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.

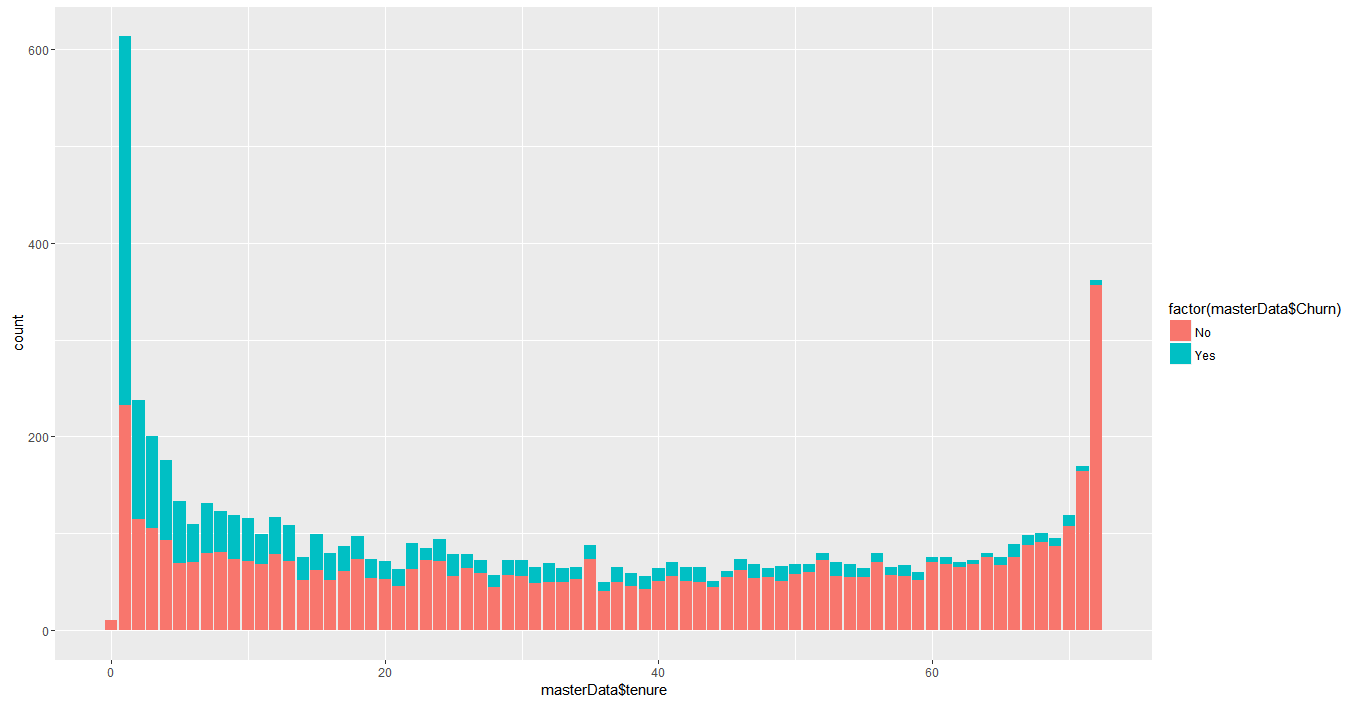
Rows: 7043

Columns: 21

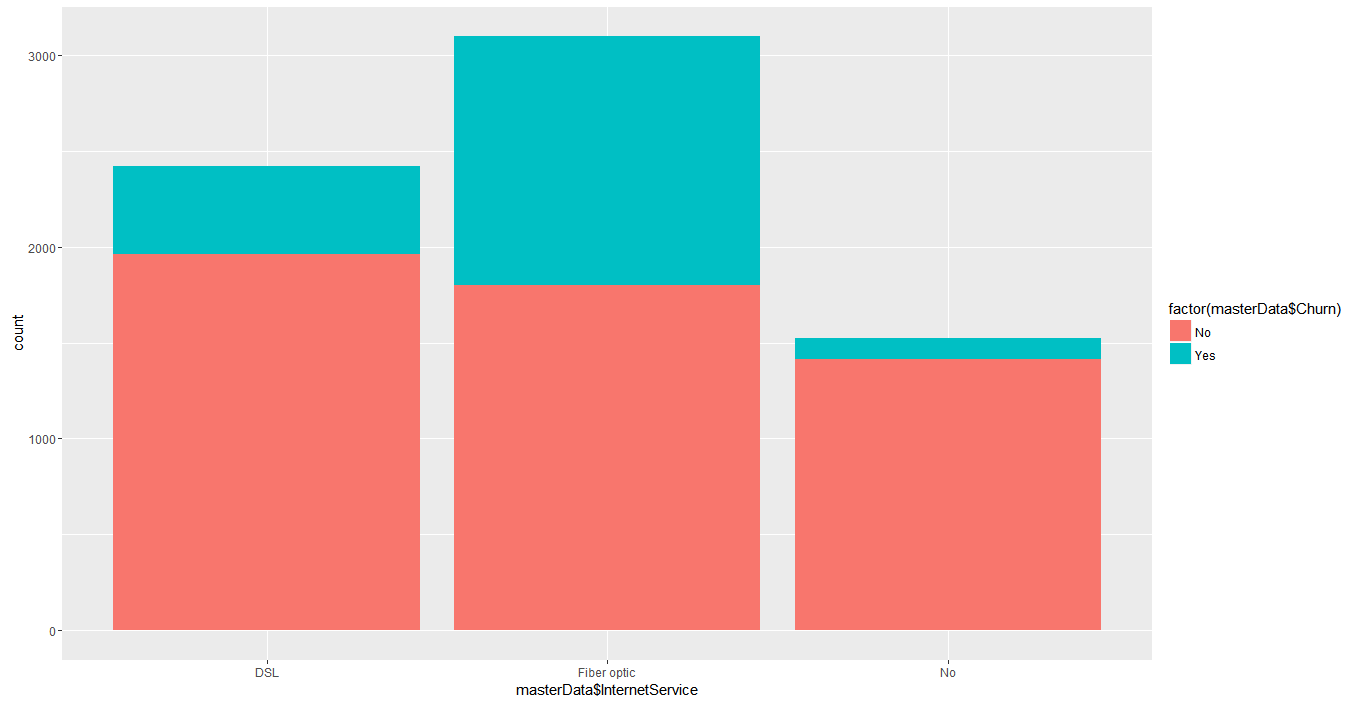
# Checkpoint 2: Exploratory Data Analysis

* Display the plots obtained and report the derivable insights.

