CREDIT RISK SCORING ANALYSIS

Kelompok 8

ANGGOTA

Zahra Avia

Ilmu Ekonomi

Universitas Padjadjaran

Dhiemas Ady

Teknik Informatika

UPNVYK

Vina

Ekonomi Pembangunan

Universitas Siliwangi

DAFTAR ISI

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

BUSINESS UNDERSTANDING

1. Determine Business Objectives
2. Determine Data Mining Goals
3. Produce Project Plan

Determine Business Objectives

Background	Banyak orang mengalami masalah dalam memperoleh pinjaman karena riwayat kredit yang tidak mencukupi. Perusahaan seperti Home Credit mencoba untuk memperluas inklusi keuangan untuk mereka yang mengalami kesulitan dengan memberikan pengalaman meminjam yang aman.
Business Objectives	Mampu memprediksi apakah peminjam dapat membayar pinjaman atau tidak, sehingga penyaluran pinjaman dapat tepat sasaran.
Business Success Criteria:	Business ini dikatakan sukses apabila berhasil menurunkan Non Performing Loan (NPL)

Data Mining Goals

Data Mining Goals

Mampu memprediksi apakah pemohon pinjaman dapat membayar atau gagal membayar pinjaman. Diprediksi dengan melakukan modeling dan deployment menggunakan data tipe klasifikasi.

Produce Project Plan

Project Plan

Waktu	Kegiatan
Senin, 5 Desember 2022	Business understanding
Selasa, 6 Desember 2022	Data understanding
Rabu, 7 Desember 2022	Data preparation
Kamis, 8 Desember 2022	Modeling
Jum'at, 9 Desember 2022	Evaluasi model
Sabtu, 10 Desember 2022	Evaluasi secara menyeluruh
Minggu, 11 Desember 2022	Deployment

Initial Asessment of Tools and Techniques

Initial Asessment of Tools and Techniques:

- Data understanding, data preparation, modeling, dan evaluasi dilakukan dengan menggunakan *google colab*.
- Deployment dilakukan dengan menggunakan *streamlit*.

Data Understanding

Sumber Data

Memeriksa duplikasi pada data

Melihat distribusi data

Memeriksa missing value

Memeriksa Outlier

Memeriksa anomali pada data

Exploratory Data Analysis (EDA)

Data

Home Credit Default Risk dari Kaggle dalam bentuk CSV.

Terdiri dari 7 dataset dengan **application_train** sebagai tabel utama.

Di dalam application_train terdapat 122 kolom, 307.511 baris, dan tidak terdapat duplikasi pada data.

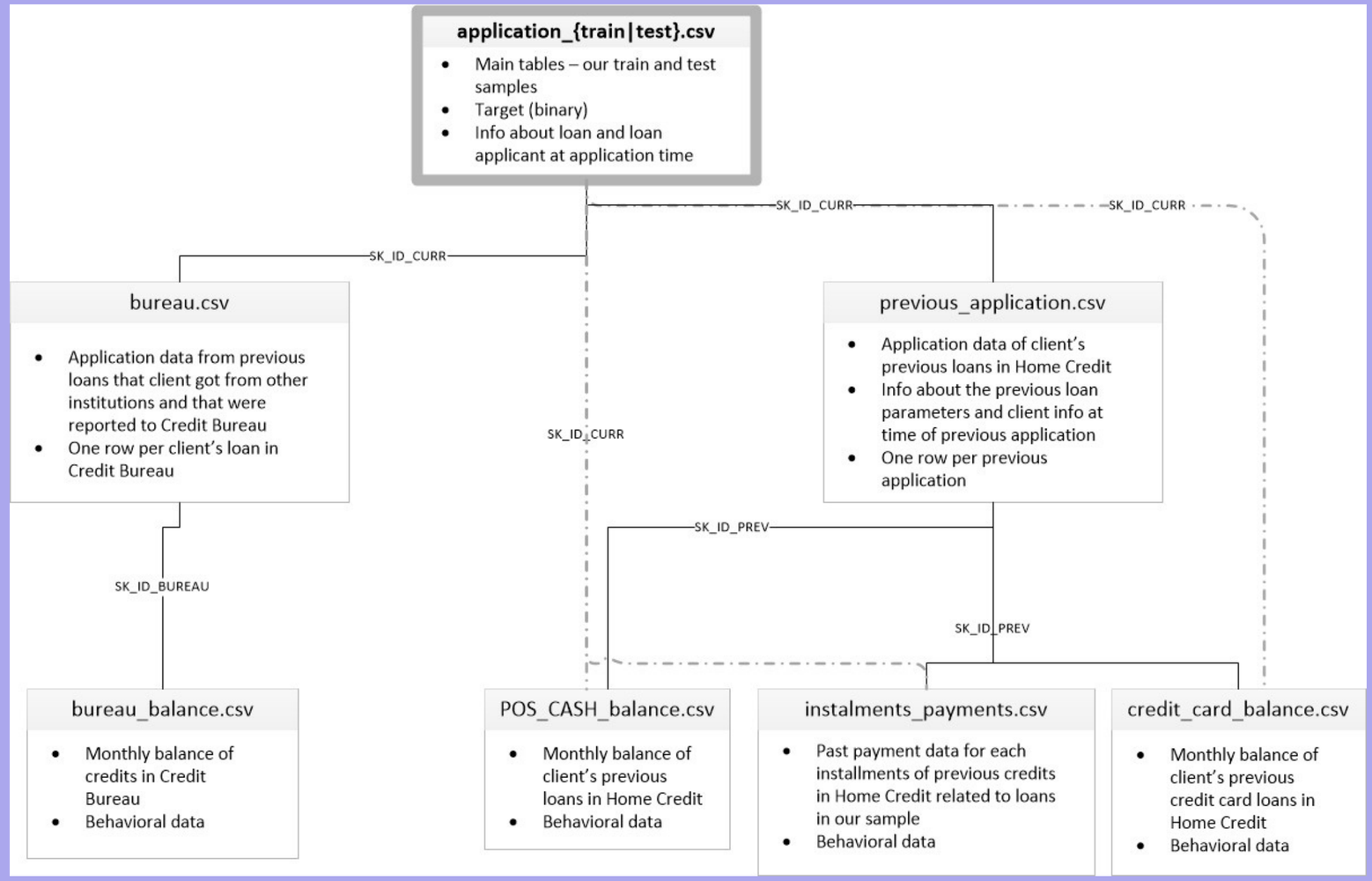
Tipe Data:

float64: 65 feature

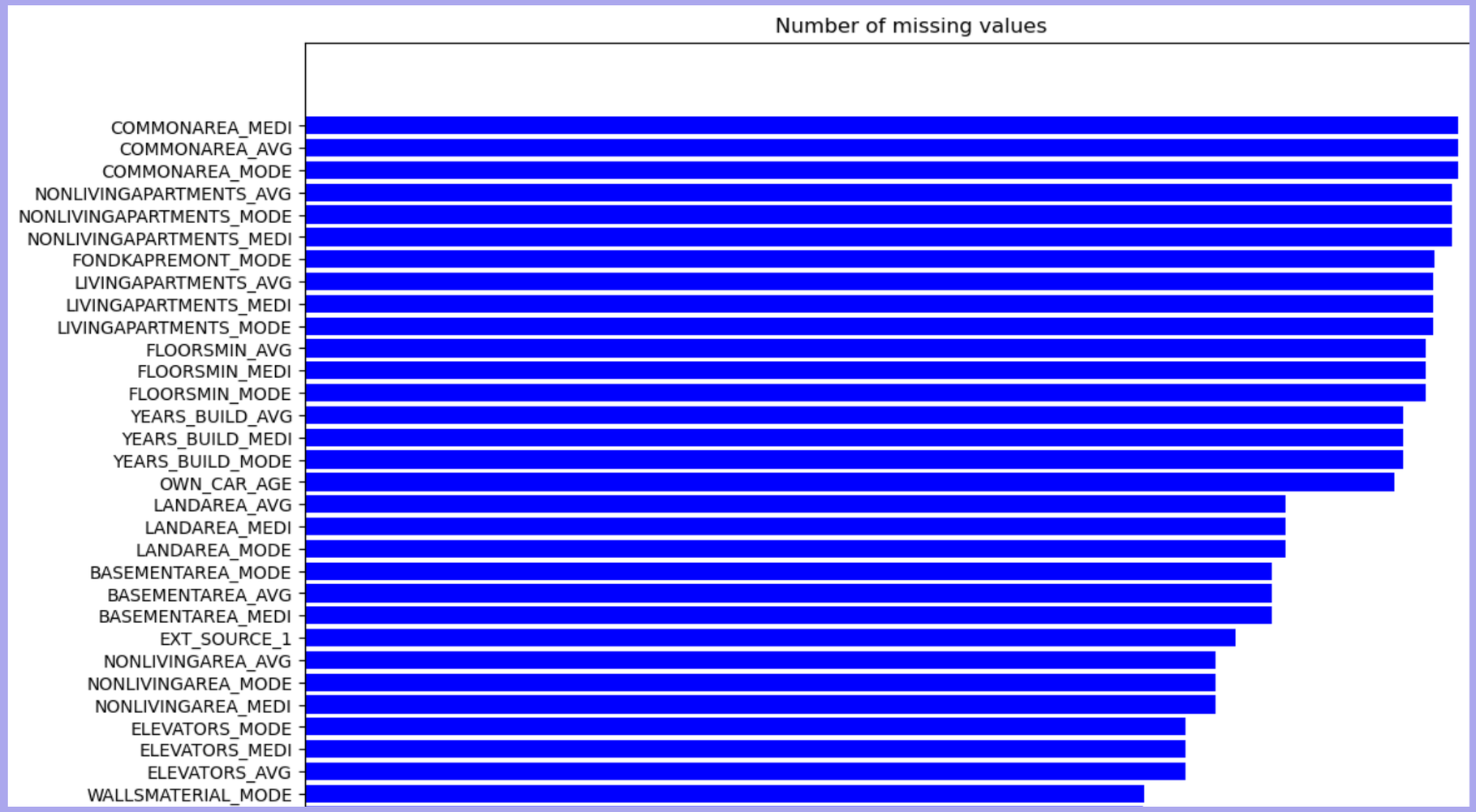
int64: 41 feature

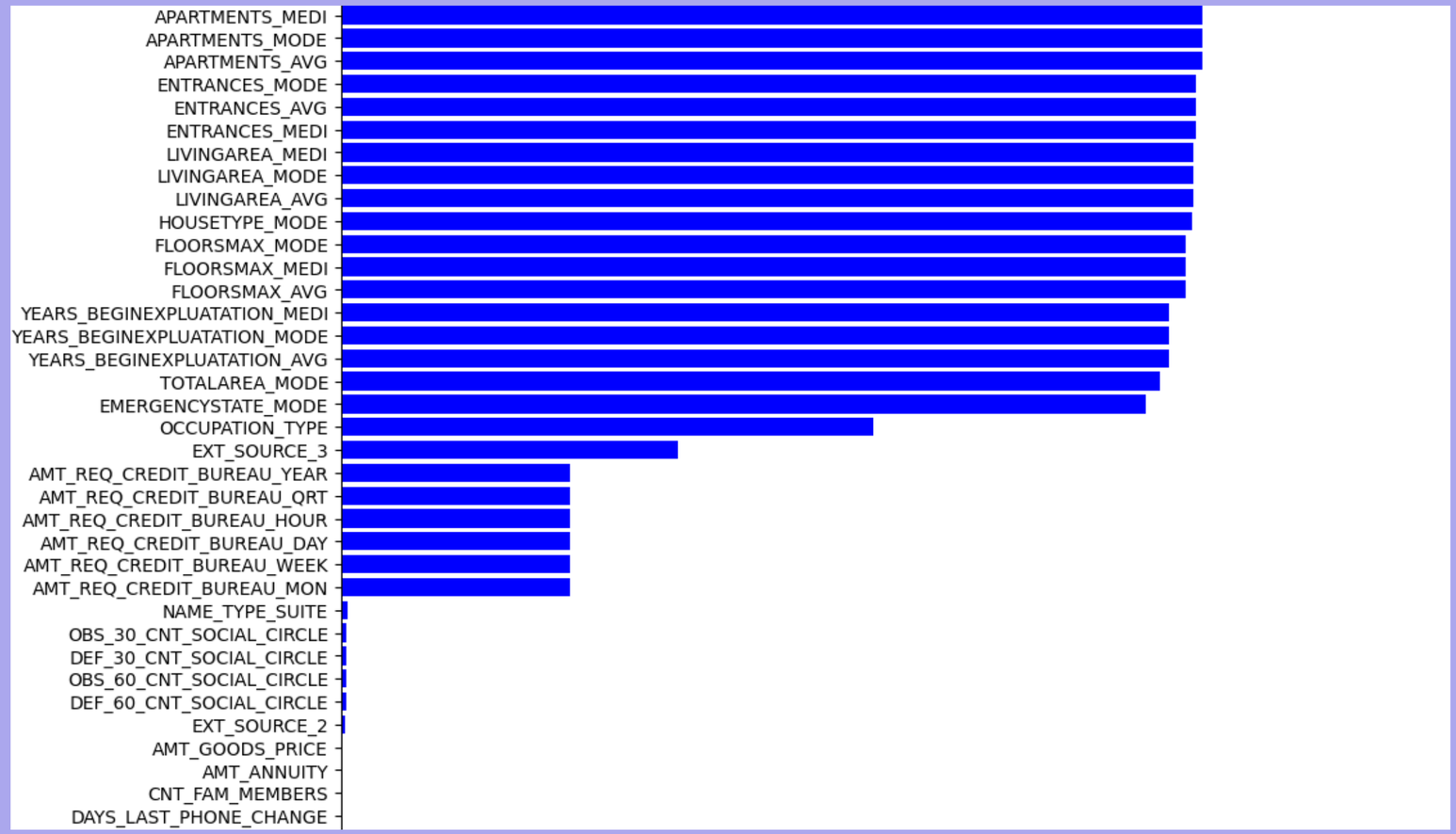
object: 16 feature

7 Dataset


application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

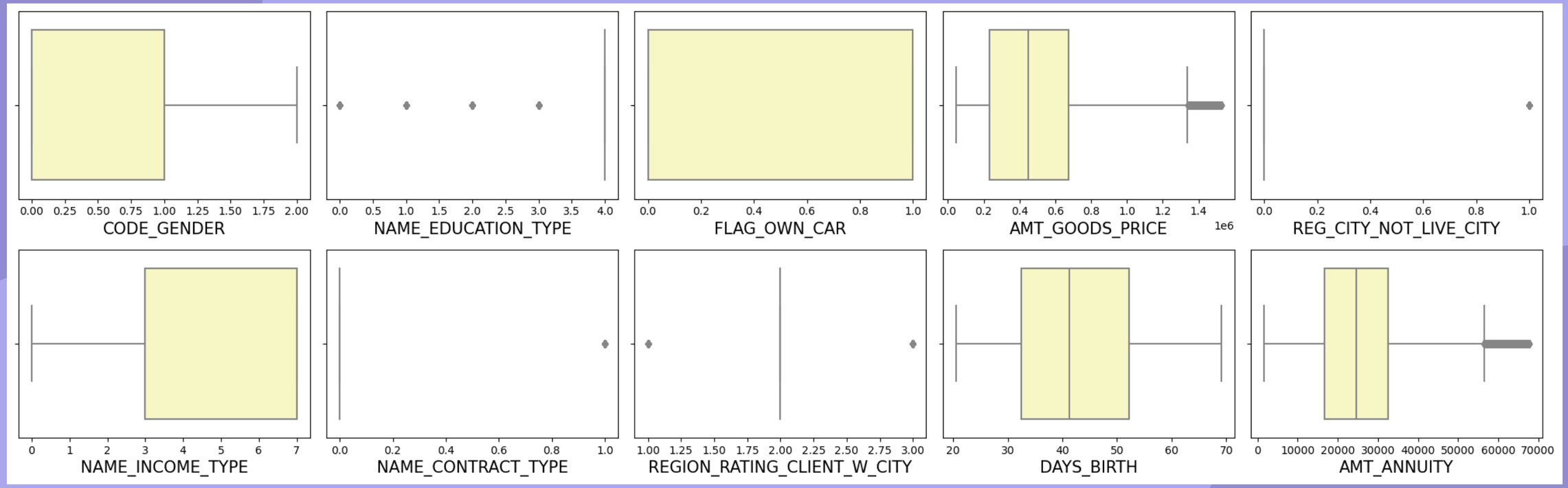
Missing Value





Terdapat *missing value* pada beberapa fitur sehingga nantinya perlu diatasi.

