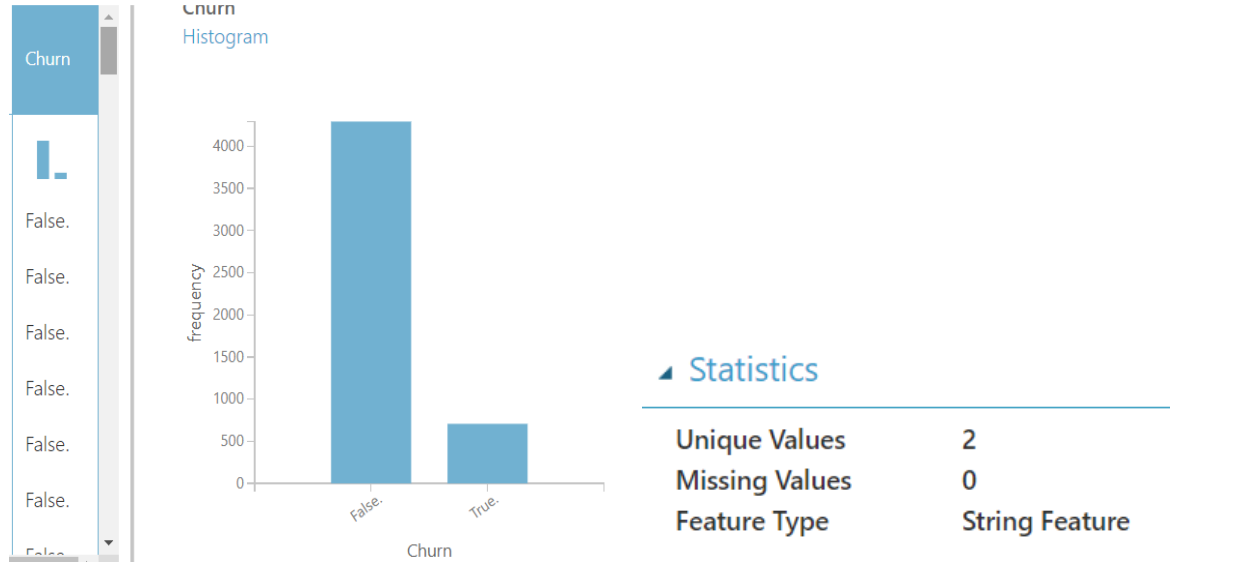
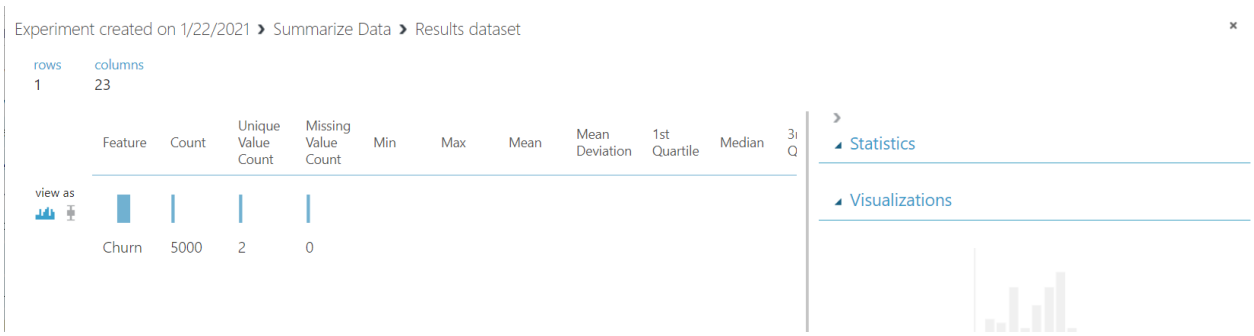
PROJECT – 2

Q1) Read the dataset and Check the target variable. State whether the data is imbalance or balance. (1points)

Ans.





