Final Project Presentation

Nomor Kelompok: 04
Nama Mentor: Aditya Bariq Ikhsan
Nama:

- Melania Justice Panggabean
- Fitrah Amalia Ramadhanti

Machine Learning Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Petunjuk

- Waktu presentasi adalah 5 menit (tentatif, tergantung dari banyaknya kelompok yang mendaftarkan diri)
- Waktu tanya jawab adalah 5 menit
- Silakan menambahkan gambar/visualisasi pada slide presentasi
- Upayakan agar tetap dalam format poin-poin (ingat, ini presentasi, bukan esai)
- Jangan masukkan *code* ke dalam slide presentasi (tidak usah memasukan screenshot jupyter notebook)

Latar Belakang

Latar Belakang Project

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

Problem: **Classification**

Tujuan:

Memprediksi Customer yang akan Churn berdasarkan AccountWeeks, ContractRenewal, DataPlan, DataUsage, CustServCalls, DayMins, DayCalls, MonthlyCharge, OverageFee, & RoamMins.

- AccountWeeks: number of weeks customer has had active account
- ContractRenewal: 1 if customer recently renewed contract, 0 if not
- DataPlan: 1 if customer has data plan, 0 if not
- DataUsage: gigabytes of monthly data usage
- CustServCalls: number of calls into customer service
- DayMins: average daytime minutes per month
- DayCalls: average number of daytime calls
- MonthlyCharge: average monthly bill
- OverageFee: largest overage fee in last 12 months
- RoamMins: average number of roaming minutes

Explorasi Data dan Visualisasi

Business Understanding

Customer churn adalah persentase pelanggan yang berhenti menggunakan produk dan layanan bisnis Anda selama jangka waktu tertentu. *Customer churn* penting diketahui bisnis karena merupakan gambaran kesuksesan suatu bisnis dalam mempertahankan pelanggan.

Bisnis akan rugi besar jika kehilangan pelanggan. Faktanya, mendapat pelanggan baru 5 kali lebih mahal daripada mempertahankan pelanggan yang sudah ada, dan membuat pelanggan baru menjadi loyal juga 16 kali lebih mahal. Jadi, diperlukan strategi untuk menghentikan *customer churn* atau kehilangan pelanggan dan menjaga pelanggan yang sudah Anda punya, karena merekalah sumber utama *revenue* bisnis!

Ketika perusahaan mampu mengurangi atau mencegah customer churn, mereka dapat meningkatkan customer lifetime value (CLV). CLV adalah jumlah total uang yang dapat Anda harapkan dari rata-rata pelanggan untuk dibelanjakan dengan bisnis Anda selama masa hidup mereka.

Data Cleansing

Pada dataset, terdapat 11 variables yang terbagi menjadi 10 independent variables dan 1 dependent variable. Serta terdapat 3333 baris pada dataset.

Independent Variables = [Accountweeks, ContractRenewal, DataPlan, DataUsage, CustServCalls, DayMins, DayCalls, MonthlyCharge, OverageFee, dan RoamMins]

Dependent Variable = Churn

	Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge	OverageFee	RoamMins
0	0	128	1	1	2.70	1	265.1	110	89.0	9.87	10.0
1	0	107	1	1	3.70	1	161.6	123	82.0	9.78	13.7
2	0	137	1	0	0.00	0	243.4	114	52.0	6.06	12.2
3	0	84	0	0	0.00	2	299.4	71	57.0	3.10	6.6
4	0	75	0	0	0.00	3	166.7	113	41.0	7.42	10.1
...	...	...	...	...	...	...	...	...	...	...	...
3328	0	192	1	1	2.67	2	156.2	77	71.7	10.78	9.9
3329	0	68	1	0	0.34	3	231.1	57	56.4	7.67	9.6
3330	0	28	1	0	0.00	2	180.8	109	56.0	14.44	14.1
3331	0	184	0	0	0.00	2	213.8	105	50.0	7.98	5.0
3332	0	74	1	1	3.70	0	234.4	113	100.0	13.30	13.7

