Assignment 3: Final Report – Week 8

Merrimack College

Machine Learning

Amol Gote

25th Oct 2020

# Executive summary

The main objective of this report is providing 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. Patient admitted beyond 30 days could be due to 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 reputation perspective as well as financially.

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

Business objective is this case was to identify if the patient will be readmitted, so target variable was readmittance and values could be “Yes” or “No”. Since this is classification problems, and predictor variable is categorical **random forest** was first natural choice to build the ML model. Apart from the random forest, Neural networks 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 : Algorithm prediction %

Figure : Algorithm prediction percentage

Neural network model prediction percentage is slightly better than that of random forest. In general model prediction percentage rate is average and not to the acceptable range. Typical acceptable range is 70-80% or 0.7-0.8.

Top 10 variables which contribute significantly towards predicting the patient readmittance are as below in 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 :Top 10 Important Variable for predicting readmittance

From a demographic attribute perspective age is only top contributing factor. Number of diagnosis and diagnosis identified play an important role in identifying the patient readmittance. From a medication variables 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 71518 records. Data comprises of patient’s demographic information which includes race, gender, age weight. From age perspective majority of the patients that were admitted were between age of 40 to 80 years old. 70-80 years old were highest in number in age segments of 10. From race perspective Caucasian were close to 75%, followed by African Americans close to 18% and remaining were Hispanic, Asian and others. Male and female ratio across 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 analysis.

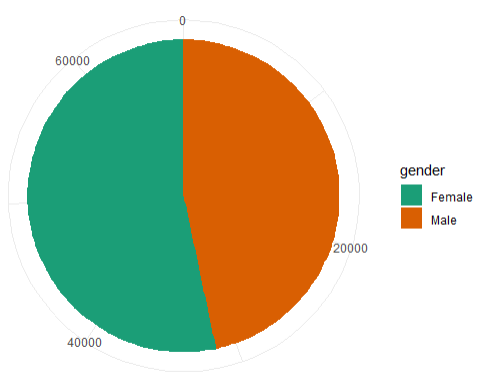
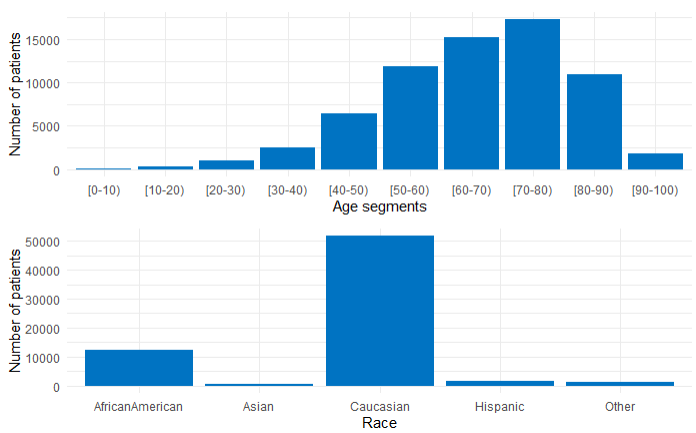
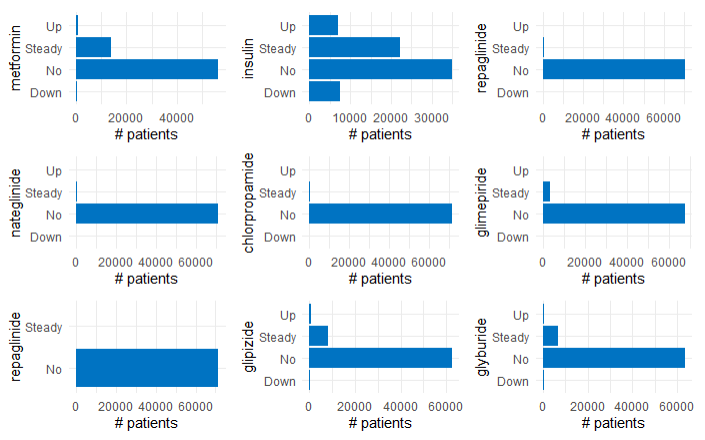


