# CREDIT CARD DEFAULT PREDICTION

Prediction of client default

# Introduction

- The goal of this project is to develop a credit card default prediction model using a given dataset. The dataset contains information about credit card clients, including their demographics, credit history, bill statements, and payment records. By analysing this data, we aim to build a predictive model that can accurately predict whether a credit card client will default on their payment or not.

# Problem Statement

- The problem statement revolves around predicting credit card default, which refers to the failure of a borrower to make timely payments on their credit card.

# Stakeholders and the Relevance of Default Prediction

Stakeholders:

- Financial Institutions.
- Risk Management Professionals.
- Credit Underwriters

Relevance:

- Risk Mitigation.
- Profitability
- Customer Satisfaction

# Impact of Accurate Predictions on Business Decisions

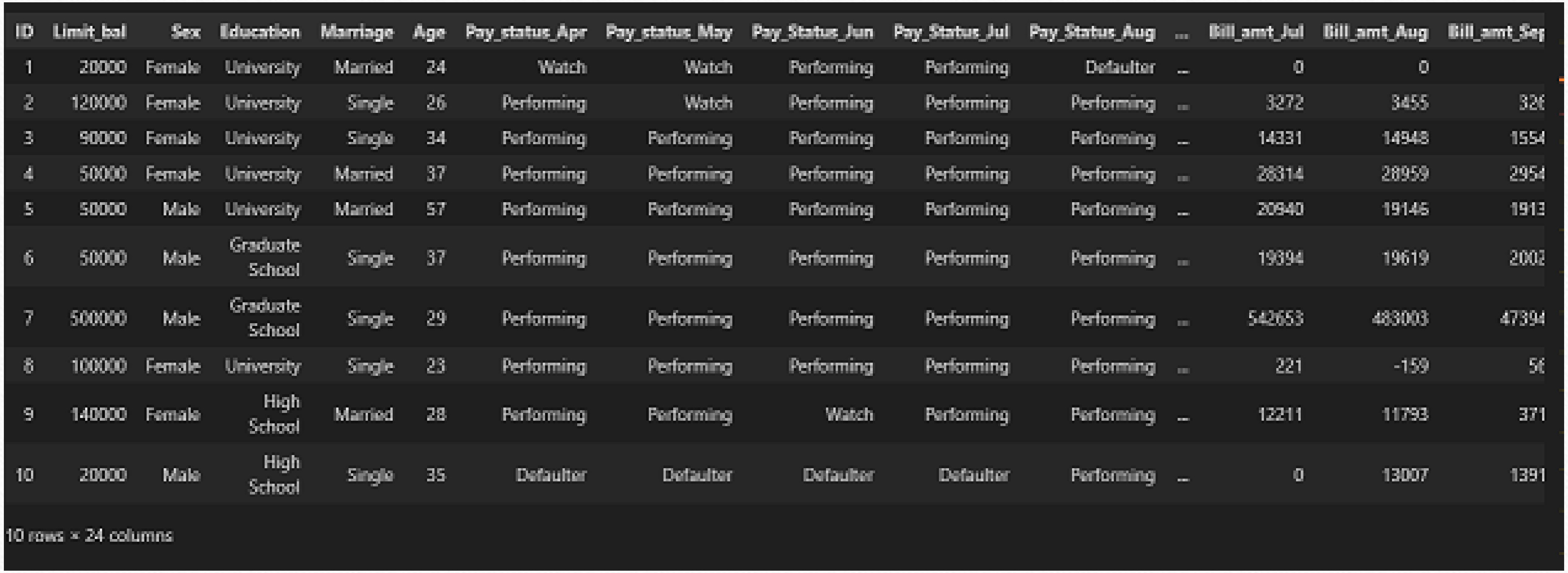
- Credit Approval and Limit Setting.
- Collection Strategies.
- Risk Management and Portfolio Optimization.

# The Dataset

The given dataset is called "default of credit card clients" and was provided by Yeh, I-Cheng (Yeh, (2016). It is available in the UCI Machine Learning Repository.

- **Response Variable:** The response variable in this dataset is "default payment," which indicates whether a credit card client defaulted on their payment or not. It serves as the target variable for our classification model. A value of 1 denotes default, while a value of 0 represents non-default
- **Explanatory Variables:** The dataset contains 23 explanatory variables that are potential predictors of credit card default. These variables encompass a range of client demographics, credit history, bill statements, and payment records. Each variable plays a unique role in predicting credit card default and contributes to the overall predictive power of the model.

