Assignment 2: Preliminary Results – Week 6

Merrimack College

Machine Learning

Amol Gote

11th Oct 2020

This report primarily focuses on the preliminary results of the analytics plans. As part of the analytics plan the problem statement was “*Will the patient be readmitted and if will it be within 30 days or greater.*”

As per the analytics plan all the feature engineering steps were performed this includes

1. Finalizing the features sets or variables to be considered for analysis
2. Cleansing the data and treating the missing values appropriately
3. Removing the rows containing NA values.
4. Following variables one hot encoded gender, change of medication, diagnosis 1, diagnosis 2 and diagnosis 3.
5. Following variable target encoded age, race, admission source, admission type, discharge disposition id, mx\_glu\_serum, A1CResult, metformin, insulin.

## Random forest

Since the business objective to is to identify if patient will be readmitted again within 30 days, this is classification problem and the predictor variable is qualitative one, hence the random forest algorithm is good fit. Random forest was run with 75/25 data split of training and test data. Sample data size for training was of 45000 and test dataset 15000.

### ROC Curve and AUC.

Below is the ROC curve for the random forest model

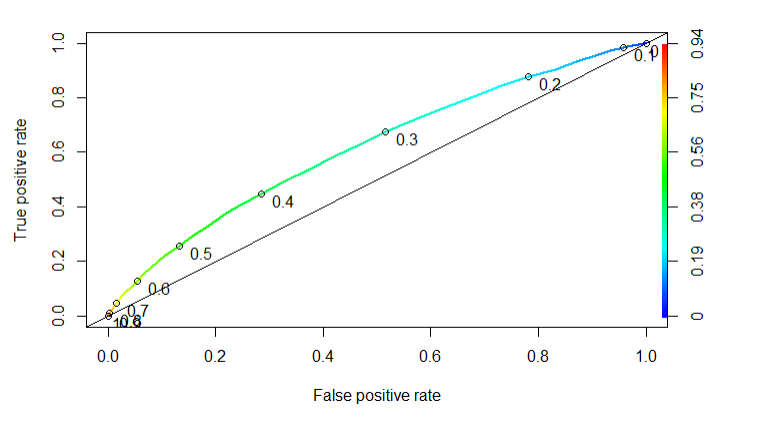


Figure : ROC Curve for random forest

AUC value that was achieved with multiple runs was **0.62**.

ROC shows the performance of the random forest classification model. A good classifier model will have the TROC curve hugging the top left corner, which clearly is not happening in this case of random forest model. AUC value of good classifier has to be 0.8 to 0.9 in this it is 0.62, so the model is not performing moderately for classifying the patient readmittance.

### Calibration Curve

Below is the calibration curve for the random forest

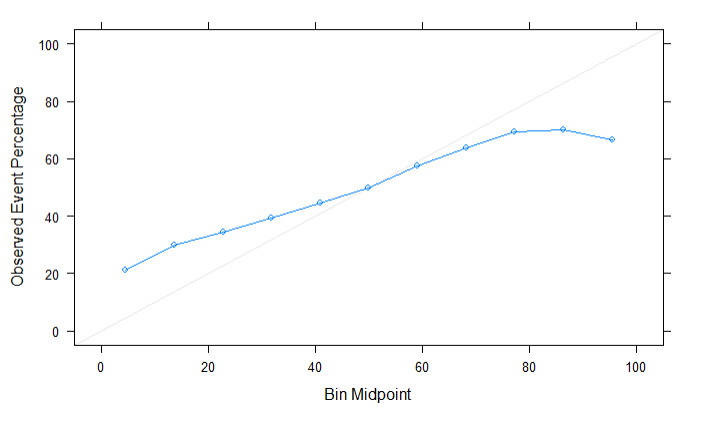


Figure : Calibration curve for random forest

In this case the calibration curve is not precisely near the diagonal line which clearly indicate the model is not performing that well and is below average.

