Credit Risk Analysis Model

A Machine Learning Model

Created by

## Milindh R Kashyap

## Prashanth Talwar

Contents

[Milindh R Kashyap 1](#_Toc152712230)

[Prashanth Talwar 1](#_Toc152712231)

[Introduction 3](#_Toc152712232)

[Background 3](#_Toc152712233)

[Motivation 3](#_Toc152712234)

[Goal 3](#_Toc152712235)

[Methodology 4](#_Toc152712236)

[Logistic Regression: 4](#_Toc152712237)

[Decision Trees: 4](#_Toc152712238)

[Random Forests: 4](#_Toc152712239)

[Result and Analysis 5](#_Toc152712240)

[Exploratory Data Analysis 5](#_Toc152712241)

[Machine learning Model 14](#_Toc152712242)

[Logistic Regression 14](#_Toc152712243)

[Decision Tree Classifier 15](#_Toc152712244)

[Random Forrest Classifier 16](#_Toc152712245)

[Visualization 17](#_Toc152712246)

[Logistic Regression 17](#_Toc152712247)

[Decision Tree Classifier 17](#_Toc152712248)

[Random Forrest Classifier 17](#_Toc152712249)

[Conclusion 19](#_Toc152712250)

[References 20](#_Toc152712251)

# Introduction

## Background

Financial institutions often face the dilemma of offering credit to individuals with limited or no credit history, a situation that can be filled with risk. Unfortunately, this void is sometimes exploited by opportunistic individuals who strategically default [1] on their obligations by providing little or no Personal information. Enter our innovative machine learning model which is designed to fill this informational void by offering a nuanced way for banks and lending firms to appraise the true creditworthiness of an applicant. By leveraging this model, businesses can make more informed lending decisions, thereby maximizing their revenue and minimizing defaults. Furthermore, this sophisticated tool offers the added advantage of expedited application processing [2], enabling banks to serve their customers more efficiently while mitigating potential risks. Studies show that on an average an education loan takes around 2-10 weeks where the borrowers are stranded without the right amount for their needs.

## Motivation

The 2008 financial recession stands as a stark reminder of the catastrophic consequences of misguided lending decisions.[3] Rooted largely in poor-quality loans, the crisis not only devastated global economies but also underscored the need for better risk assessment tools in the banking sector[4]. Our machine learning model is inspired by these lessons from the past. With an ability to discern genuine creditworthiness, even in the absence of traditional credit data, it aims to prevent such financial catastrophes in the future. This model represents not just innovation, but a safeguard against history repeating itself in the financial world.

## Goal

The goal of our machine learning model is to harness the available data to offer an advanced risk assessment tool for financial institutions. By analyzing both traditional and alternative credit indicators, the model aims to predict an applicant's likelihood of defaulting on their loan. This will enable banks and lenders to make more informed lending decisions, minimizing the risk of bad loans and setting different interest rates based on the borrower’s credit report and previous history.

In doing so, the model not only seeks to bolster the financial stability of lending institutions but also reduces the turn-around time of every loan application, helping eligible borrowers get their loans when they need it the most.

# Methodology

Before using these models [5] we have extensively pre-processed [6] the data by removing the null values, Comparing the features and removing unnecessary features from the dataset.

The dataset at our disposal serves as an example of a binary-classification [7] dataset, primarily employed to check whether a bank should approve or reject a loan based on the borrower's features. Binary classification, a fundamental machine learning task, involves training models to distinguish between two distinct groups [8]. This method is commonly applied in medical contexts, where the objective is to determine whether a patient is afflicted or not. In parallel, our project employs these principles to predict loan approval outcomes, making it as a binary classification task.

Throughout the project, we have adopted a systematic approach utilizing three distinct methods.

## Logistic Regression:

* **Logistic Regression** is a simple and interpretable algorithm that works well for binary classification tasks.
* It models the probability of the target class and is particularly useful when the relationship between the features and the log-odds of the target class is approximately linear.

