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

Initial: Rows=7043 Col=22 Variables

Number of NA = 11 (Variable: TotalCharges) - The “NA” was removed from the analysis as it forms 0.3% of the total data.

Variable Transformation:

Scaling applied to variables : tenure, MonthlyCharges, TotalCharges

Dummy variables created for : PaymentMethod, InternetService

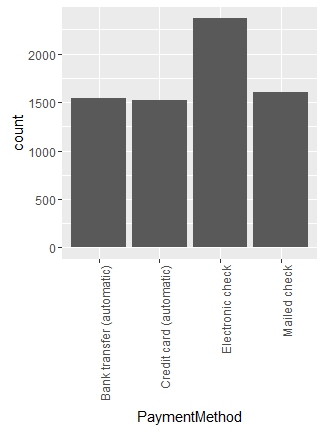
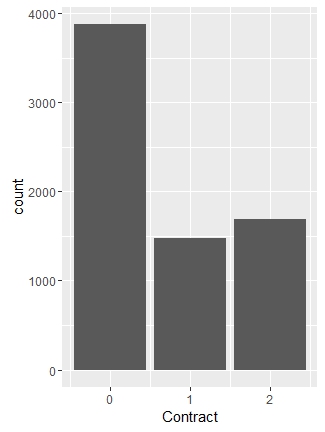
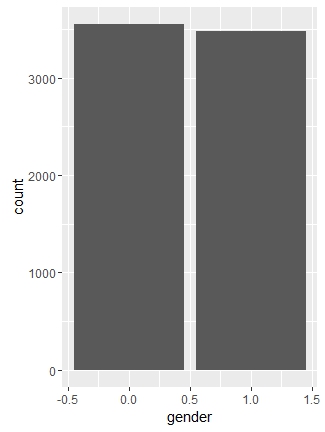
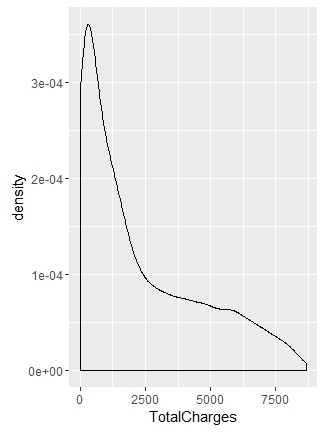
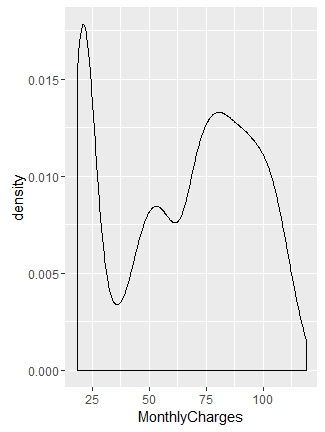
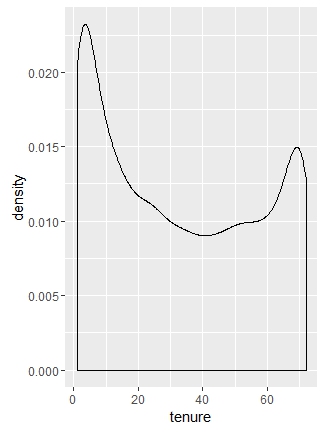
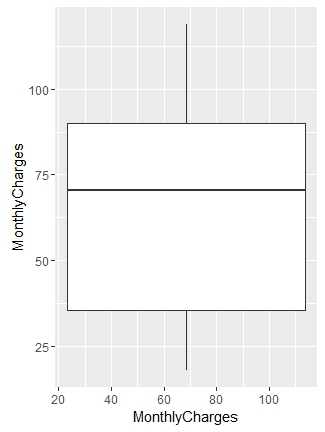
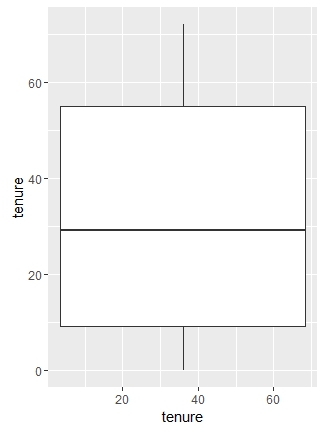
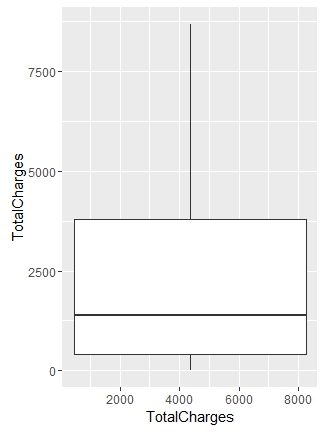
All other variables transformed to Numeric.

