TELECOM CUSTOMER CHURN PREDICTION ASSESSMENT

Group#8

Abstract

Predicting Customer Churn for a Telecom Service Provider leveraging Logistic Regression

Vinod Kumar / Jithesh / Tarique / Shivkumar

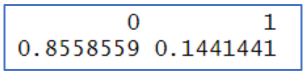
Contents

1. Data Summary

The data related to post-paid customers has the following characteristics: -

- Total number of observations – 3333
- Number of Variables – 11

Customers churned out form 14.5% of the total number of records and this could be an evidence of class imbalance.

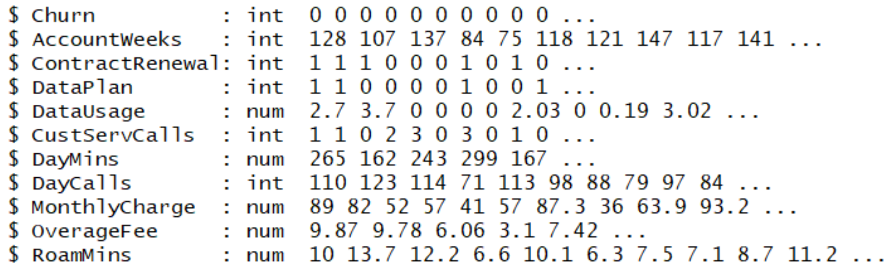


```
        0         1
0.8558559 0.1441441
```

Following tables give the variable definitions

Churn	1 if customer cancelled service, 0 if not
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

The variables have the following characteristics



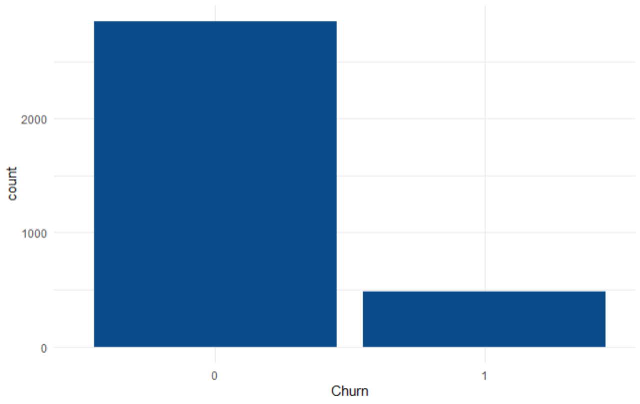
```
 $ Churn          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AccountWeeks   : int  128 107 137 84 75 118 121 147 117 141 ...
 $ ContractRenewal: int  1 1 1 0 0 0 1 0 1 0 ...
 $ DataPlan       : int  1 1 0 0 0 0 1 0 0 1 ...
 $ DataUsage      : num  2.7 3.7 0 0 0 0 2.03 0 0.19 3.02 ...
 $ CustServCalls  : int  1 1 0 2 3 0 3 0 1 0 ...
 $ DayMins        : num  265 162 243 299 167 ...
 $ DayCalls       : int  110 123 114 71 113 98 88 79 97 84 ...
 $ MonthlyCharge  : num  89 82 52 57 41 57 87.3 36 63.9 93.2 ...
 $ OverageFee     : num  9.87 9.78 6.06 3.1 7.42 ...
 $ RoamMins       : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
```

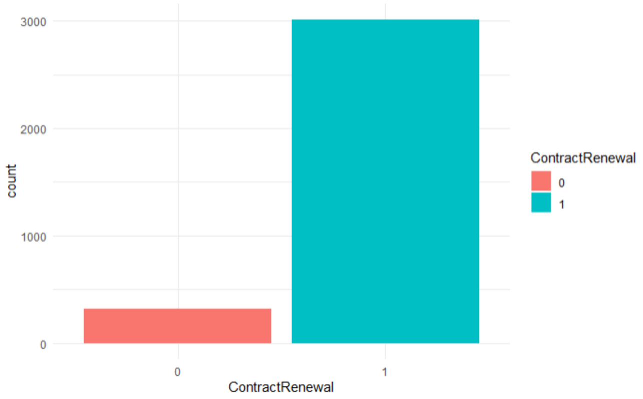
Since Churn Status, Contract Renewal and Data Plan reflect the status of customers, contracts and data plans, these variables are being converted to categorical variables (factors).

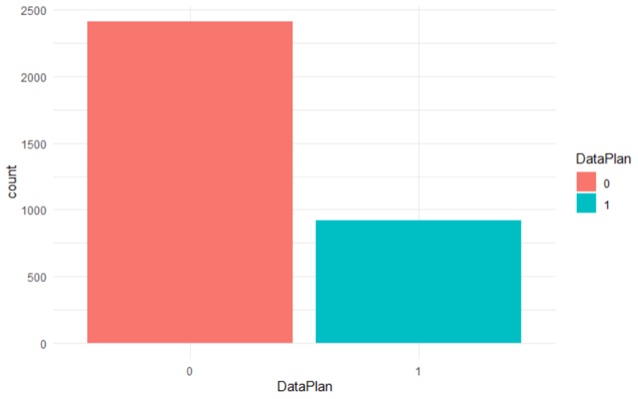
Missing Values: There are no missing values in the dataset, however variable Data Usage has 1813 0’s. This has been interpreted as those customers who do not have a Data Plan

2. Exploratory Data Analysis

1. Univariate Analysis





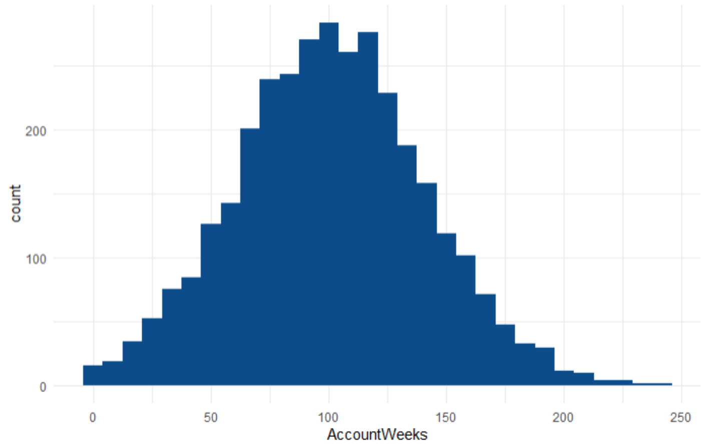


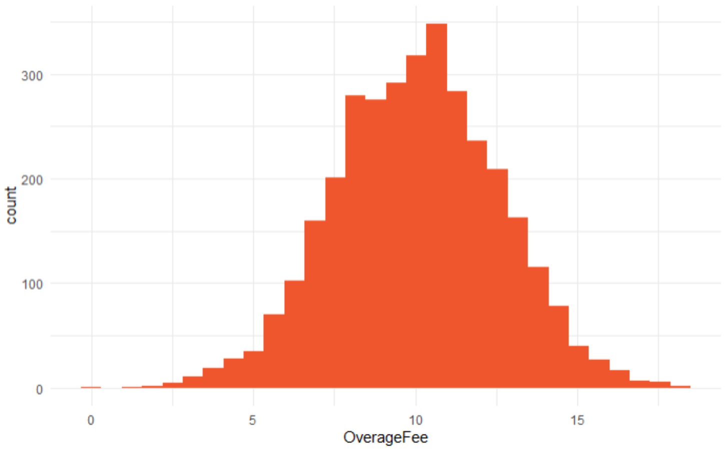
The following categorical variables show class imbalance as we can see from the graphs

