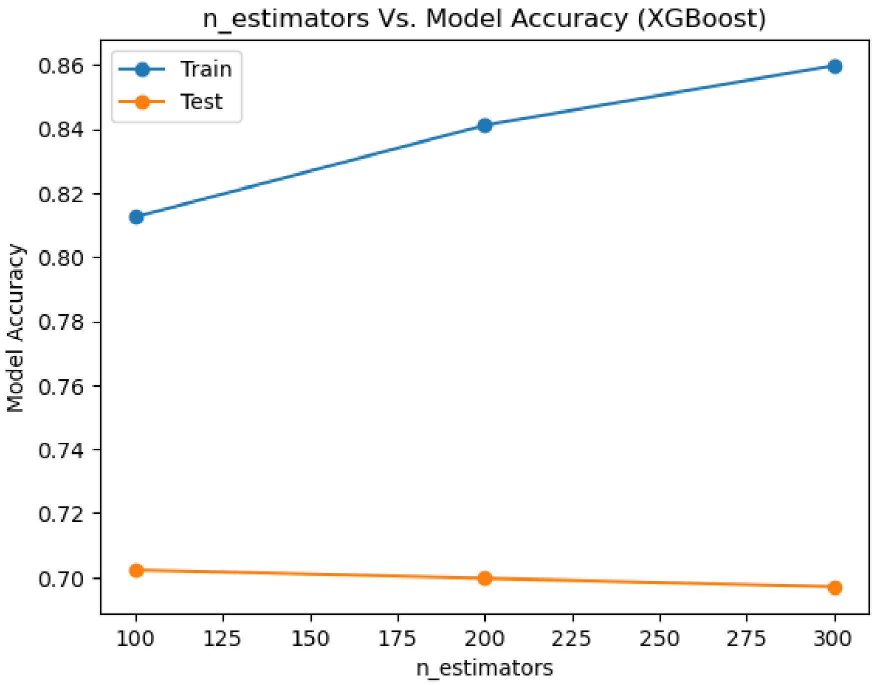


Figure 17 - n_estimators Vs. Model Accuracy of XG Boost

The above figure confirms the under-fitting nature of the model.

3. Model Evaluation, Results and Analysis

3.1 Model Evaluation

Following evaluation metrics are used to evaluate the two models.

1) Confusion Matrix
2) Precision
3) Recall
4) F-1 Score
5) Accuracy
6) ROC AUC

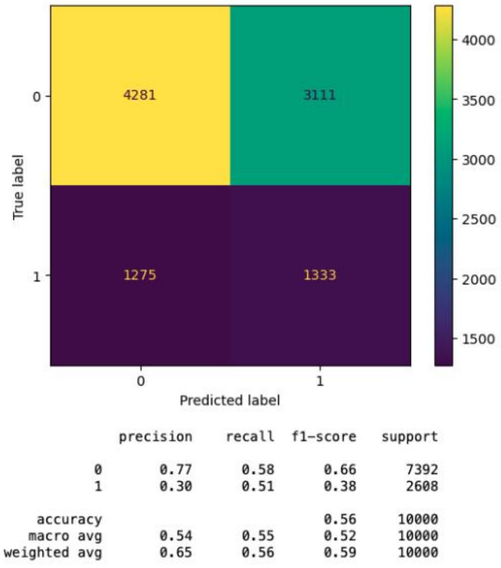
1) Confusion Matrix

Confusion Matrix is an error matrix, containing 4 types of information. True Positive, False Positive, False Negative and True Negative. Using the confusion matrix all the other metrics are calculated.

Table 4 - Structure of a Confusion Matrix

		Predicted Value	
		Positive	**Negative**
True Value	**Positive**	TP	FP
	Negative	FN	TN

Below shows confusion matrices of 2 models and their classification reports.

