

**Project Submitted By:**

**Anusha Kokkinti (Z1856785)**

**Tejaswini Mummadi (Z1859635)**

**Shruti Jituri (Z1840273)**

**Contents**

* Introduction
* Problem Description
* Data Resource
* Data Visualization
* Descriptive Analytics
* Predictive analytics
  + Logistic Regression
  + Decision Tree
  + Naïve Bayes
* Conclusion

**Introduction:**

Customer churnalso known as Customer attrition is the loss of customers. Customer analytics is a process by which data from customer behavior is used to make key business decisions via predictive analytics.

Telephone service companies or Internet service providers are using analysis of customer attrition because cost of retaining an existing customer is far less as compared to acquiring the new customer. We are trying to Predict customer behavior to retain customers.

**Problem Description:**

Business Need: The Telecom company marketing team wants to identify the customers who has the high probability of churning (Shifting to other telecom provider) so that they can offer some perks to hold back the customers.

* The information they have are the details of the customers such as gender, if senior citizen, do they have dependents, do they have partners, tenure they are with current provider, which services they are using from the provider, on what contract they are, do they enroll for paperless billing or not, what payment method they are using, monthly and total charges they are paying or used to pay.
* The team has the information of about 7043 customers with a good mixture of people who have churned and who have not left the company.

**Data Resource:**

Below are some of services which are being offered

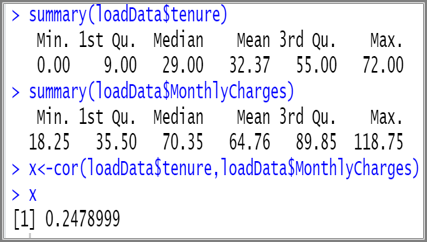
* PhoneService
* MultipleLines
* InternetService
* OnlineSecurity
* OnlineBackup
* DeviceProtection
* TechSupport
* StreamingTV
* StreamingMovies

Apart from above services or columns, our data file contains below columns

* gender
* CustomerID
* SeniorCitizen
* Partner
* Dependents
* tenure
* Contract
* PaperlessBilling
* PaymentMethod
* MonthlyCharges
* TotalCharges
* churn

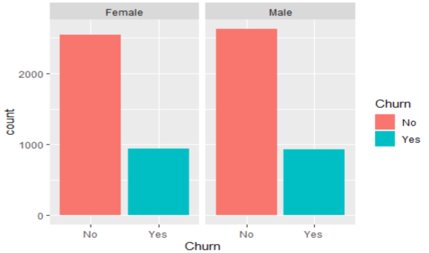
Target variable is churn we are trying to predict the risk of customer churn

**Descriptive Analytics:**

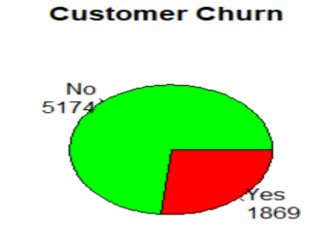


* Looking at max value of a tenure we know 1. Maximum tenure a customer could stay with the current provider2.Maximum data availability with us.
* Correlation value indicates a low correlation between Tenure and Monthly charges

**Data Visualization:**



Graph1: Gender distribution for the customer attrition.



Graph 2: Customer Churn Rate

From above graph we know total of 1869 customer churned which accounts for 26% churn rate and hence further data analysis is done

**Predictive Analytics:**

We have different models that can be used to address this type of prediction probability problem. Models we used are

