ADRISTI SHAFIYA

VISUALIZATION & MACHINE LEARNING

CHALLENGE 2:

TELECO CUSTOMER CHURN

OUTLINE

DATA PREPARATION

EXPLORATORY DATA ANALYSIS

- Univariate Analysis
- Multivariate Analysis
- Categorical Analysis
- Deep Dive Question

DATA PREPROCESSING

- Handling Outlier
- Encoding Data Train
- Normalization & Standardization

MACHINE LEARNING MODEL

- Logistic Regression
- Decision Tree
- K – Nearest Neighbor (KNN)

RECOMMENDATION

WELCOME

Business Problem

Terjadi customer churn yang bisa menyebabkan berkurangnya pendapatan bagi perusahaan telekomunikasi sehingga penting untuk ditangani.

Data Preparation

Observation

- Data memiliki 19 kolom dan 4250 baris
- Tipe data pada semua kolom sudah sesuai
- Terdapat 5 kolom kategorikal dan 15 kolom numerik

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4250 entries, 0 to 4249
Data columns (total 20 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   state                           4250 non-null   object
 1   account_length                  4250 non-null   int64
 2   area_code                       4250 non-null   object
 3   international_plan              4250 non-null   object
 4   voice_mail_plan                 4250 non-null   object
 5   number_vmail_messages           4250 non-null   int64
 6   total_day_minutes               4250 non-null   float64
 7   total_day_calls                 4250 non-null   int64
 8   total_day_charge                4250 non-null   float64
 9   total_eve_minutes               4250 non-null   float64
 10  total_eve_calls                 4250 non-null   int64
 11  total_eve_charge                4250 non-null   float64
 12  total_night_minutes             4250 non-null   float64
 13  total_night_calls               4250 non-null   int64
 14  total_night_charge              4250 non-null   float64
 15  total_intl_minutes              4250 non-null   float64
 16  total_intl_calls                4250 non-null   int64
 17  total_intl_charge               4250 non-null   float64
 18  number_customer_service_calls   4250 non-null   int64
 19  churn                           4250 non-null   object
dtypes: float64(8), int64(7), object(5)
memory usage: 664.2+ KB
```

DUPLICATED VALUE

- Tidak ada duplicated value

