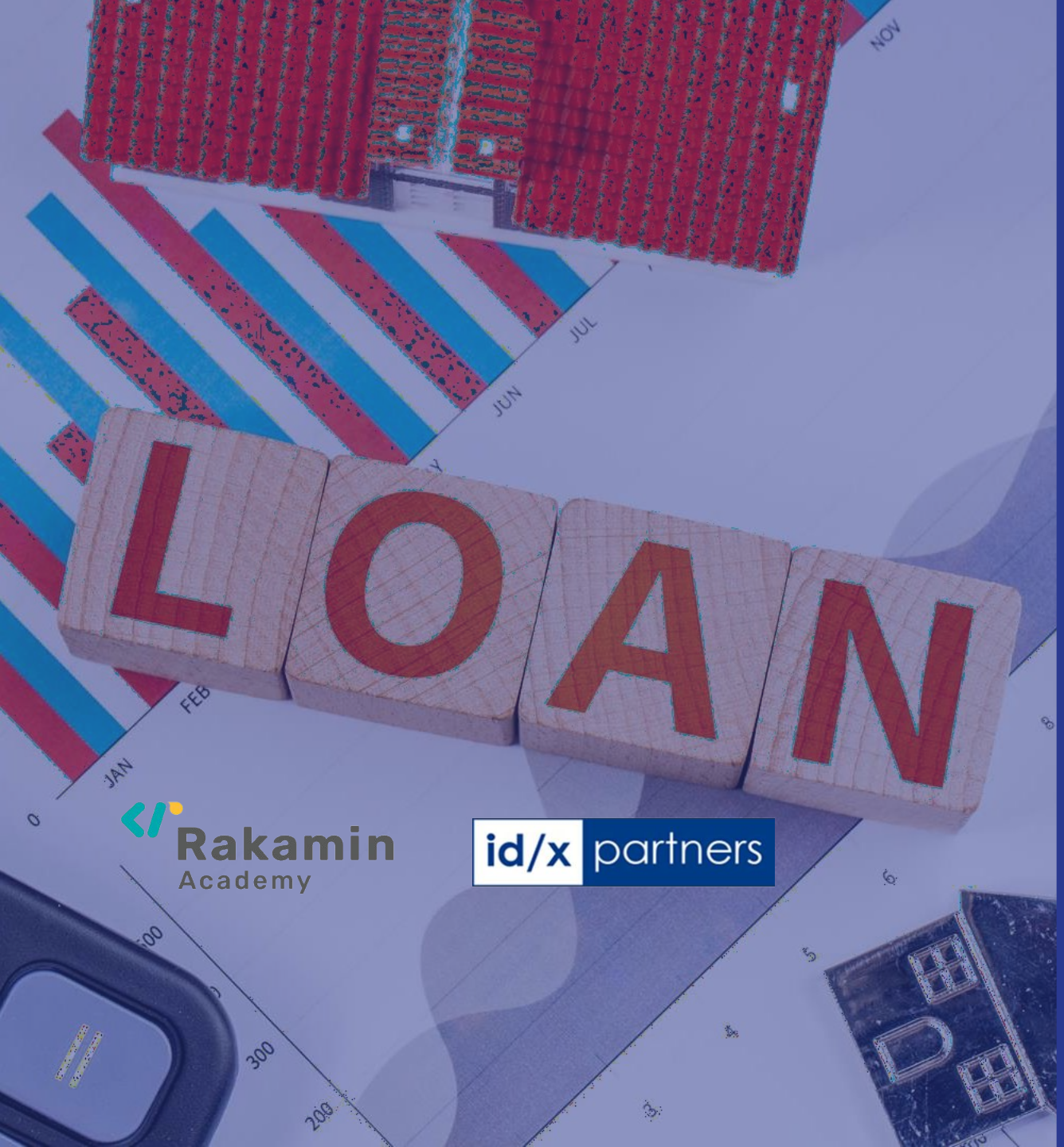


Model Klasifikasi

untuk Prediksi Risiko Gagal Membayar Pinjaman

Virtual Internship Program - Data Scientist

Oleh : Muhammad Ali Umar

Problem Statement

- *Credit Risk* adalah risiko kerugian yang dihadapi pemberi pinjaman karena kegagalan peminjam untuk membayar kembali semua jenis pinjaman atau hutang.

- Diperlukan waktu yang lama jika kita melakukan penilaian secara manual. Hal ini dapat terjadi jika kita membuat keputusan yang salah dan akan menyebabkan kerugian yang cukup besar.

- Kerugian Finansial Kredit adalah jumlah uang yang hilang oleh pemberi pinjaman ketika pemohon menolak untuk membayar atau melarikan diri dengan uang yang terutang.

GOAL & OBJECTIVE

Membuat suatu model yang dapat menentukan suatu pengguna mampu atau tidak mampu membayar kredit.

ANALYTICS APPROACH

Melihat permasalahan yang ada, aku akan membangun model *machine learning,* karena perlu membangun prediksi lebih dari sekedar menggunakan pendekatan analisis inferensial atau deskriptif.