1) Decision Tree

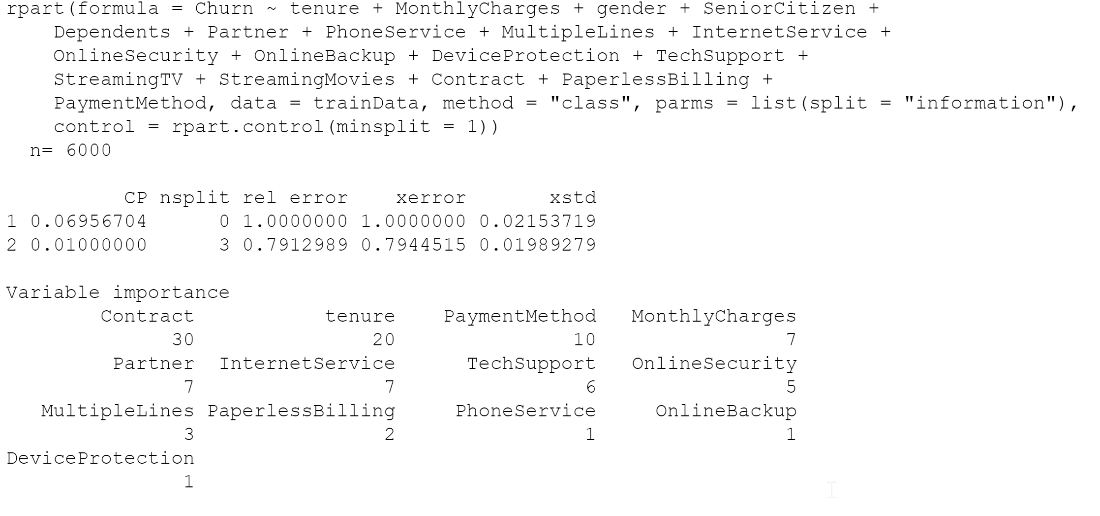
2) Naïve Bayes Algorithm

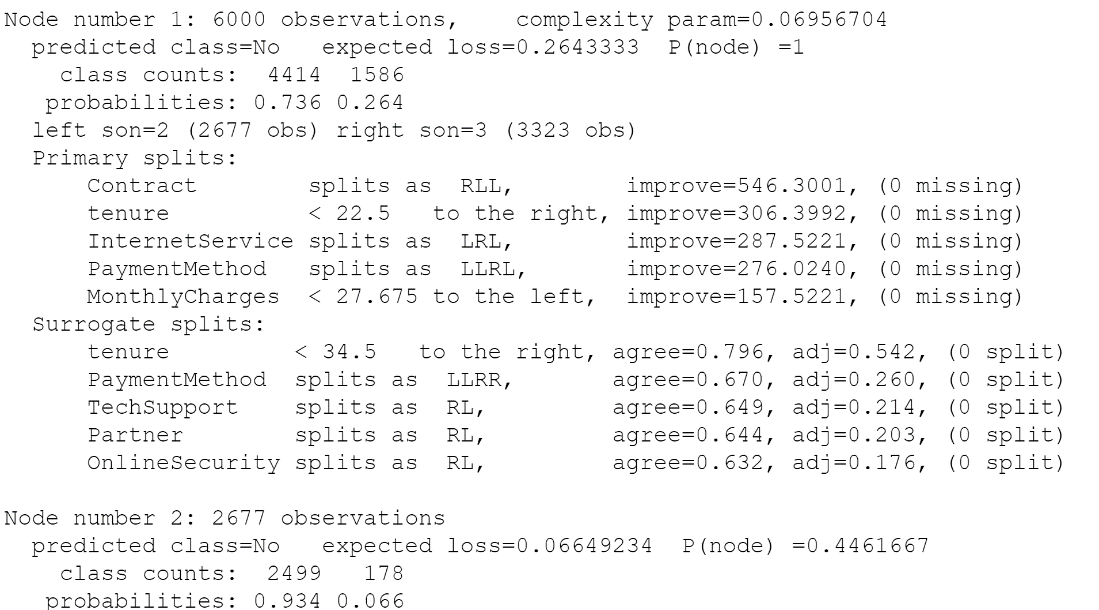
3) Logistic Regression

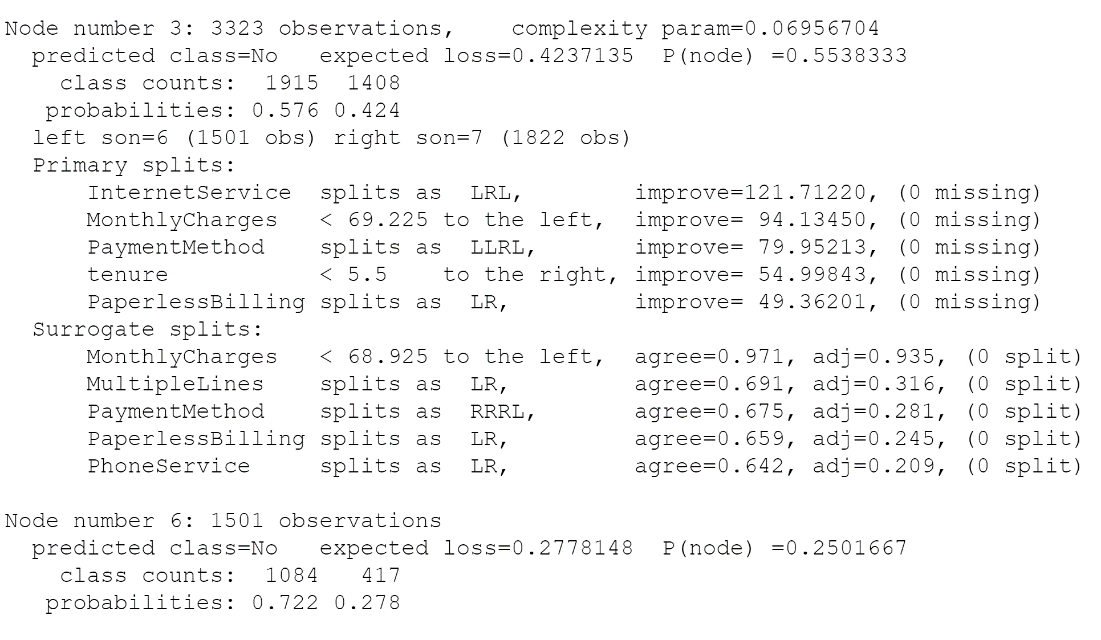
**1.Decision Tree:**

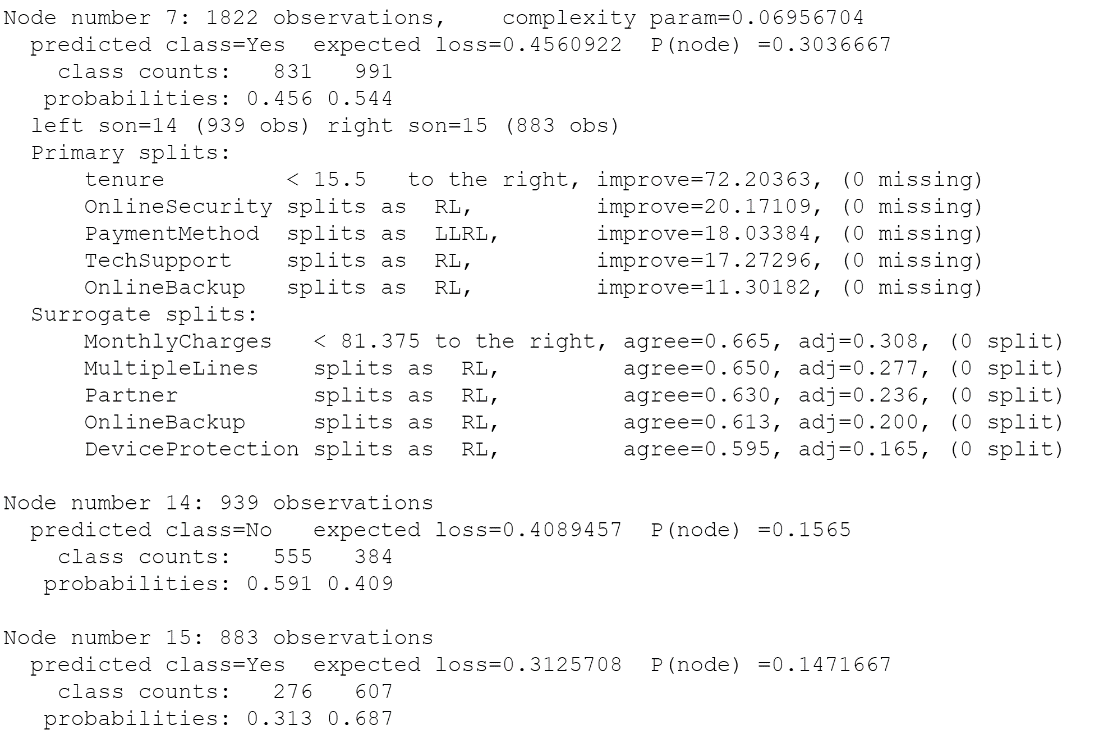
Decision Tree is implemented for the variable Churn which has classes “YES” and “NO”. The model includes all the other predictor variables.

The results are as follows:

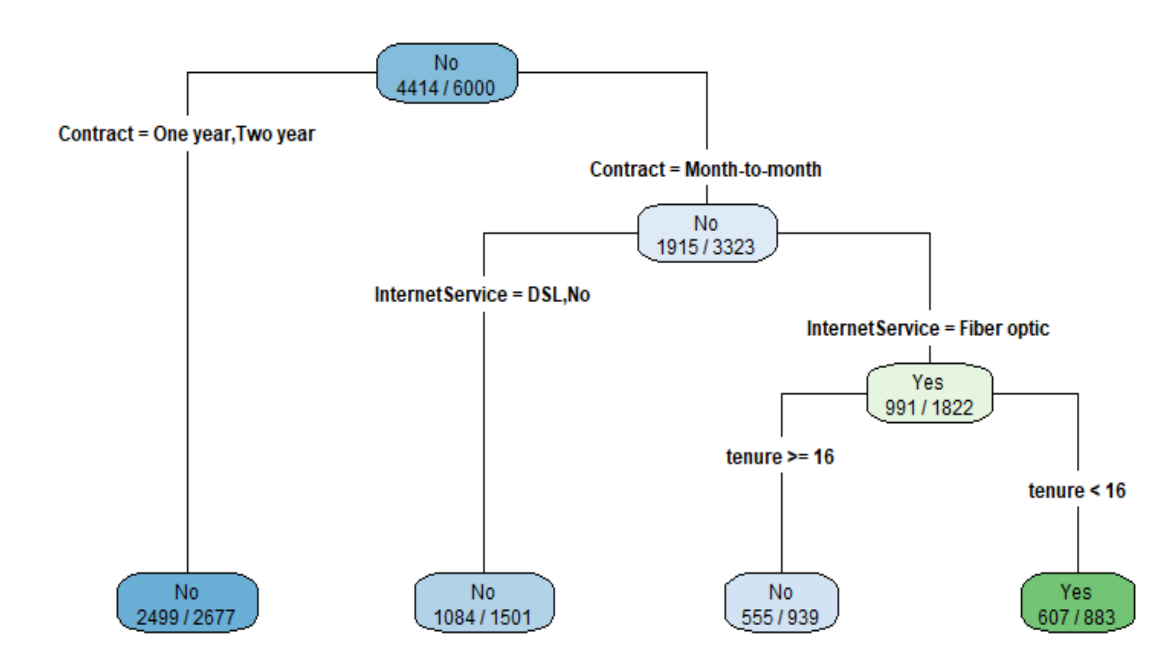








Plotting the model for a better view of the decision tree:



Interpretations:

From the above Decision Tree, Following Decision rules are made

* + **If Contract=Month-to Month, InternetService = FiberOptic, AND tenure<16 then Churn is Yes**
  + **If Contract=Month-to Month, InternetService = FiberOptic, AND tenure >=16 then Churn is No**
  + **If Contract=Month-to Month AND InternetService = DSL,No, then Churn is No**
  + **If Contract= One year, Two year then Churn is No**

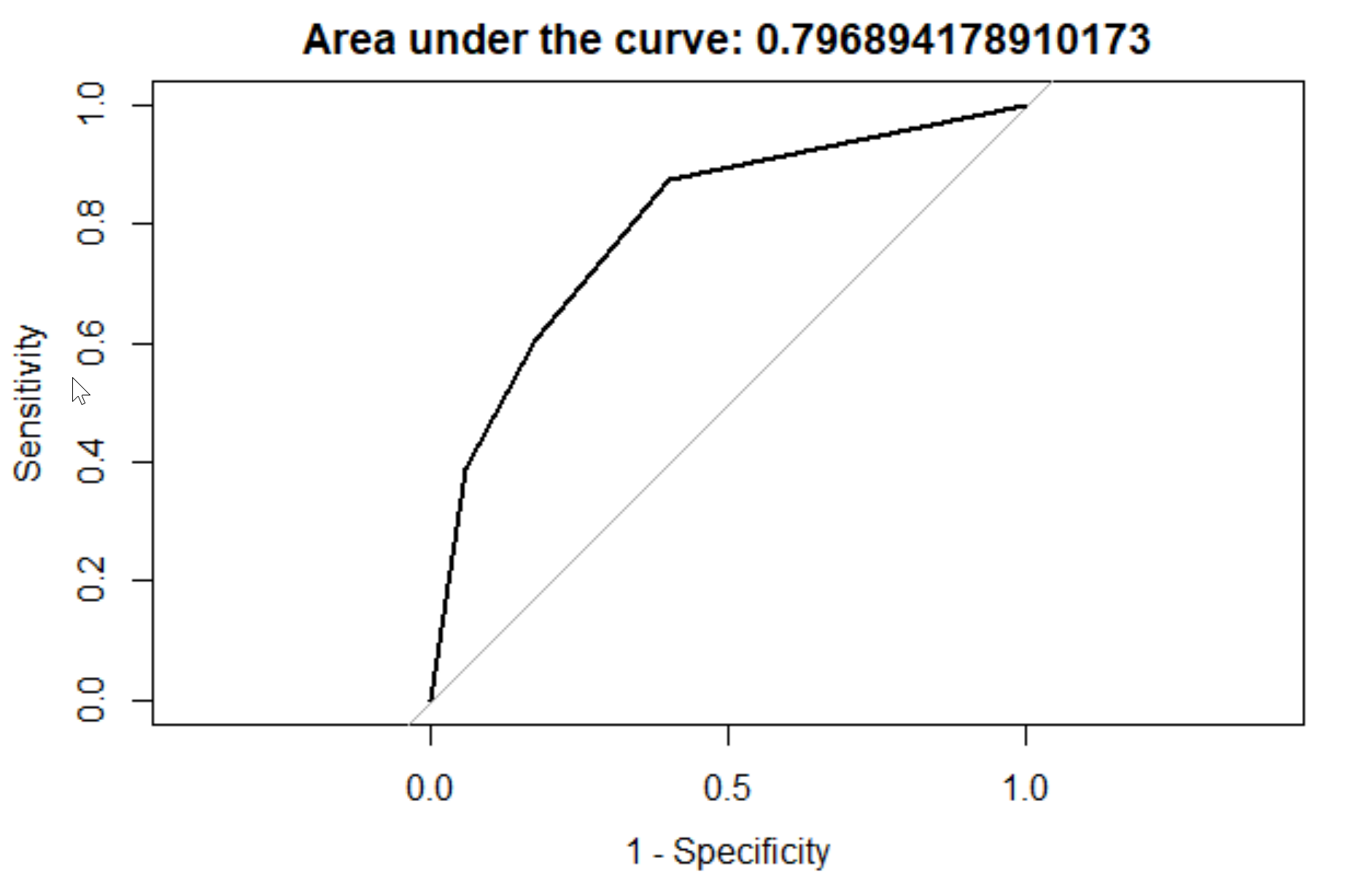
ROC curve for this model:

Testing the model performance on testData.

