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

* *The churn data has 7043 observations(rows) and 5 variables(columns)*
* *Number of NA = 11 (Variable: TotalCharges) - The “NA” was removed from the analysis as it forms only 0.3% of the total data.*
* *After merging of data there were 22 duplicated rows (not taking into account the customer ID). These were also removed from the dataset.*

Variable Transformation:

* *Scaling applied to variables: Tenure, MonthlyCharges, TotalCharges*
* *Dummy variables created for (16 variables):* 
  + - * + *gender,*
        + *SeniorCitizen,*
        + *Partner,*
        + *Dependents,*
        + *PhoneService,*
        + *Contract,*
        + *PaperlessBilling,*
        + *PaymentMethod,*
        + *MultipleLines,*
        + *InternetService,*
        + *OnlineSecurity,*
        + *OnlineBackup,*
        + *DeviceProtection,*
        + *TechSupport,*
        + *StreamingTV,*
        + *StreamingMovies*
* *All other variables transformed to Numeric.*
* *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 | 21 |
| Total missing values in the dataset | 11 |

* + *Missing data has been removed from the analysis as it only forms only 0.3% of total data.*
* Explain the methodology of Outlier treatment and fill the below table:
  + *Outlier determination was done through Boxplot, Boxplot.stats and quartile functions in R*
  + *No outliers where found in the dataset.*
* Bring the data in the correct format. Report the number of variables for which the format was changed.
  + *Format for SeniorCitizen was converted from numerical to binary (yes/no)*

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | Outliers were not found in the variables |
| Dummy creation | gender,SeniorCitizen,Partner ,Dependents,PhoneService, Contract,PaperlessBilling,PaymentMethod, MultipleLines,InternetService,OnlineSecurity, OnlineBackup,DeviceProtection,TechSupport, StreamingTV,StreamingMovies |
| Binning of variables | TotalCharges, MonthlyCharges, Tenure  *\*\* Binning was done for all continuous variables only for the Naïve Bayesian model* |

* + *Dummy variables were created for all categorical variables.*
  + *The dummy variables were created also for Boolean [Yes/No] as an easier method to convert to numeric 1/0*
  + *Binning was done for all continuous variables only for the Naïve Bayesian model*

