KELOMPOK 6

HOME CREDIT RISK SCORING

LINK GOOGLE COLAB

https://colab.research.google.com/drive/15Fm5BzQc28pidLTuVeZXsT23U86HWr58?usp=sharing

LINK GOOGLE DATA STUDIO

https://datastudio.google.com/reporting/1ce13dff-7c8a-43b0-8d49-6940df6b13da

Bussiness Understanding

- **Determine Business Objectives**

  1. Sulit mendapatkan pinjaman karena riwayat kredit yang tidak mencukupi atau tidak ada?
  2. Pengalaman meminjam yang positif dan aman jarang?
  3. Banyak yang kurang terlayani karena tidak memiliki pengalaman pinjaman yang positif?

- **Situation Assessment**

  Biaya Kesalahan Klasifikasi Tinggi: yang tidak mampu mengembalikan pinjaman diklasifikasikan sebagai mampu dan dia diberikan pinjaman dan yang mampu diklasifikasikan sebagai tidak mampu, permohonannya ditolak
  Capable = 0
  Non-capable = 1

Bussiness Understanding

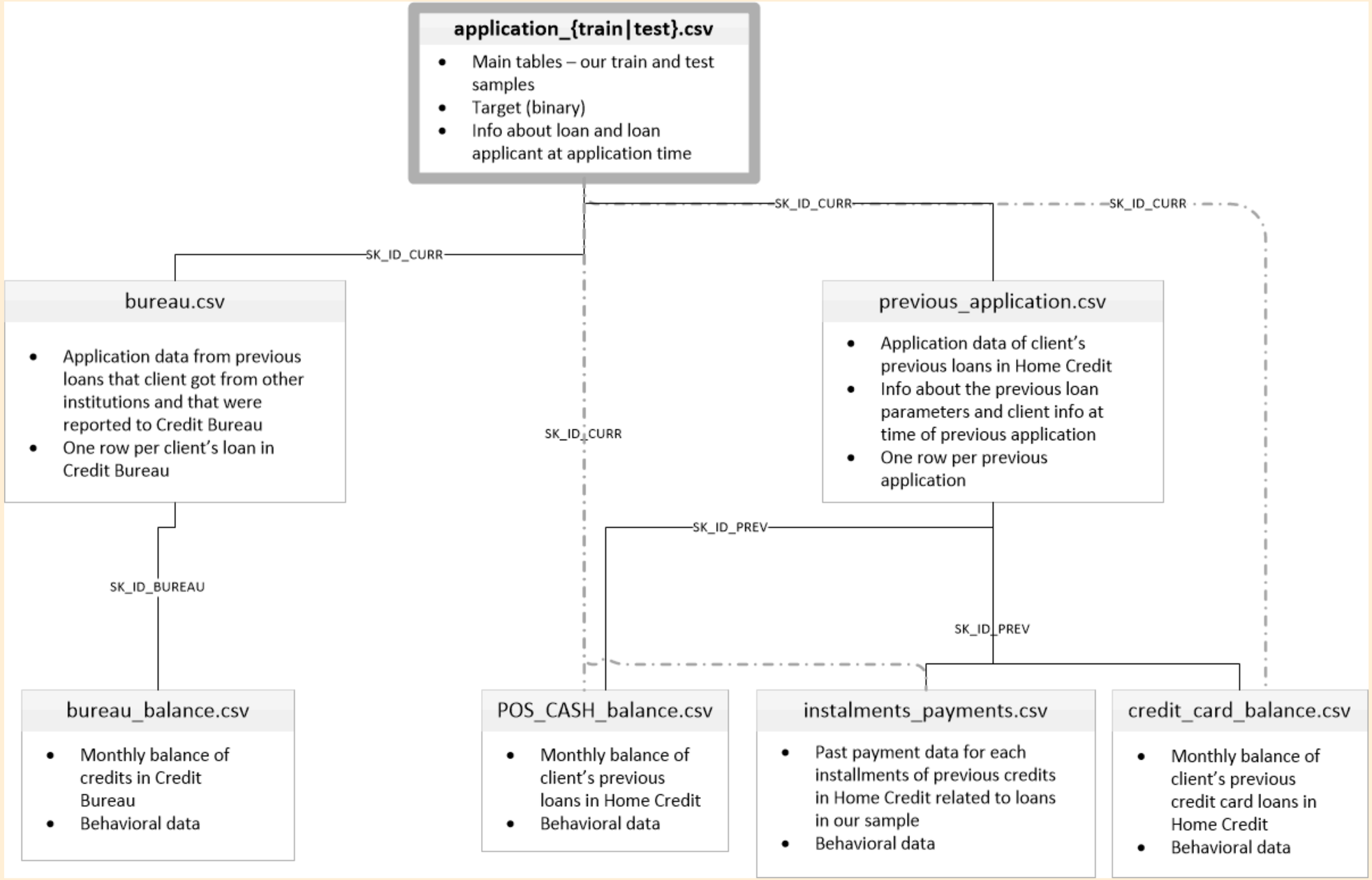
- **Determine Data Mining Goal**
  1. Melihat apakah clients capable atau non-capable dalam peminjaman home credit.
  2. Melihat jenis pinjaman yang dilakukan oleh clients.
  3. Melihat pekerjaan teratas menurut penghasilan clients.

- **Produce Project Plan**
  1. Business Understanding
  2. Data Understanding
  3. Data Preparation
  4. Modeling
  5. Evaluation
  6. Deployment (Dashboard)