Outliers

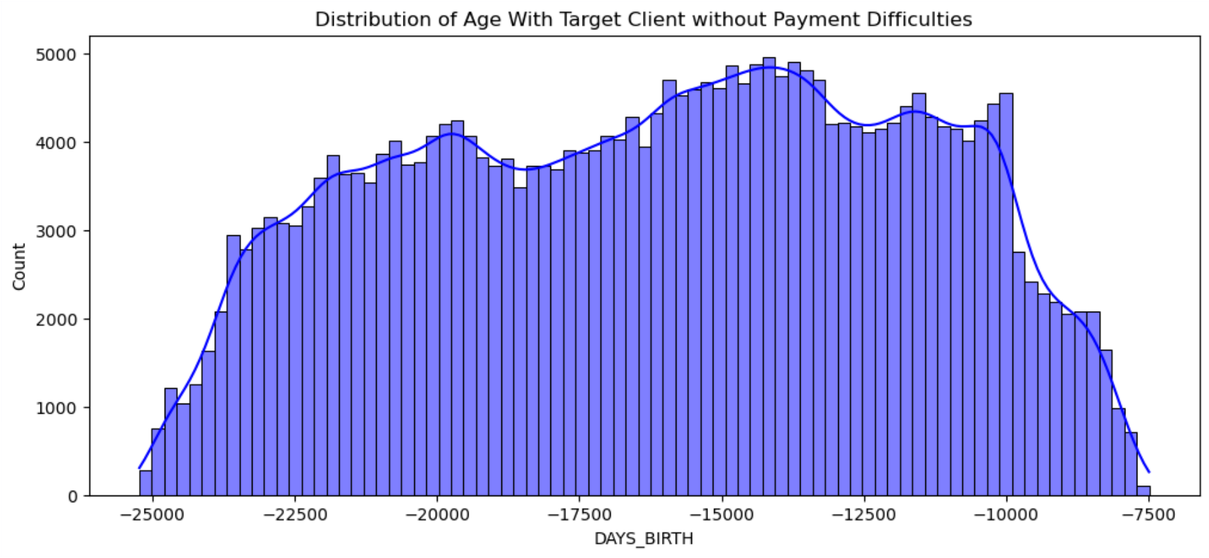


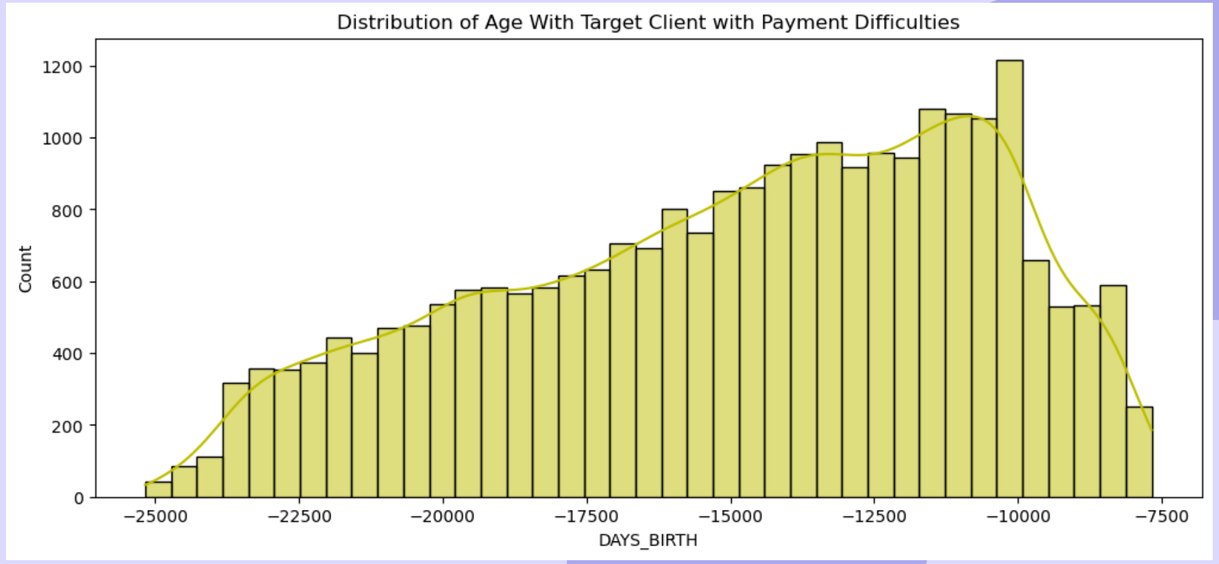
Terdapat *outliers* pada beberapa fitur sehingga nantinya perlu dihilangkan.

Data Anomali

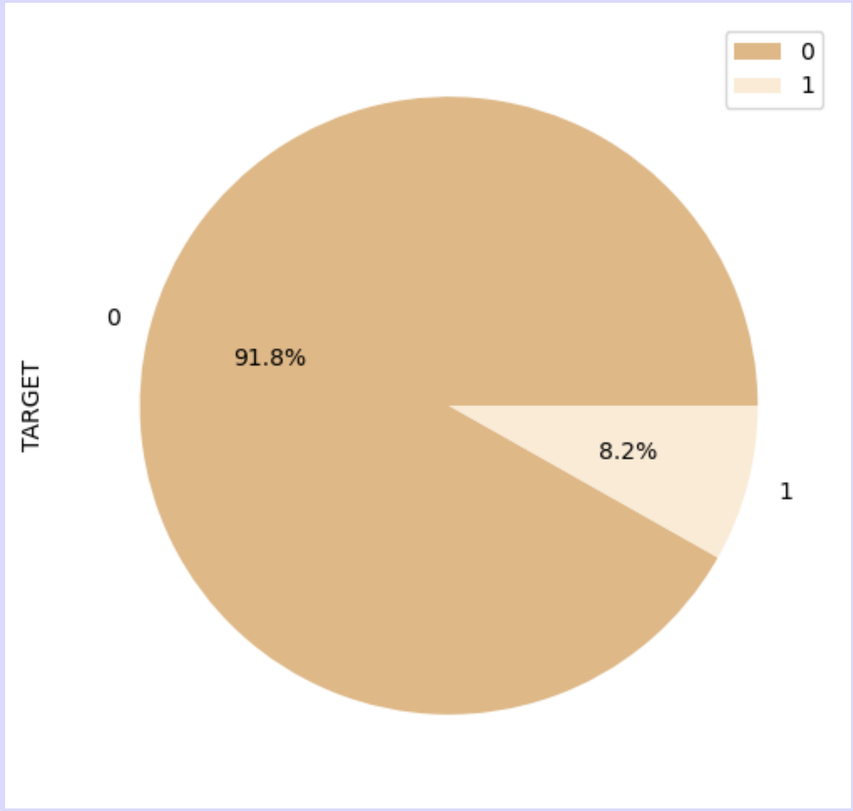
Terdapat anomali pada kolom "DAYS_BIRTH", "DAYS_REGISTRATION", "DAYS_ID_PUBLISH", ""DAYS_LAST_PHONE_CHANGE", dan "DAYS_EMPLOYED". Anomali yang kami temukan adalah data yang seharusnya bernilai positif tetapi bernilai negatif.