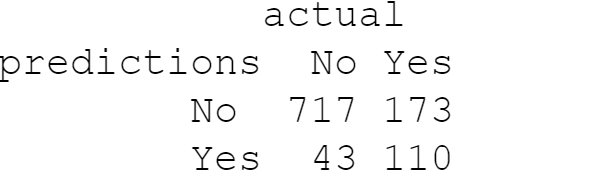
Model was run on testData, and the prediction is quite good.

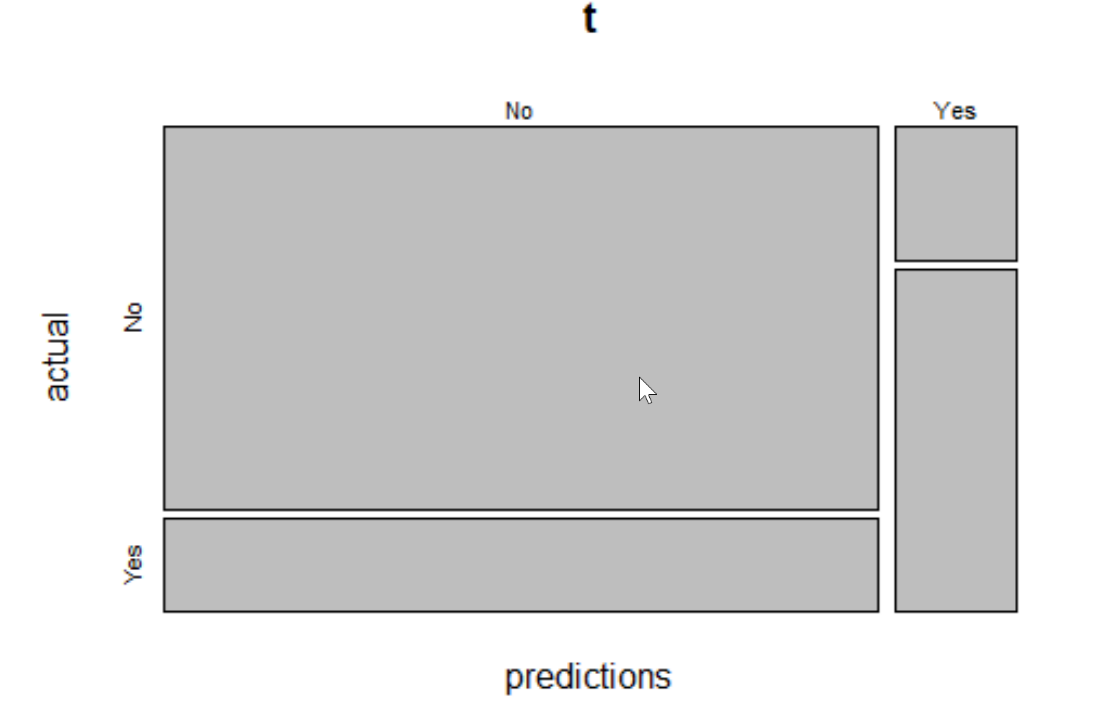
The Area Under the curve is 0.796 which is very much near to 1.

**Though the AUC is less than the AUC we got for logistic regression, It is because of less variables in the final decision model.**



Confusion matrix:





Accuracy\_Rate for Decision Tree: 79.2%

**2.Naïve Bayesian Analysis:**

From the logistic regression, we see that the model needs a proper selection of variables to explain the direction and intensity of the significance over the dependent variable “Churn”. It needs a considerable amount of time to select these variables. And, it handles missing values and high dimensional data well when compared to other classifiers. Naive Bayes model is a generative model which gives the joint distribution of the features and churn rate and then predicts the posterior probability given as P(features | Churn). And, it is computationally efficient and can handle high-dimensional data.

**Naive Bayesian as a classifier**: It estimate the likelihood of occurrences. It is a type of supervised-learning module that contains examples of the input-target mapping the model tries to learn, it uses the mathematics of Bayes' Theorem to make its predictions.

Naïve Bayes have three main equations:

1- Prior probability for class

2-Likelihood of feature(X) given class C

3.Posterior probability of X being C

That is, the Naïve Bayesian classifier predicts that tuple X belongs to the class C1 if and only if P(C1/X1)>P(C2/X2)

**Naive Bayesian Implementation Summary**: It is implemented to predict whether the customer will churn or not.

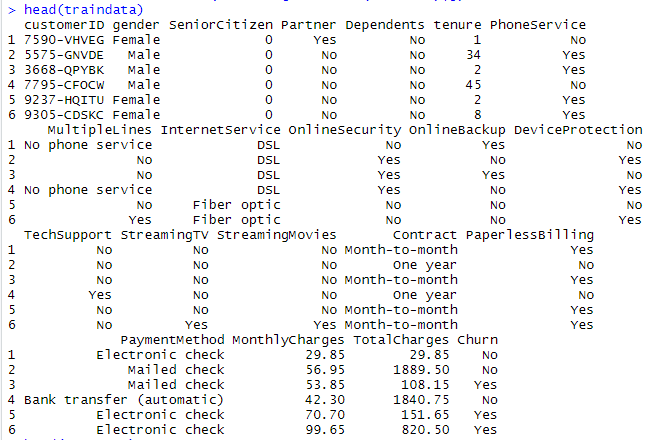
Input: Inserting all the variables into the model where it calculates the prior probabilities for the churn rate and the Likelihood of the features given the churn rate. Data is divided in Train data and Test data. Model will be implemented on the train data and is validated on the test data.

Output: Calculating the posterior probability and providing the churn prediction model, accuracy and its overall performance

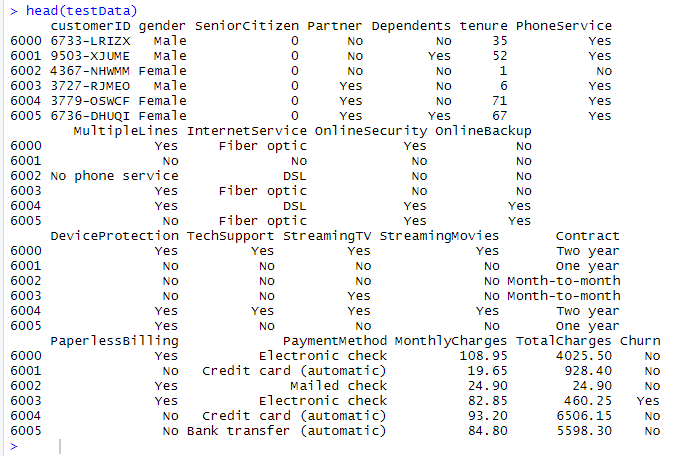
Implementation:

1. Two data frame objects called train data and test data are created for the naive Bayes classifier.

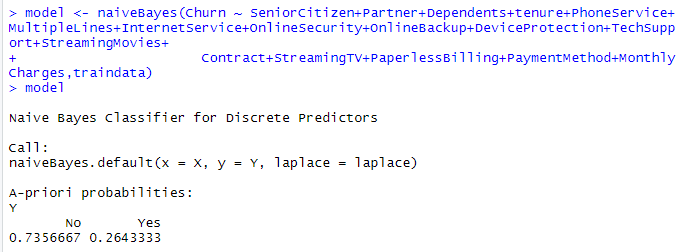




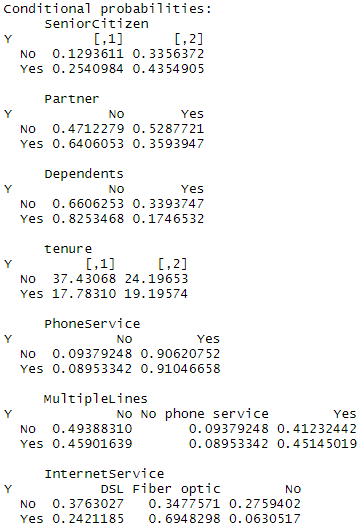


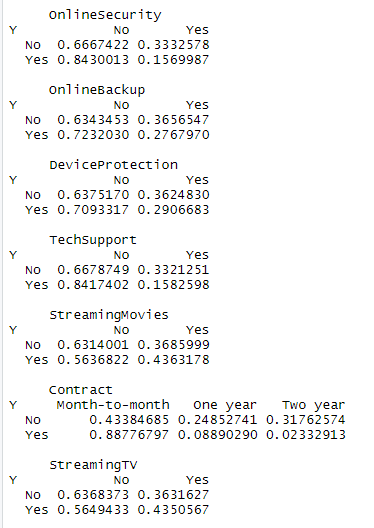


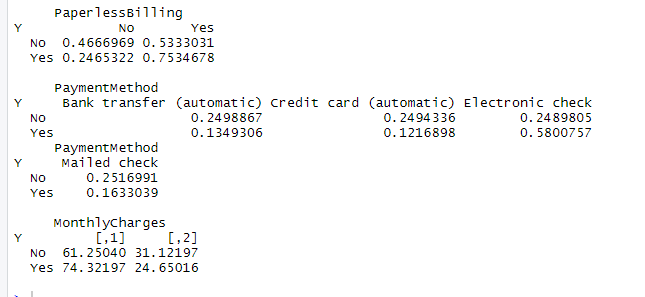
1. Implementing the model and calculating the prior probabilities of churn class and the conditional probabilities of all the variables given the churn class



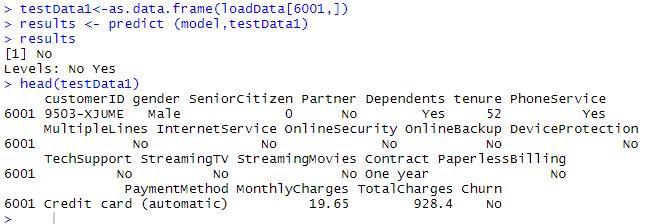
So, the probability of getting churned is 26.4% and not getting churned is 73.56% for the data. Now checking the individual conditional probabilities of the features. Count the number of "No" and "Yes" entries for each variable and normalize by the total number of “No" and "Yes" entries to get the conditional probabilities.







