

# HOME CREDIT

SCORECARD

FAHRI PUTRA HERLAMBANG

<https://www.linkedin.com/in/fahriputra/>

<https://github.com/FahriPutra00/Home-Credit-DataScience>



# Background Scorecard-Prediction

Home Credit menggunakan metode statistik dan Machine Learning untuk prediksi skor kredit. Tujuan adalah mencegah penolakan pengajuan pinjaman yang seharusnya disetujui.

## Objective :

- Meningkatkan seleksi pengajuan pinjaman klien yang sesuai
- Mengetahui resiko klien mengalami gagal bayar

## Target variable :

- 1 = Klien memiliki masalah pembayaran (Terlambat)
- 0 = Kasus Lain (Pembayaran Sesuai)

Metrik : Akurasi, F1-Score, Presisi, Recall, ROC\_AUC

# DATASET

## Application\_Train.CSV

- Columns : 122
- Rows : 307511
- Attribute : 122 Desc
  - Numerical : 106
  - Categorical : 16

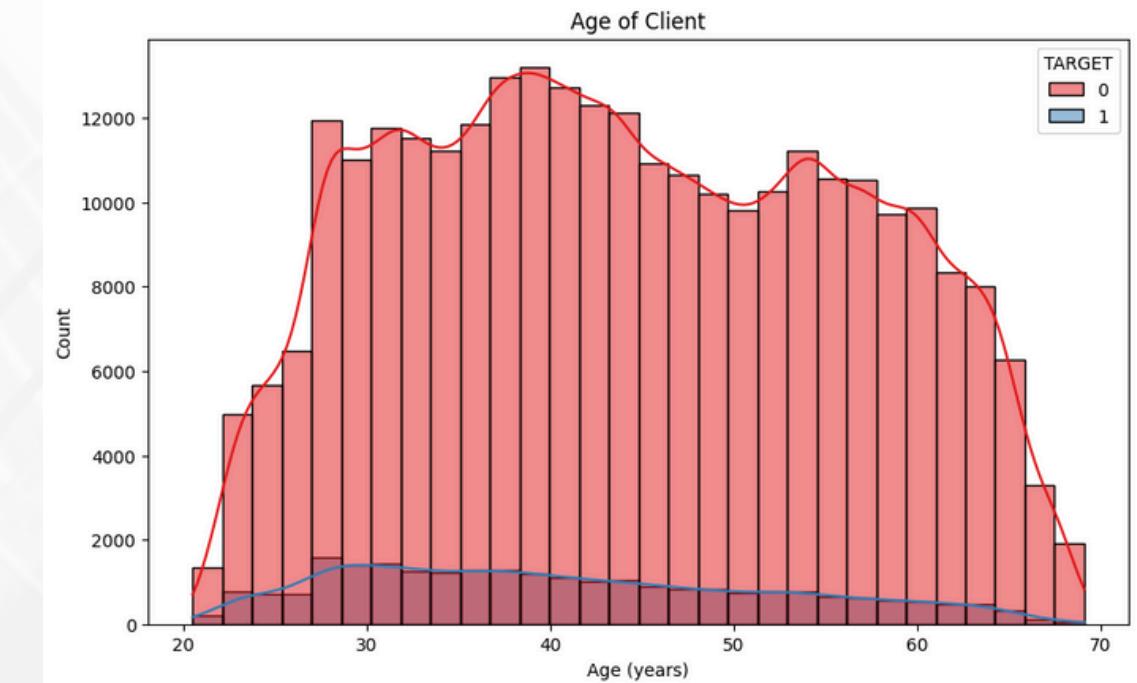
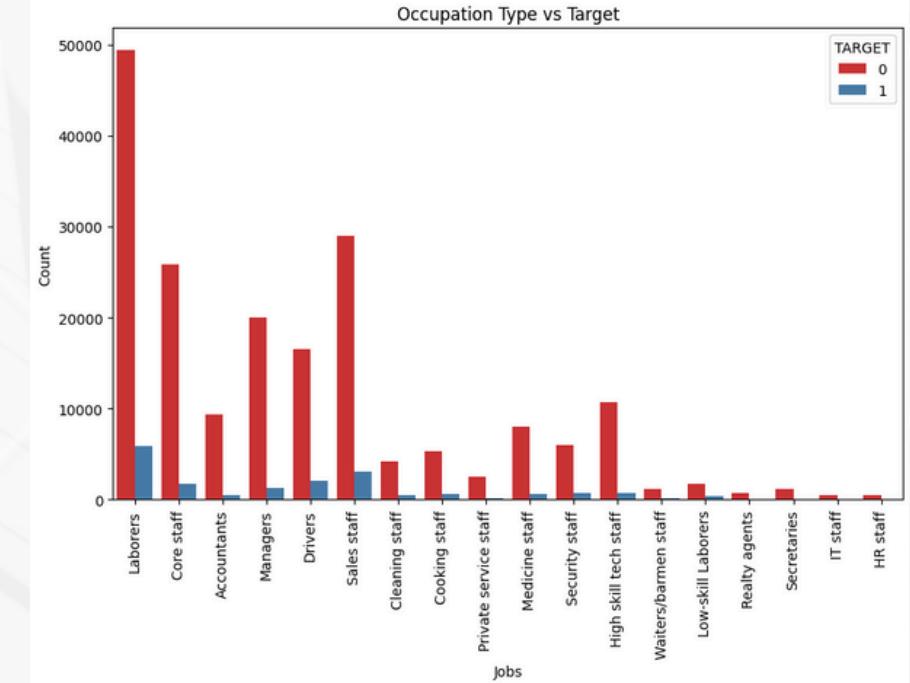
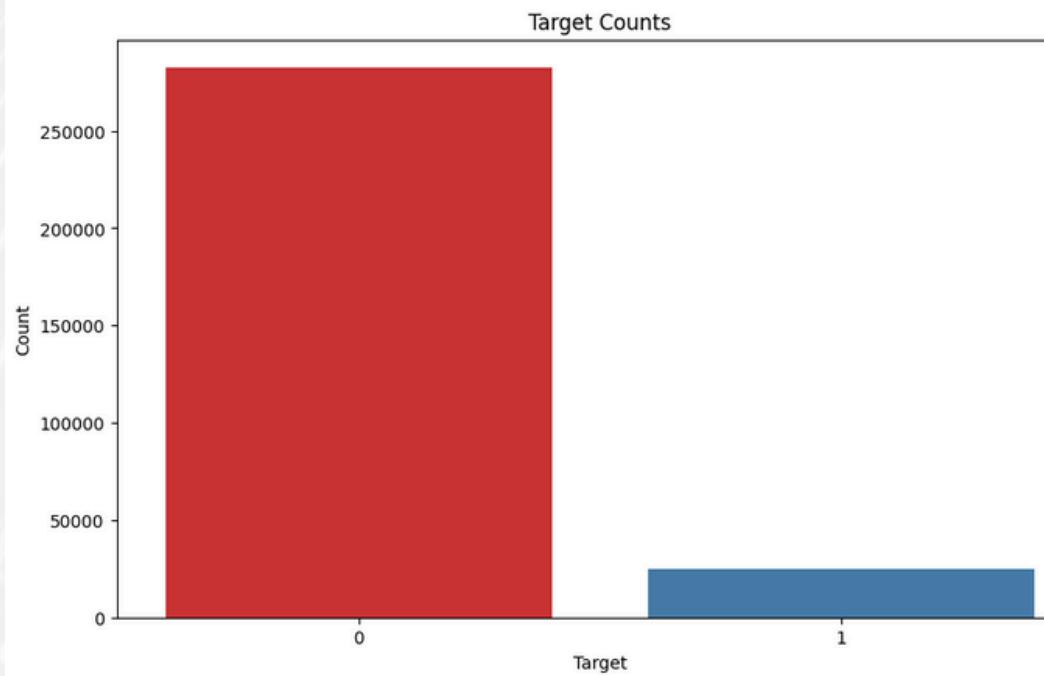
## Attribute Description

