Rakamin Academy

HOME CREDIT

Home Credit Default Risk Analysis & Prediction

Final Task - Rakamin Academy Virtual Internship
in collaboration with Home Credit

By: **Az-Zukhrufu Fi Silmi Suwondo**

Success Metric

⩽ 7%

Default Rate

The goal was to reduce defaults to 7% or lower

⩾ 70%

Approval Rate

Minimum threshold to maintain healthy loan disbursement

Problem & Objectives

Problem Statement

Many loan applicants **lack formal credit histories**, making **traditional credit scoring methods ineffective in assessing their repayment capacity**. This creates the risk of **unfairly rejecting creditworthy** individuals while also **increasing the likelihood of granting loans** to those who may default.

Goal

To develop a **credit risk prediction model** using alternative data that achieves an **AUC above 0.75** within six months, **reducing false rejections of creditworthy clients** and **supporting responsible lending**.

Objectives

- To **collect** and **preprocess alternative data** and **develop machine learning models** that achieve at least **0.75 AUC**, validated on out-of-sample data within six months.
- To provide actionable insights that **reduce false rejections** of creditworthy clients, while also **reducing the default rate to ≤ 7%** and **maintaining an approval rate of ≥ 70%**.

Dataset Overview

Rakamin Academy

The dataset consists of 7 main files that are interconnected to provide a comprehensive view of customer credit profiles:

Core Application Data

- application_train/test.csv: Main loan application data (static features)
- HomeCredit_columns_description.csv: Column descriptions for reference

Historical Credit Behavior

- bureau.csv: Credit history from other financial institutions via the Credit Bureau
- bureau_balance.csv: Monthly balances of previous credits from the Credit Bureau
- previous_application.csv: Historical loan applications submitted to Home Credit

Transaction & Payment History

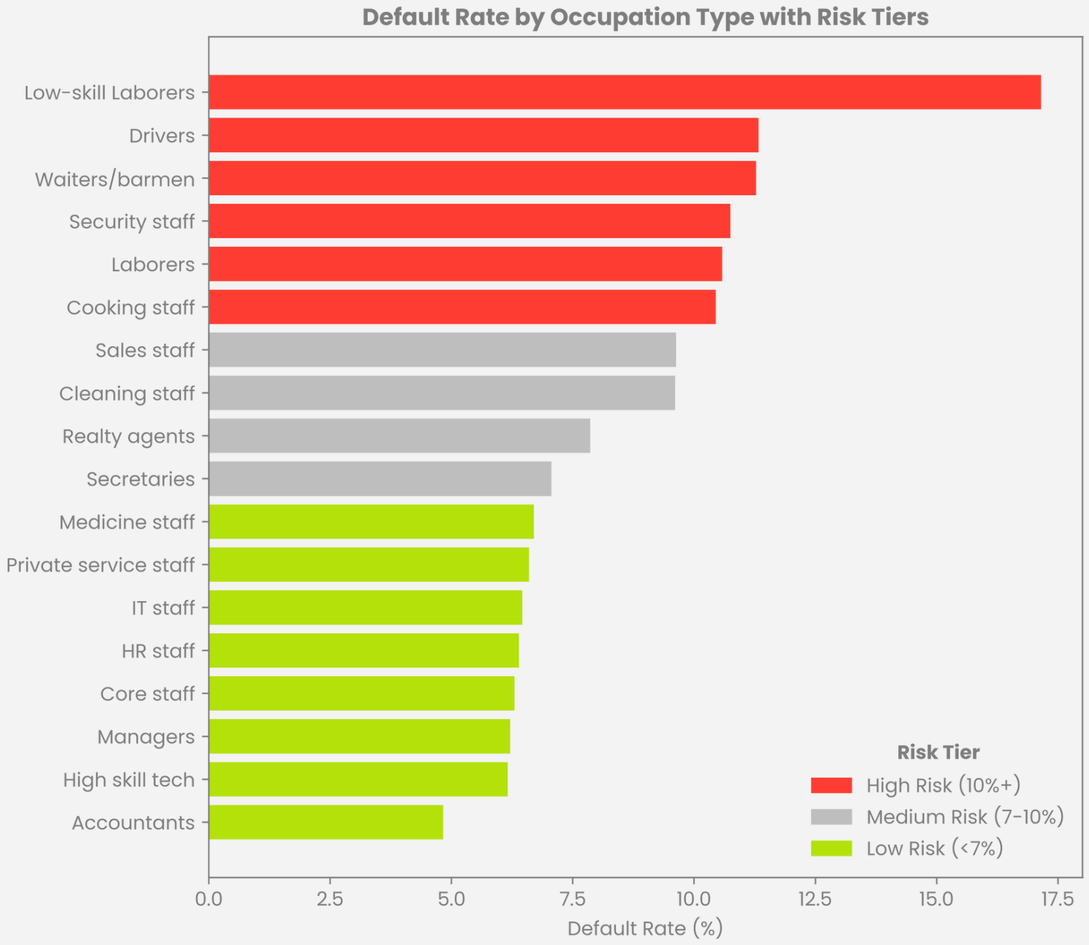
- POS_CASH_balance.csv: Monthly balances of POS and cash loans at Home Credit
- credit_card_balance.csv: Monthly credit card balances at Home Credit
- installments_payments.csv: Installment payment history (including delays)

