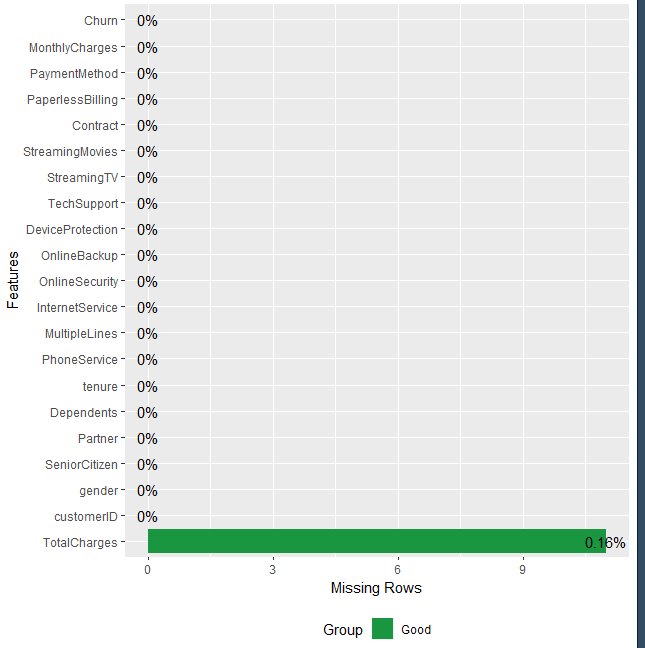
**BUS/CSC 328 Data Analytics Exam #2 – Take Home Edition**

**The Telco\_Churn dataset documents several attributes of customers of a telecommunications firms and whether they churned or not (stopped being customers). Your mission is to discover the reasons that may indicate a customer’s departure, and to make specific recommendations to senior management to reduce customer churn (departure). All output (i.e. console prints) must be placed in your Exam Report; no R script files will be accepted for partial grading. Document everything!**

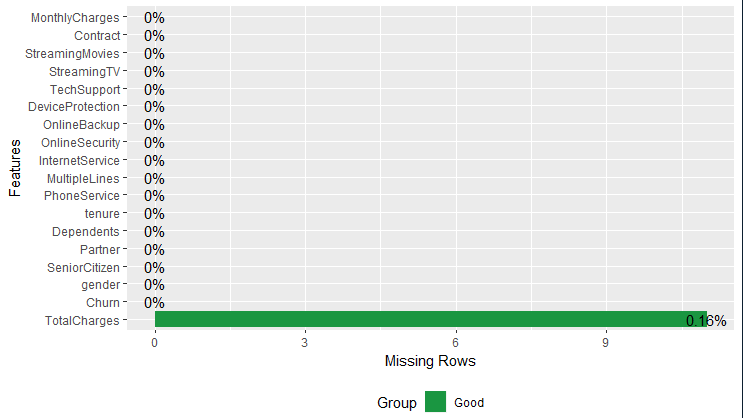
**Step 1: Fully inspect the Telco\_Churn dataset and describe in detail the important aspects.**

Ans: The Telco\_Churn is taking into consideration multiple predictors and then using those values to see if a customer would leave the customer base. This is an important aspect of business management in many industries because it leads to significant changes in the bottom line and essentially determine the future of the company.

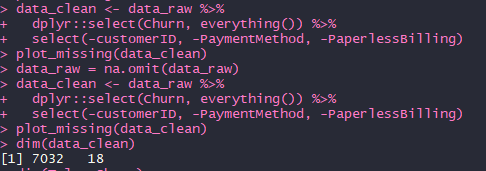
My first step was to import the dataset as it is and then visually inspect the dataset. After some consideration I decided to remove payment method, CustomerID and PaperlessBilling from the model because I do not think they provide much help in identifying customers that are not going to avail the service anymore.



