TELECOM CHURN CASE STUDY SUBMISSION

**NOTE:** This should briefly describe the important results and recommendations. The structure is suggestive; make sure to not exceed 7 pages**.**

# Checkpoint-1: Data Understanding and Preparation of Master File

3 Source File:

1. Customer: Contains 7043Records (Observations) with 5 Columns (Variables)
2. Churn: Contains 7043Records (Observations) with 9 Columns (Variables)
3. Internet: Contains 7043Records (Observations) with 9 Columns (Variables)

* Report the final number of rows and columns in the dataset.

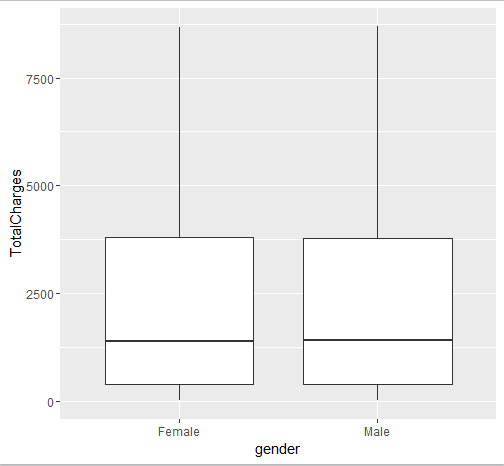
Merge File: Contains 7043Records (Observations) with 21 Columns (Variables)

After Feature Conversion of the variables (Categorical to Numerical conversion by Dummy Variables and removing unwanted Fields the below structure has been fed into the model)

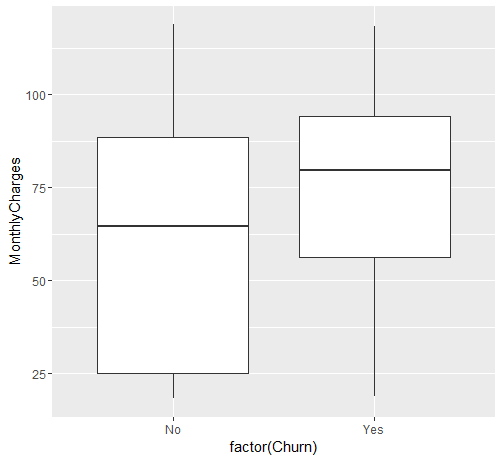
Final\_data: Contains 7043Records (Observations) with 30 Columns (Variables) where the Variable “Churn” is the target variable of factor type and remaining all are predictor variable.

# Checkpoint 2: Exploratory Data Analysis

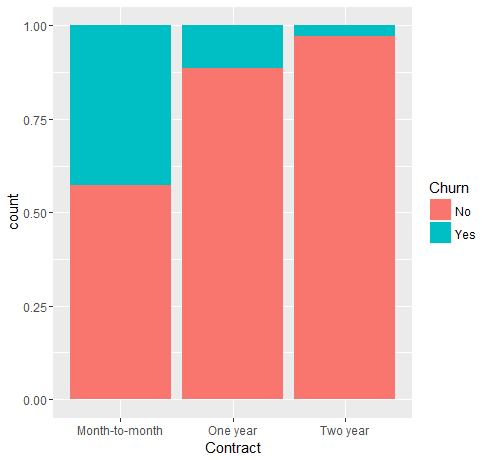
* Display the plots obtained and report the derivable insights.



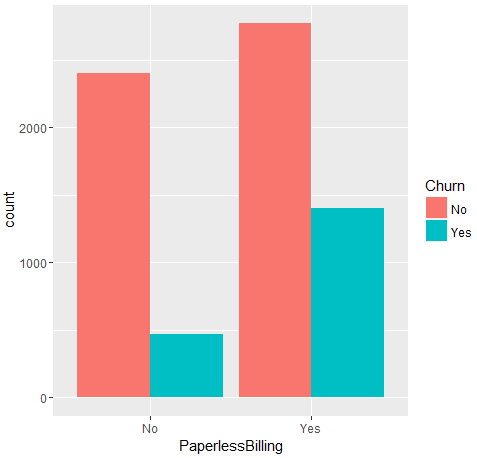
TOTAL Charges is similar across gender so there is no significant difference between male and female in Total Spending.