## Decision Trees:

* **Decision trees** are intuitive and can capture non-linear relationships in the data.
* They are capable of handling both numerical and categorical features.

## Random Forests:

* As Decision tree can overfit when the number of features are more we are using **Random Forests** which can provide better generalization performance compared to a single decision tree.
* They are robust to overfitting and can handle a large number of features.

By comparing the results of these models, we have built the best suiting model for our dataset accounting for underfitting and overfitting.

# Result and Analysis

In the result and analysis section we are going to deep dive into the project where we show every step we have taken to reach the end goal of the machine learning mode.

## Exploratory Data Analysis

We started out by importing all the necessary libraries required for the project and setting the display options for the data frame.



In the next step, we have analysed the data and detected the null values in the data frame by importing the misisngno [6] library and creating a missing number matrix.

A black and white grid

Description automatically generated

After this step we wanted to check how many features had more than 40% null values.

A screen shot of a graph

Description automatically generated

We remove the columns that have a null percent of more than 40%.

We use every column in the dataset and correlate it with the target column to check if the column has any relation with the user being a defaulter or not.

The column ‘TARGET’ has data that shows if the user or the borrower has previously defaulted on his loans. We use this column to compare it with other columns

A diagram of a number

Description automatically generated with medium confidenceA screenshot of a computer screen

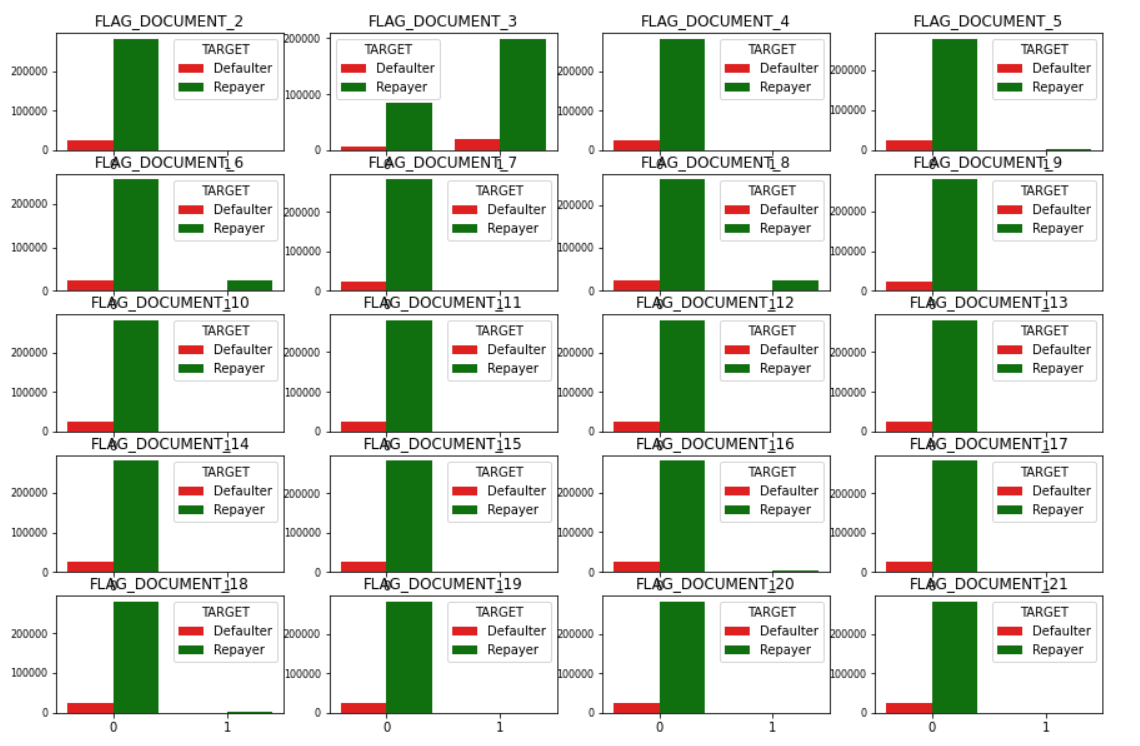
Description automatically generated

The data has features where the borrows occupation type is given and we are harnessing that data to check the occupation type of the applicants

A graph of different colored squares

Description automatically generated

There are features that include variety of documents that had to be submitted during the application process and now we will analyse these features. For this analysis we will be using the itertools library and selecting all the documents and checking which of these documents has the least number of defaulters. By this we will understand that this document is important and not many defaulters come up with this document.