3333 rows × 11 columns

Data Cleansing

Data describe ⇒ Menampilkan deskriptif statistik dari dataset

	Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge	OverageFee	RoamMins
count	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000
mean	0.144914	101.064806	0.903090	0.276628	0.816475	1.562856	179.775098	100.435644	56.305161	10.051488	10.237294
std	0.352067	39.822106	0.295879	0.447398	1.272668	1.315491	54.467389	20.069084	16.426032	2.535712	2.791840
min	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	14.000000	0.000000	0.000000
25%	0.000000	74.000000	1.000000	0.000000	0.000000	1.000000	143.700000	87.000000	45.000000	8.330000	8.500000
50%	0.000000	101.000000	1.000000	0.000000	0.000000	1.000000	179.400000	101.000000	53.500000	10.070000	10.300000
75%	0.000000	127.000000	1.000000	1.000000	1.780000	2.000000	216.400000	114.000000	66.200000	11.770000	12.100000
max	1.000000	243.000000	1.000000	1.000000	5.400000	9.000000	350.800000	165.000000	111.300000	18.190000	20.000000

Data Cleansing

Data info ⇒ Menampilkan ringkasan singkat dari dataset.

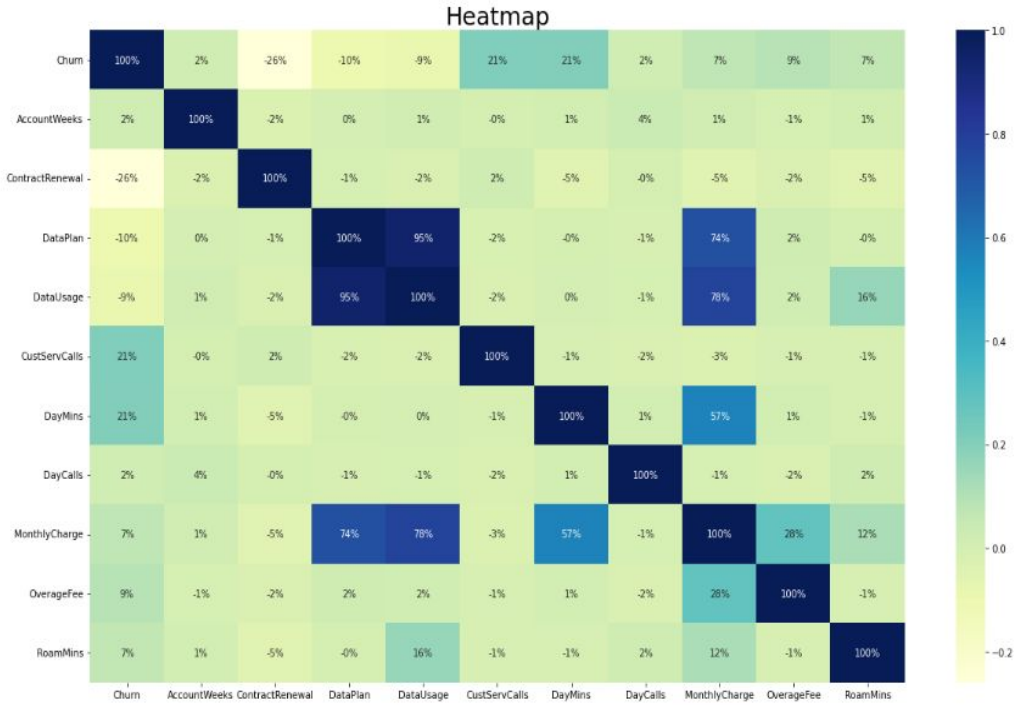
Pada dataset, tidak ditemukan adanya missing value. Sehingga dataset siap dianalisis. Serta variabel ‘Churn’ akan menjadi fitur target.

