

# HOME CREDIT RISK SCORING

[INTIP.IN/KELOMPOK25WIDYA](http://INTIP.IN/KELOMPOK25WIDYA)

# HOME CREDIT RISK SCORING

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

# Business Understanding

# Business Understanding



**Kredit** adalah kemampuan untuk melaksanakan suatu pembelian atau mengadakan suatu pembelian atau mengadakan suatu **peminjaman dengan janji pembayaran** akan dilakukan atau ditangguhkan pada suatu jangka waktu yang disepakati.

PT. **Home Credit** Indonesia (HCI) merupakan **lembaga finansial yang menawarkan pemberian** di toko-toko yang telah bergabung menjadi mitra. Pemberian non tunai bisa langsung dirasakan untuk membeli berbagai produk perangkat rumah tangga, smartphone, dan perlengkapan elektronik lainnya.

Home Credit berusaha untuk memperluas inklusi keuangan bagi populasi yang tidak memiliki rekening bank dengan memberikan pengalaman meminjam yang positif dan aman. Untuk memastikan populasi yang kurang terlayani ini memiliki pengalaman pinjaman yang positif, Home Credit memanfaatkan berbagai data alternatif - termasuk informasi telekomunikasi dan transaksional - untuk memprediksi kemampuan pembayaran peminjam mereka.

Tidak sedikit dari masyarakat indonesia yang **tidak dapat melakukan pelunasan** dalam peminjamannya. Hal yang perlu dipahami saat melakukan peminjaman yaitu **meminjam sesuai dengan kemampuan finansial** dan **hindari peminjaman untuk hal yang konsumtif**.

# Business Understanding



## BUSINESS OBJECTIVE

Dalam menjalankan fungsi perusahaannya untuk memberikan pinjaman dan pengajuan cicilan kepada peminjamnya, Home Credit memiliki tujuan untuk mendapatkan keuntungan dari peran tersebut. Keuntungan untuk Home Credit setelah menganalisis model berdasarkan data yang dimiliki, perusahaan dapat **mengetahui faktor-faktor yang mempengaruhi orang untuk melakukan default**, yaitu kegagalan untuk membayar kembali pinjaman. Default dapat terjadi **ketika peminjam tidak dapat melakukan pembayaran tepat waktu**, melewatkkan pembayaran, atau berhenti melakukan pembayaran.

## BUSINESS SUCCESS

Non Performing Loan (**NPL**) adalah pinjaman bermasalah apabila peminjam tidak melakukan pembayaran yang dijadwalkan untuk jangka waktu tertentu. NPL juga disebut sebagai kredit bermasalah. Persentase NPL yang tinggi dapat mengakibatkan suatu penyedia jasa pinjaman mengalami kesulitan dalam menyalurkan kembali kredit. **Home Credit harus menurunkan NPL agar tidak mengalami kerugian dan dapat menyalurkan kembali kredit kepada peminjam .**

# Business Understanding



## SITUATION ASSESSMENT

### ASSUMPTIONS

1. Berusaha keras untuk mendapatkan pinjaman
2. Adanya keadaan darurat
3. Bisa menjadi penundaan pembayaran dari pinjaman lain
4. Kejadian/peristiwa khusus

### TERMINOLOGY

1. Annual Percentage Rate (APR) / Tingkat Persentase Tahunan
2. Jaminan
3. Peminjam
4. Nilai Kredit
5. Pinjaman angsuran

# Business Understanding



## SITUATION ASSESSMENT

### BENEFIT

- Increased sales (peningkatan penjualan) : pembayaran menggunakan kartu kredit sangat berhubungan dengan peningkatan dorongan untuk membeli, yang meningkatkan frekuensi dan total pengeluaran jumlah pembelian.
- Customer satisfaction (kepuasan peminjam ) : 75% konsumen lebih memilih menggunakan kartu kredit dan debit untuk pembelian lebih dari uang tunai.
- Secure transactions (transaksi aman): 53% konsumen merasa bahwa pembayaran melalui kartu kredit dan debit lebih aman dibanding dengan menggunakan uang tunai.
- Safer money-handling practices (praktik penanganan uang yang lebih aman) : pembayaran digital mengurangi waktu dan biaya perhitungan, menyortir dan membawa uang tunai ke bank. Pembayaran ini juga akan membuat bisnis terhindar dari pencurian karena hanya terdapat sedikit uang di tempat itu.
- Clear accounting records (catatan akuntansi yang jelas) : dengan pembayaran digital, terdapat catatan yang jelas tentang jumlah dan waktu pembelian.

# Business Understanding



## DATA MINING GOAL

Membuat model menggunakan data historis aplikasi pinjaman untuk memprediksi apakah peminjam akan dapat membayar kembali pinjaman atau tidak.

Mengetahui faktor-faktor yang menyebabkan peminjam tidak dapat mengembalikan pinjaman dalam tempo waktu yang telah ditentukan.

# Business Understanding



## PRODUCE PROJECT PLAN

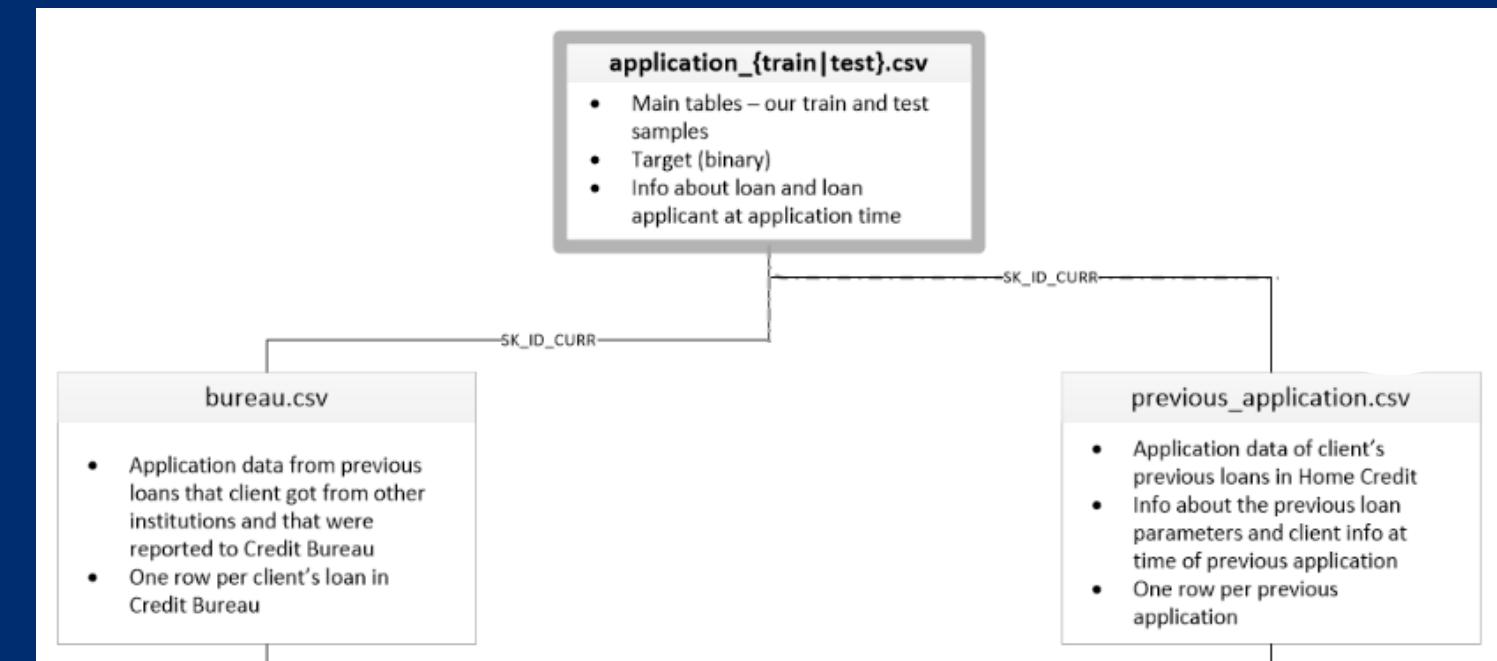
Untuk mengatasi permasalahan tersebut, maka perlu dibuat suatu model klasifikasi agar diperoleh faktor-faktor yang dapat berpengaruh terhadap kelayakan dalam memberi pinjaman uang atau kredit kepada peminjam. Pembuatan model klasifikasi ini menggunakan metode Logistic Regression, Decision Tree Classifier, Random Forest Classifier, dan Naive Bayes Classifier.