MODELLING

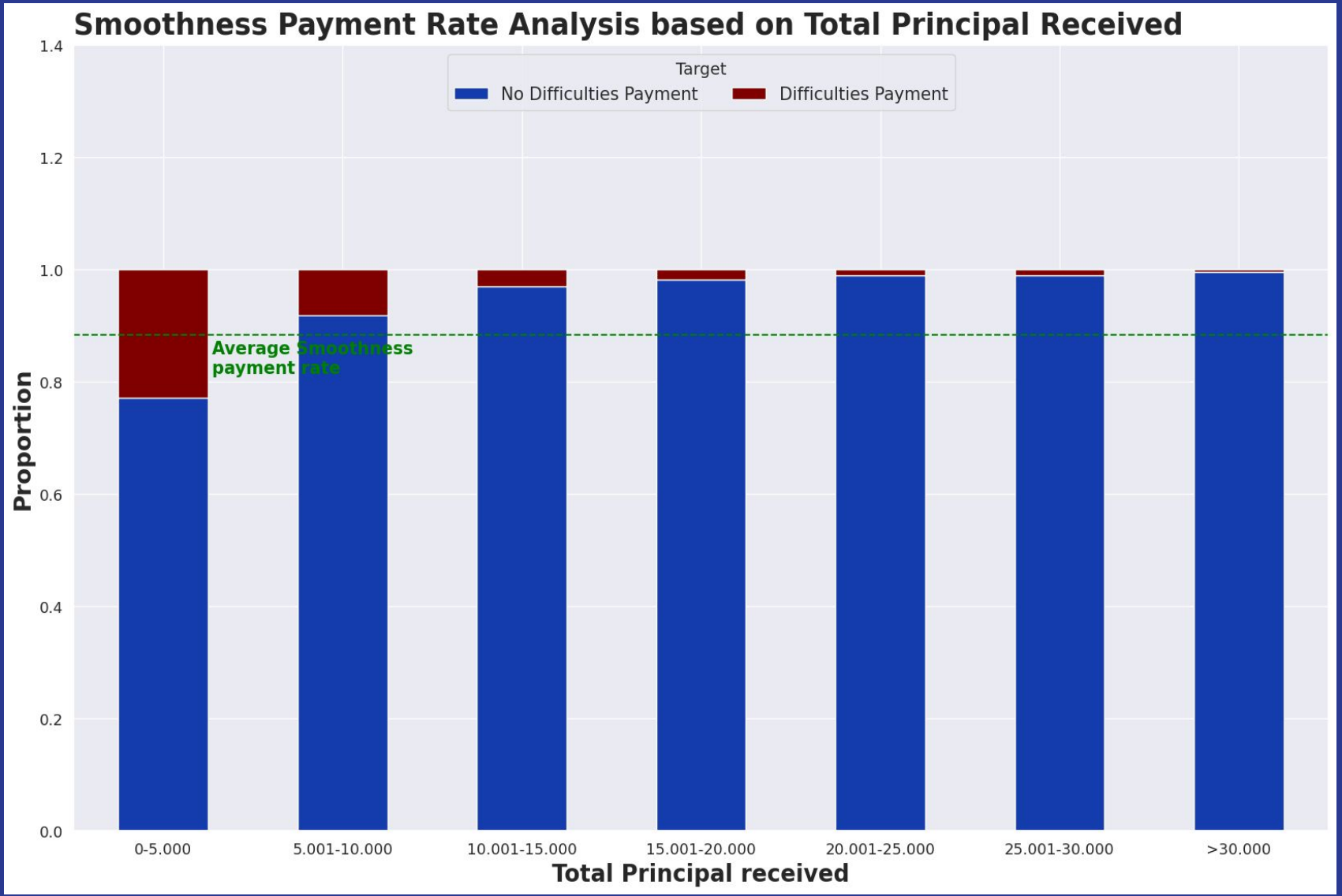
Aku akan mencoba 6 model *machine learning*: Logistic Regression, XGB, Decision Tree, Random Forest, Naive Bayes, LGBM.

