DigitalSkola

FINAL PROJECT

BANK CUSTOMER CHURN

Created by :

Naufal Abdurrahman Nafis

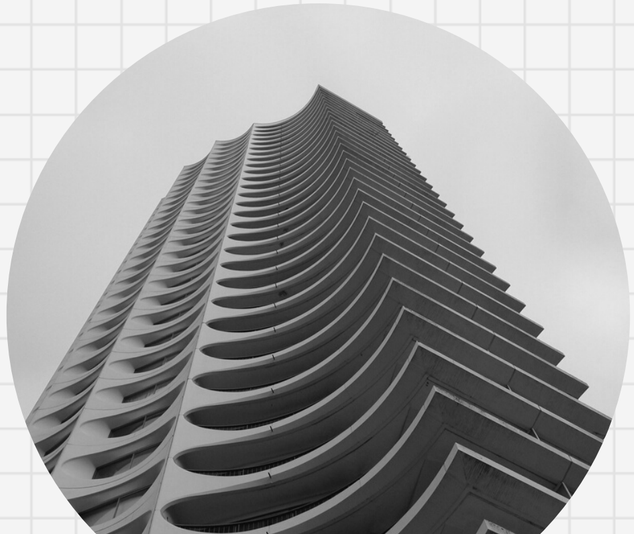


Table of Content

Business Understanding

Bank Customer Churn

Latar Belakang

Meningkatkan jumlah nasabah merupakan salah satu cara untuk mengembangkan bisnis suatu Bank. Divisi Marketing berusaha untuk terus menambah nasabah dan mempertahankan nasabah yang ada agar tidak menutup rekeningnya (churn).
Angka penutupan rekening yang tinggi tentunya akan mempengaruhi profit dan menghalangi pertumbuhan perusahaan. Prediksi yang kami lakukan bertujuan untuk memberikan insight kepada pihak terkait dalam melakukan strategi untuk mengembangkan bisnis perusahaan.

Objective

- Faktor apa yang paling penting yang mempengaruhi tingkat Churn yang tinggi?
- Model mana yang dapat memprediksi tingkat churn pelanggan dengan akurat?
- Bagaimana perusahaan dapat menggunakan analisis kami untuk meningkatkan pertumbuhan bisnisnya?

Data Understanding

Dataset ini terdiri dari 7 kolom dan 10.000 baris. Setiap kolom merepresentasikan informasi terkait nasabah. 7 kolom tersebut adalah:

	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

CustomerID : Nomor unik yang diberikan sebagai nomor identifikasi nasabah.

Gender : Jenis kelamin dari nasabah.

Age : Umur dari nasabah.

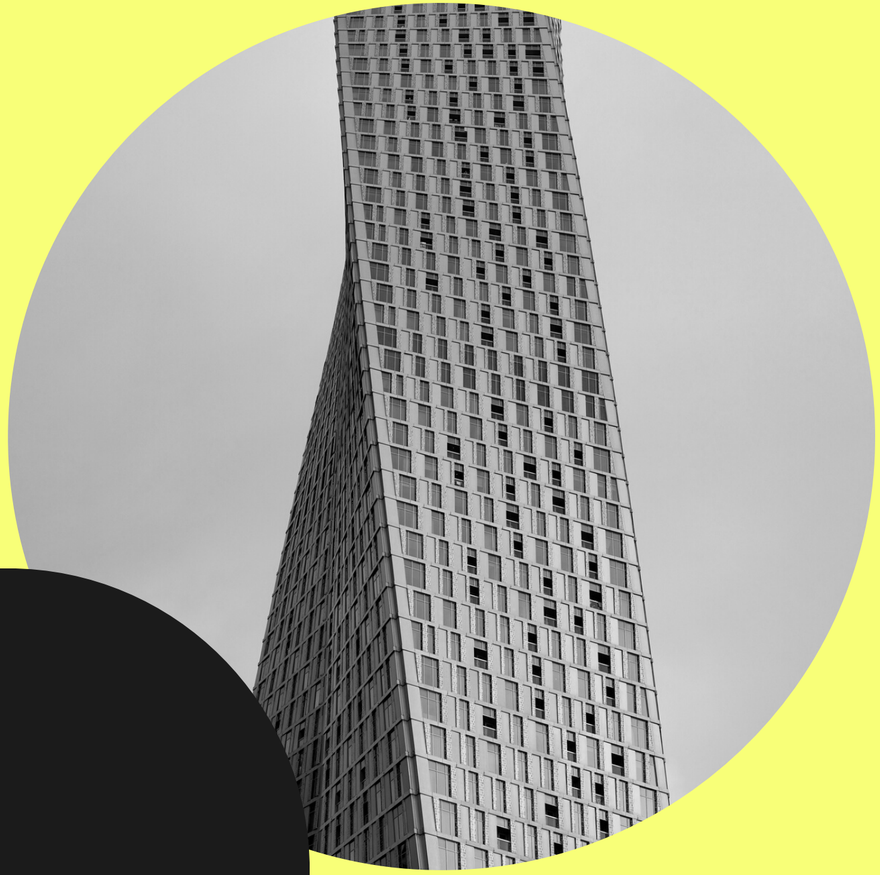
CreditScore : Nilai yang disusun atas berbagai faktor berdasarkan perilaku transaksi dari nasabah.

