Final Project Presentation

Nomor Kelompok: 9
Nama Mentor: Aditya Bariq Ikhsan
Nama:

- Chairunisa Az Zahra Arifin
- Oktaviani Nurlinda Sari

Machine Learning Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Latar Belakang

Latar Belakang Project

Sumber Data: https://www.kaggle.com/datasets/barun2104/telecom-churn?datasetId=567482

Problem: **classification**

Tujuan:

- Memprediksi dan mengelompokkan *target columnn*
- Menganalisis untuk mendapatkan *Interesting Insights* dari data set, meliputi faktor-faktor apa saja yang memengaruhi *Churn.*
- Mendapatkan model Machine Learning yang terbaik
- Memberikan rekomendasi yang dapat dilakukan perusahaan untuk mengurangi *Churn* dari para pelanggan mereka.

Explorasi Data dan Visualisasi

Business Understanding

Churn adalah pemutusan layanan jasa telekomunikasi oleh pelanggan atau perusahaan. Perusahaan lebih memutuskan mempertahankan pelanggan, karena dibutuhkan biaya lebih sedikit daripada mencari pelanggan baru.

Alasan utama mengapa *customer churn rate* penting adalah persentase pelanggan yang hilang tersebut sangat memengaruhi *growth rate* perusahaan.

Churn prediction **bertujuan** untuk memprediksi peluang seorang pelanggan untuk churn sebelum pelanggan tersebut benar-benar melakukannya, untuk menentukan strategi *marketing* yang tepat agar dapat menekan persentase *customer churn*.

Data Cleansing

- **Part 1: Check header, column and shape of data**

df.shape terdiri atas 1 target colomn dan 10 feature colomn, ukuran 3333 x 11

	Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge	OverageFee	RoamMins
0	0	128	1	1	2.7	1	265.1	110	89.0	9.87	10.0
1	0	107	1	1	3.7	1	161.6	123	82.0	9.78	13.7
2	0	137	1	0	0.0	0	243.4	114	52.0	6.06	12.2
3	0	84	0	0	0.0	2	299.4	71	57.0	3.10	6.6
4	0	75	0	0	0.0	3	166.7	113	41.0	7.42	10.1

Input data : (['AccountWeeks', 'ContractRenewal', 'DataPlan', 'DataUsage', 'CustServCalls', 'DayMins', 'DayCalls', 'MonthlyCharge', 'OverageFee', 'RoamMins'])

Label : (['Churn'])

Data Cleansing

- **Part 2: Checking Missing Value**

Dataset yang digunakan baik-baik saja. Semua kolom tidak memiliki missing value (Jika ada missing value, pasti ada kolom yang tidak tertulis 3333 non-null).

- **Part 3: Checking Column Types**

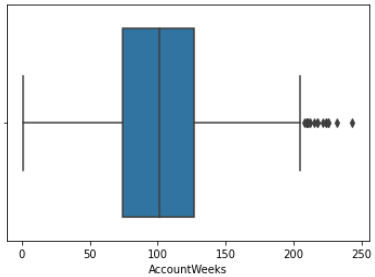
Data type 'int64' berarti kolom tersebut berisi bilangan bulat, tanpa desimal. Sedangkan, data type 'float64' berarti kolom tersebut berisi bilangan desimal.

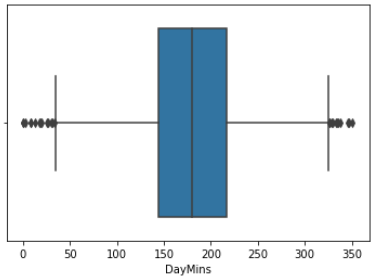
Diketahui semua kolom merupakan sudah bertipe numerik, sehingga bisa saling dibandingkan.

