Classification Model

for Loan Default Risk Prediction

Virtual Internship Program - Data Scientist

Oleh : Friska Yuliantika S

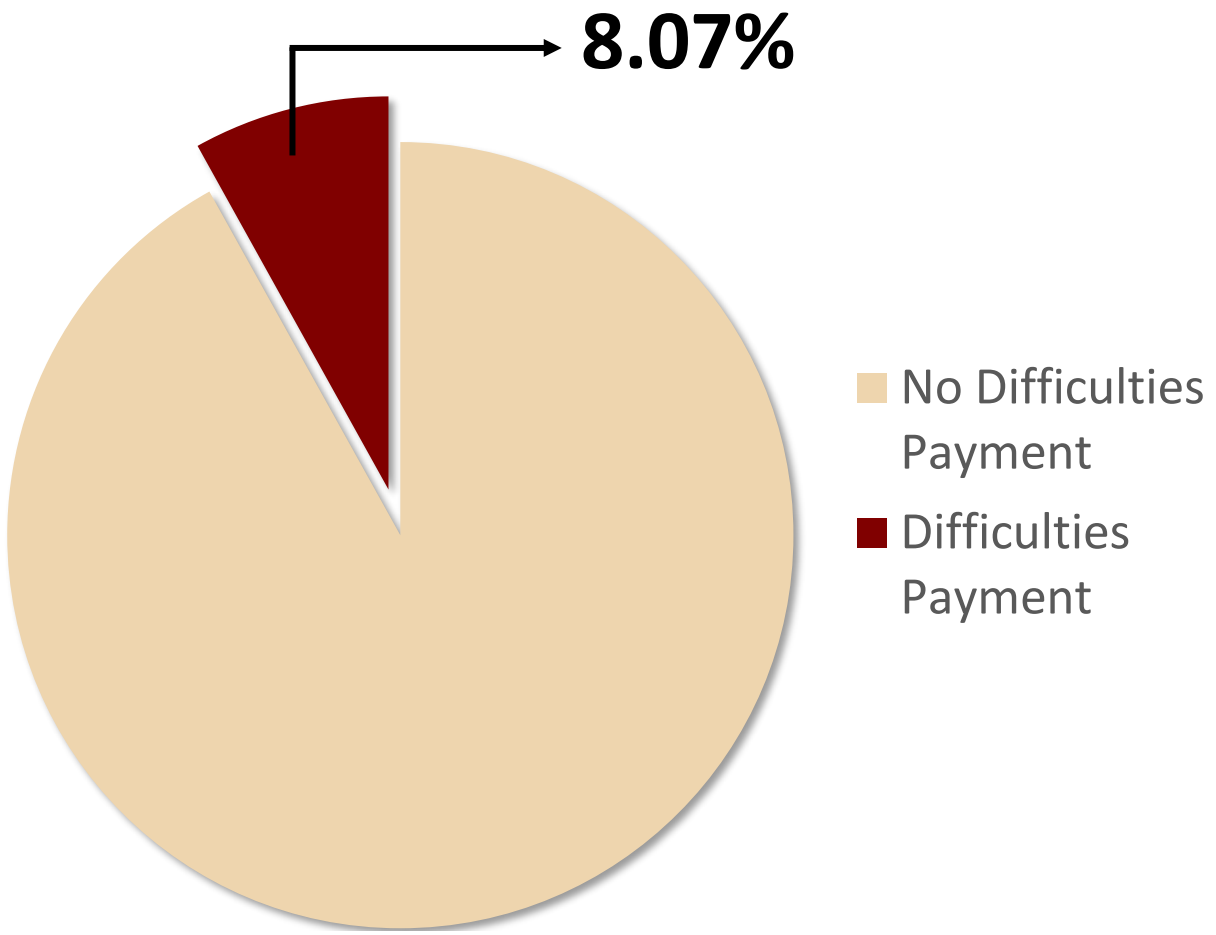
Link GitHub Repo

Problem Statement

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

- Goals : Menurunkan difficulty payment rate of customer

- Objective :
  Memprediksi customer yang lancar dalam pembayaran
  Menemukan faktor penting dari customer yang lancar dalam pembayaran