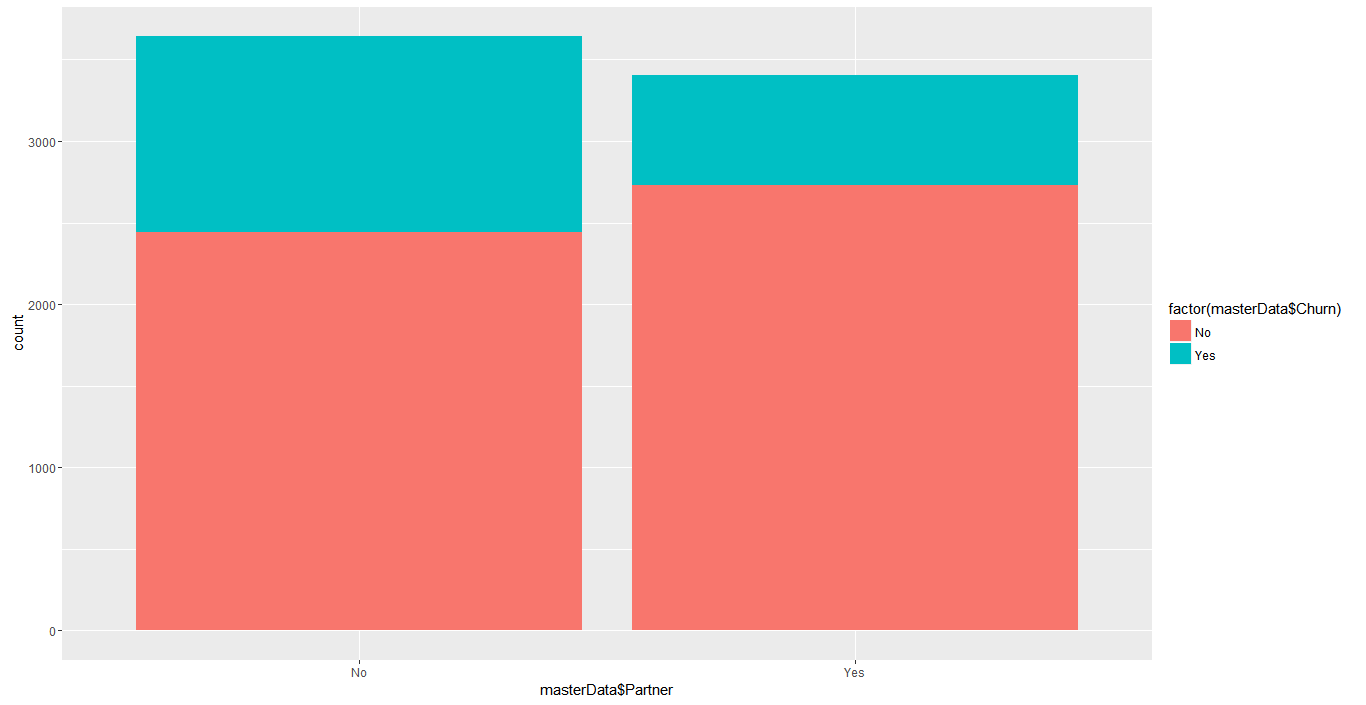
The churn rate decreases with increase in tenure. Looking at the tenure we see that 600+ customers are new or have been for least tenure and the churn rate is also pretty high at this level.



