Rakamin Academy X HOME CREDIT

SCORECARD MODEL

Home Credit Data Scientist Virtual Internship Program Batch November 2023

Presented by
Limatan Luviar

Table Of Content

Project Research

Project Background

Banyak orang yang kesulitan mendapatkan pinjaman karena riwayat kredit yang kurang atau tidak ada. Dan, sayangnya, populasi ini sering kali dimanfaatkan oleh pemberi pinjaman yang tidak dapat dipercaya. Home Credit berupaya untuk memperluas inklusi keuangan bagi masyarakat yang tidak memiliki rekening bank dengan memberikan pengalaman pinjaman yang positif dan aman. Untuk memastikan masyarakat yang belum terlayani ini mendapatkan pengalaman pinjaman yang positif, Home Credit menggunakan berbagai data alternatif - termasuk data telekomunikasi dan informasi transaksional - untuk memprediksi kemampuan pembayaran nasabah.

Define Problem

Problem yang ingin diselesaikan dari dataset ini ,yaitu Mengidentifikasi karakteristik nasabah potensial yang akan mengalami kesulitan dalam membayar pinjaman dan siapa yang tidak Misalnya, apakah ada pola tertentu dalam data nasabah yang gagal membayar pinjaman, seperti riwayat kredit buruk, penghasilan rendah, atau faktor-faktor lain seperti status pekerjaan, usia, atau jumlah tanggungan keluarga. Mengidentifikasi dan memahami nasabah potensial yang mungkin mengalami kesulitan dalam melunasi pinjaman serta memprediksi risiko pembayaran untuk meminimalkan kerugian potensial.

Data Source

Data yang digunakan adalah application train dan application test.

Goals

- Menentukan prospek mana yang akan kesulitan membayar kembali pinjamannya dan mana yang tidak.
- Memprediksi kemampuan membayar pelanggan.

Objective

- Lakukan pembersihan dan visualisasi dataPenampilan.
- Membangun model menggunakan algoritma pembelajaran mesin.
- Memberikan saran kepada perusahaan untuk meningkatkan tingkat keberhasilan pengajuan pinjaman nasabah

Data Preprocessing

Rakamin Academy

Data Application_train.csv — **Jumlah kolom 122** — **Jumlah rows 307.511**

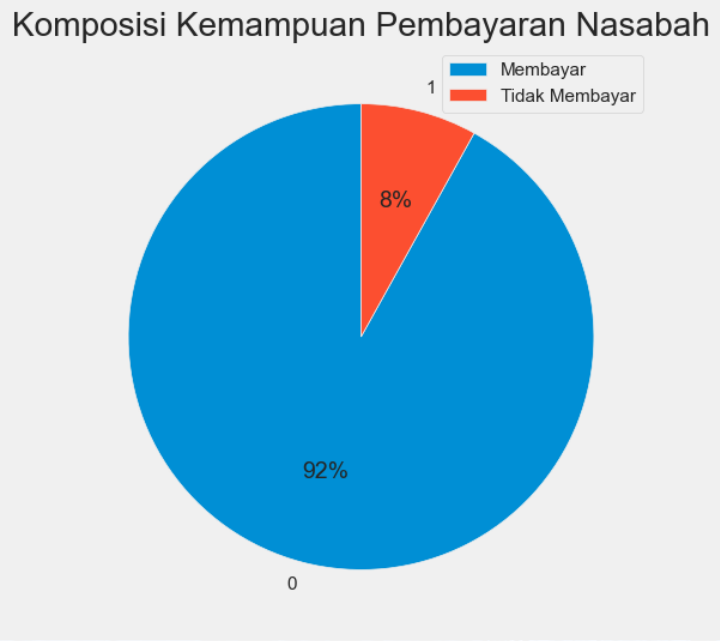
Explatory Data Analysis → **Data Cleaning** → **Model Building**

- Univariate Analysis analisis statistik yang berkaitan dengan satu variabel tunggal pada suatu waktu
- Multivariate Analysis metode statistik yang mengkaji lebih dari satu variabel secara simultan untuk memahami hubungan, pola, dan struktur yang kompleks di antara variabel-variabel tersebut

- Cek Duplikat
- Handling Missing Values
- Cek Outliers

- Label Encoding
- Feature Selection
- Handling Imbalanced Data
- Model Building
- Model Evaluation

Data Insight



:

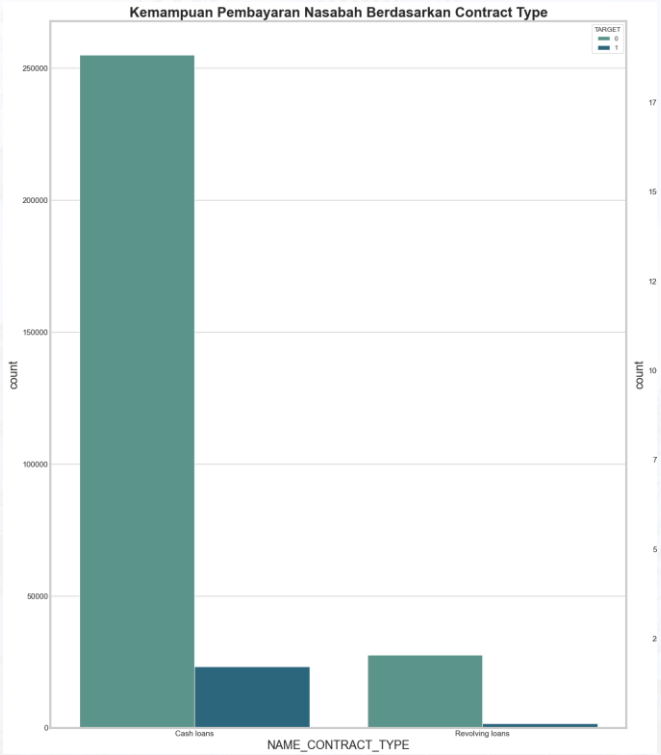
	TARGET	COUNT
0	0	282686
1	1	24825

[2]:

	index	TARGET
0	0	0.919271
1	1	0.080729

Ada sekitar 92% pinjaman yang setara dengan sekitar 282.686K dengan TARGET = 0, yang mengindikasikan bahwa nasabah tidak mengalami masalah dalam membayar kembali pinjaman dalam waktu tertentu. Sementara hanya 8% dari total pinjaman (sekitar 24.825K pemohon) dalam dataset ini yang melibatkan klien yang mengalami masalah dalam pengembalian pinjaman.

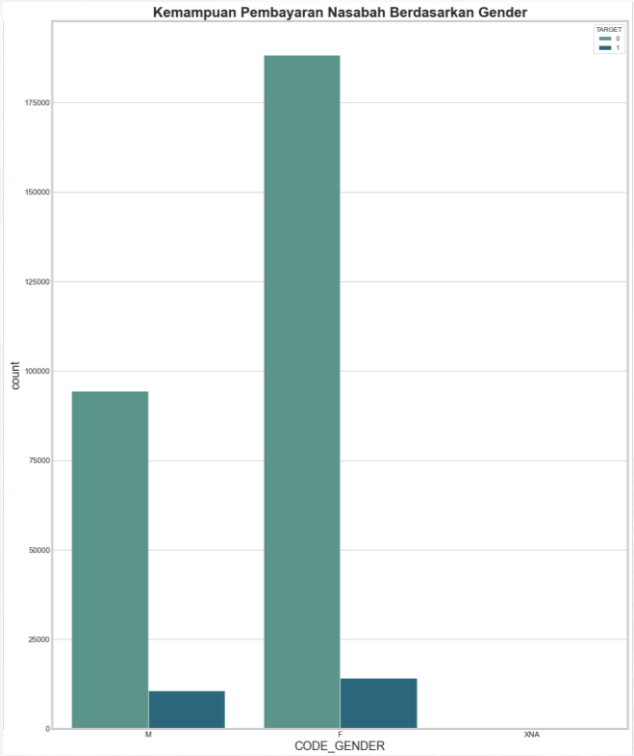
Data Insight



	NAME_CONTRACT_TYPE	TARGET	SK_ID_CURR
0	Cash loans	0	255011
2	Revolving loans	0	27675
1	Cash loans	1	23221
3	Revolving loans	1	1604

Cash Loans dengan jumlah sekitar 278 ribu pinjaman merupakan mayoritas dari total pinjaman dalam dataset ini. Revolving loans memiliki jumlah yang jauh lebih rendah, yaitu sekitar 29 ribu dibandingkan dengan pinjaman tunai

