Assignment 3: Final Report – Week 8

Merrimack College

Machine Learning

Amol Gote

25th Oct 2020

# Executive summary

The main objective of this report is to provide the findings related to building a machine learning model which will predict if the patient is going to be readmitted. From a business perspective for a hospital if the patient has been readmitted then it is an indicator that prior treatment did not work. The Patient admitted beyond 30 days could be due to the patient’s immune system, but less than 30 days is something to do with prior treatment. So, avoiding patient readmittance is critical for a hospital from a reputation perspective as well as financially.

The data that was used for analysis comprised records from 130 different hospitals across the united states. The Dataset comprises of 50 different variables and 71518 patient records. Datasets consist of patient’s demographic information (age, race, gender), procedures performed, 3 levels of diagnosis, 2 test results, and close to 24 medications that were provided to patients.

The Business objective, in this case, was to identify if the patient will be readmitted, so the target variable was readmittance and values could be “Yes” or “No”. Since this is classification problems, and the predictor variable is categorical **random forest** was the first natural choice to build the ML model. Apart from the random forest, the neural network algorithm is the most popular machine learning algorithm for classification hence was used for predicting the patient readmittance. Below is the prediction percentage accuracy rate for random forest and neural network models

|  |  |
| --- | --- |
| **Algorithm** | **Prediction Percentage** |
| Random Forest | 62.152% |
| Neural Network | 63.211% |

Table 1: Algorithm prediction %

Figure 1: Algorithm prediction percentage

The Neural network model prediction percentage is slightly better than that of random forest, hence the final **recommended model** is the **neural network**. In general, the model's prediction accuracy rate is average and not to the acceptable range. The typical acceptable range is 70-80% or 0.7-0.8.

The top 10 variables which contribute significantly towards predicting the patient readmittance are as below in a ranked manner

|  |  |
| --- | --- |
| **Sr No** | **Important Variable for predicting readmittance** |
| 1 | Number of lab procedures |
| 2 | Number of medications |
| 3 | Time in hospital |
| 4 | Number of diagnoses |
| 5 | Number of procedures |
| 6 | Diagnosis 2 |
| 7 | Diagnosis 1 |
| 8 | Diagnosis 3 |
| 9 | Age |
| 10 | Insulin |

Table 2:Top 10 Important Variable for predicting readmittance

From demographic attributes, age is the only top contributing factor. The number of diagnoses and diagnoses identified, play an important role in identifying the patient’s readmittance. From medication variables, the insulin variable is of prime significance.

# Data & Approach

The data available for analysis contains records from the year 1999-2008 at 130 US hospitals and delivery networks. It contains 50 variables; the data set contains 71,518 records. Data comprises of patient’s demographic information which includes race, gender, age weight. From an age standpoint, most of the patients that were admitted were between the age of 40 to 80 years old. 70-80 years old were highest in number in age segments of 10. From a race perspective Caucasians were close to 75%, followed by African Americans close to 18% and the remaining were Hispanic, Asian, and others. The male and female ratio across the dataset is almost equal. Out of 24 medications provided to patients, most of the medication data is non-variant except insulin and metformin, all non-variant medications variables were excluded from the analysis.