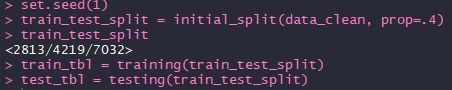
**Step 2: Document your plan for dealing with missing values.**



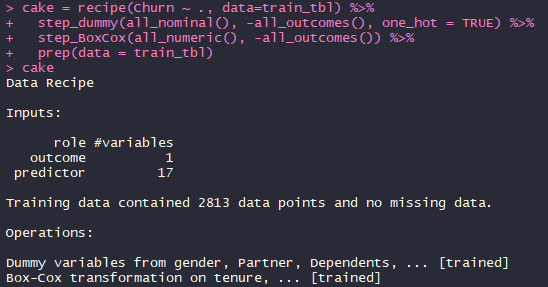
Only 0.16% of the values or simply 11 values are missing which is a small amount of data points taking into account the sheer size of the dataset. I decided to omit these values using na.omit () because I do not consider them significant enough to affect our model a whole lot.



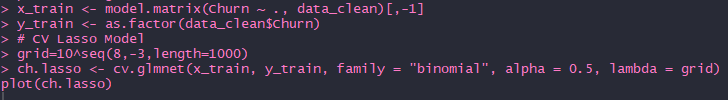
**Step 3: Create a training/test split with 40% of the data reserved for training.**

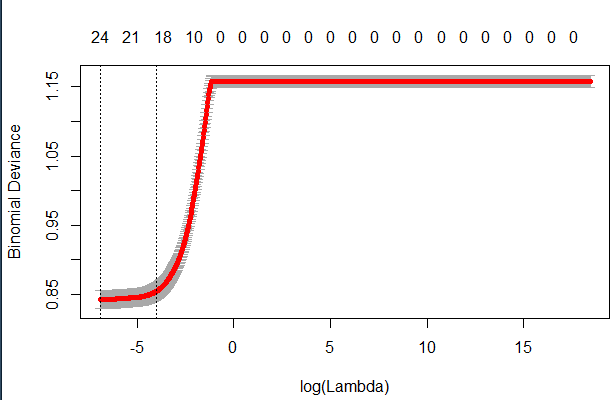


**Step 4: Create and implement a data preprocessing recipe that addresses all relevant issues.**



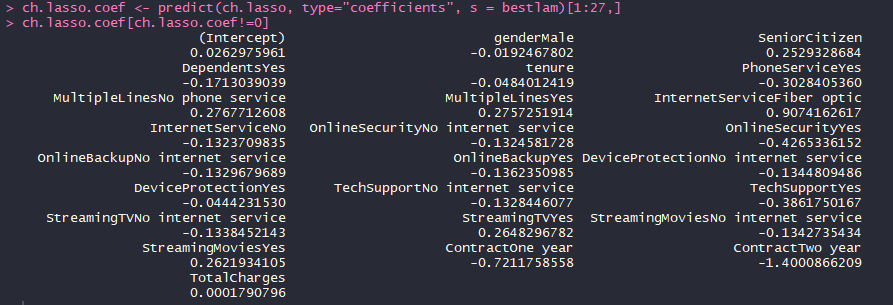
**Step 5: Prepare the training data for a cross-validated Elasticnet model and set up said model to search lambda values from 10^8 to 10^-3.**





**Step 6: Run the cross-validated Elasticnet model, plot the output, print the best value for lambda, and after predicting coefficients, print out the list of non-zero coefficients.**



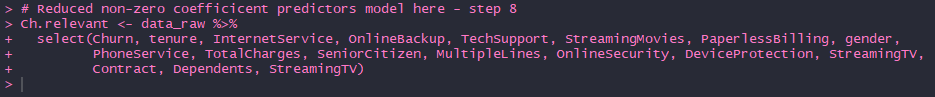


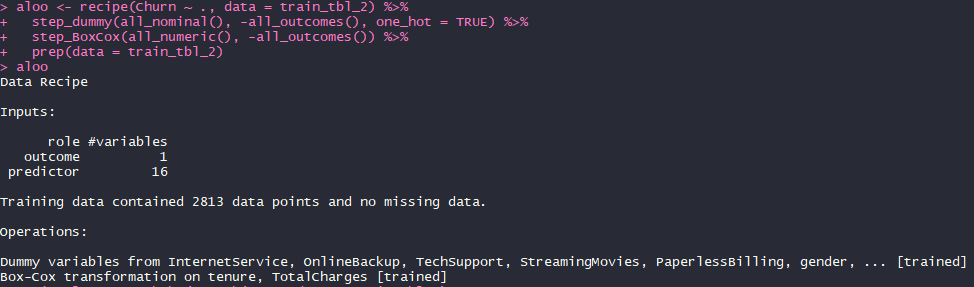
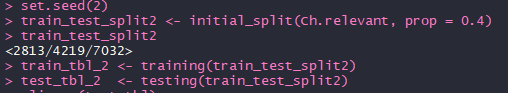
**Step 7: Using the non-zero coefficients, identify the two strongest attributes leading to churn and the two strongest attributes preventing churn.**

