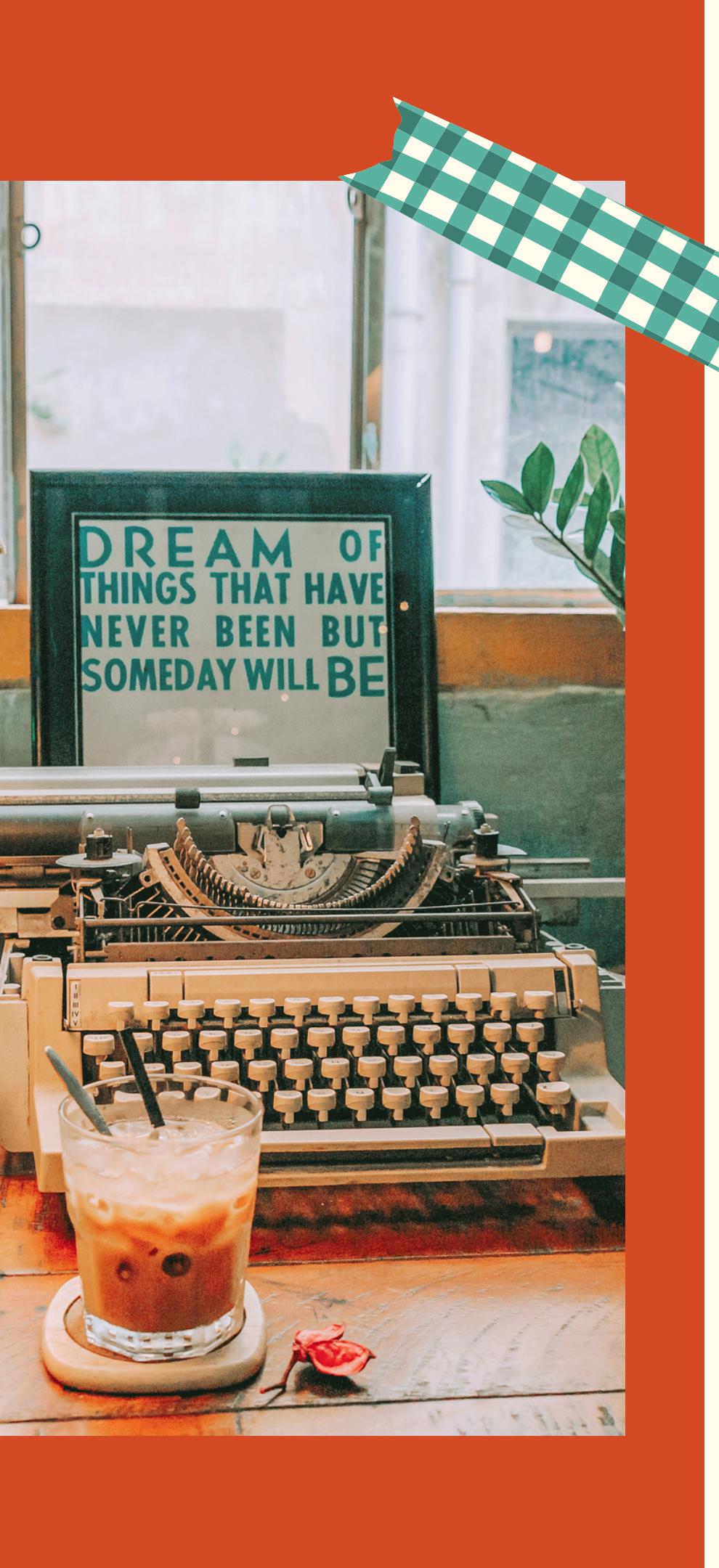


Home Credit Indonesia Data Scientist Project Based Internship Program

[Final Task] Home Credit
Scorecard Model

OLEH
Arif Putera Wijaya

Problem



Task : Menggunakan berbagai macam **metode statistik dan Machine Learning** untuk membuat **prediksi skor kredit**, 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 memotivasi pelanggan untuk sukses. menggunakan setidaknya 2 model Machine Learning dimana **salah satunya** adalah **Logistic Regression**

