Ankit Sharma

Bhavya Gupta

Rahul Mishra

Suvo Pal

|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Predictive Modelling  Group - 15 |
| |  |  |  | | --- | --- | --- | | PGP-BABI July’17 | 1/24/18 | Great Lakes | |

**Overview**

The ability to predict that a particular customer is at a high risk of churning, while there is still time to do something about it, represents a huge additional potential revenue source for every online business. Besides the direct loss of revenue that results from a customer abandoning the business, the costs of initially acquiring that customer may not have already been covered by the customer’s spending to date. (In other words, acquiring that customer may have actually been a losing investment.) Furthermore, it is always more difficult and expensive to acquire a new customer than it is to retain a current paying customer.

**Data Description**

Telecom dataset has the details for 3333 unique customers with11 feature variables where details of each customer are represented in a unique row and below is the structure of the dataset,

> str(data)

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 3333 obs. of 11 variables:

$ Churn : num 0 0 0 0 0 0 0 0 0 0 ...

$ AccountWeeks : num 128 107 137 84 75 118 121 147 117 141 ...

$ ContractRenewal: num 1 1 1 0 0 0 1 0 1 0 ...

$ DataPlan : num 1 1 0 0 0 0 1 0 0 1 ...

$ DataUsage : num 2.7 3.7 0 0 0 0 2.03 0 0.19 3.02 ...

$ CustServCalls : num 1 1 0 2 3 0 3 0 1 0 ...

$ DayMins : num 265 162 243 299 167 ...

$ DayCalls : num 110 123 114 71 113 98 88 79 97 84 ...

$ MonthlyCharge : num 89 82 52 57 41 57 87.3 36 63.9 93.2 ...

$ OverageFee : num 9.87 9.78 6.06 3.1 7.42 ...

$ RoamMins : num 10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...

From the above description,

**Response/dependent Variable:** Churn (1 being churned out and 0 being retained)

**Predictors/Independent Variables:** All other features in the dataset are the potential predictor variables which will be used for predicting the Customer if he/she would opt-out or remain in the system

**Checking if the data has any NULL Values**

> any(is.na(data))

[1] FALSE

There are four variables in that have the data type as integer when they should actually be as Factors.

1. Churn b) Contract Renewal c) DataPlan d) CustServCalls

####Converting Churn, ContractRenewal, DataPlan Variables to factors

data$Churn <- as.factor(data$Churn)

data$ContractRenewal <- as.factor(data$ContractRenewal)

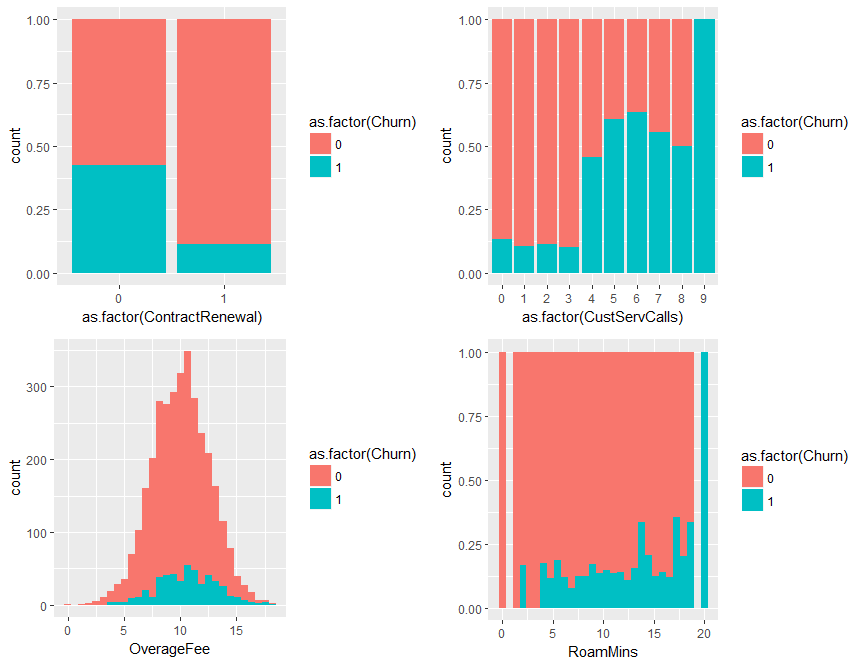
data$DataPlan <- as.factor(data$DataPlan)