Ans 7: The two strongest attributes for churning are: Internet Service and Multiple Lines.

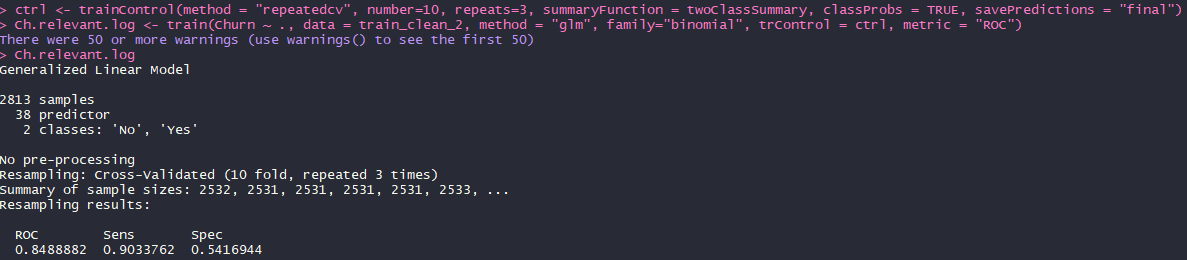
Two strongest attributes that prevents churning are: Contract length and Online Security.

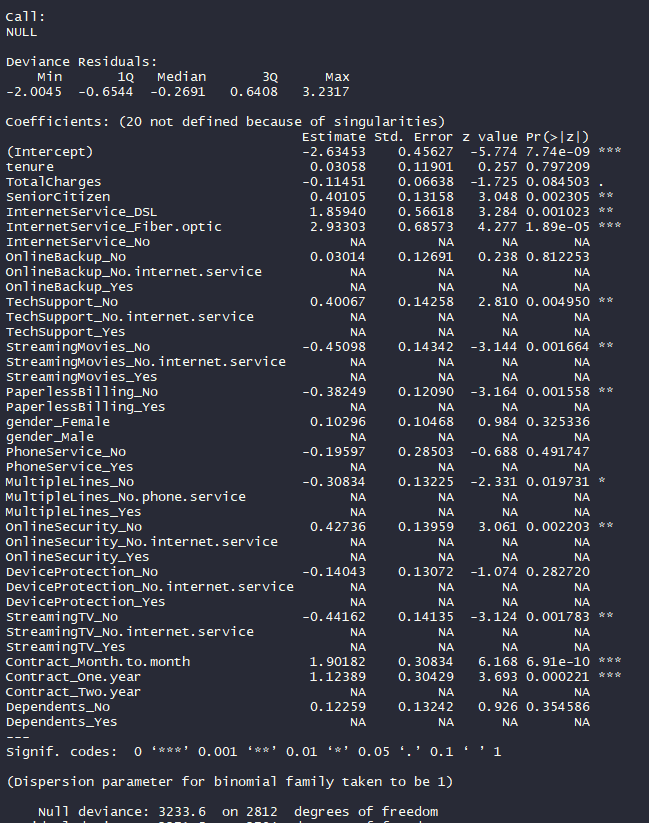
**Step 8: Build a new data set using Churn followed by the non-zero coefficients, split it into training and test sets with a 40/60 ratio as before, and rerun the data preprocessing recipe on the reduced dataset.**

