Rakamin Academy X HOME CREDIT

Home Credit Scorecat Model

Home Credit Indonesia
Data Scientist

Presented by
Muhammad Zaki

Muhammad Zaki

About Me

Computer Engineering | Passionate about Data | Seeking to Elevate Career in Data-Driven Innovations

Rakamin Academy

Experience

Research Assistant
Juli – Agustus 2021
Research Assistant on Web Scraping and Data Analyst Project (PT PROCODECG)

Data Scientist ID/X Partners, Projec Based Intern
May – Juni 2023
Create a prediction model of customers' ability to repay loans

Data Scientist Kalbe Nutritional, Project Based Intern
Juli – Agustus 2023
create a daily sales prediction model and customer segmentation

Business Understanding

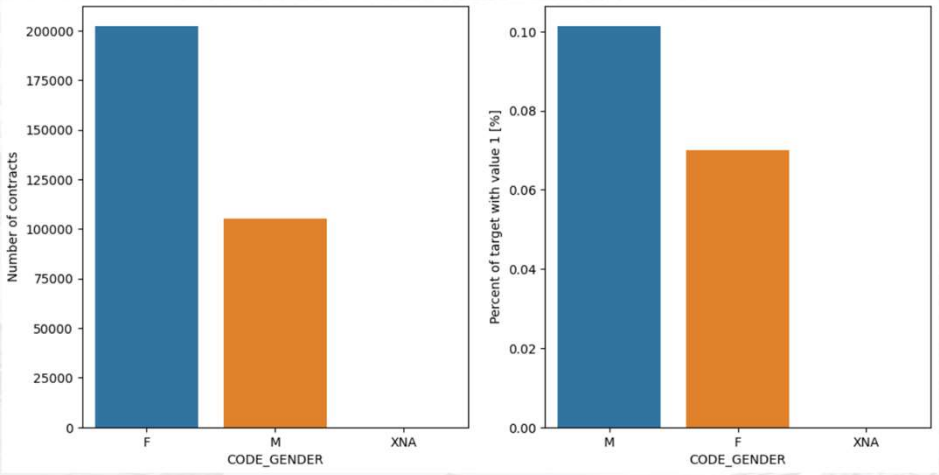
Permasalahan Umum

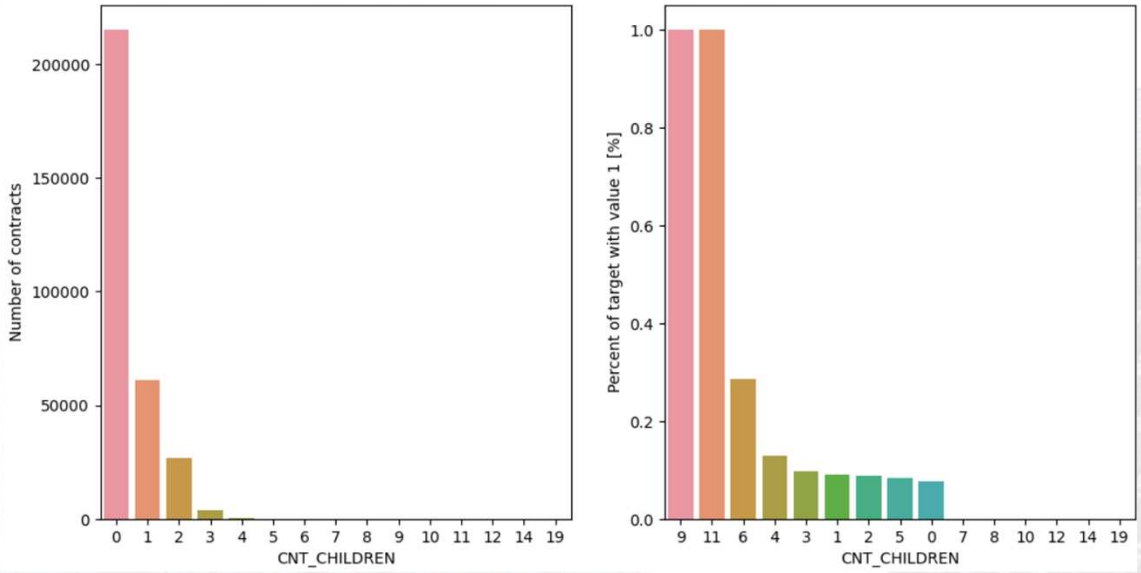
Bagaimana cara untuk memilih klien dan mengetahui bisa atau tidaknya klien membayar pinjaman yang telah diberikan

