Rakamin Academy

Virtual Internship Experience

HOME CREDIT
Anda Bisa!

Created by:
Daniel Jeffry

Outline

I. Problem Research

II. Data Visualization and Business Insight

III. Data Pre-Processing

IV. Machine Learning Implementation and Evaluation

V. Business Recommendation

Link Github
ipynb pengerjaan tugas:
https://github.com/triyoza/home credit_project/blob/main/Final_t ask_VIX_HCI.ipynb

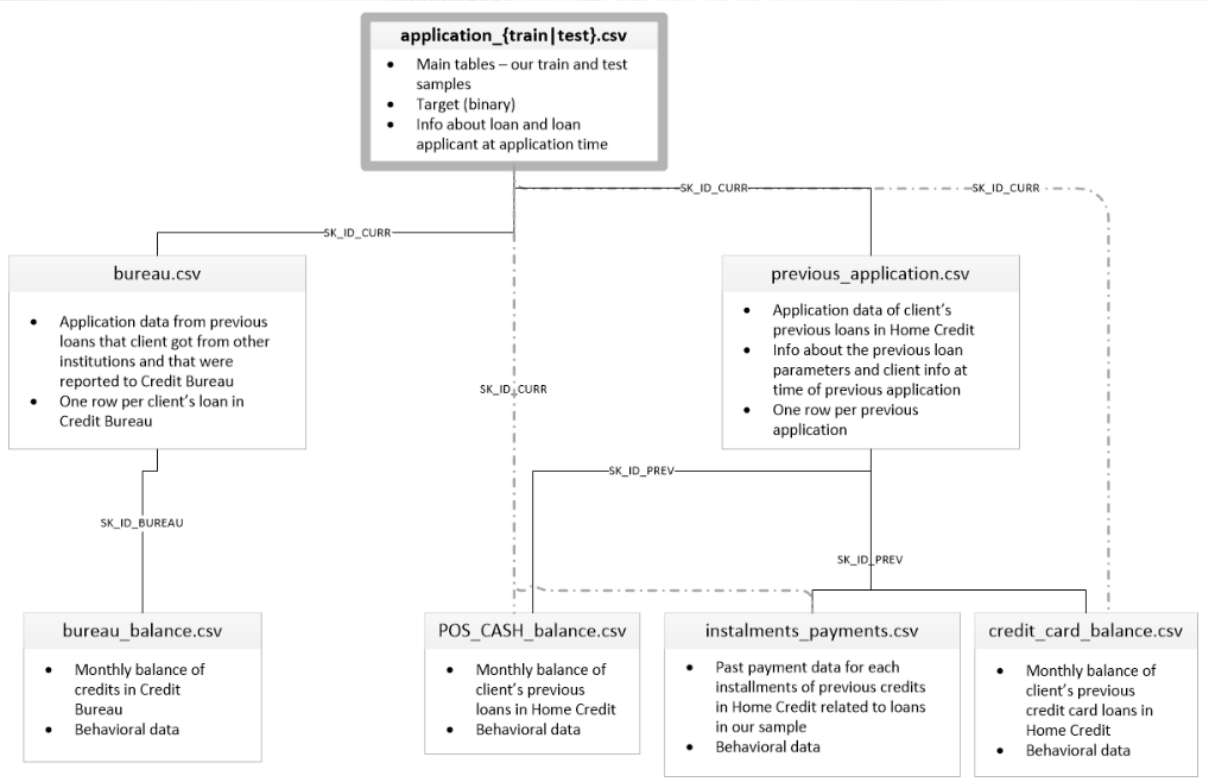
Oleh: Triyoza Aprianda

Problem Research

Pada tugas ini, kita akan membuat model klasifikasi untuk memprediksi mampu atau tidaknya pelanggan Home Credit Indonesia mengembalikan pinjaman yang sudah diajukan.

1. **Supervised**: Label terdapat dalam data train dan tujuannya adalah melatih model untuk mempelajari pola dan memprediksi label dari fitur, pada data yaitu kolom ‘TARGET’.
2. **Classification**: Kolom ‘TARGET’ terdiri dari variabel biner, 0: Tidak memiliki kesulitan pembayaran dan 1: Memiliki kesulitan pembayaran pinjaman.