Figure : Age, race, and gender Distribution



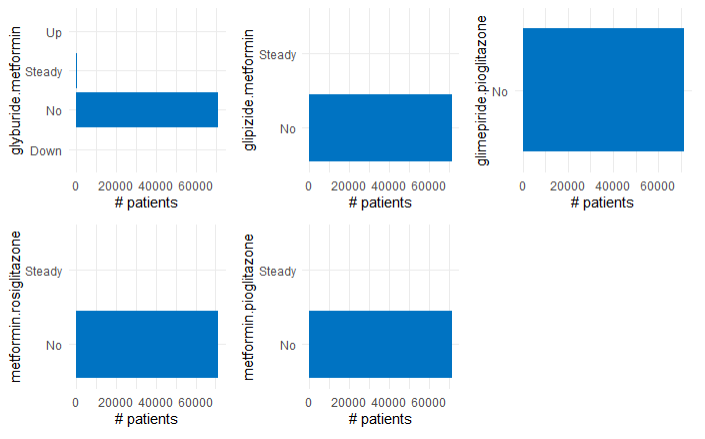
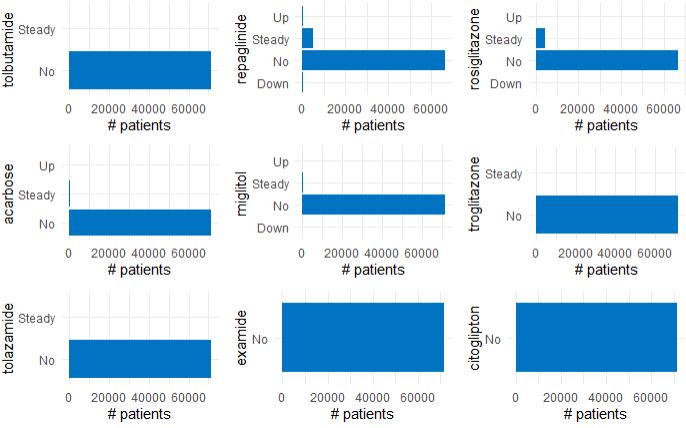


Figure : Medication Non-Variance

## Feature engineering

### Missing value treatment

* From the total records 68665 records had missing value for weight which is close 90%, so weight variable was excluded from analysis.
* 34477 missing values for medical specialty which is close 50% of the dataset size hence dropping from 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 overall dataset under analysis.
* Dropping NA values drops the dataset size to 70233 from 68358

### 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 increase in the size of the dimensions.

Diagnosis codes 1,2 and 3 which are categorical variable, had 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; race was target encoded. Admission source was consolidated from 26 unique values to 5 categorical values and was target encoded (See Figure 4). Admission type was consolidated to 4 unique values and target encoded (See Figure 4). Discharge disposition id again was consolidated to 4 categorical values and was target encoded. Business objective is to identify patient readmittance, so the patients with expired disposition id were excluded from the final dataset for analysis. Tests results like A1Cresult and max\_glu\_serum were as well target encoded.