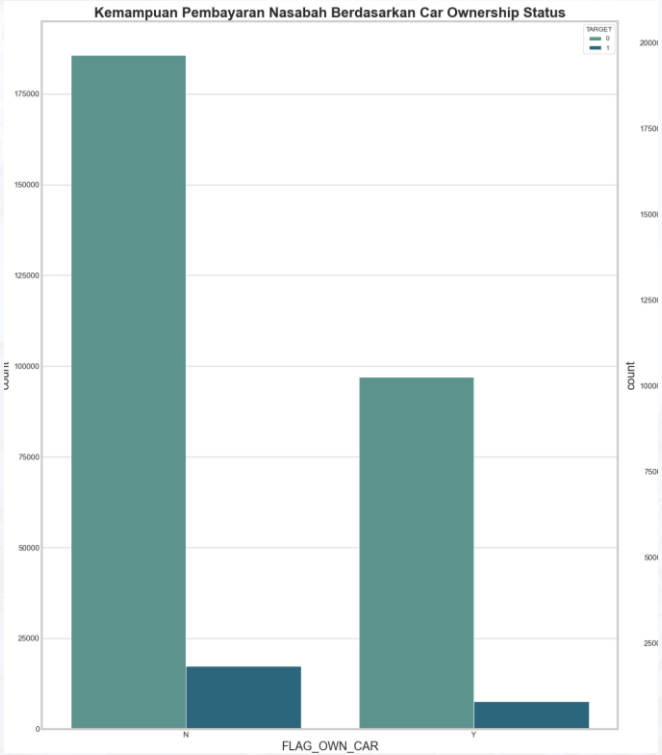
Data Insight



	CODE_GENDER	TARGET	SK_ID_CURR
0	F	0	188278
2	M	0	94404
1	F	1	14170
3	M	1	10655
4	XNA	0	4

Dapat dilihat bahwa wanita lebih banyak mengajukan pinjaman. Secara total, ada sekitar 202.448 aplikasi pinjaman yang diajukan oleh wanita, dan sekitar 105.059 pengajuan yang diajukan oleh pria.
Namun, persentase yang lebih besar (sekitar 10% dari total) pria mengalami masalah dalam membayar pinjaman dibandingkan dengan nasabah wanita (sekitar 7%).

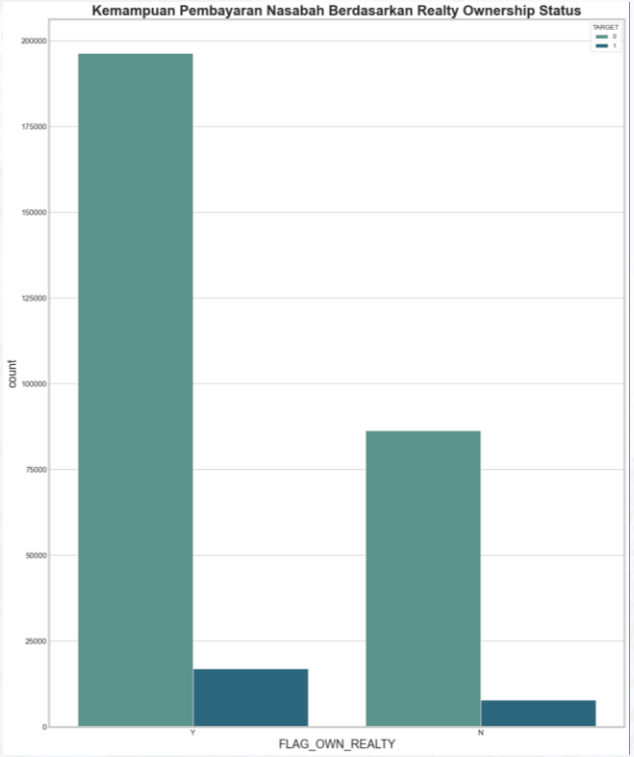
Data Insight



	FLAG_OWN_CAR	TARGET	SK_ID_CURR
0	N	0	185675
2	Y	0	97011
1	N	1	17249
3	Y	1	7576

Sebagian besar nasabah tidak memiliki mobil.
Nasabah yang memiliki mobil (sekitar 8%) memiliki masalah dalam membayar pinjaman dibandingkan dengan nasabah yang tidak memiliki mobil (sekitar 7%). Namun, perbedaannya tidak terlalu signifikan.

