Rakamin Academy

id/x partners

Classification Model

for Loan Default Risk Prediction

Virtual Internship Program - Data Scientist

Oleh : Friska Yuliantika Saputri

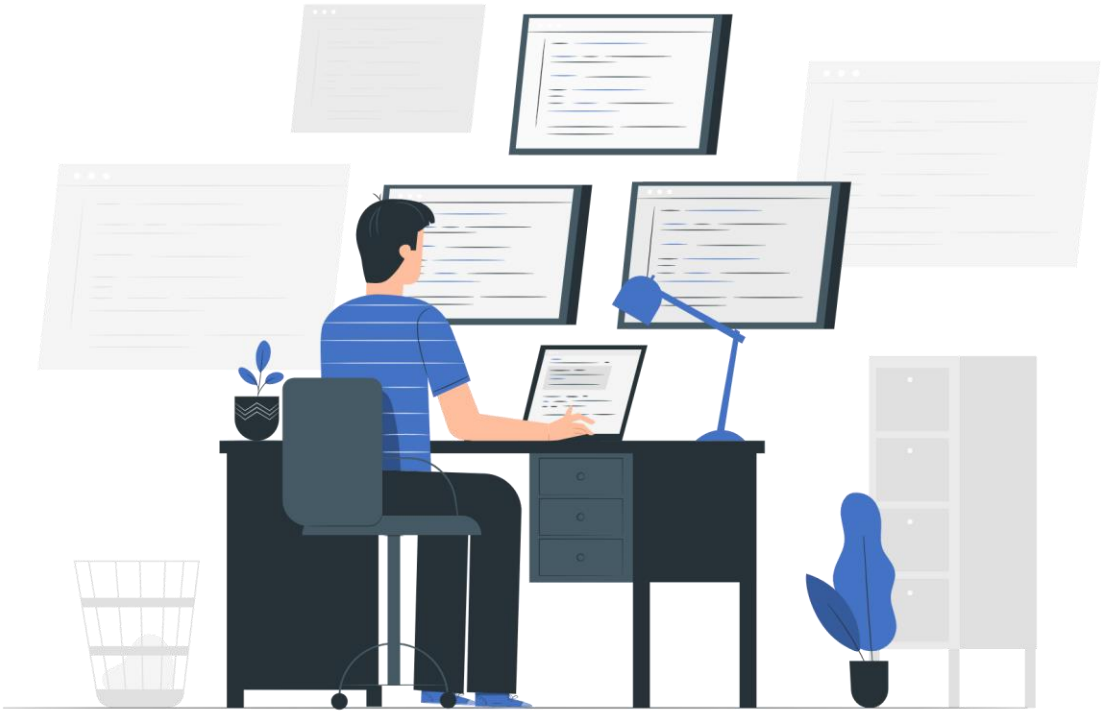
Link GiHub Repo

Problem Statement

- Masalah bisnis yang sedang dihadapi adalah terkait pinjaman yang diberikan kepada customer, sebagian diantaranya mengalami keterlambatan/ kesulitan pembayaran.

- Goals : Memperkecil resiko kredit akibat terjadinya gagal pembayaran credit dari customer

- Objective :
  - Melakukan prediktif model untuk mengetahui customer yang mengalami kesulitan bayar dan lancar dalam pembayaran.
  - Menemukan faktor penting dari customer yang lancar dalam pembayaran