## Neural networks

This is another good technique to build classification model

### ROC Curve and AUC.

Below is the ROC curve for the neural network model

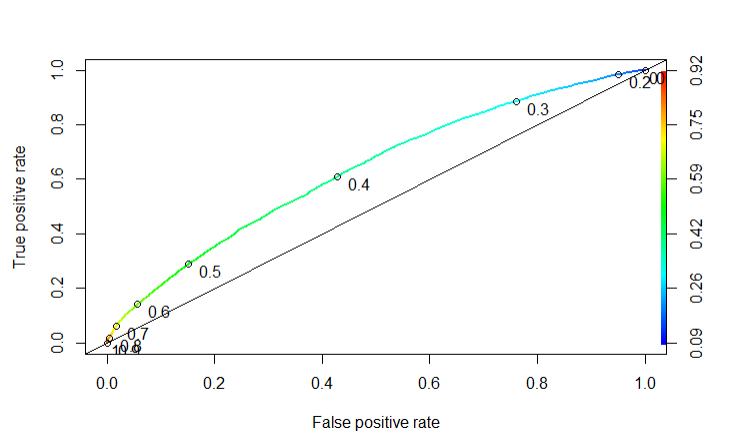


Figure 3: ROC Curve for nueral netwotk

AUC value that was achieved with multiple runs was **0.632**. This value is greater than random forest, so it is clear indicator that neural network model performs better than random forest

### Calibration Curve

Below is the calibration curve for the neural network

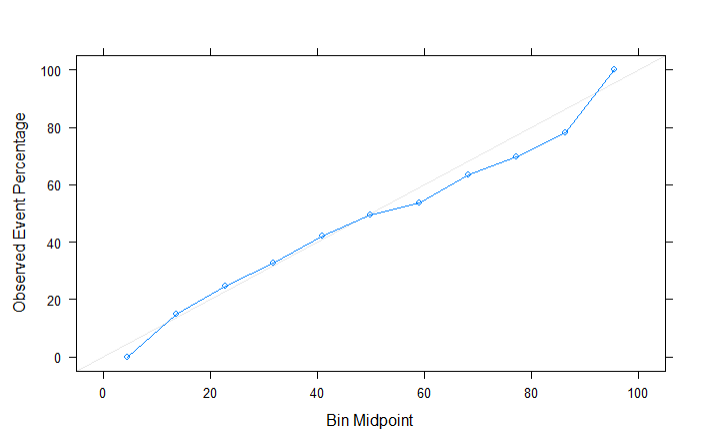


Figure 4: Calibration curve for neural network

In this case the calibration curve is near to the diagonal line compared to that of the random forest which is an indicator that neural network model performs better than random forest.

### Feature importance for random forest

Below is the feature importance plot random forest

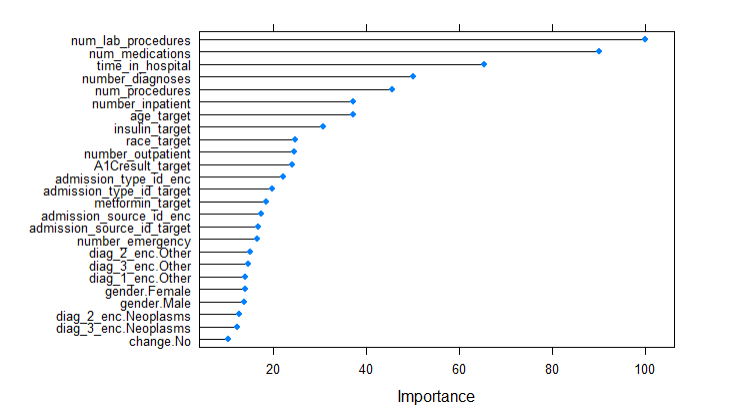
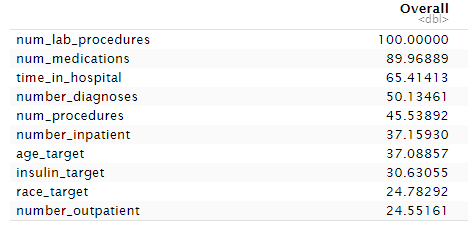
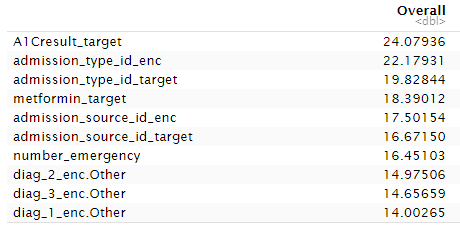


Figure : Feature Importance

Here are the feature importance values

### Additional ML algorithms and feature engineering step to improve performance