Default Risk Prediction

Nomor Kelompok: 28
Nama Mentor: Erwin Fernanda

Data Analytics Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Google colab:
https://colab.research.google.com/drive/1dTFgrDSduCW-2gCc5GNLbVoRsPyhMzh1?usp=sharing
Looker:
https://lookerstudio.google.com/reporting/d7ef68c5-8448-4641-aad2-6df96fcf6ea1

Anggota Tim

Hijriyani Mardhotillah
Institut Teknologi PLN

Nur Indah Setyaningsih
Universitas Gadjah Mada

Nur Kholifah
Universitas Negeri Malang

Sekar Kinasih
Universitas Jenderal Soedirman

Syafarizal Irgi Margiansyah
Institut Teknologi Sepuluh Nopember

Business Understanding

Business problem yang dihadapi adalah tingginya persentase gagal bayar kredit sebesar 8% atau setara dengan 20 ribu orang, yang dapat berdampak negatif pada keuangan perusahaan dan reputasi perusahaan di mata nasabah.

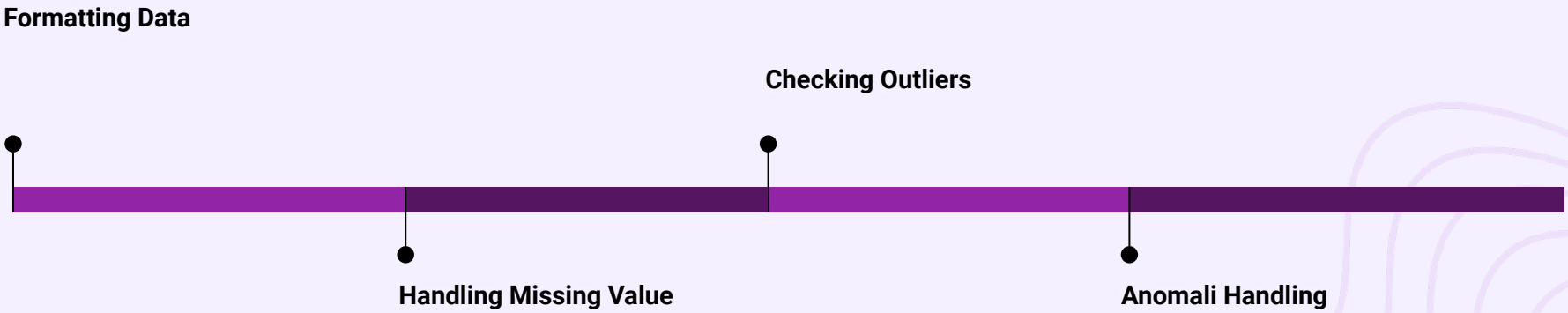
Proyek ini bertujuan untuk memprediksi aspek yang berpengaruh terhadap ketidakmampuan nasabah dalam membayar kredit menggunakan model *machine learning*.

Data Understanding

application_train

- Satu baris mewakili satu pinjaman dalam sampel data yang ditandai oleh kolom SK_ID_CURR.
- Dataset terdiri dari 307511 baris dan 122 kolom.
- TARGET
  1 : nasabah yang kesulitan/keterlambatan dalam membayar kembali pinjaman)
  0 : nasabah yang dapat/tepat waktu membayar kembali pinjaman

Data Preparation



Formatting Data

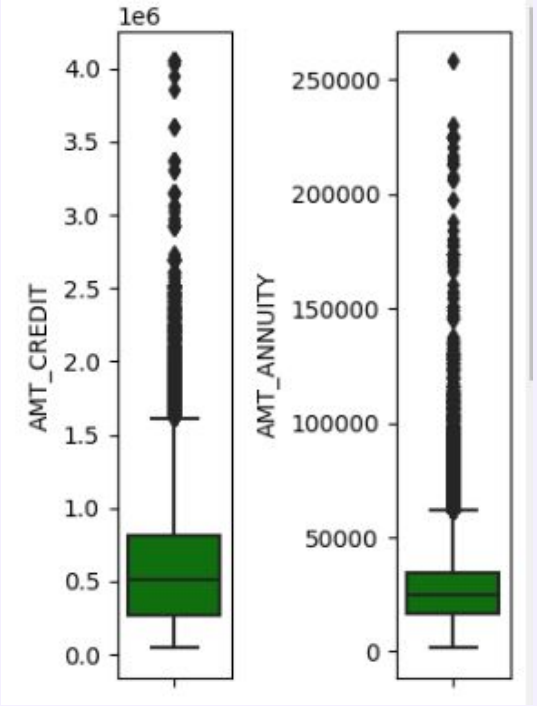
- DAYS_BIRTH, DAYS_ID_PUBLISH, DAYS_REGISTRATION, DAYS_EMPLOYED, dan DAYS_LAST_PHONE_CHANGE yang berisikan hitungan hari dengan nilai negatif sehingga perlu diganti format datanya.
- Akan diubah menjadi tahun untuk memudahkan analisis.

Handling Missing Value

```
Kolom dengan Missing Values: 67
Total Kolom: 122
Persentase dari Banyaknya Kolom dengan Missing Values: 54.9%
```