EstimatedSalary : Perkiraan pendapatan dari nasabah.

HasCrCard : Nasabah memiliki kartu kredit atau tidak.

Exited : Nasabah yang telah menutup rekening.

Data Preparation



Data Preparation

Setelah kami melihat dataset tersebut kami menganggap bahwa 'CustomerID' tidak ada kaitannya dengan apakah pelanggan akan churn atau tidak, jadi kami melakukan drop pada kolom 'CustomerID'.

```
df.drop(['CustomerId'], axis=1, inplace=True)
```

```
df.head()
```

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

Data Preparation

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   CustomerId       10000 non-null  int64  
 1   Gender           10000 non-null  object 
 2   Age              10000 non-null  int64  
 3   CreditScore      10000 non-null  int64  
 4   EstimatedSalary  10000 non-null  float64
 5   HasCrCard        10000 non-null  int64  
 6   Exited           10000 non-null  int64  
dtypes: float64(1), int64(5), object(1)
memory usage: 547.0+ KB
```

Data set ini terdiri dari 7 kolom dan 10.000 baris. Dari informasi data set dapat dilihat bahwa Data type untuk kolom **Gender** adalah **object**, dan data type untuk kolom **EstimatedSalary** adalah f**loat64.** Sedangkan sisanya memiliki data type **int64.**

Data Preparation