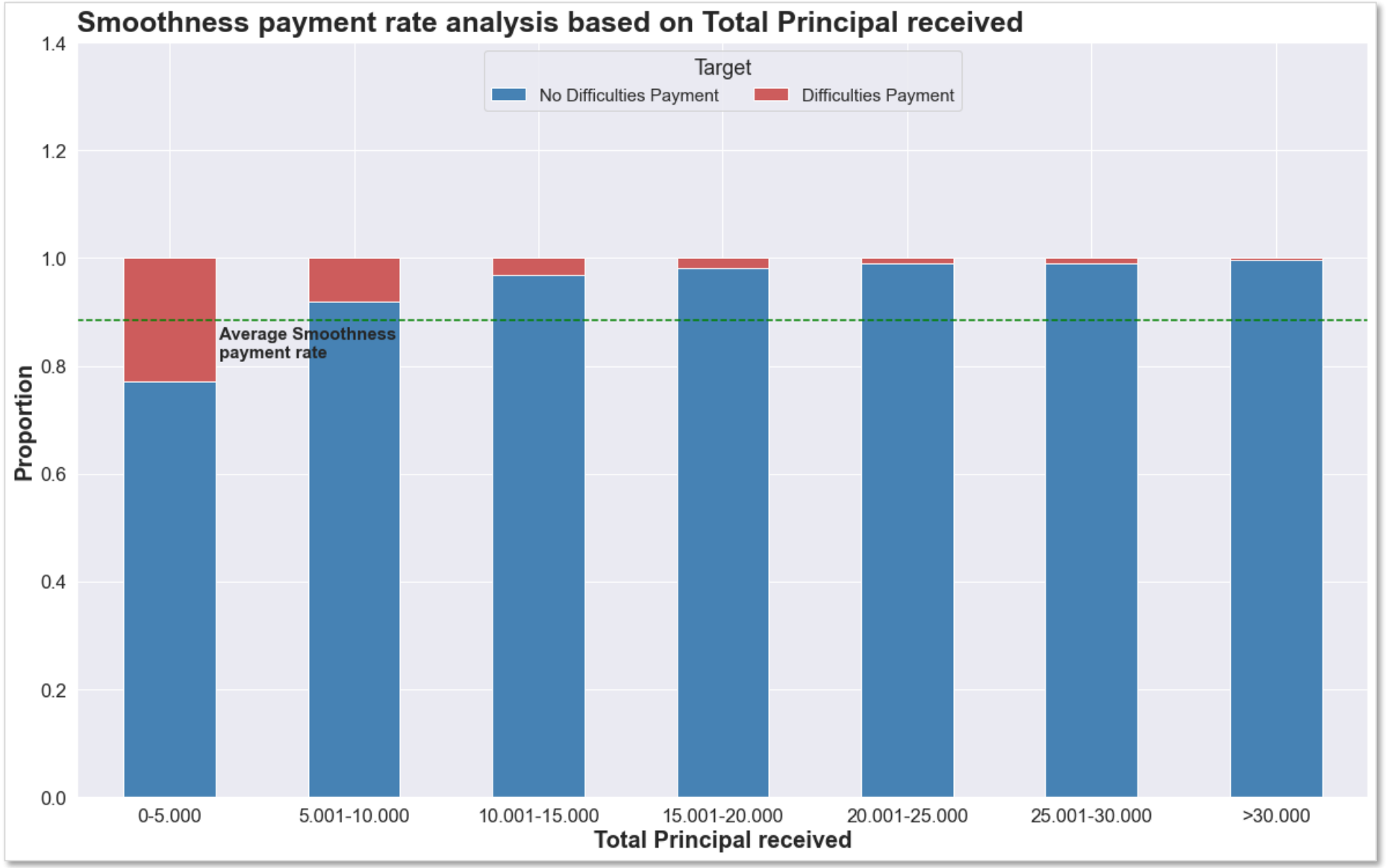


Dataset

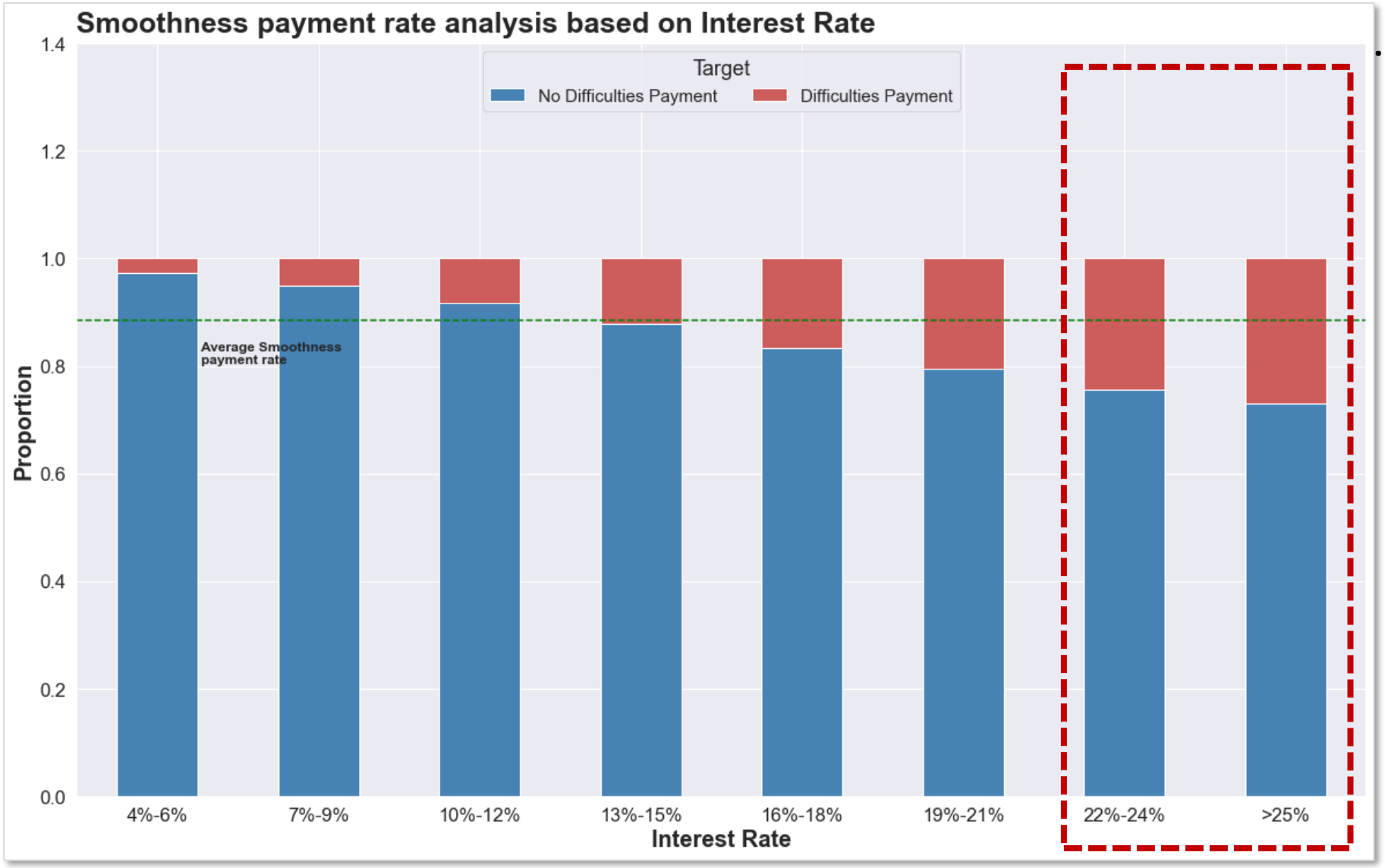
Dataset sample yang digunakan berisi data pinjaman yang diterima maupun yang ditolak dan terdiri dari

