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 have been performed this includes

1. Finalizing the features sets 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 are one hot encoded
   1. Gender
   2. Change of medication
   3. diagnosis 1, diagnosis 2, and diagnosis 3.
5. The following variable target encoded
   1. Age
   2. race
   3. admission source
   4. admission type
   5. discharge disposition id
   6. mx\_glu\_serum
   7. A1CResult
   8. metformin
   9. insulin.
6. In the analytics plan, some medication variables were dropped during the preliminary analysis as they did not have any variance (Non zero variance variables).
7. Change of medication variable had got missed in the analytics plan, have added during the preliminary analysis.

## Random forest

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

### ROC Curve and AUC.

Below is the ROC curve for the random forest model

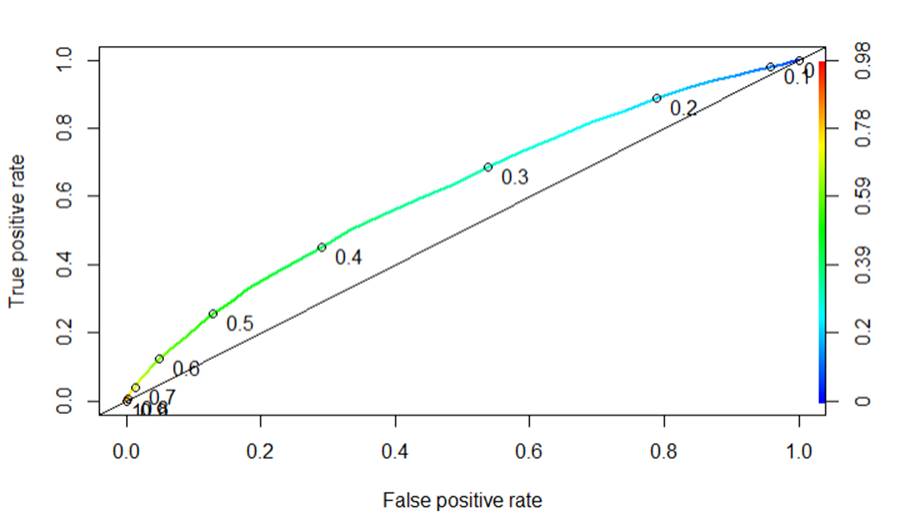


Figure : ROC Curve for random forest

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 is not happening in this case of the random forest model. AUC value of a 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

