Modeling and Predicting Churn

Data Science Assignment Chapter 5

Mardhani Dwi Novianto - DS03239

Kompi K

https://colab.research.google.com/drive/1Xo1G0ILM8SfK0hZy7_bTrnYxRXXNvGuU?usp=sharing

Models

Data yang digunakan untuk modeling

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100409 entries, 0 to 100408
Data columns (total 9 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   customer_id           100409 non-null  object
 1   payment_type          100409 non-null  object
 2   total_order           100409 non-null  int64
 3   total_installments    100409 non-null  int64
 4   total_payment         100409 non-null  int64
 5   first_segmentation    100409 non-null  object
 6   basket_size           100409 non-null  int64
 7   monetary_segmentation 100409 non-null  object
 8   segmentation_churn    100409 non-null  object
dtypes: int64(4), object(5)
memory usage: 6.9+ MB
```

Model yang dibuat

```
==========Comparison Base Modeling==========

Accuracy model Logistic Regression    : 0.5982969823722737

Accuracy model K Nearest Neighbor     : 0.7553032566477442

Accuracy model Decision Tree          : 0.8317398665471567

Accuracy model Random Forest          : 0.8403047505228562

Accuracy model Naive Bayes            : 0.4027985260432228

Accuracy model AdaBoost               : 0.8565879892440992
```

Dibuat 6 model Klasifikasi

1. Logistic Regression
2. K Nearest Neighbor
3. Decision Tree
4. Random Forest
5. Naive Bayes
6. AdaBoost

Diperoleh Akurasi tertinggi yaitu Model AdaBoost dengan akurasi 0.8565879892440992 atau 85.65%

Increase Performance

Add Features

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 100409 entries, 0 to 100408
Data columns (total 13 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   customer_id               100409 non-null  object
 1   payment_type              100409 non-null  object
 2   total_order               100409 non-null  int64
 3   total_installments        100409 non-null  int64
 4   total_payment             100409 non-null  int64
 5   first_segmentation        100409 non-null  object
 6   basket_size               100409 non-null  int64
 7   monetary_segmentation     100409 non-null  object
 8   segmentation_churn        100409 non-null  object
 9   diff_approved_purchase    100409 non-null  int64
 10  diff_carrier_approved     100409 non-null  int64
 11  diff_delivered_carrier    100409 non-null  int64
 12  diff_estimated_delivered  100409 non-null  int64
dtypes: int64(8), object(5)
memory usage: 10.7+ MB
```

Membandingkan model 1 dan 2

```
=====Comparison Model 1 dan 2=====

Accuracy model Logistic Regression
Model 1 : 0.5982969823722737
Model 2 : 0.6116422667065033

Accuracy model K Nearest Neighbor
Model 1 : 0.7553032566477442
Model 2 : 0.5934169903396076

Accuracy model Decision Tree
Model 1 : 0.8328353749626531
Model 2 : 0.8338810875410816

Accuracy model Random Forest
Model 1 : 0.8399063838263121
Model 2 : 0.8652524648939348

Accuracy model Naive Bayes
Model 1 : 0.4027985260432228
Model 2 : 0.40947116821033763

Accuracy model AdaBoost
Model 1 : 0.8565879892440992
Model 2 : 0.8631112439000099
```

Diperoleh Akurasi tertinggi yaitu Model ke-2 dari model Random Forest dengan akurasi 0.8652524648939348 atau 86.52%