```
#Checking duplicated value
df.duplicated().sum()
```

```
0
```

```
#Checking missing value
df.isnull().sum()
```

```
state                           0
account_length                  0
area_code                       0
international_plan              0
voice_mail_plan                 0
number_vmail_messages           0
total_day_minutes               0
total_day_calls                 0
total_day_charge                0
total_eve_minutes               0
total_eve_calls                 0
total_eve_charge                0
total_night_minutes             0
total_night_calls               0
total_night_charge              0
total_intl_minutes              0
total_intl_calls                0
total_intl_charge               0
number_customer_service_calls   0
churn                           0
dtype: int64
```

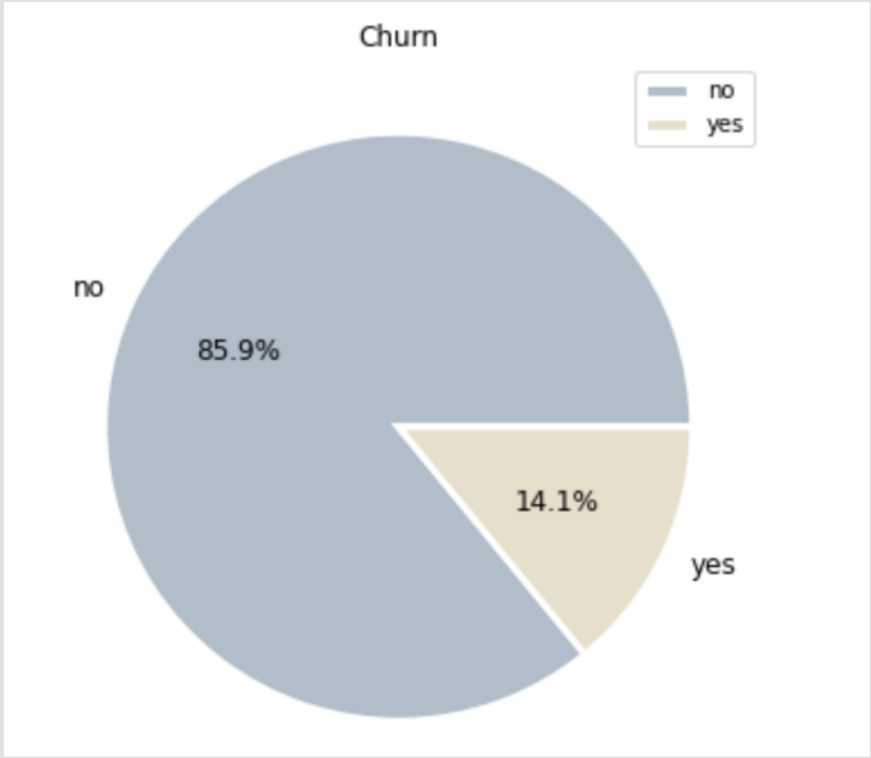
MISSING VALUE

- Tidak ada missing value

Exploratory Data Analysis (EDA)

OBSERVATION

- Dari total 4250, terdapat 85,9 % atau 3652 pelanggan tidak churn
- Terdapat 14,1 % atau 598 pelanggan memutuskan untuk churn
- Pada target variabel yaitu Churn terjadi Imbalance Data

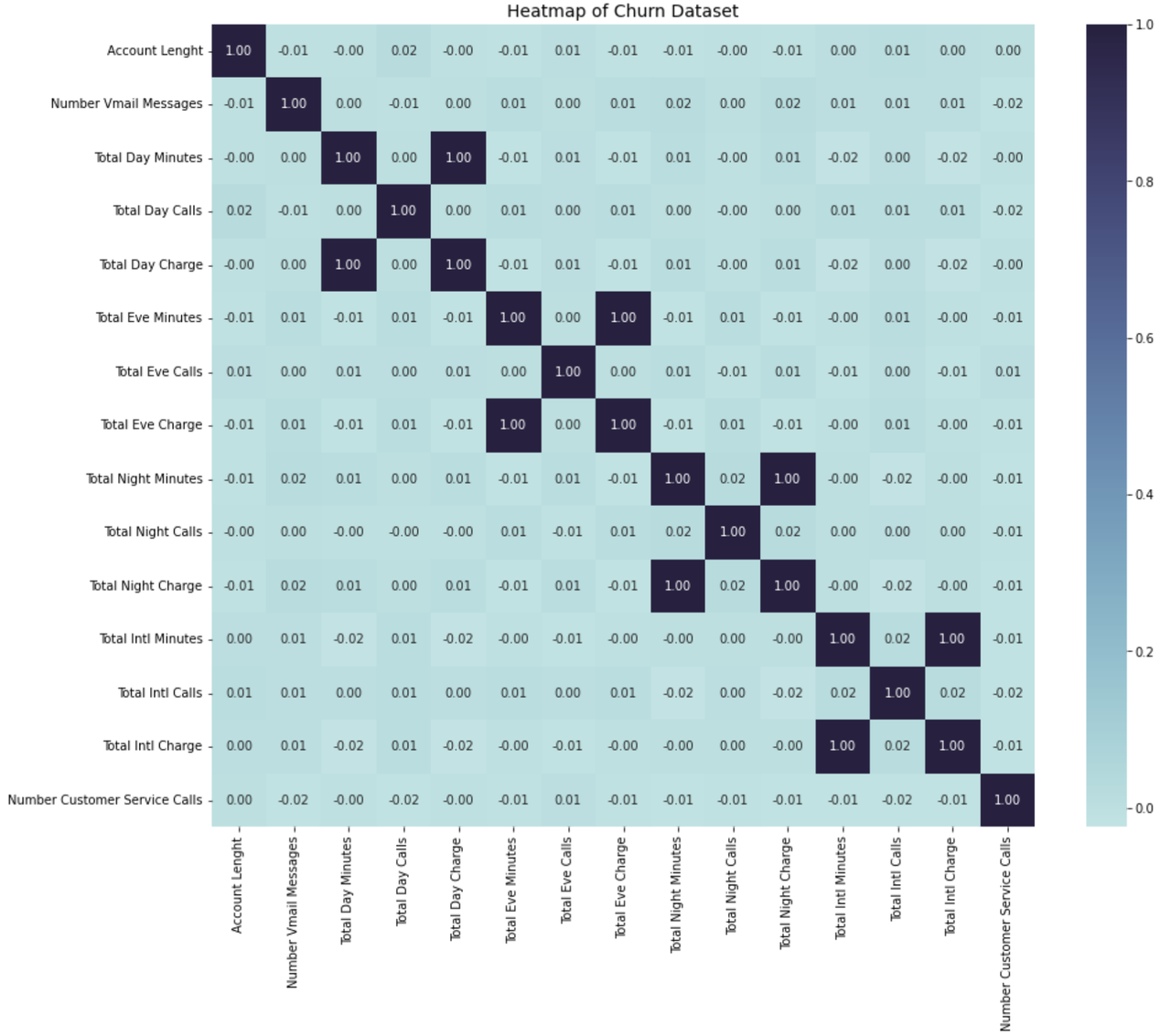


Outlier

Jumlah Outlier Berdasarkan Kolom

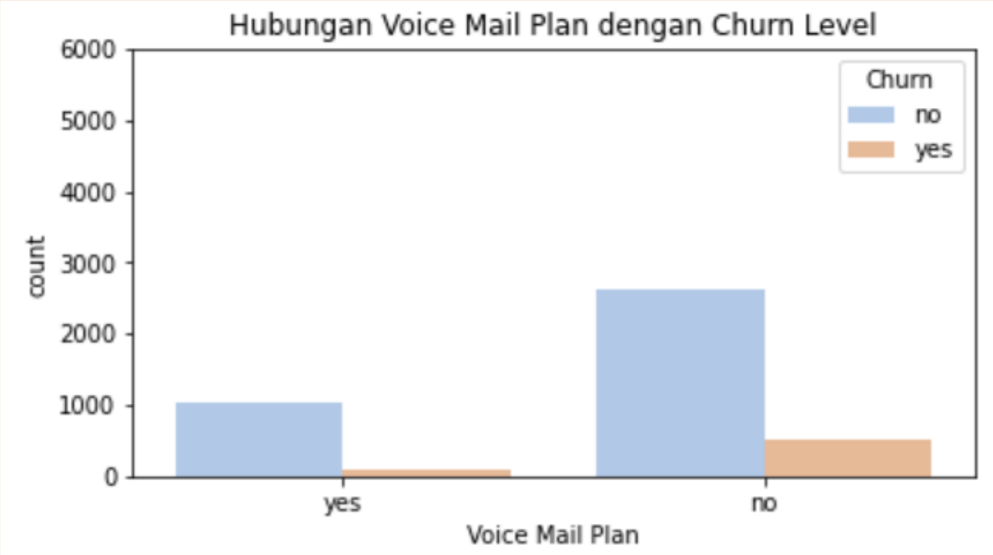
1. Account Length = 20
2. Number Vmail = 86
3. Total Day Minutes = 25
4. Total Day Call = 28
5. Total Day Charge = 26
6. Total Eve Minute = 34
7. Total Eve Call = 24
8. Total Eve Charge = 34
9. Total Night Minute = 37
10. Total Night Call = 33
11. Total Night Charge = 37
12. Total Intl Minute = 62
13. Total Intl Call = 100