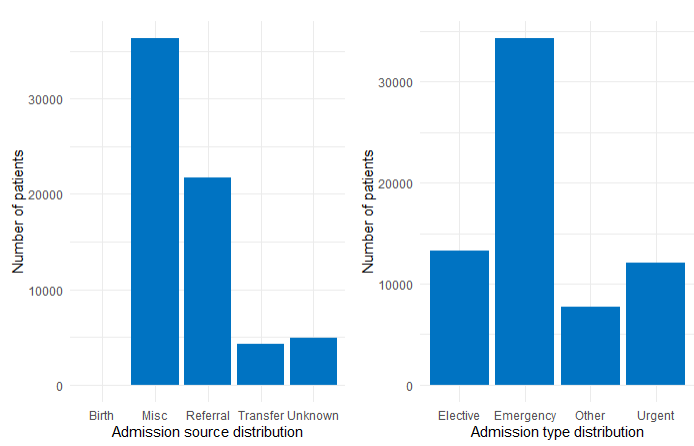


Figure : Admission type and source distribution

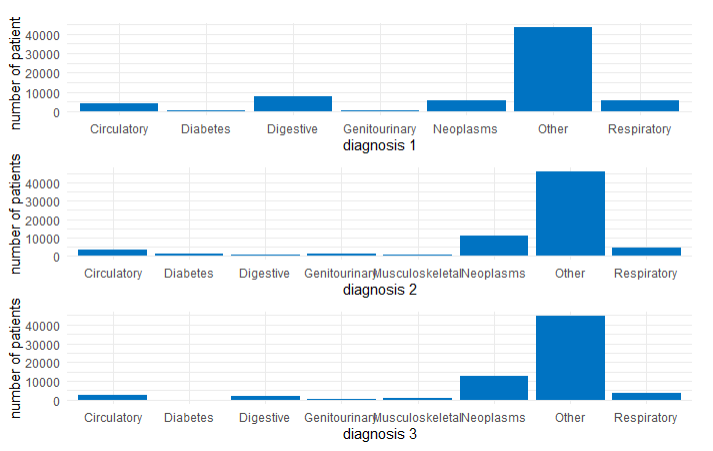


Figure : Diagnosis 1, 2 and 3 consolidation and their distribution across dataset

## 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 qualitative/categorical one, it was a classification problem hence random forest was chosen. To compare the performance with of random forest model, neural network was the second choice. Along with these 2 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, hence discarded.

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

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 e.g. is removing the records of expired patients, as they would never be readmitted.

# Detailed Findings

Neural network prediction percentage is slightly better than random forest. Neural network prediction percentage is **63.21%** while that of random forest is **62.15%.**

## Feature importance

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

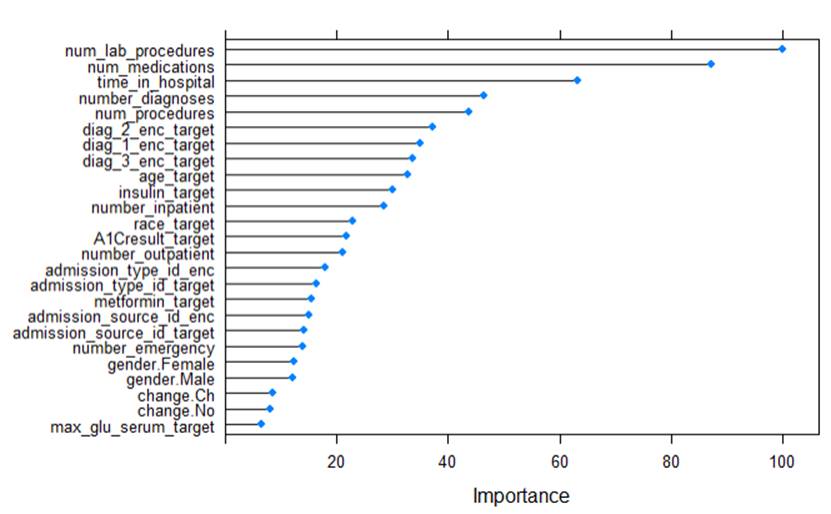


Figure 6: Feature Importance

Figure 7: Feature Importance %

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

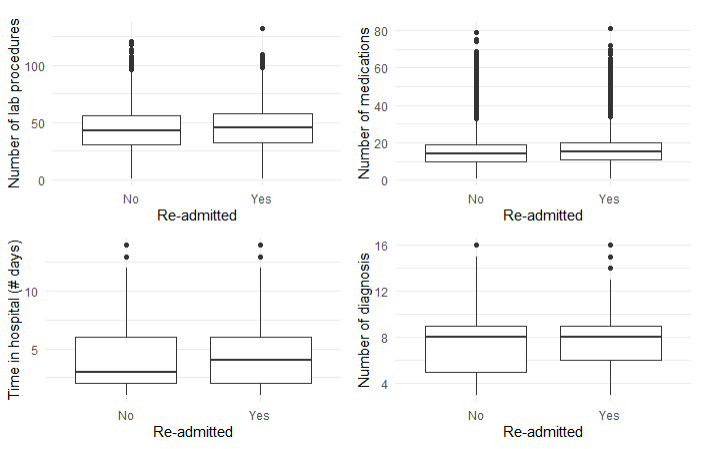


Figure 8: Box plot for important variables

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. Number of lab procedures around 45 or greater.
2. Number of medications around 15 or greater.
3. Spent 4 days or greater in hospital
4. Number of diagnosis 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. Admission type is emergency
4. Admission source is Misc.
5. 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 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.