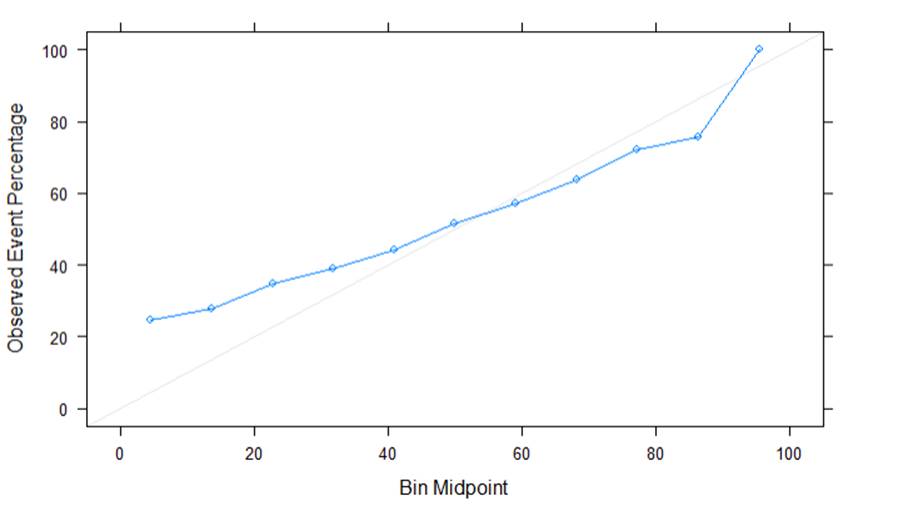


Figure : Calibration curve for random forest

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

## Neural networks

This is another good technique to build a classification model

### ROC Curve and AUC.

Below is the ROC curve for the neural network model

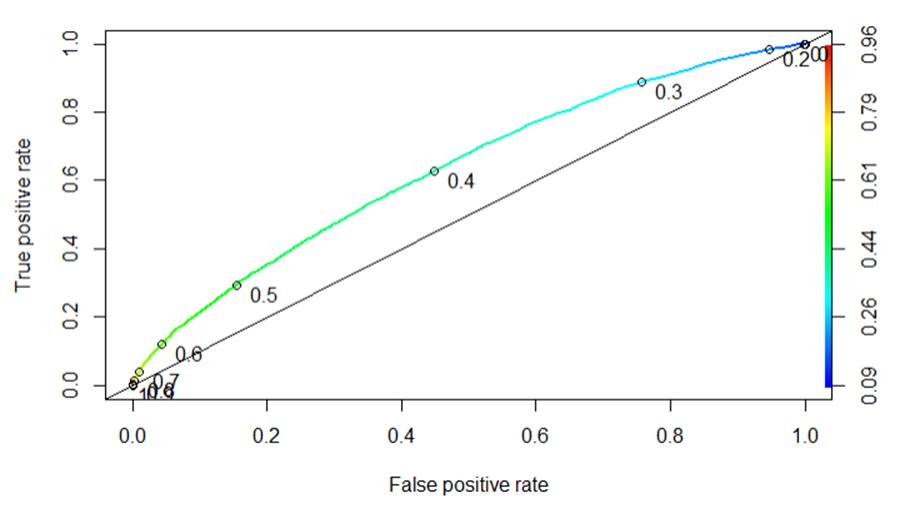


Figure : ROC Curve for neural network

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

### Calibration Curve

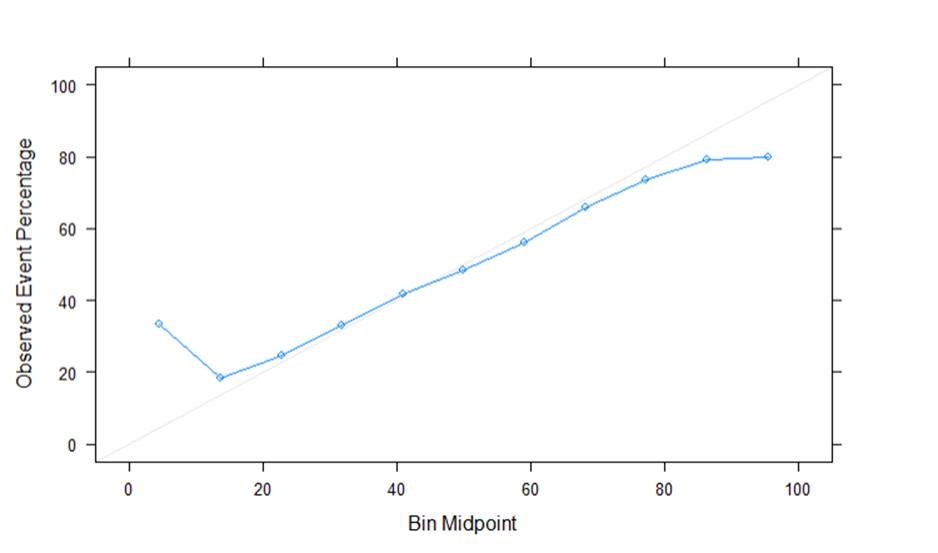


Figure : Calibration curve for the neural network

Above is the calibration curve for the 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 the neural network model performs better than the random forest.