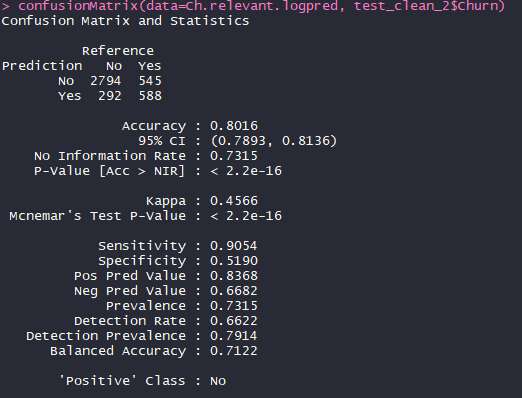




**Step 9: Using the reduced data set, run a cross-validated logistic model with the caret package and fully document the output with a standard logistic model print out (i.e., listing all coefficients) and two-class summary (confusion matrix and all related statistics).**

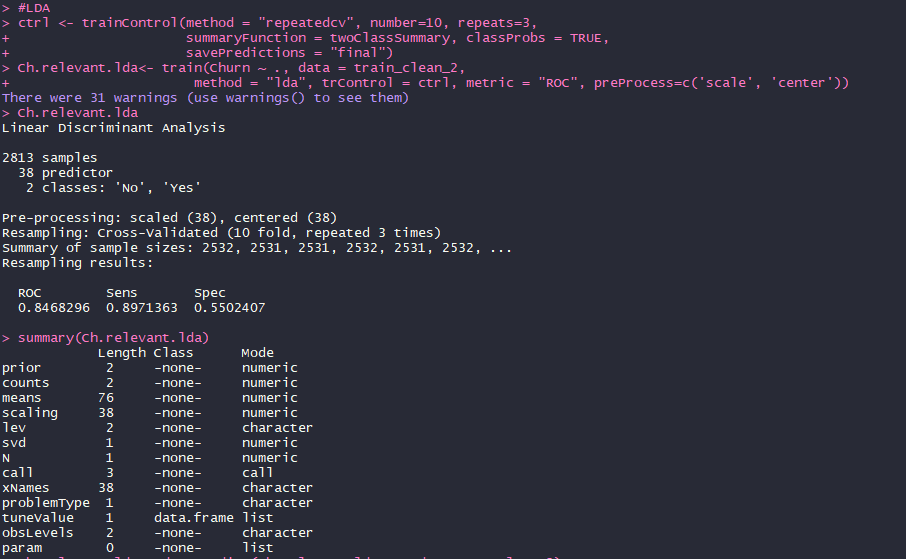


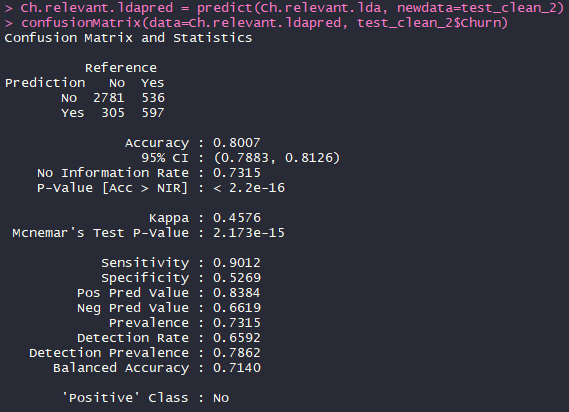




Accuracy of 80.16% and Kappa of 45.66%. Which means it is more likely that the result might be correct based on a pure chance but 45% of the time it will not be based on random chance and the accuracy will be up to 80%.

**Step 10: Using the reduced data set, run a cross-validated LDA model with the caret package and fully document the output with a two-class summary (confusion matrix and all related statistics).**





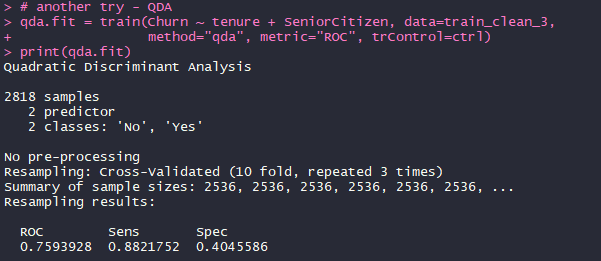
The accuracy of this model is only a fraction less than the logistic model but a fraction less likely to be a product of mere chance.