Data description



Data yang digunakan yaitu: 'application_{train|test}.csv'

- This is the main table, broken into two files for Train (with target) and Test (without target)

- Static data for all applications. One row represents one loan in our data sample.

Exploratory Data Analysis (EDA)

Univariat Numerik

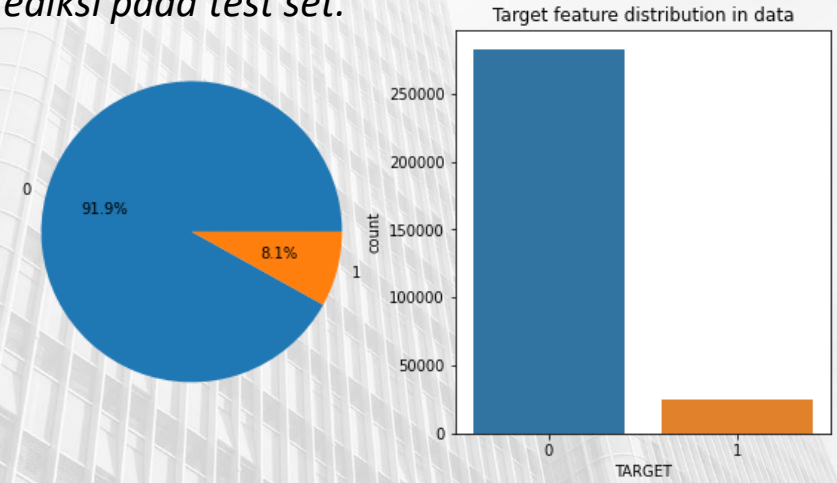
Dari semua fitur numerik dibuat histogram dan boxplot untuk melihat sebaran data

- Pada umumnya didominasi oleh fitur yang memiliki value 0, 1, 2 dan sebagainya yang mana angka-angka ini memiliki makna tersendiri seperti yes dan no, salah satunya yaitu pada fitur 'TARGET'
- Pada umumnya setiap fitur numerik memiliki outlier yang banyak, namun outlier ini tidak perlu dipermasalahkan dan tidak mempengaruhi model yang dibuat. Misalnya suatu fitur memiliki value 0 dan 1, hamper 90% nilainya 0 sedangkan selebinya 1, maka semua value 1 akan dianggap sebagai outlier. Hal ini tidak perlu dikhawatirkan karena masih masuk akal dan jika ditinjau dari setiap fitur wajar memiliki sebaran data seperti itu.

Exploratory Data Analysis (EDA)

Fitur Target

Fitur 'TARGET' merupakan fitur utama permasalahan yang menunjukkan pelanggan memiliki kesulitan pembayaran atau tidak. Fitur ini merupakan kolom label untuk pembuatan model klasifikasi dan akan diprediksi pada test set.

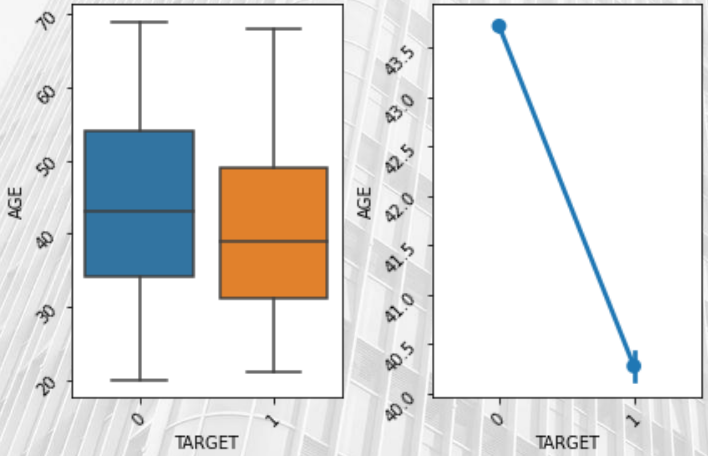


Berdasarkan pie chart dan countplot untuk fitur 'TARGET' dapat dilihat bahwa jumlah target dengan kategori 0 lebih banyak daripada kategori 1 dengan persentase 91.92% dan 8.08%. Artinya jauh lebih banyak pelanggan yang tidak memiliki kesulitan pembayaran daripada yang memiliki kesulitan pembayaran.