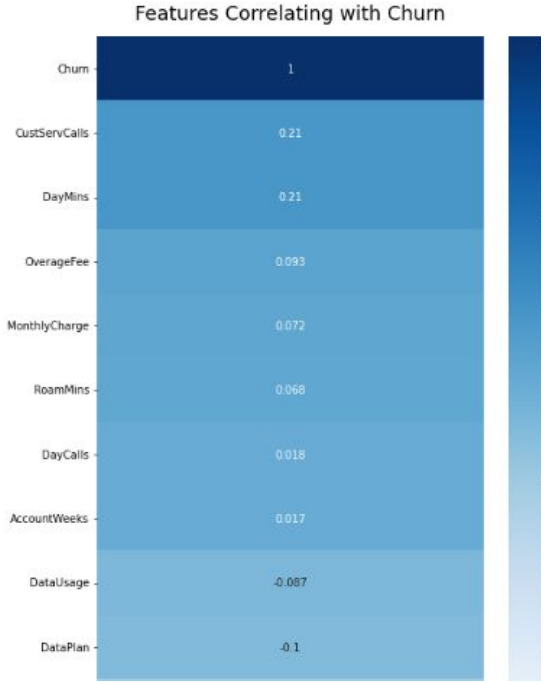
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Churn            3333 non-null   int64  
 1   AccountWeeks     3333 non-null   int64  
 2   ContractRenewal  3333 non-null   int64  
 3   DataPlan         3333 non-null   int64  
 4   DataUsage        3333 non-null   float64
 5   CustServCalls    3333 non-null   int64  
 6   DayMins          3333 non-null   float64
 7   DayCalls         3333 non-null   int64  
 8   MonthlyCharge    3333 non-null   float64
 9   OverageFee       3333 non-null   float64
 10  RoamMins         3333 non-null   float64
dtypes: float64(5), int64(6)
memory usage: 286.6 KB
```

	Description	Data Type	Data Nan
Churn	1: cancelled, 0: not cancelled	int64	0
AccountWeeks	number of weeks customer has had active account	int64	0
ContractRenewal	1: renewed contract, 0: not renewed contract	int64	0
DataPlan	1: data plan, 0: not data plan	int64	0
DataUsage	gigabytes of monthly data usage	float64	0
CustServCalls	number of calls into customer service	int64	0
DayMins	average daytime minutes per month	float64	0
DayCalls	average number of daytime calls	int64	0
MonthlyCharge	average monthly bill	float64	0
OverageFee	largest overage fee in last 12 months	float64	0
RoamMins	average number of roaming minutes	float64	0

Exploratory Data Analysis

Heatmap ⇒ Untuk melihat korelasi antar kolom pada dataset.

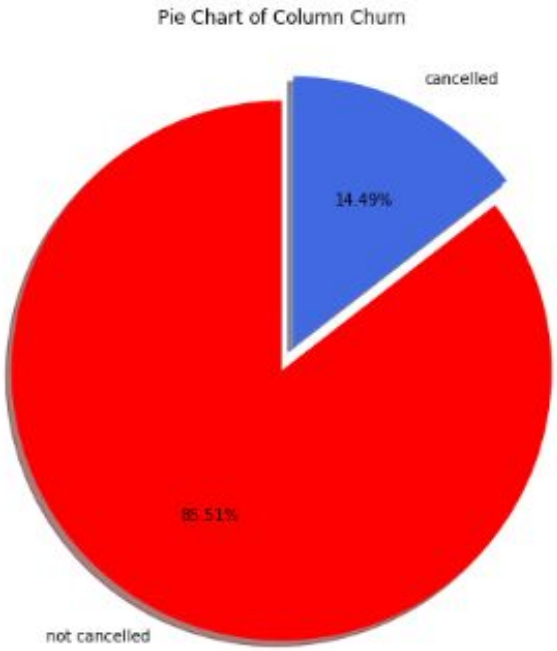




Exploratory Data Analysis

Pada kolom ‘Churn’ dapat dilihat bahwa sebanyak 2850 pelanggan (85.51%) tidak melakukan cancelled service, dan sebanyak 483 pelanggan (14.49%) melakukan cancelled service.

	Churn
not cancelled	2850
cancelled	483