Dataset

Data yang digunakan pada model ini :

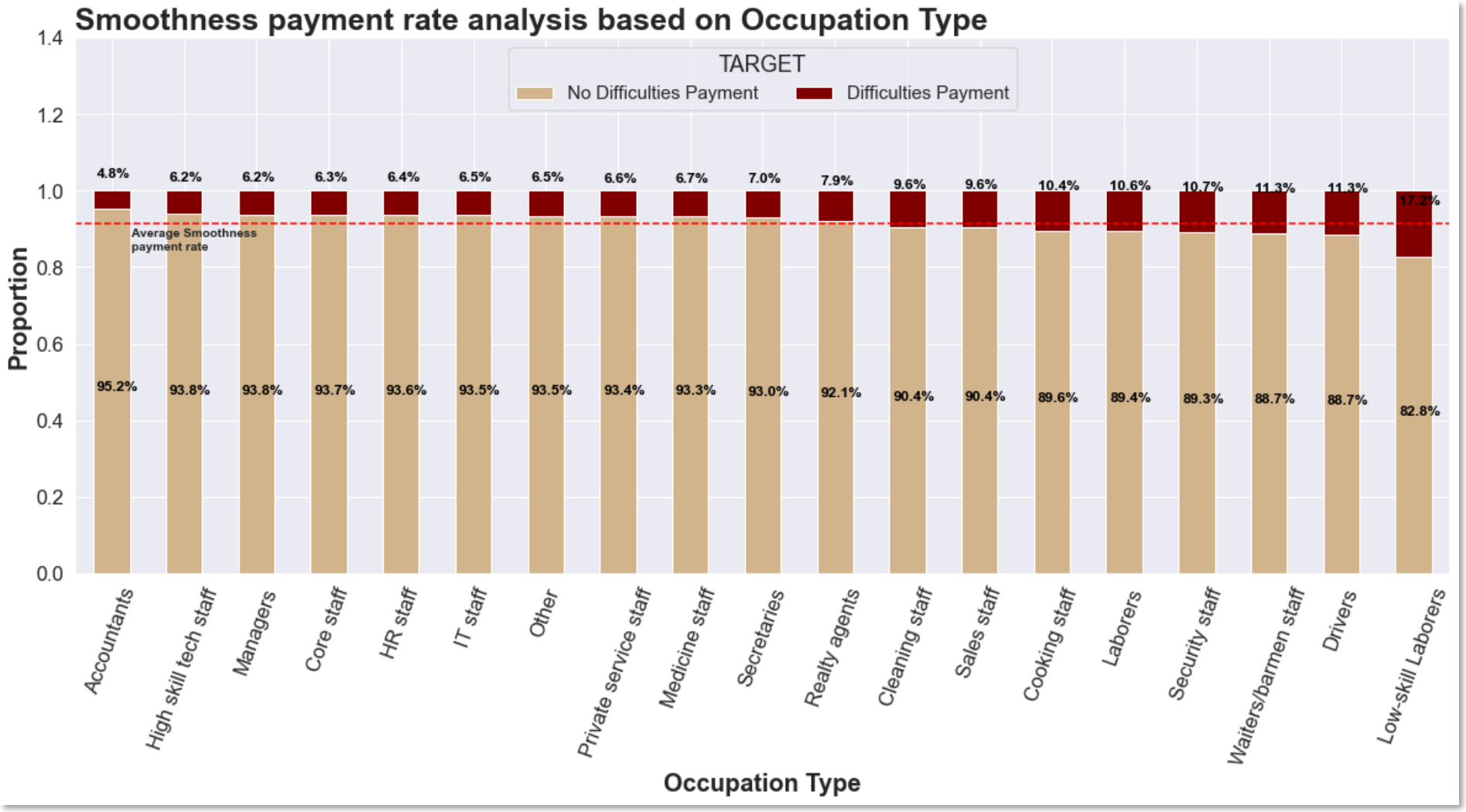
1. application_{train|test}

Data statis untuk semua aplikasi. Satu baris mewakili satu pinjaman dalam sampel data

