FALL 2022 BUSINESS ANALYTICS (MIS-64036-001)

FINAL PROJECT REPORT

Names	Contributions
NIKITHA CHIGURUPATI	R CODE , REPORT
GLORIA STEPHEN	REPORT , R CODE
SARIDINDU MEGHANA	R CODE , REPORT
AKSA TANIYA	PRESENTATION , REPORT

TABLES OF CONTENT

1. PROJECT GOAL:

A statistic where customer churn gives you information about the customer attrition rate. Its Analysis is more important since acquiring new customers is usually more difficult than keeping existing, paying customers.
Pertinent data is necessary for churn analysis.

Our Objective here is to analyze the data and the crucial variables it contains, eventually jotting down the results with the help of R language to conclude the statistic result it yields through which our main focus is understanding what can be done by the ABC Wireless company to reduce its churn percentage.

Understanding the Process of Churn Analysis-

- Customer Experience and Consumption
- Possibility of Upgrading

Understanding the Process of Churn Analysis-

- Maximize Utilization of Subscription Analytics
- Analyze Customer Segmentation
- Determine the Causes of Churn

2. OVERVIEW OF THE DATA:

Brief check of the data –
The data consists of around 20 variables which includes both numerical and categorical data, so it makes us use varied methods in analyzing the data. We mainly considered the variables which had a visible effect on the dependent variable (Churn) like Account length, Area code, international plan, Voice mail plan, Total international charge and Number of customer service calls. Understanding its effect on the churn is what we observed here through graphics, plot and tables.

Code Segmentations -
The data was cleaned and then the skewness was checked. The mean is mainly affected by any outliers or skewness of the three measures of tendency. In a symmetrical distribution, the mean, median, and mode are all equal. So here is the skewness.

Quartile segregation -
A type of percentile is a quarter. 25% of the data are below the first quartile, which is also known as the lowest quartile or Q1, making it the 25th percentile. The second quartile, known as the median or Q2, is the 50% percentile of the data, which means that 50% of the data are below the Q2 level.

Data Exploration Analysis

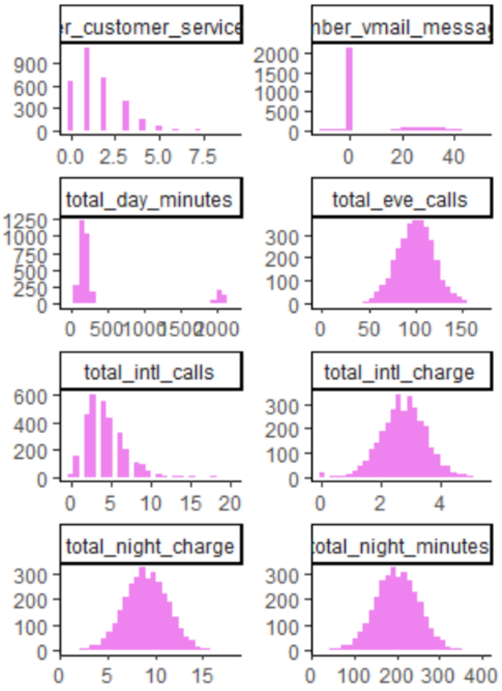
This project consists of historical data of churn train for the analysis and to build the model. It has the information about the services used by the customers of the company.

- 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

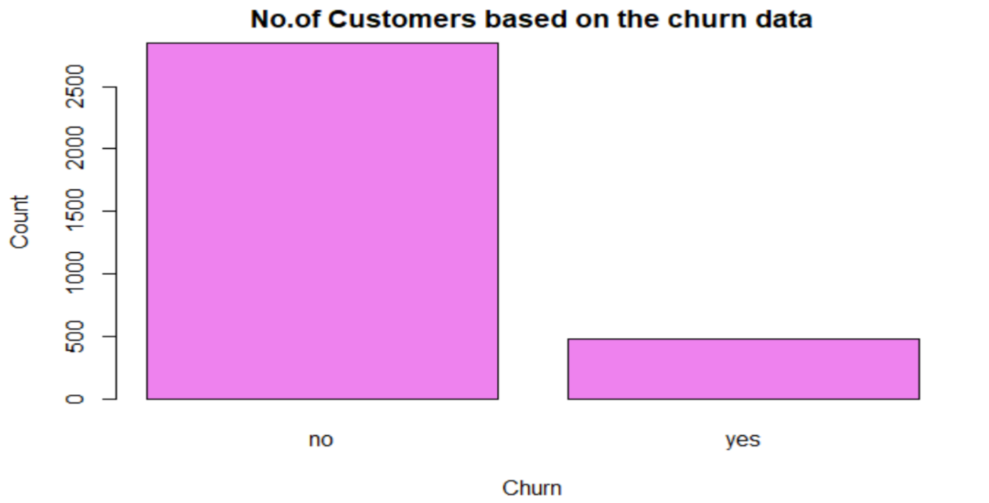
Data transformation-
Post the cleaning of the data to explore the data we need to remove the categorical variables or transform them to numeric variables as it might misinterpret the numeric command calls for the language.

