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**.**

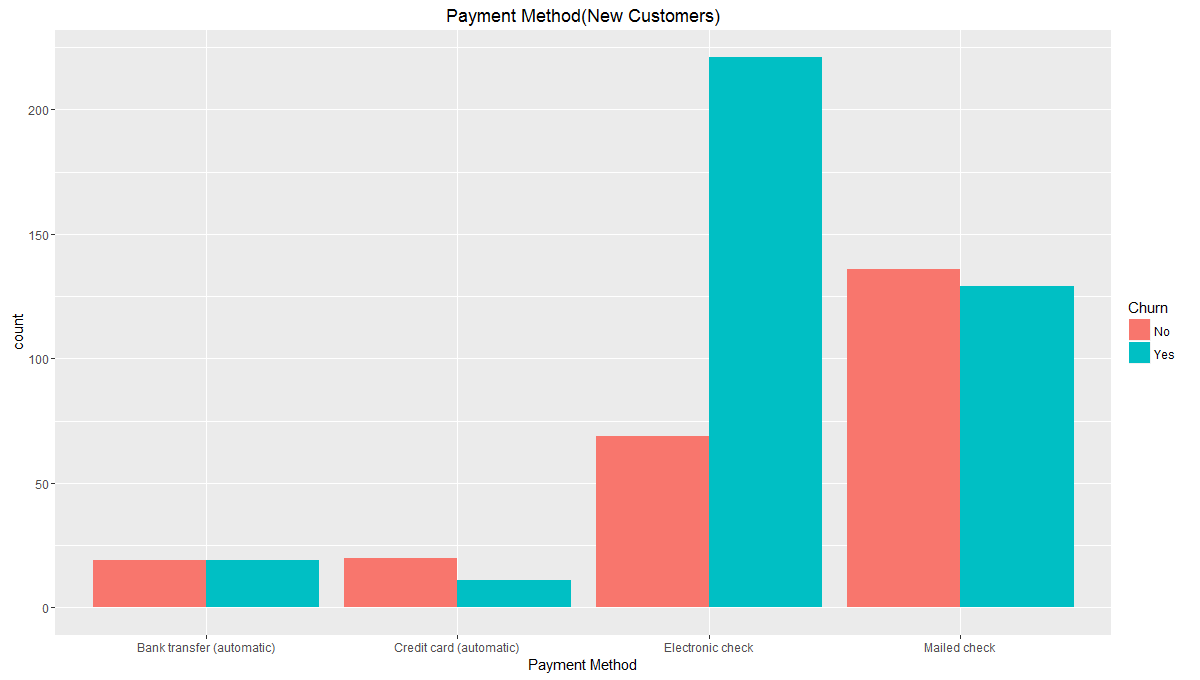
# Checkpoint-1: Data Understanding and Preparation of Master File

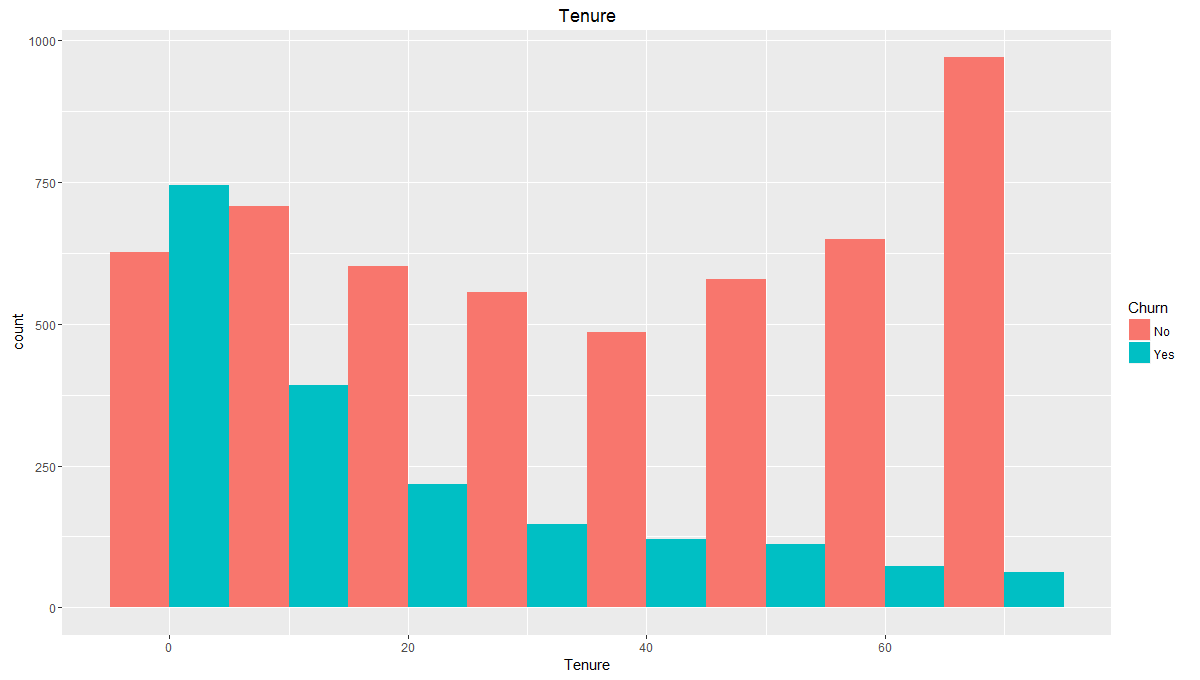
Files given:

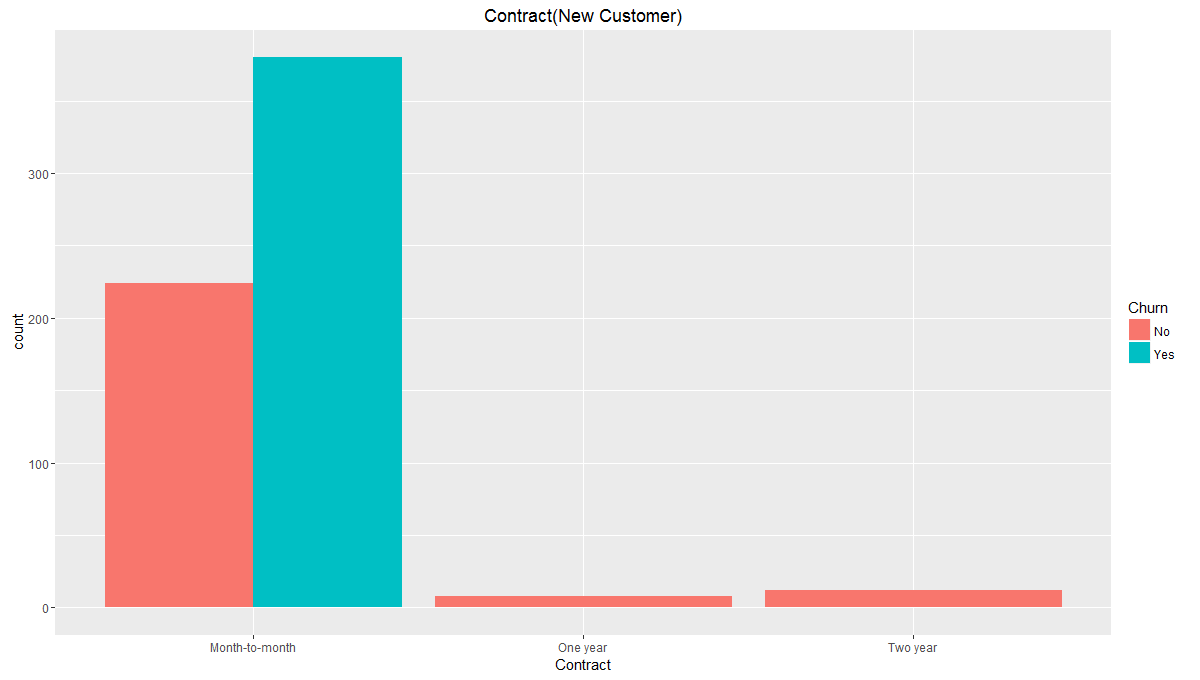
|  |  |  |
| --- | --- | --- |
|  | **# of Observations** | **Variables** |
| **customer\_data.csv** | 7043 | 9 |
| **churn\_data.csv** | 7043 | 5 |
| **internet\_data.csv** | 7043 | 9 |

