**Group Project**

**Group 1**

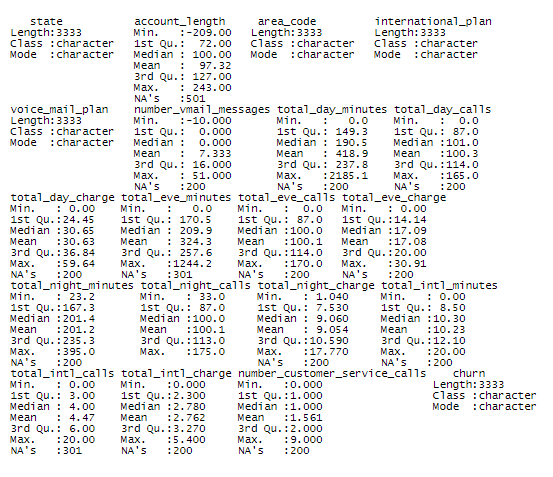
|  |  |
| --- | --- |
| Group Member | Summary of Contribution |
| Meshari Alqeryan | Model building, training, report drafting, formatting, and PowerPoint presentation. |
| Marneni Abhinay  Chiranjeeth | Model building, training, report drafting, formatting |

**Project Goal**

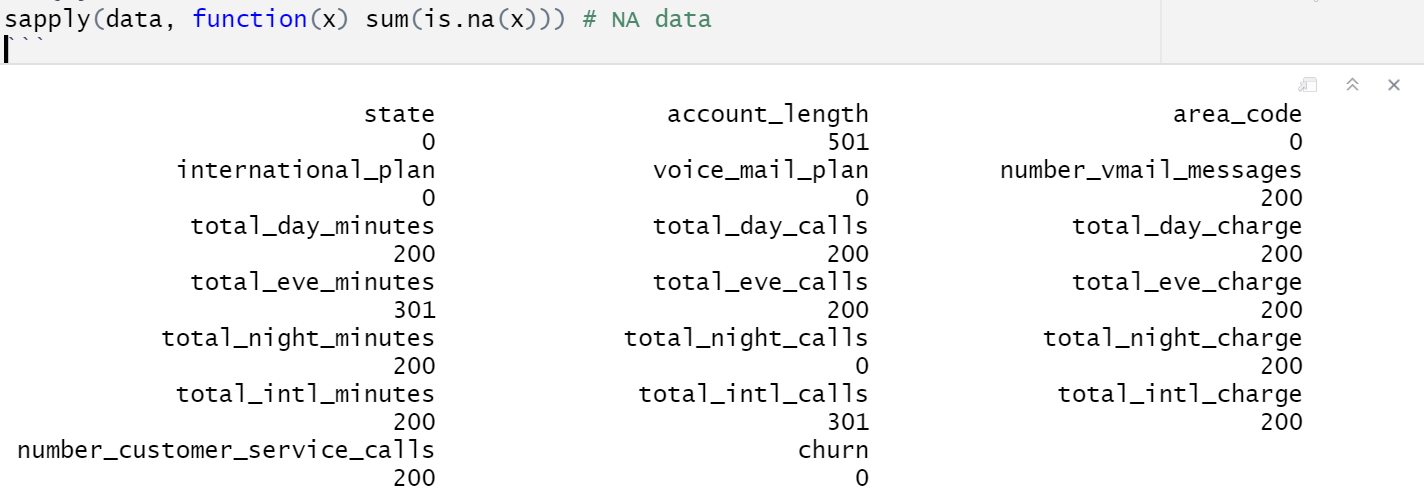
This project aims to build a model that can accurately predict the churn possibility of existing customers for ABC Wireless Inc. We will apply data science principles and analytics methods to address this issue. In this project, R is used to implement the project analysis.