DATA UNDERSTANDING

Dataset:
https://www.kaggle.com/competitions/home-credit-default-risk/data

Describe data



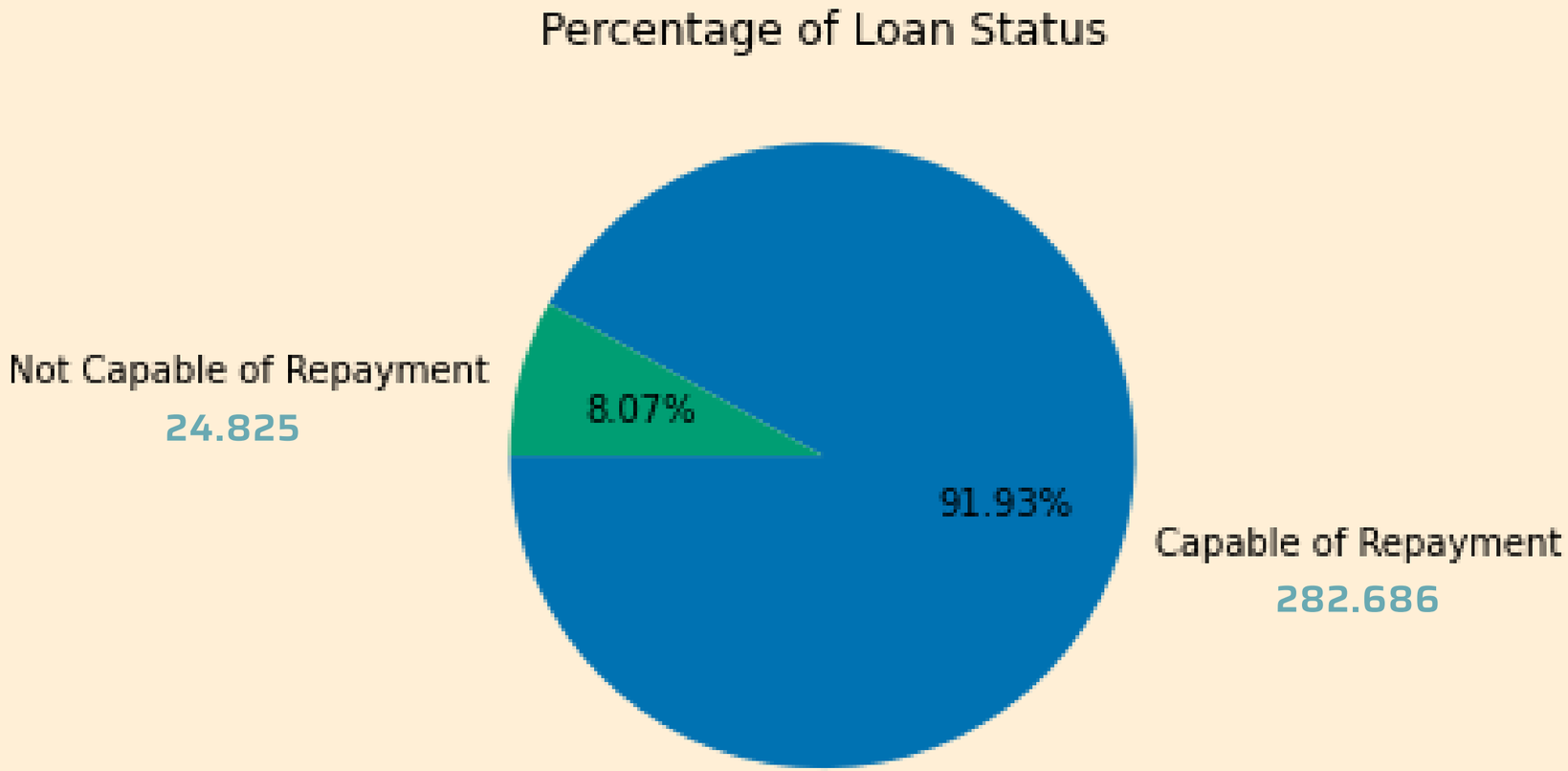
TRAIN

Data shape : (307511, 122)
Data type:
float64 : 65
int64 : 41
object : 16

TEST

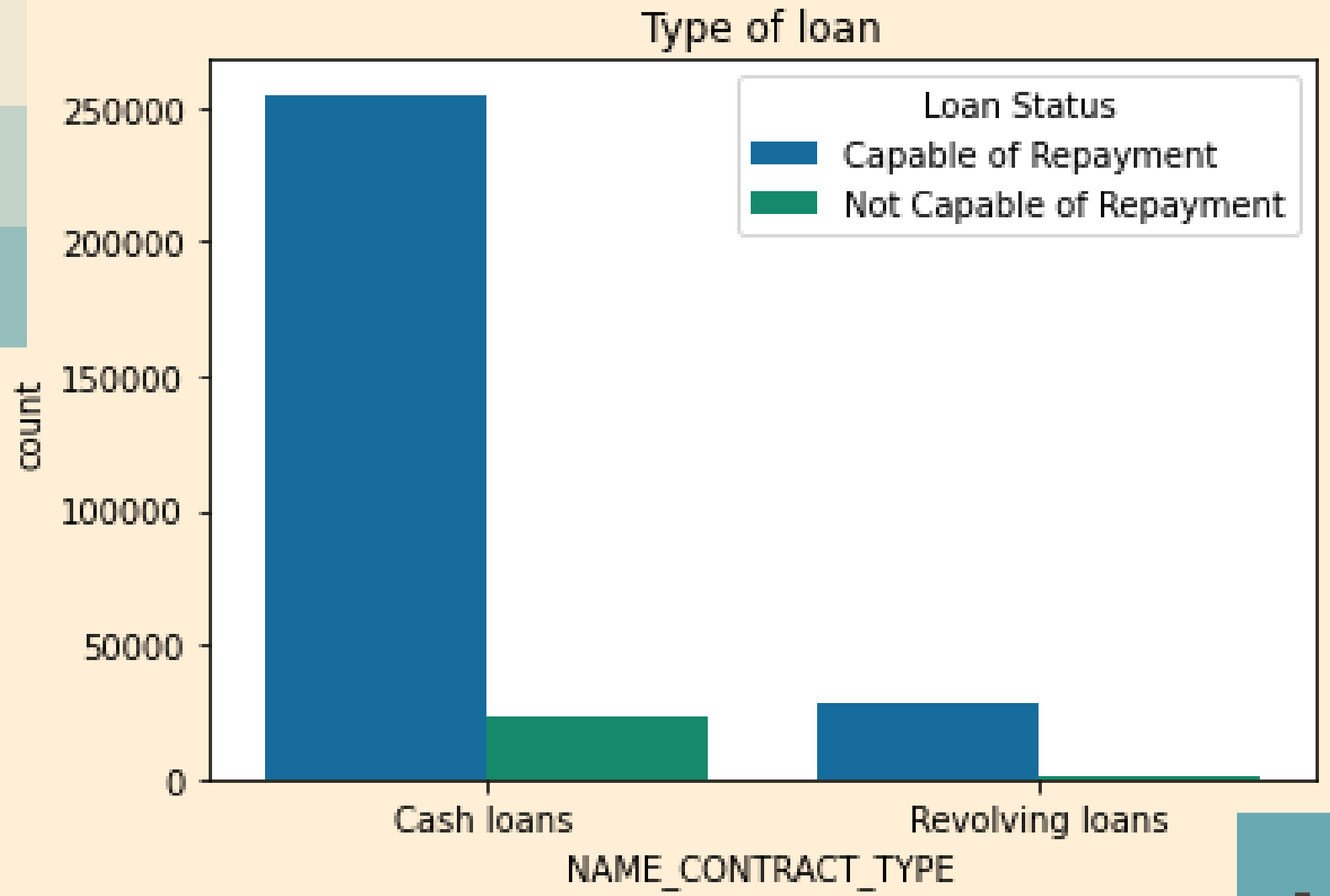
Data shape : (48744, 121)
Data type:
float64 : 65
int64 : 40
object : 16

