**Fraud data – Kmeans Cluster Report**

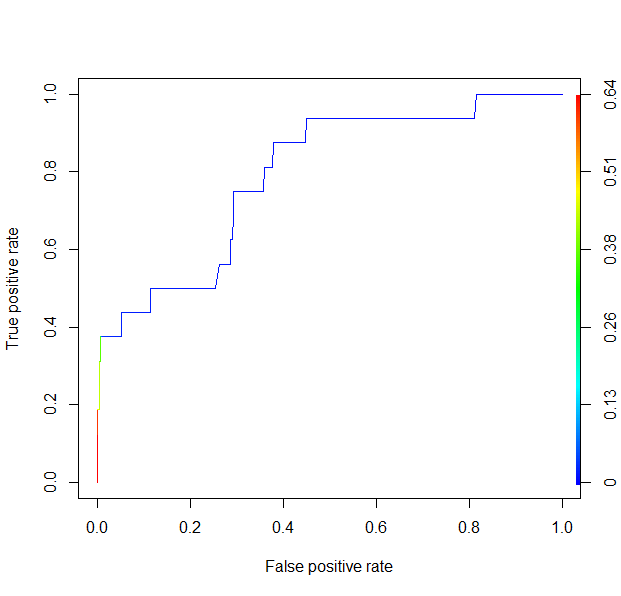
The data is transformed in excel before imported into R.

An additional 4 columns were added for each transformation; age, claim, tickets and atty.

The total number of observations in the original dataset is 5,000. However, only the first 2,000 observations were used to create and test the model. The training set is the first 1,000 observations and the test set are the second 1,000 observations.

The corresponding original fields to the four transformed fields were removed from the dataset.

The logistic model is created on the training set and the following ROC curve on the training set was generated:



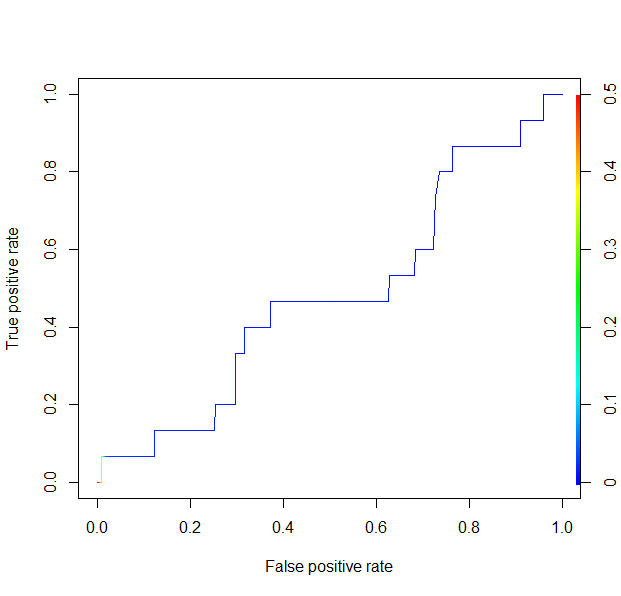
This graph shows that the best this model will do is less than 40% True positive rate and a False positive rate of almost 0 if the probability threshold for determining a fraudulent policy is approximately .2

The resulting confusion matrix for the training set at the .2 threshold is:

|  |  |  |
| --- | --- | --- |
|  | **Modeled OK** | **Modeled Fraud** |
| **Actual OK** | 977 | 7 |
| **Actual Fraud** | 10 | 6 |

The model is then applied to the test dataset:

The following ROC curve was generated on the test dataset



The resulting ROC curve demonstrates that the model is not a very good fit. A probability threshold of .1 results in a less than 10% True positive rate albeit a less than 1% False positive rate.

The resulting confusion matrix for the test set at the .1 threshold is:

|  |  |  |
| --- | --- | --- |
|  | **Modeled OK** | **Modeled Fraud** |
| **Actual OK** | 969 | 16 |
| **Actual Fraud** | 14 | 1 |

As a comparison, below I added the confusion matrix for a probability threshold at .01.

The resulting confusion matrix for the test set at the .01 threshold is:

|  |  |  |
| --- | --- | --- |
|  | **Modeled OK** | **Modeled Fraud** |
| **Actual OK** | 504 | 481 |
| **Actual Fraud** | 8 | 7 |

At a .01 probability threshold, the model accurately predicts approximately 47% of the fraudulent observations, but the accuracy of the Non-fraudulent observations is severely reduced.