### Feature importance for random forest

Below is the feature importance plot random forest

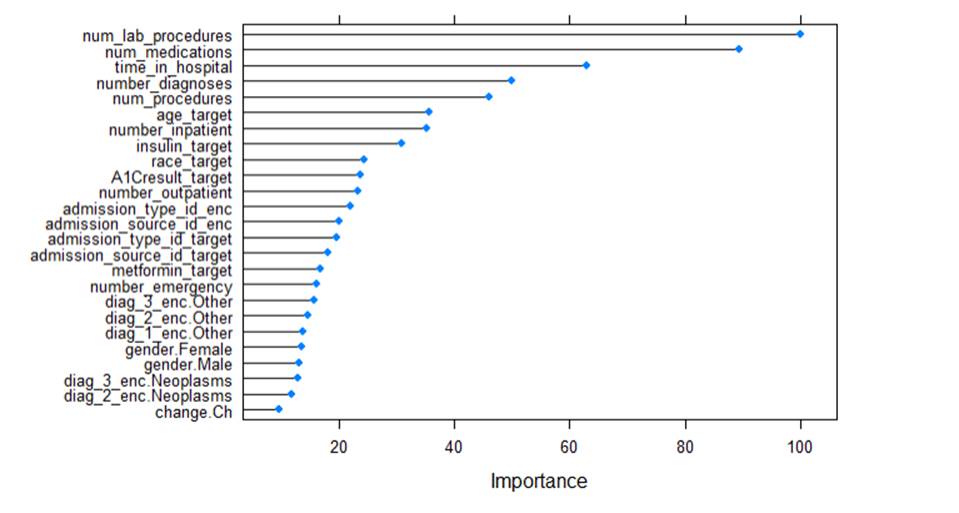


Figure : Feature Importance random forest

Here are the feature importance values



Figure : Feature importance values

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

### ML Algorithms

Random forest is already providing the feature importance list. So, based on the feature importance, select the top 10 important features and then rerun alternate machine learning algorithms

1. Logistic Regression – This is a commonly used algorithm for classification when the result has to be binary 1 or 0. In this case the target variable into discrete categories of readmittance Yes or No. Also, in this case there is already an analysis done to identify the feature importance, the same set of important features could be used to perform logistic regression.
2. Gradient boosting – This algorithm is based on the decision trees, but it is different from the random forest, the variation is in how the trees are built, random forest builds tree autonomously and combine the results at end of the process while gradient boosting builds one tree at a time and combines the result along the way.

### Class Imbalance

In this case, there is a higher representation of the patient not getting readmitted compared to that been readmitted <30: 6293, >30:22240, and No: 42985. In summary, readmitted Yes: 28533 and No: 42985, so there is a clear indication of the possibility of class imbalance. Since readmitted No value percentage is higher, both models (random forest and neural network) will do a good job in predicting the patient not readmitted value, but the value that is of interest readmitted: Yes. In this case, the specificity would be higher, and sensitivity would be lower, to improve the model performance the sensitivity value needs to be increased.

Here is the confusion matrix result

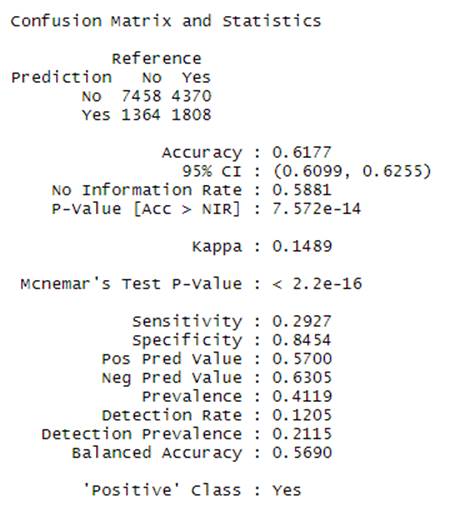


Figure : Confusion Matrix

Based on the above confusion matrix values it clear that the model is doing a good job of predicting patient nor readmitted (negative side) but not so good job of predicting readmitted: Yes (positive side). This class imbalance can be adjusted by increasing sensitivity by either under or over sampling or doing both. This can be done for increasing the model performance.

### Feature engineering - Encoding

Diagnosis related features diag\_1, diag\_2, and diag\_3 is categorical containing 9 categorical values. These 3 variables have been one hot encoded to identify which specific diagnosis has a higher impact. Based on feature importance only other and Neoplasm are only one with higher importance remaining are none, so instead of one-hot encoding, we can do target encoding. Also, one-hot encoding in this case is increasing the dimensions of the dataset, which could impact the model performance.

Various combinations of one-hot encoding and target encoding could be tried on various features and analyze if this increases the model performance.

Correlated variables – Some of the features in the dataset could be correlated with each other and might be similar in behavior, so removing one of the correlated variables could simplify the model further.

## Appendix

Attached is the R Markdown, this has been attached separately in blackboard along with this document.