Dataset Characteristics

- Relational Structure
- Time Series Component
- Comprehensive Coverage
- Rich Feature Set

Train: 307,511 row × 122 cols
Test: 48,744 row × 121 cols

Job & Credit Risk



Jobs requiring **high skills** and **stable income**, like accountants and IT specialists, show the **lowest default rates**, while **low-skill**, **fluctuating-income roles**, such as drivers and waiters, are **three times more likely to default** due to poor job security and income unpredictability.

Low Risk

- Accountants
- Core Staff
- Managers
- High Skill Tech

High Risk

- Low-skill Laborers
- Waiters/Barmen
- Drivers
- Security/Laborers

Job & Credit Risk

Rakamin Academy

$$\chi^2 = 1403.22$$

$$p < 0.001$$

Chi-square test confirms a **significant link** between profession and repayment behavior, white-collar roles are more stable, while blue-collar roles show higher volatility.

Actionable Item

B2B Payroll-Integrated Financing

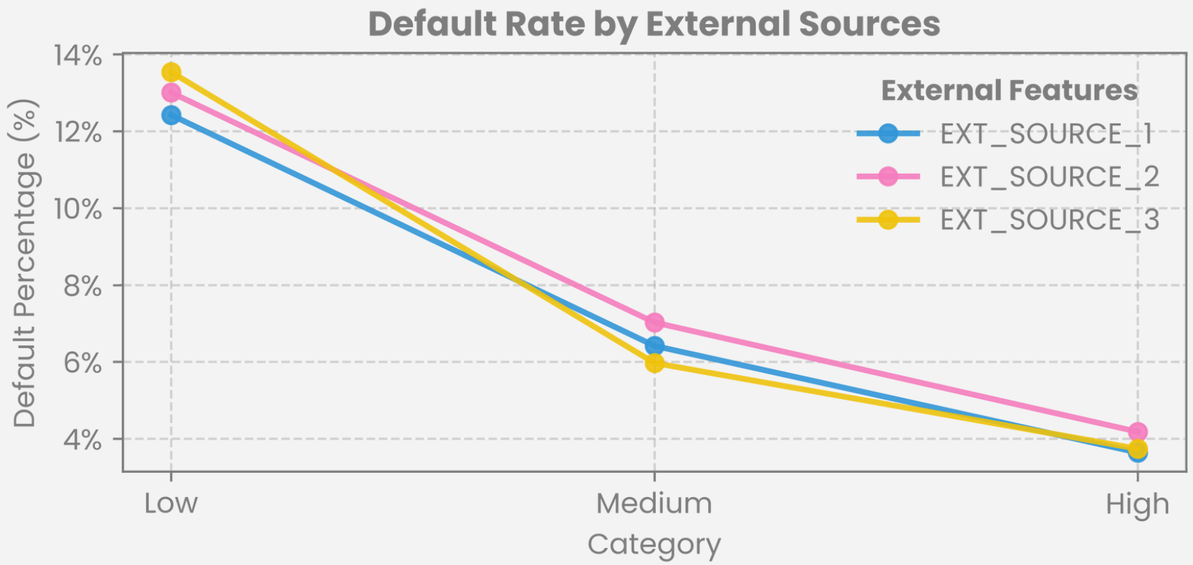
Employee loans integrated with payroll systems. Repayments auto-deducted, minimizing default risk. Ideal for stable-income sectors.

