greatlearning

# PGPDSE FT Capstone Project – Final Report

**Project Group Info:**

| | |
|---|---|
| Batch details | PGP-DSE February 2021 |
| Team members | Vikas Kagawad, Athul Gopu, Nagarjuna Gottipati, Naman Diwakar |
| Domain of Project | Healthcare Analytics |
| Proposed project title | A Data Driven Approach to Predict theReadmission of Diabetic patient into the Hospital. |
| Group Number | 01 |
| Team Leader | Vikas Kagawad |
| Mentor Name | Mr. Srikar Muppidi |

Date: 19-08-2021

Signature of the Mentor

Signature of the Team Leader

# Table of Content

## 2.0 Industry Review:

### ❖ Current practices:

- Healthcare Analytics primarily involves the exploration of actionable insights from sets of patient data collected from four areas within healthcare:
  - Claims and cost data
  - Pharmaceutical and R&D data
  - Clinical data collected from electronic medical records (EHRs)
  - Patient behavior and sentiment data
- Healthcare industry collects and process patient medical data in huge volume, diverse structure, and real-time flow of data. With the rise of technology, both from the diagnosis and monitoring, storage and analysis, novel solutions are now available to better address challenges like non-invasive screening, tailor-made treatment, and hospital readmissions
- One of the alarming concerns of health care with the current practices/protocols is the management of hyperglycemia patient. This recognition has led to the development of formalized protocols in the intensive care unit (ICU) setting with rigorous glucose targets in many institutions.
- However, the same cannot be said for most non-ICU inpatient admissions. Rather, various experience suggests that inpatient management is arbitrary and often leads to either no treatment at all or wide fluctuations in glucose when traditional management strategies are employed.
- Recent controlled trials have demonstrated that protocol driven inpatient strategies can be both effective and safe. As such, implementation of protocols in the hospital setting is now recommended. However, there are few national assessments of diabetes care in the hospitalized patient which could serve as a baseline for change.
- As the healthcare system moves toward value-based care, CMS has created many programs to improve the quality of care of patients. One of these programs is called the Hospital Readmission Reduction Program (HRRP), which 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.

## 2.1 Abstract:

Management of hyperglycemia in hospitalized patients has a significant bearing on outcome, in terms of both morbidity and mortality. However, there are few national assessments of diabetes care during hospitalization which could serve as a baseline for change. This analysis of a large clinical database (74 million

unique encounters corresponding to 17 million unique patients) was undertaken to provide such an assessment and to find future directions which might lead to improvements in patient safety. Almost 70,000 inpatient diabetes encounters were identified with sufficient detail for analysis. Multivariable logistic regression was used to fit the relationship between the measurement of HbA1c and early readmission while controlling for covariates such as demographics, severity and type of the disease, and type of admission. Results show that the measurement of HbA1c was performed infrequently (18.4%) in the inpatient setting. The statistical model suggests that the relationship between the probability of readmission and the HbA1c measurement depends on the primary diagnosis. The data suggest further that the greater attention to diabetes reflected in HbA1c determination may improve patient outcomes and lower cost of inpatient care.

## 2.2 Objectives:

- A diabetic patient suffering from Hyperglycemia is supposed to get quality health care service. Mismanagement of such patients forces them to get readmitted to hospitals within some days after their first visit. Sometimes such mismanagement can even lead to fatality.
- Readmission of patients is an ongoing real-world problem. Ripples of the problem are felt by both patients and health care service providers.
- For patients, it increases the burden of out-of-pocket expenditure, which in US has been rising yearly and currently stands at an average of 1200 dollars per capita.
- For health care providers, it damages the stature of their service and brings a dent in their efficiency.

# 3.0 Understanding the problem statement:

It is estimated that 9.3% of the population in the United States have diabetes, 28% of which are undiagnosed. The 30-day readmission rate of diabetic patients is 14.4 to 22.7 %. Estimates of readmission rates beyond 30 days after hospital discharge are even higher, with over 26 % of diabetic patients being readmitted within 3 months and 30 % within 1 year. Costs associated with the hospitalization of diabetic patients in the USA were $124 billion, of which an estimated $25 billion was attributable to 30-day readmissions assuming a 20 % readmission rate. Therefore, reducing 30-day readmissions of patients with diabetes has the potential to greatly reduce healthcare costs while simultaneously improving care.

