TELECOM CHURN CASE STUDY SUBMISSION

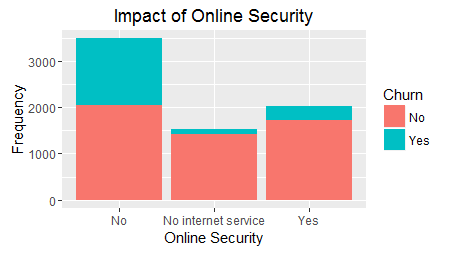
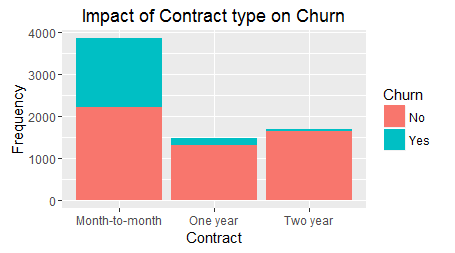
**NOTE:** This should briefly describe the important results and recommendations. The structure is suggestive; make sure to not exceed 7 pages**.**

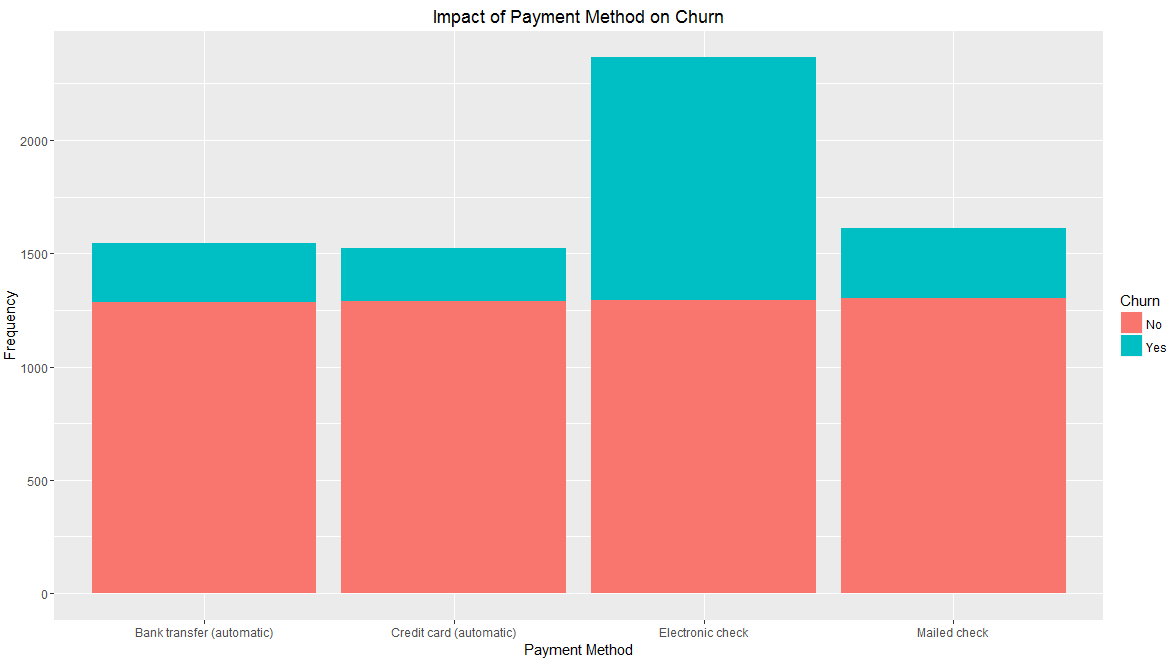
# Checkpoint-1: Data Understanding and Preparation of Master File

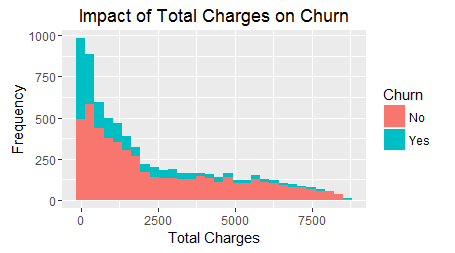
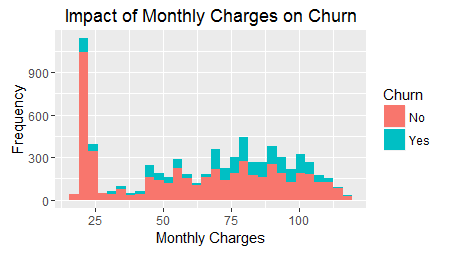
* Report the final number of rows and columns in the dataset.
  + Final churn dataframe has 7043 rows and 21 columns

