**Telecom Industry: Customer Churn Prediction**

Nikhil Kumar Sampath, Kiran Kour, Akhil Yada, Sandhya Cheepurupalli

## MIS64036-001: Business Analytics

## Professor Dr Rouzbeh Razavi

## Kent State University

## Nowadays, the telecom industry faces fierce competition in satisfying its customers. With the advent of newer technology, the services offered by telecom companies have increased from being only calls to calls, data, and web services. This means a constant struggle to strike a perfect balance between services and the pricing of these services. To survive this market, telecom companies need to innovate, offer better services and increase their customer base. With newer companies entering the market and increasing the freedom of customers to switch telecom companies, it’s becoming increasingly important to focus resources on retaining existing customers. “According to an article in Harvard Business Review, it was determined that the cost of acquiring a customer is five to twenty-five times more than retaining an existing one.” Furthermore, increasing retention by five per cent can increase profits by twenty-five to ninety-five per cent.

## Project Goal

## Customer Churn is a major problem for Telecom companies. ABC Wireless Inc. is one of these companies. The data (Churn\_Train.xlsx) has information about customer usage behaviour, contact details, and payment details. The data also indicates who were the customers who cancelled their service. Based on this past data, the goal is to build a model which can predict whether a customer will cancel their service in the future or not.

## Overview of Data, Including Data Exploration and Analysis

## Overview of Data

## Total number of rows and columns:

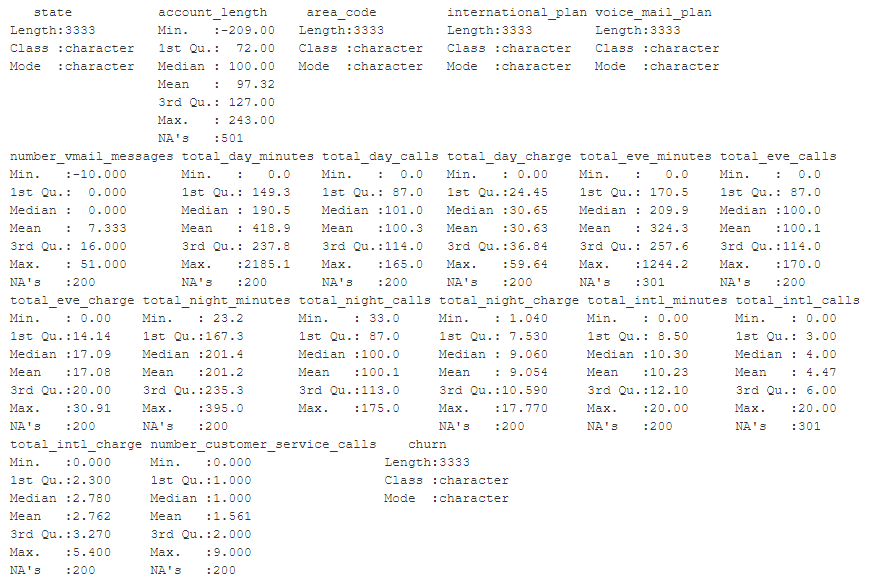
## Text Description automatically generated

## There are 3333 rows and 20 columns in the dataset

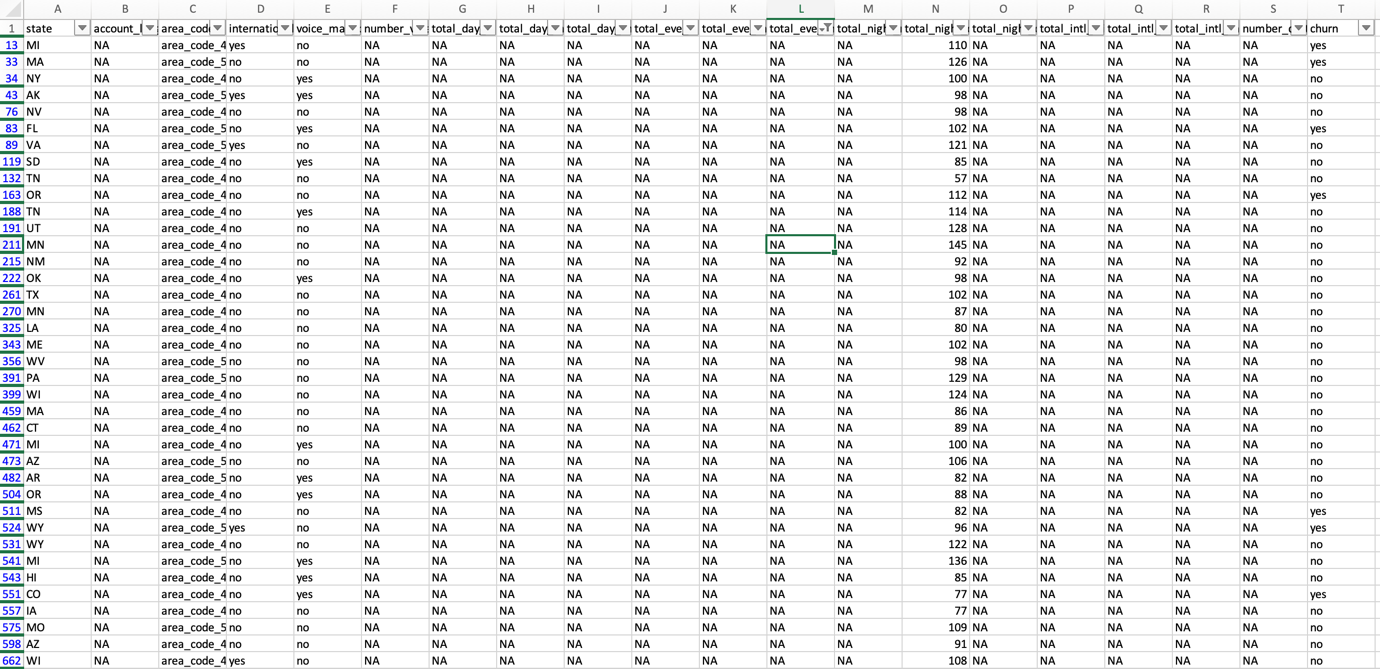
## A picture containing table Description automatically generated