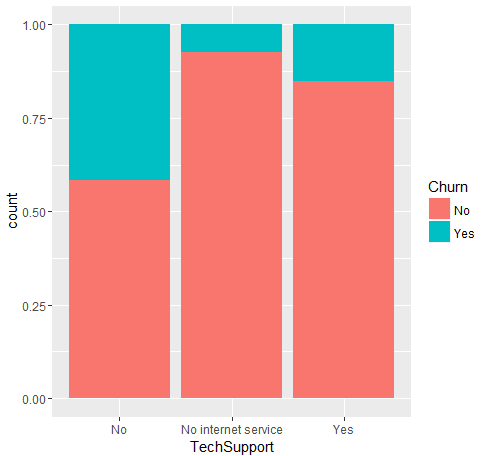
Churn rate is higher among those who are spending more on monthly basis



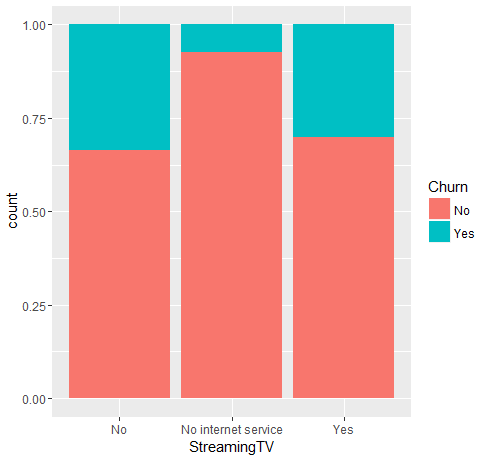
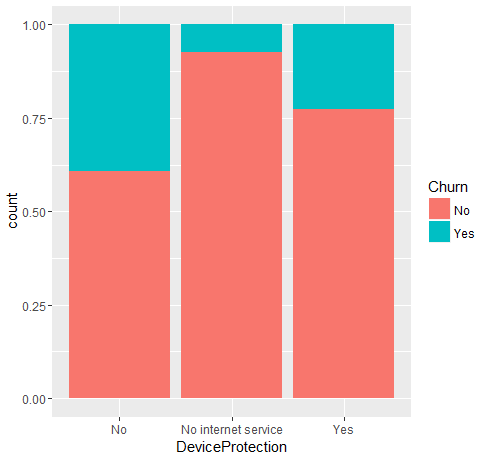
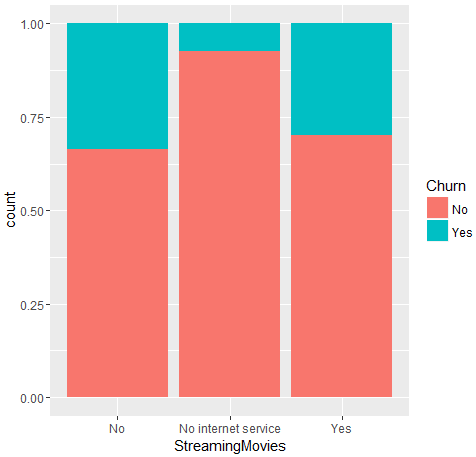
Churning tendency is higher for month to month Contract Customer rather than those who are in one year or two year contract.



