Rakamin Academy

LAPORAN FINAL PROJECT

Kelompok 2 - **Synergies**

NAMA ANGGOTA KELOMPOK

Synergies Team

- Burhanudin Yusuf Robbani
- Mellia Anggraeni
- David Melanius Nai
- Moch Agung Laksono
- Alfath Arrahman
- Dzul Wulan Ningtyas
- Zaima Syarifa Asshafa

OUTLINE PROJECT

Final Project - Synergies Team

Rakamin Academy

Stage 0
Preparation

Stage 1
EDA, insight, and visualization

Stage 2
Data Preprocessing

Stage 3
Modelling and Evaluation

Rangkuman Stage 0 : Preparation

Rakamin Academy

Problem Statement

Perusahaan Rakamin Bank Center memperoleh jumlah nasabah churn sebesar 20,37% dari keseluruhan data. Berdasarkan website *https://uxpressia.com/blog/how-to-approach-customer-churn-measurement-in-banking* toleransi nasabah churn maksimal sebesar 10%. Sementara itu, jumlah nasabah churn yang diperoleh melebihi batas toleransi.

Role

Sebagai tim data scientist dari perusahaan Rakamin Bank Center (RBC), kami bertanggung jawab, menganalisa data dan membuat model yang mana akan memprediksi nasabah mana yang akan churn.

Goals

Memprediksi nasabah yang akan churn dengan tingkat akurasi diatas 70%.

Objectives

Membuat model Machine Learning untuk membantu Perusahaan RBC dalam memprediksi nasabah yang akan churn dan membantu tim bisnis dalam menentukan strategi terhadap nasabah yang akan churn.

Business Metrics

- Churn Rate
- F1 Score sebagai metric sekunder.

Rangkuman Stage 1 : EDA, Insights & Visualization

Rakamin Academy

Descriptive Statistics

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

```
df_bank.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography        10000 non-null  object
 5   Gender           10000 non-null  object
 6   Age              10000 non-null  int64
 7   Tenure           10000 non-null  int64
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64
 10  HasCrCard        10000 non-null  int64
 11  IsActiveMember   10000 non-null  int64
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64
```

Dari dataset tersebut didapat:

- Berisi 10.000 baris dan 14 kolom.
- Kolom Exited sebagai variable targetnya.
- Semua atribut tipe datanya sudah sesuai dengan isi data, namun ada beberapa atribut yang perlu dikonversi untuk mencari pola pada data.
- Semua atribut tidak memiliki nilai kosong.

Rangkuman Stage 1 : EDA, Insights & Visualization

Rakamin Academy

Descriptive Statistics

Numerical

```
nums_df.describe()
```

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	100090.239881
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	57510.492818
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	11.580000
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	51002.110000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	100193.915000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	149388.247500
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	199992.480000

Dilihat sekilas dari Descriptive Statistics Numerical sesuai dengan rentang nilai jarak antara nilai median dan rata-rata sehingga berdasarkan angka dan sebaran datanya, sebagai berikut:

- Fitur CreditScore, Age, Tenure, NumOfProduct dan EstimatedSalary mempunyai sebaran data cenderung memiliki **distribusi normal**.
- Fitur Balance mempunyai sebaran data cenderung mempunyai sebaran data cenderung **distribusi skew** dan memiliki **nilai outlier yang ekstrim**.

Categorical

```
cats_df.describe()
```

	Surname	Geography	Gender	HasCrCard	IsActiveMember	Exited
count	10000	10000	10000	10000	10000	10000
unique	2932	3	2	2	2	2
top	Smith	France	Male	1	1	0
freq	32	5014	5457	7055	5151	7963

Dilihat sekilas dari Descriptive Statistics Categorical:

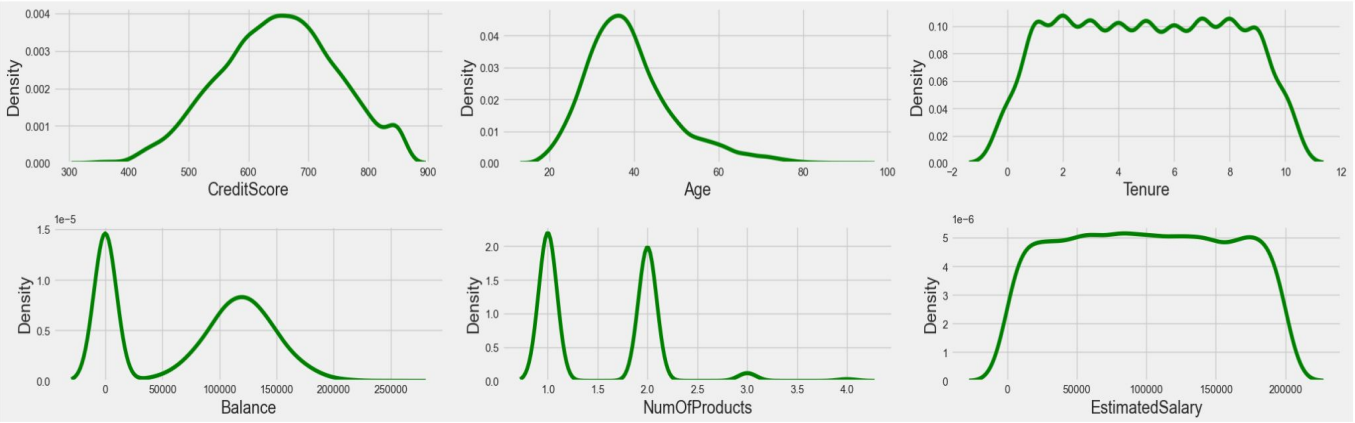
- Fitur Geography memiliki **3 nilai unik** dan negara **Prancis** merupakan negara dengan lokasi nasabah yang paling banyak.
- Fitur Gender memiliki **2 nilai unik** dan kebanyakan nasabah memiliki status jenis kelamin **laki-laki**.
- Fitur HasCrCard memiliki **2 nilai unik** dan kebanyakan nasabah **sudah memiliki kartu** kredit.
- Fitur IsActiveMember memiliki **2 nilai unik**.
- Variabel target (Exited) memiliki **2 nilai unik.**

