Final Project Presentation

Nomor Kelompok: 20
Nama Mentor: Muhammad Ramdhan Hidayat
Nama:

- Christine Amanda
- Dominikus Leki Sogen
- Lopia Santri
- Luthfita Darwis

Accelerated Machine Learning Program

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Business Understanding

Business Understanding

Business Objective :

- Banyak orang yang mengajukan permohonan kredit ke Bank dan Institusi Finansial lainnya. Akan tetapi, hanya sebagian kecil yang permohonan pengajuan kreditnya diterima.
- Umumnya, untuk menilai kelayakan peminjam, baik bank maupun institusi finansial lainnya akan mengevaluasi Riwayat kredit mereka sehingga sebagian besar pemohon yang Riwayat kreditnya tidak sesuai atau bahkan tidak memiliki akan ditolak.
- Pemohon yang ditolak cenderung mencari alternatif lain untuk meminjam dana dan seringkali mengalami kerugian oleh pihak yang tak bertanggung jawab.
- Home Credit berupaya membantu masyarat agar memperoleh pengalaman positif dalam mengajukan dana dengan memperluas prediksi melalui banyak data (seperti Telco dan Transaksi) dalam memprediksi kelayakan pemohon.

Business Understanding

Business Success Criteria

- Home Credit dapat memperluas jangkauan penerima pinjaman
- Kemungkinan terjadinya *Default Risk* rendah atau bahkan tidak ada.

Problem :

Menentukan apakah pemohon mampu atau tidak mampu untuk membayar credit (Classification)

Goals :

Dapat memprediksi kemampuan pemohon dalam membayar kredit untuk menghindari Default Risk.

Project Plan :

- Algoritma Machine Learning yang akan digunakan pada kasus ini adalah Supervised Machine Learning.

Business Understanding

Produce Project Plan :

- Mengumpulkan data yang dapat membantu mengidentifikasi kemampuan pemohon untuk membayar pinjaman.
- Memanfaatkan algoritma Supervised Machine Learning untuk membuat model yang dapat membantu mengidentifikasi kelayakan pemohon.

Data Understanding

Data Understanding

Sumber Data :

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

Dataset :

- application_train
- bureau
- bureau_balance
- POS_CASH_balance
- credit_card_balance
- previous_application
- Installment_payments

Data Understanding

application_train :

- Menyediakan informasi mengenai data lamaran, demografis, dan riwayat kredit pemohon.
- Terdapat variable ‘TARGET’ yang mengindikasikan apakah pemohon mampu membayar ataupun tidak.
- Dataset terdiri dari 122 columns (106 numerical dan 16 categorical) dan 307511 baris.

application_test :

- Memiliki informasi yang sama dengan application_train, hanya saja pada dataset ini tidak terdapat variable ‘Target’
- Dataset terdiri dari 121 columns (105 numerical dan 16 categorical) dan 48744 baris

Data Understanding

bureau

- Menyediakan informasi pemohon dari institusi-institusi lain yang dilaporkan ke Home Credit's Credit Bureau .
- Dataset terdiri dari 17 columns (14 numerical dan 3 categorical) dan 1716428 baris.

bureau_balance

- Memuat informasi jumlah dana yang harus dibayarkan setiap bulannya pada credit terdahulu di Credit Bureau
- Dataset terdiri dari 3 column (2 numerical dan 1 categorical) dan 27299925 baris

Data Understanding

POS_CASH_balance

- Memuat informasi POS (*point of sales*) sebelumnya dan *cash loans* yang pemohon miliki di Home Credit
- Dataset terdiri dari 8 columns (7 numerical dan 1 categorical) dan 10001358 baris.

credit_card_balance

- Memiliki informasi mengenai riwayat credit card terdahulu yang dimiliki oleh pemohon di Home Credit
- Dataset terdiri dari 23 column (22 numerical dan 1 categorical) dan 3840312 baris

Data Understanding

previous_application

- Memuat informasi lamaran pemohon terdahulu yang ditujukan kepada Home Credit
- Dataset terdiri dari 37 columns (7 numerical dan 1 categorical) dan 1670214 baris.

Installments_payments

- Memuat informasi riwayat pembayaran untuk kredit yang telah dicairkan sebelumnya di Home Credit
- Dataset terdiri dari 8 column (8 numerical dan 0 categorical) dan 13605401 baris

Data Exploration : application_train

Missing Data :