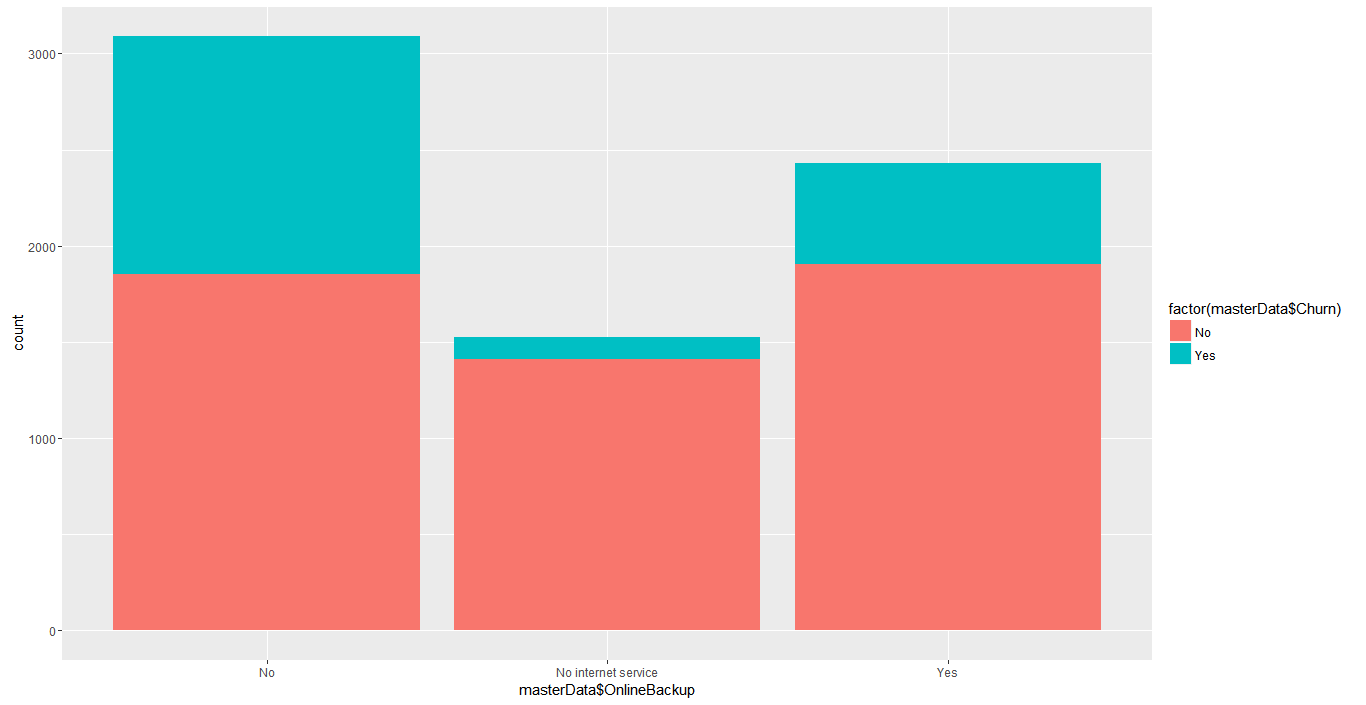
Type of internet service chosen might be a significant factor that leads to churn, we see that the number of customers churning is higher among the customers with Fibre Optic connection