**Data overview and Analysis**

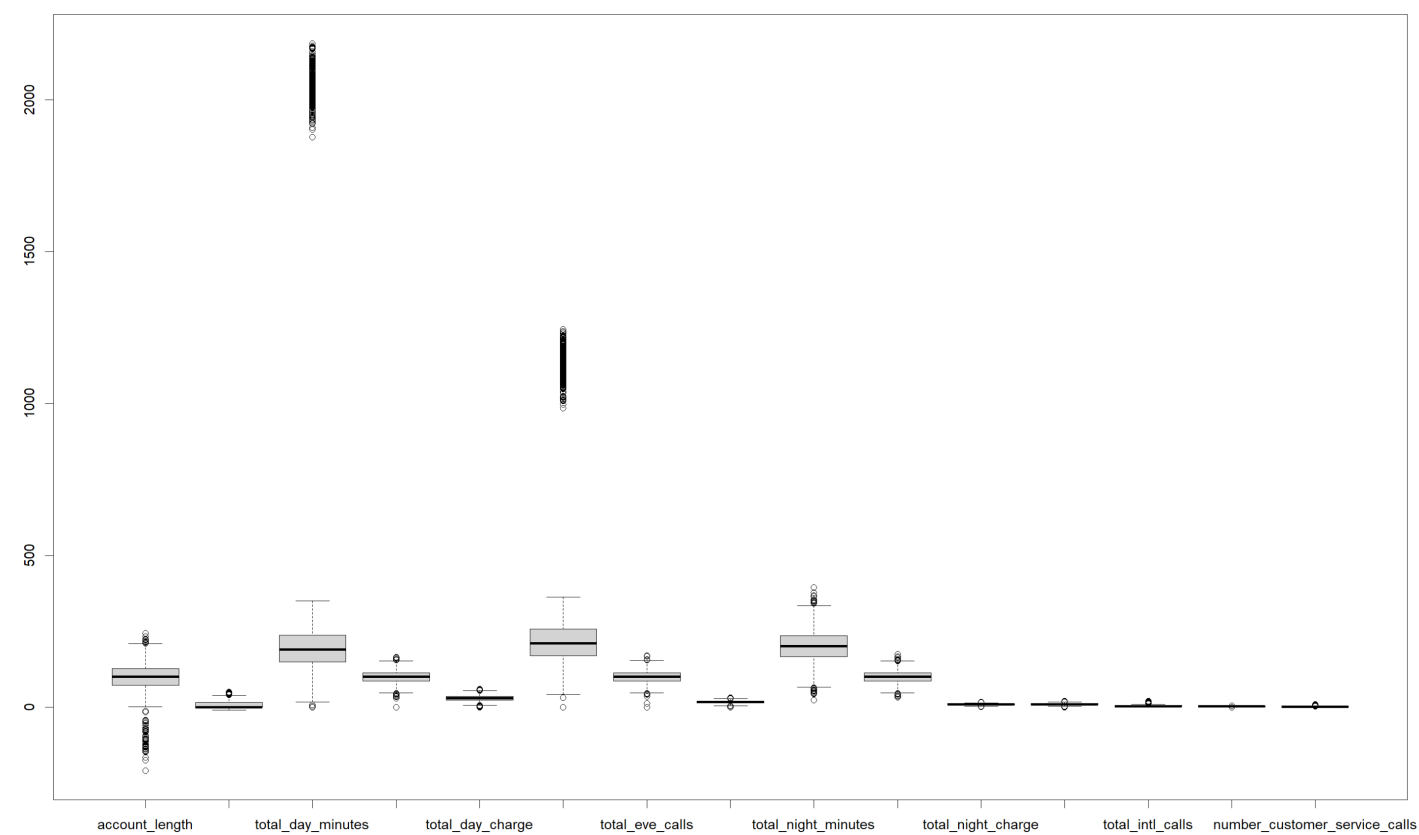
We are provided with the historical data of ABC wireless Inc. customers for this project. The dataset “Churn\_Train” has a total of 3333 observations and 20 variables, one of which is churn, our targeted variable. The churn variable appeared to be binomial and contained yes or ‘no’ factors so that will be the outcome of our project. The dataset has a huge number of observations and 20 variables so it will require preparation before we work on our model. By using the summary function in R we get the below overview:



From what we observe in the dataset summary, there are more than 200 observations recorded as NA for all the variables except “area\_code”, “international\_plan”, “voice\_mail\_plan”, “total\_night\_calls” and “churn”. Also, some negative values are present for the variables “account\_length” and “number\_vmail\_messages.