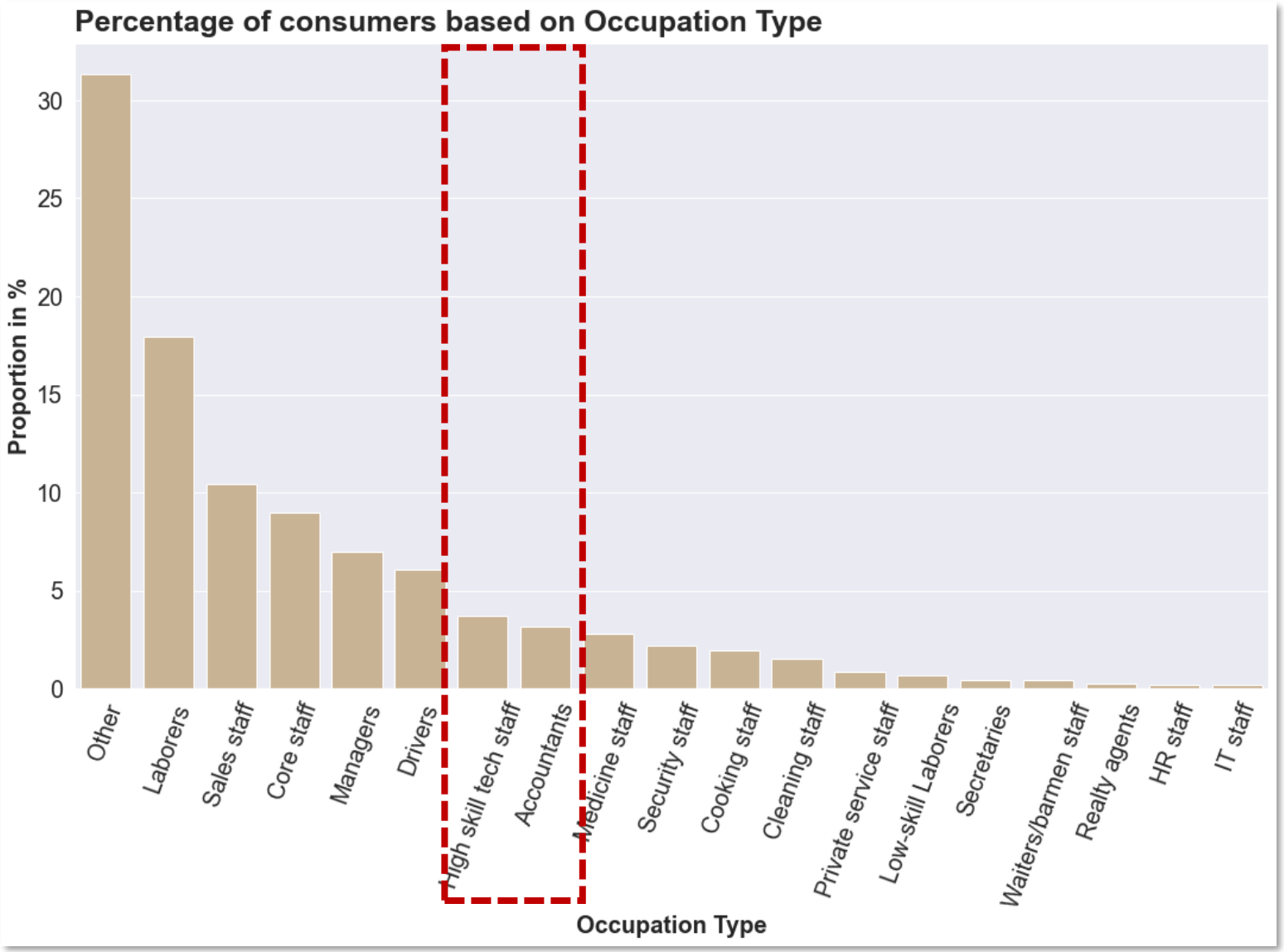
2. Bureau

History Credit customer yang tercatat di Biro Kredit

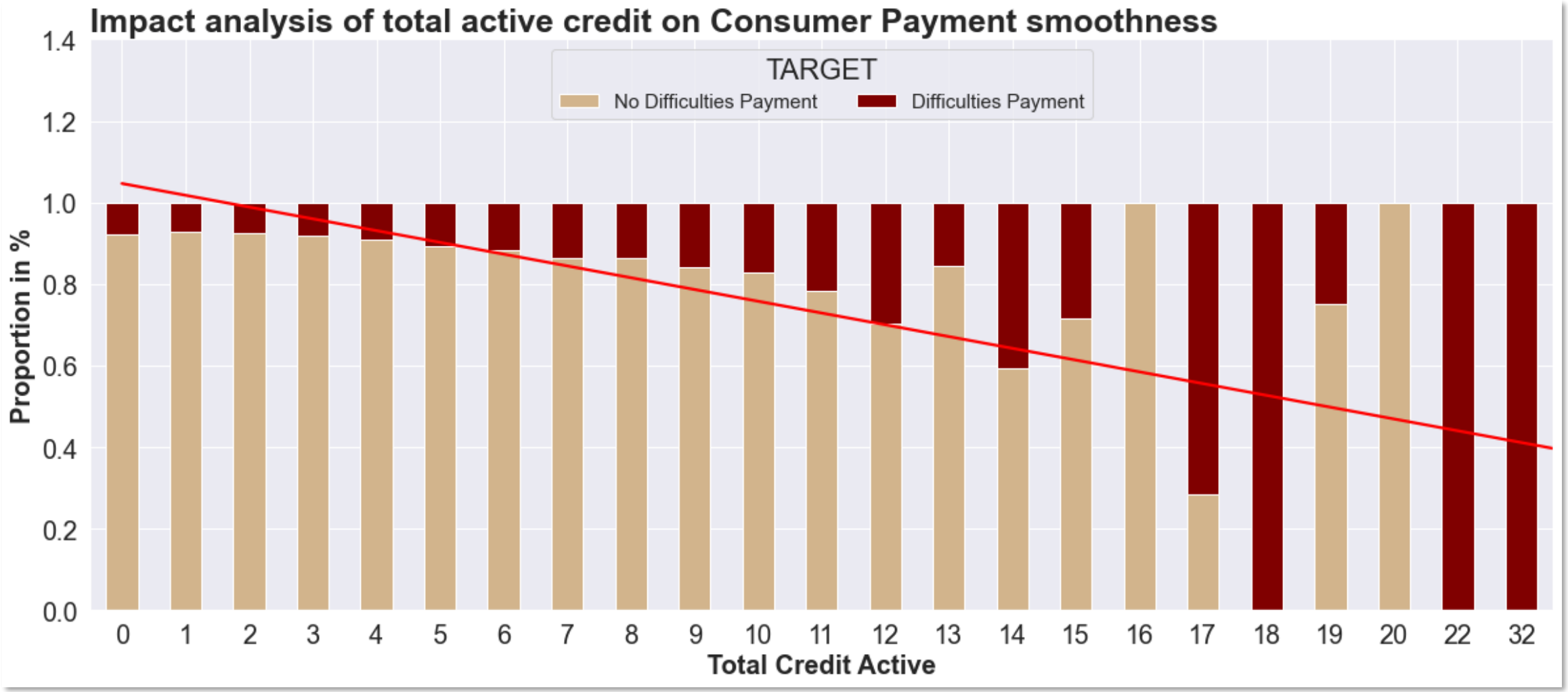
Top Insight



Accountant dan High Skill Tech Staff adalah 2 tipe pekerjaan dimana proporsi keberhasilan pembayaran paling tinggi, di atas rata-rata smoothness rate secara keseluruhan. Sedangkan Low-skill Laborers memiliki proporsi kelancaran pembayaran terendah dibandingkan tipe pekerjaan lain.