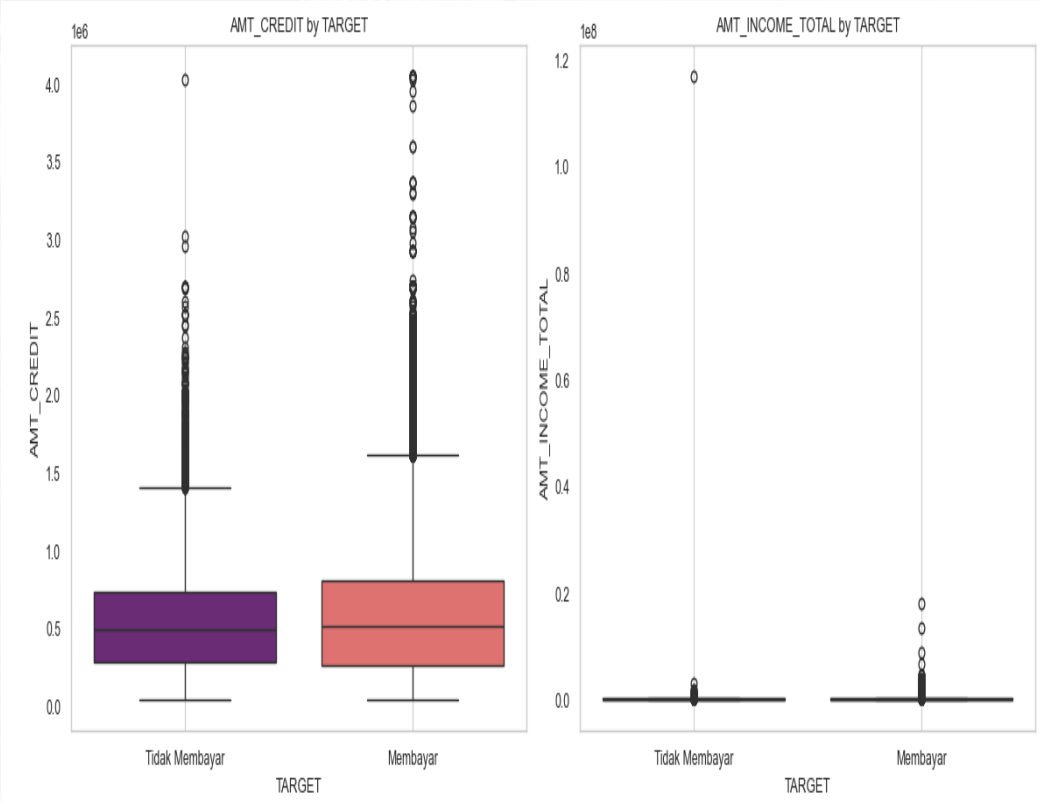
Data Insight



	FLAG_OWN_REALTY	TARGET	SK_ID_CURR
2	Y	0	196329
0	N	0	86357
3	Y	1	16983
1	N	1	7842

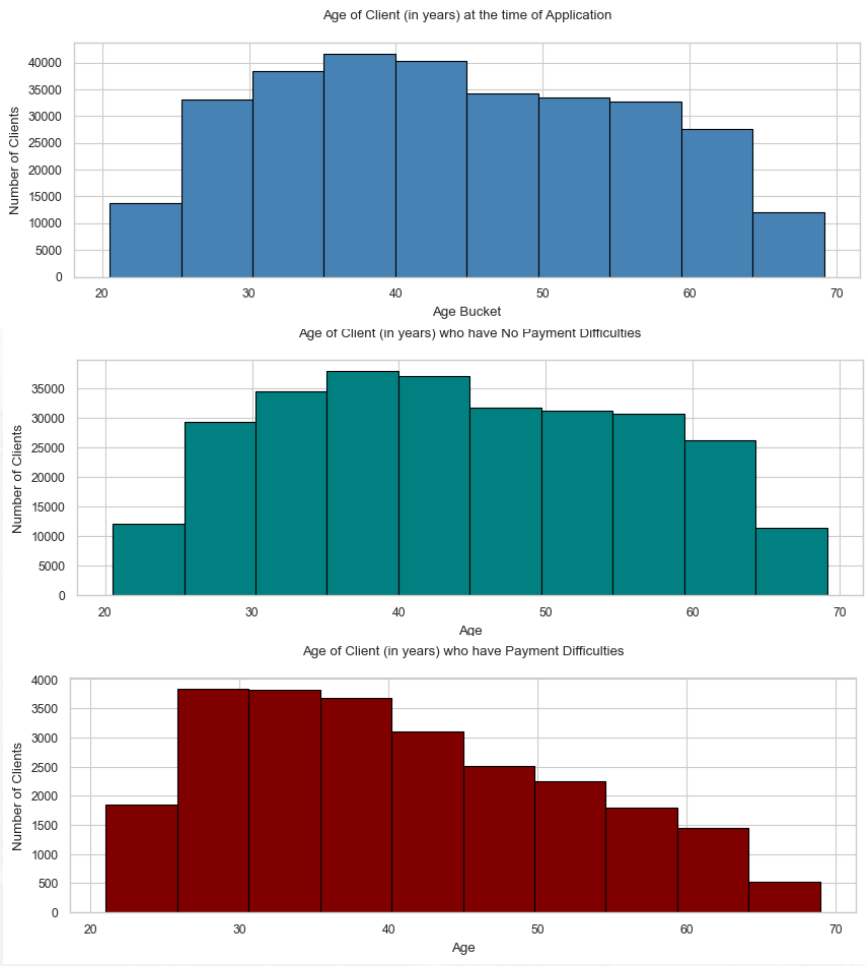
Sebagian besar nasabah memiliki rumah/apartemen. Nasabah yang memiliki rumah/flat (sekitar 8%) memiliki masalah dalam membayar pinjaman dibandingkan dengan nasabah yang tidak memiliki rumah/flat (sekitar 7%). Namun, perbedaannya tidak terlalu signifikan.

Data Insight



- nilai median dari jumlah kredit nasabah yang tidak mengalami kesulitan pembayaran sedikit lebih besar dari nilai median nasabah yang mengalami kesulitan pembayaran
- nasabah yang mengalami kesulitan pembayaran maupun nasabah yang tidak mengalami kesulitan pembayaran, sebagian besar memiliki nilai yang sama

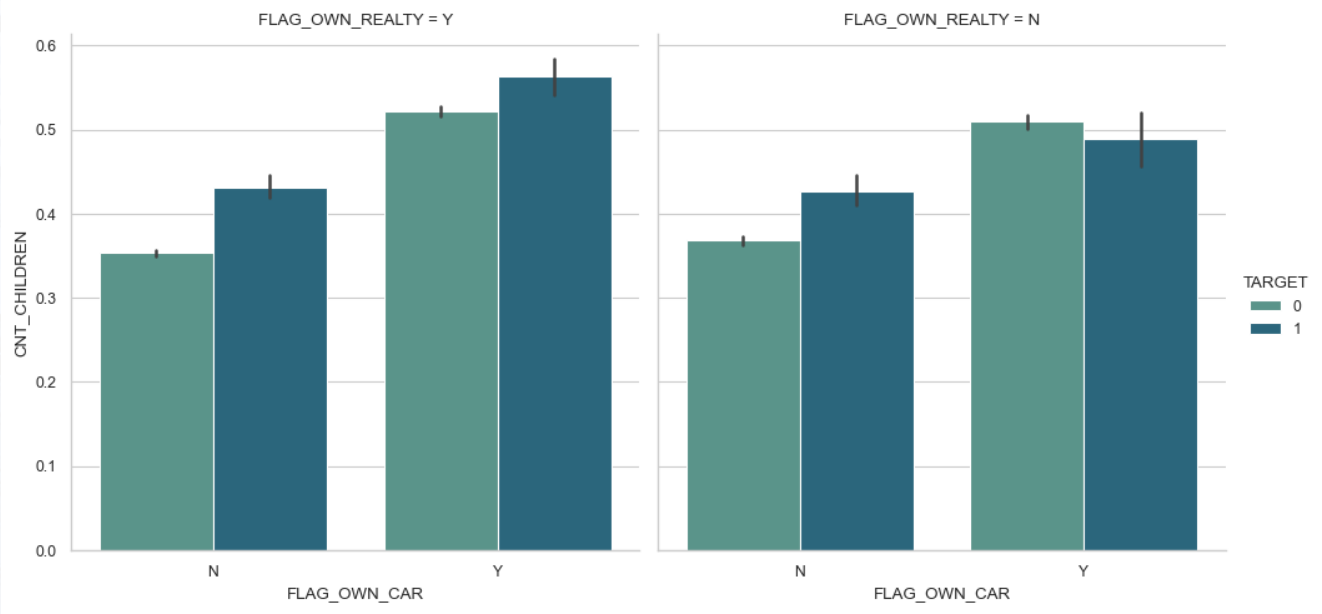
Data Insight



Sebagian besar nasabah yang mengajukan pinjaman berusia antara 35-40 tahun, diikuti oleh nasabah berusia antara 40-45 tahun. Sementara itu, jumlah pemohon untuk nasabah yang berusia <25 tahun atau usia >65 tahun sangat rendah.

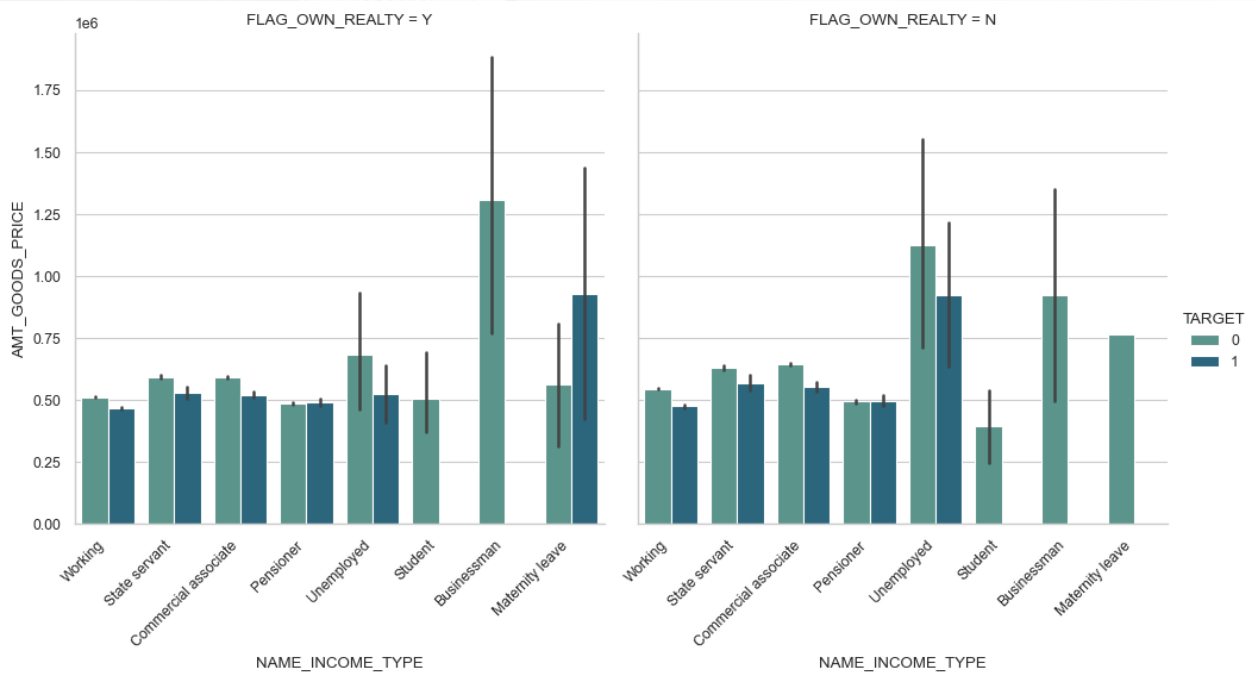
Klien yang tidak mengalami kesulitan pembayaran adalah klien dengan rentang usia 35-45 tahun. Sedangkan klien yang mengalami kesulitan pembayaran adalah klien dengan rentang usia 25-35 tahun.

