Project Proposal

Dataset Chosen and Description

Our project will use the **Credit Card Approval Prediction Dataset**, sourced from Kaggle.

This dataset consists of two related files:

- Application Record Dataset: Contains 438,557 instances and 18 features, detailing applicants' demographic and financial information such as gender, income, employment status, education level, and housing situation.
- Credit Record Dataset: Includes 1,048,575 instances and 3 features, tracking the monthly credit history of individuals, including loan repayment status and overdue payments.

Although the dataset did not specify an officially verified source, it has received high upvotes and a strong usability score on Kaggle, indicating that it is widely accepted and used by the machine learning community. Also, its structured format and large sample size make it suitable for predictive modeling.

Project Title

Our project title is **"Feature Importance in Credit Card Approval Decisions"**. In this project, we will investigate which client features play the most significant role in determining whether a credit card application is approved or rejected.

Motivation

This project is interesting and important because it directly influences both the people's financial opportunities and the stability of financial institutions. Access to credit plays a vital role in economic mobility, as many individuals rely on credit cards for daily expenses, managing cash flow, and building their financial profiles. However, the credit approval process is often complex and lacks transparency, leaving many applicants uncertain about the factors that determine their approval.

A fair and accurate assessment of creditworthiness is essential to ensure that deserving applicants receive access to credit. By identifying the key features that influence approval decisions, we can help to promote wider financial inclusion, allowing more people to access essential financial services and improve their long-term financial health.

On the other hand, financial institutions face the challenge of minimizing risk while making reasonable and consistent credit decisions. Approving unqualified applicants can lead to financial losses, while unfairly rejecting eligible individuals may limit business growth and customer trust. An accurate predictive model can help banks strike a balance between financial risk and customer accessibility, ensuring that approvals are based on objective and data-driven insights rather than subjective reasoning or inconsistent manual evaluations.

Therefore, the benefits of addressing the above problems are as follows:

- For individuals – A clearer and more transparent approval process allows applicants to understand what factors impact their chances of approval. This enables them to make informed financial decisions and improve their credit profiles. A fairer system can also help extend financial services to a broader range of people.
- For financial institutions – An accurate predictive model helps banks assess risk more effectively, reducing the likelihood of approving applicants who may default while ensuring credit is not unfairly denied to reliable individuals. By relying on data-driven decision-making, banks can make fairer, more consistent, and explainable credit approval decisions.
- For the society as a whole – Expanding fair credit access supports economic growth and financial stability by enabling more people to participate in the credit system responsibly. It can also reduce bias and improve trust in financial institutions.

By leveraging machine learning, this project aims to develop a reliable and unbiased approval system that enhances decision-making. Instead of relying on traditional approval methods that may be influenced by personal judgment, outdated policies, or systemic biases, our approach will provide transparent and consistent explanations for credit decisions. This will benefit both customers and banks, ensuring that applicants receive fair evaluations while financial institutions improve their risk assessment processes.

General Approach

All coding will be done in the Python programming language.

Exploratory Data Analysis (EDA):

- Understand the distribution of the features through methods such as bar graphs for the counts of categorical variables and histograms for continuous variables.

Data Pre-Processing and Cleaning

- Merge the 2 given datasets by respective Client ID.
- Identify duplicate entries and look for missing values in the dataset. Decide whether to drop rows with missing values or impute the values.
- Standardise data format across the columns.
- Consider encoding for categorical variables; and consider standardisation/ normalisation for numerical variables.
- Set the standard to be:
    - If the predicted loan is >= 90 days past due → bad client → reject credit card application. This is because most financial institutions flag accounts after 90+ days overdue, as it becomes a serious issue for lenders (serious delinquency with chronic non-payments, severe financial distress or even write-offs).

    - Else if the predicted loan is <90 days past due (which includes loan for the month already being paid off/no loan) → good client → approve credit card application. This is because many lenders still consider <90 overdue days as acceptable risks (mild to moderate delinquency).

- After all of the above is done, perform an 80-20 train-test split of the data, to ensure a good balance between training and testing == ensuring enough data to train, while maintaining a fair test set.
- Subsequently, for the training dataset and the test dataset, perform synthetic data generation (using SMOTE, SMOTE-NC) to ensure a balanced dataset by increasing the number of data points for the minority classes. This is especially crucial for the Logistic Regression model.

Modelling:

- Model choices:
    - Logistic Regression. Why? Because:
        - It is simple and interpretable, as the coefficient of each variable provides clear insights as to how it directly affects the log-odds of credit card approval/rejection in terms of magnitude and direction.
        - It is also computationally inexpensive, and thereby suitable for the large-sized Kaggle Credit Card dataset with 439000 data points.
        - It outputs probability scores, effectively allowing for threshold adjustments to fine-tune approval/rejection decisions.
        - It is also less prone to overfitting, compared to the Random Forest model, especially when we invoke L1/L2 regularisation.
        - HOWEVER, it still comes with certain cons, like assuming a linear relationship between the log-odds of credit card approval/rejection and the independent variables, and hence may fail to capture complex relationships. It also suffers if the different features of a client are highly correlated. We therefore also employ the use of a Random Forest model, to complement this Logistic Regression model.
    - Decision Trees (Bagging: Random Forest). Why? Because:
        - Being an ensemble / averaging of decision trees, the Random Forest model is a form of a bagging algorithm that can thereby reduce variance, offer high predictive accuracy, while still capturing complex relationships well.
        - It also provides clear insights into which factors (e.g., credit score, income, employment history) are most important for decision-making.

The most important factors offering the largest amount of Information Gain (IG) are positioned nearer the top of the Random Forest.

- HOWEVER, it still comes with certain cons, like being much more computationally expensive to make predictions, compared to the Logistic Regression model that computes only a single probability score. It also requires a greater extent of hyperparameter tuning (like the maximum depth of the Random Forest allowed, minimum samples split, and number of trees) to achieve optimal performance, compared to the Logistic Regression model.

- Model training:
    - Apply K-Fold Cross-Validation for hyperparameter tuning.

Evaluation

- Accuracy: Measures the overall correctness of the model
- False Positive: For a risk averse bank, a false positive will lead to financial loss if the person defaults on payment.
- Precision, Recall and F1-Score: These are cost-sensitive evaluation metrics, which can add another dimension to model evaluation. Can potentially be done based on the number of days the client has not paid.
- Micro-Averaging: The data is imbalanced, so micro-averaging is needed to investigate whether the above scores are too optimistic due to imbalanced data.
- Transparency: Whether one can easily tell what are the main reasons that caused the approval/rejection of one's credit card application.

Resources

1. Kaggle Credit Card Dataset Credit Card Approval Prediction
2. Secondary research on which features of a client typically/historically matter the most in determining approval/rejection of bis/her credit card application.