MIS-64036: Business Analytics

Fall 2022

ABC-Wireless Inc- Churn Prediction

Submitted By:

(Group 5)

Abinaya Sundari Panneerselvam	Introduction, Data Exploration, PPT
Joshna Katta	Introduction, Data Exploration, PPT
Venkata Naga Siddartha Gutha	Data Exploration, Model Strategy, Model Performance, Insights, PPT
Gowtham Chakri Mallepaka	Data Exploration, Insights, Conclusion, PPT

Table of Contents

Project Goal

Customers in the Telecom industry have a high tendency to switch to other service providers on a regular basis. Churn has a direct impact on the revenue of the company. When a customer leaves, the company not only loses the future revenue from that customer but also the resources spend to acquire that customer. This makes Churn one of the most problems in the Telecom industry. So, the companies in this industry try to identify factors that might affect churn directly or indirectly. Once identified they take appropriate steps to reduce the churn rate. This project builds a model to predict churn using the data provided by ABC Wireless Inc. This model will be helpful to identify customers who are likely to churn and with this information, the company could formulate strategies to retain those customers.

Overview of data

ABC Wireless Inc. has provided its historical data to use in order to help them with the customers' churn issue. We will be working on this data to build a model that can predict their customers who are likely to churn.

This dataset has both numerical and

categorical variables with 20 features and 3333 Customers. The features of the dataset

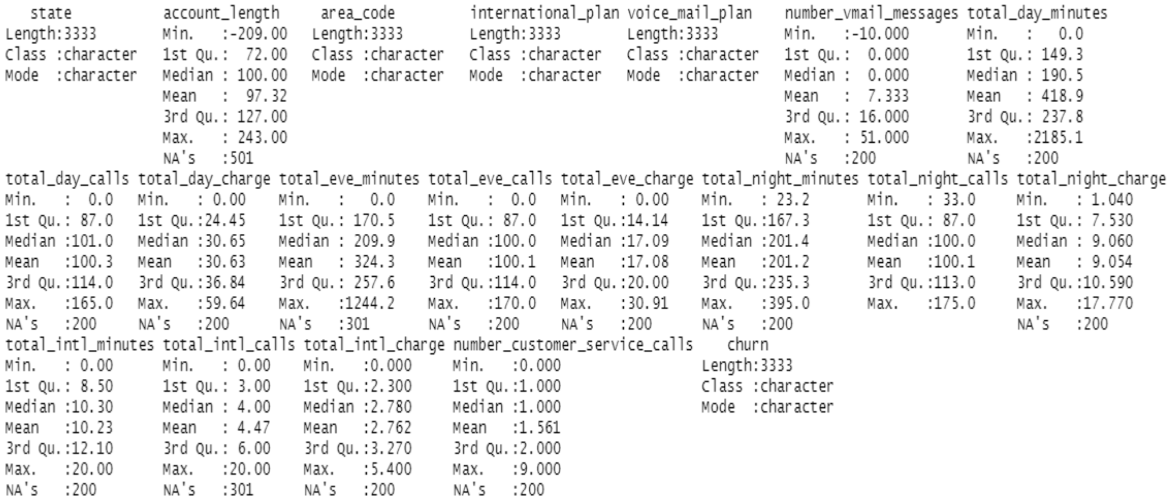
are:

- state (categorical)
- account_length
- area_code
- international_plan (yes/no)
- voice_mail_plan (yes/no)
- number_vmail_messages
- total_day_minutes
- total_day_calls
- total_day_charge
- total_eve_minutes
- total_eve_calls
- total_eve_charge
- total_night_minutes
- total_night_calls

- total_night_charge
- total_intl_minutes
- total_intl_calls
- total_intl_charge
- number_customer_service_calls
- Churn- (Target variable) which takes two values ‘no’ and ‘yes’.

Descriptive Statistics:

Descriptive statistics provide basic information about the data and this could be very helpful to understand the type of data we are dealing with. Here are the descriptive statistics of the data