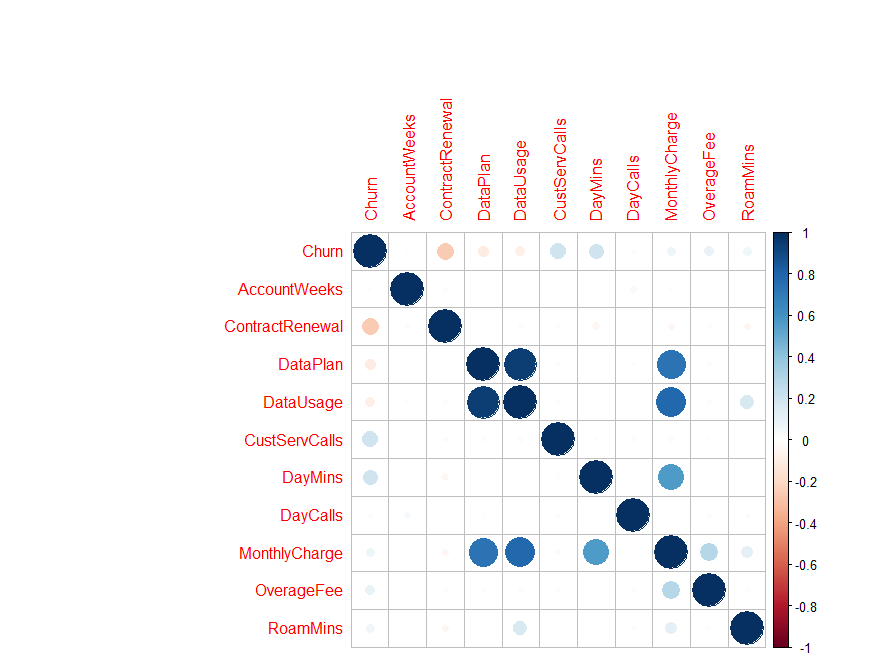
data$CustServCalls <- as.factor(data$CustServCalls)

**Exploratory Data Analysis (EDA)**

**Churn Rate:** The churn rate of the dataset is 14.5%

1. **Contract Renewal vs. Churn:** The Chrun rate is more when the customer did not renewed the contract and is almost thrice then if the customer renewed the contract
2. **ServiceCalls vs. Churn:** The Churn rate is growing as the number of ServiceCalls a customer made. It also reflects the business aspect that the customer who is dissatisifed wil make more customer service calls and hence has more chances of Churning out
3. **OverageFee vs. Churn:** The churn rate is more when the customers had to pay overage fee
4. **RoamMins:** Customers who make more Roaming calls on average have higher rate of churning out from the Telecom Service



**Correlation Plot**

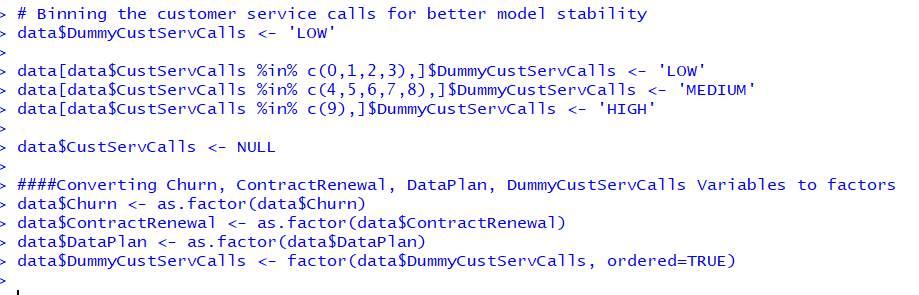
Correlation Plot would help in understanding the Multi-collinearity within the predictor variables.