Data Insight



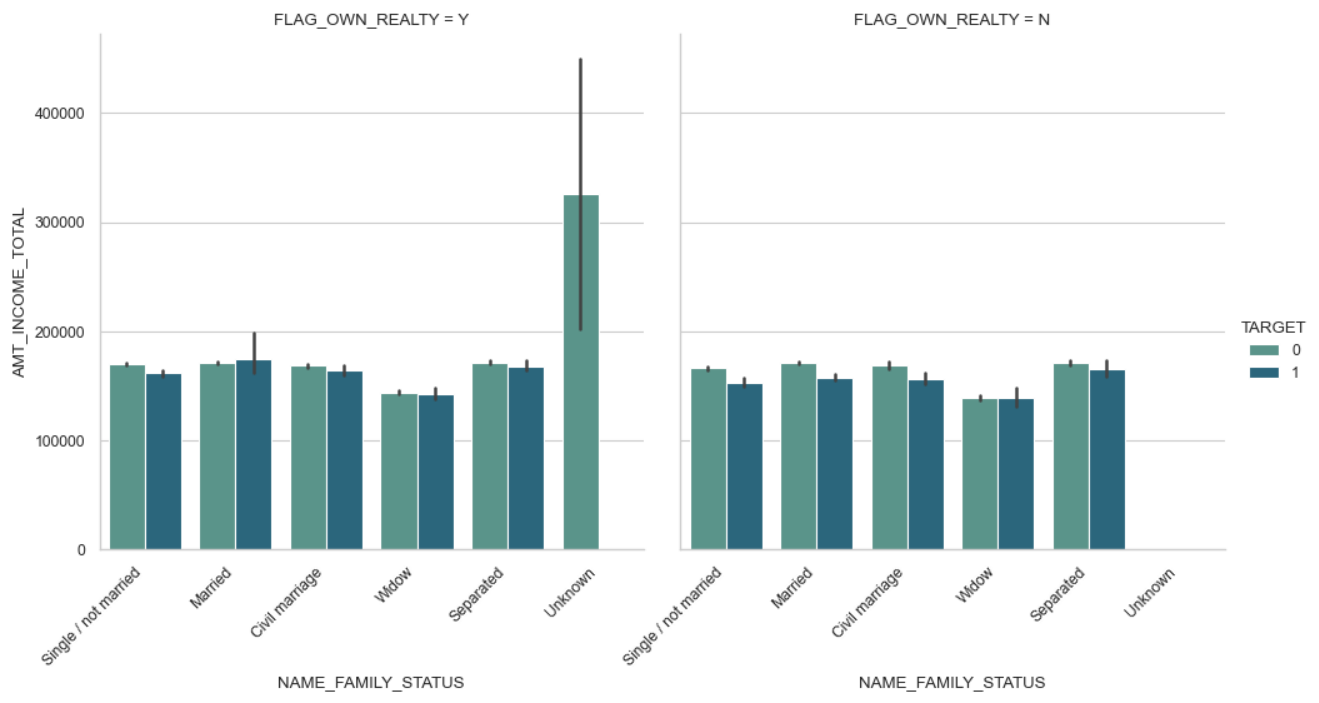
Nasabah yang memiliki mobil dan rumah/apartemen memiliki masalah dalam membayar pinjaman karena jumlah anak yang lebih banyak dibandingkan dengan nasabah yang tidak memiliki rumah/apartemen.

Data Insight



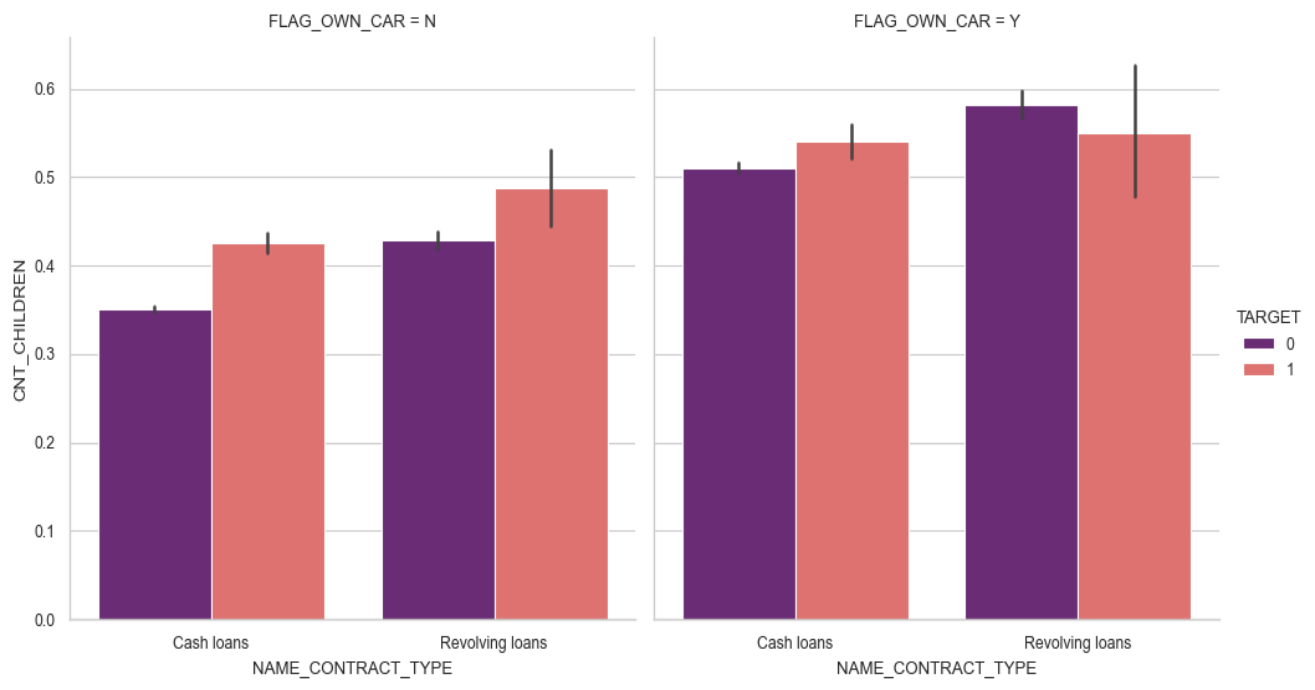
Nasabah dengan tipe pendapatan cuti melahirkan FLAG_OWN_REALTY = Ya (yaitu memiliki rumah/apartemen) mengalami kesulitan dalam membayar pinjamannya dibandingkan dengan nasabah dengan tipe pendapatan FLAG_OWN_REALTY = Tidak (yaitu tidak memiliki rumah/apartemen).

Data Insight



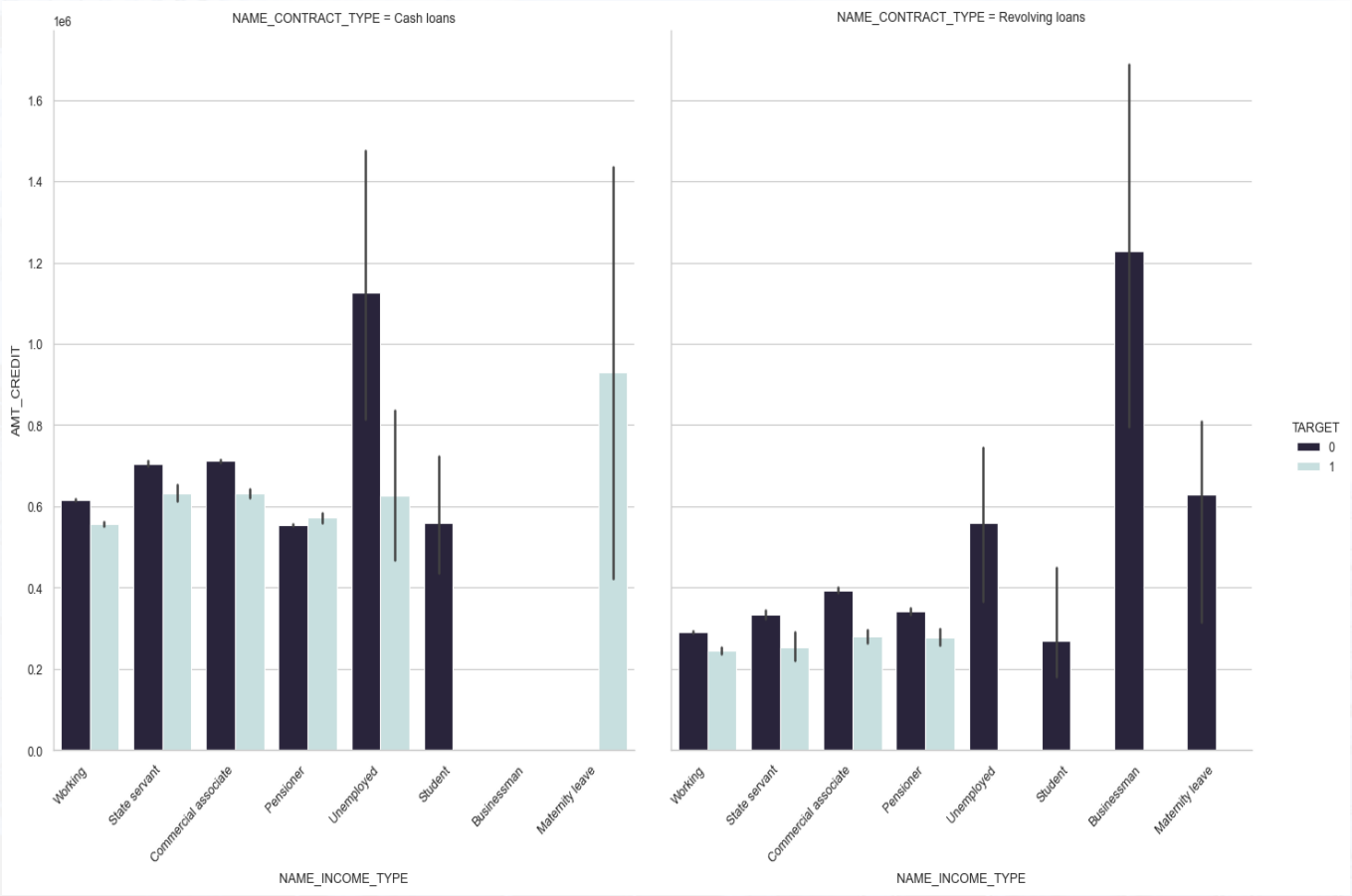
Nasabah yang sudah menikah dan memiliki rumah/apartemen (FLAG_MILIK_RUMAH = ya) mengalami kesulitan dalam membayar pinjaman berpendapatan sedang dibandingkan dengan nasabah yang tidak memiliki rumah/apartemen (FLAG_MILIK_APARTEMEN = tidak).