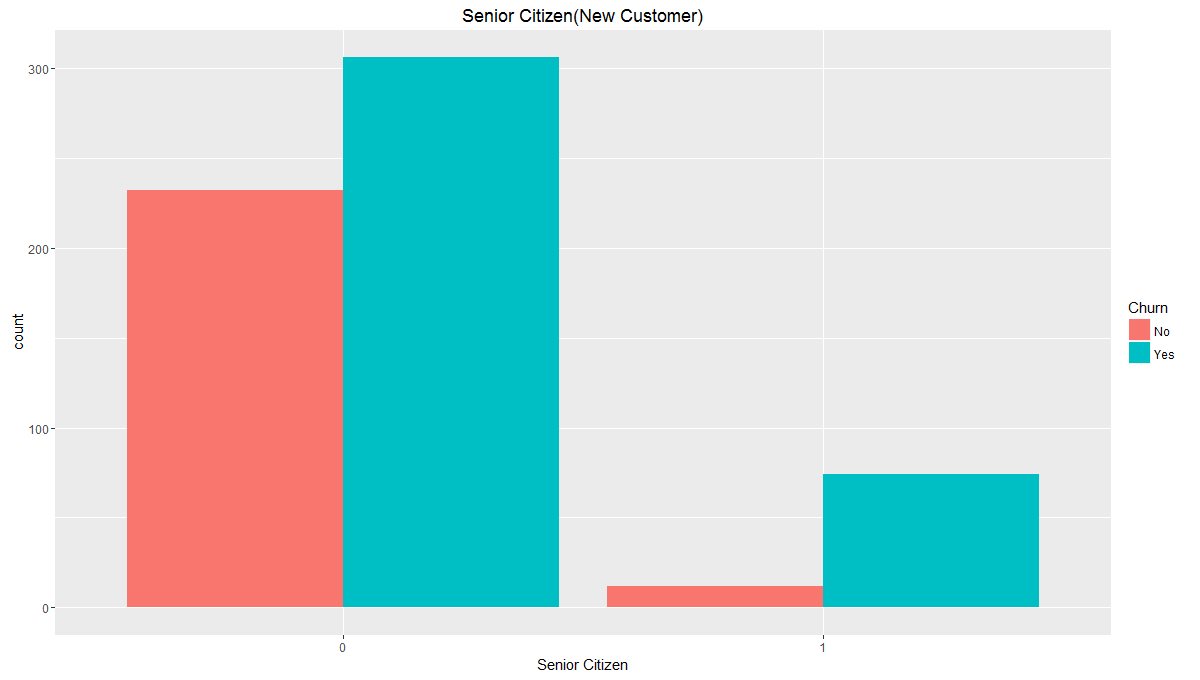
Final file also has 7043 observation when merged on customerID column

# Checkpoint 2: Exploratory Data Analysis









# Checkpoint 3: Data Preparation

* No duplicate data in the data set.
* Missing values were found only in one variable **TotalCharges**. Value was imputed with product of MonthlyCharges and tenure

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

* No outliers were found
* Bring the data in the correct format.
  + SeniorCitizen variable was changed to factor

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | None |
| Dummy creation | 1. gender 2. SeniorCitizen 3. Partner 4. Dependents 5. PhoneService 6. Contract 7. PaperlessBilling 8. PaymentMethod 9. MultipleLines 10. InternetService 11. OnlineSecurity 12. OnlineBackup 13. DeviceProtection 14. TechSupport 15. StreamingTV 16. StreamingMovies |
| Binning of variables | 1. Tenure 2. MonthlyCharges |

# Checkpoint 4: Modelling

* **Model – K-NN**

K-NN can be implemented in R using the knn() command from the ‘class’ package. Once you have built the model the next step is to evaluate it. There are four major evaluation metrics on which the performance of an algorithm can be judged. These are:

* Accuracy
* Sensitivity
* Specificity
* ROC Curve

You can directly get accuracy, sensitivity and specificity directly by using the command confusion.matrix(). It basically compares the prediction results (i.e. class labels) with the true class labels of test data.

Accuracy is defined as the percentage of correct predictions.

Accuracy = Correct Predictions/ Total Predictions

Specificity: It measures the proportion of negatives that are correctly identified as negatives. It is also known as true negative rate.

Specificity = TN / (TN + FP)

Sensitivity: It measures the proportion of positives that are correctly identified as positives. It is also known as true positive rate.

Sensitivity = TP / (TP+FN)

ROC curve plots the true positive rate vs the false positive rate i.e. it is the plot of 'sensitivity' vs '1-specificity'. The area under the curve measures how good the model in differentiating between data points belonging to different categories, more the area – better the model. A model having an area under the curve to be 50% is considered useless because there is only a 50% chance of predicting correctly. 50% area under the curve signifies that the results generated by the algorithm are only as good as random guessing. The area under the ROC curve measures the discrimination power of the algorithm i.e. whether the algorithm can differentiate between a data point belonging to a positive class from one belonging to a negative class. The greater the area, better the discrimination power.

