Credit Approval Project

By: Isabel Ernzen, Claire Linn, Cameron Porath

The purpose of this project is to gain a better understanding of the factors that influence credit approval. A person's ability to pay for school, purchase a home, or obtain a car depends on their credit approval, which is a significant aspect of daily life. This is why it's critical to examine the factors that influence these choices carefully. We aim to identify trends that explain why some applications are accepted while others are denied by looking at factors like income, credit score, loan type, and employment status. Utilizing these insights is intended to enhance the decision-making process as a whole and potentially increase the consistency and equity of credit approval.

This dataset contains information about credit applicants, including financial, demographic, and loan-specific details. Key variables include Loan Amount, Loan Duration, Interest Rate, Credit Score, and Monthly Payment, which help assess the borrower's loan terms and creditworthiness. Personal data such as Age, Income, Employment Type, and Seniority provide insight into the applicant's stability. The columns Repayment Status, Approval\_Rate, and Credit\_approval reflect whether the loan is ongoing or paid, the predicted approval score, and the actual approval decision. Together, these fields help analyze factors influencing credit approval. Credit\_approval is the target variable, showing whether their application was approved (1) or not (0).

Before cleaning, the dataset contained 200 records. After identifying and removing rows with missing values using drop, the dataset was reduced to 188 complete rows. This means 12 entries were excluded due to incomplete information in one or more fields. If any duplicates were found, they should also be removed to maintain data integrity. This cleaning process was necessary to ensure that our analysis and model training were based on accurate and complete data, free from potential biases caused by missing or duplicate entries.

This boxplot shows the distribution of three important features: Loan Amount, Income, and Credit Score. Each box represents the range where most values fall, with the line in the middle showing the median. We can see that Loan Amount and Income are widespread, meaning people borrow and earn very different amounts. Some values are much higher than others, which could affect our model. The Credit Score, on the other hand, is more tightly grouped, showing less variation among applicants. This graph helps us spot unusual values, outliers, and understand how each feature is spread out.

This histogram shows how credit scores are distributed among all applicants. The x-axis represents the credit score range, and the y-axis shows how many people fall into each range. Most applicants have scores between 500 and 800, with a noticeable peak just above 500. The distribution is uneven, meaning some credit scores are much more common than others. This helps us understand the overall profile of the applicants and tells us that credit score varies a lot in our dataset

This correlation heatmap shows how different number-based columns in the dataset are related. Each square shows a number between -1 and 1 that tells us how strong the relationship is between the two columns. If the number is close to 1 (red), the two values increase together. If it’s close to -1 (blue), one goes up while the other goes down. Numbers close to 0 (white or light color) mean there’s not much of a connection.

Looking at the graph, we see that Credit\_approval is most strongly connected to Approval\_Rate (0.95), which makes sense because they both describe the same result. To reduce redundancy and simplify the data, we decided to drop Approval\_Rate since a correlation above 0.80 typically indicates that one of the variables should be removed to make less data points; it doesn't add value but increases complexity. Interestingly, Income has a negative relationship with Credit\_approval (-0.63), which could mean that in this dataset, people with lower income were more likely to be approved, possibly because they asked for smaller loans or had other positive factors. This graph helps us understand which features might be important for predicting credit approval.

Predicting credit approval using a Random Forest Classifier, the process begins by loading and examining the dataset, where key features such as income, loan amount, and employment type are identified. Categorical variables like loan type and repayment status are encoded using pd.get\_dummies()so the model can use them. The target variable Credit\_approval is separated, and the dataset is split into training and testing subsets. After training the Random Forest model, performance is measured using a confusion matrix and a classification report, which shows high accuracy (97%), precision, and recall. This means the model is very good at correctly predicting whether a credit application will be approved or not.

The final parts of the code highlight the most important features used in making predictions. A feature importance plot shows that Income and Monthly Payment Amount are the most influential factors in determining credit approval. This tells us that financial stability, as seen in income and ability to make consistent payments, plays a key role in approval decisions. Additionally, the function plot\_confusion\_matrix\_with\_test\_size() encapsulates the entire process and visualizes the confusion matrix using a heatmap, reinforcing the model's strong predictive power and reliability across test sets. This complete workflow gives valuable insights into both how and why decisions are made in credit risk models.