14. Total Intl Charge = 62
15. Number Cust. Service Call = 335

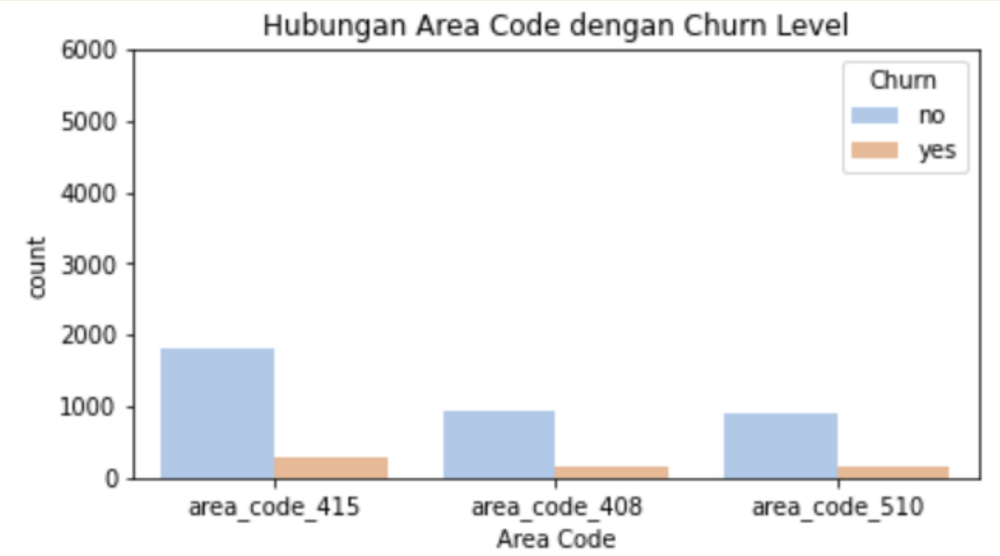
HEAT MAP CORRELATION



– Minutes dan Charge memiliki nilai yang berkorelasi dikarenakan semakin tinggi menit telepon yang dilakukan maka semakin tinggi pula chargenya

– Kali ini akan didrop untuk Total day minutes, total eve minutes, total night minutes, dan total intl minutes

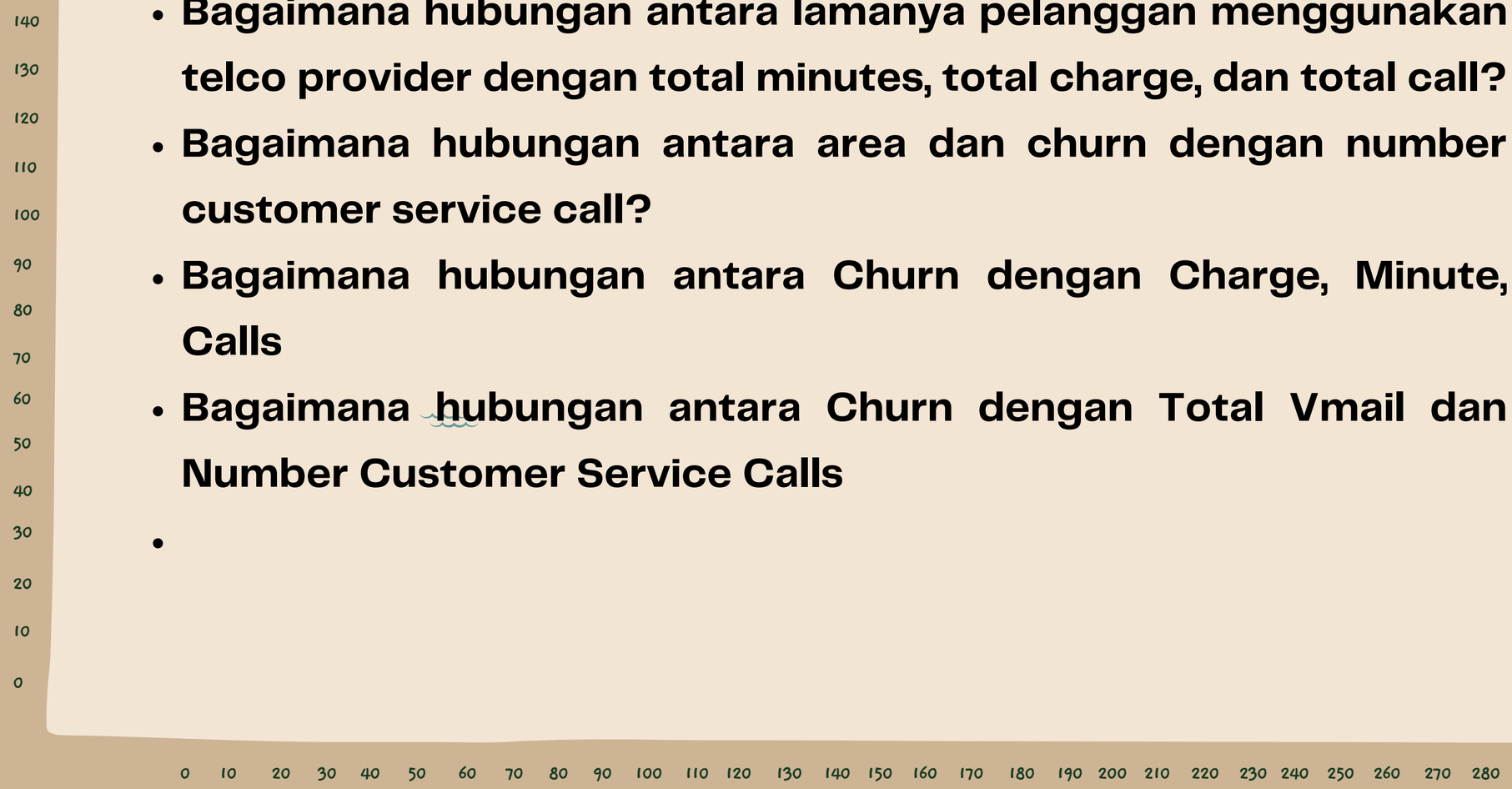




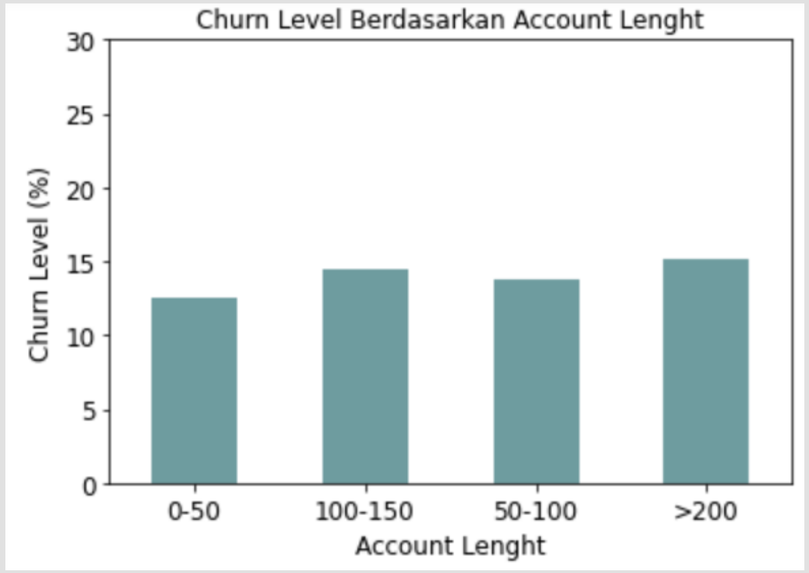
1. Pelanggan yang memiliki kecenderungan tidak churrn banyak berasal dari area_code_415 dan tidak memiliki international plan dan voice mall plan

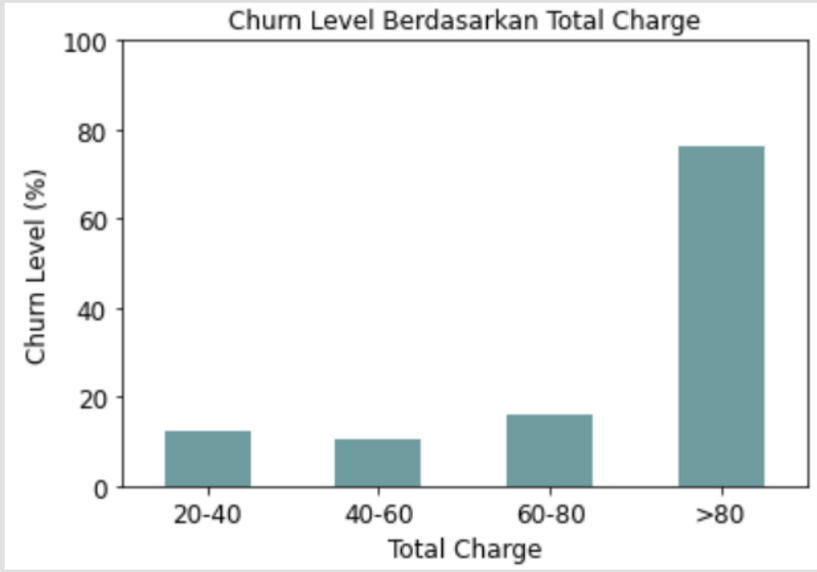
DEEP DIVE QUESTION

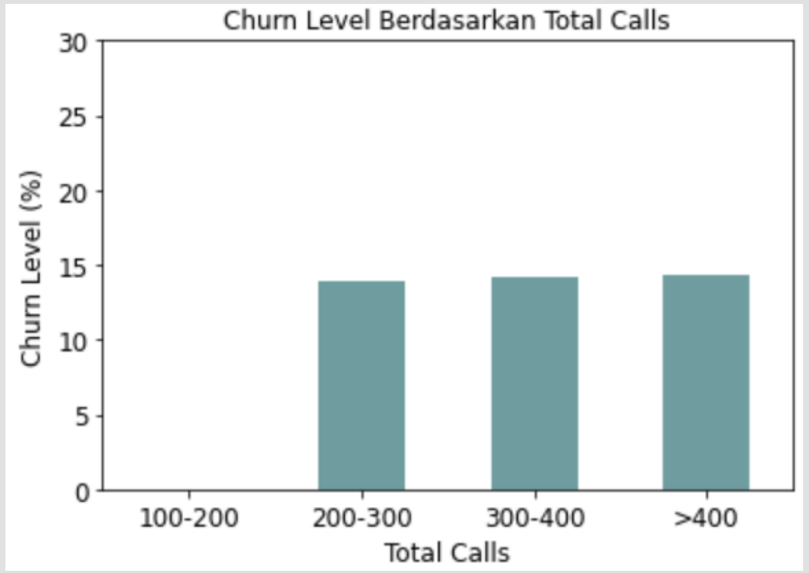