1. Next, predicting the outcome of Churn with the test data by taking a row from the test data and validating it with the actual value.

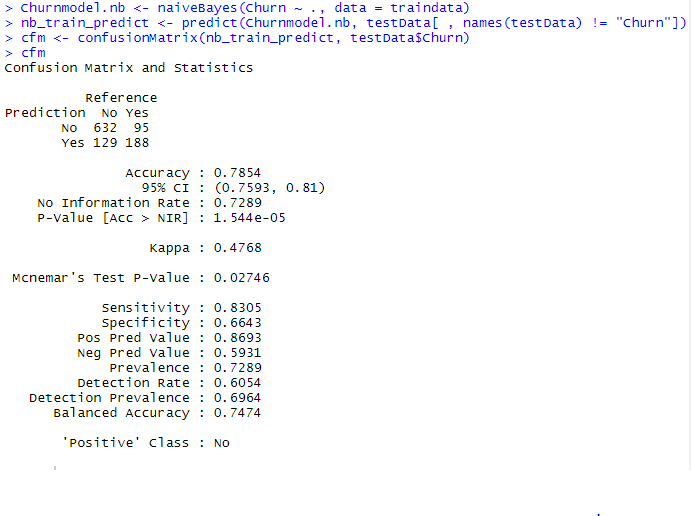


The result from the model is matching with the actual result given in the data predicting the customer getting churned as “No”

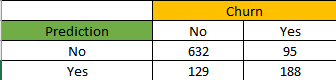
Validation:

Now validating the results of the entire test data with model using the confusion matrix.

A confusion matrix is a specific table layout that allows the visualization of the performance of the classifier. True positives (TP) are the number of positive instances the classifier correctly identified as positive. False positives (FP) are the number of instances in which the classifier identified as positive but, are negative. True negatives (TN) are the number of negative instances the classifier correctly identified as negative. False negatives (FN) are the number of instances classified as negative but, are positive. In a two-class classification, a preset threshold may be used to separate positives from negatives. TP and TN are the correct guesses. A good classifier should have large TP and TN and small (ideally zero) numbers for FP and FN.



Interpretations:



From the above confusion matrix,

True Positive (TP) instance that classified correctly as Non-churn = 632

True Negative (TN) instance that classified correctly as Churn = 188

False Positive (FP) instance that classified incorrectly as Non-churn = 129

False Negative (FN) instance that classified incorrectly as Churn = 95

The number of tuples labeled as positive P ' = TP + FP = 632+188= 820

The number of tuples labeled as negative N ' = TN + FN = 129+95 = 224

So, from prediction we have 632 correctly predicted as Non-churner customer from 727 which is (86.9%) and 188 correctly predicted as Churner customer from 317 which is (59.3%).

Accuracy of the classifier or the overall success rate is 78.4% i.e. how well a given predictors can guess the value of predicted attribute for a new data.

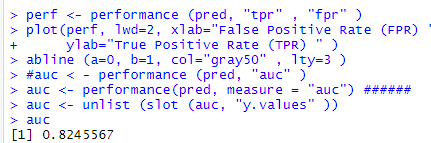
Evaluation of the classifier:

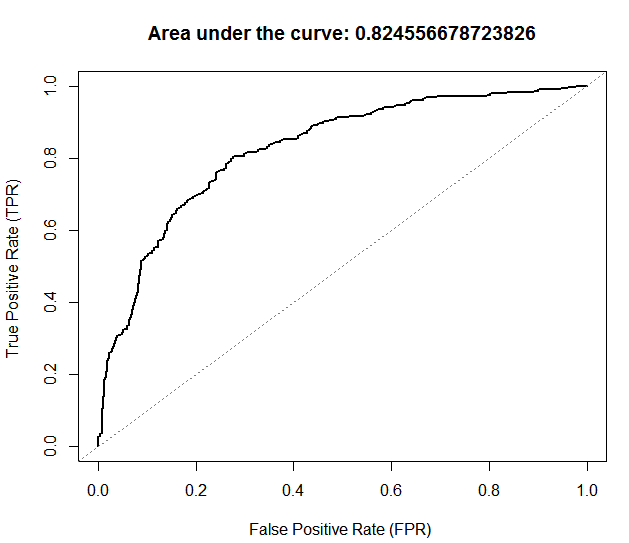
A ROC curve evaluates the performance of a classifier based on the TP and FP, regardless of other factors such as class distribution and error costs. The vertical axis is the True Positive Rate (TPR), and the horizontal axis is the False Positive Rate (FPR).

Any classifier can achieve the bottom left of the graph where TPR = FPR = 0 by classifying everything as negative. Similarly, any classifier can achieve the top right of the graph where TPR = FPR = 1 by classifying everything as positive. If a classifier performs "at chance" by random guessing the results, it can achieve any point on the diagonal line TPR=FPR by choosing an appropriate threshold of positive/negative. An ideal classifier should perfectly separate positives from negatives and thus achieve the top-left corner (TPR = 1, FPR = 0). The ROC curve of such classifiers goes straight up from TPR = FPR = 0 to the top-left corner and moves straight right to the top-right corner. It can be difficult to achieve the top-left corner. But a better classifier should be closer to the top left, separating it from other classifiers that are closer to the diagonal line.

Related to the ROC curve is the area under the curve (AUC). The AUC is calculated by measuring the area under the ROC curve. Higher AUC scores mean the classifier performs better. The score can range from 0.5 (for the diagonal line TPR=FPR) to 1.0 (with ROC passing through the top-left corner).

The following R code shows that the corresponding AUC score of the ROC curve is about 0.8246.





Interpretation:

The ROC curve shows that the area under curve is about 82.46% which means that the classifier has performed good as it relatively close to 100%

Limitations:

Although the performance is good, the Navy Bayesian classifier has few limitations when compared to other classifiers Logistic Regression and Decision tree.

1. Naïve Bayes assumes all the features to be conditionally independent. So, if some of the features are in fact dependent on each other (in case of a large feature space), the prediction might be poor. Here there exists some correlation between the variables {Internet services} and {OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies}. So, Logistic regression splits feature space linearly, and typically works reasonably well even when some of the variables are correlated.
2. Naïve Bayes cannot handle numerical data well when compared to Decision Tree classifier.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |

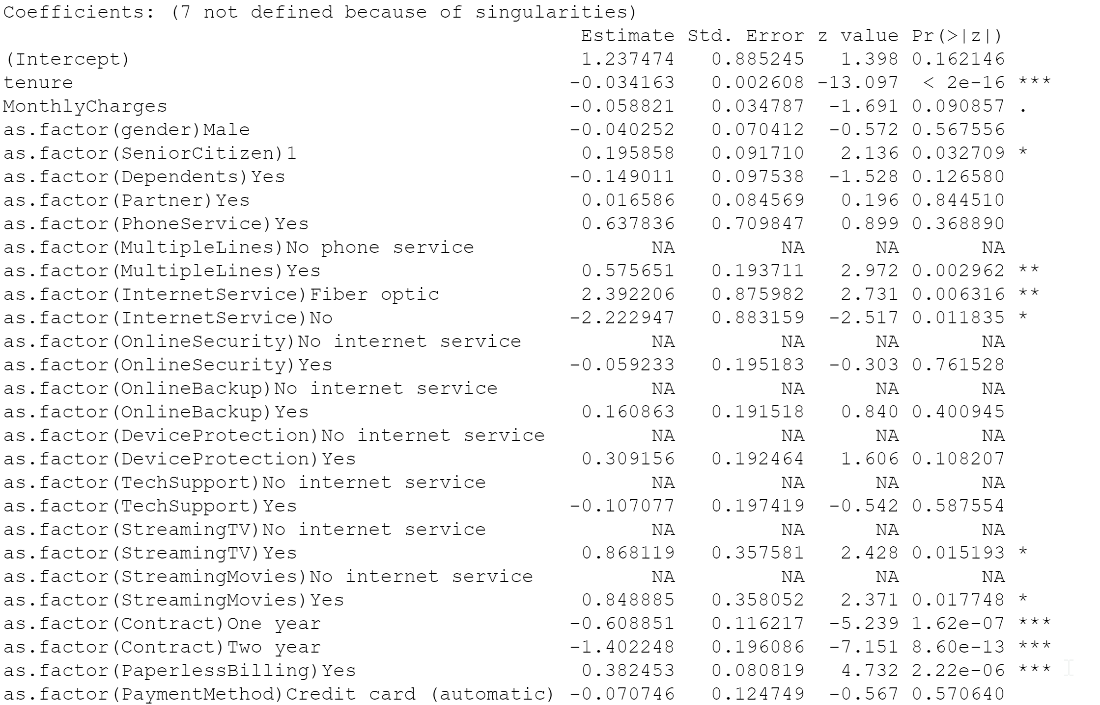
Based on this limitation, we need to consider the other classifiers as well to get the better prediction model.

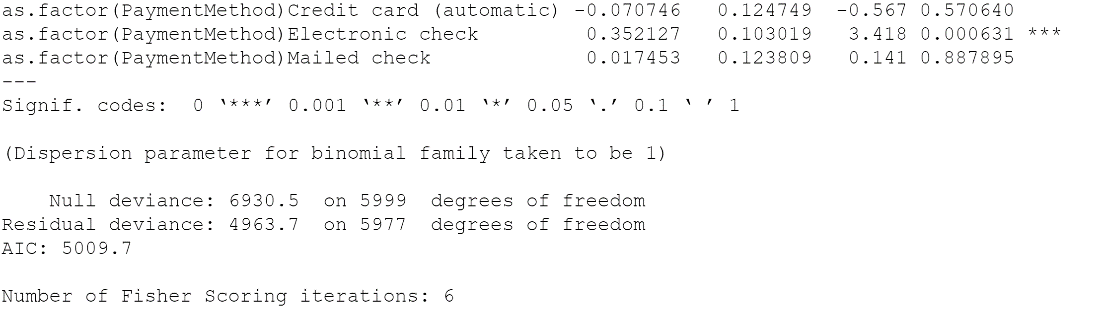
**3 .Logistic Regression:**

* Based on this existing data, **logistic regression** can be used to find the effect of each variable on the final class churn.
* **Expected outcome:** After a rigorous regression analysis with a various set of variables, we expect to see a final set of variables that are important to consider in the model along with a coefficient which provides the information on how they are impacting the probability of a customer to churn.