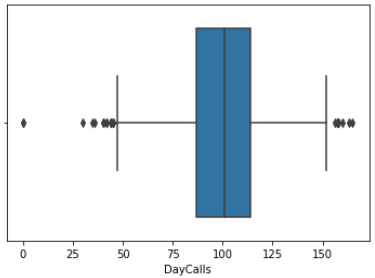
Data Cleansing

- **Part 4.1: Checking Outliers**

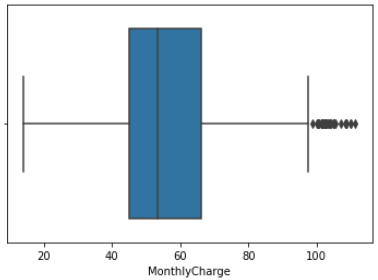
Ditemukan **Outliers**

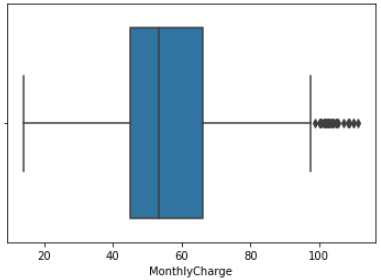


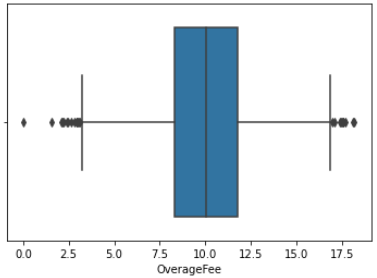




Maka dilakukan
Handling Outliers
with IQR







Data Cleansing

- **Part 4.2: Handling Outliers**

	count	mean	std	min	25%	50%	75%	max
Churn	3333.0	0.144914	0.352067	0.0	0.00	0.00	0.00	1.00
AccountWeeks	3333.0	101.064806	39.822106	1.0	74.00	101.00	127.00	243.00
ContractRenewal	3333.0	0.903090	0.295879	0.0	1.00	1.00	1.00	1.00
DataPlan	3333.0	0.276628	0.447398	0.0	0.00	0.00	1.00	1.00
DataUsage	3333.0	0.816475	1.272668	0.0	0.00	0.00	1.78	5.40
CustServCalls	3333.0	1.562856	1.315491	0.0	1.00	1.00	2.00	9.00
DayMins	3333.0	179.775098	54.467389	0.0	143.70	179.40	216.40	350.80
DayCalls	3333.0	100.435644	20.069084	0.0	87.00	101.00	114.00	165.00
MonthlyCharge	3333.0	56.305161	16.426032	14.0	45.00	53.50	66.20	111.30
OverageFee	3333.0	10.051488	2.535712	0.0	8.33	10.07	11.77	18.19
RoamMins	3333.0	10.237294	2.791840	0.0	8.50	10.30	12.10	20.00

Statistik data sebelum handling outliers (a)

	count	mean	std	min	25%	50%	75%	max
Churn	3333.0	0.144914	0.352067	0.00	0.00	0.00	0.00	1.00
AccountWeeks	3333.0	101.003300	39.644112	1.00	74.00	101.00	127.00	206.50
ContractRenewal	3333.0	0.903090	0.295879	0.00	1.00	1.00	1.00	1.00
DataPlan	3333.0	0.276628	0.447398	0.00	0.00	0.00	1.00	1.00
DataUsage	3333.0	0.816475	1.272668	0.00	0.00	0.00	1.78	5.40
CustServCalls	3333.0	1.562856	1.315491	0.00	1.00	1.00	2.00	9.00
DayMins	3333.0	179.816157	54.152190	34.65	143.70	179.40	216.40	325.45
DayCalls	3333.0	100.473597	19.863740	46.50	87.00	101.00	114.00	154.50
MonthlyCharge	3333.0	56.246655	16.263174	14.00	45.00	53.50	66.20	98.00
OverageFee	3333.0	10.052934	2.520271	3.17	8.33	10.07	11.77	16.93
RoamMins	3333.0	10.254575	2.721007	3.10	8.50	10.30	12.10	17.50

Statistik data setelah handling outliers (b)

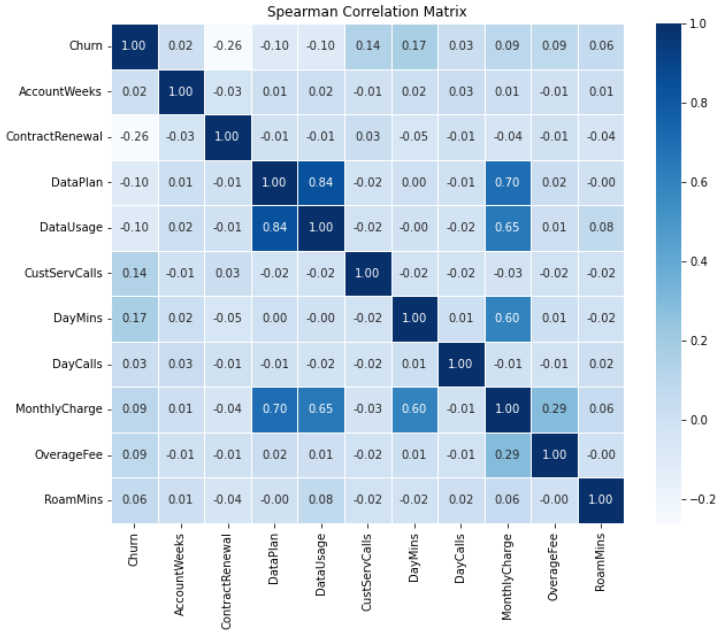
Exploratory Data Analysis

- **Part 1.1: Matriks Korelasi**
  Metode Spearman memperhitungkan hubungan monitonic kedua variable

Insight:

Hubungan korelasi diurutkan dari yang paling kuat:

- DataUsage dengan DataPlan sebesar 0.84
- MonthlyCharge dengan DataPlan sebesar 0.70
- MonthlyCharge dengan DataUsage sebesar 0.65
- MonthlyCharge dengan DayMins sebesar 0.60



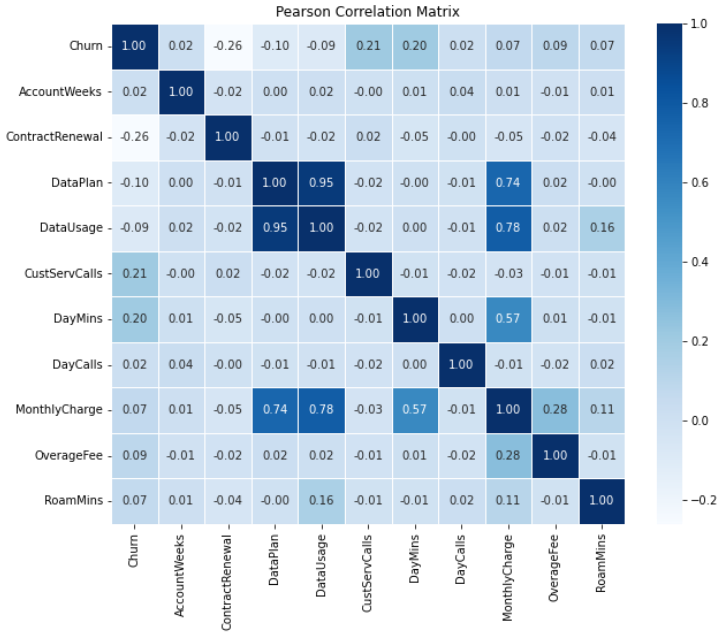
Exploratory Data Analysis

- **Part 1.2: Matriks Korelasi**
  Metode Pearson memperhitungkan hubungan linear kedua variable

Insight:

Hubungan korelasi diurutkan dari yang paling kuat:

- DataUsage dengan DataPlan sebesar 0.95
- MonthlyCharge dengan DataUsage sebesar 0.78
- MonthlyCharge dengan DataPlan sebesar 0.74
- MonthlyCharge dengan DayMins sebesar 0.57



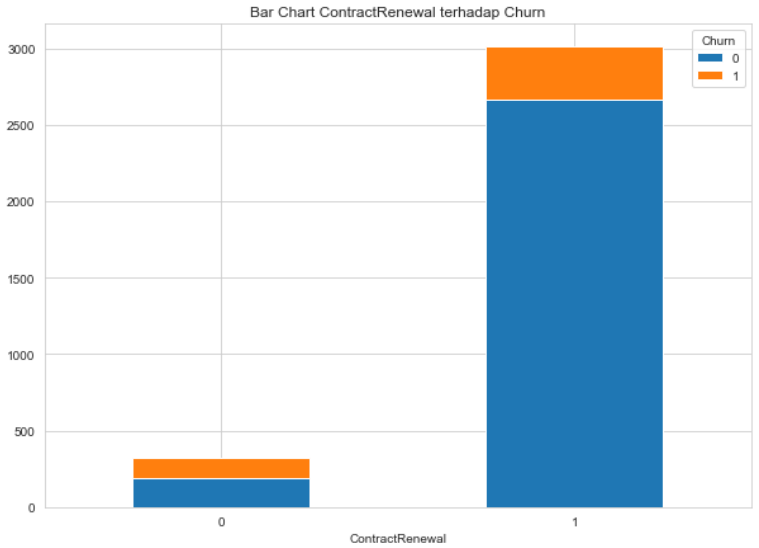
Exploratory Data Analysis

- **Part 2.1: Binary Features**

ContractRenewal terhadap Churn

Insight:

- Dari 100% contract renewal ‘no’ ada 40% churn dan 60% tidak churn.
- Dari 100% contract renewal ‘yes’ ada 10% churn dan 90% sisanya tidak churn.
- Orang yang tidak memperbarui kontrak cenderung untuk churn.