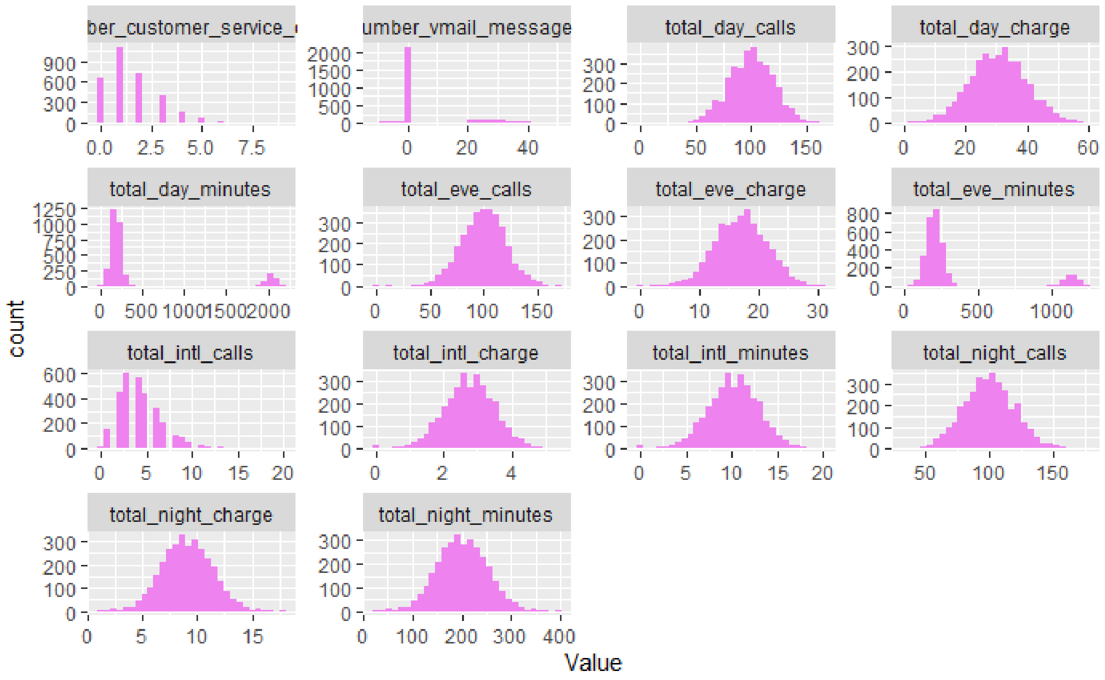
```
    state           account_length     area_code          international_plan voice_mail_plan    number_vmail_messages total_day_minutes
 Length:3333        Min.   :-209.00   Length:3333        Length:3333        Length:3333        Min.   :-10.000       Min.   :   0.0   
 Class :character   1st Qu.:  72.00   Class :character   Class :character   Class :character   1st Qu.:  0.000       1st Qu.: 149.3   
 Mode  :character   Median : 100.00   Mode  :character   Mode  :character   Mode  :character   Median :  0.000       Median : 190.5   
                    Mean   :  97.32                                                            Mean   :  7.333       Mean   : 418.9   
                    3rd Qu.: 127.00                                                            3rd Qu.: 16.000       3rd Qu.: 237.8   
                    Max.   : 243.00                                                            Max.   : 51.000       Max.   :2185.1   
                    NA's   :501                                                                NA's   :200           NA's   :200      
 total_day_calls total_day_charge total_eve_minutes total_eve_calls total_eve_charge total_night_minutes total_night_calls total_night_charge
 Min.   :  0.0   Min.   : 0.00    Min.   :   0.0    Min.   :  0.0   Min.   : 0.00    Min.   : 23.2       Min.   : 33.0     Min.   : 1.040    
 1st Qu.: 87.0   1st Qu.:24.45    1st Qu.: 170.5    1st Qu.: 87.0   1st Qu.:14.14    1st Qu.:167.3       1st Qu.: 87.0     1st Qu.: 7.530    
 Median :101.0   Median :30.65    Median : 209.9    Median :100.0   Median :17.09    Median :201.4       Median :100.0     Median : 9.060    
 Mean   :100.3   Mean   :30.63    Mean   : 324.3    Mean   :100.1   Mean   :17.08    Mean   :201.2       Mean   :100.1     Mean   : 9.054    
 3rd Qu.:114.0   3rd Qu.:36.84    3rd Qu.: 257.6    3rd Qu.:114.0   3rd Qu.:20.00    3rd Qu.:235.3       3rd Qu.:113.0     3rd Qu.:10.590    
 Max.   :165.0   Max.   :59.64    Max.   :1244.2    Max.   :170.0   Max.   :30.91    Max.   :395.0       Max.   :175.0     Max.   :17.770    
 NA's   :200     NA's   :200      NA's   :301       NA's   :200     NA's   :200      NA's   :200                           NA's   :200       
 total_intl_minutes total_intl_calls total_intl_charge number_customer_service_calls    churn          
 Min.   : 0.00      Min.   : 0.00    Min.   :0.000     Min.   :0.000                 Length:3333       
 1st Qu.: 8.50      1st Qu.: 3.00    1st Qu.:2.300     1st Qu.:1.000                 Class :character  
 Median :10.30      Median : 4.00    Median :2.780     Median :1.000                 Mode  :character  
 Mean   :10.23      Mean   : 4.47    Mean   :2.762     Mean   :1.561                                   
 3rd Qu.:12.10      3rd Qu.: 6.00    3rd Qu.:3.270     3rd Qu.:2.000                                   
 Max.   :20.00      Max.   :20.00    Max.   :5.400     Max.   :9.000                                   
 NA's   :200        NA's   :301      NA's   :200       NA's   :200                                     
```

From the above statistics, we can observe that most of the variables in the given data are numeric and there are few characters as well. There are missing values for a few variables in the data.

Data Visualization:

In this section let us visualize and observe how data is behaving under various aspects. This is an important part as it would help us to understand and analyze data in a 1much better way.

Dispersion and Skewness of the data

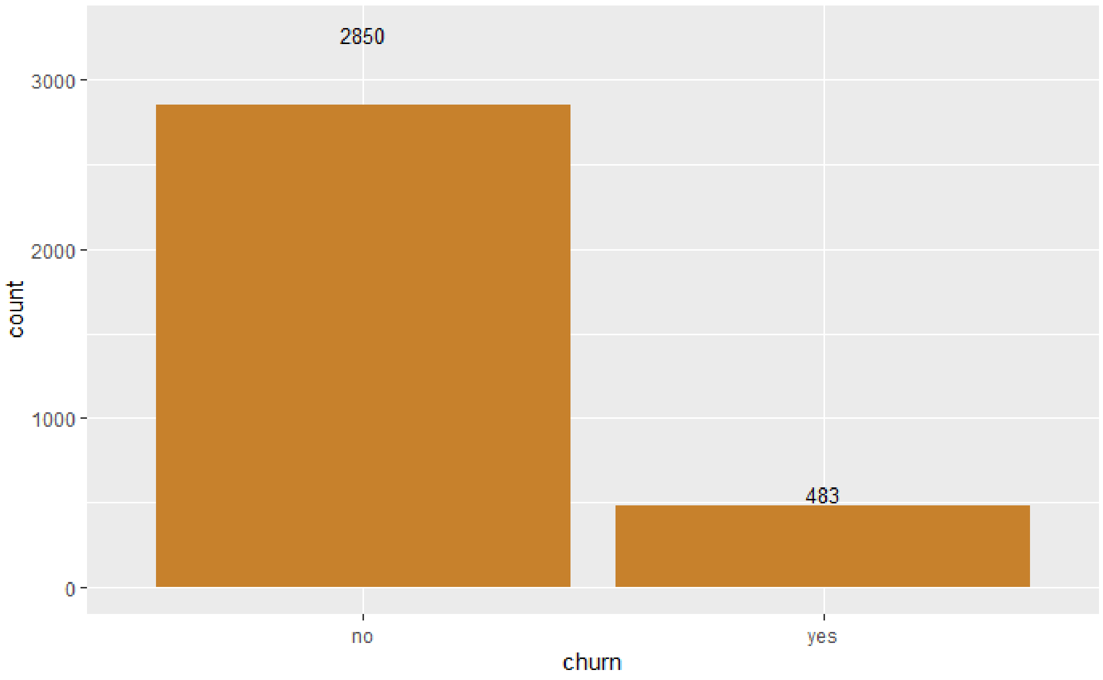


From the above graphs, we can see that most of the data is symmetrically distributed. The number of customer service calls has an irregular skewness. Total day minutes and Total evening minutes have a significant number of outliers. The majority of the voicemails are zero and there are few between 20-30.

