Capstone 2 Hospital Readmission Prediction for Diabetes Patients- Milestone Report

The Centers for Medicare & Medicaid Services, CMS which is part of the Department of Health and Human Services (HHS) has created many programs to improve the quality of care of patient as the healthcare system moves toward value-based care. Hospital Readmission Reduction Program (HRRP), which is one of them, reduces reimbursement to hospitals with above average readmissions as penalization. This enforces the hospitals under this program to take steps to create interventions to provide additional assistance to patients with increased risk of readmission.

One patient population that is at increased risk of hospitalization and readmission is that of diabetes. Diabetes is a medical condition that affects approximately 1 in 10 patients in the United States. So it would be beneficial to use predictive modeling from data science to help identify patients with a likelihood of hospital readmission.

Dataset:

Data is available in UCI machine learning repository (link). The data consists of over 100000 hospital admissions from patients with diabetes from 130 US hospitals between 1998 and 008.

Data Wrangling:

The Centers for Medicare & Medicaid Services, CMS which is part of the Department of Health and Human Services (HHS) has created many programs to improve the quality of care of patient as the healthcare system moves toward value-based care. Hospital Readmission Reduction Program (HRRP), which is one of them, reduces reimbursement to hospitals with above average readmissions. For those hospitals which are currently penalized under this program, one solution is to create interventions to provide additional assistance to patients with increased risk of readmission. I propose to use predictive modeling from data science to help identify patients with a risk for hospital readmission.

Datasets that are available for this project are 1) Diabetic data with all the details of the patients getting admitted and 2) IDS Mapping that has mapping values for some of the columns from diabetic data

I. Cleaning and Consolidating the Data

As part of consolidating patient data some unnecessary attributes were dropped to reduce dimensionality. There were a significant number of categorical data types and some of the numerical data were also categorical data. One Hot encoding was implemented to convert those to numerical data with significance in predictive modelling.

One specific categorical variable 'medical_specialty' has around 75 values, when one hot encoded would add close to 75 columns to the feature list. Since the number of records

pertaining to a majority of these values is very less, I decided to leave the top 20 variables for encoding and aggregate the remaining to a single value 'others'.

The dependent variable 'readmission_label' is a classifier with '0' or '1' value. I created 'readmission_label' based on the value of 'readmitted' variable which has three values 'No','<30' and '>30'.

II. Missing Values

Some columns in the dataset had missing values and there were a few inconsistencies in notation that were adjusted for ease of future analysis. Dropna and fillna were used to drop if the number of rows were insignificant to the data and to replace using mean value as applicable for each of the specific case.

