HOME CREDIT X Rakamin

Project Based Internship

HOME CREDIT DEFAULT RISK PREDICTION: Machine Learning Classification Model

Using Application and Credit History Data

by: Yodha Pranata

HOME CREDIT X Rakamin

BACKGROUND

Many potential borrowers *especially underbanked groups* ***have limited credit history****, making traditional credit risk assessment difficult and* ***increasing the chance of either rejecting creditworthy applicants*** *or approving high-risk loans. To support financial inclusion while controlling losses, lenders increasingly combine application data with broader "alternative" or* ***behavioral credit signals to estimate default risk more consistently****.*
The Home Credit Default Risk dataset provides historical application and credit-history tables designed ***to predict whether an applicant will experience payment difficulties*** *(TARGET=1)* ***versus not*** *(TARGET=0), enabling risk estimation at application time.*

Problem Statement

How accurately *can payment-difficulty risk be* ***predicted from application and credit history data****, and which* ***features contribute*** *most to the increase in* ***predicted default risk****?*

Objective

To ***develop a machine learning*** *model that* ***predicts the probability of payment difficulties*** *using applicant characteristics and credit history features, and to identify the most influential risk drivers affecting default risk.*

DATA SOURCE & PIPELINE OVERVIEW

HOME CREDIT X Rakamin

SQL (POSTGRESQL)

build a single dataset using **CTE aggregations** across application attributes, bureau summary, previous application history, installments behavior, POS & credit card behavior

Python

prepare EDA-ready and ML-ready datasets (**356,255 rows, 45 features**).

EDA

visualize distributions and group differences (bar/violin/box), run **statistical tests**.

Modeling

train models, run **hyperparameter tuning**, **evaluate** with AUPRC, and **interpret** results using feature importance

DATA PREPROCESSING for EDA

1 Feature Aggregation

- **EXT_SOURCE_1/2/3** → ext_median
- **DAYS_BIRTH** → age_years
- **DAYS_EMPLOYED** → employed_years.

2 Handling Duplicate

No Duplicate Data Found

3 Handling Missing Value

Numeric missing values identified → filled using median imputation.

DATA PREPROCESSING
for Machine Learning

HOME CREDIT X Rakamin

1 Feature Aggregation

- **EXT_SOURCE_1/2/3** → ext_median
- **DAYS_BIRTH** → age_years
- **DAYS_EMPLOYED** → employed_years.

SPLIT TRAIN & TEST

2 Handling Duplicate

No Duplicate Data Found

3 Handling Missing Value

Numeric missing values identified → filled using median imputation.

4 Outlier Screening

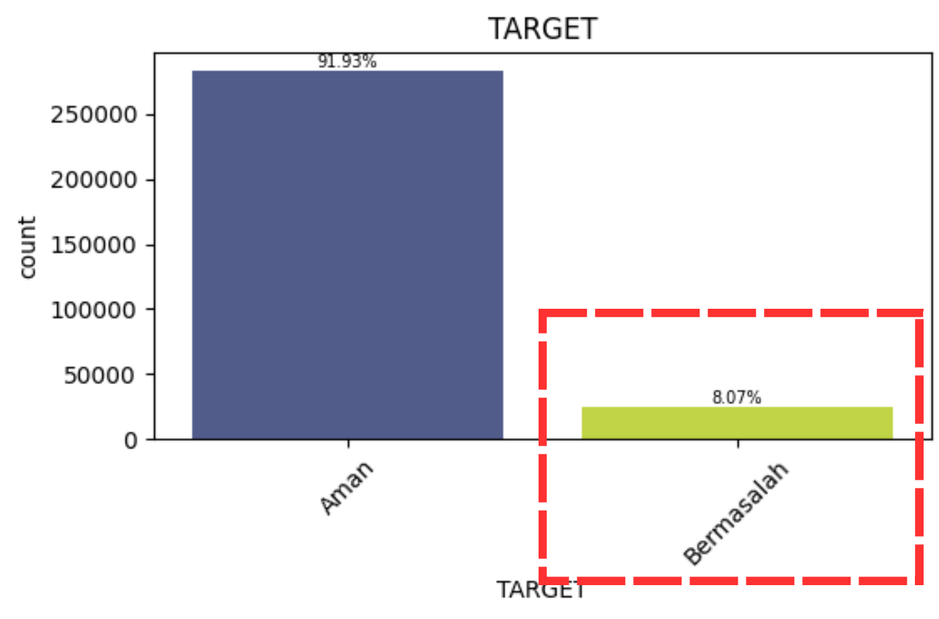
Applied **np.log1p** to reduce **skew/outlier** impact on:
AMT_INCOME_TOTAL, AMT_ANNUITY, AMT_GOODS_PRICE, bureau_cnt_total, bureau_cnt_active, prev_cnt, prev_cnt_approved, prev_cnt_refused, prev_avg_amt_credit, prev_avg_amt_annuity, inst_cnt, inst_late_cnt, inst_late_days_avg, inst_late_days_max, inst_paid_ratio, pos_rows, pos_dpd_avg, pos_dpd_max.