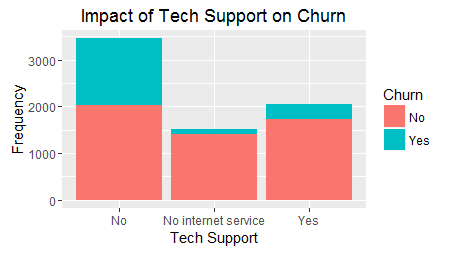
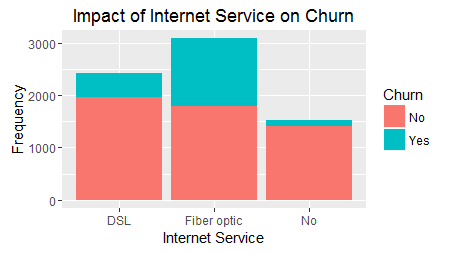
# Checkpoint 2: Exploratory Data Analysis

* Display the plots obtained and report the derivable insights.









# Checkpoint 3: Data Preparation

* Report the number of duplicated in the data.
  + No duplicate records were found in the dataset. Duplicates were checked using the duplicated function
* Explain the methodology of Missing value treatment and additionally fill the below table:
  + 11 Missing values were found in the TotalCharges Column of the dataset.
  + Since the no of missing values were small with respect to the no or rows in the dataset, these 11 records were removed from the dataset

|  |  |
| --- | --- |
| **Questions** | **Results(Numeric)** |
| Total number of observations in the dataset | 7032(After removal of Missing values) |
| 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:
  + No Outliers were found in the Numeric varaibles namely Tenure, MonthlyCharges and TotalCharges
* Bring the data in the correct format. Report the number of variables for which the format was changed.
  + Tenure was binned into 6 categorical levels namely
    - 1 – 0-12 months
    - 2 – 13-24 months
    - 3 – 25-36 months
    - 4 – 37-48 months
    - 5 – 49-60 months
    - 6 – 61-72 months

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | None |
| Dummy creation | PhoneService, Contract, PaperlessBilling, PaymentMethod, Gender, SeniorCitizen, Partner,Dependents, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, tenureYears |
| Binning of variables | Tenure was binned into 6 categorical levels and the binned variable was names tenureYears |

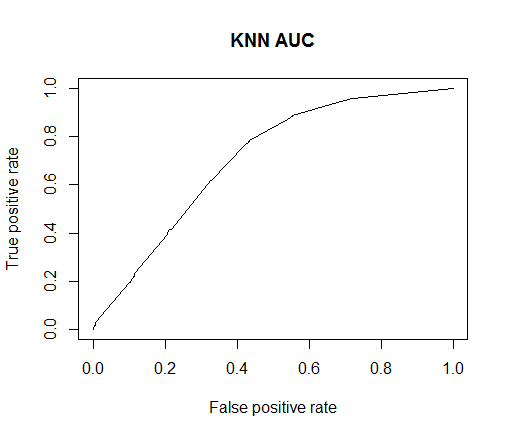
# Checkpoint 4: Modelling

* **Model – K-NN**
  + Explain the Data Preparation step for K-NN modelling.
    - All factor variables were converted to dummy variables
    - All Numeric variables were scaled.
    - Dataset was split into train and test data using the sample.split function to ensure that equal proportions of both the types(churn = yes and churn=no) are present in both the datasets
  + Explain the methodology of building the model with optimal value of K?
    - Cross-validation was performed to find the optimum value of k. the optimum value of k was found to be 10. Which is in accordance with the norms.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7867 |
| Sensitivity | 0.5740 |
| Specificity | 0.8638 |
| AUC | 0.70 |

* + Display the AUC curve.