Tujuan

Meningkatkan efisiensi dalam pemilihan klien dan menghindari peminjaman kepada klien yang berpotensi gagal dalam pembayaran

EDA





Jumlah klien wanita hampir 2 kali jumlah klien peria, tetapi klien pria memiliki persentasi lebih tinggi dalam tidak bisa melunasi pinjaman.

pada umumnya klien memiliki jumlah kurang dari 3 anak. untuk anak dengan jumlah 1-5 memiliki presentasi kurang dari 15% gagal bayar. untuk anak dengan jumlah 6 memiliki kemungkinan gagal bayar 25% dan untuk jumlah anak 9 dan 11 memiliki tingkat kegagalan bayar 100%

Data Preparation

Rakamin Academy

Feature Importance

Fungsi korelasi menampilkan variabel apa saja yang memiliki hubungan dengan kemampuan klien.
Dengan melihat hasil dari perhitungan korelasi tesebut, kita akan menggunakan 4 variable yang memiliki korelasi dengan kemampuan pelunasan pinjaman klien atau “TARGET”.

Yaitu:
Umur dan sumber dari luar(
EXT_SOURCE_1,
EXT_SOURCE_2,
EXT_SOURCE_3) yang merupakan hasil standarisasi yang telah dibuat sebelumnya

```
Most Positive Correlations:
 DAYS_REGISTRATION                                0.041975
OCCUPATION_TYPE_Laborers                          0.043019
FLAG_DOCUMENT_3                                   0.044346
REG_CITY_NOT_LIVE_CITY                            0.044395
FLAG_EMP_PHONE                                    0.045982
NAME_EDUCATION_TYPE_Secondary / secondary special 0.049824
REG_CITY_NOT_WORK_CITY                            0.050994
DAYS_ID_PUBLISH                                   0.051457
CODE_GENDER_M                                     0.054713
DAYS_LAST_PHONE_CHANGE                            0.055218
NAME_INCOME_TYPE_Working                          0.057481
REGION_RATING_CLIENT                              0.058899
REGION_RATING_CLIENT_W_CITY                       0.060893
DAYS_BIRTH                                        0.078239
TARGET                                            1.000000
Most Negative Correlations:
 EXT_SOURCE_3                                    -0.178919
EXT_SOURCE_2                                     -0.160472
EXT_SOURCE_1                                     -0.155317
NAME_EDUCATION_TYPE_Higher education             -0.056593
CODE_GENDER_F                                    -0.054704
NAME_INCOME_TYPE_Pensioner                       -0.046209
ORGANIZATION_TYPE_XNA                            -0.045987
DAYS_EMPLOYED                                    -0.044932
FLOORSMAX_AVG                                    -0.044003
FLOORSMAX_MEDI                                   -0.043768
FLOORSMAX_MODE                                   -0.043226
EMERGENCYSTATE_MODE_No                           -0.042201
HOUSETYPE_MODE_block of flats                    -0.040594
AMT_GOODS_PRICE                                  -0.039645
REGION_POPULATION_RELATIVE                       -0.037227
```

Data Preparation

Feature Engineering

Melihat dari sumber diskusi pada dataset Home Credit Default Risk mereka membahkan 4 fitur, yaitu:

CREDIT_INCOME_PERCENT	-> persentase jumlah kredit relatif terhadap pendapatan klien.
ANNUITY_INCOME_PERCENT	-> persentase anuitas pinjaman relatif terhadap pendapatan klien
CREDIT_TERM	-> lamanya pembayaran dalam bulan (karena anuitas adalah jumlah bulanan yang harus dibayar
DAYS_EMPLOYED_PERCENT	-> persentase hari kerja relatif terhadap usia klien

Modeling

Random Forest Classifier

Model yang digunakan yaitu random forest classifier yang memiliki peforma yang baik dalam menangani hubungan yang kompleks antar variabel.