◆ Data : application_test.csv

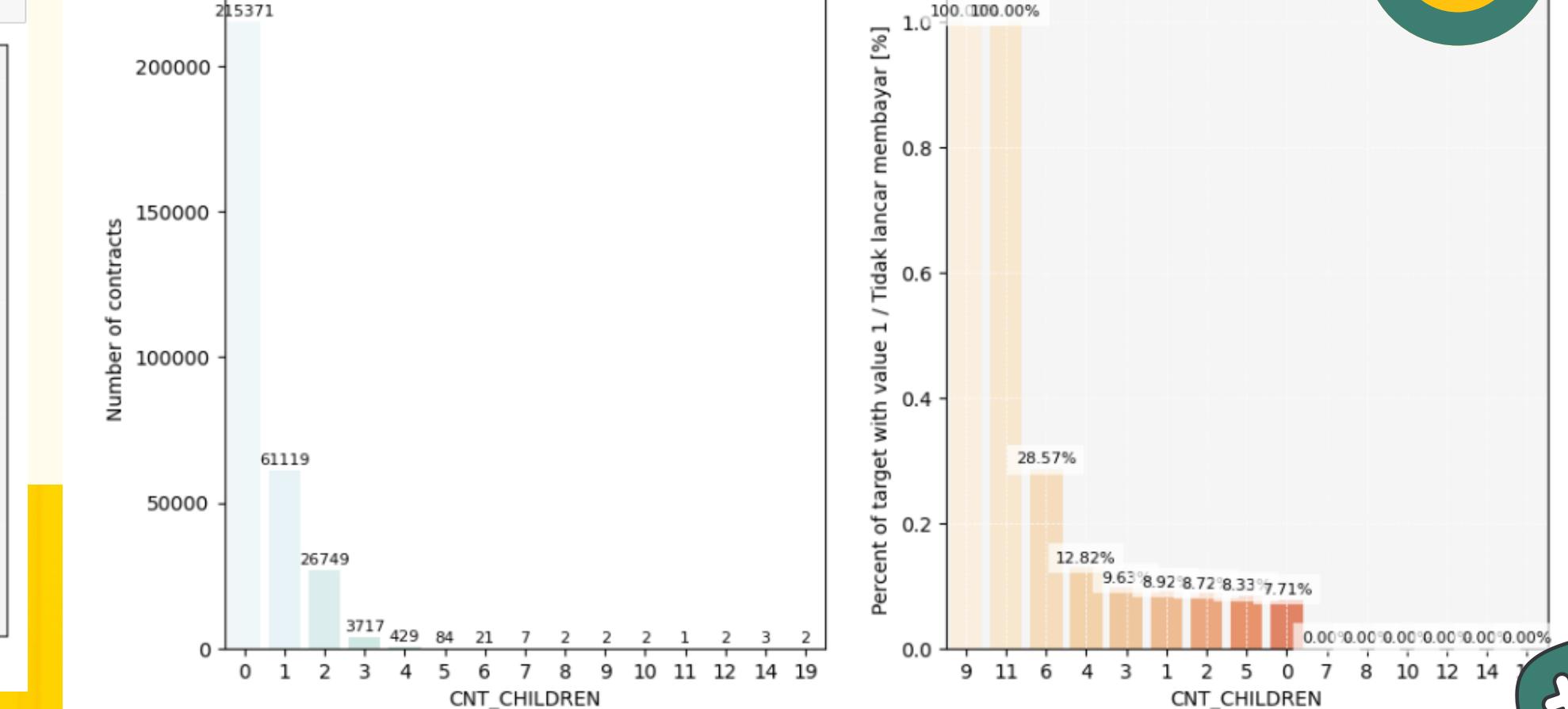
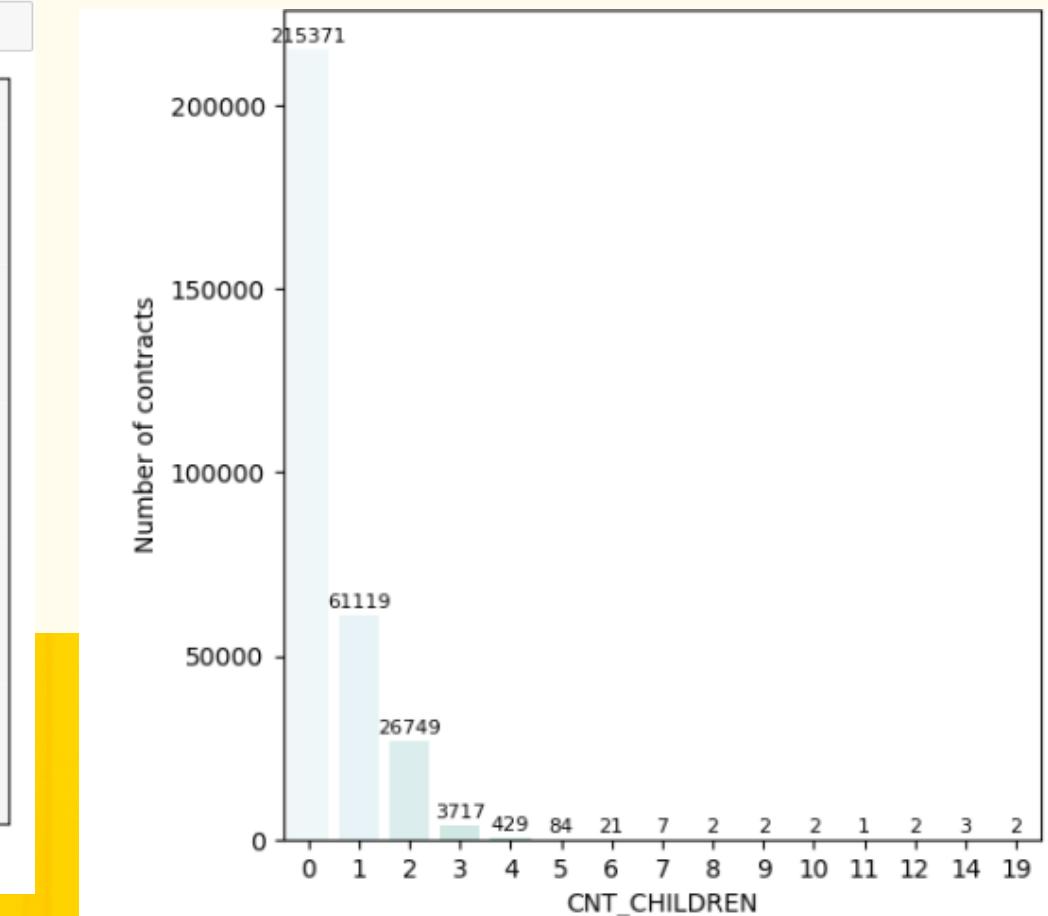
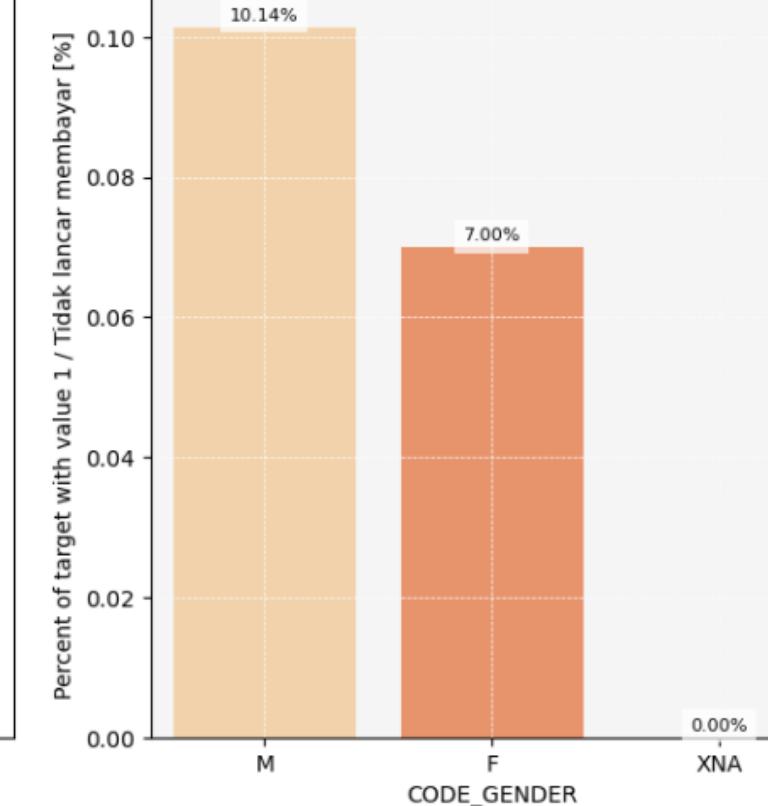
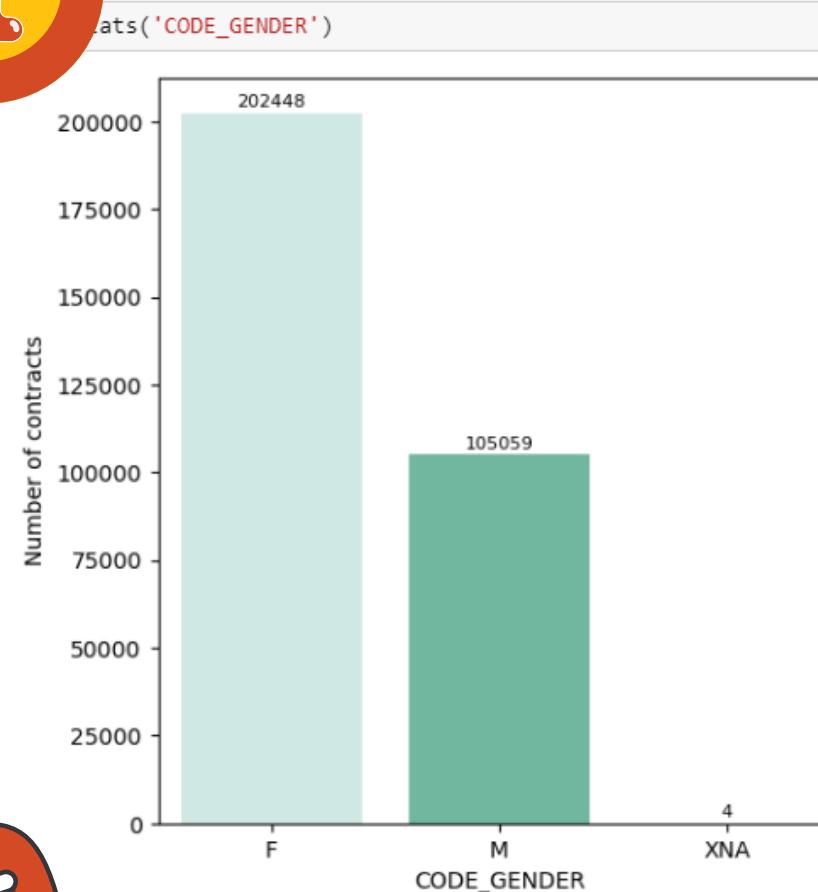
Untuk memenuhi tugas diatas, saya menggunakan Data application_test yang mana berisi data pengajuan kredit. serta dapat terlihat kelancaran pembayaran pada var TARGET



