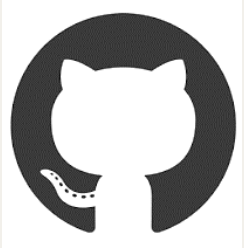
Perbandingan Logistic Regression dan K-Nearest Neighbor pada Klasifikasi Data Pelanggan Home Credit Indonesia dengan Optimasi Random Search CV

Ali Mahmudan

Project-Based Intern: Data Scientist Virtual Internship Experience
Home Credit Indonesia

OVERVIEW

- **Problem Research**
- **Data Pre-Processing**
- **Data Visualization and Business Insight**
- **Machine Learning Implementation and Evaluation**
- **Business Recommendation**



Github Repository
https://github.com/alimhdan/Home-Credit-Indonesia/blob/main/Machine%20Learning/Klasifikasi_Pelanggan.ipynb

PROBLEM RESERACH

Home Credit saat ini sedang menggunakan berbagai macam metode statistik dan Machine Learning untuk membuat prediksi skor kredit. Sekarang, kami meminta anda untuk membuka potensi maksimal dari data kami. Dengan melakukannya, kita dapat memastikan pelanggan yang mampu melakukan pelunasan tidak ditolak ketika melakukan pengajuan pinjaman, dan pinjaman dapat diberikan dengan principal, maturity, dan repayment calendar yang akan memotivsi pelanggan untuk sukses. Evaluasi akan dilakukan dengan mengecek seberapa dalam pemahaman analisa yang anda kerjakan. Sebagai catatan, anda perlu menggunakan setidaknya 2 model Machine Learning dimana salah satunya adalah Logistic Regression.

1. Cek data duplikat

```
[ ]  #data duplikat
     data_train.duplicated().sum()

     0

[ ]  data_test.duplicated().sum()

     0
```

2. Pemilihan variabel pemodelan

```
[4]  train_pakai=data_train[['TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER',
          'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN',
          'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE',
          'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE',
          'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE',
          'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH']]

     test_pakai=data_test[['NAME_CONTRACT_TYPE', 'CODE_GENDER',
          'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN',
          'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE',
          'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE',
          'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE',
          'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH']]
```

3. Penanganan missing value

```
[5]  train_pakai=train_pakai.dropna()
     test_pakai=test_pakai.dropna()
```

4. Cek word spelling

```
print(train_pakai['TARGET'].unique(),'\n',
    train_pakai['NAME_CONTRACT_TYPE'].unique(),'\n',
      train_pakai['CODE_GENDER'].unique(),'\n',
      train_pakai['FLAG_OWN_CAR'].unique(),'\n',
      train_pakai['FLAG_OWN_REALTY'].unique(),'\n',
      train_pakai['NAME_TYPE_SUITE'].unique(),'\n',
      train_pakai['NAME_INCOME_TYPE'].unique(),'\n',
      train_pakai['NAME_EDUCATION_TYPE'].unique(),'\n',
      train_pakai['NAME_FAMILY_STATUS'].unique(),'\n',
      train_pakai['NAME_HOUSING_TYPE'].unique())
```

5. Penambahan variabel

```
[10] AGE_TR=(train_pakai['DAYS_BIRTH']/-365).astype(int)
     AGE_TS=(test_pakai['DAYS_BIRTH']/-365).astype(int)
```

DATA PRE-PROCESSING

Berikut adalah tahapan data pre-processing yang dilakukan:

1. **Cek data duplikat**
2. **Pemilihan variabel pemodelan**
3. **Cek dan penanganan missing value**
4. **Cek word spelling**
5. **Penambahan variabel**

DATASET

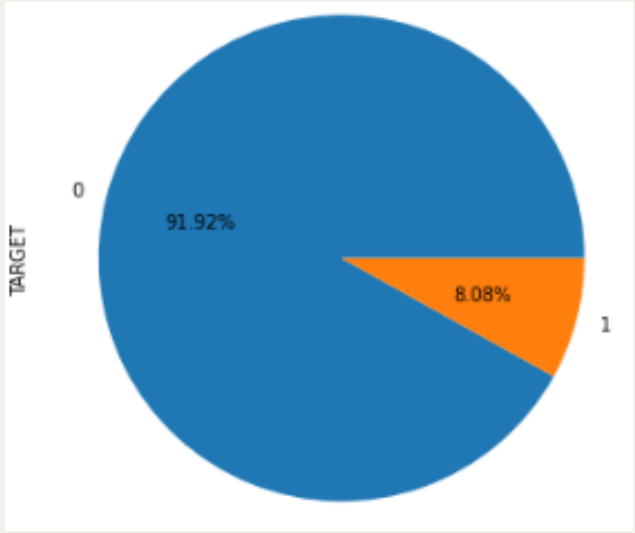
```
[11] train_pakai=train_pakai.assign(AGE=AGE_TR).drop('DAYS_BIRTH',axis=1)
     train_pakai.head()
```