Client Loan Status



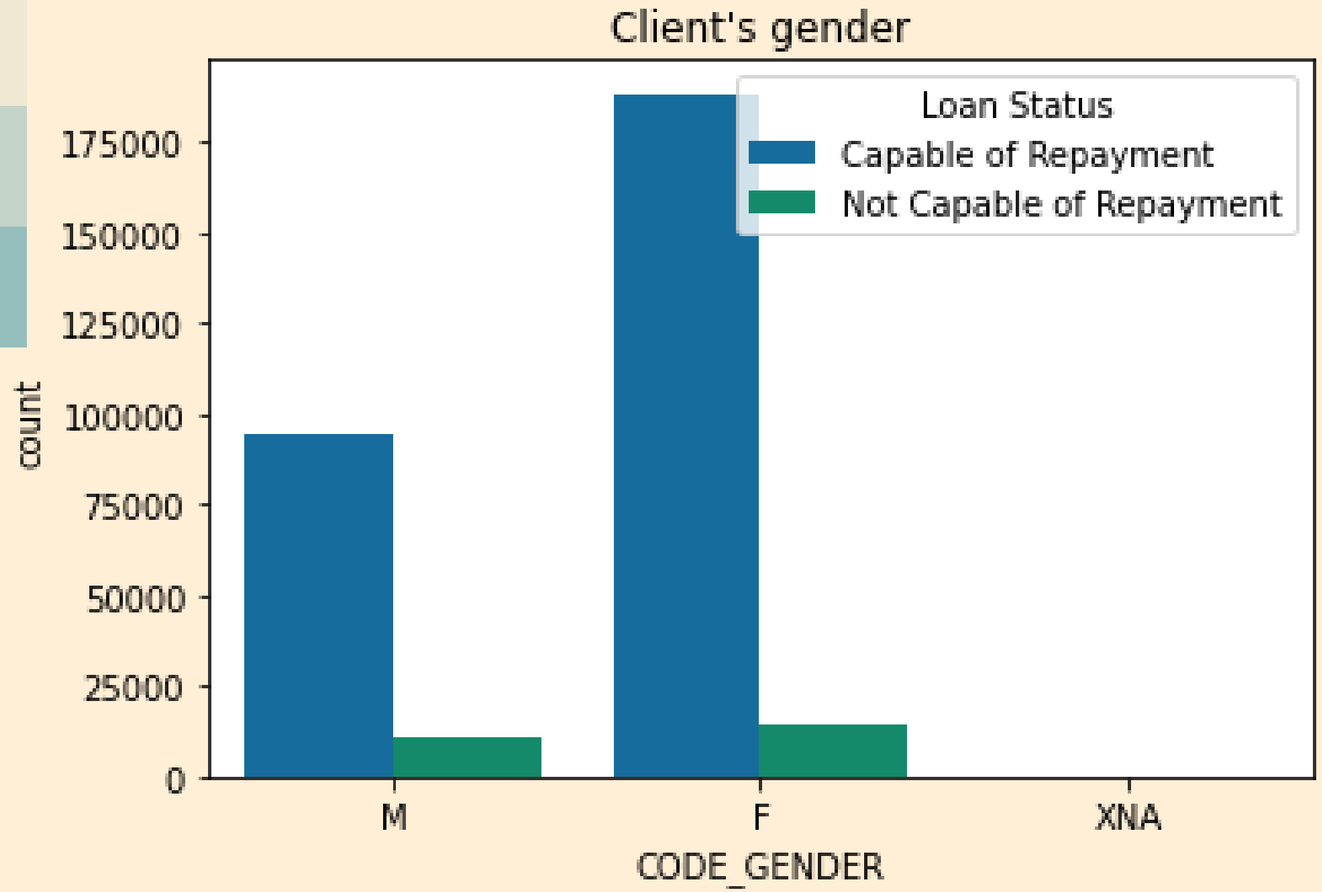
Data Visualization

Type of Loan



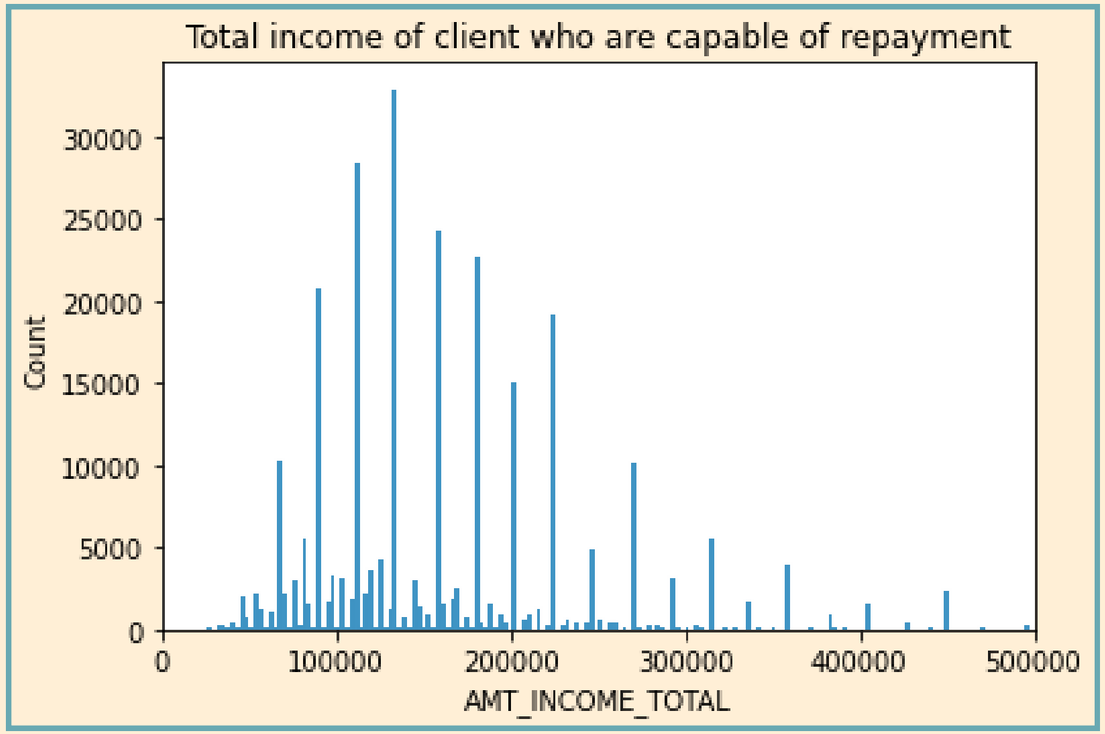
Type of loan	TARGET (Not Capable of Repayment)	TARGET (Capable of Repayment)	Total
Cash loans	23.221	255.011	278.232
Revolving loans	1.604	27.675	29.279

Client's Gender

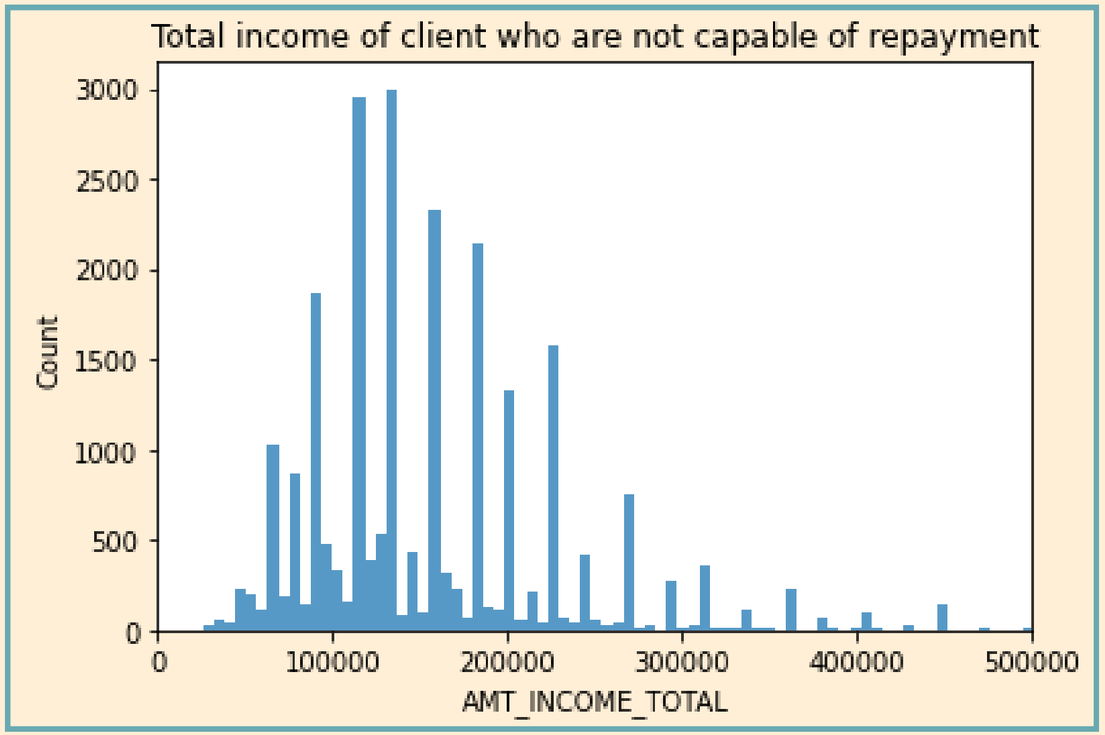


Gender	TARGET (Not Capable of Repayment)	TARGET (Capable of Repayment)	Total
Female	14.170	188.278	202.448
Male	10.655	94.404	105.059
XNA	0	4	4