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

The best way to find the optimal k is by using cross-validation. In cross-validation, the data set is divided into various combinations of training and testing data. The algorithm is implemented on all those combinations for different values of parameter i.e. K in this case. The mean accuracy of runs of every K on different combinations of data is then calculated. Optimal K is that value of K for which the accuracy is maximum.

For example, suppose you have a data set consisting of 1000 observations. Cross Validation will work in this way for that data set:

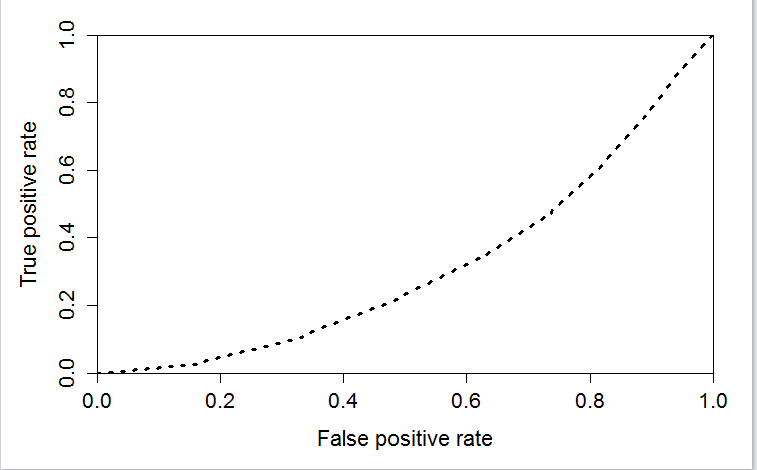
1. Break the data set into different bins of training and testing data. Say Bin 1 has the first 900 points as training data and the last 100 points as testing data; Bin 2 will have some other 900 points as training data and the rest 100 as testing data;Bin 3 will have yet another 900 points as training data the rest 100 as testing data and so on.
2. For each value of K, in this case 1 to 50, the algorithm is implemented on each of the bins. So if you have 10 such bins, K-NN will run 10 times and return 10 values of accuracy. Notice how you have generated 10 models with 1000 points to compute 10 accuracy values (while usually, we make only one) - this is the real benefit of cross-validation.
3. For each K, the average accuracy of the 10 Bins is calculated.
4. Optimal K is the K for which the average accuracy is maximum.

In short, rather than comparing only one accuracy value for each K, cross-validation smartly utilizes the data to generate multiple accuracies and compares the average.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.77 |
| Sensitivity | 0.4087 |
| Specificity | 0.9055 |
| AUC | 0.3147923 |