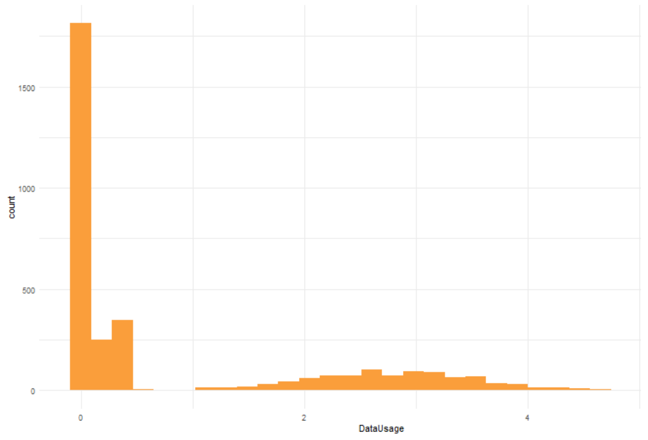
- Churn: 14.5%

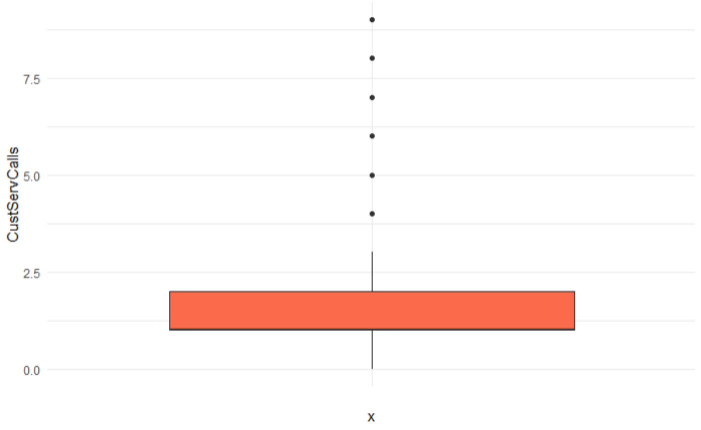
- Contract renewal: 9.03%

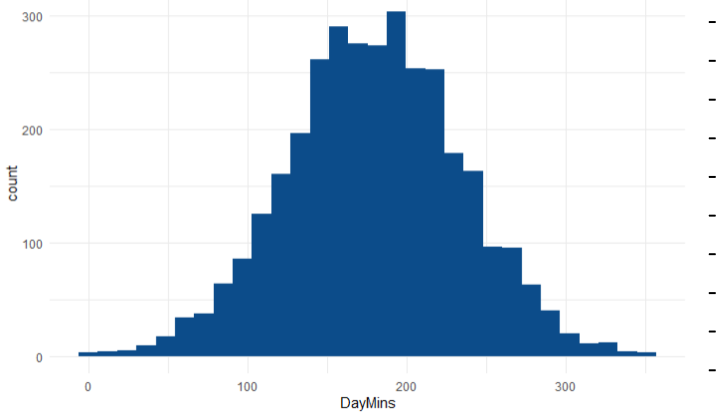
The dependent variable 'Churn shows higher amount of class imbalance which is assumed to bias the predictions in the majority class direction.