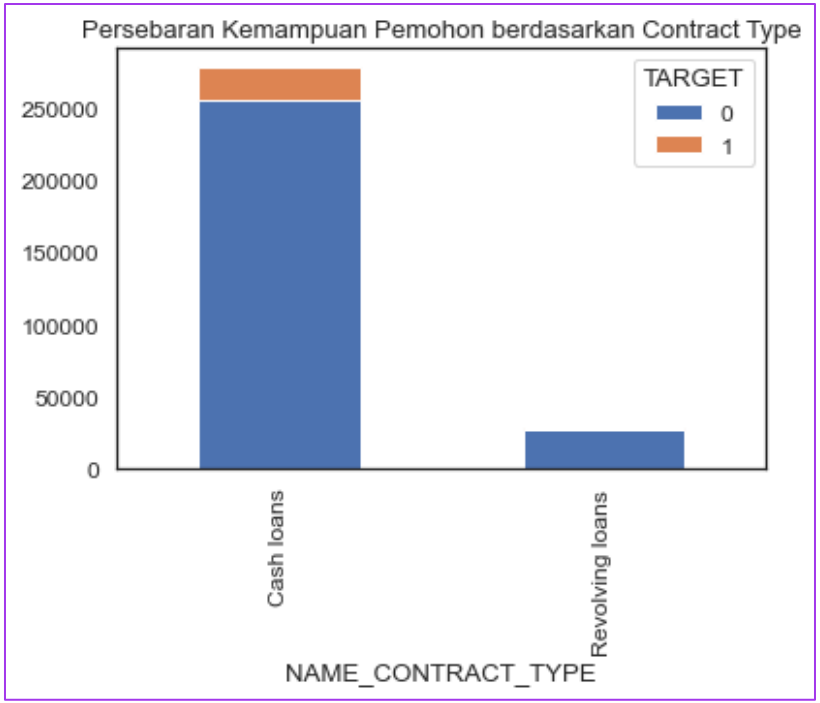
- Terdapat 67 dari 122 columns yang memiliki missing data.
- Terdapat 41 columns yang kisaran persentase missing datanya 70%-50%, diikuti 9 columns dengan persentase missing data 50%-30%, 7 columns dengan persentase missing data 30%-10%, dan 17 sisanya pada persentase di bawah 10%

Data Exploration : application_train

- Persentase pemohon yang mampu mebayar adalah 91.9% sedangkan 8.07% sisanya tidak mampu untuk membayar kembali
- Persentase ini menunjukkan nilai imbalance yang cukup besar

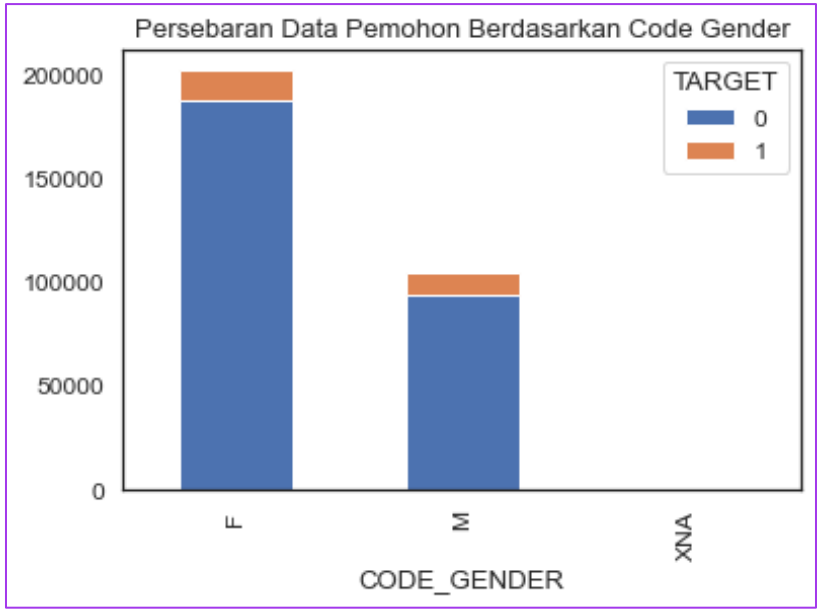
Data Exploration : application_train

- Jumlah pemohon yang mengajukan tipe kontrak berupa Cash Loans jauh lebih besar dibandingkan yang mengajukan kontrak tipe Revolving Loans



