CS3244 Project Proposal
Group 9

Group members: Zhong Zhu Chen, He James, Lee Melisa, Liu Mingcheng, Syarwina Ridwan, Nur Diyanah Binte Hasan Malik
Mentor's name: Wang XinTong

Project Title

Optimising Accuracy in Credit Risk Prediction: Evaluating Supervised and Unsupervised Learning Techniques for Credit Classification and Credit Scoring

Dataset Description

Credit risk is an important factor of consideration for banks when considering loan approvals and even interest rates. For consumers, this is tracked via a metric called credit scores which quantifies one's credit risk based on past behaviours. This dataset contains a variety of features (e.g. educational level, income, number of children etc) and leaves us to define our own credit metric based on the credit card delinquency status. The dataset originates from Kaggle and it consolidates real-world credit-related data. However, data cleaning and pre-processing are still required to handle missing values and transform categorical features for our machine learning models.

Motivation

The 2008 Financial Crisis was largely triggered by widespread defaults on Mortgage-Backed Securities (MBS), and one of the main causes was due to inadequate tracking of credit scores and evaluating borrower risk before loan approvals (Loo, 2023). This thus not only shows the importance of identifying credit scoring, but also shows the importance of implementing robust risk assessment models that can detect potential defaults early.

Hence, identifying high-risk applicants early via predictive modelling can help institutions proactively manage potential defaults, ultimately contributing to a more stable financial system. This project aims to enhance the accuracy of credit risk assessment by effectively distinguishing between good and bad credit, thereby improving bank loan robustness and contributing to greater financial stability.

General Approach

We aim to train and test both supervised and unsupervised learning techniques to help compare different approaches to identifying risky borrowers. Each method has their own unique advantages and by testing both, it allows us to gain a deeper insight into what really affects credit behaviours.

<u>Data Pre-processing and Exploratory Data Analysis (EDA)</u>

The initial phase involves exploratory data analysis (EDA) to gain insights into the dataset by examining distributions, identifying skewed or unbalanced features, and detecting potential anomalies. This will be done through histograms, summary statistics, and correlation analysis to understand feature relationships and their impact on credit risk.

Following this, we perform general data cleaning and pre-processing to prepare the dataset for both supervised and unsupervised learning models. This includes:

- Handling missing values through imputation or removal where necessary.
- Encoding categorical variables (e.g., one-hot encoding or label encoding).
- Scaling numerical features using normalization or standardization to ensure consistency across different magnitudes.
- Detecting and addressing outliers, particularly in key financial variables such as income and number of dependents.

As supervised and unsupervised models have different feature requirements, additional pre-processing will be tailored accordingly. If necessary, we will also apply resampling techniques to balance the distribution of creditworthiness labels.

Supervised:

Data pre-processing

As we aim to maximise the accuracy of the relation between credit scores and their likelihood of default, we plan to create a continuous credit score based on delinquency records, scaling the longest overdue payment history to a normalized range of [0,1] and mapping it to a credit score between [0,100]. Additionally, we plan to engineer new features such as **c**redit utilisation rate, delinquency trends over time, and length of employment relative to age to increase data size which would theoretically improve our models.

Model Selection and Training

Regression Models:

- Linear Regression & Lasso Regression: To establish a baseline predictive performance and evaluate linear relationships.
- Decision Tree & Random Forest: To capture non-linear relationships and interactions between features.
- LightGBM: A gradient boosting model that is efficient for large datasets and robust in handling missing values.

Implement baseline models and refine them through hyperparameter tuning. Compare model performances and select the best approach based on evaluation metrics.

Model Evaluation

Use Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to measure prediction accuracy. Calculate R^2 score to assess how well our model explains variance in credit scores. Perform residual analysis to ensure model assumptions hold. Conduct feature importance analysis to determine the most influential factors affecting credit scores.

Iteration and Optimization

Apply ensemble learning techniques such as bagging and boosting to improve performance. Conduct extensive hyperparameter tuning to optimize model performance. Validate the model using cross-validation techniques to ensure generalizability. Test different feature engineering strategies to enhance predictive accuracy. Combine multiple models to enhance generalization and reduce overfitting.

Unsupervised:

Data pre-processing

Since unsupervised learning does not require labelled data, minimal pre-processing is needed beyond scaling and encoding categorical variables. However, outliers can negatively impact K-Means clustering, potentially distorting cluster formation. To mitigate this, we will identify and address extreme outliers. Conversely, Isolation Forest is designed to detect anomalies, so we might explore resampling of the dataset closer to its true population distribution based on FRED which is where only about 3% of the population defaults.

Model Selection and Training

Clustering:

- K-Means: Helps to group similar individuals together in order to categorize these individuals into K different risk groups based on continuous variables, making it useful for exploratory risk segmentation without predefined labels.

Anomaly Detection:

- Isolation Forest: An anomaly detection algorithm and by using it, we are able to detect high risk borrowers or anomalies, as credit delinquencies typically account for only a small percentage of all credit card users. Given that credit card delinquency rates have ranged between 1.5% and 3.2% over the past decade (FRED, 2025), it can be shown that real world datasets are highly imbalanced toward good credit. Isolation Forest's ability to handle such skewed distributions thus makes it a robust tool for identifying at-risk borrowers in imbalanced datasets.

Model Evaluation and Hyperparameter Optimisation

For K-Means Clustering, the primary objective is to optimize the number of clusters (K) to ensure meaningful and effective groupings. This will be achieved using methods such as the Elbow Method (minimizing Within-Cluster Sum of Squares, WCSS), Silhouette Score, and Davies-Bouldin Index (DBI) to evaluate cluster quality.

For Isolation Forest, anomaly scores will be used to classify 'Good' and 'Bad' credit based on how easily a data point is isolated. The contamination parameter set based on the true population, will determine the threshold for identifying anomalies, with scores ranging from -1 (anomalous) to 1 (normal). Additional hyperparameters, such as the number of trees, maximum samples per tree, and the number of features used per split, will be optimized for performance. The model's effectiveness will be assessed using Area Under the ROC Curve (AUC-ROC), maximising the True Positive Rate (TPR) against the False Positive Rate (FPR). Additionally, we will visualise the decision score distribution to better understand the model's anomaly detection boundary.

Evaluation

Supervised:

- RMSE (Root Mean Squared Error) to measure prediction errors.
- MAE (Mean Absolute Error) for assessing the average deviation of predictions.
- R^2 to evaluate how well our regression model explains the variance in credit scores.
- Feature Importance Analysis to identify the most influential predictors.

Unsupervised:

K-Means:

- Inertia (Within-Cluster Sum of Squares) to measure how orderly the clusters are within themselves.
- Silhouette Score compares the distance of a point to its own cluster as compared to the distance of the same point to other clusters.
- Davies-Bouldin Index (DBI) measures the average similarity between clusters.

Isolation Forest:

- PCA (Principal Component Analysis) to plot the anomalies in 2D/3D space to visualise decision boundaries and see if they form separate clusters.
- Histograms to check if there is a clear separation between normal and anomalous points by plotting the distribution of the anomaly scores
- Area Under the ROC Curve (AUC-ROC) to maximise TPR and FPR

Overall Unsupervised VS Supervised:

Evaluate models in terms of following criterion:

- Respective accuracy metrics
- Interpretability
- Ability to handle imbalance

Resources

- Python, Jupyter Notebook, Scikit-learn, XGBoost, LightGBM, Tensorflow/Keras (if we end up using deep learning), KMeans, IsolationForest
- Google Colab, local CPU or SoC Cluster
- Kaggle discussions, academic papers on credit scoring materials, ML course materials