MODELLING CLIENTS' PAYMENT DIFFICULTIES

Yayang Nurul Aulia Azizah (Statistics Student)
Virtual Internship Experience Data Scientist at Home Credit Indonesia by Rakamin Academy

Referensi: https://github.com/yangswei112/fsb-project2

Link Github: https://github.com/yangswei112/vixrakamin

Problem Research

Business Question

- Berapa persen client yang mengalami kesulitan pembayaran di Home Credit Indonesia?
- Kriteria client seperti apa yang tidak mengalami kesulitan pembayaran dengan syarat pinjaman sebelumnya diterima di Home Credit Indonesia?

Goals

- Mengidentifikasi penyebab seorang client mengalami kesulitan pembayaran di Home Credit Indonesia.
- Membangun Machine Learning untuk mendeteksi client yang akan mengalami kesulitan pembayaran di Home Credit Indonesia.

Data Preprocessing
1. Check & Handling Missing Values
2. Check & Handling Duplicated Values
3. Check & Handling Outliers
4. Check & Handling Inconsistent Data

Dataset yang digunakan adalah **application_{train|test}.csv** dan **previous_application.csv**

```
# drop column with missing values > 50% of rows
for col in apl:
    if apl[col].isna().sum() > (apl.shape[0])/2:
        apl = apl.drop(col, 1)
```

✓ 14.8s

```
# check duplicated values
print(apl.duplicated().sum())
```

9] ✓ 1.7s

```
0
```

```
# check inconsistent data
for col in apl_filtered[catcol]:
    print(apl_filtered[col].value_counts(), '\n')
```

✓ 0.9s

```
# handle missing values categorical col using mode
for col in apl[catcol]:
    apl[col].fillna(apl[col].mode()[0], inplace=True)
```

✓ 1.1s

```
# handle missing values numerical data using median
for col in apl[numcol2]:
    apl[col].fillna(apl[col].median(), inplace=True)
```

✓ 2.3s

```
# handle outlier using z-score
from scipy import stats
print(f'jumlah baris sebelum difilter : {len(apl)}')

filtered_entries = np.array([True]*len(apl))
for col in numcol2:
    zscore = abs(stats.zscore(apl[col]))
    filtered_entries = (zscore < 3) & filtered_entries
apl_filtered = apl[filtered_entries]

print(f"jumlah baris setelah difilter: {len(apl_filtered)}")
```

✓ 1.1s

```
jumlah baris sebelum difilter : 307511
jumlah baris setelah difilter: 248189
```

Feature Engineering
- Feature Scaling
- Feature Encoding
- Feature Selection

```
# feature scaling for numerical feature
from sklearn.preprocessing import MinMaxScaler, StandardScaler

for col in numcol2:
    apl_filtered[col] = MinMaxScaler().fit_transform(apl_filtered[col].values.reshape(len(apl_filtered), 1))
```

✓ 3.2s

```
# feature encoding for categorical feature
for col in catcol2:
    onehots = pd.get_dummies(apl_filtered[col], prefix=col)
    apl_filtered2 = apl_filtered.join(onehots)
```