Graphical Representation-
The graph shows the skewness of the variables where the mean and median relation is to be understood. The Irregular formation shows the dependency of to the particular on the variable of churn.



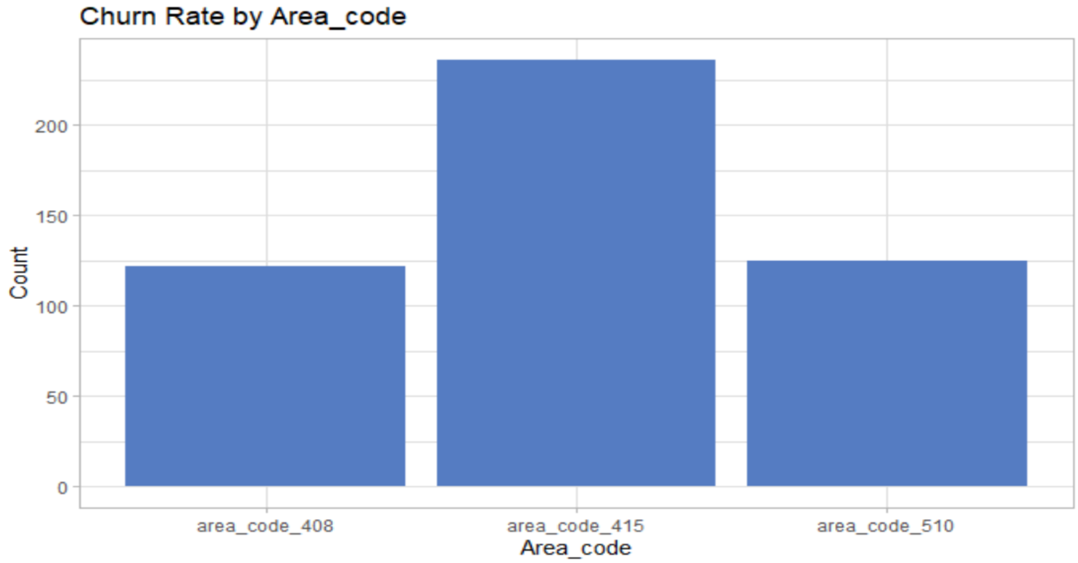
The above skewness shows a bell curve distribution of data for maximum amount of the data or variables. It is seen that the “Total day minutes” and “Total evening minutes” have a significant number of outliers. It is evident that “Customer_Service_calls” has an irregular skewness.

Determining the number of customer churn through a bar graph-



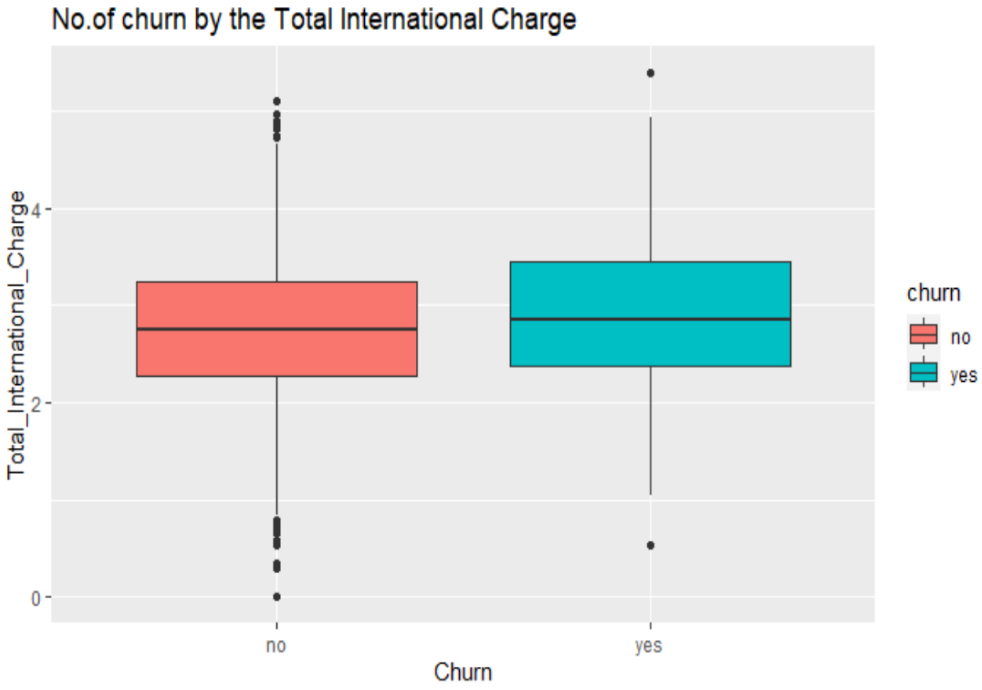
The customer churn analyzed is 483 out of 3333.Which is 2850 customers have stayed with the current provider.