Churning is little bit higher for Paperless Billing Customer



Churning is higher for customer without Tech support Service

StreamingTV, StreamingMovies, DeviceProtection, OnlineBackup etc services does not have any significant effect on Customer churning

# Checkpoint 3: Data Preparation

* Report the number of duplicated in the data. 🡪 **There is no Duplicate Record in the dataset.**
* Explain the methodology of Missing value treatment and additionally fill the below table:

**Explanation for Missing value Treatment:**

Presence of missing value has been identified using saaply function on each field of merge data frame and sum(is.na()) has been applied on each variables to find the missing values. It has been observed that missing value is present for one variable only named “TotalCharges”. Since it is a numeric field so missing values has been replaced by Mean of the remaining values for the same variable.

|  |  |
| --- | --- |
| **Questions** | **Results(Numeric)** |
| Total number of observations in the dataset | 7043 |
| Total number of variables in the dataset | 21 |
| Total missing values in the dataset | 11 |

* Explain the methodology of Outlier treatment and fill the below table:

**Explanation for outlier Treatment:**

Since the merge data frame is having only 3 numeric variables so outlier checkhas been performed on those 3variable only called tenture, monthlycharges and TotalCharges using boxplot.stat function. There was no outlier detected for these fields so no treatment has been done.

* Bring the data in the correct format. Report the number of variables for which the format was changed.

All factor type categorical variables for which more than 2 levels of distinct values are present , those have been converted to dummy variable using model.matrix command.

For Naïve-Bayes model , 3 numeric field has been converted to categorical values using binning method by cut function.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | The following Variable has been checked for outlier existence:   1. Tenure 2. MonthlyCharges 3. TotalCharges   No Variable is having outlier value. |
| Dummy creation | 1. Contract  2. PaymentMethod  3. MultipleLines  4. InternetService  5. OnlineSecurity  6. OnlineBackup  7. DeviceProtection  8. TechSupport  9. StreamingTV  10.StreamingMovies |
| Binning of variables | The following Variable has been converted to bins for Naïve Bayes Modelling:   1. Tenure 2. MonthlyCharges 3. TotalCharges |

# Checkpoint 4: Modelling

* **Model – K-NN**
  + Explain the Data Preparation step for K-NN modelling.

**Explanation:**Standard data preparation steps like below are taken:

1. All Factor type categorical variables having more than 2 levels has been converted to dummy variables.
2. All numeric variables has been standardised using scale function
3. All Factor type categorical variables with 2 levels has been converted to numeric values 0 and 1.
4. All dummy variables has been merged to form final dataset and original categorical field has been removed from final dataset.
   * Explain the methodology of building the model with optimal value of K?

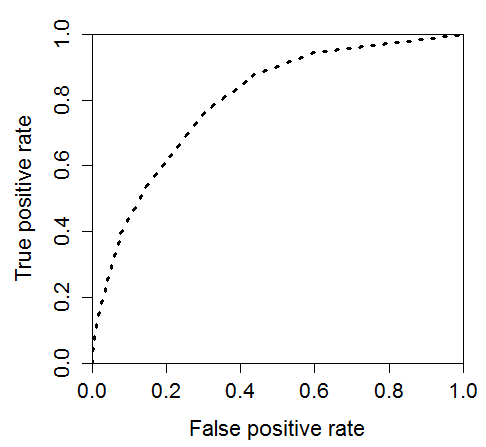
**Explanation:**

* The data has been distributed in two bucket; 70% kept as training data and 30% kept as test data using Sample function
* Knn function from class package has been executed for K=1,2,3 etc values.
* ConfusionMatrix function from caret package has been executed each time to get the various modelling statistics like accuracy, sensitivity and specificity
* To get optimal K, train function has been executed with K value range from 1 to 50 and the graph of the output of train function has been plotted using plot function to get optimal value of k as 9.
* Then final knn model has been executed with k=9 and all statistical parameter has been listed down in the below table.
* ROC curve has also been attached.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7714 |
| Sensitivity | 0.5188 |
| Specificity | 0.8623 |
| AUC | 0.7979 |

* + Display the AUC curve.



* **Model – Naïve Bayes**
  + Explain the Data Preparation step for **Naïve Bayes** modelling.