Data Insight



Untuk pinjaman bergulir dengan syarat FLAG_OWN_A CAR = Tidak (tidak ada mobil), lebih sulit melunasi pinjaman dibandingkan dengan syarat FLAG_OWN_MOBIL = Ya (memiliki mobil).

Data Insight

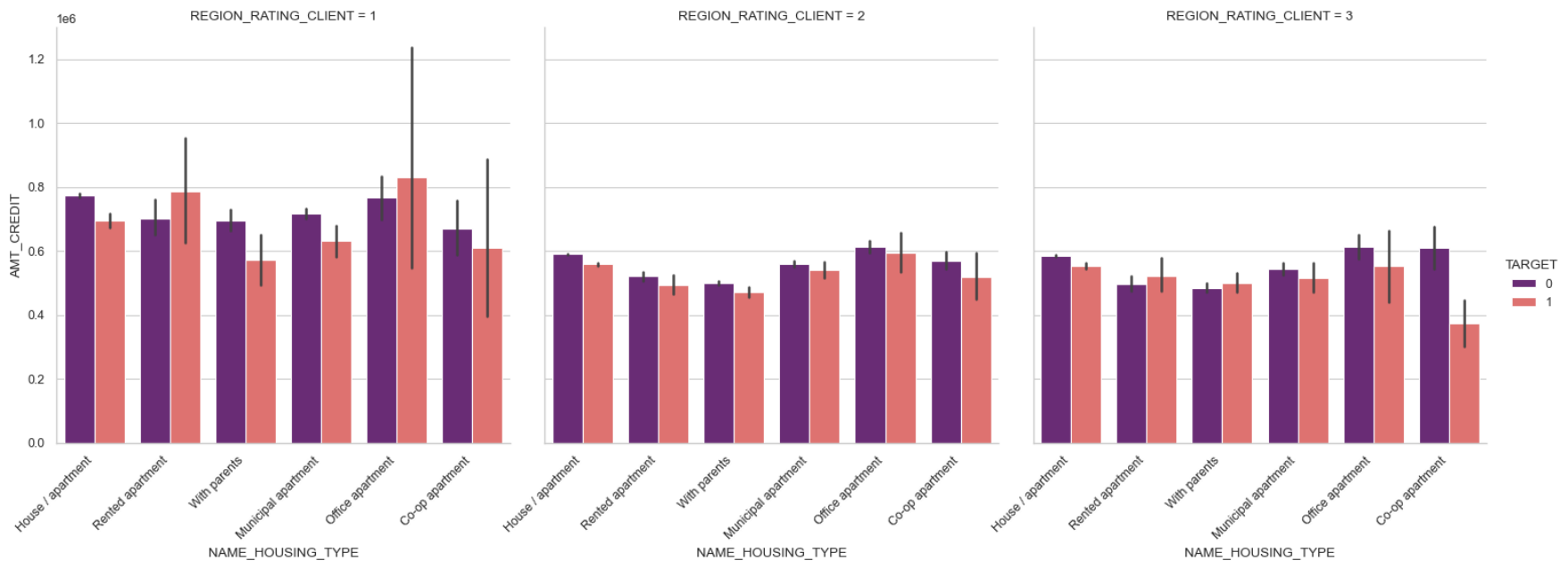


- Untuk pendapatan bersalin dengan pinjaman tunai seperti ini, semua nasabah akan kesulitan membayar pinjaman dengan batas kredit menengah. Pada saat yang sama, seluruh nasabah yang sedang cuti hamil dan pinjaman bergulir tidak mengalami kesulitan dalam membayar kembali pinjamannya

- Pada nasabah pengangguran yang memiliki pinjaman tunai, lebih dari 50% mengalami kendala pada pinjaman dengan batas kredit menengah. Sementara itu, seluruh nasabah pengangguran yang memiliki pinjaman bergulir tidak mengalami kesulitan dalam membayar kembali pinjamannya.

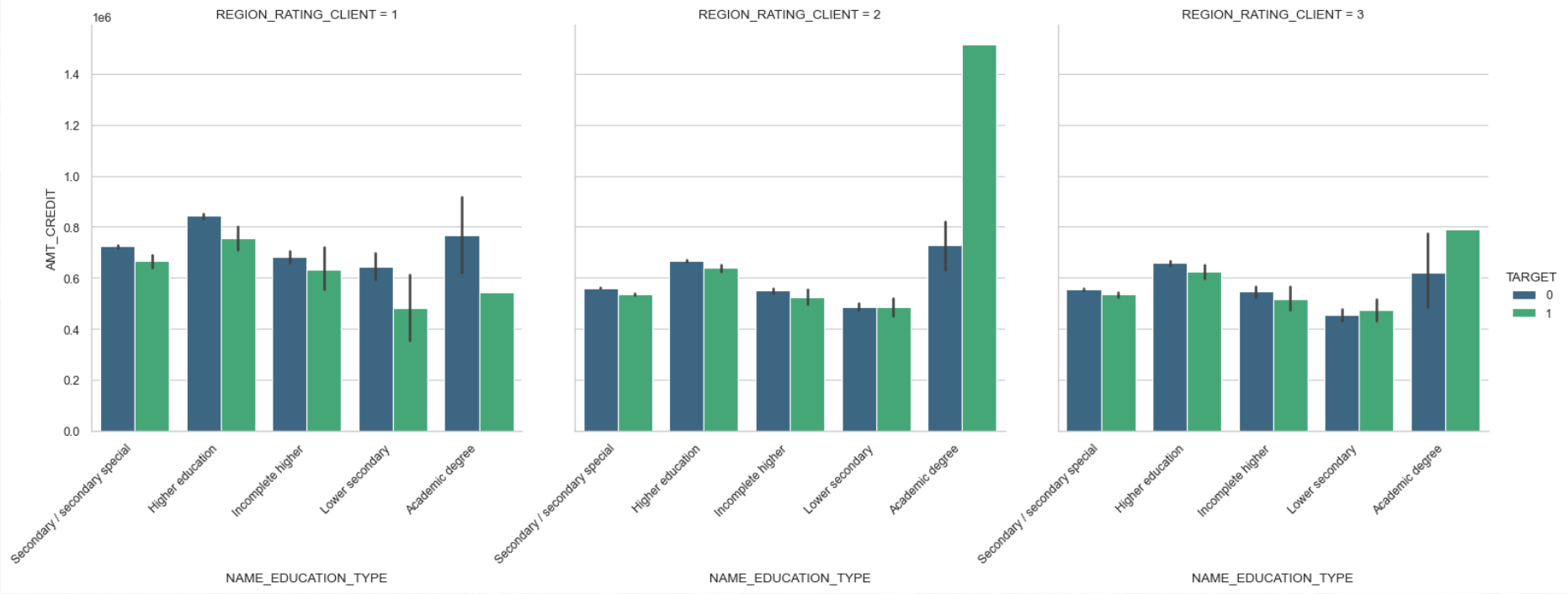
- Semua nasabah pelajar dapat dengan mudah melunasi pinjamannya melalui pinjaman tunai atau pinjaman bergulir dengan batas kredit rendah hingga menengah.

