

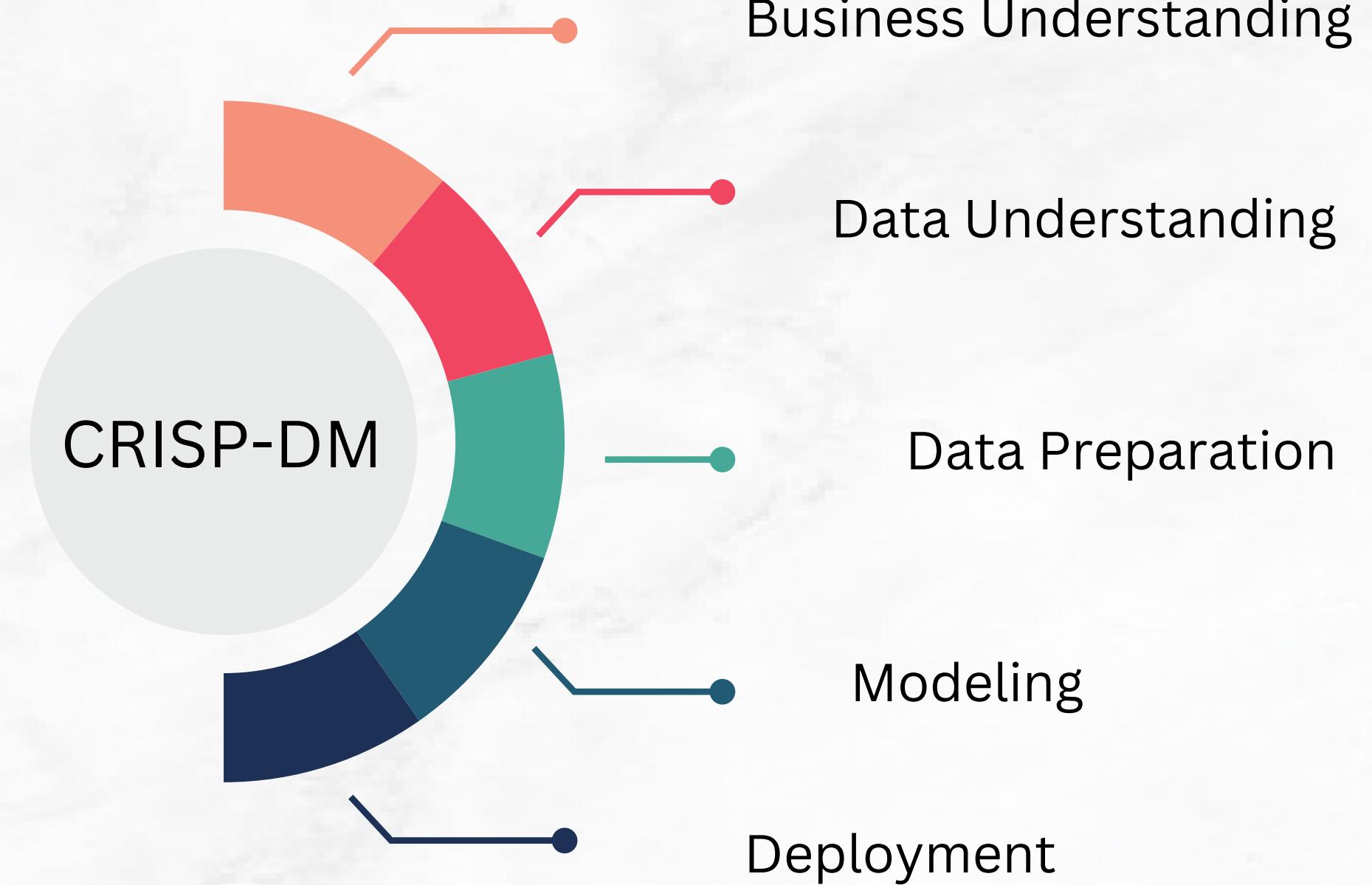


Home Credit Risk Scoring

Kelompok 8

- Annisa Riza Utami
- Mohammad Rizki Kurniawan
- Syarafina Dewi
- Zurida Alisha Muntaz







Business Understanding

...

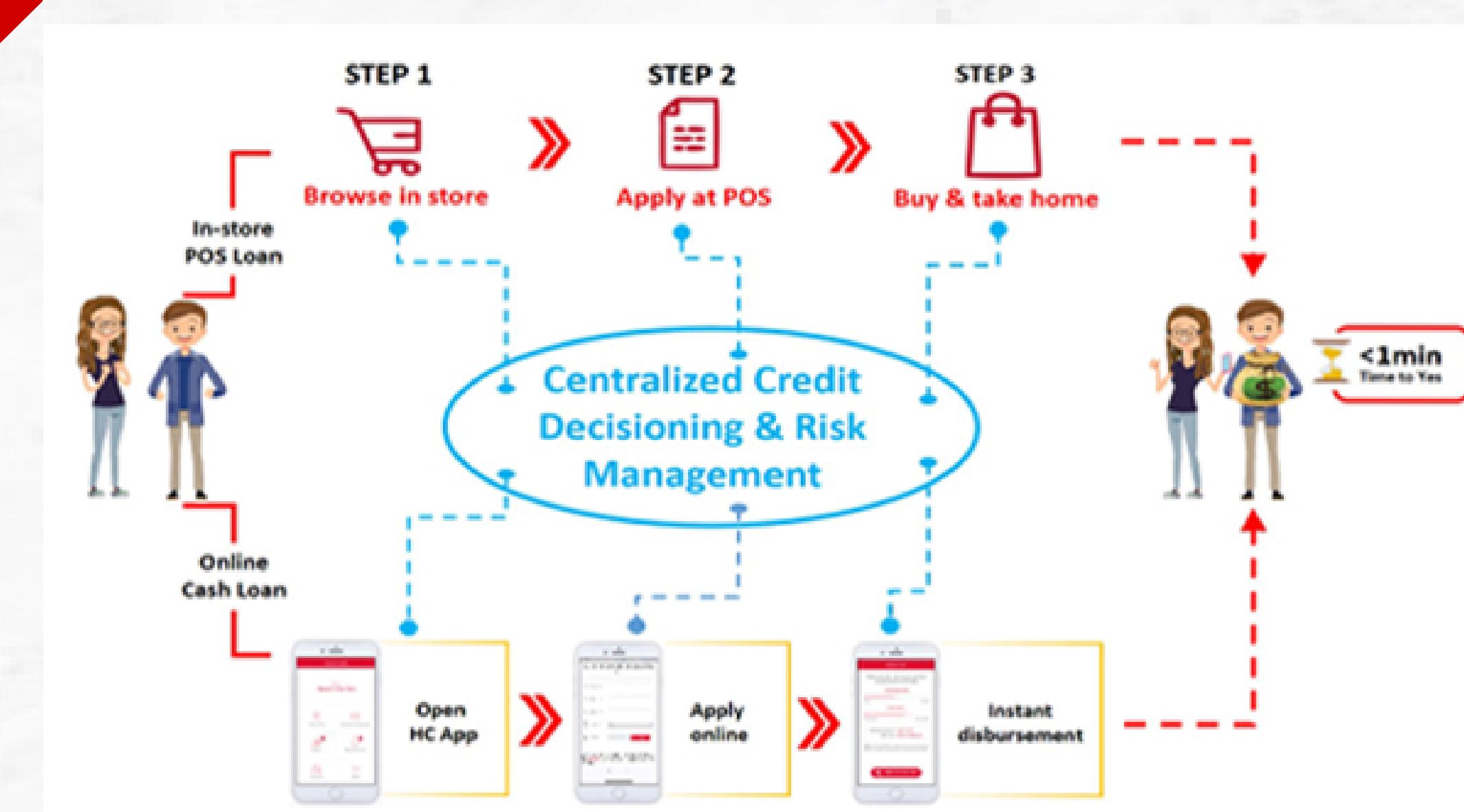
Determine Business Objectives

- **Background**

Home Credit merupakan perusahaan penyedia keuangan konsumen internasional yang beroprasi di 8 negara. Perusahaan ini berfokus pada pinjaman yang bertanggung jawab terutama kepada orang-orang dengan sedikit atau tanpa riwayat kredit. Home Credit bertujuan untuk menyediakan layanan keuangan ritel yang inovatif dengan focus pada pinjaman massa-ritel.

- **Business Objectives**

Dalam menjalankan fungsi perusahaannya dengan memberikan pinjaman dan pengajuan cicilan kepada pelanggan. Keuntungan bagi 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 pelanggan tidak dapat melakukan pembayaran tepat waktu, melewati pembayaran, atau berhenti melakukan pembayaran. Sebagai pemberi pinjaman yang bertanggung jawab, Home Credit menilai kesesuaian dan keterjangkauan kredit konsumen untuk pelanggan. Penilaian mencakup beberapa faktor seperti profil risiko pelanggan individu, jenis produk yang dibeli dan karakteristik pinjaman (seperti ukuran, jatuh tempo, dan faktor lainnya)



• Business Success Criteria

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 pelanggan.