Area code variable shows us the churn rate of customers

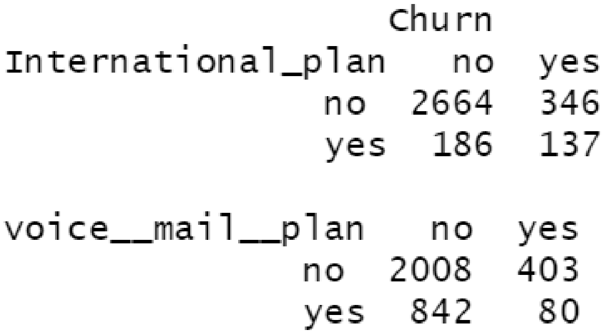


The area_code_415 has the highest of customer churn rate.

Customer churn due to the international charge



Customer churn due to international plan and voice mail plans.

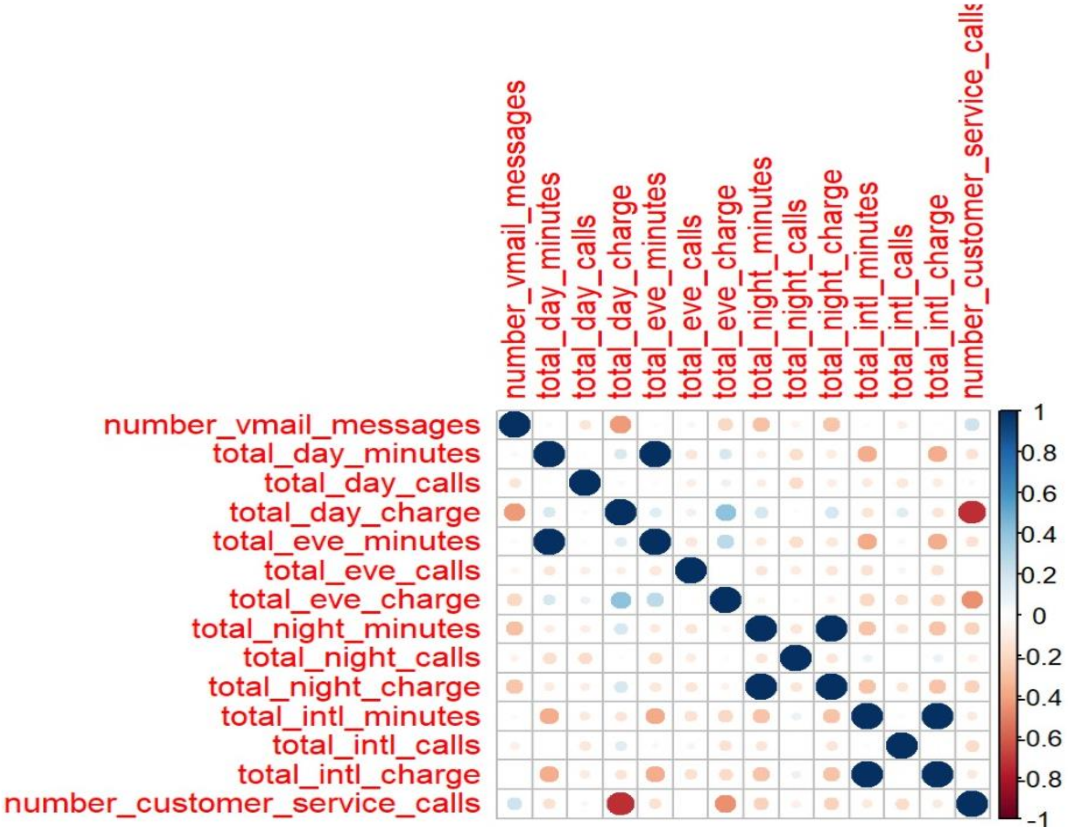


```
                      Churn
International_plan   no  yes
                no  2664  346
               yes   186  137

voice__mail__plan   no  yes
              no  2008  403
             yes   842   80
```

28% of customers are lost by the international plan and 16% of customers are lost by the voice mail plan.

