Rakamin Academy × HOME CREDIT

Oleh : Devin Syafa

Home Credit Payment Difficulty Prediction

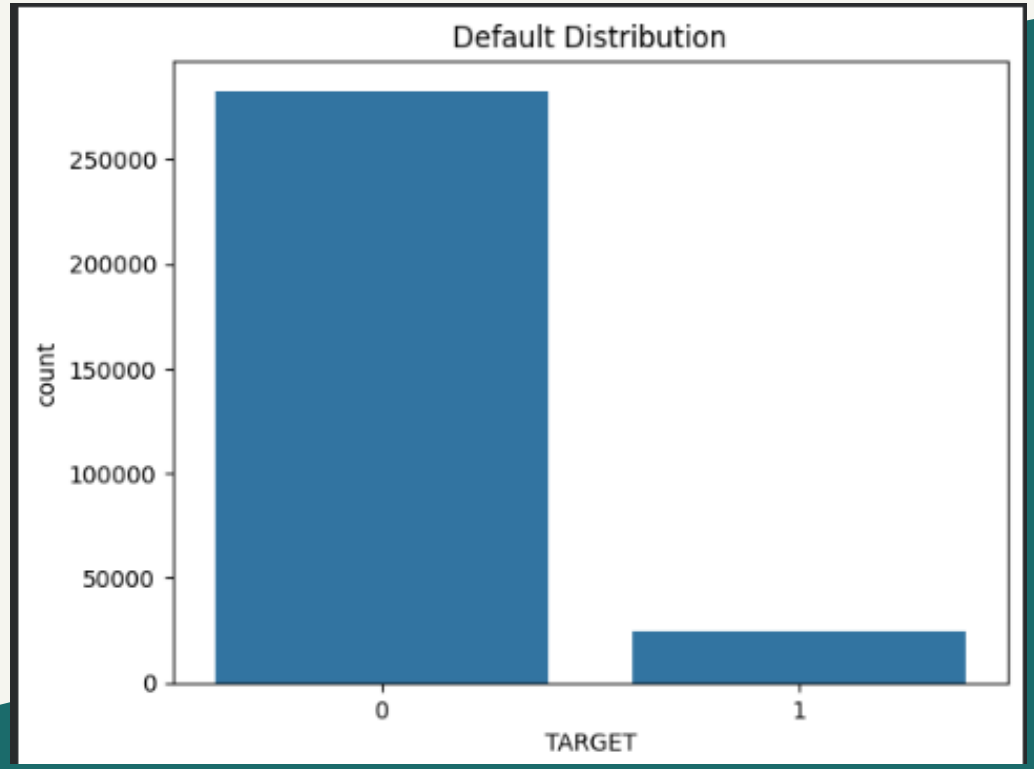
Problem Research

•Masalah

•Kasus gagal bayar (TARGET = 1) hanya ±8% → data tidak seimbang (imbalanced).
•Risiko gagal bayar menyebabkan kerugian finansial.
•Perlu sistem screening awal berbasis data

•Tujuan

- Memprediksi probabilitas gagal bayar.
- Membandingkan:
  - Model baseline (data aplikasi saja)
  - Model dengan external credit history
- Mengidentifikasi faktor risiko utama.



Dataset Overview

Sumber Data

- application_train (dengan TARGET)
- application_test (tanpa TARGET)
- bureau
- previous_application
- installments_payments
- pos_cash_balance
- credit_card_balance

Jumlah data:

- 307.511 baris

Jumlah fitur awal:

- 233 kolom

Target variable:

- TARGET 0 → Tidak gagal bayar 1 → Gagal bayar

Tahap Pengerjaan

1

Data Understanding
- Memahami konteks bisnis dan definisi TARGET (imbalanced ~8%).
- Analisis distribusi target, tipe data, dan missing values.

2

Feature Engineering
- Agregasi data riwayat kredit (bureau, previous, installments, POS, credit card).
- Menghasilkan ±45 fitur customer-level (1 baris per SK_ID_CURR).

3

Preprocessing
- Imputasi median & hapus fitur dengan missing tinggi.
- Feature transformation (ext_median, age, log transform).
- One-Hot Encoding & scaling (khusus Logistic Regression).

4

Modeling
- Baseline: Logistic Regression & XGBoost (data aplikasi saja).
- Dengan External Data: Logistic Regression & XGBoost + tuning (Optuna).