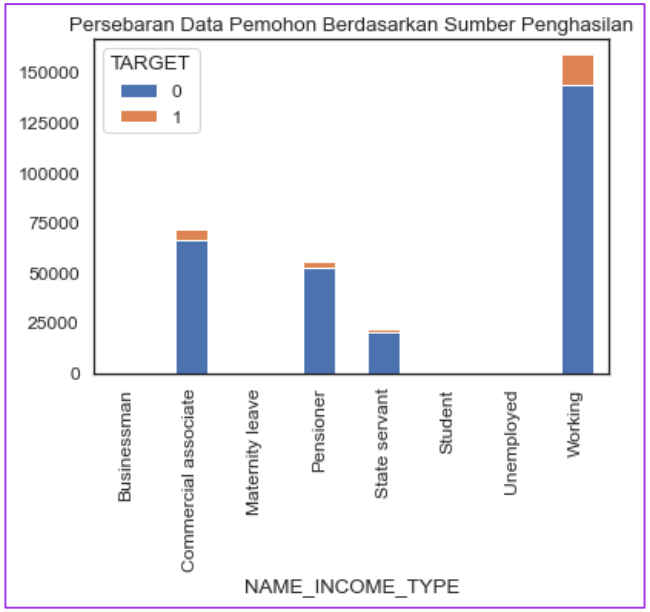
Data Exploration : application_train

- Sebagian besar pemohon yang mengajukan permohonan adalah Perempuan



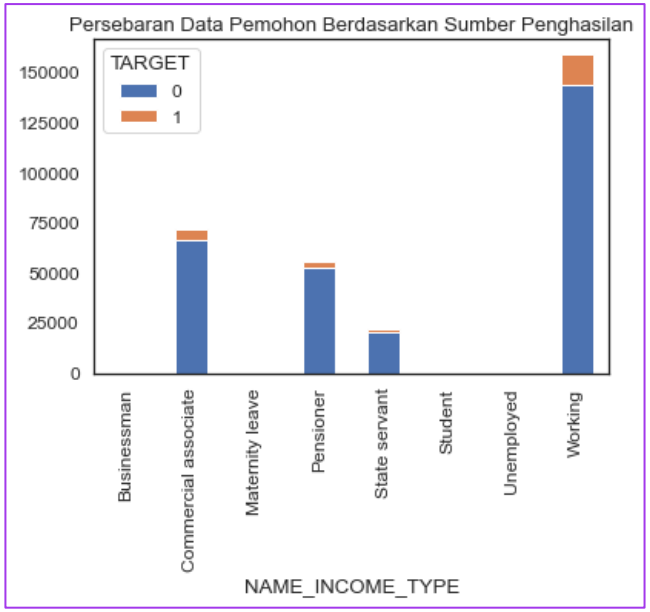
Data Exploration : application_train

- Sebagian besar pemohon memperoleh sumber penghasilannya dari bekerja. Akan tetapi, data tidak menunjukkan informasi spesifik profesi yang dimiliki pemohon.
- Comercial Associate menempati posisi kedua, diikuti oleh Pensioner, dan State Servant



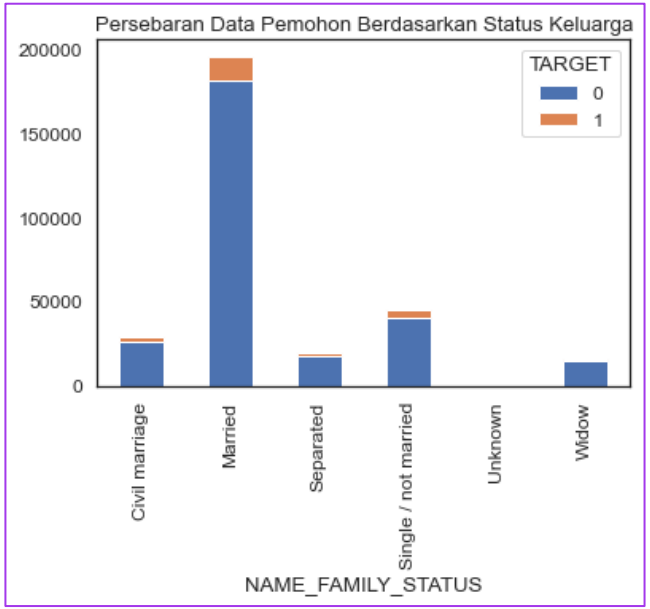
Data Exploration : application_train

- Sebagian besar pemohon memperoleh sumber penghasilannya dari bekerja. Akan tetapi, data tidak menunjukkan informasi spesifik profesi yang dimiliki pemohon.
- Comercial Associate menempati posisi kedua, diikuti oleh Pensioner, dan State Servant



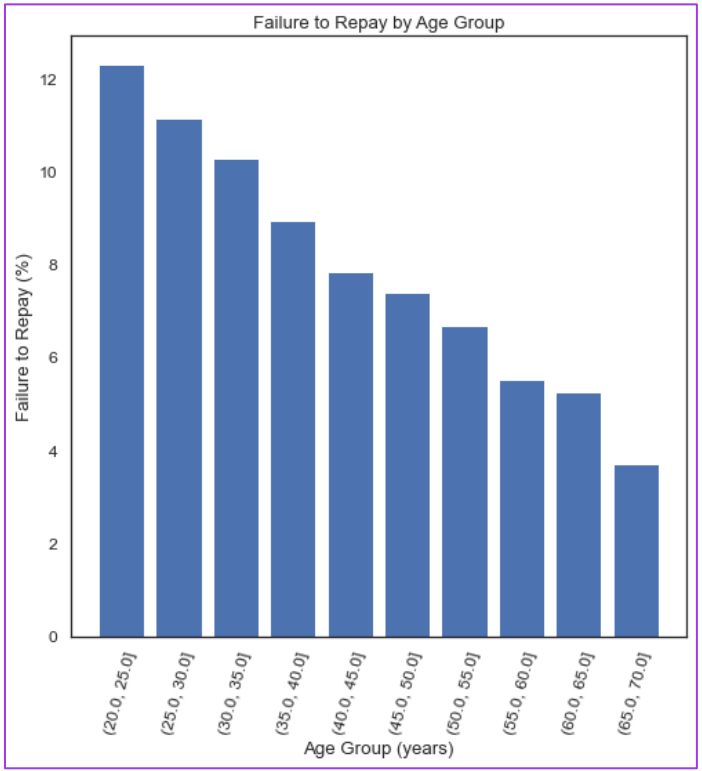
Data Exploration : application_train

- Pemohon yang sudah menikah paling banyak mengajukan credit.



