Final Project

Customer Churn

DigitalSkola

Yona Lotusia Pradana - DS18

OBJEKTIF BISNIS

Kasus :

Profitabilitas perusahaan diberbagai industri dipengaruhi oleh banyak hal, tetapi tidak ada yang lebih penting daripada retensi pelanggan. Kemampuan untuk mengembangkan dan mempertahankan basis pelanggan yang setia adalah tujuan utama bagi setiap perusahaan. Salah satu industri yang dipengaruhi oleh customer churn yaitu industri perbankan.

Jadi, para pengambil keputusan di industri perbankan sangat diharapkan untuk melacak, menganalisis, dan memprediksi potensi dari customer churn.

Solusi :

Memprediksi potensi customer churn, dan mengelompokkannya ke dalam kategori "Churn" atau "Tidak Churn", dengan cara menganalisa data dan menggunakan beberapa algoritma, sehingga dapat memberi masukan kepada para pengambil keputusan dalam meningkatkan profit perusahaan.

DESKRIPSI DATASET

	CustomerId	Gender	Age	CreditScore	EstimatedSalary	HasCrCard	Exited
0	15634602	Female	42	619	101348.88	1	1
1	15647311	Female	41	608	112542.58	0	0
2	15619304	Female	42	502	113931.57	1	1
3	15701354	Female	39	699	93826.63	0	0
4	15737888	Female	43	850	79084.10	1	0

Dataset ini mencakup atribut penting dari potensi customer churn atau tidak yang telah dikumpulkan secara manual. Link of the dataset https://drive.google.com/file/d/1QYip6pS6CBwZGJJe6nASTfyf3xx3dzVV/view?usp=sharing

Attribute	Description	Type
CustomerID	Nomor Identitas Nasabah	Integer
Gender	Jenis Kelamin Nasabah	Integer
Age	Usia Nasabah	Integer
CreditScore	Skor dari Penggunaan Kartu Kredit Nasabah	Integer
EstimatedSalary	Perkiraan Besaran Gaji Nasabah	Float
HasCrCard	Apakah Nasabah Memiliki Kartu Kredit atau Tidak	Integer
Exited	Indikasi Nasabah Churn atau Tidak	Integer

MEMBERSIHKAN DATA

```
#MISSING VALUE:
total = df.isnull().sum().sort_values(ascending=False)
percent = (df.isnull().sum()/df.isnull().count()).sort_values(ascending=False)
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
missing_data
```

	Total	Percent
Gender	0	0.0
Age	0	0.0
CreditScore	0	0.0
EstimatedSalary	0	0.0
HasCrCard	0	0.0
Exited	0	0.0

```
df.select_dtypes(np.number).columns
```

```
Index(['CustomerId', 'Age', 'CreditScore', 'EstimatedSalary', 'HasCrCard',
       'Exited'],
      dtype='object')
```

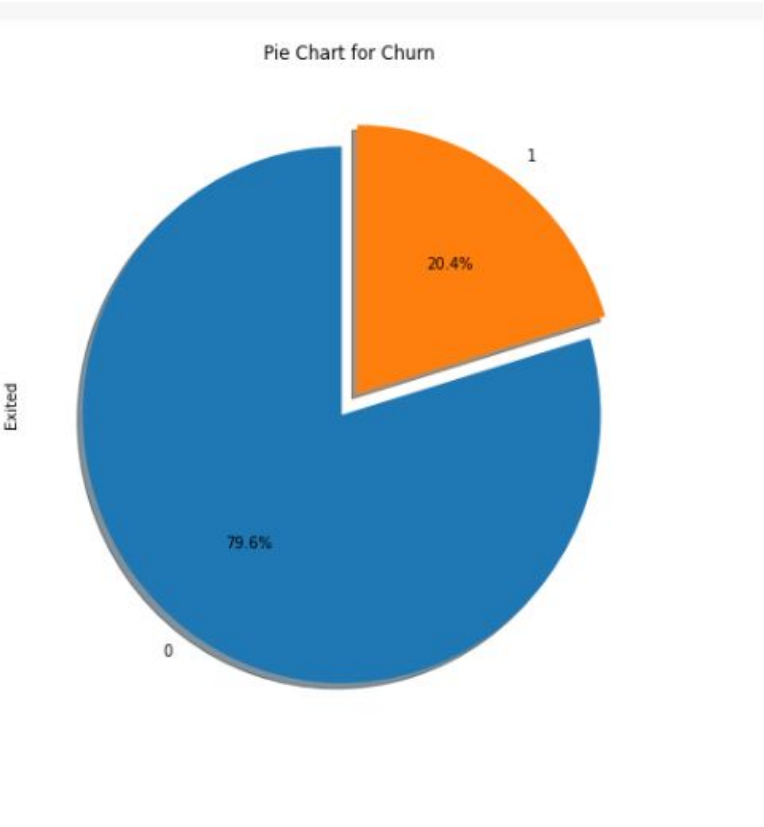
```
#Mengubah Tipe data Numerik menjadi Kategorik
df.select_dtypes('object').columns
```

```
Index([], dtype='object')
```

```
#Mengecek Duplikat
len(df[df.duplicated()])
```

```
0
```