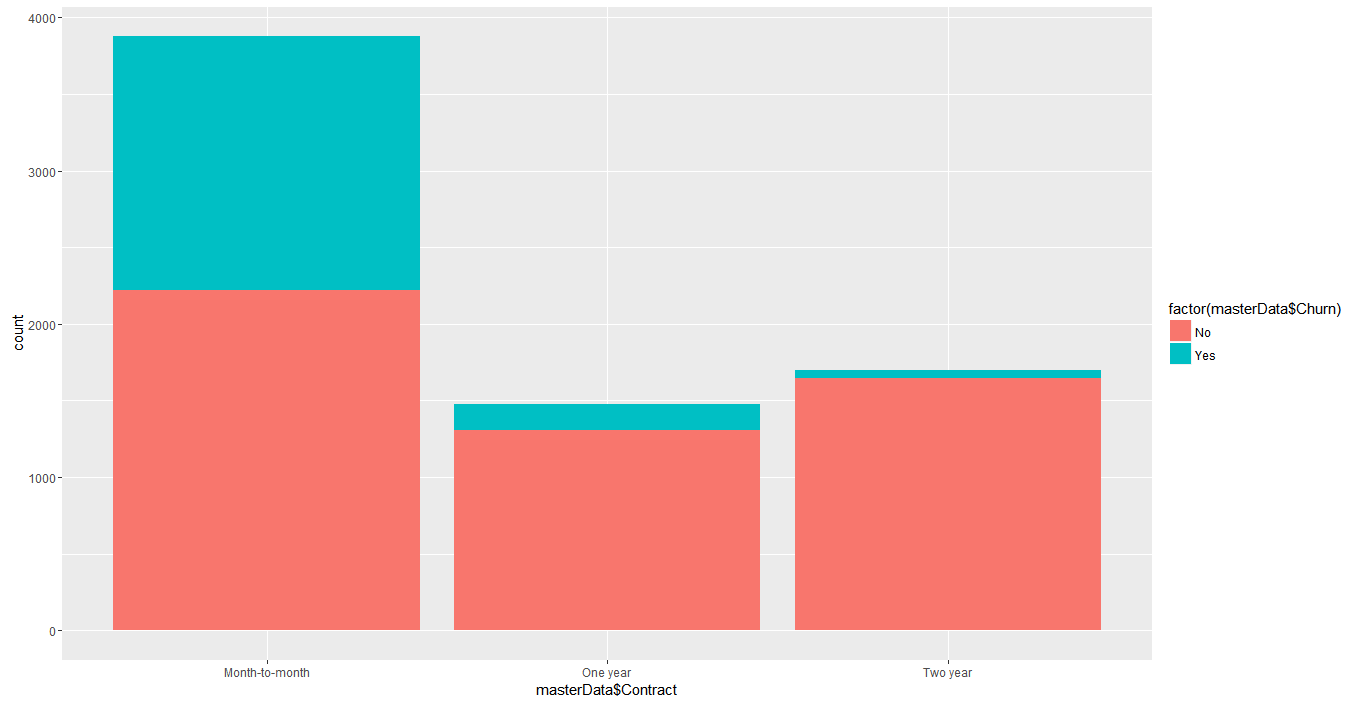
It looks like the churn rate is higher among the people who are single when compared to married/people in live in relationships

.