AUC curve



* **Model – Naïve Bayes**

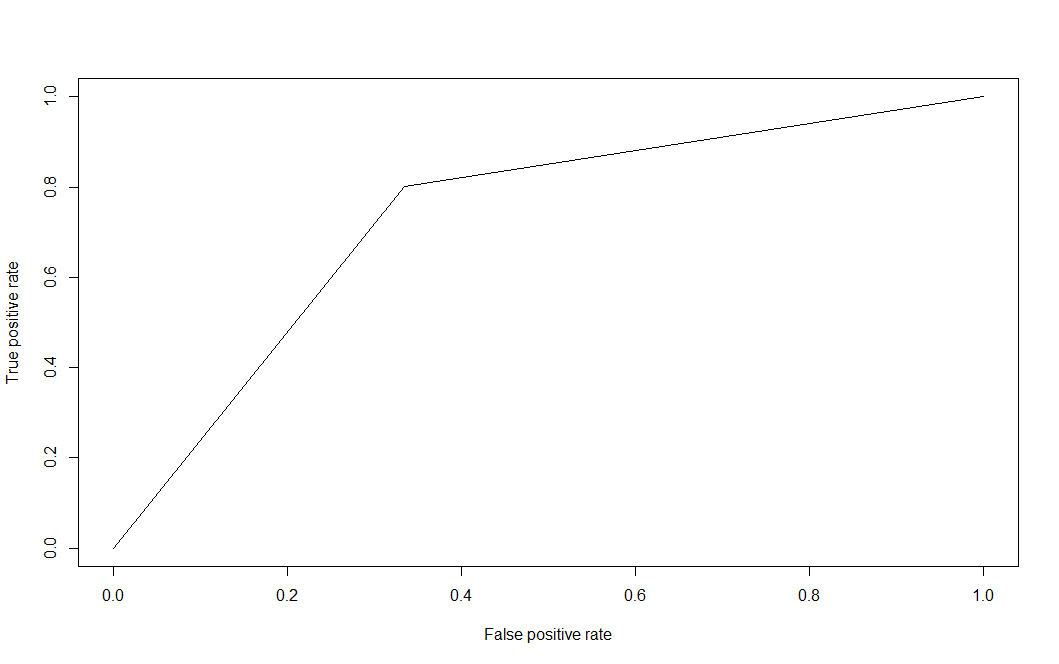
**Steps followed:**

* 1. Converted all the char variables into factor. Done the binning of numerical variables.
  2. Done the missing value treatment in Total charges variable, divided the data into train and test in 70:30.
  3. Developed the model with training data using NaïveBayes function of e1071 package.
  4. Done the prediction on test data using the above developed model.
  5. Shown the confusion matrix with the predicted data using positive class as "yes"
  6. Drawn the ROC curve between True positive rate and false positive rate

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7018 |
| Sensitivity | 0.8014 |
| Specificity | 0.6660 |
| AUC | 0.7337 |

* + Display AUC curve.



* **Model – Logistic Regression**
* Explain the methodology of building the model?
* First the model is built with all the independent variables and their relationship with dependent variables.
* Then, the resulting model is fed into stepAIC function where the best model based on minimum possible AIC value and retaining most significant variables.
* The model resulting from stepAIC function is then manually examined for significance of variables and AIC value. VIF is also checked to see multi-collinearity between variables.
* Variables with less significance and low VIF are retained till model cannot be bettered.
* The final model is tested against test data to verify its accuracy.
* In the final model, interpret what the coefficients of the variable imply. Check if the coefficients make business sense

Additionally, fill the below table:

|  |  |  |
| --- | --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeric)** | **Interpretation** |
| tenure | -0.036846 | Negative impact on churn variable. |
| SeniorCitizen | 0.354546 | Positive impact on churn variable. |
| ContractOne.year | -0.751674 | Negative impact on churn variable. |
| ContractTwo.year | -1.606095 | Negative impact on churn variable. |
| PaperlessBilling | 0.345957 | Positive impact on churn variable. |
| PaymentMethodElectronic.check | 0.455848 | Positive impact on churn variable. |
| MultipleLinesYes | 0.343756 | Positive impact on churn variable. |
| InternetServiceFiber.optic | 0.811923 | Positive impact on churn variable. |
| InternetServiceNo | -0.595472 | Negative impact on churn variable. |
| StreamingMoviesYes | 0.408441 | Positive impact on churn variable. |

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

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

High value of C-Statistics confirm that model is able to accurately predict customer churn.

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 | 0.8424 | C-statistic | 0.8411 |
| KS-statistic | 0.5268551 | KS-statistic | 0.5265818 |
| Model Evaluation (write Accept or Reject) | | Accept | |

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7602 |
| Sensitivity | 0.7607 |
| Specificity | 0.7628 |
| AUC | 0.84 |

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

The data preparation step was common for all the models and hence no explicit data preparation was done for SVM Modelling.

* + Explain the methodology of building the model.

The most critical parameters to create a sophisticated SVM model is cost function and gamma function. For our case we use only cost function as deciding factor and choosing appropriate cost function will give me better SVM model.

Initial run was done for extreme range of C values eg. C=.1 and C =10 . Then tune () was used to

choose appropriate Cost function with sequence of values c(0.001,0.01,0.1,1,5,10,100)) and we get the result as c = 0.01 as best cost function. This indicates the cost function in range of 0.01 .

Next attempt is made to choose better cost function in range .01 and .1 using sequence cost=c(0.01,0.02,0.03,0.04,.05,.06,.07,.08,0.09). Tune function choose .06 as best cost function.

Cost and misclassification relationship:

* Cost is small -> Many points can violate the margin
* Cost is large -> Margin Narrow and hence few points can violate the margin.

Additionally, fill the below table:

**Positive class = YES**

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 80% |
| Sensitivity | 54 % |
| Specificity | 90 % |
| AUC | 0.695 |

* + **Report the best model and its performance metrics.**

Since the objective for telecom company was to identify which customer which churn, that is churn = YES. Therefore **Logistic Regression is the chosen model considering relatively high Sensitivity and accuracy.**

**Positive Class = YES**

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7602 |
| Sensitivity | 0.7607 |
| Specificity | 0.7628 |
| AUC | 0.84 |

# Checkpoint 6: Threshold value

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

**Most appropriate threshold value was found to be 0.3.**

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value(0.3)** | **Values (Numeric)** |
| Overall Accuracy | 0.7602 |
| Sensitivity | 0.7607 |
| Specificity | 0.7628 |