Machine Learning Project

Churn Modeling

Business Objective:-

Customer churn is a concerning problem for large companies (especially in the Telecom field) due to its direct effect on revenues. Companies often seek to know which customers are likely to churn in the recent future so that timely action can be taken to prevent it

Problem Statement

Building Logistic Regression Machine Learning model that predicts which of their customers are likely to churn (stop using their service in future).

Data Health

- The dataset provided for this activity consists of 11 features where 10 are independent features and 1 is a target variable.
- There are **3333 data instances** distributed across 11 variables.
- Variable datatypes

  - 5 variables are of float64 datatype

  - 6 variables are of int64 datatype
- DataFrame does **not have** any **duplicate** instances

Missing Values

The DataFrame is **devoid** of any missing values

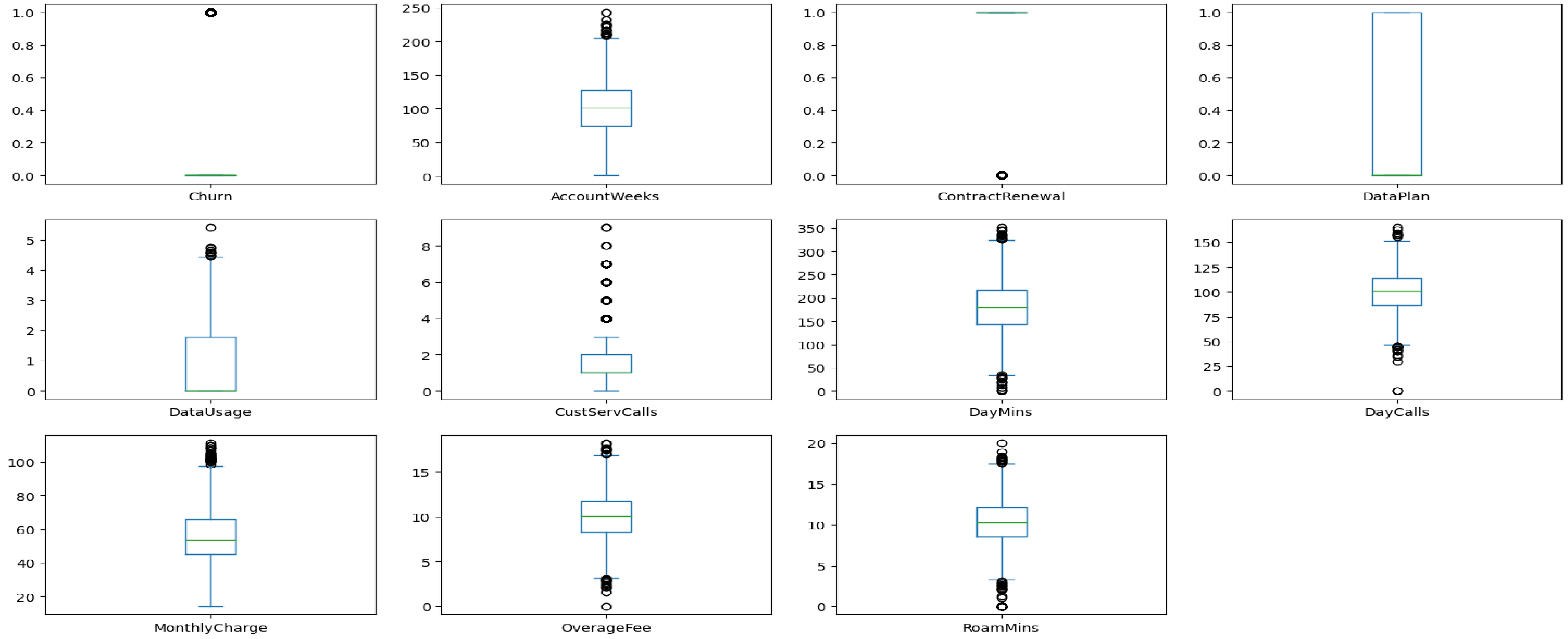
Checking for Missing Values

```
data.isnull().sum()
```

```
Churn              0
AccountWeeks       0
ContractRenewal    0
DataPlan           0
DataUsage          0
CustServCalls      0
DayMins            0
DayCalls           0
MonthlyCharge      0
OverageFee         0
RoamMins           0
dtype: int64
```

Outliers

The DataFrame has outliers.



EDA Univariate

Churn

Churn is the target variable

Data is heavily imbalanced

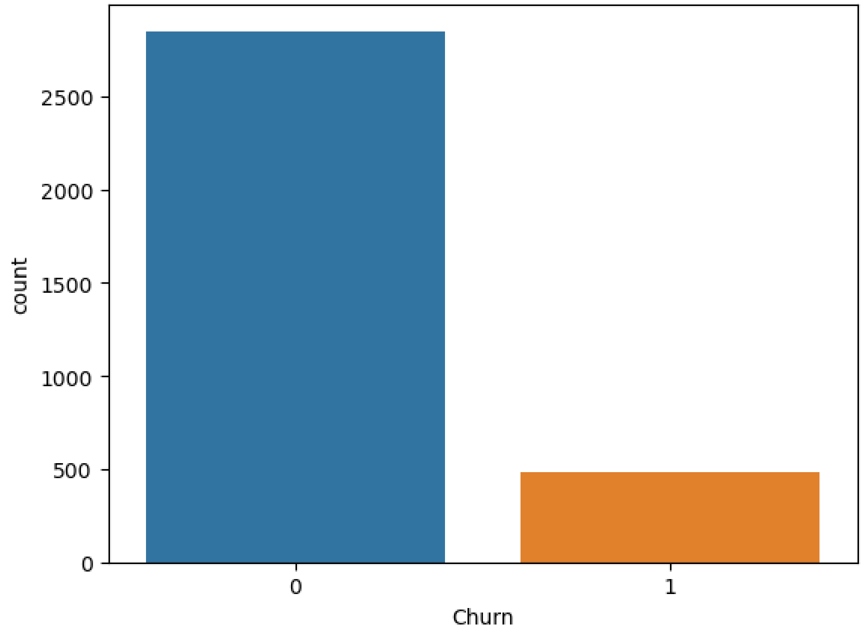
2580 data instances belongs to negative class{0} and 483 data instances belongs to positive class{1}.

```
data.Churn.value_counts()
```

```
0    2850
1     483
Name: Churn, dtype: int64
```

```
data.Churn.value_counts()/3333
```

```
0    0.855086
1    0.144914
Name: Churn, dtype: float64
```