- Num. Missing Value : 9152465
- Num. Duplicate : 0
- Outlier Columns : 2
- Imbalanced : True

## Application\_Test.CSV

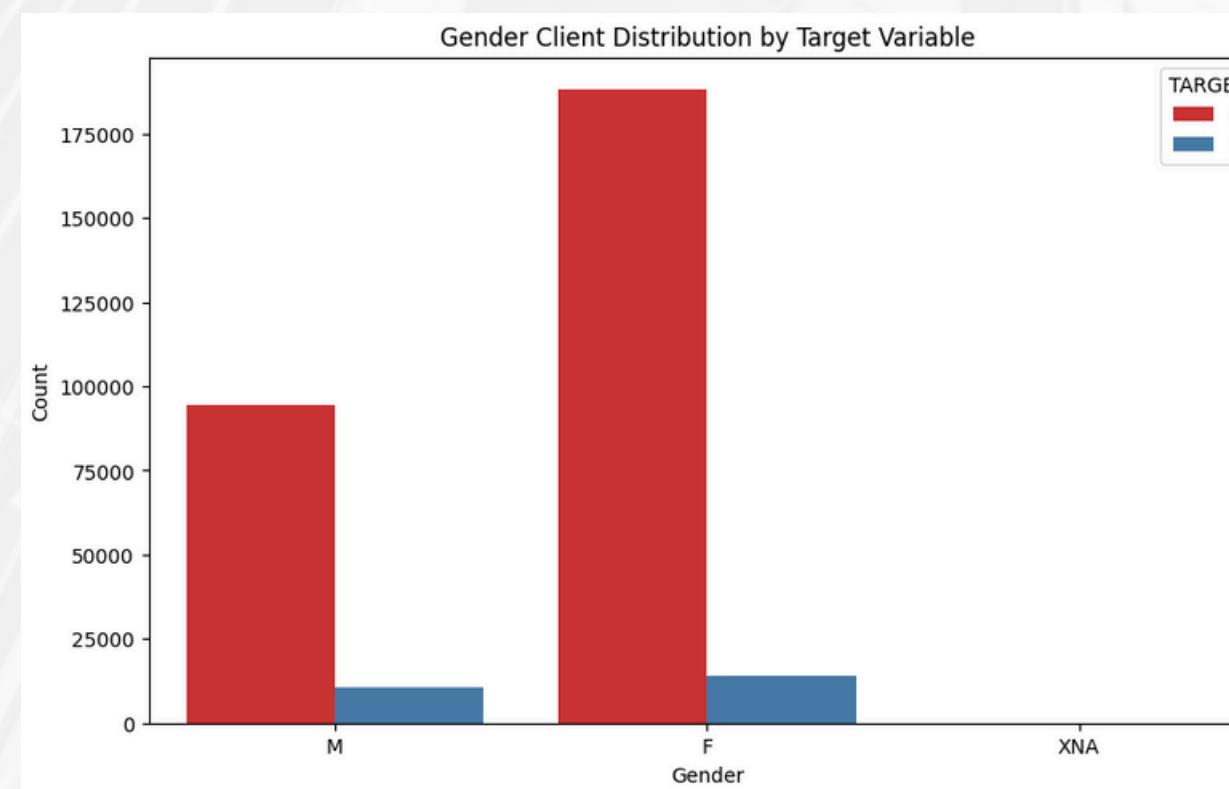
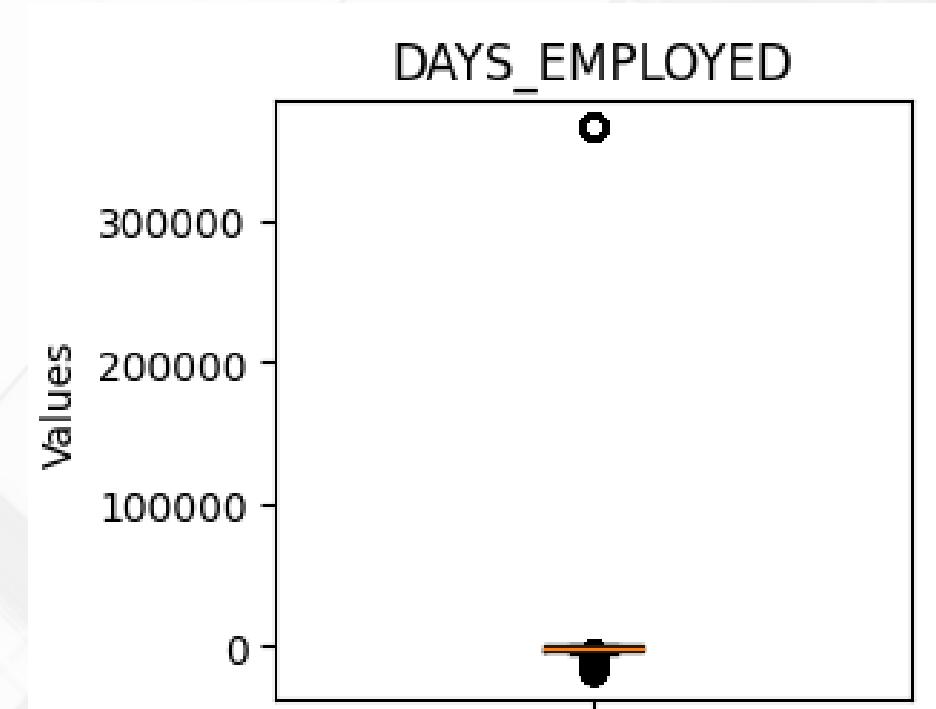