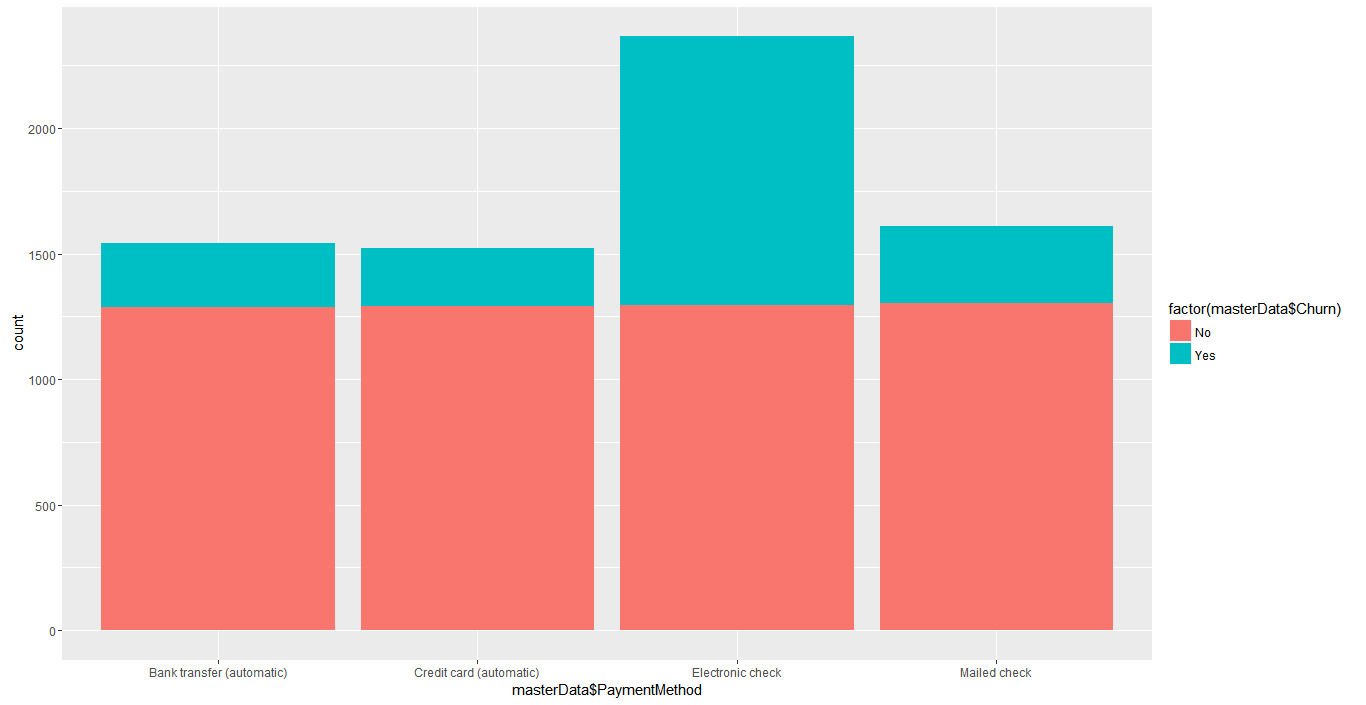
The churn is more among customers who do not opt for online backup.



The churn rate is highest among people choosing month-to-month contracts.



The churn is higher among customers paying through electronic check



# Checkpoint 3: Data Preparation

* Report the number of duplicated in the data.

0 – no duplicates.

* Explain the methodology of Missing value treatment and additionally fill the below table:

Missing values were seen in the Total Charges variable which is continuous in nature. Hence these values were imputed with the mean value.

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

* Explain the methodology of Outlier treatment and fill the below table:

Using boxplot and quantile functions we could see that there are no outliers are present in any of the numeric variables.

* Bring the data in the correct format. Report the number of variables for which the format was changed.