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

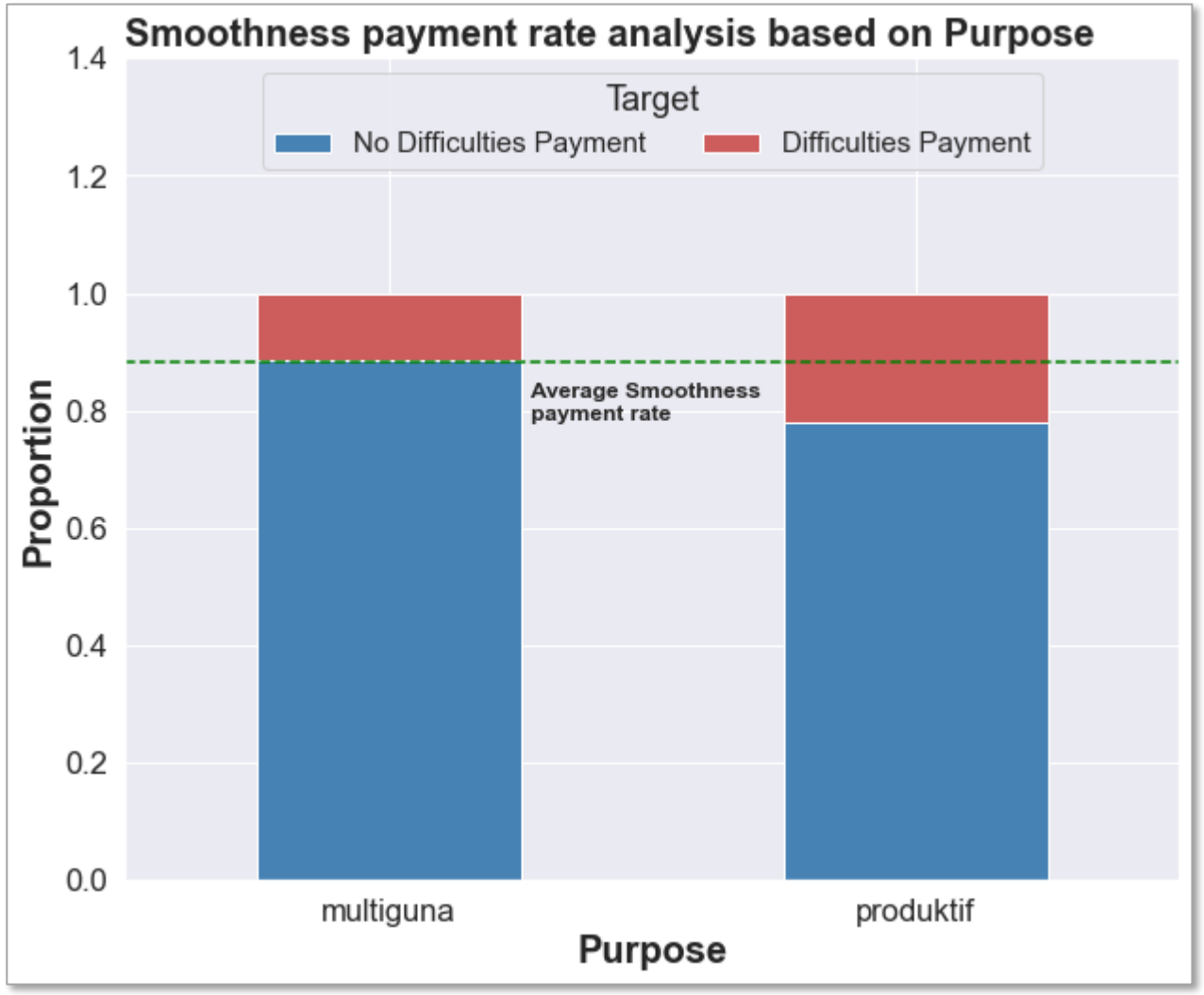
Top Insight



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



- Semakin besar interest rate atau bunga pinjaman, customer cenderung mengalami kesulitan bayar. 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% diinput dengan median.
- Missing Value <1 % akan didrop dari baris dataset
- 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

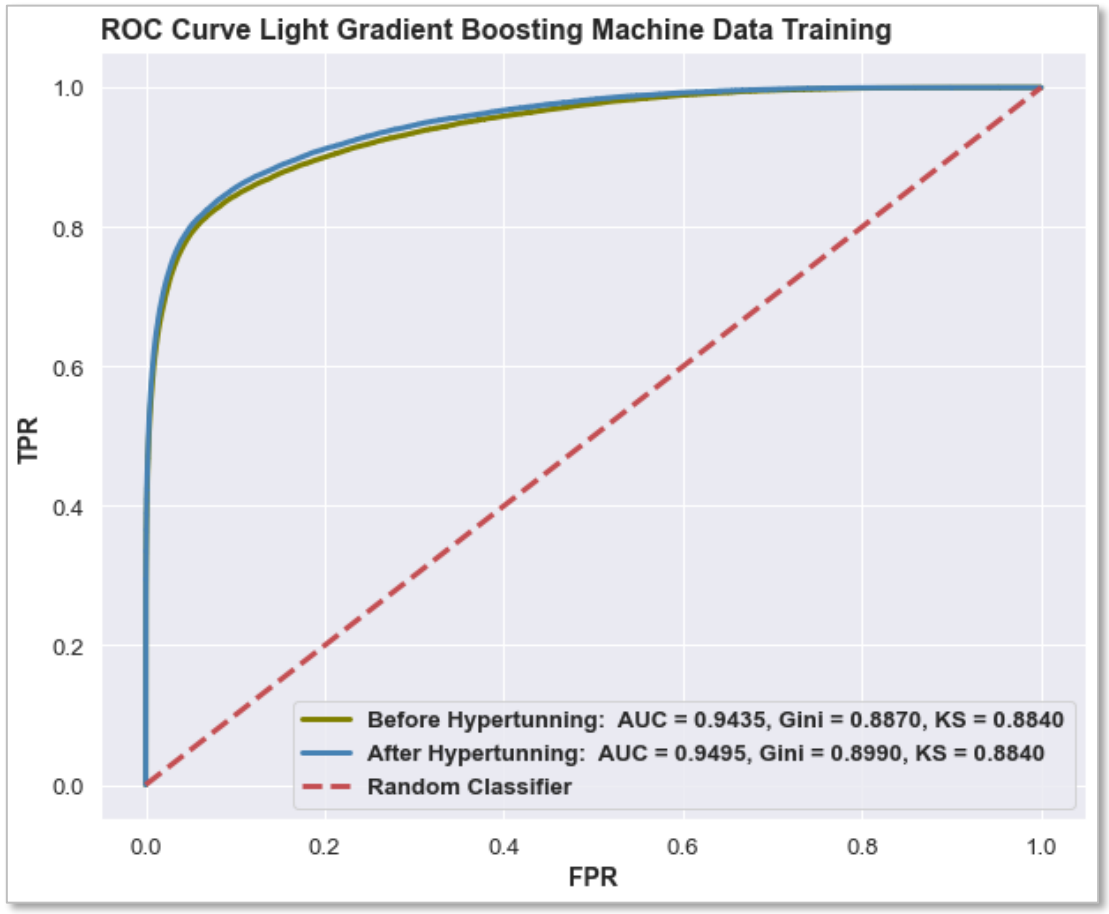
- Imbalanced data menggunakan Class Weight

