TELECOM CHURN CASE STUDY SUBMISSION

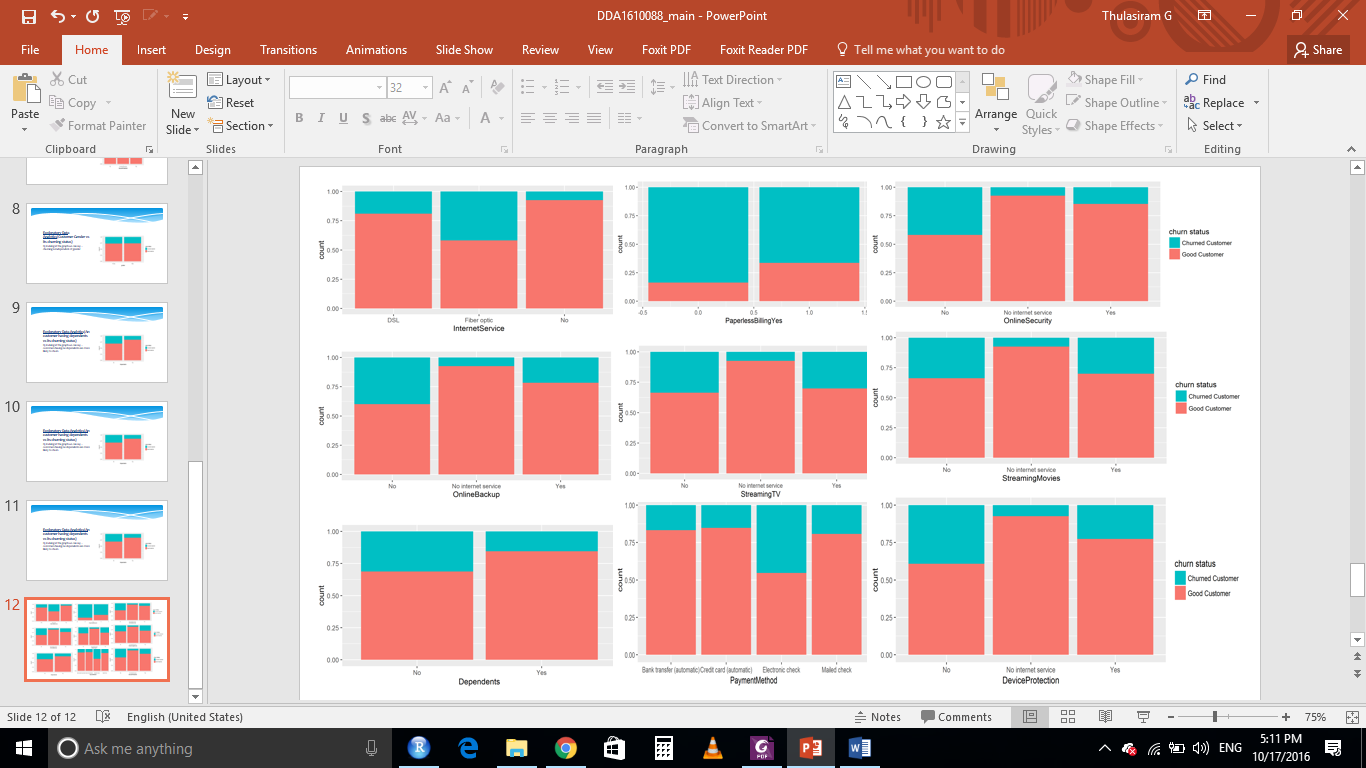
# Checkpoint-1: Data Understanding and Preparation of Master File

* There are three files named **Churn\_Data, Internet\_Data** and **Customer\_Data** for analysis
* Merged all the three data sets on **customer id** which is common across all the three files
* We named merged file **Churn**, below table has the details on merged file no of attributes and rows

|  |  |
| --- | --- |
| Total No. of Attributes | 21 |
| Total No. of rows | 7043 |

# Checkpoint 2: Exploratory Data Analysis

**Various variables and its impact on churning**



**Insights from above graph**

* Monthly plan customers are more likely to churn
* Person having no dependents are churning more
* No device protecting - more likely to churn
* Whoever using fibre optic internet services, they are more likely to churn
* Person having no online backup, online security is more likely to churn
* Unmarried people are churning more as compare to married people
* Percentage of churning on senior citizen is higher
* No tech support, Streaming movies, Streaming TV more likely to churn