Model ini akan dibuat dengan bahasa pemrograman Python yang ada di Google Colab. Kemudian, model akan dievaluasi menggunakan confusion matrix. Apabila model yang dibuat sudah cukup baik maka model dapat dianalisis untuk mengetahui faktor-faktor yang berpengaruh terhadap kelayakan dalam memberi pinjaman uang atau kredit kepada peminjam . Hasil analisis dari model tersebut akan ditampilkan sebagai dashboard menggunakan Google Data Studio.



# Data Understanding

# Data Understanding



Terdapat 3 **Dataset** yang digunakan:

- **Application train** : data untuk aplikasi kredit saat ini. Setiap pinjaman memiliki barisnya sendiri dan diidentifikasi oleh feature `SK_ID_CURR`.
- **Bureau** : Semua kredit peminjam sebelumnya disediakan oleh lembaga keuangan lain yang dilaporkan ke Biro Kredit untuk setiap pinjaman dalam sampel. Ada baris sebanyak jumlah kredit yang dimiliki peminjam di Biro Kredit sebelum tanggal pengajuan.
- **Previous\_application** : semua ajuan pinjaman sebelumnya untuk Home Credit loans dari peminjam yang memiliki pinjaman dalam sampel. Ada satu baris untuk setiap ajuan sebelumnya yang terkait dengan pinjaman dalam sampel data.

# Data Understanding



## Application Train

total rows & columns

- 30.7511 rows
- 122 columns

total type data

- float64 = 65
- int64 = 41
- object = 16

## Previous Application

total rows & columns

- 1.670.214 rows
- 37 columns

total type data

- object 16
- float64 15
- int64 6

## Bureau

total rows & columns

- 1.716.428 rows
- 17 columns

total type data

- float64 8
- int64 6
- object 3



# Data Understanding

Jumlah peminjaman Home Credit sebelumnya pada tiap peminjam

	SK_ID_CURR	TOTAL_PREV
0	100001	1
1	100002	1
2	100003	3
3	100004	1
4	100005	2

Jumlah kredit yang dimiliki peminjam di Biro Kredit

	SK_ID_CURR	TOTAL_BUREAU_LOAN
0	100001	1
1	100002	1
2	100003	3
3	100004	1
4	100005	2

Join dataset previous application dan bureau kredit ke application train dengan left join by SK\_ID\_CURR

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE	NAME_INCOME_TYPE
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	24700.5	351000.0	Unaccompanied	Working
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	1129500.0	Family	State servant
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0	135000.0	Unaccompanied	Working
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.5	29686.5	297000.0	Unaccompanied	Working
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.0	21865.5	513000.0	Unaccompanied	Working



# Data Understanding



## EXPLORATORY DATA ANALYSIS

### Memeriksa adanya missing value pada dataset

terdapat 69 variable missing value dengan 17 diantaranya memiliki presentase missing value lebih dari 60%

### Memeriksa adanya duplikat data

tidak ada data yang duplikat

### Melihat tipe data dan jumlah pada dataset yang telah di join

float64	67
int64	41
object	16

		index	Total	Percent
0	COMMONAREA_MODE	214865	69.872297	
1	COMMONAREA_AVG	214865	69.872297	
2	COMMONAREA_MEDI	214865	69.872297	
3	NONLIVINGAPARTMENTS_AVG	213514	69.432963	
4	NONLIVINGAPARTMENTS_MODE	213514	69.432963	
5	NONLIVINGAPARTMENTS_MEDI	213514	69.432963	
6	FONDKAPREMONT_MODE	210295	68.386172	
7	LIVINGAPARTMENTS_MEDI	210199	68.354953	
8	LIVINGAPARTMENTS_AVG	210199	68.354953	
9	LIVINGAPARTMENTS_MODE	210199	68.354953	
10	FLOORSMIN_MEDI	208642	67.848630	
11	FLOORSMIN_MODE	208642	67.848630	
12	FLOORSMIN_AVG	208642	67.848630	
13	YEARS_BUILD_AVG	204488	66.497784	
14	YEARS_BUILD_MODE	204488	66.497784	
15	YEARS_BUILD_MEDI	204488	66.497784	
16	OWN_CAR_AGE	202929	65.990810	