EDA Univariate

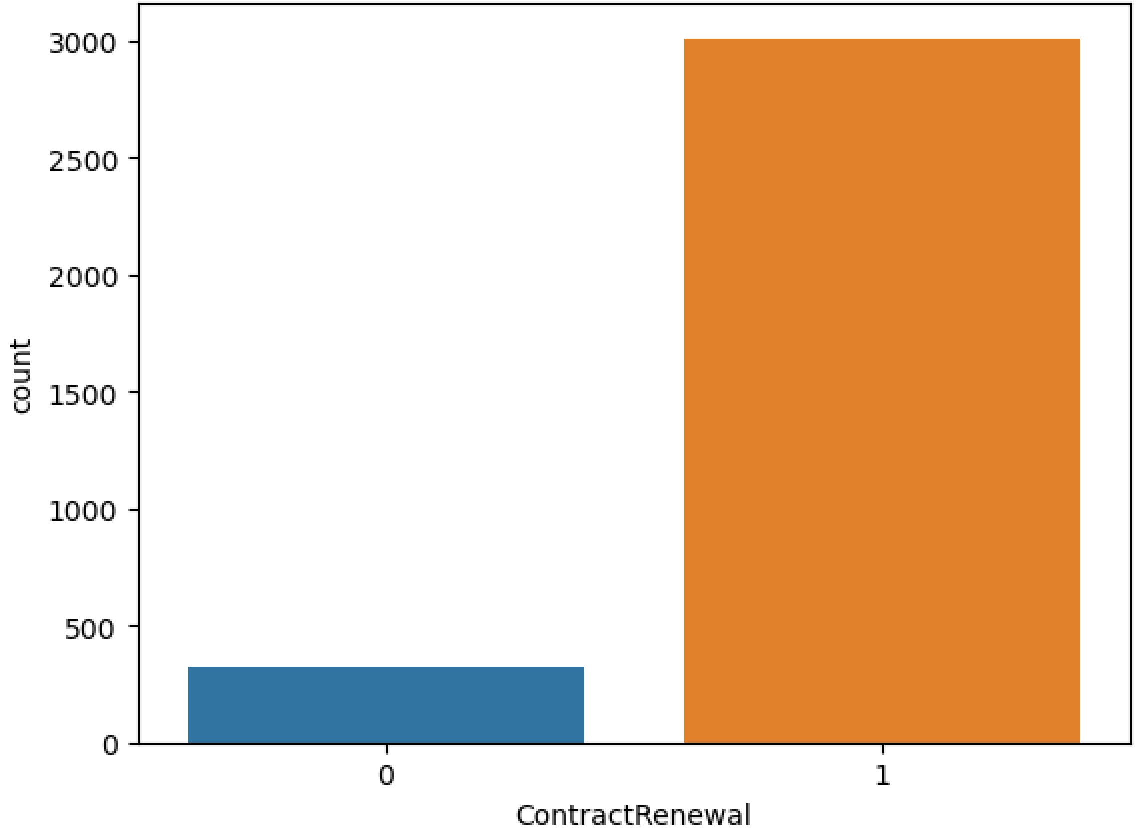
ContractRenewal

3010 customers has **recent renewal** of contract

323 customers do **not opt** for contract renewal

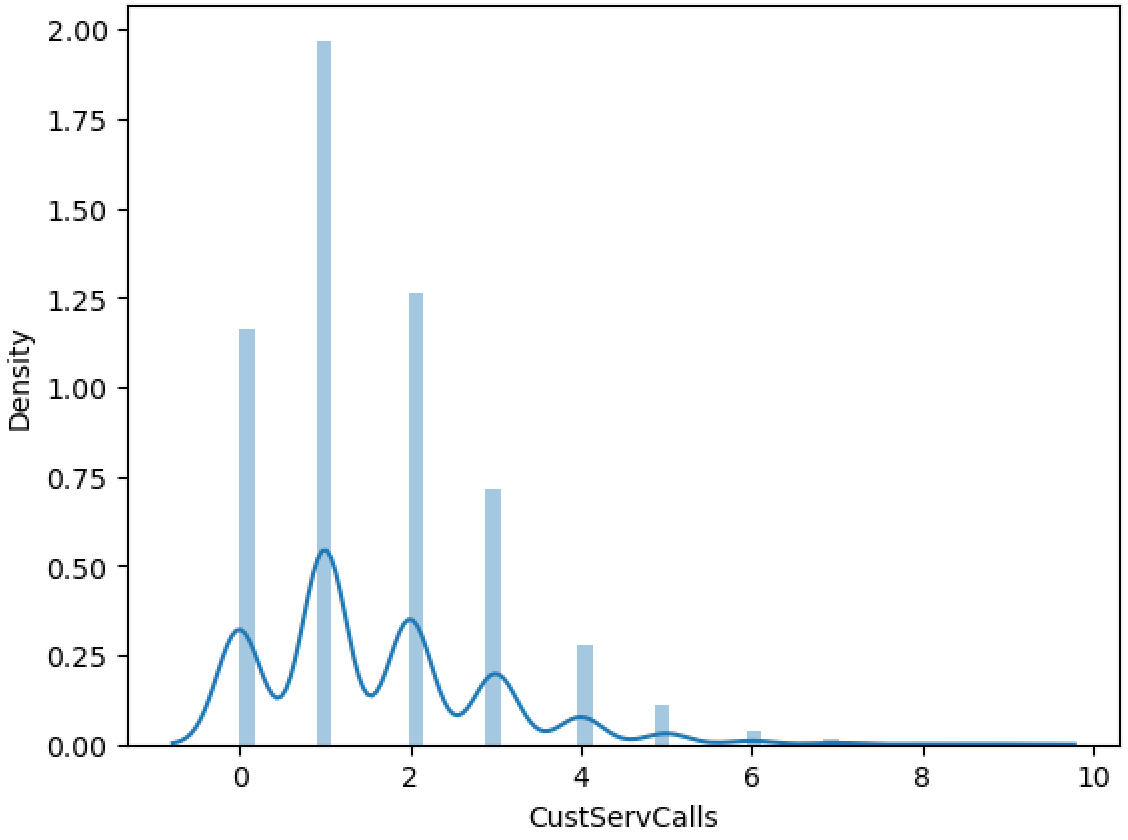
```
data.ContractRenewal.value_counts()
```

```
1    3010
0     323
Name: ContractRenewal, dtype: int64
```