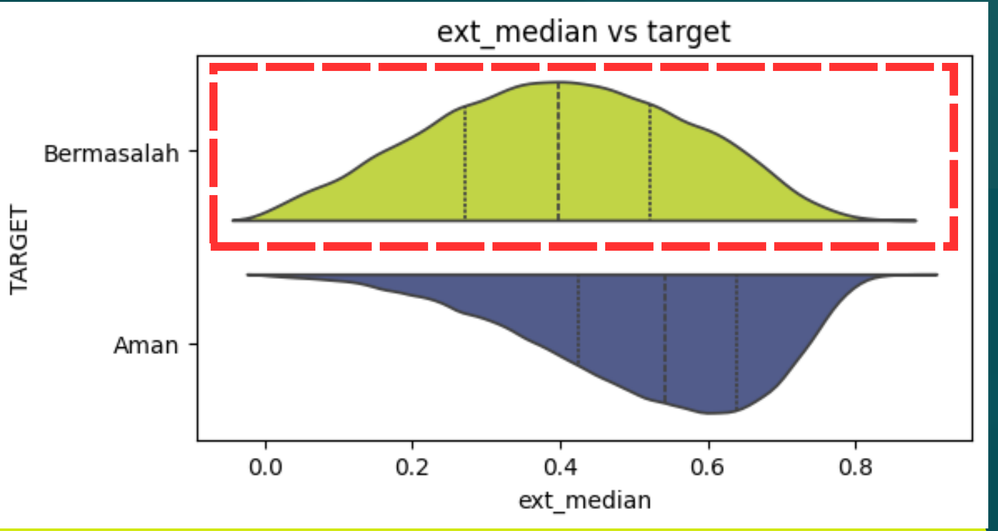
5 Feature Scalling

One-hot encode categorical features; **standard scaller for Logistic Regression**, while **XGBoost** uses OHE only (**no scaling**) for consistent feature-importance interpretation.

Correlation heatmap showed highly correlated features, so redundant pairs were reduced by dropping *AMT_GOODS_PRICE, REGION_RATING_CLIENT_W_CITY, CNT_CHILDREN, prev_cnt, prev_avg_amt_annuity, inst_late_days_max, and pos_dpd_max* to **minimize redundancy** and **improve model stability** and interpretability.

EXPLORATORY DATA ANALYSIS





***Target distribution** is **highly imbalanced** (Safe 91.93% vs Payment Difficulty 8.07%), so **AUPRC is a more appropriate** metric than accuracy or ROC-AUC because it focuses on performance for the minority (positive) class.*

- ***ext_median** is the median of three external risk scores (EXT_SOURCE_1/2/3), providing a **robust combined** signal that is less sensitive to outliers or missing values from a single source.*
- ***The violin plot** suggests **lower ext_median values** are **associated with higher payment-difficulty risk**, and the **Mann–Whitney U** test ($p < 0.001$) confirms a **statistically significant** difference between the two classes.*

MODELING & HIPERPARAMETER TUNING

*AUPRC **summarizes precision–recall across many thresholds**, which is useful for selecting an operating threshold that matches business needs; it also provides a more balanced view than reporting recall alone, since recall can be increased by flagging too many cases while **AUPRC** penalizes low precision.*

Model	Stage	AUPRC (Train)	AUPRC (Test)	Notes
Logistic Regression	Before tuning	0,2264	0,2319	Baseline linear model
Logistic Regression	After tuning	0,2269	0,2323	Small, stable improvement
XGBoost	Before tuning	0,4428	0,2308	Clear overfitting (train ≫ test)
XGBoost	After tuning	0,2338	0,2364	Better generalization, best test AUPRC

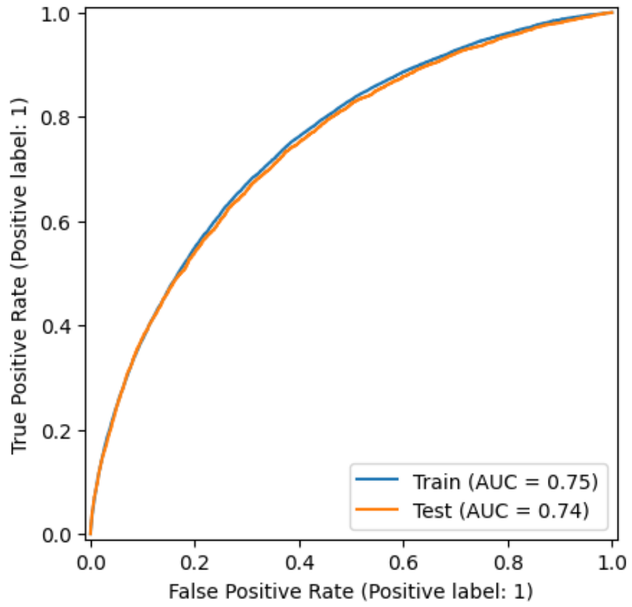
*Based on the highest test AUPRC (0.2364) and improved generalization after tuning, **XGBoost** was selected as the **best-performing model for this project**.*