Before we start with data processing, we need to take a wider look to notice if there are any outliers in our data and determine the degree of their effect on our model. We will apply outlier analysis on the numeric values only by using the “Boxplot” function we get:



Looking at the plot we can see some outliers in the “total\_day\_minutes” and “total\_eve\_minutes” variables. Calculating the outlier’s percentage shows that outliers count for approximately 12% of both variables.

**Data Preparation**

After our analysis of the dataset is complete we can prepare the data for our model. The preparation goes as follows:

* **Missing data analysis**: More than 200 missing data are in almost every column, yet we did not drop any columns or rows with missing values. All the missing values are imputed.
* **Outliers analysis**: Outliers existed on two variables at approximately 12% which is considered minor and did not predict to affect the result of our model so we haven’t removed them completely.
* **Negative value analysis**: some negative values were present for the variables “account\_length” and “number\_vmail\_messages” which we decided to normalize them using the abs function.
* **Converting data into factors:** to covert categorized data into factors to build the model
* **Data splitting:** Splitting data into 70% training and 30% validation

**Model Building and Discussion**

Since the target variable (churn) is classed as binomial, we chose the logistic regression model that can accurately predict the desired outcome based on the independent variables that are provided in the “Churn\_Train” file. We have to go through a series of steps to build an accurate model for this project:

**Step1:**

Import the “Churn\_Train.csv” file into R.



**Step2:**

After importing the dataset into R, a filtering and cleaning process must take place to handle any missing values, and to deal with the negative value we used the **abs function**. By doing so, now the data is ready to build the logistic regression model.