Client Income

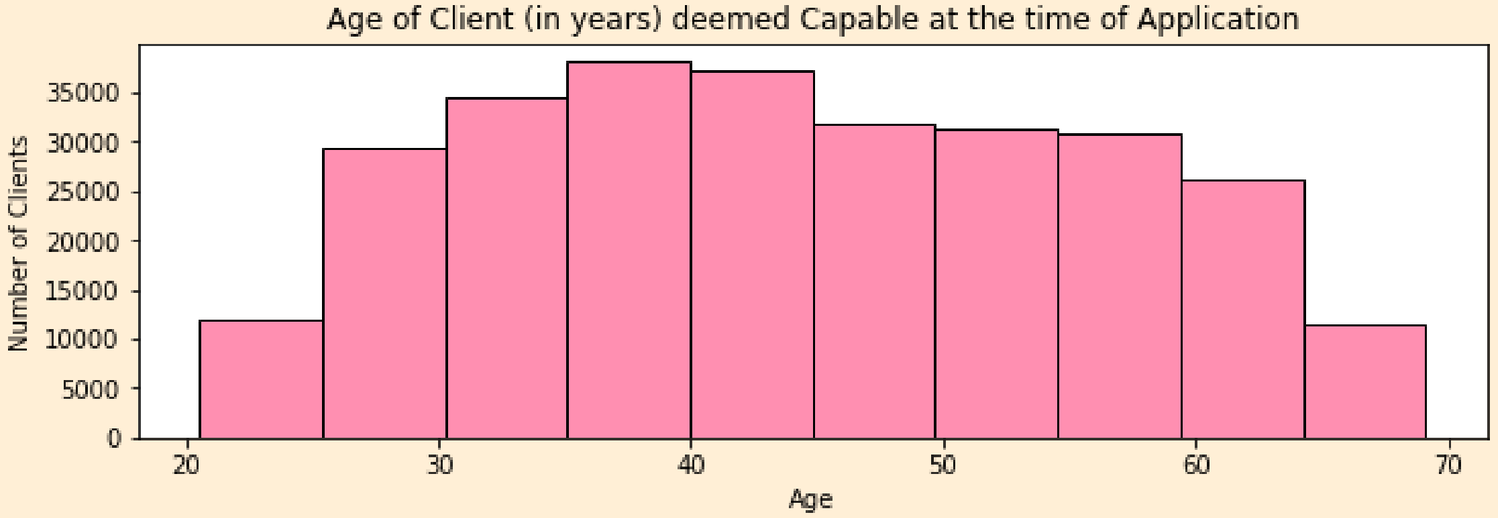


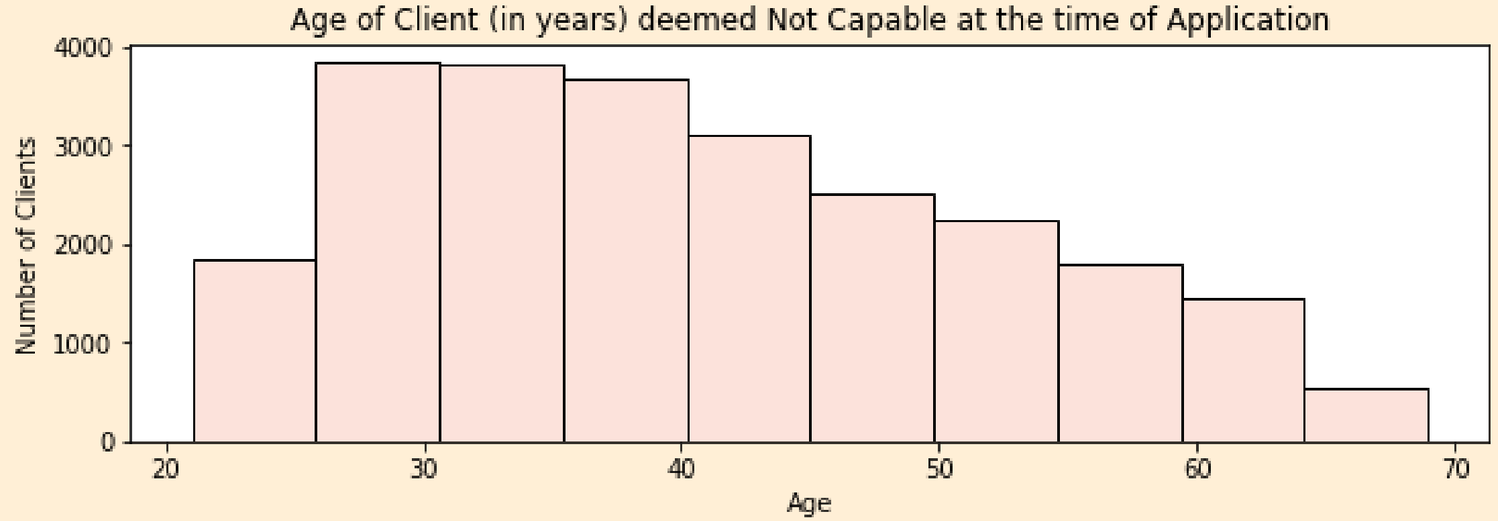
Mean : 169.077,722



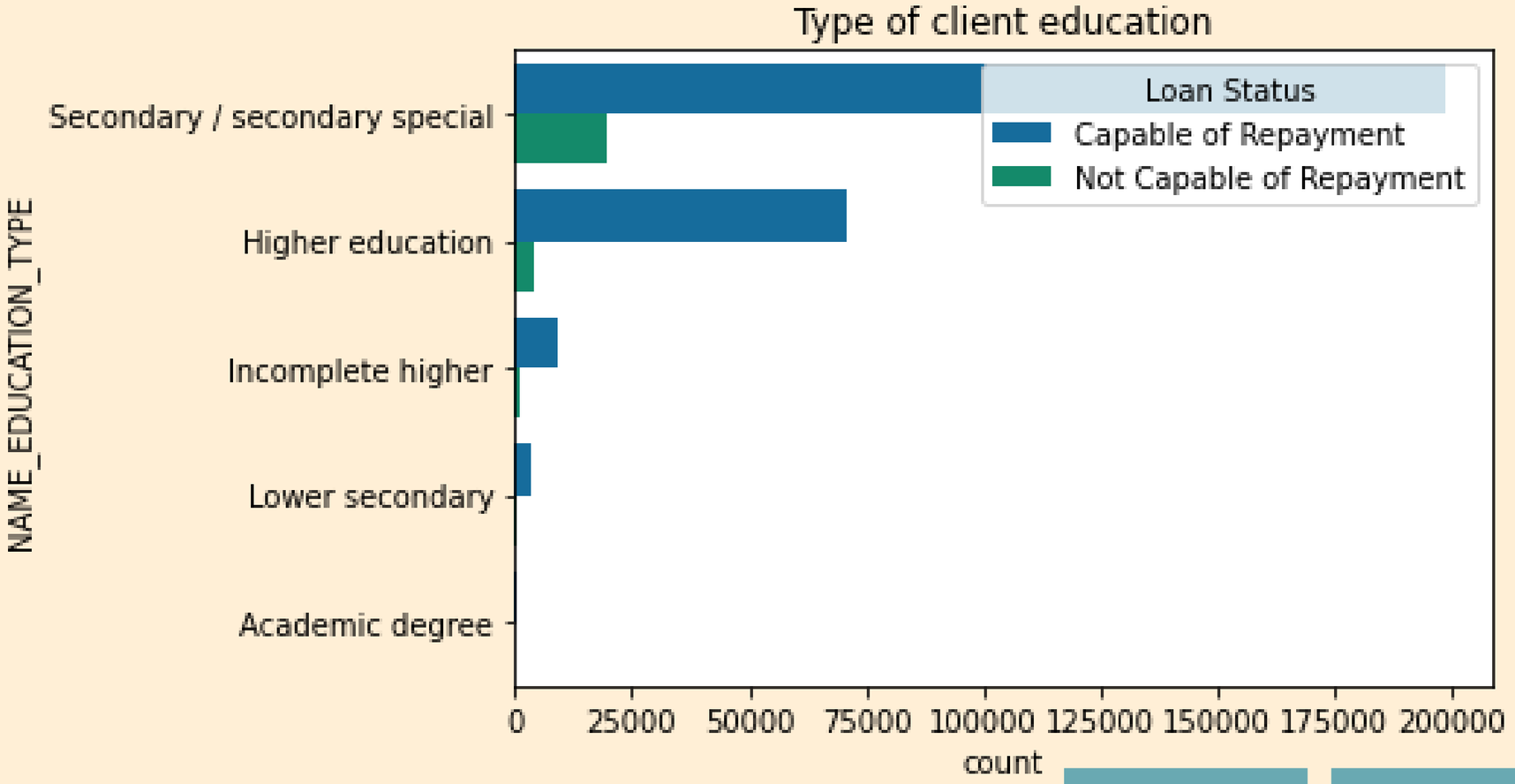
Mean : 165.611,761

Age


Age of Client (in years) deemed Capable at the time of Application
Number of Clients
0
5000
10000
15000
20000
25000
30000
35000
20
30
40
50
60
70
Age