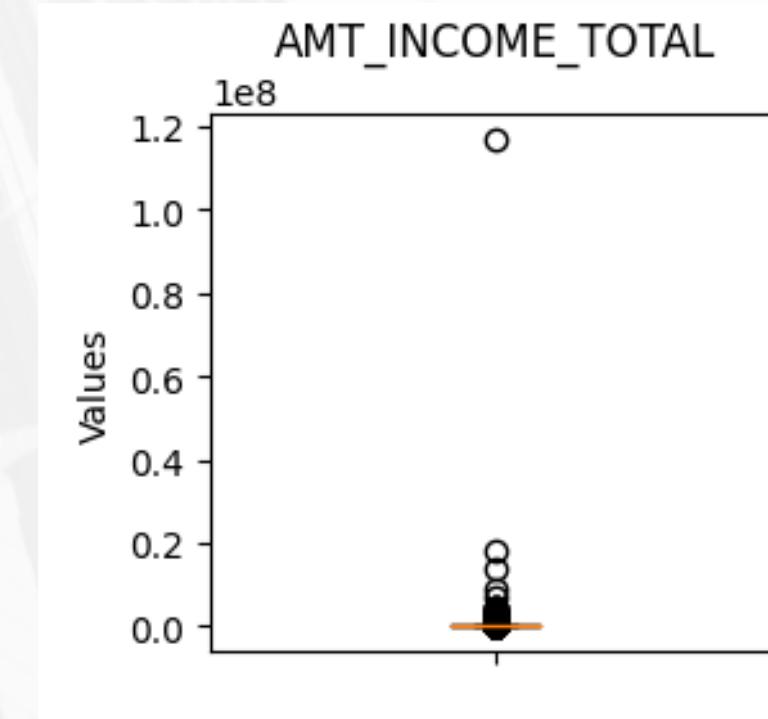
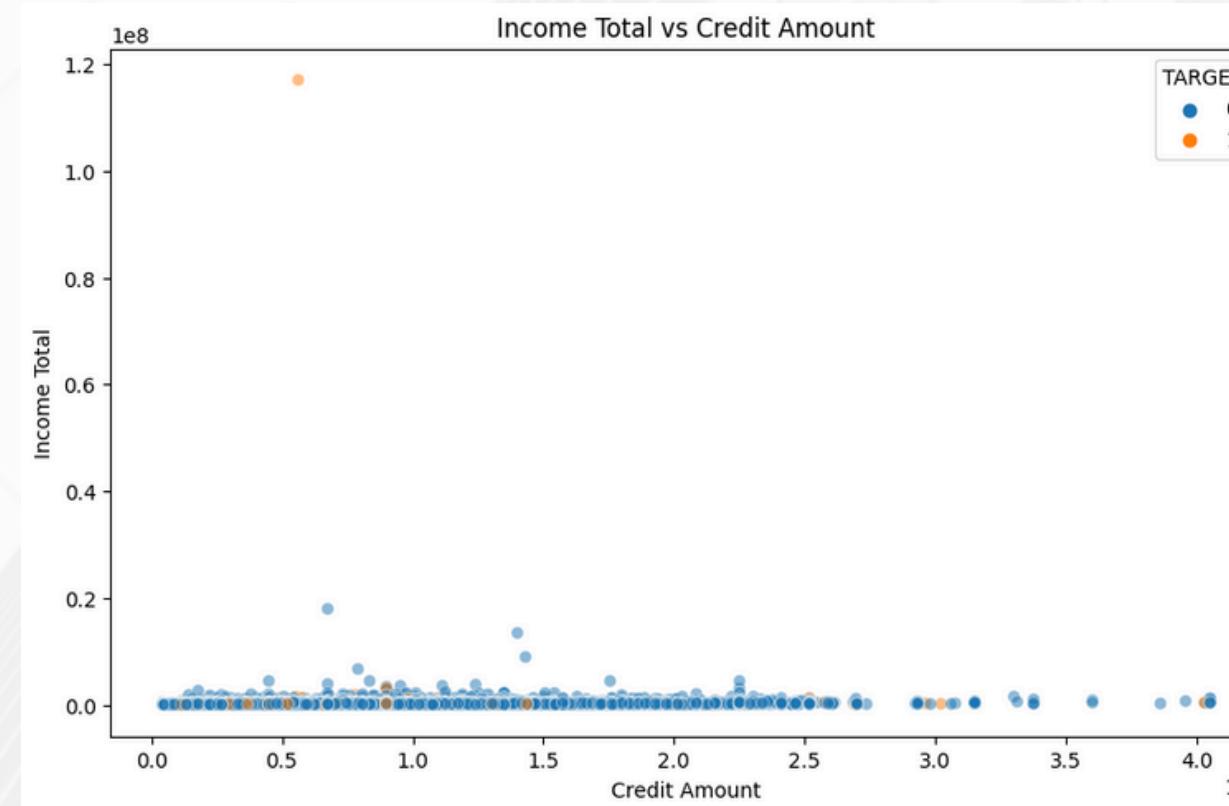
- Columns : 121
- Rows : 48744
- Attribute : 121 Desc
  - Numerical : 105
  - Categorical : 16