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

Final : Rows=7032 Col=25 Variables

# Checkpoint 2: Exploratory Data Analysis

* There are not outliers noted in the continuous variables , a few categorical variables also have been analysed.



# Checkpoint 3: Data Preparation

* Report the number of duplicated in the data.
* Explain the methodology of Missing value treatment and additionally fill the below table:

|  |  |
| --- | --- |
| **Questions** | **Results(Numeric)** |
| Total number of observations in the dataset | 7043 |
| Total number of variables in the dataset | 22 |
| Total missing values in the dataset | 11 |

Missing data has been removed from the analysis as it forms 0.3% of total data.

* Explain the methodology of Outlier treatment and fill the below table:
* Bring the data in the correct format. Report the number of variables for which the format was changed.

Additionally, fill the below table:

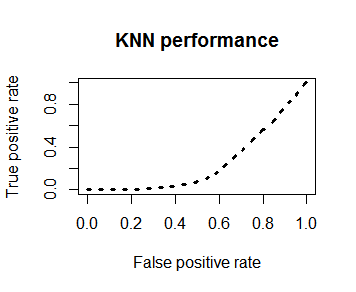
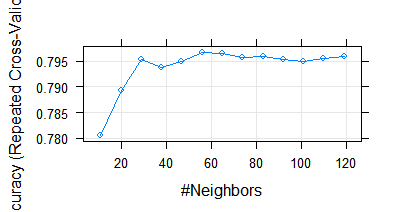
|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | Outliers were not found in the variables |
| Dummy creation | Dummy variables created for : PaymentMethod, InternetService |
| Binning of variables | NIL |

# Checkpoint 4: Modelling

* **Model – K-NN**
  + Explain the Data Preparation step for K-NN modelling. omit NA , check for outliers , scale variables to std unit ,
  + Explain the methodology of building the model with optimal value of K?   
    Find the optimal K by using *train* function from 11 to 121 steps 6

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7885 |
| Sensitivity | 0.5365 |
| Specificity | 0.8795 |
| AUC | 0.248 |

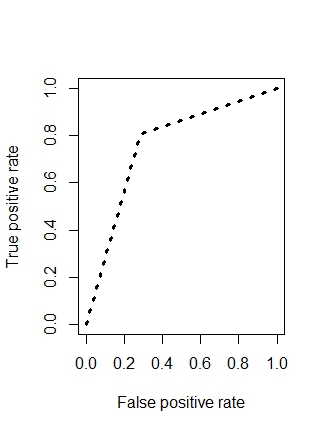


* + Display he AUC curve.
* **Model – Naïve Bayes**
  + Explain the Data Preparation step for Naive modelling.
  + Explain the methodology of building the model.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7412 |
| Sensitivity | 0.7174 |
| Specificity | 0.8045 |
| AUC | 0.7609306 |

* + Display he AUC 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

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeeric)** |
| (Intercept) | -2.81273 |
| `PaymentMethodElectronic check` | 0.36058 |
| `InternetServiceFiber optic` | 1.05078 |
| StreamingTV | 0.30355 |
| StreamingMovies | 0.42086 |
| Dependents | 0.37096 |
| Contract | -0.92897 |
| PaperlessBilling | 0.29985 |
| TotalCharges1 | -0.79789 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 4225.5 |
| Null deviance | 5664.6 |
| Residual Deviance | 4207.5 |

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

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 | 7.027808e-01 | C-statistic | 7.116373e-01 |
| KS-statistic | 0.4055616 | KS-statistic | 0.4232746 |
| Model Evaluation (write Accept or Reject) | | Accept | |

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7972 |
| Sensitivity | 0.9008 |
| Specificity | 0.5225 |
| AUC | 0.7116373 |

* **Model – SVM**
  + Explain the Data Preparation step for SVM modelling.
  + Explain the methodology of building the model.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.8085 |
| Sensitivity | 0.9040 |
| Specificity | 0.5554 |
| AUC |  |

* Report the best model and its performance metrics.

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy |  |
| Sensitivity |  |
| Specificity |  |
| AUC |  |

# Checkpoint 6: Threshold value

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

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy |  |
| Sensitivity |  |
| Specificity |  |