From the visualization we can see that the ‘FLAG\_DOCUMENT\_3’ has the least number of defaulters and hence we will remove all the other features.

Now we detect the outliers for the features like AMT\_INCME\_TOTAL, AMT\_CREDIT, AMT\_ANNUITY, AMT\_GOODS\_PRICE, and analyse the values with other features.

A group of blue and black graphs

Description automatically generated

A graph on a black background

Description automatically generated

We will use the different features to analyse the data.

Here we will use Age, Contract type, Income type, Education type, Family Status, Housing type, Gender and Work Experience

A screenshot of a computer screen

Description automatically generated

We analyse the features and divide each group into a group with payment difficulties and a group with no payment difficulties

A screenshot of a computer screen

Description automatically generated

"In the initial phases of our project, our strategy was to train a machine learning model with the TARGET variable as the target array. Our objective was to achieve a high accuracy, and our initial model appeared to perform well, yielding an accuracy of 85%. However, upon closer inspection, we identified a significant issue with false positives.

Upon further investigation, we discovered a substantial imbalance in the TARGET column, with 92% of the data corresponding to 0 and only 8% to 1. Recognizing the impact of this class imbalance on our model's performance, we took deliberate steps to address it.

To enhance the robustness of our model, we augmented our dataset by incorporating crucial values from an additional CSV file related to the application data. This enriched dataset allowed us to better capture the intricacies of the underlying patterns and relationships. Subsequently, we trained a refined machine learning model, aiming to strike a balance between sensitivity and specificity.

A blue and orange pie chart

Description automatically generated

We also analyse which income group has the most loan application and split it into the genders to see which gender group has more applications

A graph of blue and orange bars

Description automatically generated

Now we create a correlation matrix between the important features to check the correlation between them for the Defaulters and Repayors