## DATA TRAIN VISUALIZATION



Dari visualisasi data tersebut, dapat terlihat data latih memiliki target yang *imbalance* atau tidak sama jumlahnya. kemudian persebaran peminjam terbanyak dari pekerja keras atau buruh, yang memiliki tingkat gagal bayar paling tinggi. Dan yang terakhir persebaran umur peminjam yang memiliki rekam gagal bayar tertinggi adalah berumur akhir 20-an hingga awal 30-an.

# More Visualization



Dari visualisasi data tersebut, dapat terlihat data latih memiliki data outliers atau pencilan yang berlebih pada dua kolom yaitu kolom Pendapatan dan Lama Bekerja, sehingga membutuhkan metode IQR untuk menyelesaikan permasalahan ini. Kemudian pada data Gender yang seharusnya hanya terdapat data Male & Female, terdapat data XNA yang bukan merupakan bagian dari Gender.

# Data Pre-Processing

## Cleansing Data

- Mengubah nilai NaN dengan **Modus** untuk data **Categorical**.
- Mengubah nilai NaN dengan **Median** untuk data **Numerical**.
- Menghapus dari kolom '**CODE\_GENDER**' baris yang bernilai selain 'M' & 'F'.

## Feature Selection

- Melakukan **Feature Selection** pada dataset Train menggunakan **SelectKBest** dengan **score\_func=f\_classif**.
- Jumlah kolom yang diseleksi berjumlah 60 kolom.

## Data Transformation

- Melakukan **Normalisasi** kolom **Categorical** menjadi **Numerical [0,1]**.
- Melakukan **Feature Scaling** pada kolom **Numerical** kecuali kolom 'SK\_ID\_CURR' & 'TARGET' menggunakan **StandardScaler**.
- Melakukan **Handling Outliers** pada kolom 'AMT\_INCOME\_TOTAL' & 'DAYS\_EMPLOYED' menggunakan metode IQR

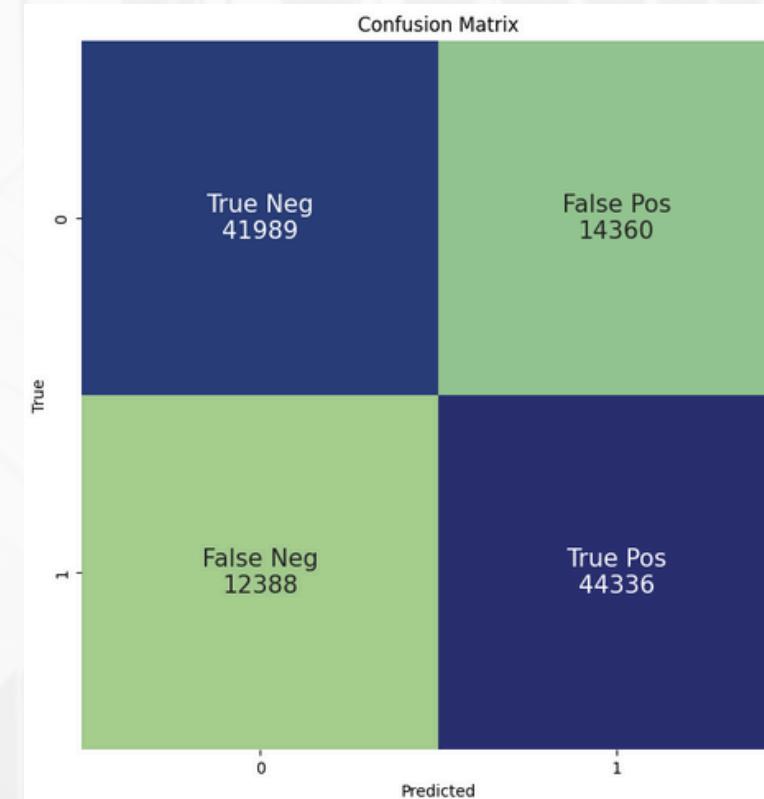
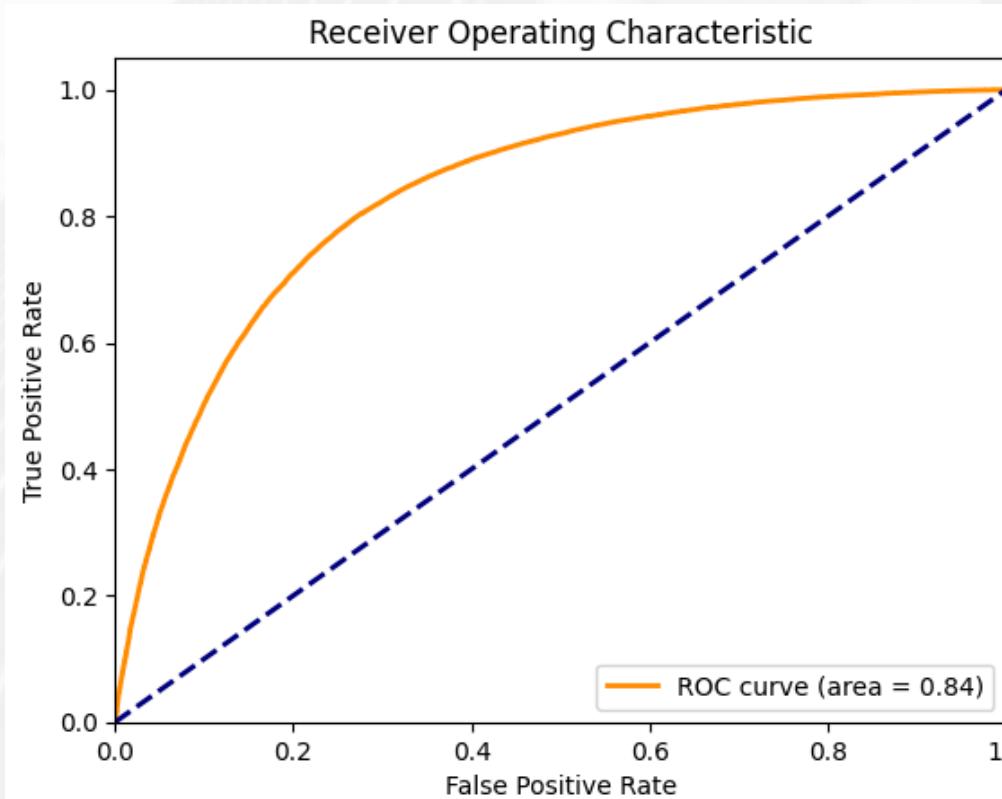
## Handling Imbalance Data