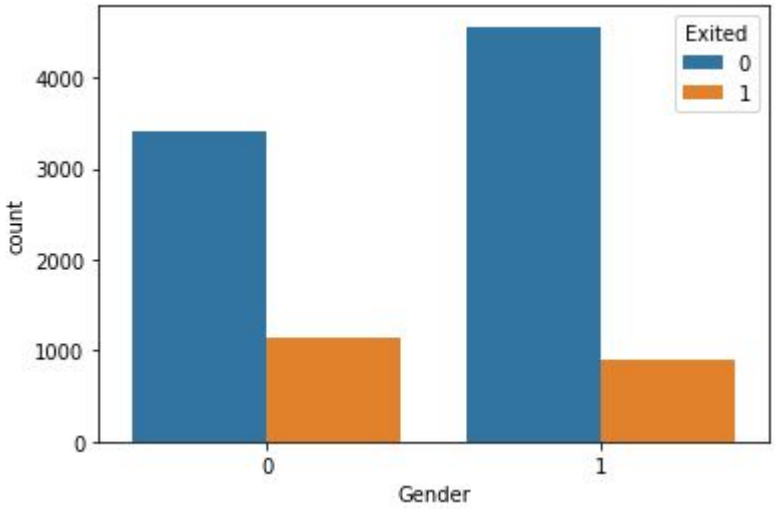
TAHAPAN ANALISA

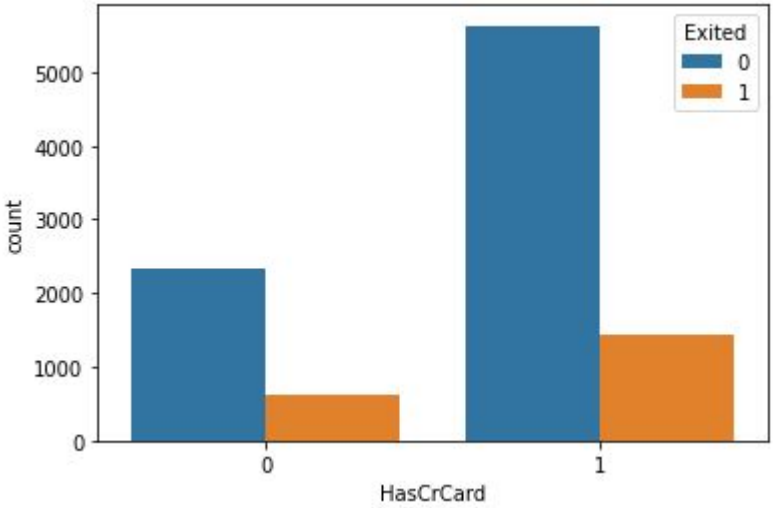


Sebanyak 2.037 nasabah churned/ tidak lagi menggunakan layanan dari bank dan sebanyak 7.963 nasabah masih menggunakan layanan bank tersebut

Adanya ketimpangan (skewed) antara total No Exited dan Exited. Hal ini mengakibatkan proporsi minority class (Exited) sebesar 20.4%, yang dimana masuk ke dalam kategori mild imbalanced data. Maka kita dapat melakukan teknik SMOTE.

