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**.**

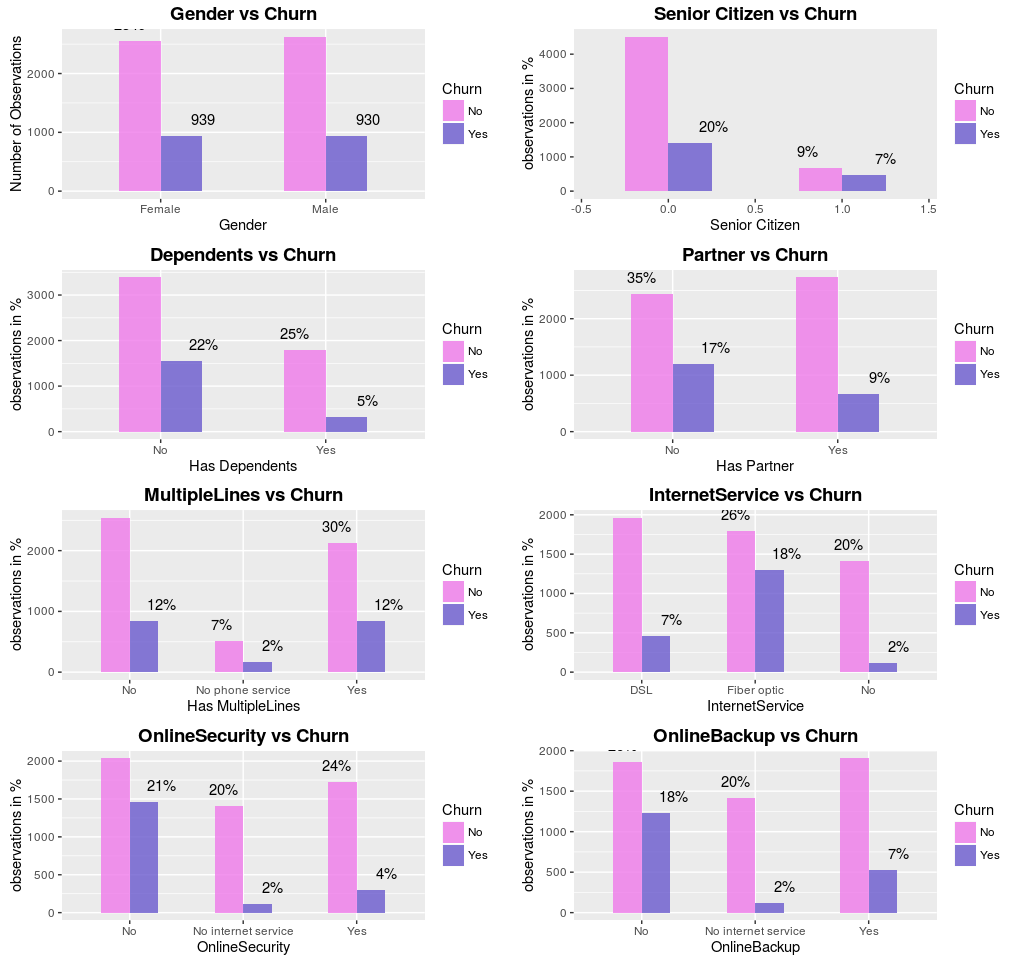
# Checkpoint-1: Data Understanding and Preparation of Master File

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

|  |  |  |
| --- | --- | --- |
| DataFile | No: of Rows | No: of columns |
| Internet\_data.csv | 7043 | 9 |
| Churn\_data.csv | 7043 | 9 |
| Customer\_data.csv | 7043 | 5 |