## Data Cleaning:

## From the below image, we can see that the data has a significant amount of NA values in the majority of the variables which we would be using in the model building,

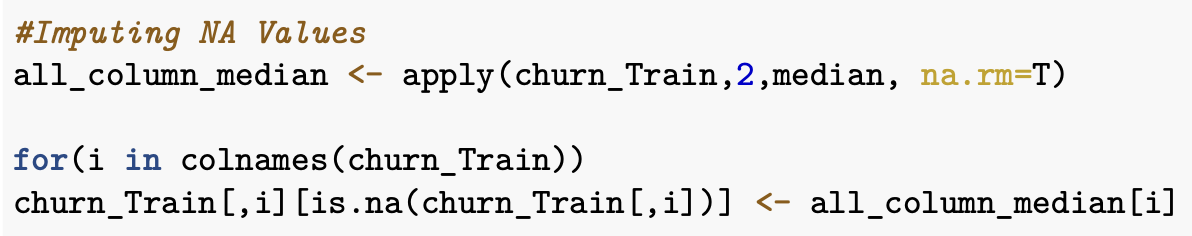


As we get to see below, when we tried to apply a filter on the column which has NAs in it we get to visualize a scattered pattern across all the variables, given that we couldn’t afford to lose the data, we would want to impute them instead. As removing the missing values would make the model fall short of data and it “***may not be a good predictive model”***, since it’s not seen a diversified amount of data.



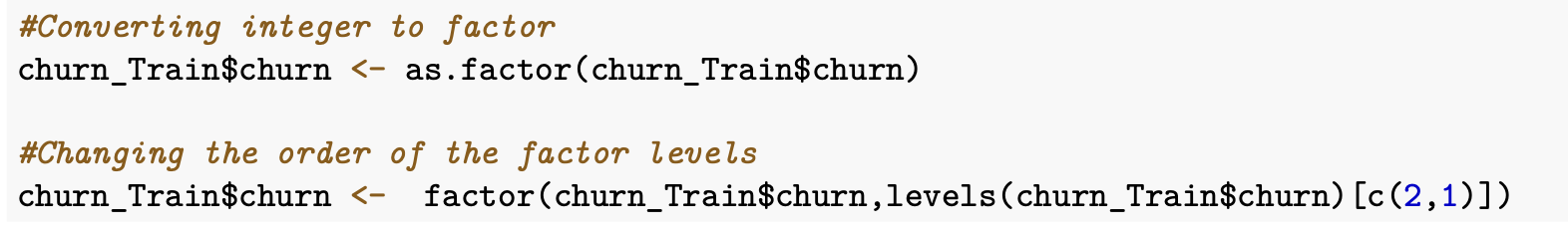
In order to impute the NAs we did use the median function to calculate the median value of a column and add it as a value to the NAs. Why Median? Why not Mean?

The Median is one of the functions of measuring central tendency and the beauty of using the median is that it doesn’t get affected by the outliers or extreme points whereas the mean is highly sensitive to extreme points and can easily get affected by it so we go with the median.