- Bagaimana hubungan antara lamanya pelanggan menggunakan telco provider dengan total minutes, total charge, dan total call?
- Bagaimana hubungan antara area dan churn dengan number customer service call?
- Bagaimana hubungan antara Churn dengan Charge, Minute, Calls
- Bagaimana hubungan antara Churn dengan Total Vmail dan Number Customer Service Calls
-

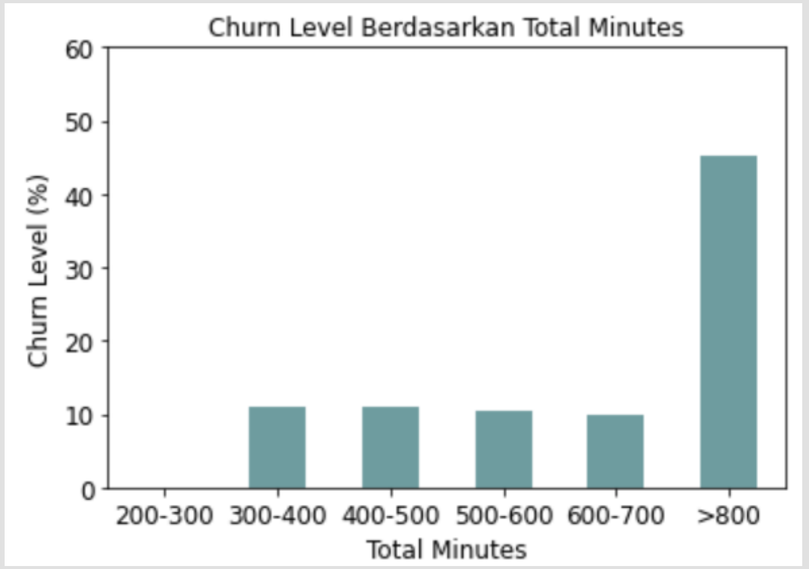


Hubungan antara churn level dengan lamanya pelanggan menggunakan telco provider, total minutes, total charge, dan total call



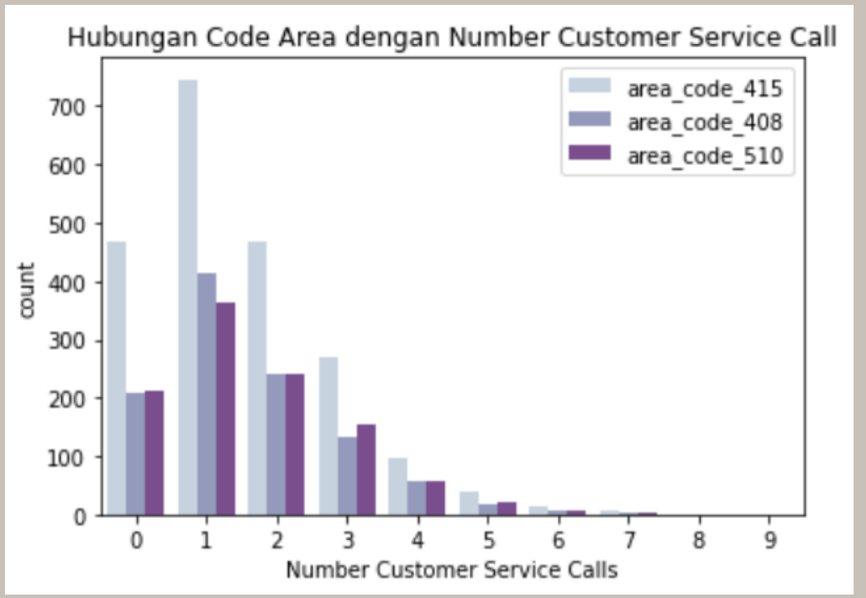


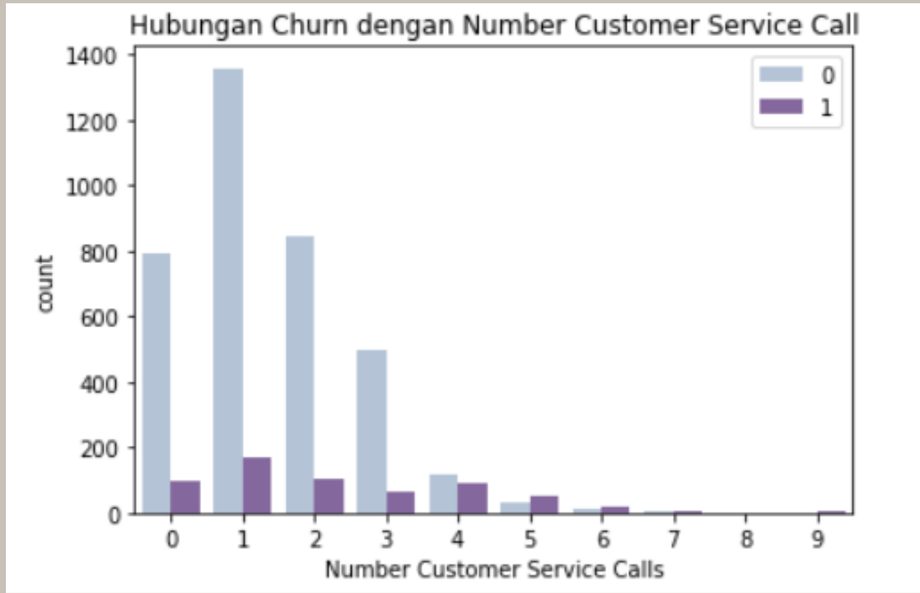




Hubungan antara area dengan number customer service call

1. Kode area 415, 408, dan 510 banyak yang menghubungi customer service sebanyak 0,1, , 2, dan 3 kali

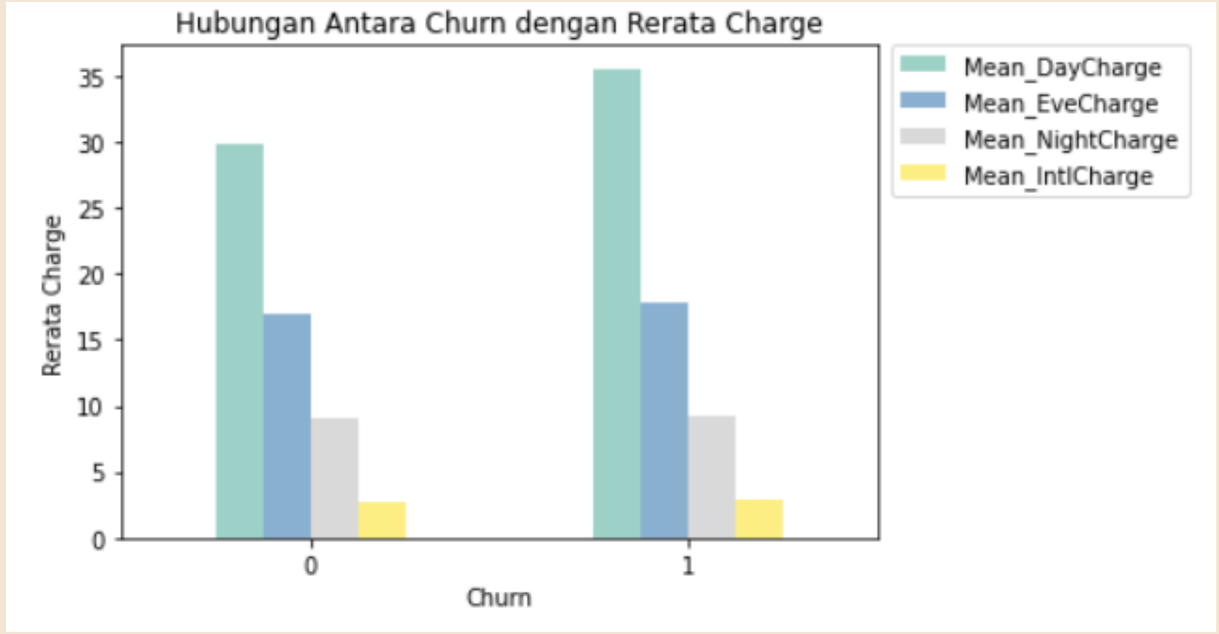


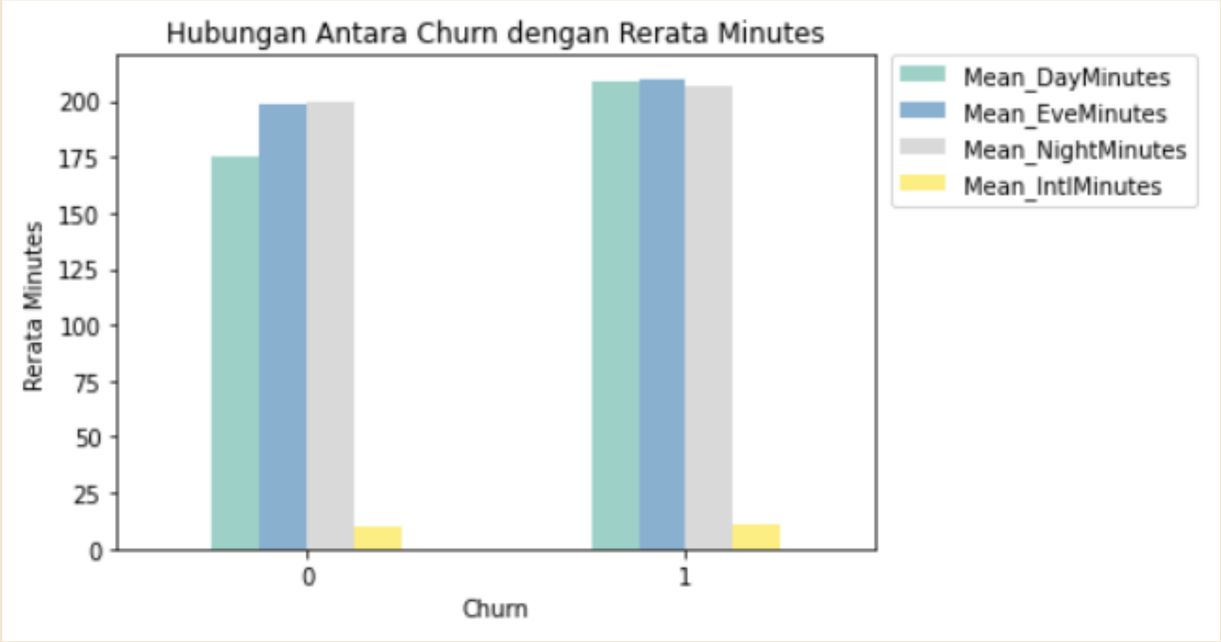