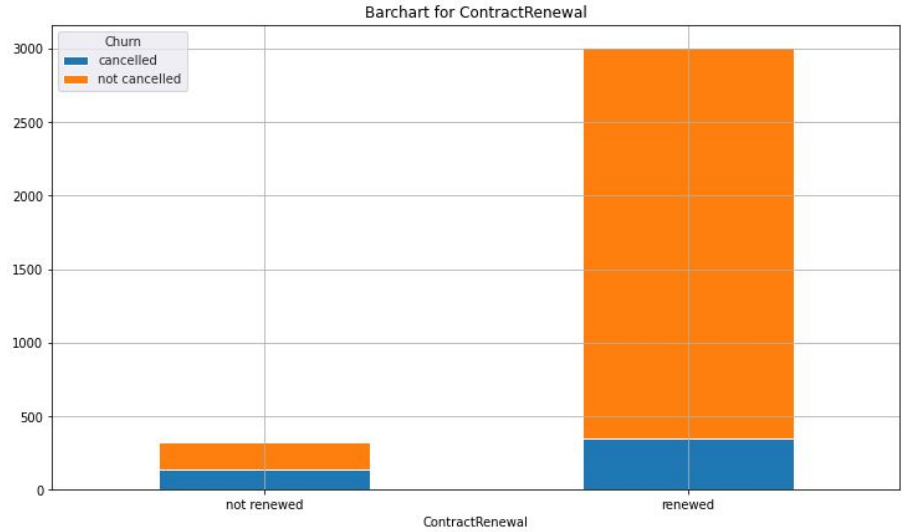
Exploratory Data Analysis

Binary Features: 'ContractRenewal' terhadap 'Churn'

Insight: 1. Sebanyak 3010 pelanggan (90%) memperpanjang kontrak mereka dan sebanyak 323 pelanggan (10%) tidak memperpanjang kontrak

2. Proporsi pelanggan yang melakukan cancelled service jauh lebih tinggi pada kelompok pelanggan yang tidak memperpanjang kontrak

	ContractRenewal
renewed	3010
not renewed	323



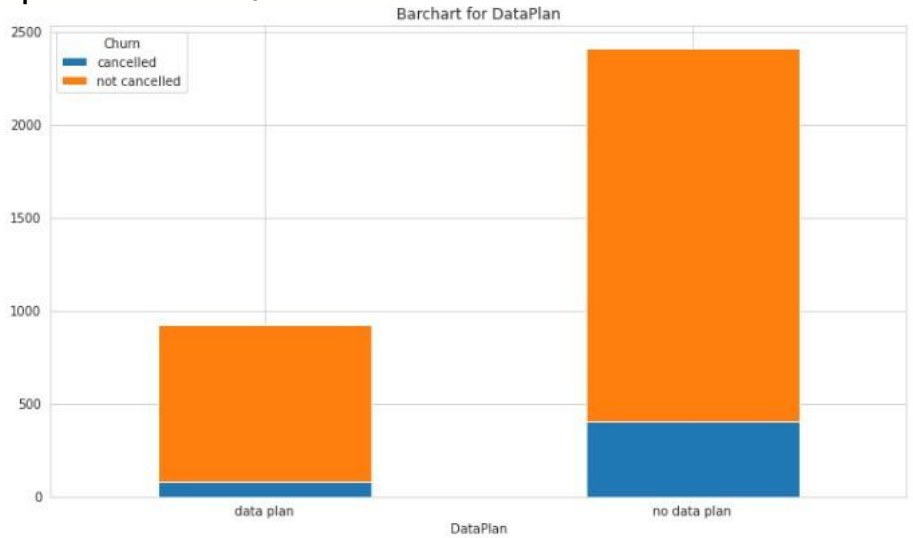
Exploratory Data Analysis

Binary Features: 'DataPlan' terhadap 'Churn'

Insight: 1. Sebanyak 2411 (72%) pelanggan tidak memiliki data plan. Dan sebanyak 922 (28%) pelanggan memiliki data plan.

2. Pada kelompok no data plan sekitar 16% (400 pelanggan) dan pada kelompok lainnya 11% (110 pelanggan) meninggalkan perusahaan / churn.

	DataPlan
no data plan	2411
data plan	922

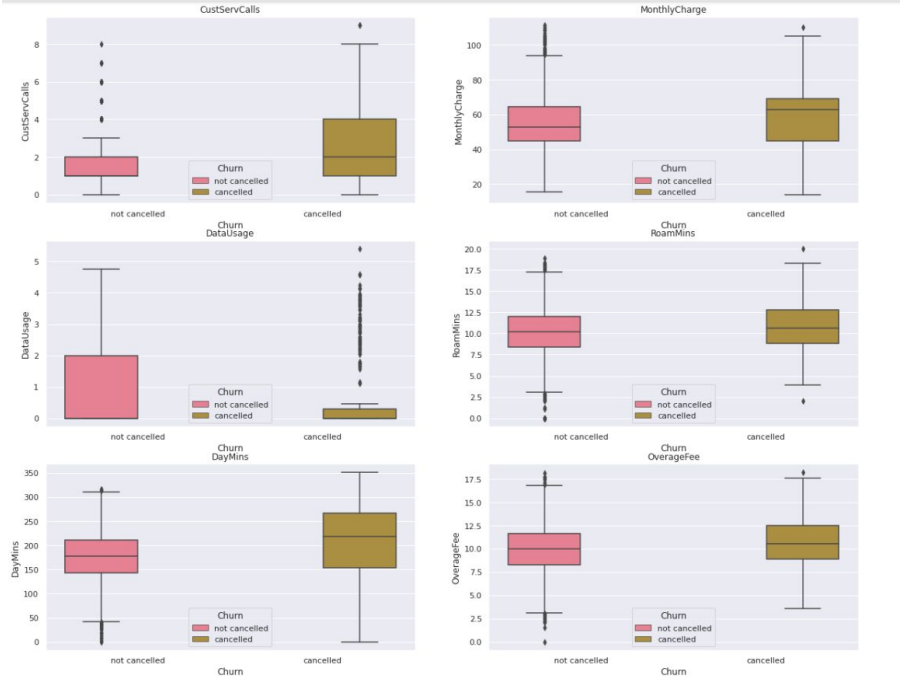


Exploratory Data Analysis

Continuous variable columns

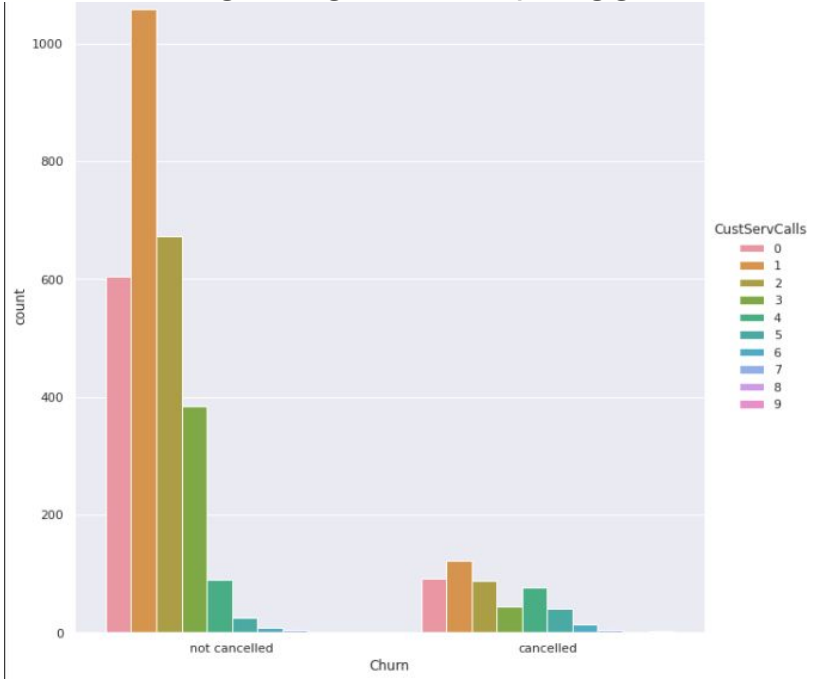
Insight: 1. Untuk pelanggan yang melakukan DayMins lebih dari 250 menit cenderung mengalami churn atau berpindah ke operator lain.

2. Untuk pelanggan yang churn memiliki biaya monthly charge sekitar 15 dollar lebih tinggi daripada yang tidak churn



Exploratory Data Analysis

Insight: Dapat dilihat untuk kasus cancelled dan not cancelled service paling banyak pelanggan melakukan panggilan sebanyak 1 kali. Pelanggan yang not cancelled service kebanyakan melakukan panggilan sebanyak 1-3 kali. Sedangkan pelanggan yang melakukan cancelled service paling banyak melakukan panggilan sebanyak 4 kali dibanding dengan 3 kali panggilan.



Exploratory Data Analysis