**Explanation:**Standard data preparation steps like below are taken:

1. Since Naïve Bayes modelling can be applied either on categorical or numerical variables ,since in the given data set most of the variables are categorical in nature so it has been decided to convert 3 numerical variables to categorical one using binning technique via cut function.
2. All categorical field has been converted factor type.
3. Target variable is also has been converted to factor type for this type of modelling.
   * Explain the methodology of building the model.

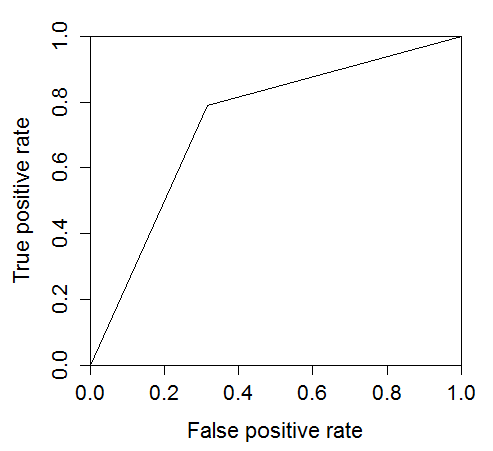
**Explanation:**

* The data has been distributed in two bucket; 70% kept as training data and 30% kept as test data using Sample function
* naiveBayes function from e1071 package has been executed for training data.
* ConfusionMatrix function from caret package has been executed to get the various modelling statistics like accuracy, sensitivity and specificity
* AUC, TPR, FPR has been calculated using prediction and performance function to get the AUC value and ROC curve.
* ROC curve has also been attached.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7123 |
| Sensitivity | 0.7907 |
| Specificity | 0.6840 |
| AUC | 0.7373 |

* + Display the ROC curve.



* **Model – Logistic Regression**
* Explain the methodology of building the model? In the final model, interpret what the coefficients of the variable imply. Check if the coefficients make business sense

**Explanation of methodology:**

* The data has been distributed in two bucket; 70% kept as training data and 30% kept as test data using Sample function
* glm function with family parameter as “binomial” is used to build the initial model.
* Then Step function has been used to reduce insignificant variables from the model and both forward and backward method has been used using the parameter direction = “both”.
* In 3rd stage VIF function from car package has been used to check multi-collinearity among the independent variable but all the variables with high VIF value is significant statistically so we can’t remove any variable based on the VIF value.
* Then based on P-value of the variable, one by one some variable has been removed based on their least significance level until we get a model with all highly significant variable.

**Interpretation of co-efficient:**

1. Some of the coefficients are +ve means those variables are proportionately related with the target variable and some of the coefficients are -ve means those variables are inversely related with the target variable.
2. As an example MonthlyCharges are inversely related with Churn which make business sense.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeric)** |
| (Intercept) | -2.3143 |
| tenure | -1.4529 |
| MonthlyCharges | -1.1501 |
| TotalCharges | 0.7574 |
| `ContractOne year` | -0.7567 |
| `ContractTwo year` | -1.6419 |
| `PaymentMethodElectronic check` | 0.3933 |
| MultipleLinesYes | 0.5073 |
| `InternetServiceFiber optic` | 1.7624 |
| InternetServiceNo | -1.7599 |
| StreamingTVYes | 0.5237 |
| StreamingMoviesYes | 0.6325 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 4240 |
| Null deviance | 5784 |
| Residual Deviance | 4216 |

* Calculate c-statistic and KS-statistic. What can you tell about the model based on their values?

Both the C and KS statistics are having similar values between training and test dataset. That signifies the model is robust and it can predict the values of unseen data pretty well.

Since KS statistics is greater than 0.5 and it exists in 4th decile for both training and testing data so that means the model has very good discriminative power as well.

