client payment prediction with
CLASSIFICATION MODELING

data science portfolio

Lutfia Husna Khoirunnisa

WHAT DOES THIS PROJECT DO?

data and problem understanding

HOME CREDIT

PROBLEMS:

PT. HOME CREDIT merupakan suatu perusahaan perusahaan pembiayaan multiguna multinasional.yang memberikan layanan pembiayaan bagi pelanggan. Terdapat banyak pengajuan credit yang masuk tetapi akan sangat menghabiskan banyak waktu jika harus dilakukan review pengajuan satu per satu secara manual.

SOLUTION:

Dibentuk suatu model yang dapat melakukan prediksi *client payment* secara otomatis yang dapat digunakan untuk mengambil keputusan penerimaan credit.

BUSINESS METRIC:

Banyak pengajuan credit yang dapat direview per harinya.

CLIENT PAYMENT PREDICTION with CLASSIFICATION MODEL

KLASIFIKASI merupakan salah satu metode dari *supervised learning*, yang dapat diartikan sebagai suatu algortima atau teknik yang dapat digunakan untuk membuat suatu skema atau kategori data yang berlabel.

Role 'model klasifikasi' dalam *project* ini adalah **melakukan prediksi *client payment*** dengan memberikan label 1 untuk ***client* yang dianggap akan memiliki kesulitan pembayaran**, dan 0 untuk ***client* yang dianggap tidak memiliki kesulitan pembayaran** untuk tiap pengajuan credit dengan melihat profil pengaju tersebut.

STEPS

01

Data Understanding

Melihat data untuk mendapatkan pemahaman awal dan menganalisis data apa saja yang dibutuhkan untuk memecahkan masalah

02

Data Preparation

Melakukan *data cleaning* dan splitting data

03

Modelling&Evaluation

Melakukan training model dan testing model, dan mengevaluasi metode mana yang paling tepat untuk digunakan

04

Model Implementation

Mengimplementasikan model yang telah dibuat dalam kegiatan perusahaan

NOTE:

Analisis dan pemodelan pada project ini menggunakan bahasa pemrograman python, di mana data dan coding yang digunakan dapat diakses di sini.

DATA UNDERSTANDING

Data yang digunakan merupakan data sample pengajuan credit untuk PT. HOME CREDIT yang dibagi menjadi dua dataset.

```
#Data Understanding
datamodel.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

Dataset pertama yang disimpan dalam variable datamodel memiliki 307511 baris data dan 122 kolom.

Data ini selanjutnya akan dibagi menjadi dua sebagai data train yang digunakan dalam proses *learning of mapping*, dan data testing untuk keperluan evaluasi model.

Dataset kedua yang disimpann dalam variable newapplication memiliki 48744 baris data dan 121 kolom. Data ini selanjutnya yang akan diprediksi label targetnya.

```
newapplication.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48744 entries, 0 to 48743
Columns: 121 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(40), object(16)
memory usage: 45.0+ MB
```

Perbedaan dataset pertama dan kedua adalah dataset pertama memiliki kolom label target yang akan digunakan dalam proses modeling, sedangkan dataset kedua tidak memiliki kolom tersebut.

DATA PREPARATION

1. DATA CLEANSING

Kedua dataset memiliki missing value yang harus diatasi, dimana missing value pada kolom OWN_CAR_AGE diisi dengan angka 0, dan baris data yang memiliki missing data lain akan dihapus. Kolom yang tidak digunakan dalam proses modeling juga dihapus.

2. DATA SPLITING

Data splitting ini dilakukan kepada dataset datamodel menjadi data train dan data test yang digunakan dalam proses modeling.

3. DATA TRAIN CLASS IMBALANCE HANDLING

Karena class 0 memiliki lebih banyak record di data train dibandingkan class 1, maka perlu untuk dilakukan oversampling data dengan metode SMOTE untuk menangani masalah class imbalance ini.

4. DATA NORMALIZATION

Data perlu untuk dinormalisasikan agar data yang digunakan tidak memiliki penyimpangan yang terlalu besar

DATA MODELING & EVALUATION

(with Logistic Regression, Random Forest, and Suport Vector Machine)

Logistic Regression model report