Data Exploration

In data exploration, we try to analyse how various attributes are affecting the churn rate. This is done through data visualization and analysing the results of it.

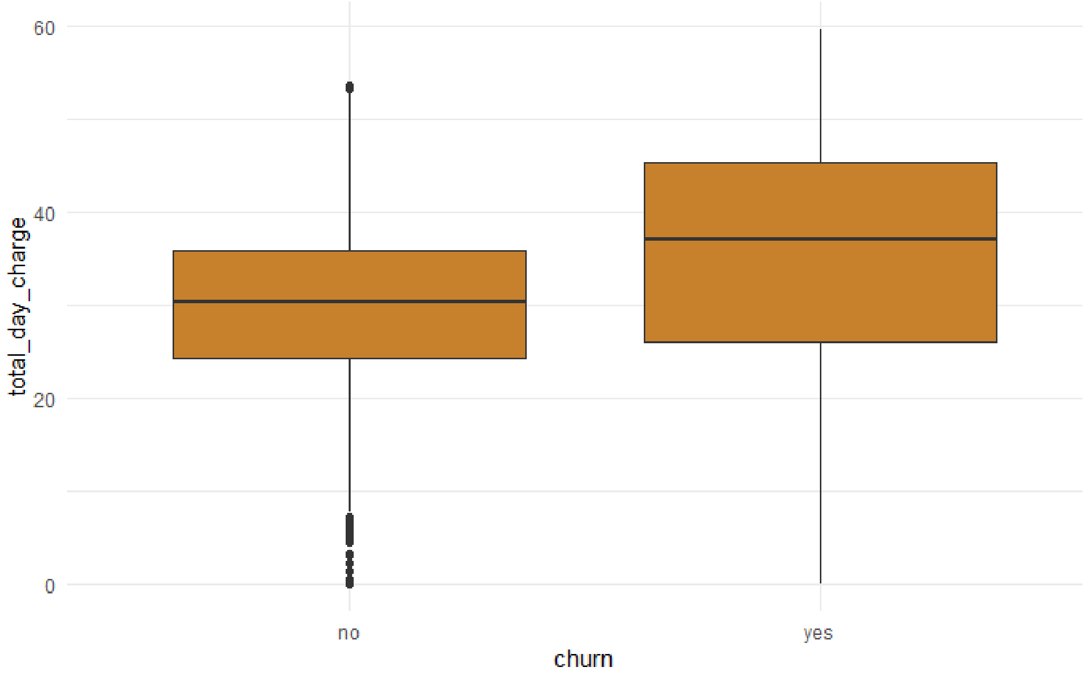
Churn in the data



out of 3333 customers in the data, a total of 483 customers have moved out of ABC wireless Inc and switched to other service providers.

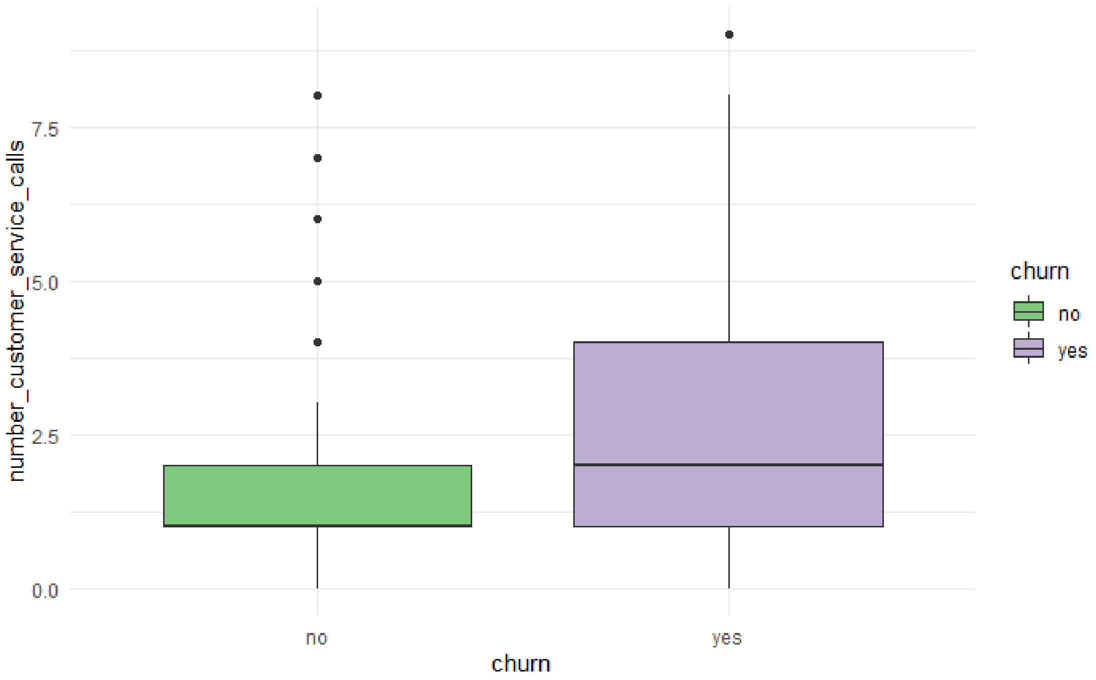
Let us understand how different variables are affecting the churn