# Checkpoint 3: Data Preparation

* Number of duplicated in the data - NIL
* Missing Value treatment –
  + Found missing value in Total Charge column.
  + Only 11 rows were getting affected out of 7043 rows which is almost 1 percent.
  + Moreover all these rows belong to those customers who are not getting churned.
  + As the no of rows are minimal and of those customers who are not getting churned, so all 11 observations were removed from the dataset.

|  |  |
| --- | --- |
| **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 |

* Outlier identification and its treatment - **Not required**
* We considered three columns for outlier identifications
* Following is the name of those columns – Tenure, Monthly charges and Total charges
* We plotted box plot to see the outliers
* No outliers have been observed in those columns

Below is the summary on outlier treatment, Dummy variable creation and binning of variables –

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | Not required |
| Dummy creation | Contract, Dependents, Device protection, gender, internet service, Multiple lines, online backup, online security, partner, payment method, phone service, senior citizen, streaming movies, streaming TV, Tech support, Tech charges |
| Binning of variables | Not required |

# Checkpoint 4: Modelling

**KNN**

**Model Development Steps –**

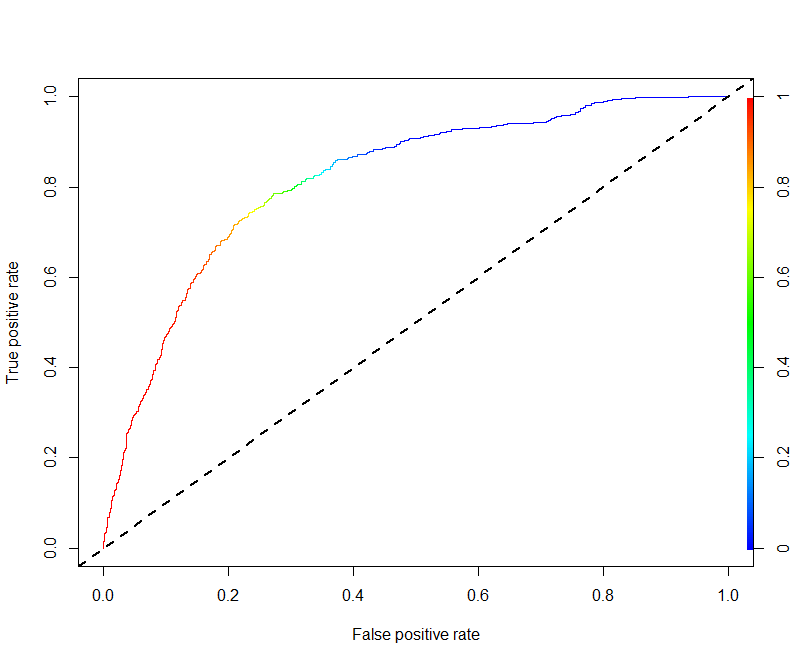
* Categorical variables were converted to numeric by creating dummy variables
* Converted the data type of response variable to factor
* Scaled the numerical variables
* Split the data into train and test data in the proportion of 70 and 30 percent ensuring equal proportion of class labels in both train and test data set
* A grid of 1 to 50 were used to find the optimal value of k
* Repeated cross validation is selected with 10 folds and 15 repetitions
* After repeated cross validation, model suggested 31 as optimal value of K

.

**Model evaluation parameters**

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7957 |
| Sensitivity | 0.5579 |
| Specificity | 0.8819 |
| AUC | 83.33% |

