

HOME CREDIT DEFAULT RECOGNITION



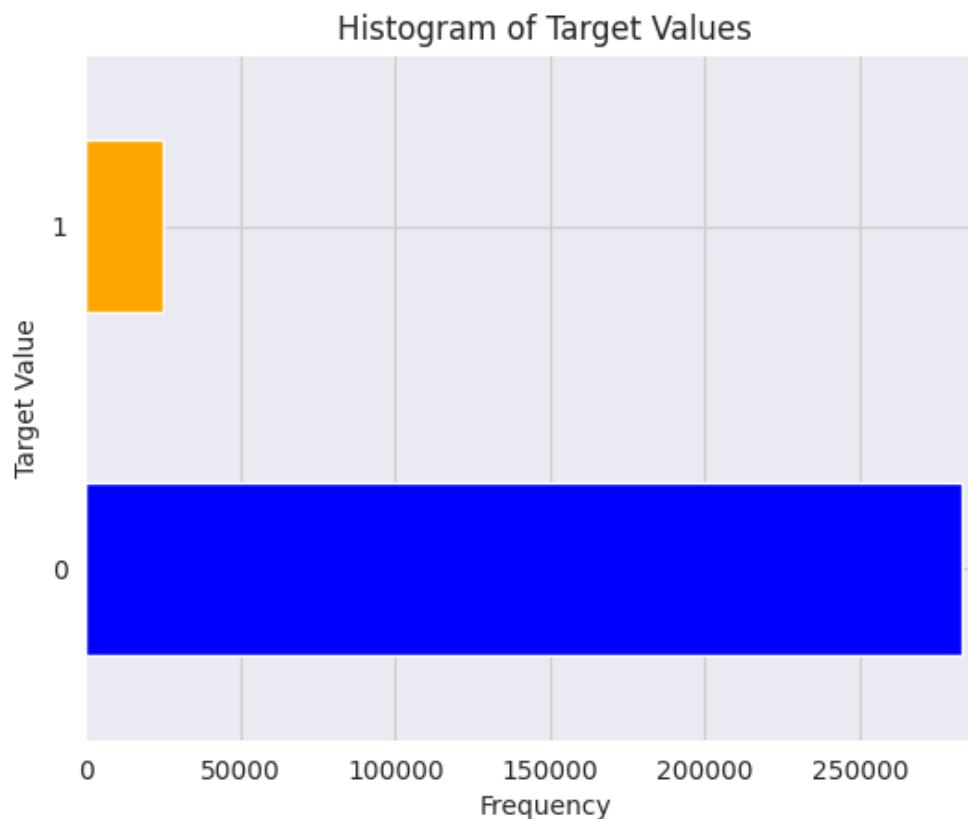
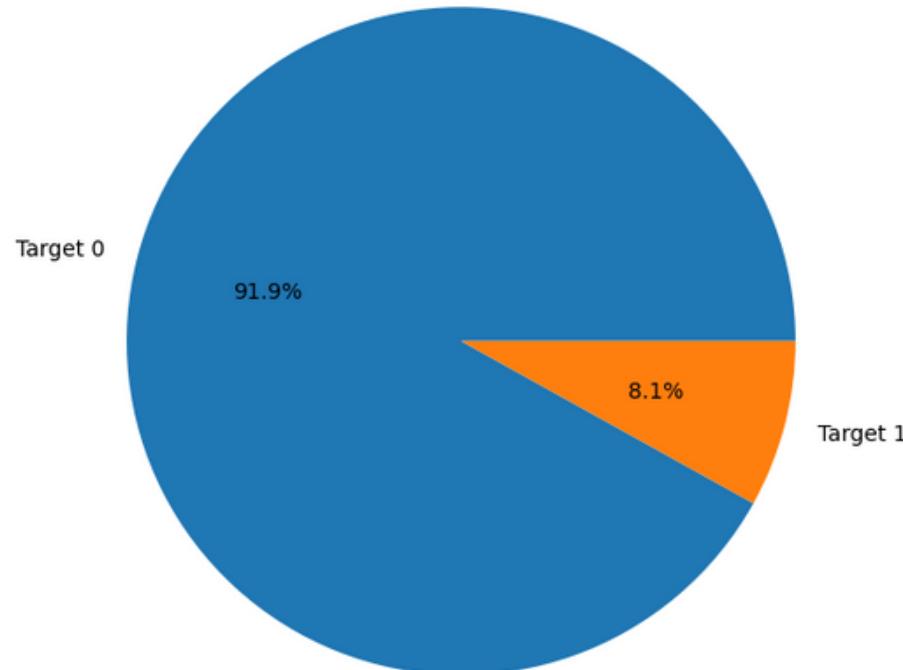
-BY FITYAN SETYAWAN

fityansetyawan27@gmail.com
github repo : <https://github.com/Seth1495/Final-Project-Home-Credit-seth.git>
github : Seth1945

PROBLEM OVERVIEW



Saat ini perusahaan memiliki tingkat gagal bayar peminjam mencapai 8.1%, dalam sektor keuangan angka tersebut tergolong besar dengan jumlah kerugian yang tentunya berdampak besar pada perusahaan



GOALS

Mengurangi tingkat gagal bayar peminjam pada perusahaan, melalui rekomendasi bisnis yang aplikatif

OBJECTIVE

1. Membuat model machine learning yang mampu mengklasifikasikan peminjam dengan resiko gagal bayar
2. Memberikan rekomendasi dan insight business.





DATA SET UNDERSTANDING

TABLES:

Application train : Tabel utama yang digunakan untuk melatih dan menguji model machine learning

Application test : Tabel utama yang digunakan untuk mengevaluasi kinerja model machine learning

Bureau : Semua kredit klien sebelumnya disediakan oleh lembaga keuangan lain yang ada dilaporkan ke Biro Kredit (untuk klien yang memiliki pinjaman).

Bureau Balance : Saldo bulanan kredit sebelumnya di Biro Kredit.

POS cash balance : Saldo bulanan POS (point of sales) sebelumnya dan pinjaman tunai itu yang dimiliki pemohon dengan Home Credit.

Credit card balance : Saldo bulanan dari kartu kredit sebelumnya yang dimiliki pemohon Home Credit.

Previous application : Semua ajuan sebelumnya untuk pinjaman Home Credit dari klien yang memiliki pinjaman di Home Credit.

Installments payment : Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.