# DATA UNDERSTANDING

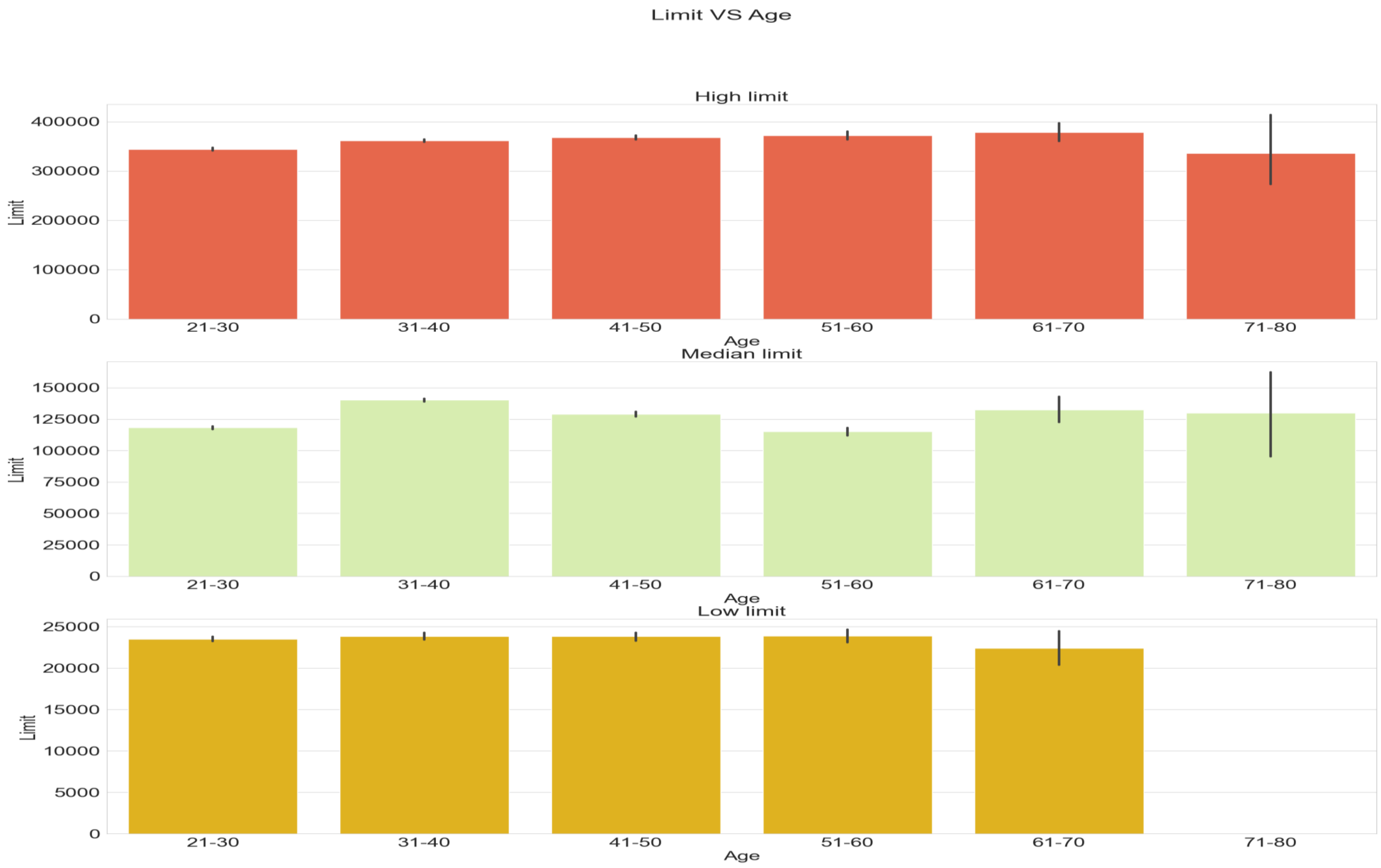


| ID | Limit_bal | Sex | Education | Marriage | Age | Pay_status_Apr | Pay_status_May | Pay_Status_Jun | Pay_Status_Jul | Pay_Status_Aug | ... | Bill_amt_Jul | Bill_amt_Aug | Bill_amt_Sep |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 20000 | Female | University | Married | 24 | Watch | Watch | Performing | Performing | Defaulter | ... | 0 | 0 | |
| 2 | 120000 | Female | University | Single | 26 | Performing | Watch | Performing | Performing | Performing | ... | 3272 | 3455 | 326 |
| 3 | 90000 | Female | University | Single | 34 | Performing | Performing | Performing | Performing | Performing | ... | 14331 | 14948 | 1554 |
| 4 | 50000 | Female | University | Married | 37 | Performing | Performing | Performing | Performing | Performing | ... | 28314 | 28959 | 2954 |
| 5 | 50000 | Male | University | Married | 57 | Performing | Performing | Performing | Performing | Performing | ... | 20940 | 19146 | 1913 |
| 6 | 50000 | Male | Graduate School | Single | 37 | Performing | Performing | Performing | Performing | Performing | ... | 19394 | 19619 | 2002 |
| 7 | 500000 | Male | Graduate School | Single | 29 | Performing | Performing | Performing | Performing | Performing | ... | 542653 | 483003 | 47394 |
| 8 | 100000 | Female | University | Single | 23 | Performing | Performing | Performing | Performing | Performing | ... | 221 | -159 | 56 |
| 9 | 140000 | Female | High School | Married | 28 | Performing | Performing | Watch | Performing | Performing | ... | 12211 | 11793 | 371 |
| 10 | 20000 | Male | High School | Single | 35 | Defaulter | Defaulter | Defaulter | Defaulter | Performing | ... | 0 | 13007 | 1391 |