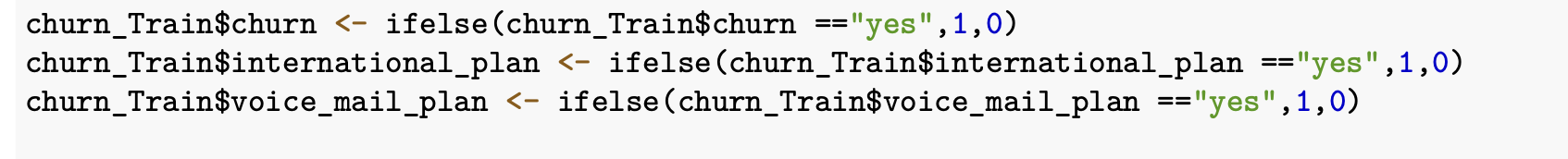


**Data Transformation:**

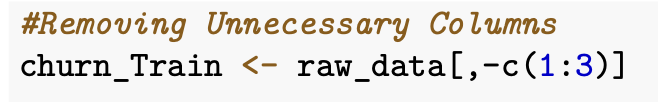
* Churn attributes is a predictor of whether a customer will cancel their service in the future or not. As Churn is a predictable variable we have converted it into a factor.
* Also, we are changing the order of the factor levels so that the prediction probabilities would correspond to “yes”.

****

* We are re-coding a few variables which have the binary values of “yes” and “no” to “1” and “0” so that they can be used in the analysis.

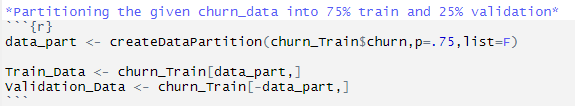
****

* The initial 3 variables in the dataset are categorical variables, these variables don’t have theoretical importance and we aren’t going to perform any analysis where we are going to differentiate between states. With this as the basis, we have removed the categorical variables.



**Details of Modelling Strategy:**

Before applying any kind of modelling strategy to the data it’s always advised to split the data into train vs validation. Here we split the data into a 3:1 ratio i.e., 75% into the **“Training Set”** and 25% into the ***Validation Set.***



After the data partition has been done we then start to build the models over the training set.

**Predictive Models:**

Here in this project, we have chosen four algorithms out of the many existing ones, to understand which among the four is recommended for predicting the churn rate.

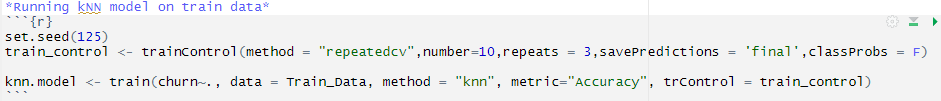
**Reasons for choosing Logistic Regression:**

A picture containing text

Description automatically generated

* Logistic Regression is an appropriate predictive analysis to conduct when the dependent variable is binary.
* Logistic regression is used to describe data and to explain the relationship between one dependent variable binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.
* Logistic regression assumes a linear relationship between each variable. However, if the customer is paying less for a particular service compared to a competitor, the chances of the customer leaving the service provider will vary exponentially.

**Reasons for choosing K-NN:**



* The k-NN algorithm is very simple to understand and equally easy to implement. To classify the new data point K-NN algorithm reads through the whole dataset to find out K nearest neighbours.
* K-NN is a non-parametric algorithm which means there need not be any assumptions to be met to implement K-NN. Parametric models like logistic regression have lots of assumptions to be met by data before it can be implemented which is not the case with K-NN.
* Given its instance-based learning; k-NN is a memory-based approach. The classifier immediately adapts as we collect new training data. It allows the algorithm to respond quickly to changes in the input during real-time use.

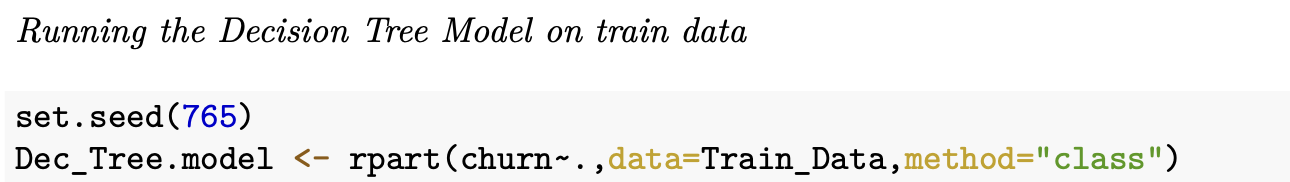
**Reasons for choosing Naïve Bayes:**

Graphical user interface, text, application

Description automatically generated

* Naive Bayes follows the concept of conditional probability, saying that it’s a better algorithm for ascertaining the probability of a given variable upon satisfying a given condition.

**Reasons for choosing Decision Tree:**



* One of the best models to run over a classification analysis, a single decision tree won’t work that great over a regression analysis but they work the best for the classification problems.
* It makes it easier to look at the classification analysis since the fancyrpartplot() helps us in knowing the flow of the variables from the root to the end leaf node.