Insight: Dapat dilihat bahwa durasi rata-rata pelanggan yang melakukan panggilan di pagi hari (DayMins) selama 206.91 menit akan melakukan churn, hal ini mungkin saja terjadi karena jaringan di pagi hari tidak stabil sehingga pelanggan berpindah ke operator lain. Kemudian untuk pelanggan yang melakukan total panggilan (DayCalls) lebih dari 100 kali cenderung mengalami churn.

not cancelled service:0

Mean

Feature	
Churn	0.000000
AccountWeeks	100.793684
ContractRenewal	0.934737
DataPlan	0.295439
DataUsage	0.862151
CustServCalls	1.449825
DayMins	175.175754
DayCalls	100.283158
MonthlyCharge	55.816246
OverageFee	9.954618
RoamMins	10.158877

Cancelled service:1

Mean

Feature	
Churn	1.000000
AccountWeeks	102.664596
ContractRenewal	0.716356
DataPlan	0.165631
DataUsage	0.546957
CustServCalls	2.229814
DayMins	206.914079
DayCalls	101.335404
MonthlyCharge	59.190062
OverageFee	10.623085
RoamMins	10.700000

Modelling

Data Preparation

1. Split data for train and test

Metode train test split / cross validation yang digunakan ⇒ Data Testing sebanyak 20% dan Data Training sebanyak 80%.

	Data Training	Data Testing	Total
Proporsi	80%	20%	100%
Jumlah	2.666	667	3.333

Data Preparation

2. Resampling Data

Pada diagram pie chart, fitur target ‘Churn’ memiliki data yang imbalance. Dimana sebanyak 2850 (85.51%) pelanggan tidak melakukan cancelled service sedangkan sebanyak 483 (14.49%) pelanggan melakukan cancelled service. Oleh karena itu, kami melakukan Upsampling dan Downsampling data pada data training sebelum menyesuaikan dengan machine learning models yang akan dibuat nantinya. Sehingga kita dapat melihat model dan metode resampling mana yang bekerja paling baik pada data testing.