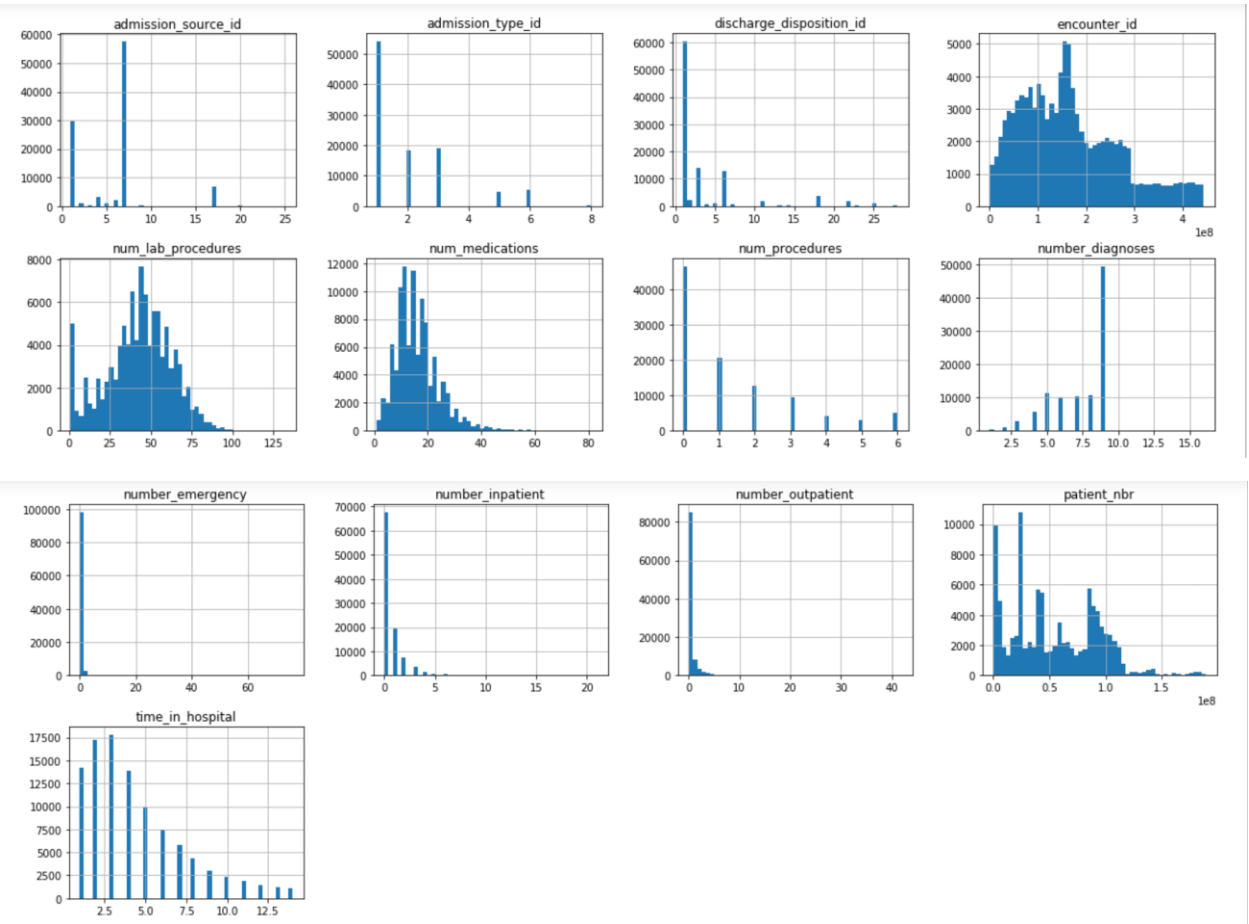
III. Outliers

There were not much significant outliers to be worked upon

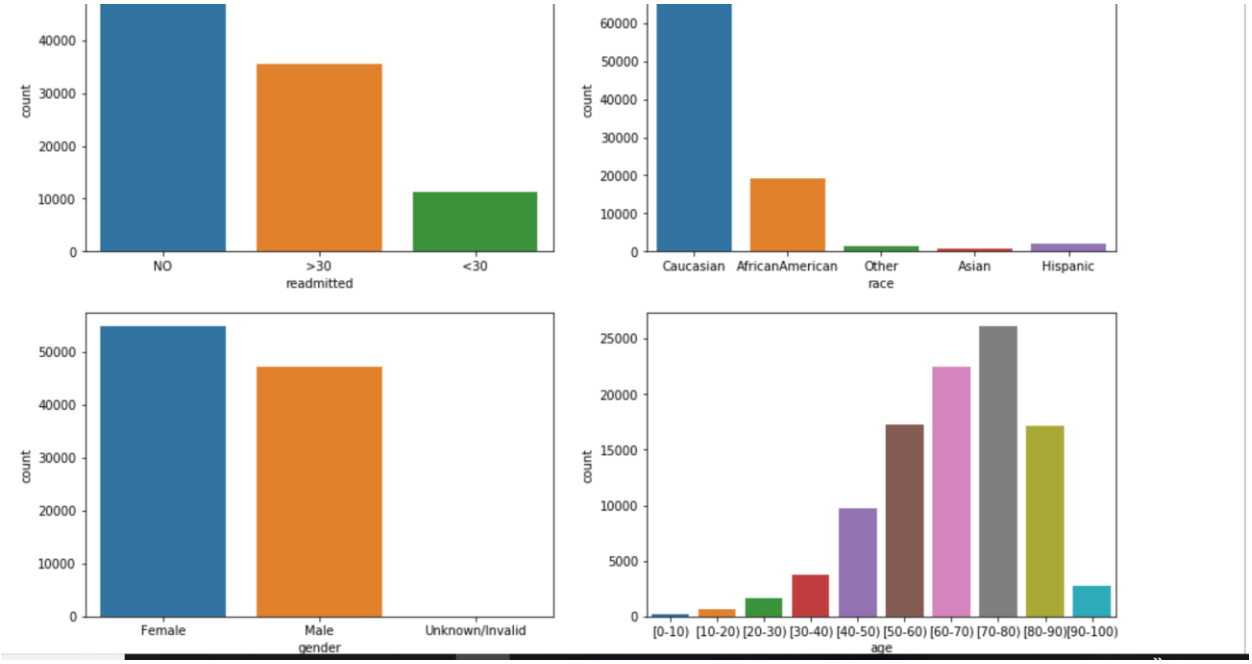
Data Visualization and Inferential Statistics