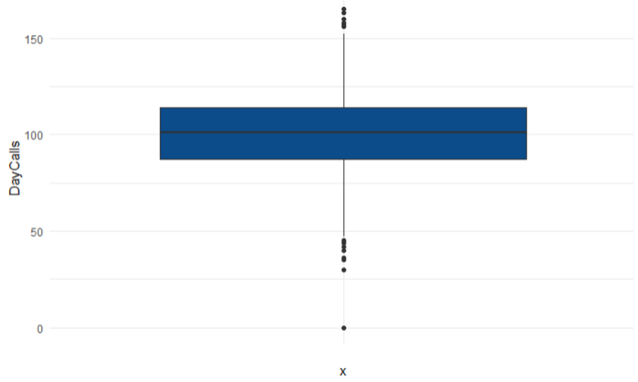


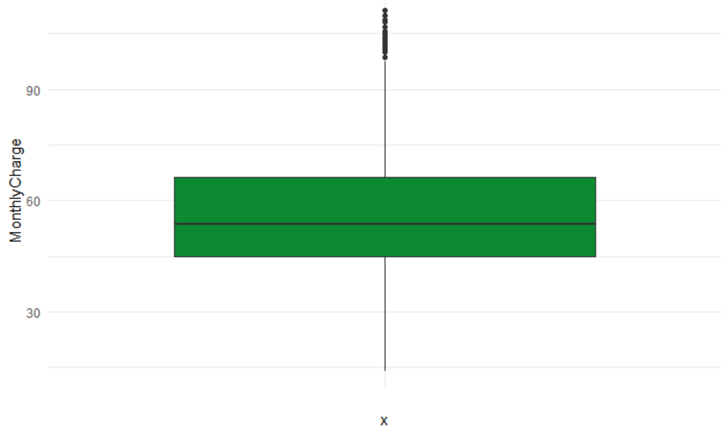


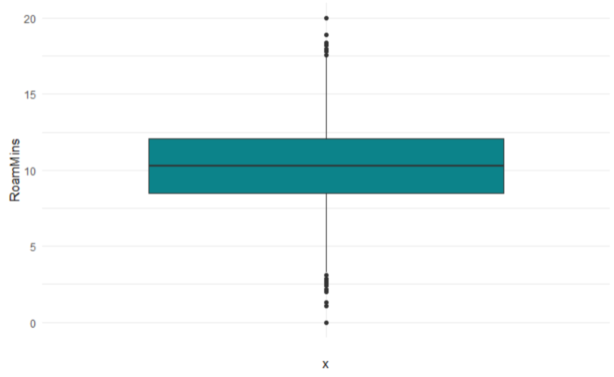








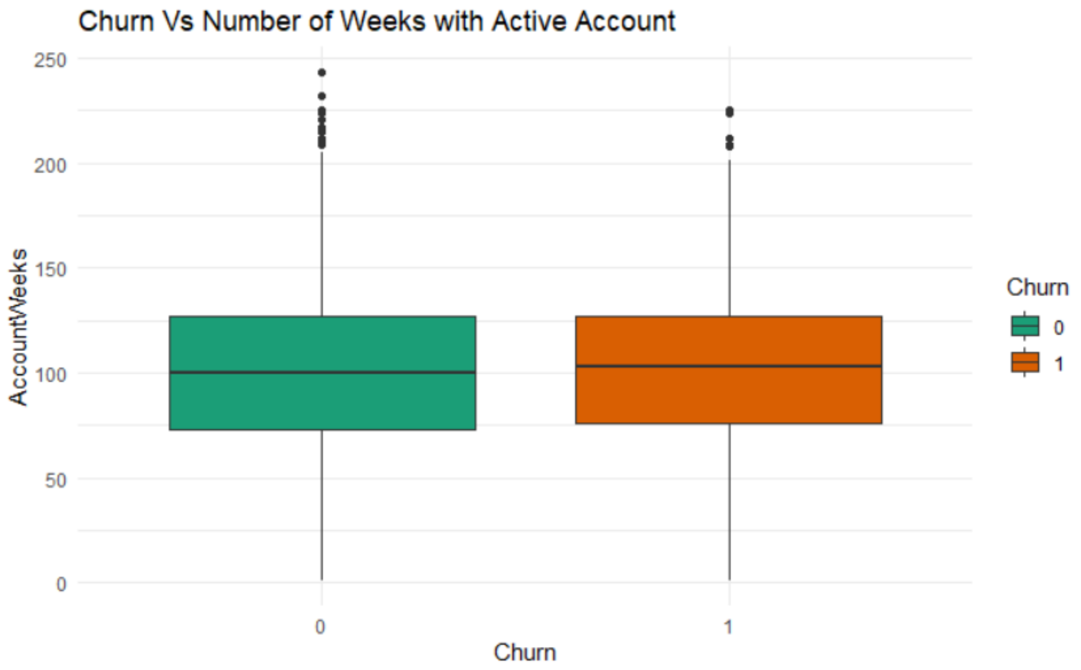




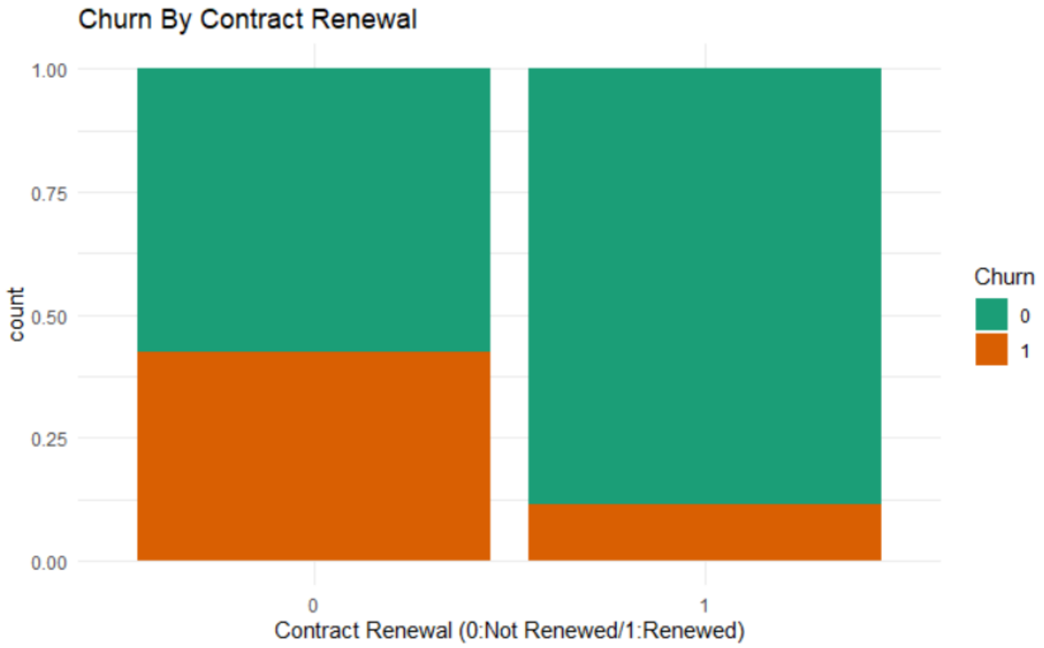
- ‘CustServCalls’ and ‘DataUsage’ are not normally distributed.
- Variable ’DataUsage’ have ‘0’ values & these correspond to those customers who do not have a data plan.
- Though there are **outliers** in Data Usage, this does not require treatment as they reflect customers without DataPlan
- All other variables by and large follow normal distribution with some outliers towards both extremes

2. Bivariate Analysis

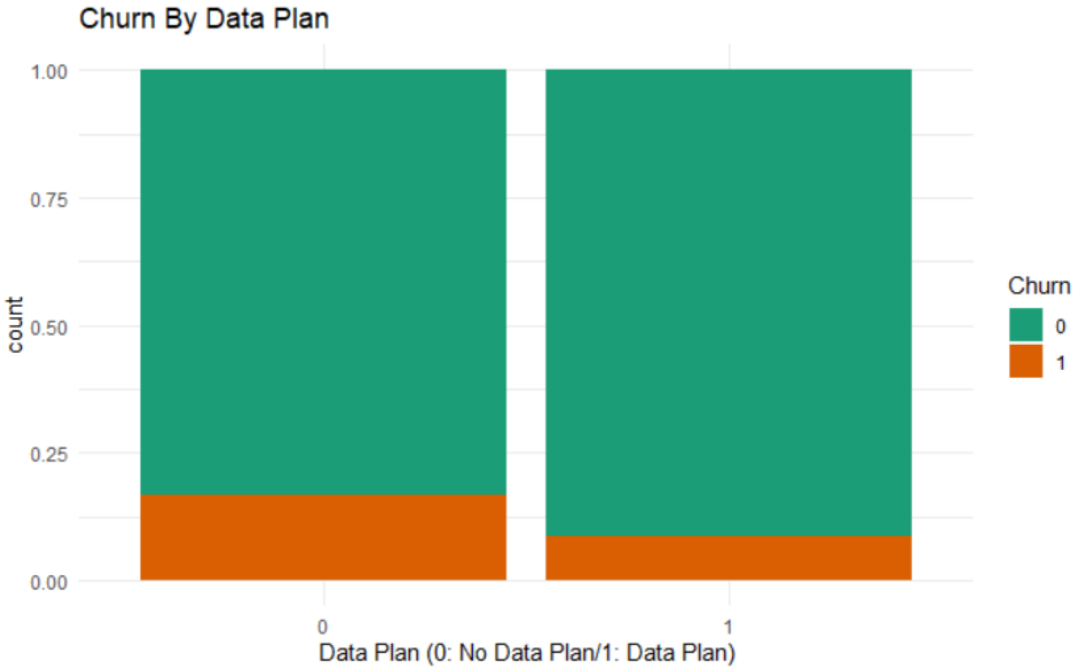
The number of weeks for which customers have had an active account are approximately the same. There is no significant relationship between Number of Weeks and Churn status as the p-value is greater than 0.05, i.e. 0.34



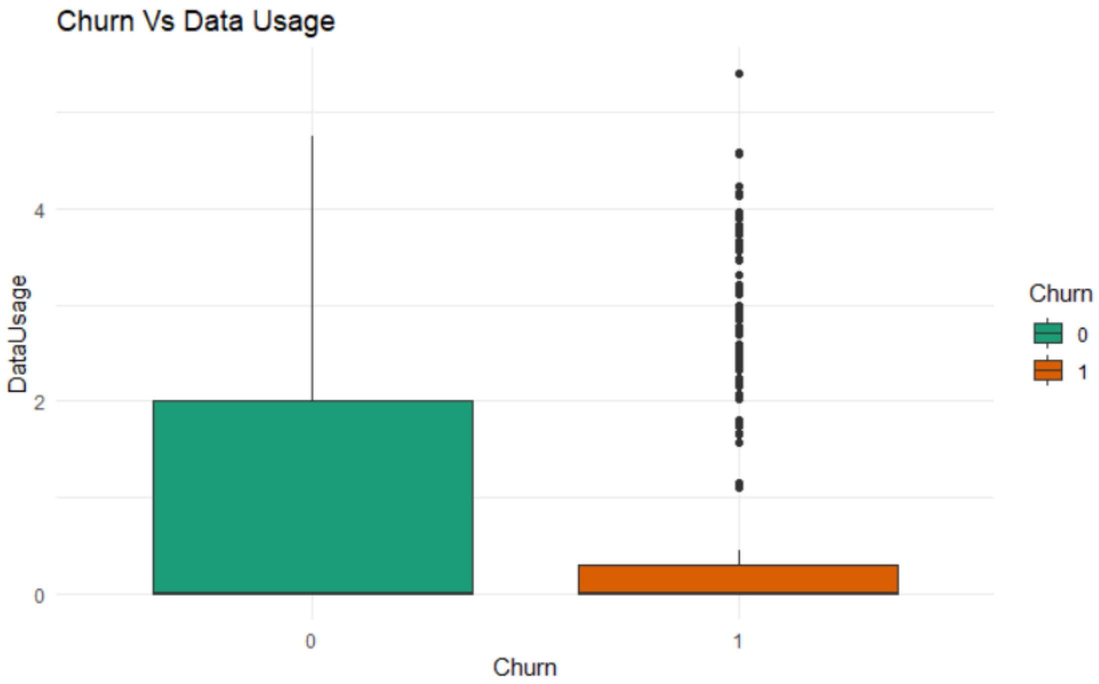
Those customers who have NOT RENEWED recently, have a HIGHER CHURN RATE than those who have renewed recently. There are customers, though they have not renewed the contract recently, have still not cancelled the service.