CREDIT

Situation Assessment

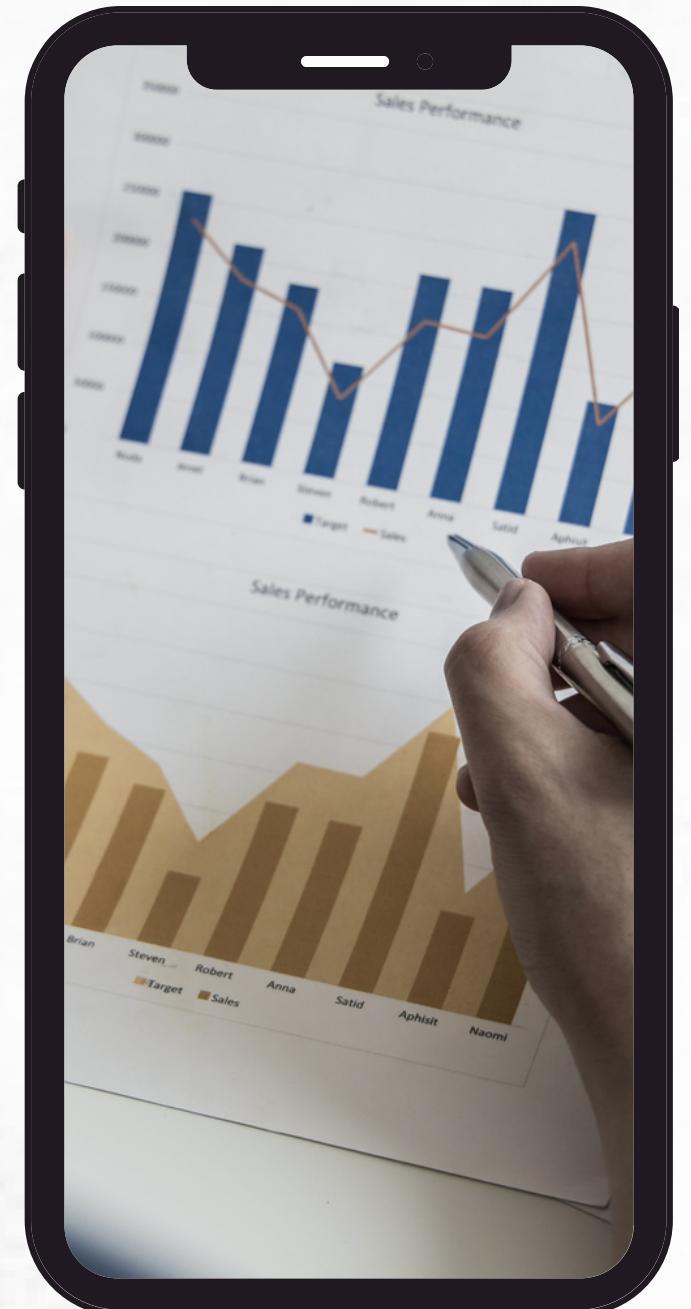
- **Assumption & Constraint**

Asumsi : Asumsi data yang akan digunakan adalah nama pelanggan, jumlah pinjaman, tujuan pinjaman, lama pinjaman, tanggal pinjaman, dan tanggal jatuh tempo dan pengembalian, pekerjaan, riwayat peminjaman

Kendala : Penggerjaan proyek yang terpisah tiap tahapnya

- **Risks & Contingencies**

Risiko	Kontingen
Tidak mendapatkan hasil model yang memuaskan	Mengulangi langkah analisis dari awal, business understanding



• Terminology

1. Annual Percentage Rate (APR)
2. Collateral
3. Borrower
4. Credit Score
5. Installment Loan

Manfaat :

- Peningkatan penjualan
- Kepuasan pelanggan
- Transaksi aman
- Praktik penanganan uang yang lebih aman
- Catatan akuntansi yang jelas

• Costs & Benefits

Biaya pemrosesan kartu kredit meliputi:

1. Biaya penyiapan akun merchant
2. Biaya pemrosesan actual
3. Perangkat keras, seperti terminal dan sistem point-of-sale
4. Tanggung jawab penipuan dan langkah-langkah keamanan

Determine Data Mining Goal

- Data Mining Goal

1. Membuat model menggunakan data historis aplikasi pinjaman untuk memprediksi apakah peminjam akan dapat membayar kembali pinjaman atau tidak.
2. Mengetahui faktor-faktor yang menyebabkan peminjam tidak dapat mengembalikan pinjaman dalam tempo waktu yang telah ditentukan

- Data Mining Success Criteria

Tingkat akurasi prediktif dalam memprediksi apakah peminjam dapat membayar kembali pinjaman sebesar >95%

Produce Project Plan

- **Project Plan**

Tahapan	Durasi
Membuat Business Understanding	3 hari
Melakukan data understanding	4 hari
Melakukan data preparation	4 hari
Membuat model	3 hari
Membuat model evaluation	3 hari
Membuat dashboard	2 hari

Produce Project Plan

- Tools & Techniques

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 pelanggan. Pembuatan model klasifikasi ini menggunakan metode (Logistic regression, Random Forests, Multi Layer Perceptron.)

Model ini akan dibuat dengan bahasa pemrograman Python yang ada di Google Colab. Kemudian, model akan dievaluasi menggunakan confusion matrix. Confusion matrix ini berbentuk tabel matriks yang menggambarkan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui. Tabel matriks tersebut memberikan informasi tentang True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Hal ini sangat berguna sebab hasil klasifikasi umumnya tidak dapat diekspresikan dengan baik dalam satu angka saja. 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 pelanggan. Hasil analisis dari model tersebut akan ditampilkan sebagai dashboard menggunakan Google Data Studio.



Data Understanding

COLLECT INITIAL DATA

Data yang digunakan untuk analisis lebih lanjut sistem home credit default risk berasal dari nasabah yang melakukan peminjaman ke home credit group. Data itu termasuk ke dalam jenis data kualitatif.

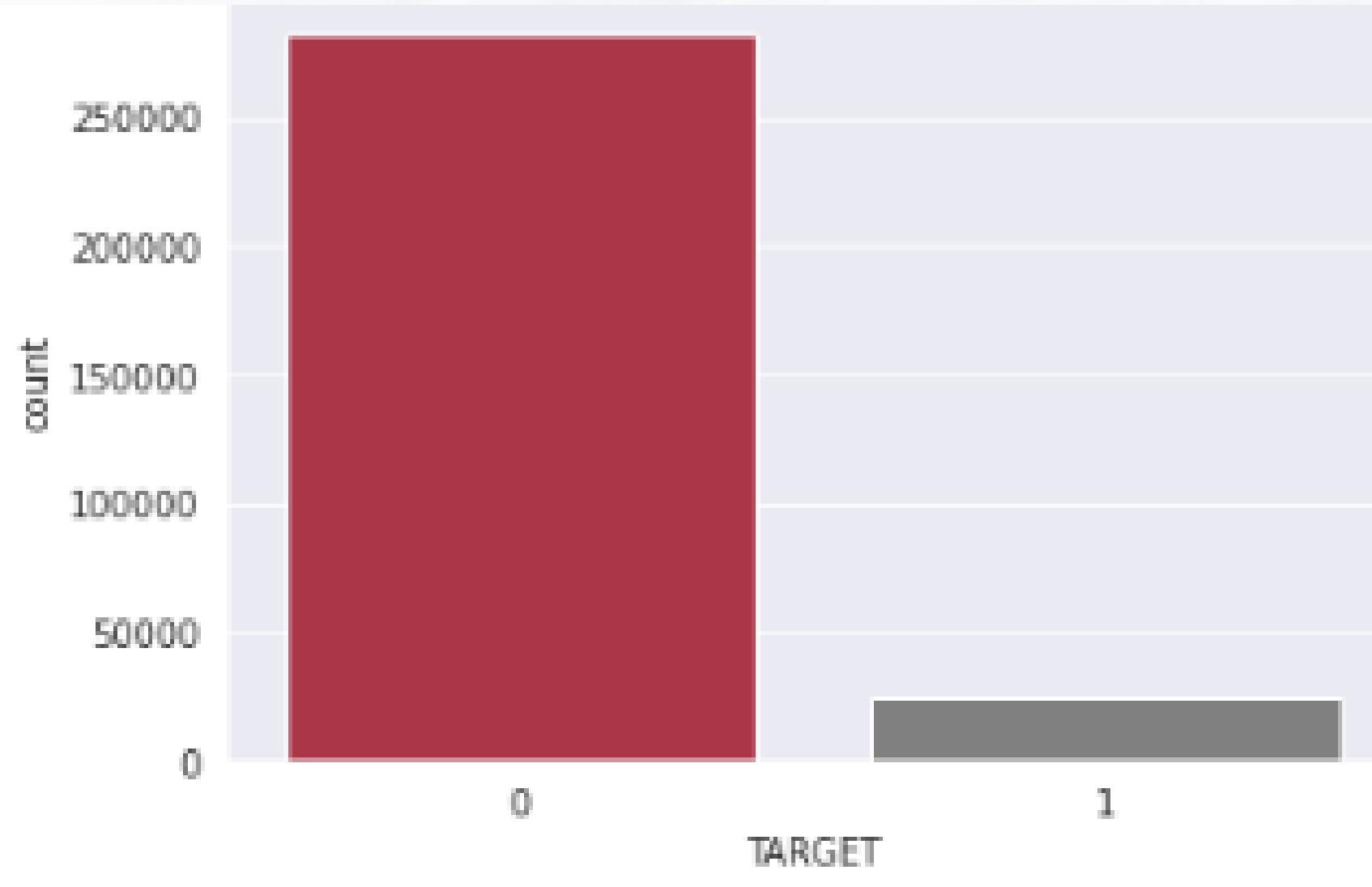
Data yang ada di home scoring default risk termasuk ke dalam jenis data statis, serta setiap basis yang ada di data ini mewakili satu pinjaman dari nasabah hal ini dapat dilihat dari riwayat data nasabah yang default maupun non default sehingga dari riwayat itu dapat diperoleh jumlah nasabah yang default maupun non default.

