I

FINAL PROJECT

Kelompok 22

15 Desember 2022

MSIB BATCH 3 –ZENIUS DATA ANALYTICS

Member of Group 22

PUTRI AGUSTINA RIADI

INDIRA MEUTIA KHAIRUNNISA

ANNISA SYIFA SUGARYADI

MADE BAIHAQI AJI KUMUDA

KEVIN AVICENNA WIDIARTO

SUMMARY

- BUSINESS UNDERSTANDING
- DATA UNDERSTANDING
- DATA PREPARATION
- MODELLING
- EVALUATION
- DEPLOYMENT

RT

SENTRA ZAMRUD | RAPAT TERBUKA

Business Understanding

- Business Objectiives
- Situation Assesment
- Data Mining Goals
- Project Plan

V

Business Objectives

BU

Background

Kredit merupakan fasilitas keuangan yang memungkinkan pelanggan meminjam uang dan membayarnya kembali dalam jangka waktu yang ditentukan. Homekredit merupakan perusahaan pembiayaan yang menyediakan pinjaman bagi pelanggan untuk berbelanja kebutuhan baik itu online maupun offline. Pada dasarnya, ditentukan beberapa kriteria untuk dapat menilai kelayakan peminjam. Dalam penentuannya, digunakan pendekatan statistik dan metode machine learning untuk menilai kemampuan calon peminjam.

XII

Business Objectives

BU

Objectives

Dapat mengelompokan dua jenis peminjam, yaitu peminjam yang layak dan tidak layak diberikan pinjaman berdasarkan riwayat perilaku keuangan calon peminjam. Analisis dilakukan untuk meminimalisir penyetujuan kredit yang tidak tepat sasaran.

XII

Business Objectives

- Criteria

BU

Adapun kriteria kesuksesan bisnis dari kelompok kami yaitu dapat meningkatkan kepercayaan customer terhadap bisnis kami, dapat memberikan customer service yang baik, dapat mendapatkan metode dengan nilai akurasi terbaik untuk menentukan customer yang layak menggunakan pinjaman dari home credit.

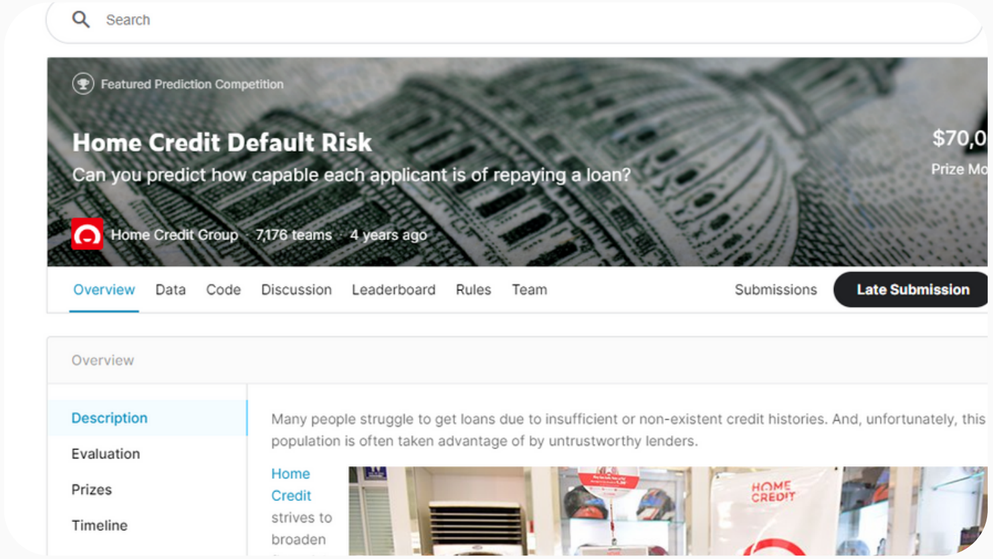
XII

Business Objectives

BU

- Situation assesment

Data yang digunakan untuk analisis adalah data riwayat perilaku calon peminjam yang diakses pada laman website Kaggle.com, yaitu https://www.kaggle.com/competitions/home-credit-default-risk/overview.