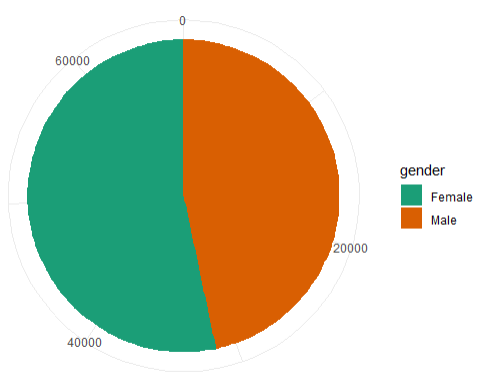
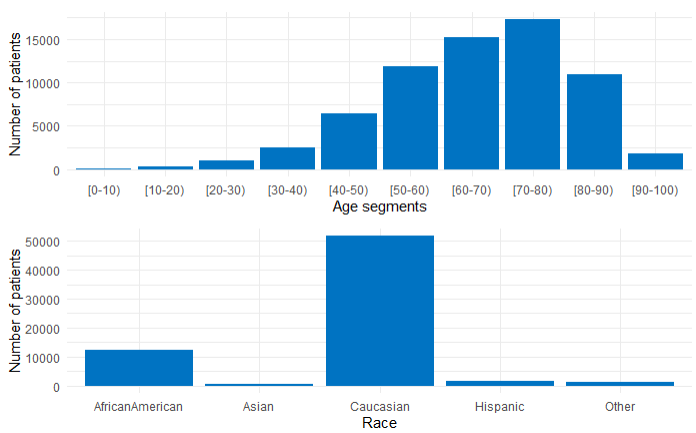
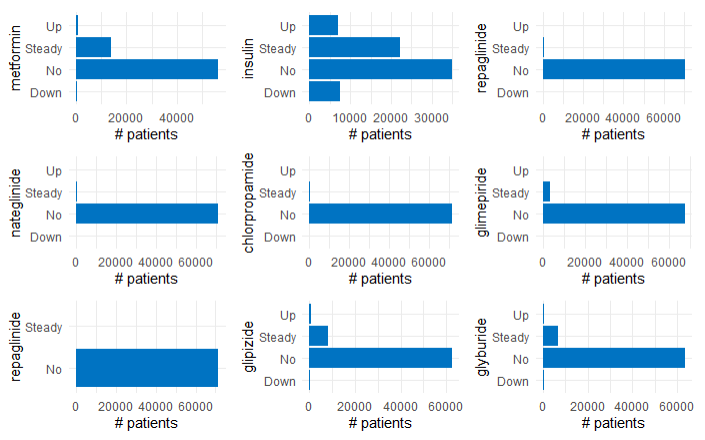


Figure 2: Age, race, and gender Distribution



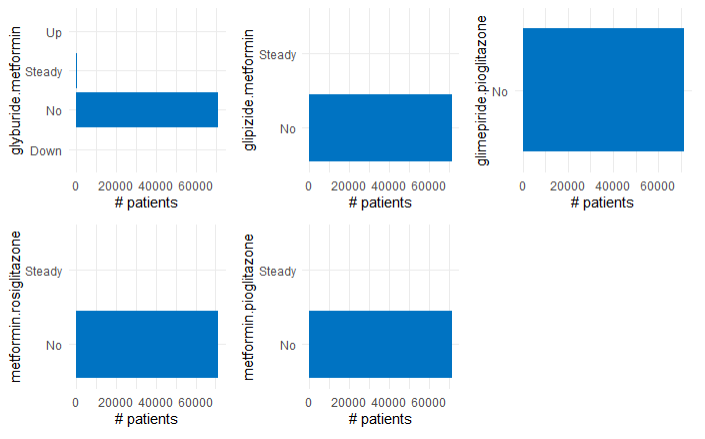
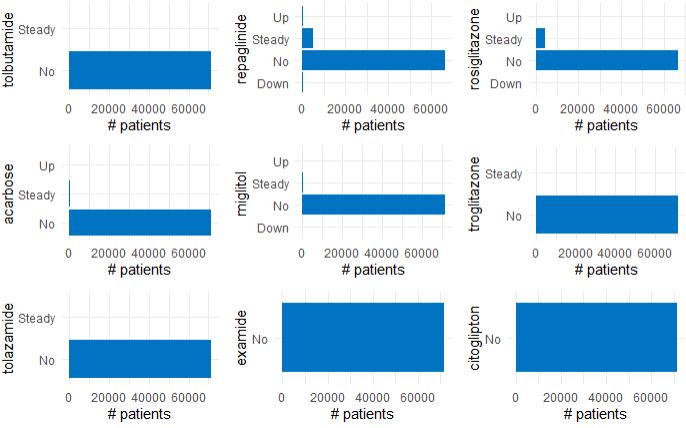


Figure 3: Medication Non-Variance

## Feature engineering

### Missing value treatment

* From the total records, 68,665 records had a missing value for the weight which is close to 90%, so the weight variable was excluded from the analysis.
* 34,477 missing values for the medical specialty which is close to 50% of the dataset size hence dropping from the analysis.
* Payer\_code - 31043 missing values which is a large proportion hence dropping the variable
* Race, Gender, diag\_1, diag\_2, diag\_3 – For all these variables there were few records with missing values, so these records were excluded from the overall dataset under analysis.
* Dropping NA values drops the dataset size to 68,358

### Encoding

Following variables were one hot encoded

* Gender
* Change of medication

Diagnosis 1,2 and 3 were also one-hot encoded, but as part of model performance improvement, they were target encoded, as one-hot encoding was leading to an increase in the size of the dimensions.

