# Predict Customer Churn by Using Logistic Regression in R

**Introduction**

**Customer churn happens when customers stop using the services of a company. This problem is very important for different companies because every company needs to be aware of the customers churn rate. It is very crucial to know the number of customers who are stop using the company’s services because the companies should find out the reason of customer churn to make proper decisions in this matter. In this project, we are going to predict customers behavior based on the Telcom dataset of a company in California and the logistic regression model In R will be performed.**

**Data Preprocessing**

**First, the preprocessing of data is the most important part of the data analysis which has been performed. Before fitting the proper model, we need to filter the data and perform data preprocessing like cleaning, transformation and model selection. The data we are using in this project is downloaded from Kaggle.com. This data included the response variable which is customer churn with different independent variables which are going to be used in this prediction. In this dataset, each row indicates a customer and each column is a feature. The target and Independent variables have been described as below:**

**Target variable:**

**Churn - Whether the customer churned or not**

**Response Variables:**

**Gender - Whether the customer is a male or a female**

**Seniorcitizen - Whether the customer is a senior citizen or not**

**Partner - Whether the customer has a partner or not**

**Tenure – Number of months the customer has stayed with the company**

**Dependents - Whether the customer has dependents or not**

**Phoneservice - Whether the customer has a phone service or not**

**Multiplelines - Whether the customer has multiple lines or not**

**Internetservice - Customer’s internet service provider**

**Onlinesecurity - Whether the customer has online security or not**

**Onlinebackup -Whether the customer has online backup or not**

**Deviceprotection - Whether the customer has device protection or not**

**Techsupport - Whether the customer has tech support or not**

**Streamingtv - Whether the customer has streaming TV or not**

**Streamingmovies - Whether the customer has streaming movies or not**

**Contract - The contract term of the customer**

**Paperlessbilling - Whether the customer has paperless billing or not**

**Paymentmethod - The customer’s payment method**

**MonthlyCharges - The amount charged to the customer monthly**

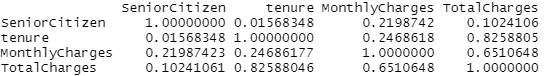
**TotalCharges - The total amount charged to the customer**

**Missing Values**

**In the first step of data preprocessing, we removed the columns which we do not need in our data analysis which is “customerID”. Then we check the data for missing values which we realized that we have missing values for the column of “TotalCharges” and in this case, we have chosen the complete cases of the data to have a data without any missing values.  There are only 11 rows and deleting them will not affect the data. So, after removing the missing values, now we have 7032 data points in dataset.**

**Exploratory Data Analysis and Model Selection**

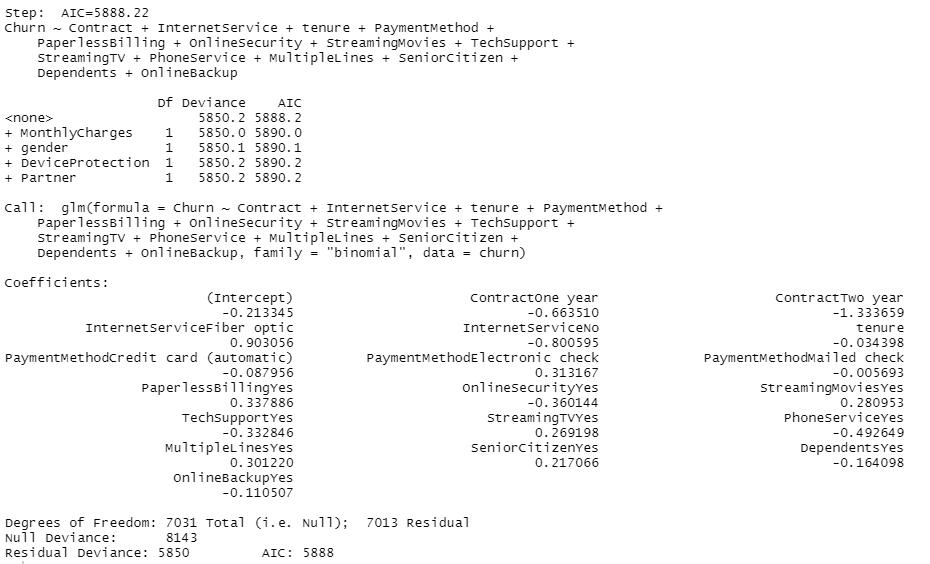
**Based on the correlation table above, the “TotalCharges” and “MonthlyCharges” are correlated, so we can remove one of them from data set. Here, we delete the “TotalCharges”.**