Rangkuman Stage 1 : EDA, Insights & Visualization

Rakamin Academy

Univariate Analysis (Numerical)

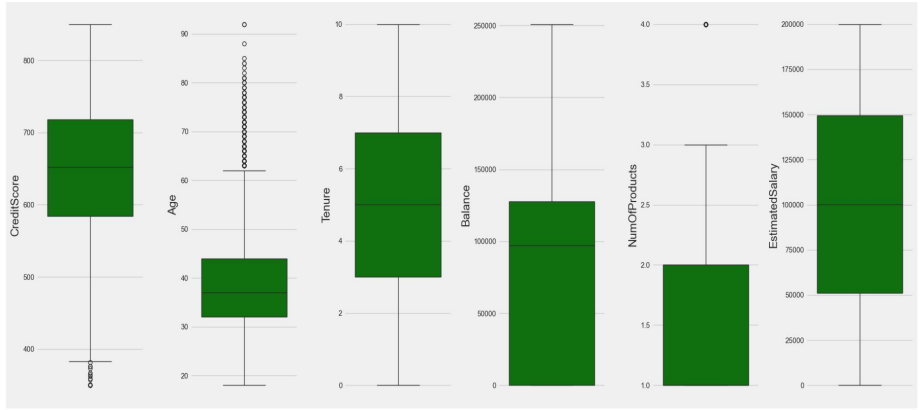
KDEplot



Dilihat dari grafik kdeplot di atas:

- Fitur CreditScore memiliki bentuk **distribusi normal**.
- Fitur Age memiliki bentuk **distribusi normal.**
- Fitur Tenure memiliki bentuk **uniform bertipe diskrit**.
- Fitur Balance memiliki kecenderungan **berdistribusi normal.**
- Fitur NumOfProducts memiliki **distribusi bimodal.**
- Fitur EstimatedSalary cenderung memiliki **distribusi normal.**

Boxplot



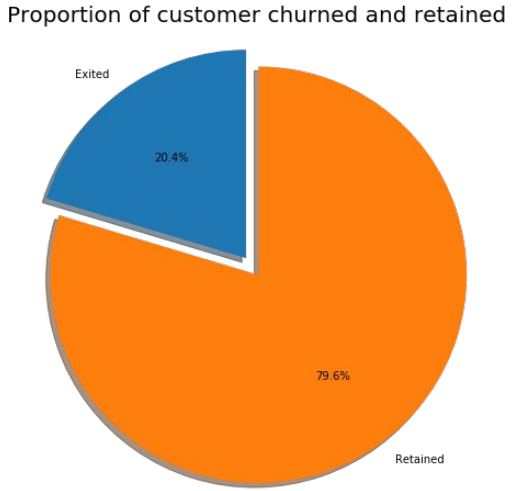
Dilihari dari grafik boxplot di atas:

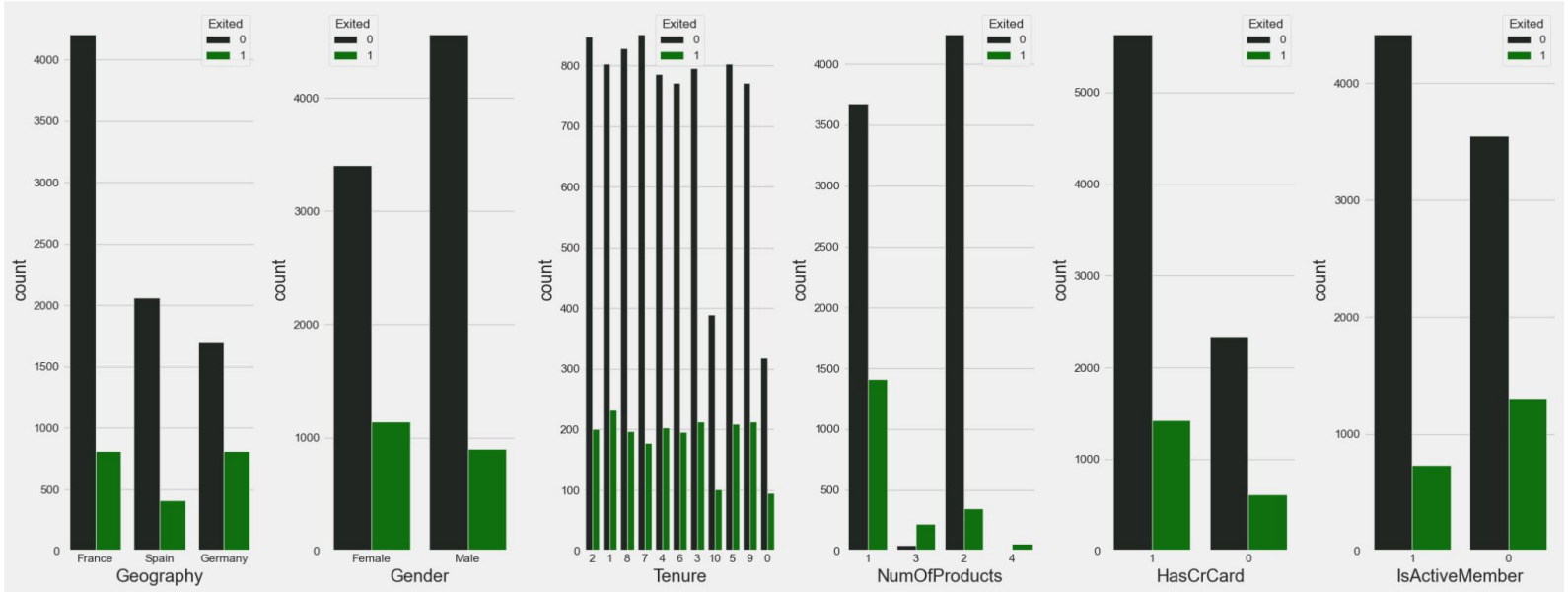
- Fitur yang memiliki nilai outlier yaitu CreditScore, Age dan NumOfProducts.
- Fitur yang tidak memiliki nilai outlier yaitu Tenure, Balance and EstimatedSalary.

