CS3244 Project Proposal
Group 35

Group members:

1. Olivia (Alex) Xiao
2. Shen Tianwei
3. Joshua Harsha Dass
4. Rayan Maknojia
5. Roderick Kong Zhang
6. Tan Jian Hing, Edwin

Mentor’s name: Qihao Liang

To help you plan for your project for timely completion, please fill out the following details on what you plan to do as a group. We have included some prompts for you to consider in your response to the various sections.

Dataset chosen and description Write in your own words (a few sentences) what this dataset is about.

The "Credit Card Approval Prediction" dataset, available on Kaggle, is designed for machine learning tasks and focuses on predicting credit card approval outcomes based on applicants’ financial profiles. It includes features such as demographic details (e.g., age and gender), employment information, and financial attributes like income, debt levels, and credit history. The dataset comprises instances representing individual credit card applicants, with each row providing a unique record of their personal and financial details. The main objective is to use these features to predict whether a credit card application should be approved or denied, making it suitable for classification tasks in machine learning.

Sourced from Kaggle and uploaded by user “rikdifos”, it has been exposed to community use and scrutiny. The dataset’s origin is not explicitly disclosed, but it is described as a synthetic or anonymised collection, possibly derived from financial simulations or historical data aggregations, which helps protect sensitive information. Its reliability is further supported by a perfect usability score of 10.00 from Kaggle, with 100% ratings in Completeness, Credibility, and Compatibility. However, while Kaggle datasets are often valuable for learning and experimentation, it’s important to review any accompanying documentation or metadata provided by the contributor to assess the data’s reliability and any inherent biases.

Project Title: Analysing Factors Affecting Credit Card Approval

Motivation Explain why this project is interesting and important and the problem you aim to address.

This project is centred on understanding and enhancing the factors that influence credit card approval decisions—a challenge that lies at the heart of modern financial risk management. At its core, the project addresses the shortcomings of traditional credit evaluation methods, which can sometimes result in inconsistent or biased decisions. Such limitations not only increase the risk for financial institutions, by potentially leading to higher default rates, but also deny deserving individuals fair access to credit.

For financial institutions, a refined credit approval process can lead to markedly improved risk management. By harnessing data-driven insights, banks and credit card companies can more accurately pinpoint key indicators of creditworthiness. This improved precision reduces the likelihood of defaults, optimises credit portfolios, and streamlines resource allocation, ultimately lowering administrative costs. These advancements contribute to a more stable financial ecosystem and enhanced profitability.

Similarly, individuals benefit from a more equitable and transparent evaluation process. A data-informed approach minimises biases, ensuring that creditworthy applicants receive the consideration they deserve. This broader access to credit, coupled with clearer insights into how credit profiles are assessed, empowers consumers to make more informed financial decisions and work towards improving their credit standing. In doing so, both financial institutions and individuals stand to gain from a system that is fairer, more efficient, and ultimately more inclusive.

General Approach A high-level description of the general approach you'll use to address the questions. Sketch out how you plan to run the necessary analysis and experiments.

1. **Data Acquisition and Initial Inspection** – Import the dataset and review its structure, including the features and instances. The definition of 'good'/'bad' clients is not given, so there is a need for further analysis to self-define the labels; techniques such as vintage analysis could be used to construct the label.
2. **Data Cleaning** – Handling missing values, addressing outliers, fixing data types, removing duplicates
3. **Exploratory Data Analysis (EDA)** – Visualise data distributions and examine relationships between variables using statistical plots and summary metrics, checking for imbalances in data
4. **Feature Engineering** – Create, select, or transform variables to better capture the underlying factors that influence credit card approval decisions, possibly use L1 Lasso regularisation for feature selection
5. **Data Splitting** – Divide the dataset into training and testing sets (and a validation set if necessary) to ensure robust model evaluation (e.g., randomly splitting the data into 30% test and 70% train)
6. **Baseline Modelling:** – Implement logistic regression as an initial, interpretable model to establish a baseline for performance
7. **Advanced Modelling Techniques** – We will explore the machine learning models to enhance predictive accuracy: Linear Regression, Polynomial/Ridge Regression, Random Forest, SVM, and K-means clustering. Ensemble modelling such as bagging and boosting will be conducted if necessary to decrease variance and bias respectively.
8. **Model Evaluation** – We will evaluate the models using mean squared error, mean absolute error and R-squared values. Utilise cross-validation alongside evaluation metrics such as accuracy, precision, recall, and F1-score to assess model performance.
9. **Interpretation and Insights** – Analyse feature importance and model outcomes to identify the key factors affecting credit card approvals, ultimately informing both financial institutions and individuals

Evaluation Indicate how you will evaluate your project, i.e., how will you evaluate how your experiments have turned out?