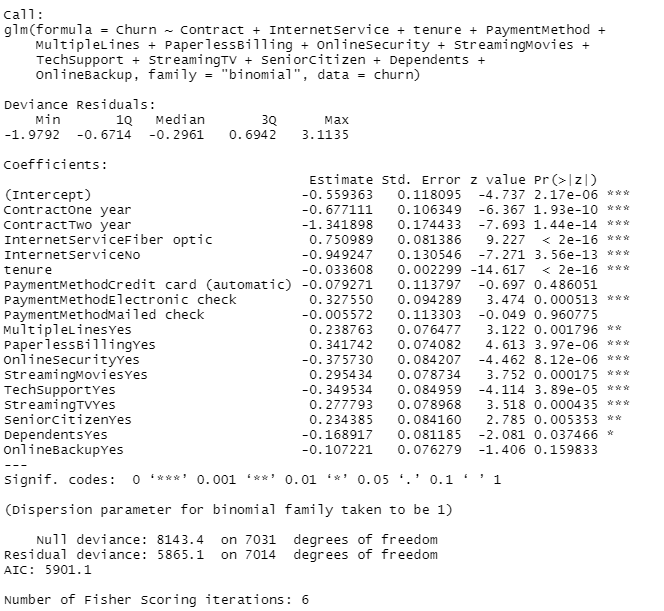
Then, we have noticed that some columns like “OnlineSecurity”, “OnlineBackup”, “DeviceProtection”, “TechSupport”, “streamingTV”, “streamingMovies” had three values as “No”, “Yes” and “No internet service”. So, we changed the “No internet service” to “No”. Also, for the column “MultipleLines” we have changed the ““No phone service” to “NO”. Now, we have a two factor for these variables. In addition, we have changed the values in column “SeniorCitizen” from 0 or 1 to “No” or “Yes”.

**Model Selection**

In this project, stepwise procedure has been performed for selecting the best possible models.

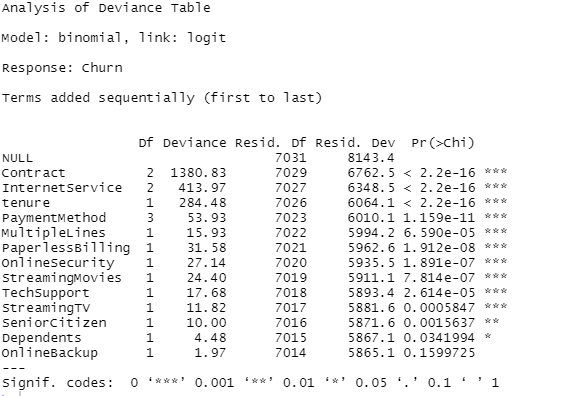


Then we fitted the model with all the possible variables which have been selected by stepwise procedure.



**Goodness of fit**

Below is the ANOVA table for the performed model:



Also, we have performed the cook’s distance on the dataset and there were no influential points in the dataset because we didn’t find any large values based on the cook’s distances. Besides, we have performed the Hosmer-Lemeshow Test for investigating the goodness of fit of the model. Based on the results, p value is 0.7102 which means that the suggested model is a good fit.

