SPRINT CODE

- Sprint Code 1
- Sprint Code 2

DASHBOARD

Model Selection and Evaluation for Credit Risk Prediction: A Home Credit Case Study

Nomor Kelompok: 29
Nama Mentor: Erwin Fernanda
Nama Tim: Analysis Alliance

Data Analytics Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Athiyah Fitriyani Basuki
UPN Jawa Timur

Desi Noviyanti
IPB University

Ridat Maulana
Universitas Suryakancana

Rosyida Ishma Mardhiyyah
Institut Teknologi Sepuluh Nopember

Muhammad Zainil Mubarok
Universitas Diponegoro

Daftar Isi

BUSINESS UNDERSTANDING

Business Understanding

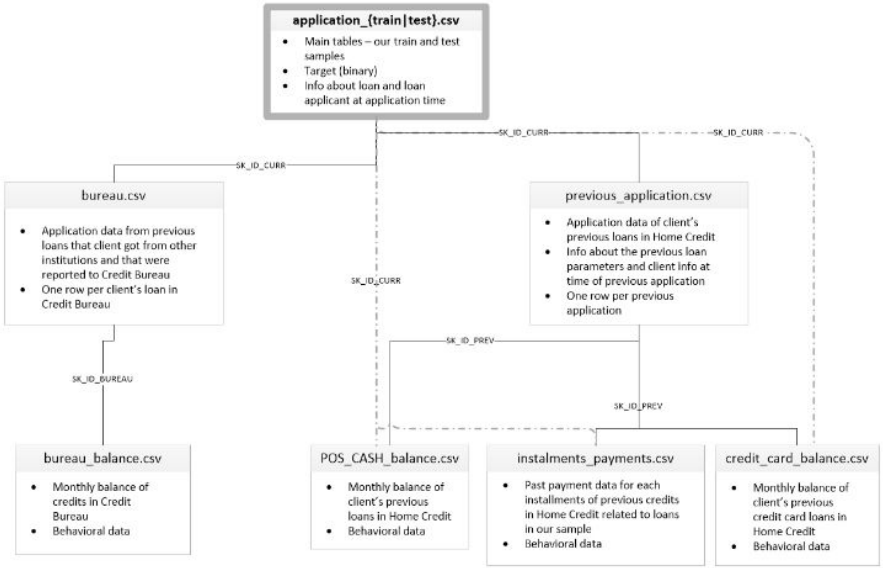
Banyak orang mengalami kesulitan untuk mendapatkan pinjaman karena riwayat kredit yang tidak memadai atau tidak ada sama sekali. Sayangnya, populasi ini sering menjadi korban dari pemberi pinjaman yang tidak dapat dipercaya.

Home Credit merupakan perusahaan keuangan yang **menyediakan layanan kredit kepada konsumen yang memiliki keterbatasan akses** ke layanan keuangan tradisional. Perusahaannya, menawarkan layanan kredit kepada individu yang tidak memiliki riwayat kredit atau yang berisiko tinggi.

Tujuan dari analisis ini yaitu, **memahami faktor-faktor yang memengaruhi risiko gagal bayar kredit**. Berdasarkan data yang diperoleh, **sulitnya nasabah dalam mengembalikan pinjaman** mencapai angka **8%** yang artinya, dalam dunia keuangan **8% menunjukkan angka yang sangat tinggi dan akan berisiko** untuk perusahaan tersebut. Dengan demikian, menganalisis data kredit yang ada dapat membantu perusahaan dalam **mengidentifikasi pola atau tren yang dapat membantu perusahaan dalam mengambil keputusan dalam menilai risiko kredit pelanggan**.

DATA UNDERSTANDING

Describe Data



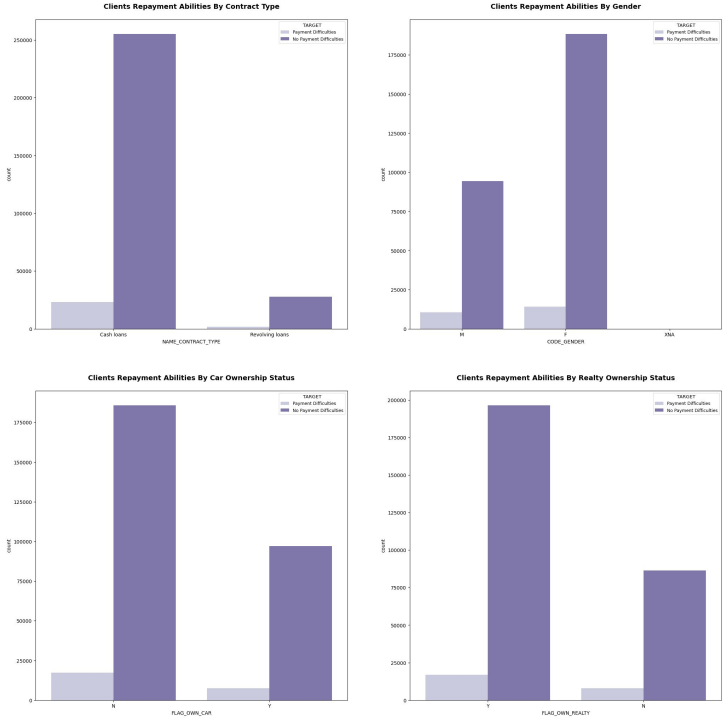
Data yang digunakan dalam Final Project ini yaitu, application_train dengan informasi tentang setiap aplikasi pinjaman di Home Credit. Setiap pinjaman ditandai dengan fitur SK_ID_CURR. Data aplikasi pelatihan dilengkapi dengan TARGET.

Exploratory Data Analysis

Berdasarkan Pie Chart di samping, terdapat **92% customer yang tidak memiliki masalah dalam melunasi pinjaman** pada waktu tertentu, sedangkan **8% customer bermasalah dalam melunasi pinjaman.**

Contract Type, Gender, Car Ownership Status, and Realty Ownership Status

- **Contract Type**
  Mayoritas peminjam mengajukan pinjaman tipe cash loans yaitu sekitar **278 ribu peminjam**.
- **Gender**
  Pengajuan pinjaman sebagian besar dilakukan oleh perempuan yaitu **sekitar 202.448 yang diajukan**.
- **Car Ownership Status**
  Perbedaan tidak terlalu signifikan, dengan **8% mengalami kesulitan dalam pengembalian sedangkan 7% tidak mengalami kesulitan dalam pengembalian**.
- **Realty Ownership Status**
  Perbedaan tidak terlalu signifikan, dengan **8% mengalami kesulitan dalam pengembalian sedangkan 7% tidak mengalami kesulitan dalam pengembalian.**



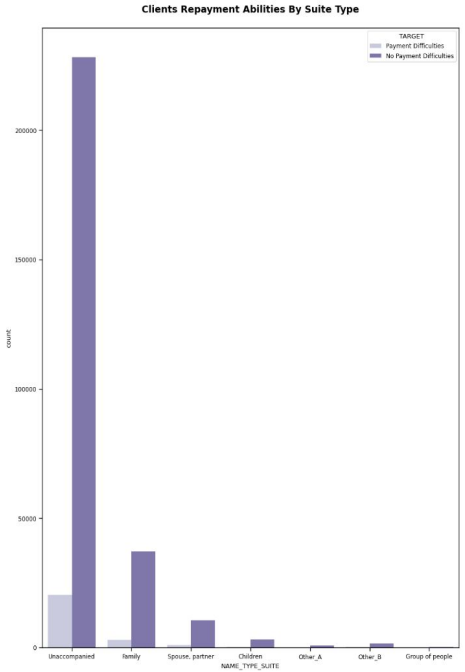
Suite Type and Income Type

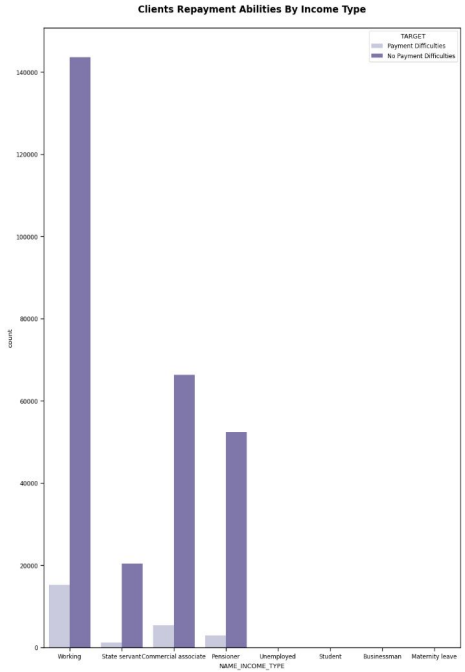
- **Suite Type**

  Nasabah yang didampingi dengan **Other_B** saat mengajukan pinjaman memiliki **kesulitan dalam mengembalikan pinjaman yaitu sekitar 10%.**