Rangkuman Stage 1 : EDA, Insights & Visualization

Rakamin Academy

Univariate Analysis (Categorical)





Beberapa informasi yang dapat diperoleh dari informasi di atas adalah:

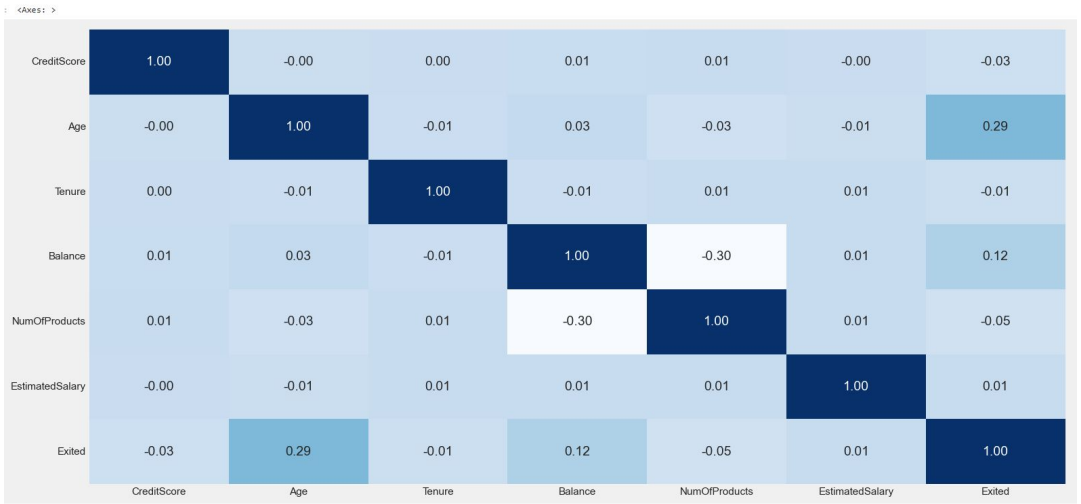
- Variabel target (Exited) memiliki **bentuk data yang tidak seimbang**.
- Semua fitur distribusinya **terlihat tidak seimbang ketika data dipecah berdasarkan variabel targetnya**. Diperlukan pemerataan data.

Rangkuman Stage 1 : EDA, Insights & Visualization

Rakamin Academy

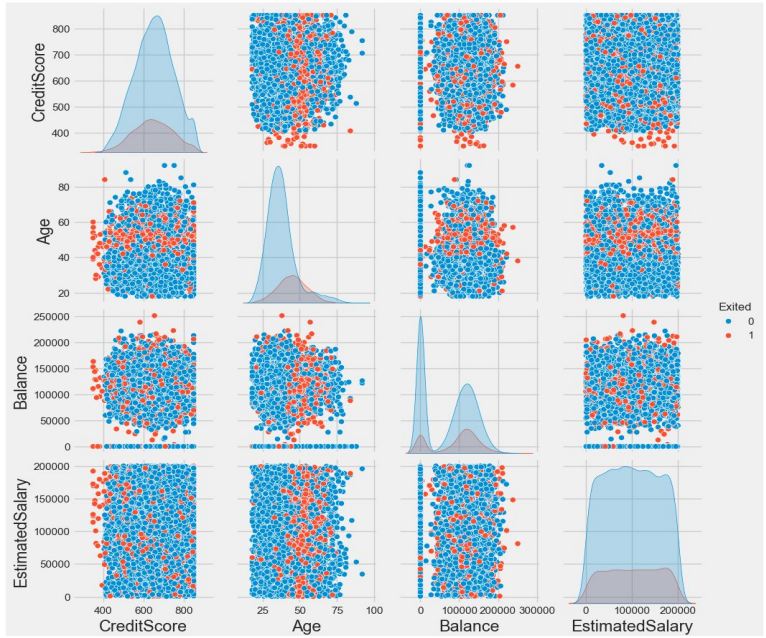
Multivariate Analysis

Heatmap



Informasi yang diperoleh dari grafik heatmap adalah sebagai berikut: **Semua fitur numerikal** cenderung memiliki **hubungan korelasi yang lemah** sehingga **tidak ada hal yang mengindikasikan adanya multikolinearitas.**

Pairplot



Informasi yang diperoleh dari grafik pairplot adalah sebagai berikut:

- **Tidak terlihat adanya segmentasi tertentu** pada distribusi data antara nasabah churn dan non-churn terhadap hubungan antar fitur (Scatter plot).
- Begitu pula data Categorical – Numerical, hasilnya **tidak terlihat adanya segmentasi tertentu**.

Rangkuman Stage 1 : EDA, Insights & Visualization

Rakamin Academy

Follow up for Pre-processing

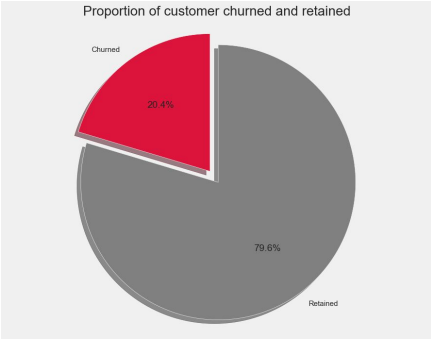
Kesimpulan dari EDA sebelumnya, maka diperoleh beberapa follow-up yang perlu dilakukan ketika melakukan data cleansing yaitu;