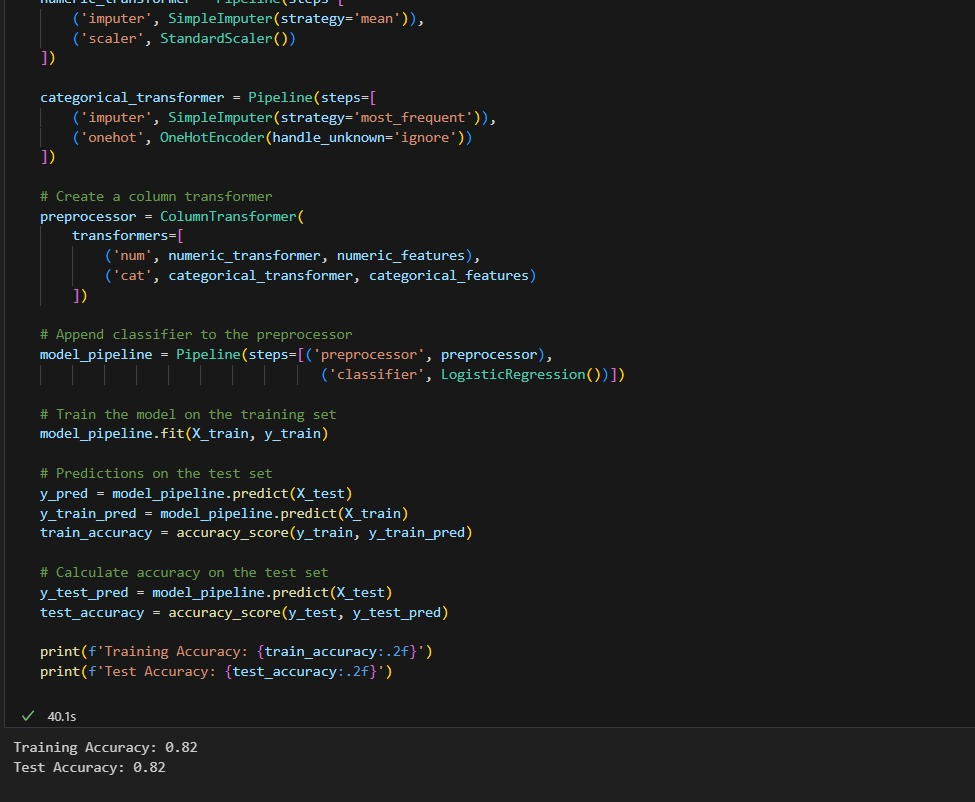
A screenshot of a graph

Description automatically generated A screenshot of a graph

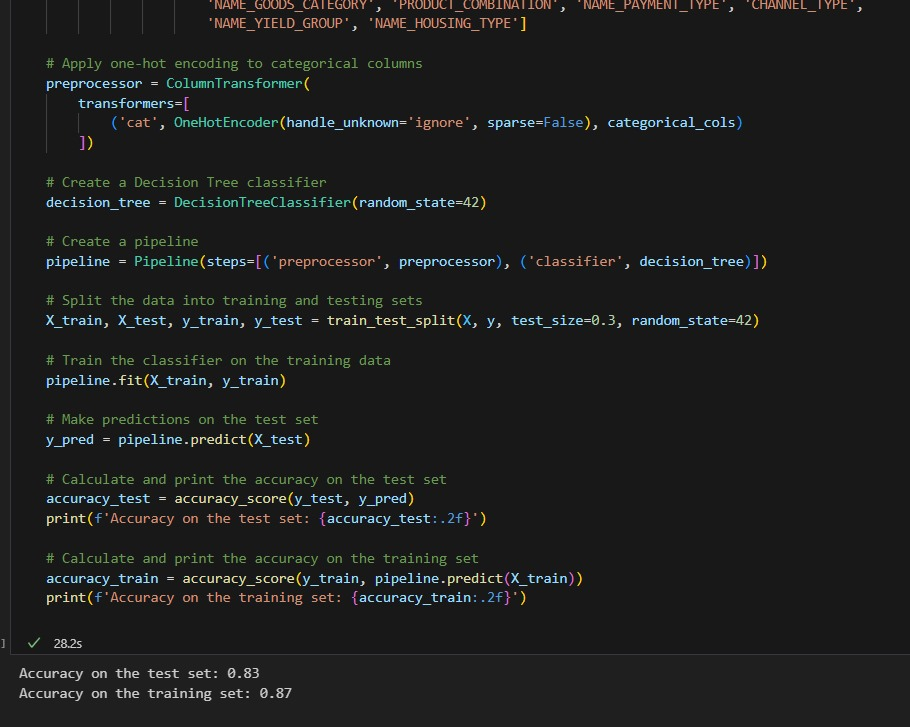
Description automatically generated

## Machine learning Model

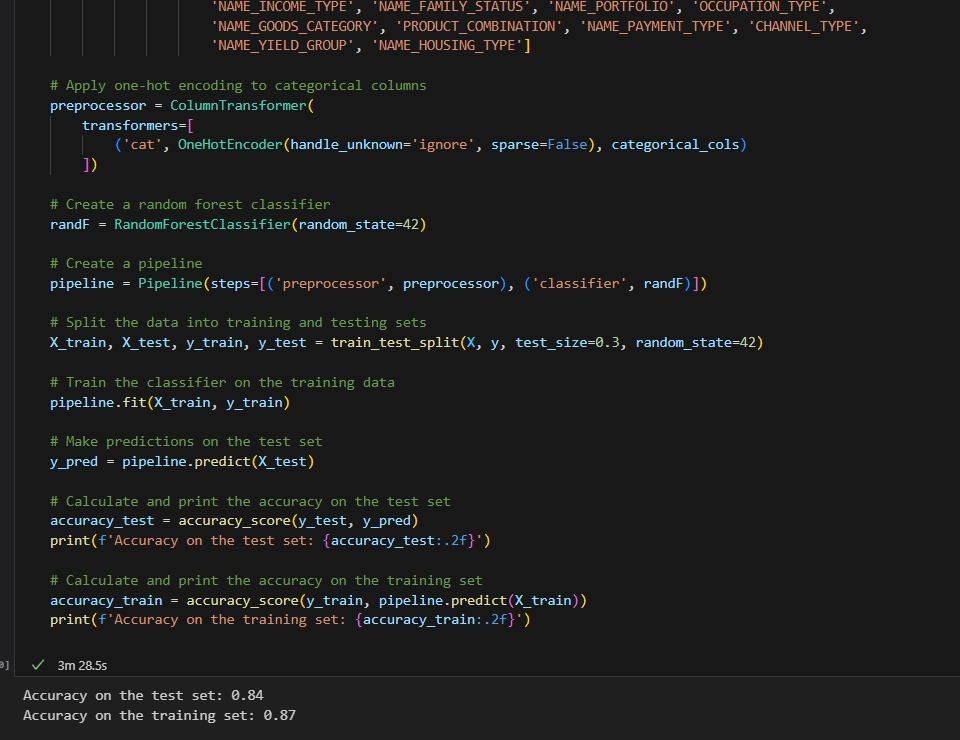
### Logistic Regression



### Decision Tree Classifier

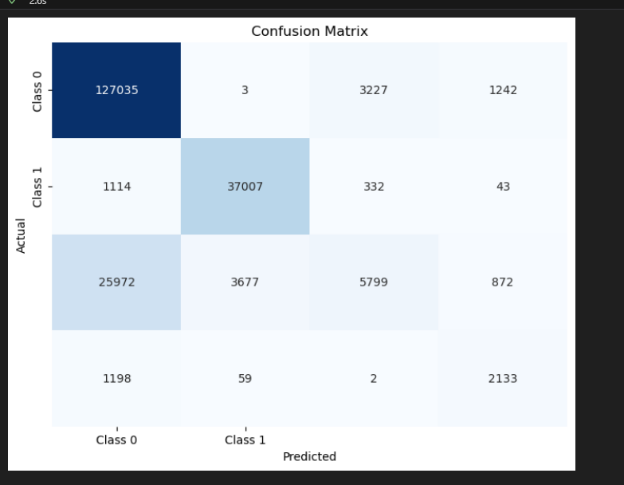


### Random Forrest Classifier

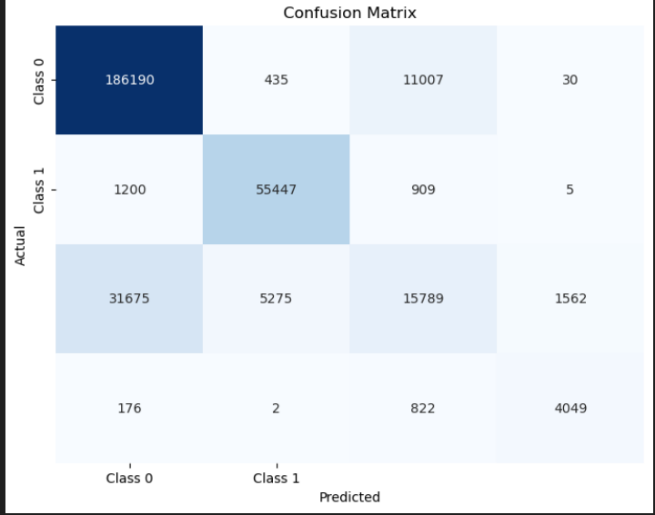


## Visualization

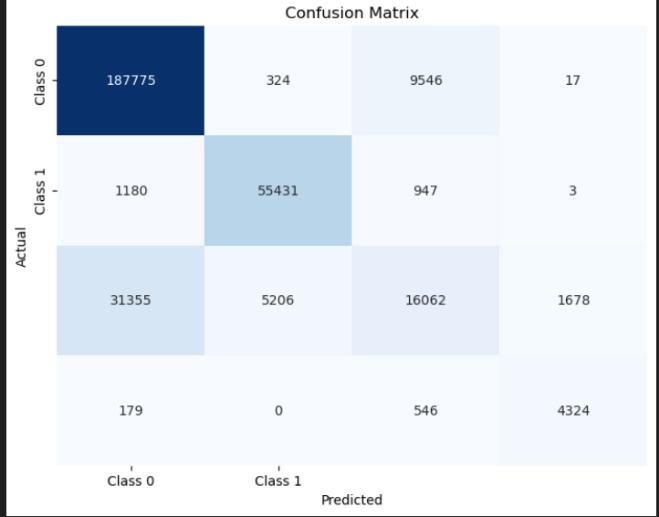
### Logistic Regression



### Decision Tree Classifier



### Random Forrest Classifier



# Conclusion

In conclusion, our machine learning project focused on leveraging a comprehensive dataset containing crucial information about borrowers to develop a robust eligibility assessment system for loans. The success of our project relied on the strategic use of various Python libraries, including NumPy for numerical operations, Pandas for efficient data manipulation, Matplotlib and Seaborn for insightful data visualization, and Scikit-Learn for implementing machine learning models.