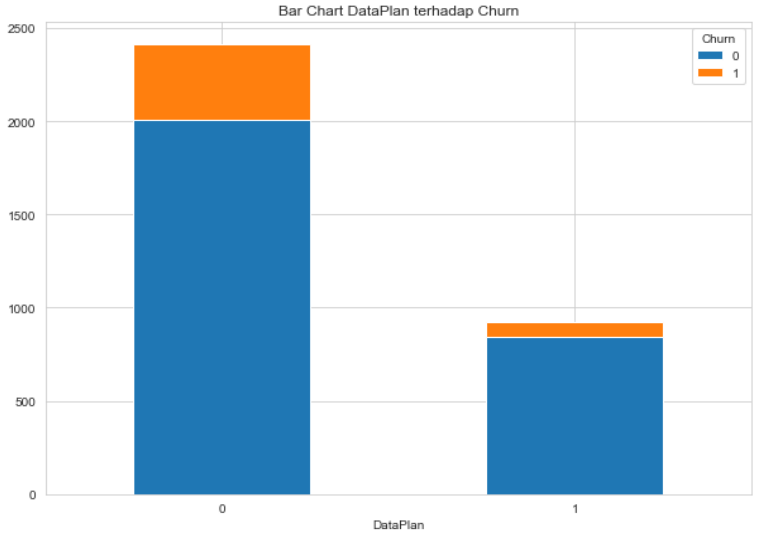
Exploratory Data Analysis

- **Part 2.2: Binary Features**

DataPlan terhadap Churn

Insight:

- Dari 100% dataplan ‘no’ ada 20% churn dan 80% tidak churn.
- Dari 100% dataplan ‘yes’ ada 10% churn dan 90% tidak churn.
- Orang yang tidak menggunakan dataplan cenderung untuk churn.



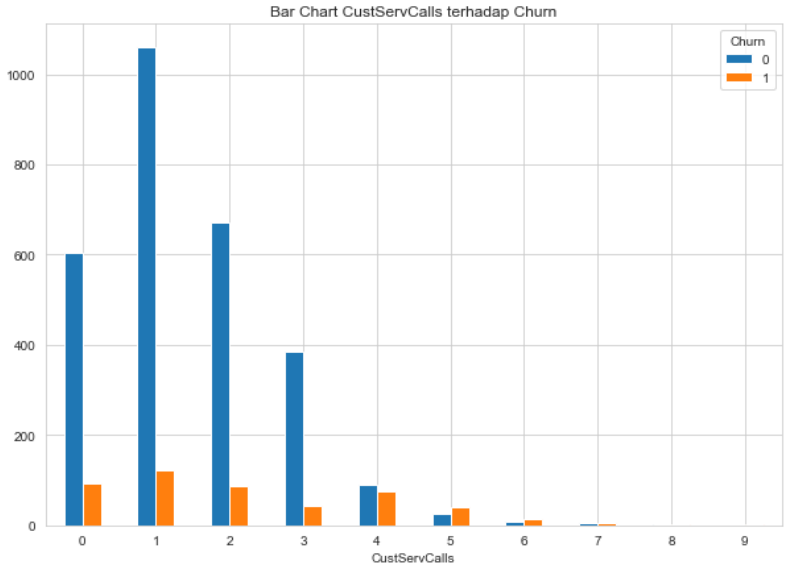
Exploratory Data Analysis

- **Part 2.3: Bar Chart**

CustServCalls terhadap Churn

Insight:

- Orang yang lebih banyak melakukan churn adalah customers dengan jumlah panggilan custumer service sebanyak 1-3 kali.
- Namun pada jumlah panggilan 4-7 kali, jumlah antara churn dan tidak churn cenderung sama besarnya.



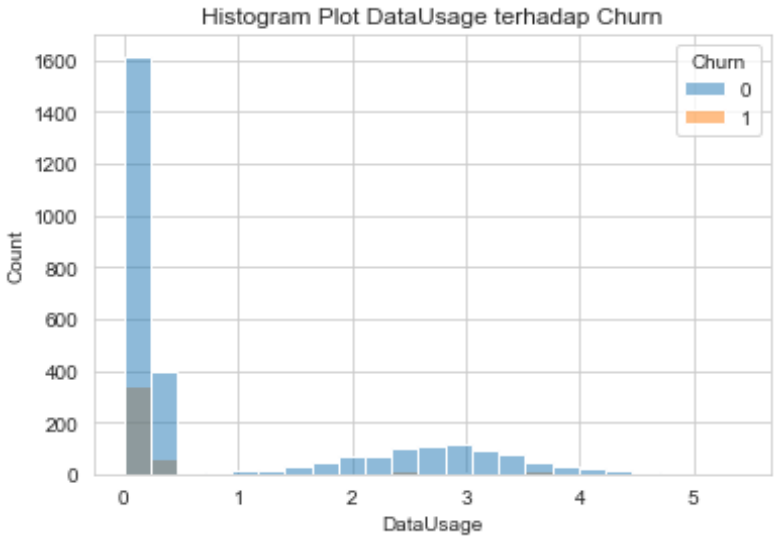
Exploratory Data Analysis

- **Part 2.4: Histogram Plot**

DataUsage terhadap Churn

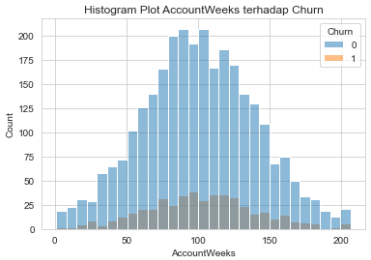
Insight:

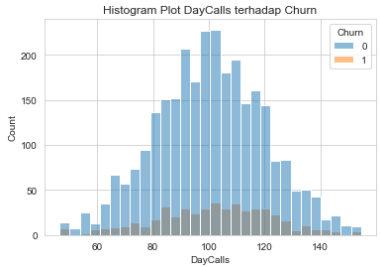
- Yang paling banyak melakukan churn adalah customers yang tidak ada pemakaian data atau berjumlah 0 data usage.