We can see from the ratios the number of True and False the 2 values have drastic difference in the number. False is having 4293 i.e., 86% elements and True having 14% i.e., 707 elements. Yes, data is imbalance as we have a greater number of readings having target value false i.e., it's a biased data.

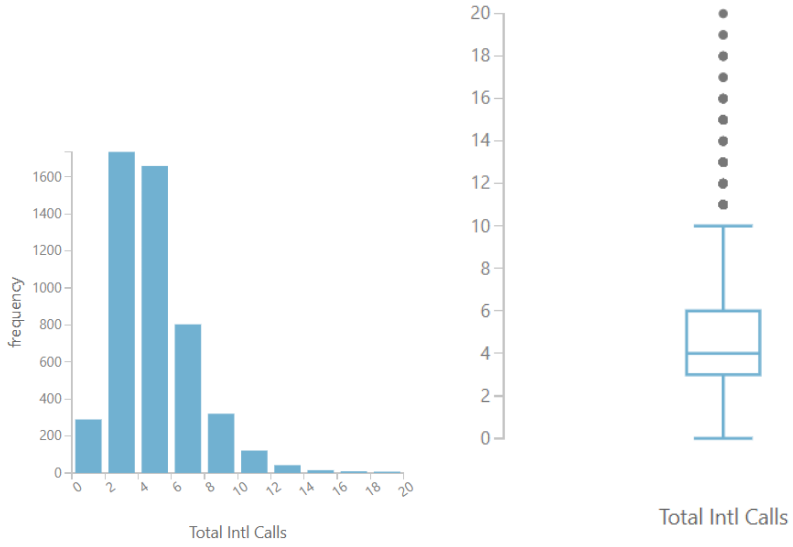
Q2) Check the distribution of Total international Calls. Convert it to normal distribution if it not in normal distribution. Please share the screenshot of distribution before and after doing the transformation. (3 points)

Ans.

Statistics

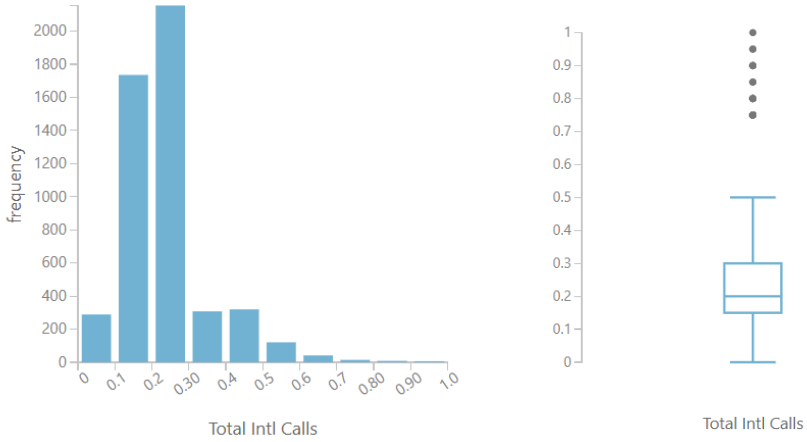
Mean	4.4352
Median	4
Min	0
Max	20
Standard Deviation	2.4568
Unique Values	21
Missing Values	0
Feature Type	Numeric Feature

Data before normalizing



Feature	Count	Unique Value Count	Missing Value Count	Min	Max	Mean	Mean Deviation	1st Quartile	Median	3rd Quartile	Median	3rd Quartile	Mode	Range	Sample Variance	Sample Standard Deviation	Sample Skewness	Sample Kurtosis	P0.5	P1	P5
Total Intl Calls	5000	21	0	0	20	4.4352	1.865836	3	4	6	4	6	3	20	6.035808	2.456788	1.360692	3.268184	1	1	1