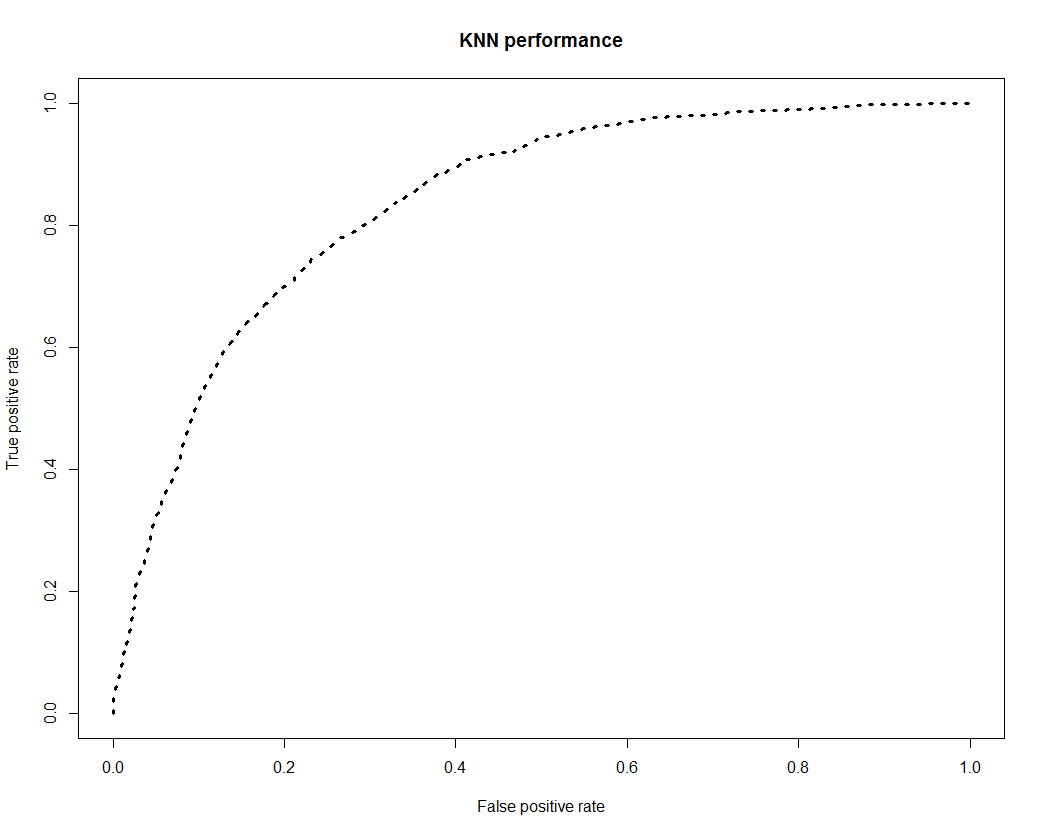
# Checkpoint 4: Modelling

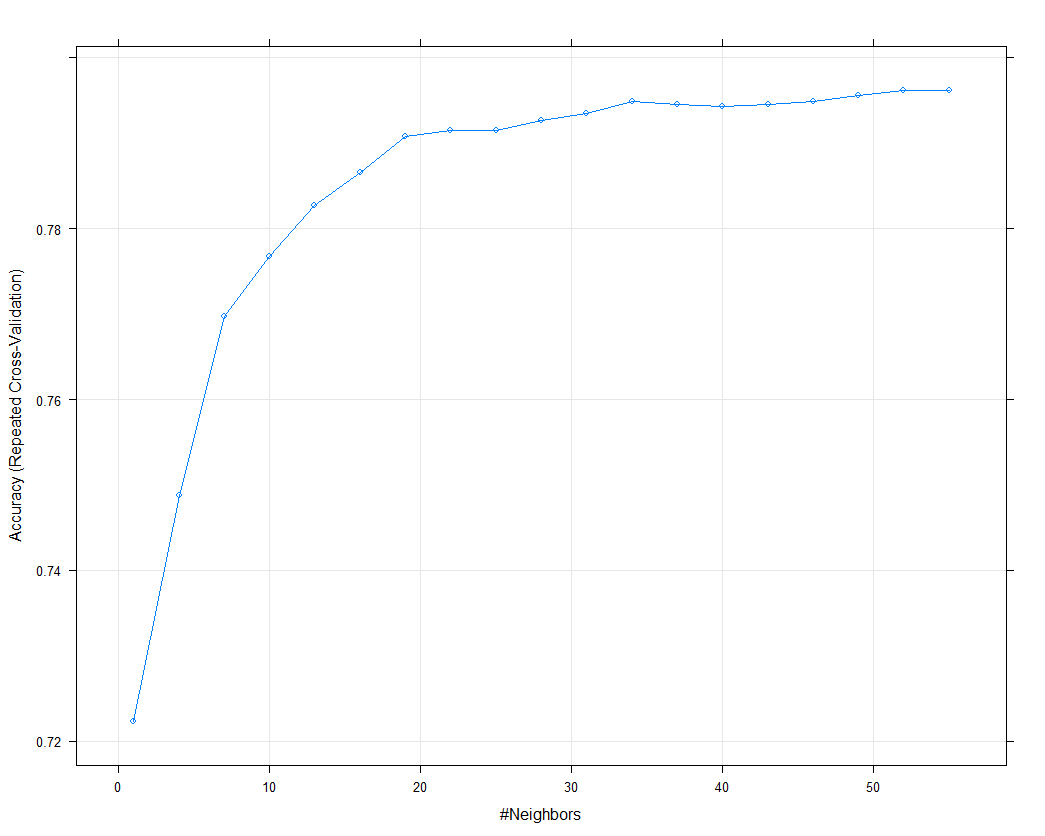
* **Model – K-NN**
  + Explain the Data Preparation step for K-NN modelling.
    - *Omit NA*
    - *Check for outliers*
    - *Scale variables to std. unit*
    - *Split data into training(70%) and test (30%) data*
  + Explain the methodology of building the model with optimal value of K?
    - *Find the optimal K by using train function from 1 to 55 steps3*
    - *Evaluated the model with 30% (test) of the data.*

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.796 |
| Sensitivity | 0.5745 |
| Specificity | 0.8758 |
| AUC | 0.8359187 |
| 'Positive' Class | Yes |

