**Project Goal**

ABC wireless Inc. wants to achieve churn reduction . So to enable that we apply data science principles and analytics to address the customers’ churn issue.

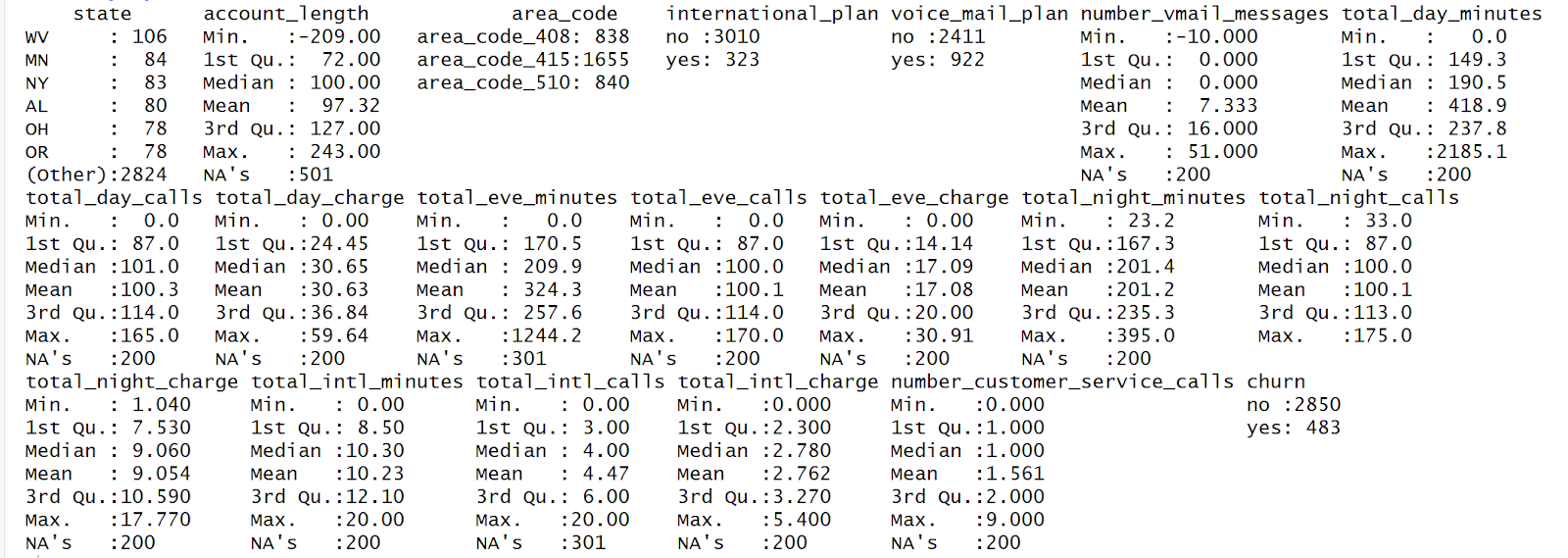
The goal is to develop a model that can best predict the probability of a discrete outcome (defined as 1 or 0, for the “yes” or “No” of churn variable) based on a set of explanatory input variables related to that outcome using a set of past inputs and outcomes.

**Data Overview & Exploration**

**First Impression** :  Given data is the historical data of customers of  ABC wireless Inc. It has total 3333 records and 20 variables. One of them is the **churn** ,which is our target variable. It is binomial -’yes’ or ‘no’ . The data contains some negative values as well as NA values.The data will require preprocessing before we use it to build the model.

Given data is stored in dataframe **‘mydata’** and for a closer look at all variables and their distribution we use the **summary** function.

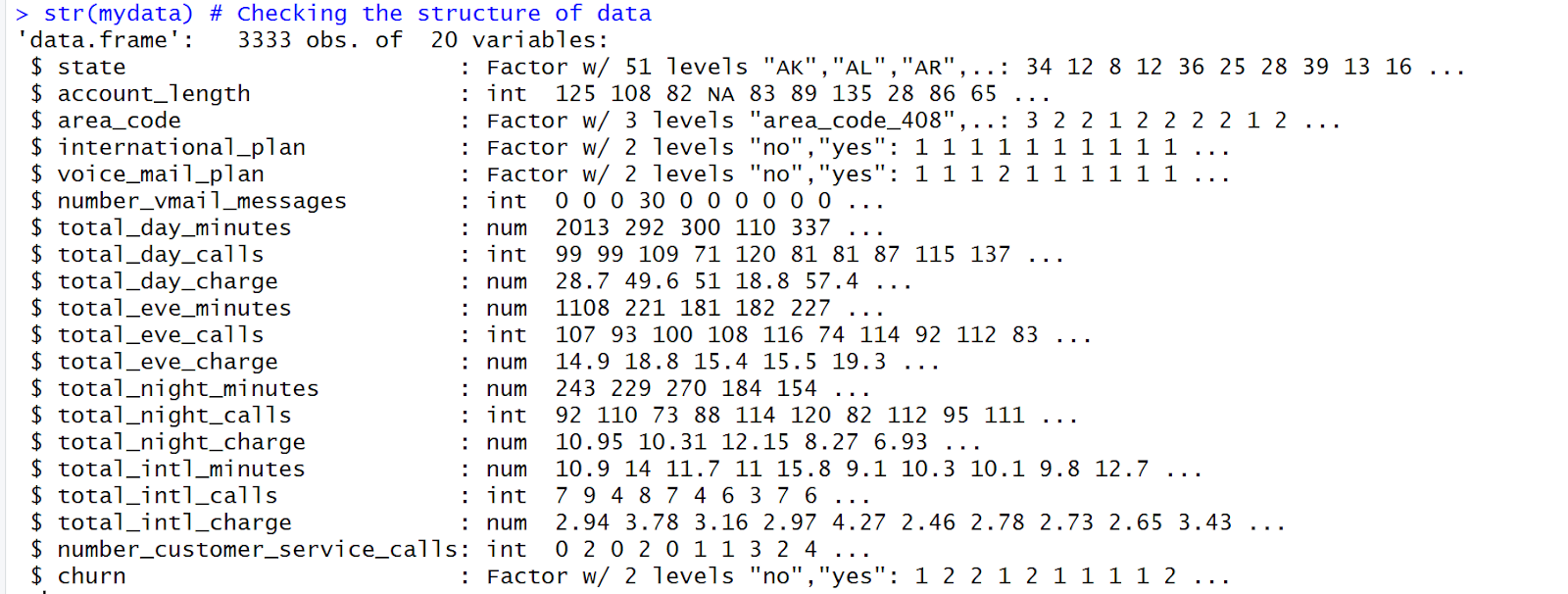
* **Summary** : the given data(historical data)



The summary of data gives the Statistical overview of each column or variable . It also shows the NA or missing values of each column. The data is not very suggestive at this step but, for us to draw more inferences from it we need to process the data . There are 200 records that have NA for all the columns. There are Negative values present  for the variables account\_length and number\_vmail\_messages, which is to be taken care of .