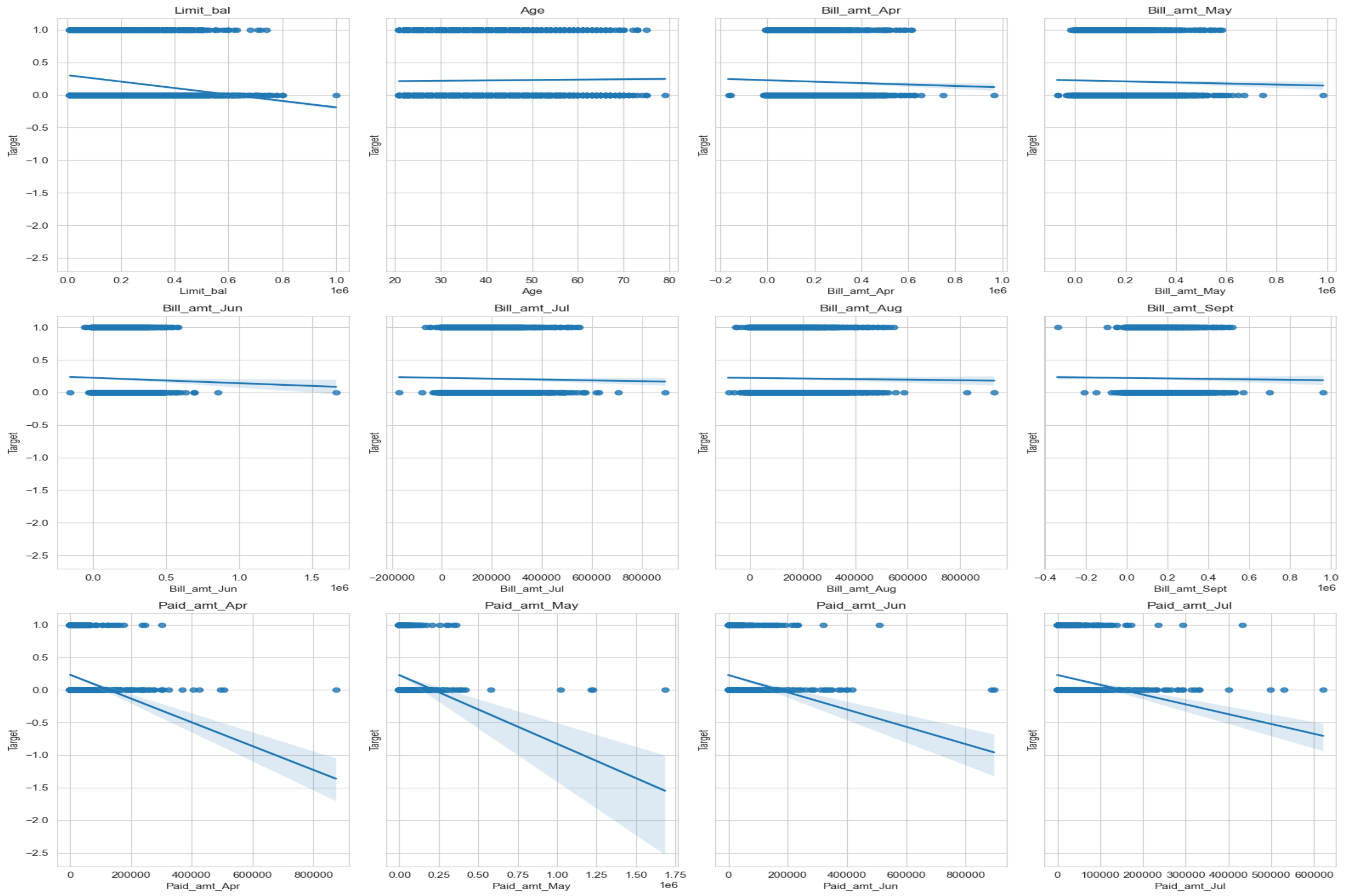
10 rows × 24 columns

# Loan Limit and Age

- We can look at the relationship between loan limit and Age using the graph below.
- We observe that for the distribution of age, the limit is almost evenly distributed, although for the bracket 71-80, they all have loan limits above twenty-five thousand, with the rest having almost an equal number of limits.
- We can still deduce that the outliers present in the data are in the age group 71 - 80. and it would be appropriate to assume the highest limit is also in this bracket.