	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	REGION_POPULATION_RELATIVE	AGE
0	1	Cash loans	M	N	Y	0	202500.0	406597.5	24700.5	351000.0	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	0.018801	25
1	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	1129500.0	Family	State servant	Higher education	Married	House / apartment	0.003541	45
2	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0	135000.0	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	0.010032	52
3	0	Cash loans	F	N	Y	0	135000.0	312682.5	29686.5	297000.0	Unaccompanied	Working	Secondary / secondary special	Civil marriage	House / apartment	0.008019	52
4	0	Cash loans	M	N	Y	0	121500.0	513000.0	21865.5	513000.0	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	0.028663	54

```
[12] test_pakai=test_pakai.assign(AGE=AGE_TS).drop('DAYS_BIRTH',axis=1)
     test_pakai.head()
```

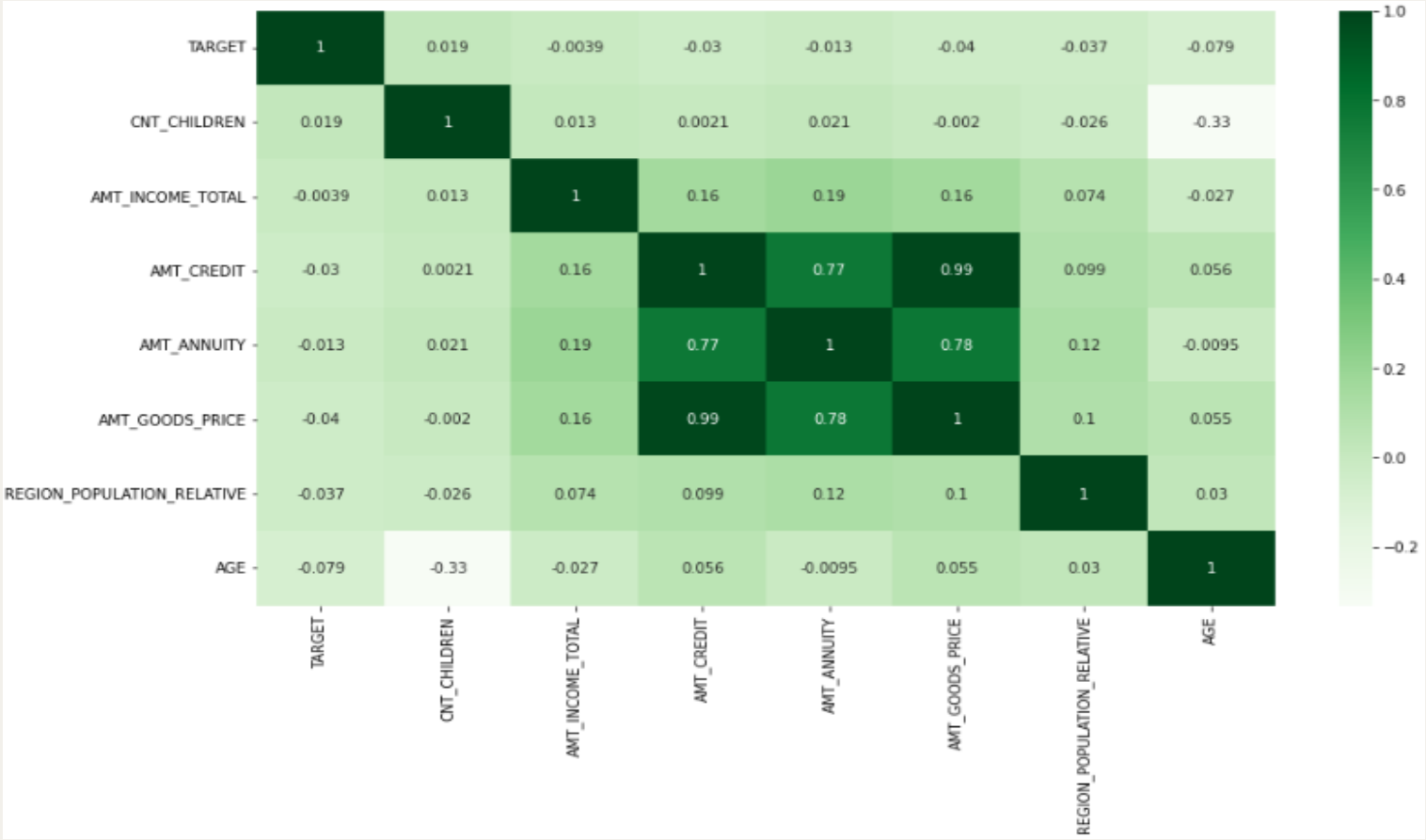
	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	REGION_POPULATION_RELATIVE	AGE
0	Cash loans	F	N	Y	0	135000.0	568800.0	20560.5	450000.0	Unaccompanied	Working	Higher education	Married	House / apartment	0.018850	52
1	Cash loans	M	N	Y	0	99000.0	222768.0	17370.0	180000.0	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	0.035792	49
3	Cash loans	F	N	Y	2	315000.0	1575000.0	49018.5	1575000.0	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	0.026392	38
4	Cash loans	M	Y	N	1	180000.0	625500.0	32067.0	625500.0	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	0.010032	35
5	Cash loans	F	Y	Y	0	270000.0	959688.0	34600.5	810000.0	Unaccompanied	State servant	Secondary / secondary special	Married	House / apartment	0.025164	50