Total day charges



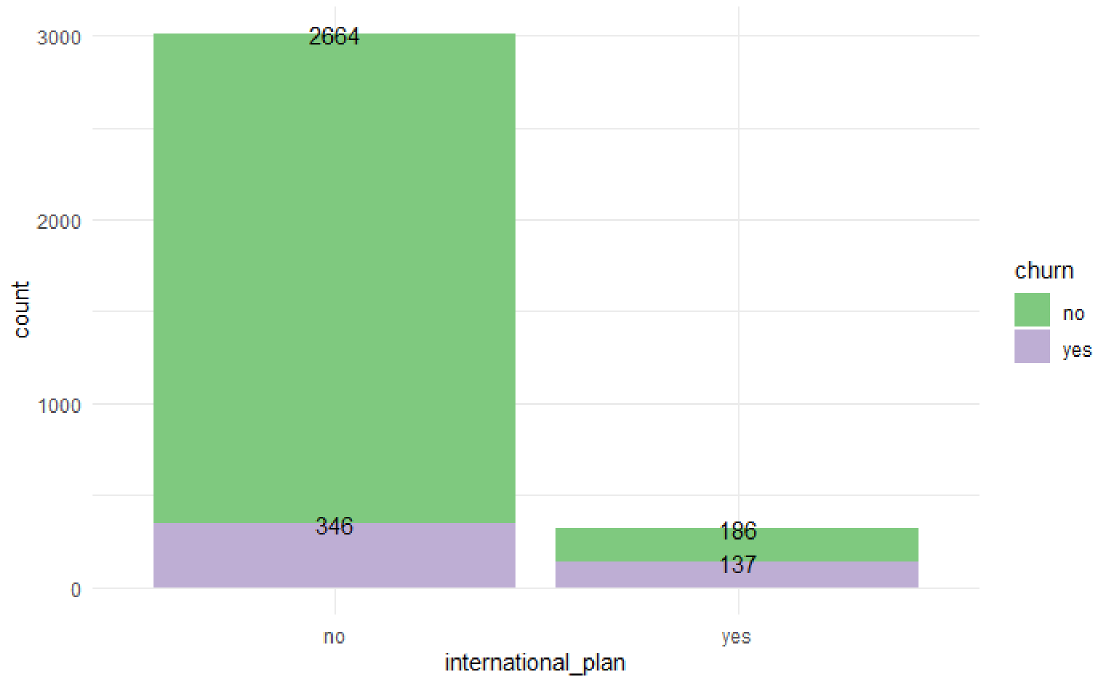
The above box plot explains how churn is affected by the total day charge. We can notice that the median line of the 1st box plot is lower than that of the 2nd one. On analyzing it we can understand that when the total day charges are higher than 30 then the customer is more like to churn.

Customer service calls:



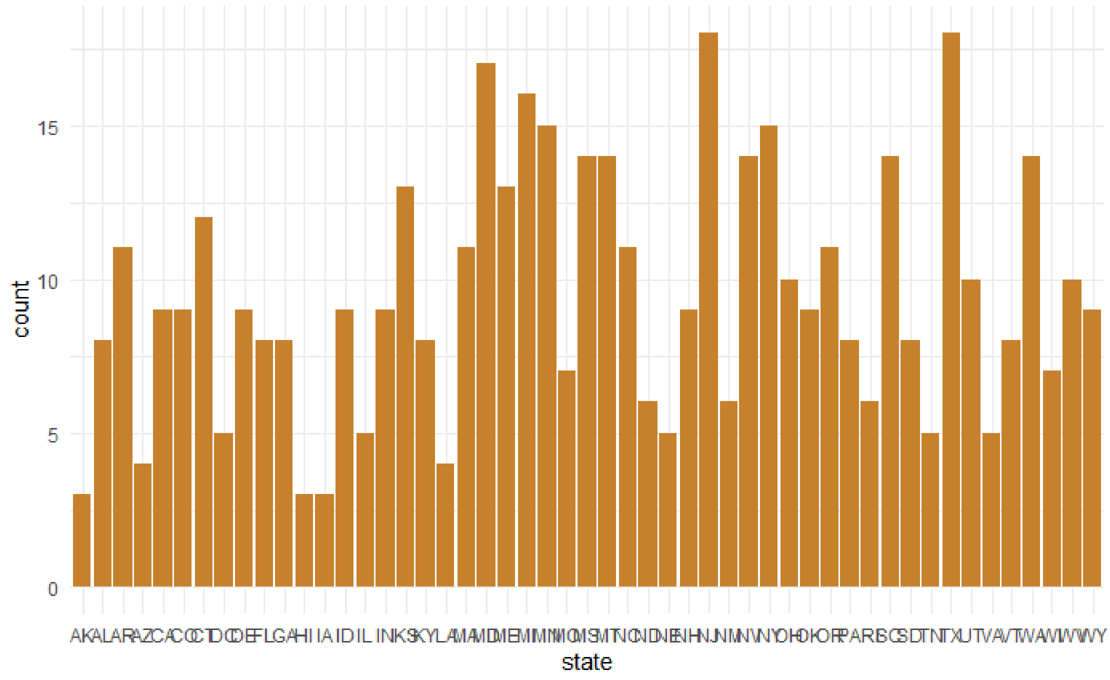
A customer is more likely to make a call to the service provider when he/she faces any problem or is unsatisfied with the service provided. So, the number of customer service calls is directly related to the satisfaction level of the customers. From the above graph, we can understand that customers who have churned have called more than 2 times to customer service. About 76% of customers who have called customer services have switched to other service providers.