Data Insight



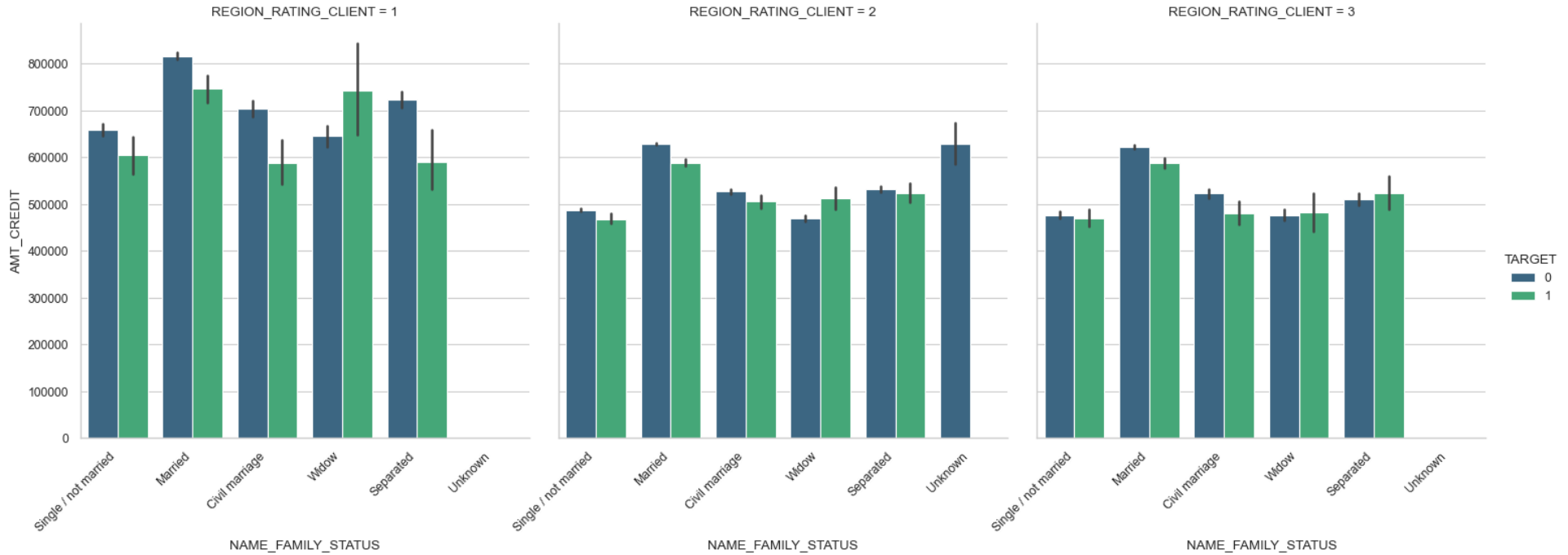
Nasabah yang tinggal di apartemen sewa atau perkantoran di wilayah dengan peringkat 1 lebih sulit membayar kembali pinjaman dengan jumlah pinjaman sedang dibandingkan dengan nasabah di wilayah dengan peringkat 2.

Data Insight



Untuk nasabah yang memiliki gelar akademis dan tinggal di wilayah dengan peringkat 2, mereka memiliki masalah dalam membayar pinjaman untuk jumlah pinjaman yang lebih tinggi. Dan, nasabah dengan gelar yang sama tetapi tinggal di wilayah dengan peringkat 3 memiliki masalah dalam membayar pinjaman untuk jumlah pinjaman menengah.

Data Insight



Nasabah yang berstatus janda/duda, baik yang tinggal di wilayah rating 1, 2, atau 3, memiliki masalah dalam membayar pinjaman untuk jumlah pinjaman sedang hingga tinggi.

Nasabah yang berstatus janda/duda, dan tinggal di daerah peringkat 3, memiliki lebih banyak masalah dalam membayar pinjaman untuk jumlah pinjaman sedang dibandingkan dengan nasabah yang tinggal di daerah peringkat 1 atau 2.

Data Modeling

Rakamin Academy

Menggunakan Logistic Regression dan Decision Tree Sebagai Model

For Train data

LogisticRegression

```
# train the model
log_m = LogisticRegression().fit(X_train, y_train)
print(log_m)
```

```
LogisticRegression()
```

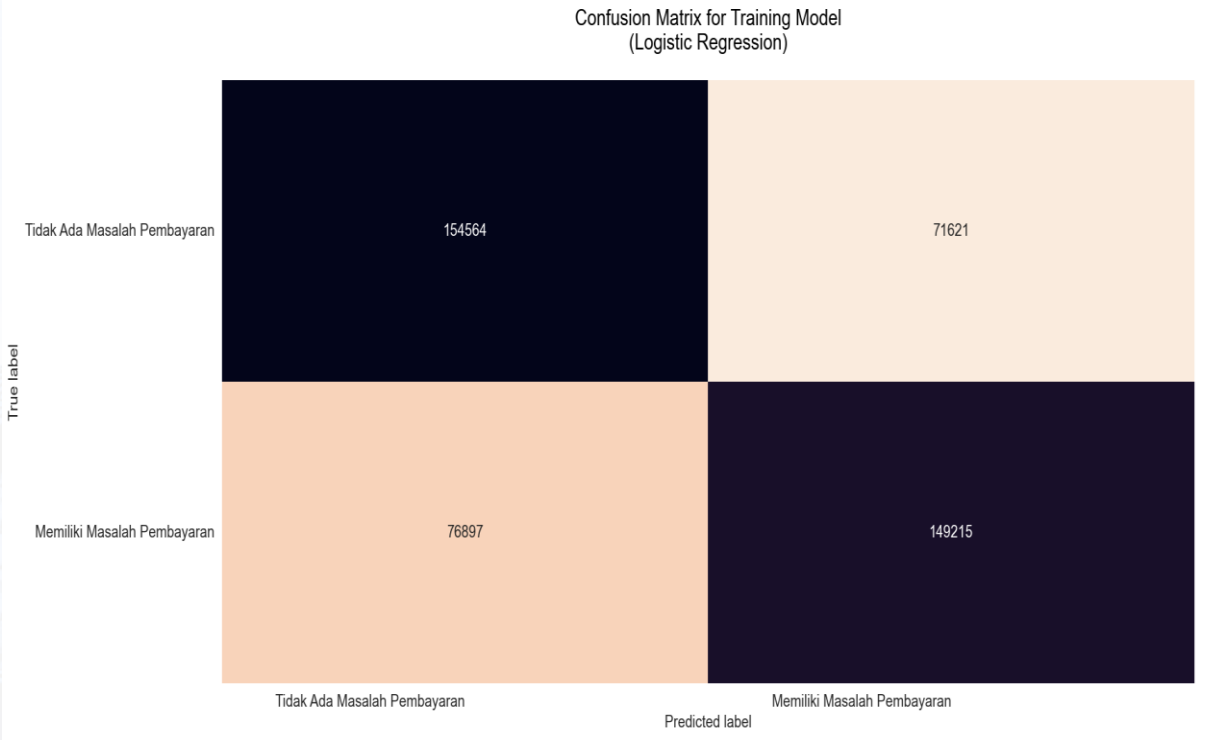
```
# predict data train
y_train_pred_log = log_m.predict(X_train)

# print classification report
print('Classification Report Training Model (Logistic Regression):')
print(classification_report(y_train, y_train_pred_log))
```

```
Classification Report Training Model (Logistic Regression):
              precision    recall  f1-score   support

           0       0.67      0.68      0.68    226185
           1       0.68      0.66      0.67    226112

    accuracy                           0.67    452297
   macro avg       0.67      0.67      0.67    452297
weighted avg       0.67      0.67      0.67    452297
```