Dataset

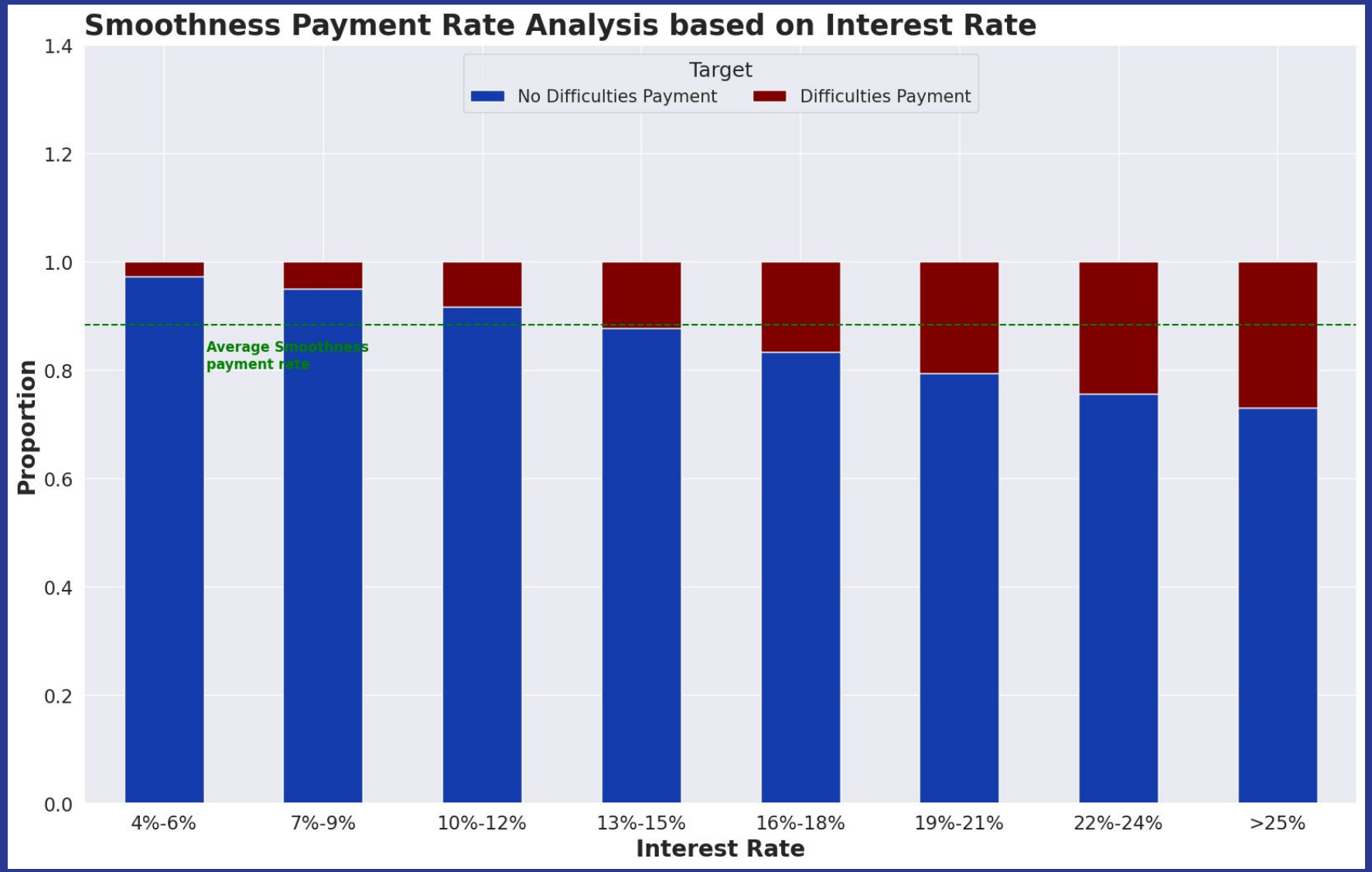
Dataset sample yang digunakan berisi data pinjaman yang diterima maupun yang ditolak dan terdiri dari:

- 466.285 Baris
- 75 Kolom
- 22 Tipe Data Kategorik
- 53 Tipe Data Numerik

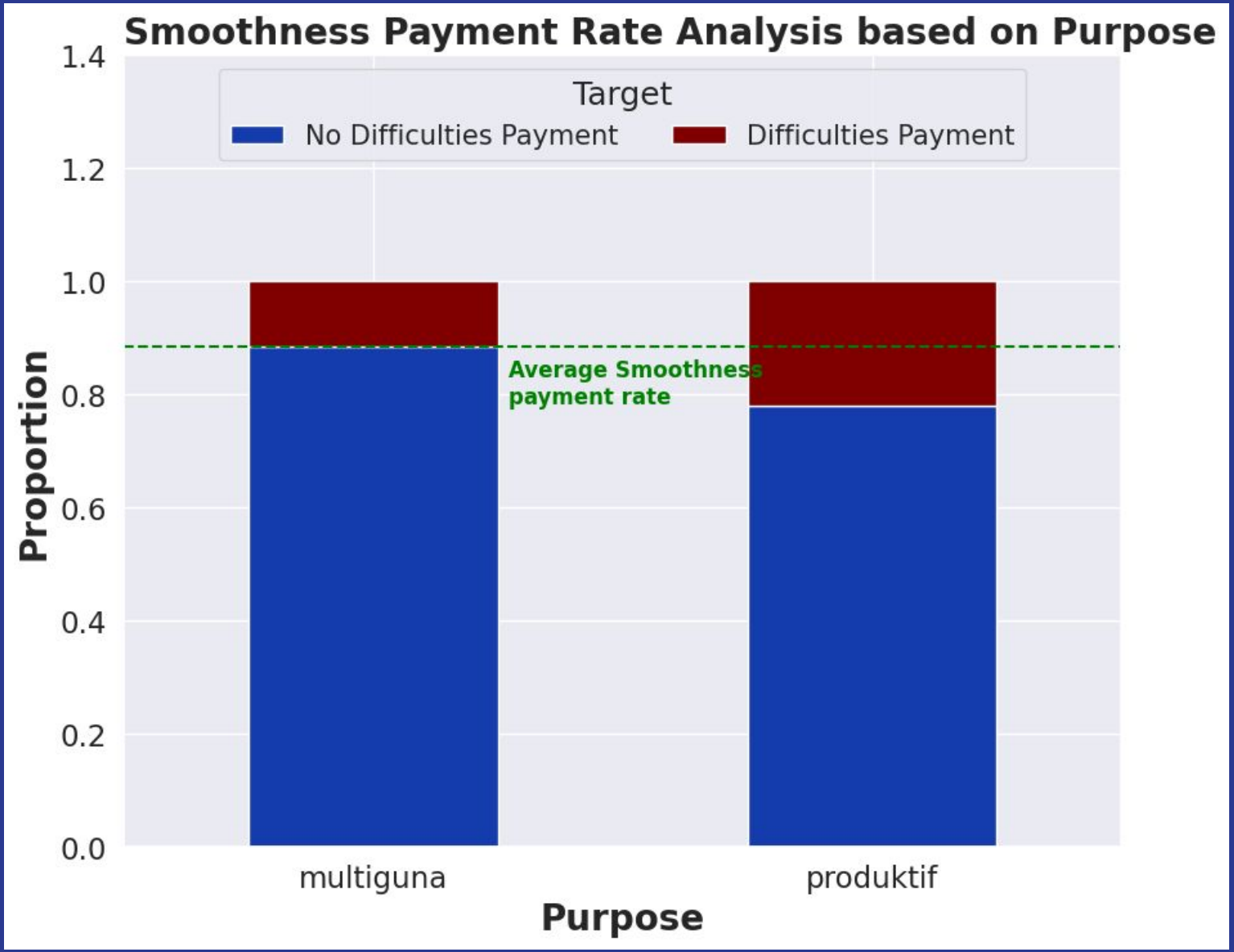
Insight



Semakin banyak *total principal received* atau total pokok hutang yang telah dibayarkan *customer*, proporsi keberhasilan pembayarannya juga cenderung tinggi dibandingkan total pokok hutang yang nominalnya lebih kecil.



- Semakin besar i*nterest rate* atau bunga pinjaman, *customer* cenderung mengalami kesulitan membayar. Terlihat dari proporsi keberhasilan pembayaran, dimana bunga pinjaman lebih dari **22%**, proporsi *customer* mengalami kesulitan bayar lebih dari **20%**.



Customer yang melakukan pengajuan pinjaman dana dengan tujuan produktif, proporsi mengalami kesulitan bayar cenderung lebih tinggi dibandingkan dengan tujuan multiguna (kebutuhan konsumtif).

Oleh karena itu perlu dilakukan analisis terhadap jenis bisnis *customer*, agar diketahui jenis bisnis seperti apa yang memiliki potensi mengakibatkan *customer* mengalami kesulitan pembayaran *credit*.

Data Cleansing & Preprocessing

- *Drop* kolom dengan *Missing Value* ≥ 50%.
- *Missing Value* kurang dari 50% dan di atas 1% di *input* dengan *median*.
- Tidak ada data *duplicated* di *dataset* ini.

- *Feature Selection* dilakukan dengan menggunakan Uji Statistik dan *Heatmap Correlation*. Hasilnya 21 fitur digunakan (tidak termasuk TARGET).

- Deteksi *outlier* dengan *box-plot* dan *QQPlot*. *Outlier* yang bertipe *global outlier* akan dihapus dari baris.

- Melakukan *feature engineering* pada variable/kolom **home_ownership, purpose, term**, dan **earliest_cr_line**.
- Menghapus kolom **earliest_year** (hasil *feature engineering* **earliest_cr_line**) yang memiliki nilai *negative*.

- Standard Scaler untuk Variabel numerik.
- Ordinal Encoding : Data ordinal dan data nominal yang memiliki 2 kelompok.
- OneHot Encoding : Data nominal > 2 kelompok.