Before we begin with processing the data for the identified issues , we need to find the datatype of all variables. We check how many of them are numerical and categorical variables and if all are in the required formats. We do that using the **str function.**

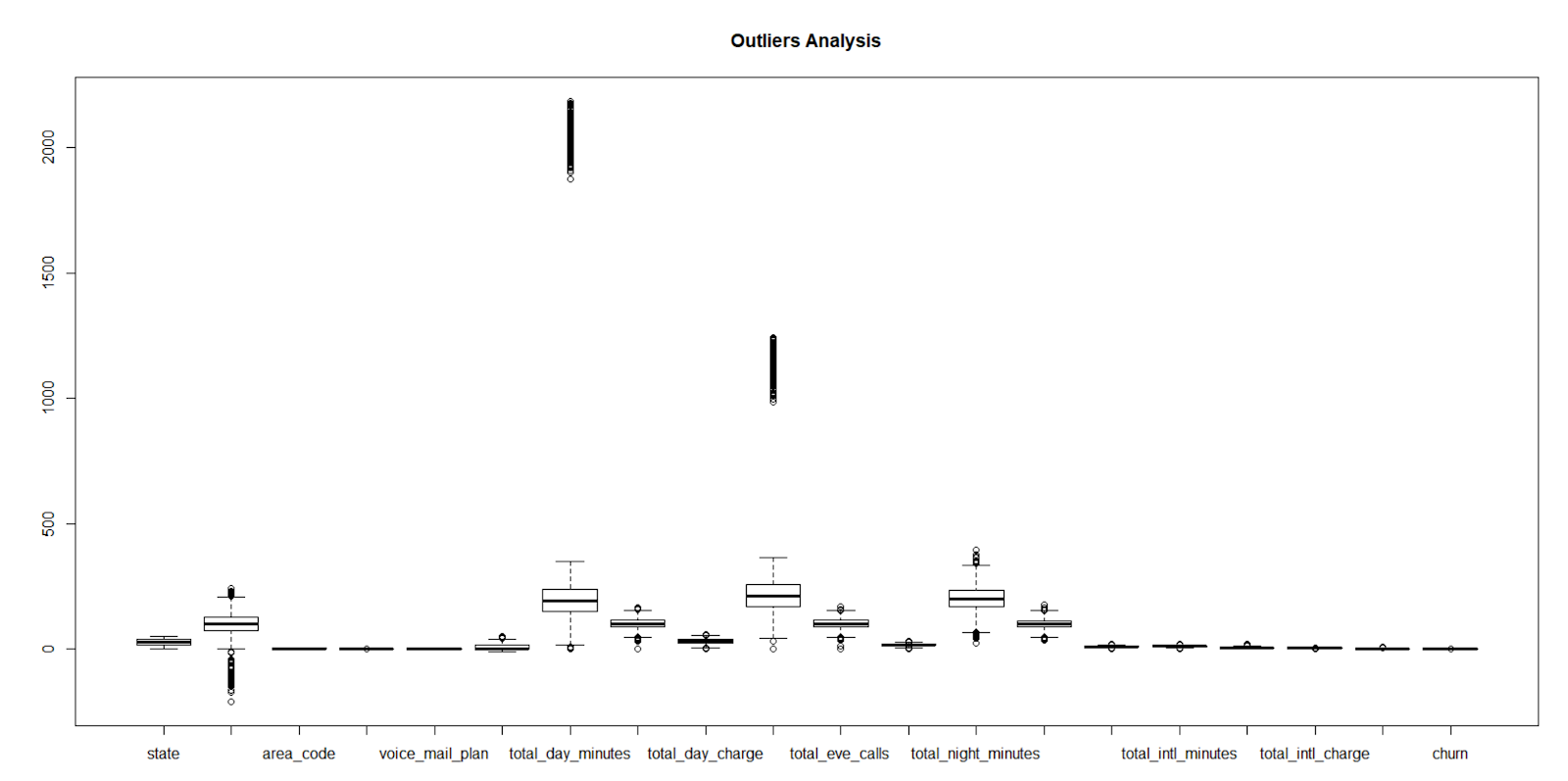
* **Structure**

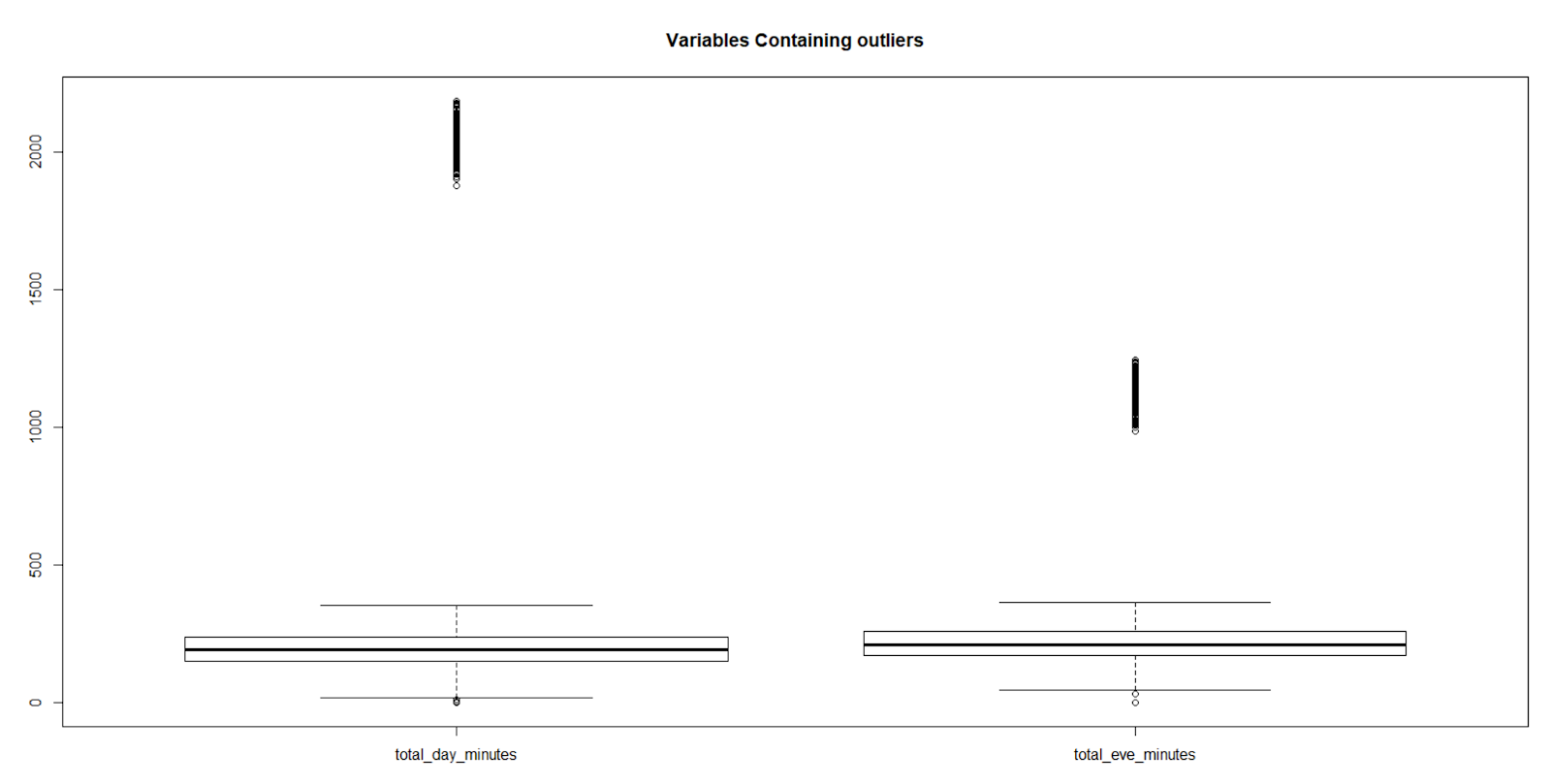


There are mostly numeric and integer data types in most variables and that seems about right . There are 5 variables with factor data types as follows:

1. State  - 51 levels - two letter code
