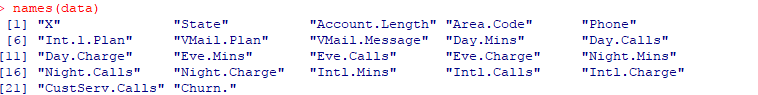
The Task:

The purpose of this analysis and case study is to predict the customers who are likely to stop using the servicebya telecom company strategize accordingly so that it can retain their customers.Customer churn is when an existing customer, user, player, subscriber or any kind of return client stops doing business or ends the relationship with a company. So, to solve this problem, the nature of the relationship between of each variable with Churn must be understood along with the individual characteristic of each variable. A statistical model is then adopted to further the analysis and arrive at the results and interpretation.

The Dataset:

The dataset contains the following variables:

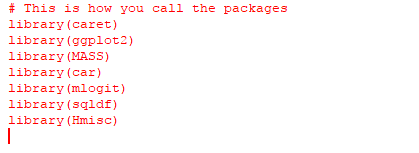


Variables “X” and “Account.Length” has been removed as these weren’t necessary.

The statistical model:

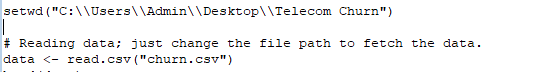
We have adopted the logistic regression model analysis in this case. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). In the following pages of documentation, the approach steps have been clearly outlined.

Setting up the R model by loading the required libraries:



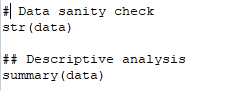
Data:

In the next step, the data is read into the R environment from the file.



Data selection and data type modification:

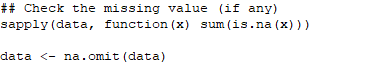
In the given dataset, no modification required.



Data Cleaning:

In this step all the numeric variables are checked for the presence of outliers and none was found.

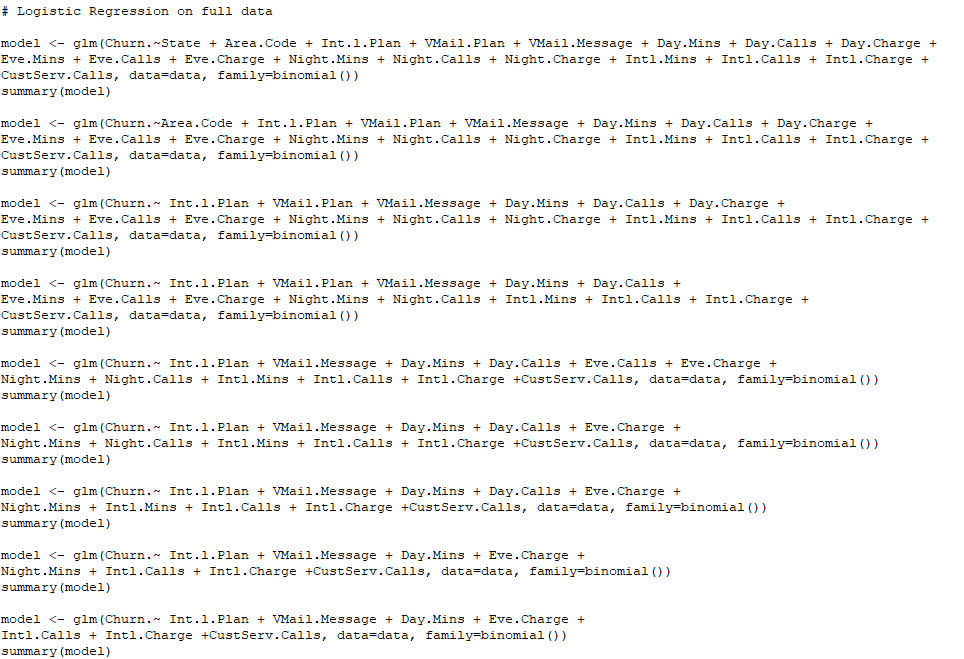
Checking for missing values:



After the data has been cleaned off all the outliers, it is then checked for any missing values in the following manner. 2217 missing values were found.

**Running the linear regression model:**

Once the data has been cleaned, a logistic regression has been performed with the value as the dependent variable. Once the model has been run, the anova value for each individual variable is checked. The variables with p value< 0.05 are removed from the model one by one such that only the statistically significant ones remain.

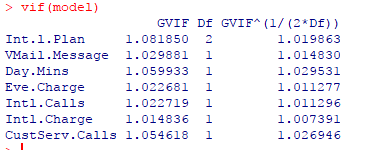


Goodness of fit of the model:

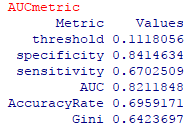
Calculate the C statistic (equivalent to the area under the Receiver Operating Characteristic Curver ROC) for a logistic regression model, a measure of goodness of fit for binary outcomes in a logistic regression model. Values for this measure range from 0.5 to 1.0. A value of 0.5 indicates that the model is no better than chance at making a prediction of membership in a group and a value of 1.0 indicates that the model perfectly identifies those within a group and those not. Models are typically considered reasonable when the C-statistic is higher than 0.7 and strong when C exceeds 0.8. Here it is 82% which indicates a high amount of goodness of fit of the model.