1. **Menghapus beberapa fitur yang tidak relevan.**
2. **Mengecek apakah ada sebuah data duplikat**.
3. **Menghandle outliers** terhadap CreditScore dan Age.
4. **Melakukan fitur engineering** terhadap penambahan fitur baru.
5. **Melakukan data scaling** terhadap beberapa fitur.
6. **Melakukan fitur encoding**.
7. **Melakukan proses menangani data yang tidak seimbang.**

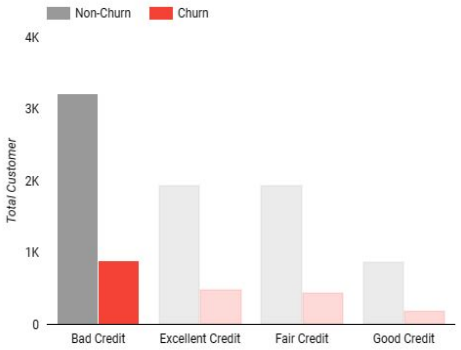
Rangkuman Stage 1 : EDA, Insights & Visualization

Rakamin Academy

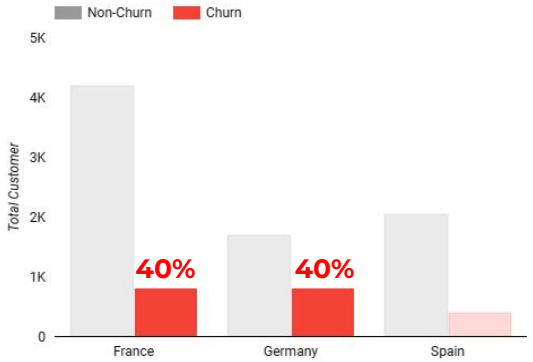
Business Insight



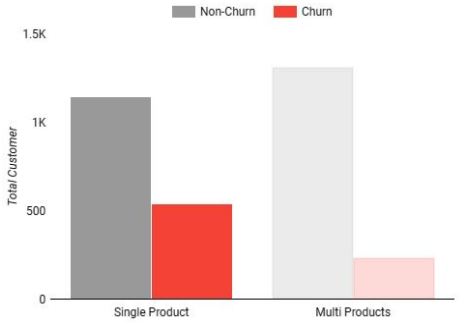
Jumlah nasabah churn sebanyak 20% dari keseluruhan nasabah.



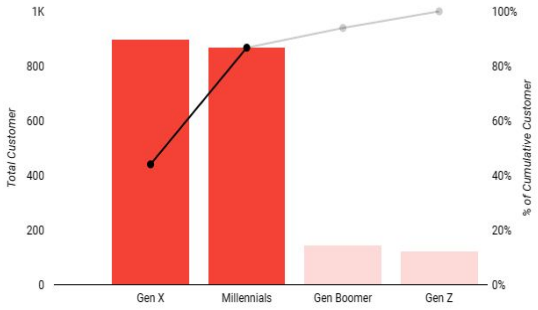
45% nasabah churn memiliki status **skor kredit yang buruk**.



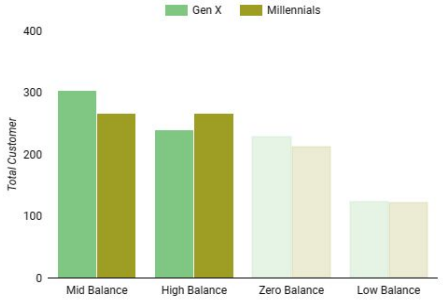
80% nasabah churn berasal dari **negara prancis** dan **jerman**.



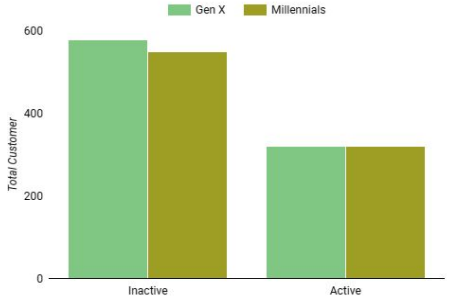
69% nasabah churn hanya memiliki **satu jenis produk saja**.



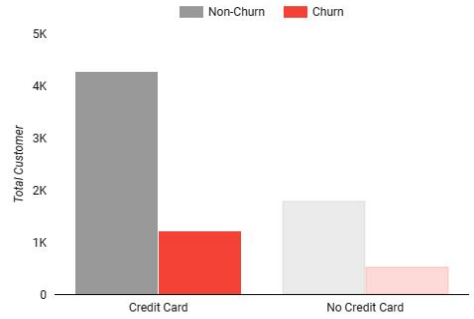
Akumulasi persentase dari nasabah churn sekitar **86%** yang berasal dari **generasi Millennials** dan **generasi X**.



Akumulasi persentase dari nasabah churn sekitar **61%** yang berasal dari nasabah dengan **rentang jumlah balance menengah ke atas**.



64% nasabah churn memiliki riwayat **tidak aktif dalam melakukan aktivitas transaksinya**.



70% nasabah churn **memiliki kartu kredit**.

Rangkuman Stage 1 : EDA, Insights & Visualization

Rakamin Academy

Insight Summary

Beberapa rangkuman insight yang diperoleh dari hasil analisa data lainnya adalah;

1. Perbandingan nasabah yang masih menggunakan produk bank antara berstatus aktif dan non-aktif adalah **56 : 44 persen.**
2. Sebanyak **86%** nasabah churn mayoritas berasal dari **generasi Millennials** dan **generasi X**.
3. Pada generasi tersebut, mereka memiliki **jumlah balance dengan rentang menengah ke atas**.
4. Mereka churn karena **banyak yang sudah tidak aktif** dalam melakukan aktivitas transaksi.
5. Dampak dari banyaknya kasus nasabah churn karena tidak aktifnya mereka dalam melakukan aktivitas transaksi **bisa terjadi kembali apabila tidak dilakukan strategi khusus** terhadap nasabah yang masih menggunakan produk bank.
6. Dan sebaiknya juga **lakukan strategi berbeda terhadap nasabah yang aktif** dalam aktivitas transaksinya.
7. Mayoritas dari nasabah churn tersebut **memiliki kartu kredit**.
8. Namun mayoritas juga **skor kartu kredit mereka berstatus buruk**.
9. Kemungkinannya salah satu penyebab adalah **produk yang dimiliki hanya satu jenis saja**.