1. **DataUsage and DataPlan**: The two features seems to have a very high collinearity among them. This is evident and expected as a person with the DataPlan add-on will tend to use the data more
2. **MonthlyCharge with DataPlan & DataUsage**: MonthlyCharge also shows high collinearity with DataPlan and DataUsage which is also relevant from business sense as Charges incurred by the customers will be more if they have DataPlan activated or if they use more data without the Data Plan activated in their accounts

### Data Pre-processing

Data cleansing and preparation will be done in this step. Transforming continuous variable into meaningful factor variable will improve the model performance and help understand the insights of the data.

* For example, in this dataset, the Contract Renewal/ Data Plan variable is converted to factor variable with binomial responses of 1 or 0. Also CustServCalls are binned into High/Medium/Low category converted based on the data and converted to factor variable. Thus, understanding the type of customers with tenure value to perform churn decision.



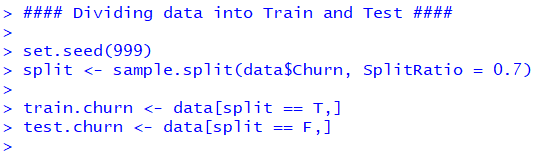
* As part of data cleansing, the missing values are identified using the missing map plot. The telecom dataset has no missing value record.



**Model Preparation**

In the predictive modelling, the data need to be partitioned into train and test sets. 70% of the data will be partitioned for training purpose and 30% of the data will be partitioned for testing purpose.

In this dataset, 2333 customer records are used for training purpose and 1000 records are used for testing purpose.



**Running Logistic Regression**

Classification algorithms i.e., Logistic Regression is used to predict churn using glm package that are available in R.

Running the first regression model, including all the predictor variables in the first iteration.

#### Running GLM on Train dataset ####

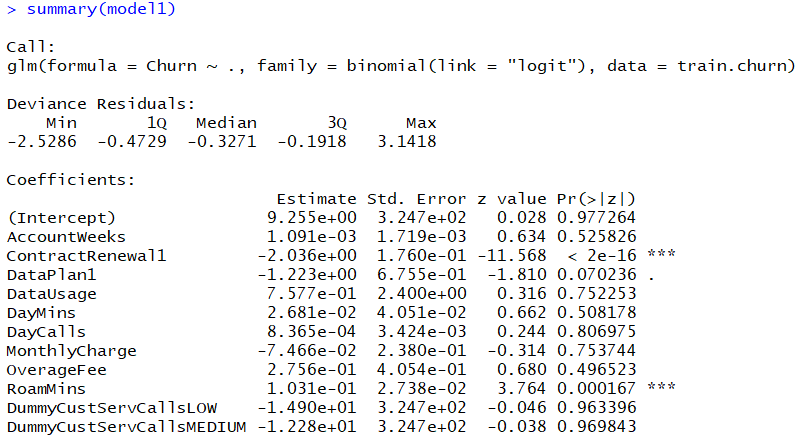
model1 <- glm(Churn ~ . , family = binomial(link = "logit"),

data = train.churn)

**Model Summary**

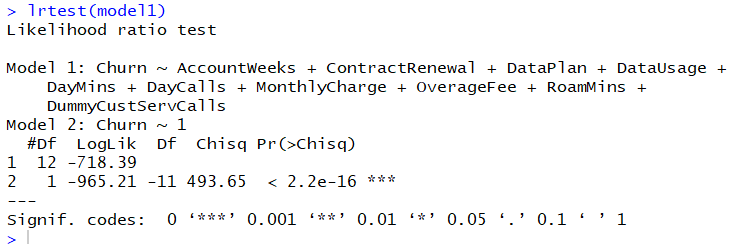
The model summary signifies that **Contract Renewal, Data Plan & Roam Mins** are significant. The first iteration of the model is being run only to see if the logistic regression model is here to stay or not.