Contoh:

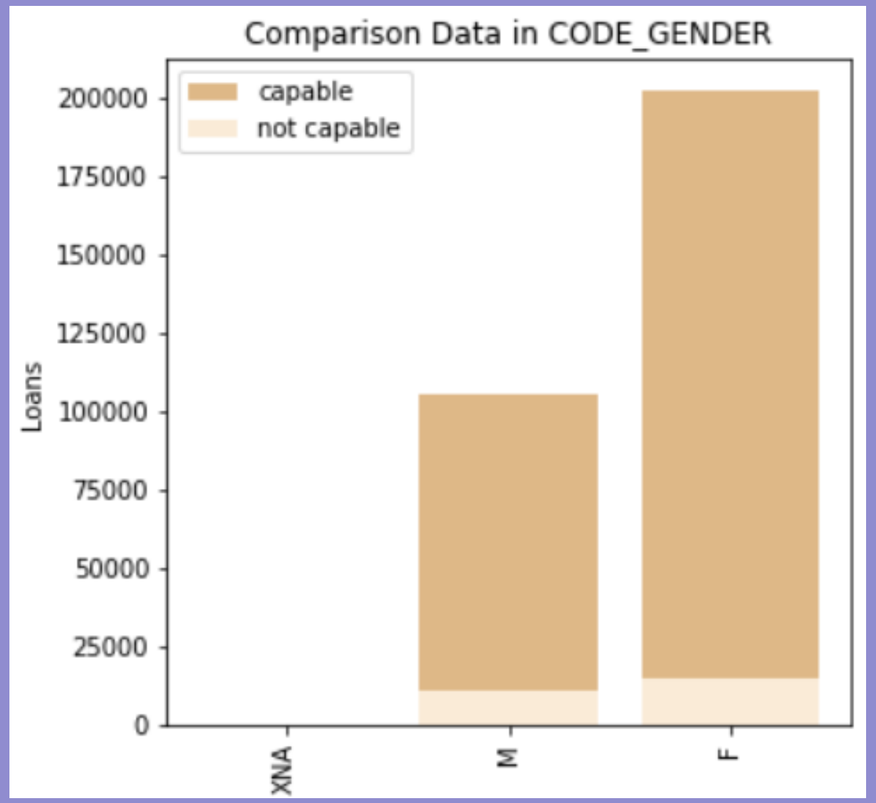




Exploratory Data Analysis (EDA)

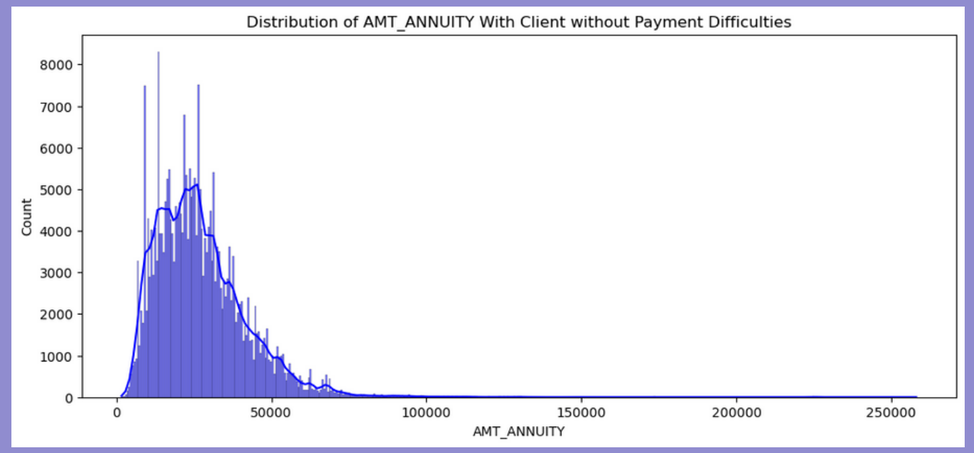


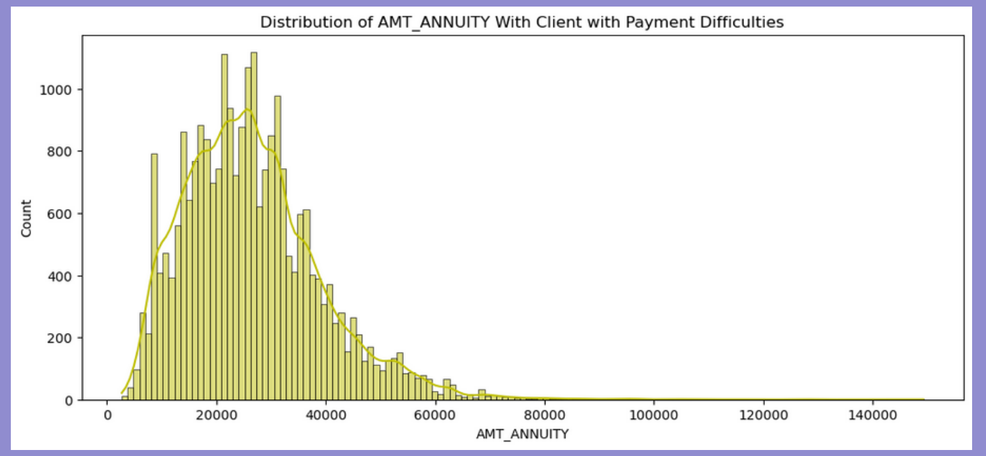
Di sini terdapat imbalance data antara 2 target di mana kita dapat melihat bahwa hanya 8.2% klien Home Credit memiliki kesulitan dalam membayar kredit.

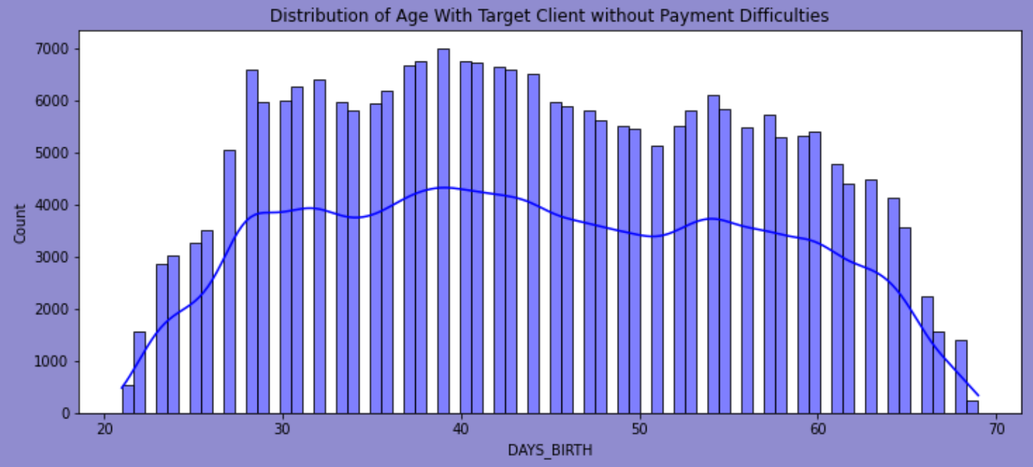


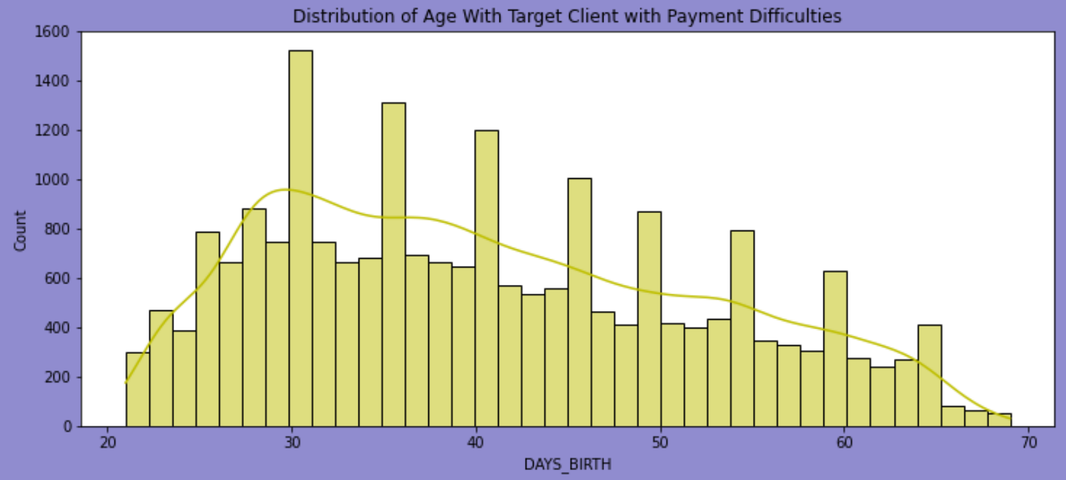
CODE_GENDER	TARGET	TOTAL	% CANT REPAY
XNA	0	4	0.000000
M	10655	105059	0.101419
F	14170	202448	0.069993

Dibandingkan laki-laki, perempuan relatif meminjam dalam jumlah yang lebih besar. Namun, Kapabilitas *repayment* perempuan lebih tinggi sebesar 3% dibandingkan laki-laki.