XII

Business Objectives

- data mining goals

BU

Mengelompokan jenis customer menjadi defaulted customer (peminjam tidak layak) dan not defaulted customer (peminjam layak). Selain itu, tujuan dari analisis ini adalah untuk dapat membentuk model terbaik dengan nilai akurasi tertinggi dan kecenderungan eror terendah. Model digunakan untuk menentukan kelayakan customer berdasarkan kriteria yang telah ditentukan. Serta dengan model ini harapan kami dapat menumbuhkan kembali kepercayaan masyarakat kepada bisnis home kredit.

XII

Business Objectives

BU

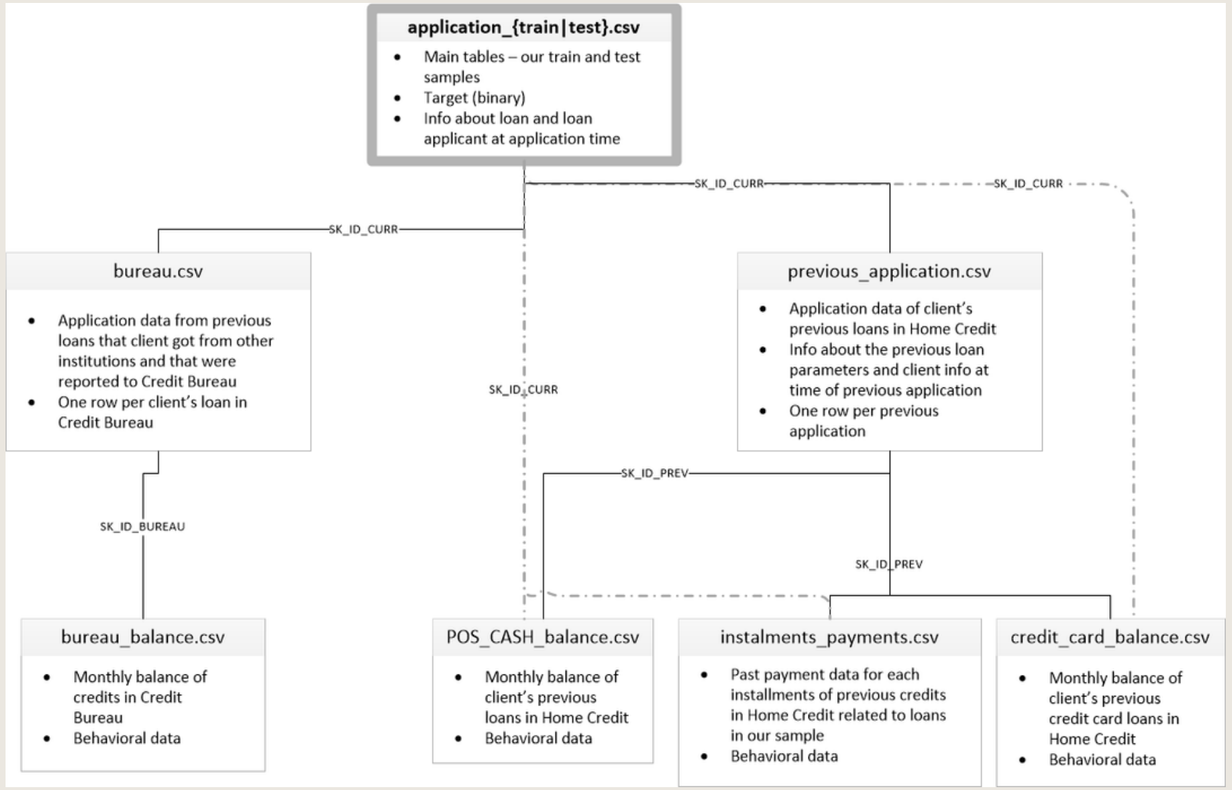
Project Plan Timeline

VII

DATA UNDERSTANDING

RT

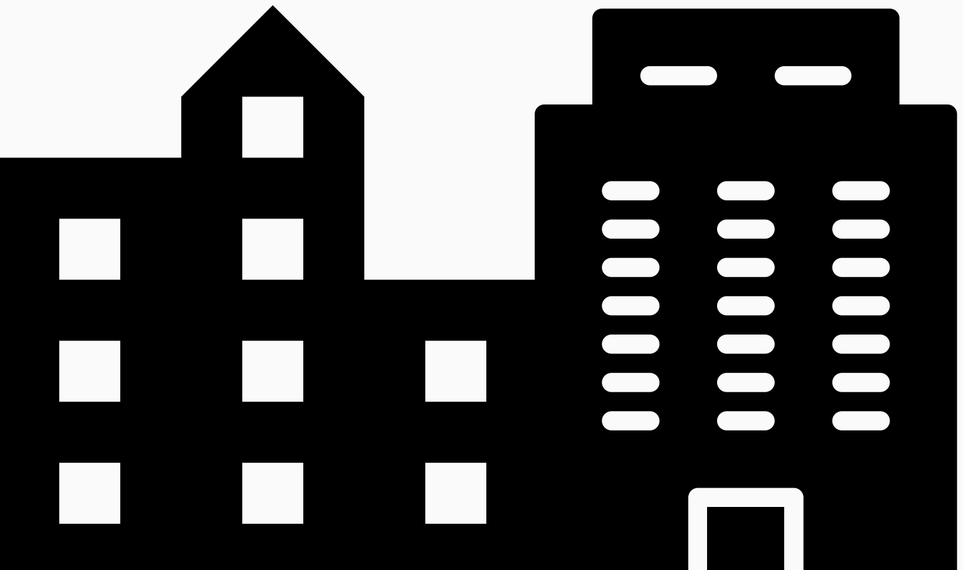
Data



Data disini disediakan oleh Home Credit, merupakan sebuah layanan yang didedikasikan untuk menyediakan jalur kredit (pinjaman) kepada populasi yang tidak memiliki rekening bank. Memprediksi apakah klien akan melunasi pinjaman atau mengalami kesulitan adalah kebutuhan bisnis yang penting, dan Home Credit mengadakan kompetisi ini di Kaggle untuk melihat model seperti apa yang dapat dikembangkan oleh komunitas pembelajaran mesin untuk membantu mereka dalam tugas ini.

VI

Collect Initial Data



Adapun data yang digunakan oleh kelompok 22 adalah

Application_train & Application_test

Data pelatihan dan pengujian utama dengan informasi tentang setiap aplikasi pinjaman di Home Credit. ·Data aplikasi pelatihan dilengkapi dengan TARGET yang menunjukkan 0: pinjaman telah dilunasi atau 1: pinjaman tidak dilunasi.

Bureau

Berisi data mengenai kredit klien sebelumnya dari lembaga keuangan lain. Setiap kredit sebelumnya memiliki barisnya tersendiri di biro, tetapi satu pinjaman dalam data aplikasi dapat memiliki beberapa kredit sebelumnya.

Previous Application

Berisi data pengajuan pinjaman sebelumnya di Home Credit oleh klien yang memiliki pinjaman dalam data aplikasi. Setiap pinjaman yang ada saat ini di aplikasi data dapat memiliki beberapa pinjaman sebelumnya.