We will now run some test to justify if the Logistic Regression model is actually applicable and of any importance in this case or not



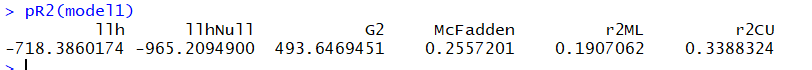
1. **Likelihood Ratio Test**

The overall test of model significance based on the Chisq test above is overwhelmingly significant indicating the likelihood of Churn depends upon predictor variables. Using the statistical language, this implies that the null hypothesis of all Betas are zero is rejected and we conclude that at least one Beta is nonzero



1. **Pseudo R Square Test**

Based on McFadden R Square, we conclude that based on McFadden value 25.57 percent of the uncertainty of the Intercept only Model (Model 1) has been explained by the Full Model(Model 2). Thus the goodness of fit is moderate.



**Running the Final Model**

Variable selection is the important aspect for finalising the model for:

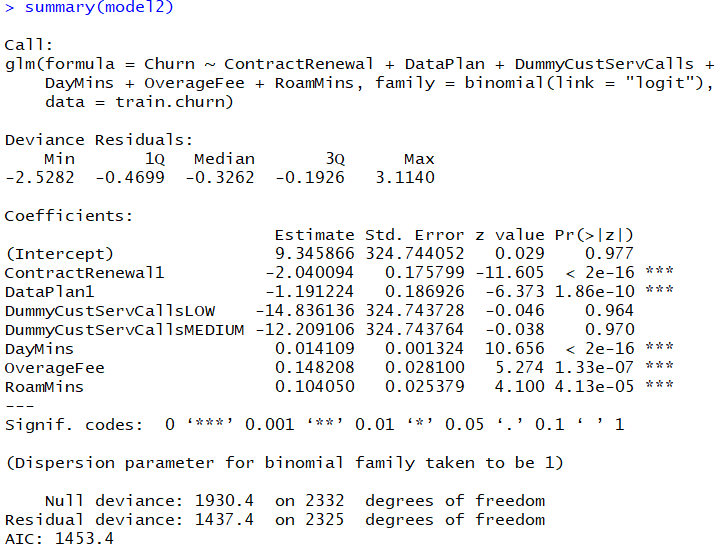
* Variable Selection
* Remove Multi Collinearity Effect
* Increase model performance

*We will run the step function as it will help in the variable selection for the model. The Step function resulted in the following features which comes out to be important with significant p values*

**

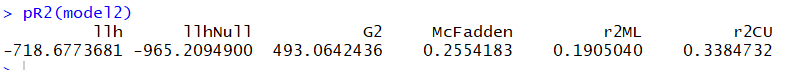
**Model Summary**

the model summary shows that all the variables selected have significant p values which explains that the model independent variables are significant in explaining if the customer would leave or not. AIC value also improved which indicates that this model is more stable.



**Running Diagnostics**

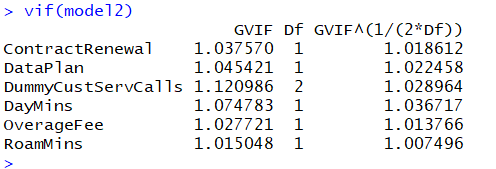
Based on McFadden R Square, we conclude based on McFadden R Square value that 25.54 percent of the uncertainty of the Intercept Only Model (Model 1) has been explained by the Full Model (Model 2). Thus the goodness of fit is moderate.

****

**Multicollinearity Check**

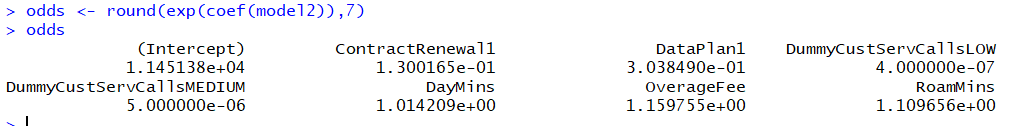
We can check the Multi collinearity b/w independent variables through Variance Inflation Factor (VIF)

*The VIF values for the Independent variables are around 1 which signifies that the multicollinearity problem in the model is absent. This signifies that the model is robust and does not get affected by the multi collinearity problem*