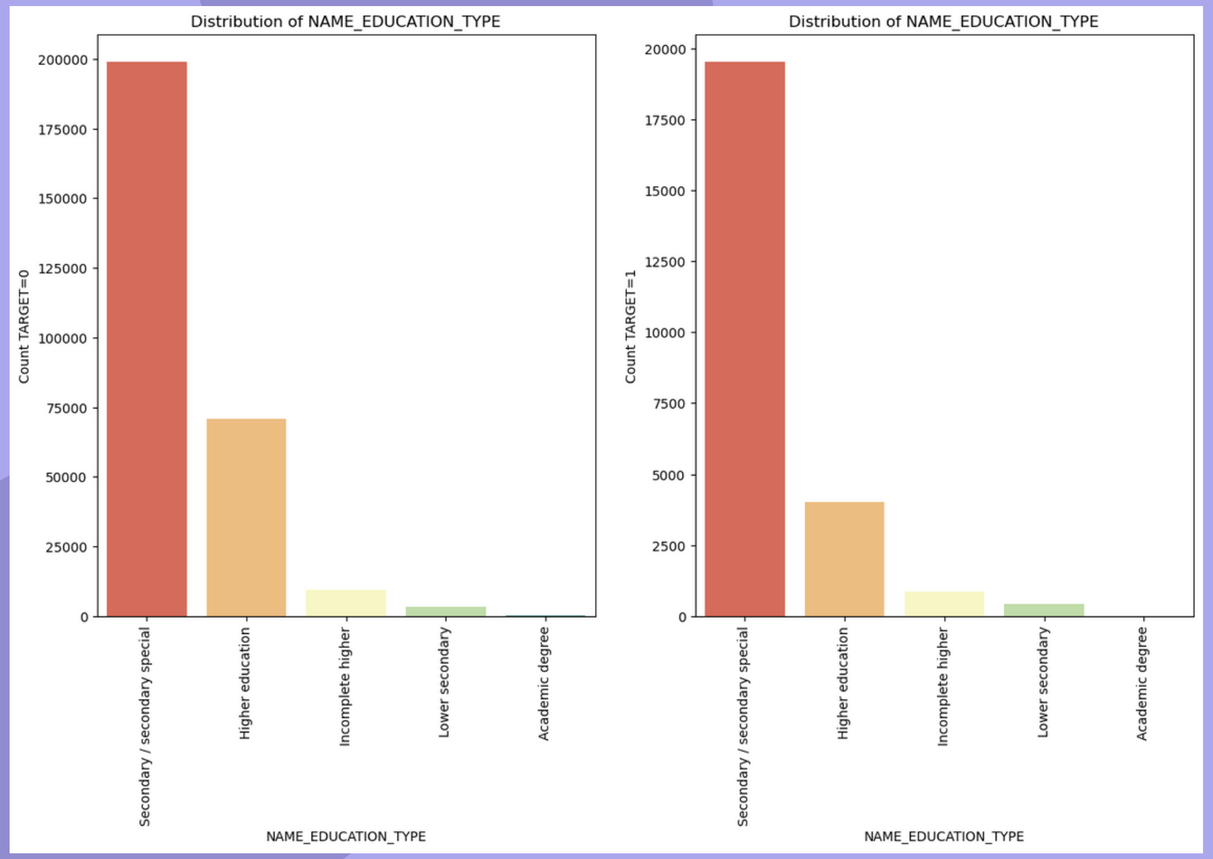






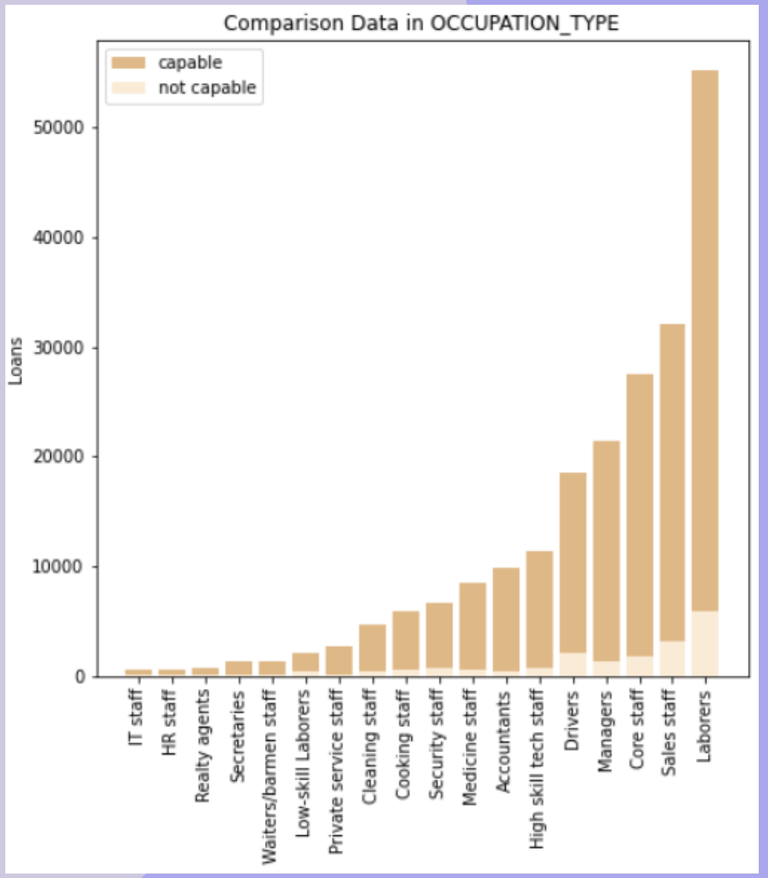


- Pinjaman per tahun para klien biasanya ada di kisaran $20.000 -$30.000 dan jarang ada yang lebih dari $50.000.
- Di sini kita melihat bahwa kebanyakan klien yang meminjam berusia sekitar 30-an diikuti dengan 40-an, di mana umur tersebut termasuk umur produktif seseorang.



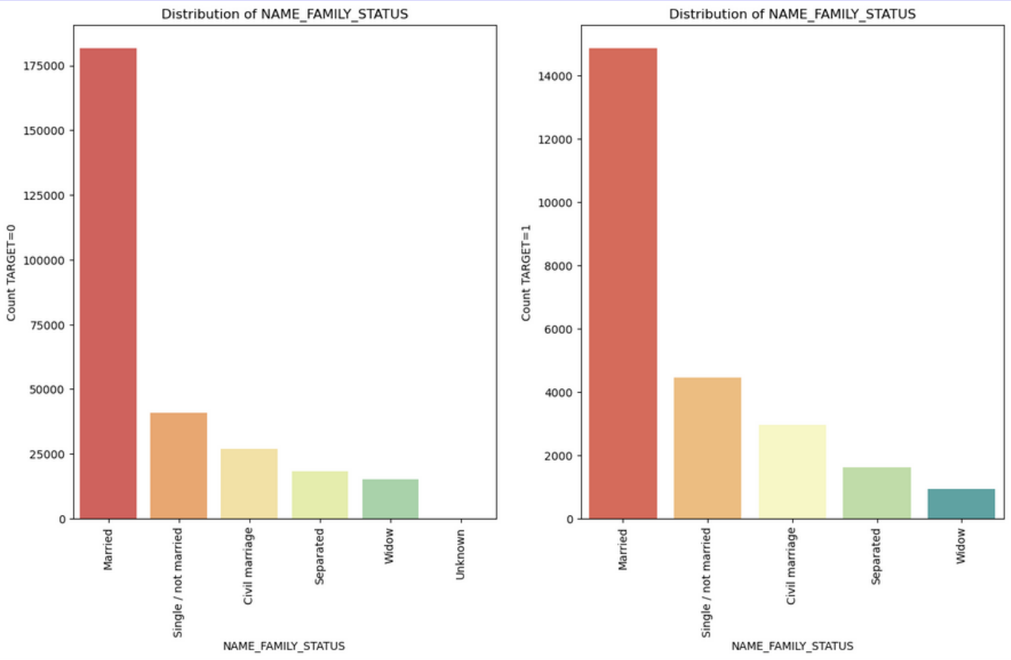
NAME_EDUCATION_TYPE	TARGET	TOTAL	% CANT REPAY
Academic degree	3	164	0.018293
Lower secondary	417	3816	0.109277
Incomplete higher	872	10277	0.084850
Higher education	4009	74863	0.053551
Secondary/secondary special	19524	218391	0.089399