- Melakukan **Oversampling** pada dataset Train dikarenakan data 'TARGET' yang tidak seimbang menggunakan **SMOTE**.

# Logistic Regression

## Data Split (Application\_Train.csv)

- Train : 80%
- Val : 20%
- Random\_State ; 42



## Parameter

- random\_state=42
- verbose=1
- n\_jobs=-1
- max\_iter=200
- C=0.5
- penalty='l1'
- solver='liblinear'
- class\_weight='balanced'

## Result

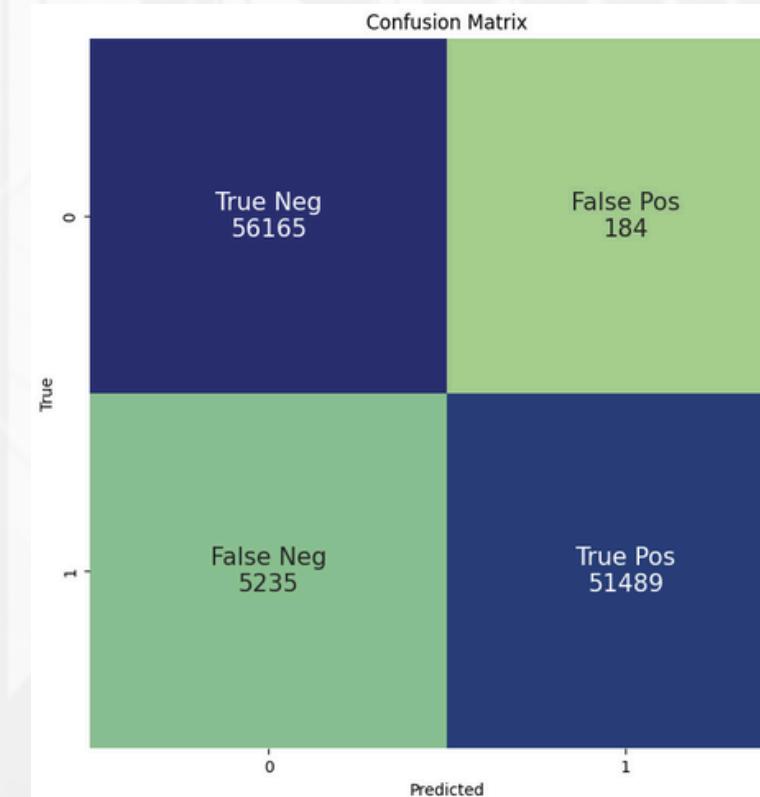
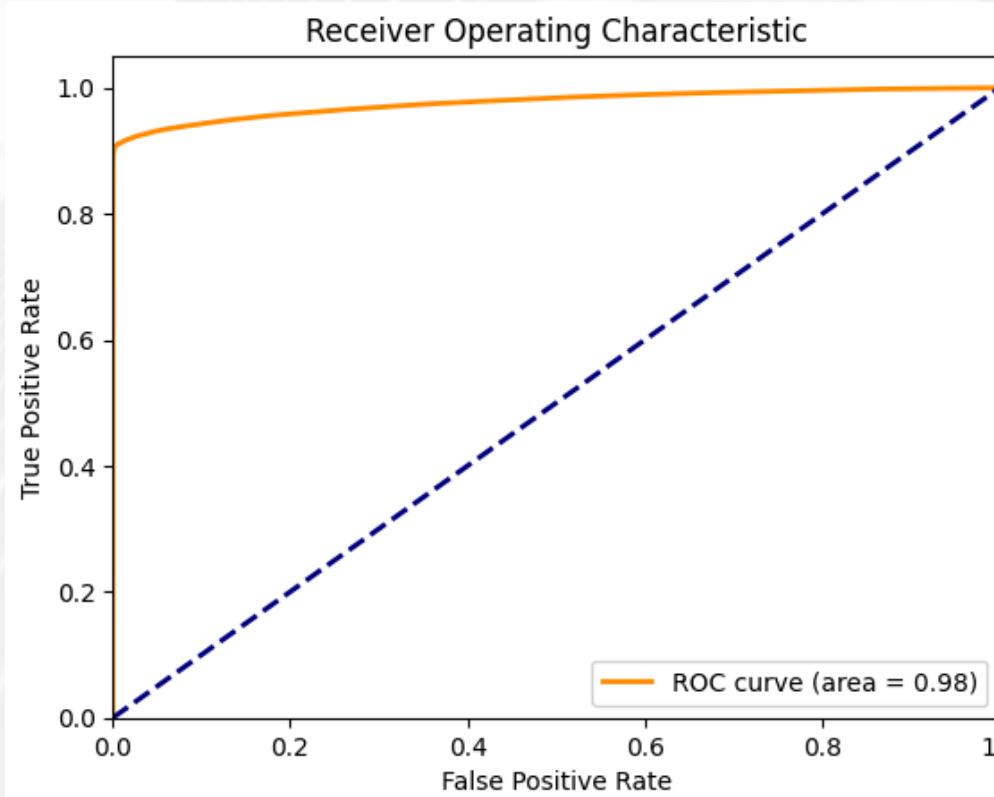
- Accuracy : 0.76
- Precision : 0.76
- Recall : 0.78
- F1- Score : 0.77
- ROC\_AUC : 0.84