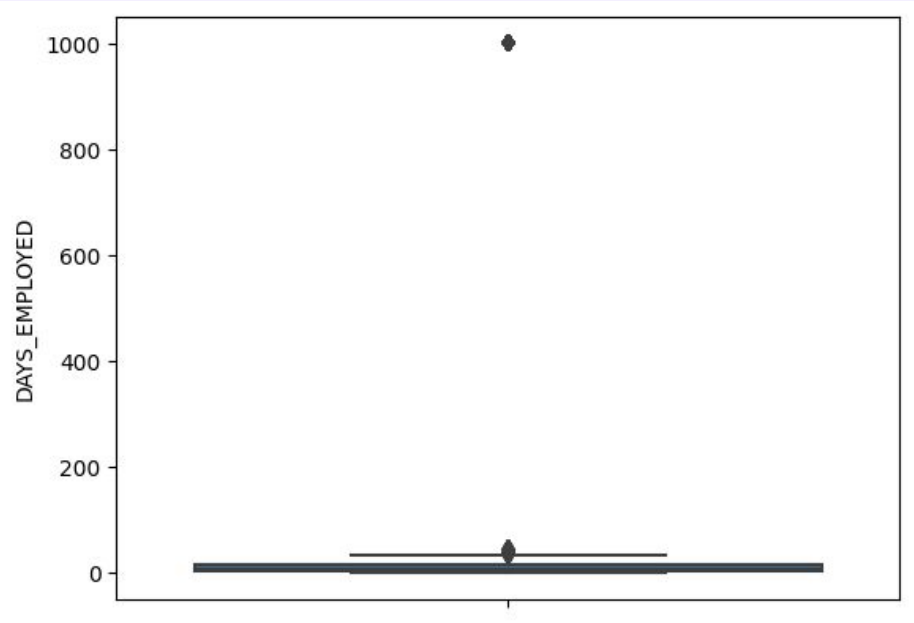
- Kolom dengan jumlah missing value lebih dari 50% akan di-*drop.*
- Impute dengan nilai mediannya untuk data numerik dan nilai modus untuk data categorical.
- Ukuran data setelah *handling missing value* adalah 307511 baris dan 81 kolom.

Checking Outliers



Outliers masih masuk akal

Anomali Handling

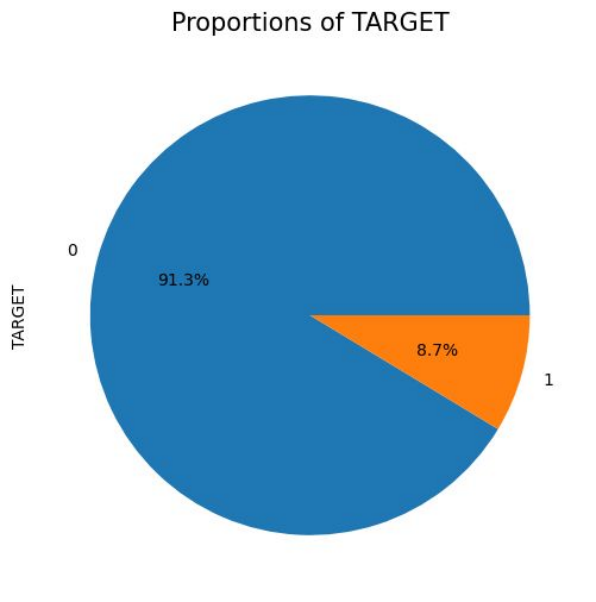


Nilai anomali lama waktu bekerja mencapai 1000 tahun akan didrop

Data Analysis

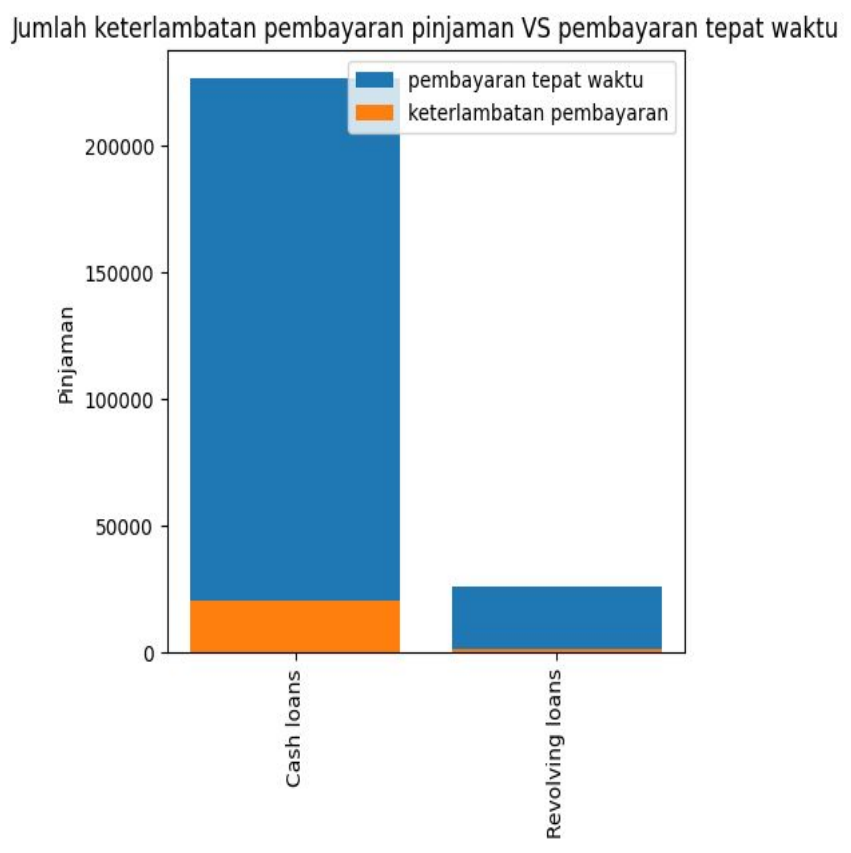
Univariate Analysis

- **Data Target**



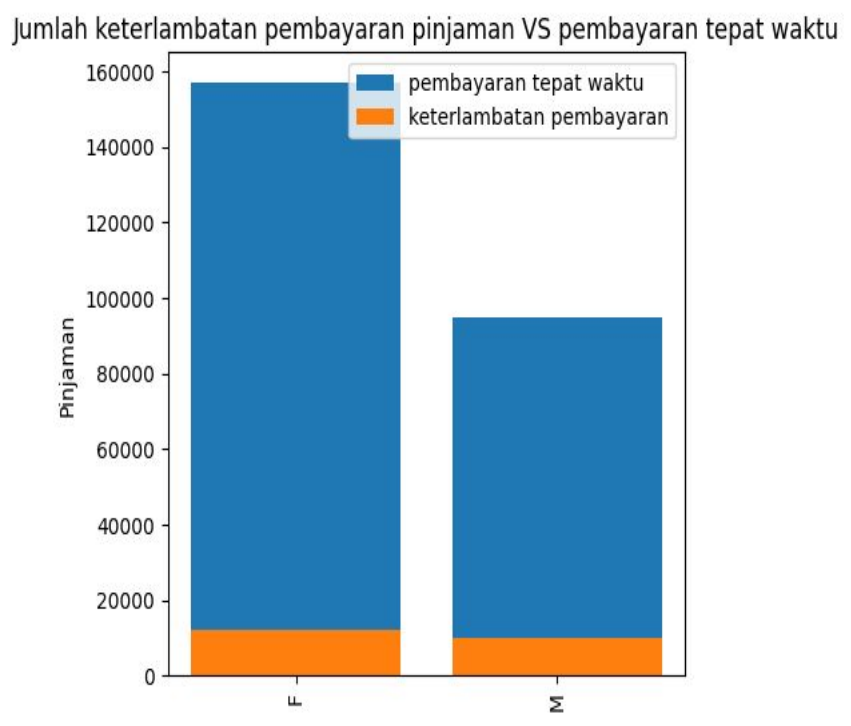
Terdapat ketidakseimbangan data target.