Selection of Predictor Variables:

* We are using Log- likelihood ratio test in a forward/backward stepwise method to add/ remove the variables to/from the Proposed Logistic Regression Model.
* glm() method is used with family =”binomial” for performing logistic regression.
* Churn is our dependent variable.
* Initially all the other variables are considered as the predictor variables.
* After running the glm() method with this setting, obtained results gives the coefficients for each variable. These coefficients are examined further for their significance value (p-value) which denote the p-value for the hypothesis test to determine that the model parameter is significantly different than zero.
* The results of the model with all the predictor variables are as follows:





* 7Coefficients are not defined because of singularities which represents as correlation of those variables with other variables.
* Logistic Regression doesn’t handle the correlation variables well. Thus, going ahead for some data preparation for logistic regression.

Data Preparation:

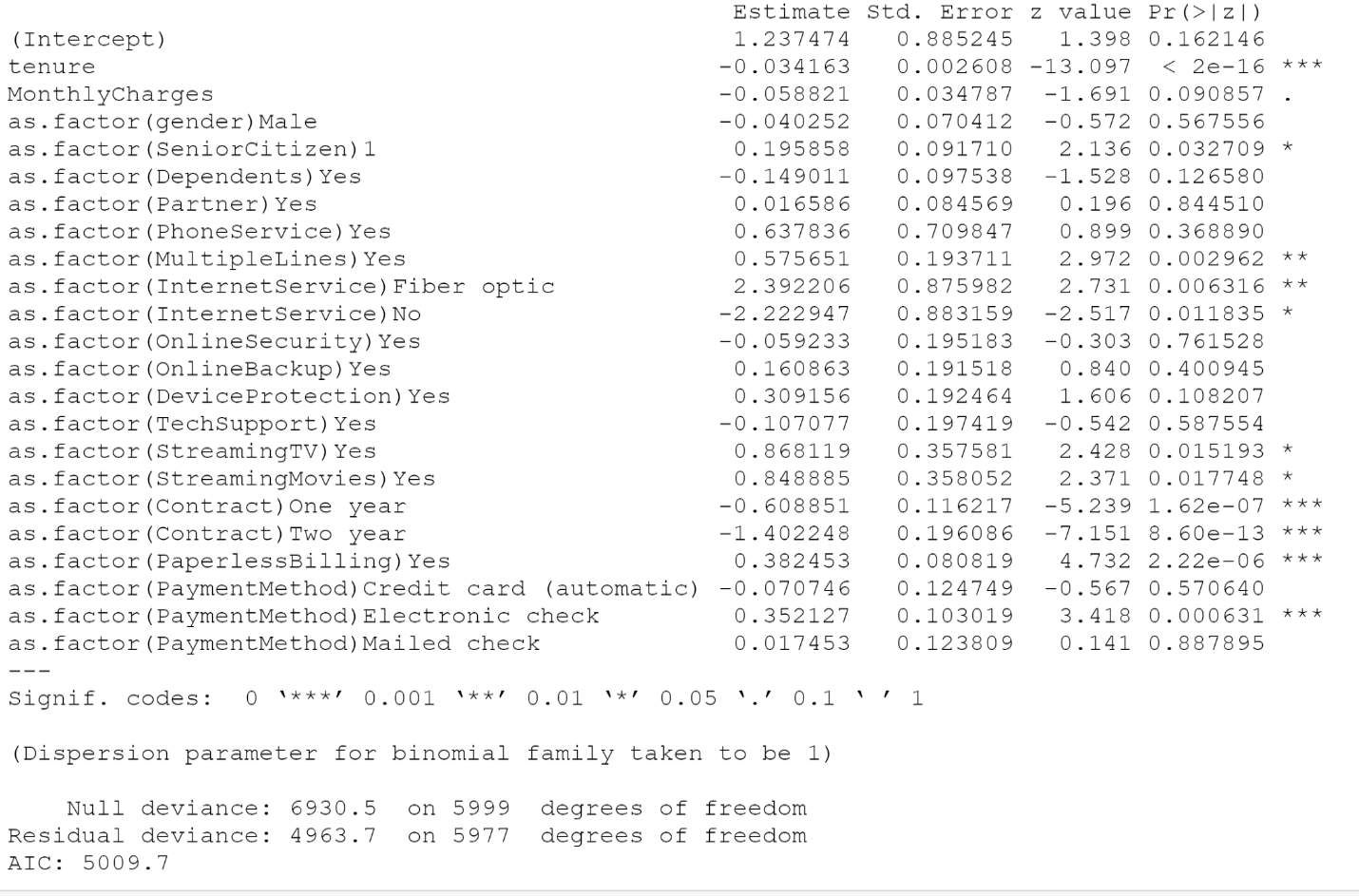
* Cleaned data by assigning “No” to the values “No Internet Service” and “No Phone service” thus removing the correlation. This change is not affecting any of the information in the data.

**Model:**

* Logistic Regression model with all the variables except Total Charges. Total Charges variable is excluded because

a) It has some missing values, and logistic regression cannot handle missing value well.

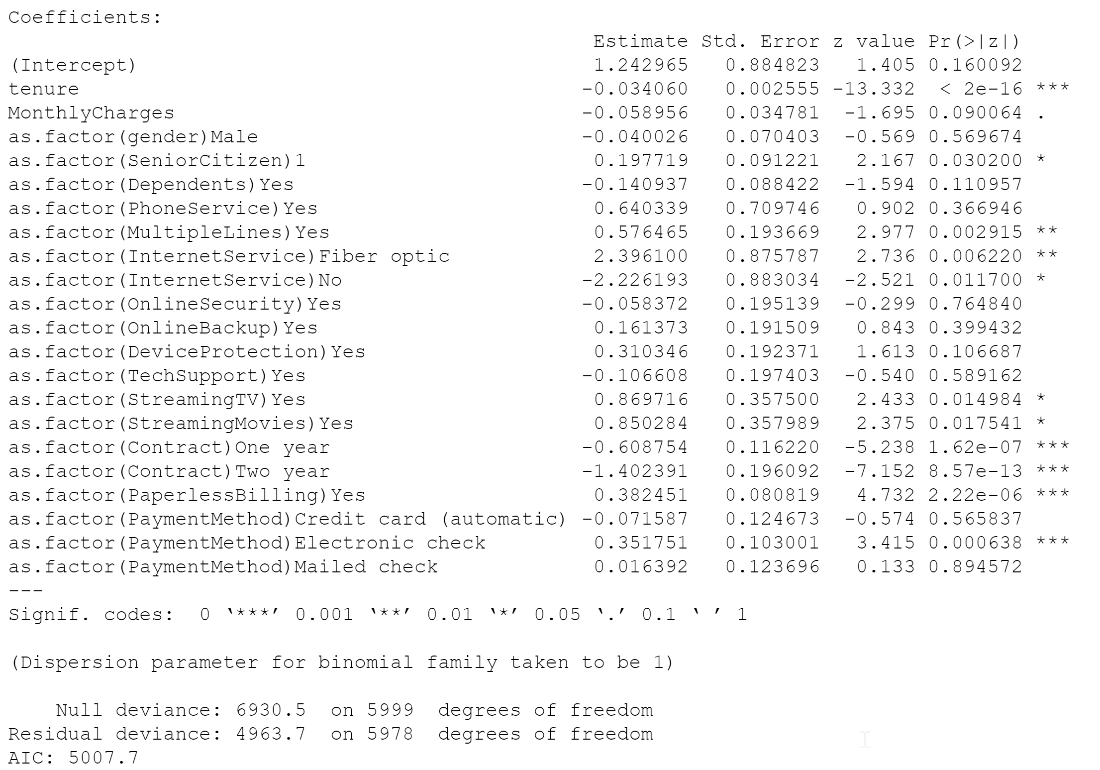
b) it is proportional to MonthlyCharges, thus considering MonthlyCharges would give same information to the model as TotalCharges.



**Rerunning the model by removing the variables whose coefficients have very high p-value.**

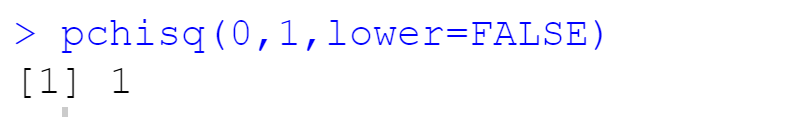
1. First remove the variable highest p-value first- i.e Partner.