Dari model Logistic Regression tersebut memperlihatkan akurasi yang cukup baik namun, belum menjadi parameter yang terbaik untuk memprediksi nilai creditcard dikarenakan nilai akurasi yang hanya 76%.

# Random Forrest

## Data Split (Application\_Train.csv)

- Train : 80%
- Val : 20%
- Random\_State ; 42



## Parameter

- random\_state=42
- verbose=1
- n\_jobs=-1
- n\_estimators=200
- max\_features='log2'
- class\_weight='balanced'

## Result

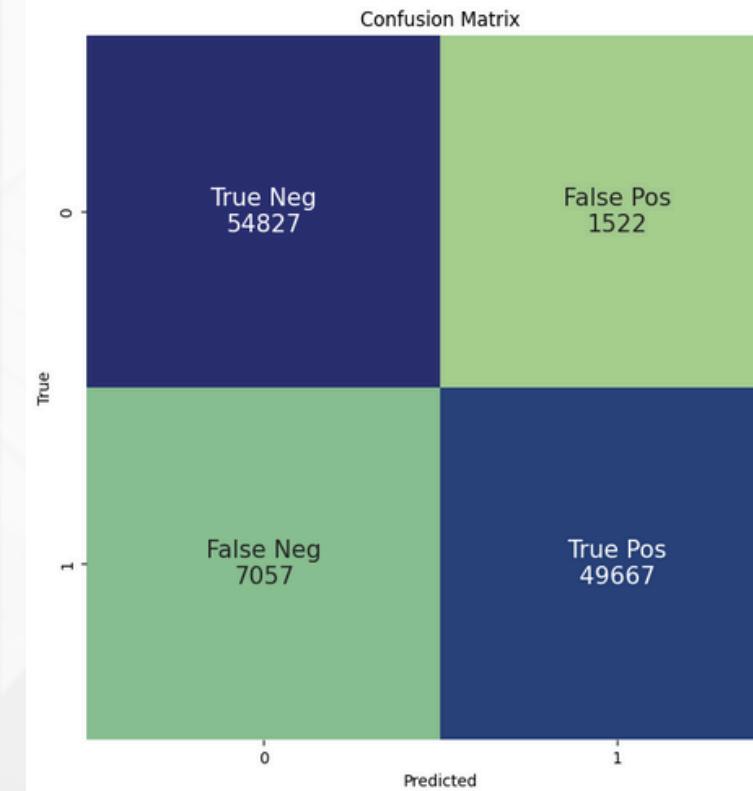
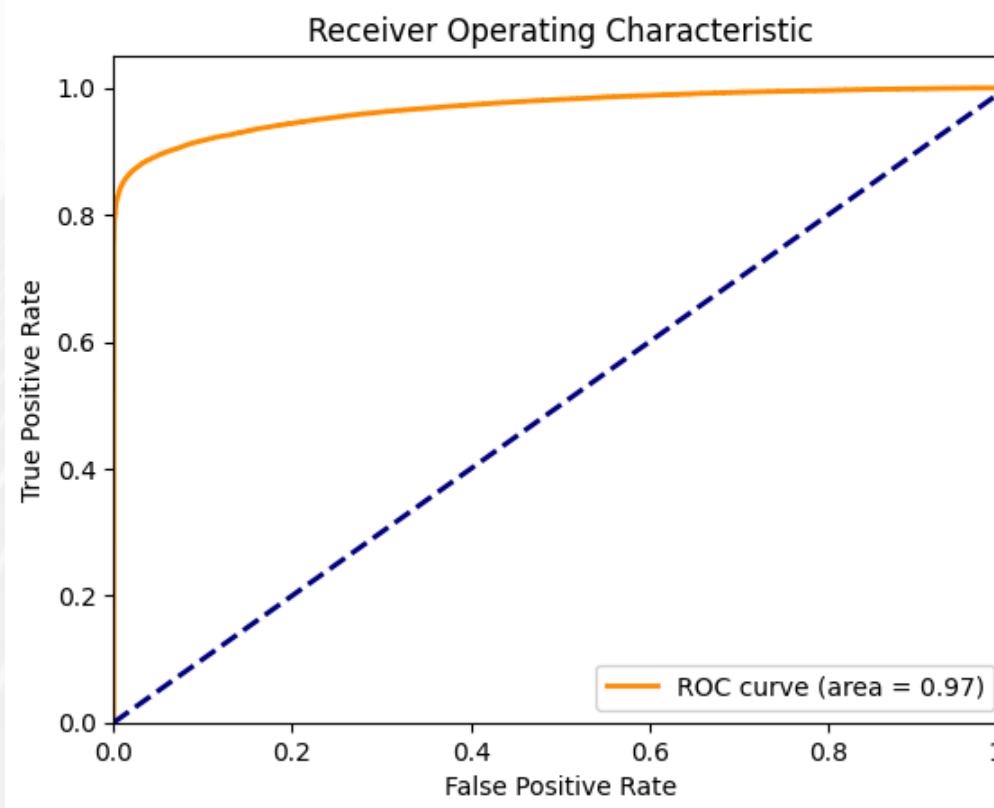
- Accuracy : 0.95
- Precision : 1.00
- Recall : 0.91
- F1- Score : 0.95
- ROC\_AUC : 0.98

Dari model Random Forrest tersebut memperlihatkan hasil akurasi yang sangat baik yaitu 95% dengan ROC\_AUC sebesar 0.98. hal ini menunjukan bahwa model Random Forrest dapat dijadikan model untuk memprediksi calon peminjam credit card.