A. Hubungan Data Kategorikal terhadap Target





Berdasarkan gender,

jumlah nasabah yang belum menutup akun = pria > wanita

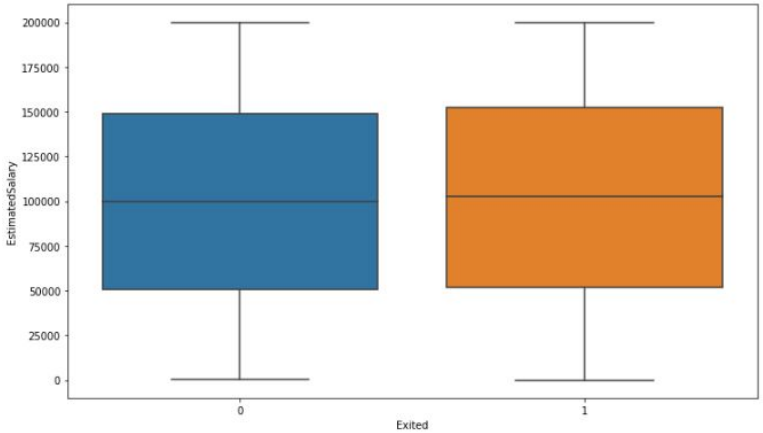
jumlah nasabah yang sudah menutup akun = wanita > pria

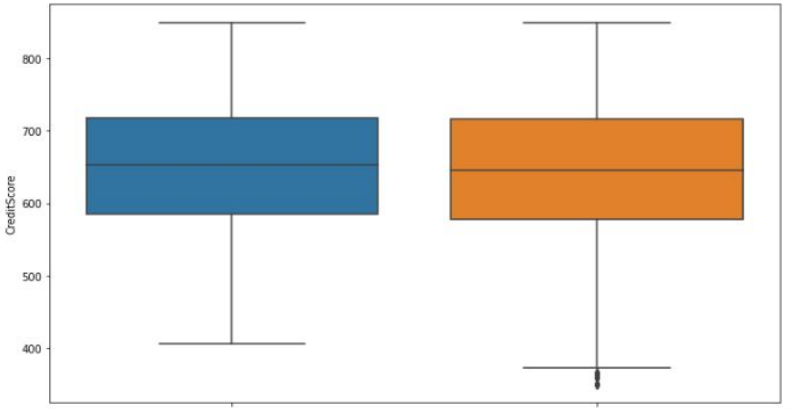
Berdasarkan kepemilikan kartu kredit,

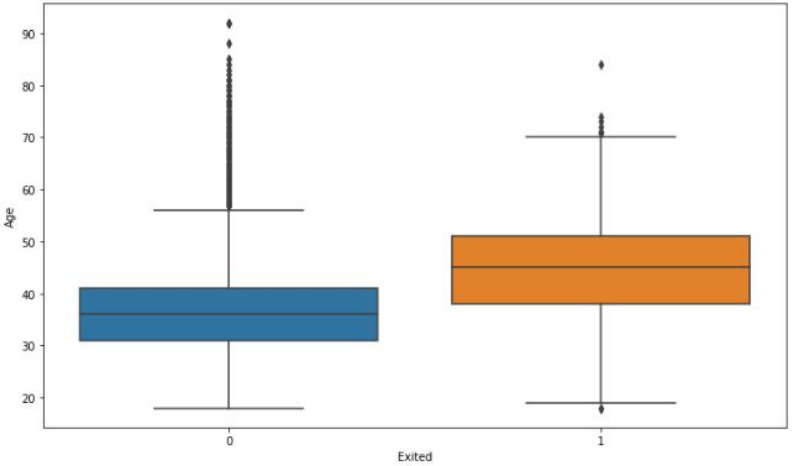
jumlah nasabah yang belum menutup akun = NoCrCard > HasCrCard

jumlah nasabah yang sudah menutup akun = HasCrCard > NoCrCard

B. Hubungan Data Numerikal terhadap Target







Berdasarkan perbandingan **EstimatedSalary, CreditScore** dan **Age** dengan Exited terlihat bahwa potensi untuk customer churn atau tidak churn dipengaruhi oleh usia.

