Credit Card Approval Prediction

Data overview:

Credit risk analysis is a fundamental process in the financial industry, allowing lenders to evaluate the probability of a borrower repaying their debts on time. By analyzing historical data and key applicant characteristics, financial institutions can minimize risk while ensuring fair and responsible lending practices. Understanding the factors that contribute to overdue credit payments is essential for developing models that can improve loan approval processes and reduce defaults

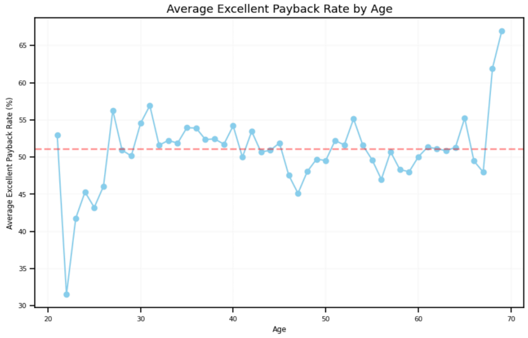
In this project, we aim to identify the factors that influence whether a credit card applicant is classified as a "good" or "bad" client. Using historical data, we analyze various applicant attributes, including years employed, total income, age, education level, family or marital status, and "properties owned". By leveraging machine learning techniques, we built a predictive model to assess creditworthiness, providing valuable insights into the characteristics associated with timely or overdue payments.

To understand how banks determine whether an applicant qualifies for credit, they rely on the 5 C's of Credit:

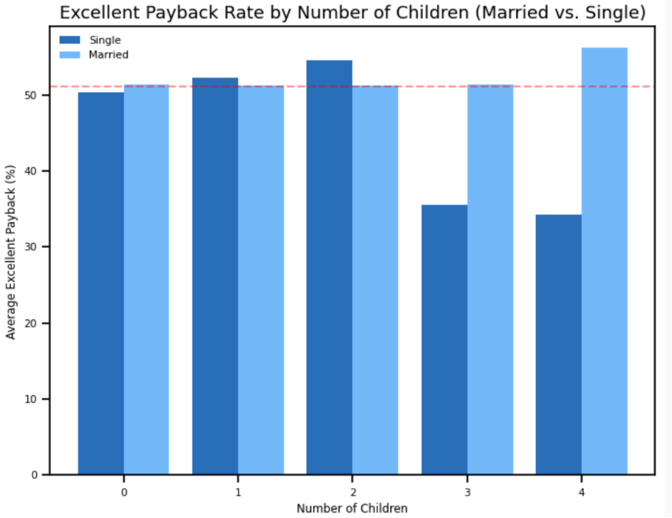
- Character – A borrower's financial history, including credit scores and past payment behavior, which reflects their reliability in repaying debts.
- Capacity – The borrower's ability to repay, assessed through income, employment stability, and existing debt obligations.
- Capital – The money a borrower invests upfront (such as down payments), demonstrating commitment to the loan.
- Collateral – Assets like homes or cars that can secure the loan, reducing risk for lenders.
- Conditions – External factors like interest rates, economic trends, and loan purpose, which impact the likelihood of repayment.

Data exploration:

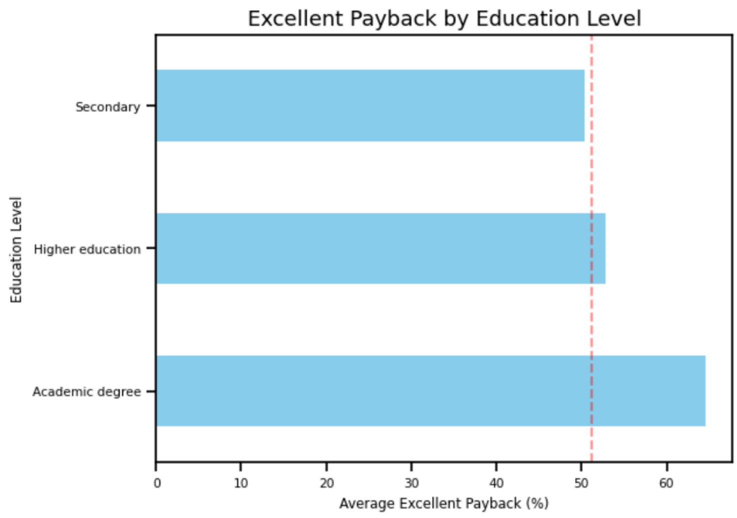
One of the key aspects of credit analysis is understanding the factors that influence an individual's ability to repay their debts. Our analysis revealed several patterns in credit behavior based on demographic and financial characteristics. Age played a significant role, with younger applicants (under 30) exhibiting lower payback rates, while older individuals (above 70) showed the highest repayment consistency. This trend suggests that financial maturity and experience contribute to responsible credit behavior.



Family dynamics also influenced repayment likelihood. While married and single individuals had similar payback rates, those who were single with children had significantly lower repayment rates. Households with dual incomes, such as married individuals with stable employment, were more likely to make payments on time. The graph below illustrates this trend, showing that single individuals with three or more children had a sharp decline in excellent payback rates compared to their married counterparts. This suggests that the financial burden of raising children on a single income may contribute to higher repayment difficulties, reinforcing the importance of household stability in credit risk assessment.