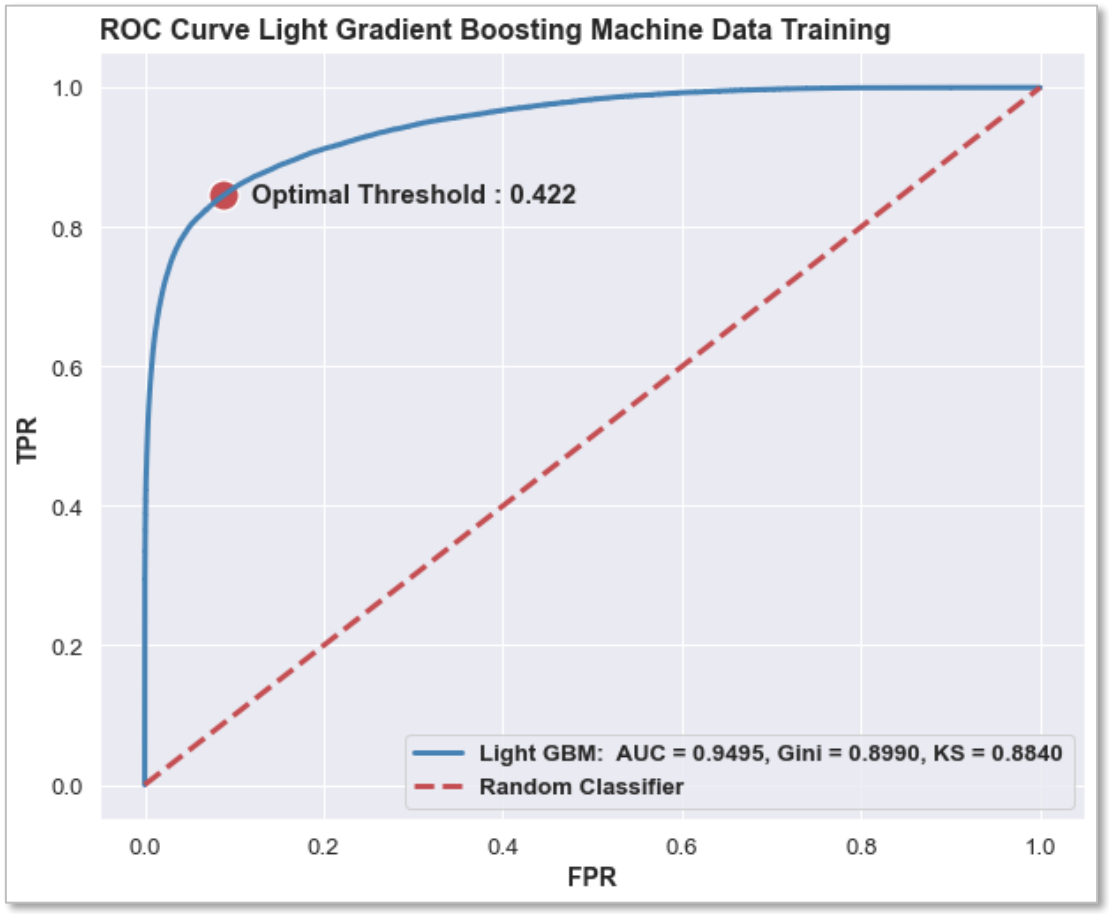
Modeling

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	0.8606	0.9459	0.0016	0.0853
Decision Tree	1.0000	0.8240	0.0020	0.1760
Random Forest	0.9999	0.9139	0.0027	0.0860
Naive Bayes	0.5809	0.7973	0.0038	0.2164
LGBM	0.8731	0.9361	0.0013	0.0629