- **Name_Contract_Type**



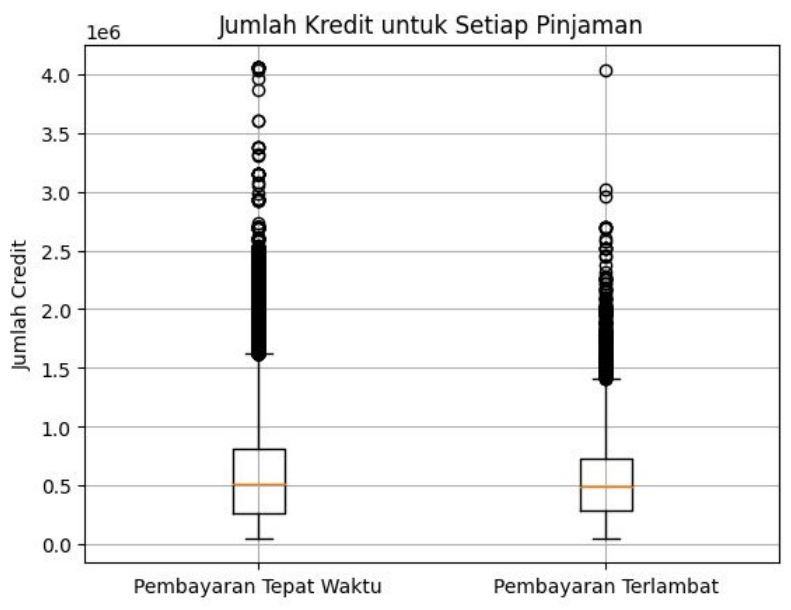
Sebagian besar orang mengambil pinjaman dalam bentuk pinjaman tunai (*cash loans*) daripada pinjaman bergulir (*revolving loans*) seperti kartu kredit.

- Code_Gender



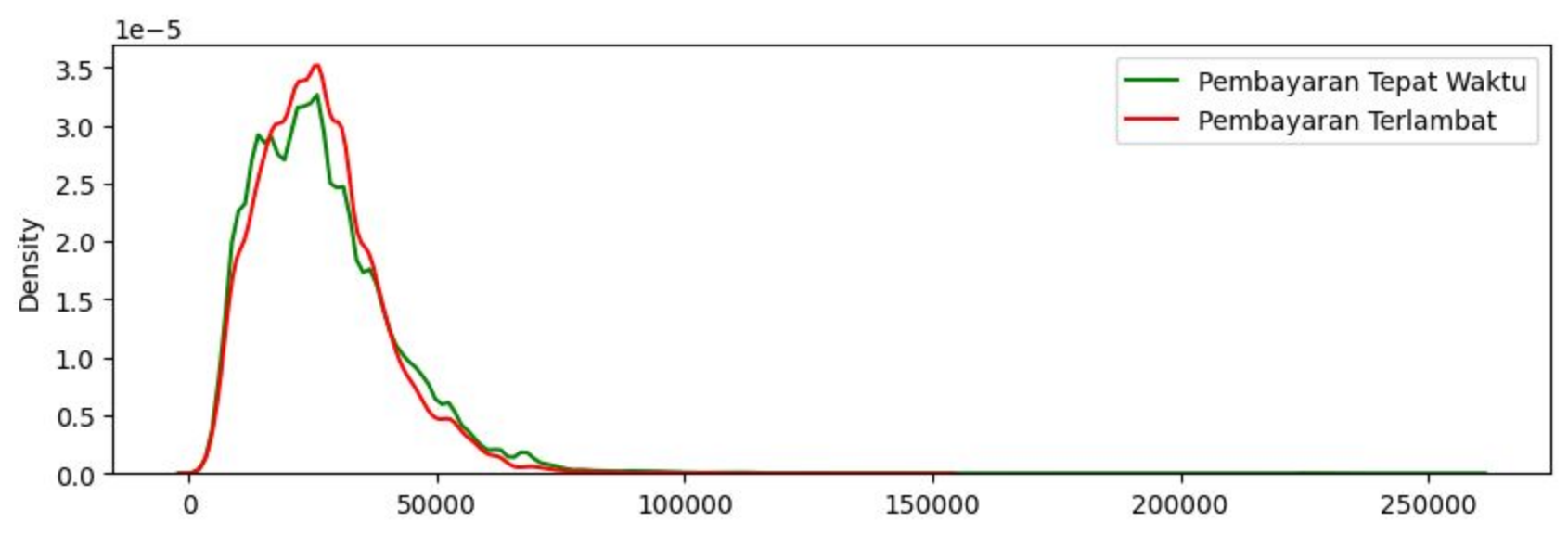
- Wanita mengambil lebih banyak pinjaman dibandingkan dengan pria.
- Dan pada saat yang sama, wanita sedikit lebih mampu membayar kembali pinjaman dibandingkan dengan pria.

- Amt_Credit



Nilai median dari jumlah kredit peminjam yang mampu mengembalikan pinjaman tepat waktu sedikit lebih besar dari nilai median peminjam yang terlambat.

