MAFS 5440 Project 1: Home Credit Default Risk

Team Member:Kedeng QIU ,Yifei SONG ,Jiarui JIANG ,Zewen WAN

Hong Kong University of Science and Technology

1. Introduction

This study uses the Kaggle Home Credit dataset to predict clients' default probabilities through LightGBM with Extra tree. Exploratory data analysis and feature visualization (e.g., scatterplot matrix) help reveal relationships among credit attributes and guide model design.

2. Data analysis

To examine the relationships among key variables used in the credit risk model, a scatterplot matrix was generated for six selected features: `EXT_SOURCE_1`, `predicted_ir`, `nn_oof_single_past`, `AMT_ANNUITY`, `OCCUPATION_TYPE`, and `ORGANIZATION_TYPE`.
The diagonal plots display the kernel density estimation (KDE) of each variable, while the off-diagonal panels visualize pairwise scatterplots. Several patterns can be observed:

- **Positive correlations** are evident among `EXT_SOURCE_1`, `predicted_ir`, and `nn_oof_single_past`, suggesting that both internal and neural-network-based risk scores are consistent with external credit evaluations.
- `AMT_ANNUITY` shows a moderate relationship with the predicted risk, indicating potential links between repayment burden and default probability.
- Encoded categorical variables, `OCCUPATION_TYPE` and `ORGANIZATION_TYPE`, reveal visible clusters, implying different risk distributions across occupations or organization types.

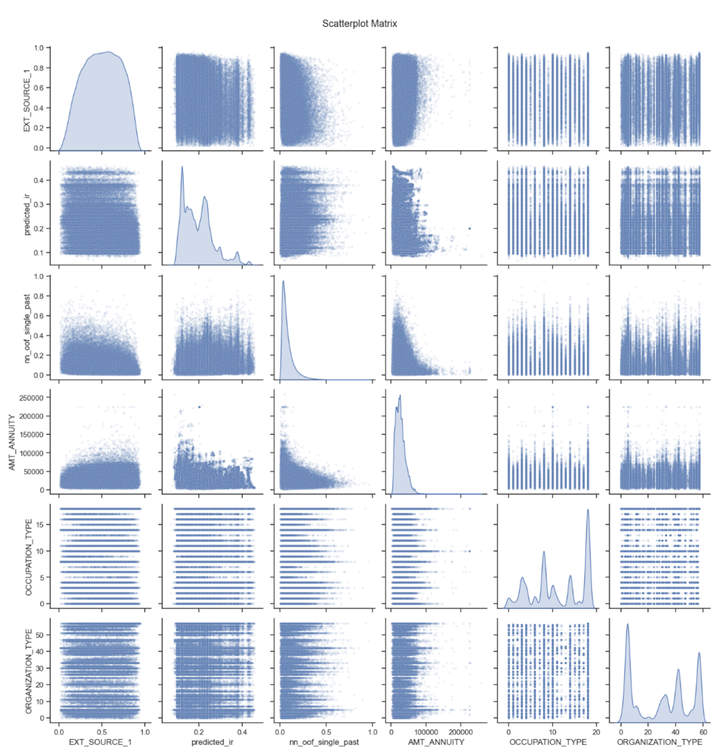

Fig. 1: Scatterplot Matrix

This visualization helps identify inter-feature dependencies, detect potential multicollinearity, and understand the overall data structure, which informs subsequent feature selection and modeling.

Feature Engineering

Data preprocessing

Main table: Cleaned abnormal values, deleted 4 rows with invalid gender codes, and used LabelEncoder to convert text columns.
Subtables:

- **Bureau and balance analysis:** Bureau and balance data were aggregated through SK_ID_BUREAU and then SK_ID_CURR using selected statistical metrics to create new meaningful features.
- **Credit card analysis:** Engineered utilization rates, payment ratios, drawing shares, and delinquency flags from credit card balance data.
- **Previous applications analysis:**Created application-to-credit ratios, approval rates, and estimated interest rates from previous application data.
- **Payment behavior analysis:**Built installment completion ratios, DPD flags, and payment consistency metrics from POS cash and installments payments.
- **Innovation:**Memory-efficient encoding strategy with domain-aware financial ratio engineering, generating 500+ features while maintaining computational efficiency for subsequent machine learning models.