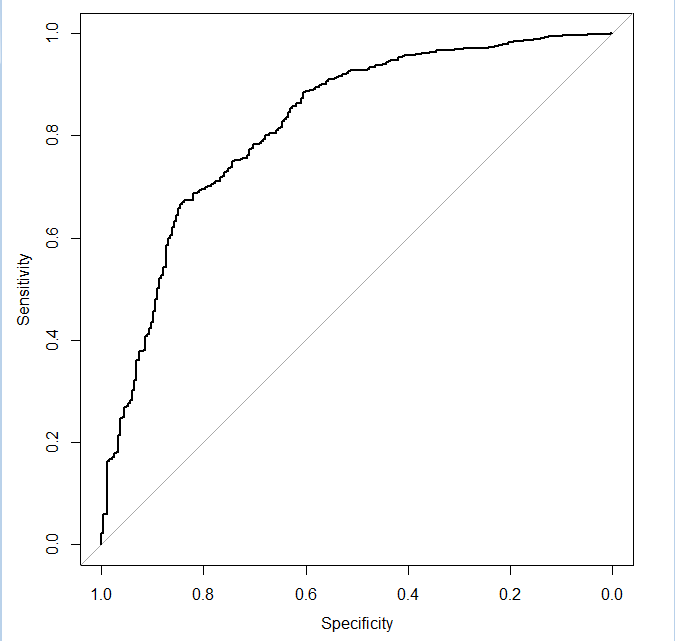
**Performance of the Logistic Regression :**

1. ***Assumption of multicollinearity:*** - This is the most important of the assumptions of a linear model and it states that there should be no perfect linear relationship between two or more of the predictors or independent variables. This is tested with the vif function and any variable with a value of GVIF should be within 2. If it is greater than 10 then serious problem. In our case multicollinearity between independent variables was absent.

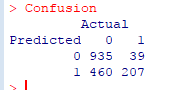


1. ***Gini Coefficient:-*** It is the most commonly used measure of inequality. The range of the Gini coefficient goes from 0 (no concentration) to √(n−1n) (maximal concentration). The bias corrected Gini coefficient goes from 0 to 1. The small sample variance properties of the Gini coefficient are not known, and large sample approximations to the variance of the coefficient are poor. In this case it is approximately 0.62.
2. ***Accuracy Rate:-*** Accuracy is one of the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Higher accuracy means model is preforming better. In this case it is approximately 70% which implies to be a better model.





***4>Confusion Matrix:-*** It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. This is how it looks like:

****

**Validation of the model:**

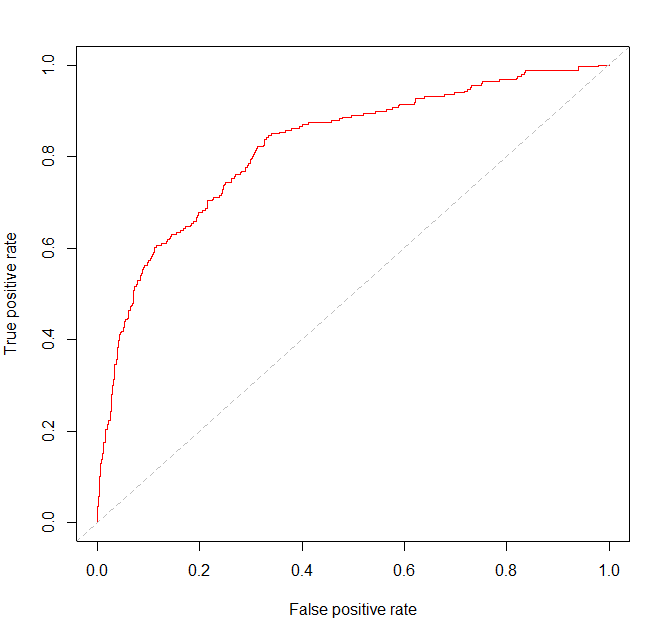
The predictions for the value variable are done using the validation part of our data. The predictions are then saved in a csv file for reference.

names(data)[23] <- "pred"

write.csv(data,"churn\_result.csv")

***KS Statistics:*** Kolmogorov-Smirnov (KS) statistics is one of the commonly used measures to assess predictive power for marketing or credit risk models. The KS statistic is usually published for logistic regression problems to give an indication of the quality of the model.The value for KS ranges from 0 to 1 and the closer to 1 the KS, the better is the model. The calculation is :

C:\Users\ADMIN\Desktop\Telecom Churn\KS statistics.PNG



We can assume the model to be ok because of above result.

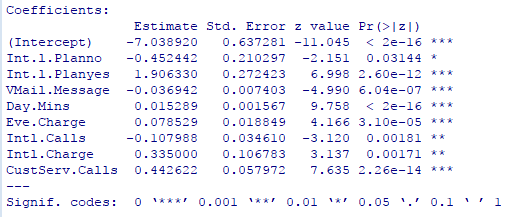
The significant variables and their significance:

The following image shows the variables that are significant to our model and the image below shows us the relationship of these variables with the dependent variable. The following are the significant variables with a positive relationship with customer lifetime value:

* Int.l.Planyes
* Day.Mins
* Eve.Charge
* Intl.Charge
* CustServ.Calls

However, the following are the significant variables with a negative relationship with the dependent variable:

* Int.l.Planno
* VMail.Message
* Intl.Calls



The Business Interpretation

For the Telecom Company to prosper they should focus on the following targets:

* The company should focus on customers who have a **International line Plans**. These customers are potential goldmines when it comes to churn as they have positive relationship as inferred by the model.
* **Day.Mins** should also be targeted as how long a customer is on a call at day time must be noted. They also have a strong positive relationship with churn.
* The company must look into the **Eve.Charge** as it has a significantly positive relationship with churn.
* Higher **International Charge** can be fruitful and this variable should be kept within the fold.
* Better the **customer service calls** higher the chance if retaining customers.
* Those customers with **no International line Plans** must not be pursued as it has significantly negative relationship with churn.
* Customers who use frequent **voice mail messages** are not so important as it have much negative relationship.
* Customers who makes **International calls** have significantly negative relationship with churn.