Most variables in the dataset are factorial in nature. So, while importing these, we had set the argument “StringasFactors as True”. Note that, we did check that the number of levels in these variables are fine after import. So, the only variable which had to be converted was SeniorCitizen.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | None of the variables need outlier treatment |
| Dummy creation | Contract,paymentMethod,multipleLines,internetService,onlineSecurity,onlineBackup,deviceProtection,techSupport,streamingTv,streamingMovies,Gender,partner,seniorCitizen,dependents,phoneService,paperlessBilling |
| Binning of variables | We have not done any binning here. We could have done binning for Tenure variable but however we chose not to do it. |

# Checkpoint 4: Modelling

* **Model – K-NN**
  + Explain the Data Preparation step for K-NN modelling.

Feature transformation, Scaling and splitting of the dataset was done in common. We used the sample.split() function from the caTools package for splitting the data in the ratio of 70:30.

* + Explain the methodology of building the model with optimal value of K?

In order to find the Optimal K-value, we used the 10-fold cross-validation method.

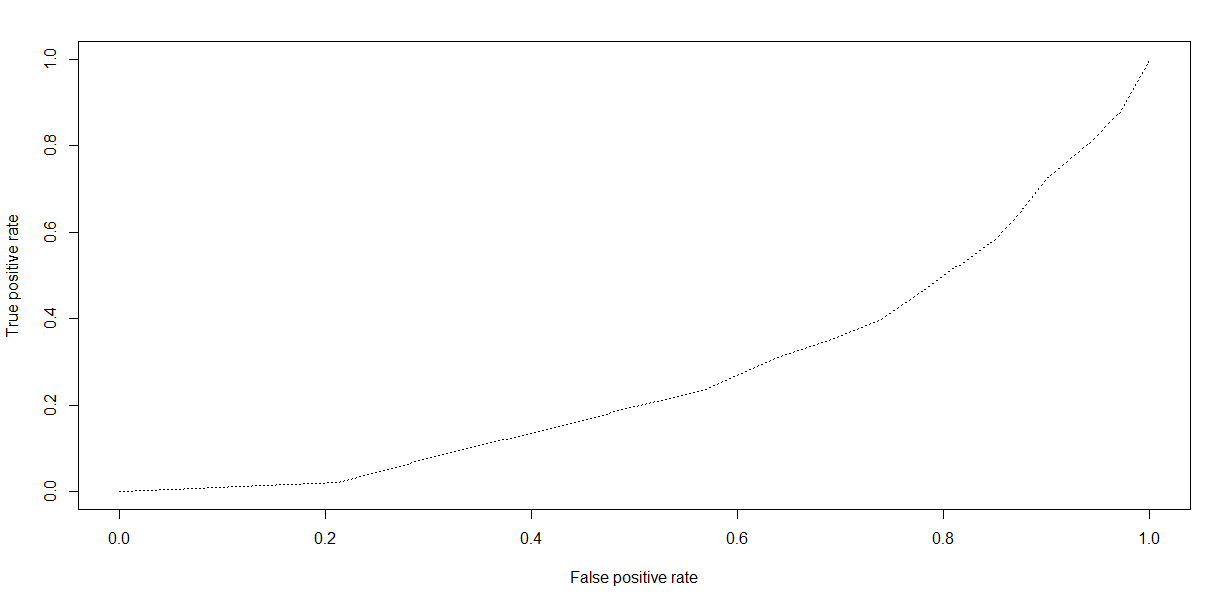
The result of cross validation was obtained as k=35. Accuracy was used as the prime measure for the model building.

Additionally, fill the below table:

'Positive' Class: Yes (Churn = Yes)

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7932 |
| Sensitivity | 0.5330 |
| Specificity | 0.8872 |
| AUC | 0.2513972 |

* + Display the AUC curve.



* **Model – Naïve Bayes**
  + Explain the Data Preparation step for Naïve Bayes modelling.

Feature transformation, Scaling and splitting of the dataset was done in common. We used the sample.split() function from the caTools package for splitting the data in the ratio of 70:30.

* + Explain the methodology of building the model.