**Collinearity**

In this section, we have checked the collinearity by using variance inflation factors. Based on this method, if any of VIFs are greater than 10, then we can say that there is collinearity in the model. But, here we did not have any VIFs greater than 10, so there is no collinearity between the predictor variables.

**Power**

Another metric to investigate the accuracy of our model is the power of the model using McFadden R2. Here, the McFadden R2 is 0.2797 which is between 0.2 and 0.4, so we can conclude that the model is a good fit for predicting the customer churn.

**Cross Validation**

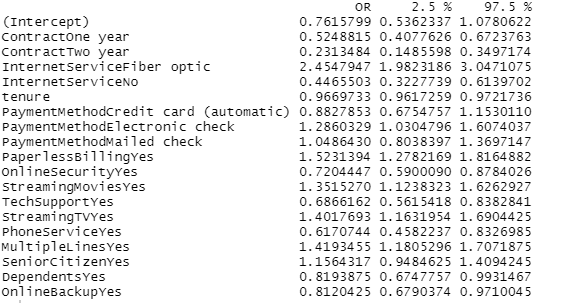
We split the data into training and testing sections and then we fitted the Logistic Regression Model on the training data.

Then we assessed the predictive ability of the Logistic Regression model by finding the accuracy rate. Here the Logistic Regression Accuracy is 0.80. Also, we have found the Confusion Matrix for Logistic Regression which is as below:



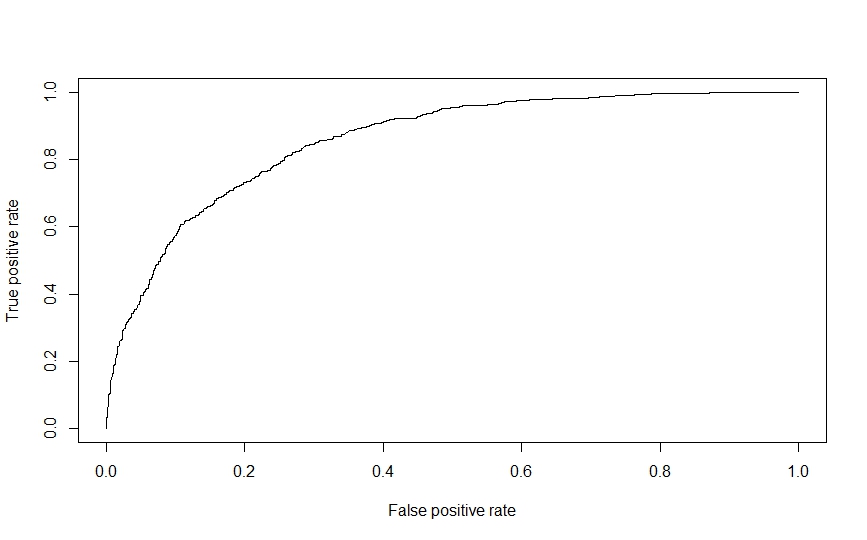
**Odds Ratio**

Another performance measurement in logistic regression which we have used is Odds Ratio. Odds ratio indicates the odds that an outcome will happen which has been represented as below:



**ROC analysis**

After fitting the logistic model, Roc analysis has been performed, and ROC curve has been plotted. As can be seen, plot reached to 1 and stays flat which means that the model is accurate. Besides, the area under the curve is 0.8401815 which is a good result because more area under the ROC curve, the greater the accuracy. So, we can say that our model is a good fit for predicting the customer churn.



**Conclusion:**

**This project was about analyzing the customers behavior based on the Telcom dataset of a company in California and the logistic regression model In R has been performed to predict if the customers will churn or not. As we can see in the results, some variables like “tenure”, “Contract”, “PaperlessBilling”, “MonthlyCharges” and “InternetService” were the most important ones which have an effect on customer churn. On the other hand, there was not any connection between gender and customer churn.**

**Data Reference:**

**https://www.kaggle.com/datasets/blastchar/telco-customer-churn**