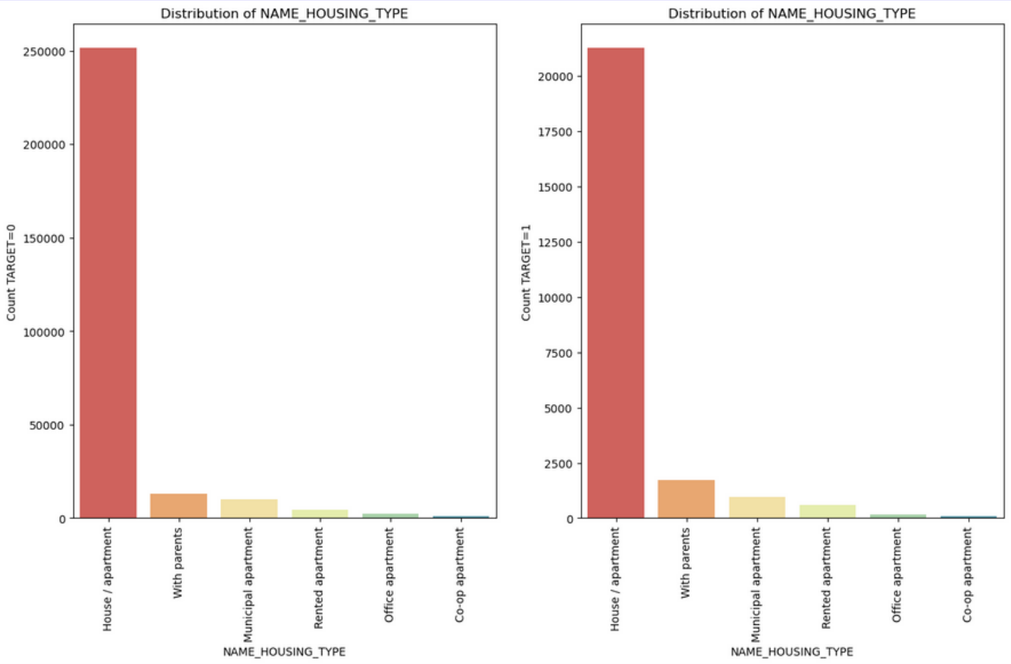
Pendidikan terakhir dari klien Home Credit mayoritas adalah SMA. Terlihat bahwa semakin rendah pendidikan klien Home Credit, maka presentasi kemunngkinan tidak membayar pinjaman juga semakin tinggi dan berlaku sebaliknya.



OCCUPATION_TYPE	TARGET	TOTAL	% CAN'T REPAYMENT
IT Staff	34	34	0.064639
HR staff	36	563	0.063943
Realty agents	59	751	0.078562
Secretaries	92	1305	0.070498
Waiters/barmen staff	152	1348	0.112760
Low-skill Laborers	359	2093	0.171524
Private service staff	175	2652	0.065988
Cleaning staff	447	4653	0.96067
Cooking staff	621	5946	0.104440
Security staff	722	6721	0.107424
Medicine staff	572	8537	0.067002
Accountants	474	98	0.048303
High skill tech staff	701	11380	0.061599
Drivers	2107	18603	0.113261
Managers	1328	21371	0.062140
Core staff	1738	27570	0.063040
Sales staff	3092	32102	0.096318
Laborers	5838	55186	0.105788

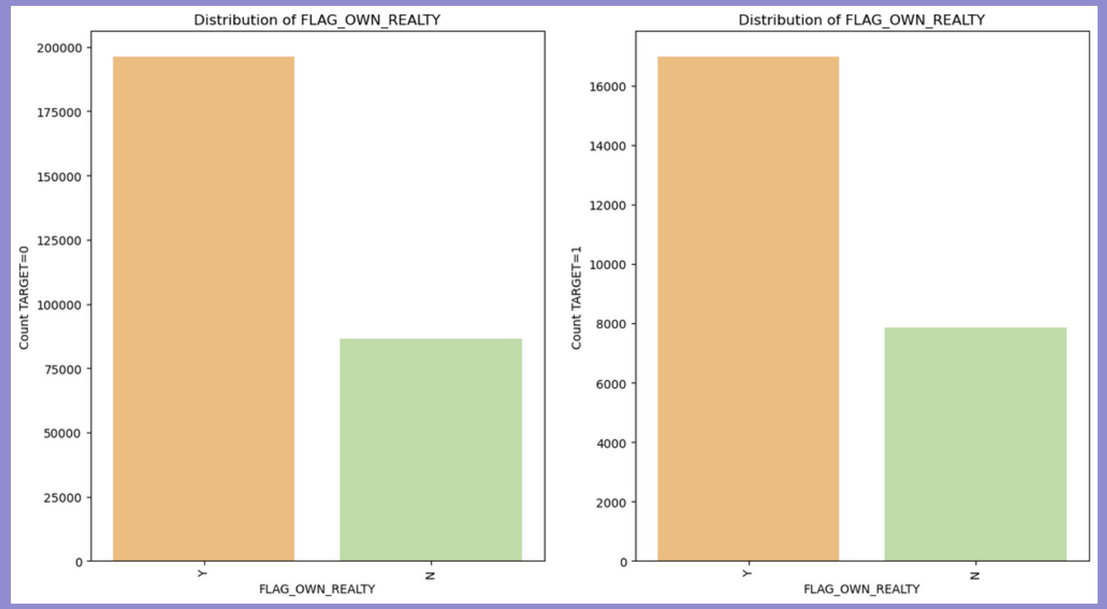
Di sini kita melihat bahwa buruh adalah klien yang sering melakukan pinjaman. Dan disini kita juga melihat bahwa *low-skill laborer* memiliki presentase tidak dapat membayar yang lebih tinggi dibandingkan dengan akuntan. Dapat kita simpulkan bahwa pekerjaan salah fitur yang penting dalam melakukan *scoring* di model.