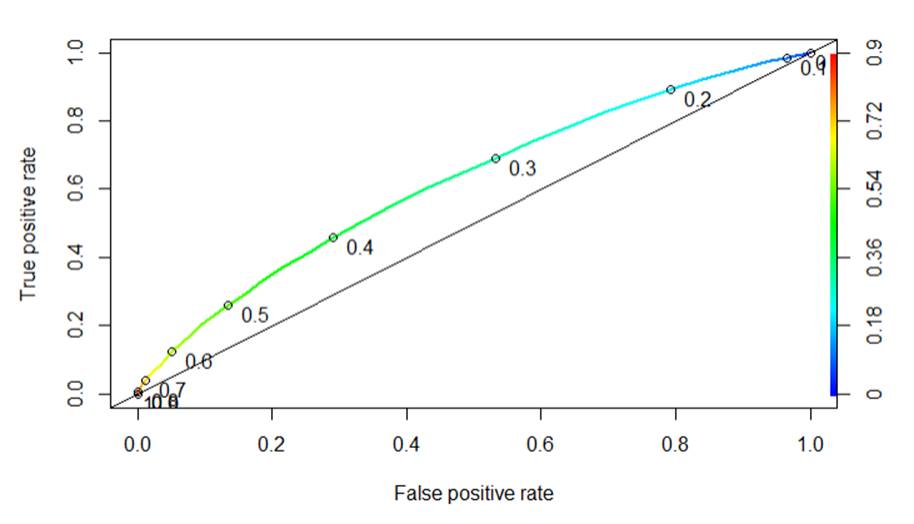


Figure : ROC Curve for random forest

Below is the 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.

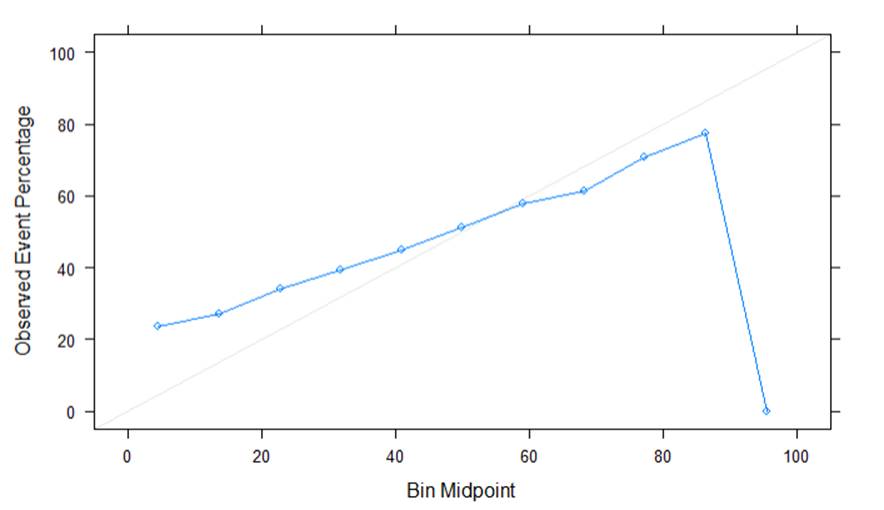


Figure : Calibration curve

## Neural network

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.