International plan



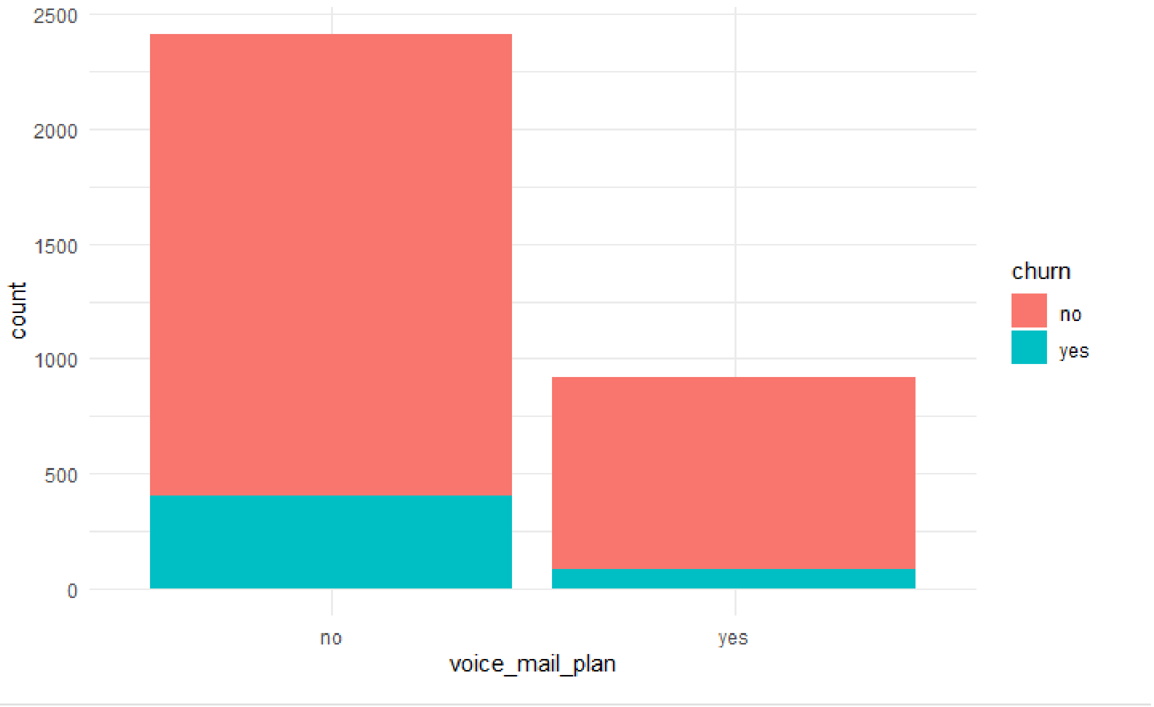
The above graph shows how churn is related to an international plan. From this, we can notice that about 42% of customers with the international plan have churned. This means that the customer with the international plan is either not satisfied with the service provided or has found a better deal with other service providers.

Churn in each state



The graph shows that the States of Maryland, New Jersey, Michigan, and Texas have high churn rates. This implies that customers from these states have a high chance to churn. So, the company should improve its marketing strategies in these states to retain customers.

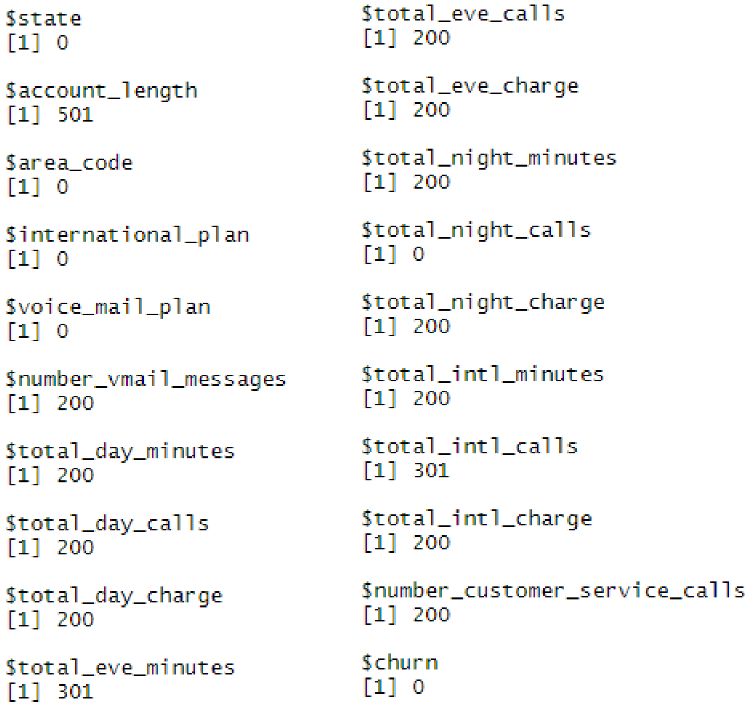
voice mail plan



The above graph shows that very few people with voice mail plans churned. This means that churn is not affected by the voicemail plan.

Data Cleaning:

It is important to have clean data before we start the analysis. Let us check if there are any missing values in the data.



```
$state                         $total_eve_calls
[1] 0                          [1] 200

$account_length                $total_eve_charge
[1] 501                        [1] 200

$area_code                     $total_night_minutes
[1] 0                          [1] 200

$international_plan            $total_night_calls
[1] 0                          [1] 0

$voice_mail_plan               $total_night_charge
[1] 0                          [1] 200

$number_vmail_messages         $total_intl_minutes
[1] 200                        [1] 200

$total_day_minutes             $total_intl_calls
[1] 200                        [1] 301

$total_day_calls               $total_intl_charge
[1] 200                        [1] 200

$total_day_charge              $number_customer_service_calls
[1] 200                        [1] 200

$total_eve_minutes             $churn
[1] 301                        [1] 0
```

It can be seen that there are multiple columns with missing values.

There are multiple methods to deal with missing values like

1. Delete the rows that have NAs

2. Impute the NAs using the Median of the columns

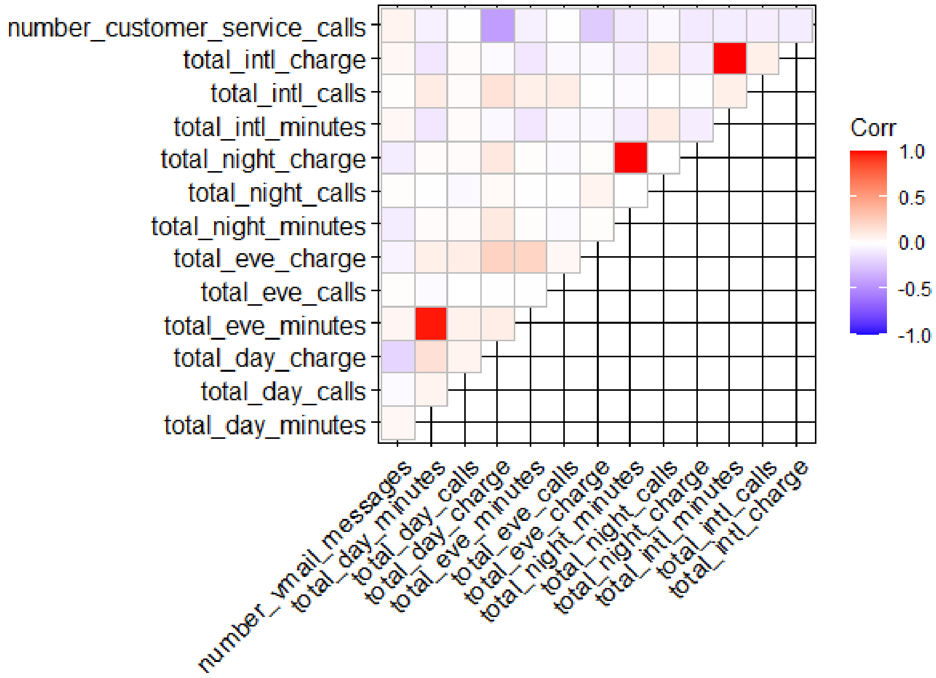
3. Impute the NAs using the Mice package (Multivariate Imputation by Chained

Equations)

4. k-Nearest Neighbour (k-NN) method

Deleting the rows that have NAs led to the loss of a significant amount of data. We used the k-Nearest Neighbour (k-NN) method to impute missing values as this method would result in better results while building the predictive model.