- **Income Type**

  Nasabah dengan Income Type **pengusaha dan siswa tidak mengalami kesulitan dalam mengembalikan pinjaman.**





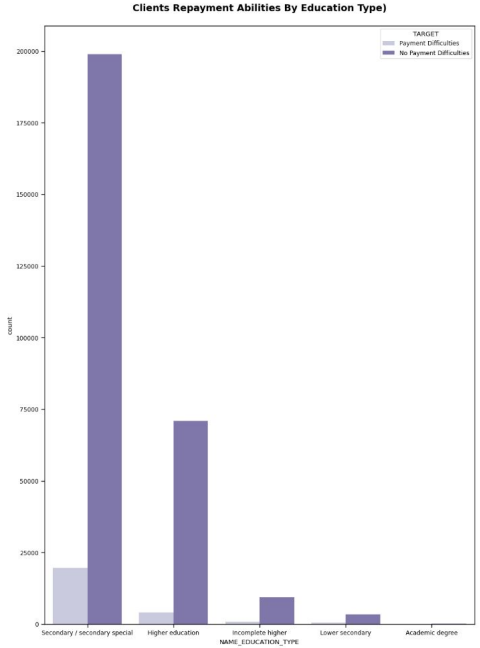
Education Type and Family Status

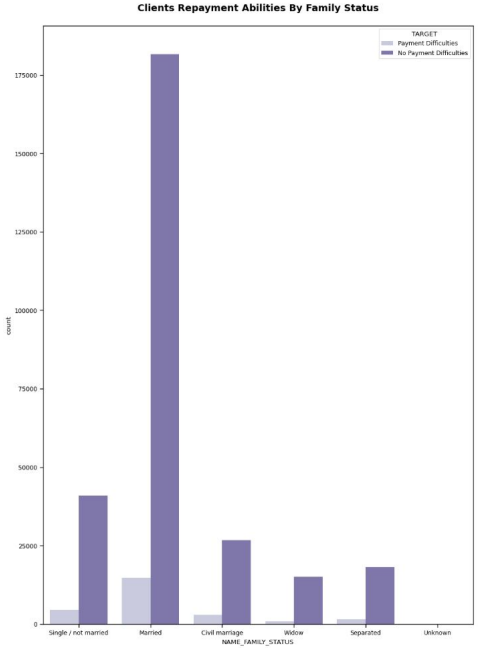
- **Education Type**

  Nasabah dengan jenis **pendidikan SMP memiliki kendala dalam mengembalikan pinjaman kredit** yaitu **sekitar 10%**.

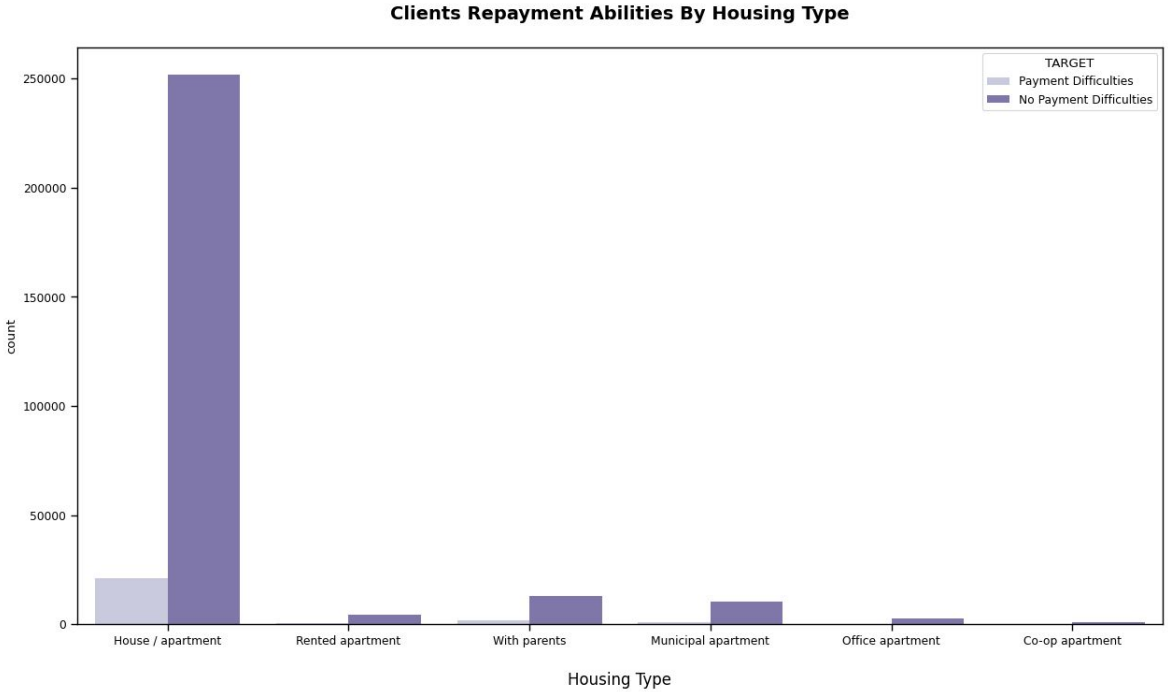
- **Family Status**

  Family status **"Married" tidak memiliki kesulitan** dalam mengembalikan pinjaman kredit.



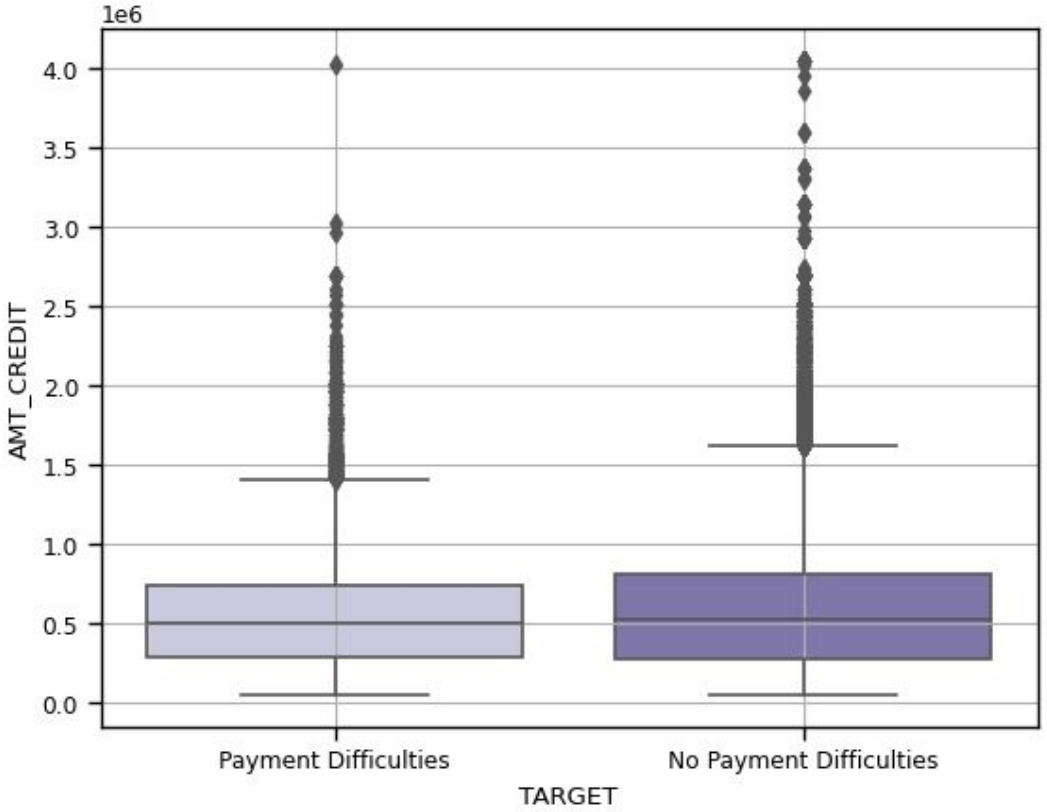


Housing Type



Jumlah pengajuan kredit terbanyak dilakukan oleh nasabah yang **memiliki rumah atau apartemen yaitu sebesar 272 ribu.**