```
Downsampled Dataset Class Counts:
0    386
1    386
Name: Churn, dtype: int64

Upsampled Dataset Class Counts:
1    2280
0    2280
Name: Churn, dtype: int64
```

Machine Learning Models

1. Hasil Tuning Parameter Metode Logistic Regression

Hasil yang didapatkan yaitu Upsampled Minority Class menghasilkan hasil yang terbaik. Terutama pada bagian AUROC Score.

```
IMBALANCED CLASSES
Best roc_auc-score during GS: 0.8097791097791097
Best params from GS:
 {'max_iter': 449, 'penalty': 'l2', 'tol': 0.00019699098521619944}
Accuracy Score:  0.8665667166416792
AUROC Score: 0.832627961656719

DOWNSAMPLED MAJORITY CLASS
Best roc_auc-score during GS: 0.814696991320368
Best params from GS:
 {'max_iter': 402, 'penalty': 'none', 'tol': 0.00019507143064099162}
Accuracy Score:  0.7616191904047976
AUROC Score: 0.8278169650931453

UPSAMPLED MINORITY CLASS
Best roc_auc-score during GS: 0.8142938211757464
Best params from GS:
 {'max_iter': 630, 'penalty': 'l2', 'tol': 0.00018661761457749353}
Accuracy Score:  0.7661169415292354
AUROC Score: 0.8347983360463014
```

Machine Learning Models

2. Hasil Tuning Parameter Metode Decision Tree

Pada metode Decision Tree didapatkan hasil akurasi baru yang tinggi.