Increase Performance

Hyperparameter Tuning

Untuk model Decision Tree

Menggunakan max_depth dan max_leaf_nodes

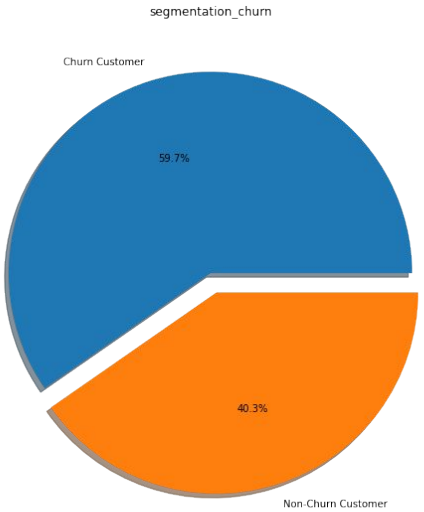
Untuk model Random Forest

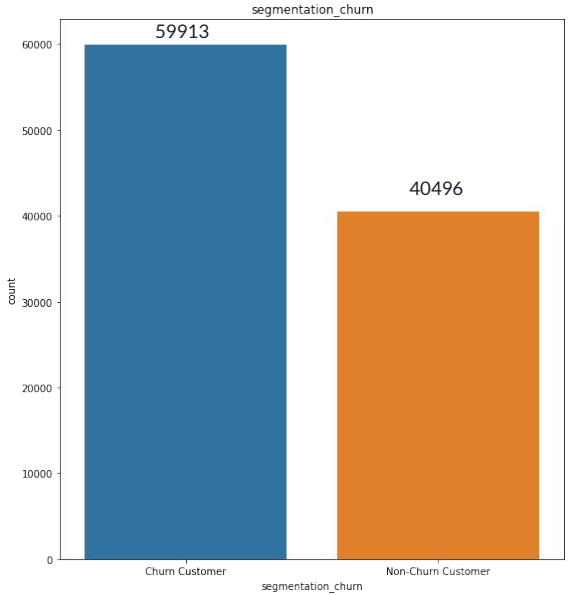
Menggunakan n_estimators dan max_features

Diperoleh Akurasi tertinggi yaitu Model ke-2 dari model **Random Forest** dengan akurasi 0.8652524648939348 atau 86.52%

```
==================================Comparison Model==================================

------------------------------------------------------------------------------------
Accuracy model Logistic Regression
1. Base Modeling 1                                          : 0.5982969823722737
2. Modeling 2 (After add features)                          : 0.6116422667065033
------------------------------------------------------------------------------------

Accuracy model K Nearest Neighbor
1. Base Modeling 1                                          : 0.7553032566477442
2. Modeling 2 (After add features)                          : 0.5934169903396076
------------------------------------------------------------------------------------

Accuracy model Decision Tree
1. Base Modeling 1                                          : 0.8328353749626531
2. Modeling 2 (After add features)                          : 0.8338810875410816
3. Modeling After Hyperparameter Tuning using max_depth     : 0.8565879892440992
4. Modeling After Hyperparameter Tuning using max_leaf_nodes : 0.8617667562991734
------------------------------------------------------------------------------------

Accuracy model Random Forest
1. Base Modeling 1                                          : 0.8399063838263121
2. Modeling 2 (After add features)                          : 0.8652524648939348
3. Modeling After Hyperparameter Tuning using n_estimators  : 0.8500647345881884
4. Modeling After Hyperparameter Tuning using max_features  : 0.8637585897818942
------------------------------------------------------------------------------------

Accuracy model Naive Bayes
1. Base Modeling 1                                          : 0.4027985260432228
2. Modeling 2 (After add features)                          : 0.40947116821033763
------------------------------------------------------------------------------------

Accuracy model AdaBoost
1. Base Modeling 1                                          : 0.8565879892440992
2. Modeling 2 (After add features)                          : 0.8631112439000099
------------------------------------------------------------------------------------
```

Segmentation





Model Random Forest sebagai model yang terbaik dalam percobaan ini sehingga yang dapat digunakan untuk **modeling dan predicting Churn.**

Berdasarkan model yang dibangun, diperoleh segmentasi Churn sebagai berikut :

Segmentasi Churn Customer, sebanyak 59.7% atau 59913 orang, sementara sisanya 40.3% atau 40496 orang termasuk dalam segmentasi Non-Churn Customer

Solution

Berdasarkan Segmentasi Churn yang diperoleh,
Dengan meninjau fitur-fitur yang digunakan untuk modeling,
Maka menurut saya, terdapat beberapa cara yang dapat digunakan untuk meminimalisir jumlah pelanggan yang melakukan churn :

1. Jarak waktu antara produk yang dipesan dengan pesanan yang dikonfirmasi, dapat dipercepat, dengan kata lain pelayanan harus ditingkatkan.
2. Jarak waktu antara pesanan dikonfirmasi dengan pesanan dikirim atau sampai, lebih baik dipersingkat atau dipercepat sehingga pelanggan tidak menunggu lama, dan segera mendapatkan update terbaru.
3. Dapat diupayakan adanya promo/reward atau fitur-fitur baru yang dapat memberikan kepuasan terhadap pelanggan. Dengan adanya kepuasan dari pelanggan, dapat meningkatkan pembelian dan berdampak pada rate atau basket size yang meningkat, juga secara tidak langsung dapat menambah profit perusahaan.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 100409 entries, 0 to 100408
Data columns (total 13 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   customer_id               100409 non-null  object
 1   payment_type              100409 non-null  object
 2   total_order               100409 non-null  int64
 3   total_installments        100409 non-null  int64
 4   total_payment             100409 non-null  int64
 5   first_segmentation        100409 non-null  object
 6   basket_size               100409 non-null  int64
 7   monetary_segmentation     100409 non-null  object
 8   segmentation_churn        100409 non-null  object
 9   diff_approved_purchase    100409 non-null  int64
 10  diff_carrier_approved     100409 non-null  int64
 11  diff_delivered_carrier    100409 non-null  int64
 12  diff_estimated_delivered  100409 non-null  int64
dtypes: int64(8), object(5)
memory usage: 10.7+ MB
```