## 4.0 Background Research:

- Diabetes is a chronic disease where a person suffers from an extended level of blood glucose in the body. Diabetes is affected by height, race, gender, age but a major reason is a sugar concentration. Diabetes affects approximately 1 in 10 patients in the United States. According to Ostling et al, patients with diabetes have almost double the chance of being hospitalized than the general population (Ostling et al 2017).

- 18% of the US GDP is spent on healthcare and we have a similar percentage of spent in most of the developed countries. Research suggests that 1 out of 3 adults have prediabetes. Of this group, 9 out of 10 don't know they have it. About 1.4 million new cases of diabetes are diagnosed in the United States every year itself and these figures are more alarming in developing nations. In the United States, type 2 diabetes is more prevalent for certain groups than for Caucasians. These people include:
    1. Native Americans
    2. African Americans
    3. Hispanics
    4. Asian Americans

- People with diabetes have twice the risk of death of any cause compared to people of the same age without diabetes. In 2014, diabetes was listed as the seventh leading cause of death in the United States. WHO estimates that 50 percent of people with diabetes die of cardiovascular diseases, such as heart disease and stroke.

- Hospital readmission is a high-priority health care quality measure and target for cost reduction. The burden of diabetes among hospitalized patients, however, is substantial, growing, and costly, and readmissions contribute a significant portion of this burden.

- The present analysis of a large clinical database was undertaken to examine historical patterns of diabetes care in patients with diabetes admitted to a US hospital and to inform future directions which might lead to improvements in patient safety. Reducing early hospital readmissions is a policy priority aimed at improving healthcare quality. In this case study we will see how machine learning can help us solve the problems caused due to readmission.

## 5.0 Application:

This problem is very crucial as it will alarm the hospital authority to take a good care of a patient. If certain medication and management process is being applied in a patient

case, the hospital or even the patient family can get the data of patient, put the data into the app and predict whether the patient's management show the symptoms of readmission of patient. If found high readmission chances, the hospital can incorporate some essential changes in their process and eventually may save a life.

Feature selection is carried out based on the relevance of the symbol. The relevance or the importance of features is based on:

- Number of unique values for categorical variables
- The categories in the categorial variables and the percentage distribution of eachcategory
- For numerical variables there is statistical test carried out. All the data is made normalso that parametric test is carried out in relation with the target variable.
- Some other variables might have to be taken out because of practical application andneeds.

The data is scaled and transformed based on the normality requirements, there is clearimbalance in the target variable. The dataset is oversampled and then Logistic Regression, Random Forest Classifier, ADABOOST Classifier, XGBOOST Classifier models are built etc.

## 6.0 Literature Survey

Here are some of the Publications, Application, past and undergoing research:

❖ Research papers:

- Paper 1: 5244347.pdf (stanford.edu)
- Paper 2: Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records (hindawi.com)
- Paper 3: Predicting 30-Day Hospital Readmission for Diabetes Patients using Multilayer Perceptron (thesai.org)

:

# 7.0 Dataset and Domain

## 7.1 Data Dictionary:

Our dataset is bank-additional-full.csv. The number of rows is 100000 and the number are columns are 50 including the target variable. The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medications, diabetic medications, number of outpatients, inpatient, and emergency visits in the year before the hospitalization, etc.

| Variable Name | Description |
|---|---|
| **Age** | Grouped in 10-year intervals: [0-10) [10-20) …… [90-100). |
| **Admission source** | Integer identifier corresponding to 21 distinct values. |
| **Admission type** | Integer identifier corresponding to 9 distinct values. |
| **Number of diagnoses** | Number of diagnoses entered the system. |
| **Change of medications** | Indicates if there was a change in diabetic medications. |
| **Discharge disposition** | Integer identifier corresponding to 29 distinct values. |
| **Diagnosis_1** | Primary diagnoses. |
| **Diagnosis_2** | Secondary diagnoses. |
| **Diagnosis_3** | Additional secondary diagnoses. |
| **Medical specialty** | Integer identifier of a specialty of the admitting physician. |
| **Gender** | Values: male, female and unknown/invalid. |
| **Readmitted** | 30 days,">30" if patient was readmitted in more than 30 days and "No" for no record of readmission |

## 8.0 Pre-Processing Data

| Variables | No. Of Missing Values | % Of Missing Values |
|---|---|---|
| **Race** | 1921 | 2.72 % |
| **Weight** | 67620 | 96.00 % |
| **Payer Code** | 30552 | 43.37 % |
| **Medical Specialty** | 33937 | 48.18 % |
| **diag_1** | 11 | 0.01 % |
| **diag_2** | 294 | 0.41 % |
| **diag_3** | 1225 | 1.73 % |

# 9.0 Project Justification- Project Statement, Complexity involved, ProjectOutcome –Commercial, Academic or Social value.