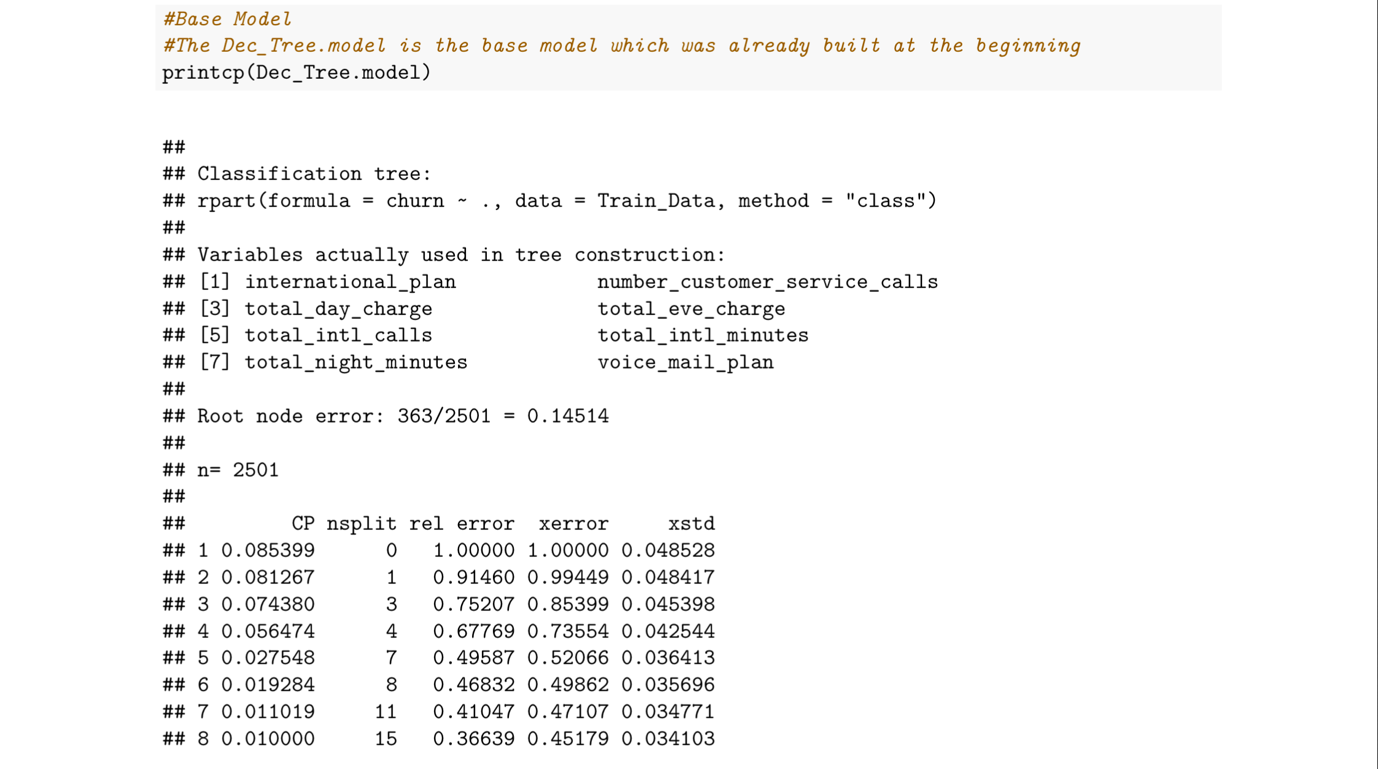
Diagram

Description automatically generated

* The decision tree model can be used for both classification and regression problems, and it is easy to interpret, understand, and visualize.
* The output of a decision tree can also be easily understood.
* Compared with other algorithms, data preparation during pre-processing in a decision tree requires less effort and does not require normalization of data.
* A decision tree is one of the quickest ways to identify relationships between variables and the most significant variable.
* New features can also be created for better target variable prediction.

**Complexity Parameter (CP):**

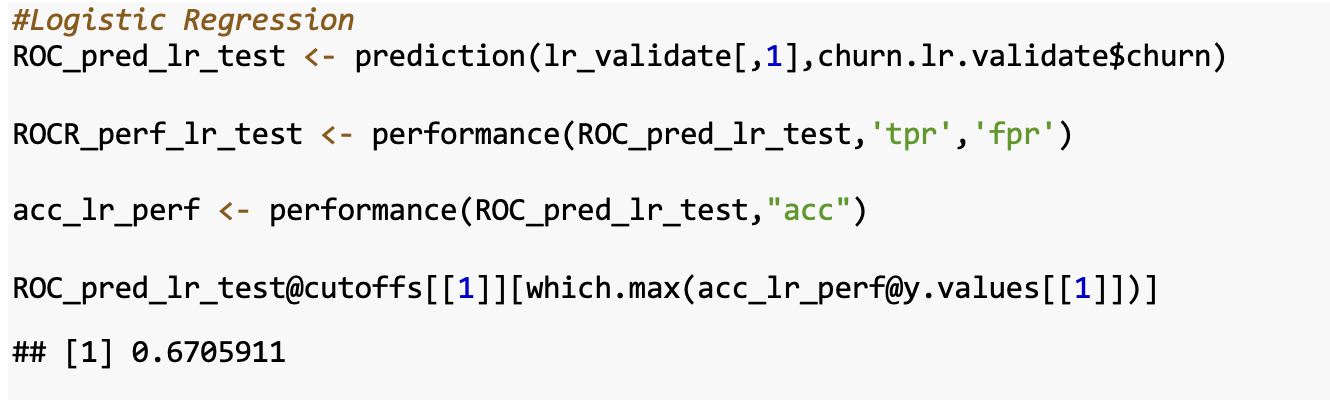
The **“printcp()”** function used on the decision tree model will henceforth give us a list of the variables which were used in taking the decisions and based on which the model will be able to predict the churn results on the unseen data,