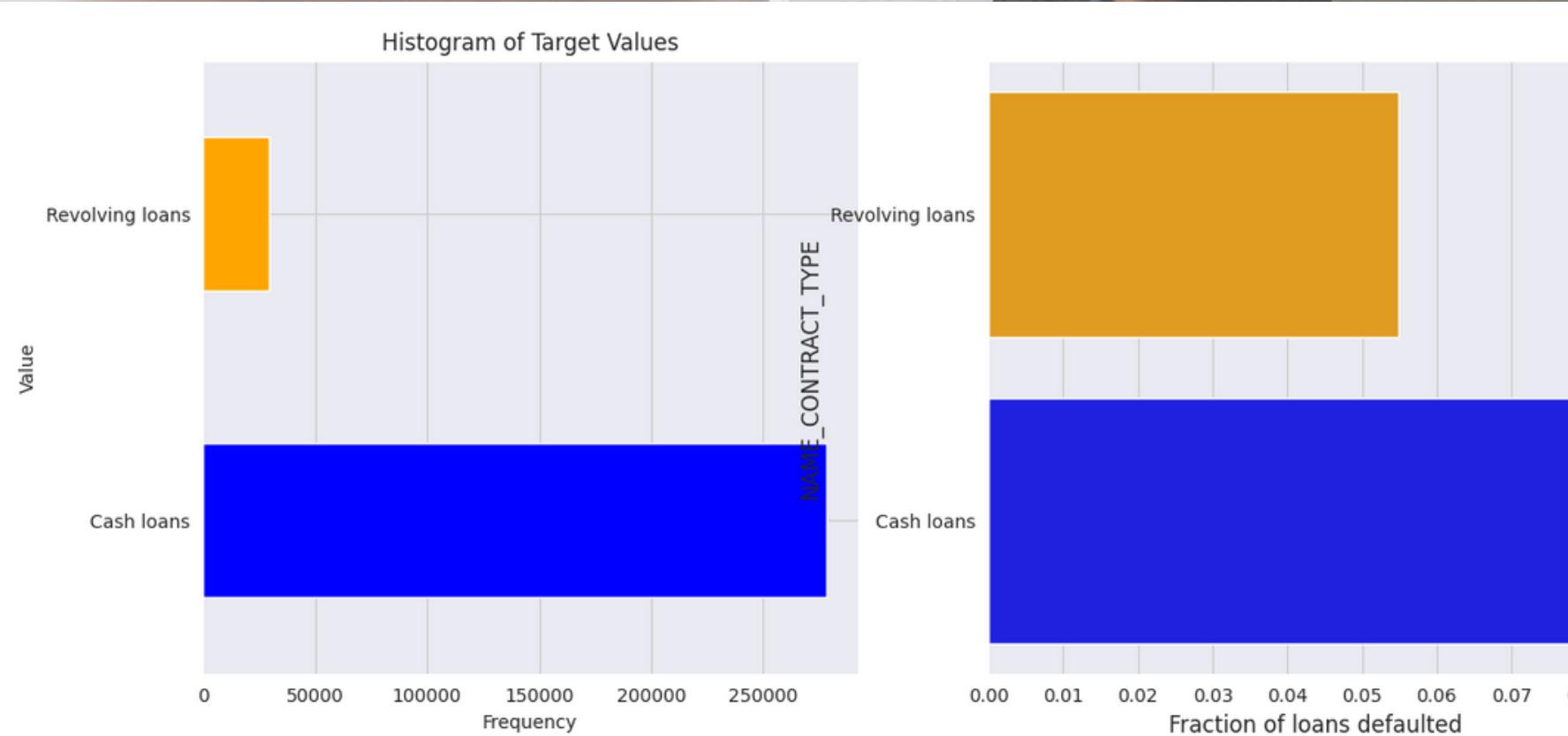
DATA INFO:

Setiap data tabel memiliki sifat yang unik atau berbeda dengan tabel lain, jumlah feature dan data rows tiap tabel juga berbeda-beda.

Tabel	Feature	Record
Application Train	122	30751
Application Test	121	48744
Bureau	17	1716428
Bureau Balance	3	27299925
Credit Card Balance	23	3840312
Installments Payments	8	13605401
POS Cash Balance	8	10001358
Previous Application	37	1670214

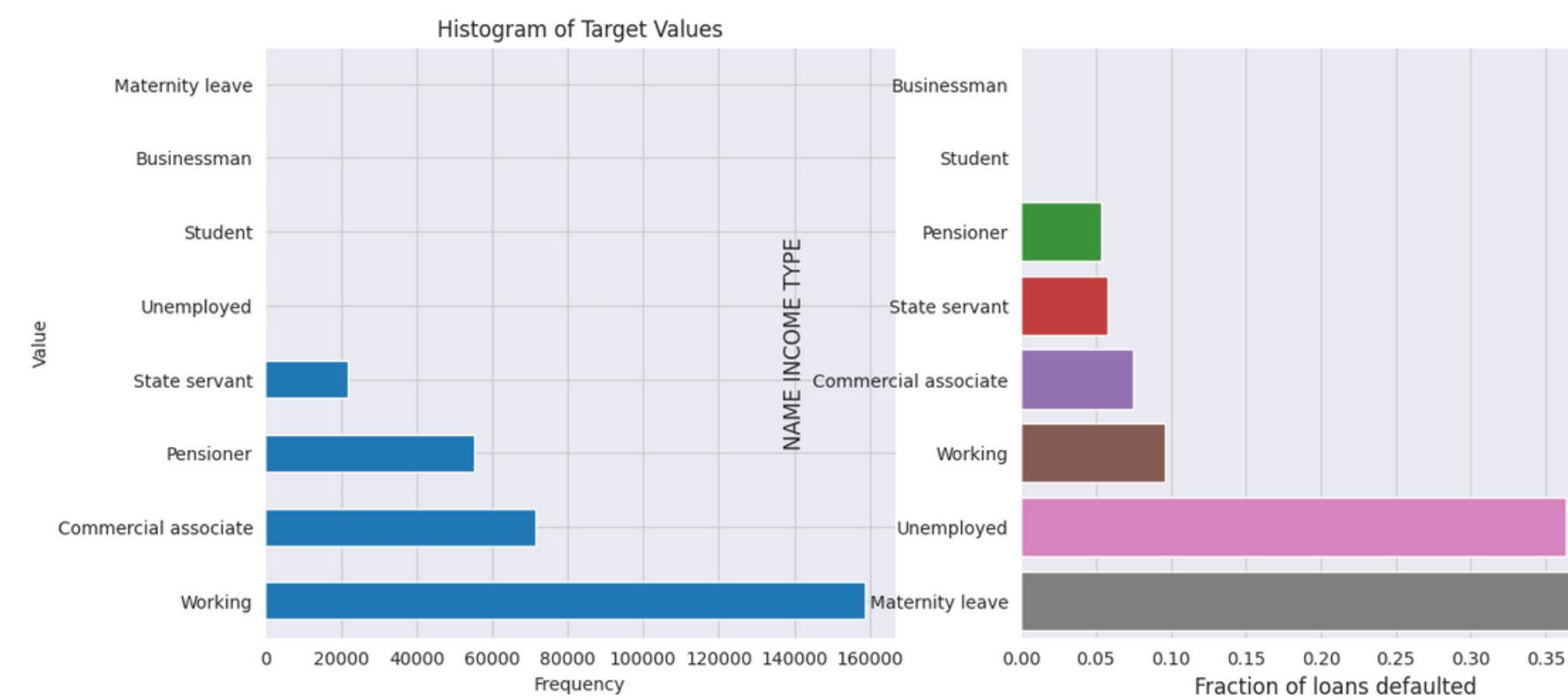
Beberapa dataset yang akan dipakai diantaranya adalah Application train