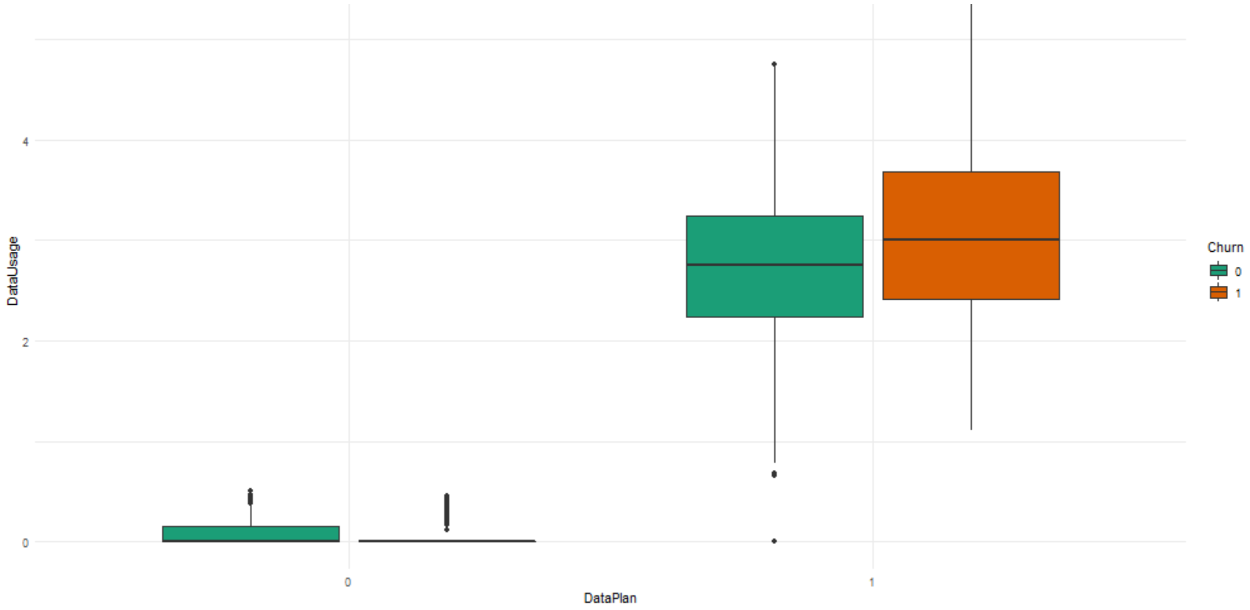
Those customers who DO NOT have a data plan have churned out more than those who have a data plan. This could perhaps mean that customers who have churned out were not happy with the talk-time services.



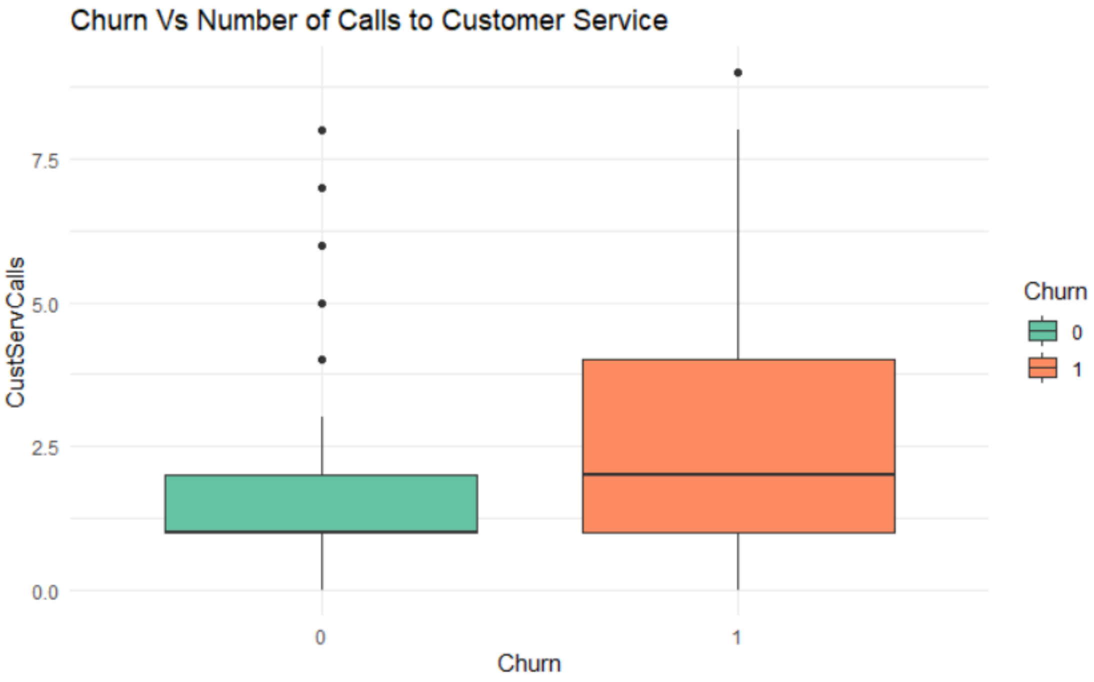
There are outliers amongst customers who have churned out with respect to monthly data usage.



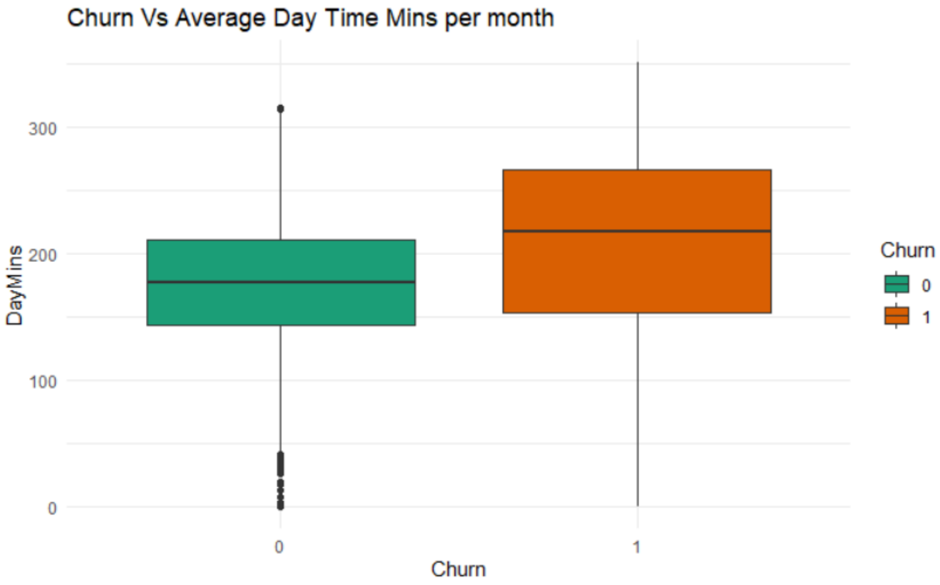
Customers who churned out and have a data plan show a higher median data usage than those who did not churn.