DATA VISUALIZATION AND BUSINESS INSIGHT



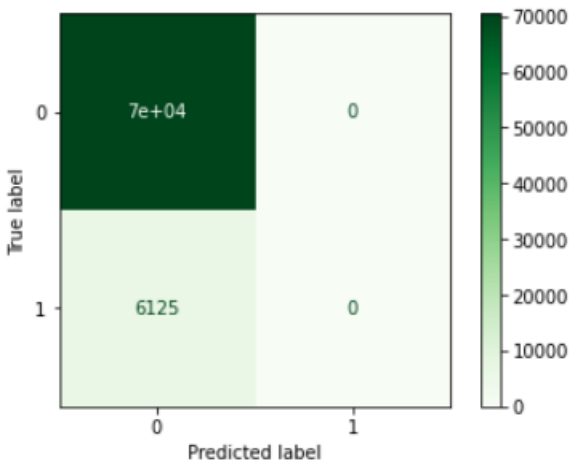
Berdasarkan visualisasi di atas, dapat dilihat bahwa jumlah TARGET dengan kategori 0 lebih banyak (281445) dibanding kategori 0 (24753) dengan persentase masing-masing adalah 91.92% dan 8.08%. Hal tersebut mengartikan bahwa pelanggan yang tidak memiliki kesulitan pembayaran jumlahnya lebih banyak (TARGET = 0).

Matrix korelasi di atas menunjukkan bahwa variabel CNT_CHILDREN, ANT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, REGION_POPULATION_RELATIVE, dan AGE memiliki nilai korelasi yang rendah terhadap variabel TARGET. Hal ini bisa dilihat pada nilai absolute koefisien korelasi < 0.8.



MACHINE LEARNING IMPLEMENTATION AND EVALUATION

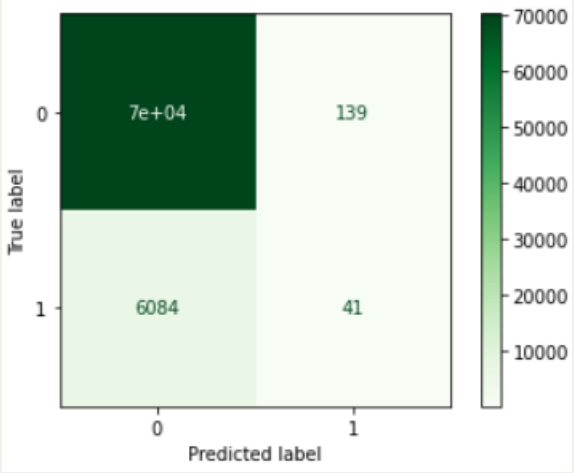
```
Logistic Regression
Best Score: 0.9188845538563705
Best Parameters: {'solver': 'liblinear', 'penalty': 'l1', 'max_iter': 2000, 'C': 0.0001}
```




```
print('Accuracy:',accuracy_score(y_test,y_prediksi_log))
print('Precision:',precision_score(y_test,y_prediksi_log,average='macro'))
print('Recall:',recall_score(y_test,y_prediksi_log,average='macro'))
print('f1-score:',f1_score(y_test,y_prediksi_log,average='macro'))
```

```
Accuracy: 0.9199869366427171
Precision: 0.45999346832135857
Recall: 0.5
f1-score: 0.47916312298009867
```