**AUC Curve**

****

**Naïve Bayes**

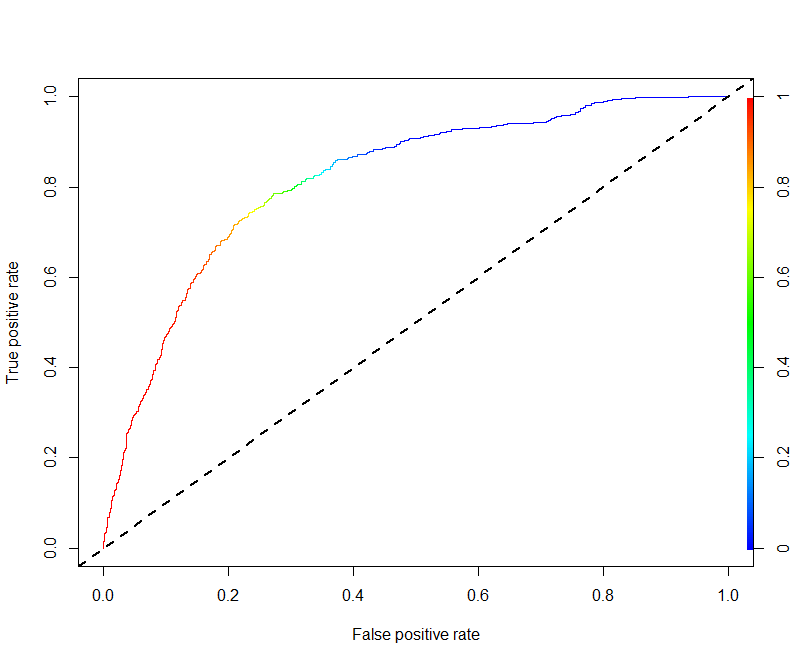
**Model development steps –**

* Split the data into train and test data in the proportion of 70 and 30 percent ensuring equal proportion of class labels in both train and test data set
* Neither scaling of numerical variables nor creation of dummy variable for categorical attributes were performed
* Train data set was used to create model
* Test data was used to evaluate the model

**Model evaluation parameters –**

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 72.59% |
| Sensitivity | 79.47% |
| Specificity | 69.87% |
| AUC | 81.61 |

**AUC Curve**



**Logistic Regression**

**Model development steps –**

* Categorical variables were converted to numeric by creating dummy variables
* Converted the data type of response variable to factor
* Scaled the numerical variables
* Split the data into train and test data in the proportion of 70 and 30 percent ensuring equal proportion of class labels in both train and test data set
* Initially a model was built with all the variables
* Then using stepwise method all insignificant variables were removed.
* Variables were checked for multicollinearity
* Variables with high VIF and less significance was removed
* KS statistics and C-statistic were computed for train and test dataset

**Model Result**

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeric)** |
| Tenure | -0.86634 |
| Senor citizen - Yes | 0.19940 |
| PhoneService-Yes | -0.47264 |
| Contract – One year | -0.79978 |
| Contract – Two year | -1.59162 |
| PaperlessBillingYes | 0.33567 |
| PaymentMethodElectronic check | 0.36067 |
| MultipleLinesYes | 0.32190 |
| `InternetServiceFiber optic` | 0.97454 |
| InternetServiceNo | -0.60998 |
| StreamingMoviesYes | 0.34022 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 4178.4 |
| Null deviance | 5699.5 |
| Residual Deviance | 4154.4 |

**Model Evaluation Parameter**

|  |  |  |  |
| --- | --- | --- | --- |
| **Train Dataset** | | **Test Dataset** | |
| C-statistic | 0.840 | C-statistic | 0.846 |
| KS-statistic | 0.52 | KS-statistic | 0.55 |
| Model Evaluation (write Accept or Reject) | | Accepted | |

|  |  |
| --- | --- |
| **Threshold value (0.3)** | **Values (Numeric)** |
| Overall Accuracy | 76.97% |
| Sensitivity | 78.25% |
| Specificity | 76.50% |
| AUC | 84.6% |

**SVM**

**Model development steps**

* Categorical variables were converted to numeric by creating dummy variables
* Converted the data type of response variable to factor
* Scaled the numerical variables
* Split the data into train and test data in the proportion of 70 and 30 percent ensuring equal proportion of class labels in both train and test data set
* Linear kernel is selected and model is tuned with various cost parameters
* The best cost value is taken and model is created from it
* Various evaluation metrics like accuracy, sensitivity and specificity for the model is checked
* Then radial kernel is selected and model is tuned with various parameters of cost and sigma
* The best cost and sigma values are taken and model was created using it

**Model evaluation parameters –**

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 80.33% |
| Sensitivity | 53.48% |
| Specificity | 90.06% |
| AUC | 84.26 |

**AUC Curve**