MODEL EVALUATION

```
AUPRC Score (Test): 0.236420523474331
Classification Report Xgboost (Test):
              precision    recall  f1-score   support

           0       0.94      0.88      0.91     56481
           1       0.24      0.41      0.30      5021

    accuracy                           0.84     61502
   macro avg       0.59      0.65      0.61     61502
weighted avg       0.89      0.84      0.86     61502
```

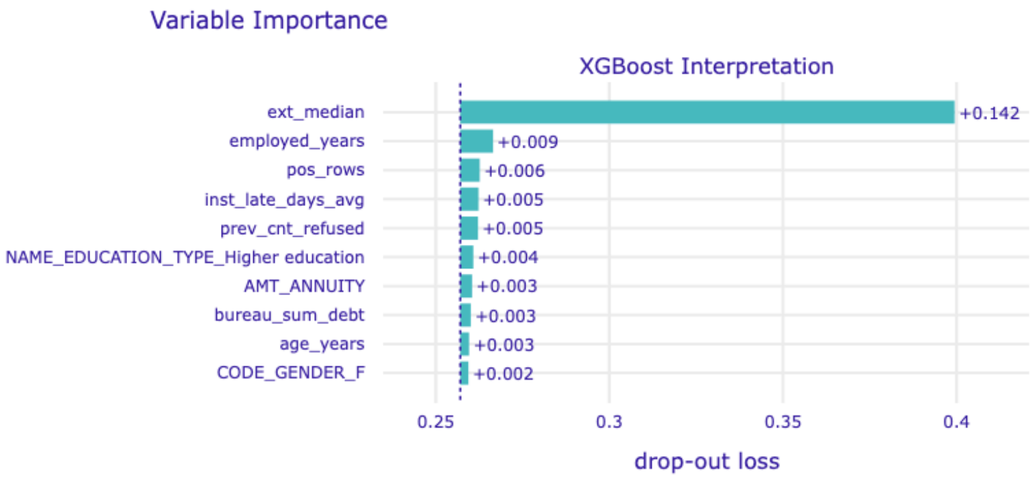


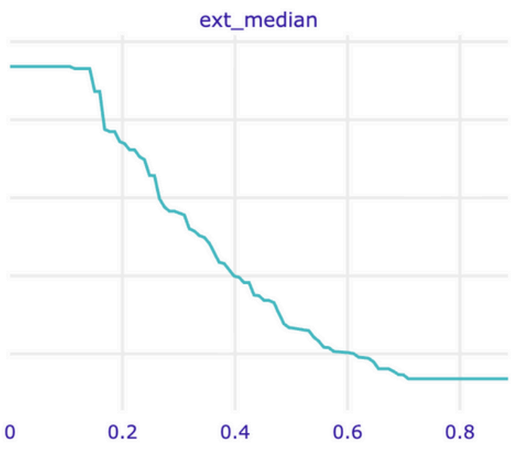
Confusion Matrix

	0	1
0	49,926 (TN)	6,555 (FP)
1	2,980 (FN)	2,041 (TP)

- AUPRC = 0.2364 exceeds the baseline (~0.08). For TARGET=1, recall = 0.41 and precision = 0.24, indicating that **the model is suitable for early screening** and **manual review**, rather than auto-rejection.
- ROC-AUC is stable with 0.75 (train) and 0.74 (test), indicating **no strong overfitting**.
- The confusion matrix shows that most s**afe borrowers are predicted correctly**. Still, **some are predicted as risky**, increasing reviews and reducing approvals. In contrast, **some risky borrowers** are missed, increasing default loss, so a risk-tier policy with manual review for high-risk cases is recommended.

FEATURE IMPORTANCE





- ***The most important driver is ext_median**, meaning external risk scores provide the strongest signal for predicting payment difficulty.*
- *The partial dependence plot shows a clear pattern: when ext_median rises from ~0.10–0.20 to ~0.60–0.70, the predicted risk drops sharply and then becomes nearly flat above ~0.70, meaning **low ext_median (<~0.30) is associated with much higher default risk.***
- *Other meaningful contributors include **employment length** (employed_years), **credit history activity** (pos_rows, bureau_sum_debt), and **repayment behavior** (inst_late_days_avg, prev_cnt_refused), which together reflect stability and past credit performance.*

BUSINESS RECOMMENDATION

1. Use the **model for underwriting triage** by assigning applicants into **low, medium, and high risk tiers** to support faster approvals and targeted manual review.
2. Apply risk mitigation on key drivers, especially **low ext_median,** by tightening verification and adjusting credit terms such as **lower limits or shorter tenors for high-risk tiers.**
3. **Avoid auto-rejection** and **control false positives** by using predicted risk to prioritize reviews rather than declining decisions directly.
4. **Set thresholds based on business costs** by **balancing missed risky borrowers versus additional review workload**, then calibrating the cutoff to match the risk policy.

PROJECT RESOURCES

LinkedIn Profile:
linkedin.com/in/yodha-pranata

GitHub
https://github.com/yodhata/Home-Credit-Default-Risk-Prediction

Link Youtube Project:
https://youtu.be/2-VzSYtL1Hw