- Amt_Anuity



Kebanyakan orang membayar anuitas di bawah 50.000 untuk pinjaman

- Name_Education_Type

NAME_EDUCATION_TYPE	TARGET	total	Avg
Secondary / secondary special	17000	173286	0.098104
Higher education	3669	66669	0.055033
Incomplete higher	848	9757	0.086912
Lower secondary	315	2287	0.137735
Academic degree	3	138	0.021739

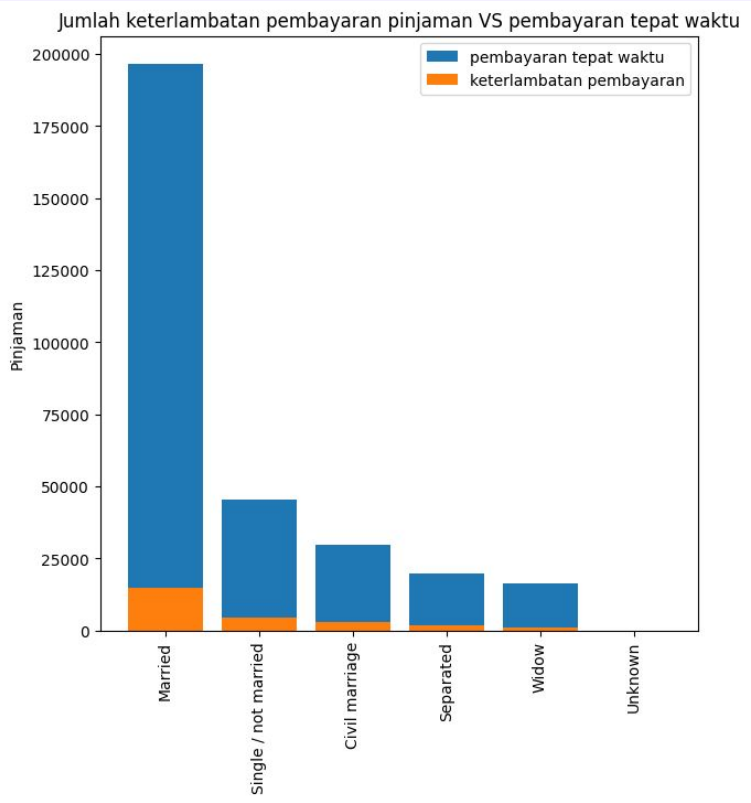
Semakin tinggi tingkat pendidikan seseorang, kemampuan pembayaran pinjamannya juga meningkat.

- Name_Income_Type

NAME_INCOME_TYPE	TARGET	total	Avg
State servant	1249	21703	0.05755
Student	0	18	0.00000
Businessman	0	10	0.00000
Pensioner	0	10	0.00000
Maternity leave	2	5	0.40000

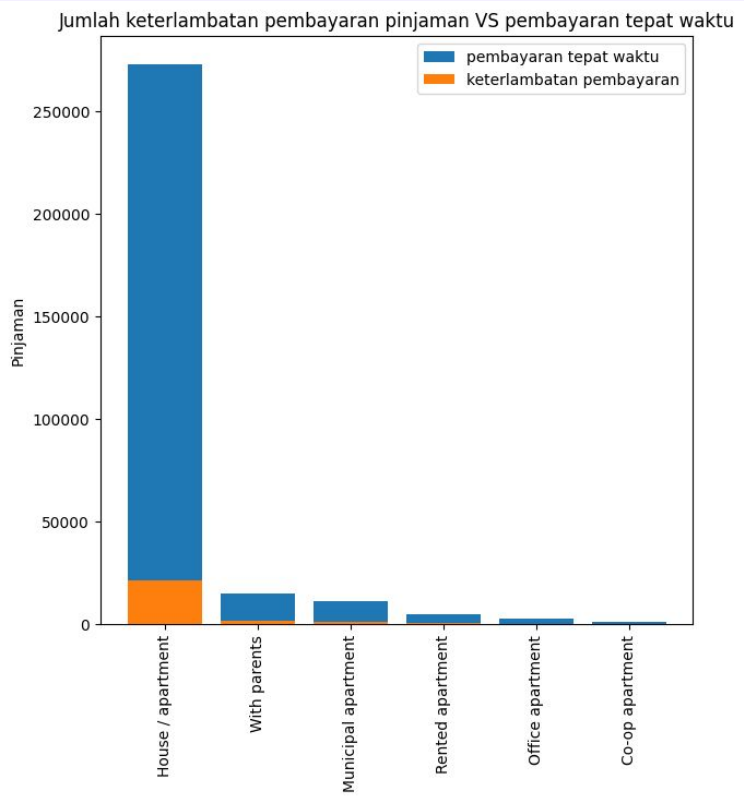
Pinjaman yang diberikan kepada siswa, pengusaha, dan pensiunan yang telah mengajukan pinjaman, mereka telah dianggap mampu membayar kembali pinjaman dengan tepat waktu.

- Name_Family_Type



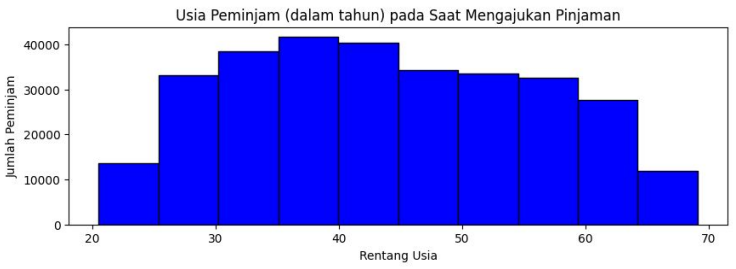
Orang yang sudah menikah (Married) mengajukan pinjaman paling banyak dan jumlah orang yang dianggap tidak mampu membayar pinjaman juga paling banyak.