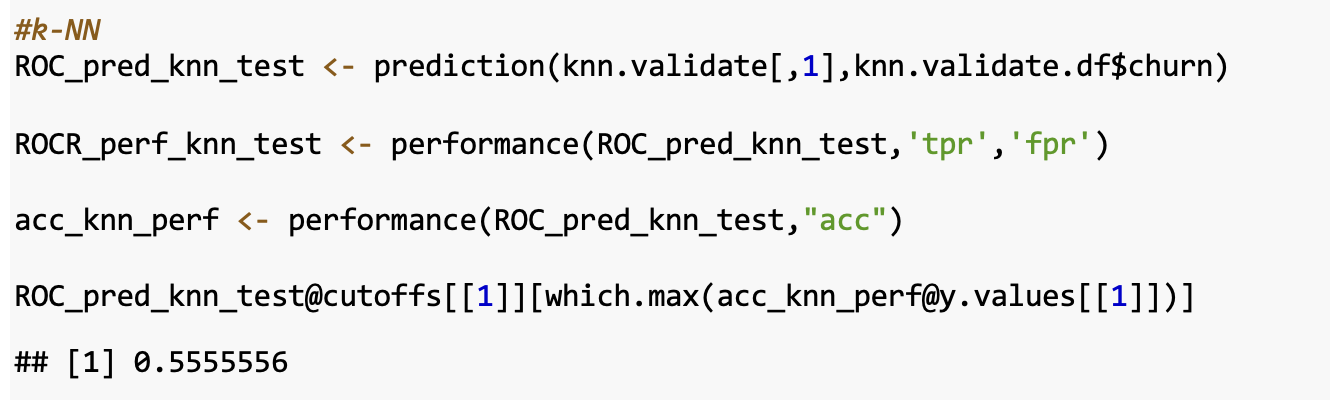


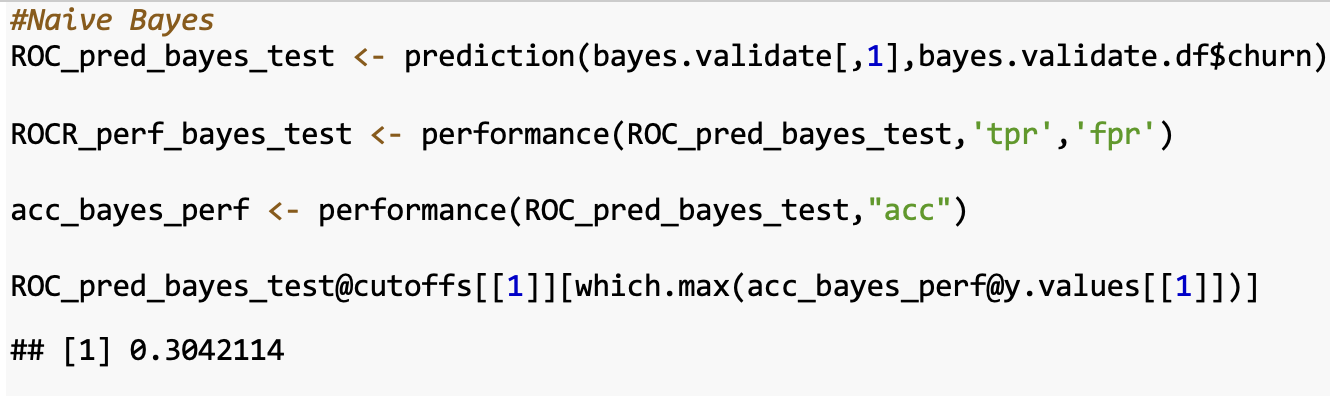
**Cut-Off Threshold:**

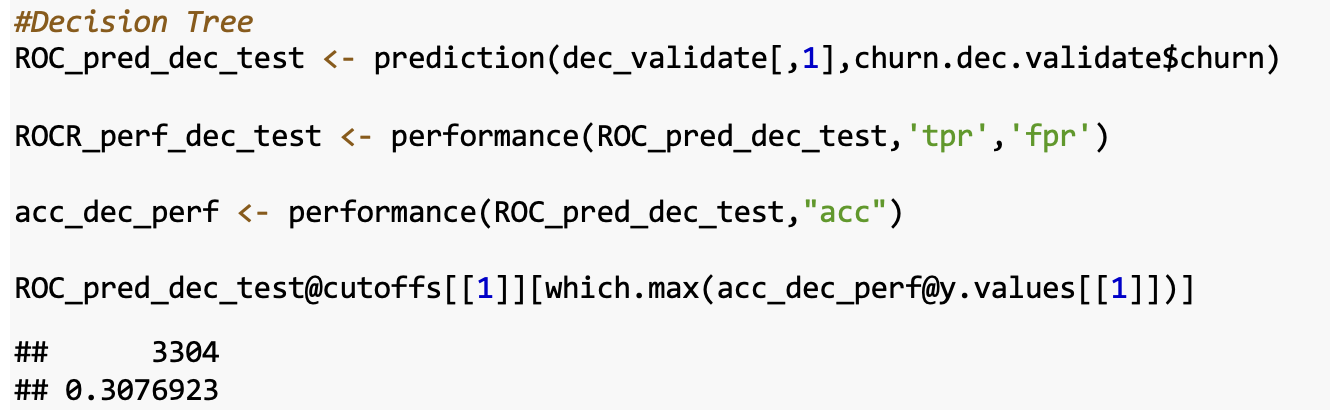
For each of the models built, we have used the ROC measure to come to the optimal point which can help us know who are the customers that are likely to churn and those who are not.

If the threshold is not set, the base probability that will be considered for cut-off is 0.5, i.e. >0.5 will be a churn and <0.5 will not be a churn.