Correlation between the variables given that the churn is equal to yes



From the above plot, it can be interpreted that for the customers who churned there is a strong positive correlation between total evening minutes and total day minutes, total night charge and total night minutes, total international charge, and total international minutes. This means that these variables are directly related to each other implying increase in one this variable will increase the other variable and vice versa. It is also evident that the total day charge and the number of customer service calls have a strong negative correlation for the customers churned.

Modelling Strategy:

Predictive Modelling can be done based on Regression and Decision Tree Models. In these models, while predicting the dependent variable, different independent variables have different levels of impact. It is key to choose the best model that suits best the data and provides better results in the predictive model.

Regression modelling can be done in two ways:

1. Linear Regression

2. Logistic Regression

For the present project on ABC Wireless Inc Logistic regression is more appropriate compared to linear regression as the dependent variable is categorical.

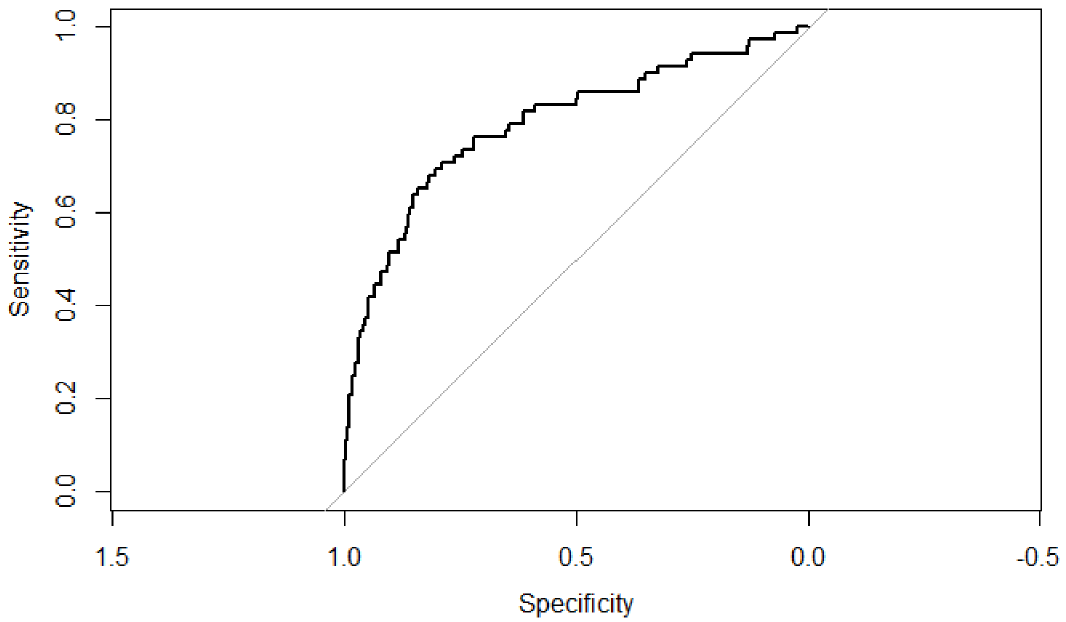
Let us build models using both Logistic Regression and Decision Tree Models and compare the performance of those to find the best one to make predictions on the test data.

Before building the model, we are dividing the model into two parts one is to train the model and the other is to measure its performance. The training set has 85% of the data and the validating set has the remaining 15% data. The model will be built on the training set and its performance will be tested on validation sets.

Building a Logistic Regression Model: -

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables and it is a method used to predict a binary outcome, such as yes or no. The model is built on the training set and validated on the validation set.

The area under the curve for Logistic Regression Model

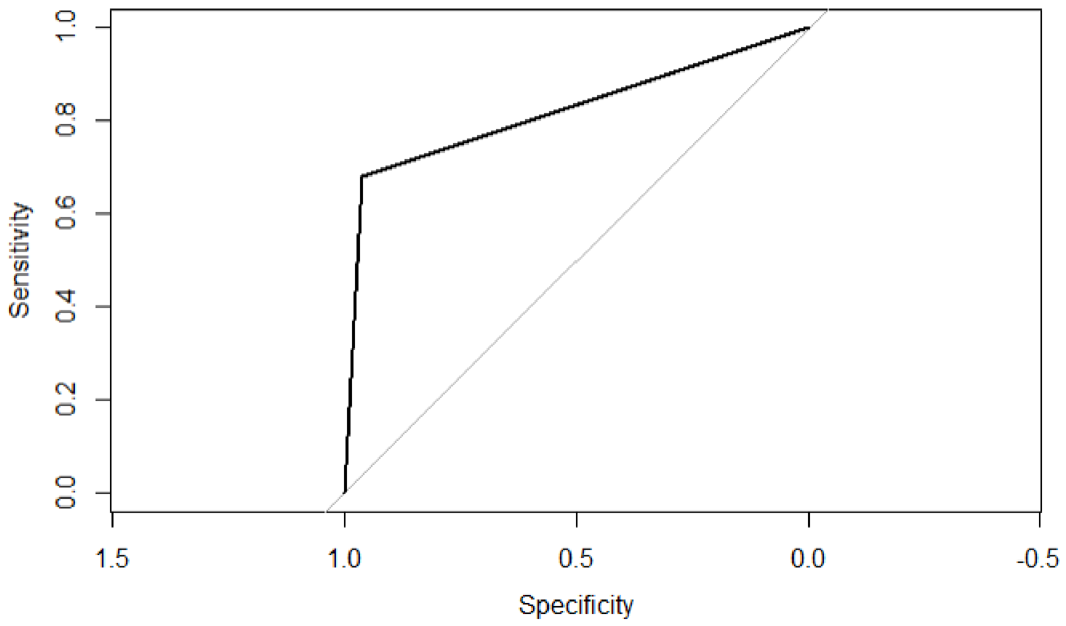


The area under the curve for the logistic regression model is 0.7914. This means that model is 79.14% correct while classifying data.