```
df.isnull().sum()
```

```
CustomerId          0
Gender              0
Age                 0
CreditScore         0
EstimatedSalary     0
HasCrCard           0
Exited              0
dtype: int64
```

Dari hasil pengecekan data **Null** dapat diketahui bahwa dari data set tersebut tidak terdapat data Null.

EXPLORATORY DATA ANALYSIS

Customer Churn Dataset

Exploratory Data Analisis terdiri dari menganalisis karakteristik utama dari sebuah set data biasanya dengan metode visualisasi dan statistik. Tujuan dari analisis ini adalah untuk memahami data, menemukan pola serta anomali, dan akan menemukan asumsi sebelum melakukan evaluasi lebih lanjut.

Menggunakan data.describe() untuk melihat karakteristik dari kumpulan data sebagai berikut :

	CustomerId	Age	CreditScore	EstimatedSalary	HasCrCard	Exited
count	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000
mean	1.569094e+07	38.921800	650.528800	100090.239881	0.70550	0.203700
std	7.193619e+04	10.487806	96.653299	57510.492818	0.45584	0.402769
min	1.556570e+07	18.000000	350.000000	11.580000	0.00000	0.000000
25%	1.562853e+07	32.000000	584.000000	51002.110000	0.00000	0.000000
50%	1.569074e+07	37.000000	652.000000	100193.915000	1.00000	0.000000
75%	1.575323e+07	44.000000	718.000000	149388.247500	1.00000	0.000000
max	1.581569e+07	92.000000	850.000000	199992.480000	1.00000	1.000000

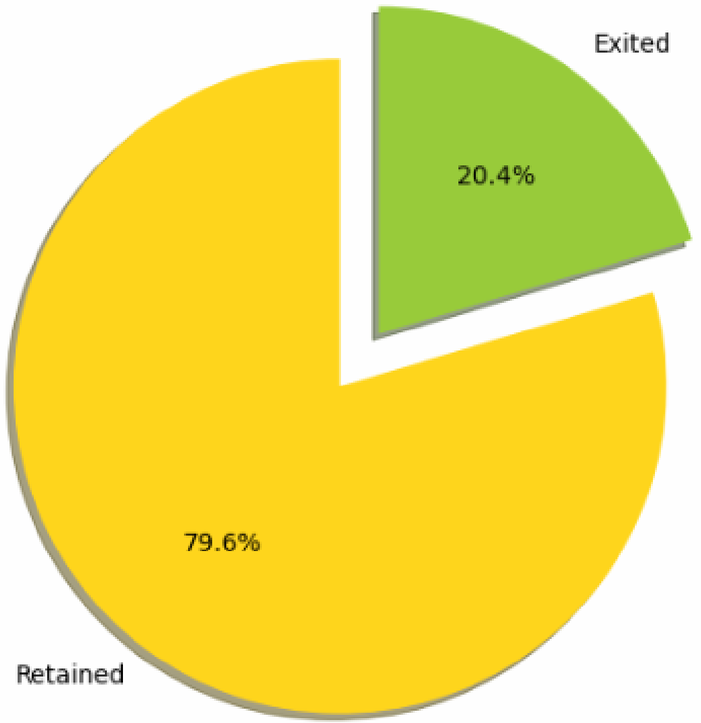
	Gender
count	10000
unique	2
top	Male
freq	5457

EXPLORATORY DATA ANALYSIS

Customer Churn Dataset

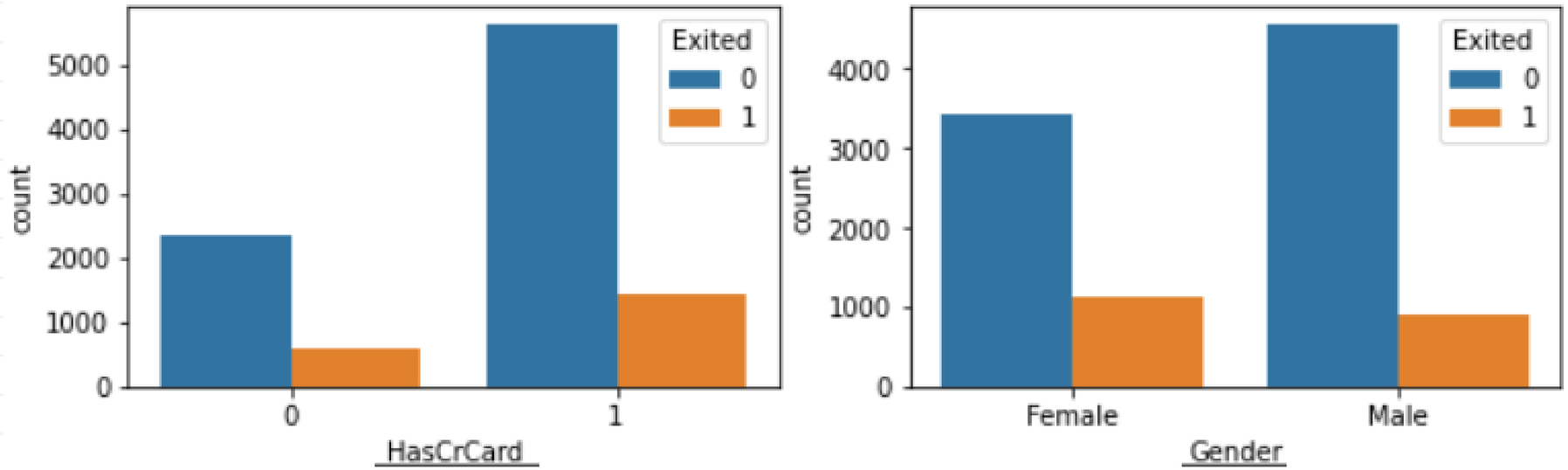