- Proporsi tipe Accountant dan High Skill Tech Staff masih di bawah 5%. Oleh karena itu, perusahaan dapat mengoptimalkan kegiatan marketing untuk konsumen yang potensial pada tipe tersebut agar tertarik mengajukan pinjaman.



Semakin banyak kontrak aktif yang dimiliki customer, kecenderungan pembayarannya semakin sulit. Oleh karena itu, perusahaan harus berhati-hati pada customer yang memiliki banyak kontrak aktif terutama jumlah kontrak aktif di atas 5.

Data Cleansing & Preprocessing

- Melakukan feature engineering pada tabel bureau
- Melakukan Merging data agregat pada Bureau dan data Application Train

- Drop kolom dengan Missing Value ≥ 50%
- Missing Value di bawah 50% diinput dengan median, mode, nilai 0 dan string 'Other'
- Tidak ada data duplicated

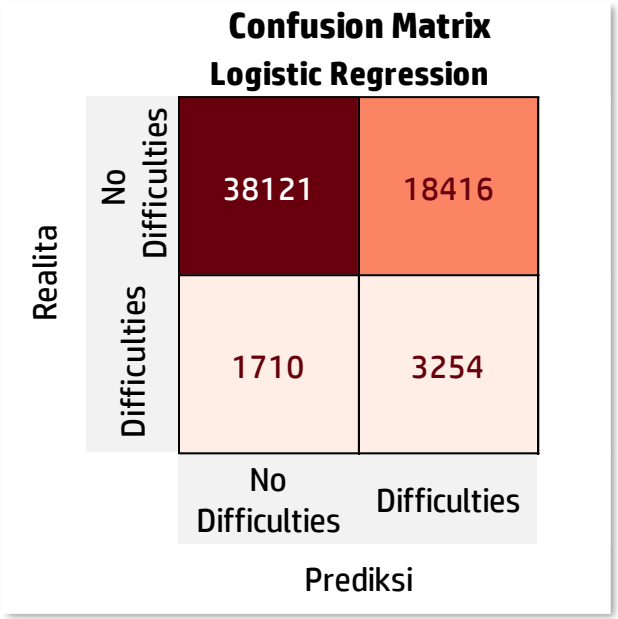
- Feature Selection dilakukan dengan menggunakan Uji Statistik dan heatmap Correlation. Hasilnya 18 fitur digunakan (tidak termasuk TARGET)

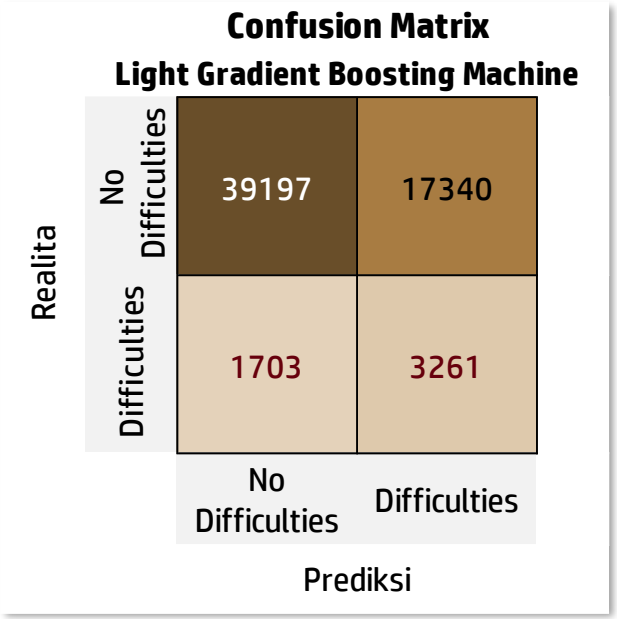
- Deteksi outlier dengan box-plot dan QQPlot, outlier yang bertipe global outlier akan dihapus dari baris