1. Dapat dikatakan bahwa pelanggan jarang untuk menghubungi customer service call hanya berkisar 0 - 3 kali saja.

Hubungan antara Churn dengan Charge, Minute, Calls

- Pelanggan yang memiliki rerata charge pada pagi memiliki kecenderungan untuk churn.
- Rerata charge pada sore hari antara pelanggan yang cenderung churn dan tidak churn memiliki nilai yang hampir sama

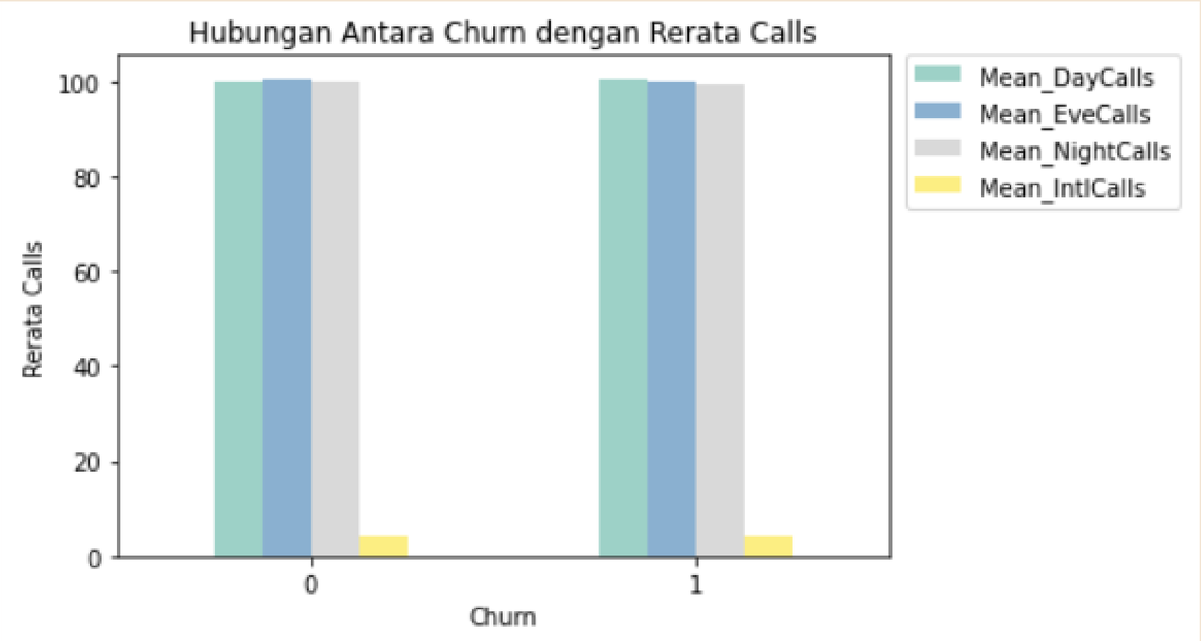




- Pelanggan yang memiliki rerata total minutes pada pagi, sore, dan malam hari tertinggi memiliki kecenderungan untuk churrn

Hubungan antara Churn dengan Charge, Minute, Calls

- Antara pelanggan yang cenderung churn dan tidak churn memiliki rerata total call pada pagi, sore, dan malam hari yang hampir sama



Data Preprocessing

HANDLING OUTLIER

