Credit Default Risk Prediction with Multi-Table Aggregation and Ensemble Boosting Models

Siqi Wang

HKUST

2025/11/18

Motivation

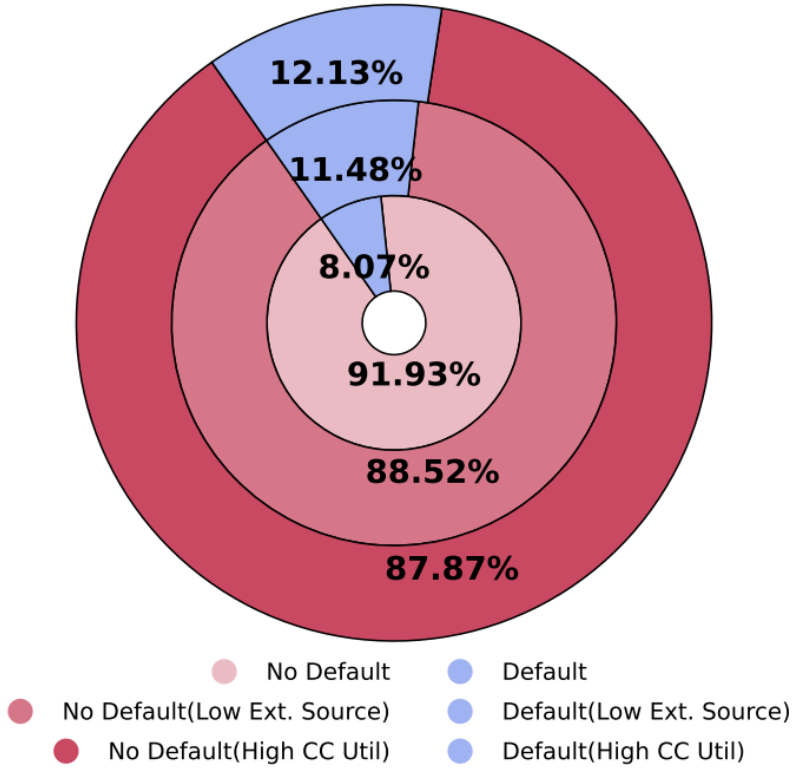
- Credit access essential for financial inclusion.
- Many applicants lack credit history → difficult risk assessment.
- Goal: interpretable and robust default prediction pipeline.

Dataset Overview

- Main application table: demographics, income, loan info.
- Six auxiliary tables: bureau, previous applications, payments, POS, credit card balance.
- Rich behavioral credit signals from historical tables.