****

****

****

****

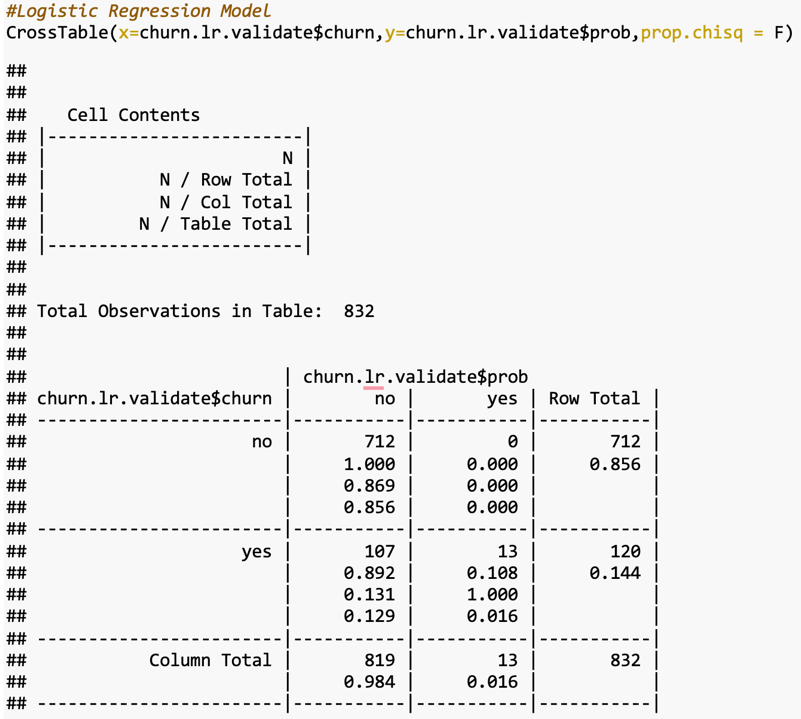
**The ideal cut-offs which we obtained for the models we built are,**

Logistic Regression – 0.6705911, k-NN – 0.555556, Naive Bayes – 0.3042114, Decision Trees – 0.3076923

**Estimation of Model’s Performance :**

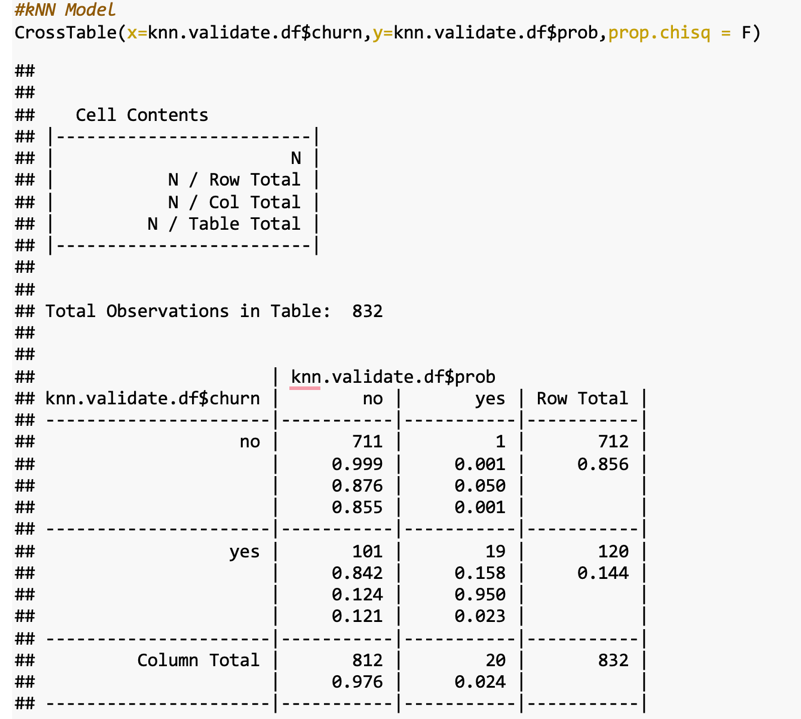
The performance of each of the models was assessed by predicting them over the validation set.