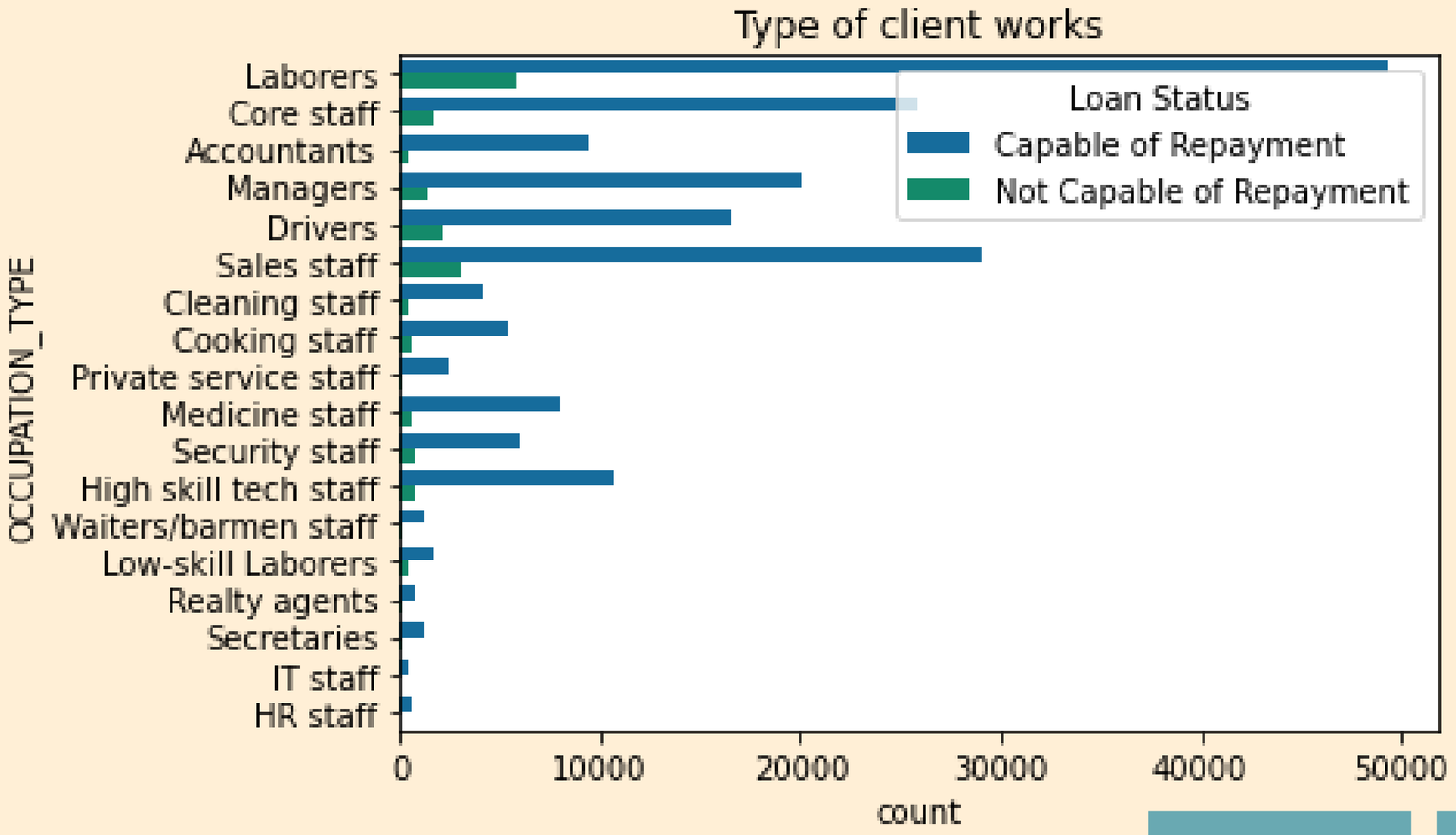

Age of Client (in years) deemed Not Capable at the time of Application
Number of Clients
0
1000
2000
3000
4000
20
30
40
50
60
70
Age

Education



Education	TARGET (Not Capable of Repayment)	TARGET (Capable of Repayment)	Total
Secondary	19.524	198.867	218.391
Academic degree	3	161	164

Work



Work	TARGET (Not Capable of Repayment)	TARGET (Capable of Repayment)	Total
Laborers	5.838	49.348	55.186
IT staff	34	491	526

Verify data quality

5 variabel dengan missing values terbanyak:

	Total	% of missing values
COMMONAREA_MEDI	92646	69.872
COMMONAREA_AVG	92646	69.872
COMMONAREA_MODE	92646	69.872
NONLIVINGAPARTMENTS_MODE	93997	69.433
NONLIVINGAPARTMENTS_AVG	93997	69.433

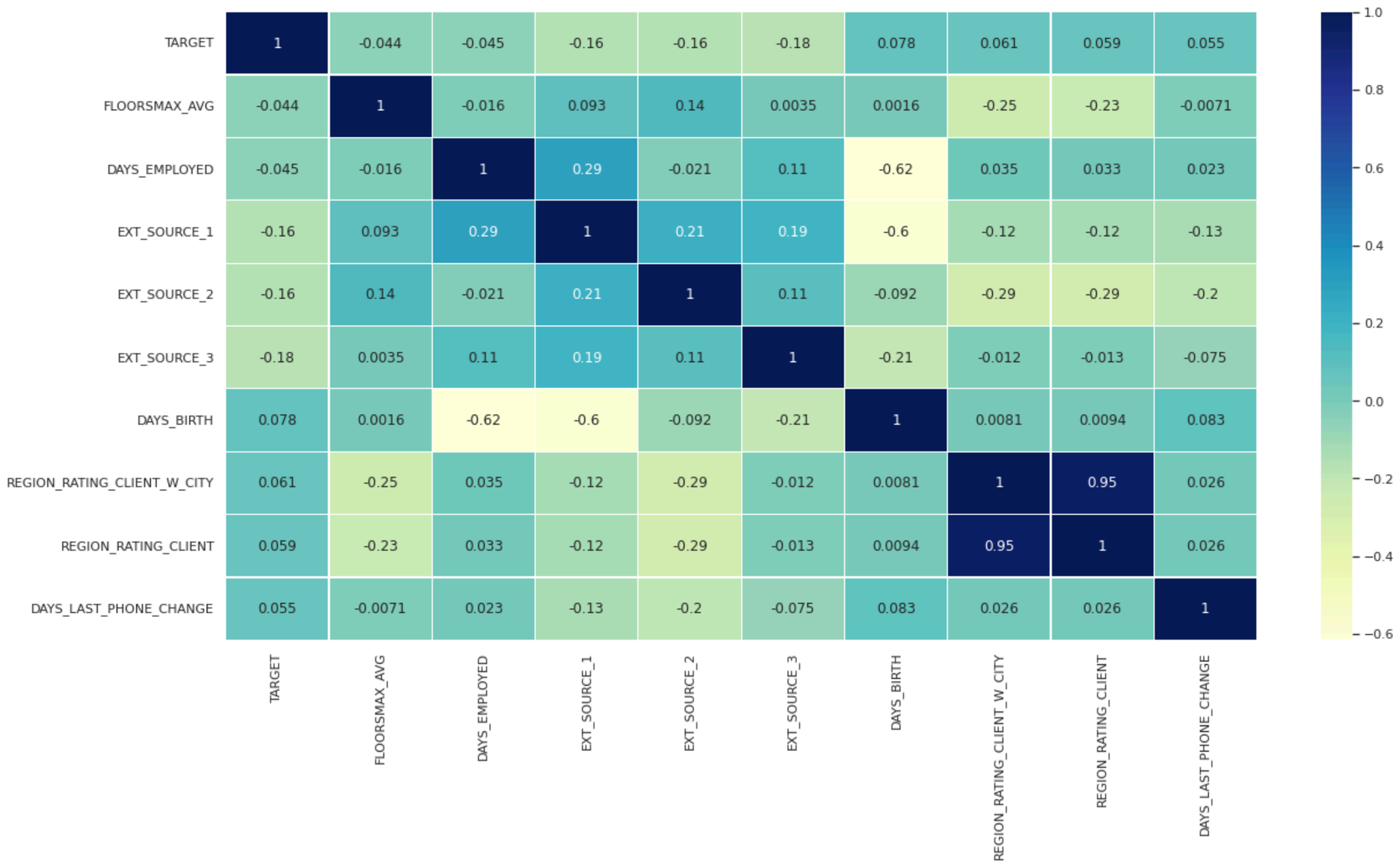
Jumlah variabel yang terdapat missing values

```
missing_val = temp_df[temp_df['% of missing values'] > 0.000]
```