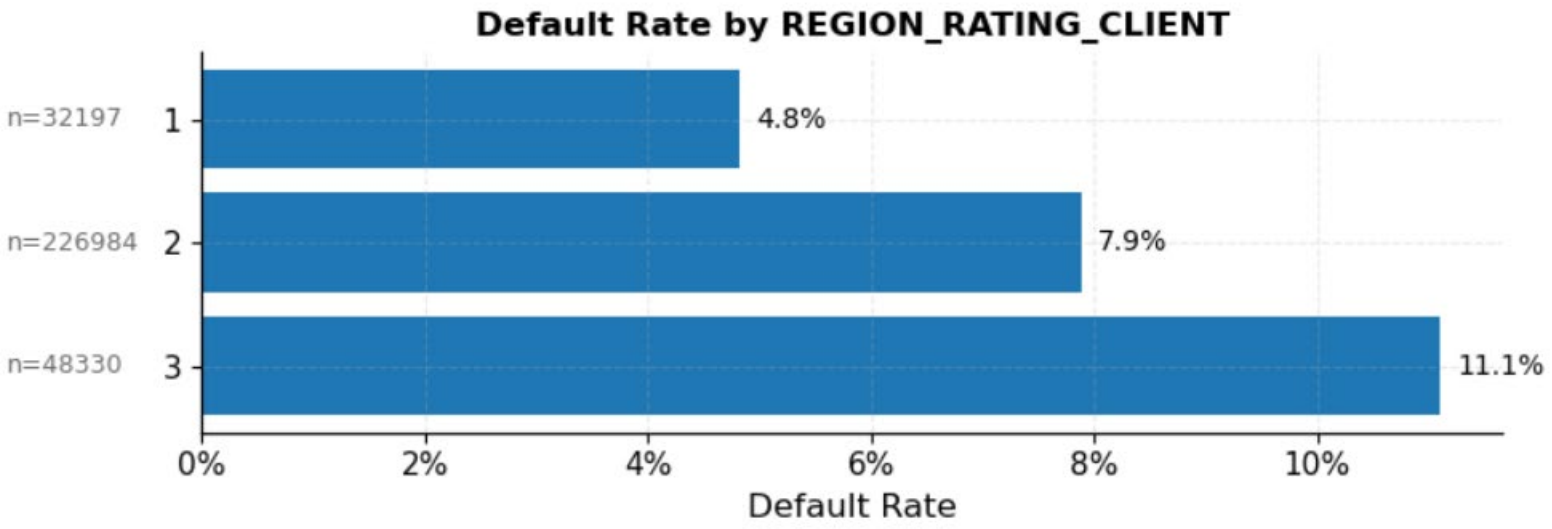
Target Distribution & Imbalance

- Only ~8% default cases → highly imbalanced dataset.
- Accuracy is misleading → use AUC, Recall, F1.
- Heavy missingness in many features.



Key Categorical Features

- Categorical: gender, living city, region, education, occupation type, organization type, refusals strongly predictive.
- Lower education-level, city mismatch, unstable occupation, etc. → higher default risk.



Key Continuous Feature

- Continues feature: age, EXT_SOURCE, delays, past refusal, etc.
- Younger age, lower EXT_SOURCE, overdue, past refusal→ higher default risk.

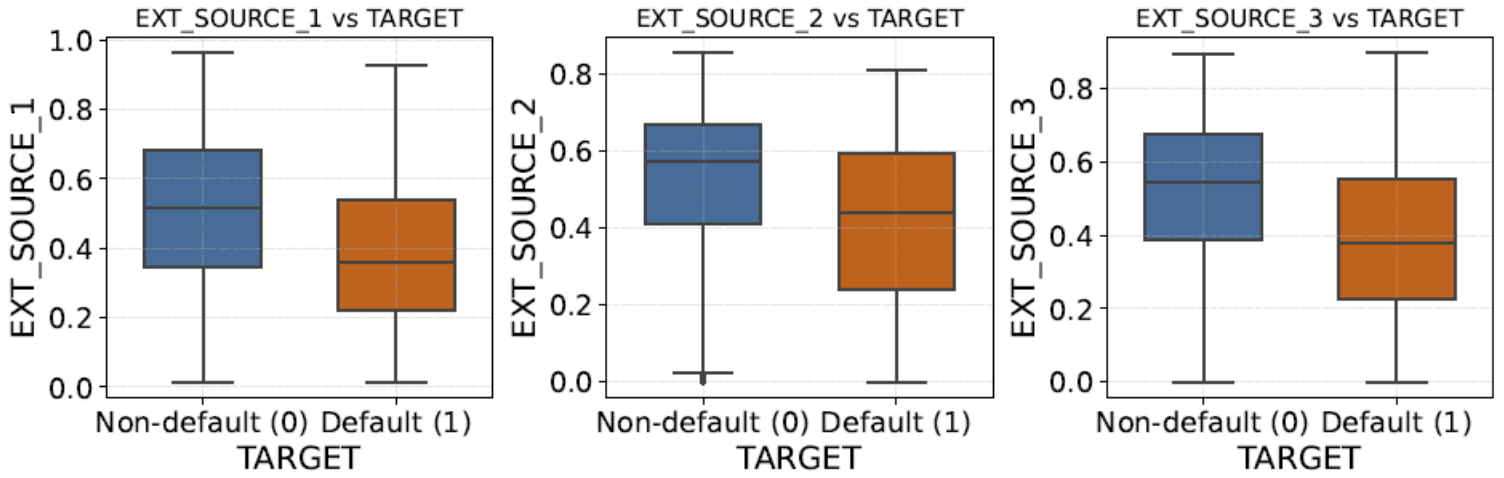
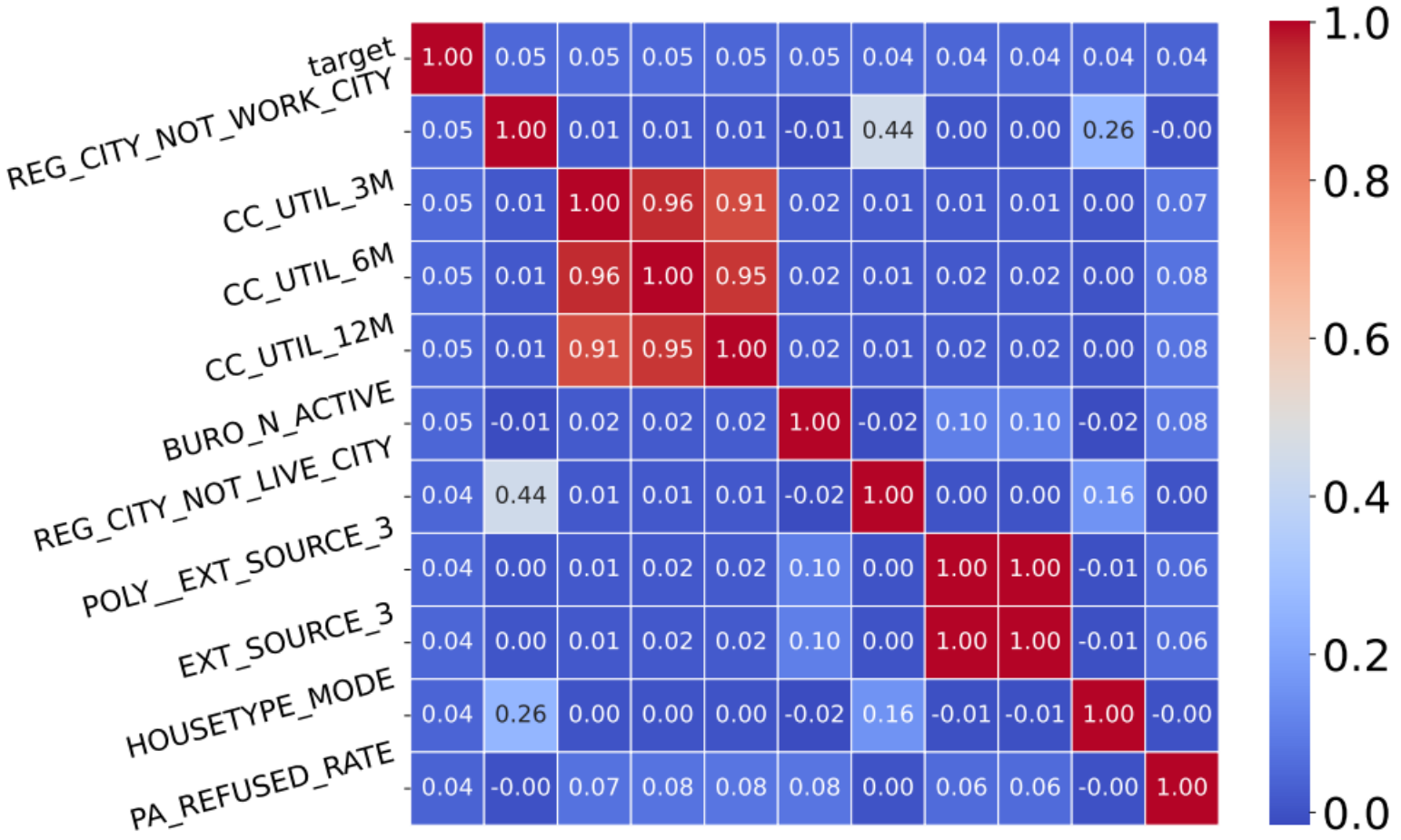