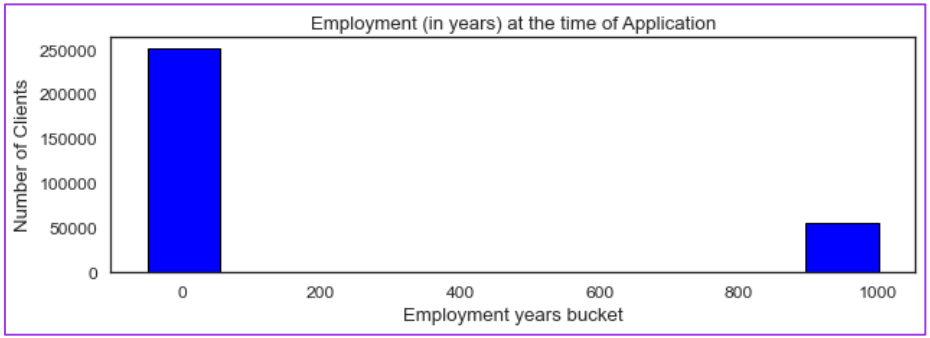
Data Exploration : application_train

- Pemohon yang berusia muda cenderung lebih berpotensi untuk tidak mampu membayar kembali pinjaman mereka.
- Sementara itu, semakin tuanya pemohon semakin rendah kecendrungan mereka untuk tidak mampu membayar pinjaman.



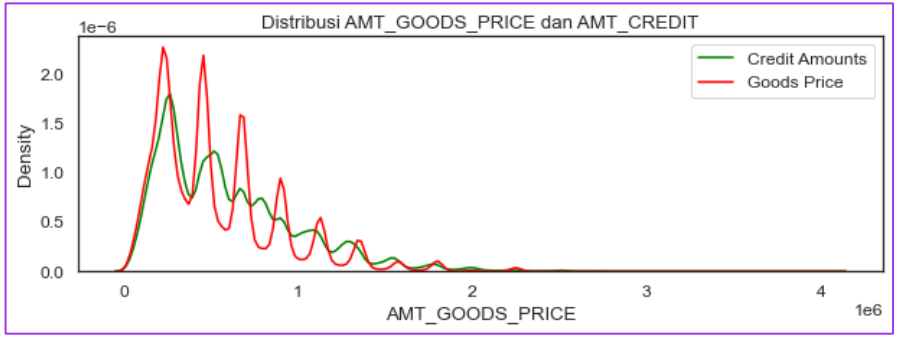
Data Exploration : application_train

- Ada pemohon yang bekerja kurang lebih 1000, hal ini tentunya tidak mungkin sehingga dianggap sebagai outliers
- Outliers yang ada pada column ‘DAYS_EMPLOYED’ akan ditangani nanti.



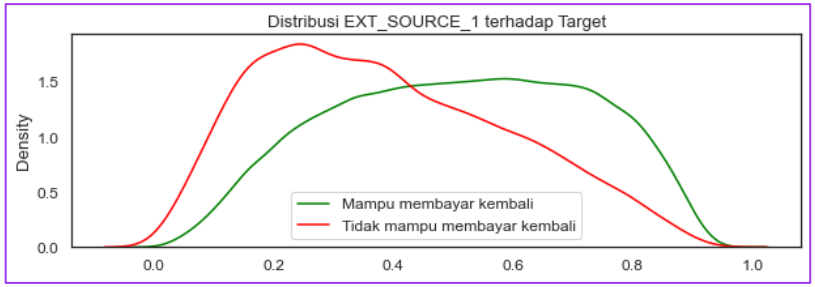
Data Exploration : application_train

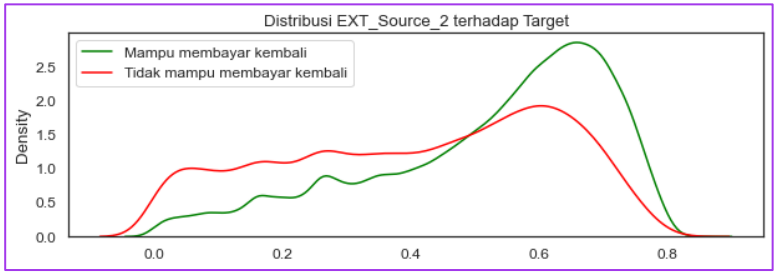
- Distribusi AMT_GOODS_PRICE dan AMT_CREDIT cenderung berimbang.
- Akan tetapi pada beberapa kondisi nilai Credit Amounts jauh di atas AMT_GOODS_PRICE dan begitu pula sebaliknya.