Diagnosis codes 1,2 and 3 which are categorical variables, had a lot of categorical values (700 unique ICD codes), these were consolidated to 9 unique groups/value (See Figure 5 for consolidation and distribution). Age was ordinal-encoded to 4 categorical values, this includes young (0-40 years), middle-aged (40-60 years), old (60-80 years), and older (80+ years). The race was target encoded. The admission source was consolidated from 26 unique values to 5 categorical values and was target encoded (See Figure 4). The admission type was consolidated to 4 unique values and target encoded (See Figure 4). The discharge disposition id again was consolidated to 4 categorical values and was target encoded. The business objective is to identify patient readmittance, so the patients with expired disposition id were excluded from the final dataset for analysis. Test results like A1Cresult and max\_glu\_serum was target encoded.

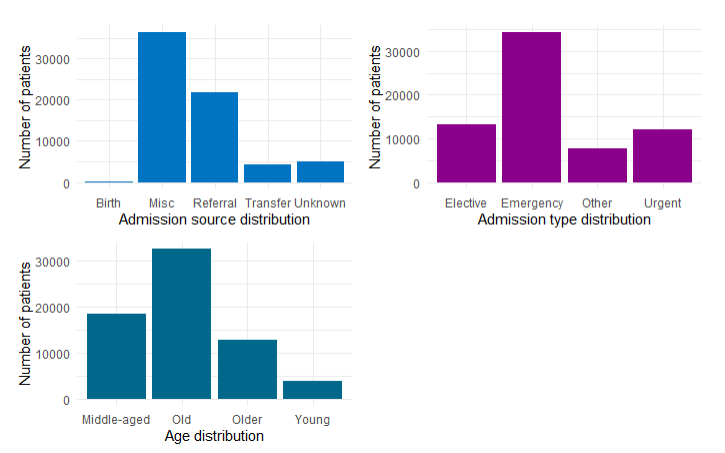


Figure 4: Age, Admission type, and source distribution

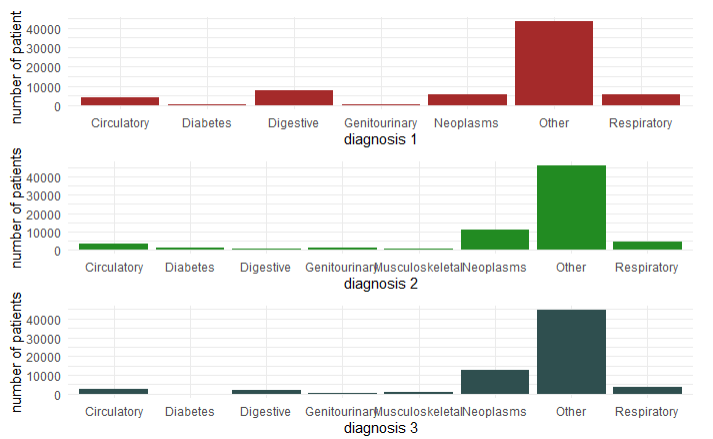


Figure 5: Diagnosis 1, 2, and 3 consolidations and their distribution across the dataset

## Approach

For building the machine learning model the business objective is identified which is to minimize the patient readmittance, hence supervised learning was chosen. Since the target variable is a qualitative/categorical one, it was a classification problem hence random forest was chosen. To compare the performance with the random forest model, the neural network was the second choice as it is the most popular machine learning algorithm for classification, and the third choice was gradient boosting machine (GBM). Reason for choosing gradient boosting (GBM) as the third choice was, it is based on the decision trees, but it is different from the random forest, the difference is in how the trees are built, random forest builds tree independently and combines the results at end of the process while gradient boosting builds one tree at a time and combines the result along the way. Along with these 3 algorithms logistic regression was evaluated as well to understand if the model performance improves with it. Model built with logistic regression results had poor performance results compared to that of random forest or neural network, it had an AUC value of 0.5570 (accuracy = 55.70%) which is comparatively less hence the model was discarded.