DATA INSIGHT



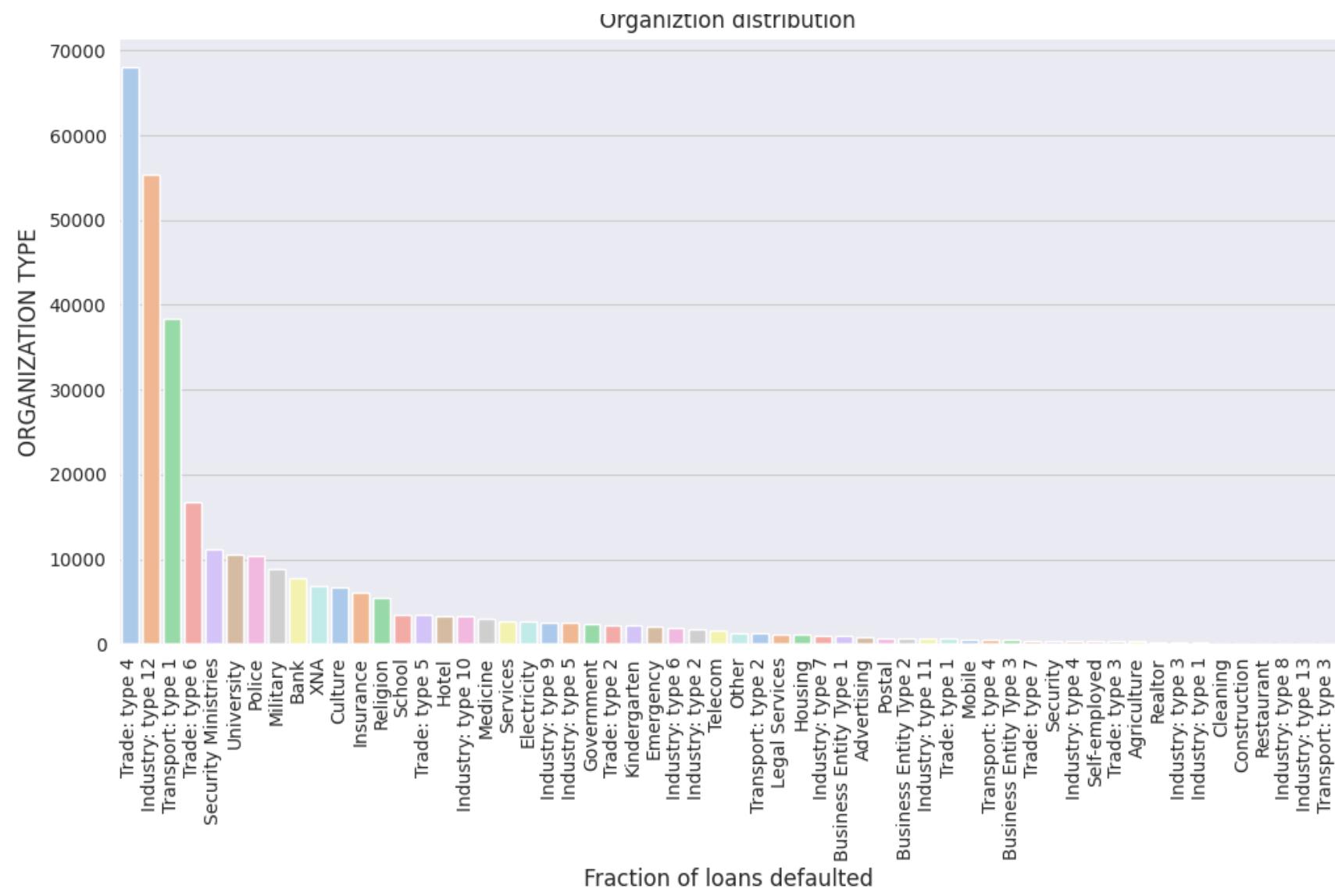
Perusahaan cenderung lebih banyak memberikan pinjaman secara cash dibanding revolving loans (bar kiri), akan tetapi tingkat gagal bayar pinjaman secara cash lebih tinggi (bar kanan)

Untuk mengurangi tingkat resiko gagal bayar pinjaman perusahaan dapat meningkatkan presentase pinjaman secara bergulir (revolving loans)