Dari Pie Chart tersebut dapat dilihat bahwa sebanyak 20,4% Nasabah Churn dan 79.6% Retained. Selanjutnya kita akan mengkategorikan beberapa fitur menjadi dua kategori yaitu categorical variables dan continues data attribute untuk menganalisis dengan melihat distribusi churn dari semua variable untuk dilihat alasan mengapa nasabah churn.

Dengan adanya data 20,4% nasabah Churn maka kami membuat model yang dapat memprediksi dengan sangat akurat sehingga dapat meminimalisir churn rate.



Categorical Variables

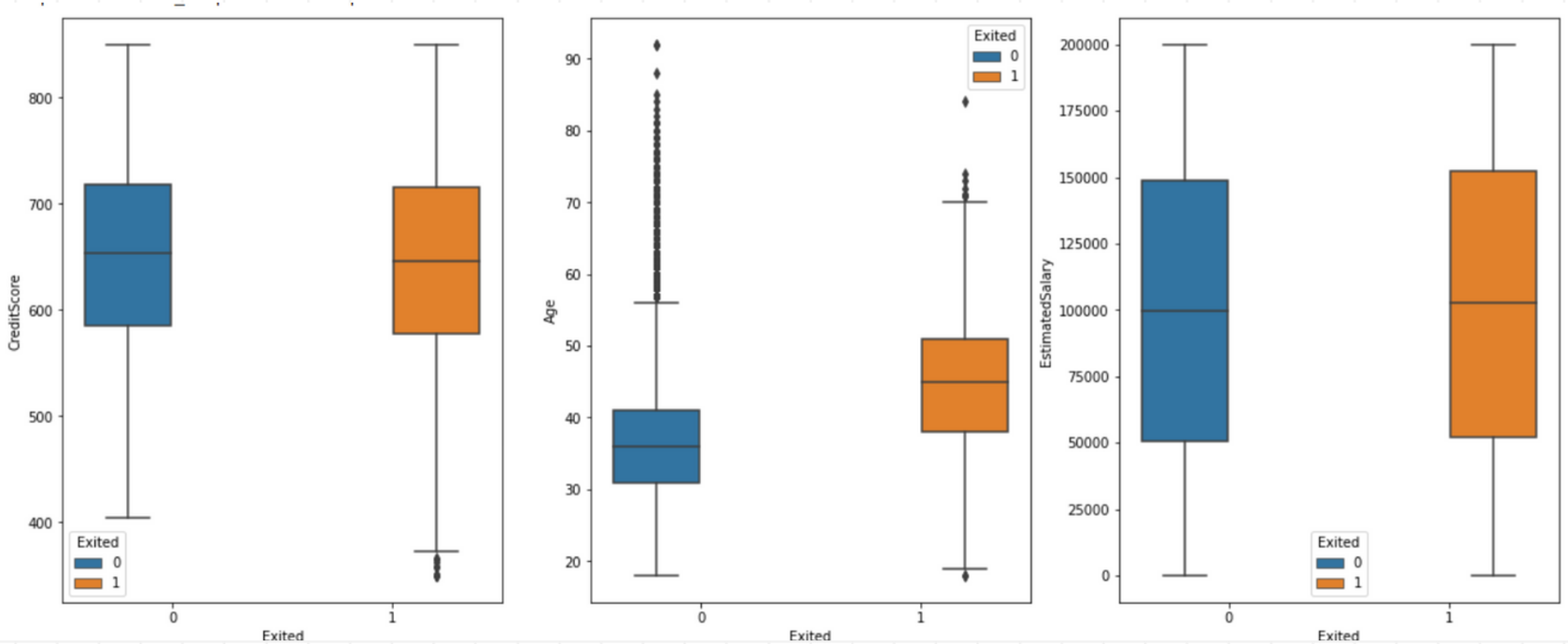
Berikut merupakan sebaran pelanggan berdasarkan categorical variables (HasCrCard dan Gender).



- Sebagian besar nasabah yang berhenti adalah mereka yang memiliki kartu kredit.
- Proporsi nasabah wanita yang berhenti lebih besar daripada nasabah pria.

Continuous Data Attributes

- Bisa kita lihat tidak ada perbedaan signifikan dalam distribusi skor kredit antara nasabah yang churn dan tidak churn.
- Nasabah yang lebih tua churn lebih banyak daripada yang lebih muda, ini menunjukkan adanya perbedaan preferensi layanan dalam kategori usia. Bank mungkin perlu meninjau pasar sasaran mereka atau meninjau strategi mereka.
- Tidak ada perbedaan yang signifikan dalam feature estimated salary.



Encoding Categorical Data

Pentingnya untuk melakukan encoding data karena hal tersebut digunakan untuk mengubah data kategori menjadi bentuk numerik yang dapat digunakan oleh algoritma Machine Learning. Encoding categorical data juga dapat membantu meningkatkan kinerja model.