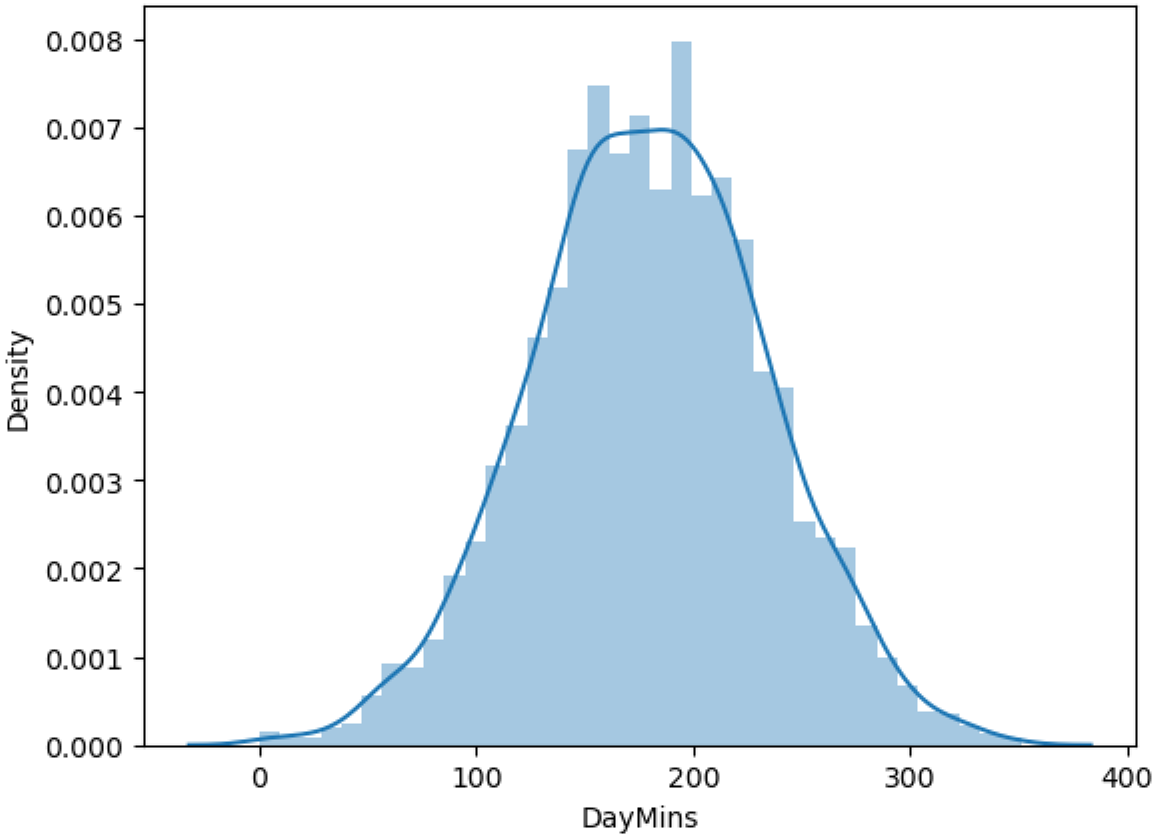


EDA Univariate

CustServCalls

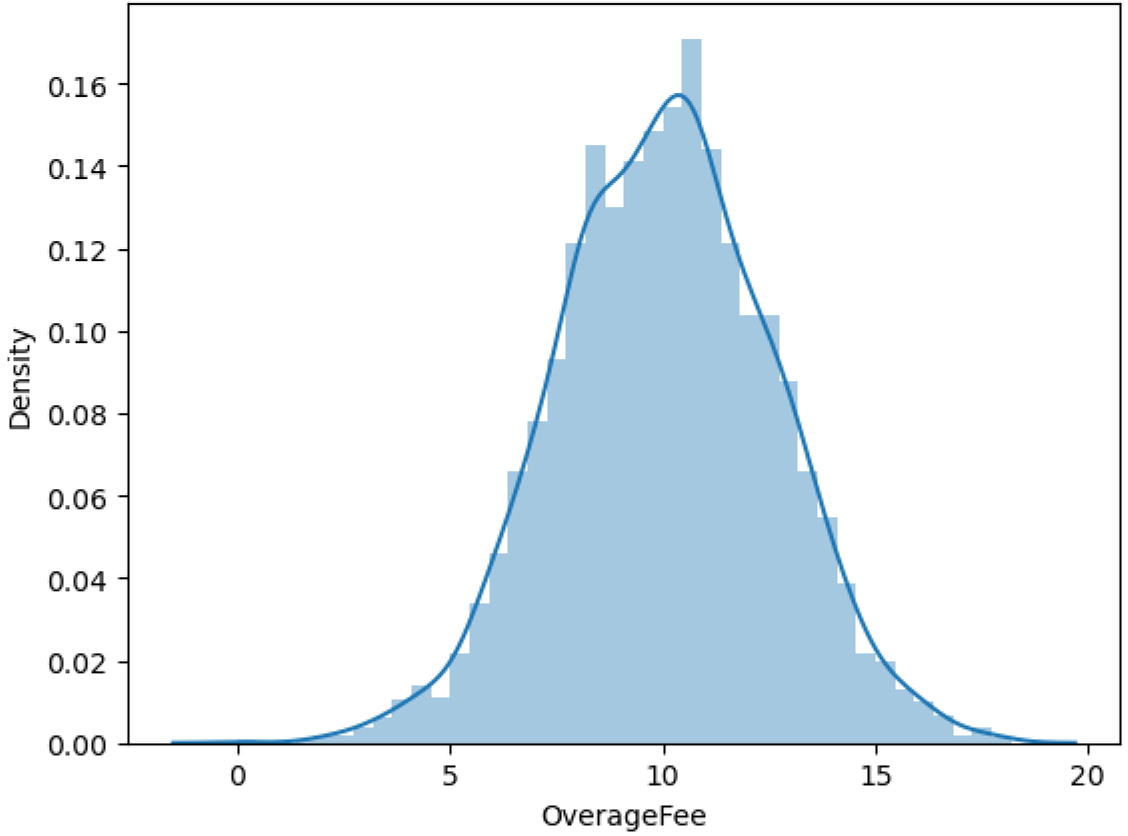


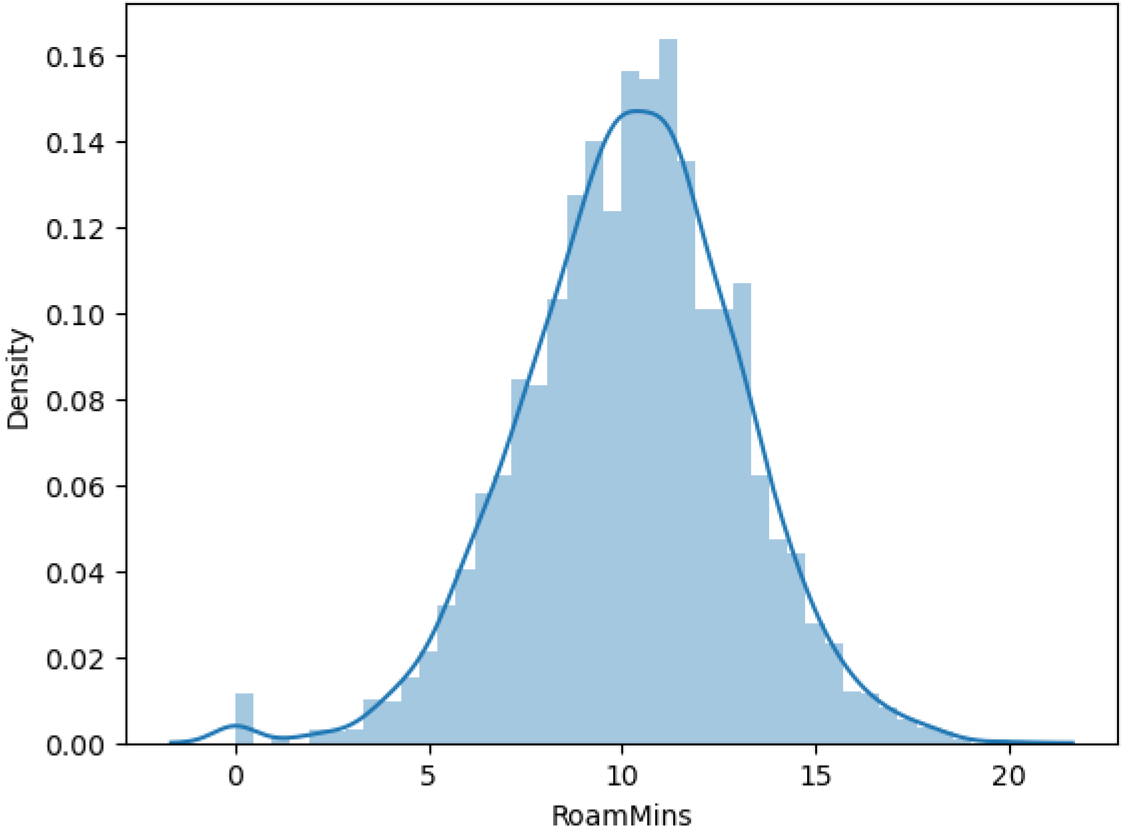
Daymins



EDA Univariate