WiiCode101

TOP INSIGHT

1



Terlihat bahwa sebagian besar pinjaman dilakukan oleh wanita yang mana hampir 2x lebih besar dibandingkan laki-laki, disamping itu persentase gagal bayar/kredit macet lebih banyak dilakukan oleh client wanita jika dilihat dari frekuensinya

Sebagian besar klien yang mengajukan pinjaman tidak memiliki anak, Semangkin banyak anak memiliki potensi gagal bayar yang semangkin besar, dari grafik bahkan klien yang memiliki 9 atau 11 anak, persentase pinjaman yang tidak dilunasi adalah 100%.



Pre-processing



Import Library

Data Cleaning

(Missing Value - Column Dropping - Feature Engineering - Save Clean Data)

Outlier Fraction

Split Data

Feature Selection

MODEL

```
['SK_ID_CURR',
'FLAG_OWN_REALTY',
'CNT_CHILDREN',
'AMT_INCOME_TOTAL',
'AMT_CREDIT',
'AMT_ANNUITY',
'REGION_POPULATION_RELATIVE',
'DAYS_BIRTH',
'DAYS_EMPLOYED',
'DAYS_REGISTRATION',
'DAYS_ID_PUBLISH',
'OWN_CAR_AGE',
'CNT_FAM_MEMBERS',
'REGION_RATING_CLIENT',
'HOUR_APPR_PROCESS_START',
'EXT_SOURCE_1',
'EXT_SOURCE_2',
'EXT_SOURCE_3',
'APARTMENTS_AVG',
'BASEMENTAREA_AVG',
'YEARS_BEGINEXPLUATATION_AVG',
'YEARS_BUILD_AVG',
'COMMONAREA_AVG',
'ENTRANCES_AVG',
'FLOORSMAX_AVG',
'FLOORSMIN_AVG',
'LANDAREA_AVG',
'NONLIVINGAREA_AVG',
'OBS_30_CNT_SOCIAL_CIRCLE',
'DEF_30_CNT_SOCIAL_CIRCLE',
'DAYS_LAST_PHONE_CHANGE',
'AMT_REQ_CREDIT_BUREAU_MON',
'AMT_REQ_CREDIT_BUREAU_QRT',
'AMT_REQ_CREDIT_BUREAU_YEAR',
'NAME_FAMILY_STATUS_Married',
'CREDIT_INCOME_PERCENT',
'ANNUITY_INCOME_PERCENT',
'CREDIT_TERM',
'DAYS_EMPLOYED_PERCENT']
```