Throughout the project, we meticulously processed the raw data, extracting and engineering features that played pivotal roles in determining loan eligibility. This meticulous feature engineering aimed at providing our machine learning models—Logistic Regression, Random Forest Classifier, and Decision Tree Classifier—with rich and relevant information to make accurate predictions.

The implementation of diverse machine learning models allowed us to explore and compare their performances, providing insights into which algorithms were most effective for our specific use case. Logistic Regression, Random Forest, and Decision Tree models were chosen based on their suitability for binary classification tasks, with each contributing unique strengths to the overall predictive power of our system.

As a result of our collaborative efforts, the machine learning models demonstrated commendable accuracy and efficiency in evaluating loan eligibility. The combination of these models with the extensive feature engineering process enabled us to create a reliable and versatile system capable of aiding financial institutions in making informed decisions about borrower eligibility.

Our project not only showcased the effectiveness of machine learning in the lending domain but also emphasized the importance of careful data preprocessing and feature selection in enhancing model performance. Moving forward, this project lays the foundation for continued exploration and refinement, encouraging the integration of more sophisticated techniques and models to further improve the precision and generalizability of our loan eligibility assessment system.

# References

[1] <https://www.forbes.com/sites/robertfarrington/2018/09/25/student-loan-defaulters-strategic-default/?sh=43370930229c>

[2] <https://www.businessinsider.com/personal-finance/how-long-to-get-a-student-loan>

[3] <https://www.federalreservehistory.org/essays/great-recession-and-its-aftermath>

[4] <https://en.wikipedia.org/wiki/Subprime_mortgage_crisis>

[5] <https://kaggle.com/code/venkatasubramanian/credit-eda-case-study-analysis>

[6] <https://www.kaggle.com/code/amritachatterjee09/eda-bank-loan-default-risk-analysis>

[7[] https://www.learndatasci.com/glossary/binary-classification/#:~:text=each%20binary%20classifier-,What%20is%20Binary%20Classification%3F,Application](%5d%20https:/www.learndatasci.com/glossary/binary-classification/%23:~:text=each%20binary%20classifier-,What%20is%20Binary%20Classification%3F,Application)

[8] <https://www.mathworks.com/campaigns/offers/next/choosing-the-best-machine-learning-classification-model-and-avoiding-overfitting.html>