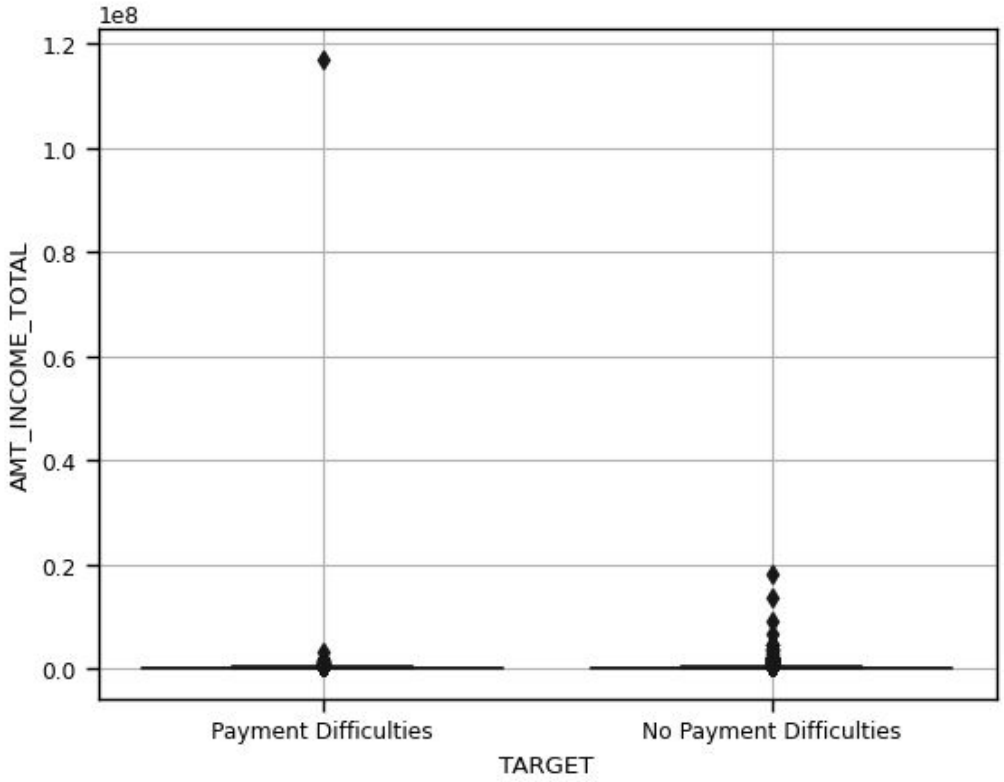
Amount Credit of The Loan



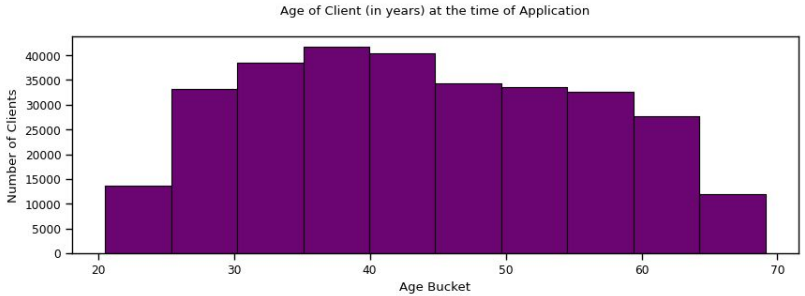
Berdasarkan nilai median, **jumlah kredit nasabah yang tidak mengalami kesulitan pembayaran sedikit lebih besar** dibandingkan dengan nasabah yang mengalami kesulitan pembayaran.

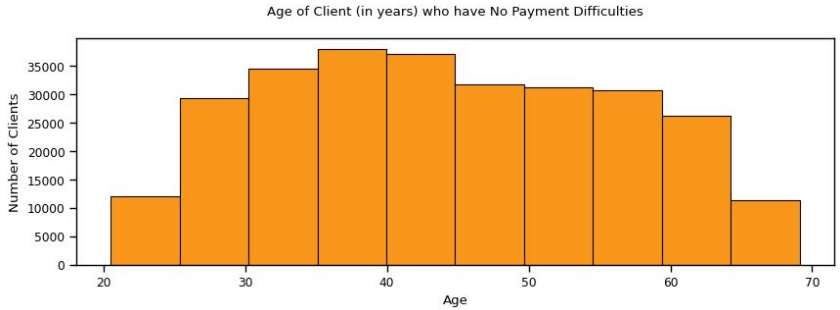
Amount Income

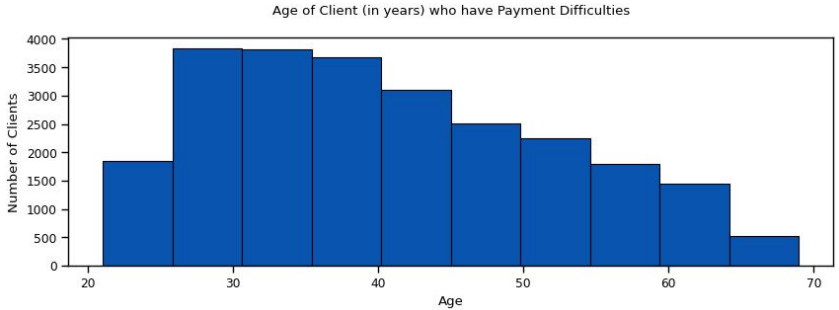
Berdasarkan chart Amount Income VS Target, **tidak ada perbedaan yang signifikan**. Hanya aja, terdapat **outlier dimana seseorang mempunyai Amount Income yang tinggi tetapi sulit dalam mengembalikan pinjamannya**. Akan tetapi dengan meningkatnya pendapatan nasabah, kemungkinan tidak akan mengalami kesulitan membayar kembali pinjaman tersebut.