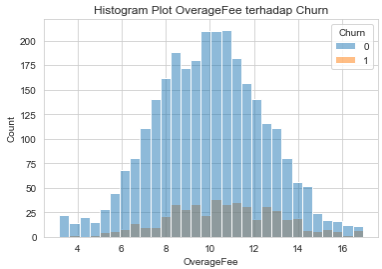


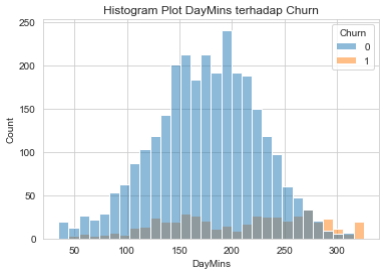
Exploratory Data Analysis

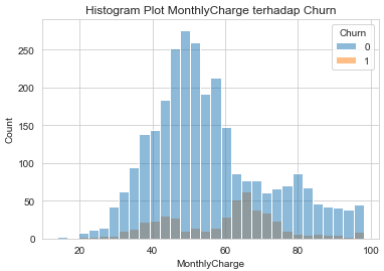
- **Part 2.5: Histogram Plot**

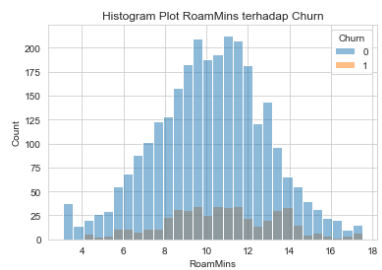












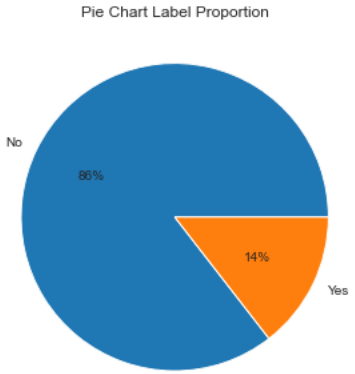
Modelling

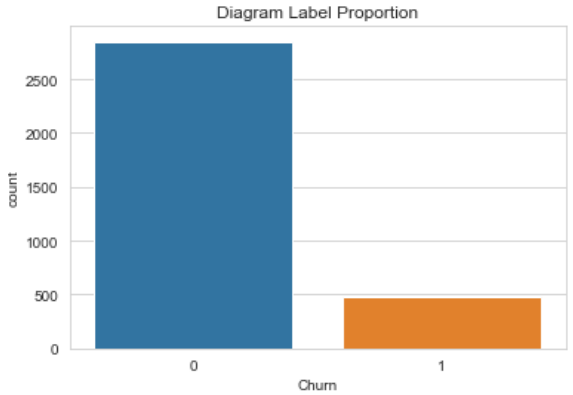
Data Preparation

- **Part 1: Check the proportion of 0 and 1 in Churn label**

Dapat terlihat bahwa sebaran data secara mayoritas customer tidak melakukan churn, dengan detail:

- Churn sebanyak 483 (14%)
- No Churn sebanyak 2850 (86%)



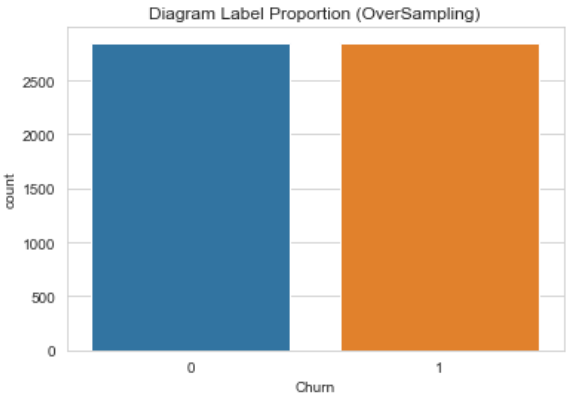


Data Preparation

- **Part 2: Handling the imbalanced class distribution**

  - After OverSampling, counts of label '1': [2850]
  - After OverSampling, counts of label '0': [2850]

  Sebaran data customer yang tidak melakukan churn dengan yang churn sudah balance setelah dilakukan oversampling.