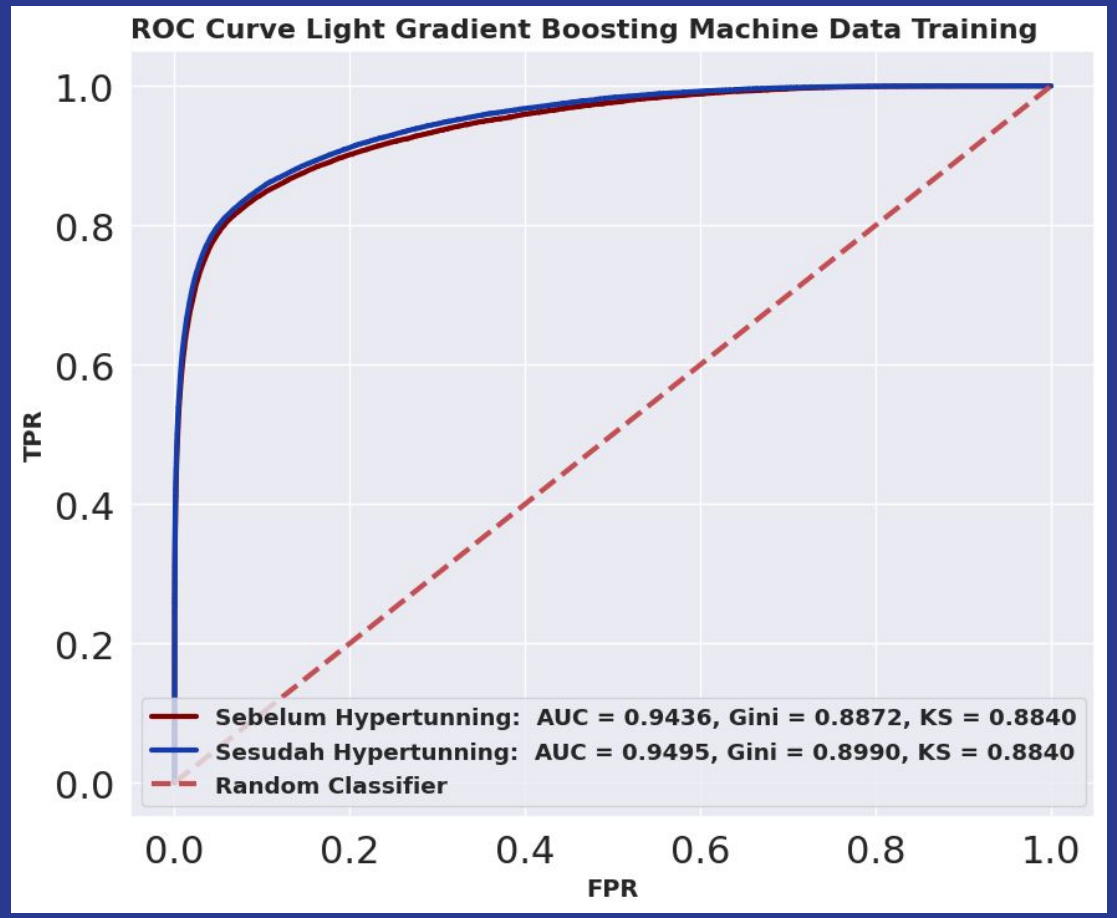
- Mengatasi *imbalanced data* menggunakan Class Weight.

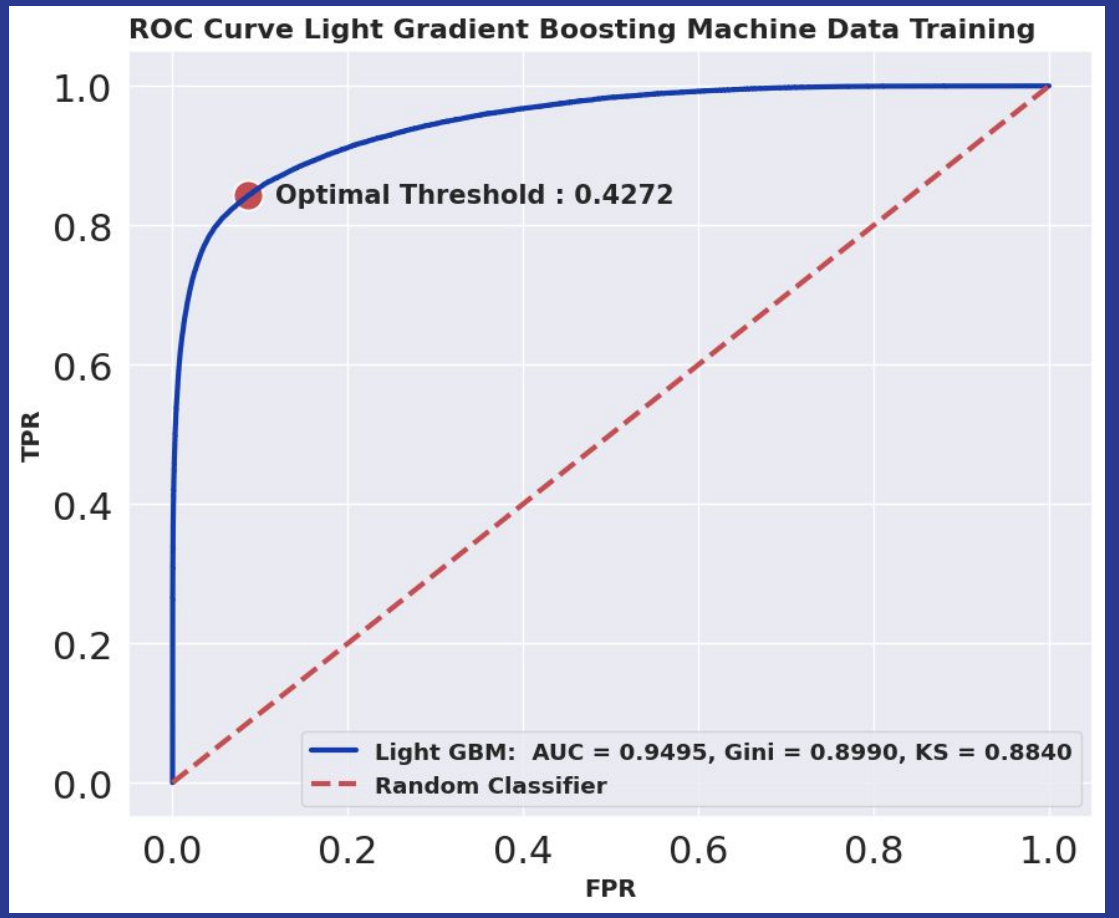
Modelling

Model	Training AUC_ROC	CV AUC_ROC (mean)	CV AUC_ROC (std)	Gap AUC_ROC
Logistic Regression	0.7863	0.8632	0.0024	0.0769
XGB	0.8369	0.9345	0.0018	0.0976
Decision Tree	1.0000	0.8238	0.0029	0.1762
Random Forest	0.9999	0.9141	0.0032	0.0857
Naive Bayes	0.5809	0.7973	0.0038	0.2163
LGBM	0.8733	0.9365	0.0014	0.0632