Age

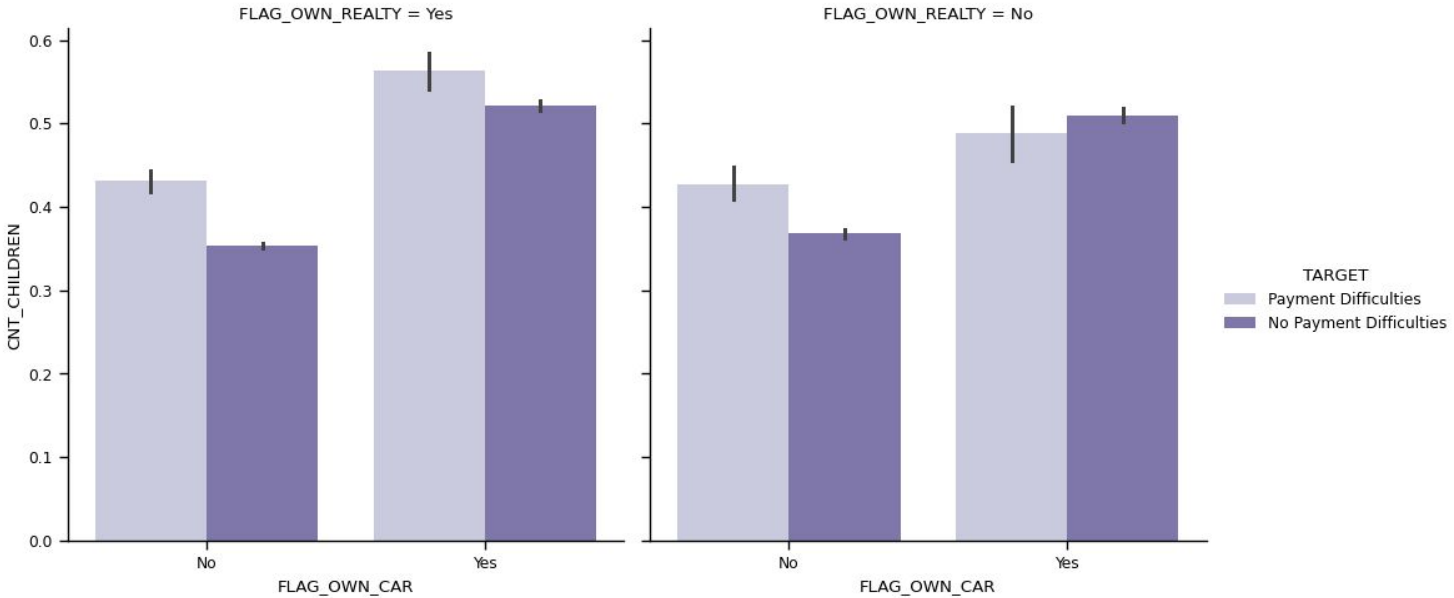






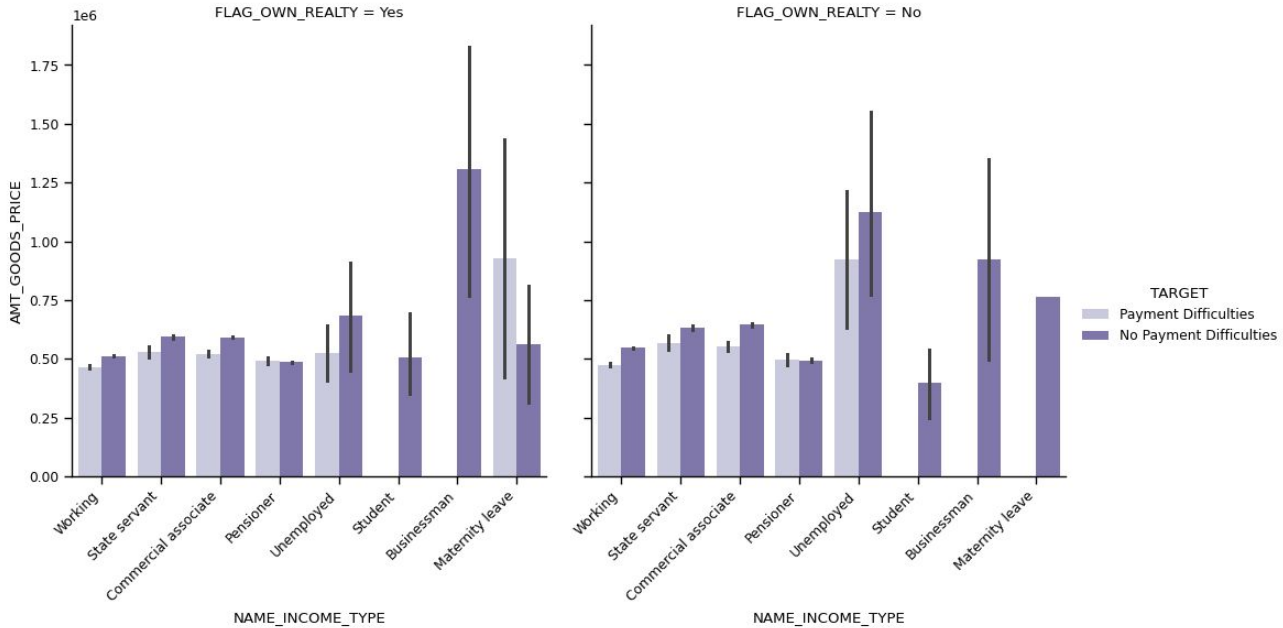
Nasabah yang **tidak mengalami kesulitan pembayaran** berada di rentang **usia 35-45 tahun**, sedangkan nasabah yang mengalami kesulitan pembayaran berada di rentang usia 25-35 tahun.

Car Ownership Status, The Number of Children, Target, and House/Flat Ownership Status



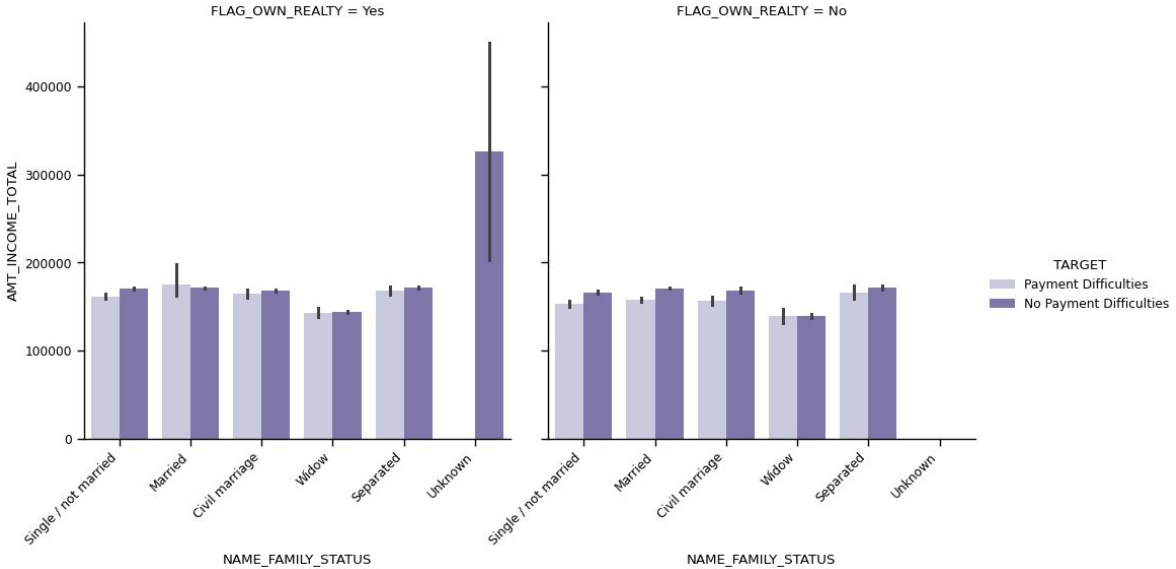
Nasabah yang **memiliki mobil dan rumah/flat memiliki masalah dalam mengembalikan pinjaman untuk jumlah anak yang tinggi** dibandingkan nasabah yang tidak memiliki rumah/flat.

Income Type, Amount of Goods Price, Target, and House/Flat Ownership Status



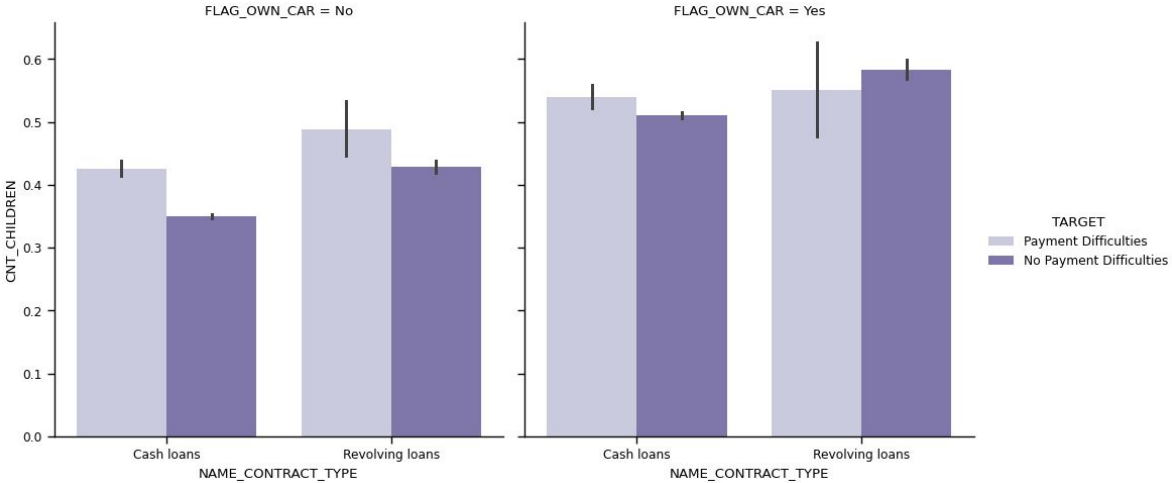
Nasabah dengan **jenis penghasilan cuti hamil dan memiliki rumah/flat mengalami masalah dalam membayar pinjaman** dibandingkan ketika nasabah tidak memiliki rumah/flat.