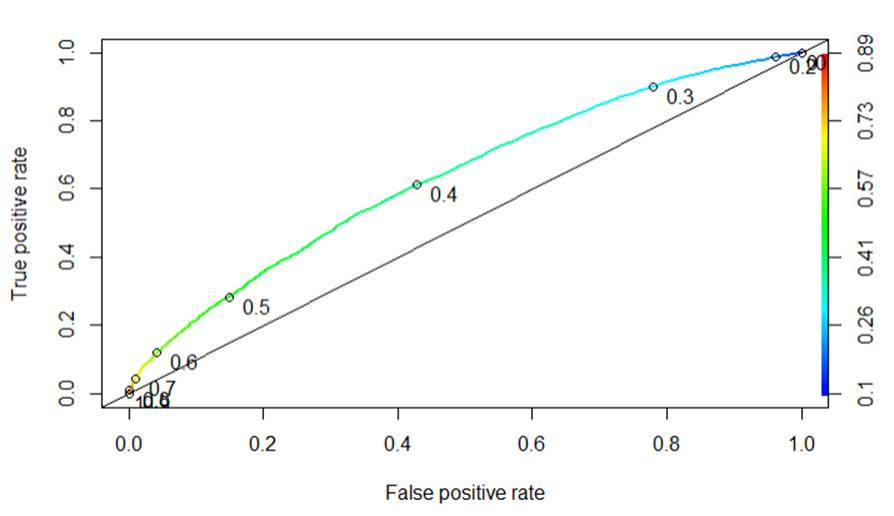


Figure : Roc curve for neural network

Below is the calibration curve with neural network.

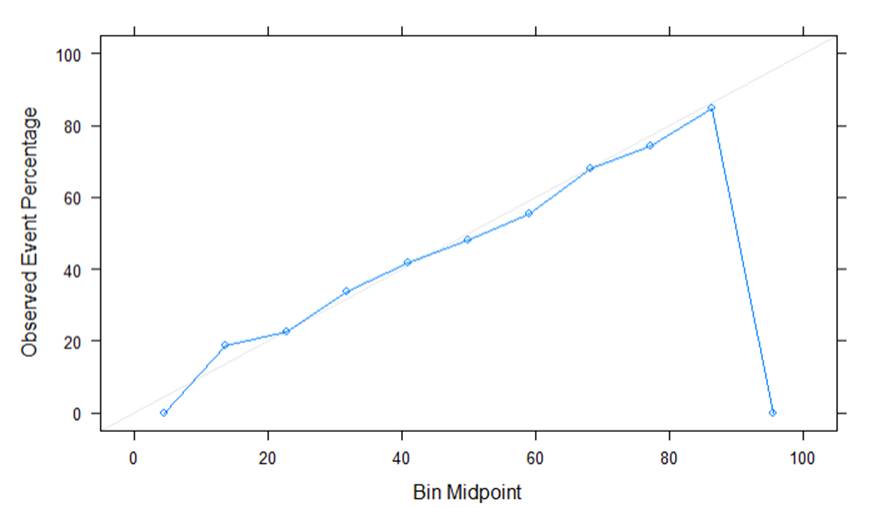


Figure : Calibration curve

## Model improvement steps

### Feature Engineering

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

### Class Imbalance

For patient readmittance representation in the dataset there is 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: 28533 and No: 42985. 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. 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 over sampling or doing both.

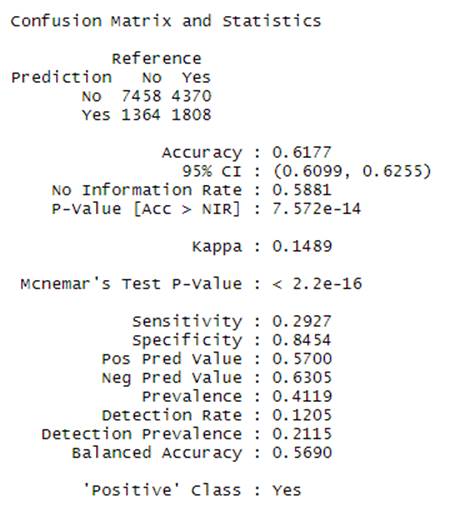


Figure 13: Confusion matrix

To improve the model performance for both neural network and random forest, methods of under, over sampling and SMOTE were used. These methods were used with 2 different packages, first one was ROSE and other one was caret. So, in general there were 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 bar plot and comparison across various above six options along with original run. Bar plot shows the AUC values across various options. All the above six options were run with targeted encoding for diagnosis 1,2 and 3.