## 9.1 Project Statement:

This data has been prepared to analyze factors related to readmission as well as other outcomes pertaining to patients with diabetes. Information was extracted from the database for encounters that satisfied the following criteria.
(1) It is an inpatient encounter (a hospital admission).
(2) It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
(3) The length of stay was at least 1 day and at most 14 days.
(4) Laboratory tests were performed during the encounter.
(5) Medications were administered during the encounter.
The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

## 9.2 Complexity Involved:

- Interpretability of model is very important: Interpretability is always important in health care domain if model predict that some patient will readmit but can't explain why it came to this conclusion the doctor will be clueless about such decision and doctor won't be able to tell the patient why he needs to readmit practically it will create lots of inconvenience to doctor as well as patient.
- Latency is not strictly important: Most of the health care related applications are not latency dependent**.**
- The cost of misclassification is high: If the patient that doesn't need to readmit if model says "yes to readmit" that will put financial burden on the patient. If patient need to readmit but model say "no to readmit" then that will cause readmission cost to the hospital so, misclassification rate should be as low as possible**.**

## 9.3 Project Outcome:

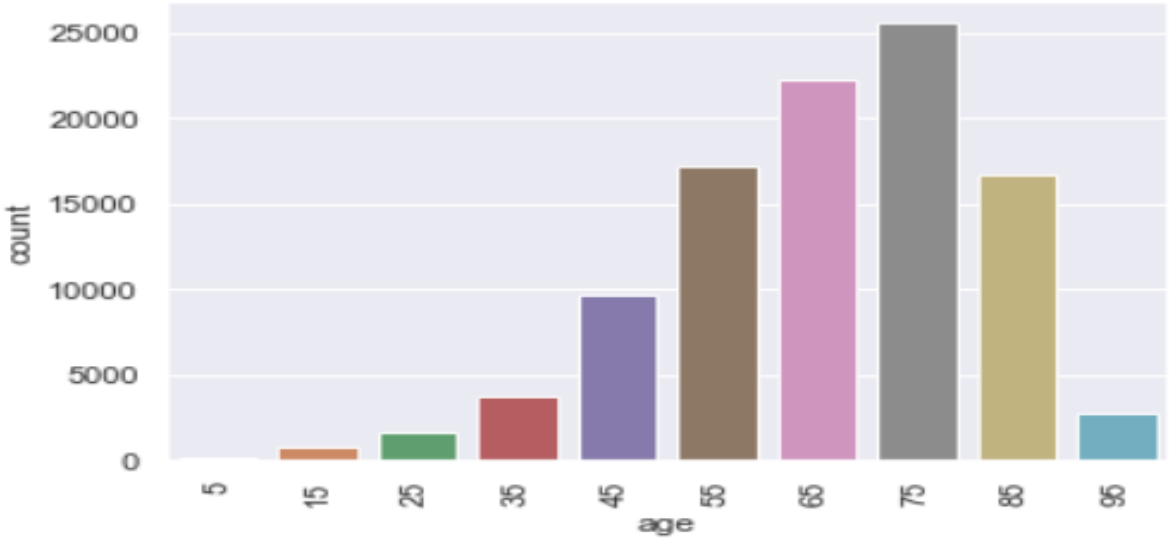
The aim of this project is to provide reducing early hospital readmissions of patients and thus improving healthcare quality. In this case study we will see how machine learning can help us solve the problems caused due to readmission. This predictive model is built on the previous data available with useful attributes to make the model efficient and practically applicable so that resource allocation can take place based on its results.