Building a Decision Tree Model: -

A decision tree model is a graph which uses a branching method to explain every possible output for a specific input. The model is built on the training set and validated on the validation set. Let us examine the results of validation.

The area under the curve for the Decision Tree Model

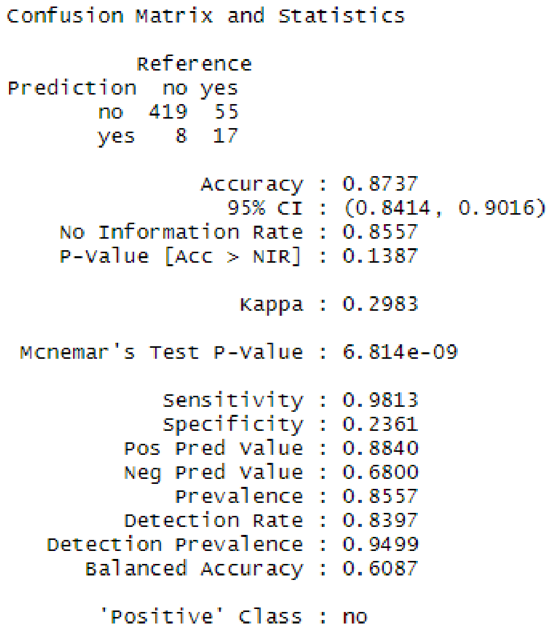


The area under the curve for the logistic regression model is 0.8215, implying that the decision tree model has correctly classified the data 82.15% of the time.

Comparing the performance of both models:

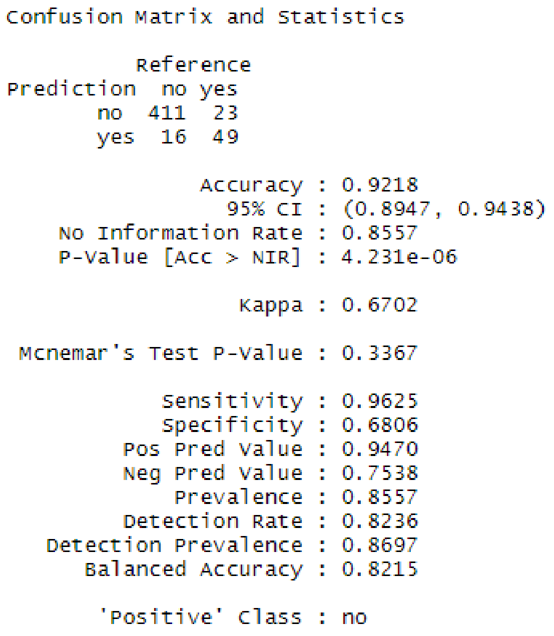
We used a confusion matrix to analyse the performance of both models. Here are the results of it

Logistic Regression Model



```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  419  55
       yes   8  17

               Accuracy : 0.8737
                 95% CI : (0.8414, 0.9016)
    No Information Rate : 0.8557
    P-Value [Acc > NIR] : 0.1387

                  Kappa : 0.2983

 Mcnemar's Test P-Value : 6.814e-09

            Sensitivity : 0.9813
            Specificity : 0.2361
         Pos Pred Value : 0.8840
         Neg Pred Value : 0.6800
             Prevalence : 0.8557
         Detection Rate : 0.8397
   Detection Prevalence : 0.9499
      Balanced Accuracy : 0.6087

       'Positive' Class : no
```

Decision Tree Model



```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  411  23
       yes  16  49

               Accuracy : 0.9218
                 95% CI : (0.8947, 0.9438)
    No Information Rate : 0.8557
    P-Value [Acc > NIR] : 4.231e-06

                  Kappa : 0.6702

 Mcnemar's Test P-Value : 0.3367

            Sensitivity : 0.9625
            Specificity : 0.6806
         Pos Pred Value : 0.9470
         Neg Pred Value : 0.7538
             Prevalence : 0.8557
         Detection Rate : 0.8236
   Detection Prevalence : 0.8697
      Balanced Accuracy : 0.8215

       'Positive' Class : no
```