Exploratory data analysis is done to get a basic sense about the data. Since there are numerous variables, finding out any correlation will help in reducing dimensionality of data.

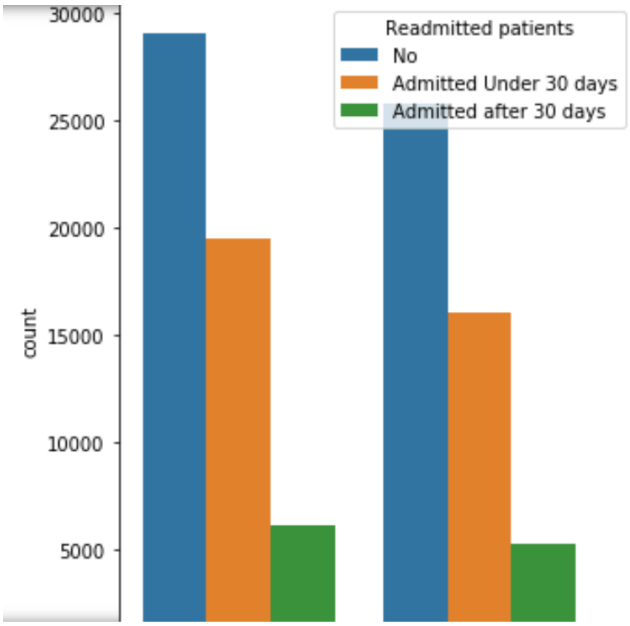
Below given visualization will help understand the dataset better.