```
[ ]  df = pd.get_dummies(df, drop_first=True)
```

```
[ ]  df.head()
```

	Age	CreditScore	EstimatedSalary	HasCrCard	Exited	Gender_Male
0	42	619	101348.88	1	1	0
1	41	608	112542.58	0	0	0
2	42	502	113931.57	1	1	0
3	39	699	93826.63	0	0	0
4	43	850	79084.10	1	0	0

Correlation Matrix sebelum FE

Dari heatmap correlation matrix dapat ditemukan bahwa adanya hubungan positif antara Exited dengan Age dan hubungan negatif antara Exited dan Gender.

Age dan Exited memiliki koefisien korelasi (0.29) artinya usia nasabah berhubungan positif dengan potensi customer churn. Semakin bertambah usia nasabah maka semakin besar potensi customer churn.

Gender dan Exited memiliki koefisien korelasi (-0.11) artinya perempuan yang melakukan churning lebih signifikan.

Feature Engineering

Mengingat fitur yang ada dalam dataset hanya 4 fitur yaitu Age, CreditScore, HasCrCard dan EstimatedSalary, maka kami memanfaatkan fitur yang ada untuk membentuk fitur baru dengan Feature Engineer untuk memperkuat model yang akan dibuat.

```
[ ]  # Credit score given age to take into account credit behaviour visavis adult life
     df_fe['Credit Score Given Age'] = df_fe["CreditScore"]/(df_fe["Age"])
```

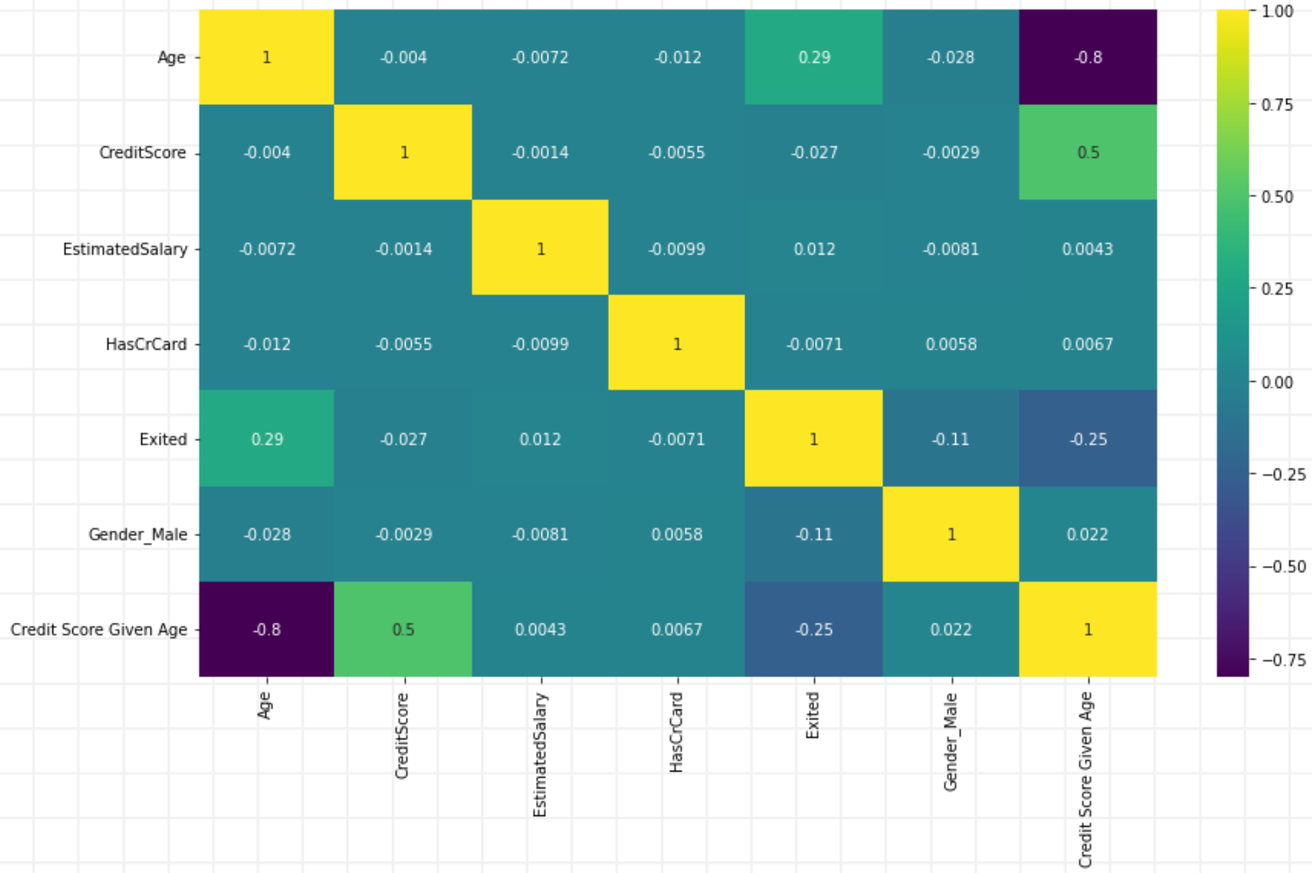
```
[ ]  # Resulting Data Frame
     df_fe.head()
```

	Age	CreditScore	EstimatedSalary	HasCrCard	Exited	Gender_Male	Credit Score Given Age
0	42	619	101348.88	1	1	0	14.738095
1	41	608	112542.58	0	0	0	14.829268
2	42	502	113931.57	1	1	0	11.952381
3	39	699	93826.63	0	0	0	17.923077
4	43	850	79084.10	1	0	0	19.767442

Correlation Matrix setelah FE

Setelah melakukan Feature Engineering, kami membuat correlation matrix dan ditemukan perubahan bahwa adanya korelasi terhadap feature baru yaitu Credit Score Given Age dengan Exited

Credit Score Given Age dan Exited memiliki koefisien korelasi (-0.25).



Modelling & Evaluation

Model Fitting & Selection

Kami mencoba model prediksi berikut sebelum nantinya memilih model prediksi terbaik, yaitu :

- Gradient Boosting Model
- K - Nearest Neighbor
- Random Forest
- Decision Tree Analysist

Split Data

Sebelum membuat model, kami melakukan split data dengan membagi data menjadi beberapa bagian yang digunakan untuk melakukan pemodelan. Dalam hal ini kami membagi data test sebesar 25% dan data train sebesar 75%.