✓ 1.7s

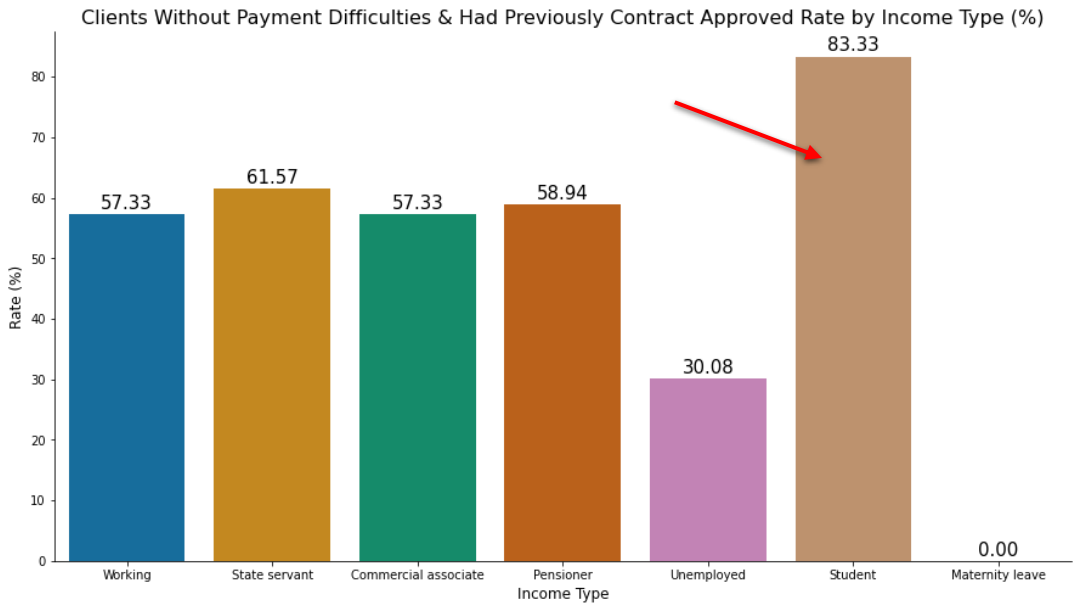
```
# drop SK_ID_CURR column
apl = apl.drop(["SK_ID_CURR"],1)
```

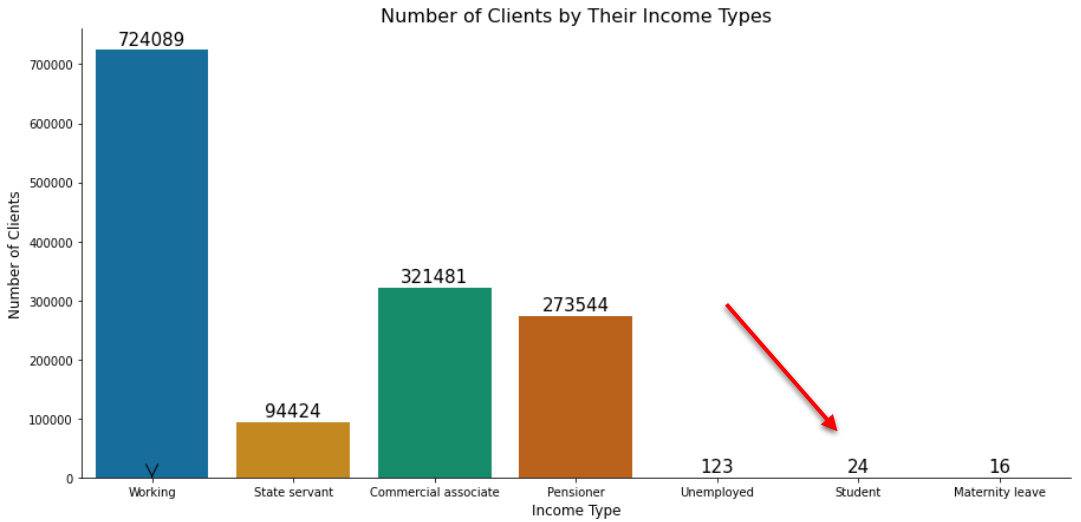
✓ 0.4s

```
# drop ORGANIZATION_TYPE cause it has many unique values
apl_filtered = apl_filtered.drop(["ORGANIZATION_TYPE"],1)
```