Data Preparation

- **Part 3: Train Split Test**
  Dengan proporsi data training 60% dan data test 40%, didapat:

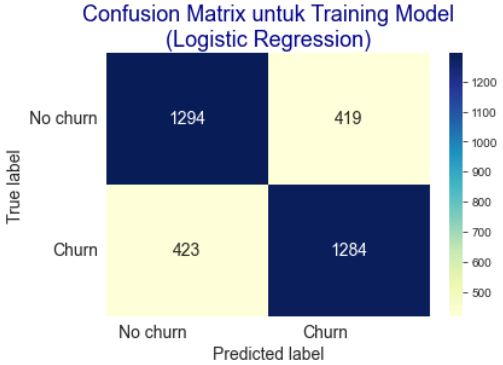
  - Before OverSampling, the shape of train_X: (1999, 10)
  - Before OverSampling, the shape of test_X: (1334, 10)
  - Before OverSampling, the shape of train_y: (1999,)
  - Before OverSampling, the shape of test_y: (1334,)

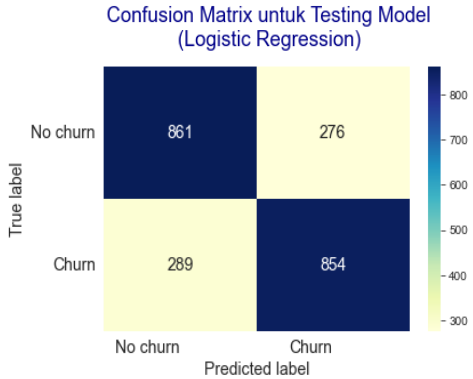
  - After OverSampling, the shape of train_X: (3420, 10)
  - After OverSampling, the shape of test_X: (2280, 10)
  - After OverSampling, the shape of train_y: (3420,)
  - After OverSampling, the shape of test_X: (2280,)

Classification Model

- **Part 1: Logistic Regression**

  - Data train: prediksi churn yang sebenarnya churn adalah 1284, prediksi tidak churn yang sebenarnya tidak Churn adalah 1294, prediksi tidak churn yang sebenarnya Churn adalah 423 dan prediksi churn yang sebenarnya tidak churn adalah 419.
  - Data test: prediksi churn yang sebenarnya churn adalah 854, prediksi tidak churn yang sebenarnya tidak Churn adalah 861, prediksi tidak churn yang sebenarnya Churn adalah 289 dan prediksi churn yang sebenarnya tidak Churn adalah 276.





Classification Model

- **Part 2: Decision Tree**

  - Data test: prediksi churn yang sebenarnya churn adalah 1127, prediksi tidak churn yang sebenarnya tidak churn adalah 1058, prediksi tidak churn yang sebenarnya churn adalah 16 dan prediksi churn yang sebenarnya tidak churn adalah 79.

```
[[1058   79]
 [  16 1127]]
```

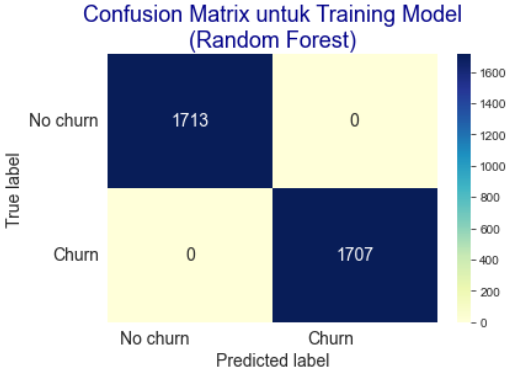
```
text_representation = tree.export_text(dtree)
print(text_representation)
```

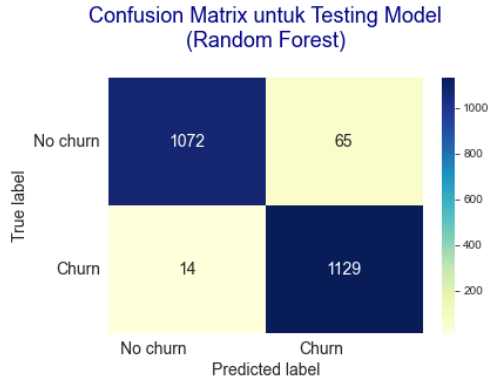
```
|--- feature_5 <= 236.15
|   |--- feature_4 <= 3.50
|   |   |--- feature_1 <= 0.50
|   |   |   |--- feature_9 <= 13.10
|   |   |   |   |--- feature_3 <= 3.24
|   |   |   |   |   |--- feature_6 <= 134.50
|   |   |   |   |   |   |--- feature_5 <= 233.80
|   |   |   |   |   |   |   |--- feature_0 <= 35.00
|   |   |   |   |   |   |   |   |--- feature_3 <= 0.27
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- feature_3 >  0.27
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- feature_0 >  35.00
|   |   |   |   |   |   |   |   |--- feature_9 <= 10.05
|   |   |   |   |   |   |   |   |   |--- feature_3 <= 2.13
|   |   |   |   |   |   |   |   |   |   |--- feature_6 <= 60.00
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |--- feature_6 >  60.00
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 8
|   |   |   |   |   |   |   |   |   |--- feature_3 >  2.13
|   |   |   |   |   |   |   |   |   |   |--- feature_5 <= 186.60
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |--- feature_5 >  186.60
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |--- feature_9 >  10.05
|   |   |   |   |   |   |   |   |   |--- feature_9 <= 10.60
|   |   |   |   |   |   |   |   |   |   |--- feature_7 <= 28.00
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |   |--- feature_7 >  28.00
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |--- feature_9 >  10.60
|   |   |   |   |   |   |   |   |   |   |--- feature_5 <= 169.75
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 7
```