Additionally, education level had a strong correlation with repayment behavior—applicants with postgraduate degrees demonstrated the highest payback rates, while those with only a high school diploma or lower education levels were more likely to miss payments. These insights highlight how socioeconomic factors contribute to financial responsibility and creditworthiness.



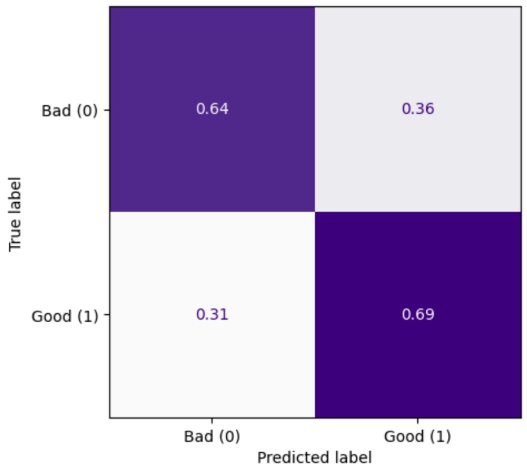
Data preparation:

Before building our predictive model, we prepared the dataset to ensure accuracy and reliability. We first merged multiple datasets using an inner join, ensuring that only applicants with complete data were included in the analysis. Next, we created a target variable by defining applicants as "Good Clients (1)" if their on-time payment rate was ≥ 75% and "Bad Clients (0)" otherwise. This classification allowed us to train the model effectively on distinguishing between high- and low-risk applicants.

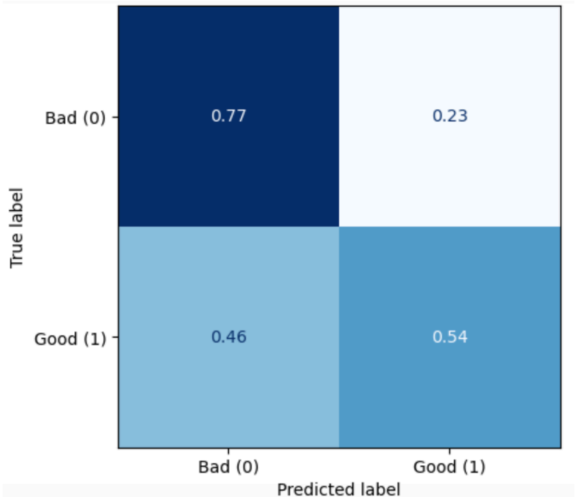
To handle class imbalance, we applied SMOTE (Synthetic Minority Over-sampling Technique), which artificially increased the number of underrepresented cases to create a more balanced dataset. We also processed categorical variables using one-hot encoding, converting features like education level, employment type, and marital status into binary columns for better model interpretation. Finally, we standardized numerical features such as years employed, total income, and age using StandardScaler, ensuring that all numerical data was on a consistent scale. These steps optimized our dataset for more accurate and meaningful predictions.

Models:

To classify applicants based on their likelihood of making on-time payments, we implemented two machine learning models: Decision Trees and Random Forests. The Decision Tree model, with an accuracy of approximately 67%, was effective at identifying true positives—individuals who were likely to repay on time. By splitting the dataset into branches based on key financial factors such as car and home ownership, total income, number of children, and years of employment, the model could isolate patterns in applicant behavior. However, a major drawback of Decision Trees is their tendency to overfit the data, making them highly sensitive to small changes and less generalizable when applied to new applicants.



The Random Forest model, on the other hand, demonstrated a slightly higher accuracy of around 70% and was more robust due to its ensemble nature. By aggregating predictions from multiple decision trees, the model was better at handling true negatives—cases where applicants were more likely to default. It also accounted for non-linear relationships between variables and helped reduce overfitting, making it a more reliable choice for real-world applications. However, the tradeoff for this improved accuracy was interpretability; as the model grew in complexity, understanding the specific decision-making process became more challenging.



Our analysis highlights key factors influencing credit repayment behavior, reinforcing the value of data-driven decision-making in financial assessments. We found that higher education levels correlated with improved on-time payments, while homeownership was a strong indicator of financial responsibility. Conversely, individuals with more children were more likely to struggle with repayments, likely due to increased financial burdens. By leveraging machine learning models like Decision Trees and Random Forests, financial institutions can refine their credit evaluation processes, improving risk assessment and lending strategies.

Conclusion:

In conclusion, credit card analysis plays a crucial role in the financial industry, ensuring that lending decisions are both responsible and data-driven. By understanding the key factors that influence repayment behavior – such as income, employment stability, education, and family structure – lenders can minimize defaults while promoting fair access to credit.

Machine learning and data analytics provide powerful tools to enhance traditional credit evaluation methods, improving accuracy and efficiency in risk assessment. While no model is perfect, leveraging data-driven insights allows financial institutions to refine their lending strategies, protect against financial losses, and support economic stability. As technology continues to evolve, the integration of advanced analytics in credit assessment will remain essential for fostering a more inclusive and responsible financial system.

Github link