DESCRIBE DATA

Aplcation_(train|test).csv

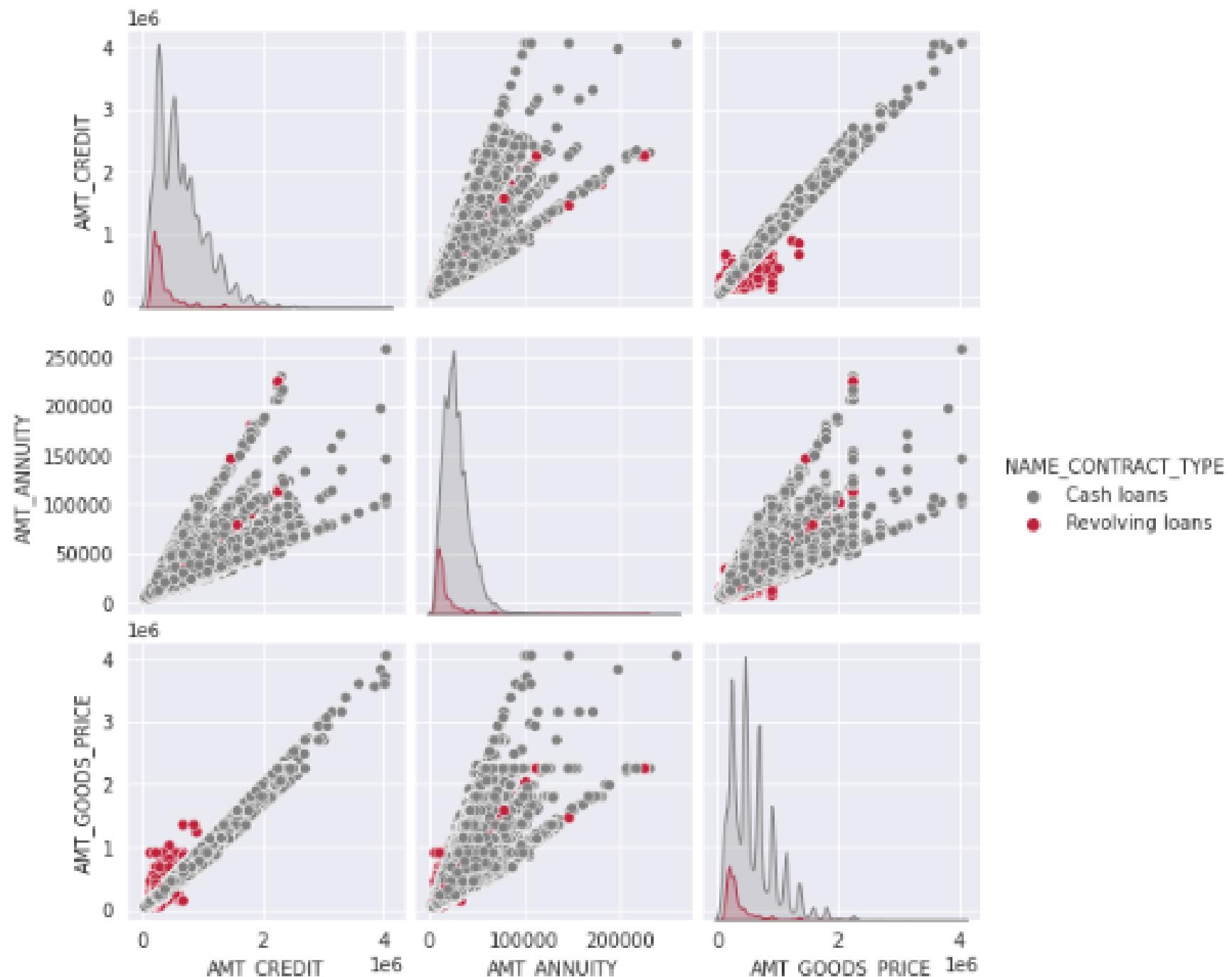
- main table yang dipecah menjadi data train dan data test
- data statis untuk seluruh aplikasi dimana satu basis mewakili satu pinjaman dalam sampel data

EXPLORATORY DATA



- Berdasarkan informasi di atas, dataset tidak seimbang. Hanya sekitar 24825 (8,0%) orang yang tidak melunasi pinjamannya
- $\text{TARGET} == 0 \rightarrow$ individu yang membayar pinjaman mereka
- $\text{TARGET} == 1 \rightarrow$ individu yang TIDAK melunasi pinjamannya

EXPLORATORY DATA

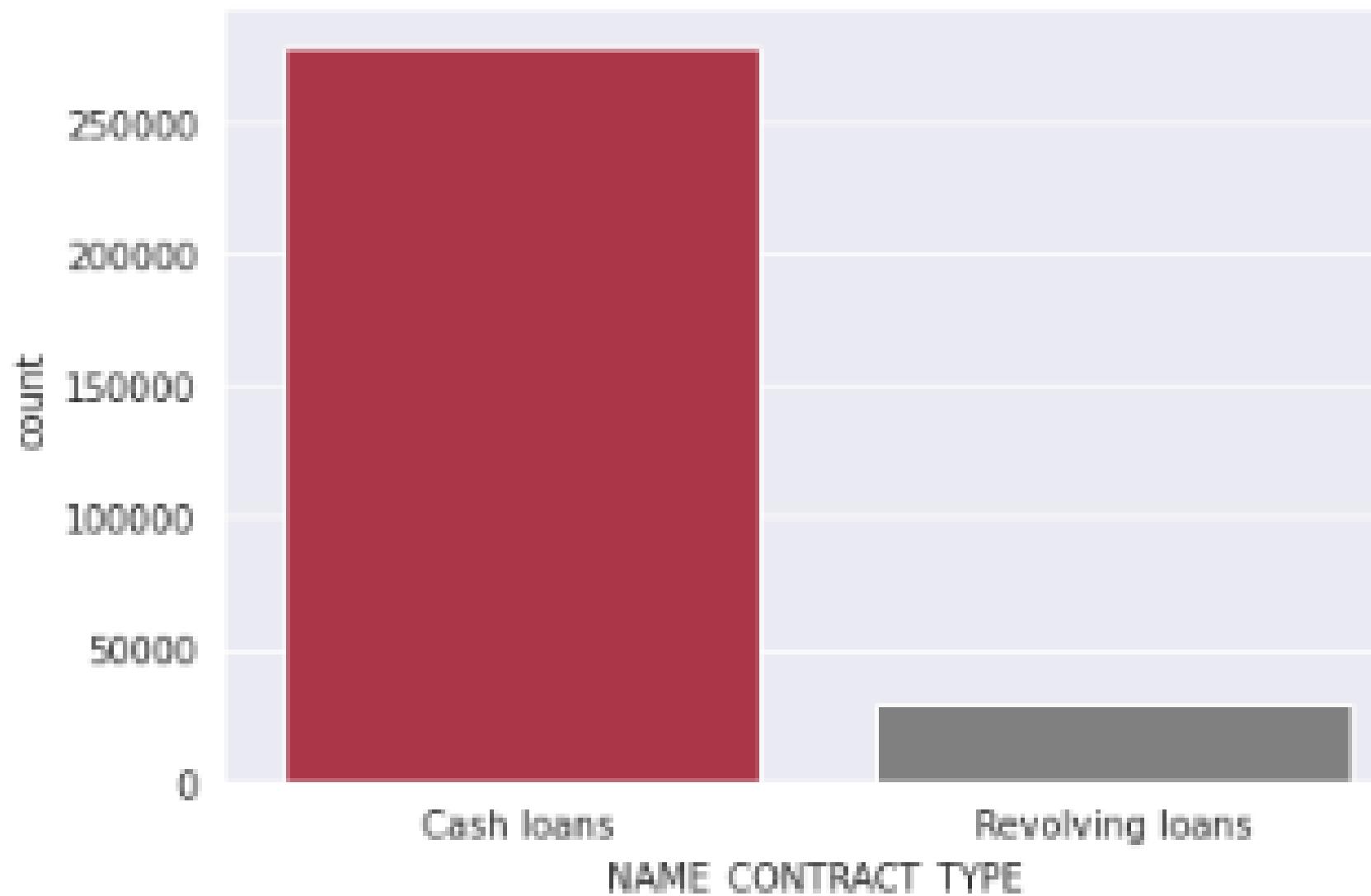


- correlation antara of credit amount dan price of good yaitu 0.99
- correlation antara annuity amount dan credit amount yaitu 0.77
- correlation antara annuity amount dan price of good yaitu 0.78
- AMT_CREDIT dan AMT_GOODS_PRICE berkorelasi tinggi (skor 0,99), dan memiliki kemiringan linier positif - yang masuk akal karena karena harga barang yang diberikan pinjaman semakin tinggi, jumlah kredit pinjaman juga semakin tinggi.
- AMT_ANNUITY juga sangat berkorelasi dengan AMT_CREDIT dan AMT_GOODS_PRICE dengan linearitas positif. Itu karena anuitas adalah jumlah jatuh tempo bulanan