Classification Model

- **Part 3: Random Forest**

  - Data train: prediksi churn yang sebenarnya churn adalah 1707, prediksi tidak churn yang sebenarnya tidak churn adalah 1713, prediksi tidak churn yang sebenarnya churn adalah 0 dan prediksi churn yang sebenarnya tidak churn adalah 0.
  - Data test: prediksi churn yang sebenarnya churn adalah 1129, prediksi tidak churn yang sebenarnya tidak churn adalah 1072, prediksi tidak churn yang sebenarnya churn adalah 14 dan prediksi churn yang sebenarnya tidak churn adalah 65.

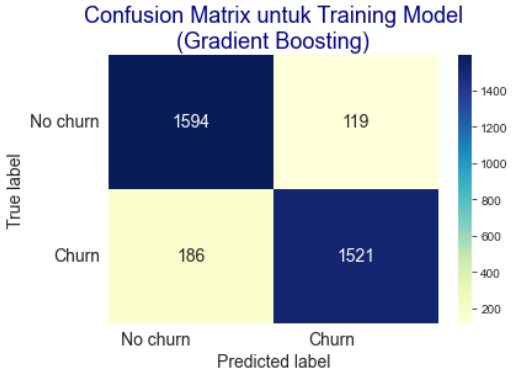


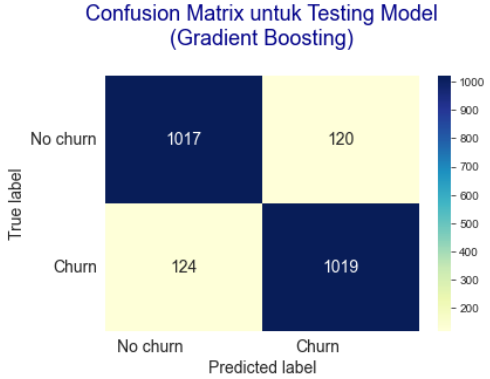


Classification Model

- **Part 4: Gradient Boosting**

  - Data train: prediksi churn yang sebenarnya churn adalah 1521, prediksi tidak churn yang sebenarnya tidak churn adalah 1594, prediksi tidak churn yang sebenarnya churn adalah 186 dan prediksi churn yang sebenarnya tidak churn adalah 119
  - Data test: prediksi churn yang sebenarnya churn adalah 1019, prediksi tidak churn yang sebenarnya tidak churn adalah 1017, prediksi tidak churn yang sebenarnya churn adalah 124 dan prediksi churn yang sebenarnya tidak churn adalah 120.





Model Evaluation

Sebelumnya kami telah mencoba memodelkan tanpa resampling, hasil evaluasi masing-masing matriks sebagai berikut:

1)

	precision	recall	f1-score	support
0	0.87	0.96	0.91	1136
1	0.44	0.16	0.23	198
accuracy			0.84	1334
macro avg	0.65	0.56	0.57	1334
weighted avg	0.80	0.84	0.81	1334

2)

	precision	recall	f1-score	support
0	0.92	0.91	0.92	1136
1	0.52	0.58	0.55	198
accuracy			0.86	1334
macro avg	0.72	0.74	0.73	1334
weighted avg	0.86	0.86	0.86	1334

3)

	precision	recall	f1-score	support
0	0.94	0.99	0.96	1136
1	0.88	0.64	0.74	198
accuracy			0.93	1334
macro avg	0.91	0.81	0.85	1334
weighted avg	0.93	0.93	0.93	1334

4)

	precision	recall	f1-score	support
0	0.94	0.97	0.96	1136
1	0.80	0.64	0.71	198
accuracy			0.92	1334
macro avg	0.87	0.81	0.83	1334
weighted avg	0.92	0.92	0.92	1334

Hasilnya, **imbalance** menyebabkan output data test pada model awal **overfitting**, the model is more biased towards majority class.