Creating a correlation plot for the numerical variables



Positive Correlation-
Total day minutes & total eve minutes
Total night calls & total night charges
Total international calls & total international charge

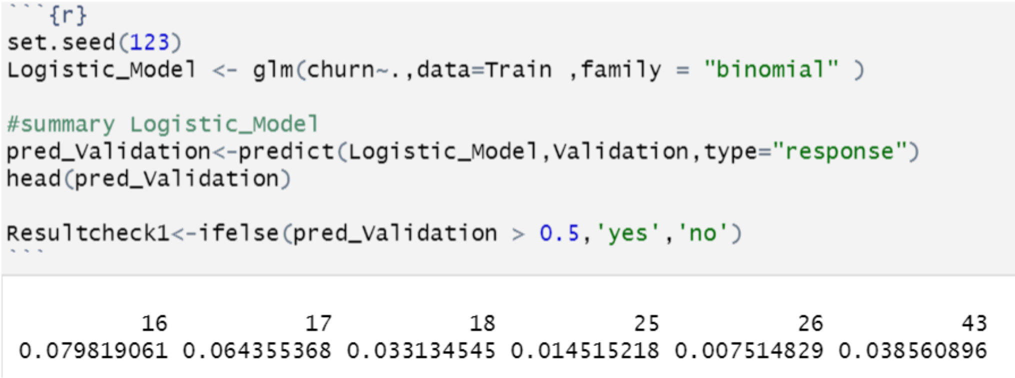
Negative Correlation-
Number of service calls & Total day charge

4. DETAILS OF MODELLING STRATEGY:

It is defined as the best way to comprehend the data- extract the main variables and the method used to obtain the best analysis through its output.

For the given data we segregated the data in a ratio of 2 portions as a Sample for training and validation set of 80 and 20. The Dataset was then evaluated under two methods i.e. Logistic regression and decision tree.

Building Logistic Regression Model



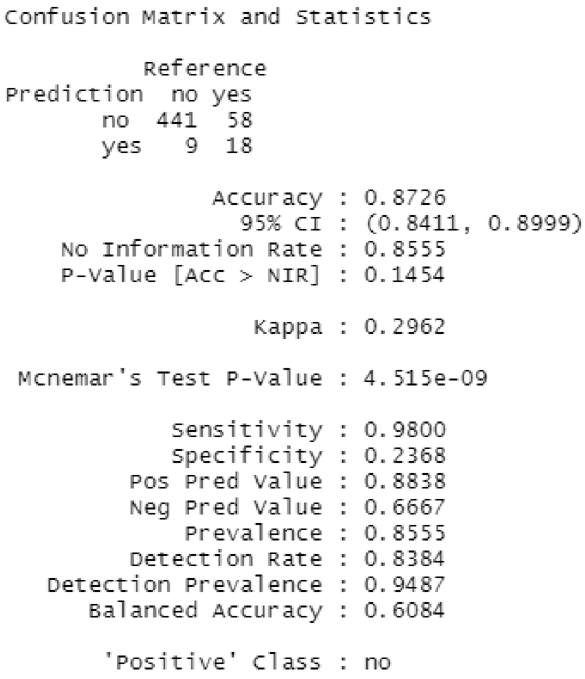
~~~
```{r}
set.seed(123)
Logistic_Model <- glm(churn~.,data=Train ,family = "binomial" )

#summary Logistic_Model
pred_Validation<-predict(Logistic_Model,Validation,type="response")
head(pred_Validation)

Resultcheck1<-ifelse(pred_Validation > 0.5,'yes','no')
```

         16          17          18          25          26          43 
0.079819061 0.064355368 0.033134545 0.014515218 0.007514829 0.038560896 
~~~