```
confusion matrix:
 [[4003 1635]
 [ 140  246]]

Model report:
              precision    recall  f1-score   support

           0       0.97      0.71      0.82      5638
           1       0.13      0.64      0.22       386

    accuracy                           0.71      6024
   macro avg       0.55      0.67      0.52      6024
weighted avg       0.91      0.71      0.78      6024
```

Model ini mendapat nilai akurasi sebesar **71%**

Suport Vector Machine model report

```
confusion matrix:
 [[4941  697]
 [ 266  120]]

Model report:
              precision    recall  f1-score   support

           0       0.95      0.88      0.91      5638
           1       0.15      0.31      0.20       386

    accuracy                           0.84      6024
   macro avg       0.55      0.59      0.56      6024
weighted avg       0.90      0.84      0.87      6024
```

Model ini mendapat nilai akurasi sebesar **84%**

Hasil evaluasi:
RANDOM FOREST MODEL mendapatkan nilai akurasi terbesar di antara model lain

Random Forest model report

```
confusion matrix:
 [[5625   13]
 [ 383    3]]

Model report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97      5638
           1       0.19      0.01      0.01       386

    accuracy                           0.93      6024
   macro avg       0.56      0.50      0.49      6024
weighted avg       0.89      0.93      0.91      6024
```

Model ini mendapat nilai akurasi sebesar **93%**

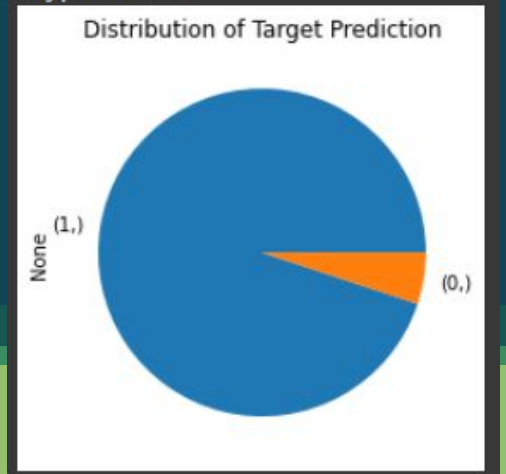
MODEL IMPLEMENTATION

Dalam modeling dan evaluasi yang telah dilakukan, diketahui bahwa model klasifikasi dengan nilai akurasi tertinggi adalah model *random forest*. Selanjutnya, model yang telah dibangun tersebut akan digunakan untuk memprediksi label target pada *credit* baru.

Data pengaju *credit* baru yang digunakan adalah dataset newapplication. Sebelum dilakukan prediksi, *missing value* dari data perlu untuk ditangani, dibersihkan dari kolom yang tidak digunakan, dan dinormalkan distribusinya.

Pada dataset tersebut, diprediksi terdapat 4493 data dengan label 1 (*client* yang dianggap akan memiliki kesulitan pembayaran), dan 240 data dengan label 0 (*client* yang dianggap tidak akan memiliki kesulitan pembayaran).

```
Target Label counts:
 1    4493
0      240
dtype: int64
```



CONCLUSION

(Business Recommendation)

Banyak pengajuan *credit* yang direview per-harinya dapat ditingkatkan dengan menerapkan prediksi *client payment* dengan model yang telah dibuat, dimana pada proses modeling dan evaluation yang telah dilakukan, diketahui bahwa model *random forest* memiliki nilai akurasi tertinggi jika dibandingkan dengan *logistic regression,* dan *support vector machine.*

THANK YOU!

CHECK OUT MY OTHER PORTFOLIOS:

https://lynk.id/lutfiahusnak

www.linkedin.com/in/lutfiahusnakhoirunnisa

lutfiahusnakhoirunnisa@gmail.com