Feature Selection

```
In [6]: from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_selection import SelectFromModel
```

```
In [7]: sel = SelectFromModel(RandomForestClassifier(n_estimators = 100, random_state=state))
sel.fit(X_train, y_train)
```

```
Out[7]:
```

```
  SelectFromModel
    > estimator: RandomForestClassifier
      > RandomForestClassifier
```

Untuk memaksimalkan akurasi model, dilakukan **feature selection** dengan bantuan **Random Forest Classifier**. Sehingga pada akhirnya didapat **39 Feature**

MODEL

Extreme Gradient Boost errors: 25491

Accuracy Score :
0.7236867778090923

Confusion matrix :
[[62014 22815]
 [2676 4749]]

	precision	recall	f1-score	support
0	0.96	0.73	0.83	84829
1	0.17	0.64	0.27	7425
accuracy			0.72	92254
macro avg	0.57	0.69	0.55	92254
weighted avg	0.90	0.72	0.78	92254

ROC AUC score is: 0.6853215625350154

Jumlah Kesalahan (Errors):

Model XGBoost menghasilkan lebih sedikit kesalahan dibandingkan dengan model Regresi Logistik sebelumnya, yang mengalami 25,491 kesalahan klasifikasi. Akurasi (Accuracy):

Akurasi model XGBoost

adalah sekitar 72.37%, meningkat dari akurasi sebelumnya dengan Regresi Logistik. Confusion Matrix:

Confusion matrix

menunjukkan bahwa model cenderung membuat kesalahan dalam mengklasifikasikan kelas 1 (fraud), yang dapat dilihat dari False Negatives (2676). Namun, jumlah False Positives (22815) juga cukup tinggi. Classification Report:

Precision

untuk kelas 1 (fraud) masih rendah (17%), menunjukkan bahwa dari prediksi positif yang dibuat oleh model, hanya sebagian kecil yang benar-benar merupakan kasus fraud.

Recall

untuk kelas 1 meningkat menjadi 64%, yang menunjukkan bahwa model dapat mengidentifikasi sebagian besar kasus fraud yang sebenarnya.

F1-score

untuk kelas 1 juga meningkat, tetapi masih tergolong rendah, menunjukkan bahwa model belum sempurna dalam menggabungkan precision dan recall.

Meskipun akurasi meningkat, kita perlu memperhatikan trade-off antara precision dan recall.

ROC AUC Score:

ROC AUC score sekitar 0.685 menunjukkan kemampuan model untuk membedakan antara kelas positif dan negatif. Skor ini menunjukkan peningkatan dari model sebelumnya, tetapi masih perlu diperhatikan untuk meningkatkannya lebih lanjut.

Analisis Keseluruhan:

Model XGBoost memberikan peningkatan yang signifikan dibandingkan dengan Regresi Logistik, terutama dalam hal akurasi dan kemampuan membedakan kelas positif dan negatif.

Target Audience Setelah berexperiment dengan berbagai Model (Random Forest, Logistic Regression, XGBoost) diputuskan untuk menggunakan Extreme Gradient Boost. Kedepannya model dapat di implementasi dalam aplikasi.

Business Recommendation



Aplikasi Optimasi Proses Persetujuan Pinjaman

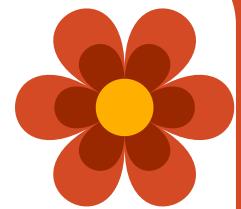
Menggunakan model prediksi risiko kredit dapat membantu perusahaan dalam mengoptimalkan proses persetujuan pinjaman



Pengembangan Program Edukasi Keuangan

Pentingnya pendekatan dengan calon client berkualitas, untuk itu dapat mempertimbangkan edukasi, dengan data data yang ada pada EDA kita dapat mengedukasi client dalam management keuangan

Segmentasi Pelanggan



Gunakan skor risiko kredit untuk membuat segmentasi pelanggan, seperti yang terlihat pada EDA



Thank you!



Wiicode101