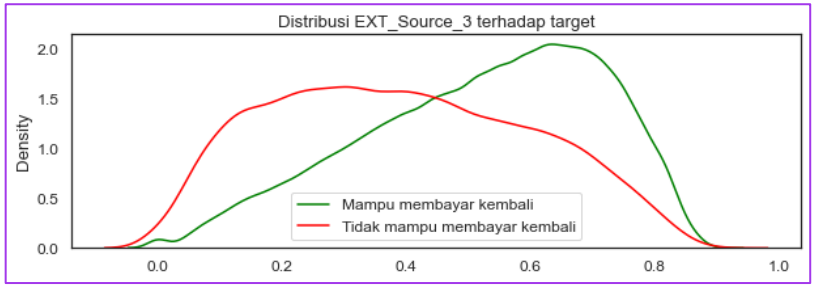


Data Exploration : application_train

- Distribusi antara EXT_SOURCE_1, EXT_SOURCE_2, dan EXT_SOURCE_3 dengan Target menunjukkan perbedaan yang cukup signifikan. Sehingga ketiganya dapat digunakan untuk memprediksi kapabilitas pemohon.

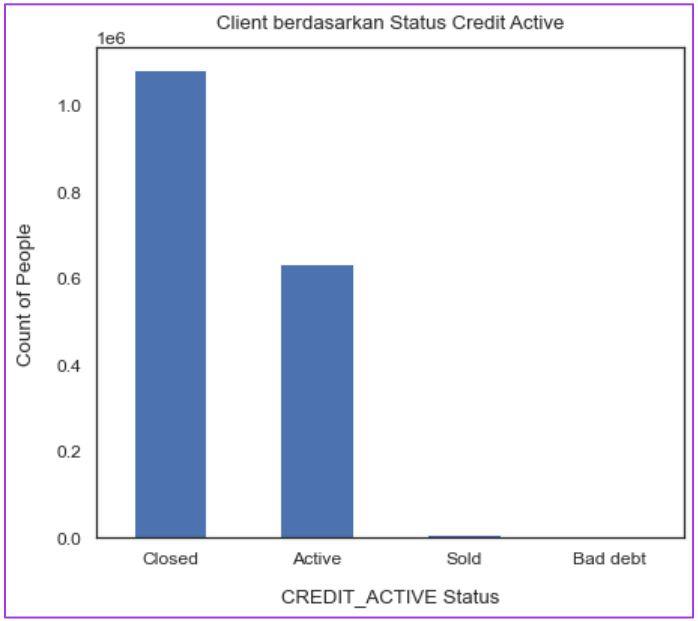






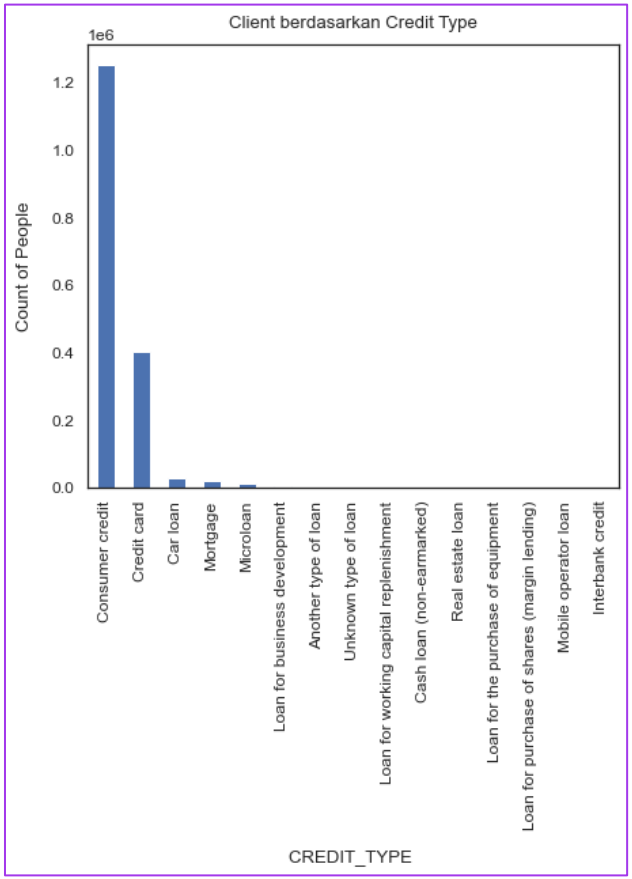
Data Exploration : Bureau

- Sebagian besar Status Credit pemohon pada Bureau adalah ‘Closed’.
- Status credit yang masih ‘Active’ menempati urutan kedua.
- Sedangkan, sebagian kecil berstatus ‘Sold’ dan ‘Bad Debt



Data Exploration : Bureau

- Sebagian besar tipe kredit yang dilaporkan ke Bureau adalah Customer Credit.



Data Exploration : Days

- Segala data yang memuat informasi jumlah hari seperti ‘DAYS_BIRTH’ bernilai negatif karena data tersebut ditinjau dari hari sebelum mendaftar ke Home Credit.