Family Status, Amount of Income, Target, and House/Flat Ownership Status



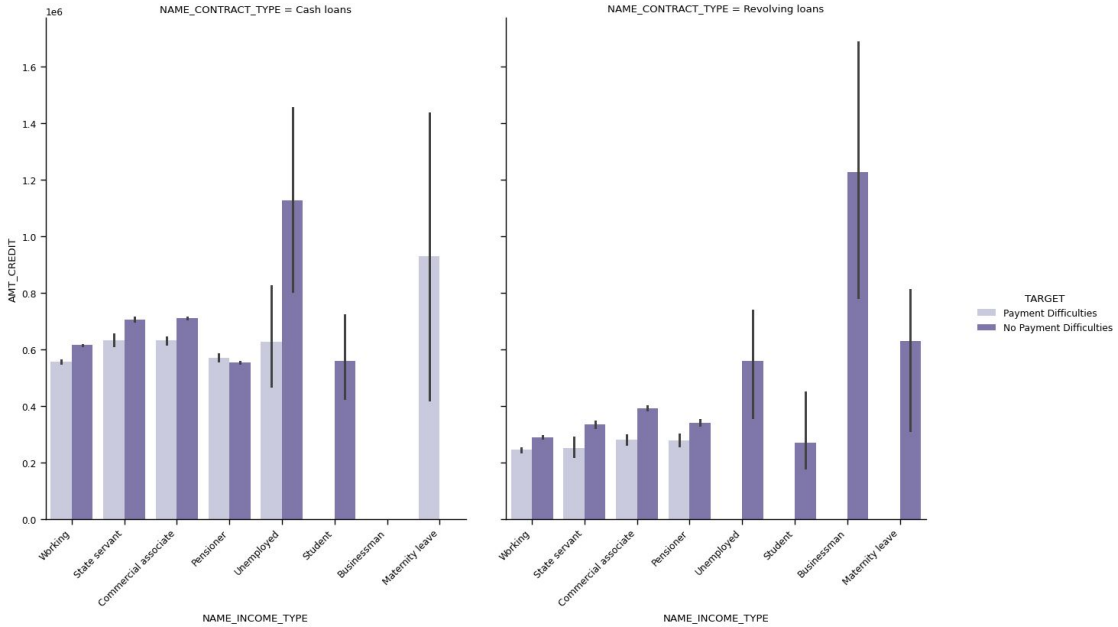
Nasabah yang **sudah menikah dan memiliki rumah/flat dengan pendapatan yang menengah mengalami masalah dalam mengembalikan pinjam** dibandingkan dengan nasabah yang tidak memiliki rumah/flat.

Contract Type, The Number of Children, Target, and Car Ownership Status



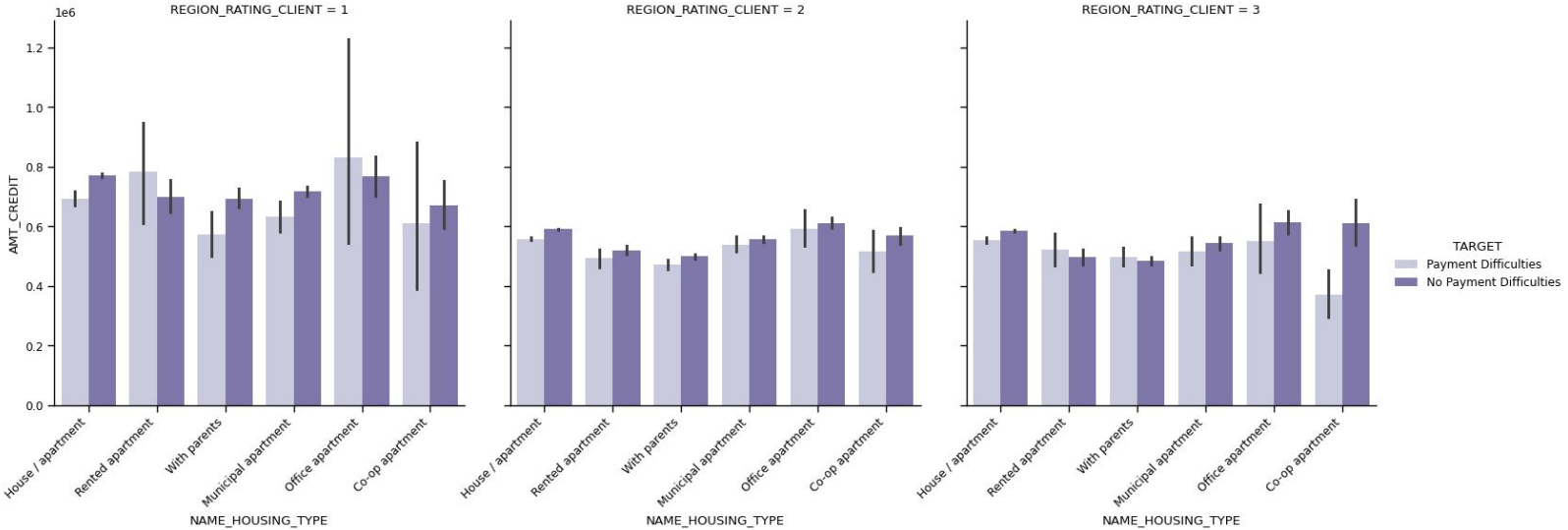
Nasabah yang mengajukan pinjaman dengan **revolving loans dan tidak memiliki mobil memiliki dengan jumlah anak yang tinggi memiliki masalah dalam membayar kembali pinjamannya**, sedangkan nasabah yang **memiliki mobil dengan jumlah anak yang tinggi tidak mengalami kesulitan dalam membayar kembali pinjamnnya.**

Income Type, Amount of Credit, Target, and Contract Type



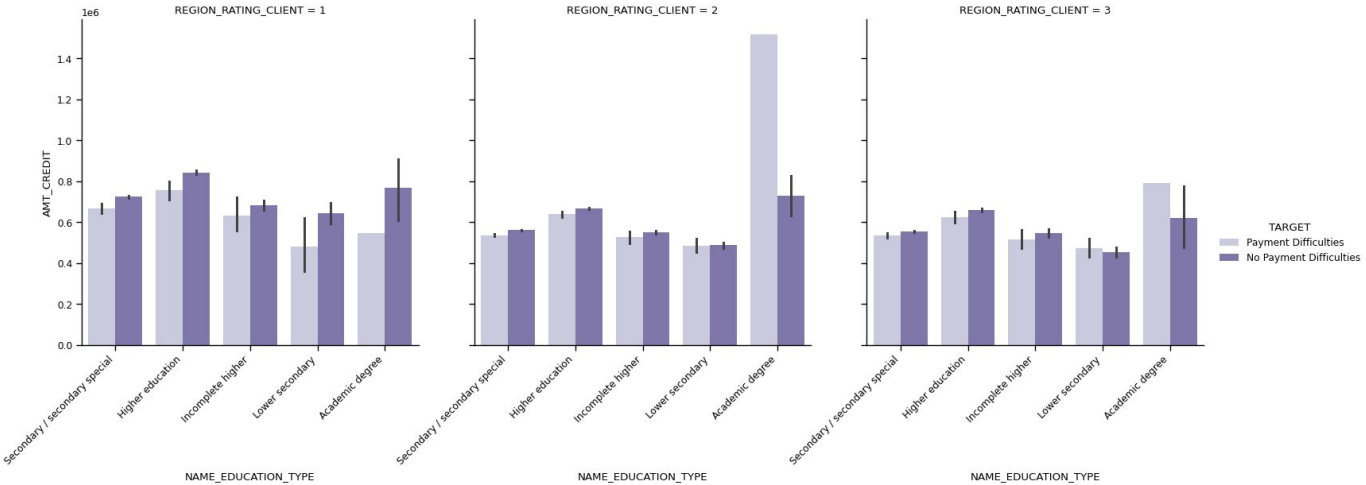
Semua nasabah dengan status **siswa tidak mengalami kesulitan dalam mengembalikan pinjaman kredit** tersebut baik **secara cash loans atau pinjaman revolving** dengan jumlah pinjaman kredit rendah hingga menengah.