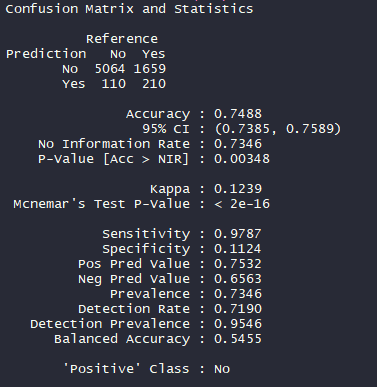
**Step 11: Using the reduced data set, run a cross-validated QDA model with the caret package and fully document the output with a two-class summary (confusion matrix and all related statistics).**

**Ans: (Using caret package)** For the QDA please refer to the code, I tried to do it with the entire dataset and it didn’t run. Then I tried to omit the missing values, which I supposed were already taken out when I created the reduced model, this step also didn’t work.

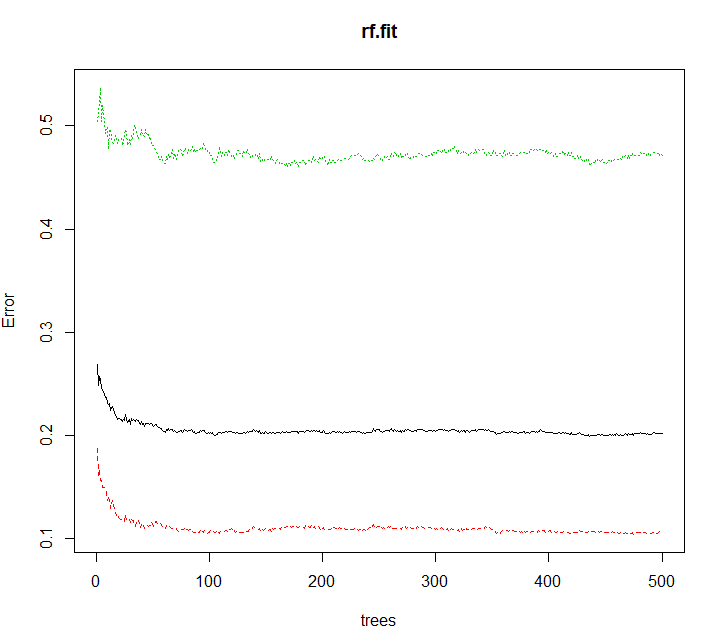
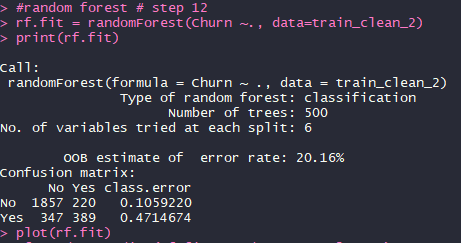
Afterwards I tried to see all the warnings and reverse engineer the issue because to no avail. I tried taking the predictors that were highly correlated using hector but it gave me a rank deficiency error so I decided to not pursue it further.

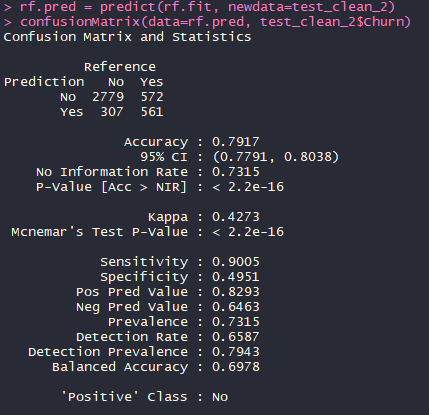
**(Without caret package)** I tried to run QDA on only a couple of predictors just to see if it runs and it did but there is not point of a model that does not take into consideration all the relevant predictors because the results aren’t trustworthy or accurate. Furthermore the Kappa value was only 12% and an accuracy of 74.88% only.