Data Preparation

Data Cleansing : application_train

Kesalahan pada dataset :

- Terdapat 67 column yang memiliki missing value. Solusinya adalah dengan menghapus column yang memiliki missing value di atas 30% dan mengisi column sisanya dengan Median dan Modus.
- Column “DAYS_EMPLOYED” memiliki nilai anomali. Solusinya adalah dengan mengubah nilai anomali menjadi missing_values dengan np.nan dan membuat column “DAYS_EMPLOYED_ANOM” untuk mengindikasikan apakah column tersebut anomali atau tidak.

Feature Engineering : application_train

- Dibuat column baru berdasarkan Domain Knowledge : “CREDIT_INCOME_PERCENT”, “ANNUITY_INCOME_PERCENT”, “CREDIT_TERM”, “DAYS_EMPLOYED_PERCENT”

Data Exploration : Bureau

- Sebagian besar tipe kredit yang dilaporkan ke Bureau adalah Customer Credit.

Data Cleansing : bureau

Kesalahan pada dataset :

- Terdapat 7 column yang memiliki missing value. Solusinya adalah menghapus column yang memiliki missing value di atas 30%

Integrate Data : Bureau and Bureau_Balance

- Dataset Bureau dan Bureau_Balance digabung dengan menggunakan ‘Inner Join’ dan ‘SK_ID_BUREAU’ sebagai key.
- Dataset gabungan memiliki 15 columns dan 24179741 baris

Correlation: Bureau and Bureau_Balance

- Correlation antara data gabungan Bureau dan Bureau_Balance dengan variable ‘Target’ memiliki nilai nyaris menyentuh 0.
- Data-data dari column ini akan dianggap tidak memberi pengaruh signifikan kepada ‘Target’ oleh karena itu tidak digabungkan dengan application_train

Modelling dan Evaluation

One Hot Encoding

- Dilakukan One Hot Encoding pada features categorical.
- Features tersebut adalah : 'NAME_CONTRACT_TYPE', 'CODE_GENDER','FLAG_OWN_CAR','FLAG_OWN_REALTY','NAME_TYPE_SUITE', 'NAME_INCOME_TYPE','NAME_EDUCATION_TYPE','NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE','WEEKDAY_APPR_PROCESS_START','ORGANIZATION_TYPE'

Oversampling Data

- Karena Data ‘Target’ imbalance dengan majority class ‘0’ bertotal 282686 dan minority class ‘1’ bertotal 24825, akan dilakukan oversampling.
- Oversampling dilakukan dengan menggunakan SMOTE
- Hasil Oversampling adalah kedua target ‘0’ dan ‘1’ bernilai sama yakni 282686

Train Test Split

- Menggunakan train test set dengan perbandingan 60:40
- Train Test Set menggunakan semua feature columns setelah cleaning kecuali ‘SK_ID_CURR’ sebagai variable x dan ‘Target’ sebagai y.

Metrik Evaluasi & Model

Metrik yang digunakan untuk melakukan evaluasi model, yaitu:

- ROC AUC

Model yang digunakan untuk melakukan prediksi data adalah:

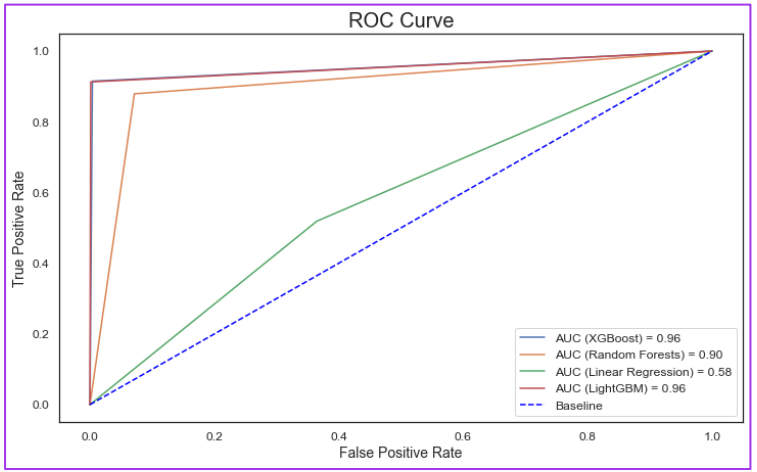
- XGBOOST Classifier
- Random Forest Classifier
- Logistic Regression
- LightGBM Classifier

Evaluasi Model

- XGBoost
  - Accuracy : 0.955511631711836
  - AUC : 0.9554970757883456
- Random Forest Classifier
  - Accuracy : 0.9038819539330265
  - AUC : 0.9038730835029606
- Logistic Regression
  - Accuracy : 0.5770841347960858
  - AUC : 0.577063313976429
- LightGBM
  - Accuracy :0.955763677929153
  - AUC : 0.9557482715464195