Housing Type, Amount Credit of Loan, Target, and Rating of Region where Client Lives



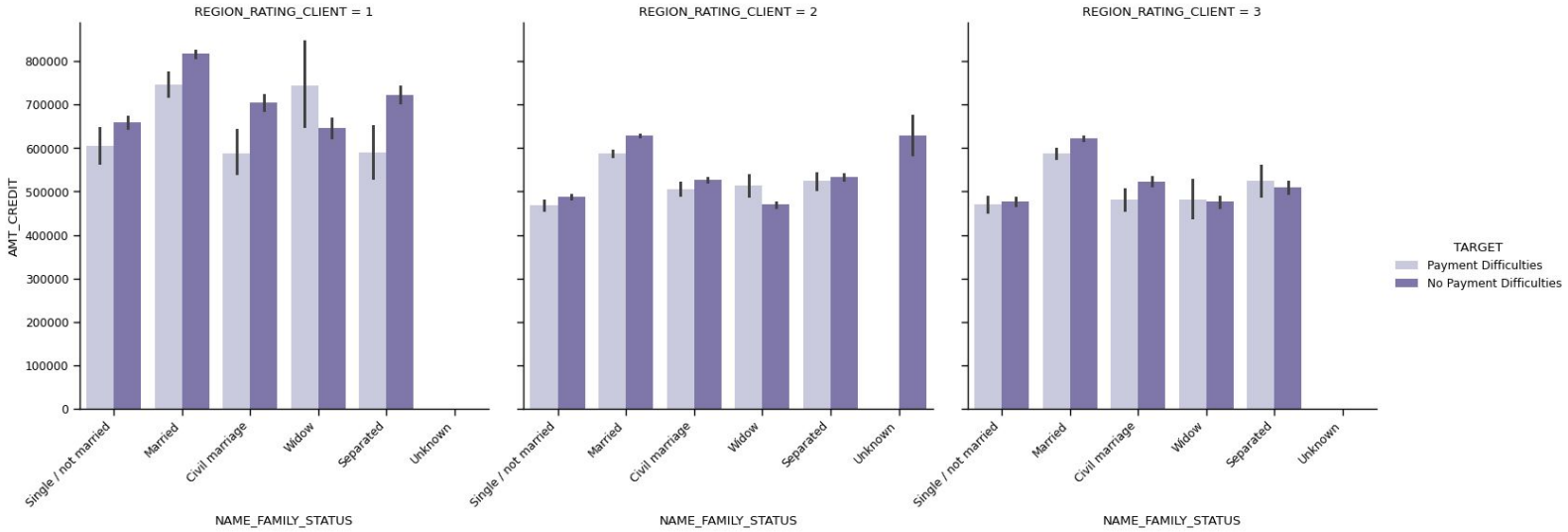
Nasabah yang tinggal di **apartemen sewaan**, **apartemen kantor**, **dan wilayahnya yang memiliki peringkat 1 memiliki masalah dalam membayar kembali pinjaman** dibandingkan dengan nasabah di wilayah yang memiliki peringkat 2 untuk jumlah pinjaman kredit sedang.

Education Type, Amount Credit of Loan, Target, and Rating of Region where Client Lives



Nasabah yang **memiliki gelar akademik dan tinggal di wilayah dengan peringkat 2, memiliki masalah dalam membayar kembali pinjaman untuk jumlah kredit pinjaman yang lebih tinggi**. Sementara nasabah dengan gelar yang sama teteapi tinggal di wilayah dengan peringkat 3, memiliki masalah dalam membayar pinjaman untuk kredit pinjaman dalam jumlah sedang.

Family Status, Amount Credit of Loan, Target, and Rating of Region where Client Lives



Nasabah yang **berstatus married, baik yang tinggal di daerah golongan 1,2, atau 3 mengalami kesulitan dalam mengembalikan pinjaman** untuk kredit pinjaman **dalam jumlah sedang hingga tinggi**.

DATA PREPARATION

Missing Value

Dataset Home Credit terdapat **67 kolom** yang memiliki null value atau data kosong. 49 dari 67 kolom tersebut memiliki **lebih dari 40% null value.**

Gambar di samping menunjukan terdapat **11 data dengan null value terbanyak**. Kolom dengan null value terbanyak memiliki persentase **hampir 70% null value.**

	Variable	Missing Value	Percentage
0	COMMONAREA_MEDI	214865	69.872297
1	COMMONAREA_AVG	214865	69.872297
2	COMMONAREA_MODE	214865	69.872297
3	NONLIVINGAPARTMENTS_MEDI	213514	69.432963
4	NONLIVINGAPARTMENTS_MODE	213514	69.432963
5	NONLIVINGAPARTMENTS_AVG	213514	69.432963
6	FONDKAPREMONT_MODE	210295	68.386172
7	LIVINGAPARTMENTS_MODE	210199	68.354953
8	LIVINGAPARTMENTS_MEDI	210199	68.354953
9	LIVINGAPARTMENTS_AVG	210199	68.354953
10	FLOORSMIN_MODE	208642	67.848630

Missing Value Handling

- Data Home Credit awalnya memiliki **307511 rows x 122 columns**.
- Setelah dilakukan handling terhadap missing value dengan cara menghapus columns yang memiliki null value **diatas 40%**, data yang tersisa yaitu **307511 rows x 73 columns.**
- Missing value yang tersisa diisi dengan **median untuk data numeric dan modus untuk data categorical.**
- Pada data home credit juga terdapat data yang memiliki isi **XNA**, lakukan handling untuk data tersebut dengan **mengganti data dengan nilai modus** pada kategori tersebut.
- Handling terakhir, yaitu **menghapus kolom FLAG_DOCUMENT** karena tidak relevan.

Outlier Handling dan Anomali Handling

- Data Home Credit yang awalnya memiliki **307511 rows menjadi 303239 rows** setelah dilakukan **outlier handling.** Outlier handling yang dilakukan adalah dengan **menghitung z-score pada data CNT_CHILDREN** kemudian **menghapus data yang memiliki z-score dibawah 3.**
- Data Home Credit memiliki beberapa value yang **anomali**, dimana terdapat beberapa data numerical dengan value yang tidak sesuai contohnya terdapat **data yang bernilai minus dalam beberapa kolom**. Hal ini harus diperbaiki dengan **mengalikan data yang bernilai minus dengan -1**.

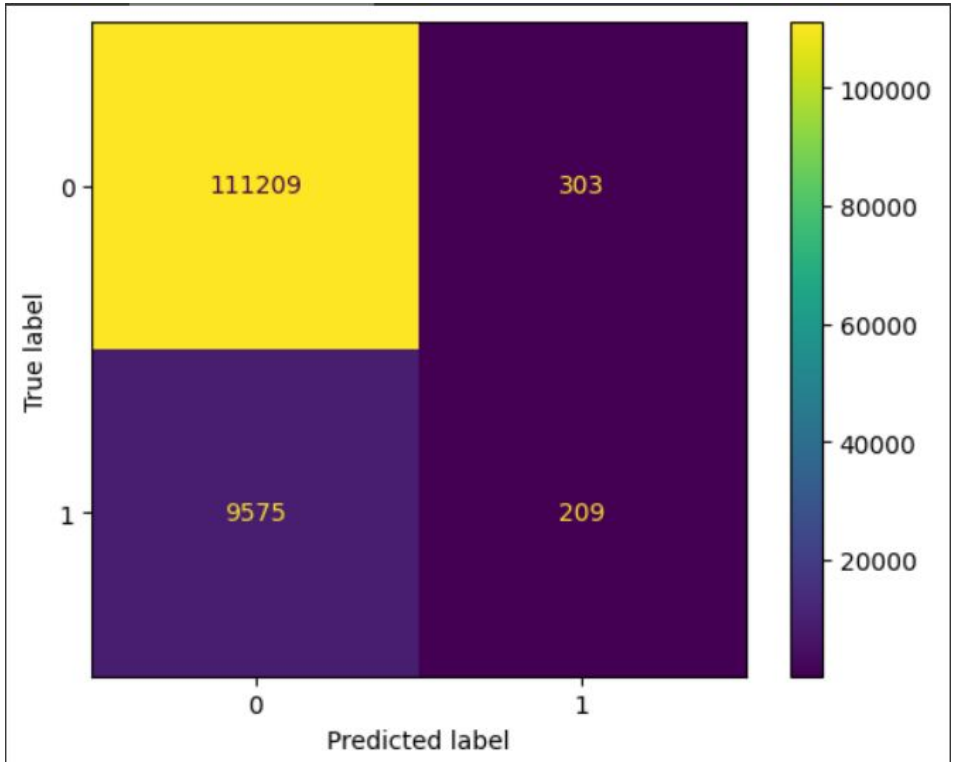
Categorical Data Encoding

Variabel kategori seringkali tidak dapat diolah secara langsung oleh model atau algoritma machine learning, sehingga data yang digunakan perlu **diubah terlebih dahulu ke dalam bentuk numerik** sebelum dapat digunakan. Pada data kali ini dilakukan **2 metode encoding yaitu one-hot-encoding dan label-encoder**. **One hot encoding** dilakukan untuk data dengan **value unique lebih dari 2** sedangkan **sisanya dilakukan label encoder**.