It was a 2-step approach, the first step was to identify the baseline model performance for both Random forest and neural network. Once the baseline performance has been identified then various techniques were applied to increase the model performance, these techniques include

1. Removing class imbalance by running with over-sampling, under-sampling, and SMOTE.
2. Reducing the dimensions.
3. Removing the data from the dataset which is of not of significance as far as target variable is concerned, one example is removing the records of expired patients, as they would never be readmitted.

# Detailed Findings

The neural network prediction percentage is slightly better than the random forest. The neural network prediction percentage is **63.21%** while that of random forest is **62.15%.** So, the final recommended model would be the neural network model. GBM prediction percentage was at par along with random forest and neural network, it was 63.57%, but since it is not dramatically different from neural network, hence the final recommended model was neural network as it better suited for classification compared to that of GBM.

## Feature importance

Below are the important features which have a significant impact on patient readmittance.

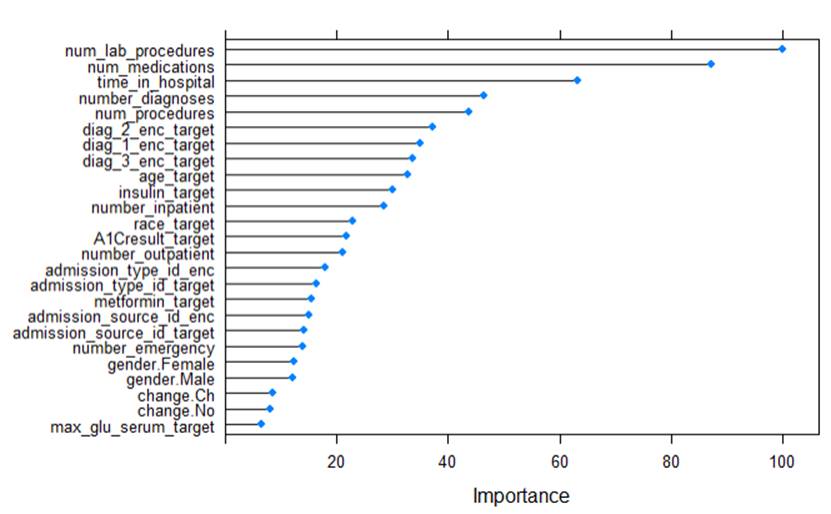


Figure : Feature Importance

Figure : Feature Importance %

Based on the variable importance below is the box plot for some of the quantitative variables.

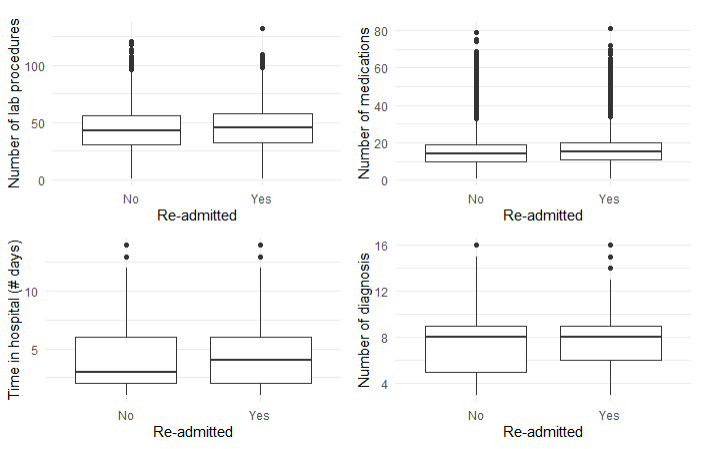


Figure : Box plot for important variables

Based on the above important variables for quantitative variables, if the patient has below data points then it likely that patient would get readmitted.

1. The number of lab procedures around 45 or greater.
2. The number of medications around 15 or greater.
3. Spent 4 days or greater in hospital
4. The number of diagnoses around 8.

For qualitative variables, below are the indicators for readmittance

1. For diagnosis 1,2 or 3, if the diagnosis is Other or Neoplasm
2. For Age between 50 to 90
3. The admission type is emergency
4. The admission source is Misc.
5. The race is Caucasian or African American

# Evaluation

Evaluation of both models was done using

* ROC Curves and AUC
* Calibration curve

## Random forest

Below is the ROC curve for the random forest, it shows the performance of the random forest model. A good classifier model will have the ROC 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.