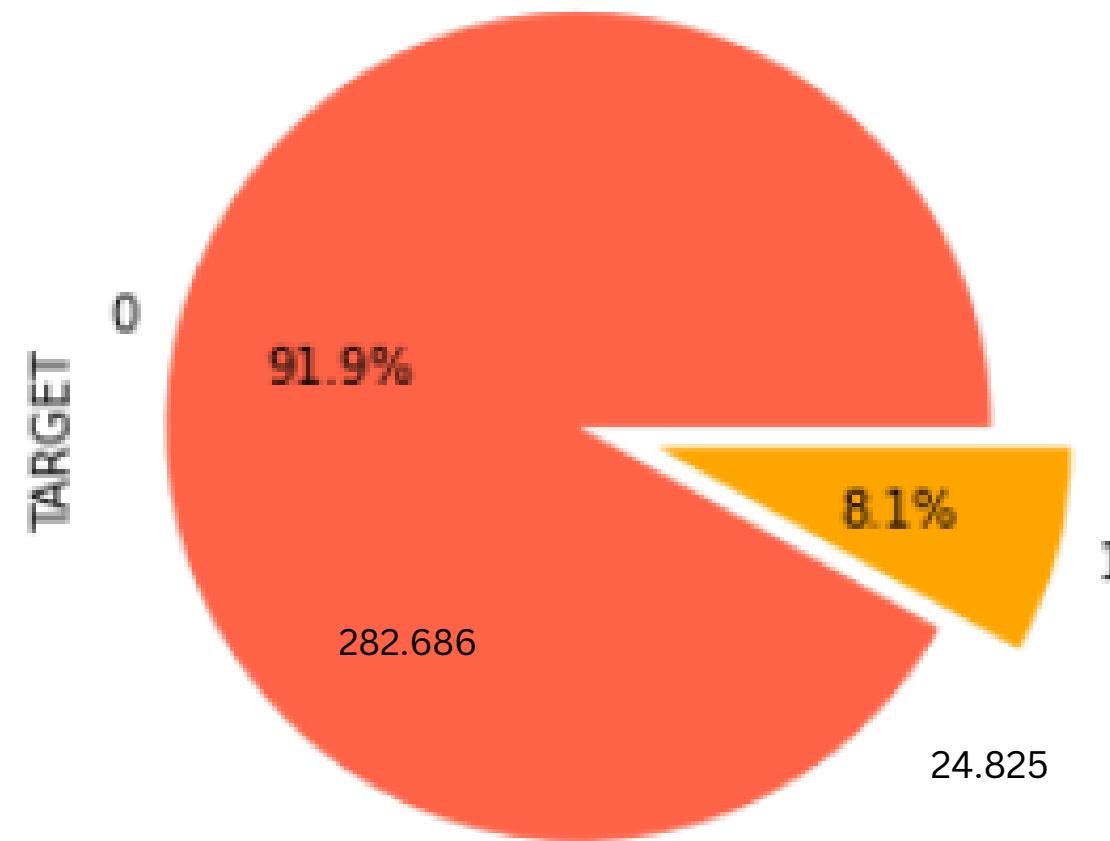


# Data Understanding



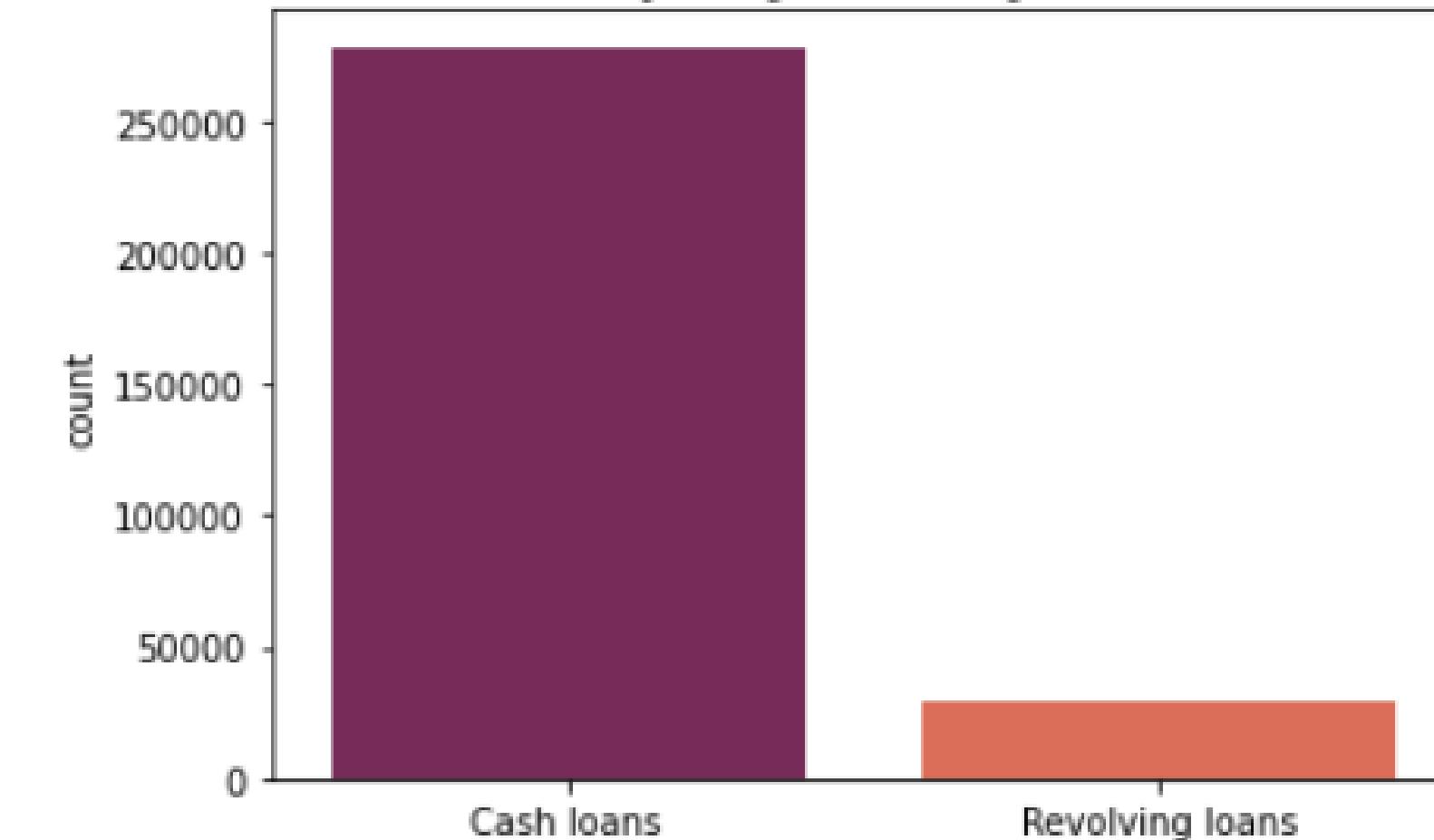
## EXPLORATORY DATA ANALYSIS

Komposisi Status Peminjaman



0= peminjaman yang dibayar tepat waktu  
1= peminjam sudah membayar pinjaman tepat waktu

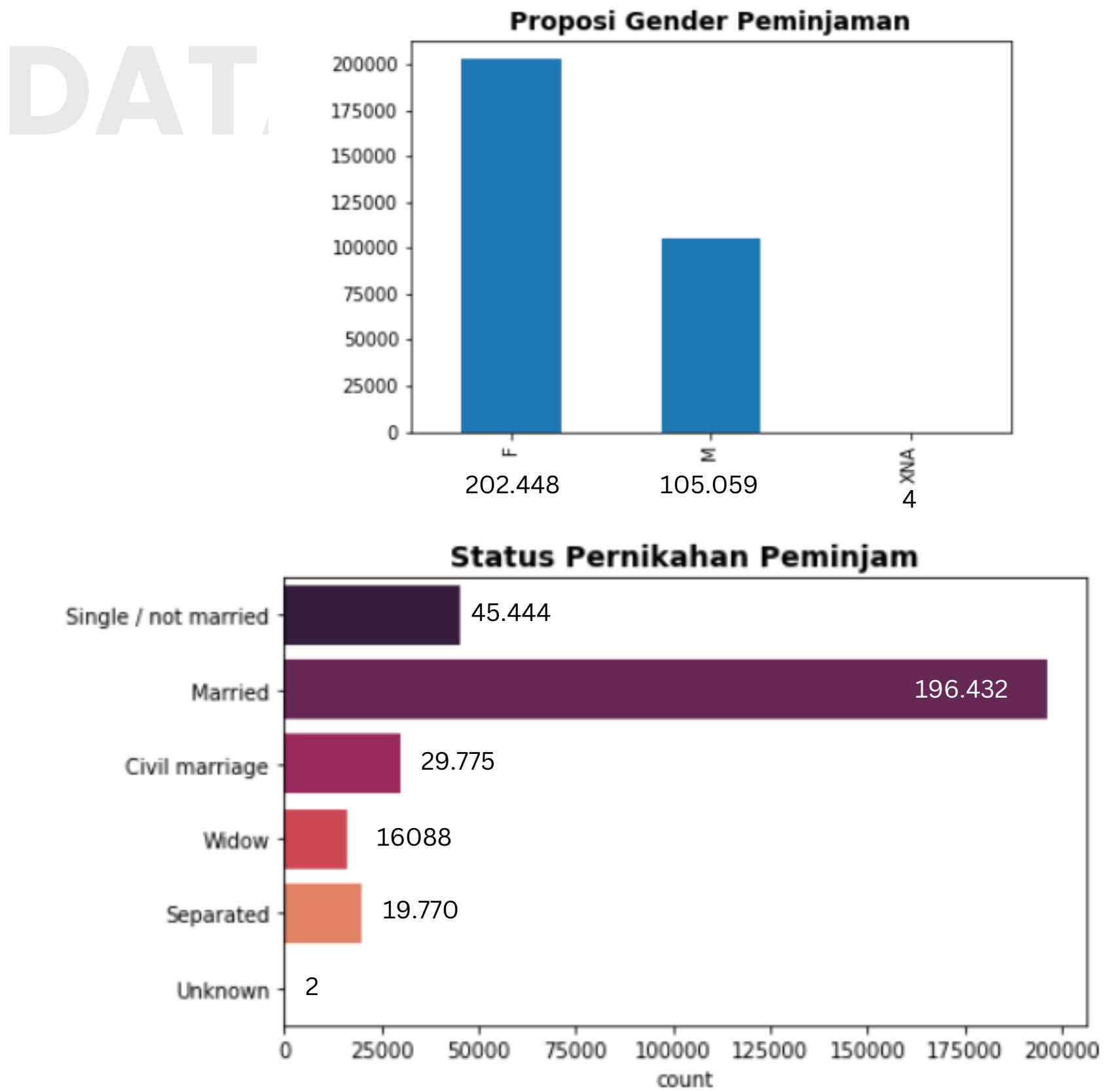
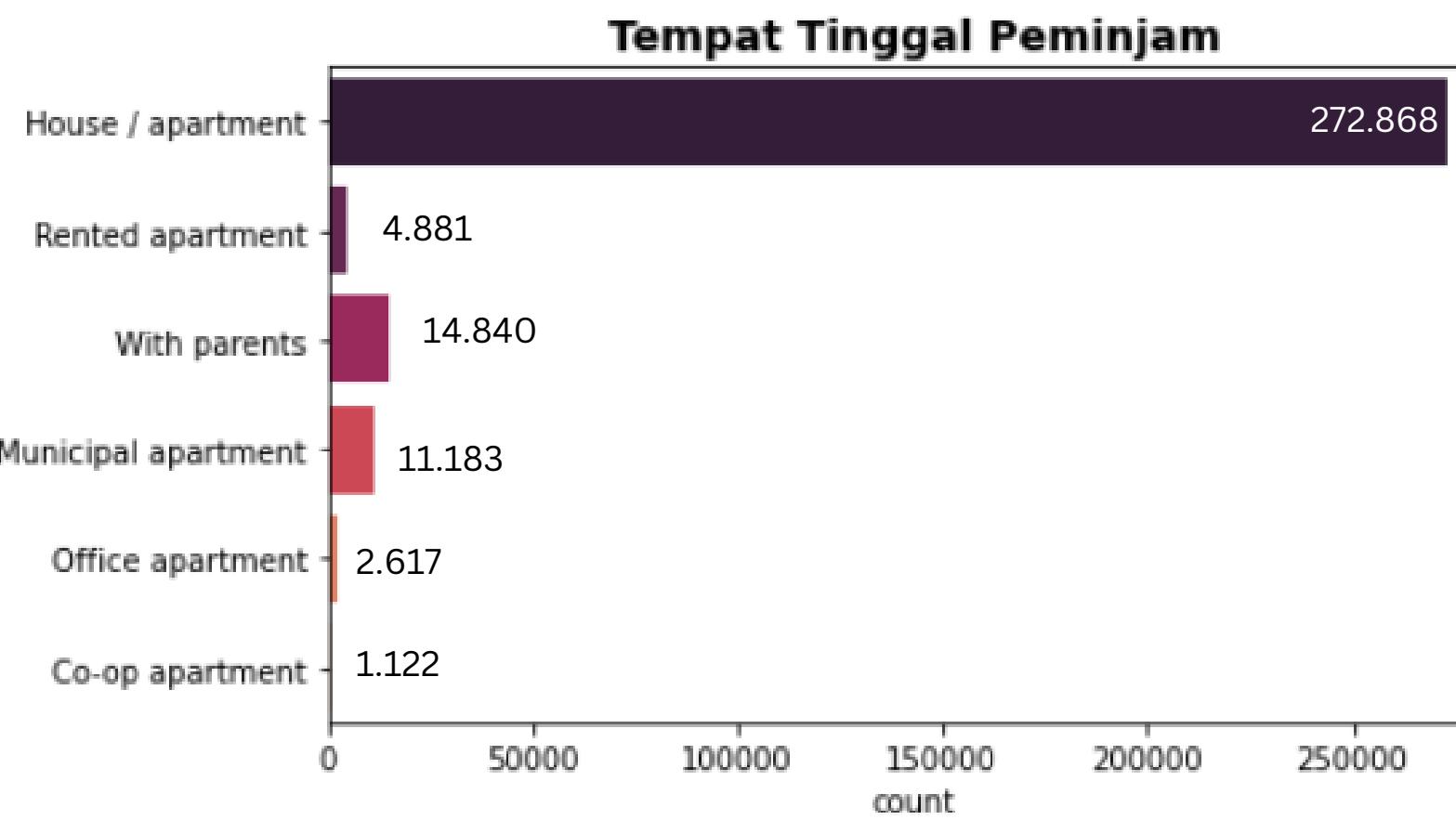
Proposi Jenis Pinjaman