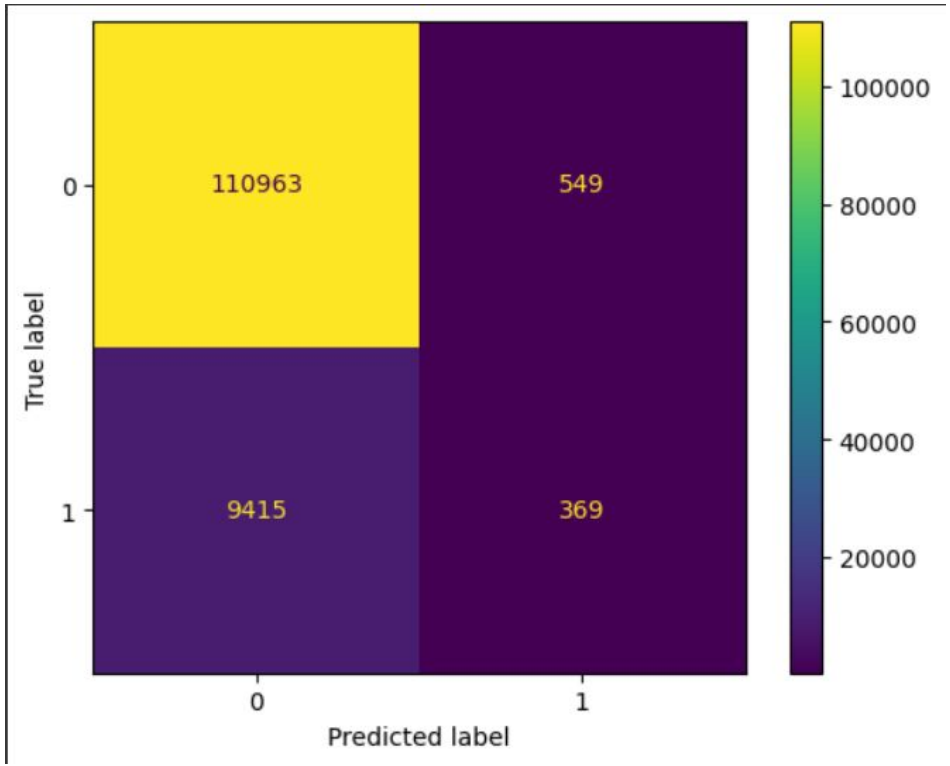
MODELLING

Logistic Regression Model

- ROC AUC : **0.7190895322244799**
- Cross-Validated ROC AUC Scores: [0.73980628, 0.74043704, 0.73800476, 0.74230201, 0.74428416]
- Average Cross-Validated ROC AUC Scores: 0.74



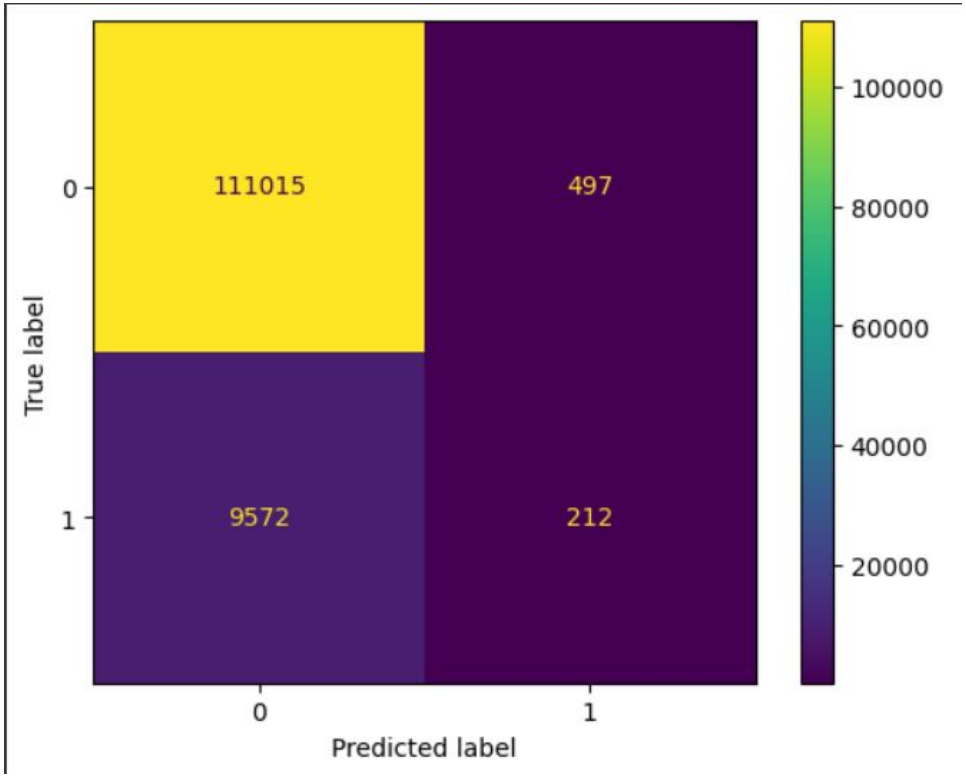
XGBoost Model



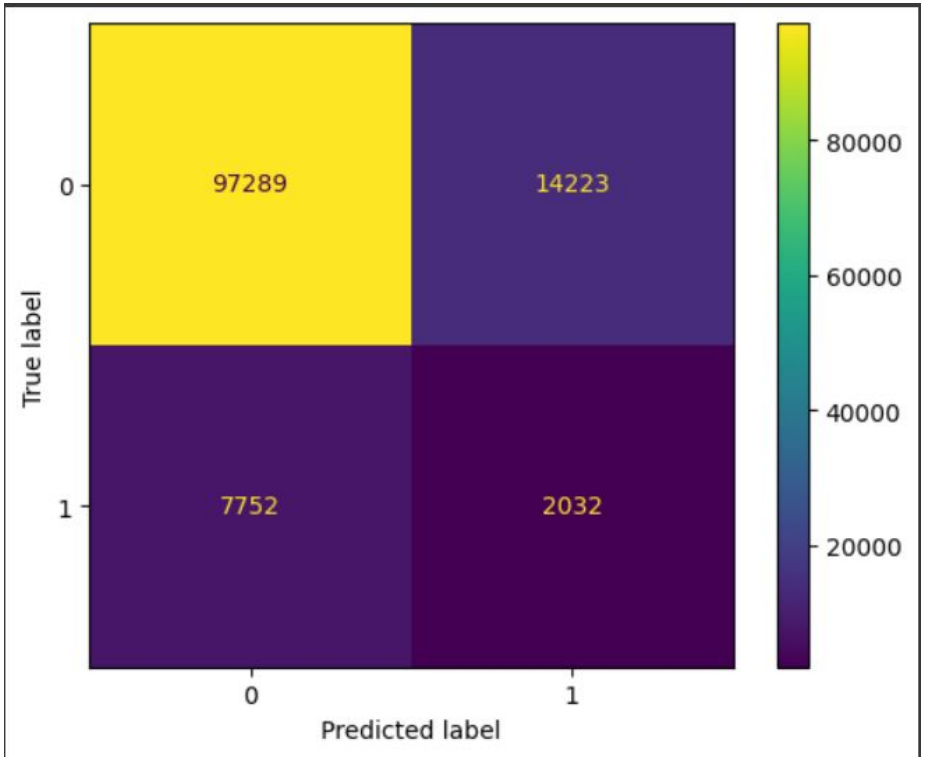
- ROC AUC: **0.7195711723797188**
- Cross-Validated ROC AUC Scores: [0.74385408, 0.74637859, 0.73947971, 0.74604294, 0.74806696]
- Average Cross-Validated ROC AUC Scores: 0.74

Random Forest

- ROC AUC: **0.6983541236346816**
- Cross-Validated ROC AUC Scores: [0.71500913, 0.71676969, 0.71318865, 0.71150339, 0.70935804]
- Average Cross-Validated ROC AUC Scores: 0.71