**Running a LR Model**



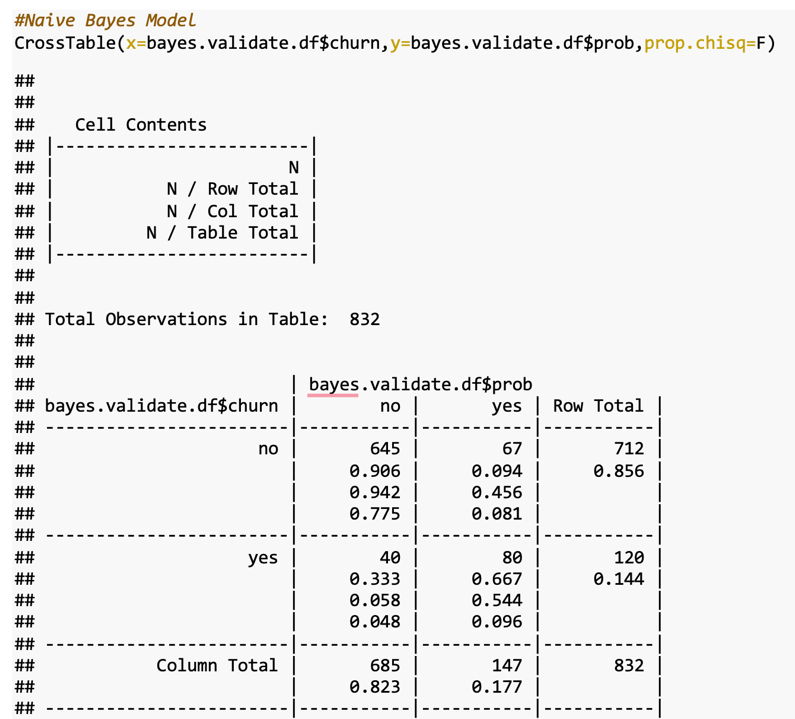
It was observed that the accuracy of the logistic regression model was 87.13% and Miscalculations/Errors are 107. The Specificity of the model is 100% and the Sensitivity of the model is 10.83%.

**Running a K-NN Model**



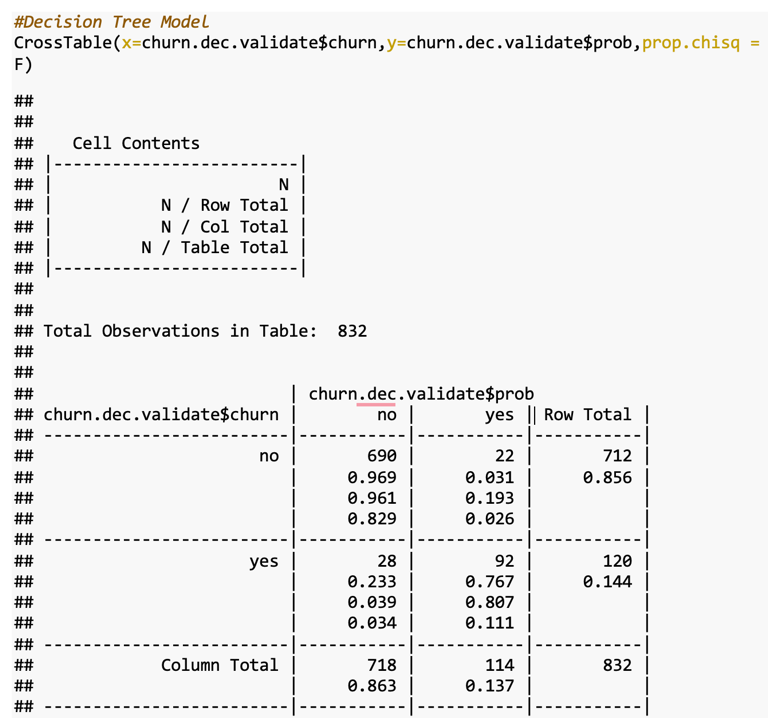
It was observed that the accuracy of the k-NN model was 87.74% and Miscalculations/Errors are 102. The Specificity of the model is 99.85% and the Sensitivity of the model is 15.83%.