5

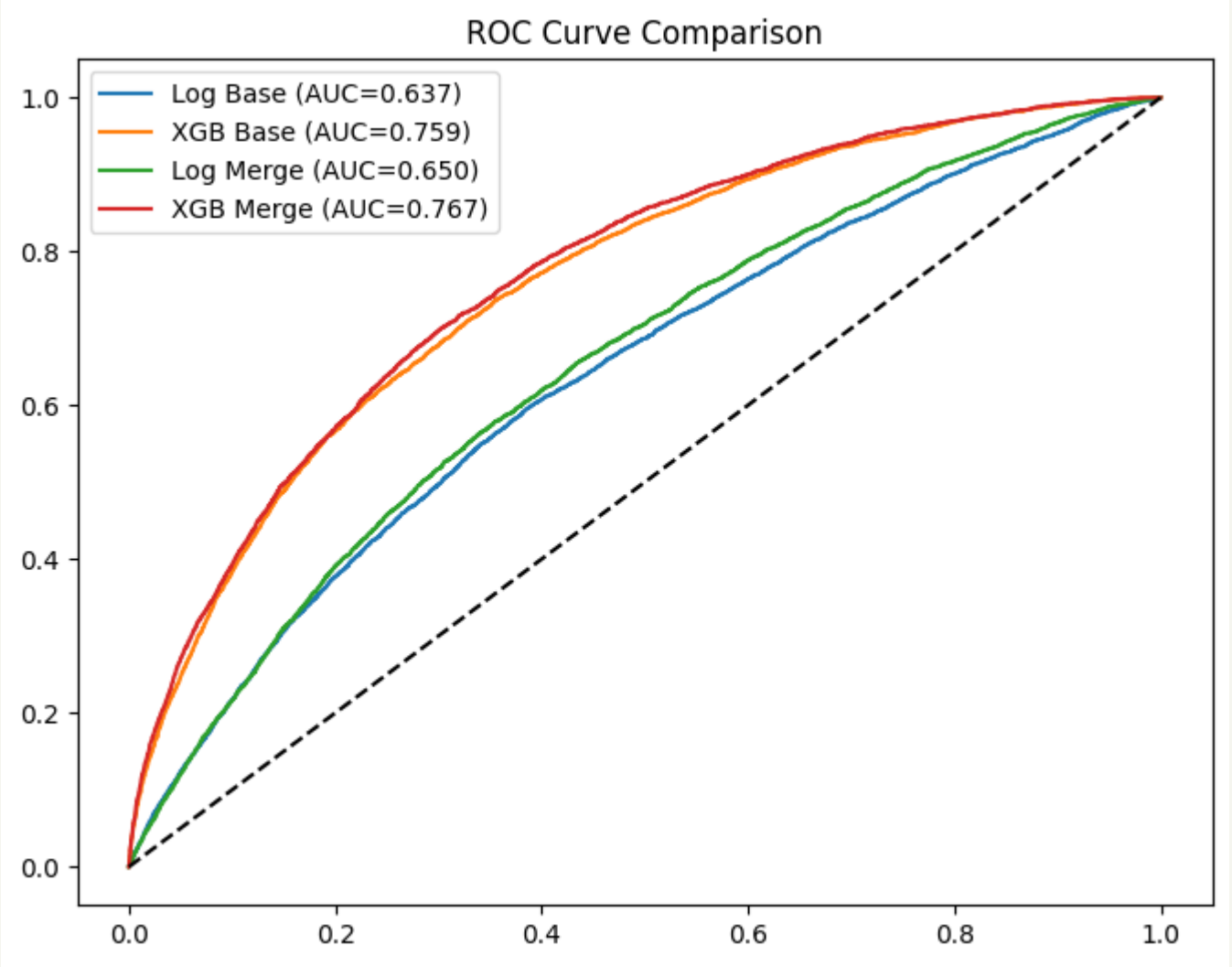
Evaluation
- Metric utama: **AUPRC** (imbalanced data).
- ROC-AUC untuk generalisasi.
- Model terbaik: **XGBoost + External Data**

Data Visualization

- **Temuan Utama dari EDA**
- **Class Imbalance**
  TARGET=1 hanya ±8% → perlu evaluasi dengan **AUPRC**.
- **Distribusi Usia**
  Nasabah usia lebih muda cenderung memiliki proporsi default lebih tinggi.
- **Distribusi Income**
  Kelompok TARGET=1 memiliki median income lebih rendah dibanding TARGET=0.
- **EXT_SOURCE (ext_median)**
  Skor eksternal rendah sangat berkorelasi dengan risiko gagal bayar.
- **Behavioral Features (DPD & Keterlambatan)**
  Riwayat keterlambatan pembayaran menjadi indikator risiko paling kuat.

Machine Learning Implementation & Evaluation

Model	Data yang Digunakan	Hasil
Logistic Regression	Application Only	0.637
Logistic Regression	Application + External Data	0.650
XGBoost	Application Only	0.759
XGBoost	Application + External Data	0.767



Business Insight

Hasil evaluasi menunjukkan bahwa penambahan external credit history secara konsisten meningkatkan performa model, baik pada Logistic Regression maupun XGBoost. Model XGBoost dengan data gabungan (application + external) memberikan performa terbaik (AUC 0.767), menunjukkan kemampuan yang lebih baik dalam membedakan nasabah berisiko dan tidak berisiko. Hal ini mengindikasikan bahwa integrasi data perilaku kredit historis sangat penting untuk meningkatkan kualitas risk screening dan mendukung pengambilan keputusan kredit yang lebih akurat.

Business Recomendation

1

Implementasikan **XGBoost + External Data** sebagai decision support dalam approval kredit.

2

Gunakan threshold berbasis risk appetite untuk segmentasi risiko.

3

Lakukan monitoring dan retraining model secara berkala untuk menjaga performa.

Thank You

Presentasi Video: https://youtu.be/ea6SiW_jCMY

Link Github : https://github.com/Devins16/Home-Credit-Final-Task.git