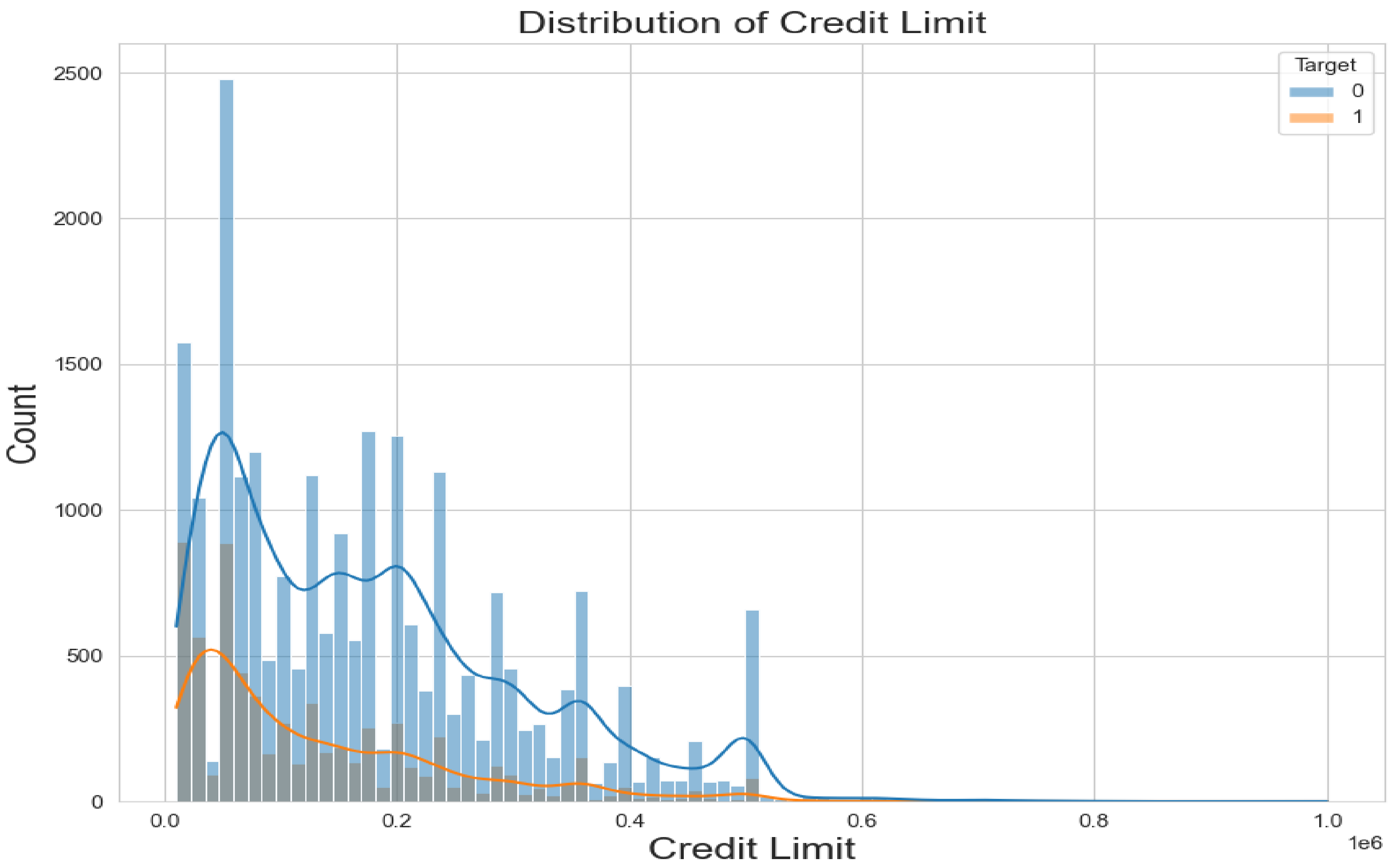

Limit VS Age
High limit
400000
300000
200000
100000
0
Limit
21-30
31-40
41-50
51-60
61-70
71-80
Age
Median limit
150000
125000
100000
75000
50000
25000
0
Limit
21-30
31-40
41-50
51-60
61-70
71-80
Age
Low limit
25000
20000
15000
10000
5000
0
Limit
21-30
31-40
41-50
51-60
61-70
71-80
Age

- The output below is a grid of regression plots, where each plot shows the relationship between a specific feature and the target variable 'Target'. The plots can help visualize the linear relationship, if any, between the features and the target variable, and provide insights into the potential predictive power of the features.


Limit_bal
Age
Bill_amt_Apr
Bill_amt_May
Bill_amt_Jun
Bill_amt_Jul
Bill_amt_Aug
Bill_amt_Sept
Paid_amt_Apr
Paid_amt_May
Paid_amt_Jun
Paid_amt_Jul
Target
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
0.0 0.2 0.4 0.6 0.8 1.0
1e6
20 30 40 50 60 70 80
−0.2 0.0 0.2 0.4 0.6 0.8 1.0
1e6
0.0 0.2 0.4 0.6 0.8 1.0
1e6
0.0 0.5 1.0 1.5
1e6
−200000 0 200000 400000 600000 800000
0 200000 400000 600000 800000
−0.4 −0.2 0.0 0.2 0.4 0.6 0.8 1.0
1e6
0 200000 400000 600000 800000
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75
1e6
0 200000 400000 600000 800000
0 100000 200000 300000 400000 500000 600000