**

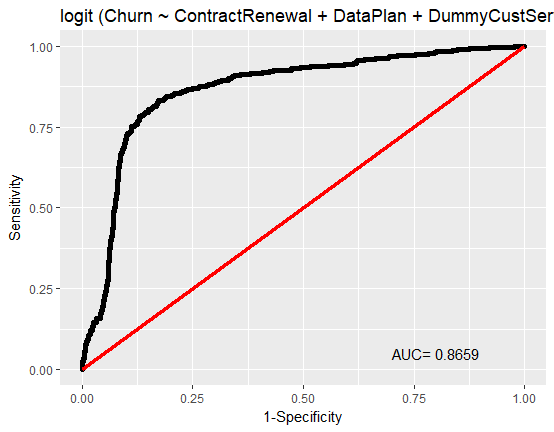
**ODDS Ratio for the Model**

The odds ratio defines how likely the odds are to increase with increase in a unit of any of the independent variables. It basically defines the relativity.



**Interpretation**

*The ODDS ratios for the variables are also significant. An ODD value > 1 signifies that one unit increase would have an impact on increasing the odds for a customer to churn. As evident from the ODDS of the features, all the independent variables are significant in explaining the Churn*

**Prediction Results and Accuracy**

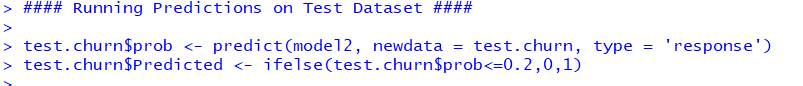
1. **ROC Curve and AUC Value**

The ROC curve is a good one with the AUC value of 86%

The AUC value signifies a strong model which is able to explain the dependent variable Churn.

1. The Model was built on the train dataset. To check the performance of the model, we will now run the model on the test dataset. The model will predict the probability value using the independent variables in the test dataset.

Here we experimented with the probability threshold values and came with 0.2 as the optimal threshold, tested the same with both train and test Dataset and accuracy is maximum at this threshold.

****

**Confusion Matrix / Misclassification Table:**

It is a table used to describe the performance of the classification model on a test data. It is used to cross-tabulate the actual value with the predicted value based on the count of correctly classified customers and wrongly classified customers.

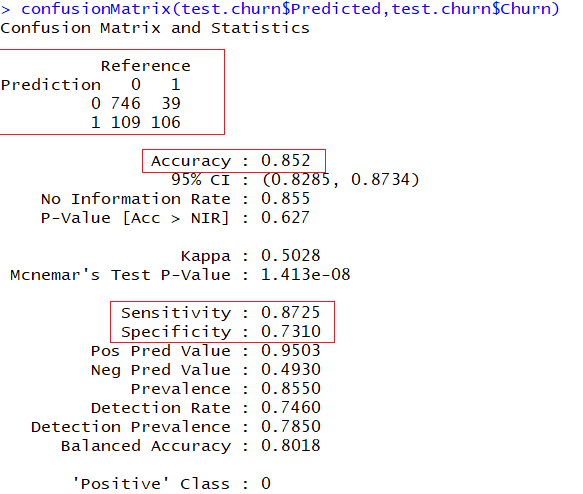
* **true positives (TP):** These are cases in which we predicted yes (They have churned), and they have churned.
* **true negatives (TN):** We predicted no, and they have not churned.
* **false positives (FP):** We predicted yes, but they haven’t actually churned. (Also known as a "Type I error.")
* **false negatives (FN):** We predicted no, but they have actually churned. (Also known as a "Type II error.")

The confusion matrix shows an excellentresult,

* **Accuracy of 85.2%**
* **Sensitivity 87.25%**
* **Specificity 73.10%**

*As per the Confusion Matrix the model is able to Accurately Predict Churn 106 times out of 145. The model not only performs in predicting the Churn value of 0 but also makes good prediction with Churn value of 1.*

*Also, from the Business perspective, we are more worried of misclassification of 1s than 0s. The model has an overall good Accuracy, Sensitivity and Specificity which makes the model robust in accurately predicting the Results*

****

**Conclusion:**

***The model has the AUC of 86.6% and predict at most approx. 85.2% accuracy that whether a person is going to leave the network or stay which is the good balanced result based on the data.***