2. Area code -  three levels - 1,2,3
3. International Plan - two levels - 1 , 2  indicating - no or yes
4. Voice mail Plan    - two levels - 1 , 2  indicating - no or yes
5. Churn -  two levels - 0 , 1  incidicating - no or yes

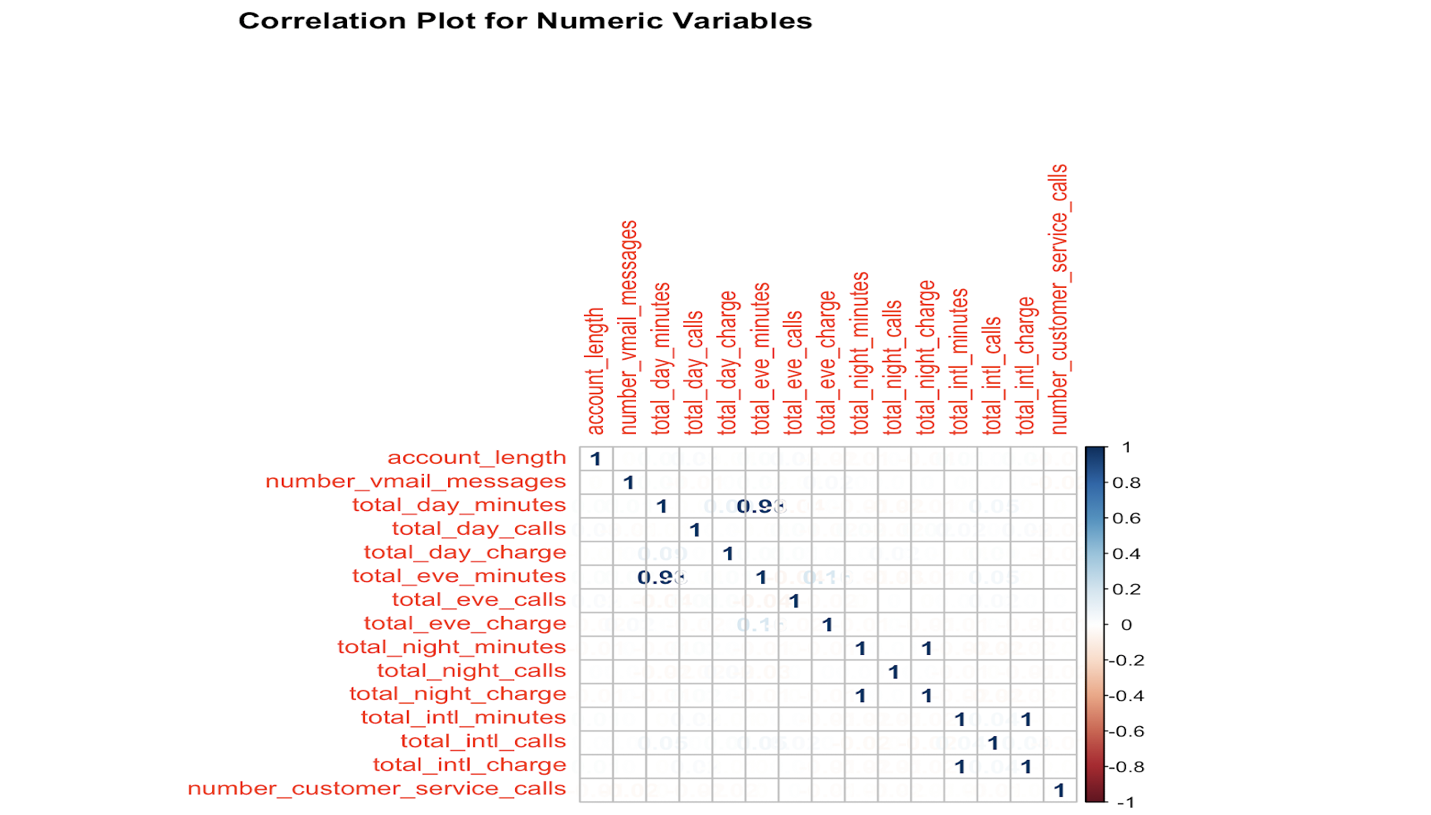
**Outlier Analysis :** To check for the outliers in the dataset, we plot them as **Boxplot.**





Looking at the plot, we can figure out that outlier are present in the columns of  **total\_day\_minutes** and  **total\_eve\_minutes** variable. There are 12% of outliers in each of these variables.