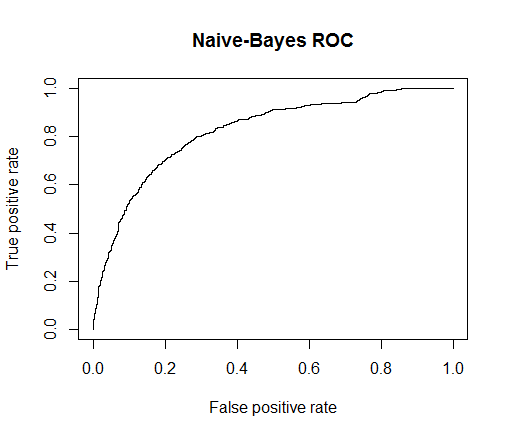


* **Model – Naïve Bayes**
  + Explain the Data Preparation step for Naïve Bayes modelling.
    - All factor variables were converted to dummy variables
    - All Numeric variables were scaled.
    - Dataset was split into train and test data using the sample.split function to ensure that equal proportions of both the types(churn = yes and churn=no) are present in both the datasets.
  + Explain the methodology of building the model.
    - Model was built using the naiveBayes function in the e1071 package

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7009 |
| Sensitivity | 0.8396 |
| Specificity | 0.6507 |
| AUC | 0.8224 |

* + Display the 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
    - All factor variables were converted to dummy variables
    - All Numeric variables were scaled.
    - Dataset was split into train and test data using the sample.split function to ensure that equal proportions of both the types(churn = yes and churn=no) are present in both the datasets

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeeric)** |
| `ContractMonth-to-month` | 2.17945 |
| `ContractOne year` | 1.15240 |
| PaperlessBillingNo | -0.31167 |
| `PaymentMethodElectronic check` | 0.36409 |
| SeniorCitizen0 | -0.20769 |
| MultipleLinesNo | -0.31296 |
| InternetServiceDSL | 1.07942 |
| `InternetServiceFiber optic` | 1.94464 |
| StreamingTVNo | -0.21082 |
| StreamingMoviesNo | -0.24313 |
| tenureYears1 | 1.60239 |
| tenureYears2 | 0.59685 |
| tenureYears4 | 0.46424 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 4173.6 |
| Null deviance | 5699.5 |
| Residual Deviance | 4145.6 |

* Calculate c-statistic and KS-statistic. What can you tell about the model based on their values?
  + The c-statistic for train dataset is 0.84 and that of test dataset is 0.84. This suggests that the model has a very good discriminative power.
  + The Ks statistic for train data is 0.53 and that of test data is 0.52. This is within the acceptable range of 0.4 to 0.7. Furthermore both the Ks statistic lie in the 1st decile of the train and test dataset which suggests that the model has a very high discriminative power.

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 | 0.8407 | C-statistic | 0.8402 |
| KS-statistic | 0.5323 | KS-statistic | 0.5273 |
| Model Evaluation (write Accept or Reject) | | Accept | |

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7479 |
| Sensitivity | 0.7790 |
| Specificity | 0.7366 |
| AUC | 0.8402 |

* **Model – SVM**
  + Explain the Data Preparation step for SVM modelling.
    - All factor variables were converted to dummy variables
    - All Numeric variables were scaled.
    - Dataset was split into train and test data using the sample.split function to ensure that equal proportions of both the types(churn = yes and churn=no) are present in both the datasets
  + Explain the methodology of building the model.
    - The initial model was built using the svm function with the parameters of cost as 0.1 and linear kernel and scale set to False
    - The model was further tuned using the tune function with the costs as 0.001, 0.01, 0.1, 0.5, 1, 10, 100.
    - Finally, the optimum value was cost was found to be 0.01

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.8033 |
| Sensitivity | 0.4545 |
| Specificity | 0.9296 |
| AUC | 0.7051 |

* Report the best model and its performance metrics.
  + The best model was found to be the model obtained by the logistic regression. The performance metric are balanced and does not overfit or underfit the data.

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7479 |
| Sensitivity | 0.7790 |
| Specificity | 0.7366 |
| AUC | 0.8402 |

# Checkpoint 6: Threshold value

* Select an appropriate threshold value and calculate the confusion matrix and overall accuracy, sensitivity and specificity
  + Appropriate threshold value was identified by making use of the roc function from the pROC package.
  + The threshold was identified as **0.294**

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7479 |
| Sensitivity | 0.7790 |
| Specificity | 0.7366 |