The results are:



**Pseudo R-squared value:**

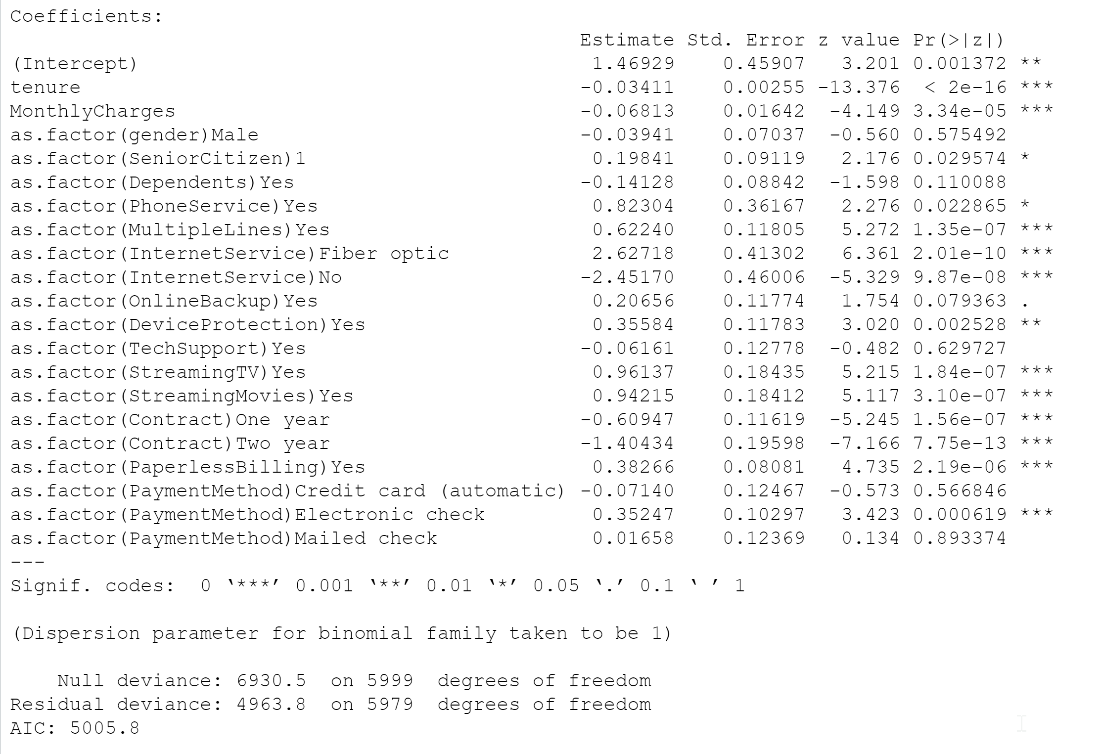
* 4963.7-4963.7, with 5978-5977 degrees of freedom.



* P-value is higher, which indicates the model with parameter is not any significant than model without parameter.
* Hence, we can safely remove Partner.

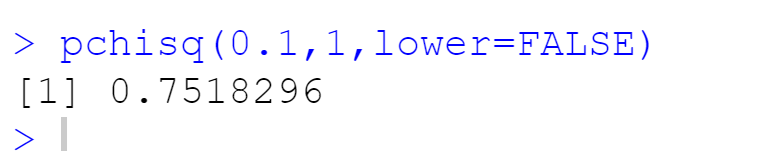
2. 

* Remove OnlineSecurity:



**Pseudo R-squared Value:**

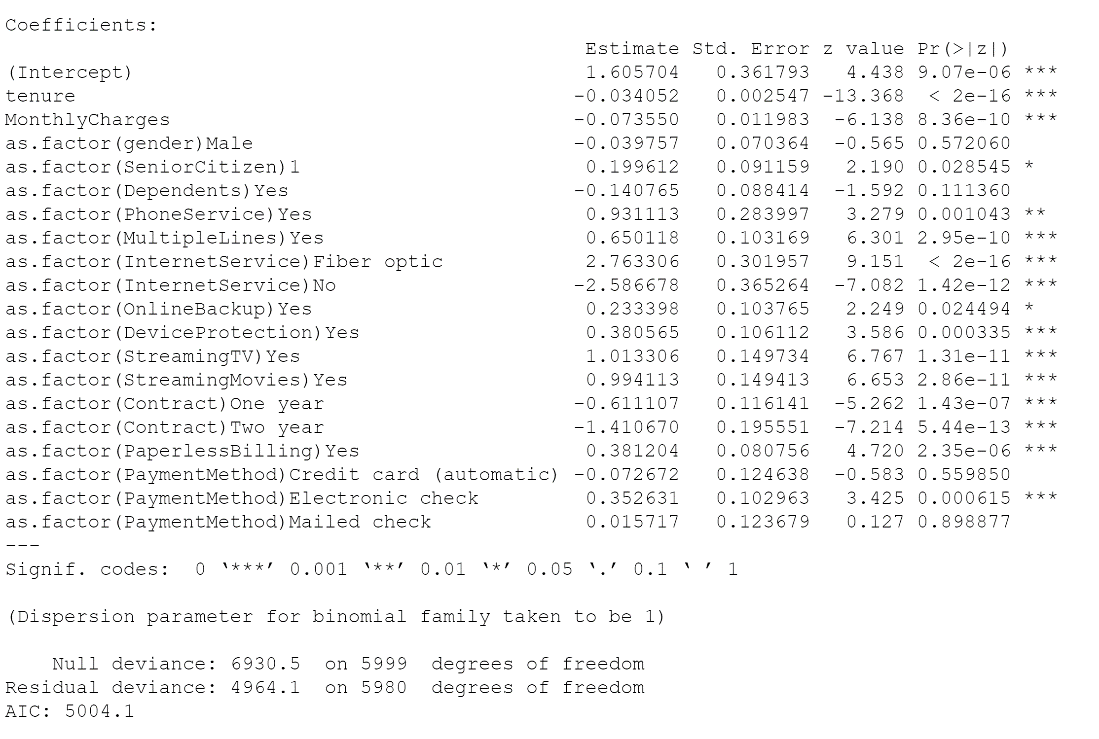
* 4963.8-4963.7 for 5979-5978 degrees of freedom



* Higher p-value for the deviance. Hence Safely remove OnlineSecurity

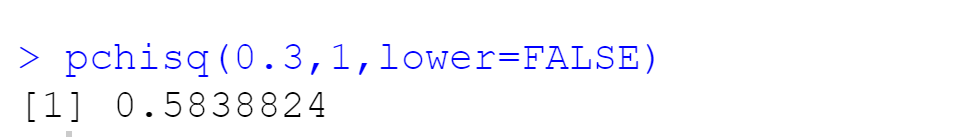
3. 

* Remove TechSupport:



**Pseudo-R-square:**

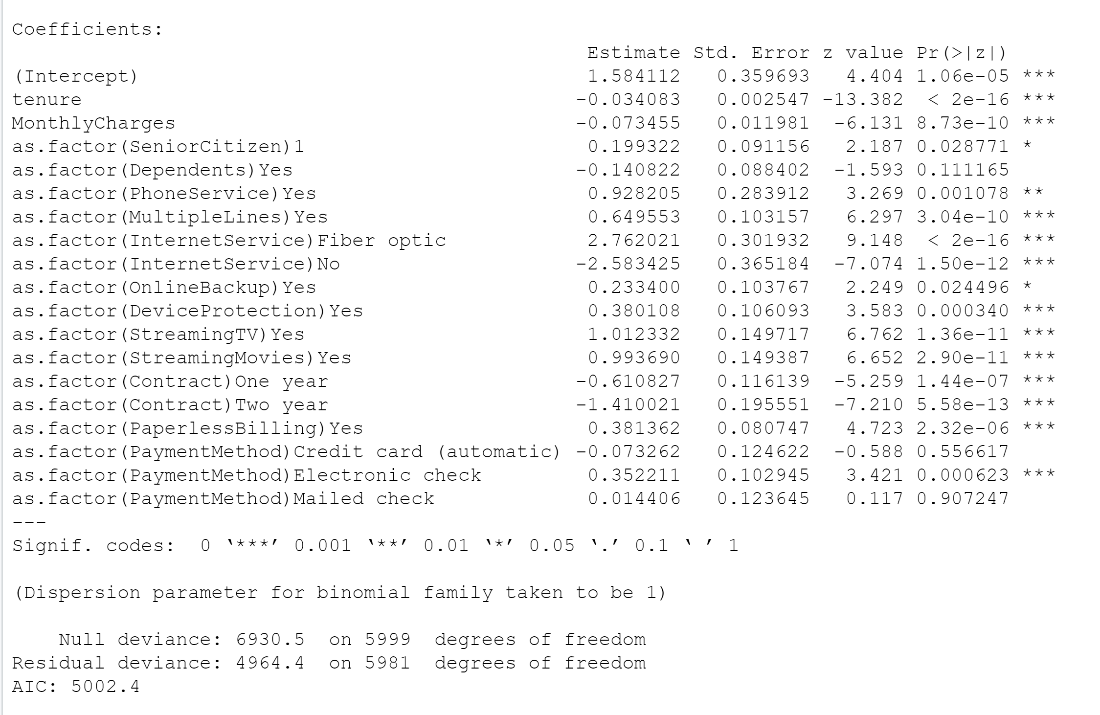
* 4964.1-4963.8 for df= 5980-5979



* P-value is large. Hence can safely proceed with the model without parameter.