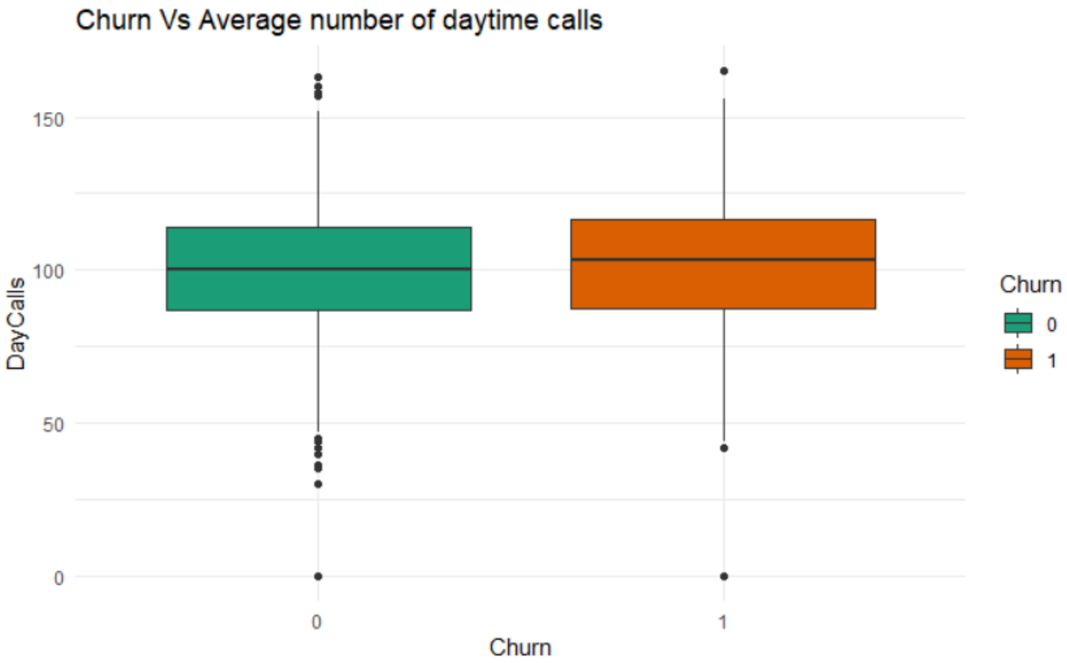
Customers who have churned out have made a higher number of calls to customer service than those who have not yet cancelled the service. This is also a reflection on the quality of the customer service provided.



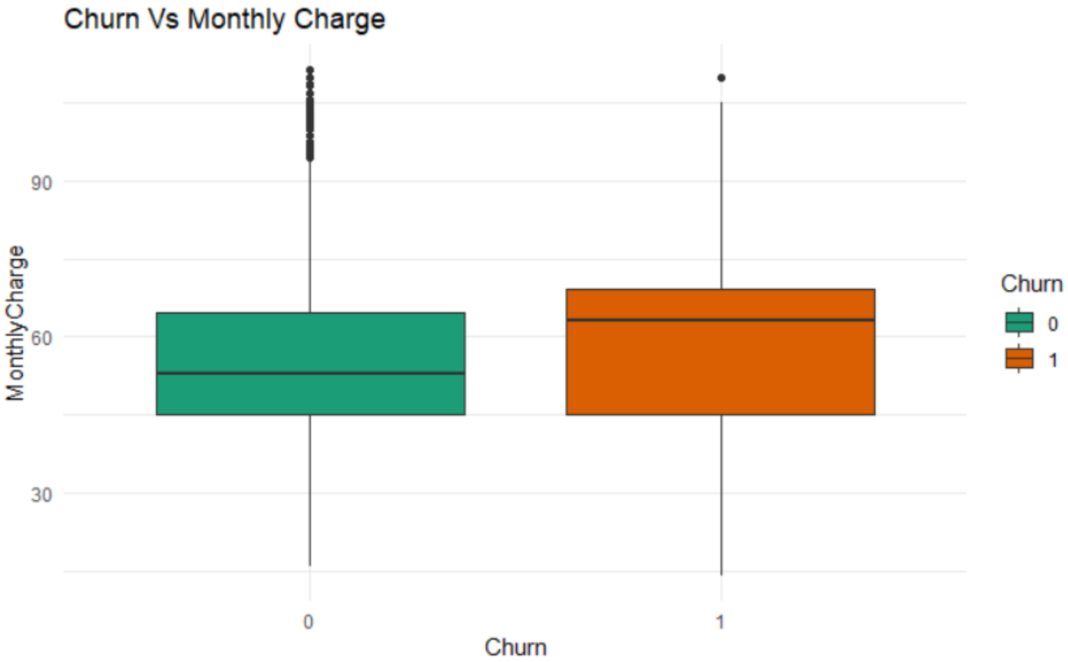
The customers who have churned out have clocked a higher average of daytime minutes per month than those who have not cancelled the service.



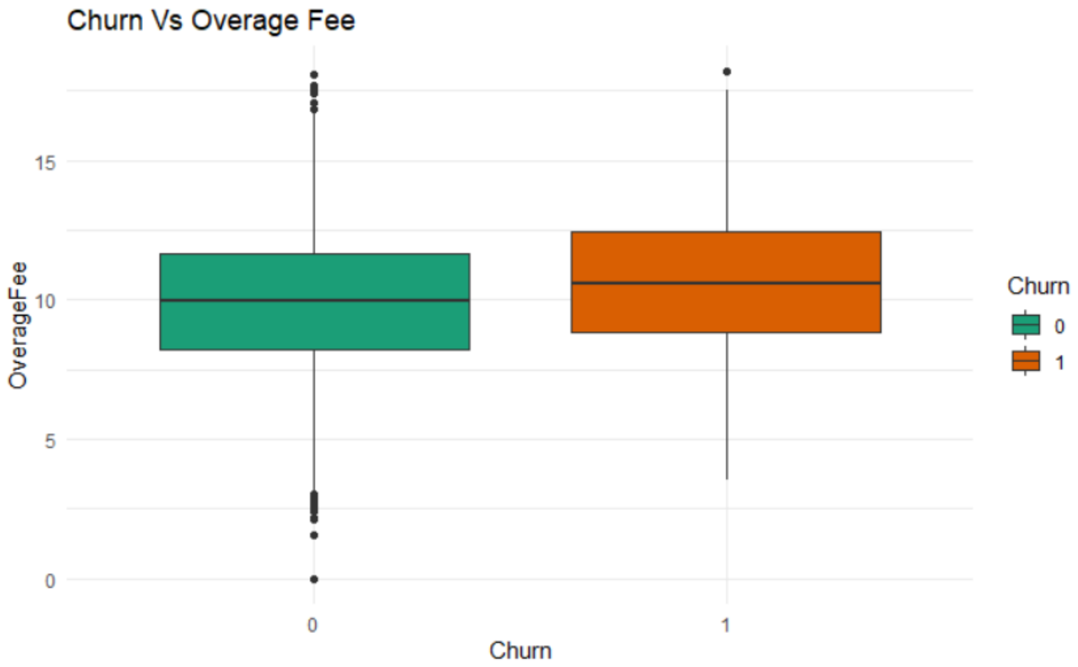
The customers who have churned out have marginally clocked a higher average number of daytime calls than those who have not churned out.



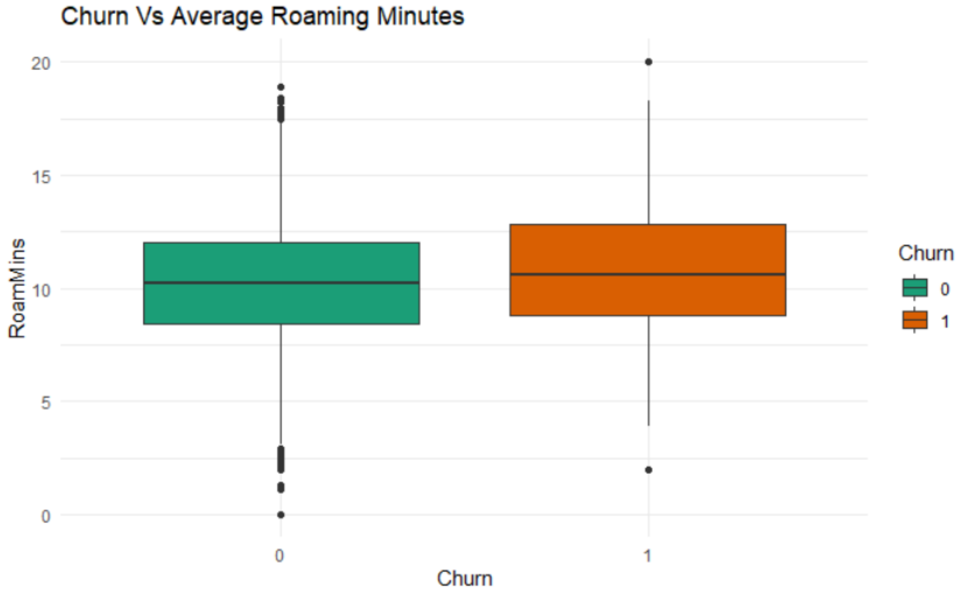
Customers who have churned out have paid a higher Monthly Charge than those who have not churned out. Higher monthly charges could be a reason for customers churning out.