Rangkuman Stage 1 : EDA, Insights & Visualization

Rakamin Academy

Business Recommendation

1. **Nasabah berstatus aktif** pada gen millennials dan gen X yang memiliki jumlah saldo menengah ke atas, **diberikan suatu program loyalti berupa reward poin** agar mereka terus melakukan aktivitas transaksi.

2. **Nasabah berstatus tidak aktif** pada generasi dan jumlah saldo yang sama dengan di atas, **diberikan push notification melalui SMS kepada nasabah yang berasal dari Perancis** dengan konten berupa **promosi diskon belanja berkategori lifestyle di merchant tertentu**, sedangkan **kepada nasabah yang berasal dari Jerman** kontennya berupa **promosi diskon belanja berkategori luxury item di merchant tertentu**.

3. **Meningkatkan kualitas produk yang dimiliki** agar nasabah tertarik menggunakan lebih dari satu jenis produk.

4. **Nasabah dengan status skor kredit yang buruk** diberikan **potongan bunga pinjaman** saat pembayaran tagihan kredit.

Rangkuman Stage 2: Data Prepocessing

Rakamin Academy

Dataset Features

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Diperoleh dataset dengan isi **10K baris** dan **14 kolom** dengan **kolom Exited sebagai variabel target** dan sisanya adalah variabel fitur. Terlihat semua atribut sudah memiliki tipe data yang sesuai dan isi datanya juga sudah sesuai.

```
df_bank.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography        10000 non-null  object
 5   Gender           10000 non-null  object
 6   Age              10000 non-null  int64
 7   Tenure           10000 non-null  int64
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64
 10  HasCrCard        10000 non-null  int64
 11  IsActiveMember   10000 non-null  int64
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64
```

Rangkuman Stage 2: Data Prepocessing

Rakamin Academy

Removing Irrelevant Features – Part 1

```
df_bank = df_bank.drop(columns = ['RowNumber'])
```

Melakukan penghapusan fitur row number untuk mengecek apakah ada data duplikat nantinya

Handling Duplicates Data

```
df_bank.duplicated().any()
```

```
False
```

Tidak ditemukan adanya data duplikat setelah menghapus fitur row number

Removing Irrelevant Features – Part 2

```
df_bank = df_bank.drop(columns = ['CustomerId', 'Surname'])
```

Penghapusan **fitur customer id dan surname**, karena kedua fitur tersebut tidak memberikan informasi yang penting untuk digunakan sebagai model klasifikasi.

Rangkuman Stage 2: Data Prepocessing

Rakamin Academy

Handling Missing Values

```
df_bank.isnull().any()
```

```
CreditScore          False
Geography            False
Gender               False
Age                  False
Tenure               False
Balance              False
NumOfProducts        False
HasCrCard            False
IsActiveMember       False
EstimatedSalary      False
Exited               False
dtype: bool
```

```
# checking if there is any irrelevant values in categorical features

print(df_bank.Geography.value_counts())
print(df_bank.Gender.value_counts())
```

```
Geography
France     5014
Germany    2509
Spain      2477
Name: count, dtype: int64
Gender
Male      5457
Female    4543
Name: count, dtype: int64
```

Setelah dilakukan penghapusan fitur sebelumnya, kami lakukan pengecekan apakah fitur yang tersedia memiliki nilai kosong di dalamnya **tidak ditemukan adanya nilai kosong** dan **semua nilai pada fitur kategorikal juga relevan** terhadap nama kolomnya.

Rangkuman Stage 2: Data Prepocessing

Rakamin Academy

Feature Encoding

```
cats_updated = ['Geography', 'Gender']

for col in cats_updated:
  print(f'value counts of column {col}')
  print(df_bank[col].value_counts())
  print('---'*10, '\n')
```

```
value counts of column Geography
Geography
France     5014
Germany    2509
Spain      2477
Name: count, dtype: int64
------------------------------

value counts of column Gender
Gender
1    5457
0    4543
Name: count, dtype: int64
------------------------------
```

```
# convert gender feature from categorical into numerical by using label encoding

mapping_gender = {
    'Female' : 0,
    'Male' : 1
}

df_bank['Gender'] = df_bank['Gender'].map(mapping_gender)
```

```
# convert Geography feature from categorical into  numerical by using one-hot encoding

from sklearn.preprocessing import OneHotEncoder

## Converting type of columns to category

df_bank['Geography'] = df_bank['Geography'].astype('category')


## Assigning numerical values and storing it in another columns

df_bank['Geo_new'] = df_bank['Geography'].cat.codes

## Create an instance of One-hot-encoder

enc = OneHotEncoder()

## Passing encoded columns

enc_data = pd.DataFrame(enc.fit_transform(
    df_bank[['Geo_new']]).toarray())

## Merge with main

df_bank = df_bank.join(enc_data)

## rename the column

df_bank = df_bank.rename(columns={0 : "is_France", 1 : "is_Germany", 2 : "is_Spain"})

## drop irrelevant column

df_bank = df_bank.drop(columns = ['Geography', 'Geo_new'])

## show the result

df_bank.head(1)
```

	CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	is_France	is_Germany	is_Spain
0	619	0	42	2	0.0	1	1	1	101348.88	1	1.0	0.0	0.0

Pada dataset kami, terdapat **2 fitur kategorikal yang perlu dikonversi** menjadi numerikal yaitu **fitur geografi dan fitur gender**. Masing-masing fitur tersebut kami tangani dengan pendekatan yang berbeda, **fitur gender menggunakan label encoding** sedangkan pada **fitur geografi menggunakan one-hot encoding**.