Confusion matrix of Logistic Regression Model



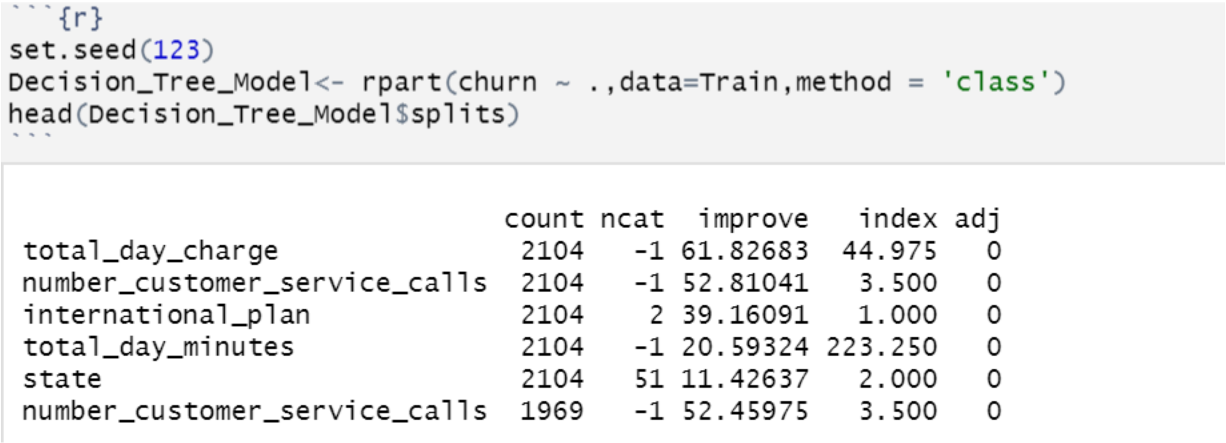
```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  441  58
       yes   9  18
                                          
               Accuracy : 0.8726          
                 95% CI : (0.8411, 0.8999)
    No Information Rate : 0.8555          
    P-Value [Acc > NIR] : 0.1454          
                                          
                  Kappa : 0.2962          
                                          
 Mcnemar's Test P-Value : 4.515e-09       
                                          
            Sensitivity : 0.9800          
            Specificity : 0.2368          
         Pos Pred Value : 0.8838          
         Neg Pred Value : 0.6667          
             Prevalence : 0.8555          
         Detection Rate : 0.8384          
   Detection Prevalence : 0.9487          
      Balanced Accuracy : 0.6084          
                                          
       'Positive' Class : no              
```

The following conclusions have been made :-

1. Accuracy - 0.87 or 87.26%
2. Sensitivity - 0.98 or 98%
3. Specificity:- 0.23 or 23.68%.

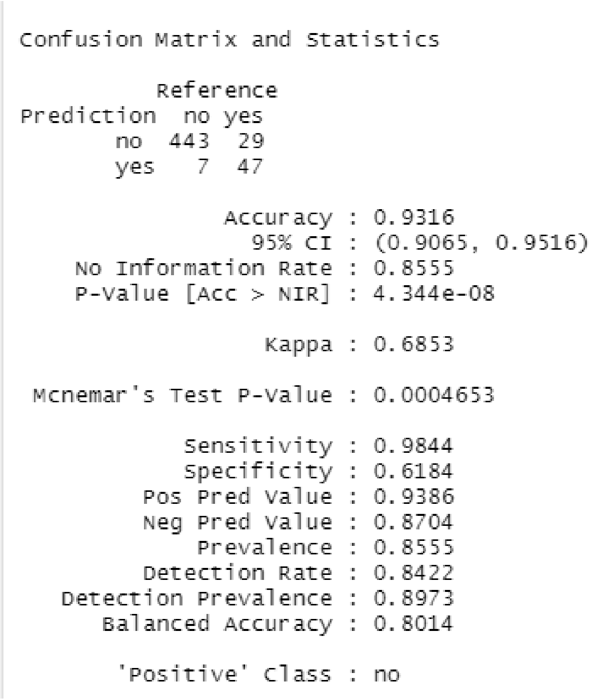
Building Decision Tree Model



~~~
```{r}
set.seed(123)
Decision_Tree_Model<- rpart(churn ~ .,data=Train,method = 'class')
head(Decision_Tree_Model$splits)
```
~~~

```
                            count ncat  improve   index adj
total_day_charge             2104   -1 61.82683  44.975   0
number_customer_service_calls 2104   -1 52.81041   3.500   0
international_plan           2104    2 39.16091   1.000   0
total_day_minutes            2104   -1 20.59324 223.250   0
state                        2104   51 11.42637   2.000   0
number_customer_service_calls 1969   -1 52.45975   3.500   0
```