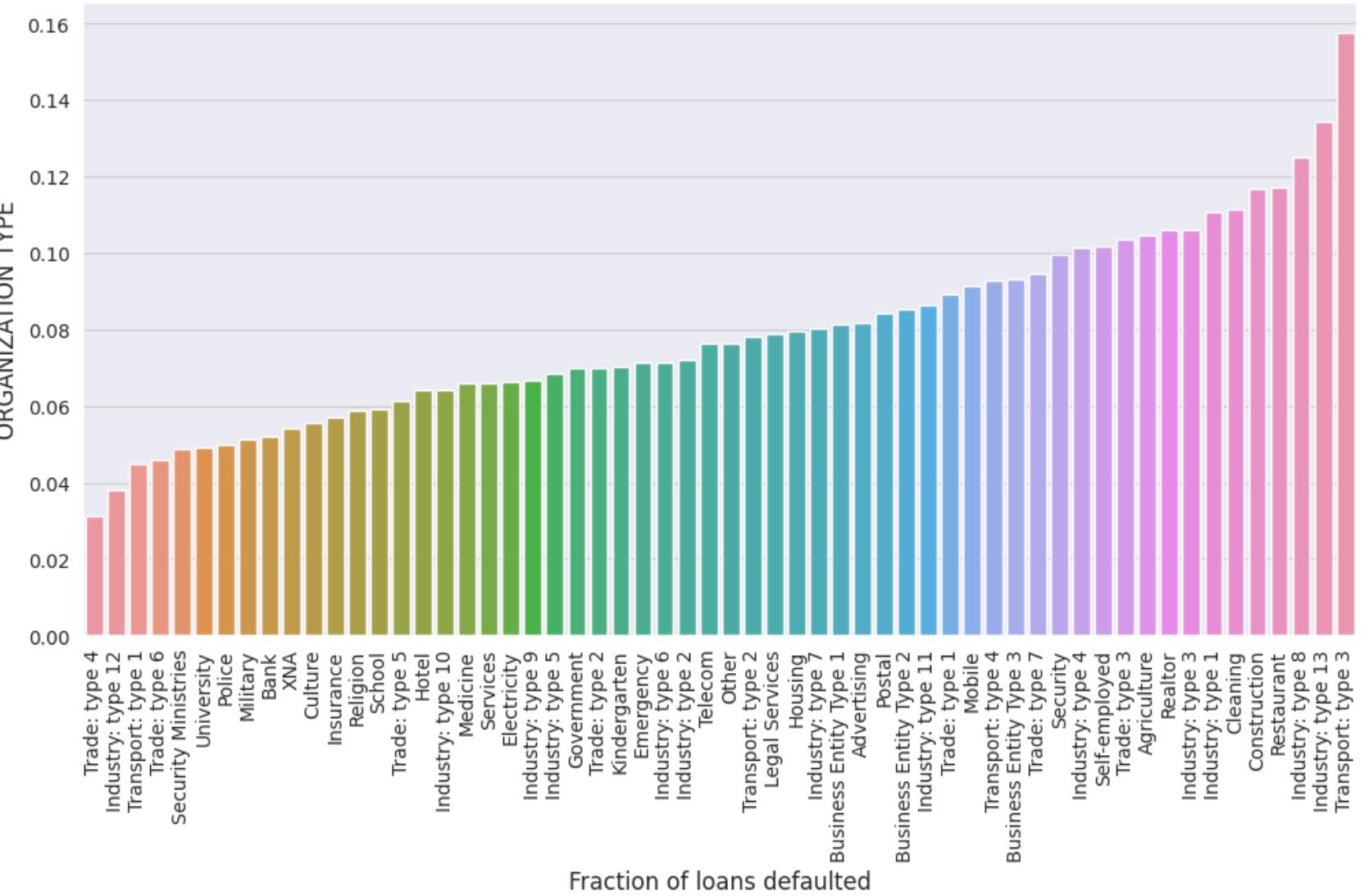


Ternyata pinjaman kredit didominasi oleh orang yang bekerja dan sangat jauh jika dibandingkan dengan orang yang tidak bekerja, Tingkat gagal bayar oleh orang yang bekerja ini pun tidak terlalu tinggi jika dibandingkan dengan tipe income lain. Perusahaan dapat meningkatkan promosi yang menargetkan pada orang-orang yang sedang bekerja, untuk meningkatkan jumlah pinjaman yang dilayani dengan resiko yang rendah.

DATA INSIGHT



Dari grafik diatas peminjam didominasi oleh pedagang type 4 yang jumlahnya lebih dari 60.000 orang, ini menjadi segmen yang menjanjikan bagi perusahaan untuk mentargetkan produk kepada pedagang-pedagang



Selain itu ternyata pedagang type 4 memiliki tingkat gagal bayar yang paling rendah dibandingkan tipe pekerjaan lain. Hal ini harus mampu dimanfaatkan oleh perusahaan dengan menciptakan strategi promosi untuk pedagang type 4 ini. Strategi nya bisa melalui penawaran khusus dengan bunga rendah dan ditawarkan pada asosiasi pedagang.



MODEL BUILDING STEPS

1

2

3

4

DATA EXPLORATION & PREPARATION

Tahap ini meliputi visualisasi dan insight data, data info, korelasi, Handling missing values, Handling outliers, One hot encoding, Rescaling data.

ALGORITHM SELECTION

Tahap ini meliputi proses feature selection, adaptive sampling (data split & imbalance corection), algorithm selection

MODEL EVALUATION

Tahap ini mengujikan model melalui score Accuracy, Recall, dan score ROC AUC.

HYPERPARAMETER TUNING

Tahap ini untuk mendapatkan parameter yang paling berpengaruh terhadap model



ALGORITHM SELECTION

HANDLING MISSING VALUES

Setelah melakukan data exploration seperti di slide sebelumnya, kemudian adalah mendapatkan informasi jumlah data rows yang missing pada table

Untuk data numerik yang kosong akan disikan nilai rata-rata pada kolom tersebut, karena data kategorikal sudah lengkap maka tidak diperlukan custom imputer. Selanjutnya data akan di drop untuk kolom yang tidak relevan dan memiliki missing values lebih dari 20%.

HANDLING OUTLIER

Dikarenakan beberapa data memiliki outlier yang tinggi seperti kolom pendapatan pada tabel application train maka perlu dilakukan suatu tindakan, karena outlier ini akan mempengaruhi hasil model nantinya

Handling outlier menggunakan metode winsorizing, dengan memotong data outlier pada bagian atas sebanyak 5% dan 20%, selain itu data juga dipotong pada batas bawah outlier sebanyak 5% dan 10%. Nilai yang lebih dari batas atas akan diganti menggunakan batas atas dan begitu juga untuk nilai bawah

ONE HOT ENCODING

One hot encoding digunakan untuk mengubah tipe data kategorikal menjadi bilangan biner 1 dan 0, sehingga jumlah feature pada data yang dipilih sebelumnya memiliki kolom tambahan

RESCALING DATA

Data kemudian akan di rescale, untuk mengatasi perbedaan skala di tiap fiturnya. Sehingga dengan teknik normalisasi min max scaling ini akan menghasilkan nilai berskala dari 0 sampai dengan 1.