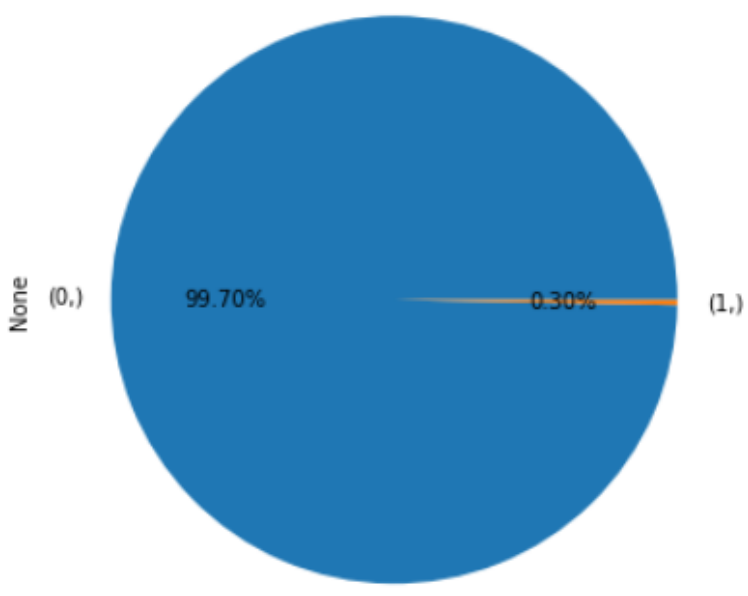
```
KNN
Best Score: 0.9174824104094215
Best Parameters: {'weights': 'uniform', 'n_neighbors': 9}
```




```
print('Accuracy:',accuracy_score(y_test,y_prediksi))
print('Precision:',precision_score(y_test,y_prediksi,average='macro'))
print('Recall:',recall_score(y_test,y_prediksi,average='macro'))
print('f1-score:',f1_score(y_test,y_prediksi,average='macro'))
```

```
Accuracy: 0.9187067276290006
Precision: 0.5740564939694179
Recall: 0.5023600733158013
f1-score: 0.4853065495454436
```

•Berdasarkan hasil klasifikasi metode KNN dengan parameter weights = 'uniform' dan n_neighbors = 9 di atas, dapat dilihat bahwa nilai akurasi data testing 0.9187067276290006 atau 91.87% sangat tinggi. Nilai tersebut hampir sama dengan best score data training yang sebesar 0.9174824104094215 atau 91.75% sehingga metode KNN tersebut layak digunakan. Sedangkan, klasifikasi Logistic Regression dengan parameter solver = 'liblinear', penalty = 'l1', max_iter = 2000, dan C = 0.0001 hasil prediksinya menunjukkan bahwa tidak ada yang diklasifikasikan ke dalam kategori TARGET=1 sehingga model regresi logistik kurang tepat jika digunakan untuk prediksi.




```
[30] y_prediksi_test.value_counts()

0    47665
1      144
dtype: int64
```

Klasifikasi metode KNN dengan parameter weights = 'uniform' dan n_neighbors = 9 menghasilkan kategori 0 (pelanggan yang tidak kesulitan membayar) sebanyak 47665 pelanggan. Sedangkan kategori 1 (pelanggan yang kesulitan membayar) sebanyak 144 pelanggan. Jika dinyatakan dalam persentase maka masing-masing persentasenya 99.70% dan 0.30%.

BUSINESS RECOMMENDATION

- Home Credit Indonesia harus memberikan perhatian khusus kepada pelanggan yang memilih pinjaman cash loans, sedang bekerja, sudah manikah, dan memiliki rumah atau apartment karena mereka adalah pelanggan dengan proporsi tertinggi yang tidak mengalami kesulitan pembayaran. Perhatian khusus bisa berupa keringanan batas waktu pembayaran, anuitas yang lebih kecil, atau peningkatan batas pinjaman.

- Disarankan untuk membentuk model klasifikasi dengan metode balancing dataset seperti SMOTE agar hasil prediksi semakin akurat. Hal tersebut dikarenakan jumlah pelanggan yang tidak mengalami kesulitan pembayaran lebih banyak dibandingkn yang mengalami kesulitan pembayaran sehingga dataset menjadi imbalance.

TERIMA KASIH