Describe Data

APPLICATION TRAIN & TEST

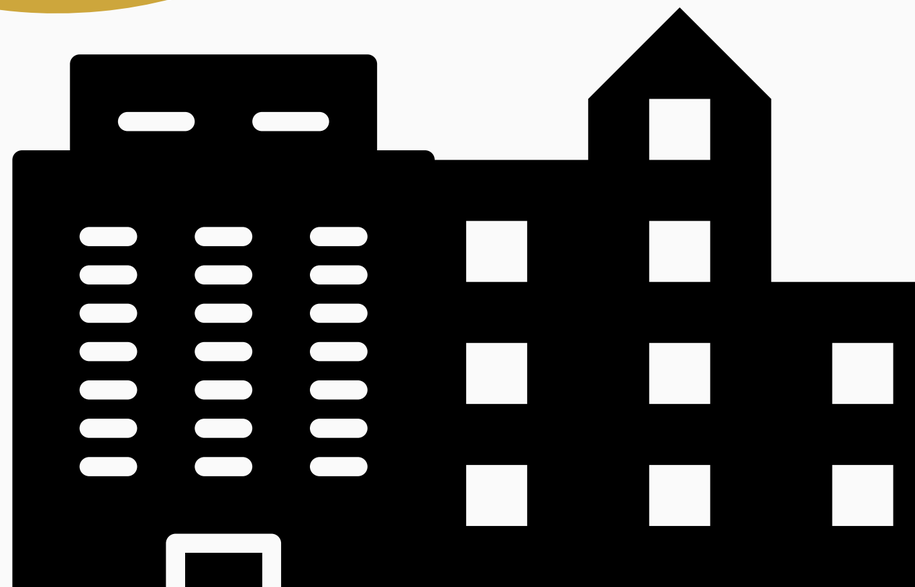
No	variabel	Deskripsi
1.	SK_ID_CURR	ID peminjam
2.	TARGET	Variable target (1=pelanggan dengan pembayaran yang sulit, 0=pelanggan dengan pembayaran yang mudah)
3.	NAME_CONTRACT_TYPE	Identifikasi apakah pinjaman itu tunai atau bergulir
4.	CODE_GENDER	Gender pelanggan
5.	FLAG_OWN_CAR	Tandai jika klien memiliki mobil
6.	FLAG_OWN_REALTY	Tandai jika klien memiliki rumah atau apartemen
7.	CNT_CHILDREN	Banyak anak yang dimiliki pelanggan
8.	AMT_INCOME_TOTAL	Penghasilan pelanggan
9.	AMT_CREDIT	Jumlah kredit dari pinjaman
10.	AMT_ANNUITY	Anuitas pinjaman
11.	AMT_GOODS_PRICE	Untuk pinjaman konsumen, itu adalah harga barang yang diberikan pinjaman
12.	DAYS_BIRTH	Usia klien dalam hari pada saat pengajuan
13.	DAYS_EMPLOYED	Berapa hari sebelum pengajuan pinjaman orang tersebut memulai pekerjaan saat ini
14.	EXT_SOURCE	Skor yang dinormalisasi dari sumber data eksternal

total ada 124 variabel

Describe Data

BUREAU

No	variabel	Deskripsi
1.	SK_ID_CURR	ID pinjaman - satu pinjaman dalam sampel dapat memiliki 0,1,2 atau lebih kredit terkait sebelumnya di biro kredit
2.	SK_BUREAU_ID	ID yang dikodekan ulang dari kredit Biro Kredit sebelumnya yang terkait dengan pinjaman kami (kode unik untuk setiap aplikasi pinjaman)
3.	CREDIT_ACTIVE	Status Credit Bureau (CB) laporan kredit
4.	CREDIT_CURRENCY	Recoded mata uang kredit Biro Kredit
5.	DAYS_CREDIT	Berapa hari sebelum pengajuan saat ini klien mengajukan kredit Biro Kredit
6.	CREDIT_DAY_OVERDUE	Jumlah hari lewat jatuh tempo kredit CB pada saat pengajuan pinjaman terkait dalam sampel kami
7.	DAYS_CREDIT_ENDDATE	Sisa durasi kredit CB (dalam hari) pada saat pengajuan di Home Credit
8.	DAYS_ENDDATE_FACT	Hari sejak kredit CB berakhir pada saat aplikasi di Home Credit (hanya untuk kredit tertutup)
9.	AMT_CREDIT_MAX_OVERDUE	Jumlah maksimum tunggakan kredit Biro Kredit sejauh ini (pada tanggal permohonan pinjaman dalam sampel kami)
10.	CNT_CREDIT_PROLONG	Berapa kali kredit Biro Kredit diperpanjang