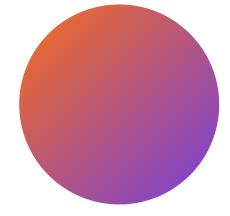


## EXPLORATORY DATA ANALYSIS

### Basic Info Costumers

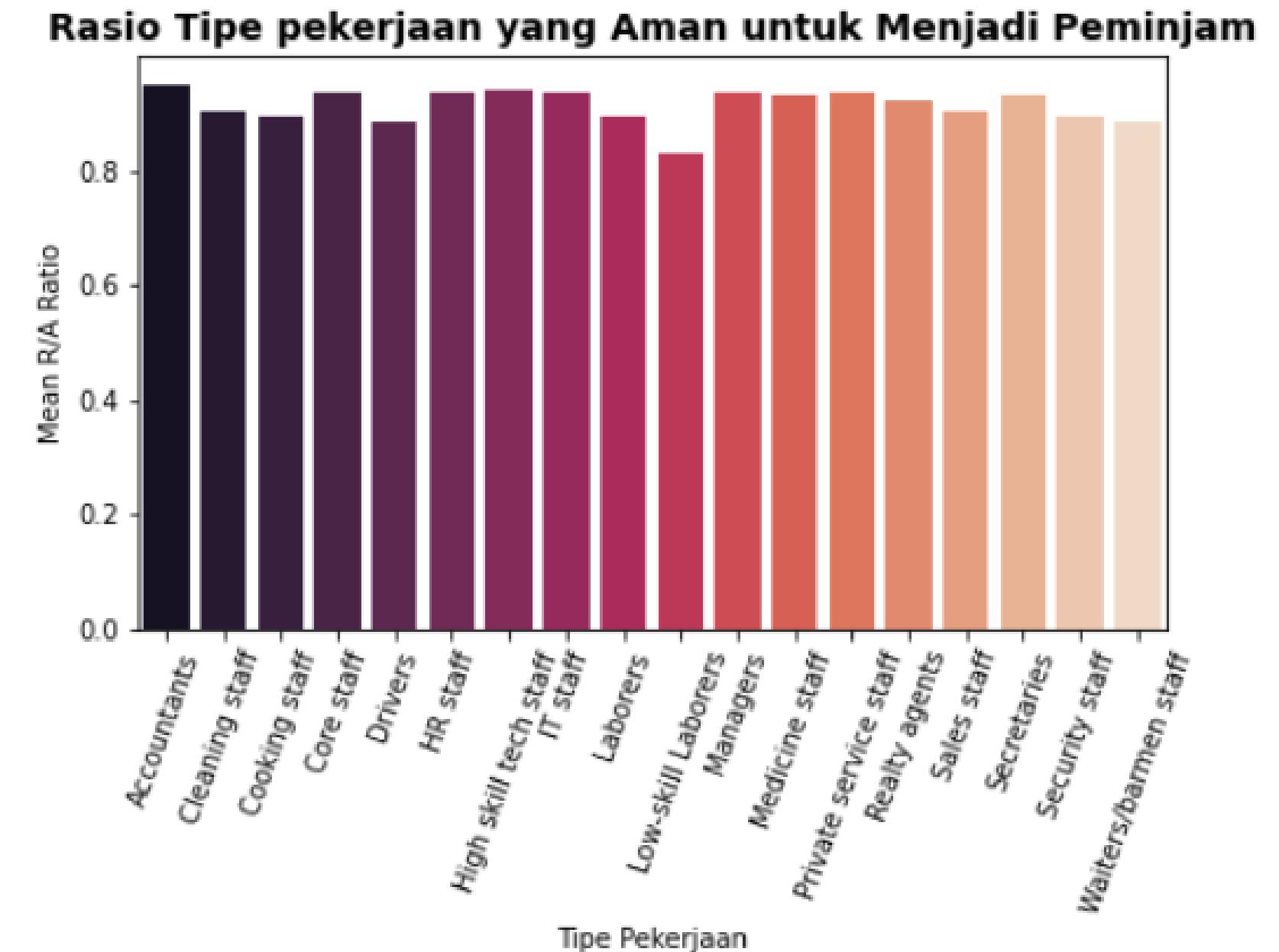
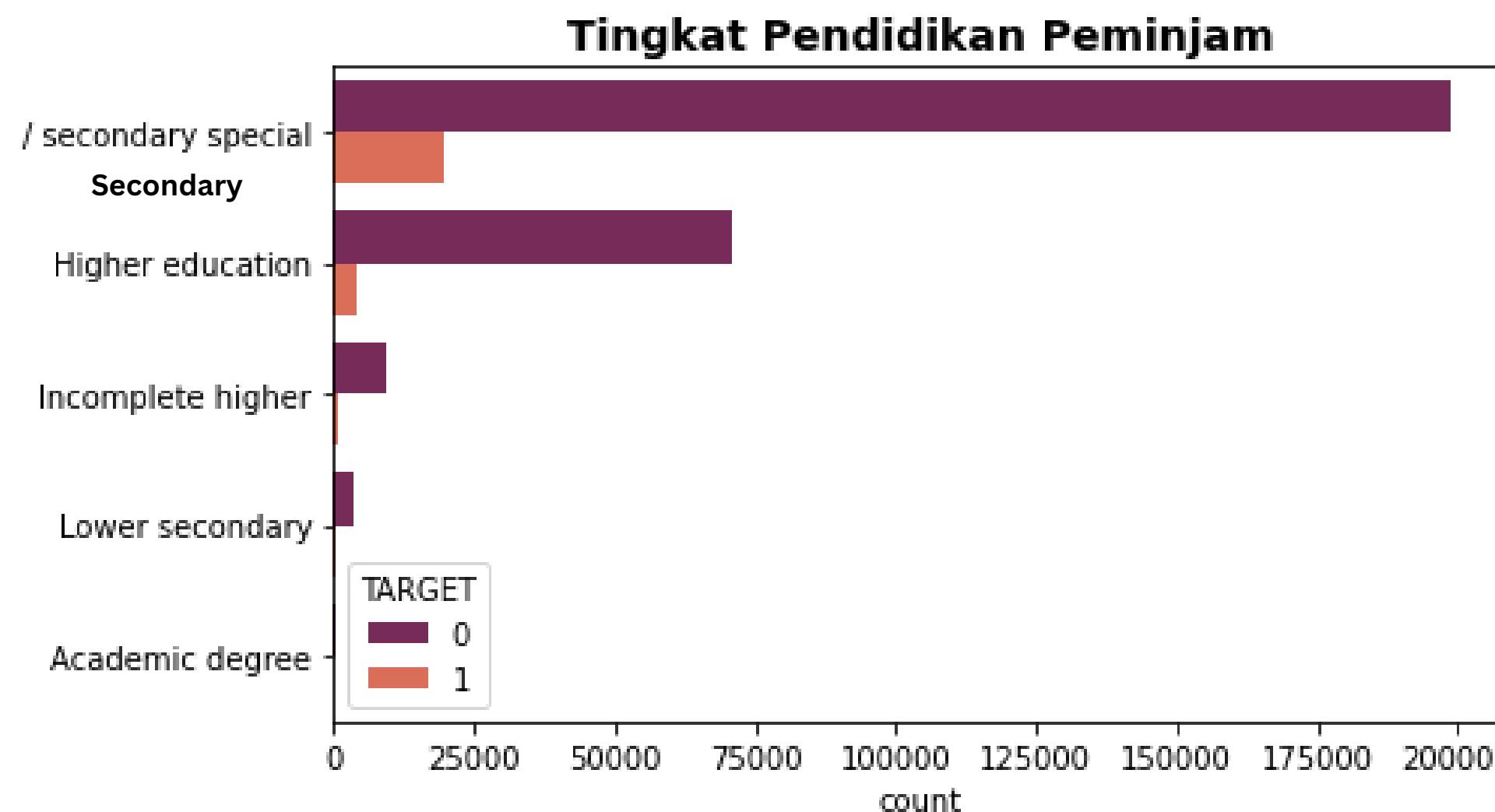