**Finding Correlation:**Looking at the correlation between the numeric variables, we tried to analyze the dependencies among the variables using **corrplot** function.

****

There exists a strong correlation between the three set of variables - **total\_eve\_minutes** and

**total\_day\_minutes** of 0.90 , **total\_night\_charge** and **total\_night\_minutes** of 1 and **total\_intl\_charge** and **total\_intl\_calls** of 1. We can later use this information while selecting features for our model.

**Data Preprocessing**

**Missing Values:** We have not dropped any columns or rows with missing values. All the missing values are imputed.

**Outlier Handling**: As the percentage of outliers is 12%, which is very less, we did not remove them completely form the variables.

**Handling Negative:** Negative values present in account\_length  and number\_vmail\_messages columns are changed to absolute values using **abs function**.

**Missing value Imputation:** The missing values are imputed with the mean of each column.For the two columns total\_day\_minutes and total\_eve\_minutes with outliers , we calculated the mean excluding outlier values to impute missing values.

**Modeling Strategy**

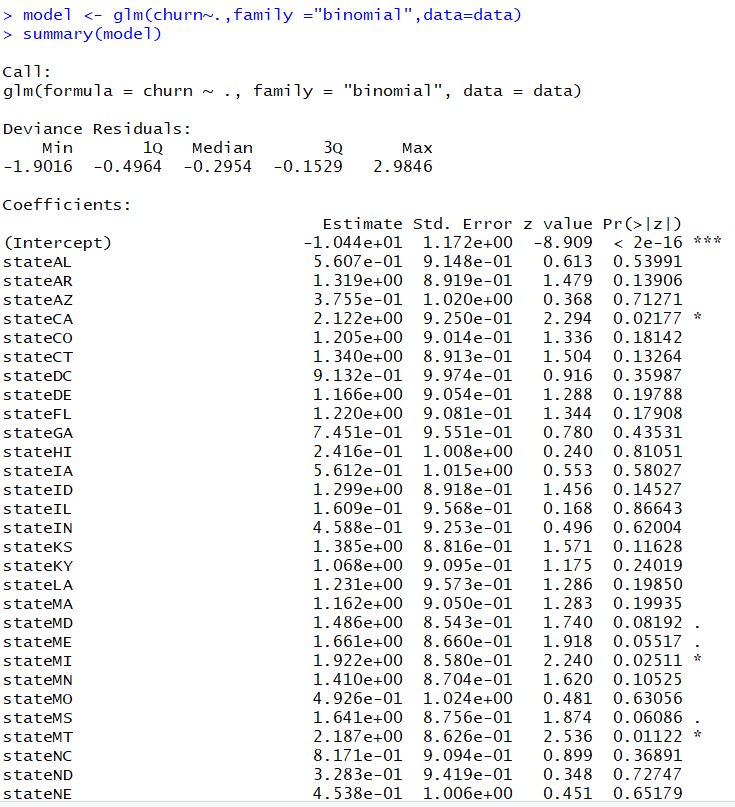
We are using Logistic regression model for prediction since the prediction data requires binomial classification.

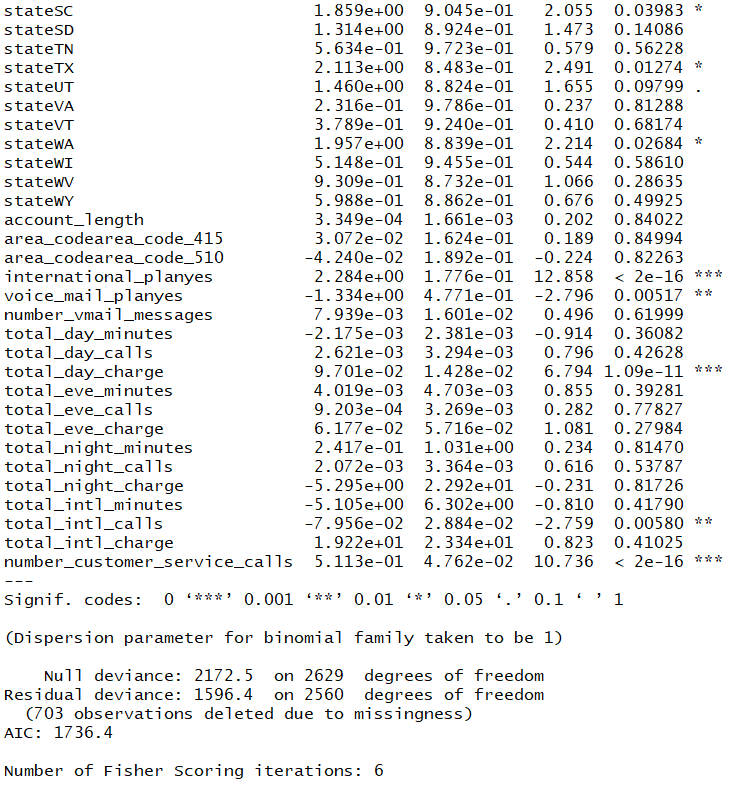
We are using the cleansed data to build the model, and consider the significant variables.

The significant variables are state, international\_plan, voice\_mail\_plan, total\_day\_charge, total\_intl\_calls, number\_customer\_service\_calls.

We also consider **state** as significant variable even though the p value is less as we think the state variable can also play a part in determining the outcome.