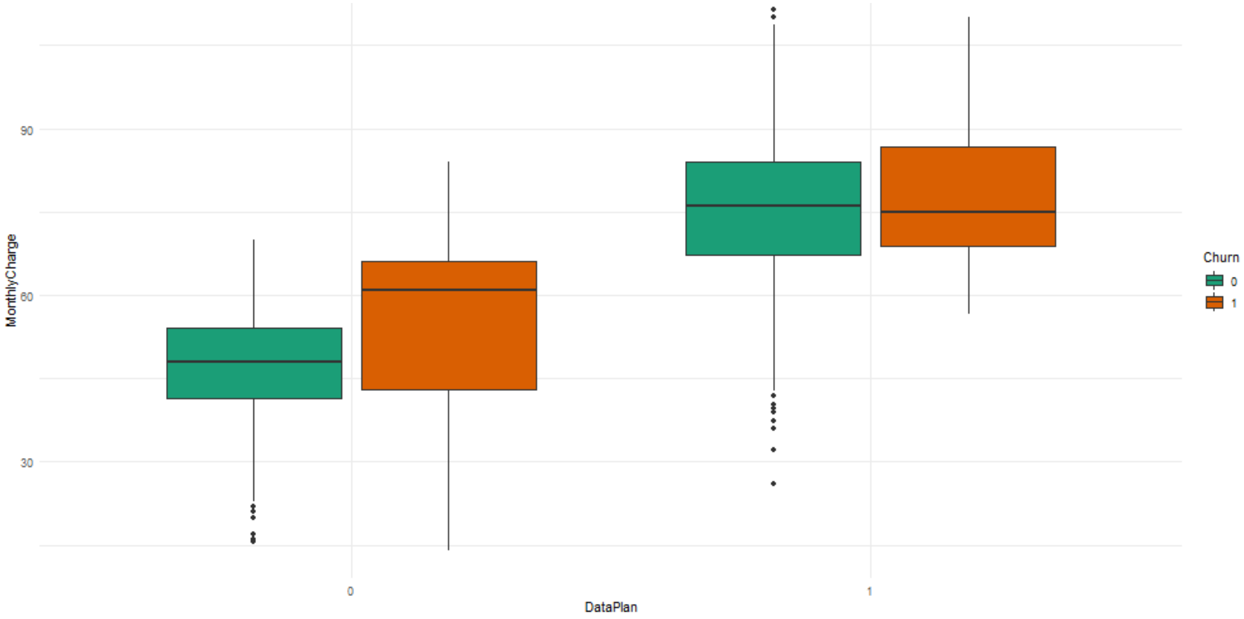
Overage charges are incurred when usage is more than the fixed quota under a post-paid plan. Customers who have churned out have paid a higher overage fee than the customers who have not churned out. Another reason for customers churning out, could be that customers may have moved to another provider who had more flexible options.



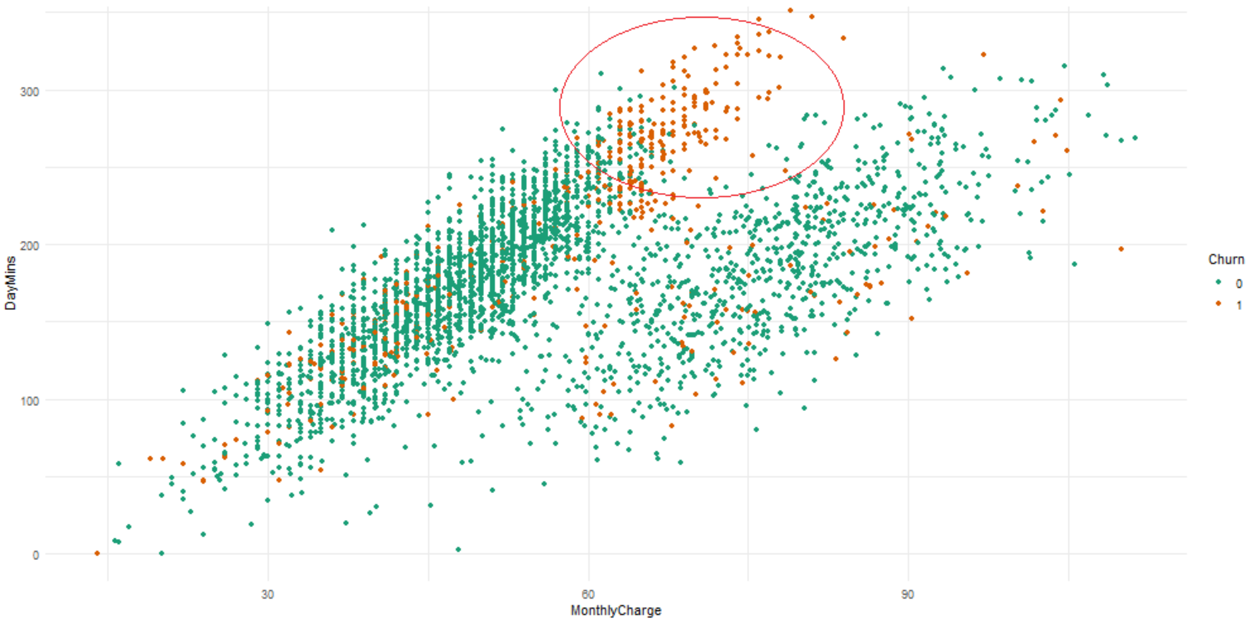
Customers who have churned out, have clocked higher average in terms of roaming minutes than the retained customers.



The monthly charge of customers who churned out having no data plan is higher than those who did not churn out



The customers who have churned out have been paying higher monthly charge on account of higher consumption in terms of DayMins



Correlation Plot

There could be multi-collinearity between the following pairs of variables: -

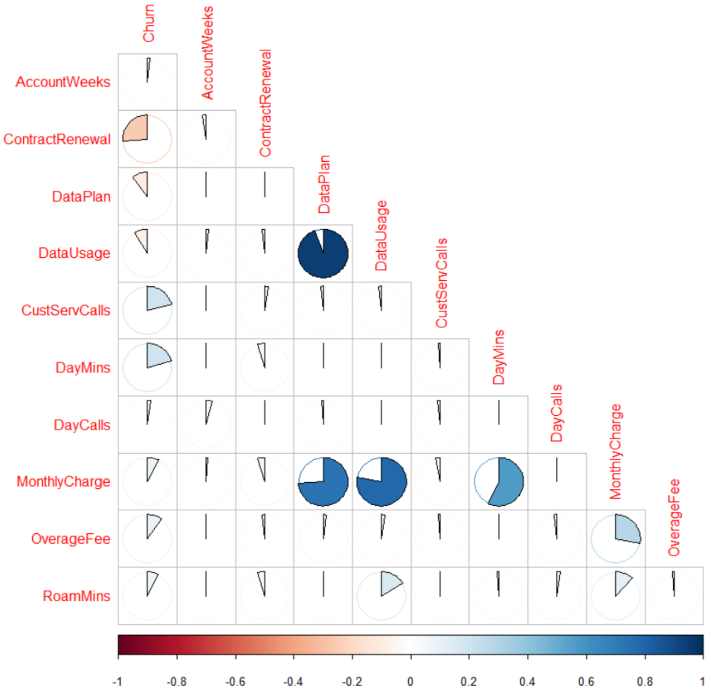
- Data Usage and Data Plan
- Monthly Charge and Data Plan
- Monthly Charge and Data Usage
- Monthly Charge and Day Minutes