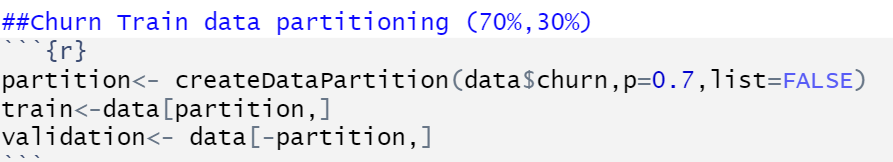


**Step3:**

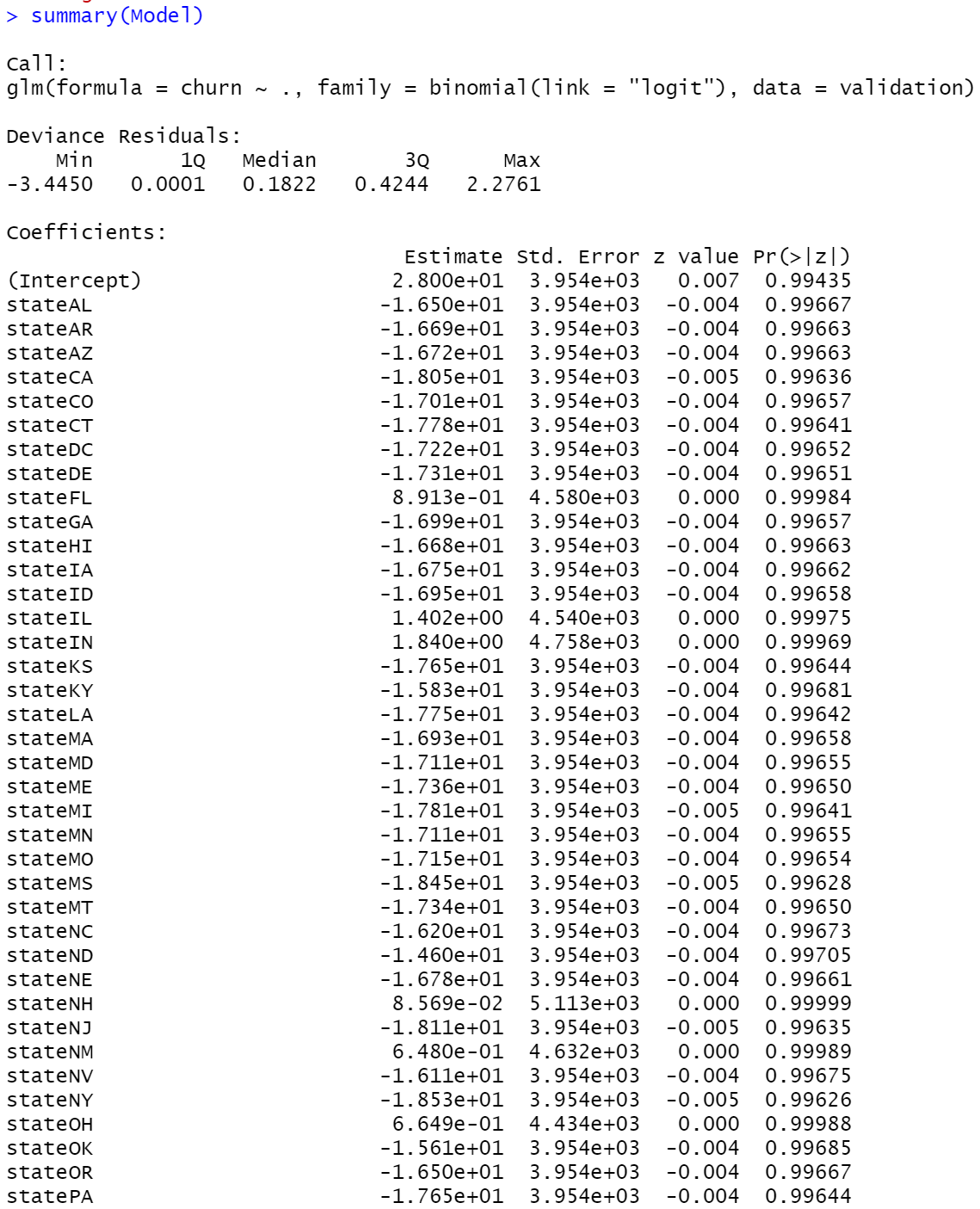
In R we can use a function called “glm” to build our logistic regression model. “glm” is used to fit logistic models to data frames. Each model will be analyzed and tested to eventually elect the most accurate model to predict the price of a house based on a certain number of predictor variables.