Actionable Item

Profession-Specific Micro-Insurance

Insurance products priced by occupational risk. Lower premiums for low-risk roles. Covers income, health, and accident with tailored protection.

Risk Gap by External Scores



The **strongest external indicator** in the credit model shows a **fivefold risk gap** between low and high scores. **External financial data** is **three times more predictive** than demographics. **Missing key external inputs signal elevated risk** and require targeted data improvements.

Actionable Item

Strategic Partnership Program with Financial Institutions

Expand collaboration with banks, fintechs, and credit bureaus to enrich external data, improve coverage, reduce gaps, and enable alternative scoring for thin-file applicants.

Data Preprocessing

Rakamin Academy

Missing Values Handling

- Redundant columns removal
- High-missing columns removal
- Imputation with median and 'Unknown'

Outlier & Error Handling

- Error handling
- Log transformation
- IQR capping

Feature Engineering

- Feature construction
- Feature scaling with RobustScaler
- Feature selection with SelectFromModel

Feature Construction

Application Features

- CREDIT_ANNUITY_RATIO

Feature Construction

Payment Behavior Features

- INST_PAYMENT_RATIO_MEAN

Feature Construction

Credit Card Features

- CC_BALANCE_MEAN

Feature Construction

Previous App Features

- PREV_AMT_APP_MEAN
- PREV_AMT_CREDIT_MEAN
- PREV_APPROVED_COUNT
- PREV_REFUSED_COUNT

Feature Construction

Bureau Features

- BUREAU_DAYS_CREDIT_MEAN
- BUREAU_CREDIT_SUM
- BUREAU_CREDIT_SUM_DEBT
- DEBT_CREDIT_RATIO

Modeling & Evaluation

Rakamin Academy

Handling Class Imbalance

scale_pos_weight was calculated based on the class ratio to balance prediction outcomes.

Categorical Feature Encoding

CatBoost automatically handles categorical features using ordered target encoding.

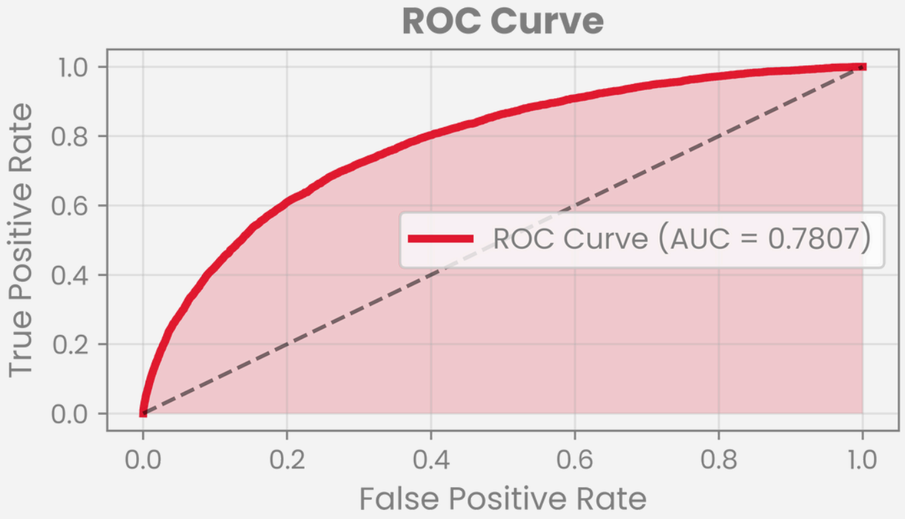
Hyperparameter Tuning

Parameters were optimized using RandomizedSearchCV with 3-fold stratified cross-validation, evaluated by ROC AUC. Best validation AUC: 0.7807.