OverageFee **RoamMins**





EDA

List of Important Variables

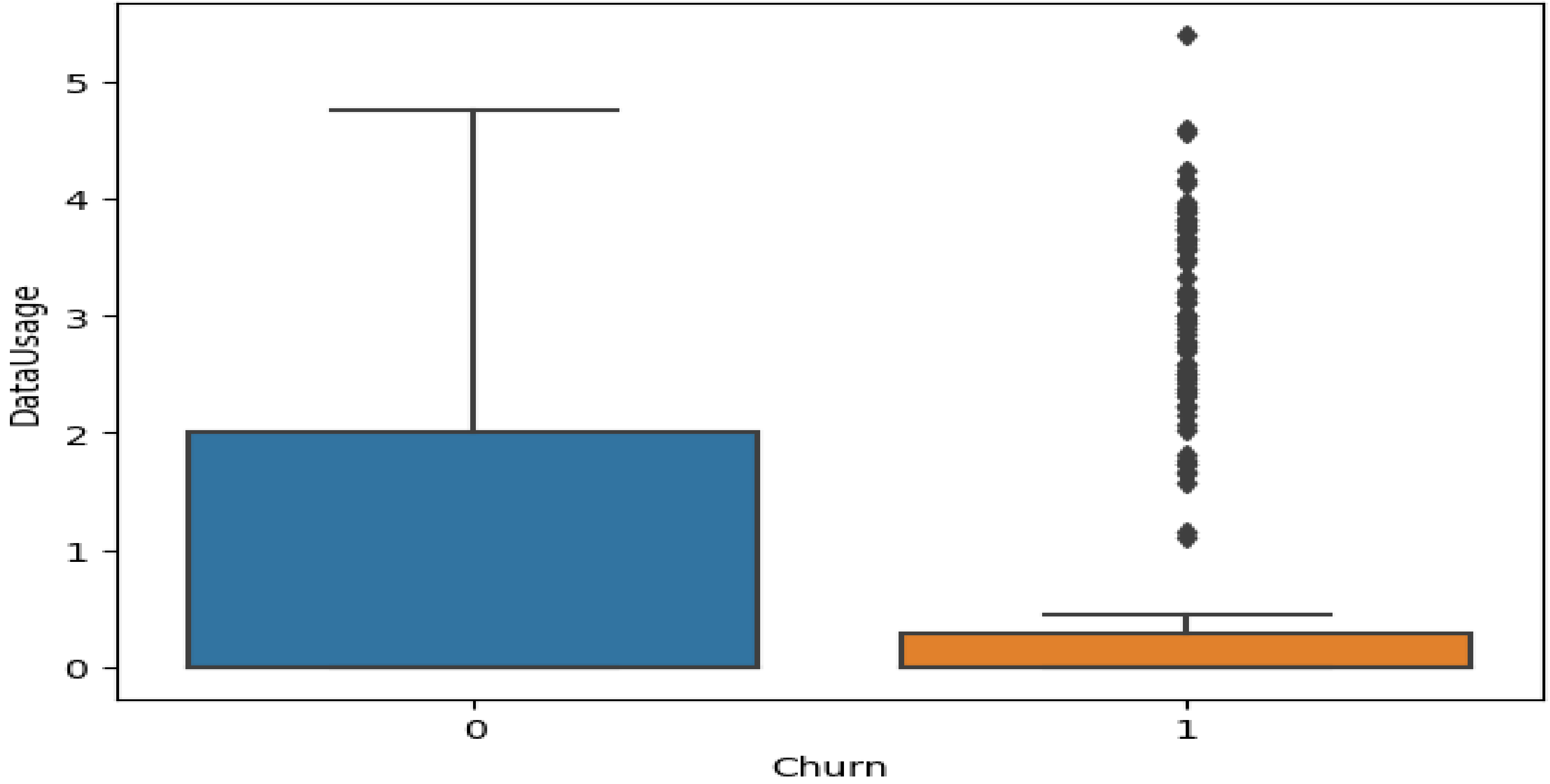
- DataUsages
- CustServCalls
- DayMins
- OverageFee
- RoamMins
- ContractRenewal