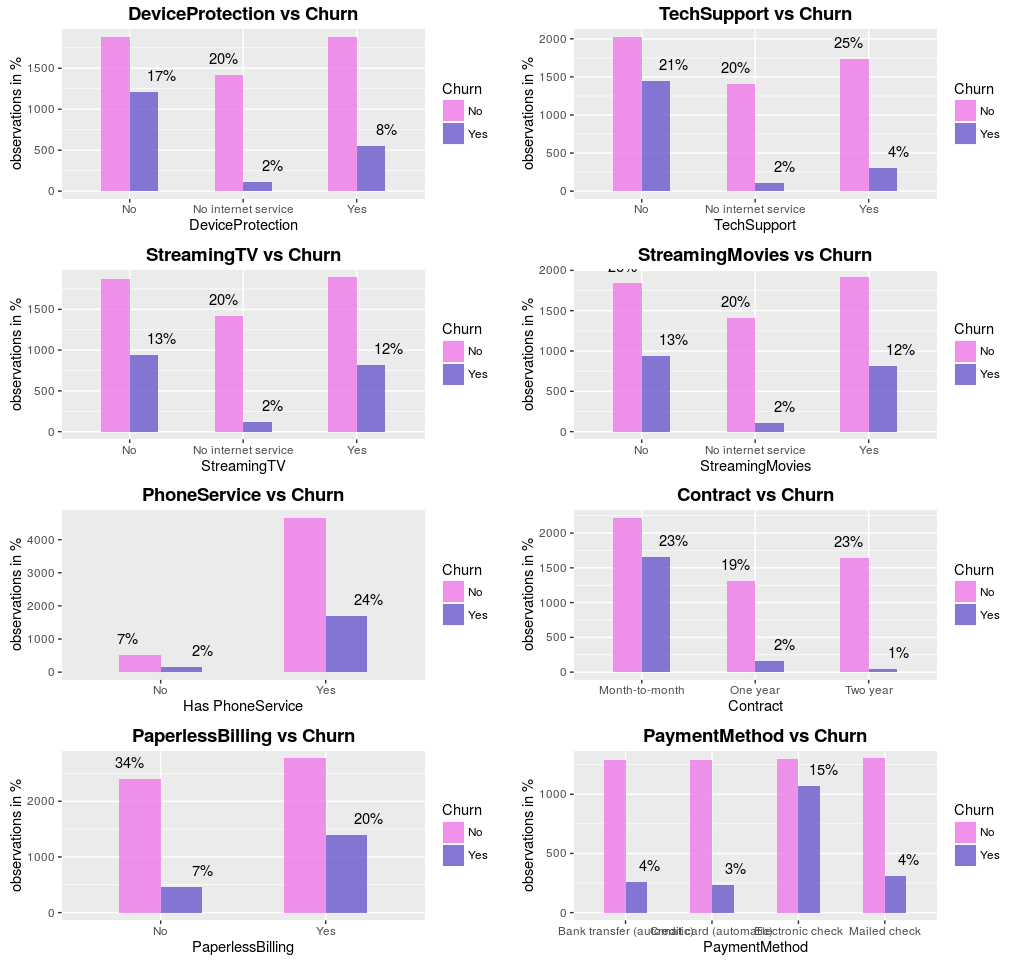
After merging of the 3 datasets into one dataset based on CustomerID column, the final data set (churn) has **7043 rows and 21 columns**.

# Checkpoint 2: Exploratory Data Analysis

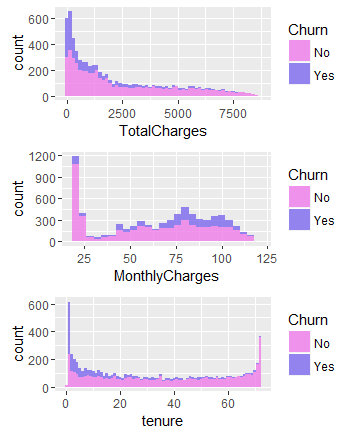


* Churn is around 45% among 16% of Senior Citizens. Senior Citizen have higher tendency to churn.
* Churn Rate is higher for those, who do not have a partner
* Lower Churn Rate for those with dependents
* 87% of the people have Phone Service.
* Slightly higher churn rate among those with no phone service
* Slightly Higher churn rate among those with Multiple Lines.

* Higher Churn Rate for those who have Opted for Fibre Optic, No Device Protection or No Online Backup or Security
* Churn Rate for Those with No Internet Service is 45%



* People who opted out of Tech Support are likely to Churn
* Those who opted out of Streaming Movies are more likely to Churn
* Highest Churn Rate for Month-to-Month Contract
* ~35% of the People who opted for Paperless Billing, churn.
* ~50% churn Rate for Electronic Check



* People who pay higher Monthly Charges, tend to churn more
* People who churn, pay lower Total Charges.
* Lower Tenure Churn is higher

# Checkpoint 3: Data Preparation

* Report the number of duplicated in the data.

No duplicate rows in the data.

However, after creation of dummy variable some of the columns had same data. Those columns were removed. Below are the columns which had same data.

1. InternetService.xNo
2. DeviceProtection.xNo.internet.service
3. OnlineSecurity.xNo.internet.service
4. OnlineBackup.xNo.internet.service
5. TechSupport.xNo.internet.service
6. StreamingTV.xNo.internet.service
7. StreamingMovies.xNo.internet.service

InternetService.xNo variable was retained and rest of the variables were removed.

PhoneService variable was also removed because PhoneService Represents Users who have Phone service, and MultipleLines.x.NoService represents users with No Phone Service. That is they represent the same thing, whether a user has or does not have phone service.