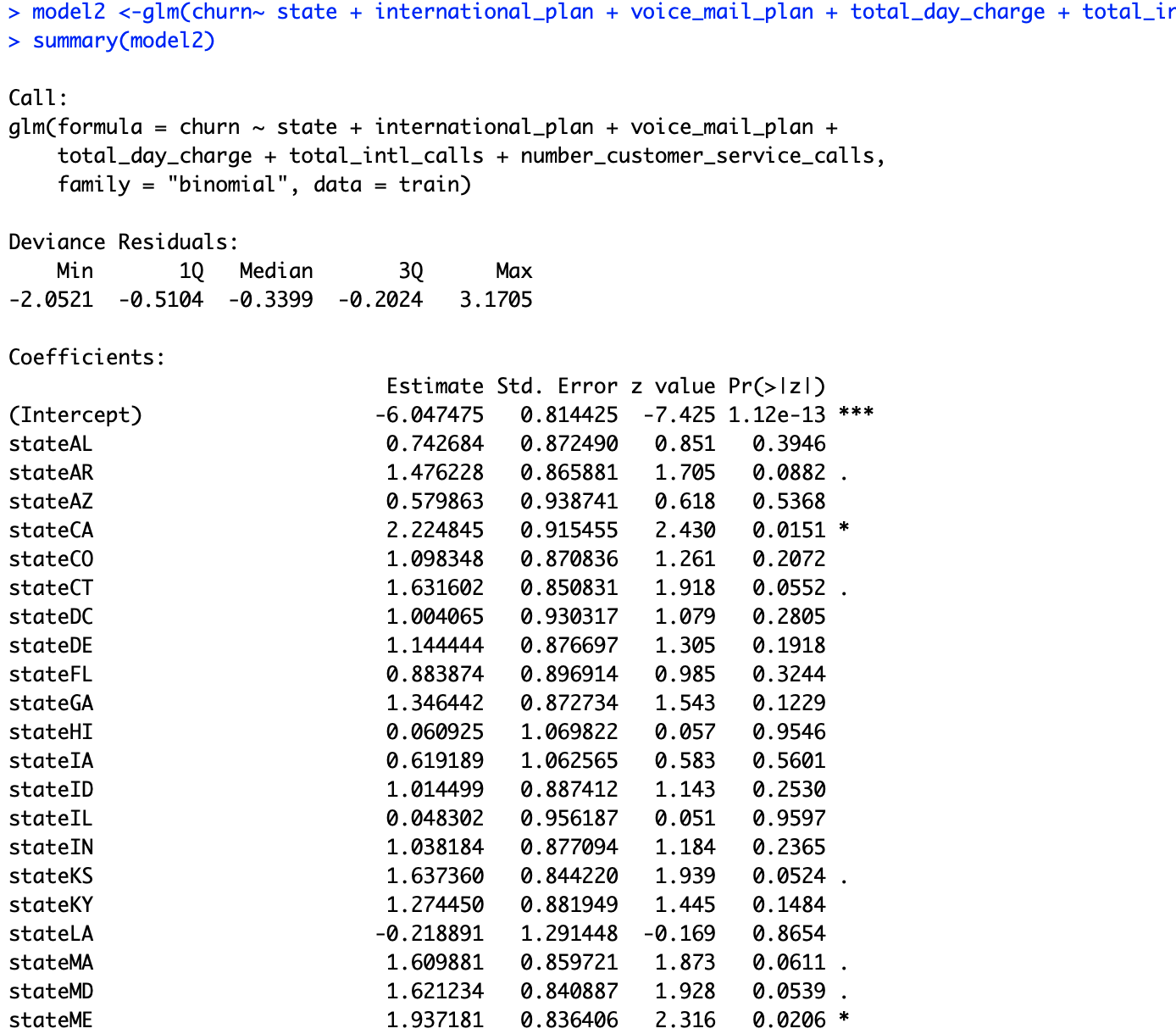
* **Finding Significant Variables:**

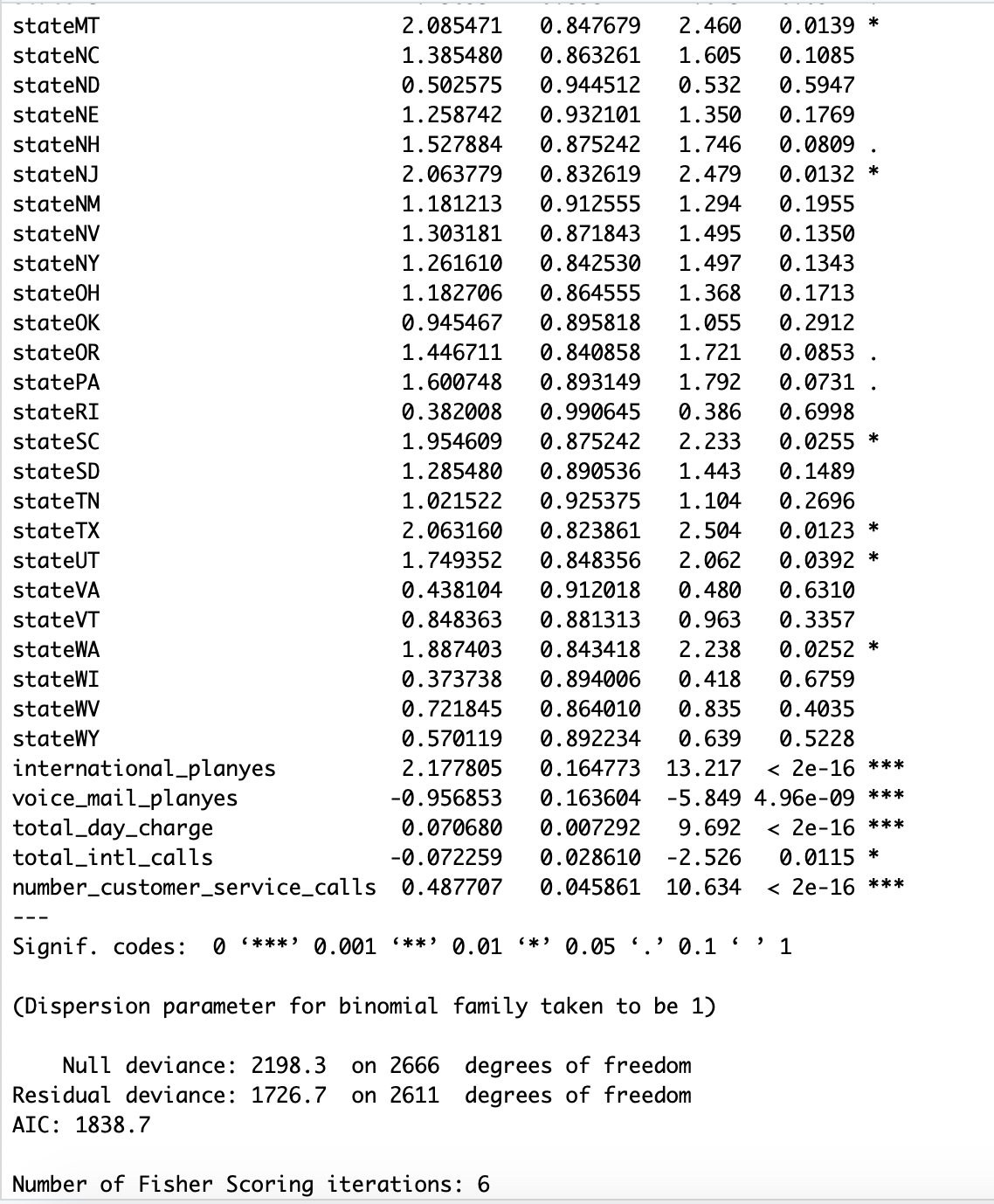




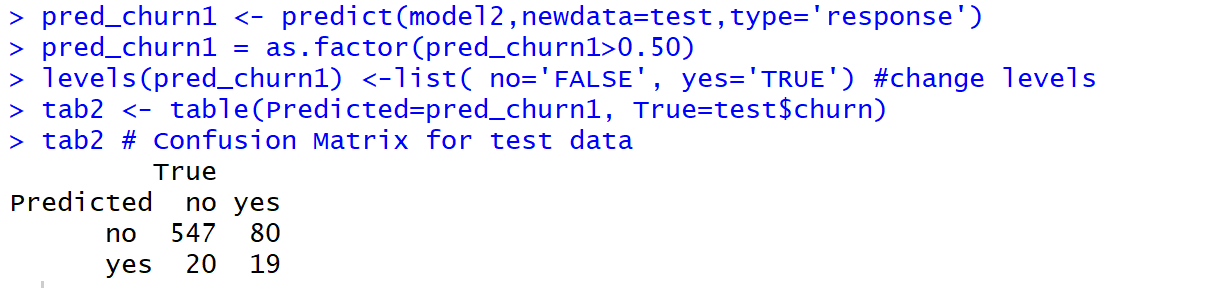
**Model building and testing:** We split the data into a training (80%) and testing (20%) data sets so that we can compare how well our model performs on test data using the model built on the training data set. We used seed = 1234 to replicate the results**.**

Model is built on the training data with significant variables to predict the target variable **churn**.





We use the model created on the train data to predict for the churn probability of the test data using **Predict function.**