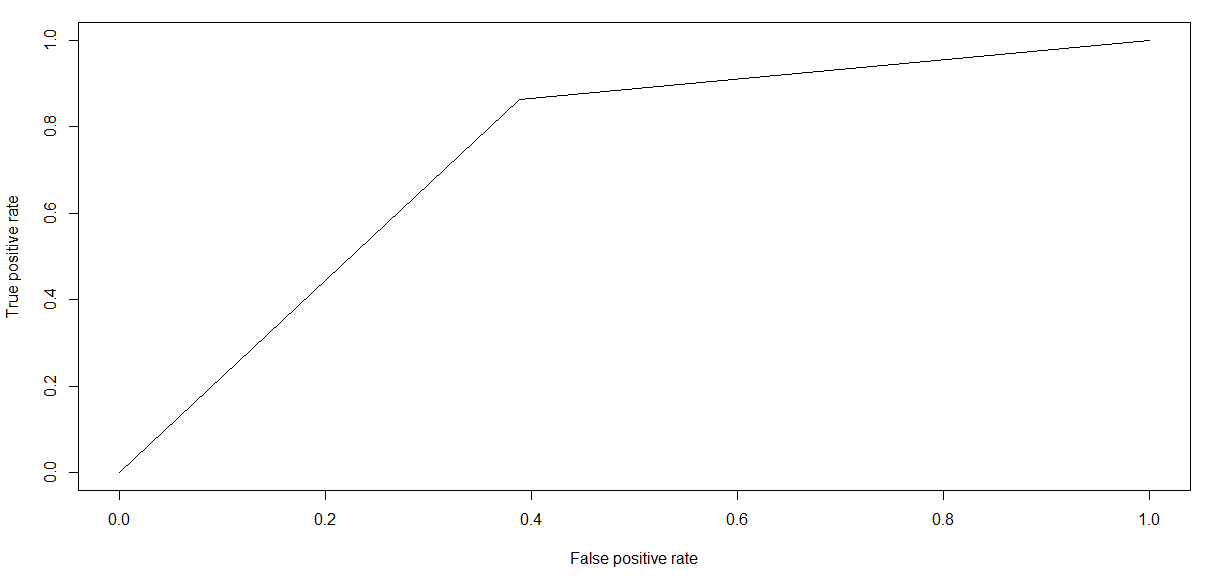
We create the model for predicting the response variable (churn) w.r.t all the other variables in the dataset. We then use the model for predicting the probability of the response variable being in either of the classes (yes or no). Details are recorded below:

Additionally, fill the below table:

# The positive class here is “No” i.e. we should look at specificity to maximize identifying the number of people who churn(churn=yes) correctly. Here, the specificity of the model is pretty high.

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.6787 |
| Sensitivity | 0.6121 |
| Specificity | 0.8627 |
| AUC | 0.7374 |

* + 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

We create an initial model with all the predictor variables in the dataset. We then use the STEPAIC method to identify the most significant variables. We further tune the model based on VIF values (<2) and also based on multicollinearity and P-Values of the variables. Hence, we arrive at a final model details of which is below:

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeric)** |
| (Intercept) | -0.598242 |
| tenure | -0.033134 |
| ContractOne.year | -0.754504 |
| ContractTwo.year | -1.657352 |
| PaymentMethodElectronic.check | 0.530221 |
| MultipleLinesYes | 0.404304 |
| InternetServiceFiber.optic | 0.958977 |
| InternetServiceNo | -0.791459 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 4206.3 |
| Null deviance | 5704.4 on 4929 degrees of freedom |
| Residual Deviance | 4190.3 on 4922 degrees of freedom |

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

As we know the **c-statistic**is an implicit measure of the fraction of concordant pairs.  More concordant pairs result in a higher value of c-statistic. The c-statistic for both the train and the test data set are 83.8% and 83.49% respectively which are very good values and it says that model is predicting well on both the train and the test data sets.

**KS-statistic** is a way to measure the goodness of fit of logistic regression models. KS- Statistic is a measure of the degree of separation between the positive and negative distributions. The KS-statistic values obtained from the final model for the test and train data sets are 53.10 and 53.07 which are also very values and it says that model is predicting well on both the train and the test data sets.

Ideally the k-statistic should be greater than 40 for a given model. Both the KS statistics lie in the 1st decile which is a very good sign as it lies in the top deciles hence we can go ahead and accept this model for the prediction of the churning customers.