- 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, dengan nilai standard deviasi yang kecil maka performa modelnya cenderung lebih konsisten

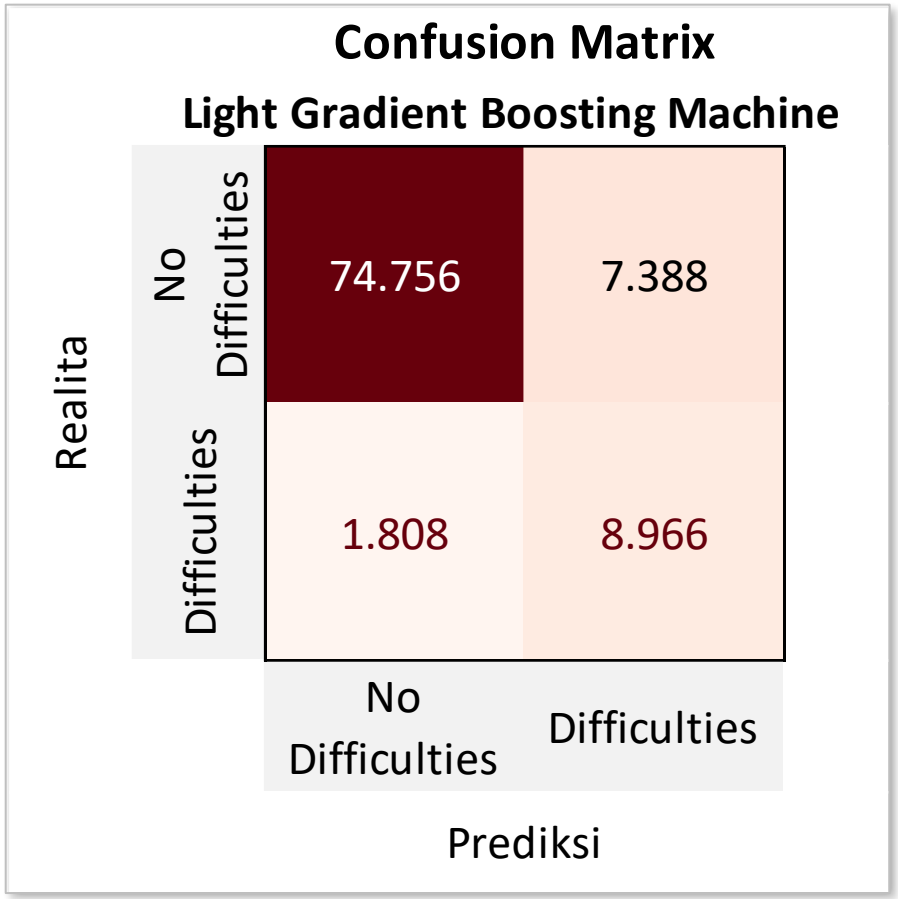
Hypertuning Parameter & Tuning Threshold Model