# Data Understanding



## Basic Info Costumers

# EXPLORATORY DATA ANALYSIS



# Data Preparation



# Data Preparation

- Dataframe utama, yaitu df memiliki 307.511 baris dan 124 kolom.
- Mengubah kolom DAYS\_BIRTH menjadi AGE untuk mendapatkan usia peminjam .
- Mengubah kolom DAYS\_EMPLOYED menjadi YEARS\_EMPLOYED untuk mendapatkan lama masa kerja peminjam dalam tahun.
- Mengubah kolom DAYS\_REGISTRATION menjadi YEARS\_REGISTRATION untuk mendapatkan berapa tahun sebelum peminjam mengubah registrasi.
- Mengganti data XNA menjadi NaN
- Mengapus kolom dengan jumlah nilai > 60 %
- Melakukan *handling missing value* menggunakan *imputation*
- Mengubah kolom DAYS\_ID\_PUBLISH menjadi YEARS\_PUBLISHED.



# Data Preparation

- Mengubah kolom DAYS\_LAST\_PHONE\_CHANGE menjadi YEAR\_LAST\_PHONE\_CHANGE untuk mengetahui berapa tahun sebelum *application peminjam mengganti handphone*
- Membuat kolom DOC\_PROVIDED untuk mengetahui apakah peminjam memberikan dokumen.
- Membuat kolom TOTAL\_DOCUMENT untuk mengetahui jumlah dokumen yang diberikan oleh peminjam .
- Melakukan scaling untuk data numerical menggunakan MinMaxScaler()
- Melakukan *encoding* terhadap data categorical. Untuk data yang memiliki 2 *unique value* diencoding menggunakan LabelEncoder(), sedangkan untuk data yang *unique valuenya* lebih dari 2 menggunakan one hot encoding.



# Data Preparation



## Kolom yang digunakan untuk Modelling

## FEATURE SELECTION

		Most Negative Correlations:	
		index	TARGET
		0	EXT_SOURCE_2 -0.160295
		1	EXT_SOURCE_3 -0.155892
		2	EXT_SOURCE_1 -0.098887
Most Positive Correlations:		3	AGE -0.078263
193	REG_CITY_NOT_LIVE_CITY	4	YEARS_EMPLOYED -0.065095
194	FLAG_EMP_PHONE	5	Higher education -0.056593
195	Secondary / secondary special	6	YEAR_LAST_PHONE_CHANGE -0.054345
196	REG_CITY_NOT_WORK_CITY	7	YEARS_PUBLISHED -0.051353
197	CODE_GENDER	8	Pensioner -0.046209
198	Working	9	YEARS_REGISTRATION -0.041820
199	REGION_RATING_CLIENT	10	AMT_GOODS_PRICE -0.039623
200	REGION_RATING_CLIENT_W_CITY	11	FLOORSMAX_AVG -0.039385
201	TARGET	12	FLOORSMAX_MEDI -0.039157

```
▶ column = ['AMT_CREDIT', 'REG_CITY_NOT_LIVE_CITY', 'FLAG_EMP_PHONE', 'Secondary / secondary special', 'REG_CITY_NOT_WORK_CITY',
            'CODE_GENDER', 'Working', 'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY',
            'EXT_SOURCE_2', 'EXT_SOURCE_3', 'EXT_SOURCE_1', 'AGE', 'YEARS_EMPLOYED',
            'Higher education', 'YEAR_LAST_PHONE_CHANGE', 'YEARS_PUBLISHED', 'Pensioner',
            'YEARS_REGISTRATION', 'AMT_GOODS_PRICE', 'FLOORSMAX_AVG', 'FLOORSMAX_MEDI', 'TARGET']
```



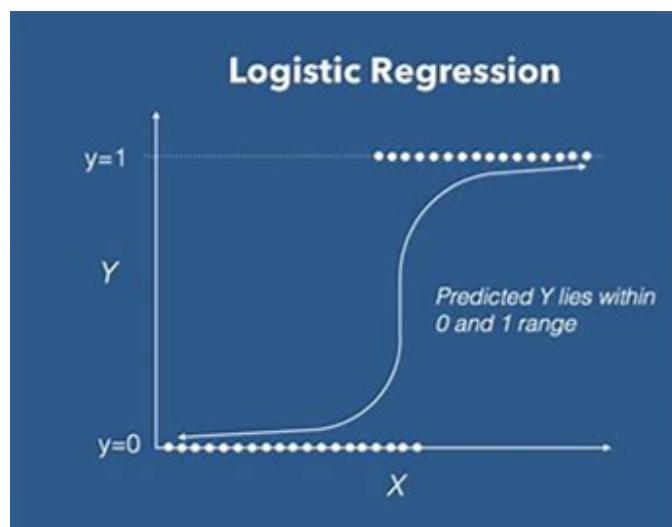
# Modeling



## Model Machine Learning Home Credit Risk Scoring

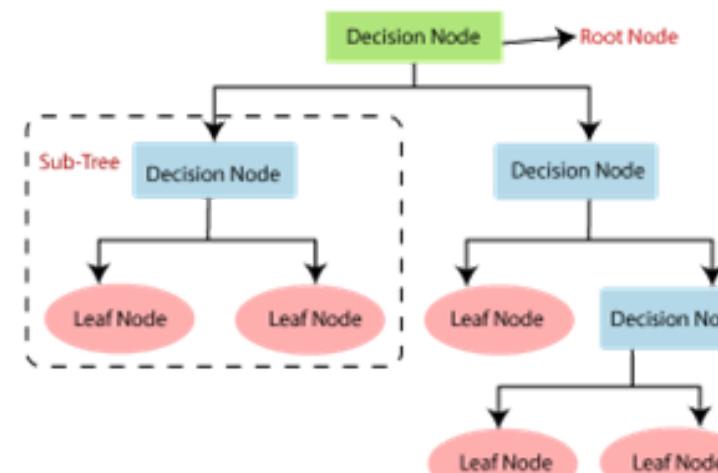
### Logistic Regression

*Logistic regression* adalah algoritma *machine learning* untuk kasus klasifikasi. Algoritma ini biasa digunakan untuk menghitung nilai probabilitas dengan data *input* yang bersifat diskrit.