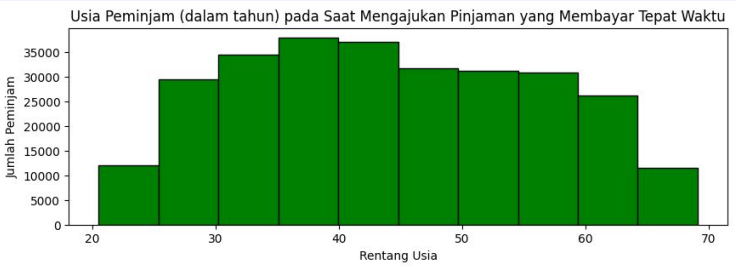
- Name_Housing_Type

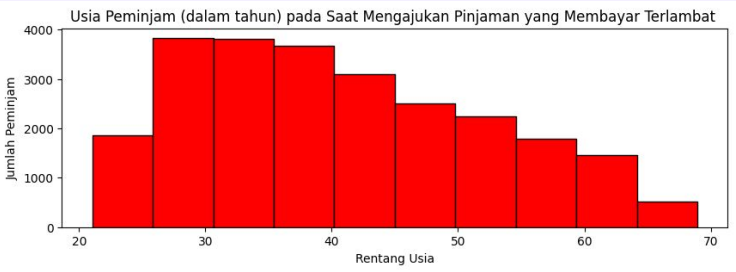


Orang yang tinggal di Rumah/Apartemen (House / apartment) mengajukan pinjaman paling banyak dan orang yang dianggap tidak mampu membayar pinjaman juga paling banyak .

- Days_Birth

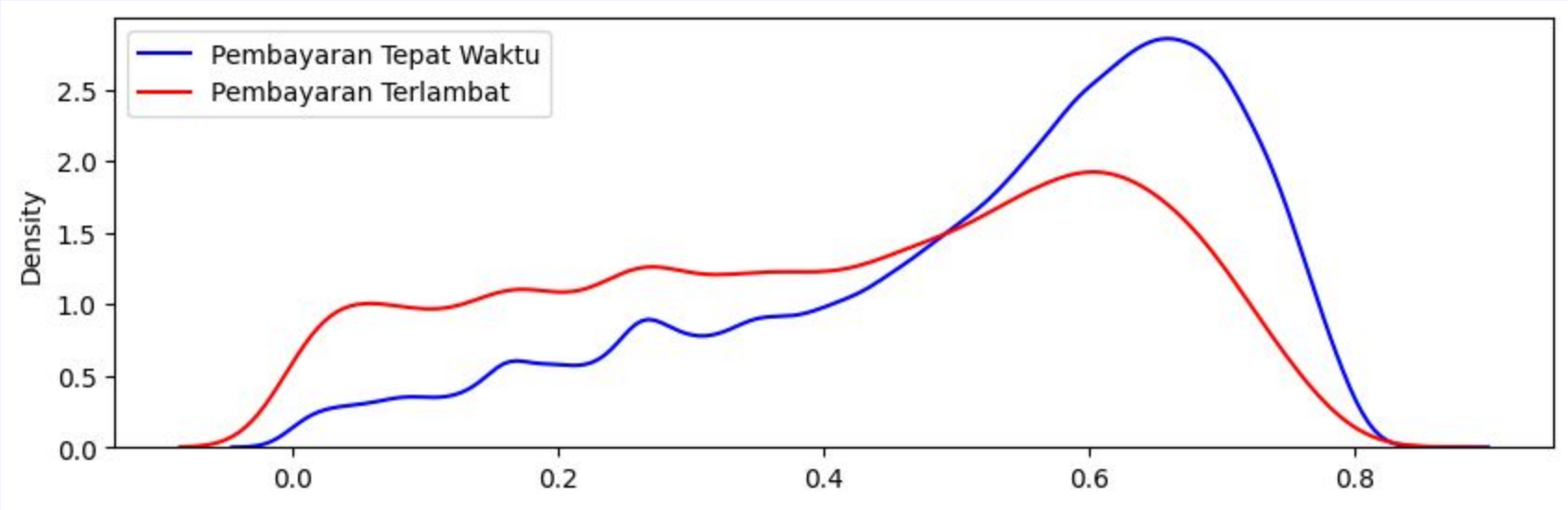






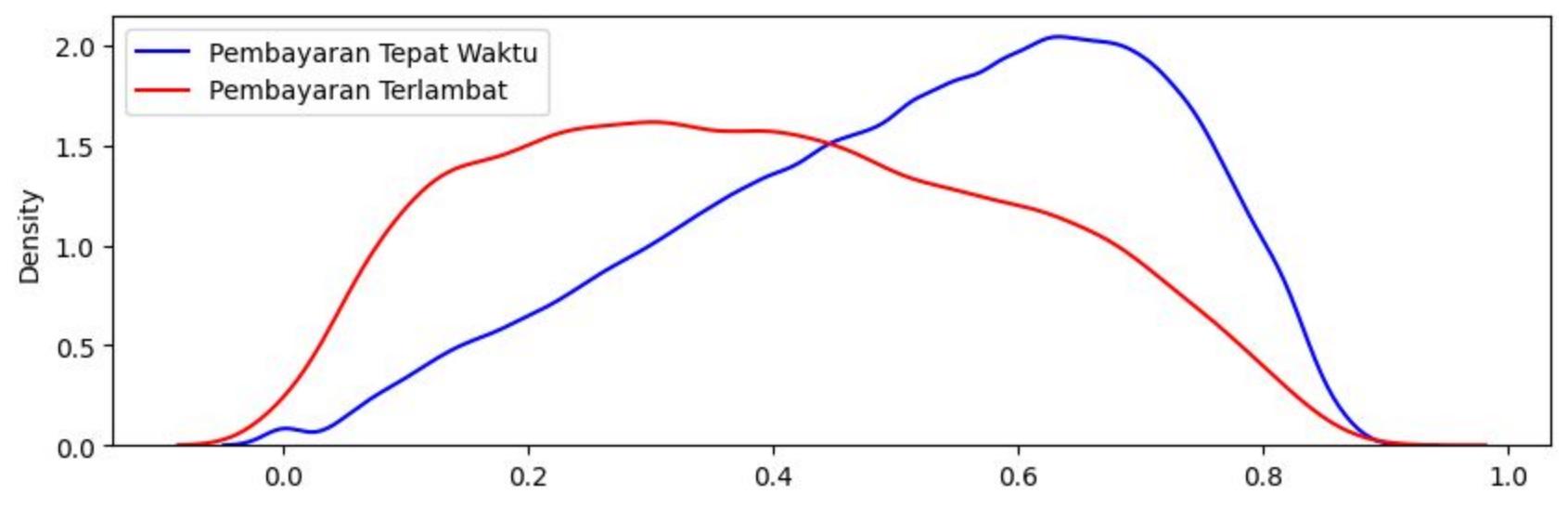
- Sebagian besar orang yang mengajukan pinjaman berada dalam kisaran 35-45 tahun.
- Orang-orang yang dianggap paling mampu mengembalikan pinjaman ada di rentang usia 35-45 tahun.
- Rentang usia 25-35 tahun memiliki peluang besar kesulitan dalam mengembalikan pinjaman.