Data Modeling

Rakamin Academy

For Testing data

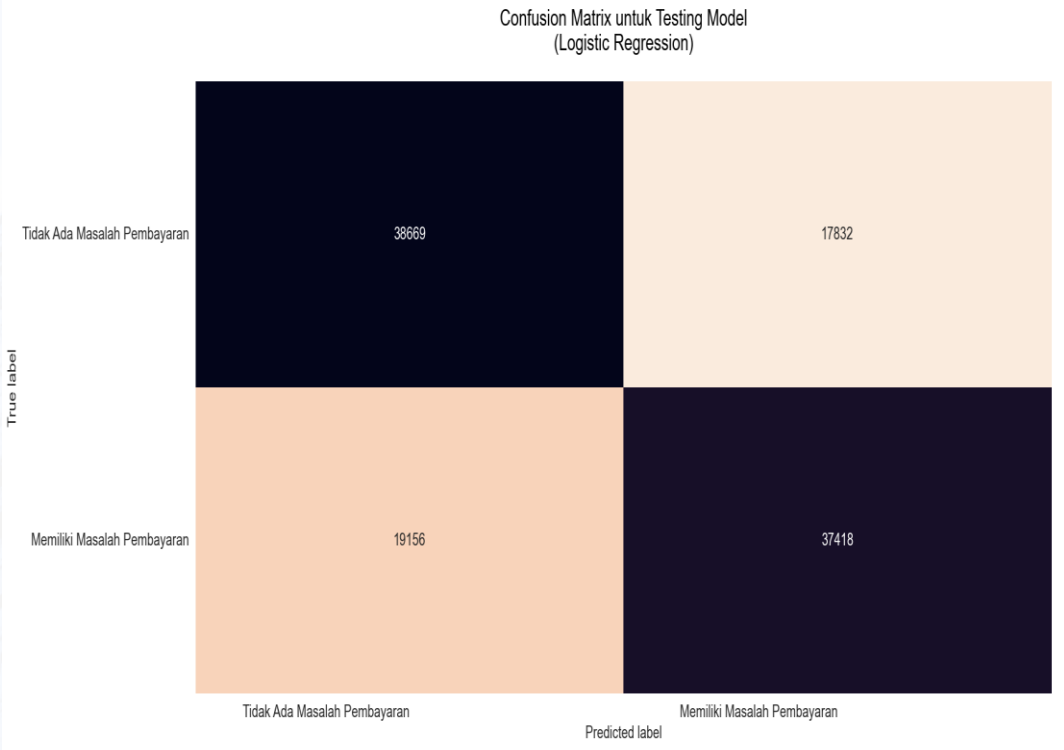
```
# predict data test
y_test_pred_log = log_m.predict(X_test)

# print classification report
print('Classification Report Testing Model (Logistic Regression):')
print(classification_report(y_test, y_test_pred_log))
```

```
Classification Report Testing Model (Logistic Regression):
              precision    recall  f1-score   support

           0       0.67      0.68      0.68     56501
           1       0.68      0.66      0.67     56574

    accuracy                           0.67    113075
   macro avg       0.67      0.67      0.67    113075
weighted avg       0.67      0.67      0.67    113075
```



Data Modeling

Rakamin Academy

For Testing data

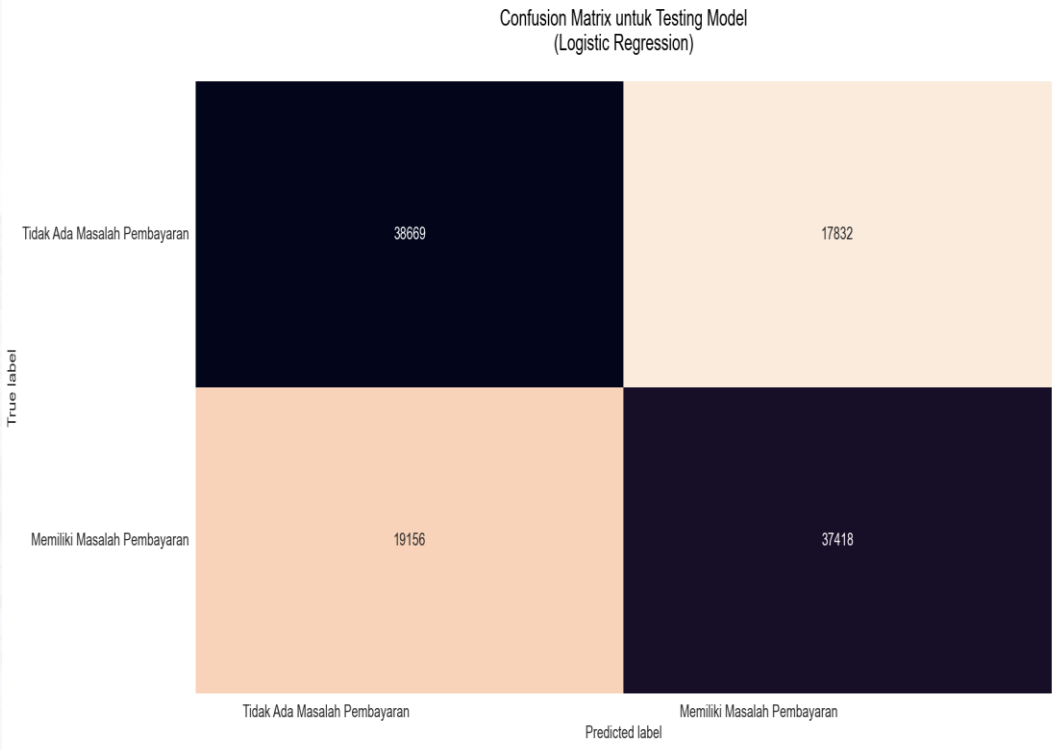
```
# predict data test
y_test_pred_log = log_m.predict(X_test)

# print classification report
print('Classification Report Testing Model (Logistic Regression):')
print(classification_report(y_test, y_test_pred_log))
```

```
Classification Report Testing Model (Logistic Regression):
              precision    recall  f1-score   support

           0       0.67      0.68      0.68     56501
           1       0.68      0.66      0.67     56574

    accuracy                           0.67    113075
   macro avg       0.67      0.67      0.67    113075
weighted avg       0.67      0.67      0.67    113075
```



Data Modeling

Rakamin Academy

Accuracy Score

```
from sklearn.metrics import accuracy_score

# Calculate accuracy for training set
acc_log_train = round(accuracy_score(y_train, y_train_pred_log) * 100, 2)

# Calculate accuracy for test set
acc_log_test = round(accuracy_score(y_test, y_test_pred_log) * 100, 2)

print("Training Accuracy: {}%".format(acc_log_train))
print("Test Accuracy: {}%".format(acc_log_test))
```

```
Training Accuracy: 67.16%
Test Accuracy: 67.29%
```

Model logistic regression memberikan hasil yang benar sebesar 67,16%. Terdapat 0,13% margin kesalahan.

```
# ROC scores
from sklearn.metrics import roc_auc_score

roc_auc_log = round(roc_auc_score(y_test, y_test_pred_log),4)
print('ROC AUC:', roc_auc_log)
```

```
ROC AUC: 0.6729
```

Data Modeling

Rakamin Academy

For Train data

DecisionTreeClassifier

```
from sklearn.tree import DecisionTreeClassifier
# train the model
dt_model = DecisionTreeClassifier().fit(X_train,y_train)
print(dt_model)
```

```
DecisionTreeClassifier()
```