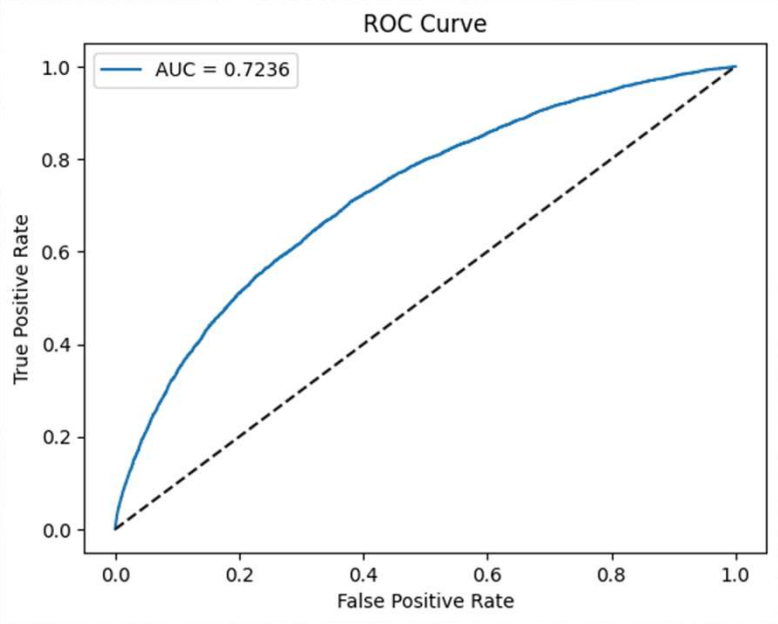
Maksimal kedalaman yang digunakan adalah 4 untuk mencegah model terlalu overfitting.

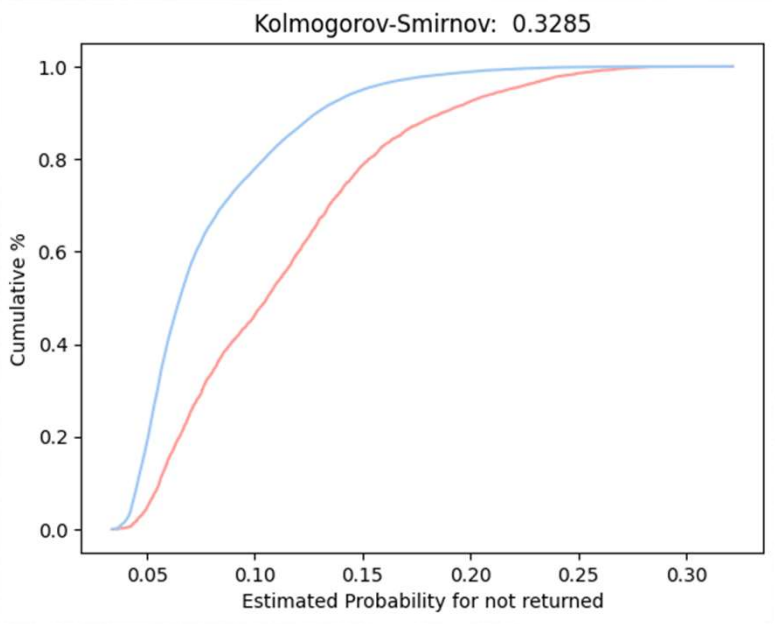
Matrik evaluasi yang digunakan adalah AUC dan ROC.

Penggunaan matriks AOC dan ROC dikareanakan data yang tidak seimbang. Data untuk berhasil bayar terlalu banyak dibanding gagal bayar, untuk data seperti itu Matriks ini memiliki keunggulan.

Uji KS atau uji Kolmogorov-Smirnov digunakan untuk memahami bagaimana perbedaan antara distribusi kumulatif prediksi "Good“ atau bisa bayar dan "Bad“ atau gagal bayar berubah seiring dengan perubahan nilai prediksi probabilitas.

Conclution





Model menghasilkan nilai AUC = 0.7236 dan KS = 0.3285, nilai tersebut menunjukkan bahwa model bisa bekerja dengan baik dalam melakukan pemisahan kelas.

Insert Your Result Here

You can add **image or link** result. You can add an explanation of how you got the result also.

https://github.com/kaniang/HCI_HomeCreditDefaultRisk


Thank You
Rakamin
Academy
X
HOME
CREDIT