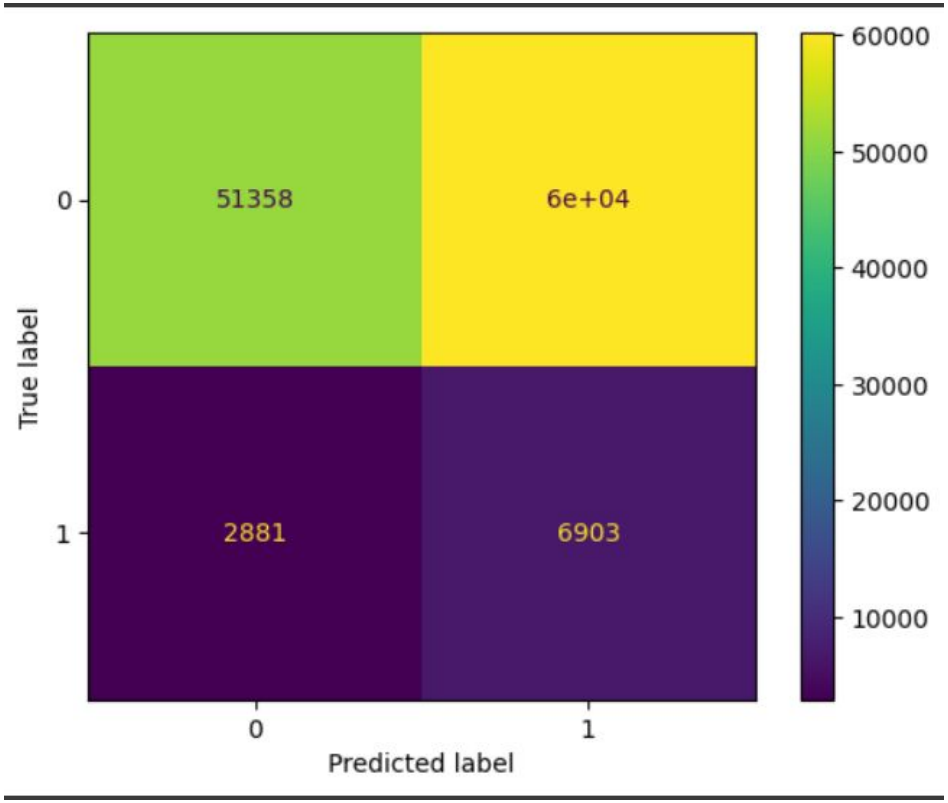
Decision Tree



- ROC AUC: **0.540069603441510 7**
- Cross-Validated ROC AUC Scores: [0.53504726, 0.53221382, 0.53728516, 0.53588006, 0.53650584]
- Average Cross-Validated ROC AUC Scores: 0.54

K-Nearest Neighbors

- ROC AUC: **0.6100349302961033**
- Cross-Validated ROC AUC Scores: [0.58955355, 0.58832464, 0.59014812, 0.58859345, 0.5905525]
- Average Cross-Validated ROC AUC Scores: 0.59



EVALUATION

Model Selection

Model klasifikasi terbaik yang didapatkan untuk memprediksi nasabah yang gagal membayar pinjaman pada data Home Credit adalah **Model XGBoost**.

	Models	ROC-AUC Score Before Cross-Val	ROC-AUC Score After Cross-Val
1	XGBoost	0.719571	0.744764
0	Logistic Regression	0.719090	0.740967
2	Random Forest	0.698354	0.713166
4	K-Nearest Neighbor	0.610035	0.589434
3	Decision Tree	0.540070	0.535386

DEPLOYMENT

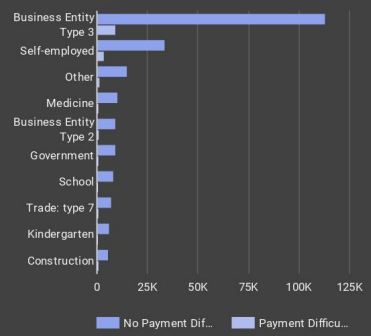
Home Credit Dashboard

Loan Application

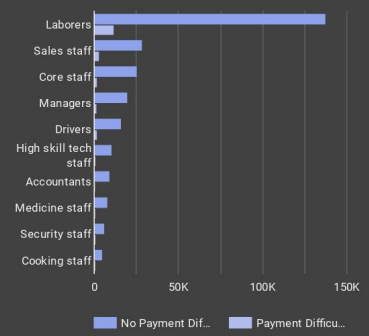
Contact Type

Total Loan Applications	Total Non-Performing Loan	Average Credit Amount	Average Income
303,239	24,396	598,910.04	168,733.46

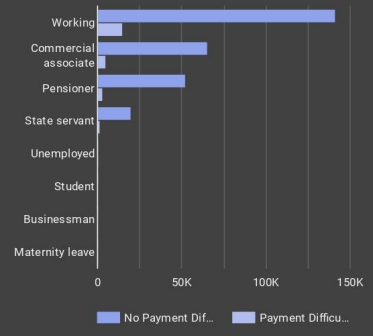
Non-Performing Loan by Organization Type



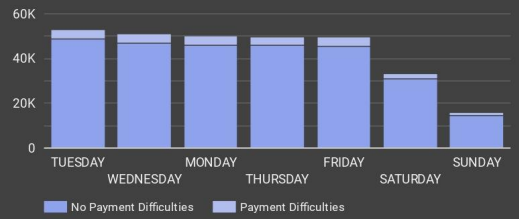
Non-Performing Loan by Occupation Type



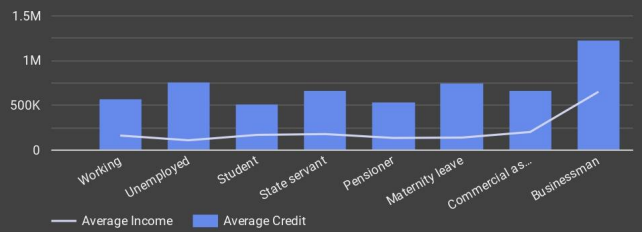
Non-Performing Loan by Income Type



Non-Performing Loan by Loan Application Day



Average Income and Credit by Income Type



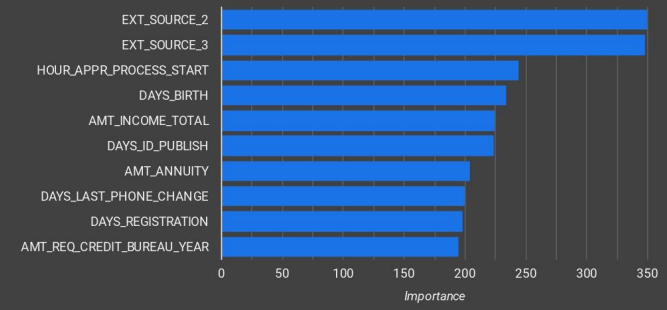
Home Credit Dashboard

Evaluation Metrics - XGBoost Model

Accuracy	Precision	F1 Score
0.92	0.4	0.07

Recall	ROC AUC Score
0.04	0.52

Top 10 Feature Importance



CONCLUSION

Conclusion

- **Model klasifikasi terbaik** yang didapatkan untuk memprediksi nasabah yang gagal membayar pinjaman pada data Home Credit adalah **model XGBoost**.
- Terdapat **10 Faktor yang berpengatuh pada prediksi gagalnya pembayaran kredit** pada nasabah antara lain **EXT_SOURCE_2**, **EXT_SOURCE_3**, **HOUR_APPR_PROCESS_START**, **DAYS_BIRTH**, **AMT_INCOME_TOTAL**, **DAYS_ID_PUBLISH**, **AMT_ANNUITY**, **DAYS_LAST_PHONE_CHANGE**, **DAYS_REGISTRATION**, dan **AMT_REQ_CREDIT BUREAU_YEAR**.
- **EXT_SOURCE_2**, **EXT_SOURCE_3** dapat menjadi faktor berpengaruh pada kegagalan pembayaran kredit karena **skor yang dinormalisasi dari sumber data eksternal** tersebut dapat **memberikan indikasi tentang kelayakan kredit peminjam**.
- Pada faktor **DAYS_BIRTH** Nasabah yang **tidak mengalami kesulitan pembayaran** berada di rentang **usia 35-45 tahun**, sedangkan nasabah yang **mengalami kesulitan pembayaran** berada di rentang **usia 25-35 tahun**.

Terima kasih!

Ada pertanyaan?