Confusion matrix of Decision Tree Model



```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  443  29
       yes   7  47

               Accuracy : 0.9316
                 95% CI : (0.9065, 0.9516)
    No Information Rate : 0.8555
    P-Value [Acc > NIR] : 4.344e-08

                  Kappa : 0.6853

 Mcnemar's Test P-Value : 0.0004653

            Sensitivity : 0.9844
            Specificity : 0.6184
         Pos Pred Value : 0.9386
         Neg Pred Value : 0.8704
             Prevalence : 0.8555
         Detection Rate : 0.8422
   Detection Prevalence : 0.8973
      Balanced Accuracy : 0.8014

       'Positive' Class : no
```

The following conclusions have been made :-

1. Accuracy - 0.93 or 93.16%
2. Sensitivity - 0.98 or 98.44%
3. Specificity:- 0.61 or 61.84%

From the above model, the Decision Tree Model is the optimal model for this dataset. It is the best model to use as it has higher accuracy than the Logistical Regression Model. Though the Sensitivities of both the models are equal, Decision Tree has a higher specificity. Hence, Decision Tree Model is the right and optimal Model to use.

Method	**Accuracy**	**Sensitivity**	**Specificity**
Logistic regression	87.26%	98%	23.68%
Decision Tree	93.16%	98.44%	61.84%

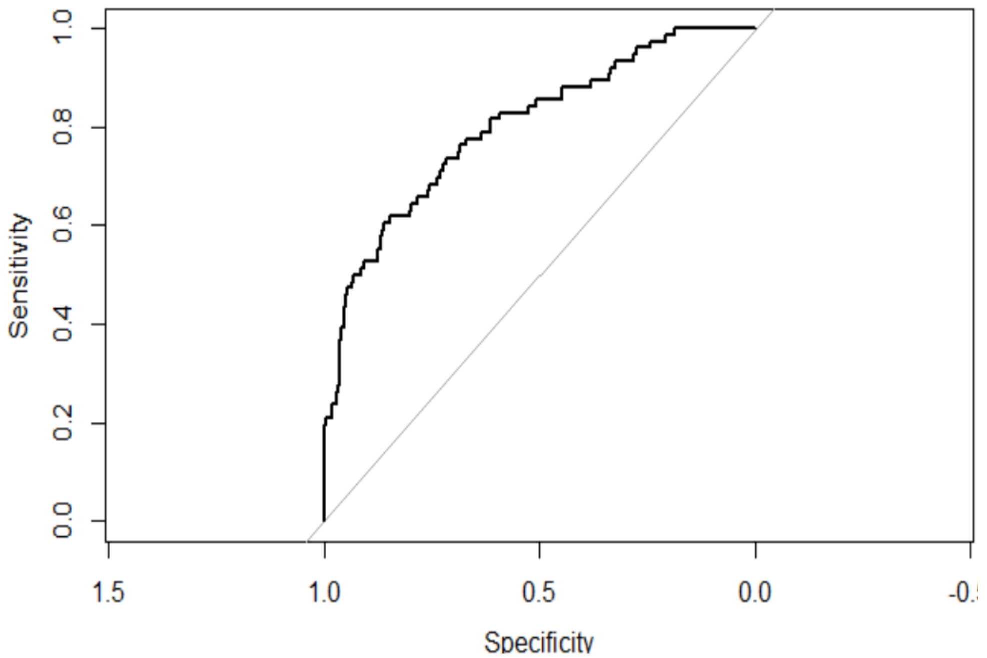
Logistic Regression gave us accuracy of 87% whereas the decision tree is of a 93% which shows a better accuracy of the churn rate reason.

Decision tree: Decision Tree Model explained in Terms of business requirements. The tree may be viewed as a form of flowchart, with you starting at the top and working your way down to a specific bucket at the bottom. The tree may be manually followed such that each client was assigned to one of the buckets. The breaks in the tree represent inquiries regarding the customer's values for each characteristic. Customers who end up in a bucket have certain traits. In this manner, we may give a churn probability to each bucket.