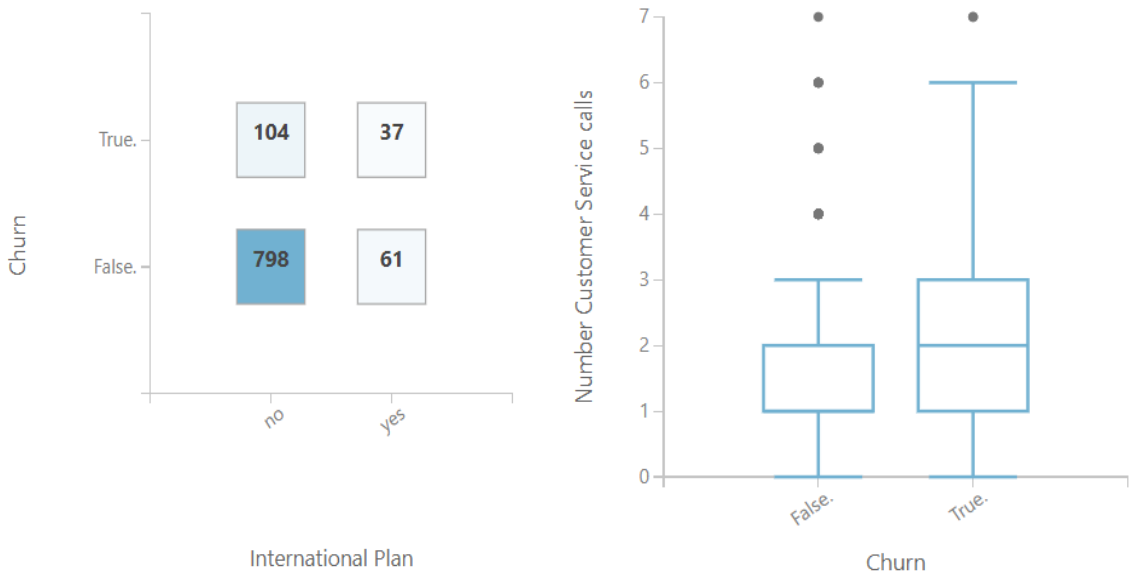
Data after normalizing



We did min max normalization on the total international calls data.

Q3) Study the relationship between International Plan Vs Churn & Churn vs Number of Customer Calls using Bivariate Analysis and state the inference clearly. (3 Points).

Ans.

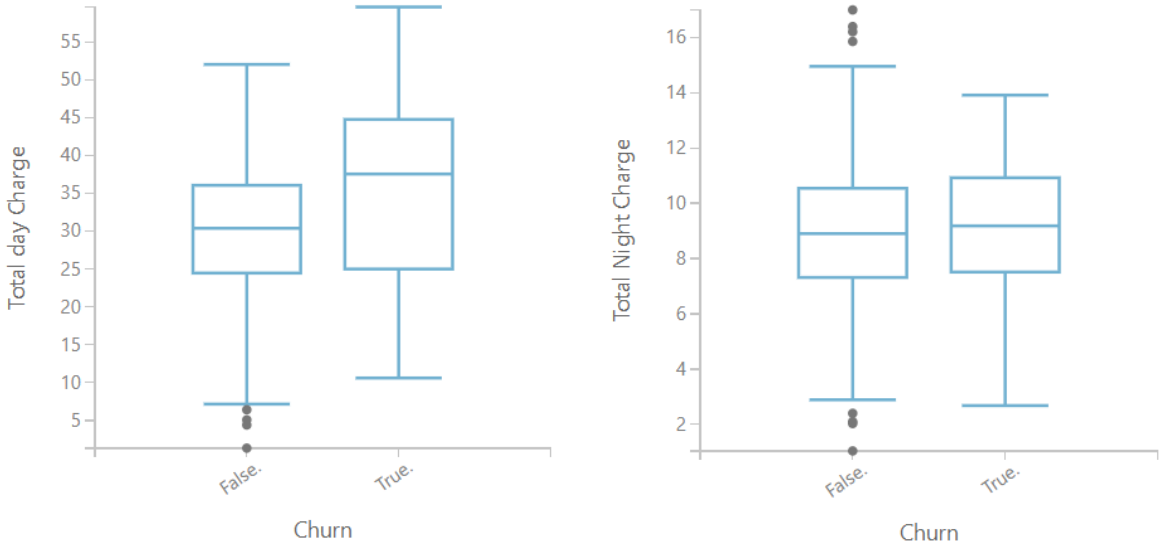


From the above graphs we can interpret that churn vs international plan gives crosstabwhile Number Customer service calls gives a box plot. Box plot is created because we have a nominal variable that is churn and a scaler variable i.e. Number Customer service calls this causes the formation off a box plot. While in first case it gives a cross tab as both the variables are nominal i.e. have only 2 values (yes/no and true/false).