total ada 19 variabel

Describe Data

PREVIOUS APPLICATION

No	variabel	Deskripsi
1.	SK_ID_PREV	ID kredit sebelumnya di Kredit rumah terkait dengan pinjaman dalam sampel
2.	SK_ID_CURR	ID pinjaman dalam sampel kami
3.	NAME_CONTRACT_TYPE	Jenis produk kontrak (Pinjaman tunai, pinjaman konsumen [POS] ,...) dari aplikasi sebelumnya
4.	AMT_ANNUITY	Anuitas aplikasi sebelumnya
5.	AMT_APPLICATION	Untuk berapa banyak kredit yang diminta klien pada aplikasi sebelumnya
6.	AMT_CREDIT	Jumlah kredit akhir pada aplikasi sebelumnya. Ini berbeda dari AMT_APPLICATION dengan cara AMT_APPLICATION adalah jumlah yang awalnya diajukan oleh klien, tetapi selama proses persetujuan kami dia dapat menerima jumlah yang berbeda - AMT_CREDIT
7.	AMT_DOWN_PAYMENT	Uang muka pada aplikasi sebelumnya
8.	AMT_GOODS_PRICE	Harga barang dari barang yang diminta klien (jika ada) pada aplikasi sebelumnya
9.	WEEKDAY_APPR_PROCESS_START	Pada hari apa klien mengajukan aplikasi sebelumnya
10.	HOUR_APPR_PROCESS_START	Kira-kira pada jam berapa klien melamar aplikasi sebelumnya
11.	FLAG_LAST_APPL_PER_CONTRACT	Tandai jika itu adalah aplikasi terakhir untuk kontrak sebelumnya. Terkadang karena kesalahan klien atau petugas kami, mungkin ada lebih banyak aplikasi untuk satu kontrak

total ada 37 variabel

Data Exploration Report

IV

Explore data adalah proses eksplorasi data yang bertujuan untuk memahami isi dan komponen penyusun data.

✓ Import data dan baca data

dengan menggunakan pandas
dengan task pd.read_csv

✓ Tampilkan data

dengan task head(), info(), columns(), shape(),dan hubungan variabel dengan target menggunakan plot

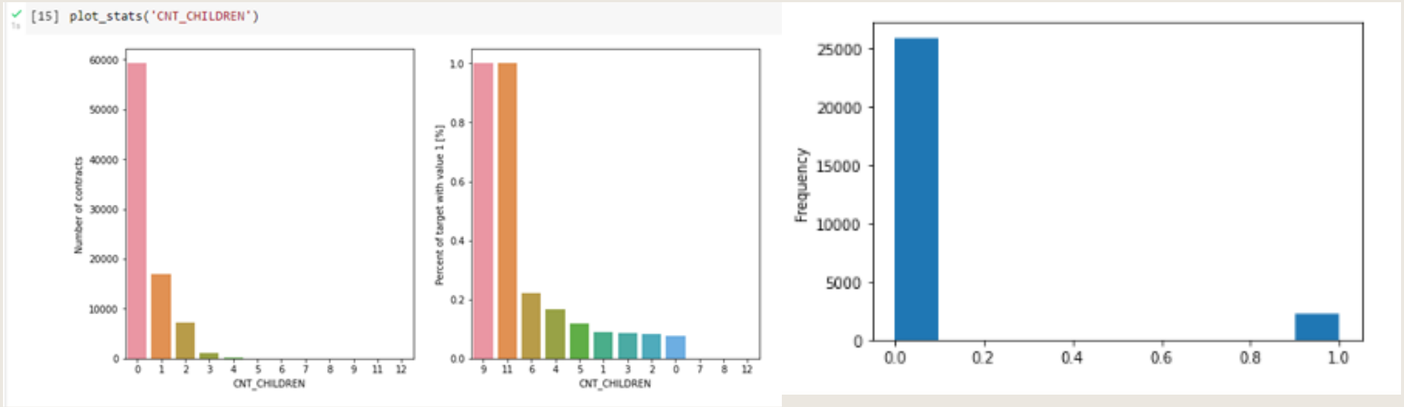
✓ periksa distribusi dari kolom target

dengan task df['TARGET'].value_counts() dan visualisasikan menggukanan plot histogram

✓ check null value

dengan menggunakan task df..isnull().sum() kemudian urutkan berdasarkan penurunan missing value dengan tabel

	Missing Values	% of Total Values
COMMONAREA_MODE	19725	70.0
COMMONAREA_MEDI	19725	70.0
COMMONAREA_AVG	19725	70.0
NONLIVINGAPARTMENTS_MODE	19559	69.4
NONLIVINGAPARTMENTS_MEDI	19559	69.4
NONLIVINGAPARTMENTS_AVG	19559	69.4
LIVINGAPARTMENTS_MODE	19279	68.4
LIVINGAPARTMENTS_AVG	19279	68.4
LIVINGAPARTMENTS_MEDI	19279	68.4



XII

RT

DATA PREPARATION

Data Preparation

Data Sets

Data set yang digunakan yaitu data application_train, application_test, bureau, dan previous application