EDA

Bivariate

Churn Vs DataUsage

There is a significant difference between the groups



EDA

Bivariate

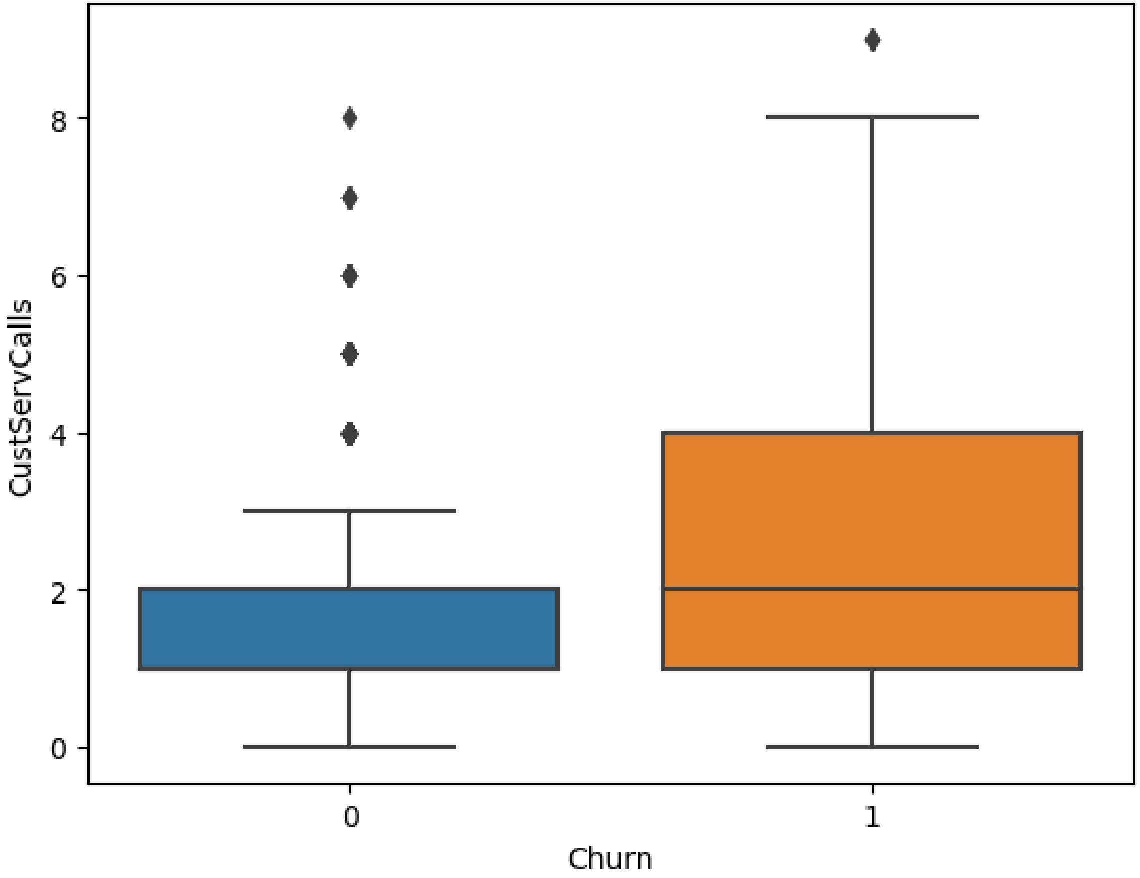
Churn Vs ContractRenawl

- Chi-Squared Statistic: **222.56575664993764**
- P-value: **2.4931077033159204e-50**
- **There is significant association**