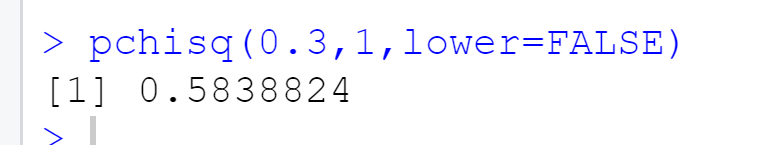
4 

* Remove Gender:



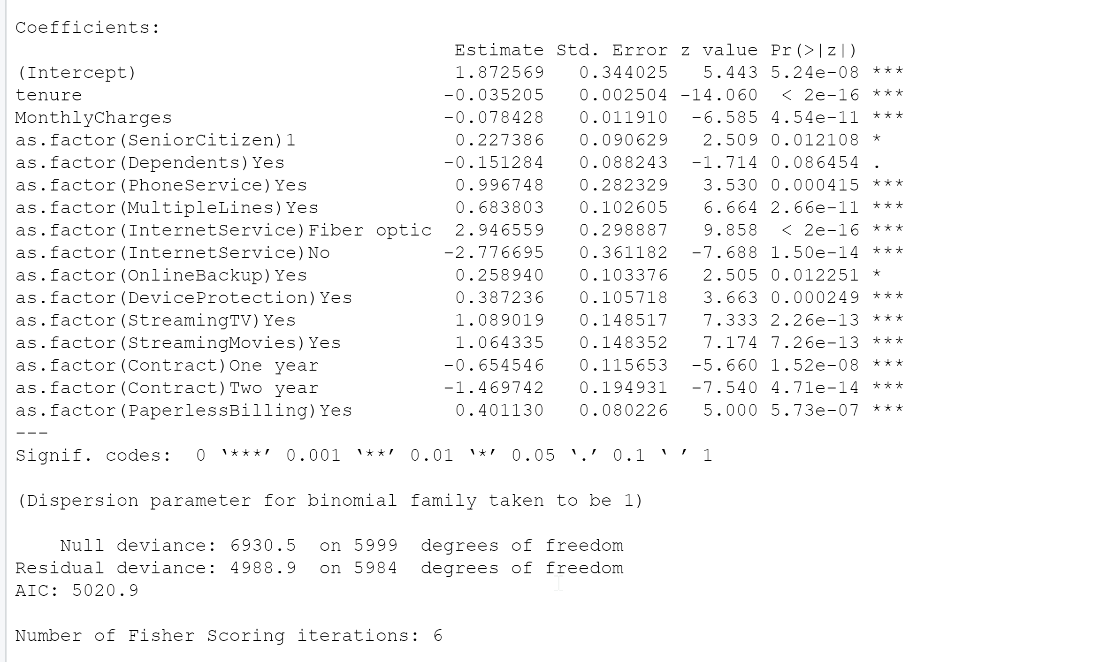
**Pseudo R-squared value:**

* 4964.4-4964.1 with df = 5981-5980

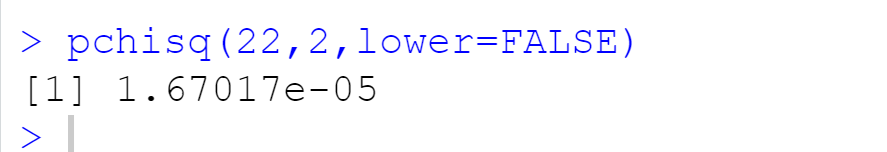


* Removal of Gender doesn’t affect the model

5. Remove payment method:

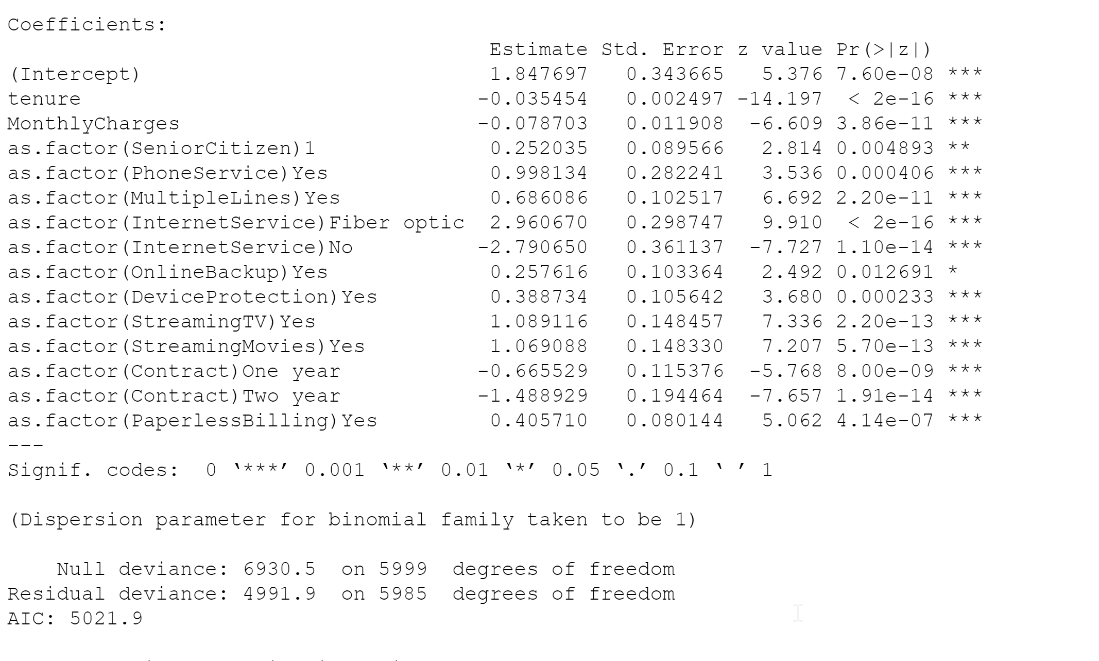


* Pseudo R-Square : 4988.9-4966.9 on 5984-5982



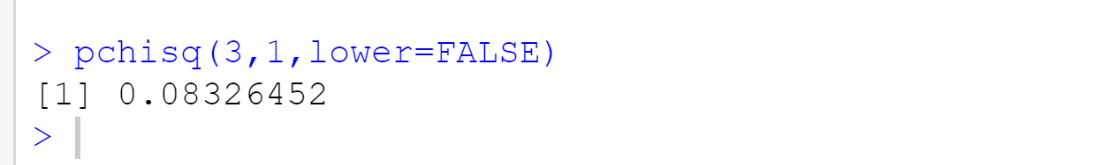
* As p-value is small, removing paymentmethod variable does not affect the model.

1. Remove dependents:



**Pseudo R-square value**

* 4991.9-4988.9 on 5985-5984 degrees of freedom

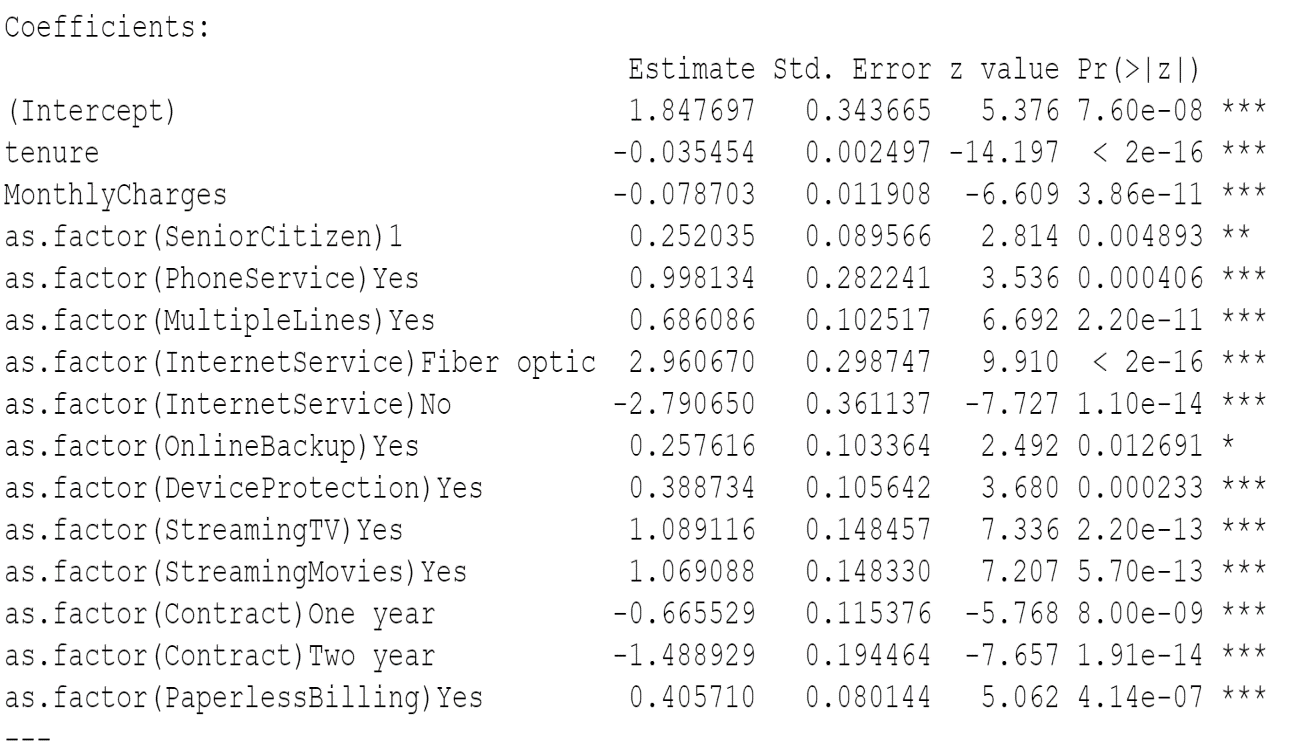


* As p-value is small, removing the Dependents parameter doesn’t affect the model.

**The final model:**

The final proposed logistic regression model contains following predictor variables:

Tenure, Monthly Charges, Senior Citizen, Phone Service, MultipleLines, Internet Service, OnlineBackup, DeviceProtection, StreamingTV, StreamingMovies, Contract,PaperlessBilling



**Interpretations:**

Interpretations from the above resulted final logistic regression model are as follows:

[Note: AS most of the predictor variables are categorical, an absolute equation to obtain the probability of the Churn = “YES” cannot be achieved]

1. For a unit(month) increase in tenure, the log odds of the customer churning (leaving the company) decreases by 0.035.
2. For a unit (Dollar) increase in MonthlyCharges, the log odds of the customer churning (leaving the company) decreases by 0.0789.
3. The log odds of the customer churning are 0.252 more when the customer is a senior citizen than he is not.
4. The log odds of the customer churning are 0.99 more when the customer is using the phone service than he is not.
5. The log odds of the customer churning are 0.687 more when the customer has multiples lines of phone service than he is not given that he is having phone service.
6. a. The log odds of customer churning who are using the internet service with FiberOptics is 2.96 more when compared to the customer who are using the internet service with DSL option.

b. The log odds of customer churning who are not using any InternetService is 2.79 less when compared to the customer who are using the internet service with DSL option.