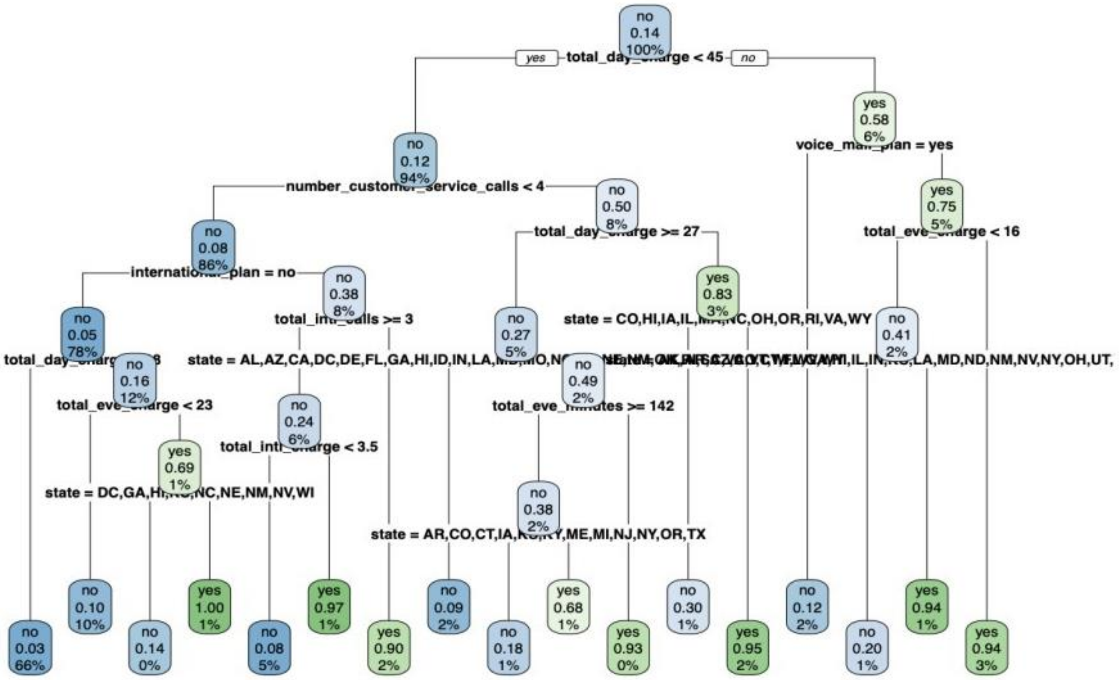
The method used to construct the tree is the most challenging part of the decision tree, yet the core notion is relatively simple. The idea is for each bucket to be as pure as possible. In a totally pure bucket, either all consumers would churn or nothing at all would. We begin with the entire dataset. The algorithm then tries to discover the attribute to divide on so that the two subsets are as pure as feasible. After picking the initial split, the process is rehashed for the portion of the data one step down in the tree, and so on.

4. **ESTIMATION OF MODEL'S PERFORMANCE**:

ROC analysis- In logistic regression, ROC curves are used to calculate the appropriate cutoff value for predicting whether a new observation is a "failure" (0) or a "success" (1). The model that has the maximum Area Under Curve (AUC) would be deemed to be the most appropriate for the given issue. The decision tree model's AUC is 0.8591, which is noticeably higher than the logistic regression model's AUC.



We can now apply all we've learned from prior data to new data from other clients. It allows us to walk through the tree for each client and estimate whether or not they are going to terminate their contract in the near future. Of course, we will not follow the tree by hand, but will instead rely on computers. This approach could be applied to an infinite number of consumers, who would all end up in one of the buckets.



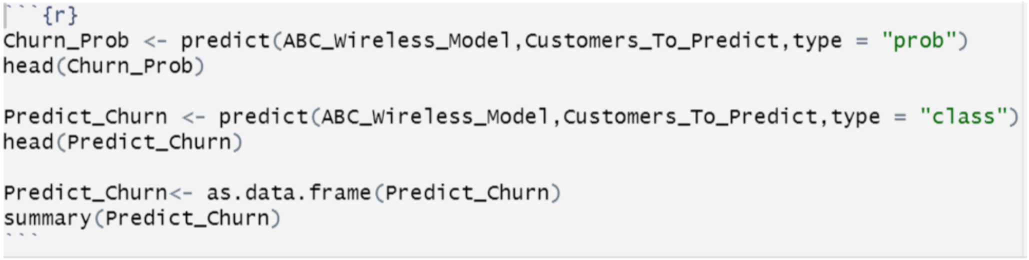
The boxes in the image represent the decision tree's nodes, which have leaves and branches. The leaves represent the class level, while the branches represent the criteria. Posted in the model a long either side, we have a "yes" and "no" that determine if the data piece meets the requirement.