**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 | 83.8% | C-statistic | 83.49% |
| KS-statistic | 53.10 | KS-statistic | 53.07 |
| Model Evaluation (write Accept or Reject) | | Accept | |

'Positive' Class: Yes (Churn = Yes)

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.6976 |
| Sensitivity | 0.8235 |
| Specificity | 0.6521 |
| AUC | 0.8349 |

* **Model – SVM**
  + Explain the Data Preparation step for SVM modelling.

Feature transformation, Scaling and splitting of the dataset was done in common. We used the sample.split() function from the caTools package for splitting the data in the ratio of 70:30.

* + Explain the methodology of building the model.

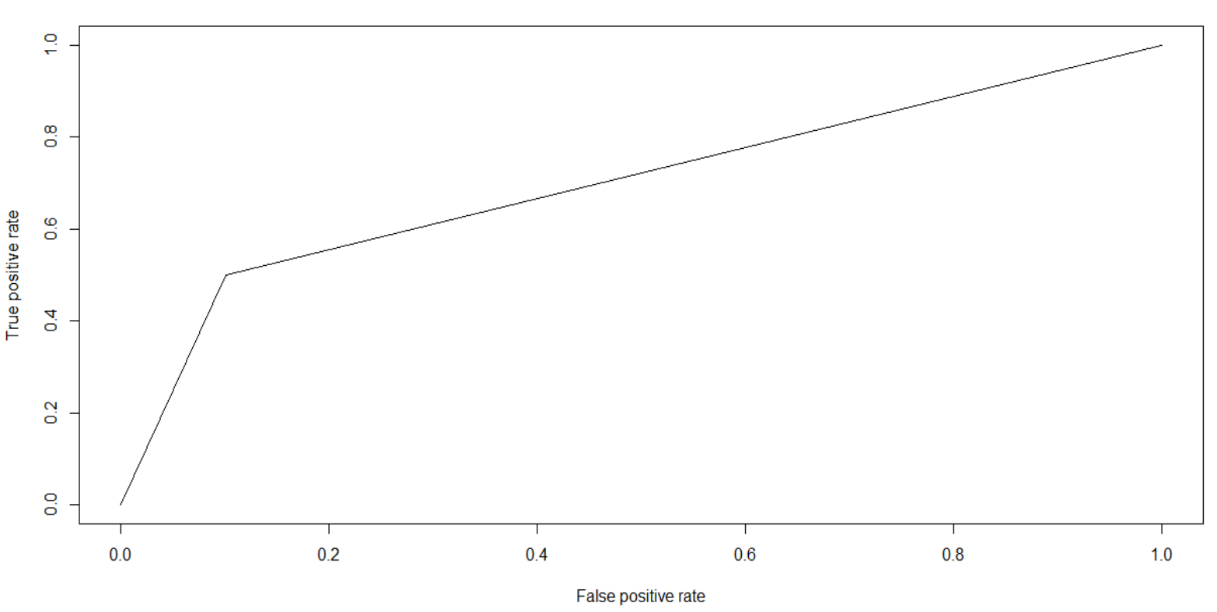
We use repeated cross validation to find the optimal cost from a range of 0.001,0.1,0.5,1,100.We used the tune function in R with above list of cost values and this function gives us the best model for the given lists of costs.

Additionally, fill the below table:

'Positive' Class: No (Churn = No)

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7979 |
| Sensitivity | 0.9001 |
| Specificity | 0.5152 |
| AUC | 0.70018 |

* + Display the AUC curve.



* Report the best model and its performance metrics.

There are two models which are performing well here. Naïve Bayes and Logistic Regression.

By looking into the accuracy, sensitivity, specificity and the AUC we select the Logistic regression model as the best model.

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.6976 |
| Sensitivity | 0.8235 |
| Specificity | 0.6521 |
| AUC | 0.8349 |

# Checkpoint 6: Threshold value

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

We have chosen a value of 0.20 as threshold as we want to increase our TPR (identify most of the customers who churn), to avoid risks. This means we will have to minimize False Negative rate, i.e. choose lower thresholds.

NOTE: We are looking at a sensitivity of 80% at least for this model. Hence the threshold.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.6976 |
| Sensitivity | 0.8235 |
| Specificity | 0.6521 |