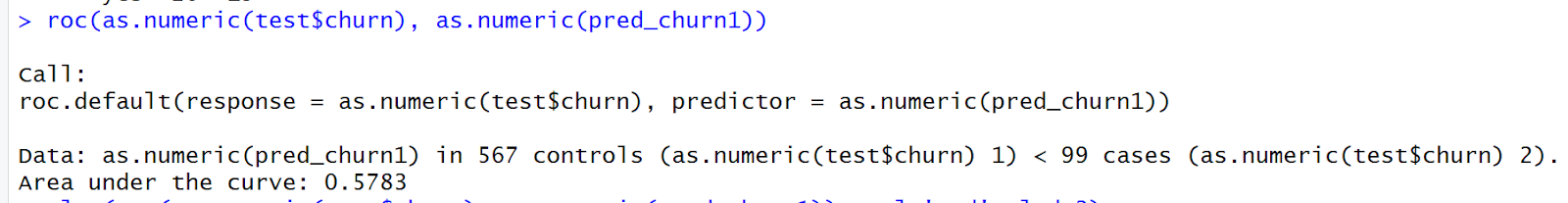
****

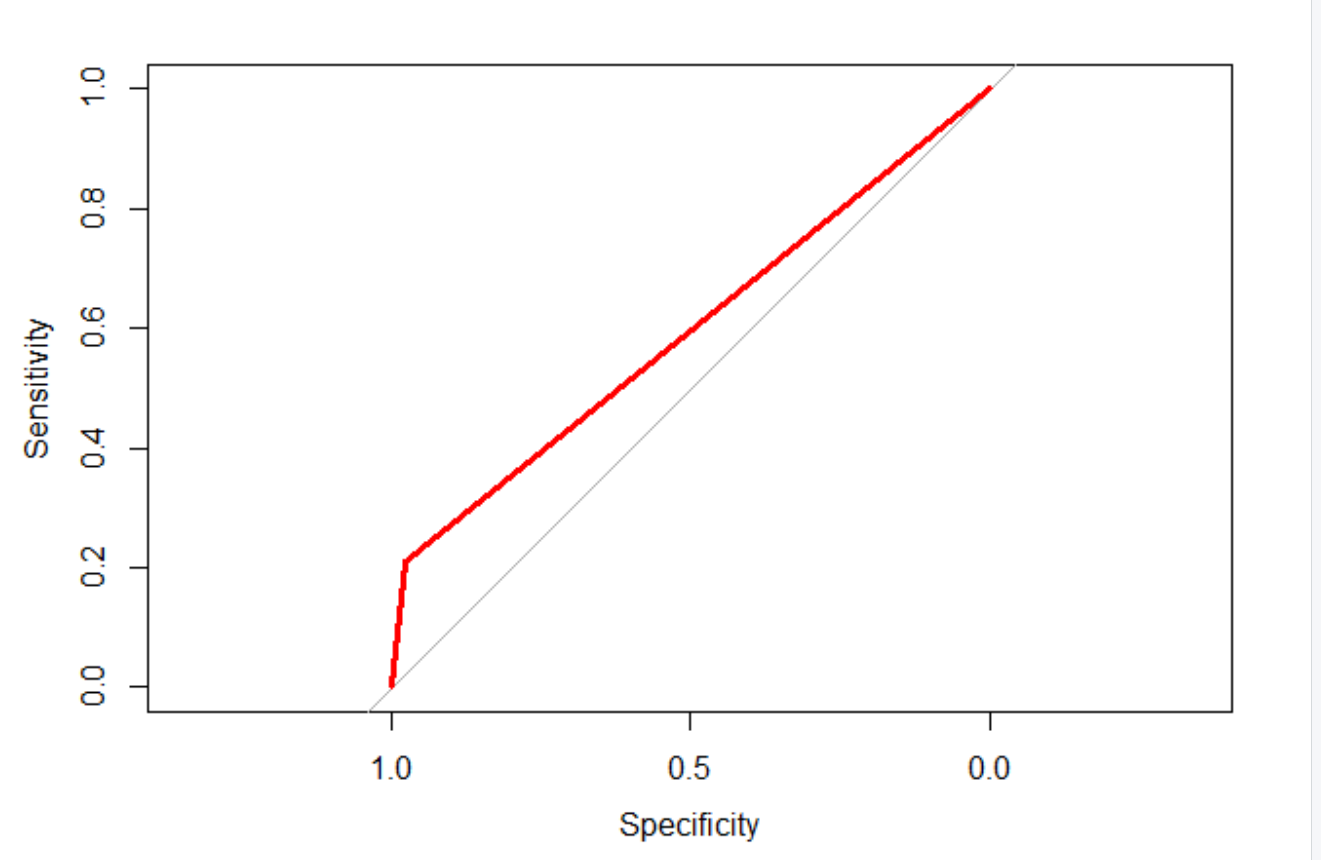
**Estimation of Model’s performance**

**Confusion Matrix**: In order to estimate the accuracy of the model, we check the confusion matrix provided by predicting the test model, and looking at the True Positive numbers being high and the False negative numbers being low.

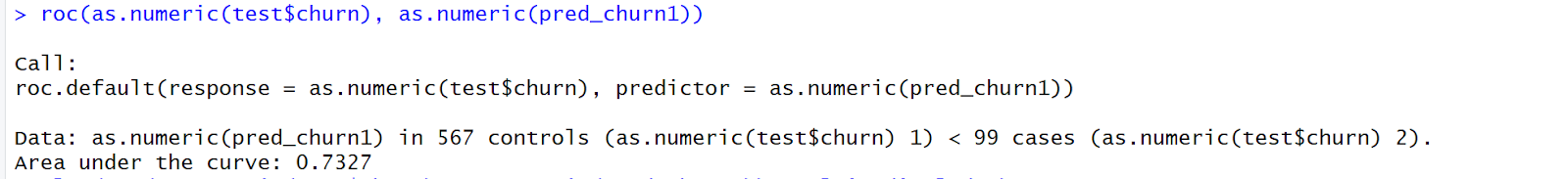
|  |  |  |
| --- | --- | --- |
| Confusion Matrix | Actual | |
| Predicted | No | Yes |
| No | 547  True negative | 80  False Positive |
| Yes | 20  False Negative | 19  True Positive |