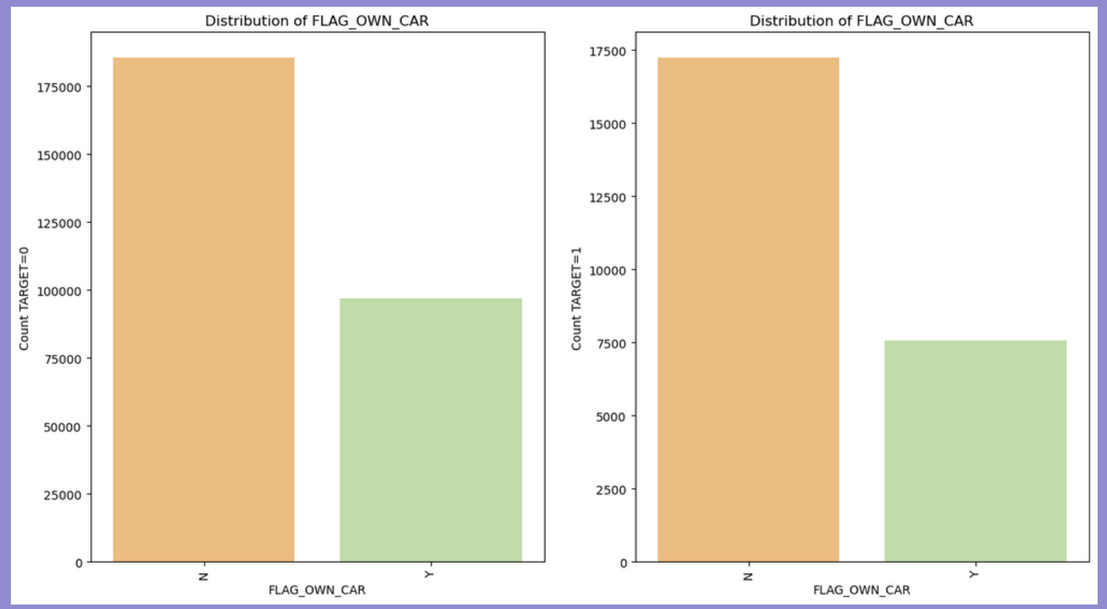




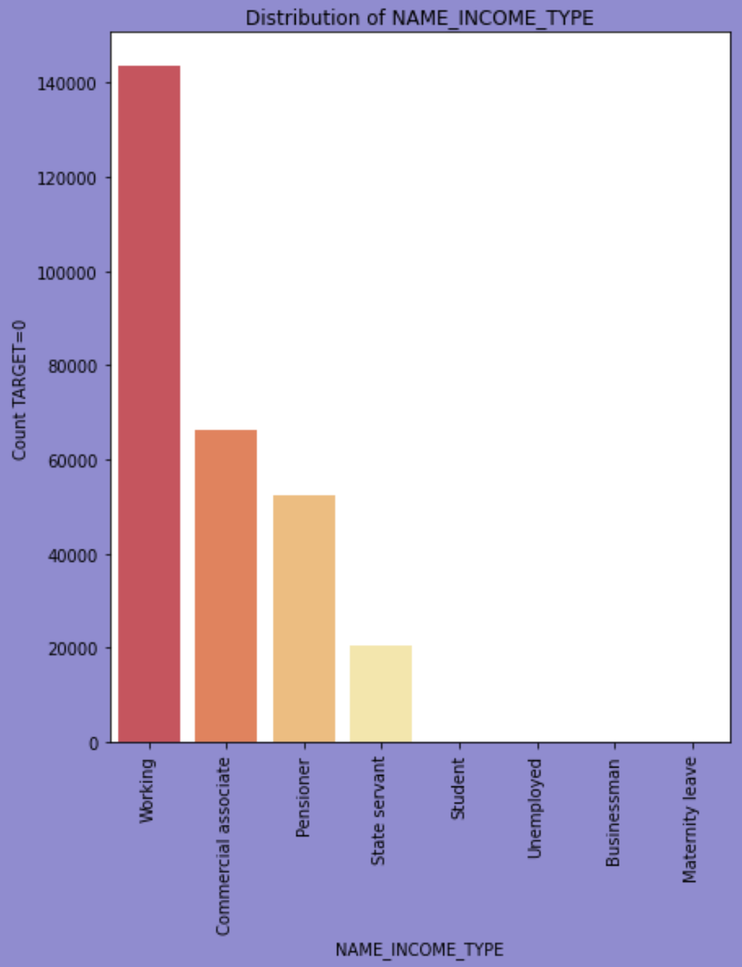
Klien dari Home Credit mayoritas berstatus *Married*. Meskipun begitu, klien yang berstatus *single* dan *Civil Marriage* justru memiliki presentase tidak dapat membayar 2.5% lebih tinggi daripada *married*.

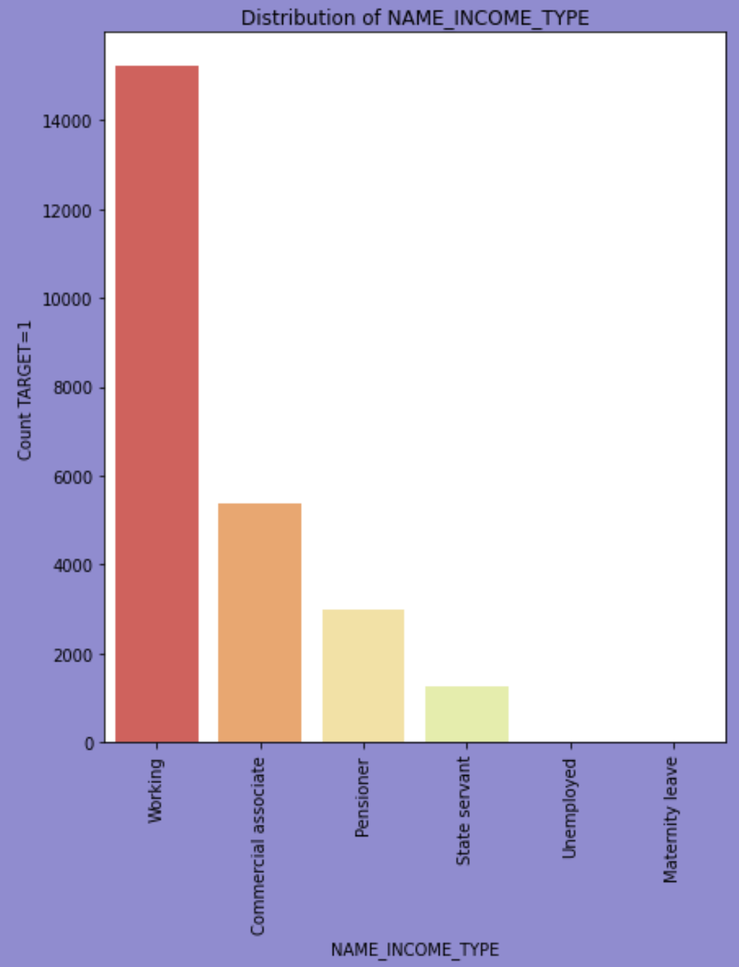
NAME_FAMILY_STATUS	TARGET	TOTAL	% CAN'T REPAY
Unknown	0	2	0.000000
Widow	937	16088	0.058242
Separated	1620	19770	0.081942
Civil marriage	2961	29775	0.099446
Single / not married	4457	45444	0.098077
Married	14850	196432	0.075599





Dilihat dari kepemilikan rumah dan mobil, klien Home Credit mayoritas memiliki rumah pribadi, dan tidak memiliki mobil.





Sumber pendapatan klien dari Home Credit mayoritas berasal dari hasil bekerja mereka. Namun, mayoritas dari mereka juga tidak dapat membayar serta terlihat bahwa pelajar dan pebisnis tidak pernah mengalami kesulitan dalam membayar pinjaman.

Melihat Korelasi

Most Positive Correlations:	
DEF_60_CNT_SOCIAL_CIRCLE	0.031276
DEF_30_CNT_SOCIAL_CIRCLE	0.032248
LIVE_CITY_NOT_WORK_CITY	0.032518
OWN_CAR_AGE	0.037612
DAYS_REGISTRATION	0.041975
FLAG_DOCUMENT_3	0.044346
REG_CITY_NOT_LIVE_CITY	0.044395
FLAG_EMP_PHONE	0.045982
REG_CITY_NOT_WORK_CITY	0.050994
DAYS_ID_PUBLISH	0.051457
DAYS_LAST_PHONE_CHANGE	0.055218
REGION_RATING_CLIENT	0.058899
REGION_RATING_CLIENT_W_CITY	0.060893
DAYS_BIRTH	0.078239

Most Negative Correlations:	
EXT_SOURCE_3	-0.178919
EXT_SOURCE_2	-0.160472
EXT_SOURCE_1	-0.155317
DAYS_EMPLOYED	-0.044932
FLOORSMAX_AVG	-0.044003
FLOORSMAX_MEDI	-0.043768
FLOORSMAX_MODE	-0.043226
AMT_GOODS_PRICE	-0.039645
REGION_POPULATION_RELATIVE	-0.037227
ELEVATORS_AVG	-0.034199
FLOORSMIN_AVG	-0.033614
FLOORSMIN_MEDI	-0.033394
LIVINGAREA_AVG	-0.032997
LIVINGAREA_MEDI	-0.032739

Data Preparation

Mengatasi Missing Value → Mengatasi Outlier → Mengatasi Data yang Anomali

Missing Value

Data Continuous

- 50-60% : Hapus kolom (46 kolom)
- 10-40 % : Diganti dengan median (9 kolom)
- <10% : Diganti dengan median

Data Categorical

- Semua *missing value* untuk data berbentuk kategorikal diatasi dengan cara diubah menjadi nilai modus dari data.