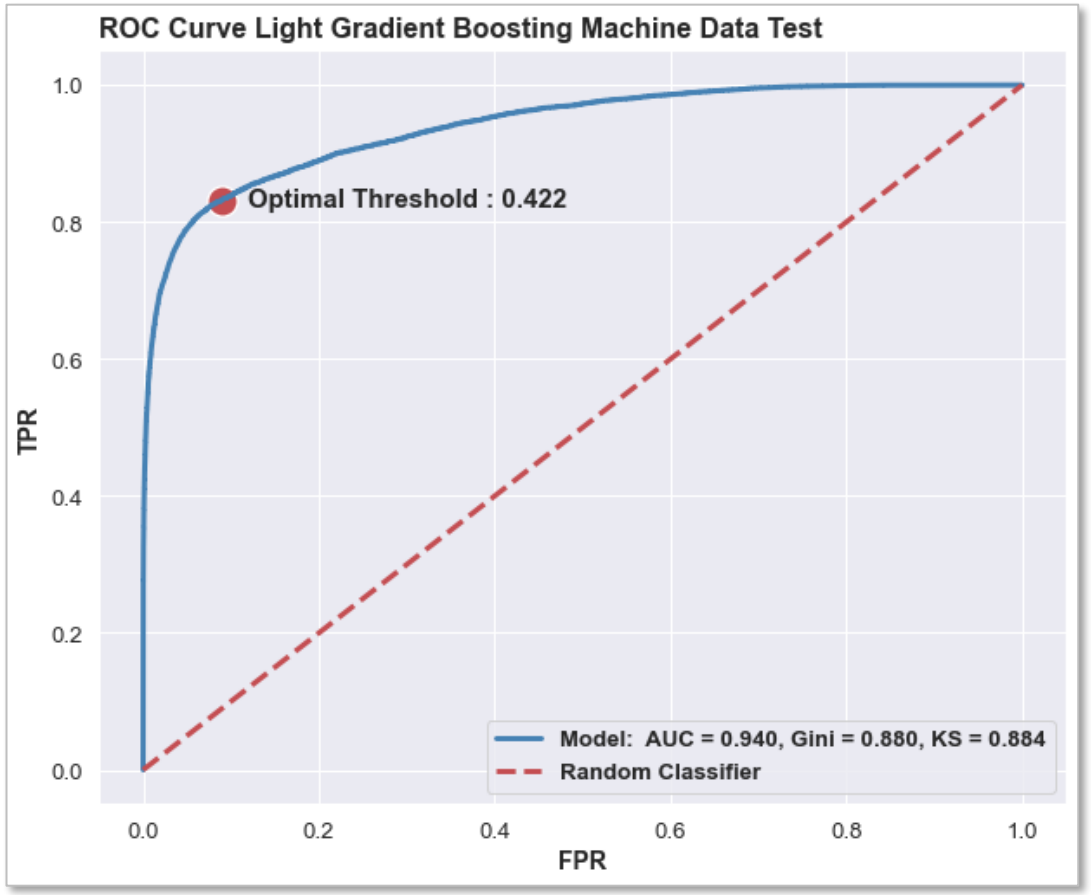




- Performa model lebih baik setelah dilakukan hypertuning parameter, terlihat dari nilai Gini yang lebih besar dibandingkan sebelum tuning hyperparameter. Selain itu AUC setelah dilakukan hypertuning juga lebih besar nilainya
- Threshold optimal yang diperoleh dengan menggunakan teknik Gmean adalah 0.422, dimana customer yang memiliki peluang lebih dari 0.422 akan terdeteksi kesulitan dalam pembayaran