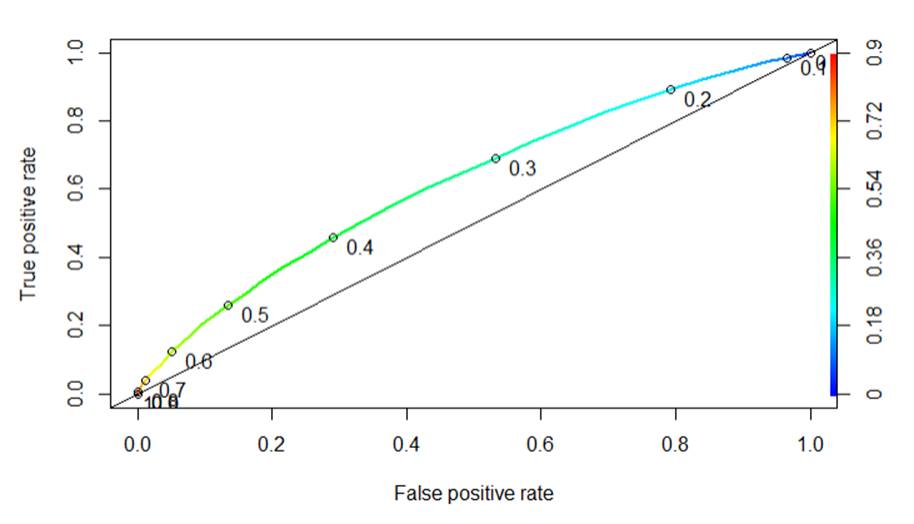


Figure 9: ROC Curve for random forest

Below is the calibration curve for the 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.

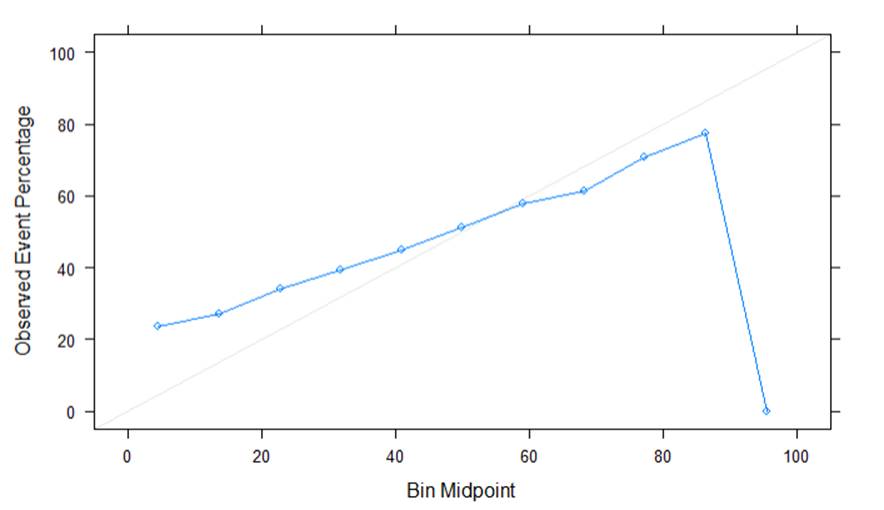


Figure 10: Calibration curve

## Neural network

Below is the roc curve for the 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.

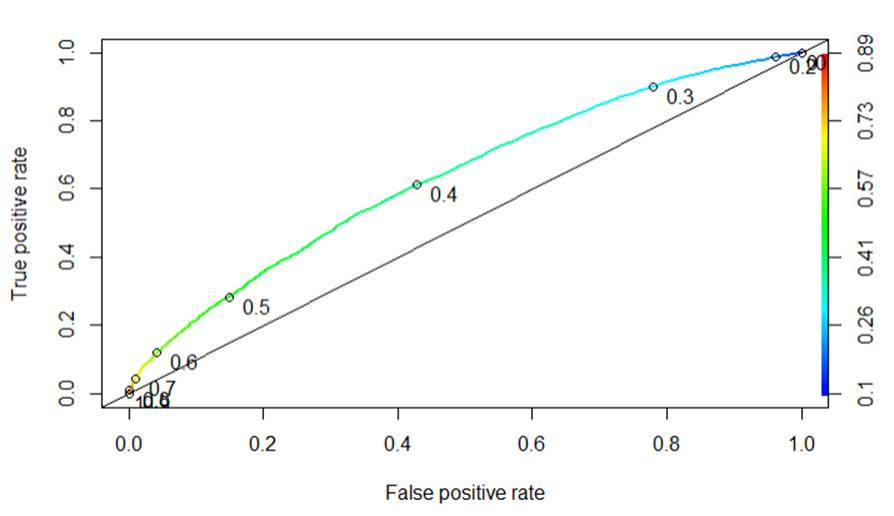


Figure 11: Roc curve for neural network

Below is the calibration curve with the neural network.