Figure 2: All three EXT_SOURCE features show clear separation across TARGET classes: borrowers with higher external source scores have lower default rates, consistent with their interpretation as creditworthiness proxies.

Continuous Feature Correlation



Summary of EDA Conclusions

- Extreme class imbalance.
- Multi-table aggregation is critical.
- EXT_SOURCE family dominates importance.
- Continuous features are weakly correlated.
- Categorical features:

  CODE_GENDER,
  REGION_RATING_CLIENT_W_CITY,
  NAME_EDUCATION_TYPE,
  OCCUPATION_TYPE,
  REG_CITY_NOT_WORK_CITY,
  ORGANIZATION_TYPE,
  NAME_CONTRACT_STATUS

Method Overview

- Multi-table aggregation.
- Merge features and add degree-2 interactive terms.
- Split data stratified into train/validation sets.
- Train boosting models with early stopping.
- Stack model outputs using logistic regression.
- Optionally prune low-gain features and retrain.
- Average multi-seed predictions for final stability.

Multi-Table Aggregation

- Missing Value Imputation: using the median values.
- Categorical Data Encoding: obtain numeric representations using mapping.
- Feature Scaling: ensure similar ranges and magnitude.
- Polynomial Feature Generation: generate interaction terms and non-linear relationships.

Models & Training

- Models: LightGBM, CatBoost, XGBoost, AdaBoost.
- Stacked using Logistic Regression.
- Early stopping, feature pruning, etc.