ContractRenewal	0	1
Churn		
0	186	2664
1	137	346

EDA Bivariate

Churn Vs CustServCall

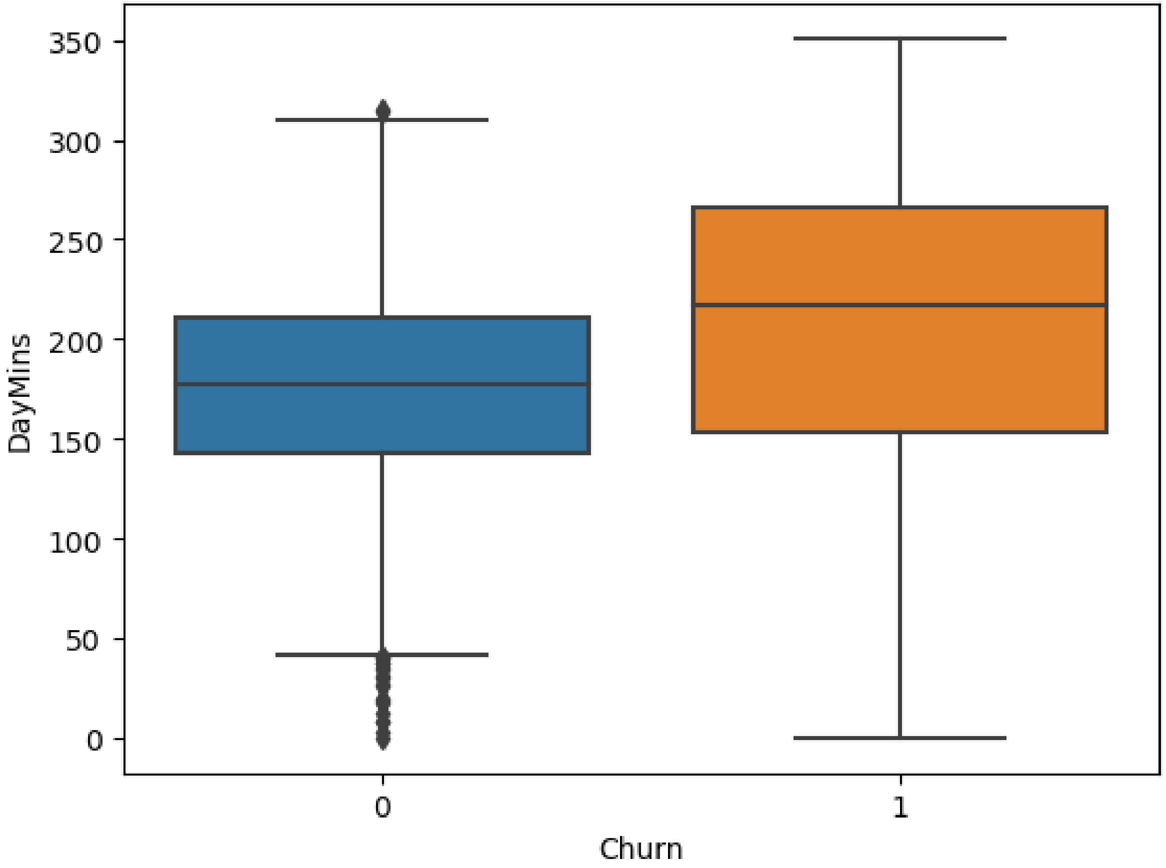


There is a significant difference between the groups

EDA

Bivariate

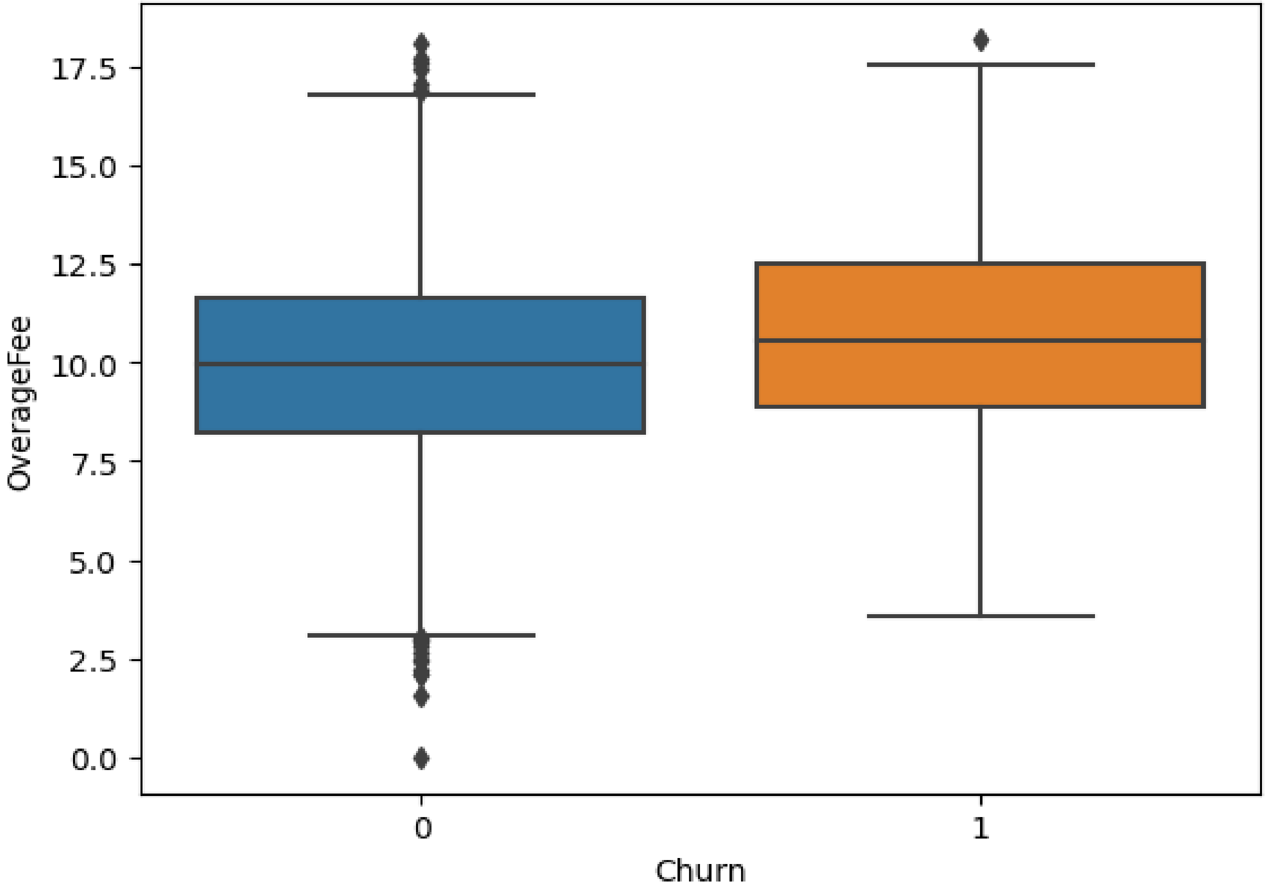
Churn Vs DayMins



There is a significant difference between the groups

EDA Bivariate

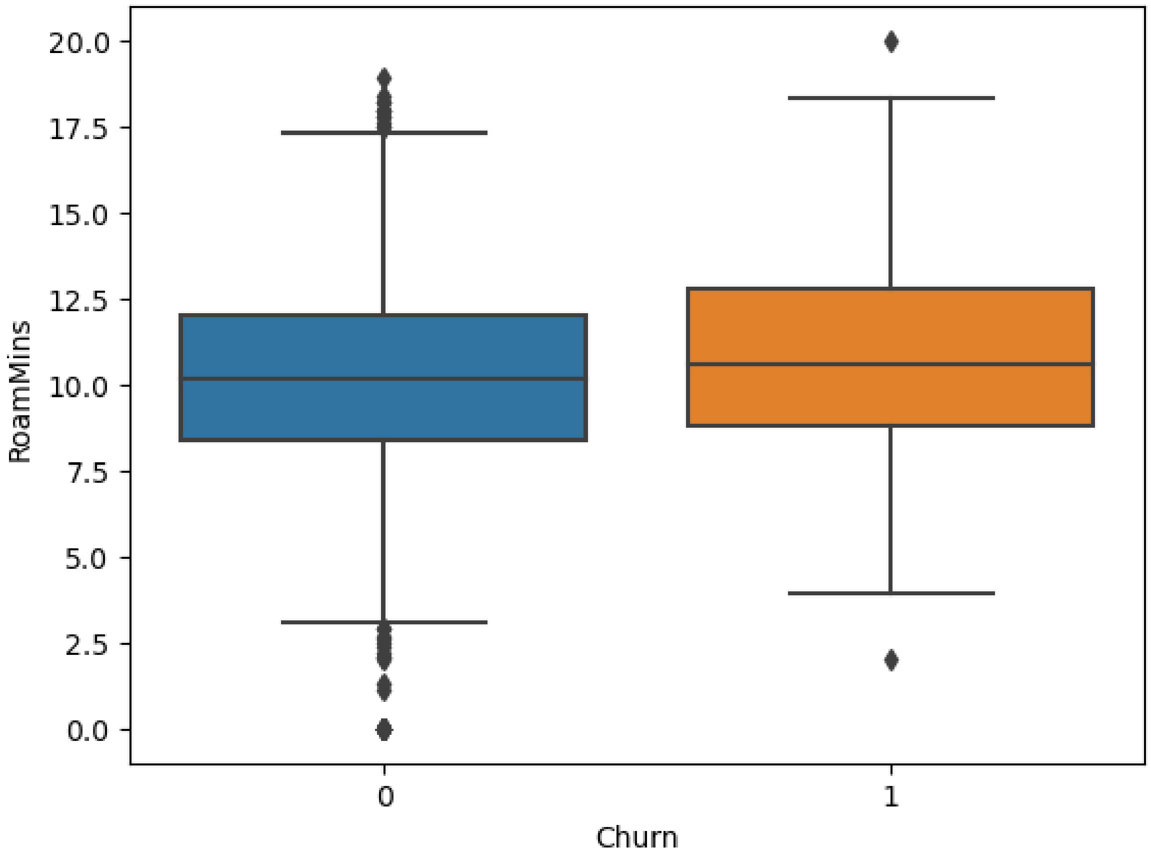
Churn Vs OverageFee



There is a significant difference between the groups

EDA Bivariate

Churn Vs RoamMins

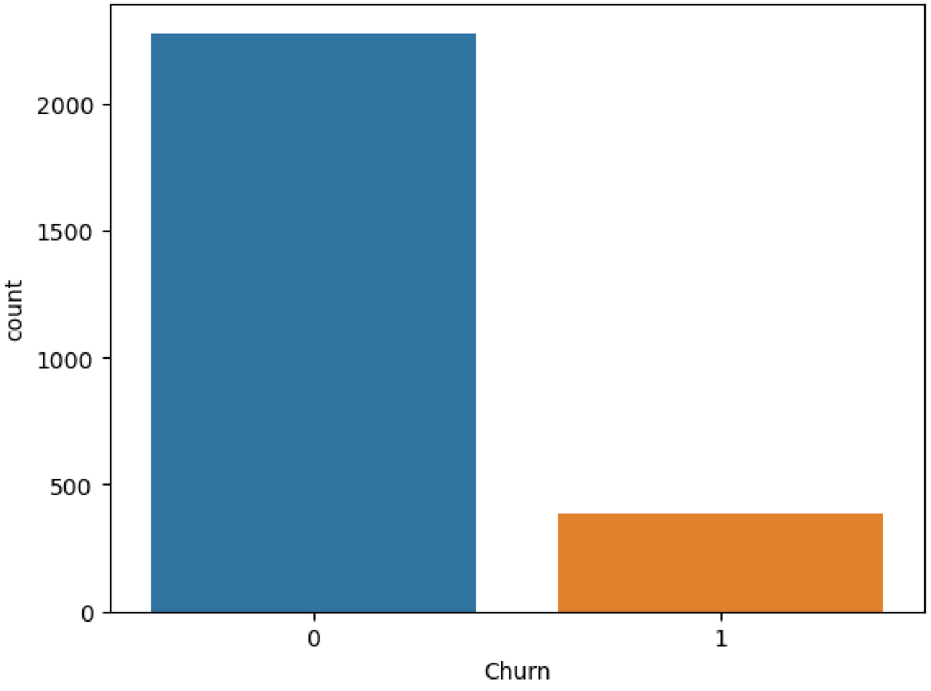


There is a significant difference between the groups

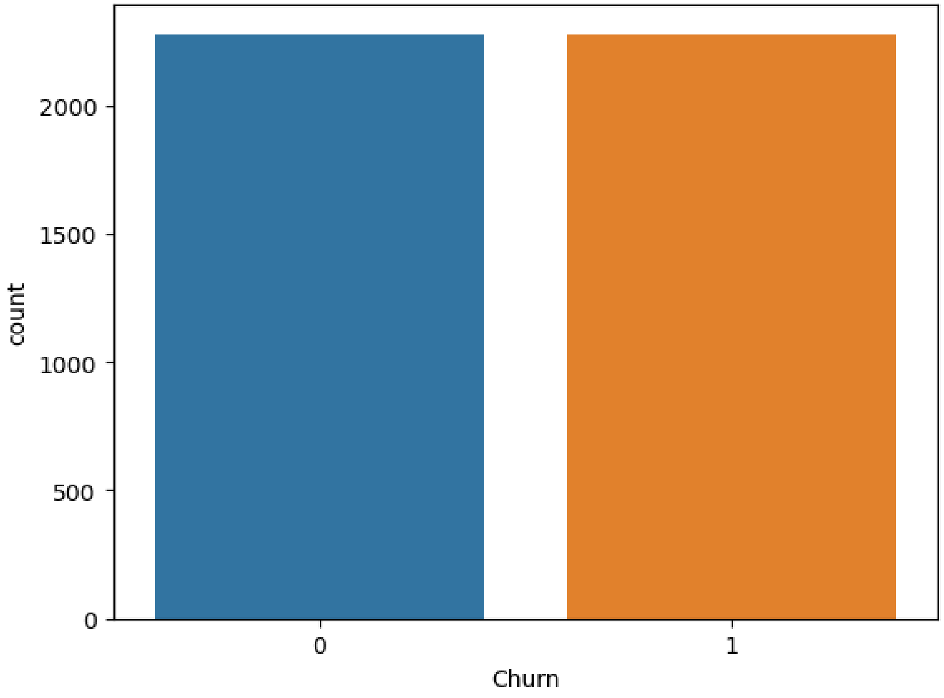
Feature Engineering

LabelEncoding has been done to bring all the variables similar Scale.

SMOTE technique has been used to get rid of the imbalanced data in Target Variable



SMOTE Resampling



ML Model {Training} Model - 1

	Negative	Positive
Negative	1775	503
Positive	528	1750

	precision	recall	f1-score	support
0	0.77	0.78	0.77	2278
1	0.78	0.77	0.77	2278
accuracy			0.77	4556
macro avg	0.77	0.77	0.77	4556
weighted avg	0.77	0.77	0.77	4556

ML Model {Training} Model - 2

	Negative	Positive
Negative	1770	508
Positive	524	1754

	precision	recall	f1-score	support
0	0.77	0.78	0.77	2278
1	0.77	0.77	0.77	2278
accuracy			0.77	4556
macro avg	0.77	0.77	0.77	4556
weighted avg	0.77	0.77	0.77	4556

ML Model {Training} Model - 3

	Negative	Positive
Negative	1770	508
Positive	524	1754

	precision	recall	f1-score	support
0	0.77	0.78	0.77	2278
1	0.78	0.77	0.77	2278
accuracy			0.77	4556
macro avg	0.77	0.77	0.77	4556
weighted avg	0.77	0.77	0.77	4556