Data Selection

Tahap selection data dilakukan dengan mencari korelasi terlebih dahulu dan dijelaskan lebih lanjut didalam proses modeling dengan menggunakan random forest untuk menemukan feature important dari data

Cleansing Data

Adapun tahapan yang dilakukan dalam cleansing data adalah sebagai berikut:

- ·Periksa missing value pada data dengan task df.isnull().sum()
- ·Drop columns yang memiliki missing value lebih dari 60% dengan task df = df.drop(columns=[])
- ·Fill missing value pada tabel yang masih diperlukan
- ·Encoding data, pada tahap ini kami menggunakan one hot encoding dengan task pd.get_dummies(df)
- ·Setelah selesai, cek shape data Kembali dan periksa korelasi data dengan target

Data Preparation

Integrating Data

Merge data yaitu menggabungkan Data digunakan untuk menggabungkan dua dataset secara horizontal, berdasarkan nilai atribut yang dipilih (kolom). Dalam input \ Merge Data, diperlukan dua set data, data dan data ekstra. Adapun task yang digunakan kelompok kami yaitu Pandas Join 3 DataFrame

XII

RT

MODELING

V

Modeling Technique

- Tujuan membuat modeling ini ialah untuk memprediksi pantas dan tidak pantas nya seseorang untuk mendapatkan pinjaman kredit.
  Teknik yang dipilih ialah *Random Forest Classifier* untuk membuat prediksi

BU

Random forest adalah Teknik modeling yang menggunakan sekumpulan pohon keputusan (decision trees). Setiap pohon keputusan dibangun menggunakan sebagian dari data training, sehingga pohon-pohon keputusan tersebut tidak saling bergantung satu sama lain.

VI

Generate Test Design

- **Identifikasi tujuan**

Tujuan dari pembuatan model ialah menentukan layak dan tidak layak nya seseorang dengan pengujian dengan berbagai feature

- **Metode Pengujian**

Menggunakan 70.000 Sample dataset yang terdapat pada application_train dengan modeling Random forest classifier

- **Skenario**
  - Menyiapkan data input Target
  - Split data Training dan data testing

VI

Building Model

Models & Parameter
Random Forest Classifier

```
# Modeling Random Forest dengan n_estimators 50 , random state 10, verbose 1 dan max_features 100
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(n_estimators=50,random_state=10,verbose=1,n_jobs=-1,max_features=100)
```

✓ 0.3s

Train Data

```
#Train data
rf.fit(X_train,y_train)
```

✓ 1m 20.2s

Prediksi Probabilitas

```
# Prediksi dengan probabilitas
rf_pred = rf.predict_proba(X_test)[:,1]
```

✓ 0.3s

VI

Building Model

Menampilkan hasil prediksi

```
# Hasil akhir akhir
submit = pd.DataFrame()
tes = pd.DataFrame(X_test)
submit['SK_ID_CURR'] = tes.index
submit['TARGET'] = rf_pred
print(submit.head())
print(submit.shape)
```

[245] ✓ 0.5s

```
   SK_ID_CURR  TARGET
0           0    0.04
1           1    0.12
2           2    0.30
3           3    0.16
4           4    0.04
(76878, 2)
```

VI

Model Evaluation

Prediksi

```
# Membuat dataframe prediksi kita dan menyeleksi 20.000 baris
rf_pred = pd.DataFrame(rf_pred)
rf_pred = rf_pred.head(70000)
rf_pred = round(rf_pred)
rf_pred.value_counts()
```

✓ 0.5s

```
0.0    69867
1.0      133
dtype: int64
```

Real Target

```
train_target_labels.value_counts()
```

1] ✓ 0.3s

```
TARGET
0         64373
1          5627
dtype: int64
```

Model mendapatkan nilai yang baik , akan tetapi nilai prediksi dan nilai Real target berbeda , mungkin dikarenakan beberapa sebab

```
                Evaluation Score
-------------------------------------------------
Score Akurasi : 0.9179714285714285
Score MSE     : 0.08202857142857142
Score MAE     : 0.08202857142857142
Score RMSE    : 0.2864063047989192
Nilai Confusion Matrix : [[64249   124]
 [ 5618     9]]
Classification Report :              precision    recall  f1-score   support

           0       0.92      1.00      0.96     64373
           1       0.07      0.00      0.00      5627

    accuracy                           0.92     70000
   macro avg       0.49      0.50      0.48     70000
weighted avg       0.85      0.92      0.88     70000
)
```