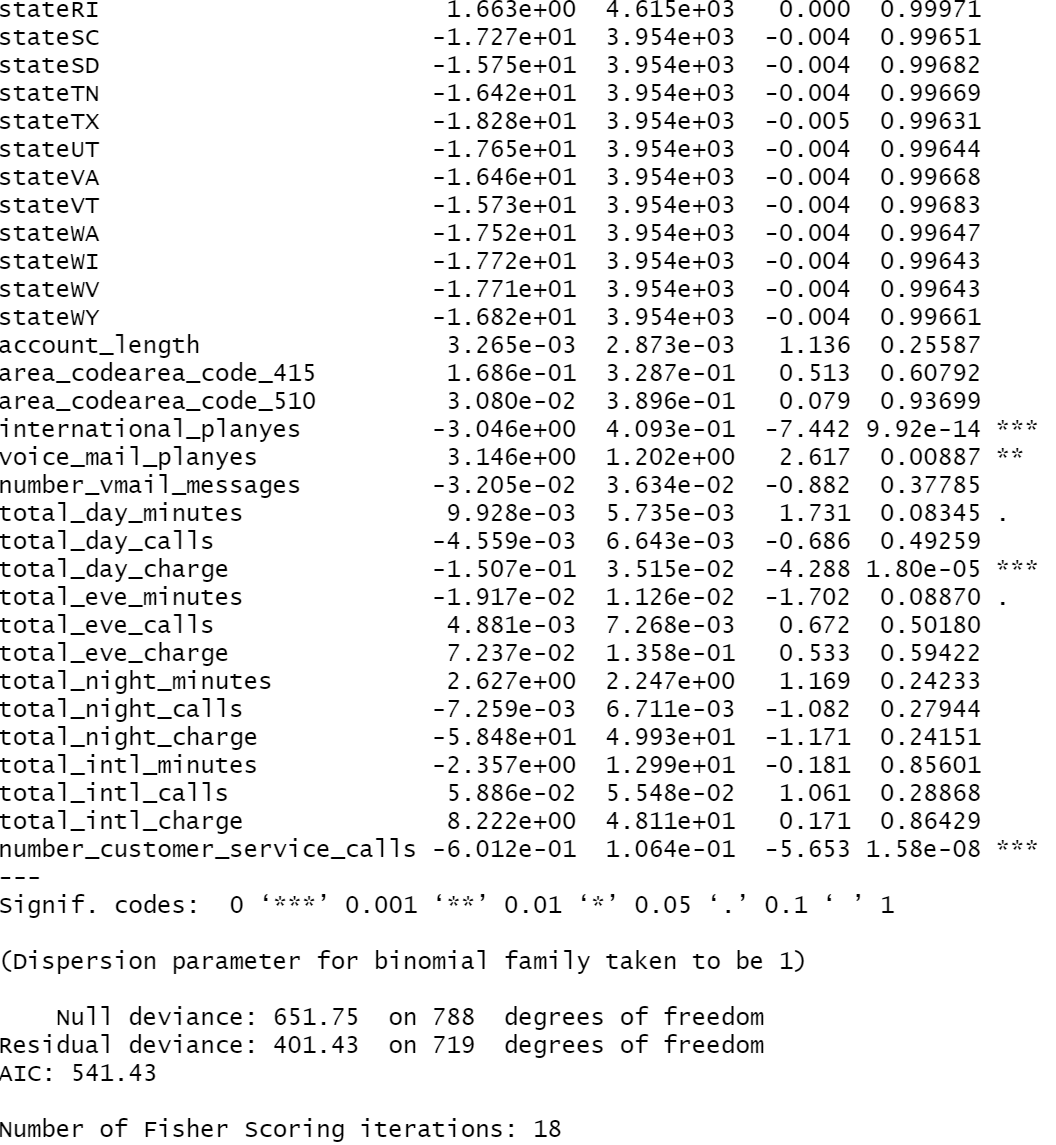
**ModelBuilding**:

Now we believe that the data set is ready for actual model building and testing. To do so, we decided to split the data into training (70%) and validation (30%) data sets so that we can compare how well our model performs on test data using the model built on the training data set.