Customers churning out has a POSITIVE correlation with the following variables (in decreasing order): -

1. Number of calls made to customer service
2. Average daytime minutes per month
3. Overage Fee
4. Monthly Charge
5. Roaming Minutes
6. Average number of daytime calls
7. Number of weeks customer has had active account

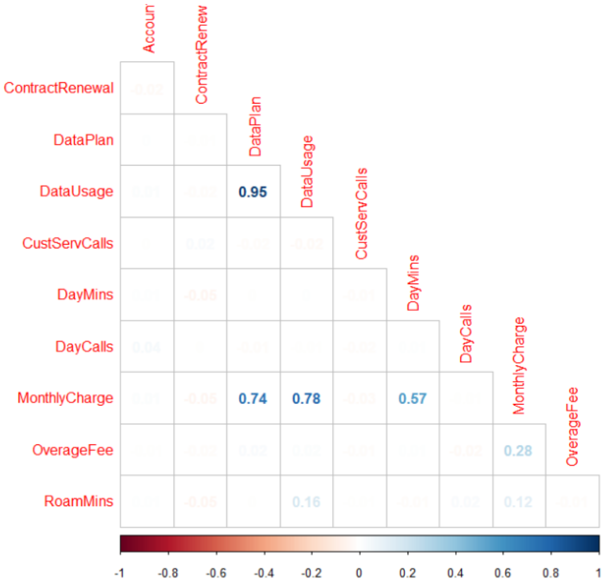
Customers churning out has a NEGATIVE correlation with the following variables (in increasing order): -

1. Contract Renewal
2. Data Plan
3. Data Usage



Correlation / Multi-Collinearity

There exists a high degree of multi-collinearity between the following predictor variables and the correlation values along with p-values give an indication of the same: -



Variable 1	Variable 2	Correlation Value	p-value
Data Usage	Data Plan	0.95	< 2.2 * 10^{-16}
Monthly Charge	Data Plan	0.7374	< 2.2 * 10^{-16}
Monthly Charge	Data Usage	0.7816	< 2.2 * 10^{-16}
Monthly Charge	Day Mins	0.5679	< 2.2 * 10^{-16}

The correlation between the above variables is significant and will impact the performance of the model.

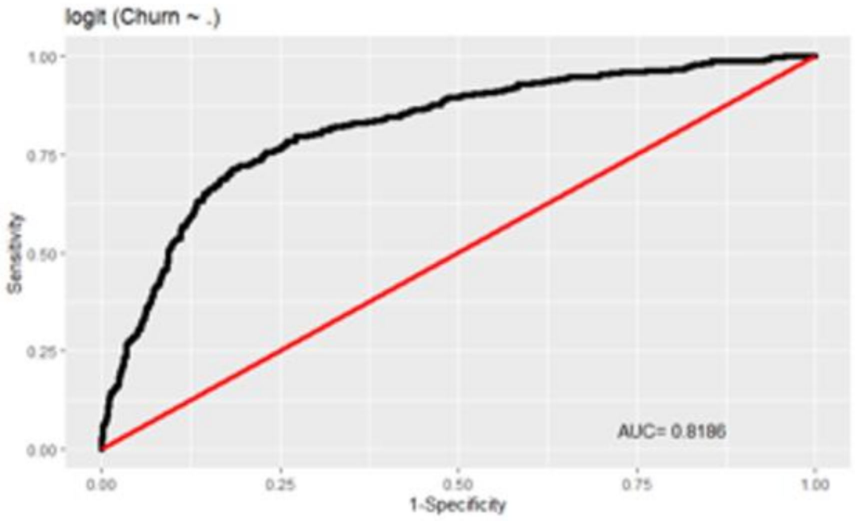
EDA Interpretation and Observations

Based on the initial analysis performed on the base data, following could be the reasons for customers churning out: -

- Dissatisfied with the customer service quality
- Higher Monthly charges and overage fees
- Unhappy with the DataUsage/Call-time/Roaming options provided by the service provider

3. Logistical Regression Model

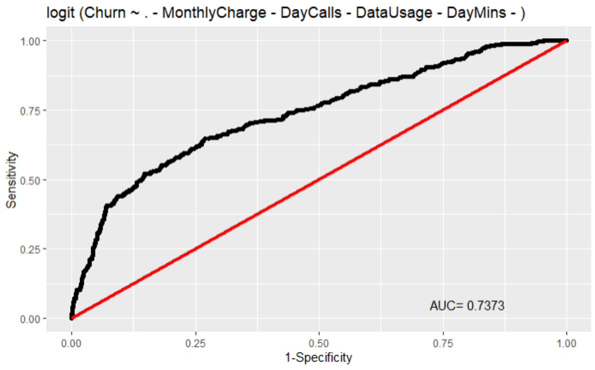
Full Model

Model Performance Metrics	Measures
McFadden R^2	0.20
Significant Variables	• Contract Renewal1 • DataPlan1 • CustServiceCalls • RoamMins
Accuracy	0.860326
Sensitivity	0.188791
Specificity	0.974436
High VIF Variables	• DataPlan • DataUsage • DayMins • Monthly Charge • Overage Fee
Area Under Curve	• 81.86%
	

Model#2 with Significant Variables

The arrows indicate the change over the Model#1

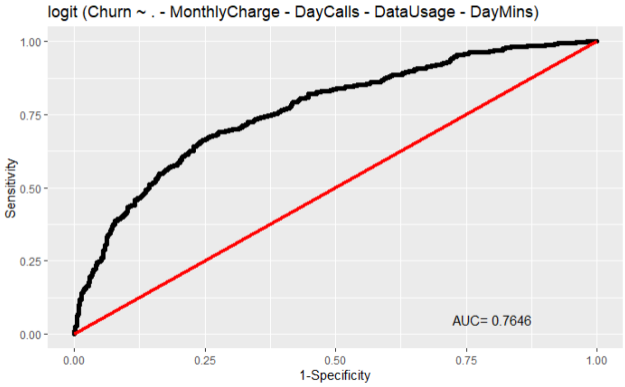
Model Performance Metrics	Measures
McFadden R^2	0.12 ⬇
Variables Dropped	1. Account Weeks 2. Data Usage 3. DayMins 4. DayCalls

	5. Monthly Charge 6. OverageFee
Significant Variables	1. Contract Renewal1 2. DataPlan1 3. CustServiceCalls **4.** RoamMins
Accuracy	0.855184 ⬇
Sensitivity	0.115044 ⬇
Specificity	0.980952 ⬆
Area Under Curve	73.73% ⬇
	

Model#3 – Reducing Multi-Collinearity

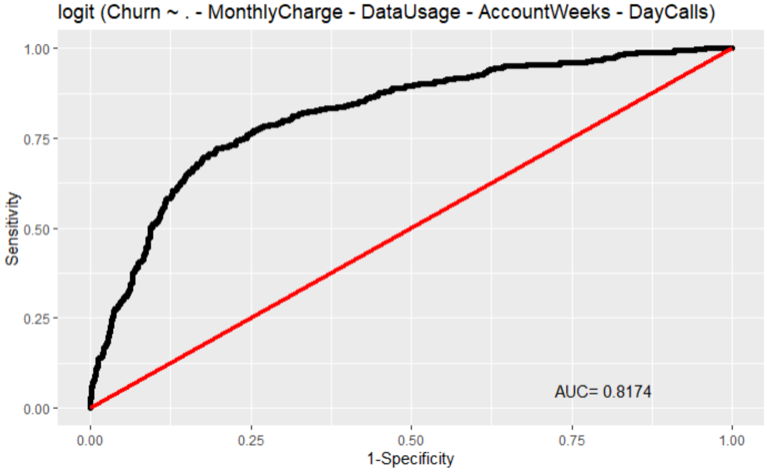
The arrows indicate the change over the Model#2

