MAFS 5440 Project 1: Home Credit Default Risk

Team Members: Zhiyuan CHENG, Anyi QIU, Hangyu ZHOU

Department of Mathematics, The Hong Kong University of Science and Technology

Introduction

This project aims to predict loan default risk using the Home Credit Default Risk dataset. We first run different models with processed data to predict the default probability in test data. And then we conducted feature engineering on income, loans, credit history, payments, and external data to enhance model prediction performance on default risk. A **LightGBM** model was built using **5-fold Stratified Cross-Validation**, evaluated with **AUC**, and retrained with the most important features before submitting predictions.

Model Selection

We first clean and enhance the raw data to provide high-quality input for machine learning models. It involves several key steps: data assessment, transformation and integration.

And then we use a flexible machine learning pipeline that supports various models (e.g., **Logistic Regression**, **Random Forest**, and **LightGBM**), which which are normally used in real-world financial credit work. It uses KFold cross-validation for model training, ensuring robust evaluation. The framework handles data preprocessing tasks such as missing value imputation and feature scaling, followed by model training, feature importance calculation, and performance evaluation using metrics like AUC. Using Kaggle scoring and performance metrics, results are shown in the figure below:

	public_score	private_score	kfold_train_score
logistic	0.67827	0.68469	0.680952
RandomForest	0.72005	0.72219	0.718835
lgbm	0.76666	0.76914	0.776297

Fig. 1: Model Performance

We can observe that **LightGBM** achieves the best performance among the models. Therefore, we will focus on using this model for further exploration.

Feature Engineering Methodolgy

In our further research work, we first focus on extracting, transforming, and constructing features from multiple data sources to improve model prediction. Here's a simplified summary of the key operations:

1. **Income & Loan Features:**
   - Calculate ratios between income and loan amounts, annuities, and goods prices.
   - Compute income per person, annuity-to-loan ratios, and income/debt ratios to assess borrower debt pressure.
2. **Credit History & Loan Status:**
   - Calculate the personal Employment-to-Age ratio, and family Adult-to-Child ratio to understand job stability and family dynamics.
3. **Credit History & Loan Status:**
   - Count loan status frequencies (e.g., bad debts), overdue amounts, and other ratios to assess financial health.
4. **Credit Card & Installment Features:**
   - Analyze overdue days, payment differences, and installment ratios to evaluate repayment behavior.
5. **External Data & Borrowing History:**
   - Aggregate external credit scores and track the number of historical loans, closed loans, and active loans.
6. **Time-related Features:**
   - Calculate differences between loan dates and process application dates (working days vs. weekends).
7. **Log Transformation & Normalization:**
   - Apply logarithmic transformations to monetary values, reducing the impact of extreme values and improving model stability.

Further Research on LightGBM

- **Model Construction:**

  We built a machine learning pipeline using the **LightGBM** model. First, we loaded the cleaned training and test data, extracted the target variable, and performed feature engineering, including handling missing values and feature scaling. Next, we used **5-fold Stratified Cross-Validation** to train the model and evaluated its performance based on AUC scores. During training, we calculated feature importance and selected the most relevant features. Finally, we retrained the model using the selected features and submitted the predictions for evaluation. Based on the results of our model, the ROC curve is shown below:

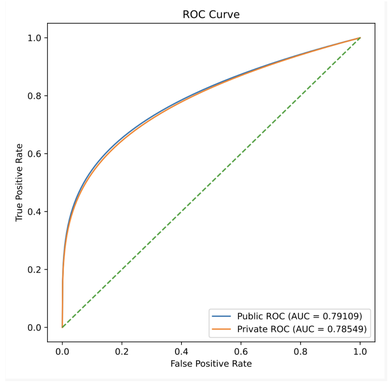

Fig. 2: ROC Curve

- **Model Comparison:**

  Using the basic variables only, cross-validated AUC ranks are consistent across folds: LGBM $>$ Random Forest $>$ Logistic Regression. After adding our engineered signals and selecting the top-500 features, the overall AUC increases slightly but consistently. The gain is most visible in the low-FPR region of the ROC (the credit-relevant zone), and fold-to-fold variance remains small, indicating stable learning rather than noise.

Analysis

Feature Importance: Across folds, the highest mean gains concentrate on external risk scores (EXT_SOURCE*), age/tenure signals (e.g., YEAR_BIRTH), and bureau/previous-application aggregates (e.g., delinquency counts, utilization/ratio features). These patterns are stable across folds. Logistic regression ranks variables differently (linear weights), while both tree models surface similar top drivers; LGBM rankings are slightly sharper after feature engineering, and trimming tail features does not reduce AUC, supporting a parsimonious final set.

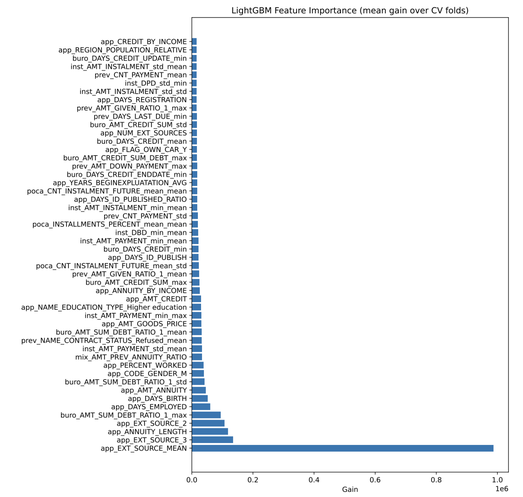

Fig. 3: Feature Importance

Conclusion & Future Work

Conclusion:

- Linear baseline is essential for reference, but underfits.
- RF improves with nonlinearity, yet plateaus.
- LGBM gives the best AUC on this credit tabular task.
- Feature engineering helps slightly; only truly novel signals move AUC.

Future Work:

- Tune LGBM (leaves, LR, regularization; partial-AUC at low FPR).
- Add targeted domain features (utilization trends, recent DPD counts).
- Calibrate probabilities (Platt/Isotonic) for PD use.
- Governance: monotonic constraints, SHAP reason codes, PSI stability & fairness checks.

Contribution

Zhiyuan CHENG: Model Construction & Analysis
Anyi QIU: Model Selection & Result Visualization
Hangyu ZHOU: Feature Engineering & Result Visualization