- **EXT_SOURCE_2**



Terlihat bahwa *feature* ini memiliki perbedaan yang cukup besar di antara 2 kelas, seperti yang terlihat pada plot PDF. Oleh karena itu, EXT_SOURCE_2 akan menjadi *feature* penting.

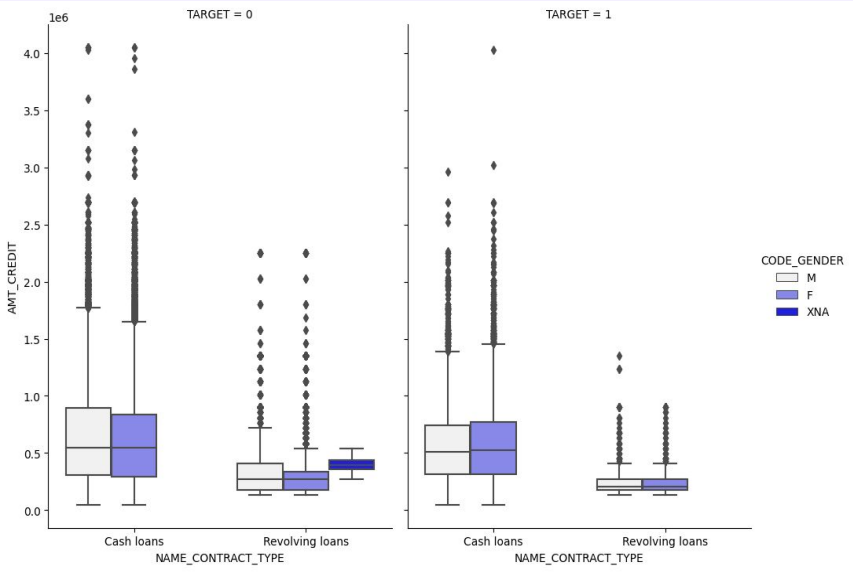
- **EXT_SOURCE_3**



Terlihat bahwa *feature* ini memiliki perbedaan yang cukup besar di antara 2 kelas, seperti yang terlihat pada plot PDF. Oleh karena itu, EXT_SOURCE_3 akan menjadi *feature* penting.