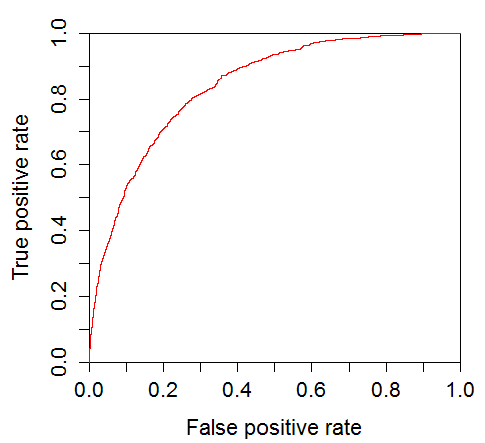
Additionally, fill the below tables:

**Note**: Write the numeric value of c-statistic and KS-statistic after applying your final model to the train dataset and test dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Train Dataset** | | **Test Dataset** | |
| C-statistic | 0.84 | C-statistic | 0.85 |
| KS-statistic | 0.52 | KS-statistic | 0.55 |
| Model Evaluation (write Accept or Reject) | | Accept | |

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.8159 |
| Sensitivity | 0.5701 |
| Specificity | 0.8964 |
| AUC | 0.8565 |

ROC Curve is as below:



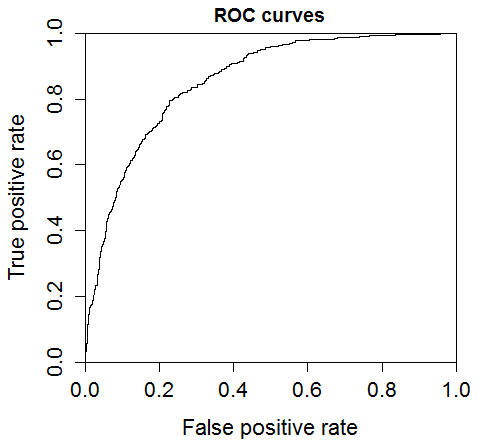
* **Model – SVM**
  + Explain the Data Preparation step for SVM modelling.

**Explanation:**Standard data preparation steps like below are taken:

1. Target variable (Dependent Variable) has been converted to factor type for SVM Modelling.
2. All numeric variables has been standardised using scale function
3. All Factor type categorical variables with 2 levels has been converted to numeric values 0 and 1.
4. All dummy variables has been merged to form final dataset and original categorical field has been removed from final dataset.
   * Explain the methodology of building the model.

**Explanation:**

* The data has been distributed in two bucket; 70% kept as training data and 30% kept as test data using Sample function
* SVM function from e1071 package has been executed for various cost parameter values like 0.1,100 etc.
* Since the data is already scaled as a part of data preparation so Scale parameter has been kept as False.
* Kernal parameter has been used as “Linear”
* To get optimal Cost parameter value : Tune function has been used to find out the best model for different values of Cost parameter; it shows best model has been given at Cost = 0.1
* ConfusionMatrix function from caret package has been executed each time to get the various modelling statistics like accuracy, sensitivity and specificity
* ROC curve has also been attached.



Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.8135 |
| Sensitivity | 0.5585 |
| Specificity | 0.8970 |
| AUC | 0.8544 |

Since the Best model has been given for Cost parameter = 0.1 so the statistical parameter values of best model will remain same as above.

* Report the best model and its performance metrics.

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.8135 |
| Sensitivity | 0.5585 |
| Specificity | 0.8970 |
| AUC | 0.8544 |

# Checkpoint 6: Threshold value

* Select an appropriate threshold value and calculate the confusion matrix and overall accuracy, sensitivity and specificity

Considering all four types of modelling technique and comparing the output statistics of all the four different model it has been proved that with this data set logistic regression model is the best fit as the modelling parameters give the highest accuracy.

Additionally, fill the below table:

So in the below table all parameters has been provided from the best model:-

|  |  |  |  |
| --- | --- | --- | --- |
| **Train Dataset** | | **Test Dataset** | |
| C-statistic | 0.84 | C-statistic | 0.85 |
| KS-statistic | 0.52 | KS-statistic | 0.55 |
| Model Evaluation (write Accept or Reject) | | Accept | |

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.8159 |
| Sensitivity | 0.5701 |
| Specificity | 0.8964 |
| AUC | 0.8565 |