# MLP-Neural Net

## Data Split (Application\_Train.csv)

- Train : 80%
- Val : 20%
- Random\_State : 42



## Model Neural Net - KERAS

- Input (Selu) : 60
- Hid. Layer1 (Selu) : 30
- Hid. Layer2 (Selu) : 40
- Hid. Layer3 (Selu) : 30
- Output (Sigmoid) : 1
- batch\_size : 3534
- loss='binary\_crossentropy'
- optimizer='adam'

## Result

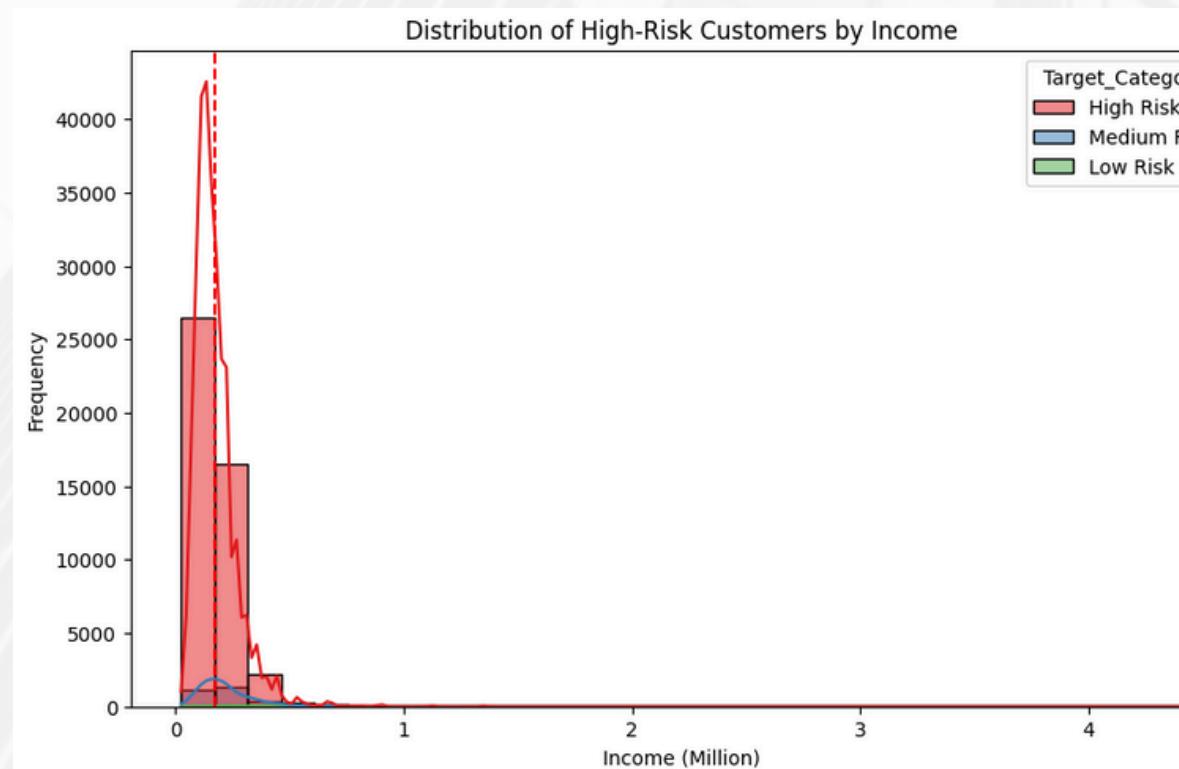
- Accuracy : 0.92
- Precision : 0.97
- Recall : 0.88
- F1- Score : 0.92
- ROC\_AUC : 0.97

Dari model MLP-Neural Network tersebut memperlihatkan hasil akurasi yang sangat baik yaitu 92% dengan ROC\_AUC sebesar 0.97. menunjukan bahwa model MLP mampu melakukan klasifikasi dengan cukup baik, namun sedikit lebih rendah dibandingkan dengan Random Forrest