- Model yang digunakan adalah *Light Gradient Boosting Machine* (**LGBM**) dikarenakan *gap* antara *score* **AUC_ROC** data *training* dan *cross validation test* cenderung lebih kecil dibandingkan model lain.

- Selain itu standard deviasi pada model **LGBM** adalah yang paling kecil. Nilai standard deviasi yang kecil maka performa modelnya cenderung lebih konsisten.

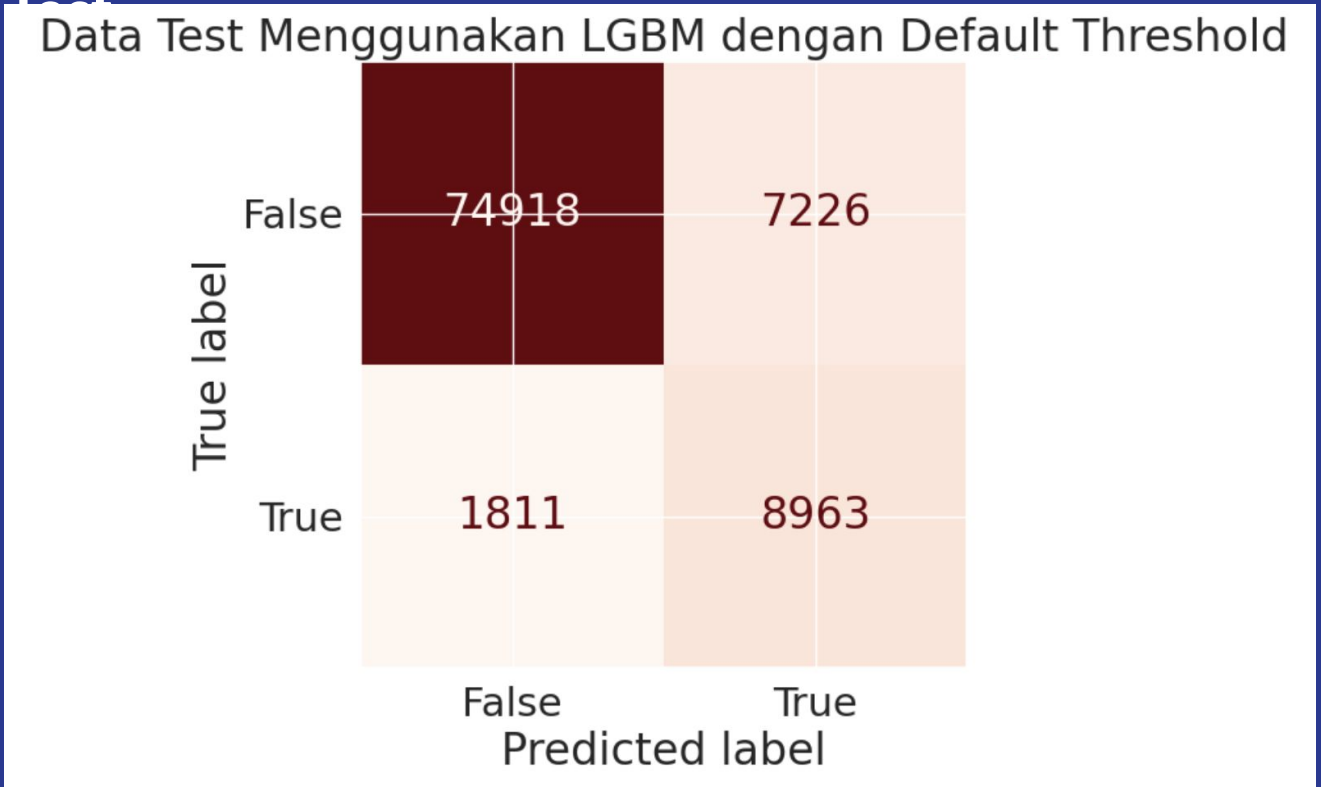


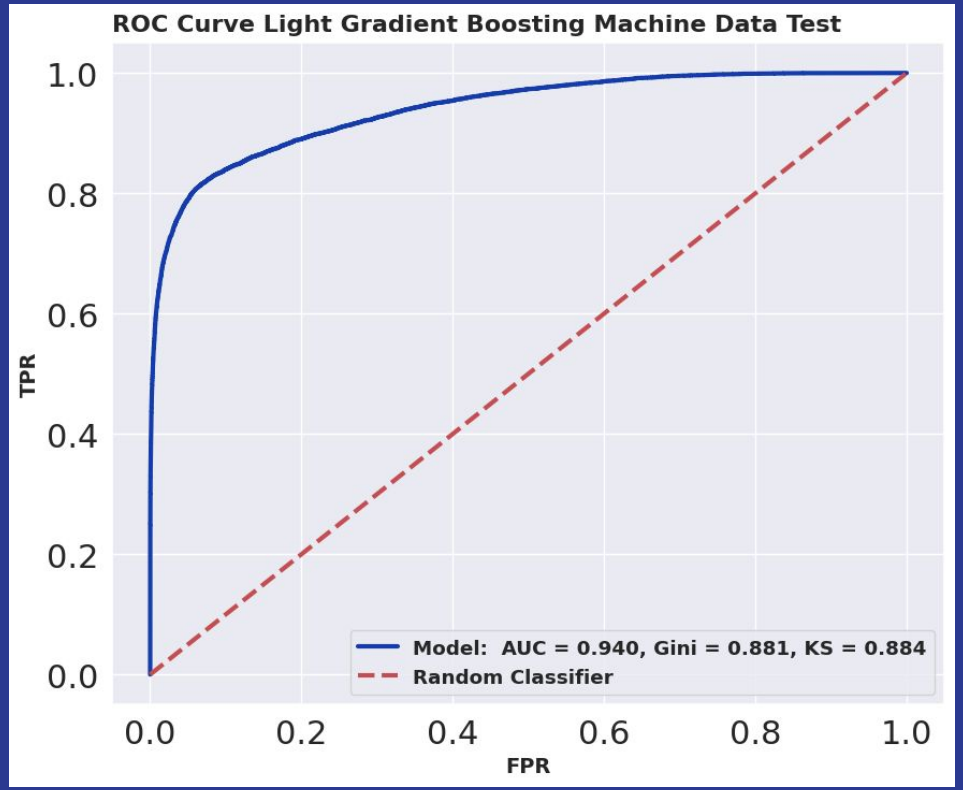


- Performa model lebih baik setelah dilakukan *hypertuning parameter*. Terlihat dari nilai *Gini* yang lebih besar dibandingkan sebelum *tuning hyperparameter*. Selain itu *AUC* setelah dilakukan *hypertuning* juga lebih besar nilainya.

- *Threshold* optimal yang diperoleh dengan menggunakan teknik *G-mean* adalah 0.427, dimana *customer* yang memiliki peluang lebih dari 0.427 akan terdeteksi kesulitan dalam pembayaran.