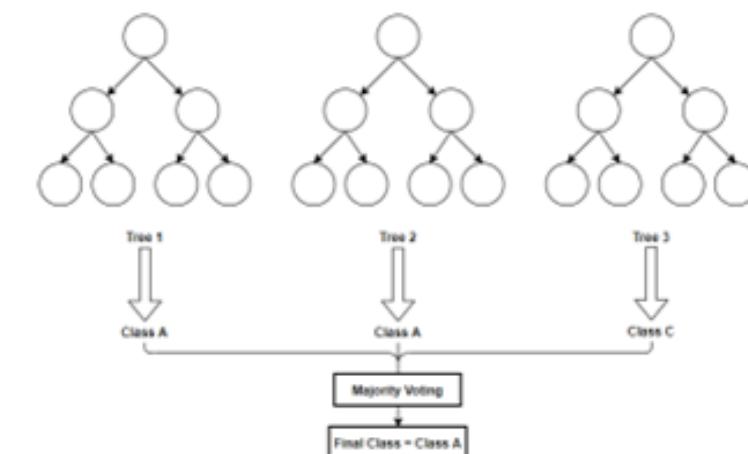
### Decision Tree Classifier

Decision Tree Classifier merupakan algoritma machine learning klasifikasi yang strukturnya mirip seperti sebuah pohon yang memiliki akar, ranting, dan daun.



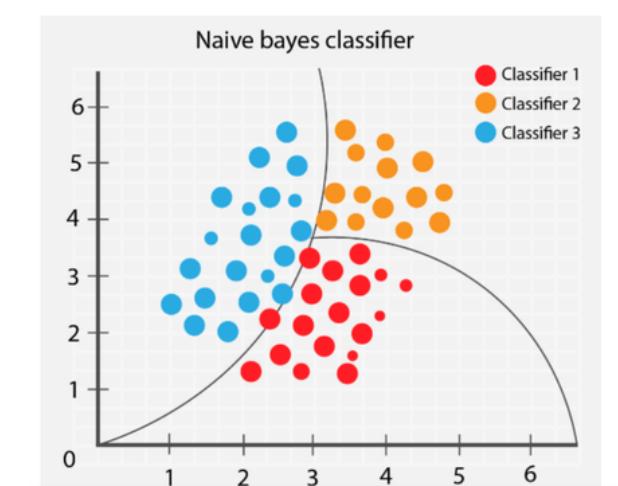
### Random Forest Classifier

*Random Forest* merupakan kumpulan dari sejumlah *decision tree*. Masing-masing *decision tree* ini akan memperoleh hasil yang mungkin berbeda-beda. *Random forest* ini akan melakukan *voting* untuk menentukan hasil mayoritas dari semua *decision tree*.



### Naïve Bayes Classifier

Naïve Bayes Classifier merupakan sebuah algoritma machine learning yang berakar pada teorema Bayes. algoritma ini menggunakan metode probabilitas dan teorema Bayes. Ciri utama dari Naïve Bayes Classifier ini adalah asumsi yg sangat kuat (naïf) akan independensi dari masing-masing kondisi.