**Running Naive Bayes Model**

****

It was observed that the accuracy of the Naive Bayes model was 87.13% and Miscalculations/Errors are 107. The Specificity of the model is 90.58% and the Sensitivity of the model is 66.66%.

**Running Decision Tree Model**



The decision tree’s accuracy was observed to be 93.99% which is higher than the other three models considered in the study. The errors and miscalculations are very less as well i.e. 50. Also, the Specificity of the model is 96.91% and the Sensitivity is 76.66%. Accuracy, Sensitivity and Specificity are comparatively high with less Error Rate so we are proceeding with the **“Decision Tree Model”** to be implemented on the **"Test Set".**

**Pruning the Decision Tree Model:**

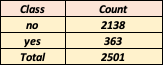
Pruning is a methodology to control the size of the tree by limiting its growth by adding some parameters.

But it was interesting to observe that pre-pruning and post-pruning both methodologies didn’t increase the accuracy and sensitivity, not even by 0.0001 per cent.

**This can let us know that pruning isn’t significant on the datasets which are imbalanced, thus we are applying the base decision tree model over the test set.**

**Findings:**

1. The predictive model was built over the “training data with 2501 observations out of which 2138 observations had the churn value as “no” and 363 observations had the churn value as “yes”. This can be a pure example of an “**Imbalanced Dataset”.**



1. It can be understood that model works well for most of the data that has a Churn value of “no”. This can be called out by the TNR/Specificity = 96.91%, since the trained model didn’t get to capture the trends much for the Churn value of “yes” the Sensitivity is lagging a bit short with 76.66%.
2. Imbalanced data can be a possible reason for less TPR and pruning the model isn't that effective as the dataset is highly imbalanced.
3. Factors having a significant impact on the customer churn responses are **“total evening charge”, “total international minutes”, and “total night minutes”** – based on decision tree flow.

**Conclusion:**

All in All, we can say that the three most important variables responsible for predicting the customer getting churned are **“total evening charge”, “total international minutes”, and “total night minutes”**.

By looking at the flow of the decisions attained by the decision trees we can see that, when the **“total evening charge”** is greater than or equal to 15, the customer is getting churned; the reason for this can be that customers who care about cost may find it expensive i.e. the telecom company might be charging high to those customers who tend to use their services abundantly during the evening's, this specific set of customer might find the services availed as expensive and will be wanting to churn to another service provider where the charges are less.

In the same way, when the “**total international minutes”** are greater than or equal to 13 and **“total night minutes”** are greater than or equal to 160, the customer is likely to churn to another service provider because the increase in call minutes will eventually make the call services expensive, this is how the telecom industry charge their customer. Eventually, this will lead to an increase in the overall cost of that particular telecom service plan to a specific customer. Most people will always want to keep their telecom bills as minimalistic as possible but when they find any rise in the charges they would want to churn to another service provider, here in our case we find the above-discussed three factors as the prominent factors leading to churn in the telecom sector.

**Recommendations:**

*“We hereby suggest to the stakeholder and executive chair of ABC Wireless Inc. that retaining the existing customers can be made possible in two ways*

* *Reducing the current pricing for the plans*
* *Providing better offers to the existing services*

*The first way can ideally be expensive to the company because, the telecom industry is ruled by three major competitors i.e. Verizon, AT&T and T-Mobile who hold 65% of the market share, for these companies reducing the price isn’t going to cost as much as it is going to cost us i.e. ABC Wireless Inc. So, this isn’t a suggestable method to retain the customers at the loss of the company’s revenue.*

*The second way is by far the best way we can think of i.e. driving the customer’s interest by providing additional offers at the same pricing plan,*

***International Minutes*** *– This is a feature which will be only used by international students or immigrants, we can drive certain offers to retain these customers by providing them –*

1. *100 MB of Data on weekends to call back to their home country at no extra cost,*
2. *3 Months Package free for initial 3 months if purchased a plan for a 1-year subscription.*

***Evening Charge*** *– All sorts of people can fall under this group, so we can suggest adding some features such as between 4 pm to 7 pm if the people are on Golden Plan they get to use data unlimited during this time range.*

**---------------------------------------------------------------------------------------------**

***Individual Student’s Contribution to the project:***

**Nikhil Kumar Sampath** – Building various Predictive Models, Running them over the validation set, Building Cross Tables of all the models, Assessing the performance of the models, Pruning Analysis, Prediction over the Test set, Understanding the decision tree flow, Project Report Summarisation, PowerPoint Ideas and voiceover for the presentation.

**Kiran Kour** -Pruning the Decision Tree, Evaluating the flow of the decision tree, Checking for the Variable Importance, Data Transformation, Framing Conclusion for the analysis, voiceover for the presentation and presentation theme.

**Akhil Yada and Sandhya Cheepurupalli** – Data Loading and Cleaning, PowerPoint Presentation and Project Report.

**---------------------------------------------------------------------------------------------**