The first "activity" in the root node (total day charge) is "No" since the majority of the data points fulfill the total day charge less than 45, implying that being less than 45 is the dominating activity. The 94% refers to the likelihood that a new data point will say "Yes" to that action, while the remaining 6% will say "No." This reasoning is repeated for each node.

In conclusion, we can see that the Total Day Charge, Number of Customer Service Calls, and so on are crucial questions to ask to identify the likelihood of consumers churning or not. As a result, ABC Wireless Inc. might obtain critical information to help with business decisions

Prediction of Test Dataset

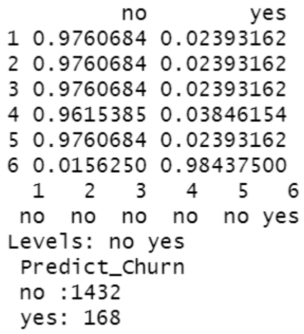
We initially reviewed the forecast data to ensure that it was in the same format as the reference data and that there were no missing values. We will offer a dataset (Customers To Predict) that will be used to forecast customer turnover using the Decision Tree Model.



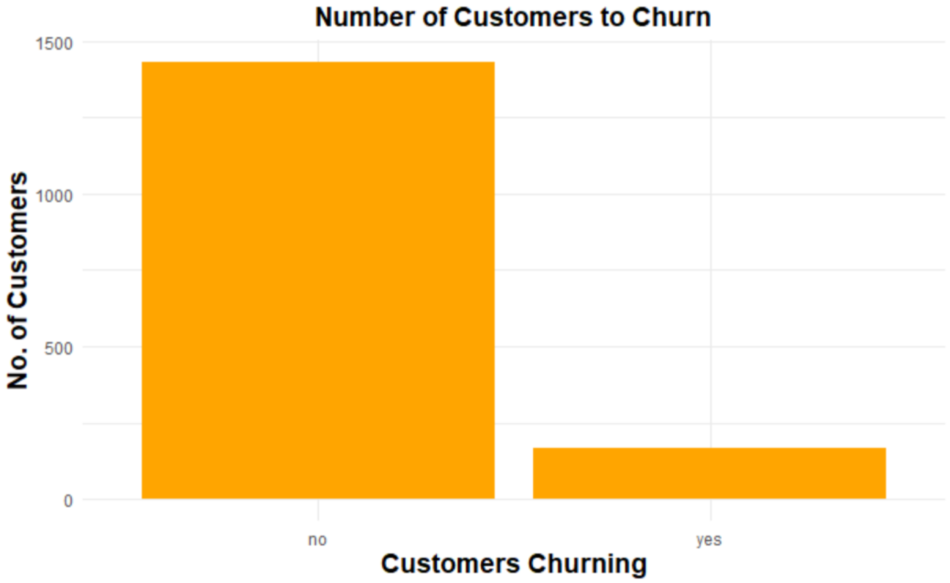
~~~
```{r}
Churn_Prob <- predict(ABC_Wireless_Model,Customers_To_Predict,type = "prob")
head(Churn_Prob)

Predict_Churn <- predict(ABC_Wireless_Model,Customers_To_Predict,type = "class")
head(Predict_Churn)

Predict_Churn<- as.data.frame(Predict_Churn)
summary(Predict_Churn)
```
~~~



```
         no        yes
1 0.9760684 0.02393162
2 0.9760684 0.02393162
3 0.9760684 0.02393162
4 0.9615385 0.03846154
5 0.9760684 0.02393162
6 0.0156250 0.98437500
  1   2   3   4   5   6 
 no  no  no  no  no yes 
Levels: no yes
 Predict_Churn
 no :1432    
 yes: 168    
```



From the above, it is concluded that 168 customers are churning out of 1600 customers.

5. INSIGHTS AND CONCLUSIONS:

The insights observed from the data are majorly from the variables like Area code, international charges, Customer Service calls and Voice mail plan. The main cause for customers to churn is represented from the above plots because of these variables.

ABC Wireless should implement the following methods to retain their customers as it is inexpensive and easier for the company.

- The company has to reduce the charges in international calls and provide a better customer service as we observed that the calls made to the customer care department is more which implies the service is not provided rightly at the initial calls from the telecom service.
- Reward the loyal customers in order to maintain the customer retention.
- Receive feedback from the customers on regular basis
- Segmentation on basis of customer plan
- Acquisition strategy for churned customers

REFERENCES

https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

https://stats.oarc.ucla.edu/r/dae/logit-regression/

https://www.statology.org/logistic-regression-in-r/