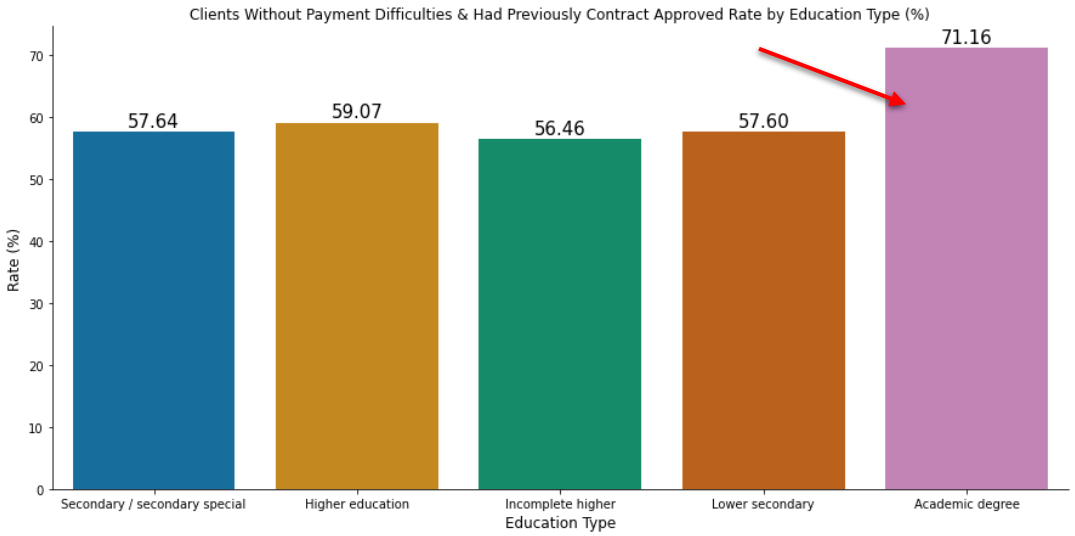
✓ 0.3s

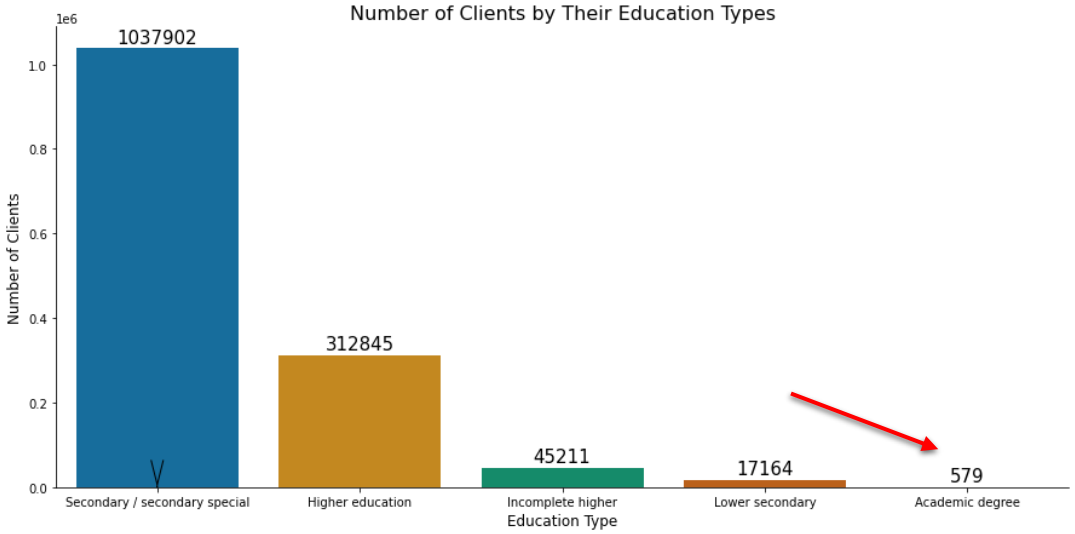
Data Visualization & Insight





Berdasarkan diagram batang di atas, Student Client memiliki rate tertinggi sebesar 83.3% untuk client yang pinjaman sebelumnya diterima & tidak mengalami kesulitan pembayaran. Namun, total Student Client yang mengajukan pinjaman hanya 24 orang atau sekitar 0.0017% dari total client yang mengajukan pinjaman. Sehingga diperlukan promosi atau campaign kepada Student Client lainnya agar semakin banyak yang tertarik untuk mengajukan pinjaman di Home Credit Indonesia.