```
[ ]  from sklearn.model_selection import train_test_split

     # Split Train, test data
     X = df_fe.drop(columns=["Exited"]).copy()
     y = df_fe["Exited"].copy()
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25,
                                                         random_state=42)
     print(len(X_train))
     print(len(X_test))
```

```
7500
2500
```

Confusion Matrix Before SMOTE

	precision	recall	f1-score	support
0	0.83	0.96	0.89	2003
1	0.57	0.24	0.34	497
accuracy			0.81	2500
macro avg	0.70	0.60	0.61	2500
weighted avg	0.78	0.81	0.78	2500

Gradient Boosting Model

	precision	recall	f1-score	support
0	0.80	0.94	0.87	2003
1	0.24	0.08	0.12	497
accuracy			0.77	2500
macro avg	0.52	0.51	0.49	2500
weighted avg	0.69	0.77	0.72	2500

KNN Model

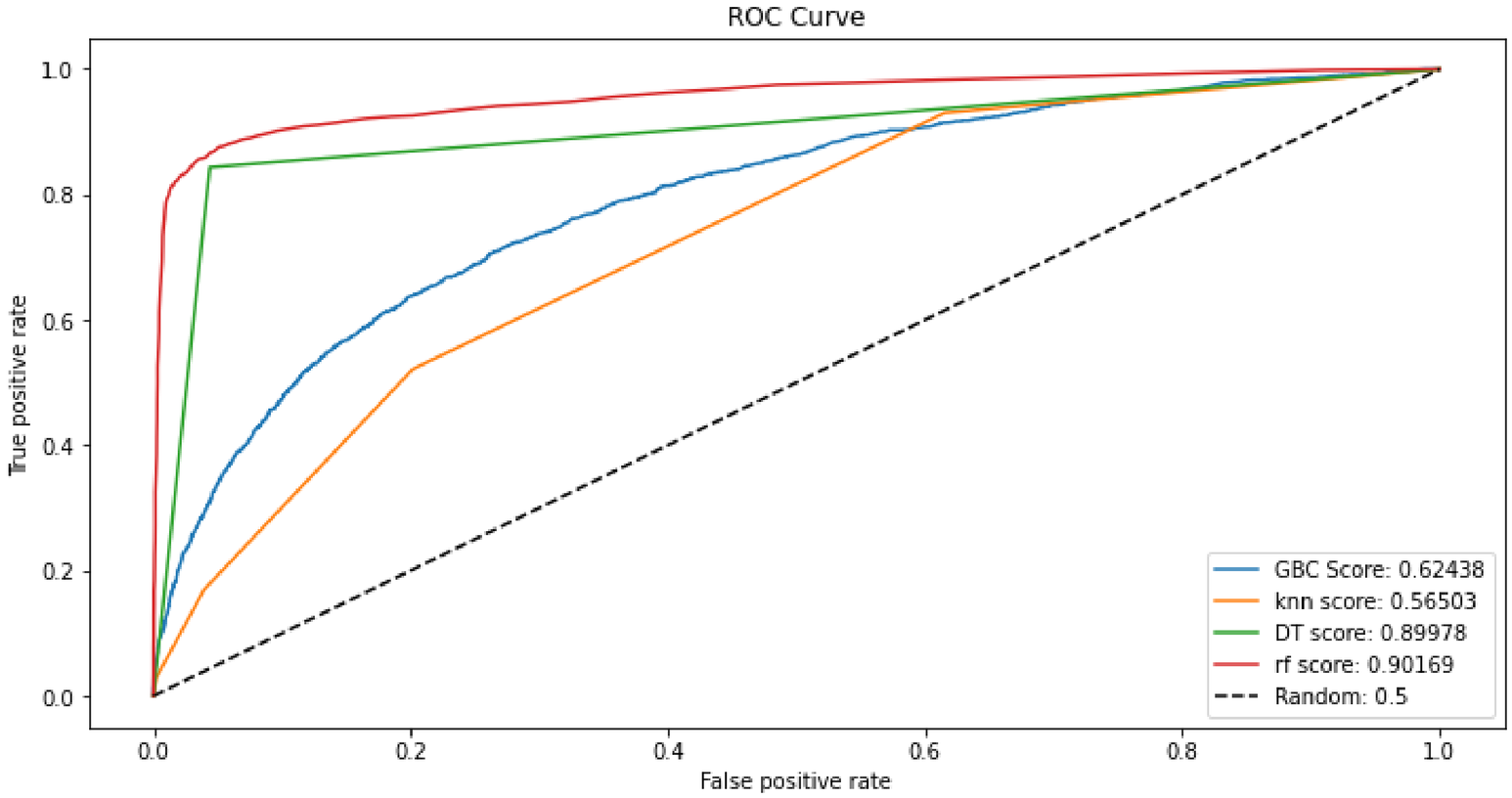
	precision	recall	f1-score	support
0	0.84	0.83	0.83	2003
1	0.34	0.36	0.35	497
accuracy			0.73	2500
macro avg	0.59	0.59	0.59	2500
weighted avg	0.74	0.73	0.74	2500

Decision Tree Analysis

	precision	recall	f1-score	support
0	0.84	0.93	0.88	2003
1	0.50	0.26	0.34	497