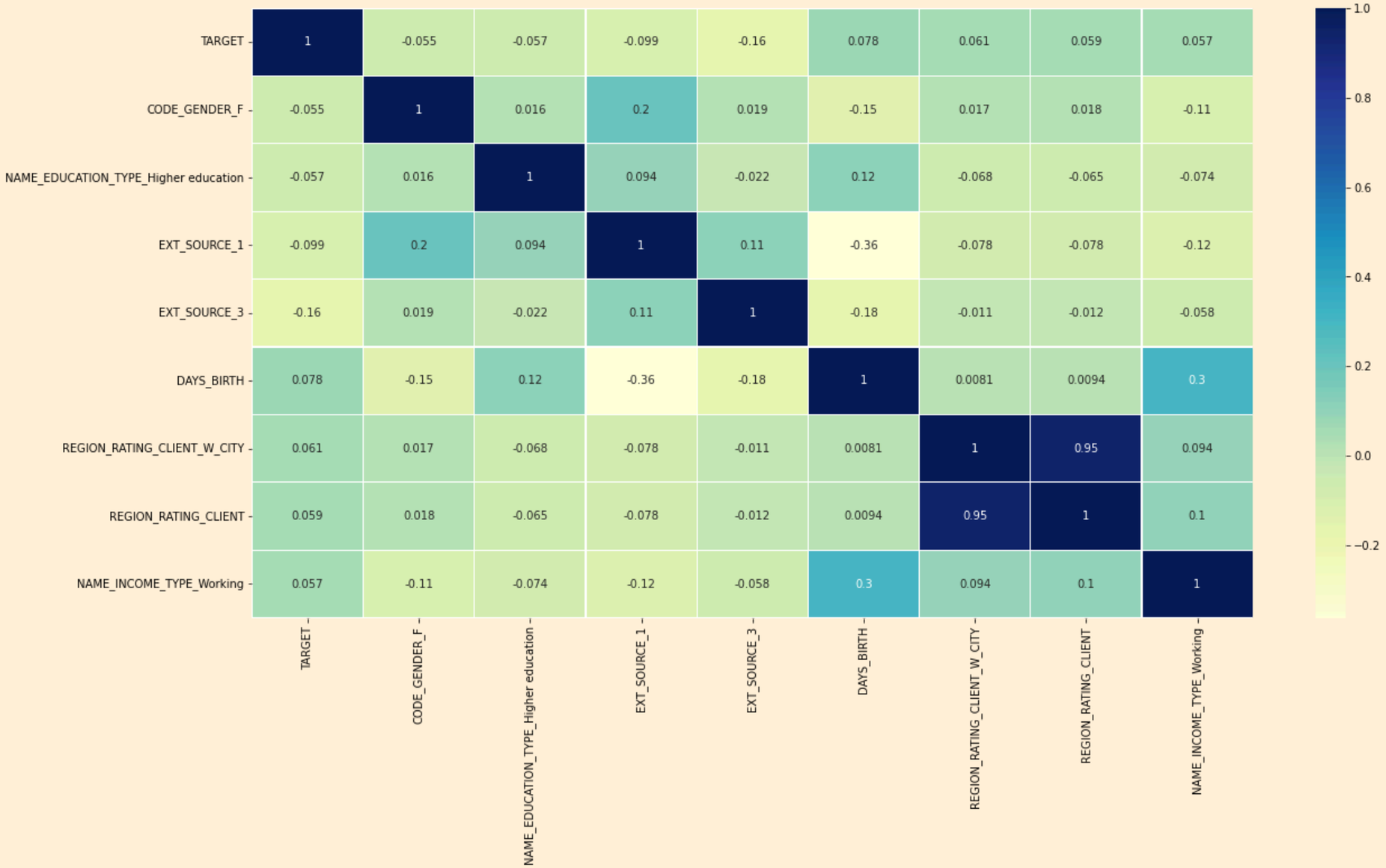
```
missing_val.count()
```

```
Total                  67
% of missing values    67
dtype: int64
```

Correlation Before Data Cleaning



Correlation After Data Cleaning



DATA PREPARATION

STEP

- Joining Dataset
- Label Encoding and One-Hot Encoding
- Feature Engineering
- Missing Values and Feature Scaling

JOINING DATASET

Dalam step joining dataset digunakan fungsi align untuk memastikan bahwa fitur yang hanya ada di kedua kerangka data digabungkan menjadi kerangka data Pandas baru **application_full.**

application_full

data shape (356255, 121)

LABEL ENCODING DAN ONE-HOT ENCODING

Machine Learning Algorithms biasanya hanya dapat memiliki nilai numerik sebagai variabel prediktornya. Oleh karena itu dilakukan Label Encoding dan One Hot Encoding untuk mengkodekan label kategorikal dengan nilai antara 0 dan 1. Setelah dilakukan pengodean fitur, jumlah fitur dalam dataset meningkat.

Size of Full Encoded Dataset (356255, 243)

FEATURE ENGINEERING

Pada step ini dilakukan feature engineering terhadap variabel FLAG_DOCUMENTS untuk melihat apakah variabel ini dapat dihapus dari dataset train dengan melihat korelasi antara variabel FLAG_DOCUMENT dan variabel TARGET.

Setelah dilakukan pengecekan korelasi, 4 Korelasi terbesar terdapat pada variabel FLAG_DOCUMENT_3, FLAG_DOCUMENT_6, FLAG_DOCUMENT_16, dan FLAG_DOCUMENT_13. Oleh karena itu, kami hanya akan menyimpan keempat variabel tersebut dalam training dataset dan membuang 16 fitur FLAG_DOCUMENT lainnya.

FLAG 3	**0,044346**
FLAG 6	**-0,028602**
FLAG 16	**-0,011615**
FLAG 13	**-0.011583**

MISSING VALUE AND FEATURE SCALLING

Pada dataset application_full terdapat 40,99% fitur memiliki lebih dari 50% *missing value*. Oleh karena itu, digunakan median untuk mengisi *missing value* tersebut.