We can conclude or interfere that Number Customer service is a statistical value while churn and international plan are nominal values.

Q4) Study the relationship between Churn vs Total Day Charge & Churn vs Total Night Charge using boxplot. State the inference Clearly. (3 Points)

Ans.



We can conclude from the box plots that the median of the first graph does not lie in the middle stating that the predicted values might be in consistent i.e. can't give accurate predictions while the second one is more consistent. While we can see from the second the Outliers also are on both sides making it more reliable. Outliers are also present in first graph but lies towards one half.

We can conclude that relationship between Total Night Charges and churn is more reliable and we can use Total Night charges for training as it gives more reliable results.

Q5) For further analysis split the data into 80:20(i.e., 80% train and 20% test) and train a logistic regression model. Evaluate the model using ROC-AUC curve, Accuracy, precision, recall, f1 score. (5 Points)

Ans.

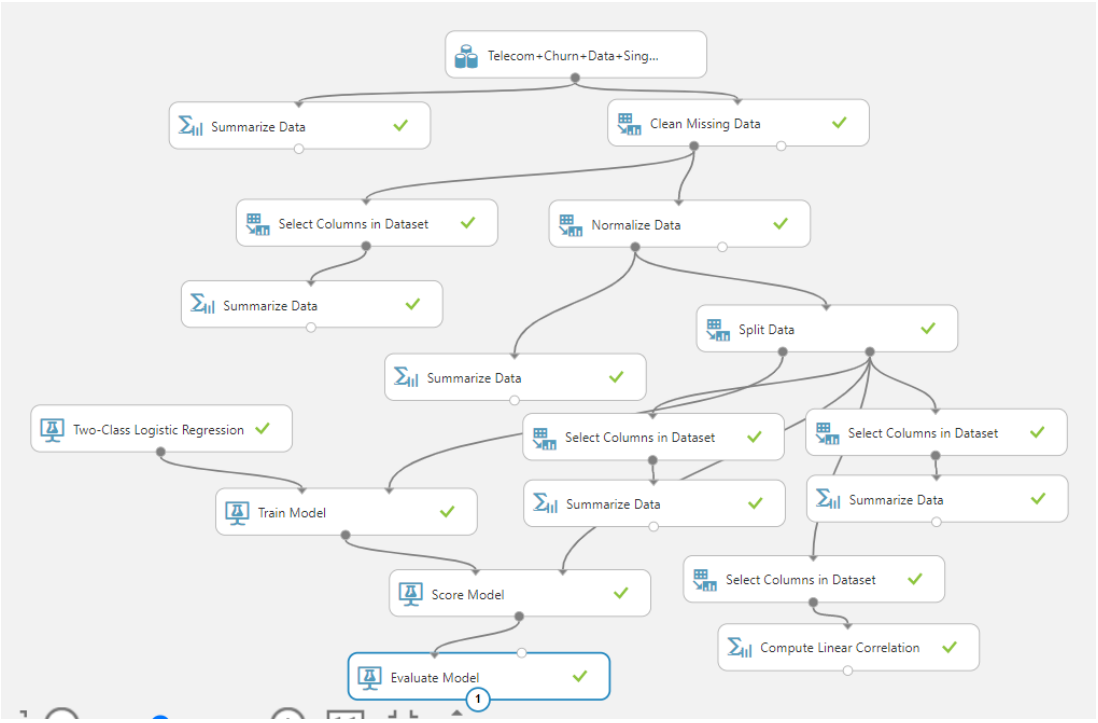
Model has been trained and results are as follows:

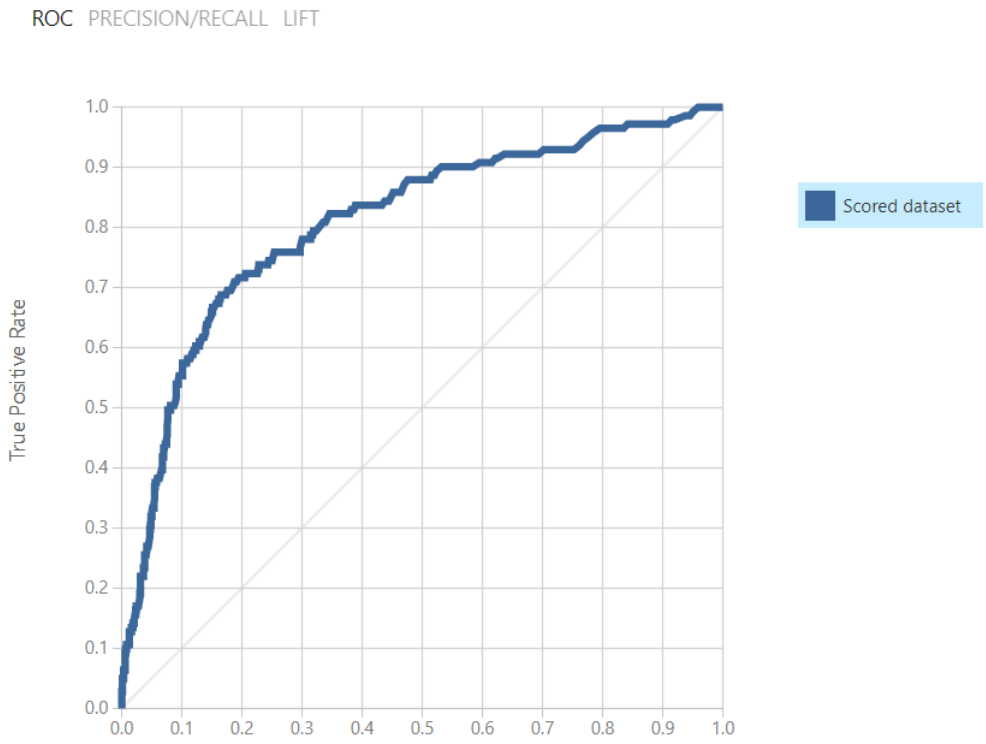
Accuracy – 0.867

Precision – 0.700

Recall – 0.099

F1 score - 0.174





True Positive	False Negative	Accuracy	Precision	Threshold	AUC
14	127	0.867	0.700	0.65	0.809

False Positive	True Negative	Recall	F1 Score
6	853	0.099	0.174

Positive Label	Negative Label
True.	False.

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	1	0	0.001	0.860	0.014	1.000	0.007	0.860	1.000	0.000
(0.800,0.900]	2	0	0.003	0.862	0.042	1.000	0.021	0.862	1.000	0.000
(0.700,0.800]	6	3	0.012	0.865	0.118	0.750	0.064	0.866	0.997	0.000
(0.600,0.700]	7	8	0.027	0.864	0.190	0.593	0.113	0.872	0.987	0.001
(0.500,0.600]	8	12	0.047	0.860	0.255	0.511	0.170	0.877	0.973	0.003
(0.400,0.500]	15	16	0.078	0.859	0.356	0.500	0.277	0.889	0.955	0.007
(0.300,0.400]	29	27	0.134	0.861	0.495	0.507	0.482	0.916	0.923	0.019
(0.200,0.300]	27	69	0.230	0.819	0.512	0.413	0.674	0.940	0.843	0.066
(0.100,0.200]	23	202	0.455	0.640	0.396	0.259	0.837	0.958	0.608	0.245
(0.000,0.100]	23	522	1.000	0.141	0.247	0.141	1.000	1.000	0.000	0.809