Outliers

- Z-score digunakan untuk mengatasi outliers. Data yang memiliki z-score >= 3 akan dibuang.

Anomali

- Mengubah data yang berisi negatif, seharusnya berisi nilai positif.
- Untuk kolom "DAYS_BIRTHDAY" diubah dari data hari menjadi tahun.
- Untuk kolom "DAYS_EMPLOYED diubah dari data hari menjadi bulan.

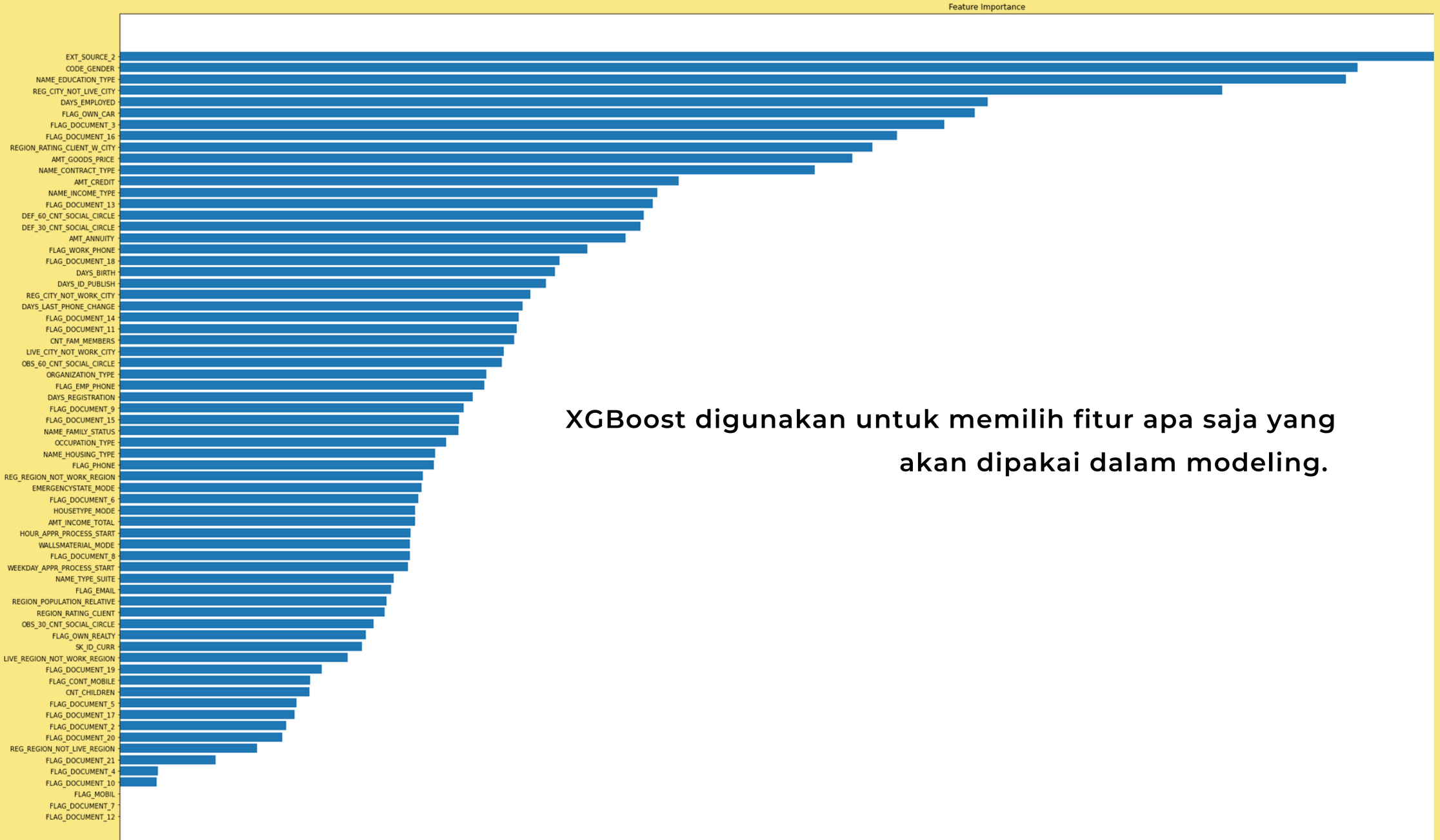
Data setelah dibersihkan:

266.969 BARIS

69 KOLOM

float64(15), int64(39), object(15)

Feature Selection:



XGBoost digunakan untuk memilih fitur apa saja yang akan dipakai dalam modeling.

Modeling

Data	Tipe data sebelum one hot encoding	Tipe data setelah one hot encoding	Keterangan
CODE_GENDER	Object	Int	Gender
NAME_EDUCATION_TYPE	Object	Int	Pendidikan terakhir
REG_CITY_NOT_LIVE_CITY	Object	Int	Alamat tetap
FLAG_OWN_CAR	Boolean	Int	Kepemilikan mobil
AMT_GOODS_PRICE	Int	Int	Harga barang yang ingin dipinjam
NAME_CONTRACT_TYPE	Object	Int	Jenis pinjaman
REGION_RATING_CLIENT_W_CITY	Int	Int	Penilaian home credit terhadap region
NAME_INCOME_TYPE	Int	Int	Sumber pendapatan
DAYS_BIRTH	Int	Int	Umur
AMT_ANNUITY	Int	Int	Jumlah Pinjaman

Modeling

Sebelum modeling, dilakukan *resampling* (upsampling) karena adanya *imbalance* pada data.

Dilakukan juga *One Hot Encoding* menggunakan Label Encoder dari *package skicit-learn*.

Modeling

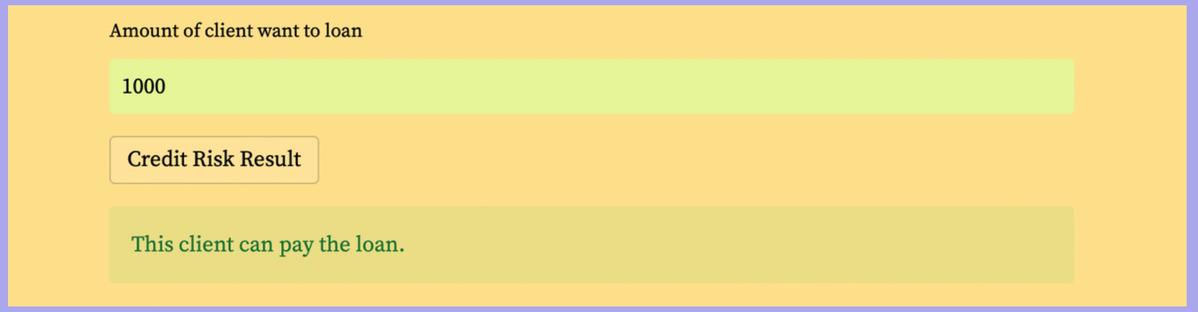
1. Logistic Regression
2. Desicion Tree
3. Random Forest
4. K-Nearest Neighbors
5. Naive Bayes

Evaluasi

Modeling	Classification Matrix	AUC	SELISIH MAE
Logistic Regression	0.56	0.55	0.002
Desicion Tree	0.89	0.88	0.061
Random Forest	0.61	0.61	0.061
K-Nearest Neighbors	0.81	0.81	0.051
Naive Bayes	0.56	0.56	0.061

Model yang terpilih ialah Desicion tree. Hal ini dikarenakan dtree memiliki nilai acurracy pada f1 score tertinggi, yaitu 89% dan nilai AUC sebesar 88%. Selisih MAE juga tidak terlalu besar yang menandakan model ini tidak overfit.

Deployment



Hasil Deployment

Terima Kasih

Link Google Collab