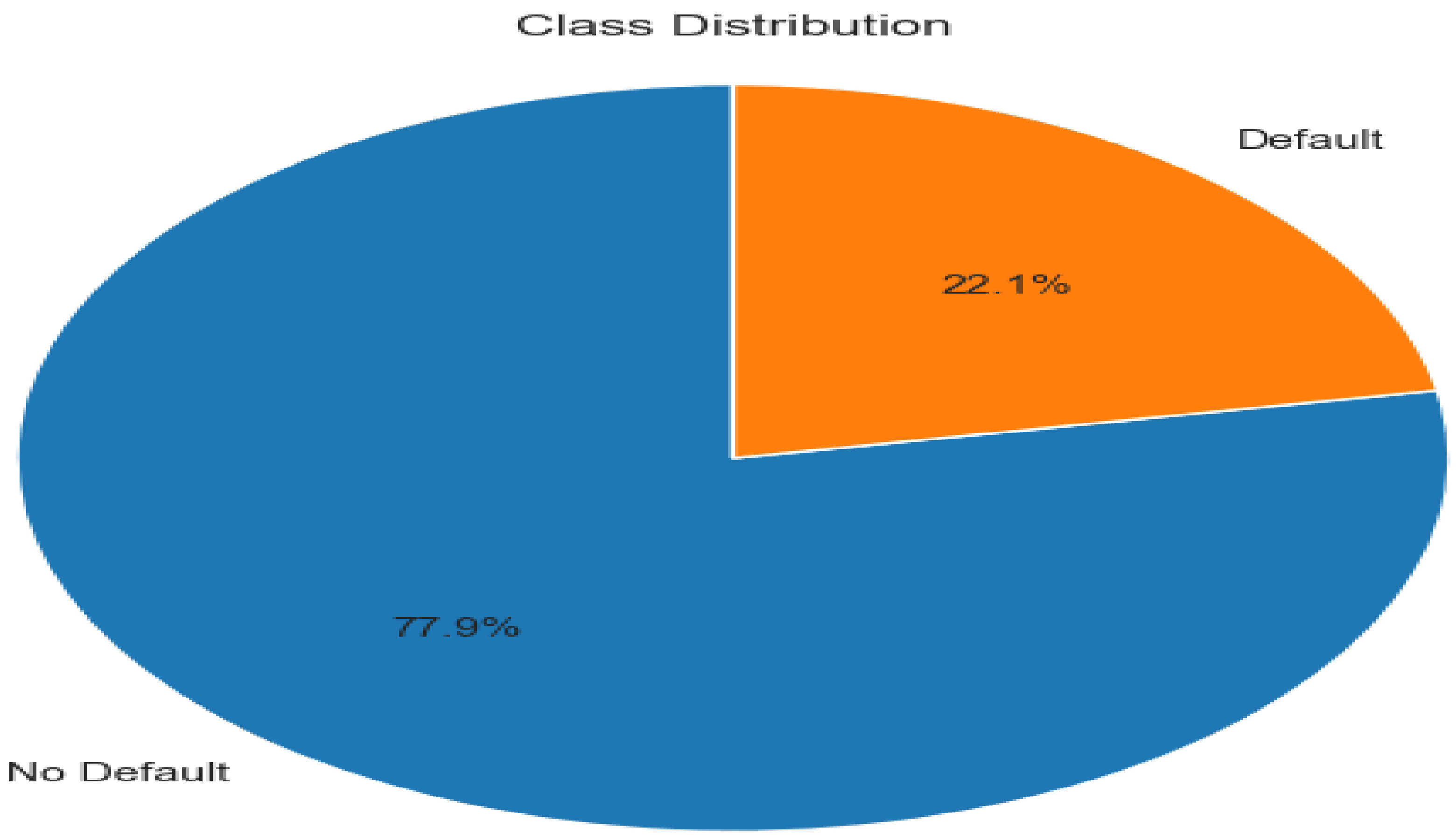
# Data Correlation

- The correlation coefficients range from -1 to 1, with -1 indicating a strong negative correlation, 0 indicating no correlation, and 1 indicating a strong positive correlation. The provided correlations indicate the strength and direction of the linear relationship between each feature and the target variable. However, it's important to note that correlation does not imply