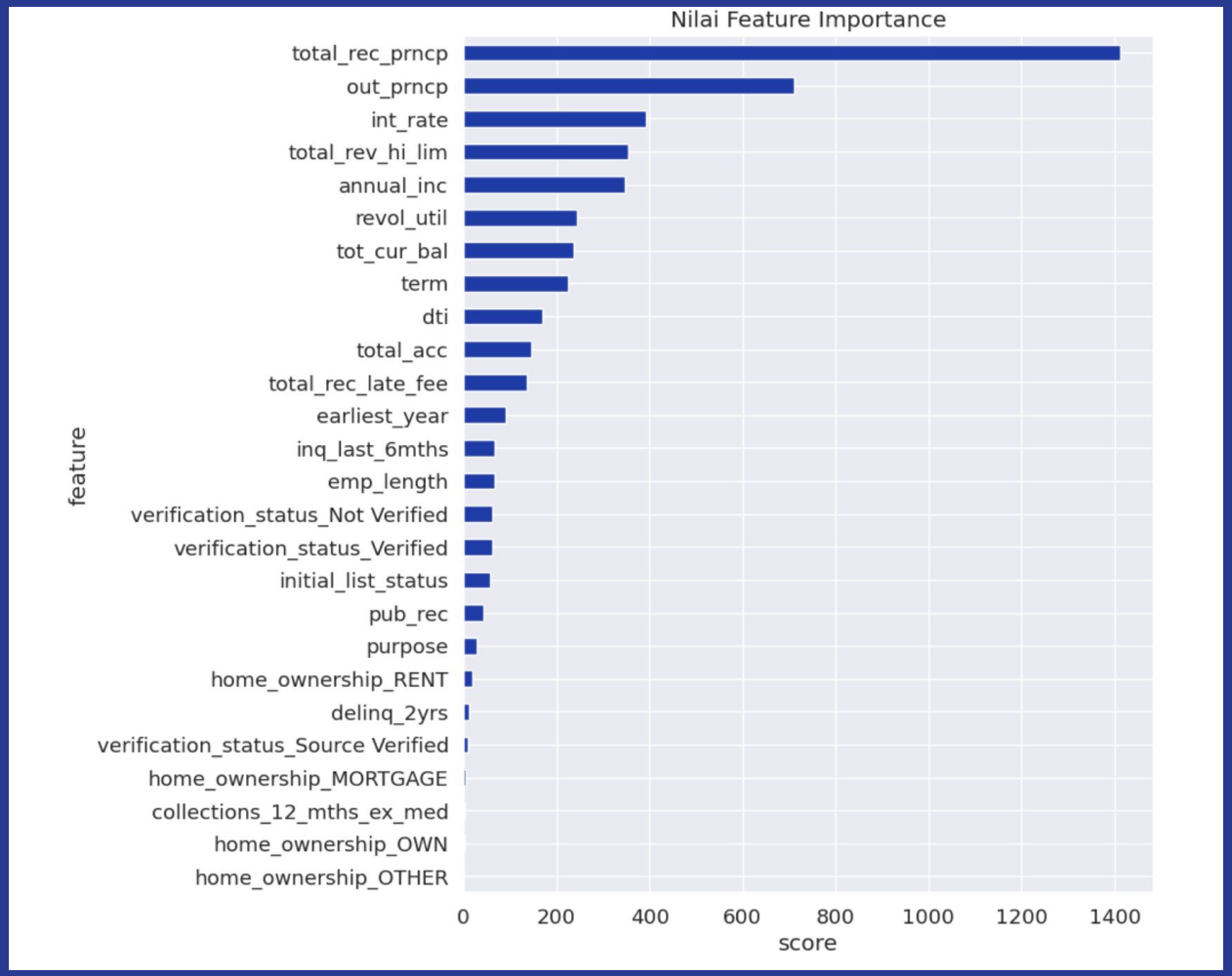
Implementasi Model pada Data Test

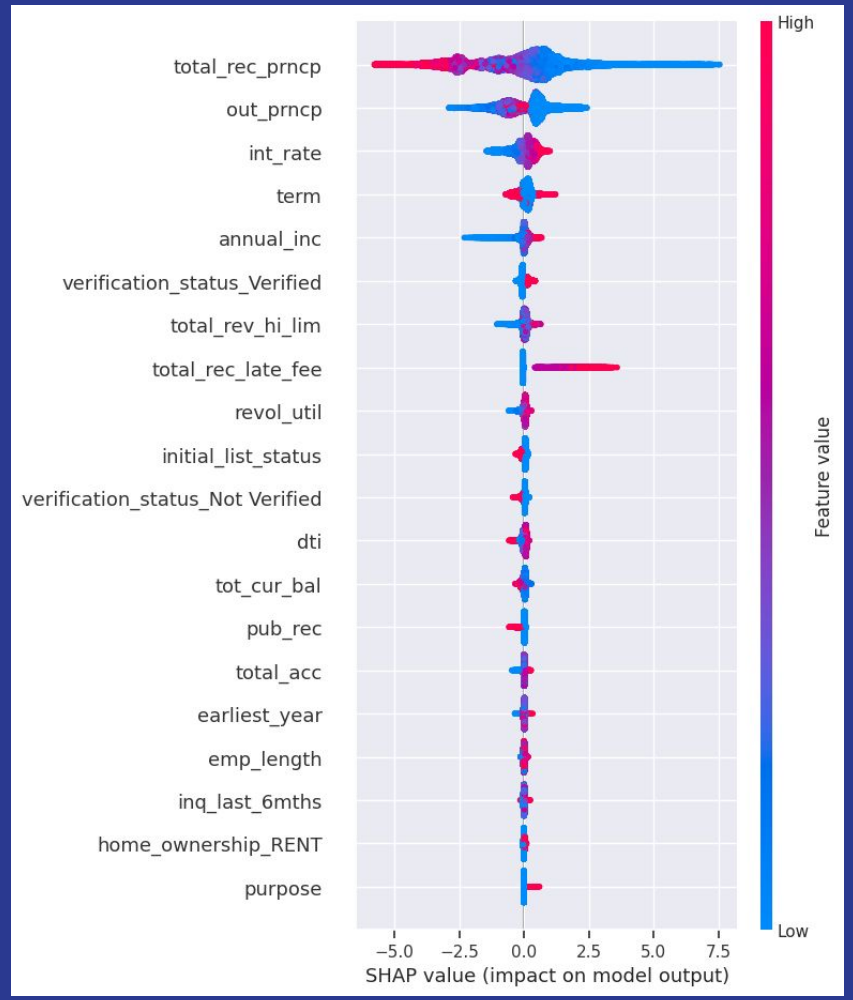




- Setelah diimplementasikan ke *Data Test* dari total **92.918** *customer*, **16.189** diantaranya diprediksi mengalami kesulitan pembayaran.

- Nilai *AUC* pada *Data Test* **0.94** dan nilai *Gini* serta *KS* masih di atas **0.8**. Performa model masih cukup baik dalam melakukan prediksi pada *Data Test*.





- Semakin tinggi *Total Principal Received* atau total pokok hutang yang telah dibayarkan *customer*, maka semakin besar peluang *customer* mengalami kesulitan pembayaran.
- Semakin tinggi *Outstanding Principal* atau sisa pokok hutang *customer,* maka peluang *customer* mengalami kesulitan pembayaran juga cenderung besar.
- Semakin tinggi *Interet Rate* atau bunga pinjaman, maka semakin besar peluang *customer* mengalami kesulitan pembayaran.

Link GitHub dan Youtube

Link GitHub :
https://github.com/contekan-si-al/Final-Task-ID-X-Partners-Data-Scientist/tree/main

Youtube :
https://youtu.be/i_RqII990_0

Rakamin Academy

Terima Kasih