```
Jumlah baris sebelum memfilter outlier: 4250
Jumlah baris setelah memfilter outlier: 3515
```

Area Code_area_code_408	Area Code_area_code_415	Area Code_area_code_510	International Plan_no	International Plan_yes	Voice Mail Plan_no	Voice Mail Plan_yes	State_frec_enc
0	1	0	1	0	0	1	0.018588
0	1	0	1	0	1	0	0.018824
0	1	0	0	1	1	0	0.014824
0	1	0	0	1	1	0	0.014588
0	1	0	0	1	0	1	0.026588
...	...	...	...	...	...	...	...
0	0	1	1	0	1	0	0.014353
0	1	0	1	0	1	0	0.016471
1	0	0	1	0	1	0	0.026588
1	0	0	1	0	1	0	0.016471
1	0	0	1	0	0	1	0.014824

ENDOCING DATA

- Melakukan one hot encoding untuk kolom Code Area
- Melakukan label encoding untuk kolom binary
- Melakukan frequency encoding untuk kolom State dikarenakan memiliki kategori yang lumayan banyak

NORMALISASI DAN STANDARISASI

- Melakukan standarisasi untuk kolom yang memiliki distribusi normal
- Melakukan normalisasi untuk kolom yang memiliki Skew pada distribusinya

Machine Learning Model

- Logistic Regression
- Decision Tree
- K-Nearest Neighbor
- Random Forest

We are going to...

Memprediksi customer churn untuk diberikan pelayanan agar bisa tetap loyal sehingga perlu meminimalisir **false negative**. hal ini dikarenakan jika pelanggan yang churn di prediksi tidak churn maka provider tidak tahu bahwa seharusnya pelanggan tersebut diberi teratment/ pelayanan agar tidak churn.

ANALYSIS SCORE

Logistic Regression

- Akurasi = 92,6%
- Recall = 42,86%

Decision Tree

- Akurasi = 94,7%
- Recall = 58,6%

K-Nearest Neighbor

- Akurasi = 91.32%
- Recall = 17,2%

Random Forest Biasa

- Akurasi = 96,02%
- Recall = 61,43%

Random Forest Undersampling

- Akurasi = 84.06%
- Recall = 78,57%

Random Forest Oversampling

- Akurasi = 95,16%
- Recall = 71,43%

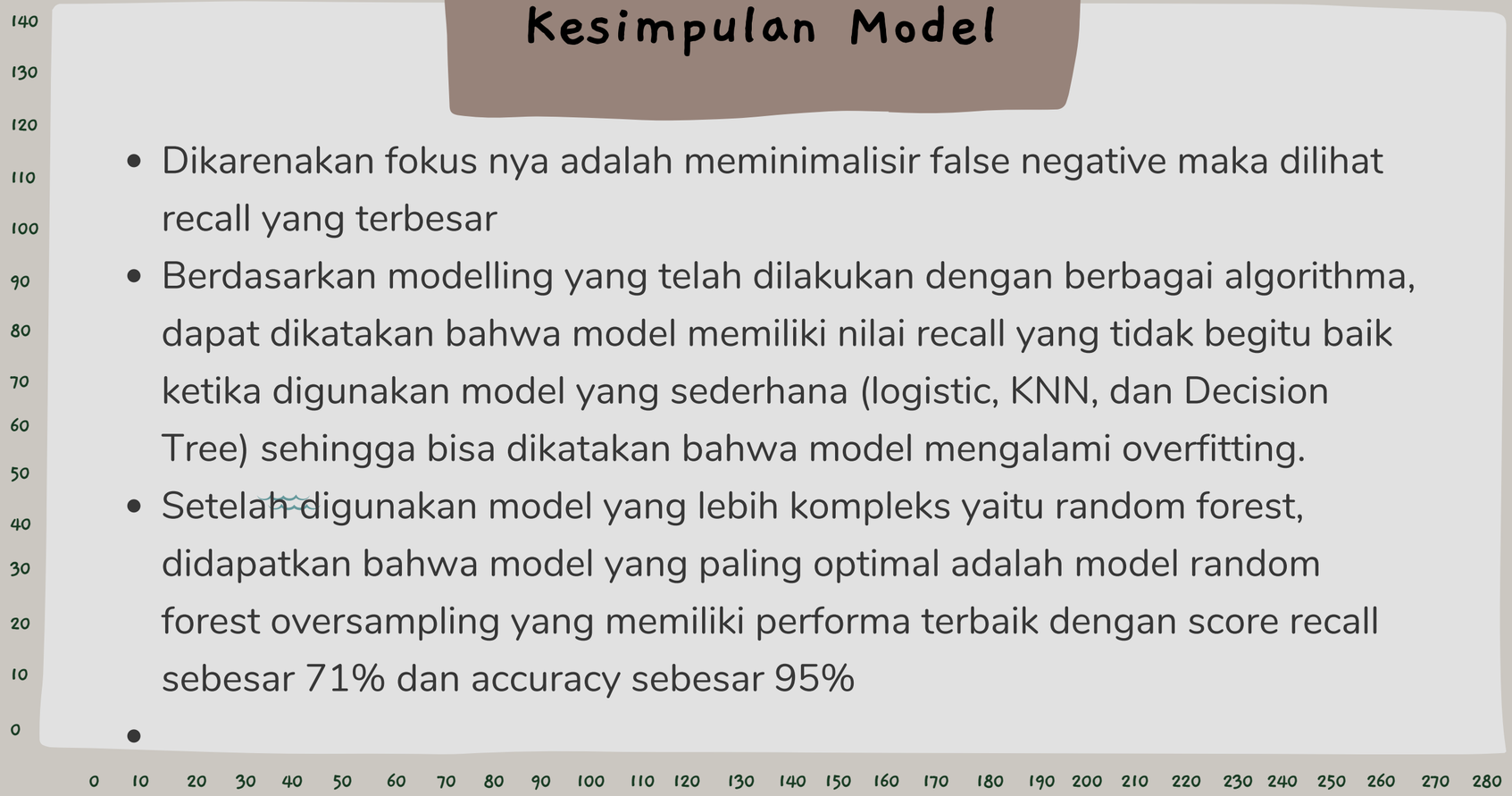
Best Model

Random Forest Combining

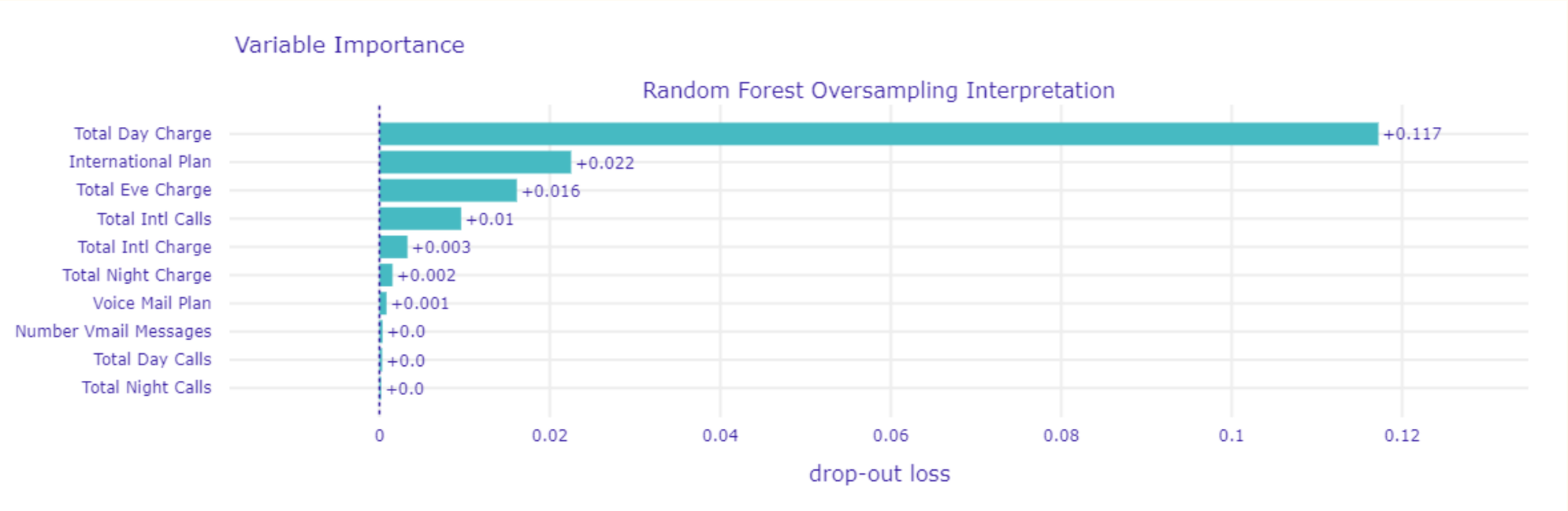
- Akurasi = 94.59%
- Recall = 71.43%

Kesimpulan Model

- Dikarenakan fokus nya adalah meminimalisir false negative maka dilihat recall yang terbesar