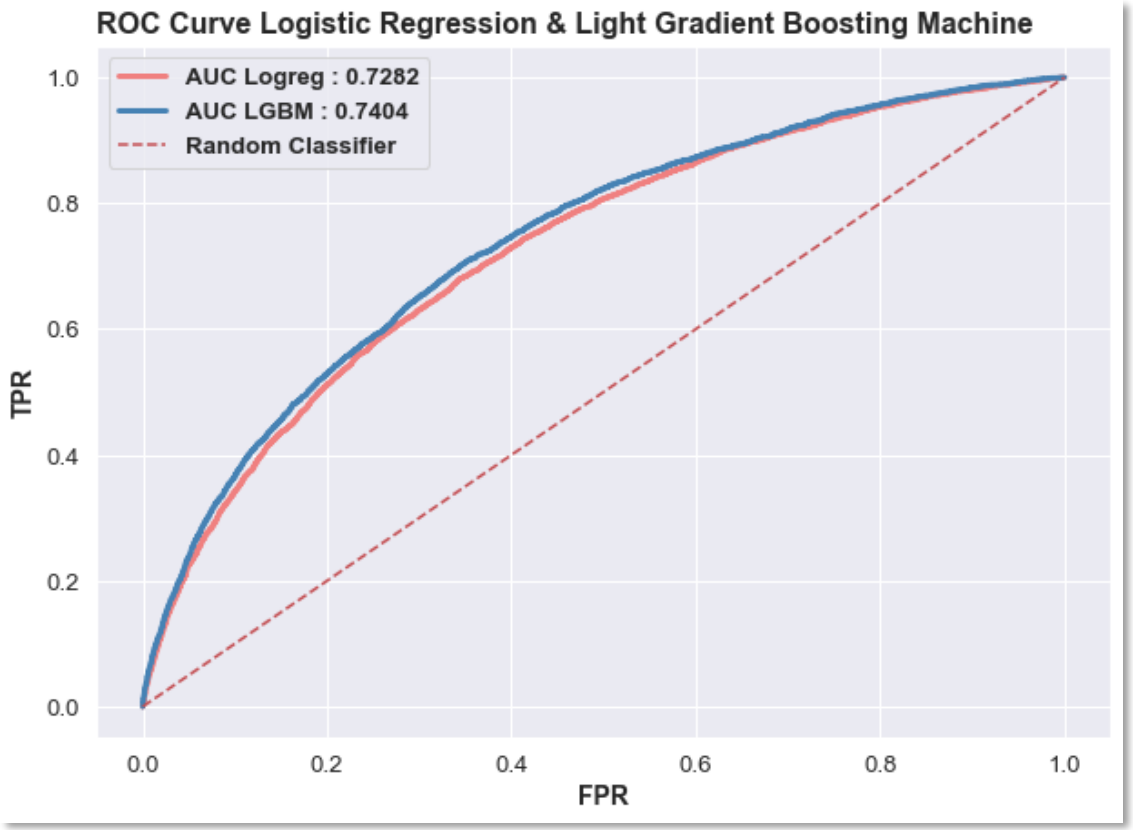
- Standard Scaler untuk Variabel numerik
- Ordinal Encoding : Data ordinal dan data nominal yang memiliki 2 kelompok
- OneHot Encoding : Data nominal >2 kelompok