Kami akan melakukan **cek kembali apakah salah satu hasil fitur tersebut memiliki pengaruh signifikan terhadap target serta apakah ada indikasi multikolinearitas** antar fiturnya. Kami gunakan **metode test statistik chi2 dan nilai VIF (Variance Inflation Factor)**.

Rangkuman Stage 2: Data Prepocessing

Rakamin Academy

Feature Encoding

```
# Checking the significant of new feature to the target by using chi2 statistic test (categorical vs categorical)

X = df_bank.drop(columns = ['CreditScore', 'Age', 'Tenure', 'NumOfProducts', 'Balance', 'EstimatedSalary', 'Exited'])
y = df_bank['Exited']
print(X.columns)
chi2(X,y)
```

```
array([7.01557451e-13, 6.98496209e-01, 1.56803624e-27, 1.25300579e-13,
       5.81457176e-51, 4.92250487e-06]))
```

Setelah dilakukan uji statistik chi2, **hanya fitur 'HasCrCard' yang tidak memiliki pengaruh signifikan** ($p > 0.05$) terhadap target, sedangkan **hasil ketiga fitur dari proses encoding semuanya memiliki pengaruh signifikan** ($p < 0.05$) terhadap target.

```
### Checking multicollinearity by VIF

from statsmodels.stats.outliers_influence import variance_inflation_factor

vif_select = df_bank
vif_data = pd.DataFrame()
vif_data["feature"] = vif_select.columns

# calculating VIF for each feature
vif_data["VIF"] = [variance_inflation_factor(vif_select.values, i)
                   for i in range(len(vif_select.columns))]

print(vif_data)
```

	feature	VIF
0	CreditScore	1.001643
1	Gender	1.013210
2	Age	1.110478
3	Tenure	1.002156
4	Balance	1.339246
5	NumOfProducts	1.123001
6	HasCrCard	1.001617
7	IsActiveMember	1.046623
8	EstimatedSalary	1.001048
9	Exited	1.177569
10	is_France	41.366649
11	is_Germany	22.804330
12	is_Spain	21.087015

Setelah dilakukan pengecekan nilai VIF, ternyata **hasil ketiga fitur dari proses encoding semuanya memilki nilai VIF > 5** yang artinya ada **indikasi multikolinearitas**, lakukan penggabungan fitur is_Germany dengan is_Spain menjadi not_France **apabila hasil model evaluasinya mengalami overfitting**.

Rangkuman Stage 2: Data Prepocessing

Rakamin Academy

Feature Selection –
Jika model overfitting

Manual - Feature Selection

```
# backing plan for feature selection if the default model is overfit

df_bank2 = df_bank.copy()
df_bank2 = df_bank2.drop(columns = ['HasCrCard', 'Tenure', 'EstimatedSalary', 'is_Spain'])
```

Automatic - Using SelectKBest for Feature Selection if the model is overfit

```
X = df_bank[['CreditScore', 'Age', 'Gender', 'Tenure', 'Balance', 'NumOfProducts', 'HasCrCard', 'IsActiveMember', 'EstimatedSalary', 'is_France', 'is_Germany', 'is_Spain']]
y = df_bank['Exited']
from sklearn.feature_selection import SelectKBest, mutual_info_classif

X_new = SelectKBest(mutual_info_classif, k=10).fit(X, y)
X_new
```

```
SelectKBest(score_func=<function mutual_info_classif at 0x000001F731C3FCE0>)
```

```
X_new.get_feature_names_out()
```

```
array(['CreditScore', 'Age', 'Gender', 'Tenure', 'Balance',
       'NumOfProducts', 'IsActiveMember', 'is_France', 'is_Germany',
       'is_Spain'], dtype=object)
```

Apabila hasil model evaluasi mengalami overfit, maka akan kami lakukan feature selection dengan 2 cara, secara otomatis menggunakan **library SelectKBest atau manual dengan menghapus fitur yang tidak berpengaruh signifikan dengan uji chi2 serta fitur yang memiliki nilai VIF tinggi di atas 5**

Rangkuman Stage 2: Data Prepocessing

Rakamin Academy

Feature Engineering –
Jika model underfitting

```
## create a copy

df_bank_new = df_bank.copy()

# balance per salary

df_bank_new['BalanceperSalary'] = df_bank_new['Balance'] / df_bank_new['EstimatedSalary']
```

Kami memutuskan akan menggunakan hasil dari fitur engineering di atas **apabila hasil model evaluasi mengalami underfitting**.

Rangkuman Stage 2: Data Prepocessing

Rakamin Academy

Handling Outlier (Untuk Data Awal Modelling)

```
# Split the data into training and testing with the proportion of 70:30

X = df_bank.drop(columns=['Exited'])
y = df_bank[['Exited']]

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Kami menghapus nilai outlier dengan **metode Z-score**. Data yang kami gunakan untuk menghapus outlier adalah data training agar tidak terjadi data leaking terhadap data testing. Diperoleh data training setelah dilakukan penghapusan outlier sebanyak 6906 baris (berkurang 1.3%).

```
# Removing outliers using Z-Score

from scipy import stats

print(f'Jumlah baris sebelum memfilter outlier {len(data_train)}')


for col in ['CreditScore', 'Age']:
    zscore = np.abs(stats.zscore(data_train[col]))
    filtered_entries = (zscore < 3)

data_train = data_train[filtered_entries]

print(f'Jumlah baris setelah memfilter outlier: {len(data_train)}')
```

```
Jumlah baris sebelum memfilter outlier 7000
Jumlah baris setelah memfilter outlier: 6906
```

Features Transformation

```
# Standardization

from sklearn.preprocessing import StandardScaler
ss = StandardScaler()

numerical_features = X.columns.to_list()
for n in numerical_features:
  scaler = ss.fit(X_train[[n]])
  X_train[n] = scaler.transform(X_train[[n]])
  X_test[n] = scaler.transform(X_test[[n]])
```

Kami melakukan scaling data training dan data testing dengan **metode standarisasi** agar semua fitur yang ada memiliki bentuk distribusi mendekati normal dan jarak nilai min-max antar feature tidak terlalu jauh.

Rangkuman Stage 2: Data Prepocessing

Rakamin Academy

Handling Class Imbalance