EXPLORATORY DATA

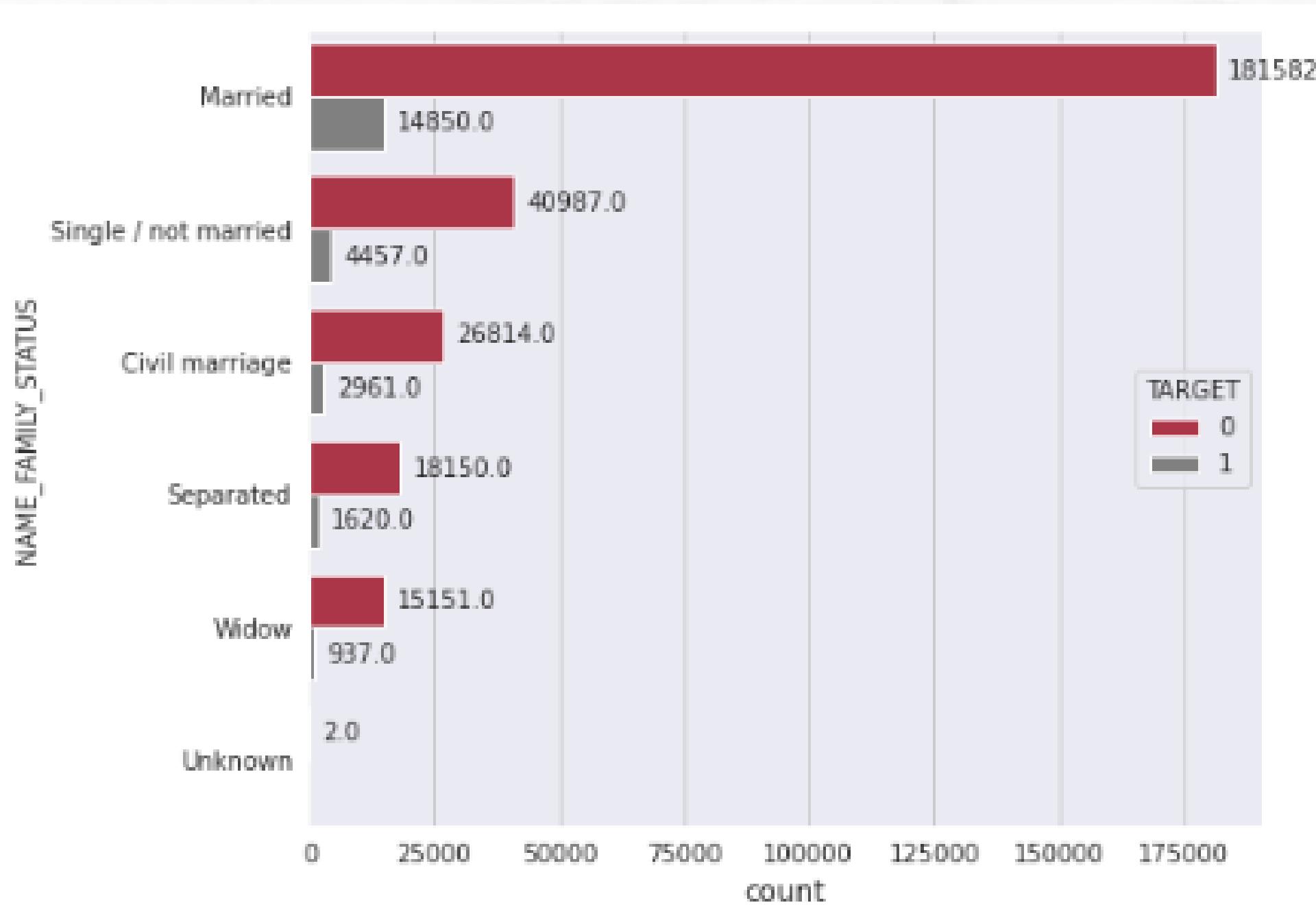
Percentage of defaulted cash loan: 8.35 %

Percentage of defaulted revolving loan: 5.48 %



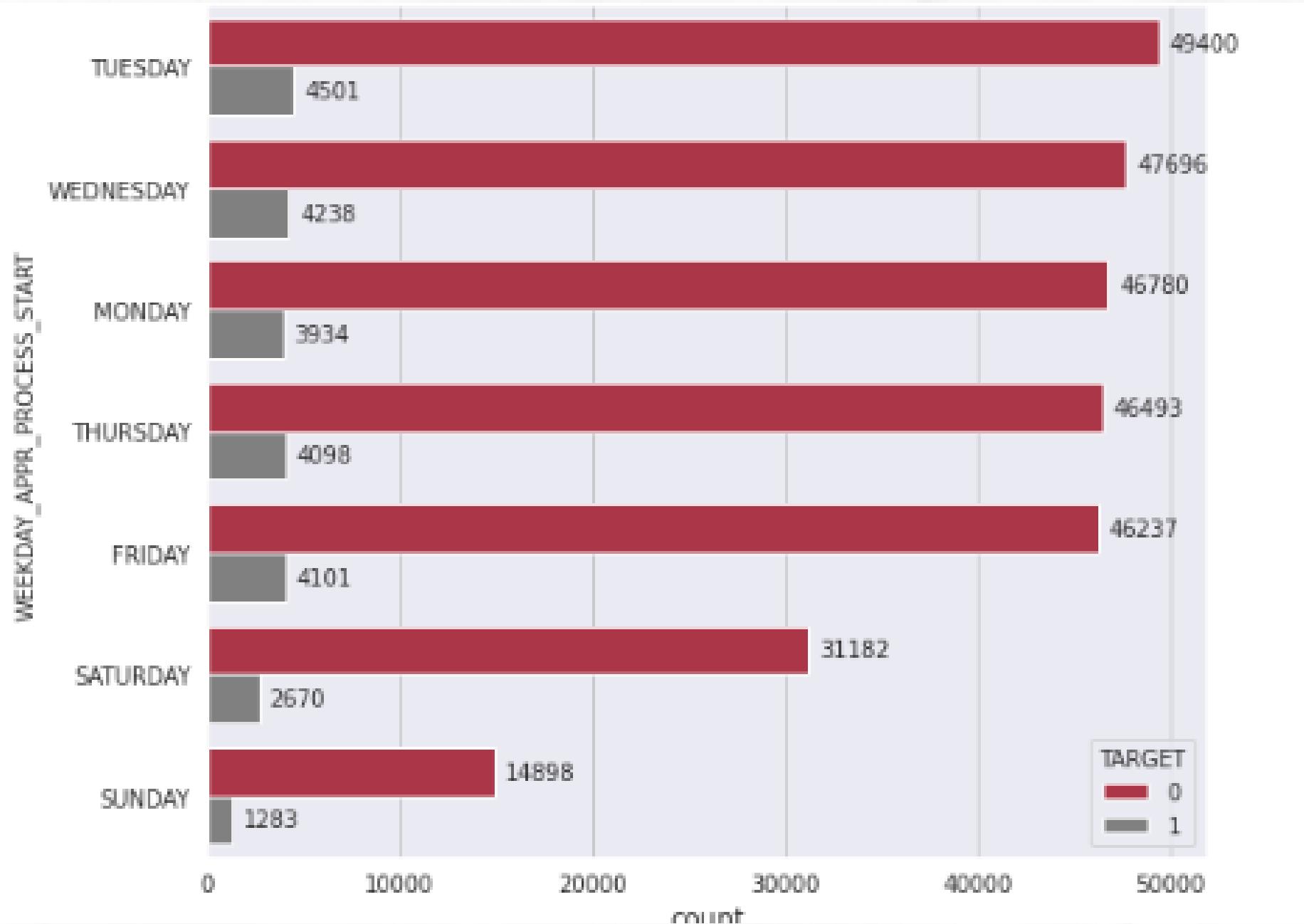
- gagal bayar jauh lebih besar dalam hal pinjaman tunai dibandingkan dengan pinjaman bergulir, namun, kita harus mencatat bahwa pinjaman tunai secara signifikan lebih populer bagi konsumen

EXPLORATORY DATA



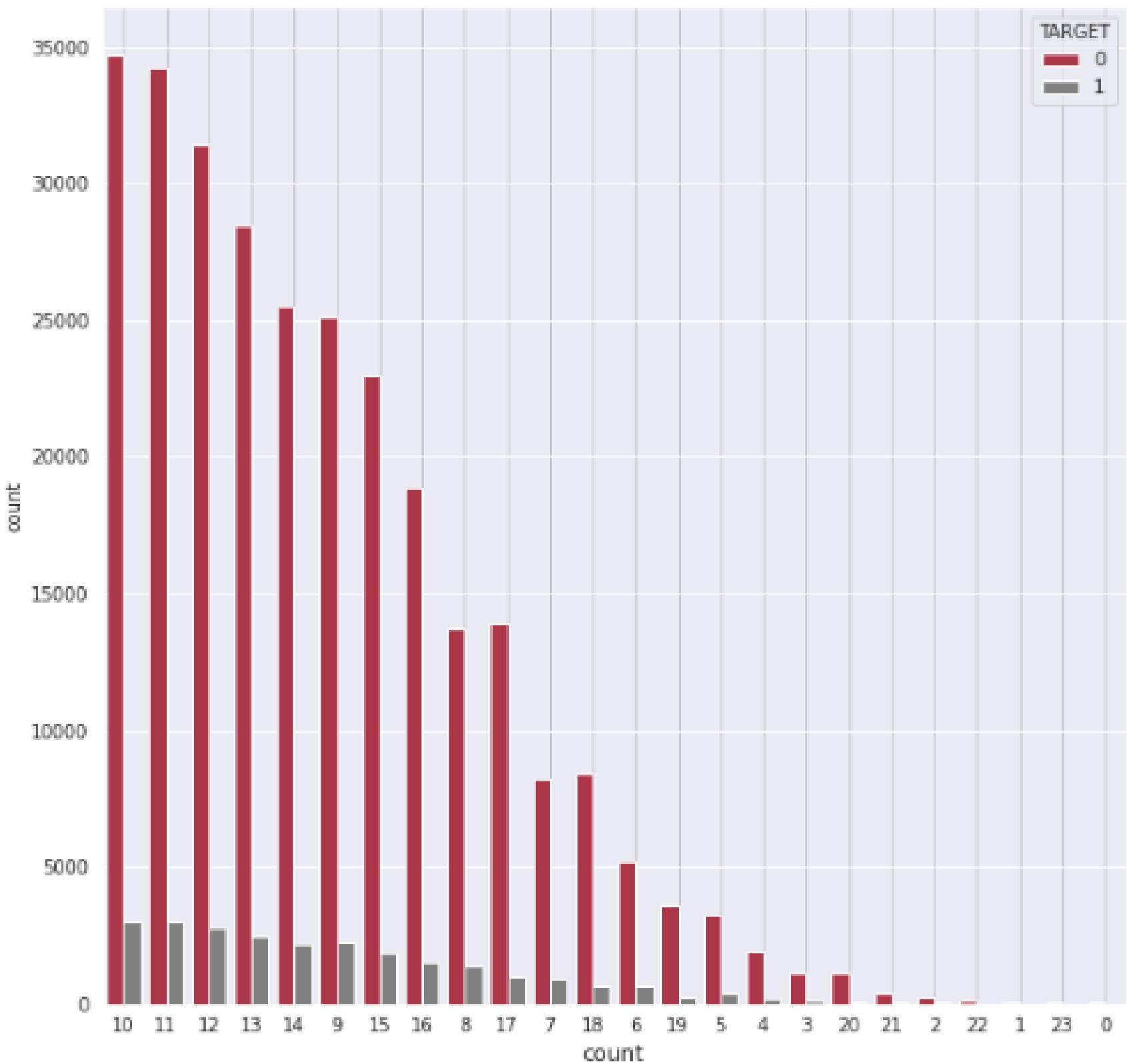
- pada data disamping ini menujukan status famili seseorang. orang yang menikah paling sering gagal bayar dibanding yang lain