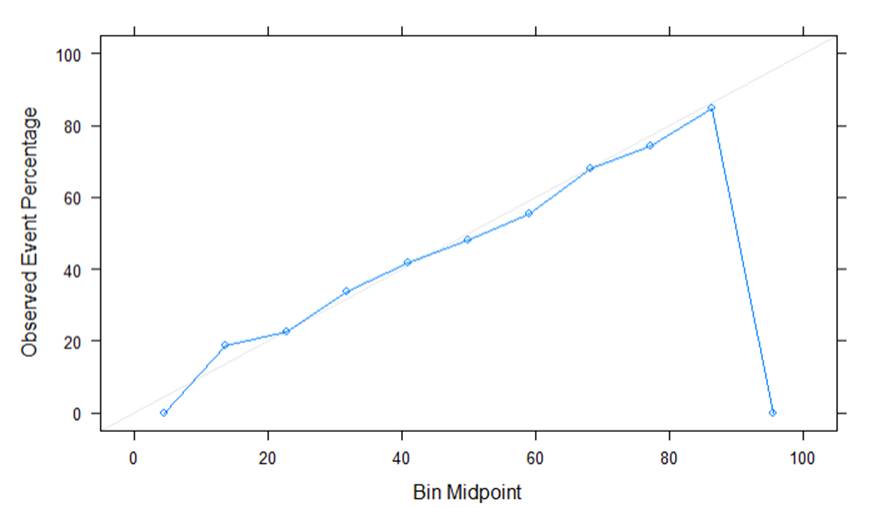
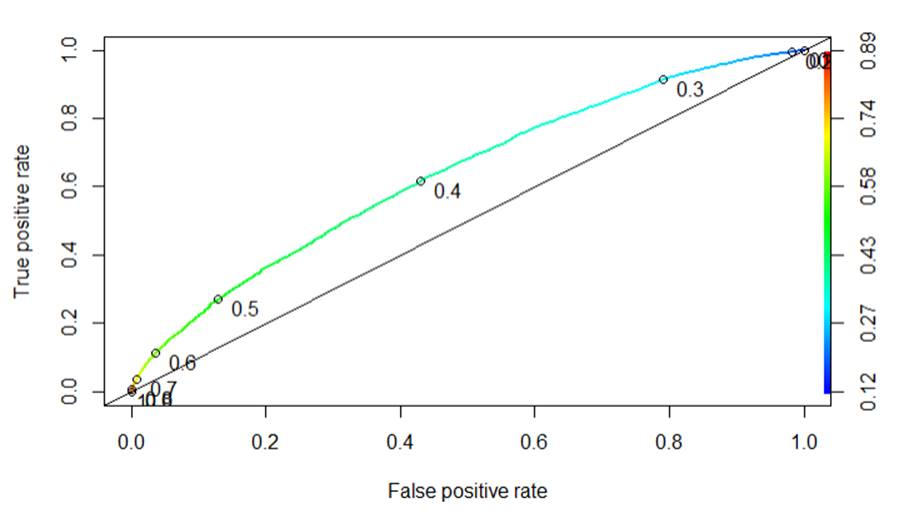


Figure 12: Calibration curve

## Gradient Boosting Machine

Below is the ROC curve for the GBM. AUC value that was achieved with multiple runs was 0.6357. The AUC value is on par with the neural network.



## Model improvement steps

### Feature Engineering

1. Removed the records for which the patient was marked expired.
2. Reverted from one-hot encoding to target encoding for diagnosis 1,2, and 3. The diagnosis code was causing an increase in the number of dimensions.

### Class Imbalance

For patient readmittance representation in the dataset, there is a clear class imbalance. 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: 28,533 and No: 42,985. 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 of interest is 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. Below figure 7 displays the 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 oversampling or doing both.