* Missing Value Treatment

Missing values were present in the Totalcharges variable. These missing values were imputed with zero, as all the users whose Total Charges were missing, had tenure zero which indicates users who have recently joined.

|  |  |
| --- | --- |
| **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 Treatment

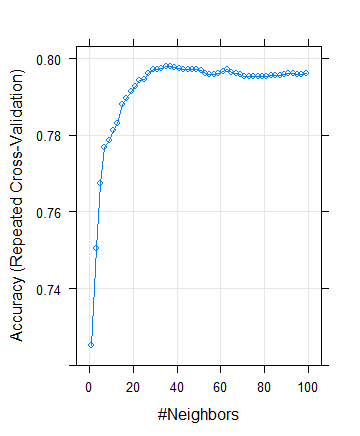
There were no outliers in the dataset. This was checked via boxplot.stats function.

* Data Type Formatting

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | None |
| Dummy creation | Gender  Partner  Dependents  PhoneService  MultipleLines  InternetService  OnlineSecurity  OnlineBackup  DeviceProtection  TechSupport  StreamingTV  StreamingMovies  Contract  PaperlessBilling  PaymentMethod  Churn |
| Binning of variables | None |

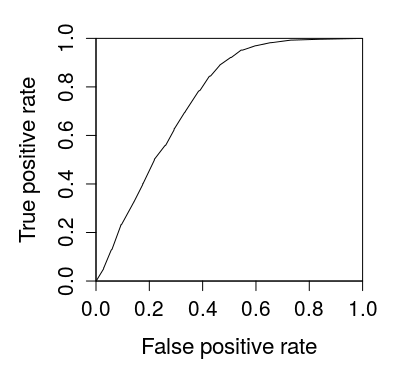
# Checkpoint 4: Modelling

* **Model – K-NN**
  + Cross Validation is used to identify the optimal k value.
  + The optimal k is found to be 37



**Figure : Optimal K Using Cross Validation**

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | **0.7965** |
| Sensitivity | **0.5383** |
| Specificity | **0.8898** |
| AUC | **0.7512381** |



**Figure : AUC Curve KNN**

* **Model – Naïve Bayes**

Naïve Bayes model was built by the use of naivebayes function which is present in the R package e1071. Apply the naivebayes function on the train dataset. This will give the naivebayes model.

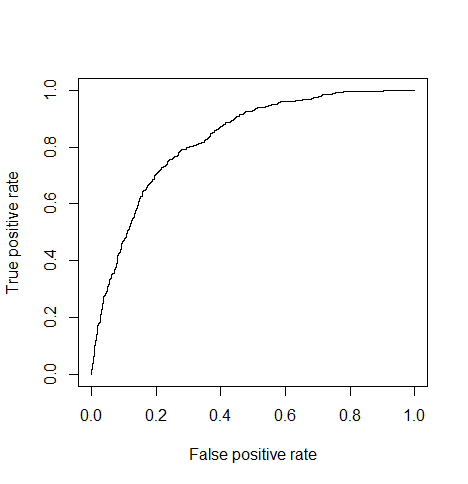
Using this model, we are predicting the values for test dataset where response/target is not present. This result was put in a dataframe pred variable. The probability of the Churning, churn values from test were taken for prediction and then performance plot of TPR vs FPR was done.

Area under the curve was calculated. AUC value : **0.826**

Naïve Bayes, Gives a good Sensitivyt (>70%) for all Threshold, but we have selected the Threshold as 0.3 as we want to minimise False Negative Rate, that is we want to identify all potential Churners, so its OK if a few Non-Churners are identified as potential churners. This is because Customer Retention is much cheaper than Customer Acquistion.

**Threshold value: 0.3**

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7331 |
| Sensitivity | 0.7914 |
| Specificity | 0.7120 |
| AUC | 0.8270526 |

Figure : Naïve Bayes AUC Curve

* **Model – Logistic Regression**

Post data preparation, Logistic Regression model building was done. Glm function was used to build the initial model. After that, stepwise logistic regression method was adopted for getting the optimal model.

Further, the variables were removed based on their insignificance level (high p-value) and high VIF (multicollinearity).

VIF value threshold was kept below 2.

Once we have reached at a model where all the variables have high significance (low p-value) and the VIF value less than 2, we can stop. This model can be concluded to be the final model.

The variables with coefficients with negative sign infront are the ones that negatively affect the churn.  
the variables with coefficients with positive sign are the ones that positively affect the churn.

|  |  |
| --- | --- |
| **Significant variables in final model** | **Coefficients value (Numeric)** |
| SeniorCitizen | 0.35455 |
| Tenure | -0.90491 |
| MultipleLines.xYes | 0.34376 |
| InternetService.xFiber.optic | 0.81192 |
| InternetService.xNo | -0.59547 |
| StreamingMovies.xYes | 0.40844 |
| Contract.xOne.year | -0.75167 |
| Contract.xTwo.year | -1.60610 |
| PaperlessBilling | 0.34596 |
| PaymentMethod.xElectronic.check | 0.45585 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 4158.2 |
| Null deviance | 5704.4 |
| Residual Deviance | 4136.2 |

**C-statistics:**

Train : C Index is 8.424092e-01

Test : C Index is 8.375956e-01

From results of C-statistics for train and test, we can infer that the model has very good discriminatory power.