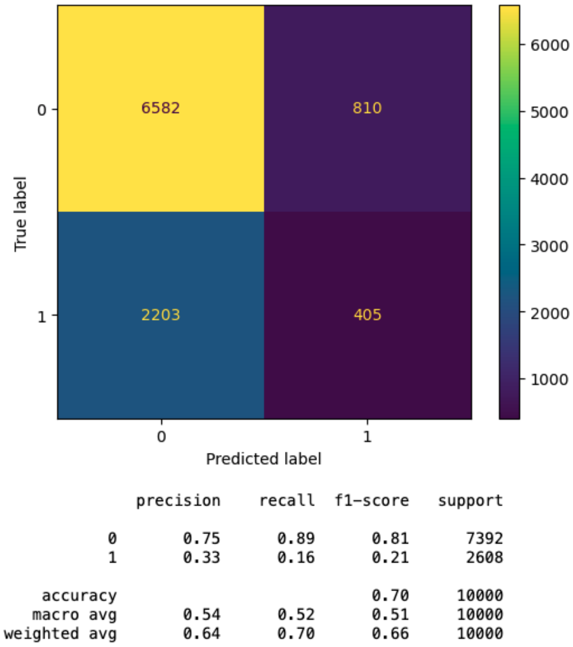


```
              precision    recall  f1-score   support

           0       0.77      0.58      0.66      7392
           1       0.30      0.51      0.38      2608

    accuracy                           0.56     10000
   macro avg       0.54      0.55      0.52     10000
weighted avg       0.65      0.56      0.59     10000
```

Figure 18 - Confusion Matrix of Random Forest (Model 1)



```
              precision    recall  f1-score   support

           0       0.75      0.89      0.81      7392
           1       0.33      0.16      0.21      2608

    accuracy                           0.70     10000
   macro avg       0.54      0.52      0.51     10000
weighted avg       0.64      0.70      0.66     10000
```

Figure 19 - Confusion Matrix of XG Boost (Model 2)

2) Precision

Precision quantifies the proportion of predictions that were truly correct. The weighted average is used to address the class imbalance.

$Precision = (True\ Positive)/(True\ Positive + False\ Positive)$

3) Recall

Recall reflects how well the model can capture all the actual positive cases. The weighted average of recall is used to address the class imbalance.

$Recall = (True\ Positive)\ /\ (True\ Positive + False\ Negative)$

4) F-1 Score

f-1 score is the weighted average of precision and recall.

$F1 = 2\ [(Recall \times Precision)/(Recall + Precision)]$

5) Accuracy

Accuracy is the proportion of correctly classified samples over the whole data set.

6) ROC AUC

Receiver Operator Curve (ROC) and Area Under the Curve (AUC) display the relationship between sensitivity and fall-out. It aggregates confusion matrices across various probability thresholds to determine an optimal probability threshold for classification.