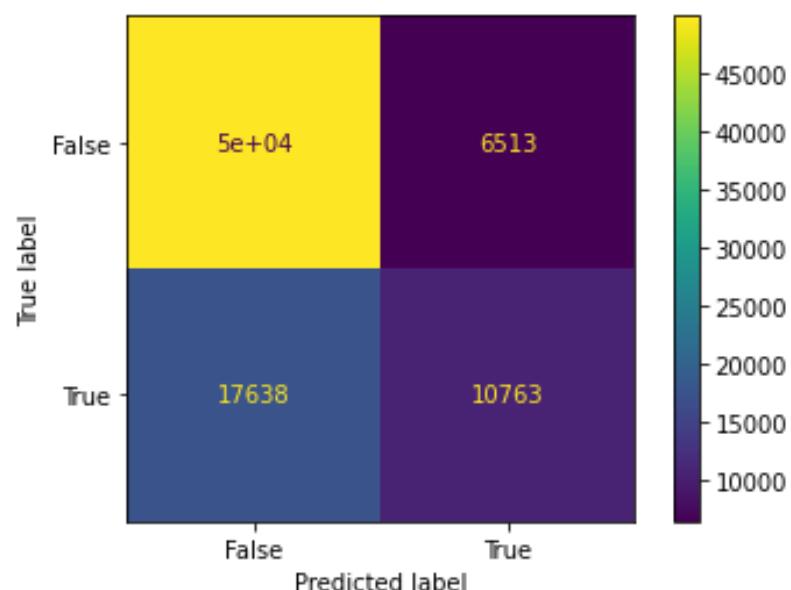


# Evaluation

# Evaluation

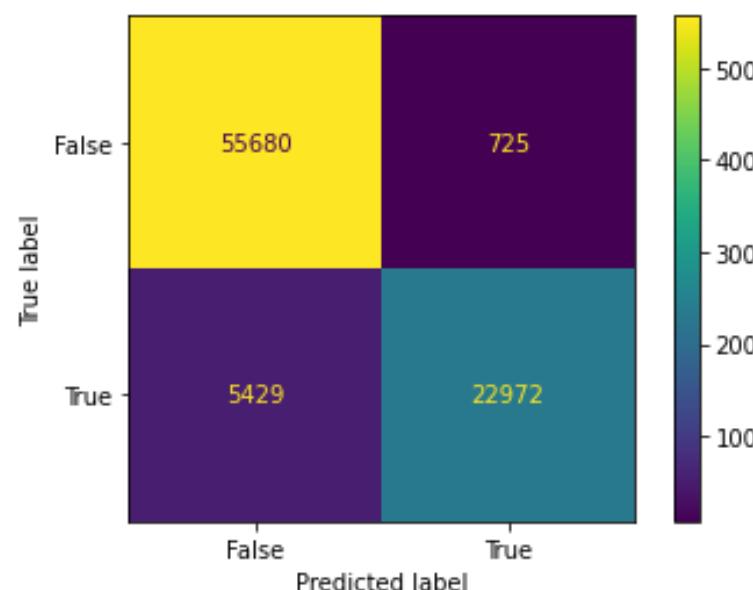
## Logistic Regression

Nilai akurasi: 0.72  
Nilai AUC: 0.74



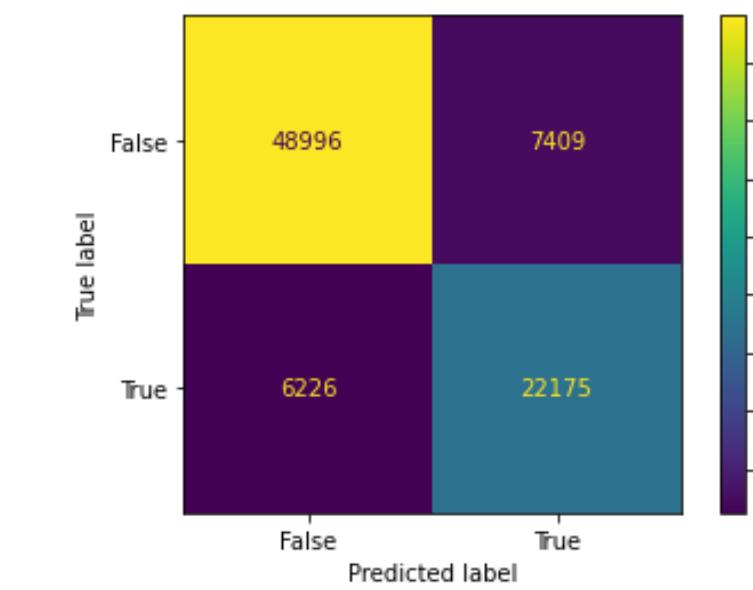
## Decision Tree Classifier

Nilai akurasi: 0.93  
Nilai AUC: 0.97



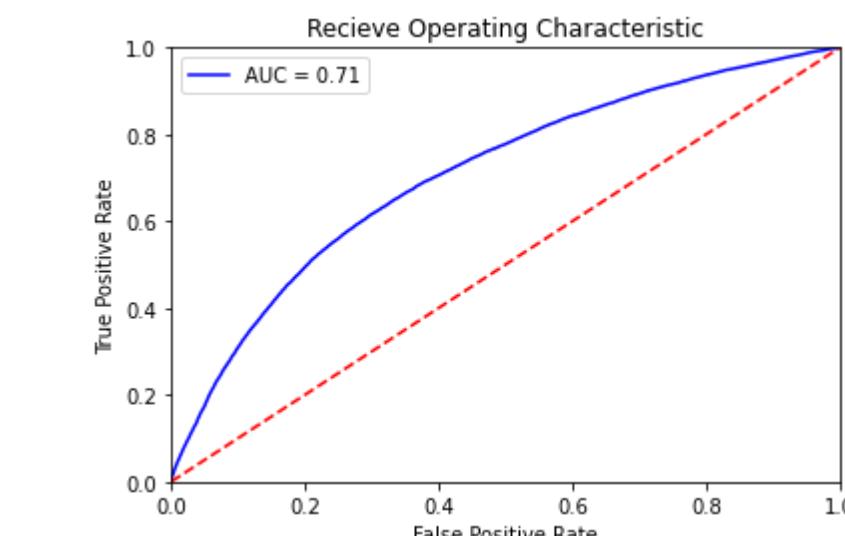
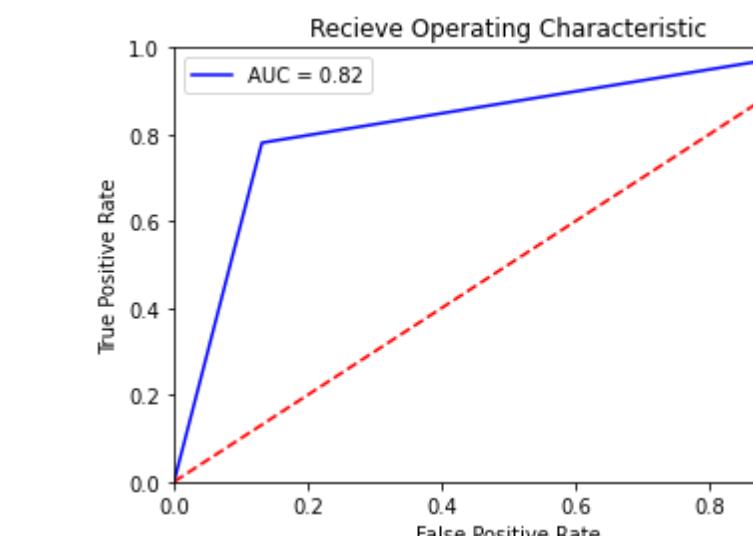
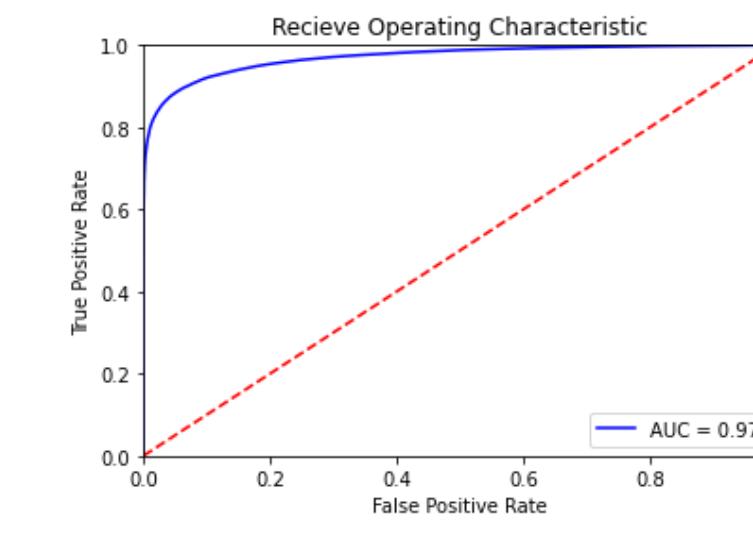
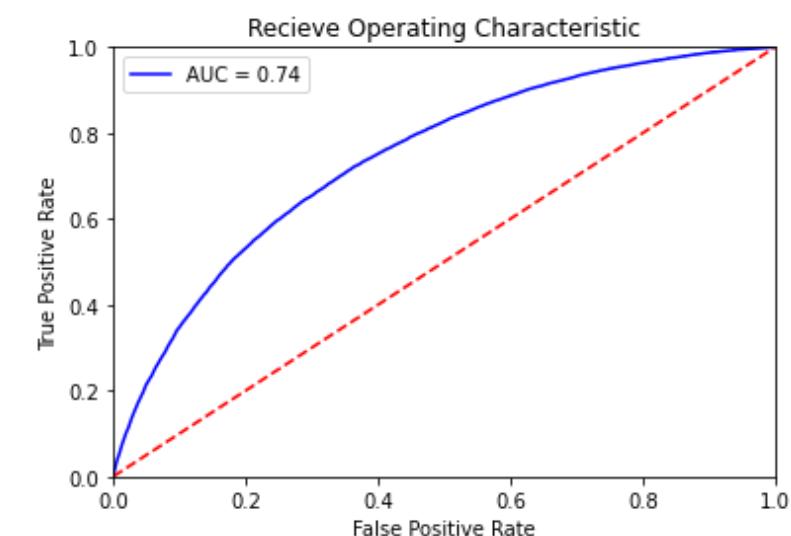
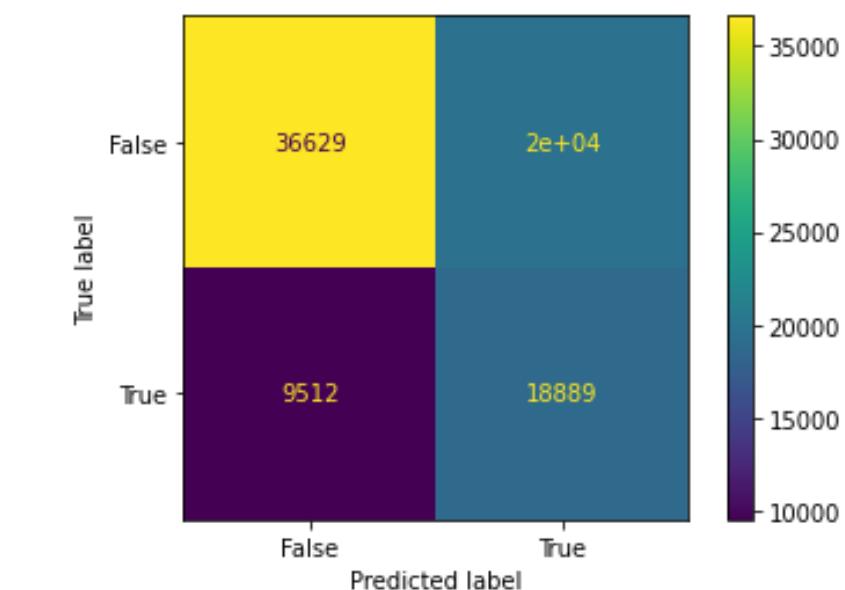
## Random Forest Classifier