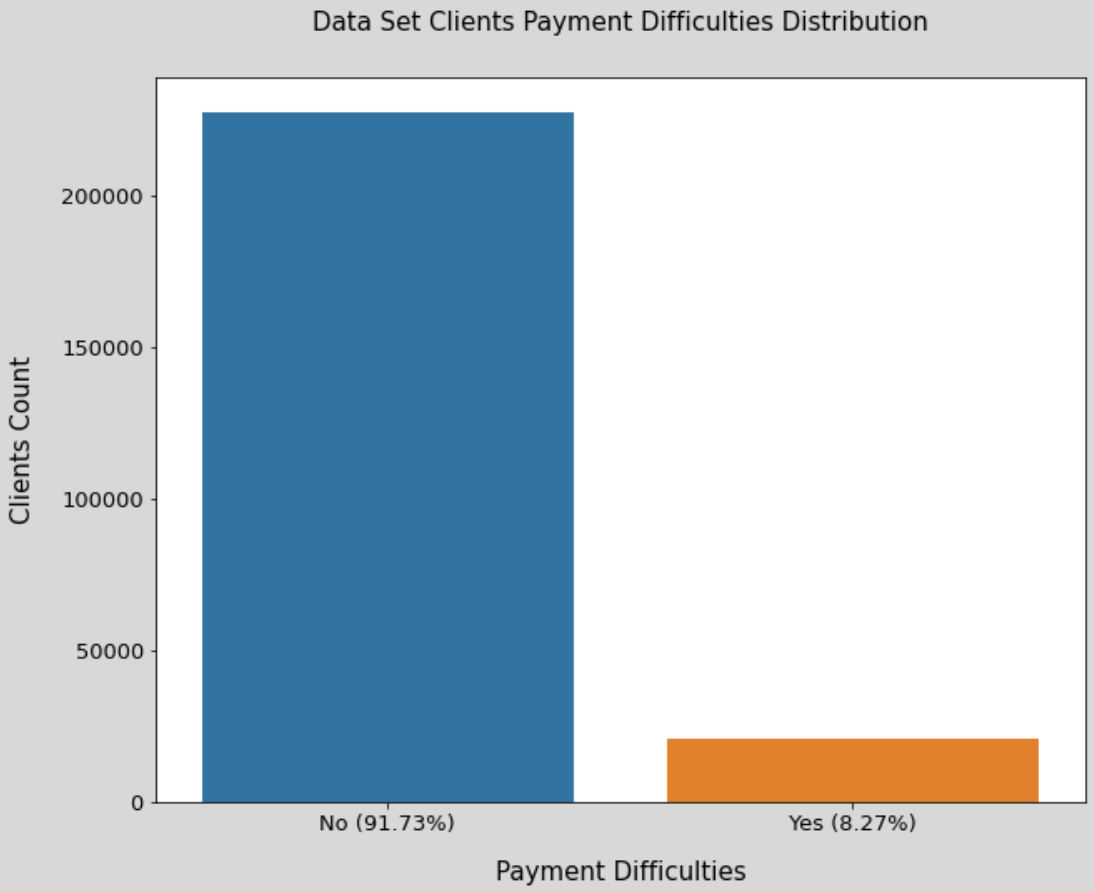




Data Visualization & Insight

Berdasarkan diagram batang di samping, Client dengan gelar akademik memiliki rate tertinggi sebesar 71,16% untuk client yang pinjaman sebelumnya diterima & tidak mengalami kesulitan pembayaran. Namun, total Client dengan gelar akademik yang mengajukan pinjaman hanya 579 orang atau sekitar 0,04% dari total client yang mengajukan pinjaman. Sehingga diperlukan promosi atau campaign kepada Client dengan gelar akademik lainnya agar semakin banyak yang tertarik untuk mengajukan pinjaman di Home Credit Indonesia.