1. The log odds of the customer churning are 0.257 more when he is using OnlineBackup service than he is not.
2. The log odds of the customer churning are 0.388 more when he is using Device Protection service than he is not.
3. The log odds of the customer churning are 1.089 more when he is using StreamingTV service than he is not.
4. The log odds of the customer churning are 1.069 more when he is using StreamingMovies service than he is not.
5. a. The log odds of customers churning who are on one-year contract is 0.665 less than those of the customers who are on month-month contract.

b. The log odds of customers churning who are on Two-year contract is 1.488 less than those of the customers who are on month-month contract.

1. The log odds of customers who have opted for paperless billing for churning is 0.405 more than those of the customer who have not opted for paperless billing.

**Inferences & Recommendations:**

1. As the tenure is increasing, customers are getting more comfortable with service and more likely to stay with the company. There are high chances of the customers leaving the company at their initial days of enrolling for the service from the company. Company can concentrate on those people who are new.
2. Two things can happen in the case of senior citizens which attributed for more probability of churning.
   1. Involuntary churning (death or moving to other place because of retirement). Involuntary churning is not of importance for the marketing team and thus can be ignored.
   2. Senior citizens are not comfortable with the services. Hence, customization for senior citizens is a research & Development item to work on from the company side.
3. Customers with Internet Service with FiberOptics tend to churn more compared to DSL or not having the internet service. So, customers can be shifted to DSL or give better Performance with FiberOptics.
4. It’s always easy for customers who are on short term contract to leave the company.

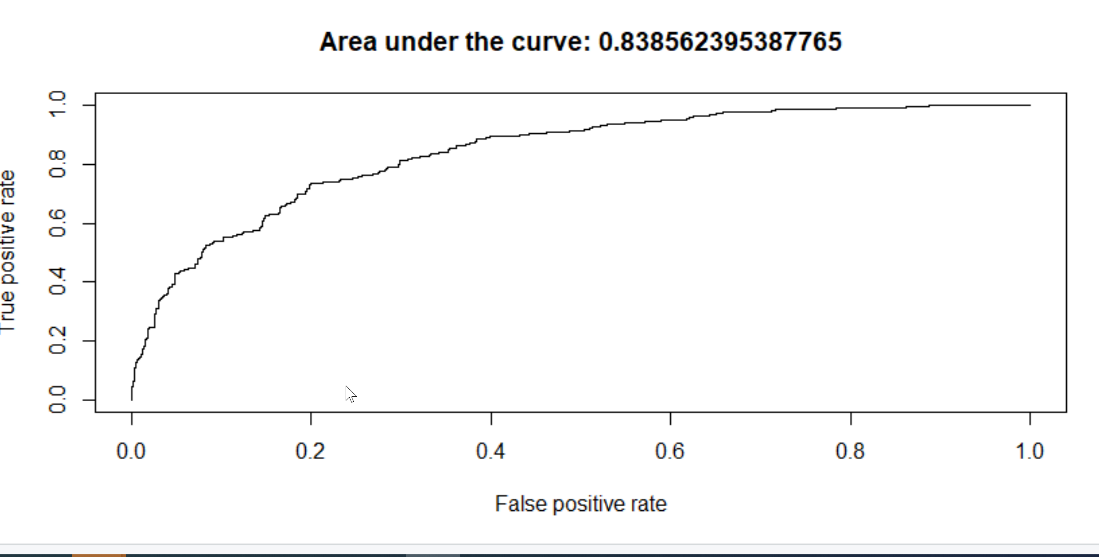
**Logistic Regression as classifier:**

Testing how the model is predicting the churn rate on test data.

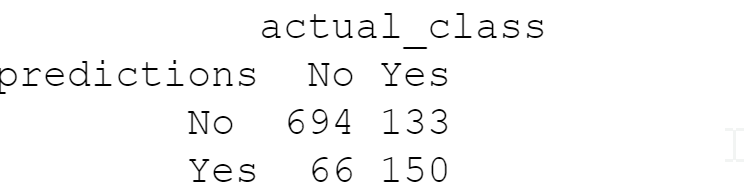
Model was run on testData, and the prediction is quite good.

The Area Under the curve is 0.838 which is very much near to 1.

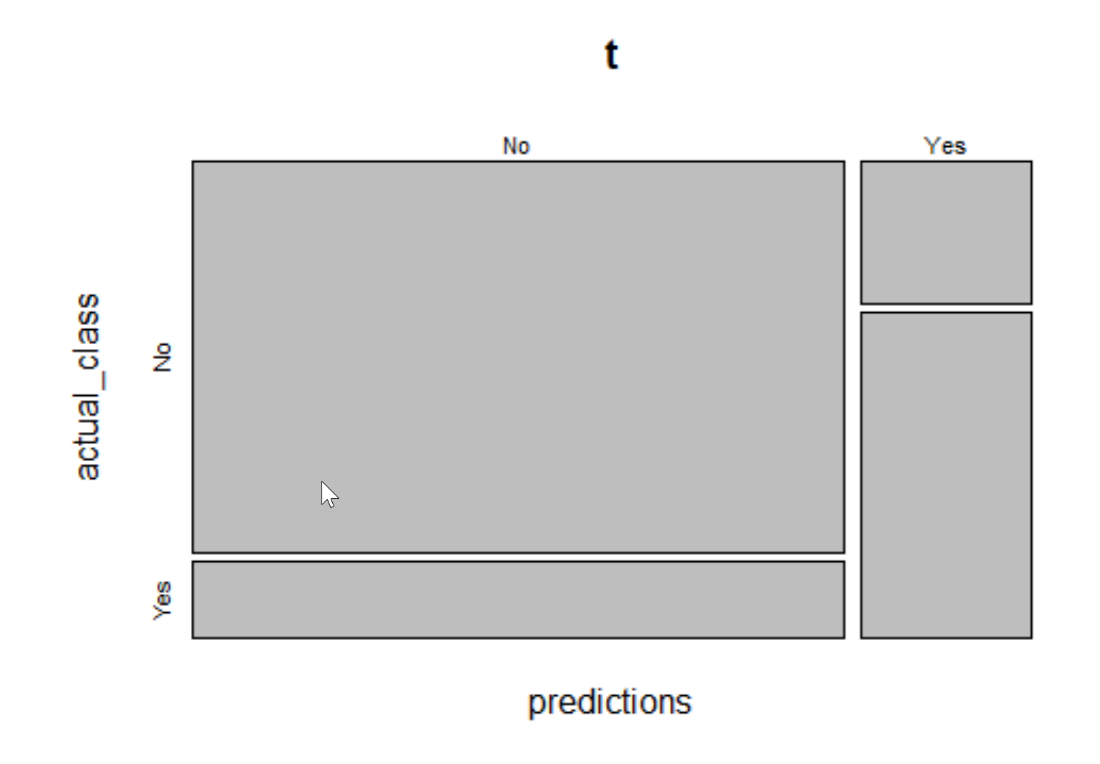
**Hence the model is performing better as a classifier which classifies whether the given customer will churn or not.**



Confusion Matrix :



Accuracy rate**: 80.9%**

Disadvantages:

* The model has only fewer parameters with a more loss of information. Though the model is performing better, loss of information might have a little impact on the predicting.
* We are confident on the variables removed are not adding any information to the model but among the variables included in the model, there is a chance of few variables being interrelated or correlated for which we don’t know have information.

Conclusion**:** With few disadvantages the current Logistic Regression model is performing good.

**Conclusion:**

Customer churn is a big issue in telecom companies especially for prepaid subscribers because it happens easily under light of strong competition in this business area, so these companies need to build a churn prediction model to identify churner and non-churner customer and avoid this churn.

Based on the AUC values from the model comparison graphs, we see that Logistic Regression is giving the best prediction model with 83.86%

From this model, we got two outputs:

1. Which Customers are likely to Churn
2. Which features have the most impact on customer leaving

Based on the model, below is table showing the best predictors of the Churn rate.

|  |  |
| --- | --- |
| Feature | Impact |
| Tenure | Positive |
| Monthly Charges | Positive |
| Senior Citizen | Negative |
| Phone Service | Negative |
| Multiple Lines | Negative |
| Internet service- Fiber Optics | Negative |

The first two values, Tenure and Monthly charges has a positive impact on the model meaning they are reducing the churn rate. The remaining values, Senior Citizen, Phone Service, Multiple Lines and Internet -Fiber Optics have a negative impact on the model meaning they are increasing the churn rate. So, it is highly recommended to share these insights with customer success team and adapt their focus to increase the business and to reduce the churn rate.

Recommendations:

Based on the features, below are the few recommendations made to reduce the churn rate:

1. Ask for feedback, communicate latest developments that might be from interest or educate them on new product features. Approach customers likely to churn, but make sure that you come up with relevant things that may fit their individual needs. It will create a feeling of being understood and bind them to you and your business.
2. Adjust the Relevant Criteria: When selling a product, the basic goal isn’t to sell the product, but rather, what the product can do for the customer. Senior citizens don’t want the same thing that a teenager wants, but at the same time the company needs to produce the product/service after learning on how they can be useful to the customer.
3. Senior Citizens prefer the quality customer service rather than automation and self-service speed up the process. So, the company needs to provide the best customer service to attract the senior citizens
4. Customer Feedback Loop: It's hard to improve your business if you don't know how your customers feel about your product/service. Customer feedback loop provides a system for collecting, analyzing, and distributing customer reviews and surveys.
5. Feedback Loop: It's hard to improve your business if you don't know how your customers feel about it. This is where a customer feedback loop provides a system for collecting, analyzing, and distributing customer reviews and surveys.
6. Onboarding: Onboarding is an effective customer retention tool because it prevents churn with new customers. It teaches new customers how to use the product or service. Rather than learning by themselves, customers are taught by a company representative who personalizes the training according to their needs. This way, customers not only save time but also bind to the company for a longer time.