Model Evaluation

Kemudian seperti pada bagian data preparation, hasil evaluasi masing-masing matriks setelah dilakukan **oversampling** sebagai berikut:

1)

	precision	recall	f1-score	support
0	0.74	0.76	0.75	1137
1	0.75	0.74	0.75	1143
accuracy			0.75	2280
macro avg	0.75	0.75	0.75	2280
weighted avg	0.75	0.75	0.75	2280

2)

	precision	recall	f1-score	support
0	0.98	0.92	0.95	1137
1	0.93	0.98	0.95	1143
accuracy			0.95	2280
macro avg	0.95	0.95	0.95	2280
weighted avg	0.95	0.95	0.95	2280

3)

	precision	recall	f1-score	support
0	0.98	0.95	0.97	1137
1	0.95	0.98	0.97	1143
accuracy			0.97	2280
macro avg	0.97	0.97	0.97	2280
weighted avg	0.97	0.97	0.97	2280

4)

	precision	recall	f1-score	support
0	0.88	0.89	0.88	1137
1	0.89	0.88	0.88	1143
accuracy			0.88	2280
macro avg	0.88	0.88	0.88	2280
weighted avg	0.88	0.88	0.88	2280

We see their **accuracy and recall** results, the recall value of minority class has also improved. This is a good model compared to the previous one. Recall is great.

Model Evaluation

Hasil evaluasi data test setiap model setelah dilakukan **oversampling** sebagai berikut:

	Recall	Precision	Accuracy
Logistic Regression	0.7478	0.7479	0.7478
Decision Tree	0.9529	0.9549	0.9530
Random Forest	0.9683	0.9689	0.9684
Gradient Boosting	0.8833	0.8833	0.8833

Conclusion

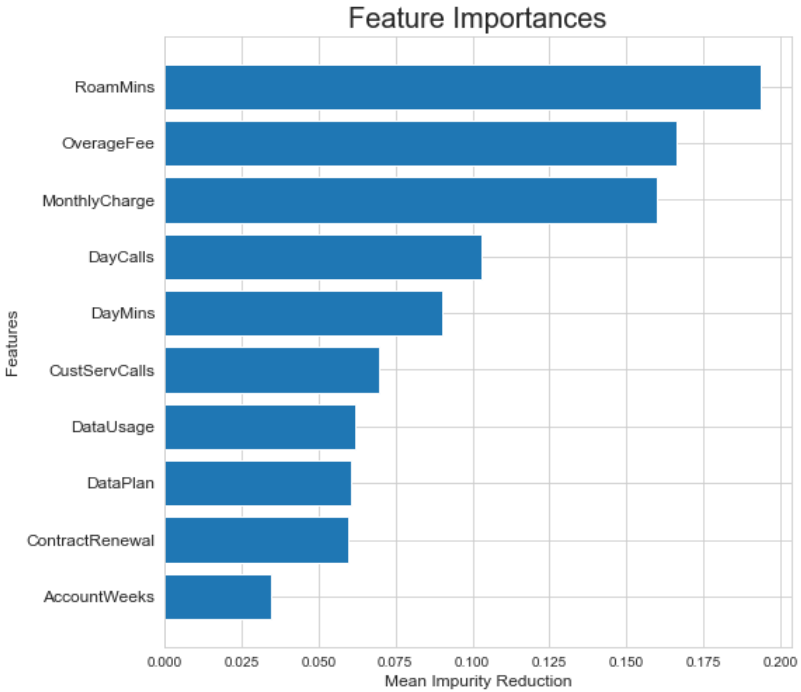
Final Model

- Model yang baik adalah model yang mampu memberikan performa bagus di fase training dan testing model. Sehingga dapat disimpulkan model yang terbaik dari keempat model di atas adalah model dengan metode **Random Forest** dengan hyperparameter tuning.
- Akurasi data train sebesar 100%, akurasi data set sebesar 97%

Random Forest	Recall	Precision	Accuracy
Tanpa Hyperparameter Tuning	0.9683	0.9689	0.9684
Dengan Hyperparameter Tuning	0.9756	0.9754	0.9754

Recommendation

- Data RoamMins merupakan data yang paling mempengaruhi data Churn. Selain itu RoamMins juga menjadi faktor utama untuk membedakan customer yang churn dan tidak churn.
- Berdasarkan insight yang diperoleh, churn terjadi pada orang-orang yang tidak menggunakan DataPlan dan DataUsage, maka dari itu perusahaan perlu memberikan paket penawaran. Tujuannya agar penggunaan data pelanggan meningkat sehingga menekan churn rate.
- Growth perusahaan juga akan meningkat jika promo berhasil membuat pelanggan memperbarui paket atau ‘Contract Renewal’.



Terima kasih!

Ada pertanyaan?