Nilai akurasi: 0.84  
Nilai AUC: 0.82

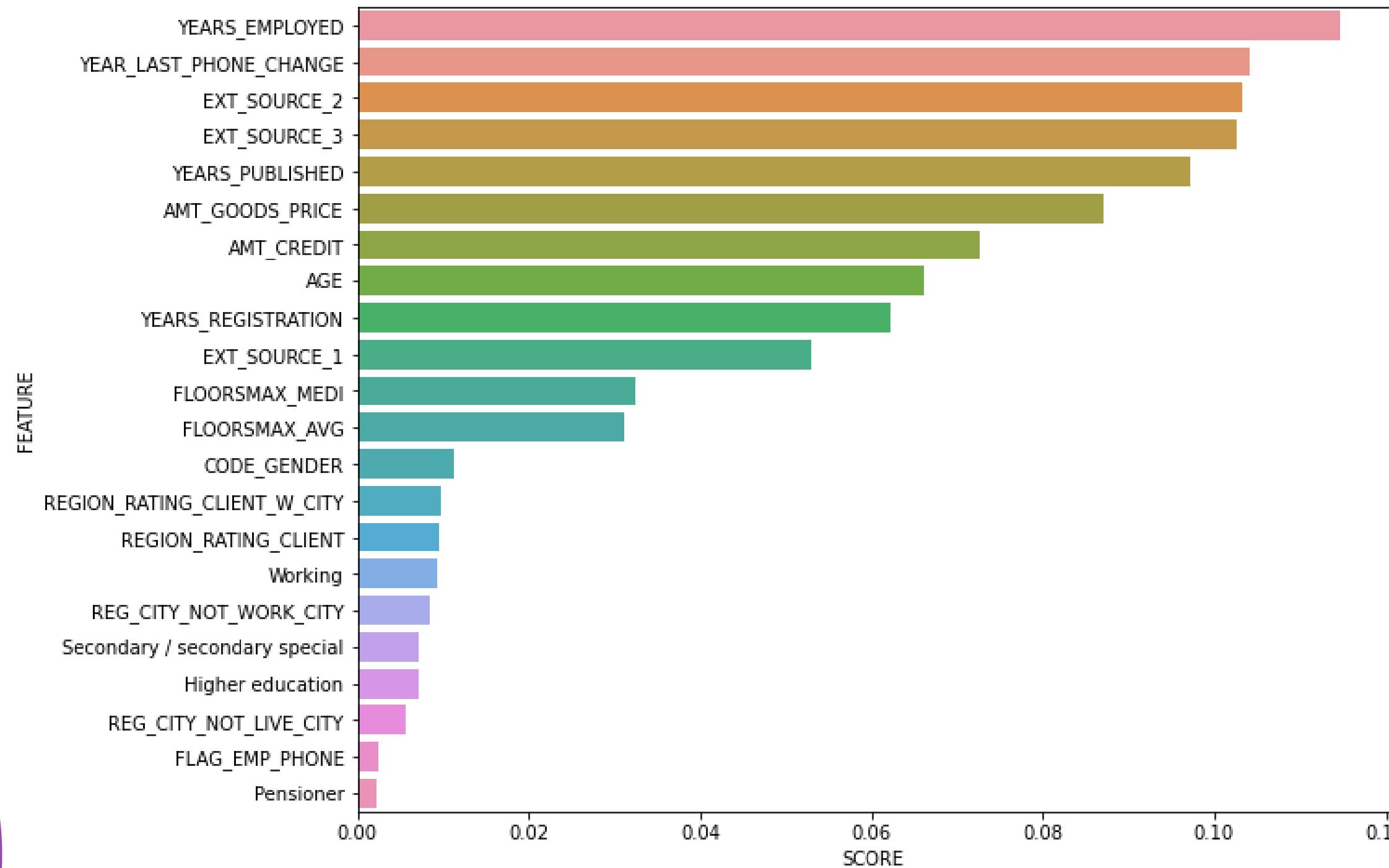


## Naïve Bayes Classifier

Nilai akurasi: 0.65  
Nilai AUC: 0.71



# Feature Importance



# Business Metric

Potential impact setelah menggunakan model yang dibangun adalah menurunnya Loss Given Default (LGD).

Loss Given Default (LGD) sebelum menggunakan model :

```
LGDBefore = Defaulters['AMT_CREDIT'].loc[Defaulters['TARGET'] == 1]
LGDBefore
1    3535.999551
```

Loss Given Default (LGD) setelah menggunakan model :

```
LGDAfter = FalseNegative['AMT_CREDIT'].sum()
LGDAfter
726.949265674276
```

Percentase penurunan LGD :

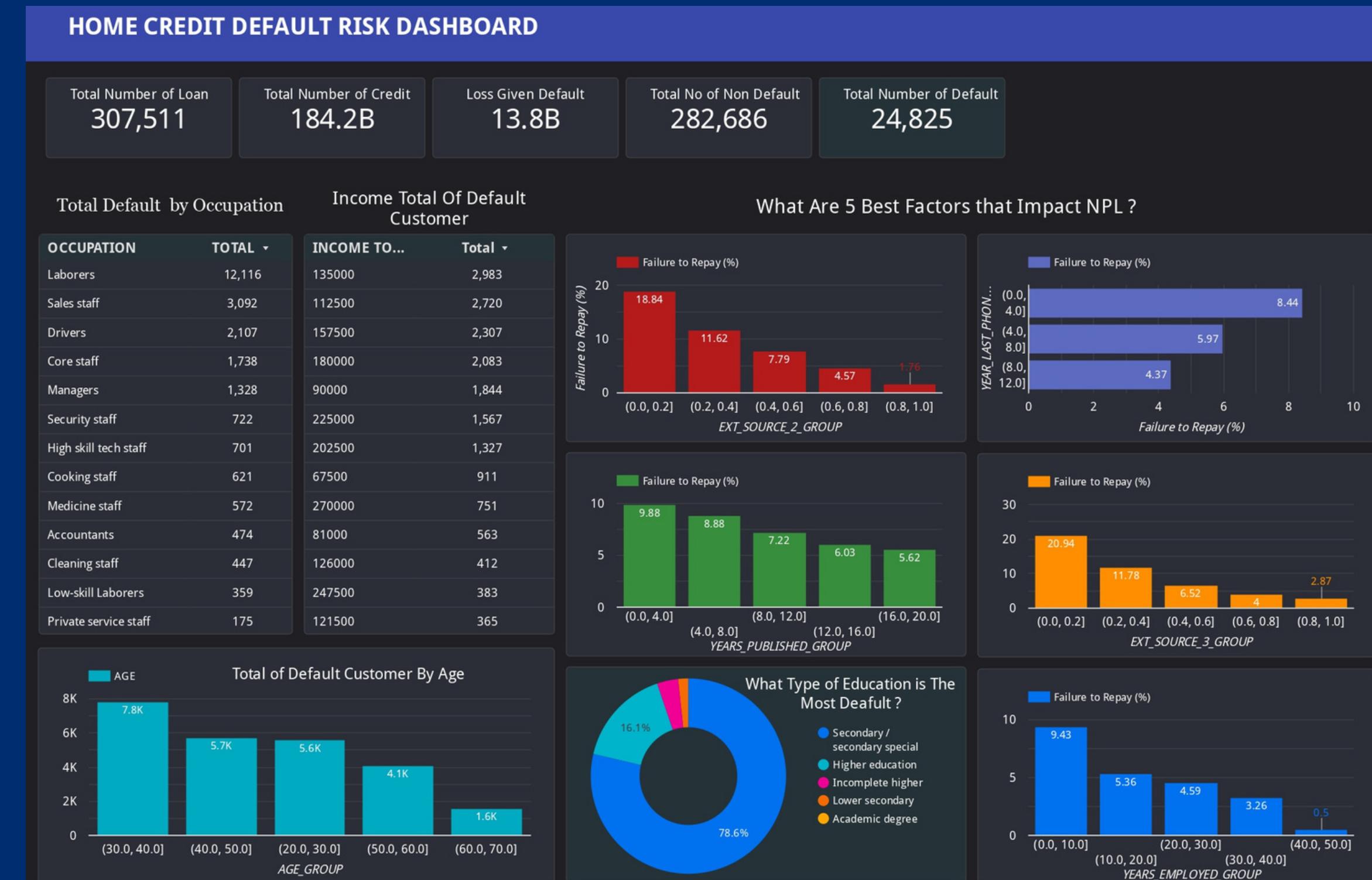
```
LGD_Decreased_Percentage
1    -79.441477
```



# Deployment

# Deployment

## HOME CREDIT - DASHBOARD


[LINK GOOGLE DATASTUDIO](#)



# Thank You