```
# checking the total amount of each label

y_train.value_counts()
```

```
Exited
0    5459
1    1447
Name: count, dtype: int64
```

```
# using undersamping for majority class with the proportion feature target is 70:30

from imblearn import under_sampling
X_under, y_under = under_sampling.RandomUnderSampler(random_state = 42, sampling_strategy = 0.428).fit_resample(X_train, y_tra
```

```
y_under.value_counts()
```

```
Exited
0    3380
1    1447
Name: count, dtype: int64
```

Kami menggunakan **metode undersampling** untuk menghandle data yang imbalance dengan **proporsi 70:30**. Namun penggunaan data tersebut kami gunakan sebagai **langkah alternatif terakhir karena pada metriks model evaluasi yang akan kami gunakan adalah metriks F1 - Score yang lebih robust terhadap data imbalance**.

Rangkuman Stage 3: Machine Learning Modelling

Rakamin Academy

Hasil Modelling

Data cross validation	
urutan skor	F1 score, recall, ROC-AUC

Default Parameter

	Undersample (default parameter)		
	Skor training	Skor testing	Kesimpulan
Logistic Regression	[0.47, 0.38, 0.77]	[0.47, 0.37, 0.77]	roc auc bagus tapi metirks lain underfitting
KNN	[0.71, 0.63, 0.92]	[0.58, 0.50, 0.79]	skor lebih baik daripada logistic regression namun overfitting
SVM	[0.66, 0.56, 0.88]	[0.63, 0.53, 0.84	skor lebih baik daripada logistic regression namun underfitting
Decision Tree	[1, 1, 1]	[0.57, 0.58, 0.7]	overfit
Random Forest	[1, 1, 1]	[0.65, 0.57, 0.86]	overfit
Adaboost	[0.66, 0.59, 0.86]	[0.65, 0.58, 0.85]	skor lebih baik namun metriks utama masih underfitting
XGBoost	[0.93, 0.9, 0.99]	[0.65, 0.59, 0.85]	overfit

Hyperparameter Tuning

	Undersample (Hyperparameter Tuning)		
	Skor training	Skor testing	Kesimpulan
Logistic Regression	[0.47, 0.38, 0.77]	[0.47, 0.37, 0.77]	skor lebih baik daripada default namun metriks utama masih underfitting, hanya ROC-AUC yang memiliki skor bagus
KNN	[0.57, 0.46, 0.86]	[0.57, 0.42, 0.82]	skor tidak lebih baik daripada default, hanya ROC-AUC yang memiliki skor bagus dan semua metriks sudah tidak overfitting
SVM	[0.7, 0.6, 0.9]	[0.64, 0.55, 0.84]	skor lebih baik daripada default namun metriks utama masih underfitting, hanya ROC-AUC yang memiliki skor bagus
Decision Tree	[0.65, 0.58, 0.85]	[0.64, 0.56, 0.83]	skor lebih baik daripada default namun metriks utama masih underfitting, hanya ROC-AUC yang memiliki skor bagus
Random Forest	[0.64, 0.72, 0.82]	[0.64, 0.72, 0.82]	skor lebih baik daripada default namun metriks utama masih underfitting, tetapi kedua metriks supporting sudah mendapatkan skor terbaik
Adaboost	[0.66, 0.59, 0.86]	[0.65, 0.58, 0.85]	tidak ada perbedaan antara sebelum dan sesudah hyperparameter tuning, hanya ROC-AUC yang memiliki skor bagus
XGBoost	[0.63, 0.53, 0.85]	[0.62, 0.53, 0.85]	skor tidak lebih baik daripada default, hanya ROC-AUC yang memiliki skor bagus dan semua metriks sudah tidak overfitting

Di modelling ini, kami menggunakan F1 Score sebagai metrics utama, recall dan roc-auc sebagai metrics sekunder yang nilainya sudah diatas 0.7 atau 70%. Kami mencoba 7 algoritma untuk machine learningnya. Di default parameter, metrics utama kebanyakan pada overfitting tapi ada juga yang underfitting. Maka dari itu kita perlu ada evaluasi pada ketiga metrics tersebut terutama pada F1 score dengan melakukan hyperparameter tuning. Dan hasil dari hyperparameter tuning menunjukan **hasil model terbaik** pada algoritma **random forest** walaupun metrics utama masih underfitting tetapi didukung dengan metrics sekundernya karena random forest adalah salah satu algoritma klasifikasi akurasi cenderung lebih tinggi daripada algoritma klasifikasi lainnya.

Rangkuman Stage 3: Machine Learning Modelling

Rakamin Academy

Hasil Modelling Random Forest setelah hyperparameter Tuning

```
Accuracy (Train Set): 0.76
Accuracy (Test Set): 0.76
Precision (Train Set): 0.58
Precision (Test Set): 0.43
Recall (Train Set): 0.70
Recall (Test Set): 0.71
F1-Score (Train Set): 0.64
F1-Score (Test Set): 0.54
roc_auc (train-proba): 0.82
roc_auc (test-proba): 0.82