**KS-statistics:**

Train:   
KS-stats: 0.5268  
Cumulative percentage value is 0.1732. This is falling in the 2nd decide.

Test:  
KS-stats: 0.5228  
Cumulative percentage value is 0.2839. This is falling in the 3rd decile.

Ideally, KS-stats are above 0.40 for train & test, and if the cumulative percentage values fall in top 3 deciles, then the model is said to be a good one.

As per the test results above, KS-stats and cumulative percentages for both train and test are falling in the desirable range. Thus, we can say the model is a good one.

|  |  |  |  |
| --- | --- | --- | --- |
| **Train Dataset** | | **Test Dataset** | |
| C-statistic | 8.424092e-01 (c index) | C-statistic | 8.375956e-01 (c index) |
| KS-statistic | 0.5268551 | KS-statistic | 0.5228858 |
| Model Evaluation (write Accept or Reject) | | Accept | |

Threshold value : 0.3 (Positive : 1)

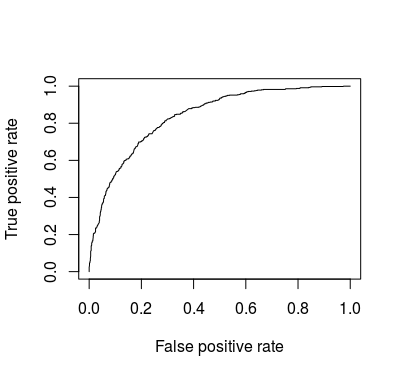
|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7623 |
| Sensitivity | 0.7607 |
| Specificity | 0.7628 |
| AUC | 0.8375956 |

Threshold value : 0.5 (Positive : 1)

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.8002 |
| Sensitivity | 0.5252 |
| Specificity | 0.8995 |
| AUC | 0.8375956 |

Threshold value : 0.7 (Positive : 1)

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7665 |
| Sensitivity | 0.17813 |
| Specificity | 0.97902 |
| AUC | 0.8375956 |

****

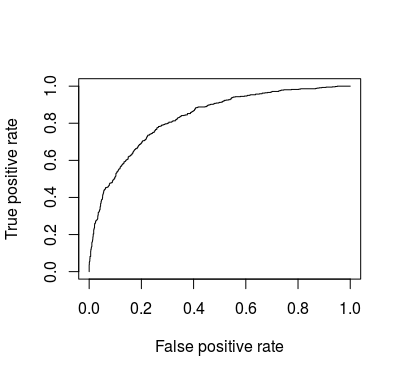
**Figure : AUC Curve for Logistic Regression**

* **Model – SVM**

Since, SVM uses dataset with numeric values, the same final dataset was used for SVM modelling where all the variables are numeric.

Tune function is used to find the optimal cost value via cross validation by providing a range of cost values. Once the best cost value is obtained, svm function was used to build the best SVM model.

Best model was achieved when the value of cost is 0.1



**Figure : AUC Curve for SVM**

**Threshold Value : 0.3.**

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7643 |
| Sensitivity | 0.7148 |
| Specificity | 0.7822 |
| AUC | 0.8302524 |

# Checkpoint 6: Threshold value

* Customer Retention is much cheaper than Customer Acquisition. This means, that Churn Rate must be decreased.
* We want to identify all Potential Churners. This means, we have to increase the True Positive Rate and minimise the probability of classifying a churner as a non-churner
* Hence, we go for a threshold which has High Sensitivity.[Churn is taken as the positive class]
* The Threshold Chosen is **0.3**
* For Threshold Value = **0.3,**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Specificity** | **Sensitivity** | **AUC** |
| KNN | 0.7965 | 0.8898 | 0.5383 | 0.7512381 |
| Naïve Bayes | 0.7331 | 0.7914 | 0.7120 | 0.8270526 |
| **Logisitic Regression** | **0.7623** | **0.7607** | **0.7628** | **0.8375956** |
| SVM | 0.7643 | 0.7148 | 0.7822 | 0.8302524 |

We, choose Logistic Regression as the best model to predict Churn. Apart from having High sensitivity and accuracy for the given Threshold, Logistic Regression also helps us understand how various variables influence churn and by how much. Logistic Regression also have the Highest AUC in comparison to other models.

**Threshold value : 0.3 (Positive : 1)**

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7623 |
| Sensitivity | 0.7607 |
| Specificity | 0.7628 |
| AUC | 0.8375956 |