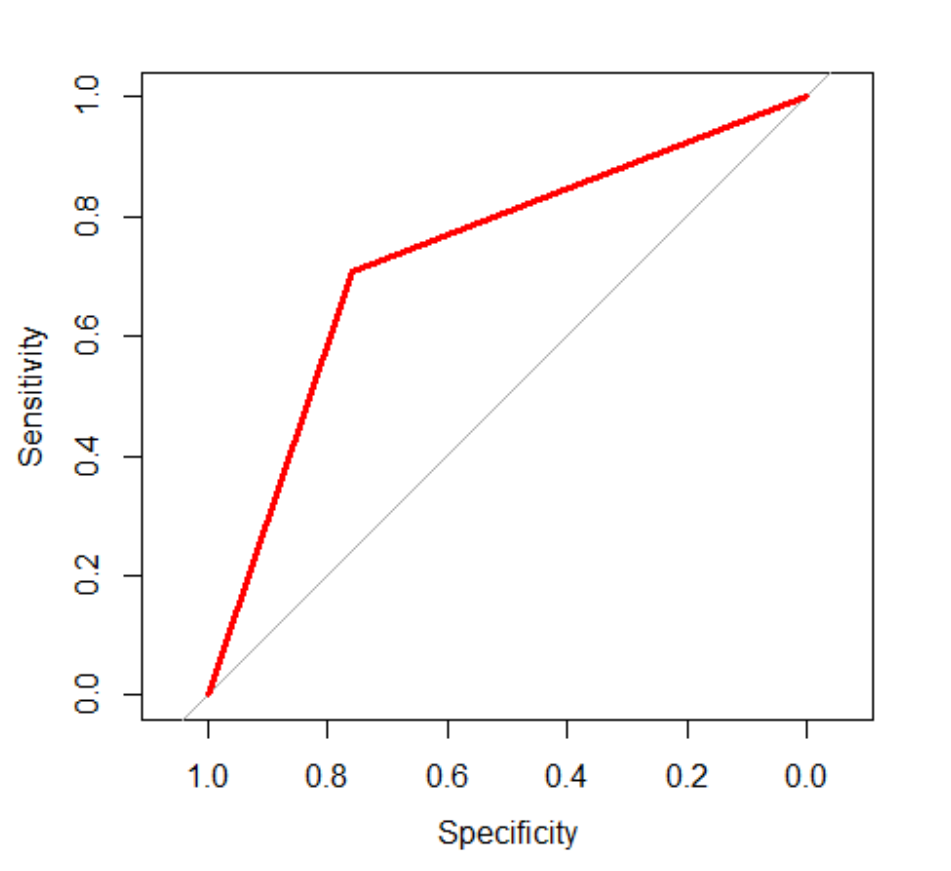
**AUC (Area under Curve):** For further checks, the second method we are using is the AUC of ROC (Receiver Operator Characteristic). We chose 0.5 as the threshold for making “yes” (or positive) predictions. The area under the curve for the model is 58% which is very good indicator of our model being very accurate.



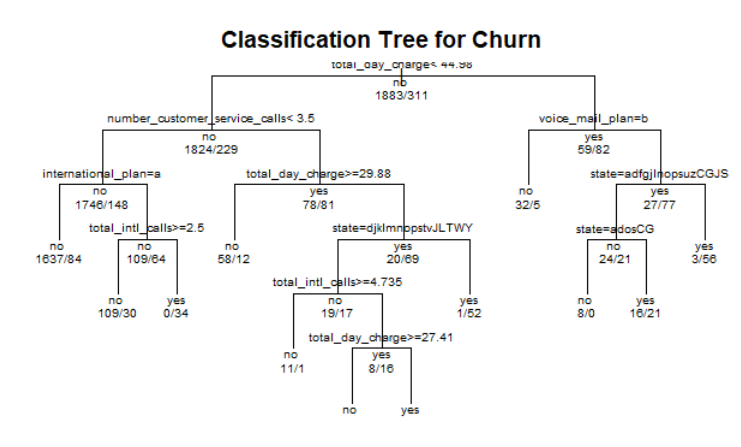
****

If  we set pred\_churn > 0.15 , It gives better AUC and more correct values of "yes" though the misclassification error is slightly higher.