- Imbalanced data menggunakan Class Weight

Modeling



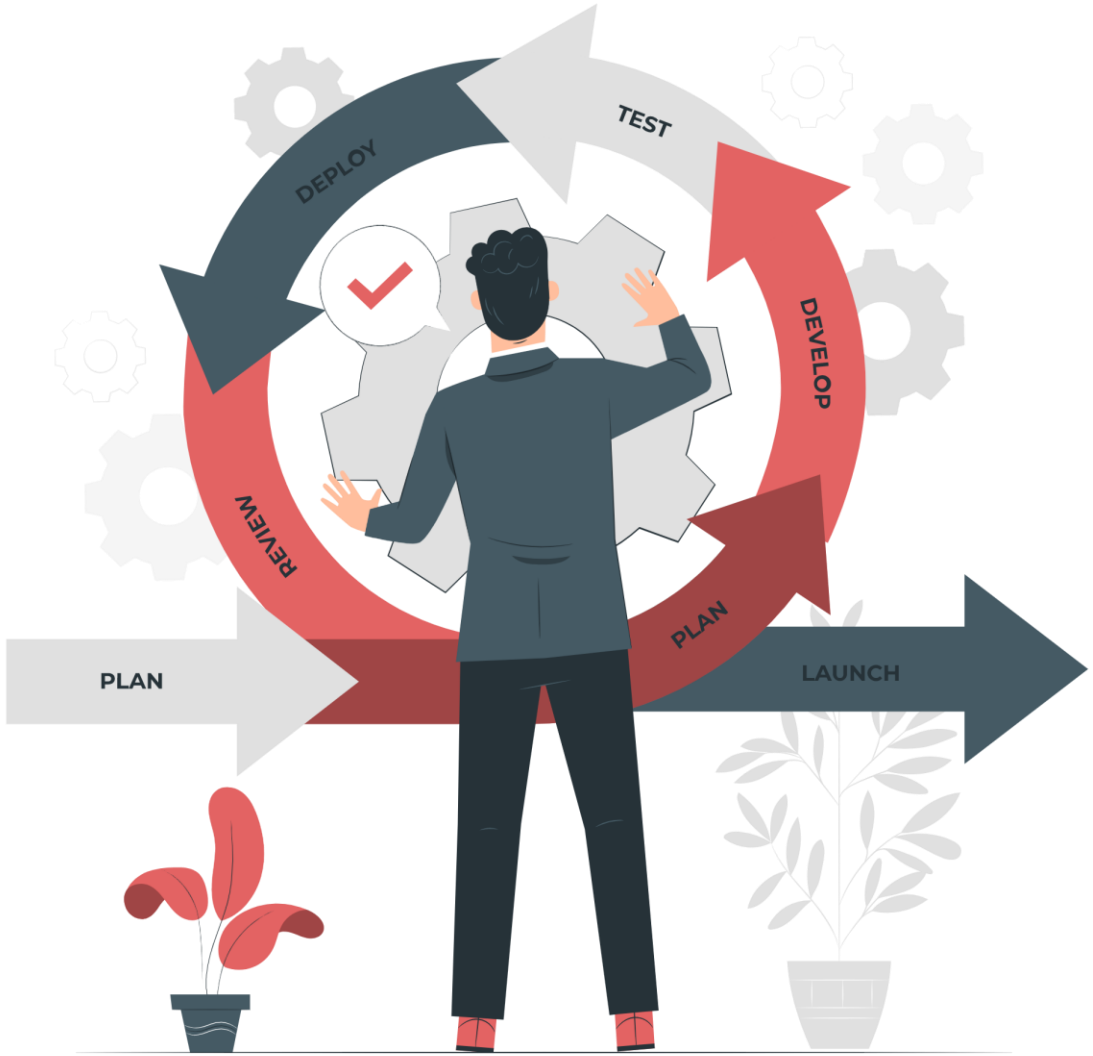




Recall dari kedua model adalah 66%. Model yang akan dipilih adalah Light GBM karena AUC nya lebih tinggi diandingkan dengan Logreg

Prediksi Data Application Test

SK ID CURR	Prediksi	Peluang Lancar Bayar	Peluang Tidak Lancar Bayar
100001	Lancar Bayar	0.597	0.403
100005	Kesulitan Bayar	0.350	0.650
100013	Lancar Bayar	0.771	0.229
100028	Lancar Bayar	0.628	0.372
100038	Kesulitan Bayar	0.394	0.606
100042	Lancar Bayar	0.554	0.446
100057	Lancar Bayar	0.646	0.354
100065	Lancar Bayar	0.531	0.469
100066	Lancar Bayar	0.879	0.121
100067	Lancar Bayar	0.590	0.410
100074	Lancar Bayar	0.515	0.485
100090	Kesulitan Bayar	0.481	0.519
100091	Kesulitan Bayar	0.430	0.570
100092	Kesulitan Bayar	0.452	0.548
100106	Kesulitan Bayar	0.464	0.536



Terima kasih