	% of Total Values
COMMONAREA_AVG	69.714
COMMONAREA_MEDI	69.714
COMMONAREA_MODE	69.714
NONLIVINGAPARTMENTS_MEDI	69.293
NONLIVINGAPARTMENTS_MODE	69.293
NONLIVINGAPARTMENTS_AVG	69.293
LIVINGAPARTMENTS_MODE	68.204
LIVINGAPARTMENTS_AVG	68.204
LIVINGAPARTMENTS_MEDI	68.204
FLOORSMIN_AVG	67.678
FLOORSMIN_MEDI	67.678
FLOORSMIN_MODE	67.678

	% of Total Values
FLAG_OWN_CAR	0.000
ORGANIZATION_TYPE_Agriculture	0.000
OCCUPATION_TYPE_Sales staff	0.000
OCCUPATION_TYPE_Secretaries	0.000
OCCUPATION_TYPE_Security staff	0.000

MODELING

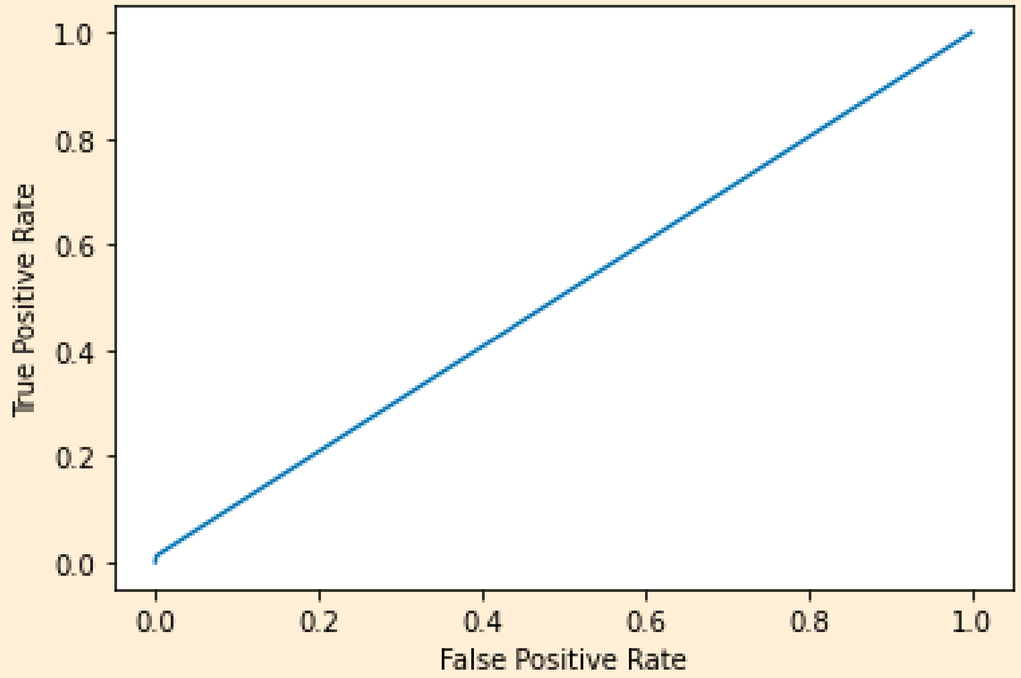
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Decison Tree Classifier HyperParameters
- Random Forest Classifier HyperParameters

LOGISTIC REGRESSION

Goals: Memprediksi probabilitas client yang kesulitan membayar (non capable) dan client yang dapat membayar pinjaman (capable) berdasarkan nilai-nilai variabel yang ada. Target merupakan variabel respons yang menjadi analisisnya.

Capable: 92069

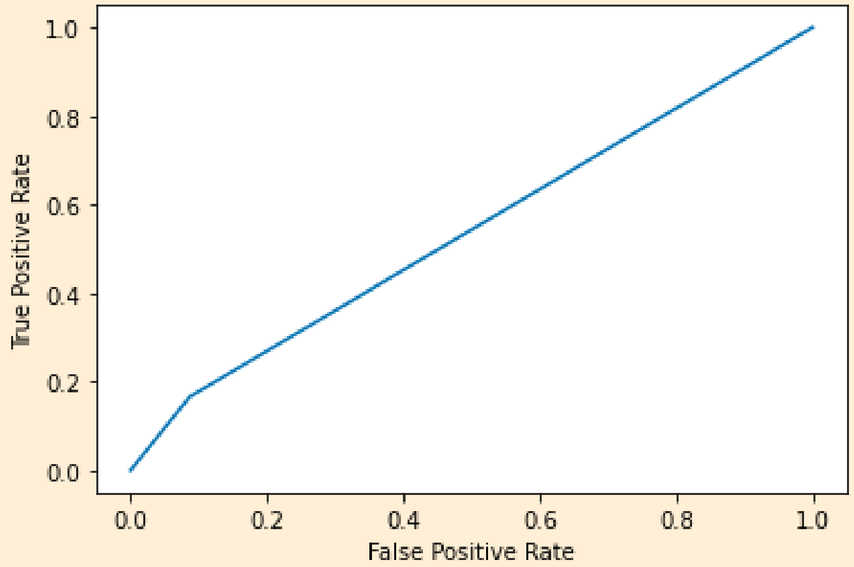
Non Capable: 185



DECISION TREE

Capable: 83700

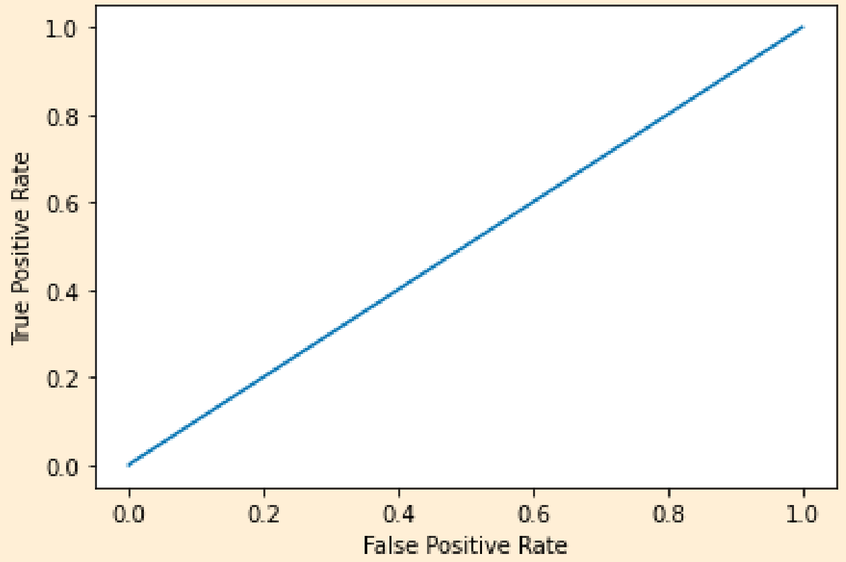
Non Capable: 8554



RANDOM FOREST

Capable: 92236

Non Capable: 18

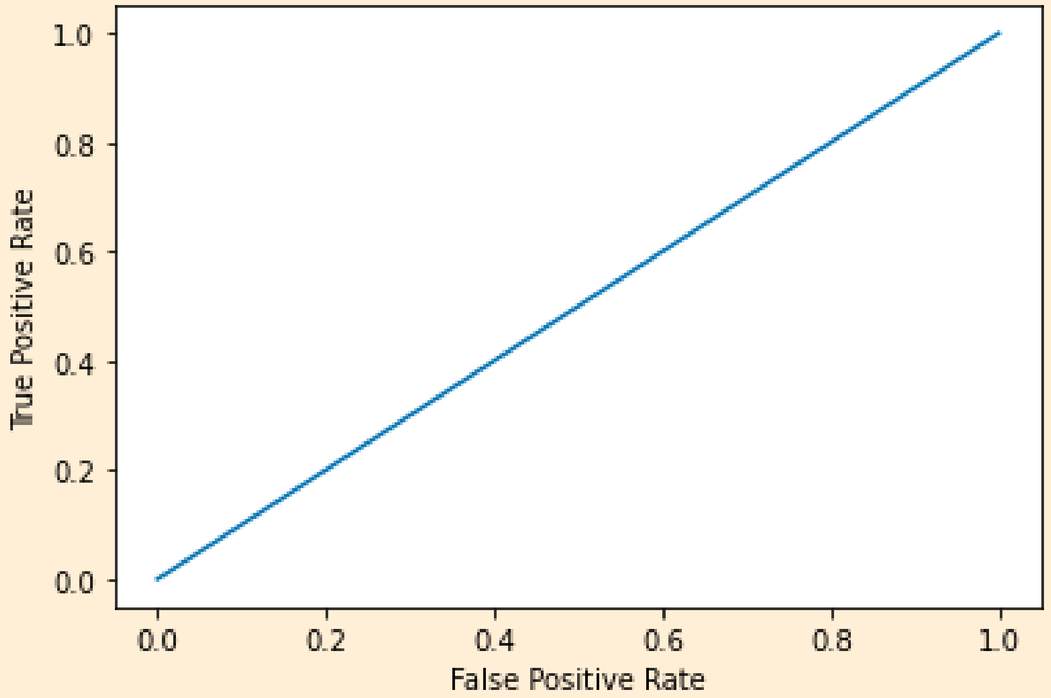


DECISION TREE HYPERPARAMETERS

Goals: Untuk menghindari overfitting, perlu membatasi parameter decision tree dalam training data.