Evaluasi Model

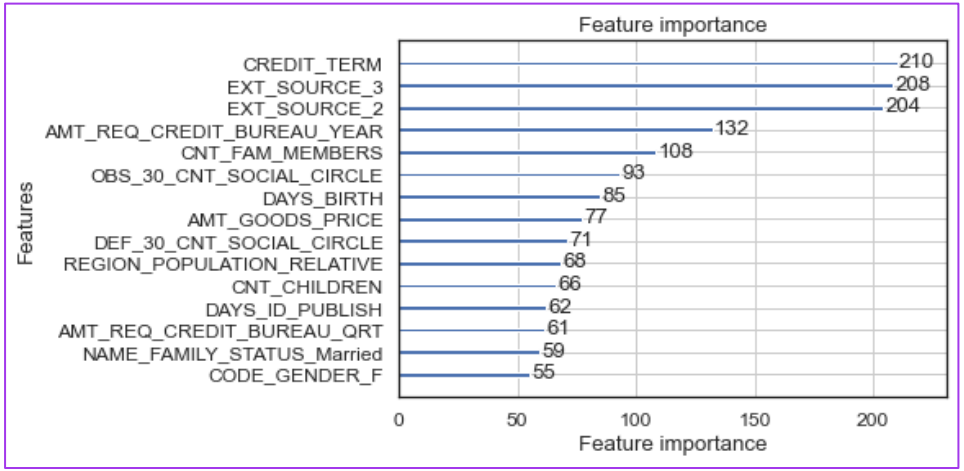
- Seluruh model tampak berada jauh dari baseline
- Akurasi XGBoost dan LightGBM menempati posisi tertinggi dari model lainnya.
- Akurasi Linear Regression berada di posisi terendah.



Feature Selection

- Menggunakan dua train test set yang berbeda. Kedua train test set memiliki perbandingan 60:40.

- Train Test set 1 menggunakan lima belas *features* yang dipilih menggunakan LightGBM Feature Importances sebagai variable X



Feature Selection

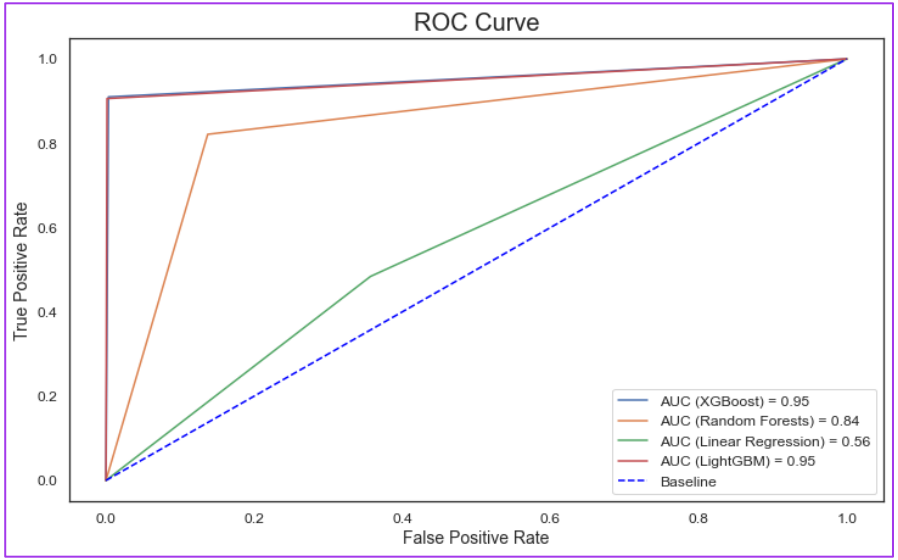
- Train Test Set 2 menggunakan 15 features yang diambil dari Random Forest Feature Importance. Features tersebut antara lain : EXT_SOURCE_3, CODE_GENDER_F, NAME_EDUCATION_TYPE_Higher education, EXT_SOURCE_2, FLAG_OWN_CAR_Y, OBS_60_CNT_SOCIAL_CIRCLE, FLAG_PHONE, NAME_FAMILY_STATUS_Married, OBS_30_CNT_SOCIAL_CIRCLE, AMT_REQ_CREDIT_BUREAU_YEAR, FLAG_OWN_REALTY_Y, FLAG_OWN_REALTY_N , NAME_INCOME_TYPE_Commercial associate, WEEKDAY_APPR_PROCESS_START_WEDNESDAY, CNT_FAM_MEMBERS

Evaluasi Model (Train Test Set 1)

- XGBoost
  - Accuracy : 0.9526639516425012
  - AUC : 0.8261809691840335
- Random Forest Classifier
  - Accuracy : 0.8419228031076856
  - AUC : 0.8261779709567808
- Logistic Regression
  - Accuracy : 0.5000862263375031
  - AUC : 0.499914766771488
- LightGBM
  - Accuracy : 0.9511693617924466
  - AUC : 0.9511528921017992

Evaluasi Model

- Keseluruhan model berada pada posisi jauh dari baseline.
- Pada Trainset1 ini, Model XGBoost dan LightGBM tetap menempati posisi terjauh dari baseline
- Accuracy Random Forest menurun jika dibandingkan pada Basic Model.