causation, and other factors may influence the relationship between variables.
- Below we will also plot a graph showing the highest corelated variable to our target.


Distribution of Credit Limit
Target
0
1
Count
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Credit Limit

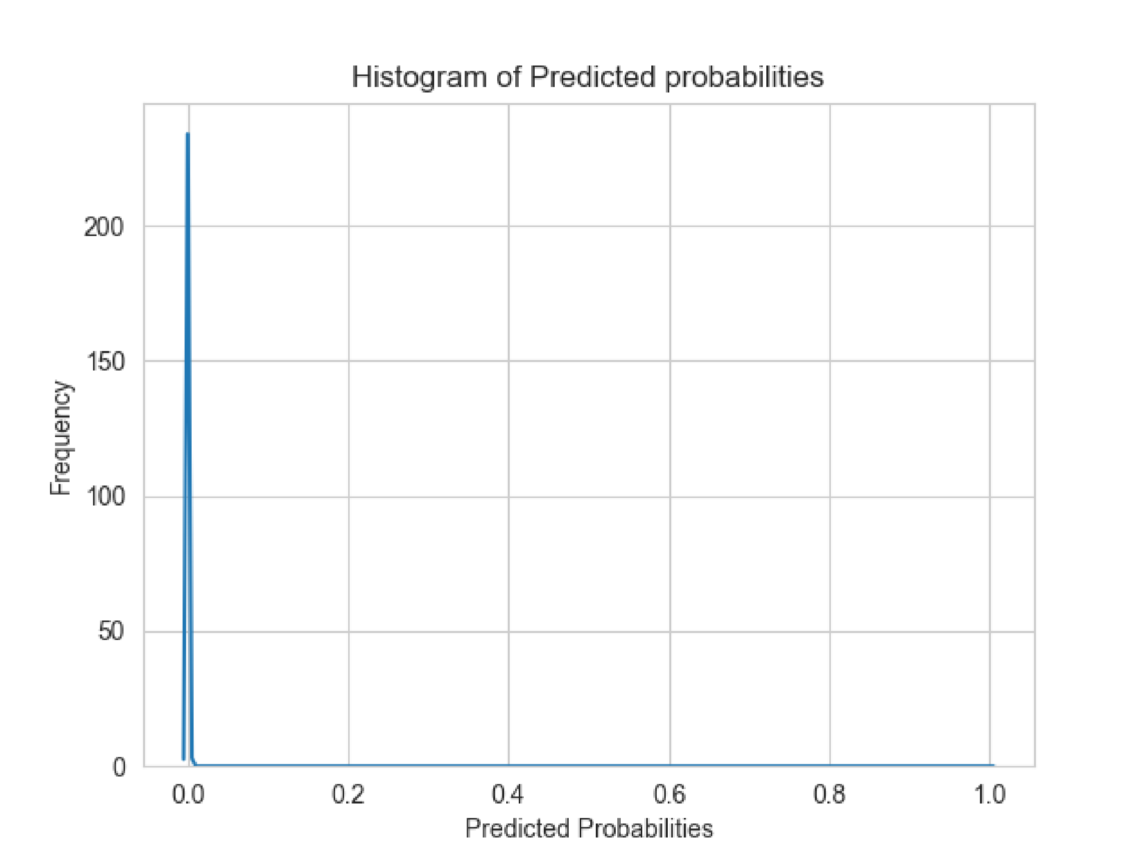
# Class Imbalance Investigation

- The class imbalance is moderate and may require addressing.
- Class Imbalance Ratio: 3.52