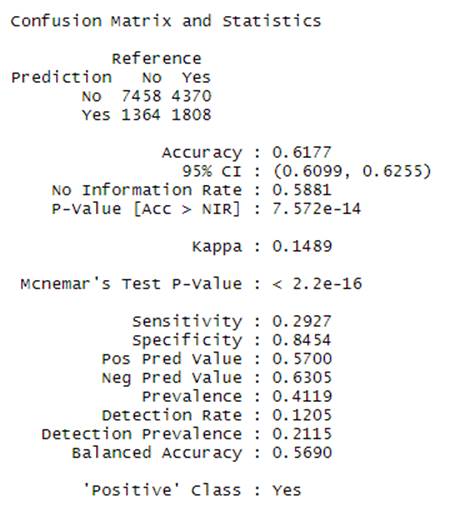


Figure : Confusion matrix

To improve the model performance for both neural network and random forest, methods of under, oversampling, and SMOTE were used. These methods were used with 2 different packages, the first one was ROSE and another one was caret. So, in general, six combinations that were tried

1. ROSE Package
   1. Up Sampling
   2. Down Sampling
   3. Both
2. Caret
   1. Up Sampling
   2. Down Sampling
   3. SMOTE

Below is the bar plot and comparison across the above six options along with the original run. The bar plot shows the AUC values across various options. All the above six options were run with targeted encoding for diagnoses 1,2 and 3.

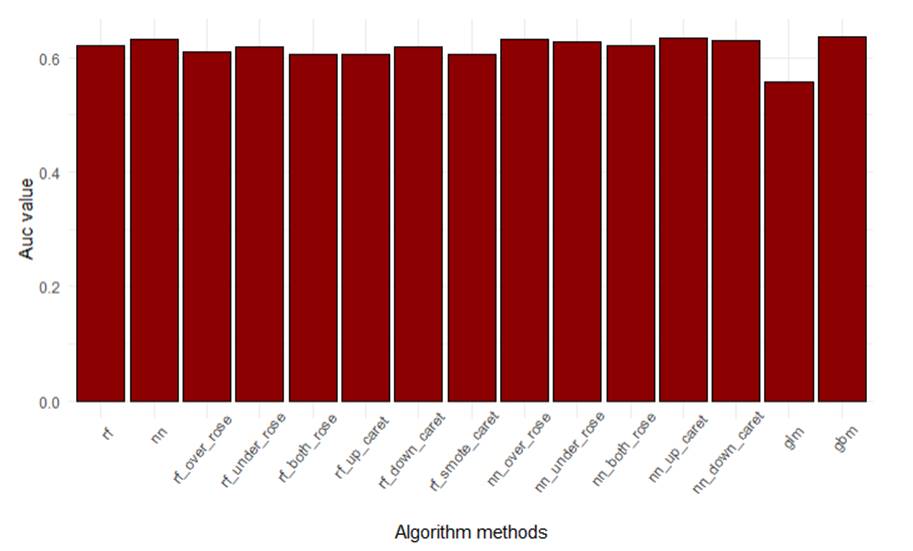


Figure 14: Model perf comparison based on class imbalance remediation

Below are various AUC values for model performance for remediating class imbalance.





Figure 15: AUC values using various approaches (Class Ibalance)

Original AUC values for the random forest and the neural network were 0.62152 and 0.63211, respectively. Clearly, random forest AUC values have not improved by under-sampling, down-sampling, or both with both ROSE and caret package approaches. There is a slight marginal improvement with a neural network with up-sampling in both approaches ROSE and caret. In general remediating class imbalance did not improve the model performance dramatically, so even if the model without class imbalance approach should work as well.

The recommended final model is the neural network model with an accuracy of 63.21%.

## Future evaluation

When new data arrives, the model needs to be periodically trained and deployed to the production system. Data tend to vary with time, so it is recommended to train the model periodically, this is sometimes referred to as continual learning, where the model gets trained and deployed continuously.

The current model was built with computing constraints of personal laptops, so the model was trained with a minimal set of records, but in a commercial approach, the model will be trained using commodity hardware or cloud-based approach where there would not be any computing constraints which will not provide any boundaries on the sample size to train the model.

# Appendix

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



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