To evaluate our project, we will assess both the overall success of our analysis and the performance of our machine learning models. Model performance will be measured using standard metrics: mean absolute error (MAE), mean squared error (MSE), R-squared for regression tasks, as these quantify prediction accuracy and explained variance, respectively. Additionally, we will utilise cross-validation alongside evaluation metrics such as accuracy, precision, recall, and F1-score to assess model performance. We will split the dataset into training (70%) and testing (30%) sets to ensure unbiased evaluation. Additionally, we will conduct a qualitative assessment by interpreting model outputs (e.g., feature importance) to validate whether they align with domain knowledge about credit card approvals.

Resources A list of resources you need to conduct the project. This includes additional reading, software, compute, additional datasets, reference code (GitHub links etc) beyond your chosen dataset.

1. **Credit Card Approval Prediction Dataset:** https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction
2. **Credit Approval Dataset:** https://archive.ics.uci.edu/dataset/27/credit+approval\
3. **Cleaned Credit Approval Dataset:** https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data
4. **K-Means Clustering**: Comprehensive Guide to K-Means Clustering (publicly available) https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/
5. **Polynomial Regression and Ridge Regression**: Code and Lecture Materials from EE2211 Introduction to Machine Learning
6. **Lasso Regularisation**: https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-complete-tutorial/
7. **Vintage Analysis**: https://www.kaggle.com/code/rikdifos/eda-vintage-analysis

Role Assignment and Schedule. Provide a schedule of work for the entire team indicating when you plan to complete components of the project. Make sure the schedule is plausible.

- **(Mar 16 - Mar 23): Data Preparation**
  - Olivia: Exploratory Data Analysis
  - Shen Tianwei: Data Cleaning
  - Rayan: Exploratory Data Analysis
  - Joshua: Data Acquisition and Initial Inspection
  - Roderick: Data Acquisition and Initial Inspection
  - Edwin: Feature Engineering
- **(Mar 24 - Mar 30): Initial Model Implementation**
  - Rayan: Logistic Regression
  - Olivia: K-Means Clustering
  - Joshua: Linear Regression
  - Roderick: Polynomial/Ridge regression
  - Edwin: Random Forest
- **(Mar 31 - Apr 6): Model Evaluation and Refinement**
  - Rayan: Logistic Regression
  - Olivia: K-Means Clustering
  - Joshua: Linear Regression
  - Roderick: Polynomial/Ridge Regression
  - Edwin: Random Forest

- **(Apr 7 - Apr 13): Final Synthesis and Report Writing**
    - All members: Compare results, discuss insights, and finalise the report
- **(Apr 14 - Apr 19): Presentation Recording**
- **(Apr 20): Final Presentation**

TA Updates:

1. **Mar 30, 2025:** Complete EDA and initial model runs
2. **April 6, 2025:** Finish model evaluation and preliminary insights

Individual Milestones

Name	Milestone	Deadline
Shen Tianwei	Data Cleaning and Analysis	23 March
Shen Tianwei	Further Data Cleaning if Required	30 March
Alex	Finish Running K-Means Clustering	30 March
Alex	Refine and Evaluate K-Means Clustering	6 April
Rayan	Finish Running Logistic Regression	30 Marth
Rayan	Refine and Evaluate Logistic Regression	6 April
Joshua	Finish Running Linear Regression	30 March
Joshua	Refine and Evaluate on Linear Regression	6 April
Roderick	Finish Running Polynomial/Ridge Regression	30 March
Roderick	Refine and Evaluate Polynomial/Ridge Regression	6 April
Edwin	Finish Running Random Forest	30 March
Edwin	Refine and Evaluate Random Forest	6 April