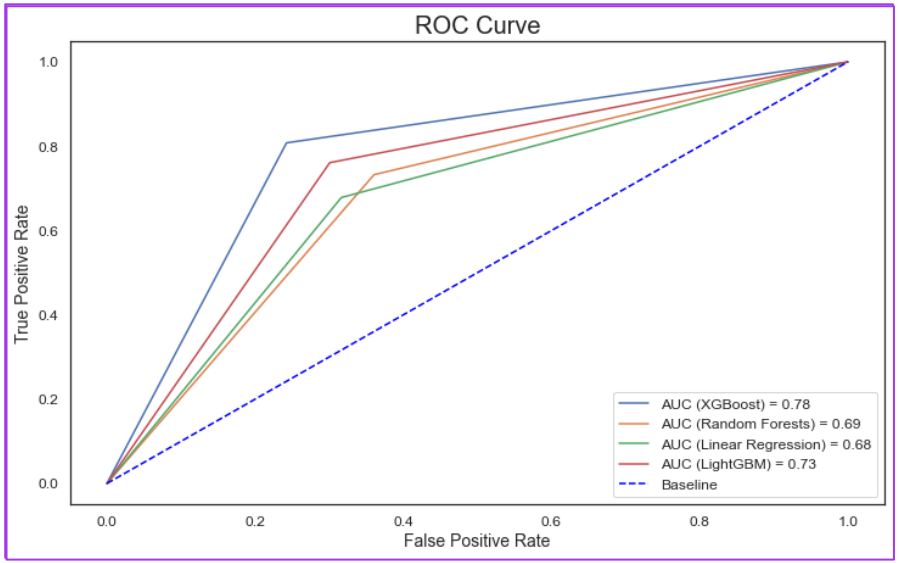


Evaluasi Model (Train Test Set 2)

- XGBoost
  - Accuracy 0.7854379192479295
  - AUC : 0.7854457266980025
- Random Forest Classifier
  - Accuracy : 0.6858929289981384
  - AUC : 0.6859093933071627
- Logistic Regression
  - Accuracy : : 0.680944863784496
  - AUC : 0.6809431725519569
- LightGBM
  - Accuracy : 0.7304387815113045
  - AUC : 0.7304494485533217

Evaluasi Model

- Model XGBoost menempati posisi terjauh dengan nilai akurasi 0.78 disusul oleh LightGBM dengan akurasi 0.73
- Accuracy Random Forest dan Linear Regression hampir berimbangan dengan perbedaan 0.01
- XGBoost, Random Forest, dan LightGBM mengalami penurunan performa.



Hyperparameter Tuning

- Untuk model LightGBM menggunakan Teknik randomized search on hyper parameters. Dengan rentang parameter :
  - 'colsample_bytree': 0.9234,
  - 'min_child_samples': 399,
  - 'min_child_weight': 0.1,
  - 'num_leaves': 13,
  - 'reg_alpha': 2,
  - 'reg_lambda': 5,
  - 'subsample': 0.855

Evaluasi Model

- LightGBM
  - Accuracy : 0.9559405524676209
  - AUC : 0.9558944353782046

Model Final

- Menggunakan Model LightGBM
- Menggunakan lima belas *features* yang dipilih menggunakan Light GBM Feature Importance sebagai variabel X (Train Test Set 1)

Deployment

Deployment

Deployment Plan :

- Pertama, model yang dipilih dapat dimanfaatkan oleh bisnis untuk menentukan kelayakan pemohon.
- Kedua, akan ditentukan lokasi hosting untuk mendeploy model. Disini direncanakan menggunakan Herokuapp.
- Ketiga, tidak hanya memprediksi tetapi deployment juga akan menyajikan dashboard yang menampilkan karakteristik dari setiap pemohon dan faktor yang mempengaruhi kelayakan mereka.

Deployment

Monitoring and Maintenance

- Monitoring dan Maintenance akan dilakukan dengan memanfaatkan Network Health Dashboard untuk mengecheck permasalahan jaringan dan bagaimana menavigasikan dashboard dengan baik.

Conclusion

Kesimpulan

- Sebagian besar pemohon yang mengajukan pinjaman adalah mereka yang sudah bekerja.
- Kemampuan pemohon dalam membayar kembali pinjaman ditentukan oleh banyak faktor.
- Informasi pribadi pemohon dapat membantu mengidentifikasi kemampuan mereka dalam membayar kembali pinjaman.
- Pemohon yang berusia tua cenderung lebih mampu untuk membayar kembali pinjamannya.

Saran

- Home Credit perlu memperkaya data mereka dengan informasi yang lebih spesifik agar proses prediksi menjadi lebih maksimal.
- Faktor-faktor kuat yang dapat digunakan Home Credit dalam penentuan kemampuan pemohon antara lain :
  - Informasi pribadi pemohon (seperti Usia, Pekerjaan, Status)
  - Demografis
  - Riwayat kredit terdahulu

Link Pengerjaan

https://colab.research.google.com/drive/116a4Yyap6Ws5Yum4BKlmtSGhxRjppMl5?usp=sharing

Terima kasih!

Ada pertanyaan?