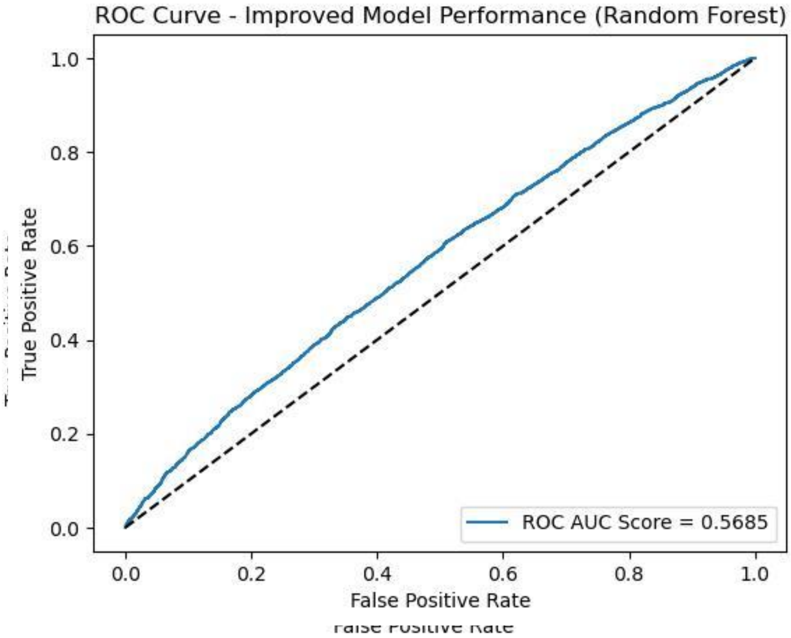


Figure 20 - ROC Curve - Improved Model Performance Model 1 (Random Forest)

Below table shows the summary of evaluation metrics for both models.

Table 5 - Model Evaluation Summary

Evaluation Metric	Model 1 (Random Forest)	Model 2 (XG Boost)
Accuracy	0.56	0.7
Precision	0.65	0.64
Recall	0.56	0.7
F-1 Score	0.59	0.66
ROC AUC	0.568	0.566

When comparing the performance of both models there is a significant difference in accuracy and recall. All the other metrics are falling closely. The main highlight is that ROC value is the same in both models.

3.2 Results Analysis

Both models were underfitting during the development stage. That means the model does not have suitable capacity for the complexity of the dataset. In other words, model cannot learn the training data set.

The purpose of this study was to build a model for financial institutions to predict customers' future potential to default, to minimize the future loss. This can be interpreted using the confusion matrix.

- False Positive (Type 1 error) are where the model predicted positive but truly negative.
- False Negative: (Type 2 Error) are where the model predicted negative but truly positive.

For this study, the most crucial aspect is False Negatives. That means, a customer is truly defaulted, but the model is predicting it as not. This is where the banks lose money. False Positives in other hand can be considered as lost, potential earning opportunities. Because lending money and earning from its interest is a main source of income for banks.

Both models are performing below the acceptable level of performance. Out of these 2 models, Model 2, with XG Boost algorithm performs rather well as it has a higher f1 score, recall and accuracy.

4. Further Analysis Possibilities

Since the current models are under performing, the next best step is to acquire more data related to the study. Both models are seemingly underfitting. As each hyperparameter increases, validation accuracy is slightly increasing as well as the training accuracy.

The quality of the data that was used to build the model could be improved. Alternative data sources also would come in handy (Social media sentiment, purchase history, etc.) depicting a more holistic view of borrowers' financial health.

The next step would be employing advanced machine algorithms like neural networks to train data which might improve overall model accuracy.

Also, dynamic models can be employed which can update themselves in real time, allowing quicker adaptation to market changes.

5. Conclusion

This study explored the development of machine learning models for credit risk assessment. The key challenges tackled include data cleaning, manipulation, addressing class imbalance and hyperparameter tunning. Two models, Random Forest and XG Boost were implemented and evaluated. While both models suffered from underfitting, the second model performed slightly better than the other model. With accuracy and recall at 0.7, f-1 score at 0.66 and AUC score at 0.56, XG Boost model stands to be the better option out of 2.

Future studies should concentrate on acquiring more appropriate data and utilize domain expertise to improve models' overall accuracy and generalizability.

References

01. Brownlee, J. (2019) *How to use learning curves to diagnose machine learning model performance*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/ (Accessed: 22 April 2024).
02. *How to evaluate classification models* (no date) *Edlitera*. Available at: https://www.edlitera.com/en/blog/posts/evaluating-classification-models (Accessed: 22 April 2024).
03. DeNicola, L. (2023) *How is your credit score determined?*, *Experian*. Available at: https://www.experian.com/blogs/ask-experian/how-is-your-credit-score-determined/ (Accessed: 22 April 2024).
04. Langager, C. (no date) *How is my credit score calculated?*, *Investopedia*. Available at: https://www.investopedia.com/ask/answers/05/creditscorecalculation.asp (Accessed: 22 April 2024).