EXPLORATORY DATA



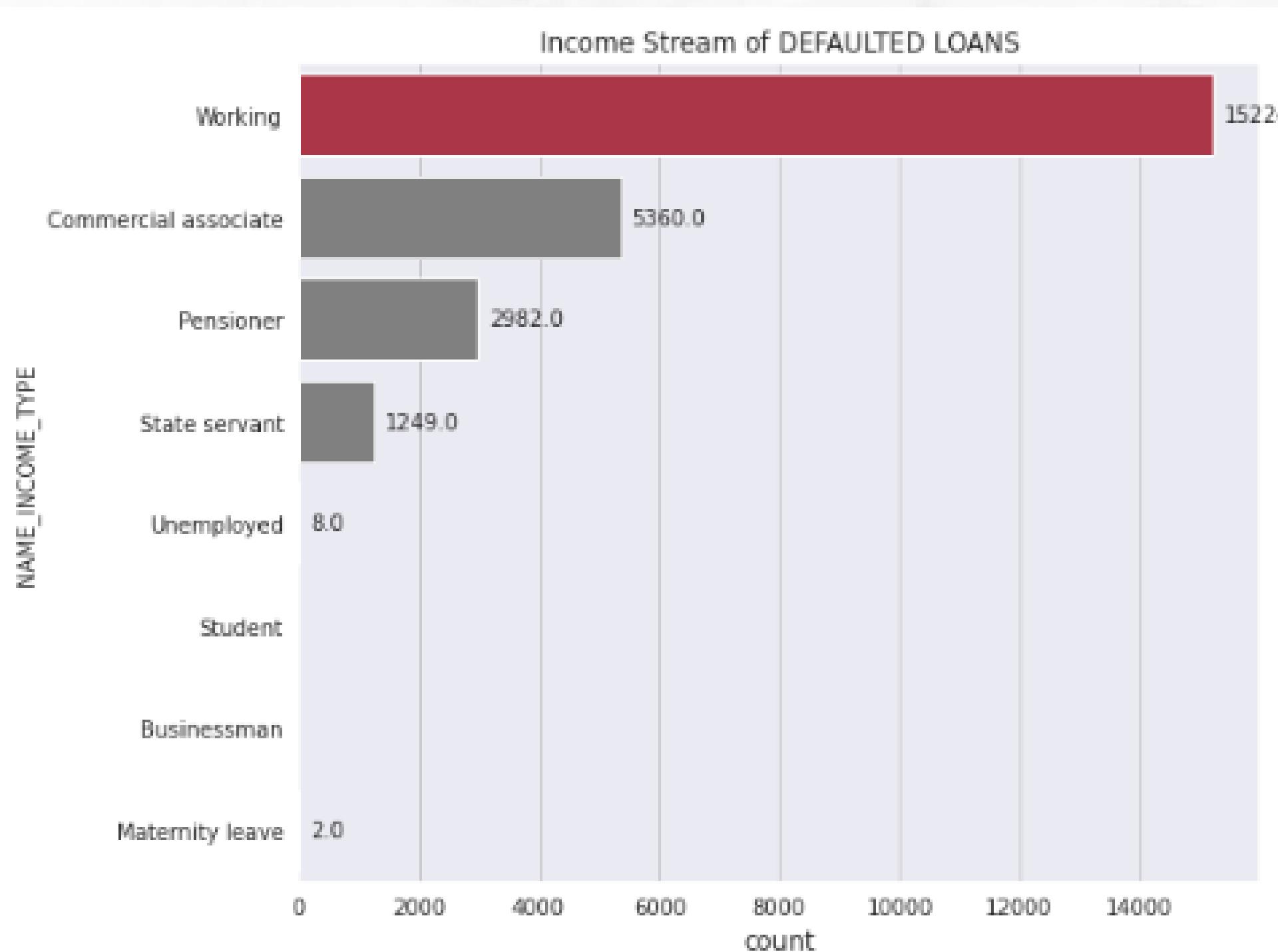
- Mayoritas pelanggan mendaftar pinjaman selama hari kerja, yaitu terbanyak pada hari kamis dan yang beberapa di akhir pekan. Kecenderungan nasabah yang tidak mampu membayar kembali pinjaman serupa dengan mereka yang melakukannya
- .

EXPLORATORY DATA



- pada data disamping ada yang mencurigakan yaitu, ada orang yang mengajukan pinjaman sejak jam 3 pagi, dan semakin padat sepanjang hari. mereka yang gagal dalam pinjaman mereka memiliki pola yang sama dengan mereka yang memiliki catatan bagus.

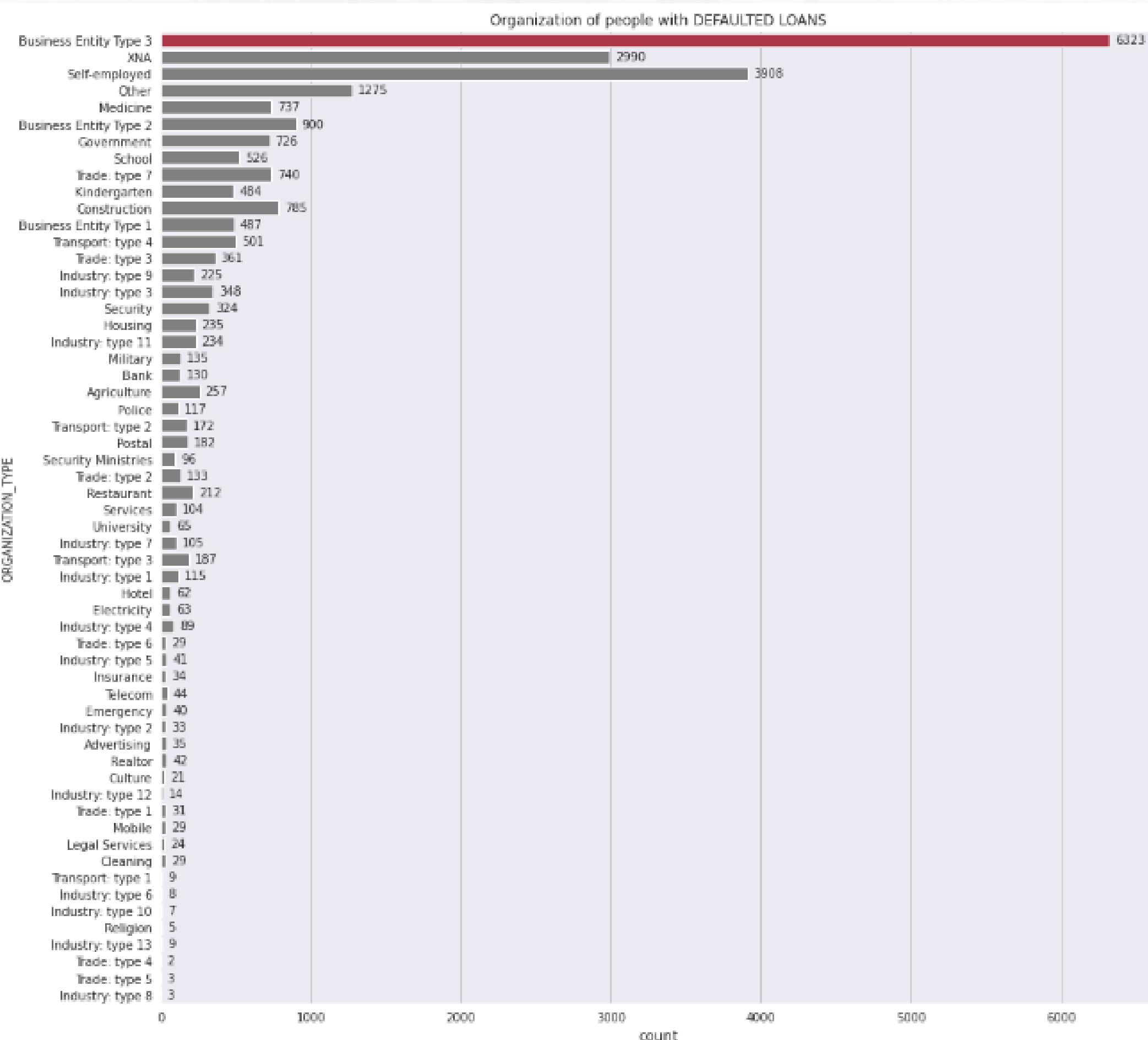
EXPLORATORY DATA



- Kategori 'bekerja' adalah yang paling padat dalam hal pelanggan gagal bayar tinggi dikarenakan upah rendah. dalam dataset ini juga memiliki sangat sedikit sampel tentang 'pengangguran', 'pelajar', 'cuti melahirkan' . kategori 'pengusaha' memiliki total pendapatan di atas rata-rata dan memiliki peluang besar untuk mempertahankan skor kredit yang baik.

EXPLORATORY DATA

...



