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.

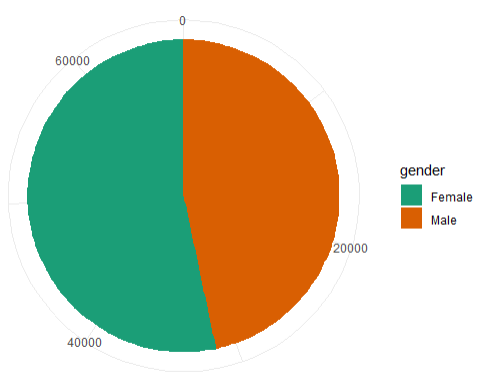
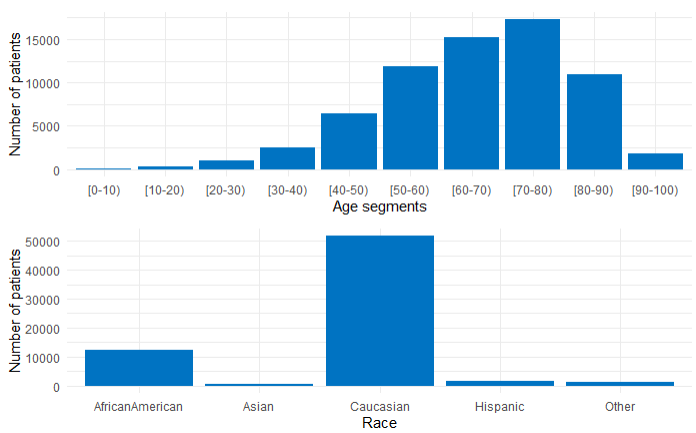
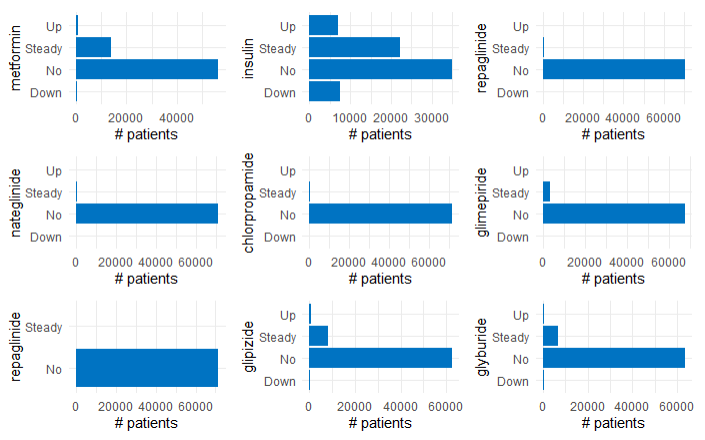


Figure : Age, race, and gender Distribution



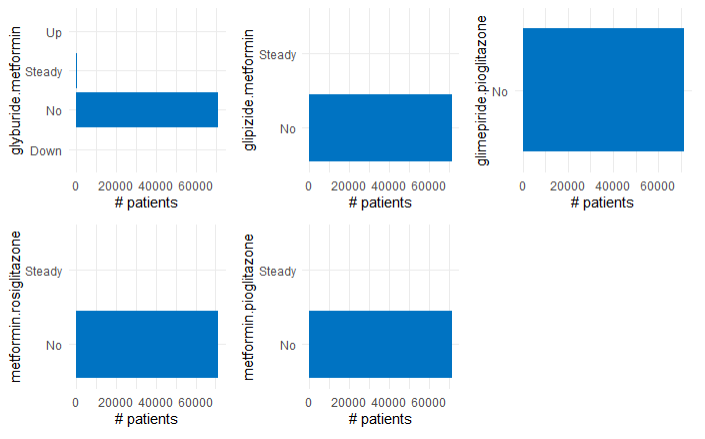
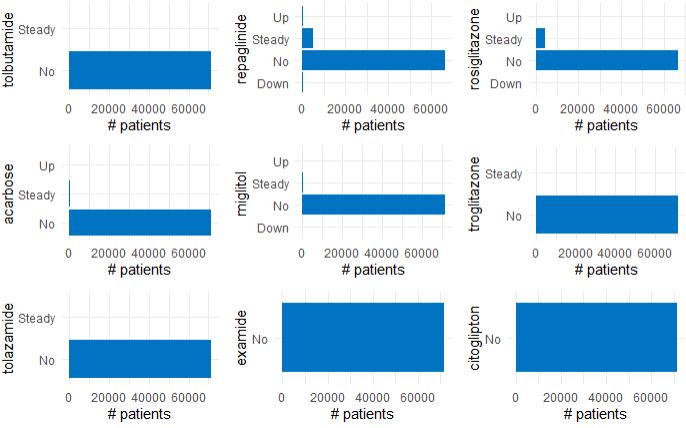


Figure : Medication Non-Variance

## 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. 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. Admission type was consolidated to 4 unique values and target encoded. 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.

# Detailed Findings

# Evaluation