Ablation Experiment

- Compared with the App-Train only baseline (AUC = 0.773), final pipeline achieves 2.4 % AUC gain and 62 % improvement in F1-score.

Table 2: Ablation study comparing different feature sets on the app_train dataset. The top-performing model in each metric is highlighted in bold.

Feature Set	AUC	Precision	Recall	F1-Score	Accuracy
App-Train Features Only	0.773	0.529	0.037	0.068	0.919
+ Polynomial Terms	0.773	**0.539**	0.035	0.066	**0.920**
+ All-Table Aggregation	0.791	0.516	**0.061**	0.109	0.919
+ Poly + Aggregation	**0.792**	0.518	**0.061**	**0.110**	**0.920**

Comparison Experiment

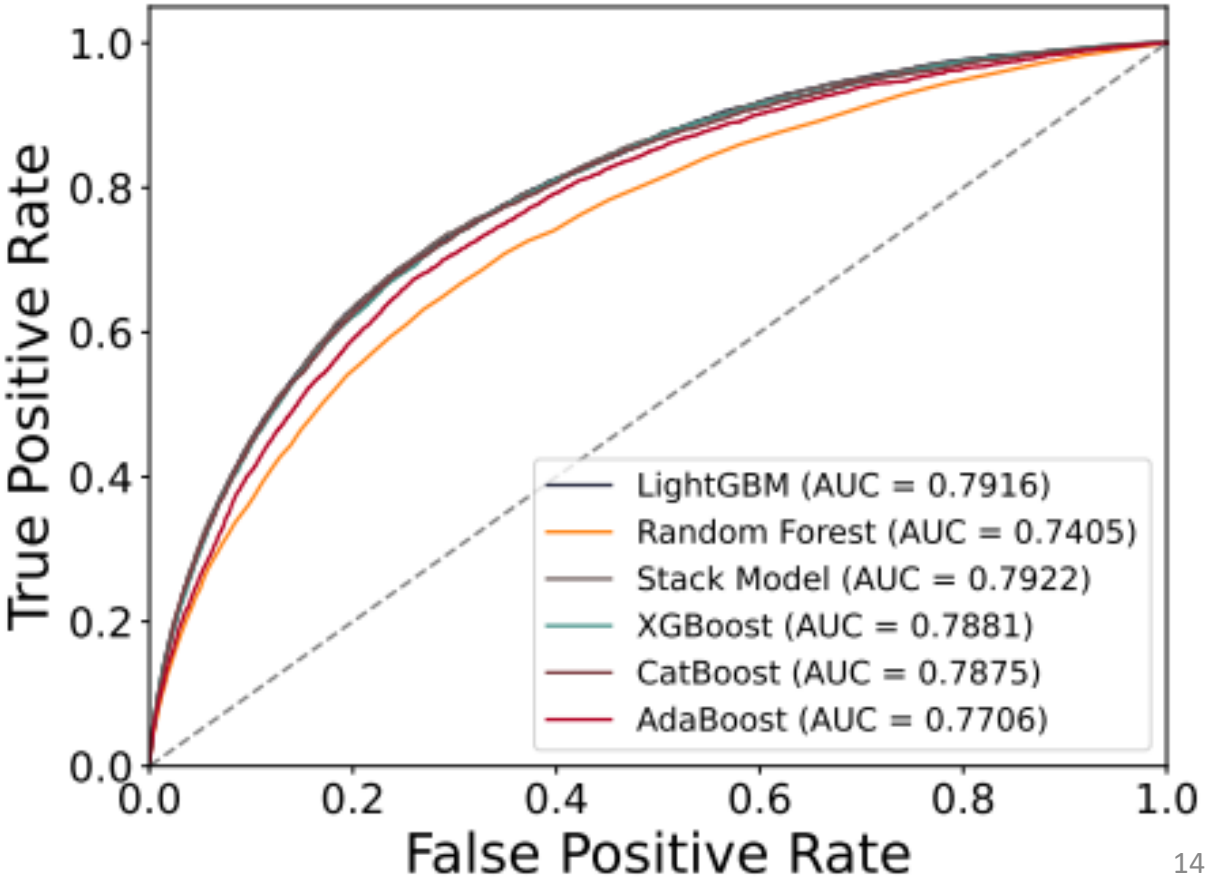
- Boosting models ~0.79 AUC.
- Stacked model achieves best AUC (0.7922).

Table 3: Model performance comparison on validation data. The top-performing model in each metric is highlighted in bold.