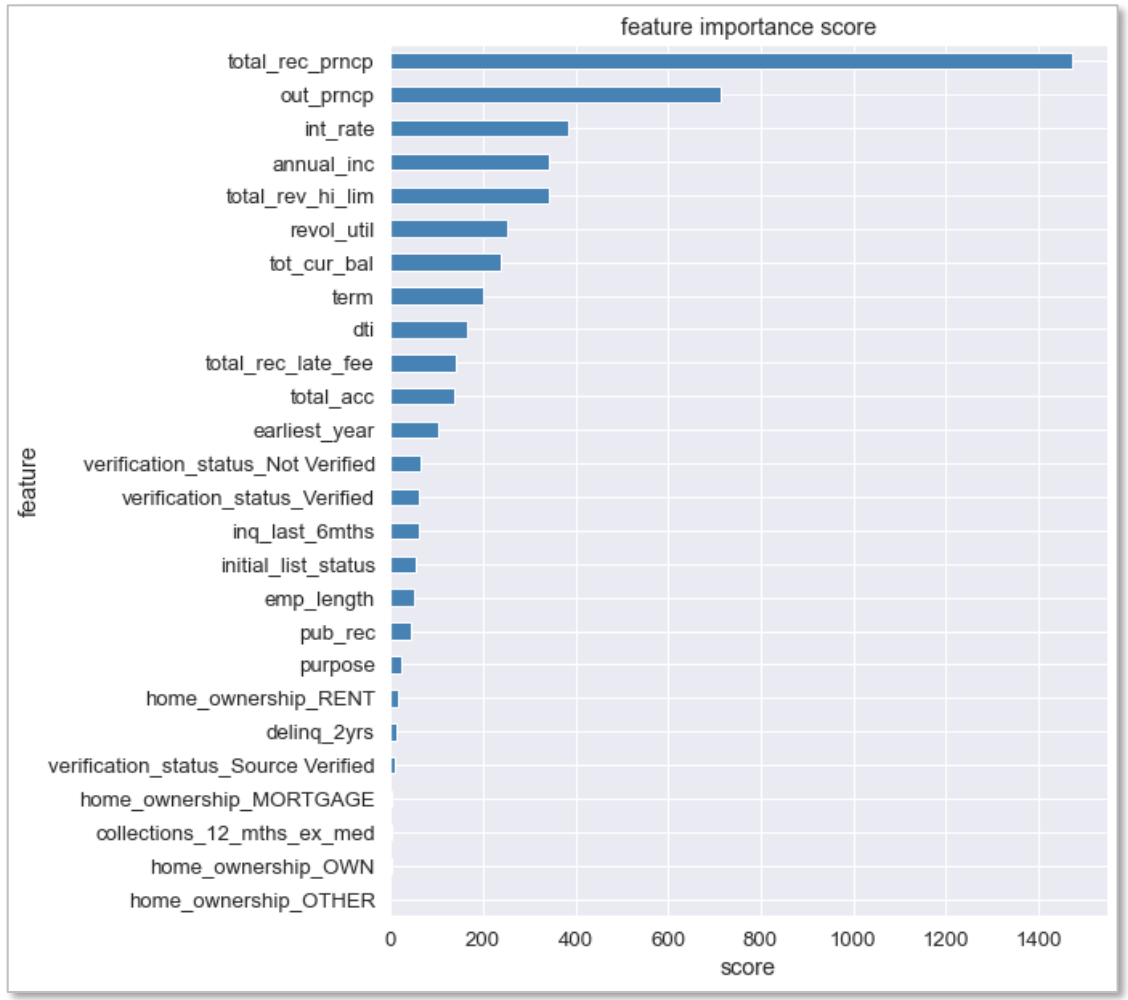
Implementasi Model pada Data Test

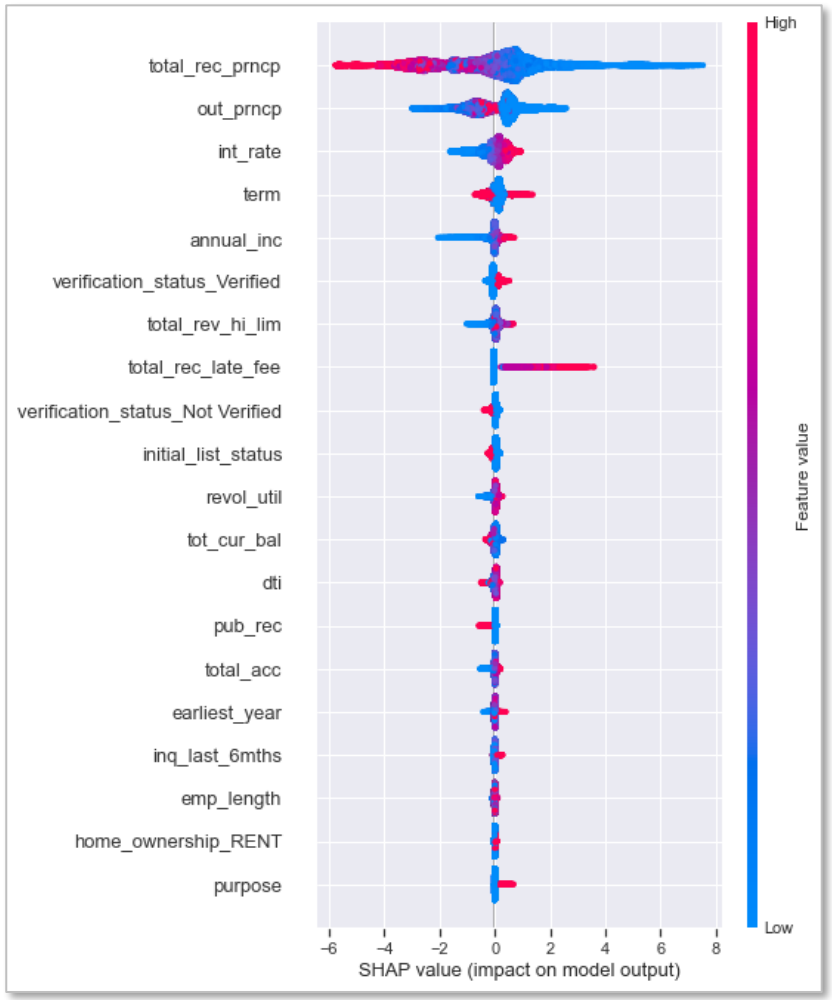




- Setelah diimplementasikan ke Data Test dari total 92.918 customer, 16.354 diantaranya diprediksi mengalami kesulitan pembayaran.
- Nilai AUC pada Data Test 0.940, dan nilai Gini serta KS masih di atas 0.8. Performa model masih cukup baik dalam melakukan prediksi pada Data Test.

Feature Importance dan Shap Value





- 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

Terima kasih