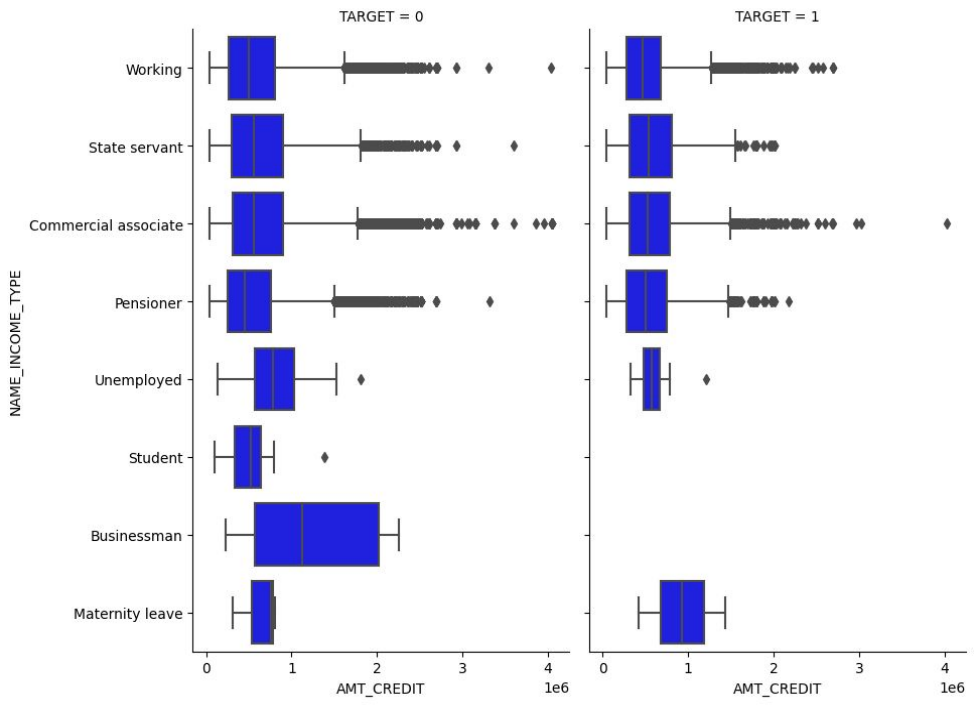
Bivariate Analysis

- NAME_CONTRACT_TYPE & AMT_CREDIT



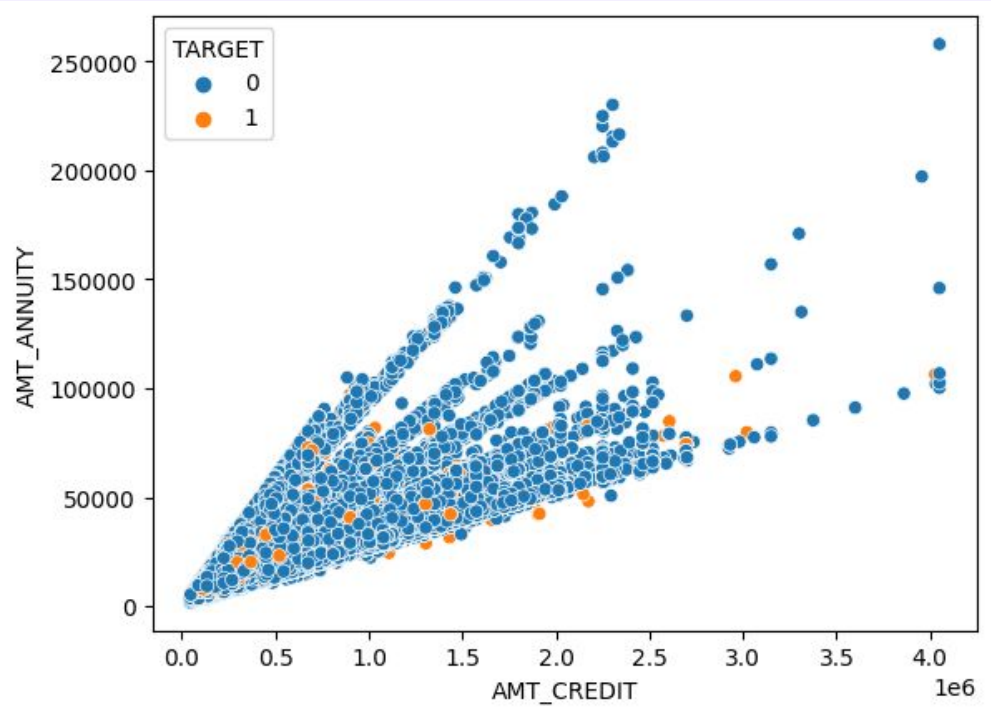
Pria & Wanita dengan Pinjaman Tunai (Cash Loan) memiliki peluang lebih tinggi untuk dianggap mampu membayar kembali pinjaman berdasarkan Jumlah Kredit (Credit Amount) mereka

- NAME_INCOME_TYPE & AMT_CREDIT



Peminjam dengan nilai jumlah kredit lebih tinggi memiliki kemungkinan tinggi di di berbagai jenis pendapatan, terutama di kasus 'Pengangguran', 'Mahasiswa' dan 'Pengusaha'

- AMT_CREDIT & AMT_ANNUITY



Jika Jumlah Kredit (AMT_CREDIT) tinggi, jumlah Anuitas(AMT_ANNUITY) yang sama juga akan tinggi.

Modelling dan Evaluation

Data Encoding

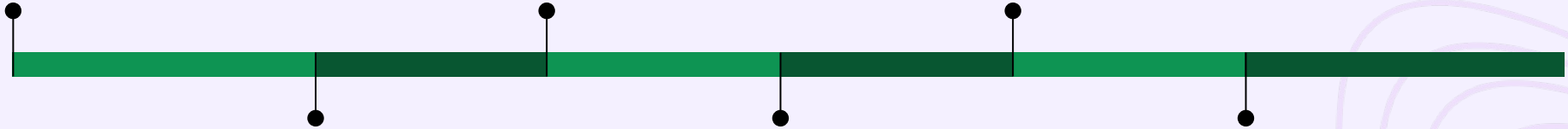
one hot encoding. Dataset menjadi 252.137 baris dan 178 kolom.

Feature Selection

Berdasarkan Mutual Information Gain diambil 20 kolom terbaik.

Standardization

membakukan data x_train dan x_test



Random Undersampling

undersampling data minor, menjadi

0: 21835, 1: 21835

Split Data

Menjadi 70% data train, 30% data test.

Modelling

- Random Forest
- Logistic Regression
- Gradient Boosting
- Light GBM

Evaluation

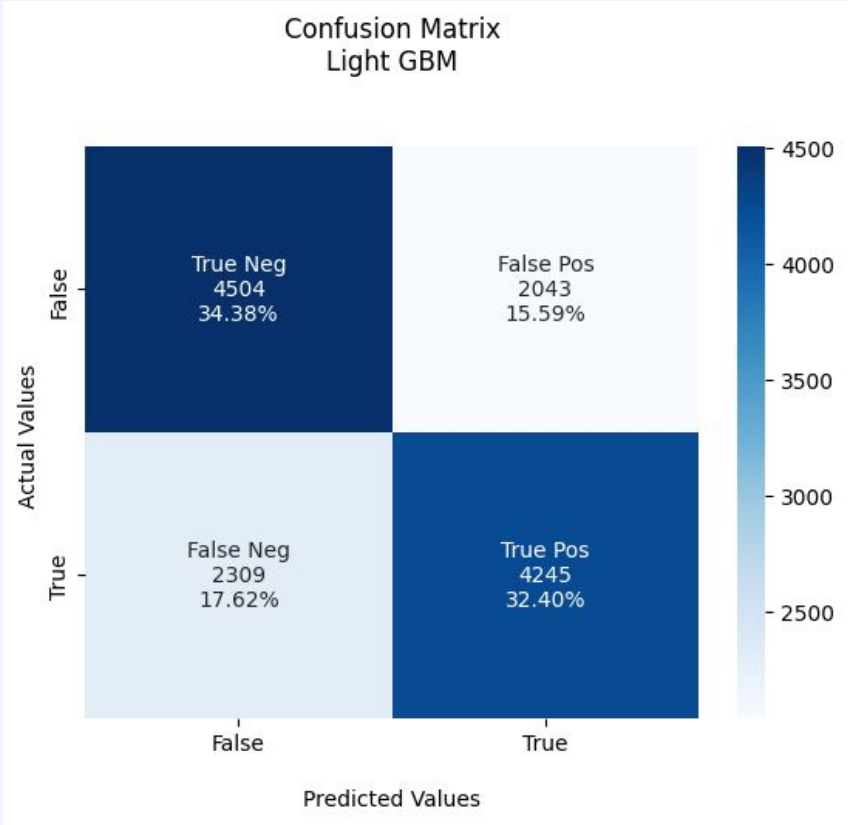
Evaluasi model adalah proses menggunakan metrik evaluasi yang berbeda untuk memahami kinerja model pembelajaran mesin, serta kekuatan dan kelemahannya. Pada kali ini, kami akan menggunakan ROC-AUC,Accuracy, Recall serta Confusion matrix.