## 10.0 Exploratory Data Analysis

### 10.1 Relationship between variables

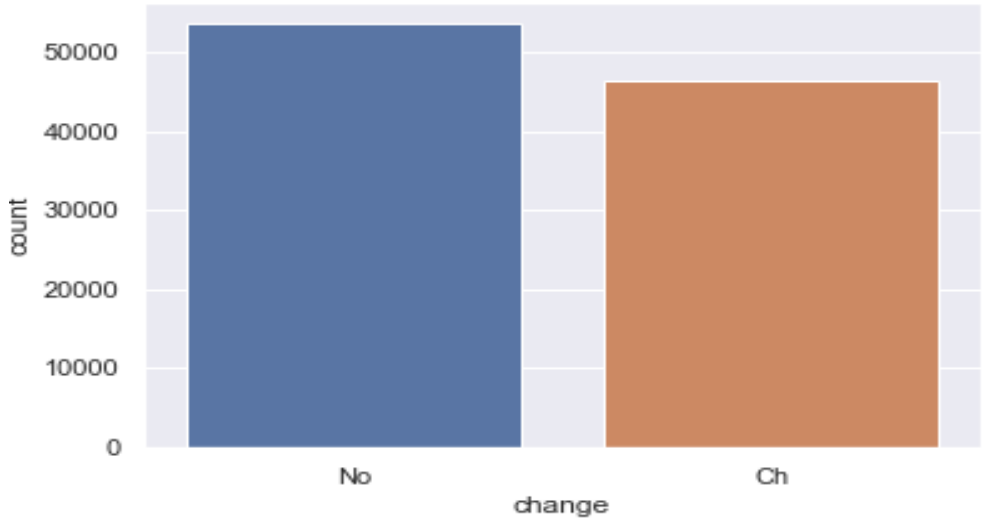
#### 10.1.1 Univariate Analysis

❖ **Age**



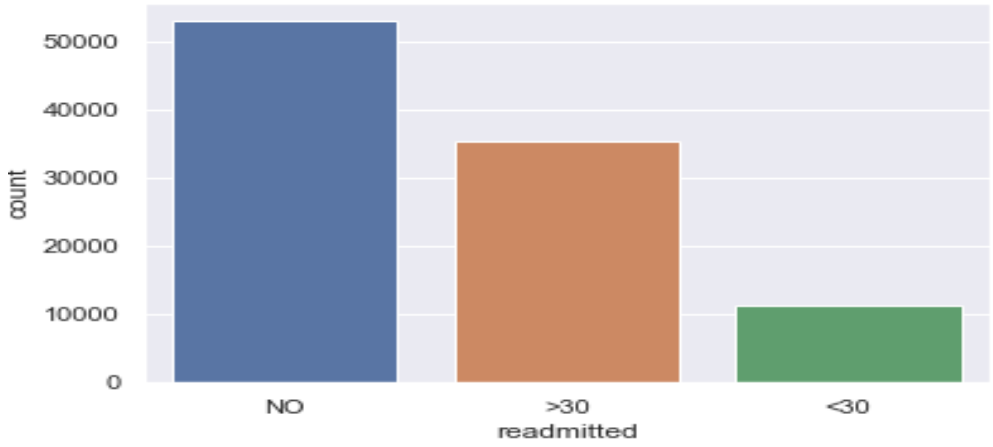

- Most of the patients falls in the age range 50-90 in our dataset.

❖ **Changes In Medications**



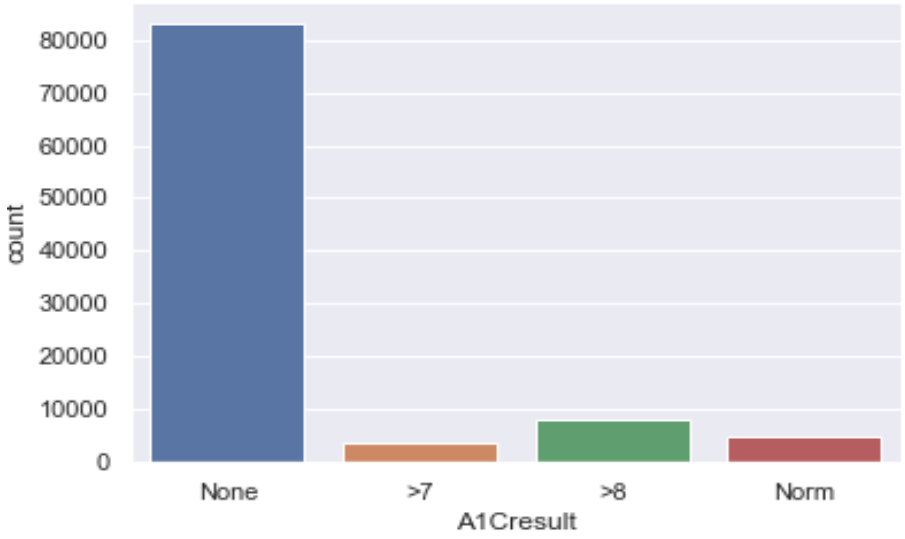

- Graph shows that around **46% of the patients** were observed with change in their diabetic routine medications.