Rescaling data ini berguna agar meningkatkan kinerja model dengan menghindari masalah skala yang tidak seimbang untuk setiap featurenya. Hal ini juga berguna untuk menghindari overfitting pada model

DATA EXPLORATION AND PREPARATION

FEATURE SELECTION

Feature yang dipilih adalah berdasarkan hasil dari tahap rescaling data dengan jumlah feature yaitu 142. Korelasi dari tiap feature ditunjukkan dengan warna heatmap yang ada, semakin berwarna gelap maka semakin kuat korelasinya.

ADAPTIVE SAMPLING

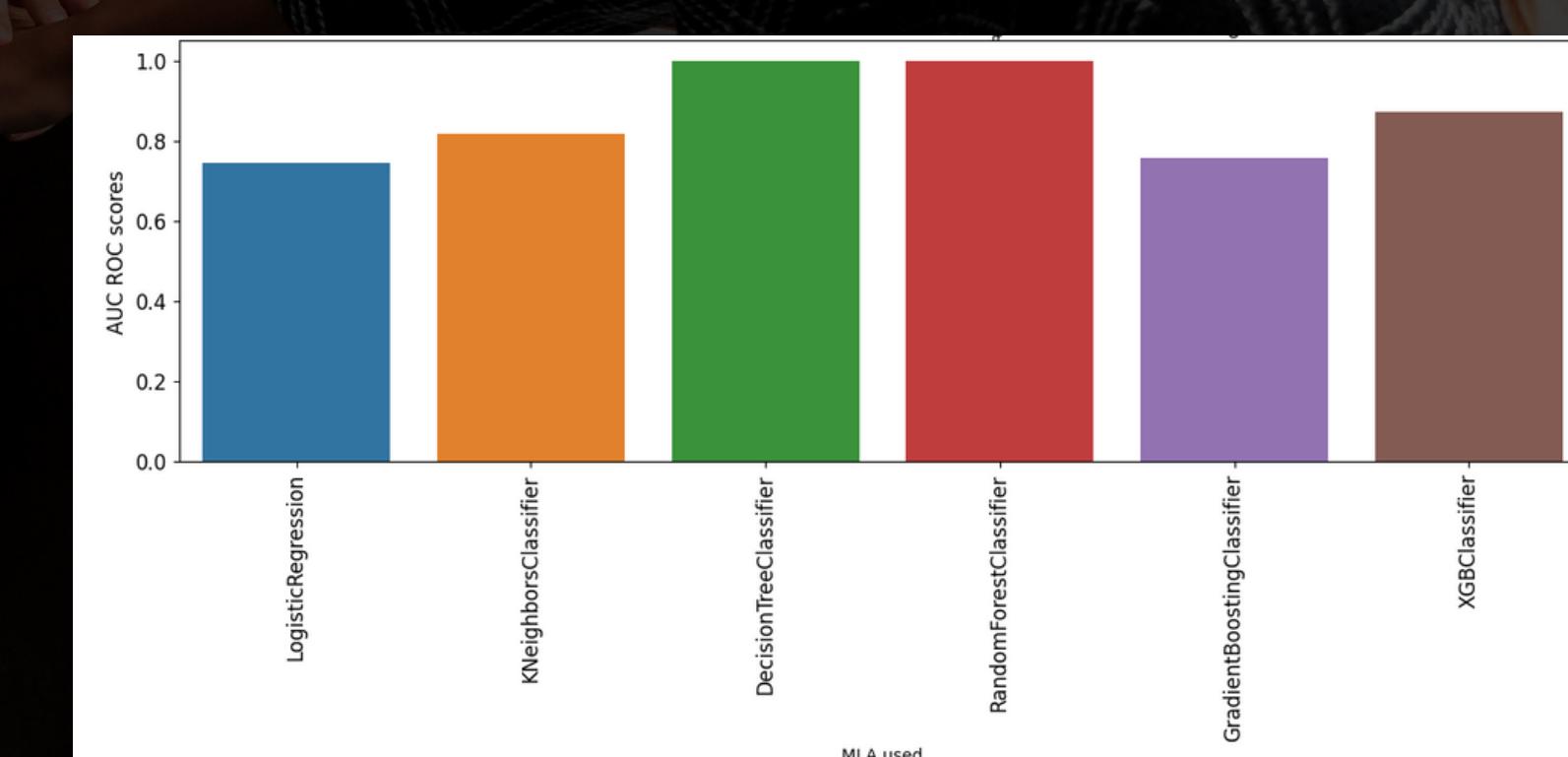
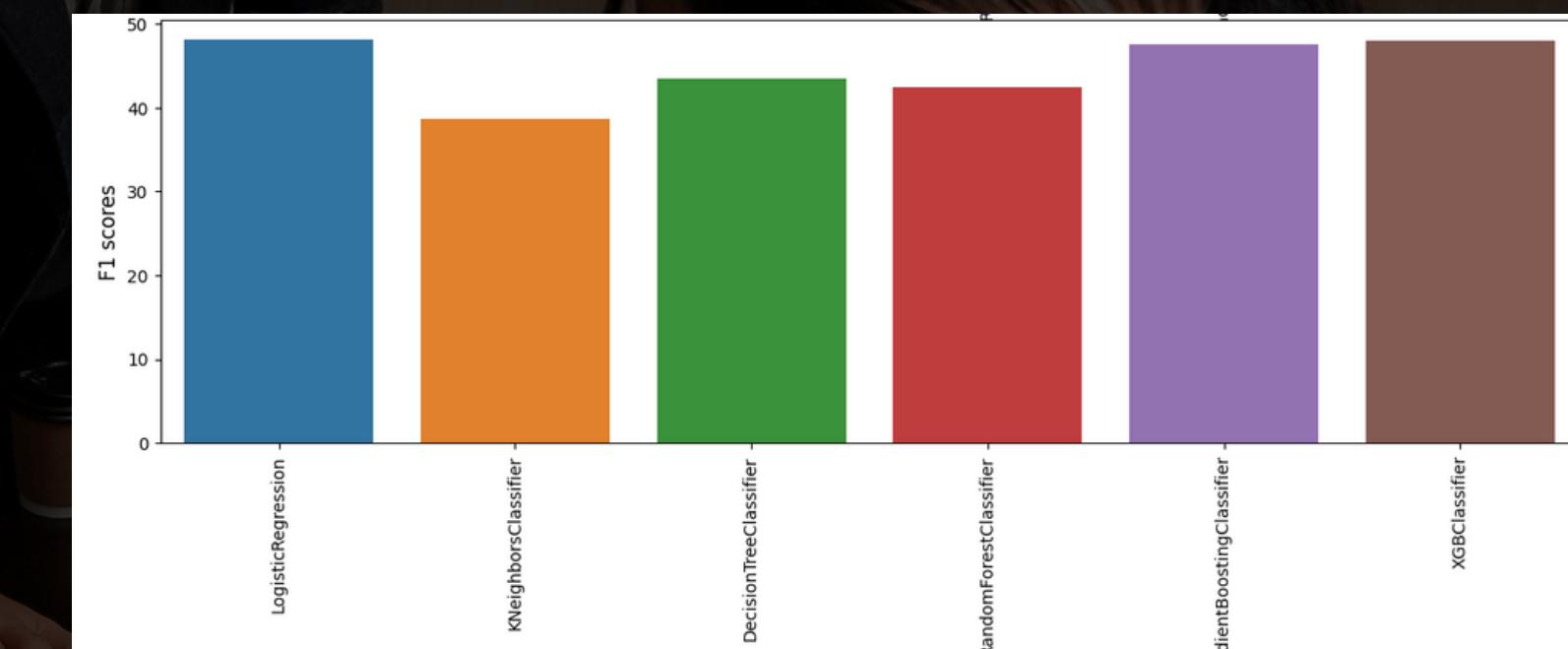
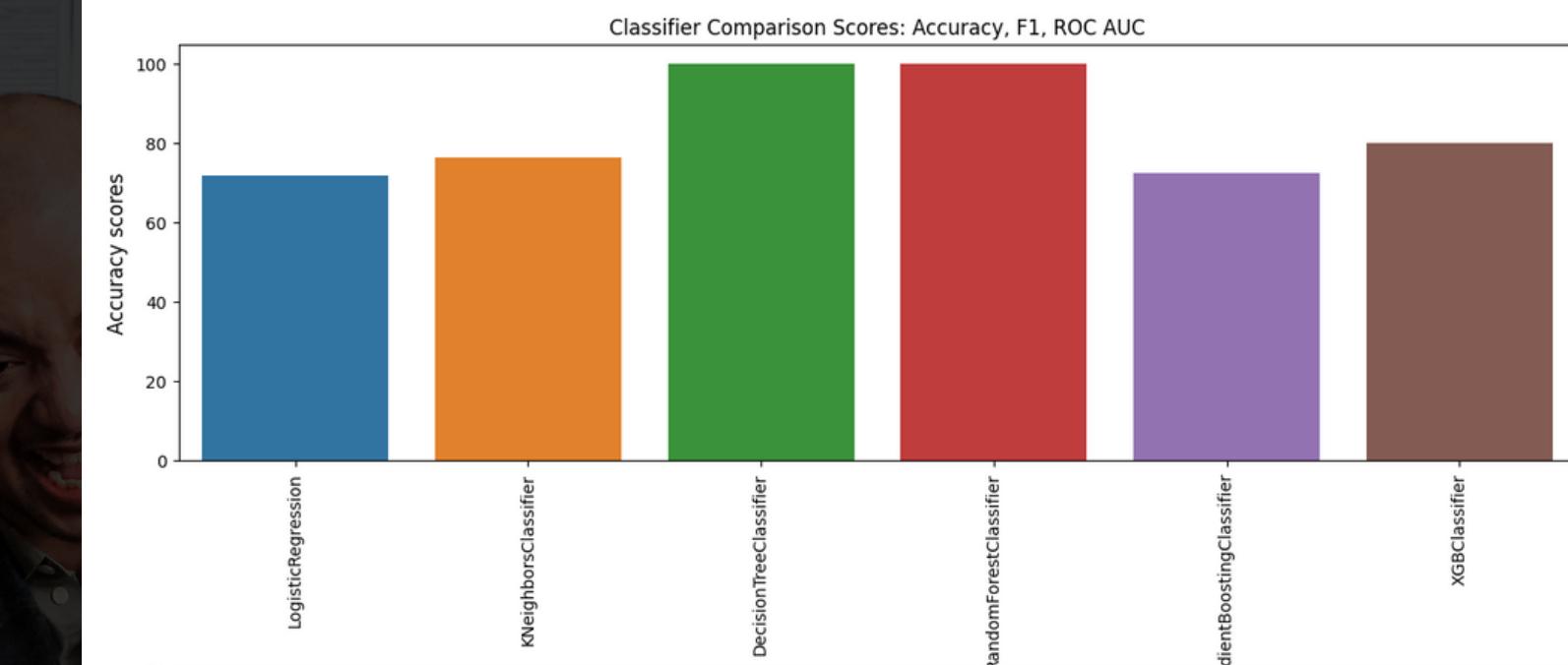
Data kemudian akan dipisahkan menjadi data train dan data test, untuk mengkoreksi dari data imbalance menggunakan metode undersampled karena data train mencapai 230633 sedangkan data test hanya 76878, sehingga data dihasilkan lebih seimbang yaitu data train 1