Threshold Tuning

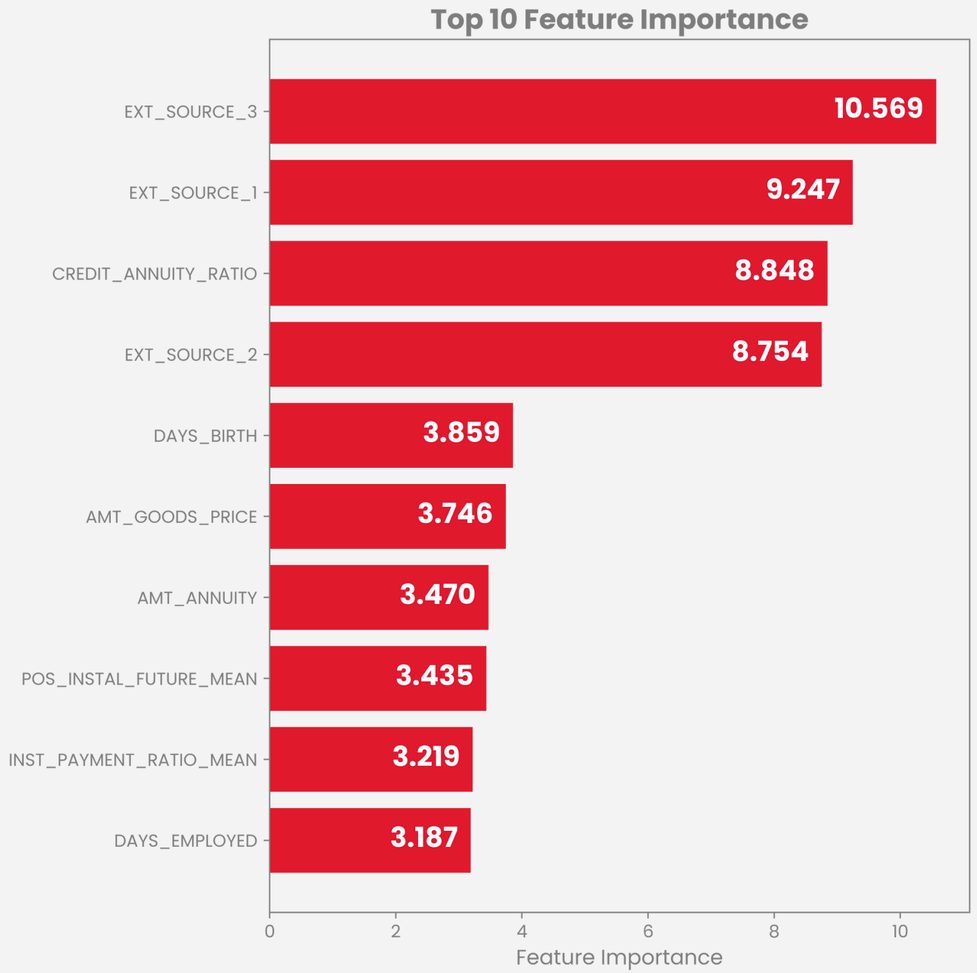
The optimal threshold (0.69) was selected to maximize F1-score (0.3313), considering the precision–recall trade-off.

The CatBoost model achieved an AUC of 0.7807, showing strong overall discrimination, but while it performs well on class 0, it struggles with class 1 due to lower precision and recall.



		Predicted	
		Negative	Positive
Actual	Negative	51,299	5,239
	Positive	2,939	2,026

Feature Importance



The model uses a total of **108 features** covering **external scores**, **demographics**, **financial metrics**, **employment history**, **credit applications**, and **repayment behavior**. The most important features are highlighted below.

Top Features Ranking

- EXT_SOURCE_3 (Highest importance)
- EXT_SOURCE_1
- CREDIT_ANNUITY_RATIO
- EXT_SOURCE_2

Business Insights

- Model heavily relies on external data sources
- Internal financial metrics are also critical
- Age and employment history play significant roles

Recommendation

Rakamin Academy

Data Quality Enhancement

- Prioritize EXT_SOURCE completeness
- Monitor top 10 features
- Implement data validation

Risk-Based Product Development

- B2B Payroll-Integrated Financing
- Profession-Specific Micro-Insurance

Partnership Expansion

- Strategic Partnership Program
- Alternative Data Sources

Model Maintenance

- Continuous Monitoring
- Feature Stability Tracking

Business Impact

- Default rate dropped to **5.42%**, beating the 7% target (from 8.07%)
- Approval rate stayed high at **88.19%**, well above the 70% minimum

Source Code *LinkedIn* *afsilmis@gmail.com*