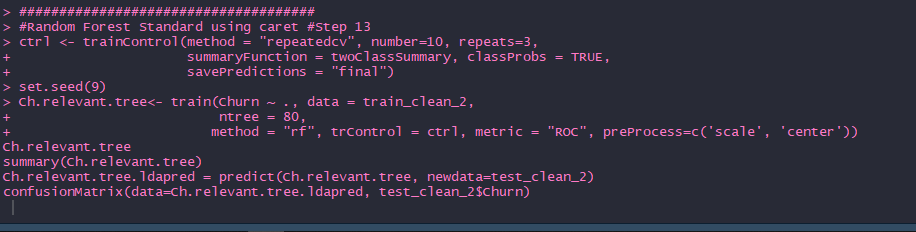


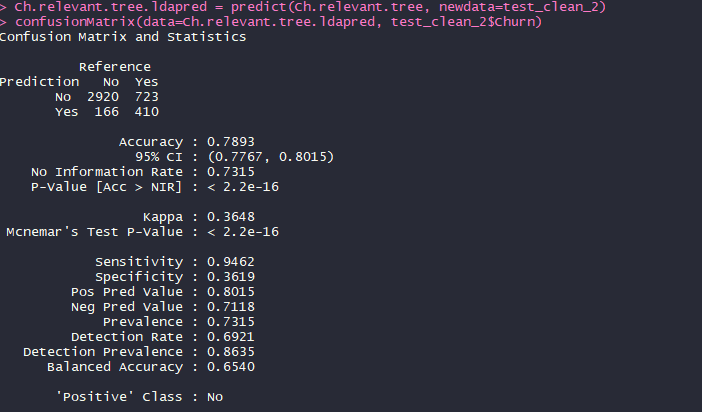
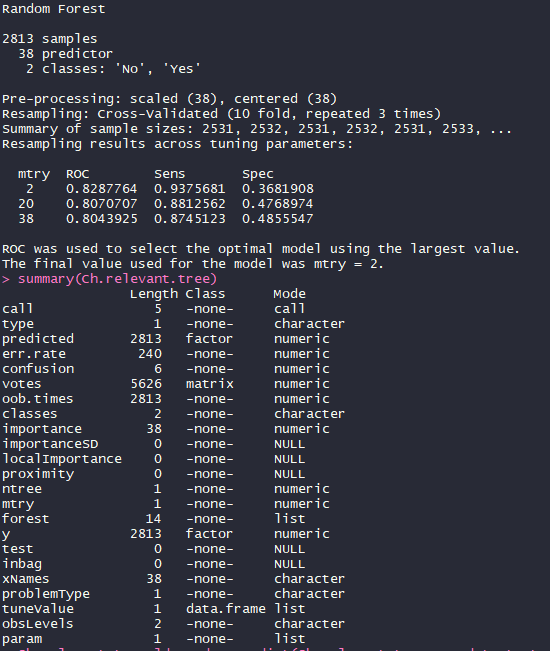
**Step 12: Using the reduced data set, run a standard random forest model to discover the optimum number of trees. You will use this number of trees in the next model.**