****

**Classification Tree:** As third method, we used classification tree for all the calls considered in churn Dataset. The decision is made on basis of number of calls  and the churn factor having values true and false.

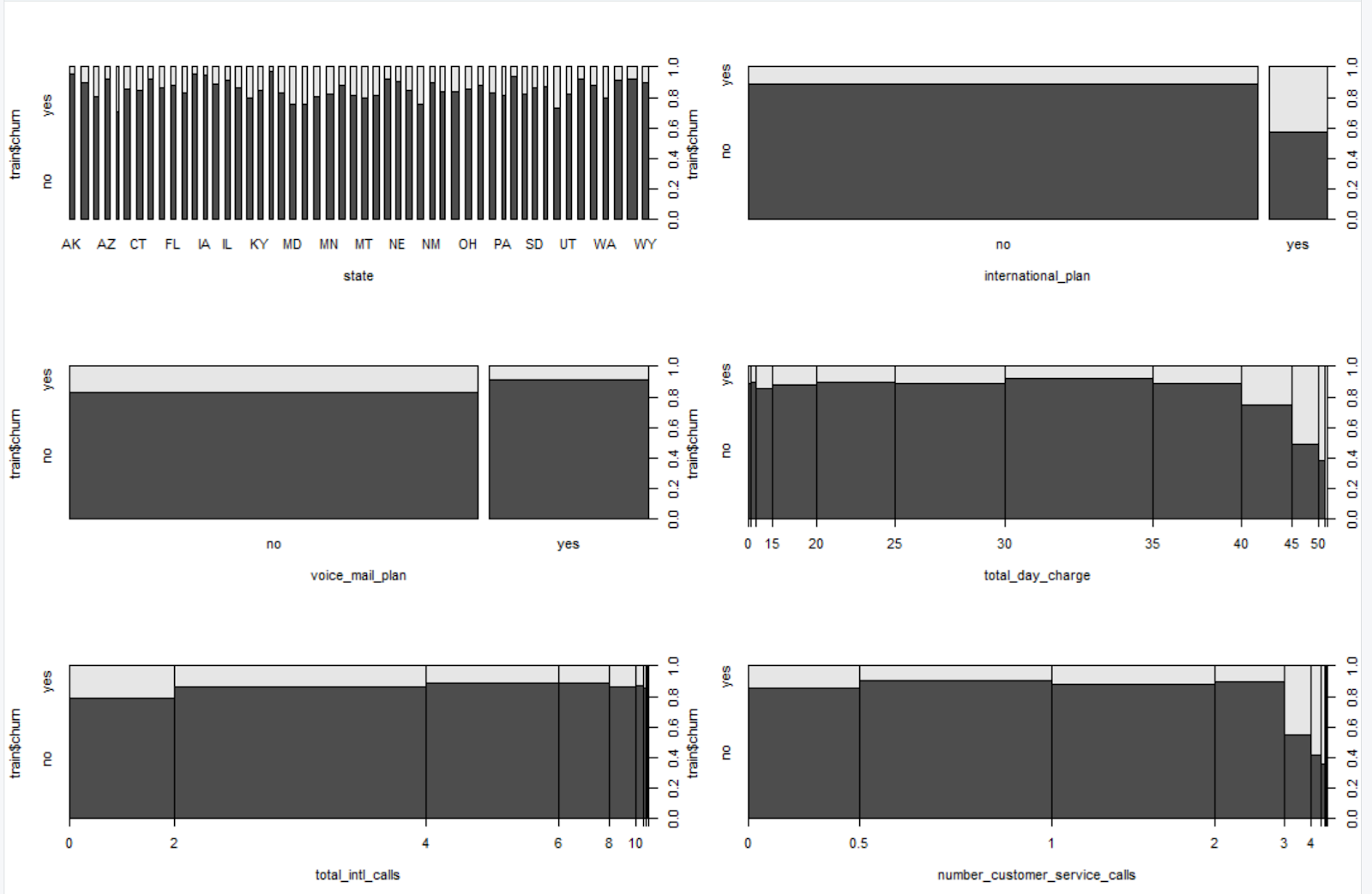


**Insights and Conclusions**

The company wants to focus on not losing the existing customers as marketing the product and gaining new customers is expensive as compared to retaining the customers. The model we developed can predict 75 % accurate results.

The logistic regression model  can help the company to determine the customers that are likely to churn. The model uses state, international\_plan, voice\_mail\_plan , total\_day\_charge , total\_intl\_calls, number\_customer\_service\_calls as predictors.

Depending on the outcome of the model, the company can provide lucrative offers only by analyzing the details of these significant variables that we have used in our model. This can help them focus only on the areas which is causing churn instead of analyzing each and every detail.The relation between the churn and these variables in the historical data is as shown below:



In addition to these, the company can have a new variable called **‘feedback’** which can significantly improve churn prediction. This can be a categorical variable(levels : Excellent, Satisfactory, Average, Unsatisfactory, Poor) where the company can have feedback from the customers about their service once in every six months. The company can analyze feedback values of variable  and decide on what aspect do they need to improve to retain the customer. And also the feedback variable will help in better prediction of churn probability.

The company should really focus on improving the after sales services in order to retain more and more customers.This could be achieved by developing a system where a customer can know the way the company is using to resolve his/her query, the tentative date by which the problem will get fixed and how to provide the reference for the issue customer reported and the methods of further inquiry.

Since we have considered **state** as significant variable, the offers should be targeted to the type of customers present in various geographic bounds , demographics and market survey of such potential and existing markets is must.

**A better prediction technique**

J48 Algorithm

J48 (formula, data, subset, control= Weka\_control ()

J48 Decision Tree Technique

library(RWeka)

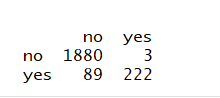
tree<- J48(train$churn~.,data=train)

tree

table<-table(train$churn,predict(tree))

table

plot(table)



J48 construction is like a flowchart. A test applied on an attribute is denoted by internal node, its effect is denoted by a branch and class labels are presented by leaf nodes. Process is divided in two levels, one is Division of root is recursively based on selection of attribute for all training examples at the tree construction and second is that the noise or outliers  branches are identified and removed by Tree pruning. Rules can be classified from the tree. If then statement is used to represent the knowledge. For each path from root to a leaf one rule is created.