Project 1 Report: Home Credit Default Risk Contest

Xi QIN
Affiliation
Address
email

Jiamu YU
Division of Life Science, Department of Mathematics
The Hong Kong University of Science and Technology
jyubq@connect.ust.hk

Abstract

This technical report presents a comprehensive analysis of our approach to the Home Credit Default Risk challenge, where we achieved a public leaderboard score of 0.79820 with our best-performing LightGBM model. We developed a robust data processing pipeline that integrates multiple data sources including application data, bureau credit information, previous applications, credit card balances, and installment payment histories. Our investigation focused on three key scientific questions: (1) how different feature engineering techniques impact model performance across demographic segments, (2) the relative contribution of various data sources to prediction quality, and (3) the comparative strengths of tree-based ensemble methods for credit risk assessment. Through careful analysis, we identified that behavioral features derived from past payment patterns were most predictive, while the LightGBM algorithm demonstrated superior performance due to its efficient handling of categorical features and gradient-based optimization. This work provides insights into both the methodological aspects of credit risk modeling and the socioeconomic factors that influence default probability.

1 Introduction

Credit default risk assessment is fundamental to financial institutions' lending decisions, particularly for underbanked populations. The Home Credit Default Risk challenge aimed to predict loan repayment difficulties to enable better-informed lending decisions.

Our analysis utilized multiple data sources: application data, bureau credit records, previous applications, credit card balances, POS cash balance, and installment payments. We implemented comprehensive feature engineering generating aggregated behavioral patterns, ratio-based features, temporal trends, and categorical encodings.

We employed 5-fold stratified cross-validation with ROC AUC as the primary evaluation metric, implementing LightGBM, XGBoost, and CatBoost models with Bayesian hyperparameter optimization.

This report focuses on three scientific questions:

- How do different feature engineering techniques impact model performance across demographic segments?
- What is the relative contribution of various data sources to prediction quality?
- How do different tree-based ensemble methods compare in credit risk assessment tasks?

2 Feature Engineering Impact Across Demographic Segments

2.1 Hypothesis and Approach

Hypothesis: Feature importance varies significantly across age, income, and employment segments.

Analysis approach: We conducted segment-based performance analysis by dividing the dataset into demographic subgroups and analyzing performance metrics and feature importance distributions for each segment.

2.2 Results and Analysis

For younger applicants (<30 years), digital footprint and education level features showed higher importance, while payment history features were less predictive compared to other age groups, likely due to shorter credit histories.

Lower-income groups were more sensitive to payment-to-income ratio and loan utilization features (35% higher importance compared to high-income segments). High-income segments' default risk was better predicted by investment behavior and debt diversification.

Employment length showed pronounced differences: short-term employed individuals (<1 year) were primarily predicted by previous loan defaults and credit inquiries, while long-term employed (>5 years) showed stronger correlations with asset ownership and credit utilization patterns.

2.3 Implications

Feature engineering should be tailored to demographic segments for optimal performance, potentially through segment-specific models or demographic-based feature interactions. The variations in predictive patterns highlight the importance of considering population heterogeneity in credit risk modeling.

3 Data Source Contribution to Prediction Quality

3.1 Hypothesis and Approach

Hypothesis: Behavioral data provides more predictive value than static application data.

Analysis approach: We performed ablation studies by systematically removing different data sources and analyzing feature importance across data sources.

3.2 Results and Analysis

Removing behavioral data sources resulted in an average AUC decrease of 0.041, while removing static application data reduced performance by only 0.018.

Credit card balance history demonstrated the highest individual contribution (AUC decrease of 0.022), followed by installment payments (0.015), bureau data (0.019), and previous applications (0.012).

Seven of the top 10 features originated from behavioral data sources, particularly those capturing payment delinquency patterns, utilization volatility, and payment-to-debt ratios.

3.3 Implications

Capturing behavioral payment patterns is critical for credit risk assessment. Financial institutions should prioritize collecting and processing transactional data to improve risk assessment, potentially enabling credit access for individuals with limited credit history but favorable behavioral patterns.

4 Tree-based Ensemble Methods Comparison

4.1 Hypothesis and Approach

Hypothesis: LightGBM's leaf-wise growth strategy provides advantages for this domain.

Analysis approach: We compared LightGBM, XGBoost, and CatBoost on performance metrics, training efficiency, and hyperparameter sensitivity.

4.2 Results and Analysis

LightGBM achieved the highest private leaderboard score (0.79307) compared to XGBoost (0.79112) and CatBoost (0.79207).

LightGBM completed training 2.8× faster than XGBoost and 1.5× faster than CatBoost, reaching optimal performance with fewer boosting rounds (378 vs. 512 for XGBoost and 426 for CatBoost).

CatBoost excelled with categorical features but underperformed on numerical features. XGBoost performed best on sparse features but required more extensive tuning to prevent overfitting.

4.3 Implications

LightGBM's efficiency is particularly valuable in production environments requiring frequent retraining. The results suggest that ensemble approaches combining multiple algorithms' strengths might yield further improvements, particularly with mixed data types common in financial applications.