Model	AUC	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.5747	0.1250	0.0004	0.0008	0.9191
Random Forest	0.7405	**0.7037**	0.0038	0.0076	0.9195
CatBoost	0.7875	0.6204	0.0342	0.0649	**0.9203**
LightGBM	0.7916	0.5178	0.0614	0.1098	0.9196
AdaBoost	0.7706	0.5152	0.0274	0.0520	0.9194
XGBoost	0.7881	0.5789	0.0421	0.0785	0.9202
LR-Stack (LightGBM + CatBoost + XGBoost)	**0.7922**	0.5053	**0.0961**	**0.1614**	0.9194

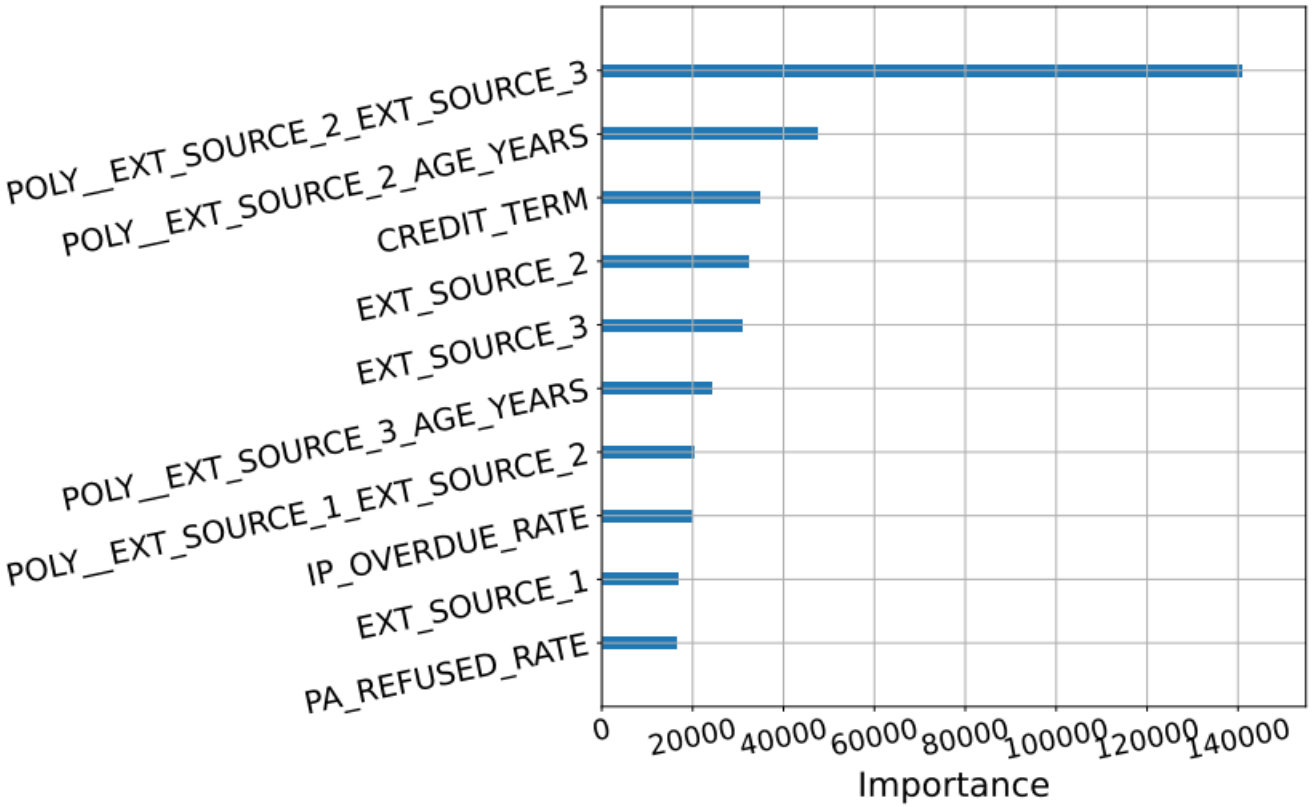
Comparison Experiment

- Boosting models ~0.79 AUC.
- Stacked model achieves best AUC (0.7922).