- ORGANIZATION_TYPE cukup beragam mengenai di mana pelanggan ini bekerja. Namun berdasarkan histogram, kategori yang dominan mengalami default atau gagal bayar adalah tipe badan usaha 3, wiraswasta, XNA



Data Preparation

Data Wrangling

- Complete the null values of the following features:
 - * 'EXT_SOURCE_1'
 - * 'EXT_SOURCE_2'
 - * 'EXT_SOURCE_3'

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   EXT_SOURCE_1 307511 non-null  float64
 1   EXT_SOURCE_2 307511 non-null  float64
 2   EXT_SOURCE_3 307511 non-null  float64
dtypes: float64(3)
memory usage: 7.0 MB
None

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48744 entries, 0 to 48743
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   EXT_SOURCE_1 48744 non-null  float64
 1   EXT_SOURCE_2 48744 non-null  float64
 2   EXT_SOURCE_3 48744 non-null  float64
dtypes: float64(3)
memory usage: 1.1 MB
None
```

Complete the null values for 'CNT_FAM_MEMBERS'.

Data Wrangling

Convert the categorical text columns to numerical ones for:

- * CODE_GENDER
- * NAME_EDUCATION_TYPE
- * ORGANIZATION_TYPE

	CODE_GENDER	NAME_EDUCATION_TYPE	ORGANIZATION_TYPE
0	1	2	6
1	2	4	40
2	1	2	12
	CODE_GENDER	NAME_EDUCATION_TYPE	ORGANIZATION_TYPE
0	2	4	29
1	1	2	43
2	1	4	55

Feature Engineering

01. Creating New Column

FLAG_ASST
FLAG_CONTACTS
FLAG_DOCS
FLAG_ADDR
DAYS_BIRTH
ETC

Out[138]:

	FLAG_CONTACTS
0	4
1	3
2	3
3	4
4	4

Out[141]:

	FLAG_DOCS
0	1
1	1
2	1
3	1
4	1

Out[144]:

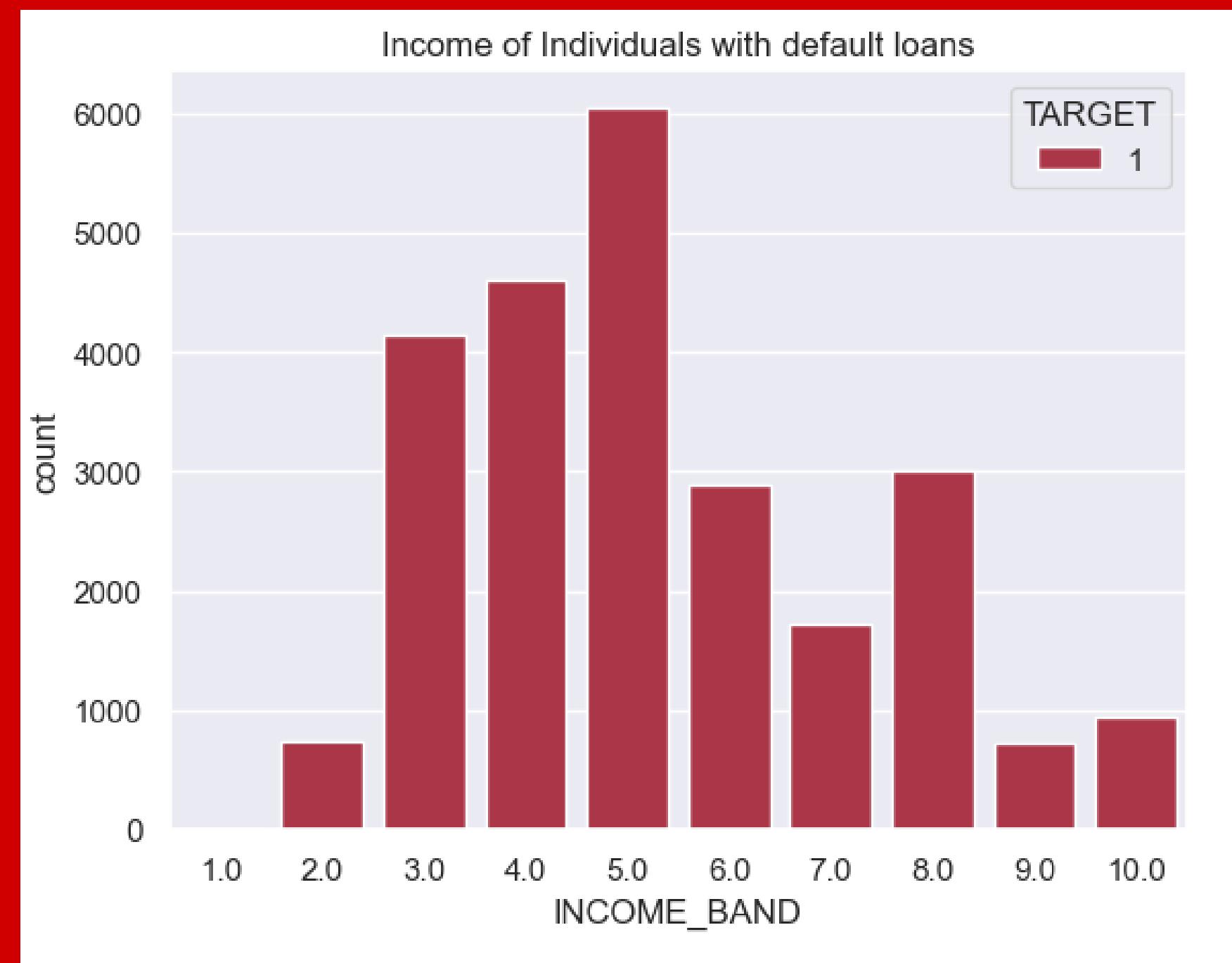
	FLAG_ADDR
0	0
1	0
2	0
3	0
4	1

Feature Engineering

02.

Creating Income Band

from using
data