	Accuracy	ROC AUC	Precision	Recall	F1
Random Forest	0.663613	0.663620	0.667971	0.651358	0.659560
Logistic Regression	0.664987	0.663620	0.667596	0.657919	0.662722
Gradient Boosting	0.666590	0.666595	0.669879	0.657614	0.663690
Light GBM	0.671933	0.671936	0.673592	0.667836	0.670702

Model yang terpilih adalah **Light GBM**

Light GBM

	precision	recall	f1-score	support
0	0.66	0.69	0.67	6547
1	0.68	0.65	0.66	6554
accuracy			0.67	13101
macro avg	0.67	0.67	0.67	13101
weighted avg	0.67	0.67	0.67	13101

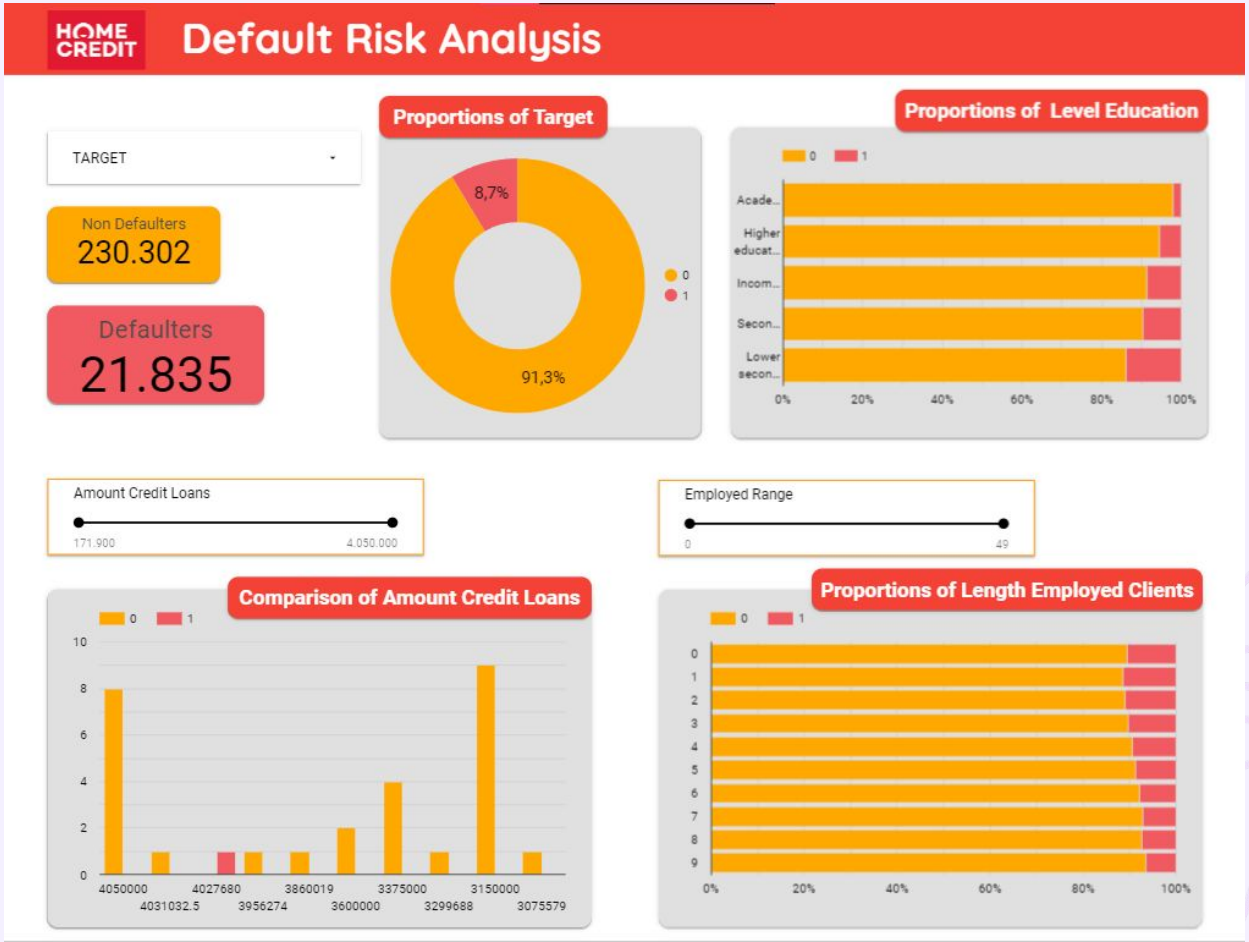


	feature	importance
0	DAYS_EMPLOYED	312
1	EXT_SOURCE_3	305
2	AMT_ANNUITY	296
3	EXT_SOURCE_2	282
4	DAYS_BIRTH	275
5	AMT_CREDIT	242
6	AMT_GOODS_PRICE	232
7	DAYS_LAST_PHONE_CHANGE	221
8	DAYS_REGISTRATION	216
9	REGION_POPULATION_RELATIVE	136
10	TOTALAREA_MODE	135
11	YEARS_BEGINEXPLUATATION_MEDI	129
12	NAME_EDUCATION_TYPE_Secondary / secondary special	58
13	FLAG_DOCUMENT_3	51
14	REGION_RATING_CLIENT	39
15	NAME_INCOME_TYPE_Working	36
16	FLOORSMAX_MODE	35
17	FLAG_EMP_PHONE	0
18	NAME_EDUCATION_TYPE_Higher education	0
19	FLAG_MOBIL	0

Berdasarkan *feature importance* yang didapat dari model Light GBM, kelompok kami memilih *feature* berikut ini yang berpengaruh dengan *default risk*:

1. DAYS_EMPLOYED
2. AMT_ANNUITY
3. DAYS_BIRTH
4. AMT_CREDIT
5. NAME_EDUCATION_TYPE

Deployment



Conclusion

- Model klasifikasi terbaik yang didapatkan untuk prediksi gagal bayar adalah Light GBM.
- Faktor yang berpengaruh untuk prediksi gagal bayar menggunakan dataset Home Credit adalah lama waktu bekerja, anuitas pinjaman, usia, jumlah kredit dari pinjaman, dan tingkatan pendidikan.
- Semakin lama pengalaman bekerja seseorang, kemungkinan dapat membayar kembali pinjaman lebih besar.
- Peminjam dengan rentang usia 25-30 tahun perlu diperhatikan karena mereka cenderung sulit membayar kembali pinjaman.
- Peminjam dengan anuitas dan credit pinjaman yang besar mempunyai kecenderungan tidak memiliki kesulitan dalam membayar pinjaman.

Terima kasih!