F1-Score (crossval train): 0.64
F1-Score (crossval test): 0.64
recall (crossval train): 0.72
recall (crossval test): 0.72
ROC_AUC (crossval train): 0.82
ROC_AUC (crossval test): 0.82
Precision (crossval train): 0.58
Precision (crossval test): 0.58
Accuracy (crossval train): 0.76
Accuracy (crossval test): 0.76
```

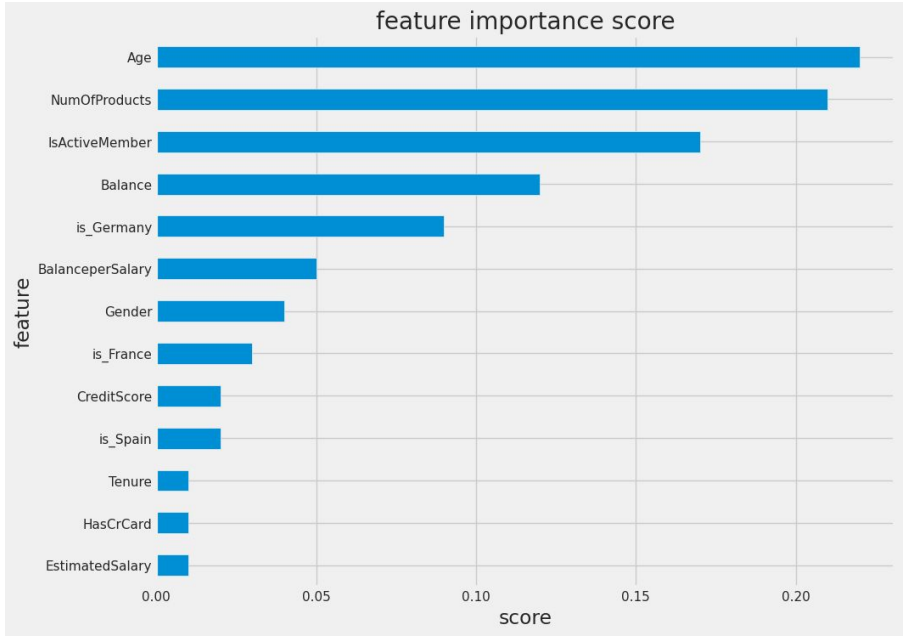
Setelah hasilnya di hyperparameter tuning, train-test set pada metric f1 score = overfitting, recall = underfitting dan roc-auc = best-fitting. Sedangkan pada cross validation semua metric miliki skor best-fitting setelah di hyperparameter tuning. Kenapa begitu? Karena ada beberapa parameter tuning yang kami rubah seperti min_samples_split nilai defaultnya 2, kami rubah antara nilai 2 sampai 350; max_samples nilai defaultnya 0, kami rubah antara 0.01 sampai 0.1; dan min_weight_fraction_leaf nilai defaultnya 0, kami rubah nilainya antara 0 sampai 0.15 supaya skor modelnya menjadi lebih baik daripada parameter defaultnya.

Rangkuman Stage 3: Machine Learning Modelling

Rakamin Academy

Feature Importance, Shap Values and Business insight

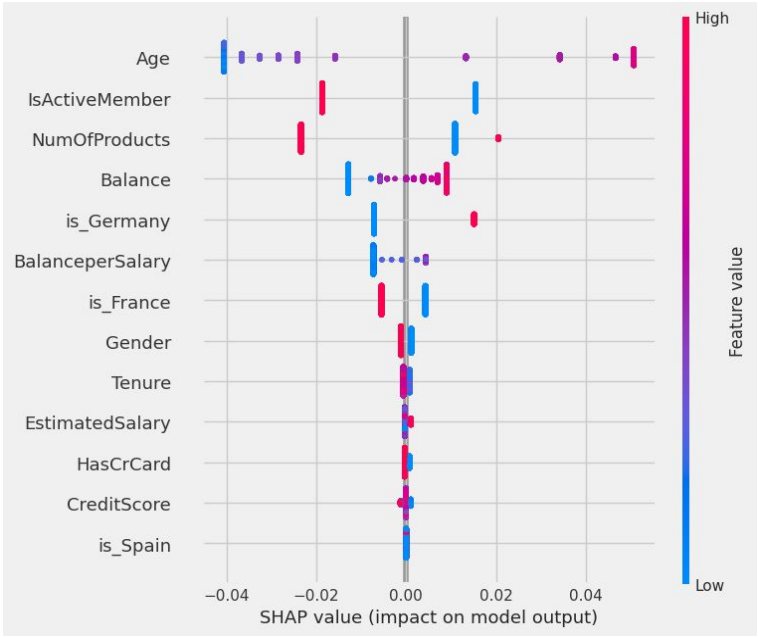
Fesature Importance



Business Insight

Di feature importance, kami akan pilih top 5 teratas untuk bisnisnya dan kami nanti akan menghilangkan top 5 bawah untuk pemakaian model machine learning selanjutnya

Shap Values



- Sisa yang feature lainnya seperti EstimatedSalary, HasCrCard, CreditScore dan is_Spain tidak bisa didefinisikan.

Business Insight

Di shap values yang dihasilkan:

- Age: Semakin tua usianya maka nasabah berpotensi churn.
- IsActiveMember: Nasabah yang tidak aktif berpotensi churn.
- NumOfProducts: Hanya sedikit produk yang dimiliki oleh nasabah berpotensi churn.
- Balance: Semakin besar saldo maka nasabah berpotensi churn.
- Is_Germany: Banyak orang asli jerman berpotensi churn.
- BalanceperSalary: tidak terlalu berdampak.
- Is_France: Banyak orang yang bukan asli perancis berpotensi churn.
- Gender: Jenis kelamin Wanita berpotensi churn.
- Tenure: Semakin cepat tenggat waktu pembayaran maka nasabah berpotensi churn.

Rangkuman Stage 3: Machine Learning Modelling

Rakamin Academy

Business Recomendation

Business Recommendation yang akan kami berikan:

- Memberikan program loyalitas reward seperti poin jika pelanggan aktif yang berasal dari generasi Millenial dan Gen X yang memiliki saldo menengah dan tinggi, sering melakukan aktivitas transaksi.
- Memberikan notifikasi push melalui pesan telepon kepada pelanggan tidak aktif asal Jerman dengan isi notifikasi mempromosikan diskon belanja di merchant tertentu jika membeli produk mewah tertentu.
- Meningkatkan kualitas produk bank sehingga nasabah tertarik untuk membeli produk lainnya.