Exploratory Data Analysis (EDA)

Numerik - Target

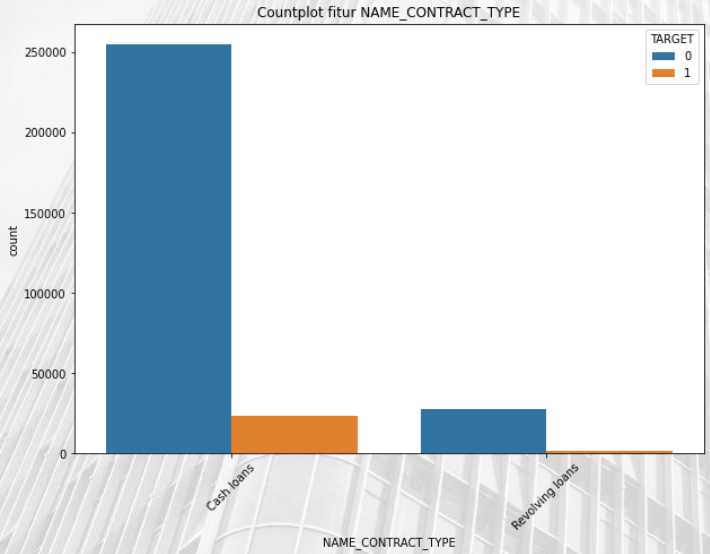
- Disini dibuat boxplot dan pointplot fitur numerik dengan sumbu x berupa ‘TARGET’
- Pada umunya di setiap fitur numerik memiliki rata-rata nilai yang lebih tinggi pada target 0 (tidak mimiliki kesulitan pembayaran.
- Seperti pada fitur ‘AGE’ berikut:

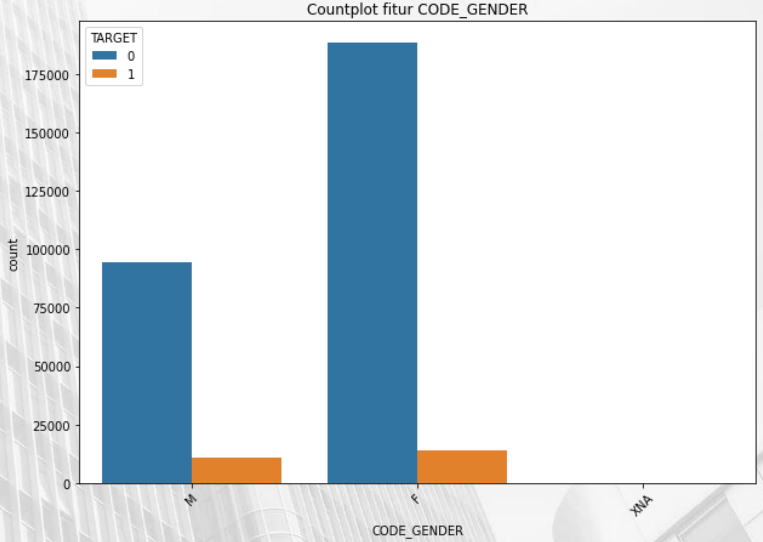


Berdasarkan boxplot dan pointplot fitur AGE dapat dilihat bahwa rata-rata umur dengan target 0 atau rata-rata umur yang tidak memiliki kesulitan pembayaran lebih tinggi daripada yang memiliki kesulitan pembayaran.

Kategorik - Target

- Sama halnya dengan Numerik-Target sebelumnya, disini dibuat countlplot untuk setiap fitur kategorik dengan hue berupa kolom ‘TARGET’
- Setiap plot menunjukkan bahwa untuk masing-masing fitur setiap kategorinya lebih banyak dengan label 0 daripada 1.