Check & Handling Imbalanced Class




```
# handling using SMOTE
from imblearn.over_sampling import SMOTE
oversample = SMOTE()
x_smote, y_smote = oversample.fit_resample(x, y)
```

✓ 30.5s

```
y_smote.value_counts()
```

✓ 0.1s

```
TARGET
0          227672
1          227672
dtype: int64
```

20517 of 248189 clients with payment difficulties and it is the 8.27% of the data set.
227672 of 248189 clients without payment difficulties and it is the 91.73% of the data set.

Data Modelling & Evaluation

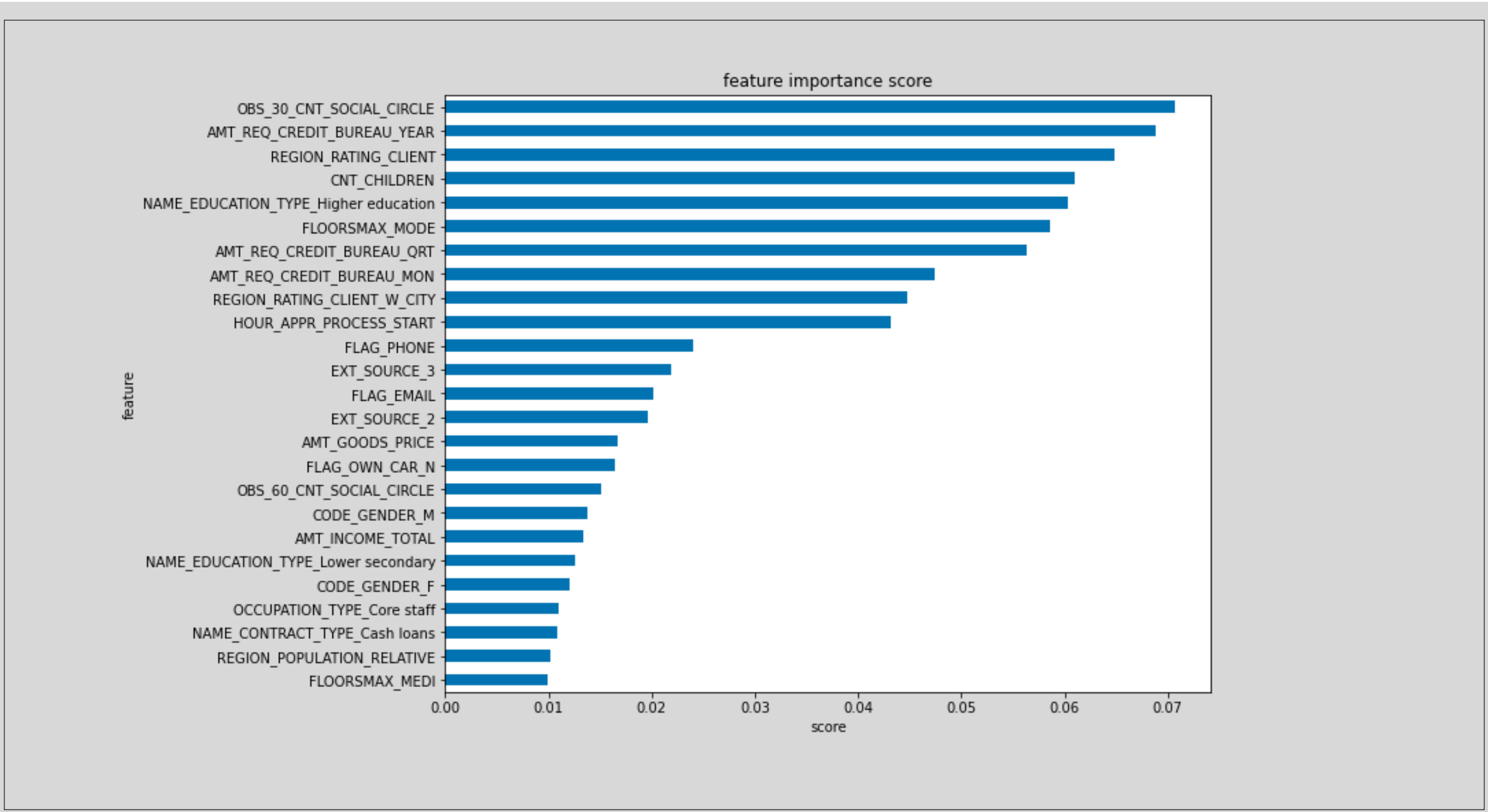
- Logistic Regression
- Random Forest
- XGBoost