C. Feature Engineering

Pengaplikasian Feature Engineering pada case ini yaitu melakukan perbandingan Credit Score dengan Usia Nasabah.

Hal tersebut dilakukan karena sebagai tolak ukur loyalitas nasabah yang mempengaruhi churn.

	Gender	Age	CreditScore	EstimatedSalary	HasCrCard	Exited	CrdScoreGivenAge
0	0	42	619	101348.88	1	1	14.738095
1	0	41	608	112542.58	0	0	14.829268
2	0	42	502	113931.57	1	1	11.952381
3	0	39	699	93826.63	0	0	17.923077
4	0	43	850	79084.10	1	0	19.767442
...	...	...	...	...	...	...	...
9995	1	39	771	96270.64	1	0	19.769231
9996	1	35	516	101699.77	1	0	14.742857
9997	0	36	709	42085.58	0	1	19.694444
9998	1	42	772	92888.52	1	1	18.380952
9999	0	28	792	38190.78	1	0	28.285714

10000 rows × 7 columns

MODELLING

Kita memecah dataset menjadi dua kelompok:

- Kelompok train set sebesar 65%
- Kelompok test set sebesar 35%

Algoritma yang kita gunakan pada model:

- KNN
- Logistic Regression
- SVC
- Decision Tree
- Random Forest

Karena data target tidak seimbang (imbalanced), maka kita melakukan dua tahap modelling, yaitu sebelum dan sesudah aplikasi SMOTE

Matriks Modelling

TANPA SMOTE

	KNN	Logistic Regression	SVC	Decision Tree	Random Forest
Accuracy	0.81	0.79	0.8	0.72	0.79
Precision	0.57	0.34	0.59	0.32	0.49
Recall	0.18	0.05	0.13	0.32	0.27
F1	0.27	0.08	0.22	0.32	0.35
ROC AUC	0.72	0.72	0.55	0.57	0.71

SMOTE

	KNN	Logistic Regression	SVC	Decision Tree	Random Forest
Accuracy	0.74	0.72	0.74	0.69	0.75
Precision	0.73	0.71	0.73	0.69	0.75
Recall	0.75	0.74	0.76	0.7	0.76
F1	0.74	0.73	0.75	0.69	0.75
ROC AUC	0.81	0.79	0.81	0.69	0.83

Feature Importance

(dilakukan pada data train setelah SMOTE)

	Features	Importance
5	CrdScoreGivenAge	0.272421
3	EstimatedSalary	0.248572
2	CreditScore	0.228183
1	Age	0.209797
4	HasCrCard	0.022673
0	Gender	0.018355

Deployed Model

https://finalprojectchurnpred.herokuapp.com/

SUMMARY

Perbandingan model tanpa SMOTE dengan SMOTE terlihat bahwa nilai persentase model dengan Accuracy, Precision, Recall, F1 dan ROC AUC terdapat perbedaan nilai imbalance. Maka perlu dilakukan SMOTE agar nilai persentase balance.

Setelah dilakukan perbandingan model maka dipilih model Random Forest dikarenakan beberapa hal yaitu :

1. Nilai akurasi yang terbaik dimana dapat dipertanggung jawabkan keakuratan datanya
2. Mempertimbangkan Recall dikarenakan nilai Recall yang semakin besar maka semakin kecil False Negative. Hal tersebut dapat memprediksi nasabah melakukan churn atau tidak dengan memperhatikan False Negative agar kecil kemungkinan kesalahan dalam memprediksi
3. Perbandingan Recall dan F1 dalam penerapan model Random Forest dan SVC saling bersaing. F1 didefinisikan sebagai mean harmonis antara Precision dan Recall. Jika nilai F1 mendekati 1 berarti nilainya lebih baik. Namun dilihat dari ROC AUC-nya, Random Forest lebih unggul. ROC itu kan menaikan True Positif dan False Postif, artinya tidak ada kesalahan dalam prediksi nasabah churn atau tidak. AUC berfungsi untuk melaraskan ROC, peran AUC untuk meninggikan True Positif dan merendahkan False Postif, sehingga prediksi lebih akurat. Maka kita memilih nilai ROC AUC yang lebih tinggi.


THANK
YOU