**Model Summary**

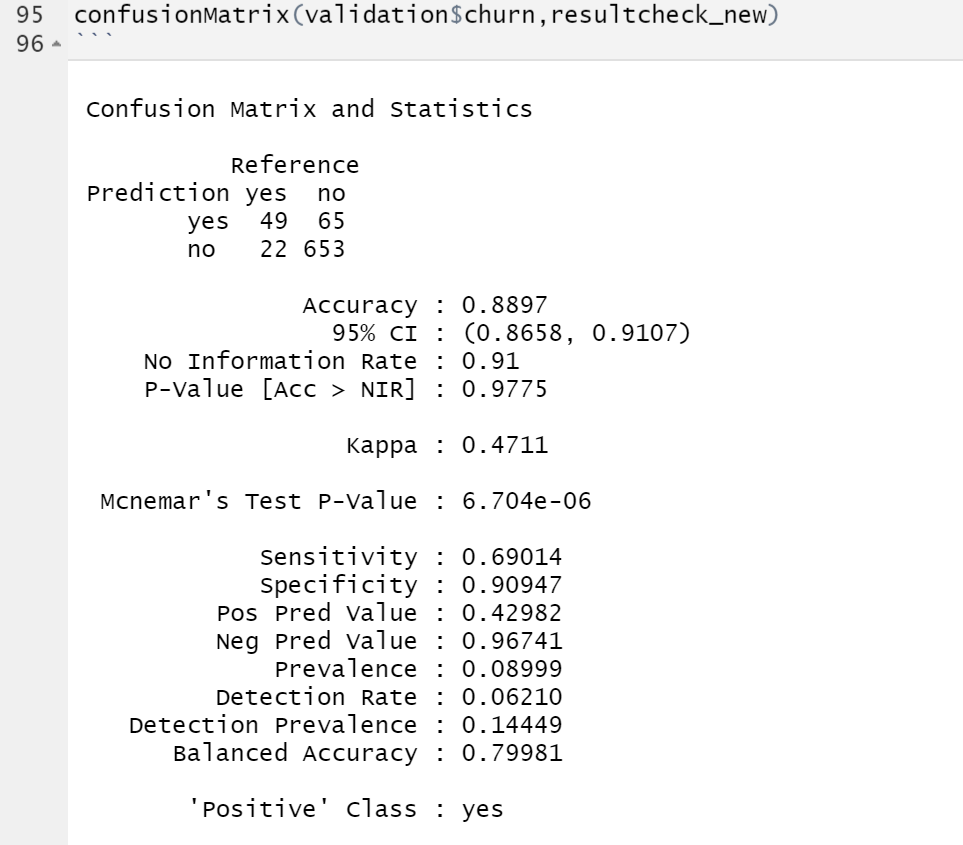




* **Before we move forward, we need to test the model and see how well it’s doing. We are going to perform 2 testing methods using a confusion matrix and AUC**

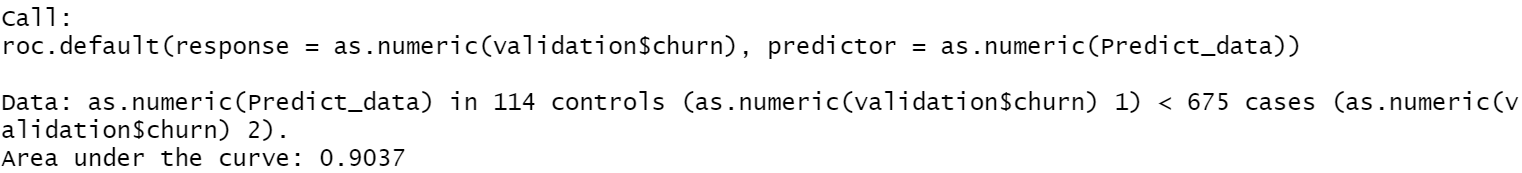
**Step5: Model Testing:**

* Confusion Matrix: we used a confusion Matrix to visualize the performance of our model.