ALGORITHM SELECTION

Percobaan model menggunakan 6 algoritma yaitu logistic regression, K Neighbour dan decision tree classifier, Random forest, Gradient Boosting, dan XGB classifier

Hasil yang akan dibandingkan adalah precision sebagai pertimbangan utama, Recall, dan ROC AUC score perbandingan dapat dilihat pada histogram disamping

Terdapat 2 model dengan accuracy mencapai 100% yaitu decision tree dan random forest classifier, algoritma tersebut bisa mengakibatkan model menjadi over-fit. Oleh karena itu akan dipilih model dengan tingkat accuracy tinggi, serta perbandingan ROC -AUC train dan test yang tidak terlalu jauh maka algoritma yang akan digunakan adalah XGB classifier



MODEL EVALUATION

HYPERPARAMETER TUNING



CONFUSION MATRIX

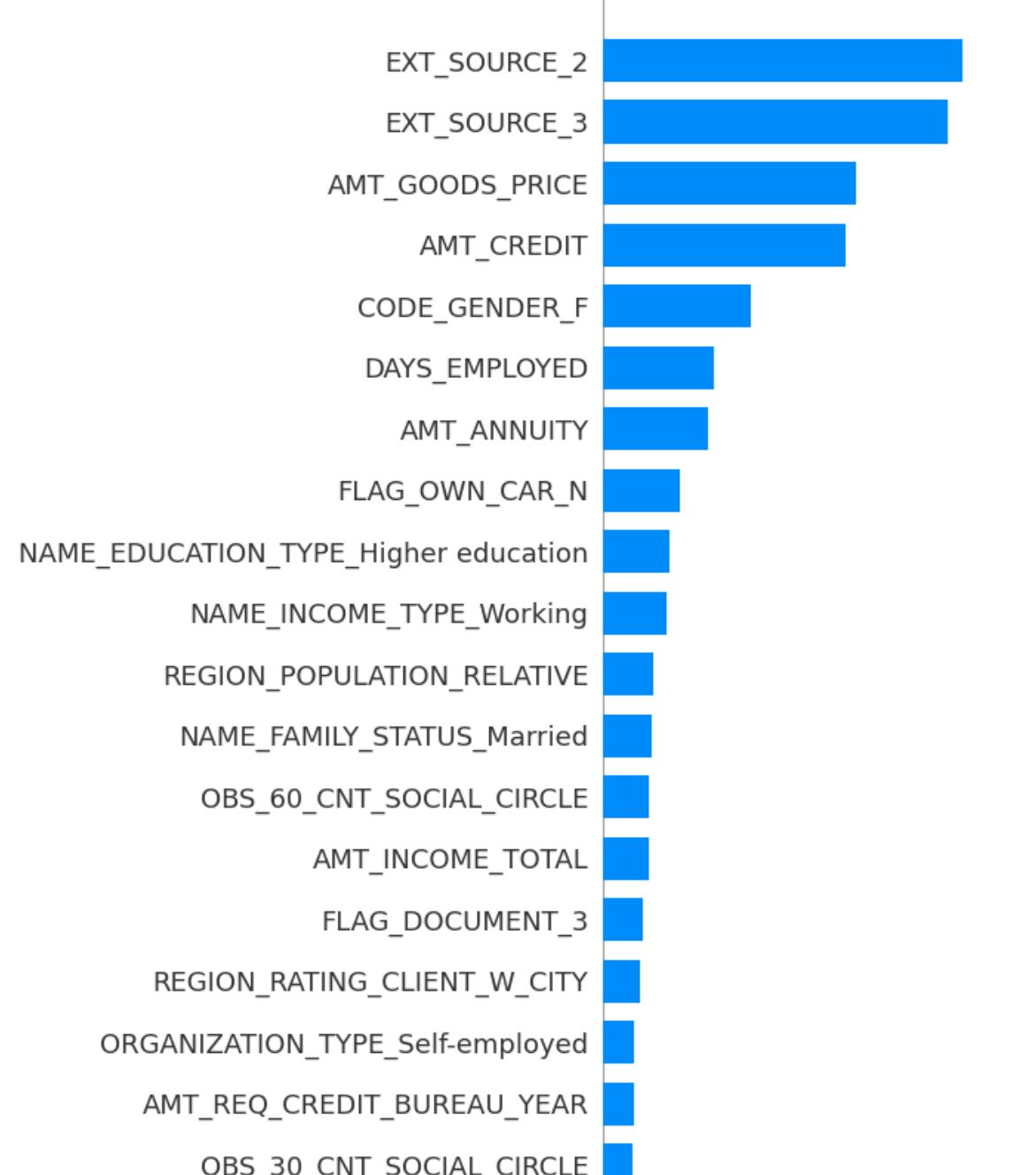
Dengan menggunakan model XGB classifier didapatkan confusion matris seperti berikut :

```
**CONFUSION MATRIX**  
[[ 7978 1438]  
 [2482 2019]]
```

Sehingga Accuracy, Recall, dan ROC AUC dapat dihasilkan sebagai berikut

```
AUC Score Train vs Test:  
AUC Score Train proba: 0.763  
AUC Score Test proba: 0.739
```