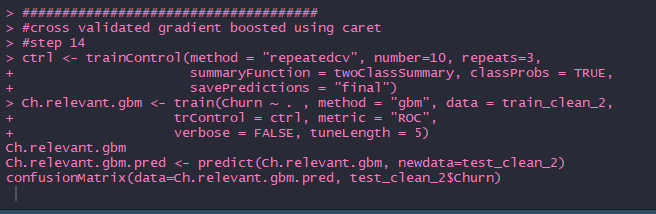


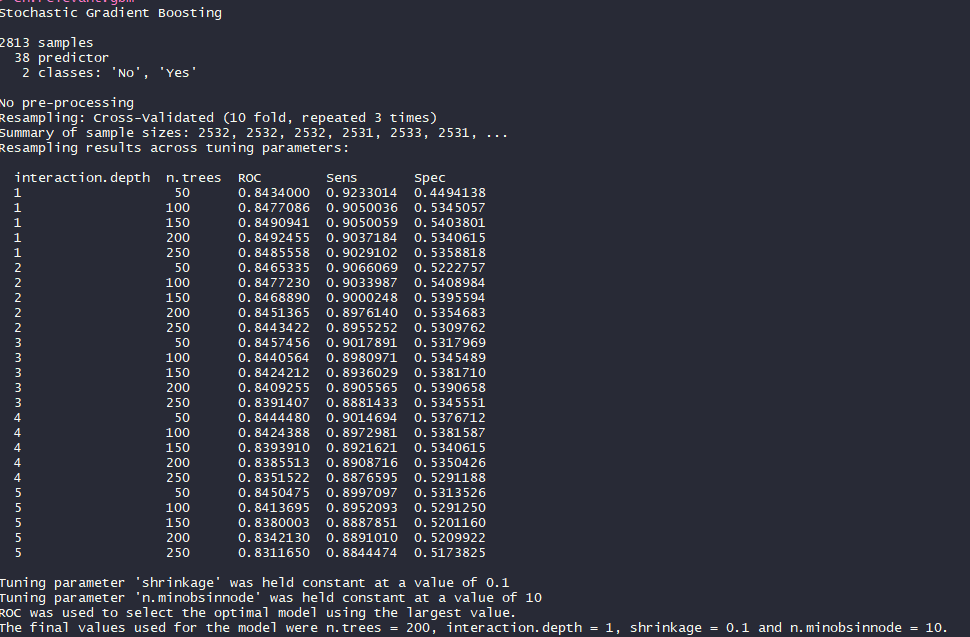
**Step 13: Using the reduced data set, run a cross-validated Random Forest model with the caret package and fully document the output with a two-class summary (confusion matrix and all related statistics).**

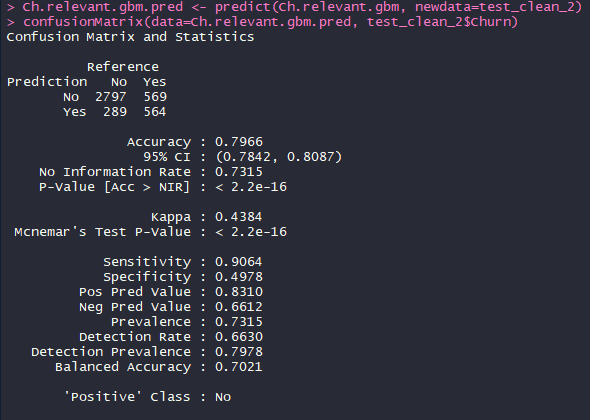




**Step 14: Using the reduced data set, run a cross-validated gradient boosted tree model with the caret package and fully document the output with a two-class summary (confusion matrix and all statistics).**







**Step 15: Identify the best model of the five algorithms and write up specific recommendations to senior management as to how to reduce churn without negatively affecting finances.**

**Ans:** Given the results of the model above it appears that the logistic model and the cross-validated LDA are both almost equally accurate in identifying the churners (80.16% vs 80.07%), followed by cross-validated gradient boost (79.66%). However I would choose cross-validated LDA over logistic model even though the LDA is relatively less accurate compared to the logistic model and my rationale for that is the fact that the LDA has a greater Kappa. In business, uncertainty is one of the most important factors for decision-making. If a business is able to limit uncertainty the management can prepare for different scenarios including unfavorable scenarios and tune their business model accordingly. The management should consider the following:

The fiber optic service of the company is the biggest contributor to the reasons why customers are leaving the company. It seems that the speed or the overall quality of the service is not up to mark because other customers who are not availing the internet service are actually less likely to leave the company. An apparent response to this problem will be to increase bandwidth or reduce latency by switching older systems with new ones or upgrade the existing ones.

After the fiber optics, multiple line users are the second biggest churners from the company. It is not easy to guess why this could be a problem but a common multiple line issue would be sustainability. Everyday customers are less concerned with the technical aspects of these lines but are more aware of the number of things they have to deal with. Which means if there are multiple lines coming into a typical household for phones, internet, TV and others services, the customer is more likely to be annoyed and instead prefer a single line which can support all the activities simultaneously. The company may be using these lines either to charge higher fees or to reduce costs for the existing systems. The company should try to merge all the services in a single package or market better to tell the customers why such steps had to be taken.

On the bright side, it appears that customers tend to like the service if they have availed it for a longer time period. This is based on the fact that people tend to be intimidated by major changes once they enter a comfort zone. The company can use this as a learning point and offer free trials of 3 or 6 months, whichever is more economically feasible. Once new customers get a hang of the services and benefits, they are more likely to get into a contract and continue the relationship for an extended period of time and are less likely to leave the company.