Capable: 92254

Non Capable: 0

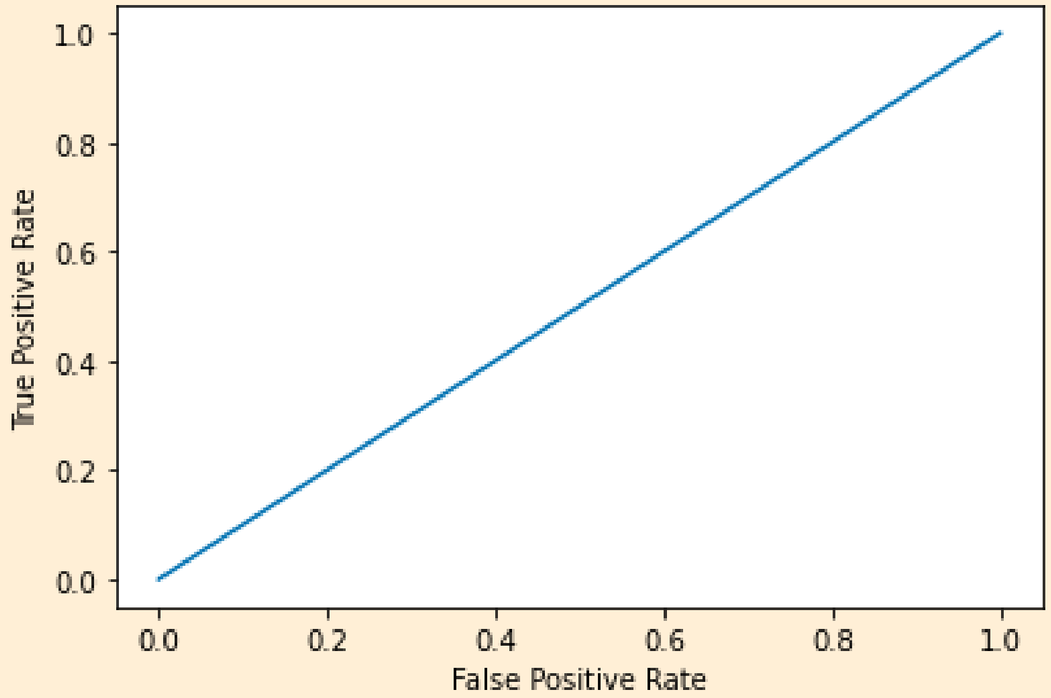


RANDOM FOREST HYPERPARAMETERS

Goals: Untuk menghindari Overfitting, running dengan peningkatan jumlah estimator.

Capable: 92253

Non Capable: 1



MODEL EVALUATION

measure the performance of all the model

ROC AUC score

Model	ROC AUC
Logistic Regression	0.505
Default Decision Tree	0.539
Default Random Forest	0.500
Tuned Decision Tree	0.500
Tuned Random Forest	0.500

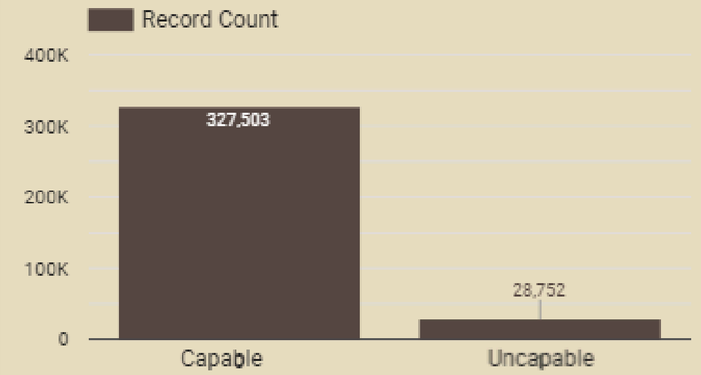
Accuracy , F1 Score, Precision, Recall

Model	Accuracy	F1 Score	Precision	Recall
Logistic Regression	0.921	0.885	0.464	0.013
Decision Tree	0.856	0.861	0.141	0.163
Random Forest	0.921	0.884	0.929	0.002
Tuned Decision Tree	0.921	0.884	0.000	0.000
Tuned Random Forest	0.921	0.884	0.000	0.000

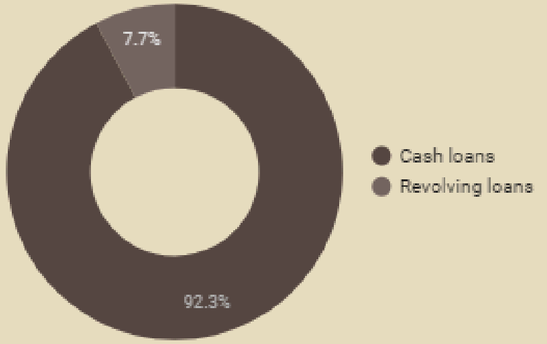
KELOMPOK 6

FINAL PROJECT DASHBOARD
HOME CREDIT RISK SCORING

How many clients are labeled Capabled?



What's the top loans type?



AGE AVERAGE	INCOME AVERAGE
43.93	$356,255.00

TARGET

EDUCATION TYPE

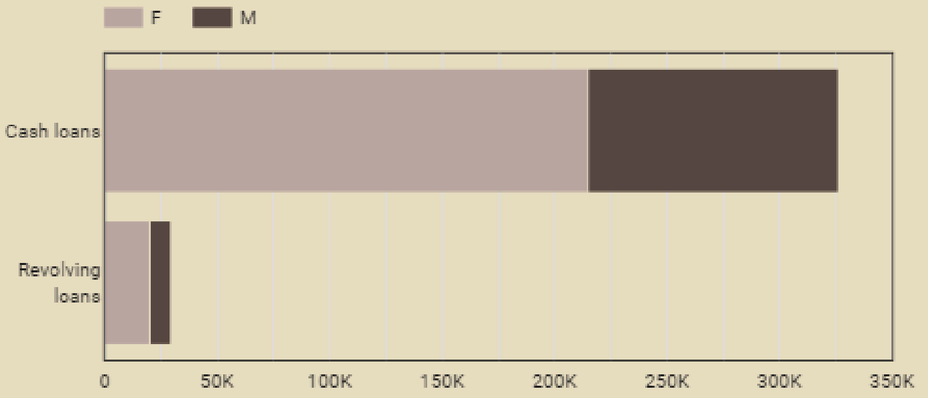
FAMILY STATUS

Top 8 Occupation by Income Client

	OCCUPATION TYPE	INCOME
1.	Waiters/barmen staff	$1,526
2.	Security staff	$7,636
3.	Secretaries	$1,518
4.	Sales staff	$37,174
5.	Realty agents	$889
6.	Private service staff	$3,107
7.	Medicine staff	$9,853
8.	Managers	$24,945

1 - 10 / 18

Type of Loan by Gender



CONCLUSION

Berdasarkan hasil yang didapatkan di dalam dashboard dapat disimpulakan jumlah client yang capable jauh lebih banyak daripada client yang non capable, dan dari gender dapat dilihat bahwa client home credit lebih banyak wanita dibandingkan pria. Dengan tipe pekerjaan yang paling pendapatan paling banyak terdapat pada pelayan (waiters).
Dari Model yang sudah dilakukan dapat disimpulkan bahwa model Logistic Regression menunjukkan performa model terbaik ditinjau dari model evaluasi ROC dan AUC score yaitu 0.505, Accuracy 0.921, dan F1 score 0.88