- Berdasarkan modelling yang telah dilakukan dengan berbagai algorithma, dapat dikatakan bahwa model memiliki nilai recall yang tidak begitu baik ketika digunakan model yang sederhana (logistic, KNN, dan Decision Tree) sehingga bisa dikatakan bahwa model mengalami overfitting.
- Setelah digunakan model yang lebih kompleks yaitu random forest, didapatkan bahwa model yang paling optimal adalah model random forest oversampling yang memiliki performa terbaik dengan score recall sebesar 71% dan accuracy sebesar 95%
-



Feature Importance



- Dari feature importance di atas dapat disimpulkan bahwa ada 10 fitur yang mempengaruhi kecenderungan churn. Fitur yang paling berpengaruh adalah total day charge, international plan dan total evening call.
- Hal itu dapat dilihat berdasarkan deep dive question yang telah dilakukan bahwa semakin tinggi total day charge maka pelanggan memiliki kecenderungan untuk churn. Selain itu karena semakin tinggi total charge maka lama nya pelanggan menggunakan telepon (Total minutes) juga semakin tinggi dan berpengaruh terhadap pembiayaan

FINISH

Recommendation

- Semakin lamanya waktu telepon dan banyaknya telepon, akan mempengaruhi total charge. dan semakin tinggi total charge maka semakin tinggi pula kecenderungan pelanggan untuk churn. Hal ini perlu diatasi dengan melakukan evaluasi terhadap pembiayaan telepon terrutama di pagi hari karena charge dan lama waktu tellepon terbesar ada di pagi hari. Bisa juga dilakukan dengan memberikan beberapa promo untuk memperkecil pembiayaan
- Meningkatkan layanan pada voice mail plan dan international plan
- Memperbaiki layanan di kode area 415 karena memiliki pelanggan terbanyak di daerah tersebut
- Dapat dikatakan bahwa telco provider memiliki pelayanan yang cukup baik karena banyak yang memiliki kecennderungan tidak churn dan penelepon customer service hanya menelepon sekitar 0–3 kali saja

THANK YOU