AMT_INCOME_
TOTAL



Feature Engineering

03. Removing unnecessary column and merge

Since we have new feature(column), we can delete some unnecessary column that useless for our analysis

from 122 columns -> 21 columns

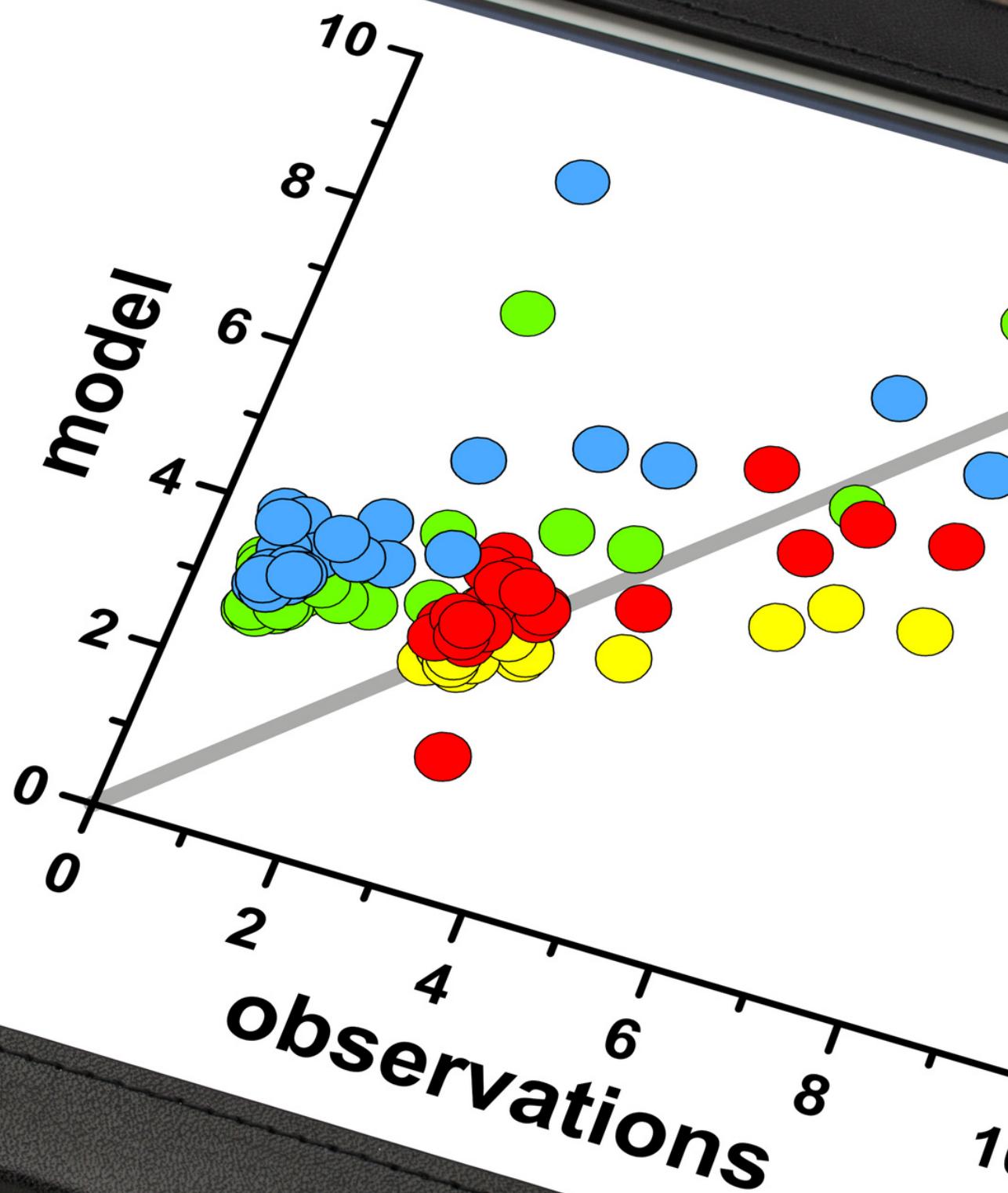
Out[180]:

	TARGET	CODE_GENDER	NAME_EDUCATION_TYPE	REGION_POPULATION_RELATIVE	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	ORGANIZATION_
0	1	1	2	0.018801	1.0	2	
1	0	2	4	0.003541	2.0	1	
2	0	1	2	0.010032	1.0	2	
3	0	2	2	0.008019	2.0	2	
4	0	1	2	0.028663	1.0	2	
...
307506	0	1	2	0.032561	1.0	1	
307507	0	2	2	0.025164	1.0	2	
307508	0	2	4	0.005002	1.0	3	
307509	1	2	2	0.005313	2.0	2	
307510	0	2	4	0.046220	2.0	1	

307511 rows × 21 columns

...

Modeling dan Evaluasi



Model Machine Learning

01. Light GBM

kerangka kerja
peningkatan gradien
berdasarkan pohon
keputusan untuk
meningkatkan efisiensi
model dan mengurangi
penggunaan memori.

02. Logistic Regression

Algoritma machine learning untuk menemukan hubungan antara dua faktor data,
Prediksi terbatas seperti ya atau tidak

03. Decision Tree

algoritma machine learning yang menggunakan seperangkat aturan untuk membuat keputusan dengan struktur seperti pohon

04. Random Forest

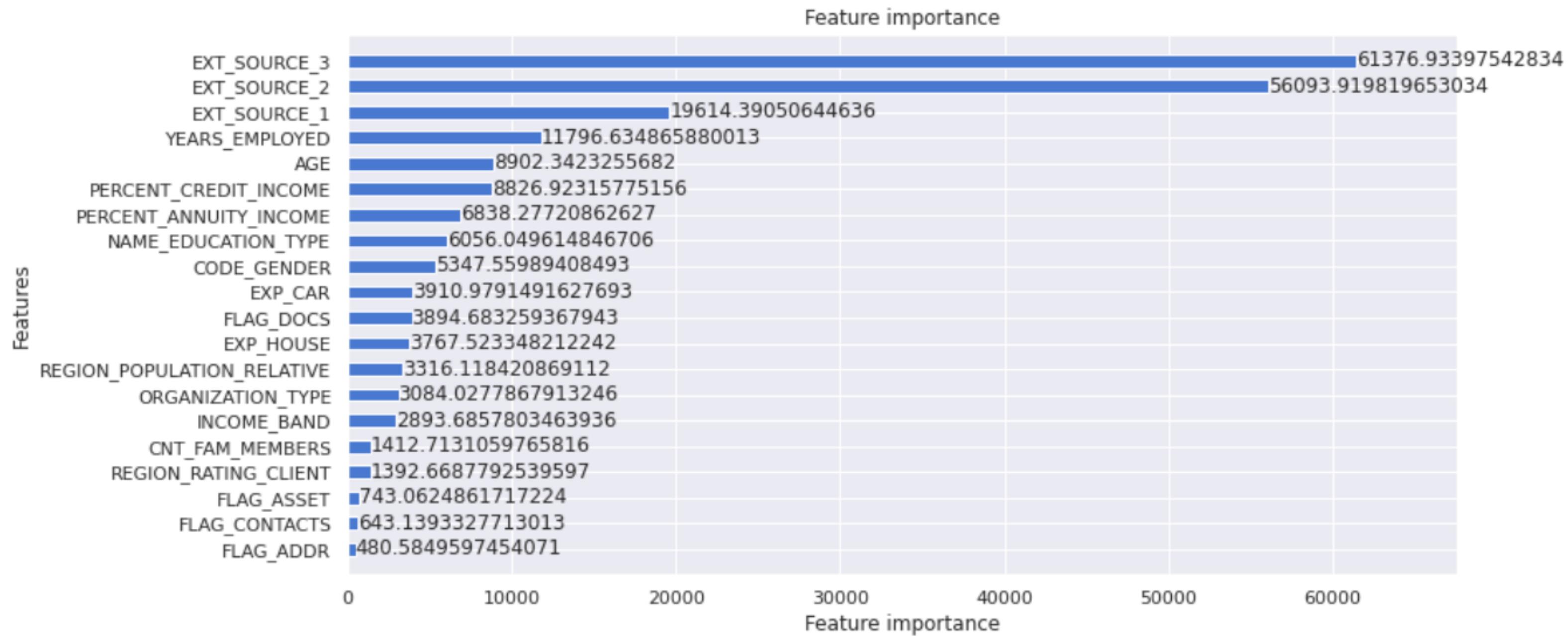
kombinasi dari beberapa tree predictors / decision trees dimana setiap tree bergantung pada nilai random vector yang dijadikan sampel secara bebas dan merata pada semua tree dalam forest tersebut

05. XGBoost

algoritma gradient boost decision tree

•••

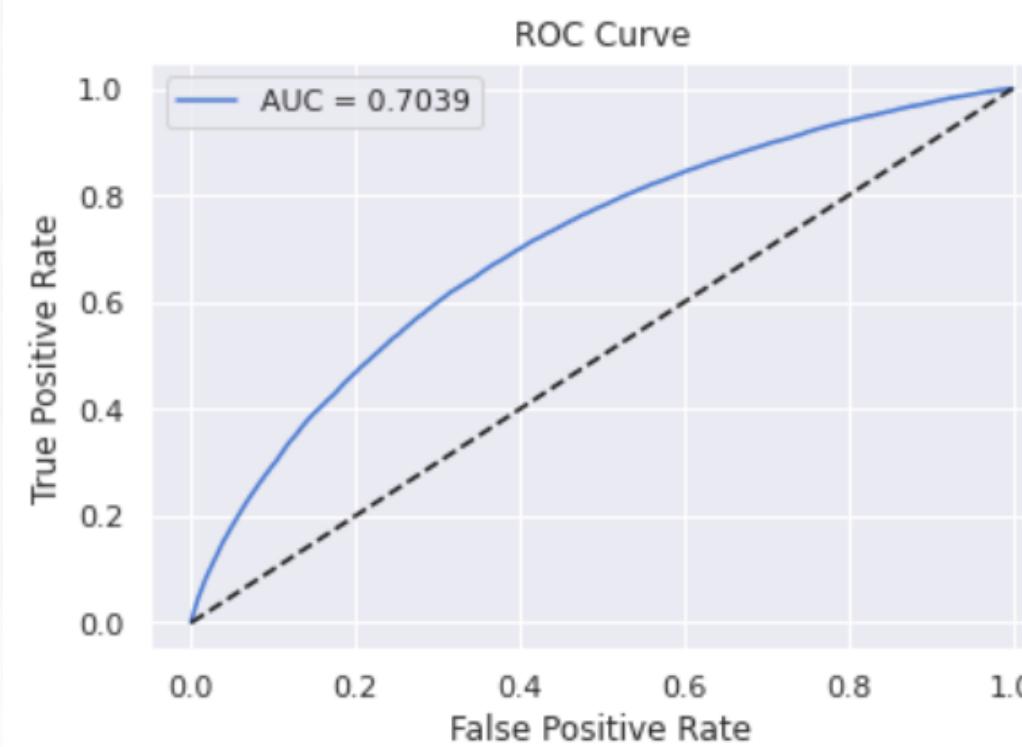
Feature Importance