accuracy			0.80	2500
macro avg	0.67	0.60	0.61	2500
weighted avg	0.77	0.80	0.78	2500

Random Forest

Model Fitting & Selection Before SMOTE



Dari kurva ROC bisa dilihat bahwa model Random Forest memiliki kinerja terbaik dalam memprediksi dataset dengan nilai 0.9.

Confusion Matrix After SMOTE

	precision	recall	f1-score	support
0	0.88	0.75	0.81	2003
1	0.37	0.60	0.46	497
accuracy			0.72	2500
macro avg	0.63	0.68	0.64	2500
weighted avg	0.78	0.72	0.74	2500

Gradient Boosting Model

	precision	recall	f1-score	support
0	0.81	0.57	0.67	2003
1	0.21	0.45	0.28	497
accuracy			0.55	2500
macro avg	0.51	0.51	0.48	2500
weighted avg	0.69	0.55	0.59	2500

KNN Model

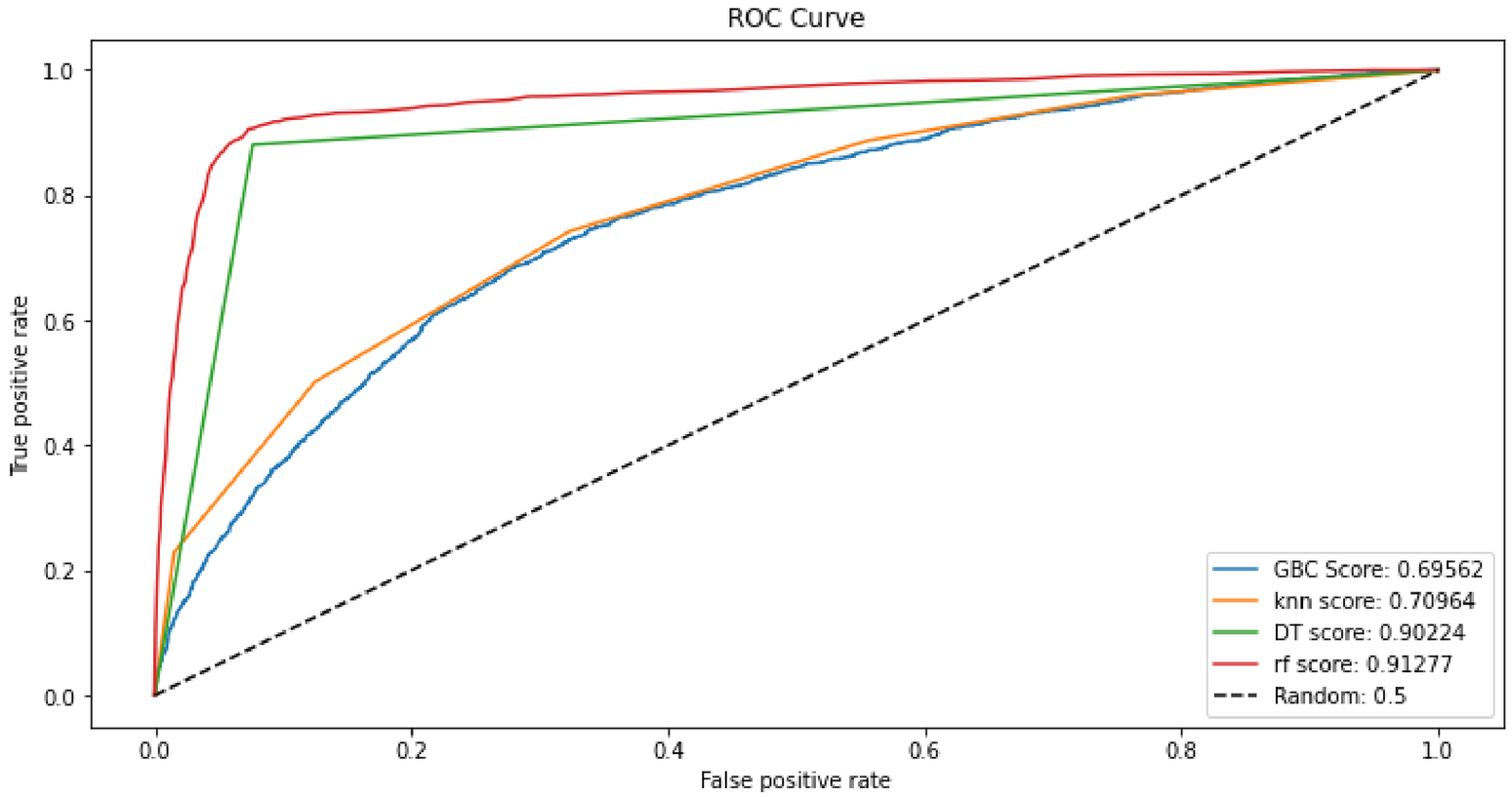
	precision	recall	f1-score	support
0	0.85	0.70	0.77	2003
1	0.30	0.51	0.37	497
accuracy			0.66	2500
macro avg	0.57	0.60	0.57	2500
weighted avg	0.74	0.66	0.69	2500

Decision Tree Analysis

	precision	recall	f1-score	support
0	0.87	0.75	0.81	2003
1	0.35	0.54	0.43	497
accuracy			0.71	2500
macro avg	0.61	0.65	0.62	2500
weighted avg	0.77	0.71	0.73	2500

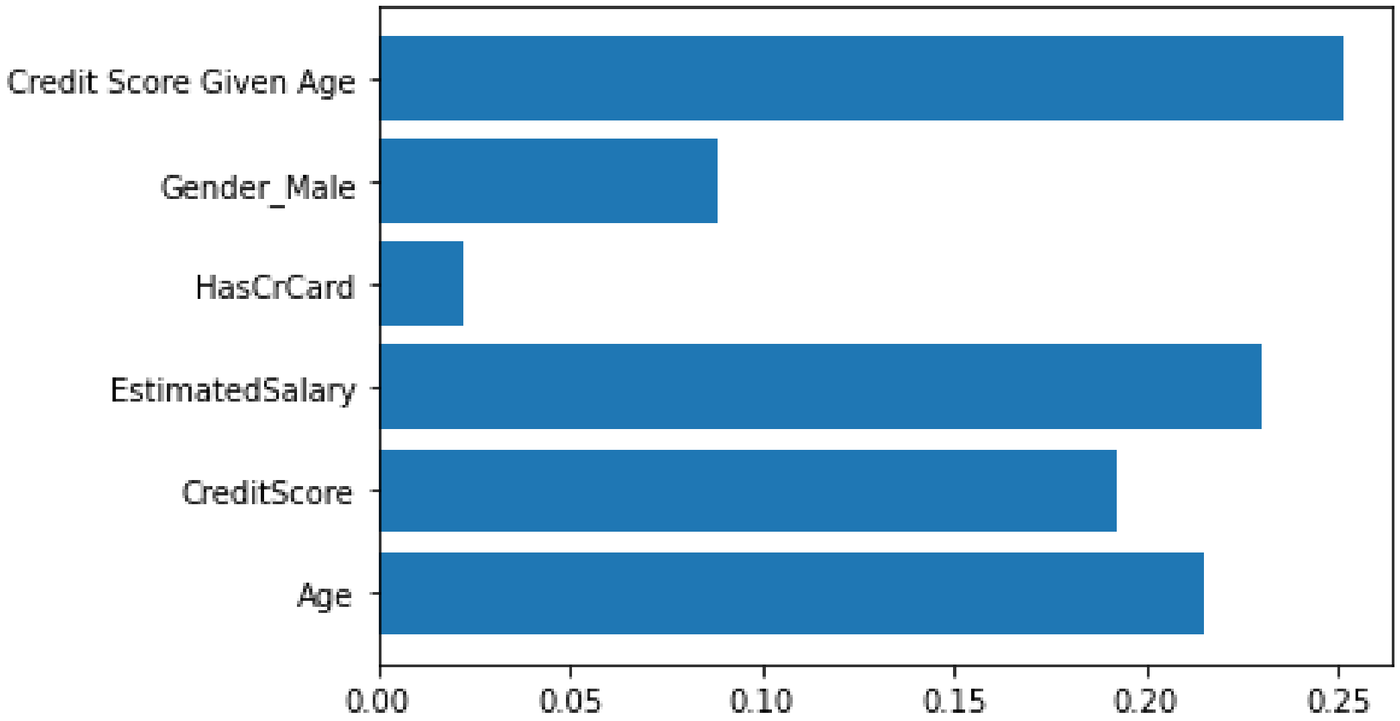
Random Forest

Model Fitting & Selection After SMOTE



Dari kurva ROC bisa dilihat bahwa beberapa kurva mengalami kenaikan akurasi setelah dilakukan SMOTE. Diperoleh model Random Forest memiliki kinerja terbaik dalam memprediksi dataset dengan nilai 0.91

Feature Importance



Grafik disamping menunjukkan fitur-fitur yang mempengaruhi model machine learning. Beberapa fitur nya adalah **Credit Score Given Age, EstimatedSalary, CreditScore dan Age.**

Kesimpulan & Rekomendasi

- Faktor gaji dan usia merupakan faktor yang paling mempengaruhi tingkat churn dengan korelasi positif.
- Model terbaik yang dapat memprediksi tingkat churn nasabah adalah Random Forest, karena memiliki nilai ROC yang paling tinggi dibandingkan model lain.
- Dari hasil prediksi yang kami pilih, maka perusahaan dapat memfokuskan pada segment usia muda dan berpenghasilan cenderung lebih rendah.

Saran

- Untuk meningkatkan akurasi model dapat dicoba melakukan SMOTE dengan perbandingan data 60:40 agar tetap ada data yang lebih dominan.

Terima kasih!

Created by:

Naufal Abdurrahman Nafis