Feature analysis

After examining the missing value distribution, features with extremely high missing ratios were first removed to ensure data quality. Subsequently, an **Information Value (IV)** analysis was performed to evaluate each feature's predictive strength with respect to the target variable (default).
To balance interpretability and feature richness, only variables with **IV > 0.02** were retained. This threshold effectively removes uninformative variables while preserving sufficient predictors for robust LightGBM modeling.

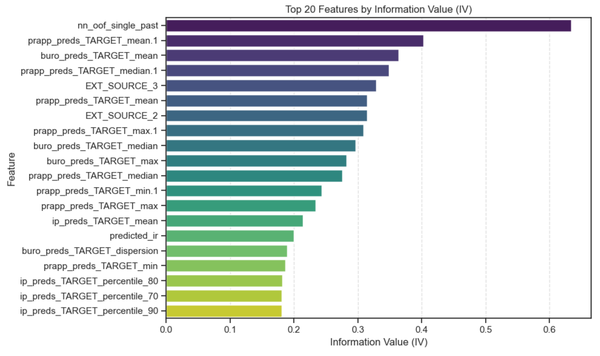

Fig. 2: Top 20 iv-value feature

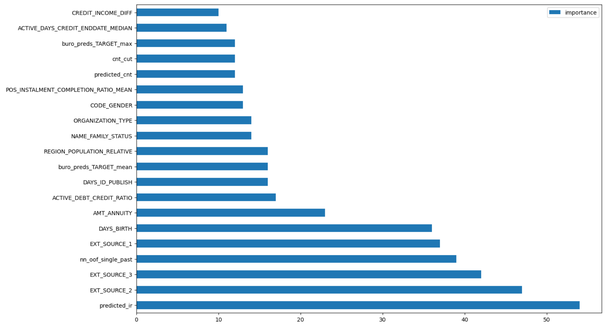

Fig. 3: Feature importance

Model Construction

Method1:LightGBM with Optuna Tuning

Gradient-boosted trees (LightGBM) tuned via Optuna to maximize OOF AUC with stratified K-fold CV, early stopping, and median-pruning. Search over `num_leaves`, `min_child_samples`, `max_depth`, `feature/bagging_fraction`, `lambda_l1/l2`, `learning_rate`. Final model retrained on full data after IV-based feature filtering (IV$>$0.02).Achieve results with validation AUC about 0.78829

Method2:AutoGluon

To improve training efficiency and model performance, we used the AutoGluon package with a multi-layer ensemble learning structure that combines Bagging + Stacking + Weighted Ensemble, achieving excellent results with a public score of 0.80599 and a private score of 0.79656.

Model	Method	Score	Weight
Lightgbm	Bagging_Fold=8	0.7947	0.143
Lightgbm+Extra Trees		0.7971	0.357
Lightgbm	Num_Stack_Levels=1	0.7971	0.214
Lightgbm+Extra Trees		0.7972	0.286
Weighted Ensemble		0.7979	1

Fig. 4: Model Weight by AutoGluon

Conclusion & Future Work

Conclusion

- **Feature engineering** plays a crucial role in enhancing model performance. Newly constructed features, when providing complementary information, significantly improve the predictive power of credit risk models.
- Models with similar architectures tend to select overlapping key variables, while different types of models (e.g., tree-based vs. linear) capture distinct data patterns.
- The integration of **AutoGluon** further demonstrates that automated ensemble frameworks can achieve competitive results with minimal manual tuning, complementing LightGBM's optimized model.

Future Work

- Extend **hyperparameter tuning** using Bayesian or multi-objective optimization to further enhance stability and interpretability.
- Explore **stacked ensemble frameworks** combining LightGBM, AutoGluon, and deep neural architectures for improved robustness.

References

[2] [1]

[1] Nick Erickson et al. *AutoGluon: Towards AutoML with Deep Learning*. `https://auto.gluon.ai`. Amazon AWS AI Labs. 2020.

[2] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017. `URL: https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf`.

Contribution

Data analysis:Zewen WAN
Feature engineering:Yifei SONG ,Jiarui JIANG ,Zewen WAN
Model Construction:Kedeng QIU ,Zewen WAN