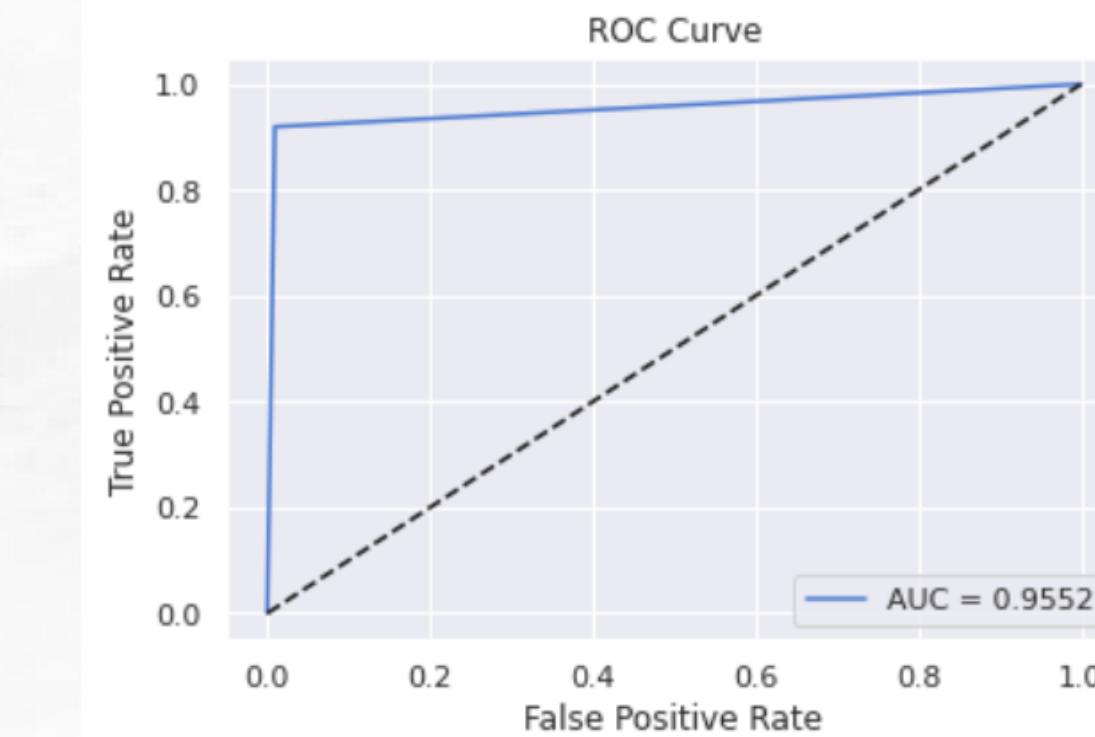
Confusion Matrix

Selected Features	Model	Class	Precision	Recall	F1-score
['CODE_GENDER', 'NAME_EDUCATION_TYPE', 'REGION_POPULATION_RELATIVE', 'CNT_FAM_MEMBERS', 'REGION_RATING_CLIENT', 'ORGANIZATION_TYPE', 'EXT_SOURCE_1', 'EXT_SOURCE_2', EXT_SOURCE_3', 'FLAG_ASSET', 'FLAG_CONTACTS', 'FLAG_DOCS', 'FLAG_ADDR', 'AGE', 'YEARS_EMPLOYED', 'INCOME_BAND', 'PERCENT_ANNUITY_INCOME', 'PERCENT_CREDIT_INCOME', 'EXP_CAR', 'EXP_HOUSE']	Logistic Regression	0	0.92	1.00	0.96
		1	0.25	0.00	0.00
	Decision Tree	0	0.93	0.91	0.92
		1	0.14	0.17	0.15
	Random Forest	0	0.92	1.00	0.96
		1	0.57	0.01	0.02
	XGB	0	0.92	1.00	0.96
		1	0.56	0.01	0.03

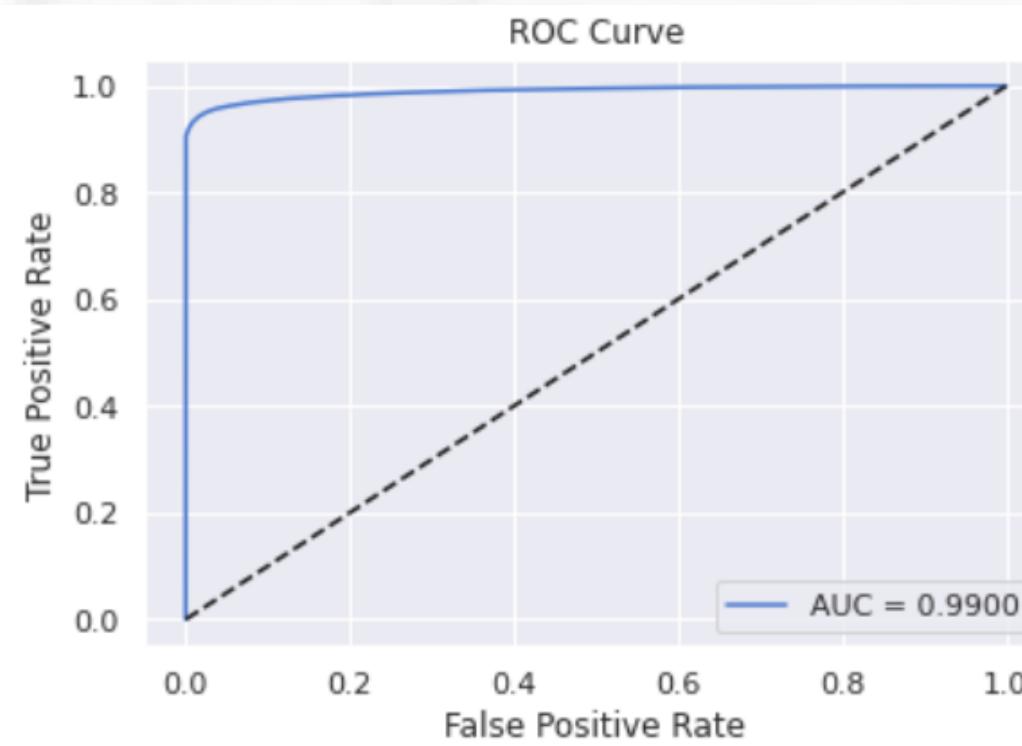
Kurva AUC



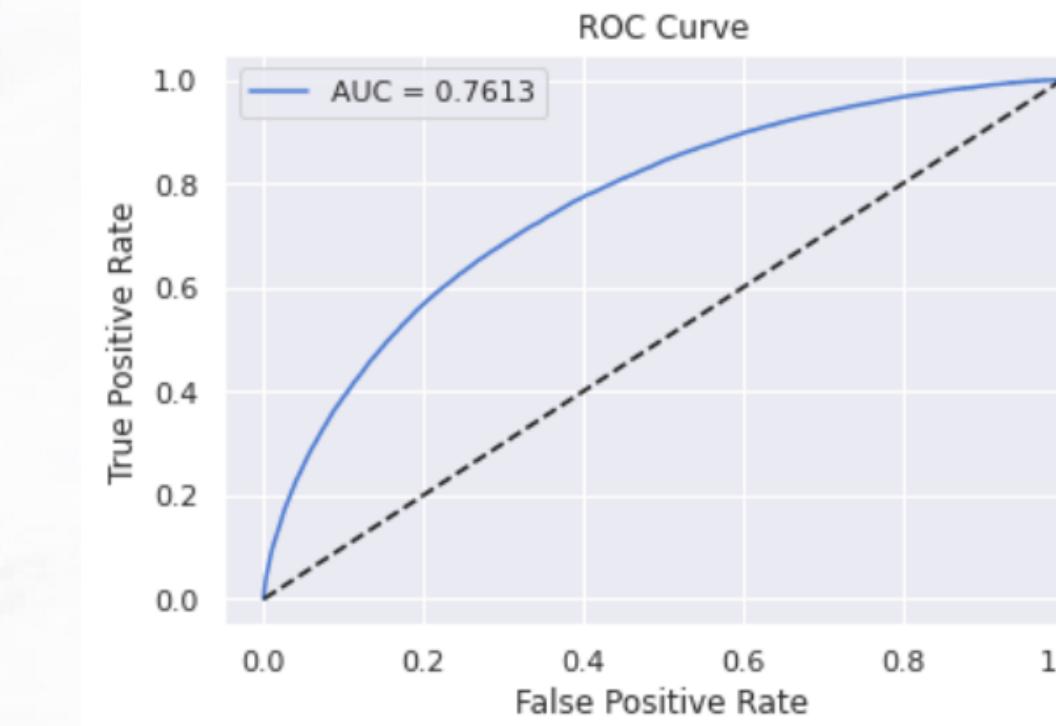
Logistic Regression



Decision Tree



Random Forest



XGB

Deployment





HOME CREDIT

DASHBOARD OF TOP FEATURES

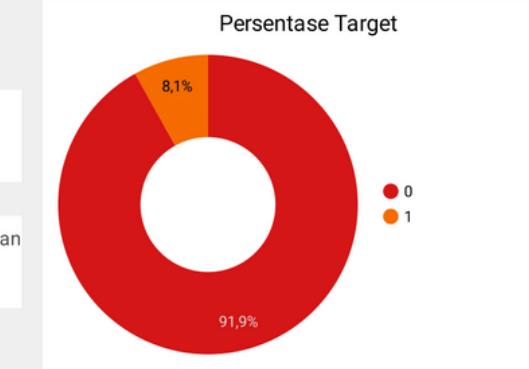
Client yang Bisa Membayar Pinjaman

91.904

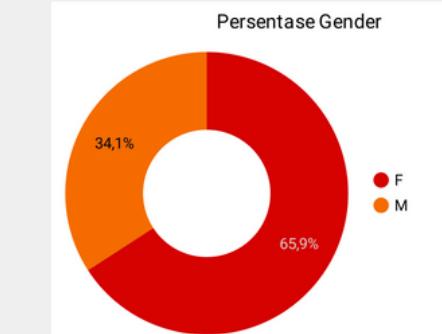
Klien yang Tidak Bisa Membayar Pinjaman

8.093

Percentase Target



Percentase Gender



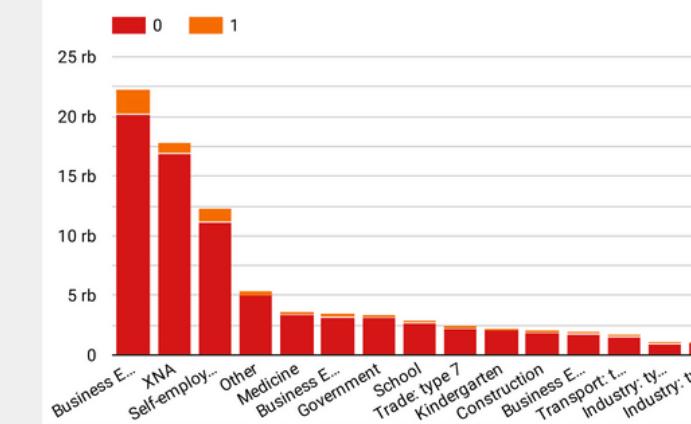
PERCENT_AN...

PERCENT_C...

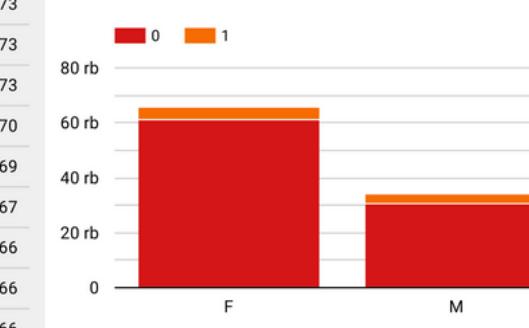
1.	2022-06-11	73
2.	2022-08-13	73
3.	2022-02-11	73
4.	2022-01-12	70
5.	2022-02-12	69
6.	2022-03-14	67
7.	2022-06-12	66
8.	2022-01-16	66
9.	2022-06-12	66

1 - 100 / 4905

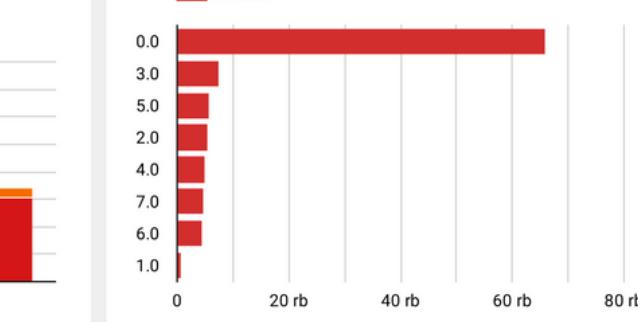
Tipe Organisasi Berdasarkan Target



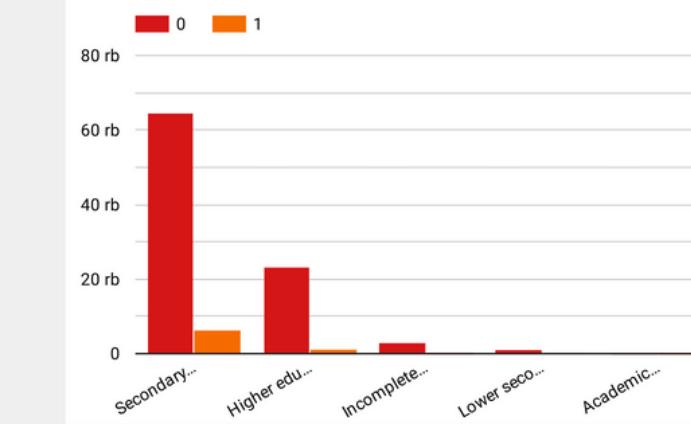
Jenis Kelamin Berdasarkan Target



TARGET



Nama Edukasi Berdasarkan Target





thank you!