ML Model {Validation} Final Model

	Negative	Positive
Negative	570	2
Positive	90	5

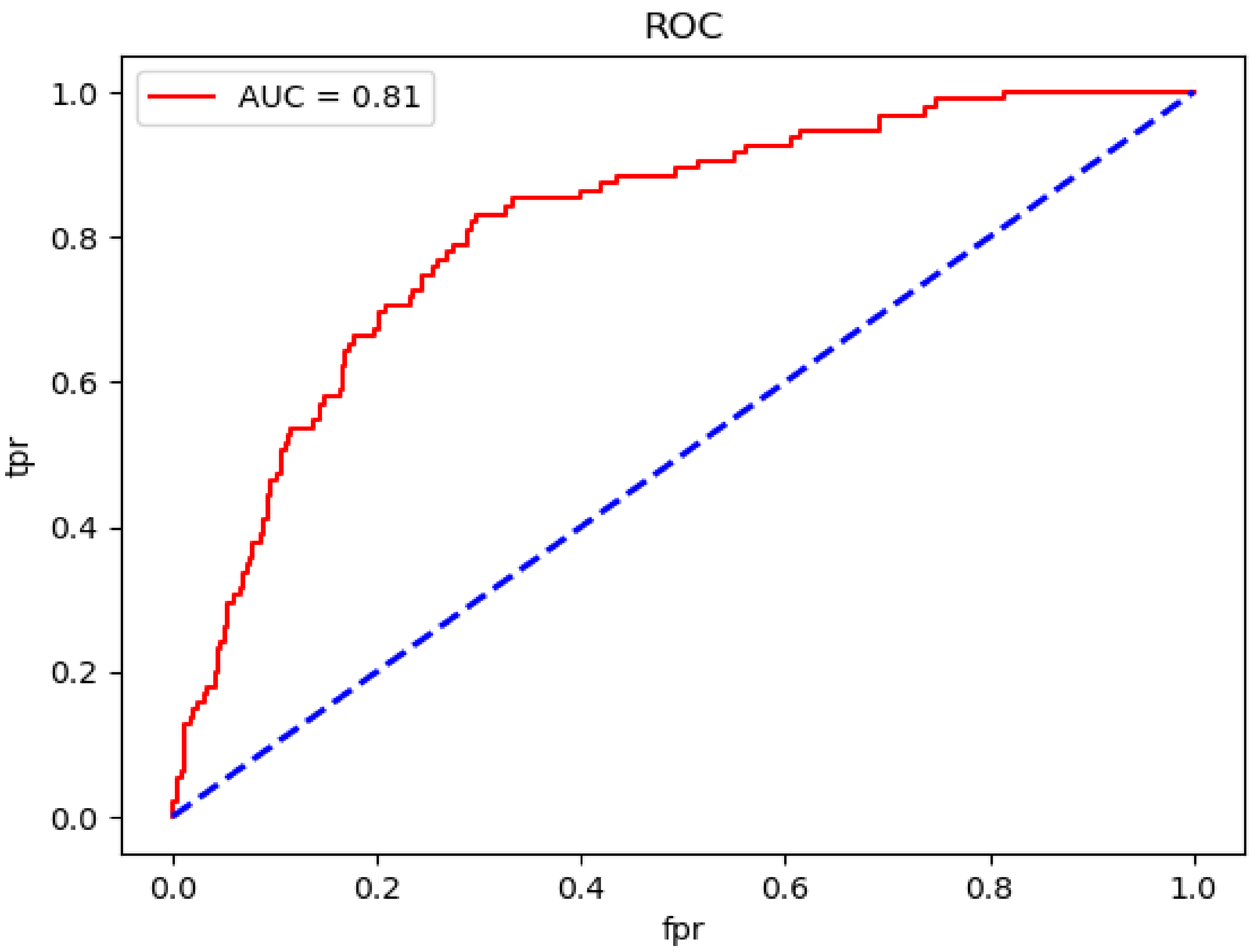
	precision	recall	f1-score	support
0	0.86	1.00	0.93	572
1	0.71	0.05	0.10	95
accuracy			0.86	667
macro avg	0.79	0.52	0.51	667
weighted avg	0.84	0.86	0.81	667

Final Model - Logistic Regression

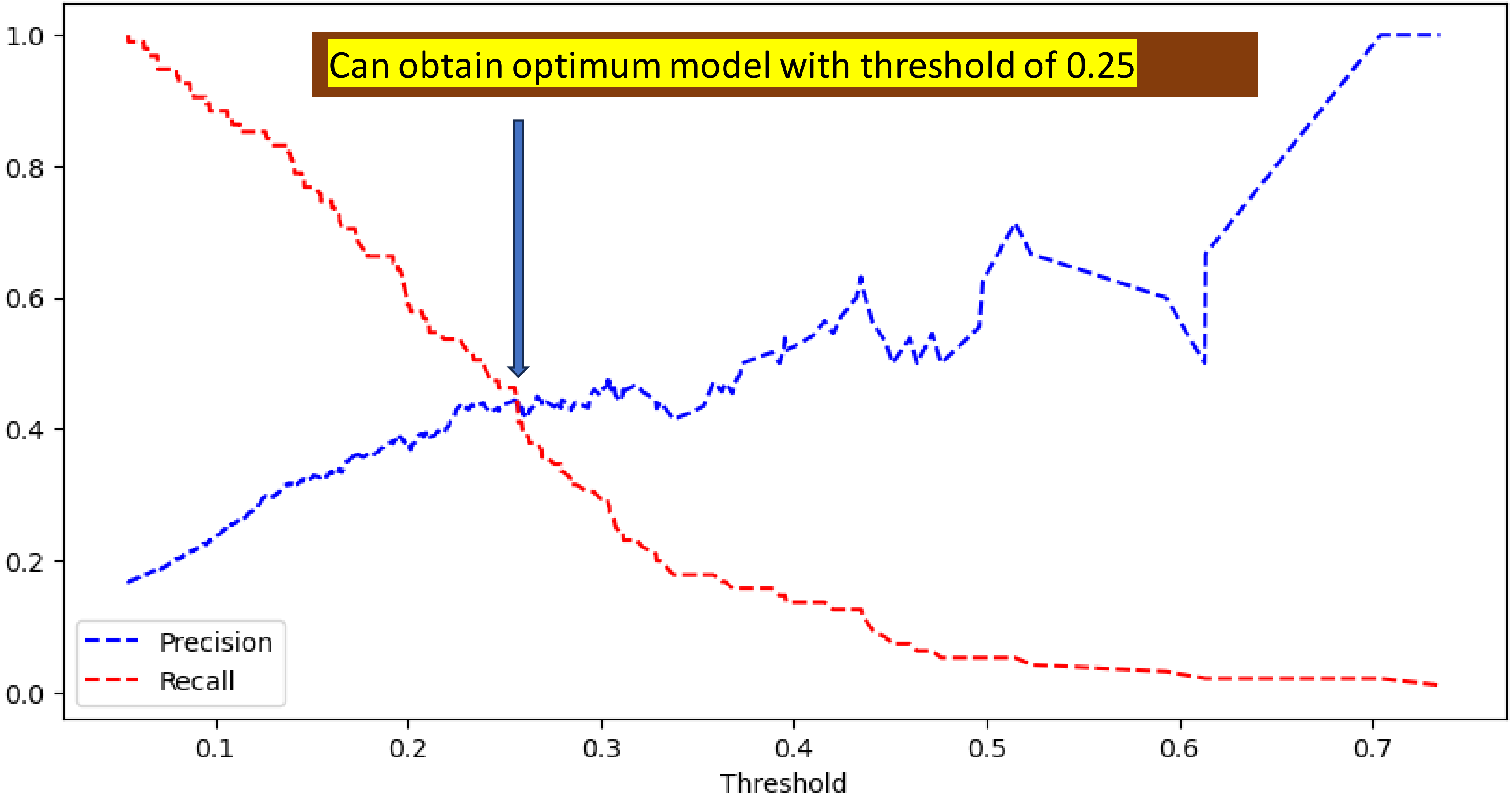
- The Logistic Regression algorithm works on probability threshold
- Default Value is 0.5
- So far all the metrics that we have analyzed works on threshold.
- To judge the performance of our model in better way, we should go for ROC_AUC score.

ROC_AUC Score & ROC Curve

ROC_AUC Score is **0.814**



Optimizing the Model using Precision and Recall Score



Optimized Model

	Negative	Positive
Negative	516	56
Positive	51	44

	precision	recall	f1-score	support
0	0.91	0.90	0.91	572
1	0.44	0.46	0.45	95
accuracy			0.84	667
macro avg	0.68	0.68	0.68	667
weighted avg	0.84	0.84	0.84	667