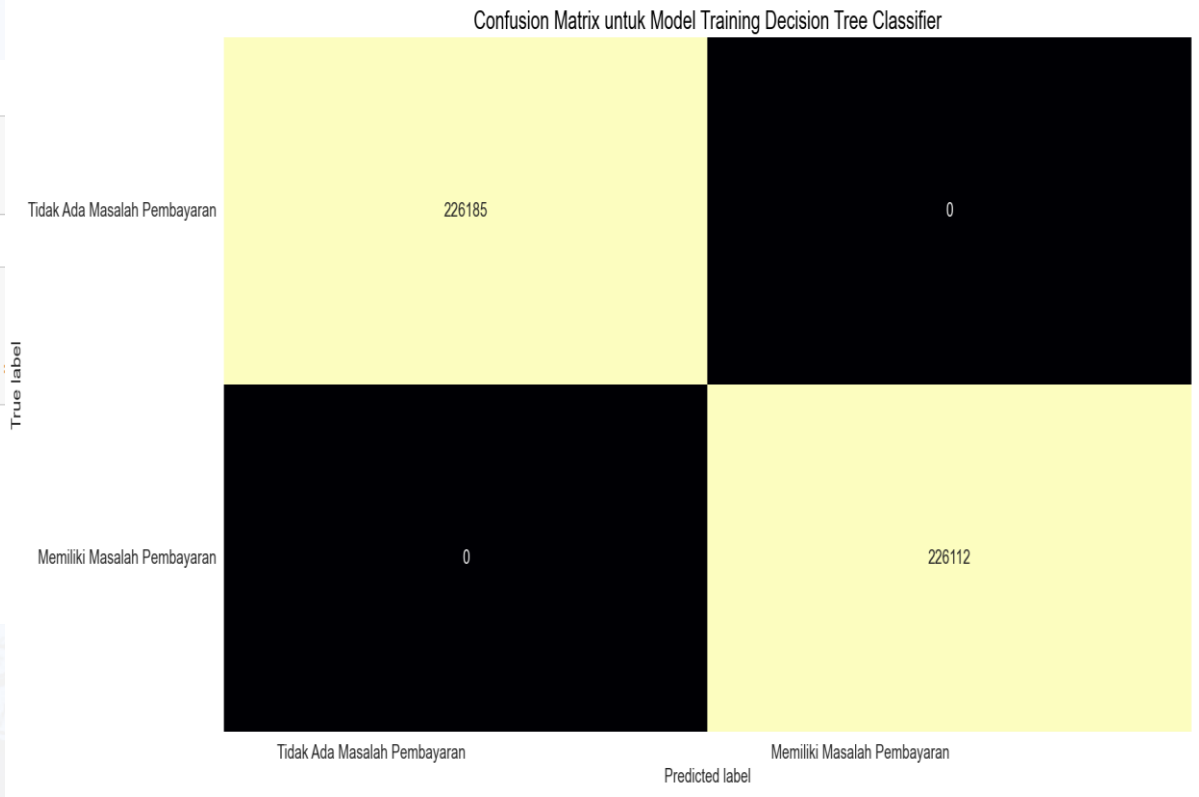
```
# predict data train
y_train_pred_dt = dt_model.predict(X_train)

# print classification report
print('Classification Report Training Model (Decision Tree Classifier):')
print(classification_report(y_train, y_train_pred_dt))
```

```
Classification Report Training Model (Decision Tree Classifier):
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    226185
           1       1.00      1.00      1.00    226112

    accuracy                           1.00    452297
   macro avg       1.00      1.00      1.00    452297
weighted avg       1.00      1.00      1.00    452297
```



Data Modeling

Rakamin Academy

For Testing data

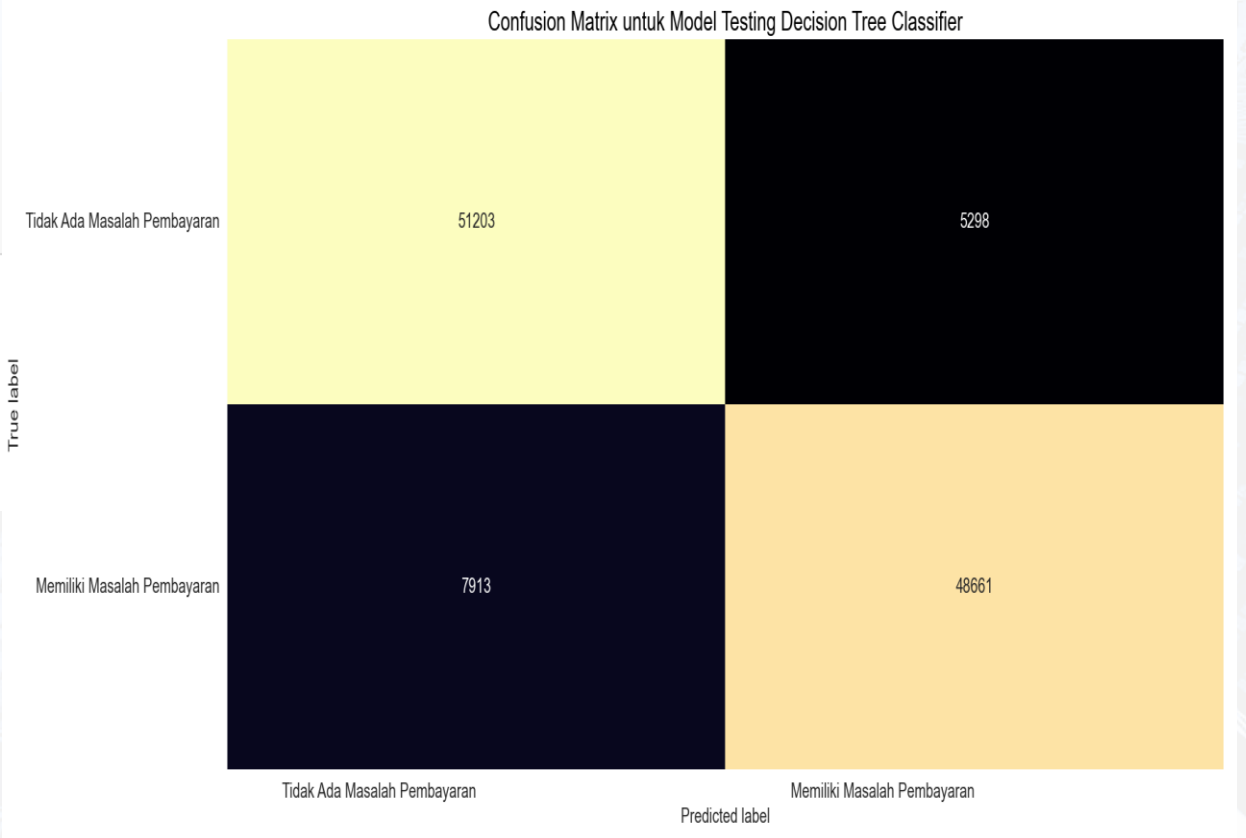
```
# predict data test
y_test_pred_dt = dt_model.predict(X_test)

# print classification report
print('Classification Report Testing Model (Decision Tree Classifier):')
print(classification_report(y_test, y_test_pred_dt))
```

```
Classification Report Testing Model (Decision Tree Classifier):
              precision    recall  f1-score   support

           0       0.87      0.91      0.89     56501
           1       0.90      0.86      0.88     56574

    accuracy                           0.88    113075
   macro avg       0.88      0.88      0.88    113075
weighted avg       0.88      0.88      0.88    113075
```



Data Modeling

Rakamin Academy

Accuray Score

```
acc_dt_train=round(dt_model.score(X_train,y_train)*100,2)
acc_dt_test=round(dt_model.score(X_test,y_test)*100,2)
print("Training Accuracy: % {}".format(acc_dt_train))
print("Test Accuracy: % {}".format(acc_dt_test))
```

```
Training Accuracy: % 100.0
Test Accuracy: % 88.32
```

dari hasil diatas keputusan memberikan hasil yang 100% benar. Terdapat 11,74% margin kesalahan. Ini tidak baik untuk data ini.

```
# ROC scores
roc_auc_dt = round(roc_auc_score(y_test, y_test_pred_dt),4)
print('ROC AUC:', roc_auc_dt)
```

```
ROC AUC: 0.8832
```

Data Modeling

Rakamin Academy

Accuray Score

```
acc_dt_train=round(dt_model.score(X_train,y_train)*100,2)
acc_dt_test=round(dt_model.score(X_test,y_test)*100,2)
print("Training Accuracy: % {}".format(acc_dt_train))
print("Test Accuracy: % {}".format(acc_dt_test))
```

```
Training Accuracy: % 100.0
Test Accuracy: % 88.32
```

dari hasil diatas keputusan memberikan hasil yang 100% benar. Terdapat 11,74% margin kesalahan. Ini tidak baik untuk data ini.

```
# ROC scores
roc_auc_dt = round(roc_auc_score(y_test, y_test_pred_dt),4)
print('ROC AUC:', roc_auc_dt)
```

```
ROC AUC: 0.8832
```

Business Recomendation

Rakamin Academy

1. Membuat kampanye untuk menarik lebih banyak pelajar, akuntan, teknisi berketerampilan tinggi, manajer yang tertarik untuk mengajukan pinjaman
2. Melakukann survei untuk melihat apakah terdapat permasalahan pada kontrak pinjaman tunai bagi nasabah yang sedang cuti hamil atau menganggur. Agar di masa depan, jika Anda memiliki klien dengan pendapatan seperti ini, Anda dapat merekomendasikan jenis kontrak yang sesuai agar permohonan mereka disetujui.
3. Pelanggan yang tidak mengalami kesulitan pembayaran adalah mereka yang berusia 35-45 tahun. Anda dapat menjadikan pelanggan ini sebagai prioritas utama Anda.

Rakamin Academy

Link Github and LinkedIn

https://github.com/LimatanL
https://www.linkedin.com/in/limatanluviar/

Thank You