```
IMBALANCED CLASSES
Best roc_auc-score during GS: 0.8657849825612984
Best params from GS:
 {'max_depth': 6, 'min_samples_leaf': 4, 'min_samples_split': 5}
Accuracy Score:  0.9415292353823088
AUROC Score: 0.921287755471152

DOWNSAMPLED MAJORITY CLASS
Best roc_auc-score during GS: 0.8449924534340119
Best params from GS:
 {'max_depth': 6, 'min_samples_leaf': 4, 'min_samples_split': 5}
Accuracy Score:  0.8860569715142429
AUROC Score: 0.8886507505878097

UPSAMPLED MINORITY CLASS
Best roc_auc-score during GS: 0.9640322022160663
Best params from GS:
 {'max_depth': 16, 'min_samples_leaf': 2, 'min_samples_split': 7}
Accuracy Score:  0.8725637181409296
AUROC Score: 0.820898896726352
```

Machine Learning Models

3. Hasil Tuning Parameter Metode Random Forest

Random Forest berkinerja lebih baik daripada Logistic Regression, dan Decision Tree. Serta Upsampled Minority Class memberikan hasil yang baik.

```
IMBALANCED CLASSES
Best roc_auc-score during GS: 0.9030325668483563
Best params from GS:
 {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 444}
Accuracy Score:  0.9565217391304348
AUROC Score: 0.9301862904684391

DOWNSAMPLED MAJORITY CLASS
Best roc_auc-score during GS: 0.8853929620163387
Best params from GS:
 {'max_depth': 11, 'min_samples_leaf': 4, 'min_samples_split': 6, 'n_estimators': 370}
Accuracy Score:  0.8875562218890555
AUROC Score: 0.9195876288659793

UPSAMPLED MINORITY CLASS
Best roc_auc-score during GS: 0.9988015543244074
Best params from GS:
 {'max_depth': 16, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 413}
Accuracy Score:  0.9460269865067467
AUROC Score: 0.924290106710074
```

Evaluation Model for Training Dataset

	Logistic Regression	Desicion Tree	Random Forest
Accuracy	75.26%	89.06%	94.90%
Precision	75.61 %	92.46%	98.07%
Recall	94,31%	94.61%	96.05%

Conclusion

Recommendations Model

Model Random Forest

Model Random Forest menghasilkan nilai AUC-Score, Accuracy, Recall, dan Precision yang terbaik.

Interesting Insights

Feature Importance yang didapatkan dari Random Forest Algorithm:

	Feature Importance
DayMins	0.168671
MonthlyCharge	0.154507
CustServCalls	0.149221
OverageFee	0.105827
RoamMins	0.092737
ContractRenewal	0.086582
AccountWeeks	0.077681
DayCalls	0.076204
DataUsage	0.069015
DataPlan	0.019555

Rekomendasi kepada *stakeholder / audience*

- Untuk menurunkan churn, maka perusahaan harus memperhatikan variabel DayMins (Total durasi panggilan siang hari yang digunakan). Karena variabel DayMins merupakan variabel yang sangat penting terhadap prediksi customer churn dilihat dari nilai feature importance yang paling besar.
- MonthlyCharge → Memiliki korelasi yang tinggi terhadap DataPlan, DataUsage, dan DayMins. Sehingga memberikan promo paket dalam jangka waktu tertentu merupakan pilihan yang terbaik
- CustServCalls → Pelanggan yang cenderung churn menggunakan layanan CS diatas rata-rata. Sehingga meningkatkan kualitas customer care merupakan pilihan yang tepat supaya customer tidak memilih pelayanan jasa kompetitor
- DataPlan → Setiap pembelian paket data, maka akan mendapatkan poin yang bisa ditukarkan menjadi paket data internet secara free.
- DataUsage → Meningkatkan *bandwidth* internet kepada pelanggan terutama di pagi hari

Conclusion

Apa saja faktor-faktor utama yang membedakan *customer* yang *churn* dan tidak *churn*?

	Day Mins	Mothly Charge	Cust Serv Calls	Overage Fee	Roam Mins	Account Weeks	Day Calls	Data Usage	Data Plan
Churn (Cancelled Service)	Long Duration Call	Pricey	High CustServ Calls	High Overage Fee	High Roaming	Long Active Period	High Calls Total	Low Bandwidth Internet	No Data Plan
Tidak Churn (No Cancelled Service)	Short Duration Call	Cheap	Low CustServ Calls	Low Overage Fee	Low Roaming	Short Active Period	Low Calls Total	High Bandwidth Internet	Data Plan

Terima kasih!

Ada pertanyaan?