- Plot di atas memberikan informasi bahwa jumlah pelanggan dengan Name_Contract_Type kategori cash loan lebih banyak daripada revolving loan.
- Pelanggan didominasi oleh pelanggan yang berjenis kelamin perempuan

Data Preprocessing

Missing Value Handling

- *Terdapat beberapa fitur dengan missing value*
- *Handling yang dilakukan: Missing Value >= 10% dilakukan drop fitur, sedangkan yang kurang dari 10 % dilakukan simple imputer*
- *Simple Imputer (strategy = most frequent) untuk kategorik dan Simple Imputer (strategy = median) untuk fitur numerik*

Transformasi

- *Pada fitur numerik dilakukan scaling, yang dipakai minmax scaler*
- *Pada Fitur Kategorik dilakukan Encoding : One Hot Encoder*

Feature Selection

- *Dilakukan seleksi fitur dengan beberapa metode:*
- *Multicolinearty Reduction: disini didrop beberapa fitur*
- *Kemudian dilanjutkan dengan Mutual Information, disini fitur dengan nilai mutual information 0 (nol) didrop.*

Modeling

- *Hanya sempat membuat dua model yaitu: Logistic Regression dan Adaboost*
- *Logistic Regression menunjukkan performance yang lebih baik:*
- *Hyperparameter:*

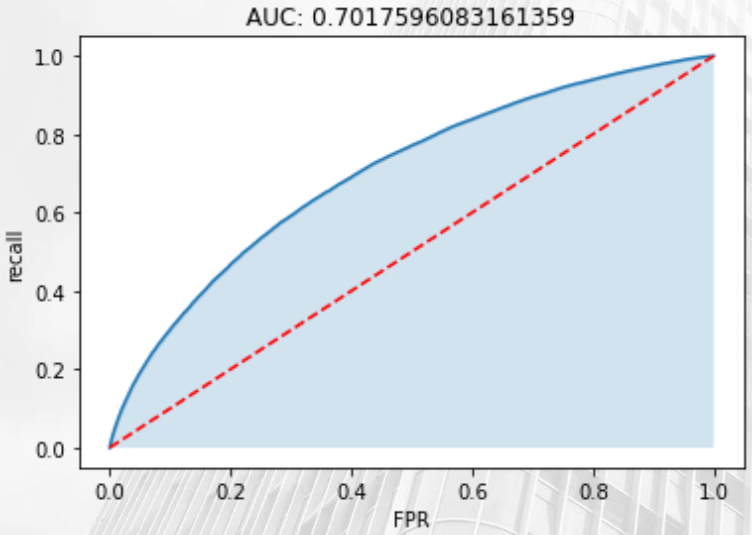
```
logreg_final = LogisticRegression(C=1, penalty = 'l2', random_state=42, solver='saga')
logreg_final.fit(X_train_sel2, y_train)
```

- *Performance:*

```
[62] classif_pr(X_train_sel2, y_train, logreg_final, thre_roc[25854] )
```

	precision	recall	f1-score	support
1	0.14	0.63	0.23	24825
0	0.95	0.67	0.79	282686
accuracy			0.66	307511
macro avg	0.55	0.65	0.51	307511
weighted avg	0.89	0.66	0.74	307511

AUC



Logistic Regression dengan AUC 70%

Evaluasi

Evaluasi

Memprediksi kolom target pada test set menggunakan model final yang dibuat: Logistic Regression

```
[71] y_pred = logreg_final.predict_proba(X_test_sel)[:,1] > thre_roc[25854]
```

```
[72] y_pred
```

```
array([False,  True, False, ..., False,  True, False])
```

```
[73] y_pred_int = y_pred.astype(int)
```

```
[74] target_test = pd.DataFrame(y_pred_int, columns= ['label'])
     target_test.head()
```

	label
0	0
1	1
2	0
3	0
4	1

HOME CREDIT
Anda Bisa!

THANKS

Link Github
ipynb pengerjaan tugas:
https://github.com/triyoza/homecredit_project/blob/main/Final_task_VIX_HCI.ipynb