Model Performance Metrics	Measures
McFadden R^2	0.14 ⬆
Variables Dropped (High VIF Values)	1. Data Usage 2. DayMins 3. DayCalls 4. Monthly Charge
Significant Variables	1. Contract Renewal1 2. DataPlan1 3. CustServiceCalls 4. Overage Fee **5.** RoamMins
Insignificant Variables	1. Account Weeks
Accuracy	0.861183 ⬆
Sensitivity	0.147493 ⬆
Specificity	0.982456 ⬆
Area Under Curve	76.46% ⬆



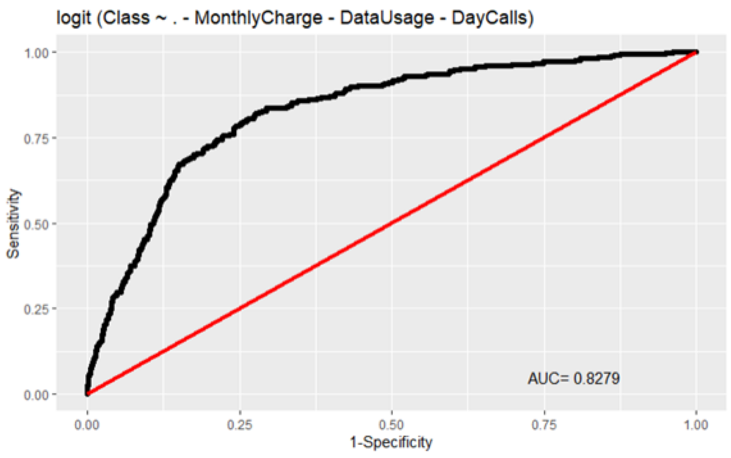
Model#4 – Reducing Multi-Collinearity and Dropping Insignificant Variables

The arrows indicate the change over the Model#3

Model Performance Metrics	Measures
McFadden R^2	0.20 ↑
Variables Dropped (High VIF Values)	1. MonthlyCharge 2. DataUsage
Variables Dropped (Insignificant)	1. Account Weeks 2. Day Calls
Significant Variables	1. Contract Renewal1 2. DataPlan1 3. CustServiceCalls 4. DayMins 5. Overage Fee **6.** RoamMins
Accuracy	0.858612 ↓
Sensitivity	0.179941 ↑
Specificity	0.973935 ↓
Area Under Curve	81.74 ↑
	

Model#5 – Up Sampling

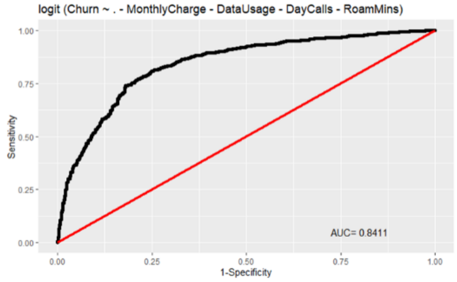
Up Sampling technique is leveraged to overcome class imbalance as 14.5% of the data is related to customer churn. The arrows indicate the change over the Model#4

Model Performance Metrics	Measures
McFadden R^2	0.25 ⬆
Variables Dropped (High VIF Values)	1. MonthlyCharge 2. DataUsage
Variables Dropped (Insignificant)	1. Day Calls
Significant Variables	1. Account Weeks 2. Contract Renewal1 3. DataPlan1 4. CustServiceCalls 5. DayMins 6. Overage Fee **7.** RoamMins
Accuracy	0. 764912 ⬇
Sensitivity	0. 754887 ⬆
Specificity	0. 774937 ⬇
Area Under Curve	82.79% ⬆
	

Model#6 – Smote Sampling

The arrows indicate the change over the Model#5

Model Performance Metrics	Measures
McFadden R^2	0.27 ⬆
Variables Dropped (High VIF Values)	1. MonthlyCharge 2. DataUsage
Variables Dropped (Insignificant)	1. Day Calls 2. RoamMins
Significant Variables	1. Account Weeks 2. Contract Renewal1

	3. DataPlan1 4. CustServiceCalls 5. DayMins 6. Overage Fee
Accuracy	0. 764855 ↔
Sensitivity	0. 6647 ⬇
Specificity	0. 8399 ⬆
Area Under Curve	84.11% ⬆
	

4. Model Performance Measures

The following model performance measures were arrived at based on a cut-off / threshold of 50%: -

Models		Train Data			Test Data		
Model#	*Model Scope*	*Sensitivity*	*Specificity*	*Accuracy*	*Sensitivity*	*Specificity*	*Accuracy*
Model 1	Full Model	0.188791	0.974436	0.860326	0.166667	0.978947	0.861862
Model 2	With Significant Variables only	0.115044	0.980952	0.855184	0.145833	0.984795	0.863864
Model 3	Reducing Multi-Collinearity	0.147493	0.982456	0.861183	0.1875	0.983626	0.868869
Model 4	Reducing Multi-Collinearity & Dropping Insignificant Variables	0.179941	0.973935	0.858612	0.1875	0.981287	0.866867
Model 5	Up Sampling	0.754887	0.774937	0.764912	0.784722	0.750877	0.755756
Model 6	Smote sampling	0.6647	0.839971	0.764855	0.597222	0.792982	0.764765

Model#5 has been tuned based on the threshold values of 0.5, 0.43 and 0.39 and the following table gives the model performance measures: -

Model#5 Thresholds	Sensitivity	Specificity	Accuracy	Model predicting Churn / Actual Retain	Model predicting Churn / Actual Churn	Number of Customers to be targeted for promotional campaign	% Increase
Threshold > 0.5	0.7847	0.7509	0.7558	213	113	326	
Threshold > 0.43	0.8333	0.6737	0.6967	279	120	399	22%
Threshold > 0.39	0.8611	0.6199	0.6547	325	124	449	38%

Model#5 with a threshold of **0.43** is the recommended model as Sensitivity increased to **83.33%**, predicting more number of customer churns as compared to a threshold value parameter of 0.5. However, the model performance for the threshold of **0.43** drops in Specificity **by 8%**.

Though promotional offers will go out to a larger customer base (increase of **22%**), it will reduce the number of customers churning out.

Model#5 gives the highest value in terms of Sensitivity for the Test Data and the following variables make up this model along with their explanatory power

Variables	**Definition**	**Probability of Customer Churn**
CustServiceCalls	Number of calls into customer service	64.61%
Overage Fee	Largest overage fee in last 12 months	53.55%
RoamMins	Average number of roaming minutes	51.64%
DayMins	Average daytime minutes per month	50.35%

AccountWeeks	Number of weeks customer has had active account	50.06%
DataPlan1	Customer has data plan	30.24%
ContractRenewal	Customer recently renewed contract	9.74%

5. Actionable Insights & Recommendations

Customers are churning out for the following reasons based on the model results: -

- Not happy with the customer service quality
- Paying a high fee when they tend to extend beyond the service provider plan limit.
- Not happy with the roaming options provided by the service provider
- Not happy with the talk-time (DayMins) options provided as a part of the plan
- Not using the data option provided as a part of the plan
- Not happy with the service post renewal, inspite of being an active customer

The Telecom Service Provider needs to: -

- Improve the customer service, perhaps by faster and effective resolution of customer complaints
    - Position promotional offerings to customers who have logged a higher number of service calls
- Segment customers based on the following and cross-sell / up-sell plans with appropriate options:
    - Data Usage
    - Talk-Time
    - Roaming Options
    - High Data Usage & Talk-Time