The data presents us with the following features for our model:

|  |
| --- |
| CUSTOMERID, CITY, STATE, ZIP, REGION, CURRENTBALANCE, CONTRACT\_FEE, TOT\_OPEN\_AMT, TOT\_INVOICE\_AMT, TOT\_PAID\_AMT, NUM\_INVOICES, ACTIVATED\_YEAR, ACTIVATED\_MONTH, AGE\_RANGE, CREDIT\_APPROVAL, RENEWAL\_YEAR, RENEWAL\_MONTH, CONTACT\_METHOD, RATE\_PLAN, CHURN |

Modeling with all the variables gives the following results:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **TP** | **FN** | **FP** | **TN** |
| Neural Network | 96% | 26603 | 1025 | 1937 | 44822 |
| SVM | 94.8% | 24666 | 2962 | 935 | 45824 |
| Decision Forest | 92.2% | 18643 | 8965 | 1582 | 45177 |
| Boosted Decision Tree | 97.2% | 26556 | 1072 | 1019 | 45740 |

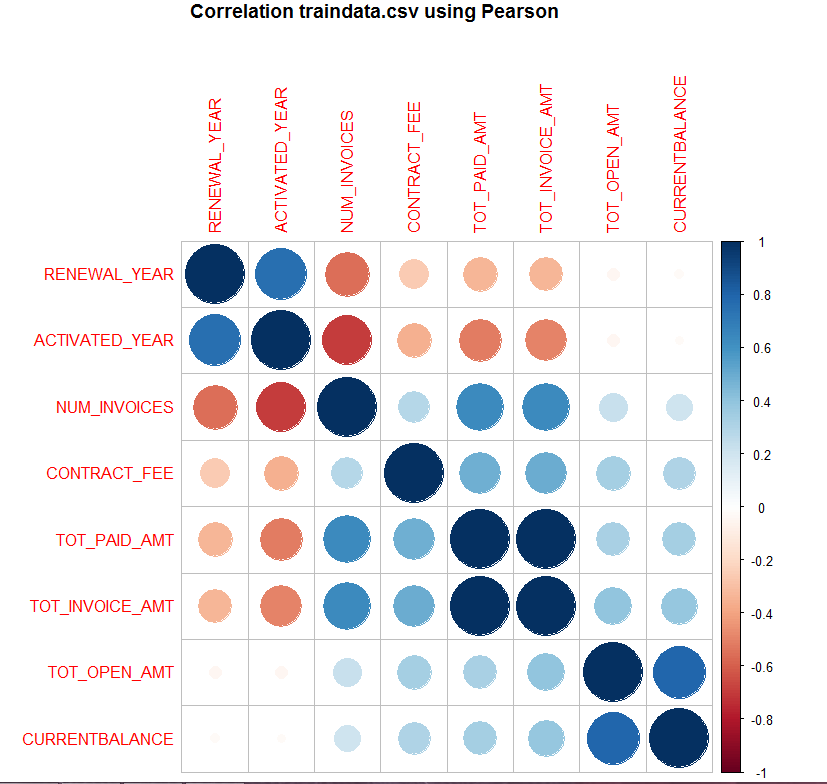
|  |  |
| --- | --- |
| **Neural Network** | **SVM** |
| **Decision Forest** | **Boosted Decision Tree** |
| **ROC** | |

I utilized the Microsoft Azure Machine Learning service to get the results above. These are considerably good predictive models. We can see that boosted decision tree performs the best in overall prediction. We can however see that SVM gives the least false positive values. The adjacent ROC supports our observations.

We can see that we have an accuracy rate of around **97.2%** from the models. A point to note would be that the FP (False Positives) are pretty low. These algorithms were able to identify the customers who most likely would have churned with high accuracy.

Performing similar functions for Random Forest in R, we have the following results:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **TP** | **FN** | **FP** | **TN** |
| Random Forest | 97.04% | 159778 | 4039 | 3344 | 93192 |

  
The models were created with all the variables. On closer observation, we must try to analyze the variables further to create a model which would work in real life. Using the Pearson’s correlation, we find strongly correlated variables in the model.

**TOT\_INVOICE\_AMT** & **TOT\_PAID\_AMT, CURRENTBALANCE** & **TOT\_OPEN\_AMT, NUM\_INVOICES** & **YEARS**.  
   
These variables can be combined using PCA for reducing dimensions.

For a telecom company, it is critical to identify customers who are most likely to churn since it is more expensive to attract new customers than to retain old ones. Low FP values show that the churn models have been successful at identifying the customers who are most likely to churn.