* + Display the AUC curve



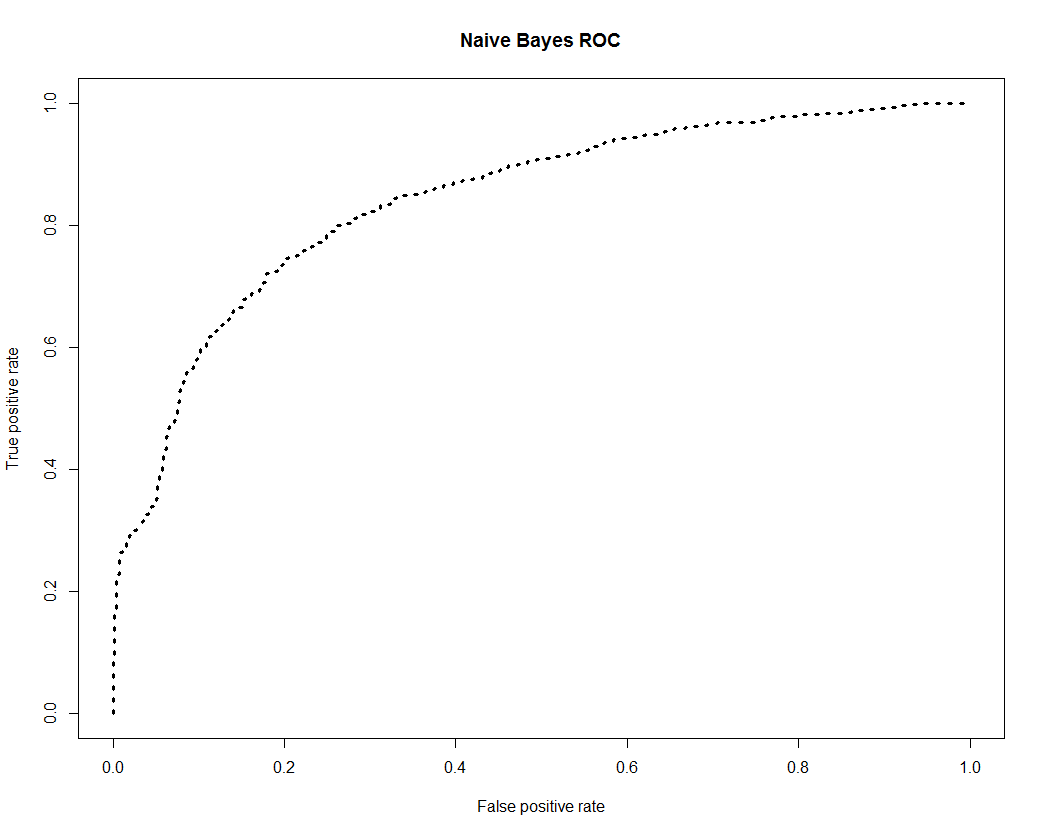


* **Model – Naïve Bayes**
  + Explain the Data Preparation step for Naive modelling.
    - *omit NA*
    - *check for outliers*
    - *bin the continuous variables*
    - *split data into training(70%) and test(30%) sets*
  + Explain the methodology of building the model.
    - *Using {e1071} package naiveBayes function to determine the model for the training data.*
    - *Validate the model with the test data*

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7356 |
| Sensitivity | 0.7038 |
| Specificity | 0.8241 |
| 'Positive' Class | No |
| AUC | 0.8409598 |

* + Display the AUC curve.



* **Model – Logistic Regression**
  + Explain the methodology of building the model?
    - *Create an initial model using the glm function of {stats} package.*
    - *Refine the model programmatically using step function of {stats} library in “both” directions. This will give us the best model with the lowest AIC values.*
    - *We further refine the model manually by removing features that have high VIF and PValues, till we have a model with mainly significant attributes and VIF values lower than 2.*
  + In the final model, interpret what the coefficients of the variable imply.
    - *The lower the coefficients the more significant are the attributes in the model.*
  + Check if the coefficients make business sense
    - *With any mobile contract the prevailing idea is customers look for how long the user has been with the provider (tenure), how long is the contract period and what the monthly charges are. These as per our data are the significant values and the rest have lesser significance which makes complete business sense.*

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeric)** |
| (Intercept) | < 2e-16 |
| Tenure | < 2e-16 |
| MonthlyCharges | < 2e-16 |
| ContractOne.year | 9.34e-14 |
| ContractTwo.year | < 2e-16 |
| PaperlessBillingYes | 8.84e-07 |
| PaymentMethodElectronic.check | 1.06e-09 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 4216.7 |
| Null deviance | 5673.9 |
| Residual Deviance | 4202.7 |

* Calculate c-statistic and KS-statistic. What can you tell about the model based on their values?
  + *C-Statistics: Used to compare the goodness of fit of logistic regression models, 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.* ***As our model has a value above 0.8 this is a strong model for predicting churns***
  + *K-Statistics: The higher the value of KS statistic, the better the model is at separating the positives from the negatives.* ***As our K-Stats values are very similar for test and training this is a fairly strong model for Churn detection****.*

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 | 8.323861e -01 | C-statistic | 8.338867e-01 |
| KS-statistic | 0.5165529 | KS-statistic | 0.5324693 |
| Model Evaluation (write Accept or Reject) | | Accept | |

* + - *Threshold value set at > 0.5*

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7908 |
| Sensitivity | 0.5045 |
| Specificity | 0.8939 |
| AUC | 0.8450588 |

* **Model – SVM**
  + Explain the Data Preparation step for SVM modelling.
    - *omit NA*
    - *check for outliers*
    - *bin the continuous variables*
    - *split data into training(70%) and test(30%) sets*
  + Explain the methodology of building the model.
    - *We have used the tune function on a set of cost values to get the optimal cost value model.*
    - *We have evaluated the model with the test data.*

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7989 |
| Sensitivity | 0.8946 |
| Specificity | 0.5332 |
| 'Positive' Class | No |
| AUC | 0.8378975 |

* Report the best model and its performance metrics.
  + - ***SVM model has the best performance metrics***

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7989 |
| Sensitivity | 0.8946 |
| Specificity | 0.5332 |
| 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:

* + - ***We have used 0.5 as threshold values***

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7908 |
| Sensitivity | 0.5045 |
| Specificity | 0.8939 |