Semua feature dimasukkan ke dalam model kecuali feature SK_ID_CURR & ORGANIZATION_TYPE

Model	Accuracy	Recall	AUC
Logistic Regression	0.8614	0.8064	0.86
Random Forest	0.9518	0.9159	0.95
XGBoost	0.9537	0.9112	0.95

Model	Training Score	Test Score
Logistic Regression	0.8604	0.8614
Random Forest	1.0000	0.9518
XGBoost	0.9557	0.9537

Pada hasil evaluasi (tabel kiri), model Random Forest memiliki nilai Accuracy, Recall, & AUC yang lebih tinggi daripada dua model lainnya. Namun, pada hasil evaluasi (tabel kanan) model Random Forest mengalami overfitting, yakni score data trainingnya lebih besar daripada score data testingnya. Sehingga model terbaik yang dipilih adalah model Logistic Regression dan XGBoost yang memiliki nilai Accuracy, Recall, & AUC cukup baik dan tidak terjadi overfitting.


feature importance score
OBS_30_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_YEAR
REGION_RATING_CLIENT
CNT_CHILDREN
NAME_EDUCATION_TYPE_Higher education
FLOORSMAX_MODE
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_MON
REGION_RATING_CLIENT_W_CITY
HOUR_APPR_PROCESS_START
FLAG_PHONE
EXT_SOURCE_3
FLAG_EMAIL
EXT_SOURCE_2
AMT_GOODS_PRICE
FLAG_OWN_CAR_N
OBS_60_CNT_SOCIAL_CIRCLE
CODE_GENDER_M
AMT_INCOME_TOTAL
NAME_EDUCATION_TYPE_Lower secondary
CODE_GENDER_F
OCCUPATION_TYPE_Core staff
NAME_CONTRACT_TYPE_Cash loans
REGION_POPULATION_RELATIVE
FLOORSMAX_MEDI
feature
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
score

Summary & Recommendation

- Sebesar 8.27% client mengalami kesulitan dalam pembayaran di Home Credit Indonesia.
- Student Client memiliki rate tertinggi sebesar 83.3% untuk client yang pinjaman sebelumnya diterima & tidak mengalami kesulitan pembayaran. Namun, total Student Client yang mengajukan pinjaman hanya 24 orang atau sekitar 0.0017% dari total client yang mengajukan pinjaman. Sehingga diperlukan promosi atau campaign kepada Student Client lainnya agar semakin banyak yang tertarik untuk mengajukan pinjaman di Home Credit Indonesia.
- Client dengan gelar akademik memiliki rate tertinggi sebesar 71,16% untuk client yang pinjaman sebelumnya diterima & tidak mengalami kesulitan pembayaran. Namun, total Client dengan gelar akademik yang mengajukan pinjaman hanya 579 orang atau sekitar 0,04% dari total client yang mengajukan pinjaman. Sehingga diperlukan promosi atau campaign kepada Client dengan gelar akademik lainnya agar semakin banyak yang tertarik untuk mengajukan pinjaman di Home Credit Indonesia.
- Model terbaik yang dipilih adalah model Logistic Regression dan XGBoost yang memiliki nilai Accuracy, Recall, & AUC cukup baik dan tidak terjadi overfitting.
- 5 feature important score tertinggi yang memengaruhi kesulitan client dalam melakukan pembayaran adalah OBS_30_CNT_SOCIAL_CIRCLE, AMT_REQ_CREDIT_BUREAU_YEAR, REGION_RATING_CLIENT, CNT_CHILDREN, & NAME_EDUCATION_TYPE_Higher education. Namun masih diperlukan uji hipotesis untuk membuktikannya.