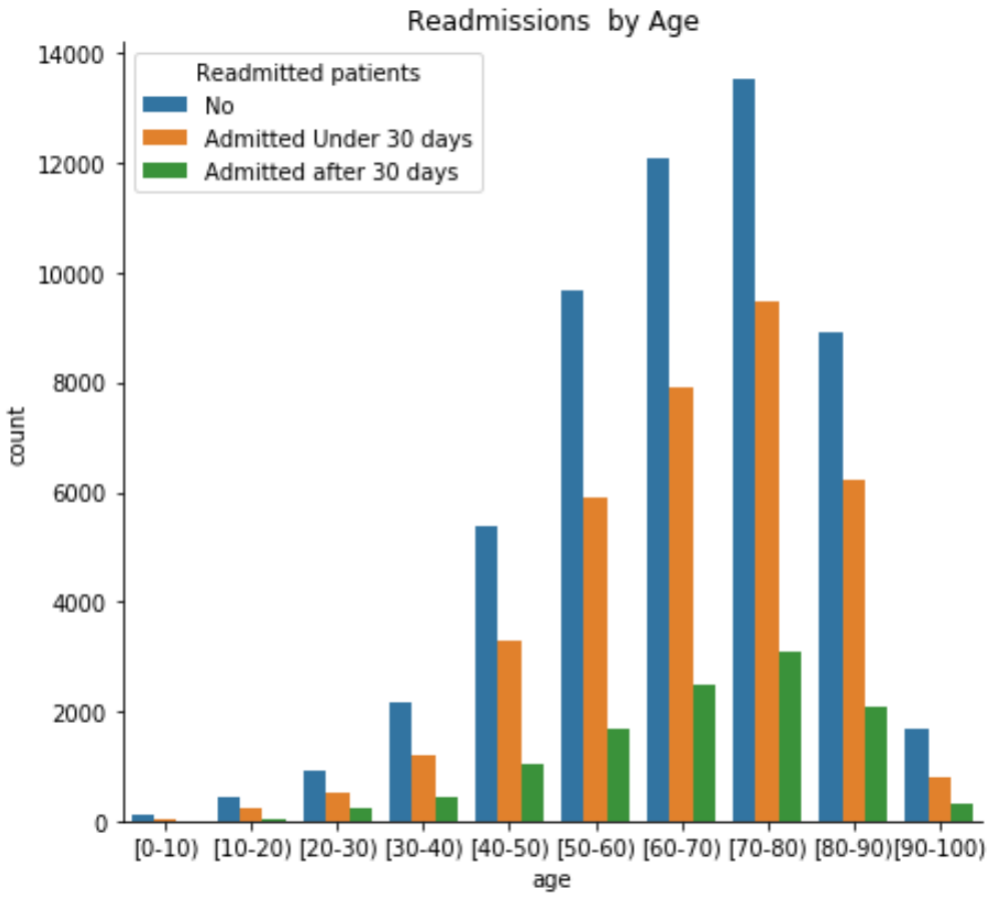
It is interesting to note that the number of lab procedures and number of medications taken, follow a bell curve in the population getting admitted to the hospital. It would be worth the effort to check for influence.



Readmission based on 'race' does not indicate that a specific race has more chances of readmission, as it could also be just reflective of ethnic distribution of population in that city. It is interesting to note that female population has more cases of re-admission. As expected, age pattern follows a bell curve.



It can be noted that readmission categorization follows similar pattern across both the genders.



All the 3 categories 'Admitted Under 30 days', 'Admitted after 30 days' and ' No readmissions' follow a bell curve peaking at 70-80s

From the exploratory data analysis and statistical inferences drawn, though some trends cane be drawn, no significant enough information is seen to proceed with the modeling. As next step, feature importance can be used to order the feature variables according to their significance, which would help the prediction.

Machine Learning Modeling

At this point , I would like to pause and ponder about class imbalance of the dataset.The training set is heavily skewed with ' no readmissions' at 88.61% and ' readmissions' at 11.38%. With a greater imbalanced ratio, the decision function favor the class with the larger number of samples, usually referred as the majority class.So the accuracy would be affected.The approach is to develop a model that helps draw out the relationship between the feature variables and the target variable. The data set would be split into training data and testing data. The training data would be used to train the different machine learning algorithmic models and and subsequently the model would be tested on the test data to see the accuracy of the prediction. Since the test data is unseen data, it would give clarity on the accuracy of the model. This type of learning falls under the category of supervised learning.

The models that are being used are

1.) Decision Trees
2.) Random Forest
3.) Logistic Regression
4.) Gradient Boosting Classifier
5.) Ada Boosted Classfier and
6.) K Nearest Neigbor(KNN)

Before training the models, a relook at the data indicates that the data set is heavily skewed with ' no readmissions' at 88.61% and ' readmissions' at 11.38%. With such highly imbalanced data, the decision function will favor the class with the larger number of samples, usually referred to as the majority class. So the accuracy would be affected.

For negating the class imbalance, both oversampling and undersampling would be done separately onlyon the training set, to balance the data and then train the model on both the resultant balanced data set. Based on the result, one set would be selected and the models would be trained on the resampled training set. The test data would be the original test data .

The following would be used to analyse the performance of the model:

1. Accuracy
2. Classification summary (recall, precision, etc.)
3. Confusion matrix
4. ROC curve