# Business Recomendation

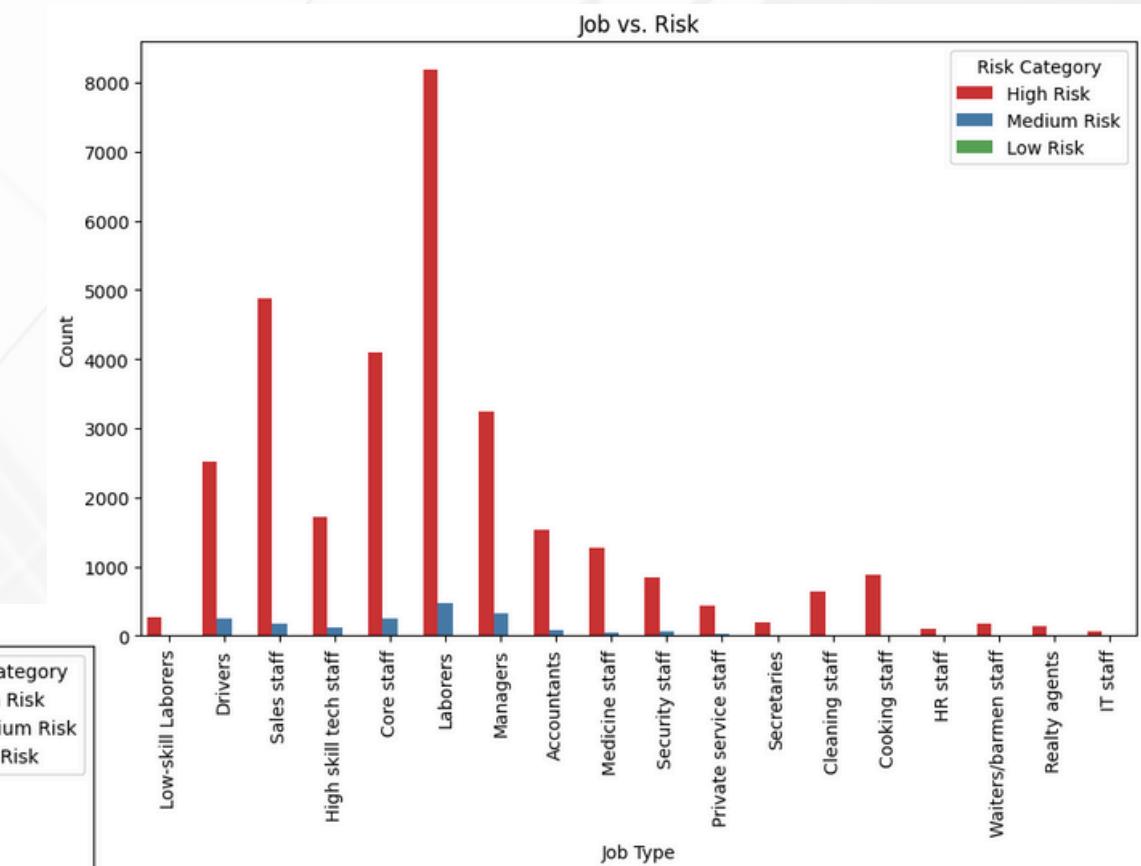
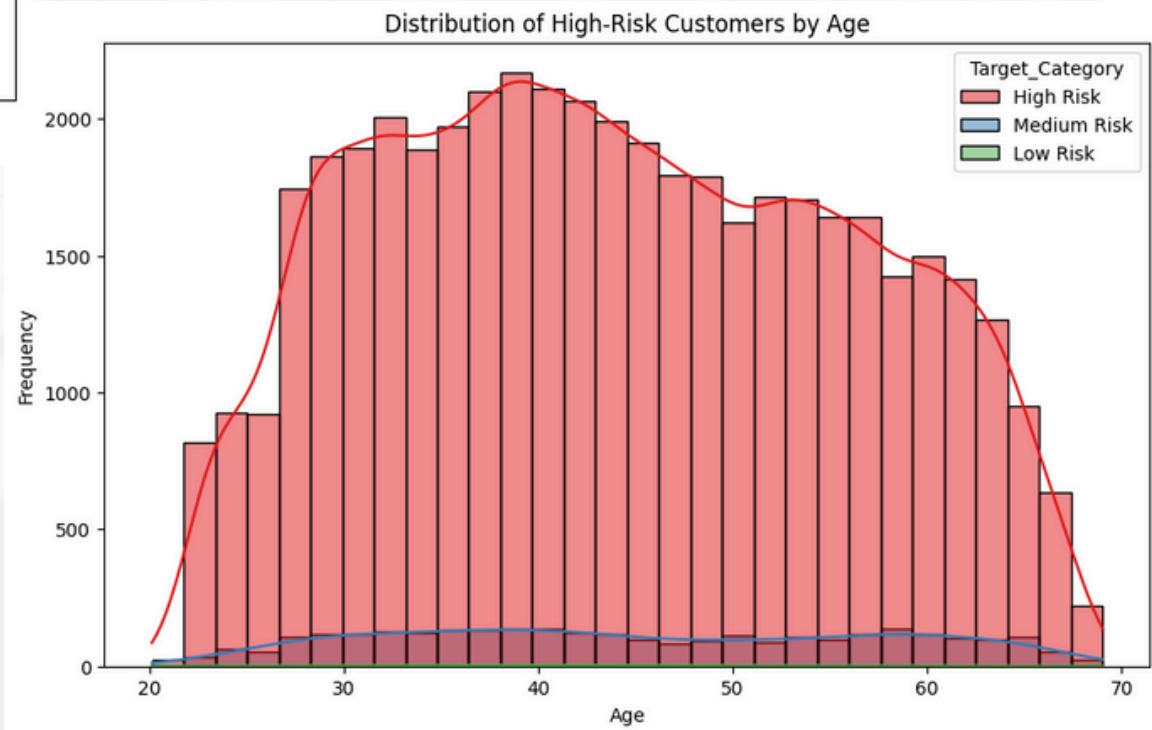
Menggunakan **Random Forrest Classifier** kita melakukan prediksi pada Application\_Test.csv

Dengan Rule Jika Prediksi  $\geq 0.75$  menjadi High Risk, Prediksi  $\geq 0.5$  menjadi Medium Risk, dan lainnya menjadi Low Risk



- Rata-rata Pendapatan High Risk = 0.18 Juta atau 175,336.48 Ribu

- Rata-rata Umur High Risk: 43.96 Tahun



- Pekerjaan Dengan High Risk Terbanyak :
  - Laborers
  - Sales Staff
  - Core Staff

# Kesimpulan

- Hasil Model Machine Learning dapat digunakan untuk mencegah penolakan pengajuan pinjaman yang seharusnya disetujui, dengan melihat skala resiko yang di prediksi.
- Berdasarkan Data Test tersebut klien yang memiliki profil resiko tinggi memiliki hasil prediksi diatas 0.75
- Pekerjaan seperti Laborers, Sales Staff, Core Staff memiliki potensi profil resiko yang tinggi lebih besar.



Github Project

Thank You