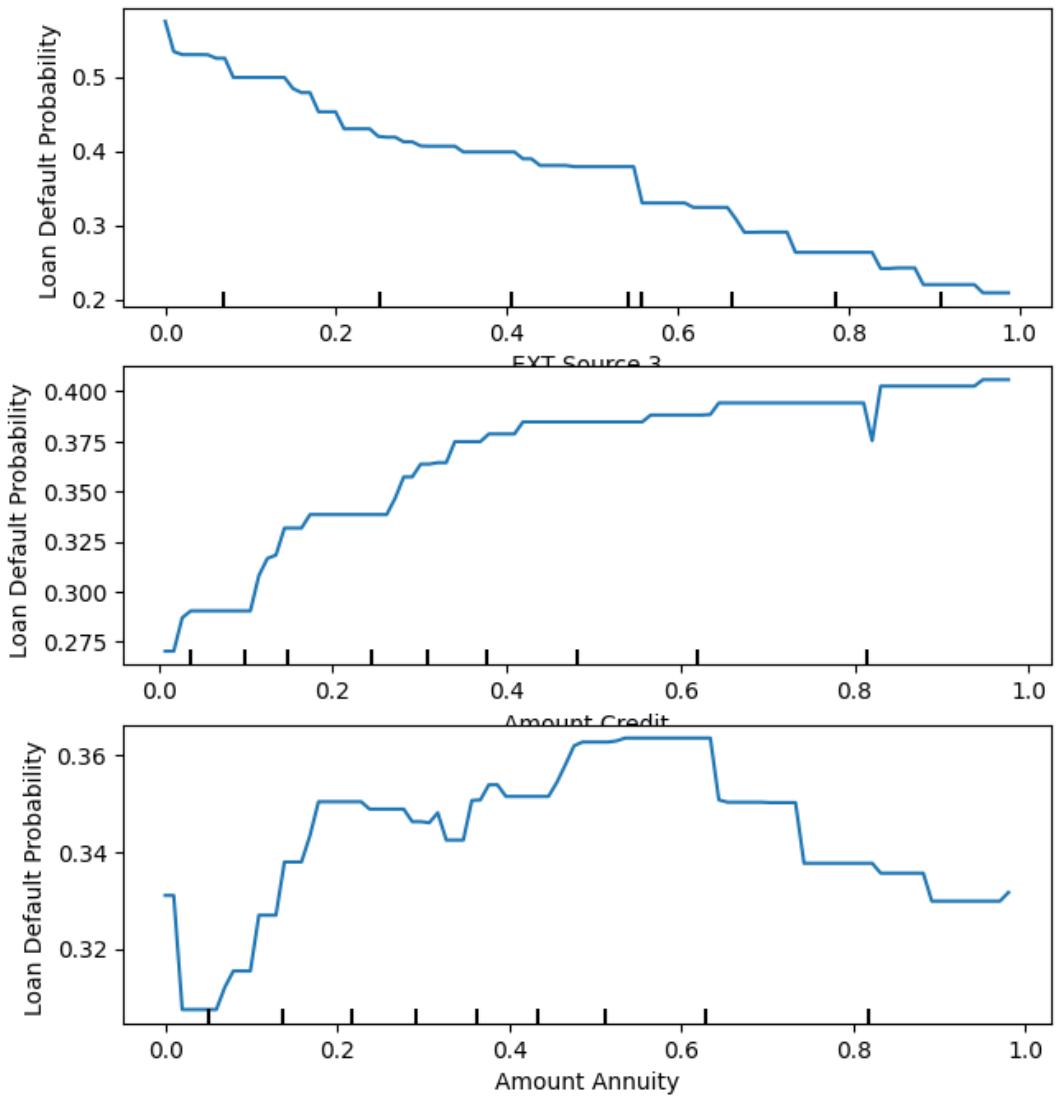
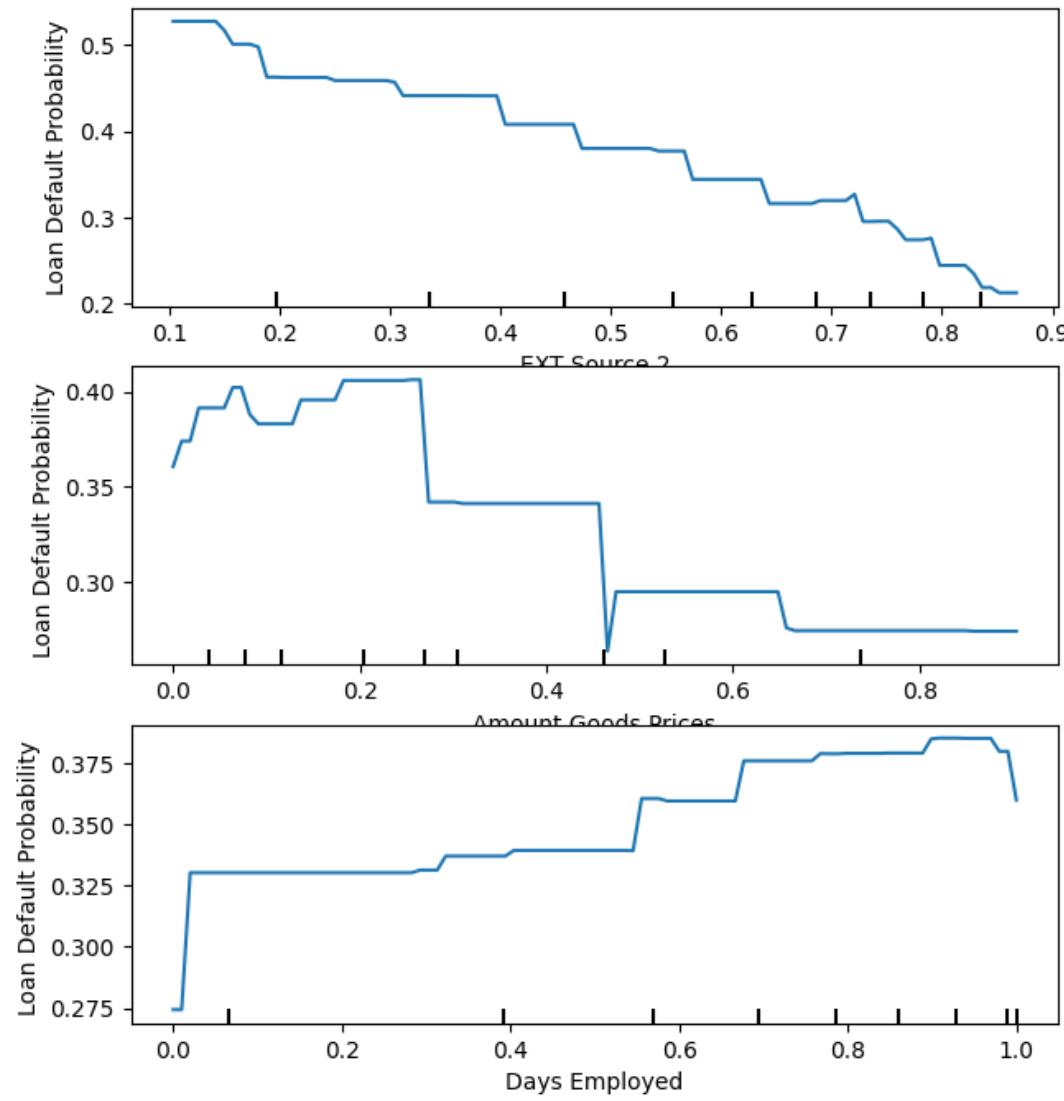
```
=====  
Others Metrics Evaluation:  
Train Accuracy Score : 0.729  
Test Accuracy Score : 0.718  
Precision Score Test: 0.584  
Recall Score Test : 0.449  
F1 Score Test : 0.507
```



Dengan menggunakan modul Shap dapat diketahui bahwa nilai paling tinggi atau yang memiliki pengaruh kuat pada model diantaranya adalah EXT_Source_2 dan EXT_SOURCE_3, kemudian parameter tersebut akan digunakan untuk melatih model

BUSINESS RECOMENDATION

Hyperparameter compare with loan default probability



Business recomendation

1. Perusahaan mempertimbangkan penuh pada skor kredit borrower sebelum memberikan kredit pinjaman
2. Borrower yang telah lama bekerja lebih banyak di setujui pinjamannya, disisi lain ternyata memiliki resiko gagal bayar yang tinggi. Hal tersebut bisa disebabkan oleh berbagai faktor seperti jumlah gaji yang sebenarnya tidak mencukupi
3. Income borrower yang tinggi tidak sepenuhnya memiliki kemampuan gagal bayar yang rendah, perusahaan perlu meninjau aspek lain seperti amount credit yang diajukan.



cityanseyawan27@gmail.com

github repo : <https://github.com/Seth1495/Final-Project-Home-Credit-seth.git>

Github : Seth1495