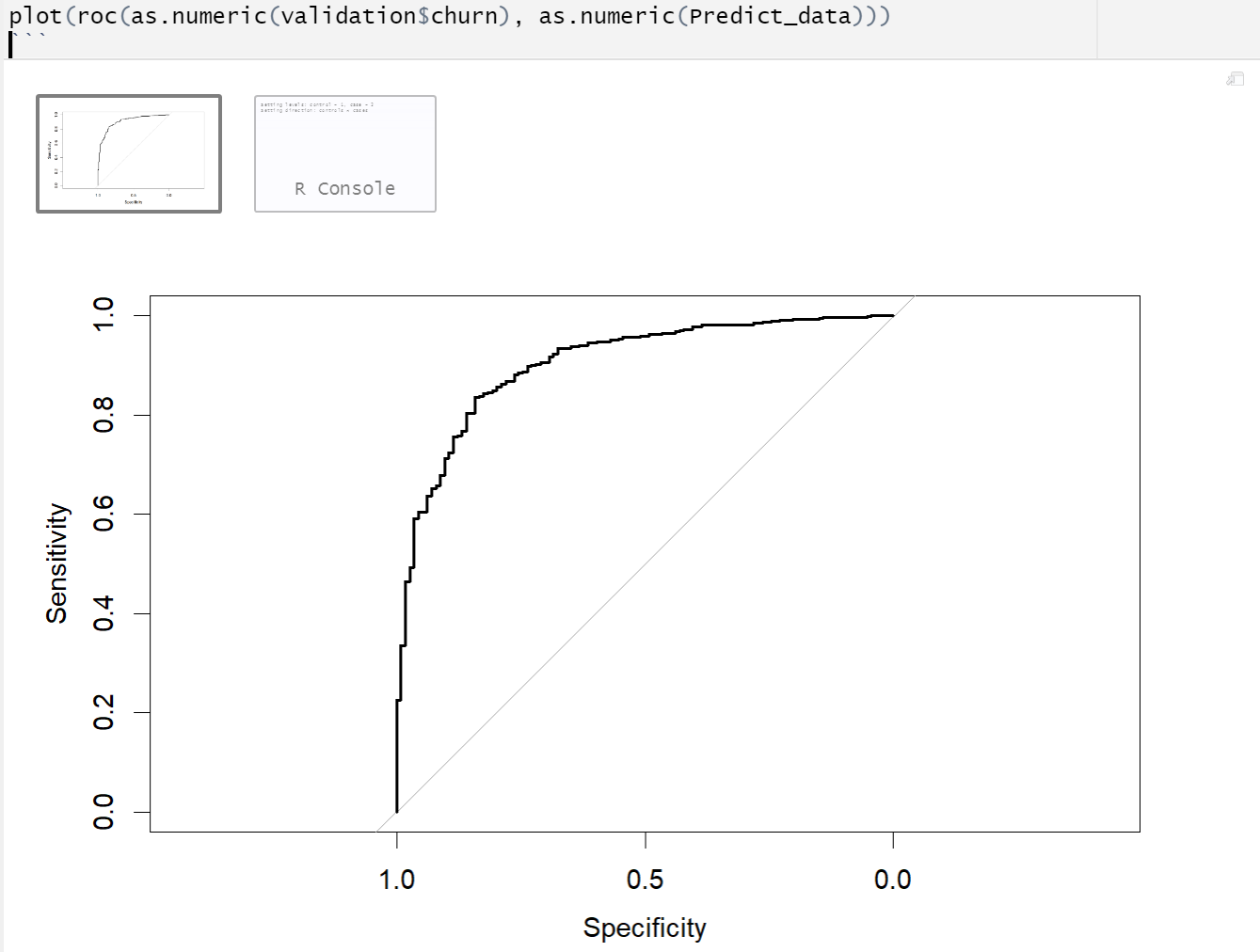


|  |  |  |
| --- | --- | --- |
| Confusion Matrix | Actual | |
| Predicted | NO | YES |
| NO | 653 | 22 |
| True negative | False Positive |
| YES | 65 | 49 |
| False Negative | True Positive |

* **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 90% which is a very good indicator of our model being very accurate.



AUC graph:



**Overview and Discussion**

* 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 87 % accurate results.
* The logistic regression model can help the company determine the customers likely to churn. The model displays that international\_plan, voice\_mail\_plan, total\_day\_charge, total\_intl\_calls, and number\_customer\_service\_calls are more significant than other variables. As per the model prediction, the company can offer suitable offers only to those that customers are going to churn to other competitors to prevent that from happening.
* In addition to the model prediction the company might add the below to maintain and reduce the churn rate.
* After-sales services to be improved to maintain the customer happy and loyal to the company
* Feedback option which will significantly improve churn prediction in case any customer leaves for another provider we can use their feedback and determine the area immediately and try to improve it to avoid losing other customers.
* Advertising is one of the key aspects of success, as more publicity for the company the better the company progress and stays competitive with others in the same field.