Class Distribution
Default
22.1%
77.9%
No Default

# MODELLING

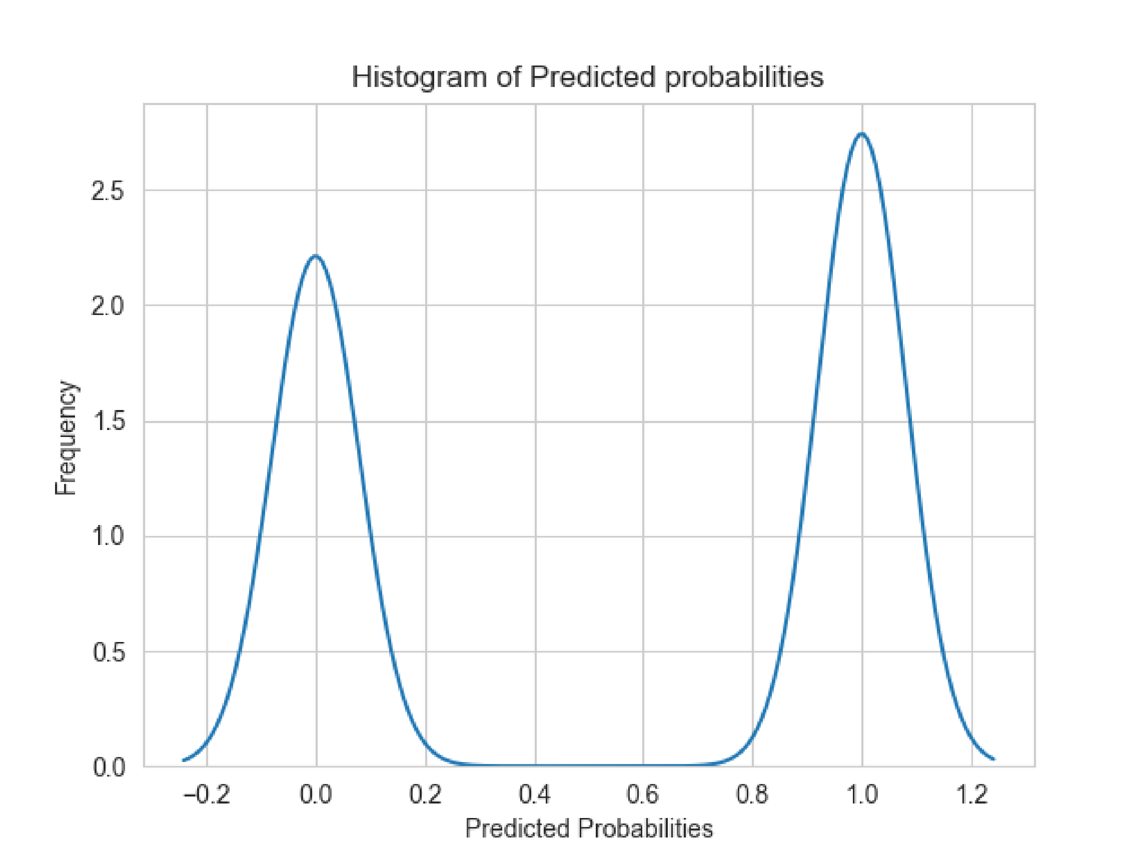
Overall, the results suggest that the model's performance is subpar. It has low recall, indicating that it fails to identify a significant portion of positive instances. The precision is also low, suggesting a high rate of false positives. The F1-score and ROC AUC further confirm the poor performance of the model. Further analysis and improvement of the model may be necessary to achieve better results.