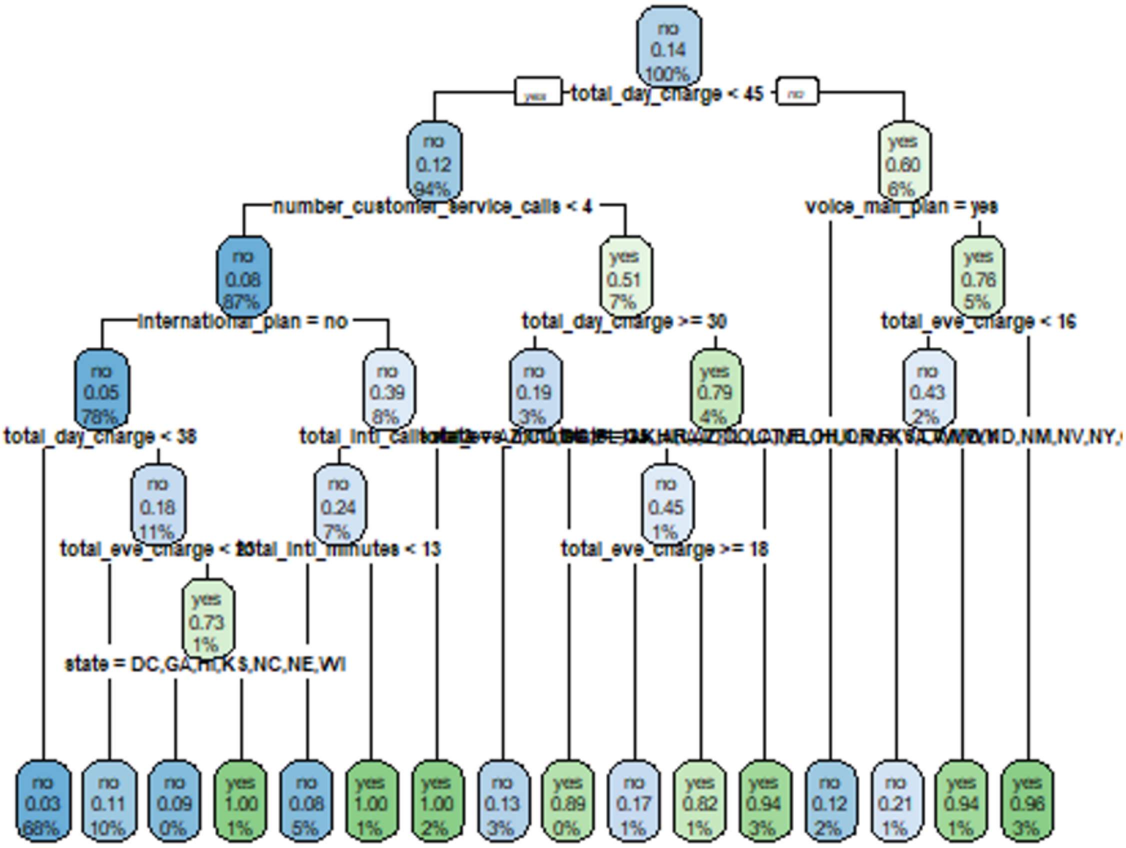
From the results of the confusion matrix, it can be seen that the Logistic regression model has an Accuracy of 87.37%, Sensitivity of 98.13% and Specificity of 23.61%. Whereas the Decision tree model has an Accuracy of 92.18%, Sensitivity of 96.25% and Specificity of 68.06%.

On comparing the performance of both models, we notice that the Decision tree has done better in terms of accuracy and specificity. The sensitivity of logistic regression is slightly higher than the Decision tree model. Over Decision tree has performed better on the validation set compared to the logistic regression model. Therefore, we are choosing the Decision tree model as the best model to make predictions of the test data.

Prediction of Churn in Test data

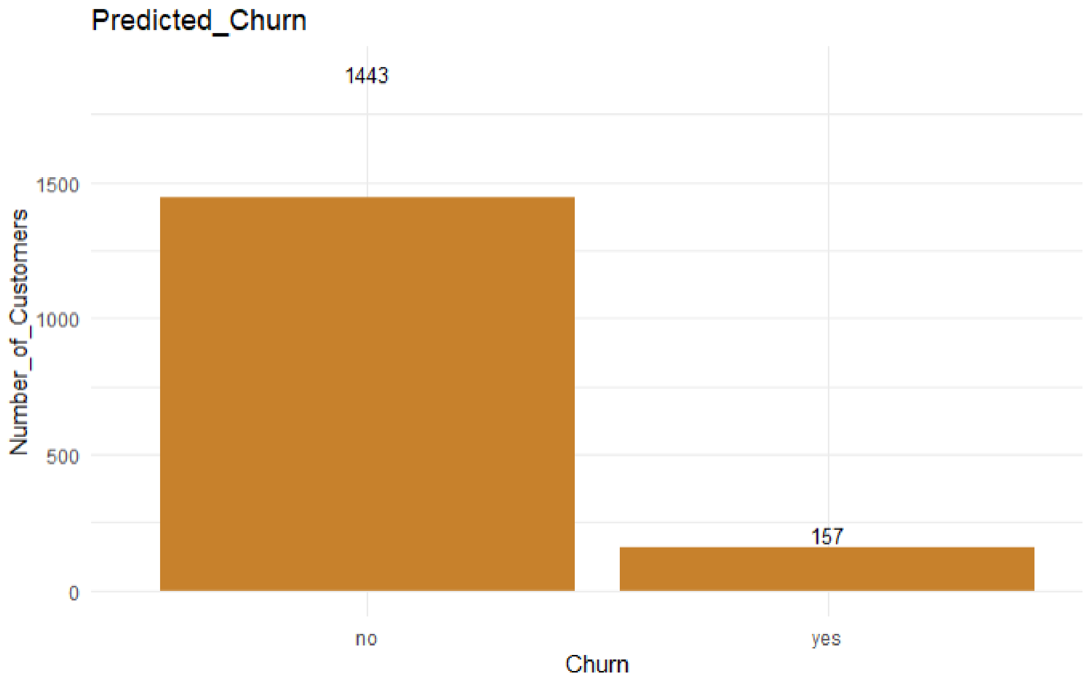
Decision Tree Model: -

We have built a Decision Tree using the entire data set to predict the churn of test data.



We have used this model to predict churn on the test data provided.

Here are the predicted results of the Decision tree model on test data



So, the Decision tree predictive model has predicted that out of 1600 customers in test data 157 customers are likely to churn.

Insights:

Following are the conclusions made from the Data Exploration:

1. customers who are paying a total day charge of more than 30 are more likely to churn

2. Customers who call customer service more than once are likely to churn.

3. Customers with international plans are more probable to switch to other carriers.

4. Customers from the States of Maryland, New Jersey, Michigan and Texas have a high churn rate

Suggestions and Recommendations:

An overall company need to take the following steps in order to reduce the churn rate:

1. Try to reduce the Total day charge. As we have noticed that when the total day charge is high churn rate is also high. We can also interpret that customers who have churned might have found better deals with other service providers.

2. Company need to improve customer satisfaction as low customer satisfaction leads to customer service calls and it is directly related to churn. They should immediately identify the problems or troubles faced by customers immediately and try to solve them as the delay in it might lead to loss of customers.

3. Company need to provide better deals for customers with an international plan.

4. Company need to come up with better marketing strategies for Maryland, New Jersey, Michigan and Texas States. These states have high churn rates, which might imply that customers are getting better options from other service providers. So, the company should take immediate action like improving their marketing strategies in order to improve their brand loyalty.

Conclusion:

ABC Wireless Inc should try to target those 157 customers as they are a high chance to churn. The company need to do strategic marketing to those customers to improve their brand loyalty of those customers. The company need to implement all suggestions mentioned as soon as possible to stop existing customers from switching to other service providers.