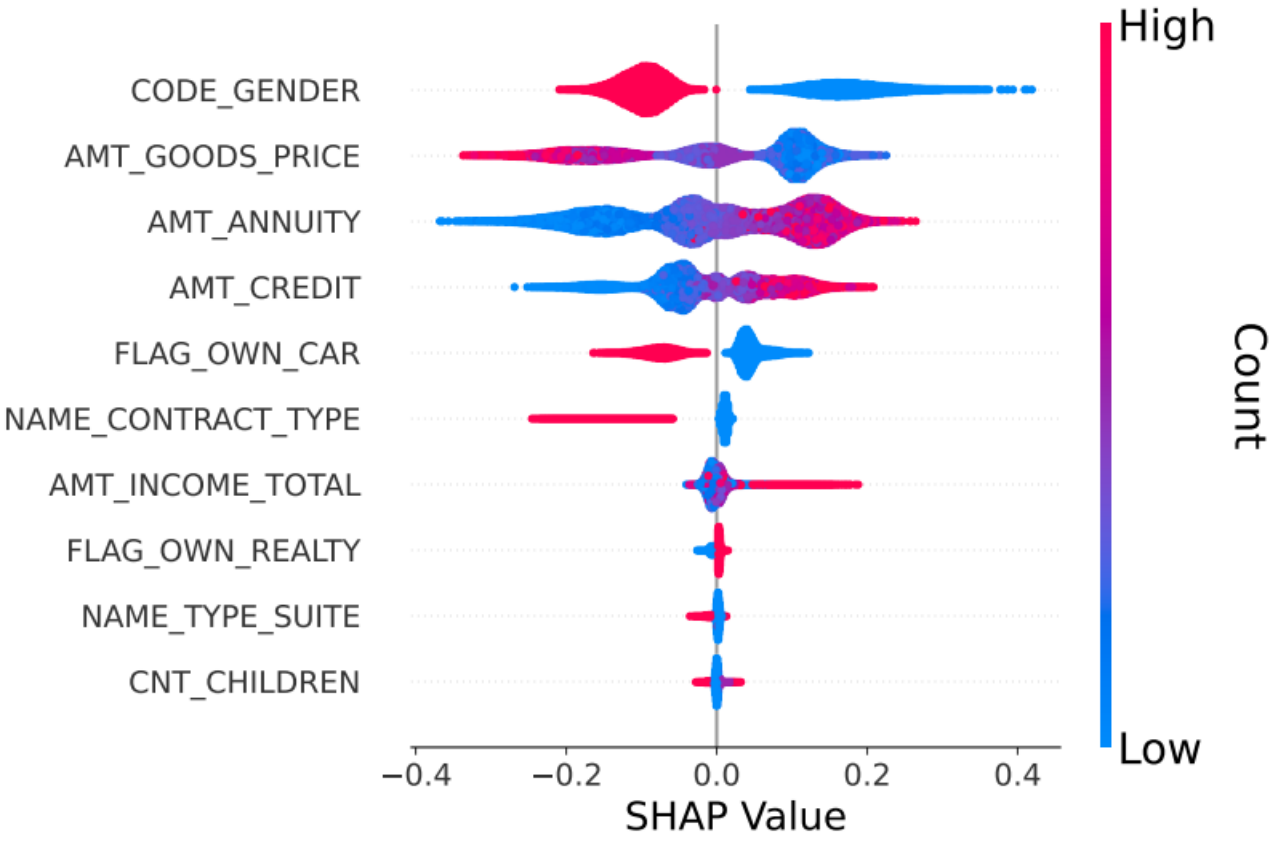
Feature Importance

- Top: EXT_SOURCE family, credit term, annuity.
- Polynomial EXT×Age improves predictions.



Feature Importance

- SHAP: higher credit amount → higher risk; higher income → lower risk; smaller household size -> lower risk.



Discussion

- Feature ratios enhance interpretability.
- Table aggregation provides the largest performance gains.
- Polynomial interactions offer small but consistent improvements.
- Ensembling improves generalization and stability.

Conclusion

- Built a reproducible and interpretable pipeline.
- Performed extensive EDA to identify key feature–target relationships.
- Built a robust preprocessing pipeline: encoding, scaling, imputation, interactions.
- Developed a stacked ensemble (LightGBM + CatBoost + AdaBoost).

Thank You

- Questions?
- GitHub: github.com/siqi-wang25/Home-Credit-Default-Risk-Project
- **Kaggle Leaderboard:**
  - *Private Score:* 0.78857
  - *Public Score:* 0.79108

Model	**AUC**	**Notes**
LightGBM	0.7916	Single best model
CatBoost	0.7875	Handles categorical features well
XGBoost	0.7881	Competitive baseline
Stack (LGBM + CB + XGB)	**0.7922**	Final ensemble model