# Classification Model

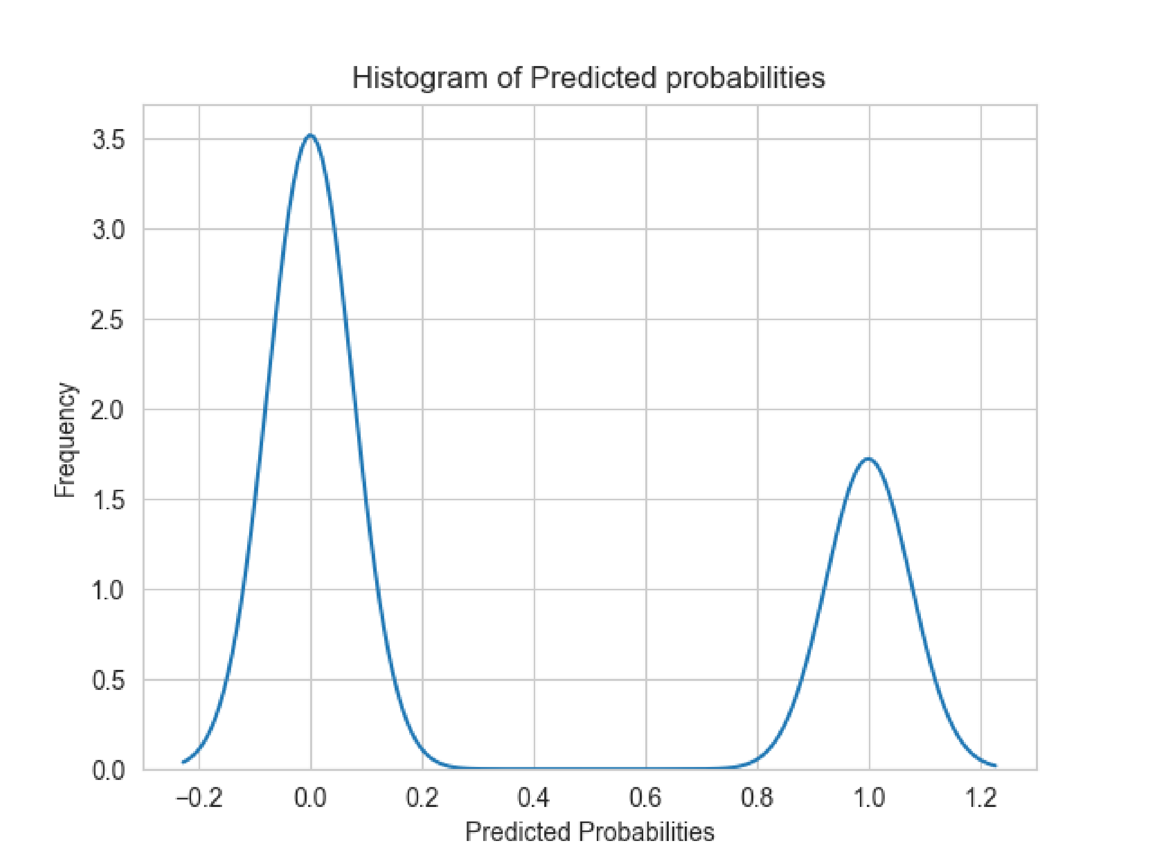
The classifier model has a higher ROC AUC score compared to the baseline model, It suggests that the model can rank positive instances higher than negative instances more consistently than the baseline model.

The Model does improve in performance, but it is not near the score we would want to use as a determiner for policy changes.



# Final Model

By incorporating cross-validation and ensemble methods like bagging and boosting with a Random Forest model, you can potentially improve its performance, reduce overfitting, and enhance its predictive power. Performing the above-mentioned steps, we ascertained that the best our model could predict was 63% only which is not significant enough to make sustainable business decisions.



# RESULTS

The results of our credit card default prediction model indicate that the model's performance is subpar. The logistic regression model achieved an accuracy of 77.35%, which means that 77.35% of the instances were classified correctly. However, the precision, recall, and F1-score are relatively low, indicating room for improvement

- The precision of the model is low, suggesting a high rate of false positives. This means that the model incorrectly identifies a significant number of individuals as likely to default on their credit card payments. The recall score is also low, indicating that the model fails to identify a considerable portion of actual positive instances (individuals who will default). The F1-score, which combines precision and recall, further confirms the poor performance of the model.
- The ROC AUC score, which measures the model's ability to distinguish between positive and negative instances, is close to 0.5. This indicates that the model has poor discriminatory power and is not effectively capturing the underlying patterns in the data.

# RECOMENDATIONS

- Explore additional feature engineering techniques to enhance the predictive power of the model.
- Expand the dataset by incorporating additional relevant data sources
- Experiment with different machine learning algorithms and optimization techniques to find the best-performing model.

- As observed in our analysis, class imbalance can affect model performance.
- Validate the developed credit card default prediction model using an external dataset or real-world implementation.
- Continuous monitoring and updating: credit risk and default patterns can change over time.