RT

EVALUATION

- Evaluating Results
- Process Review
- Next Step

SENTRA ZAMRUD | RAPAT TERBUKA

EVALUATION

Evaluating Results

Berdasarkan analisis, didapatkan nilai RMSE sebesar 0,286. Hal ini menunjukkan bahwa model Random Forest dapat melakukan klasifikasi dengan akurasi yang sangat baik hingga mencapai 92% dan nilai eror yang kecil yaitu dengan skor MAPE 8%. Ditemukan defaulted customer sebanyak 69.887 orang dan non defaulted customer sebanyak 113 orang.

Process Review

Business Understanding

Pada tahap ini kami mendeksripsikan latar belakang masalah, bisnis objektif, tujuan bisnis, kriteria sukses dan plan dari project kami. Bisnis homekredit merupakan perusahaan yang menyediakan pinjaman bagi pelanggan untuk memenuhi kebutuhan hidup yang memiliki kesulitan dalam menentukan customer mana yang layak mendapatkan pinjaman kredit. Dataset yang digunakan diambil dari kaggle.com. Kriteria untuk dapat menilai kelayakan peminjam yaitu 0 (defaulted/tidak layak) dan 1 (not defaulted/layak). Dalam penentuannya, digunakan pendekatan statistik dan metode machine learning untuk menilai kemampuan calon peminjam. Planning project dilakukan agar dapat memanage waktu dengan baik serta memudahkan proses pengerjaan.

Data Understanding & preparation

pada tahap proses ini, data dideskripsikan agar mudah dipahami dan dikelola seperti di explor, cleansing dan merge sehingga siap digunakan pada tahap selanjutnya

Modeling

Tujuan membuat modeling ini ialah untuk memprediksi pantas dan tidak pantas nya seseorang untuk mendapatkan pinjaman kredit. Teknik yang dipilih ialah Random Forest Classifier untuk membuat prediksi

EVALUATION

V

Determine Next Steps

Dari hasil model yang diterapkan dan feature important yang didapat maka dengan ini kita bisa melanjutkan ke step selanjutnya yaitu "Deployment"

VI

DEPLOYMENT

VI

Deployment Plan

Pada deployment plan proyek kali ini memiliki rencana untuk menampilkan visulasisasi antara variabel top feature. Top feature variabel merupakan variabel yang memiliki pengaruh besar dalam skoring kredit untuk menilai kelayakan client dalam menerima kredit. Untuk top feature diambil dari 2 tabel yaitu “train” dan “brueu”

VI

Plan Monitoring and Maintenance

Pada monitoring dan maintenance untuk dashboard kita menggunakan persektif dari pihak penyedia jasa kredit. Yang dimana data yang ditampilkan pada dashboard nanti memuat berbagai informasi (top feature) tentang client. Sehingga harapannya pihak penyedia jasa kredit menilai dan memutuskan clientnya akan menerima kredit atau tidaknya.

VI

Final Report

Link Google Data Studio :

https://datastudio.google.com/reporting/7563a095-180b-48ef-abfc-d68fa81e9f81

VI

Review Project

Hasil pembuatan dashboard dapat dibagi menjadi beberapa bagian untuk bisa dijelaskan secara proses seperti berikut :

- Flied ID Customer berfungsi untuk memanggil data client mana yang akan kita lihat dengan memanggil IDnya
- Perbandingan antara income total dan amt_annuity berfungsi untuk melihat apakah client sanggup tidaknya membayar cicilan berdasarkan pemasukannya
- Perbandingan usia, lama bekerja, dan lama registrasi menunjukan umur dan rekam jejak client dalam pengajuan kredit
- Keterediaan jumlah aset menunjukan nilai aset yang dimiliki oleh client yang menjadi salah satu nilai plus dalam pengajuan kredit
- Penggunaan pie chart untuk menampilkan riwayat dari peminjaman client, dimana menjelaskan saat ini client sedang menerima banyaknya peminjaman atau tidaknya
- Penggunaan barchart selanjutnya berfungsi untuk melihat durasi berapa lamanya jika client menunggak kredit