❖ **Readmitted column**



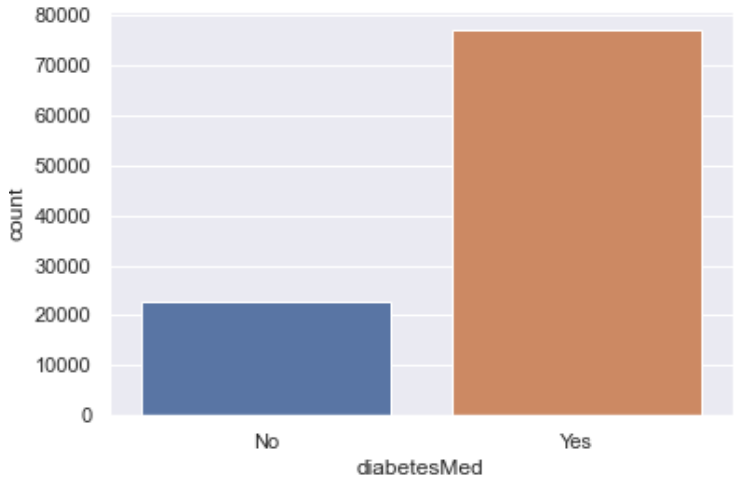

- From our data only **11% of the patients** were readmitted within **30 days** and **35% after 30 days**. Also, there is **53% of population** which remained **unaffected of readmission** due to diabetic conditions.

❖ **A1Cresult column**




- A1C test results denotes the average blood sugar level of the patient. Here, it is distributed in ranges where **above 6.5 denotes** that the **patient is diabetic**. It also indicates that around **12% of samples from our dataset are diabetic**. Apart from that, we have majority of unmeasured values in the dataset imputed as none.

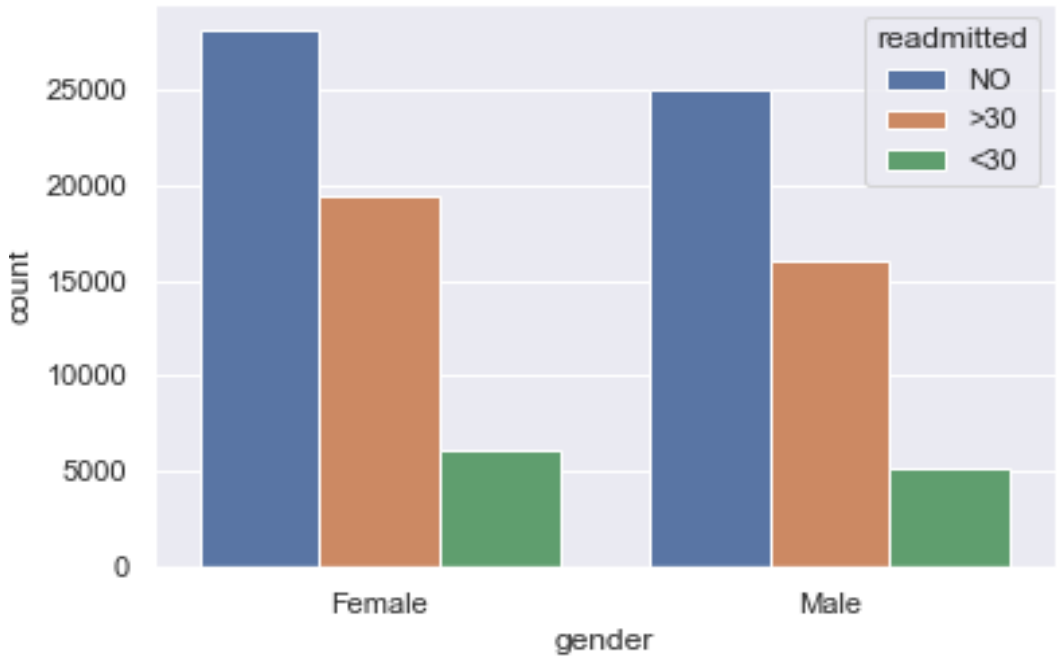
❖ **Diabetes Med**




- We can observe from plot that around **77% of samples from our data take diabetes medicines**.

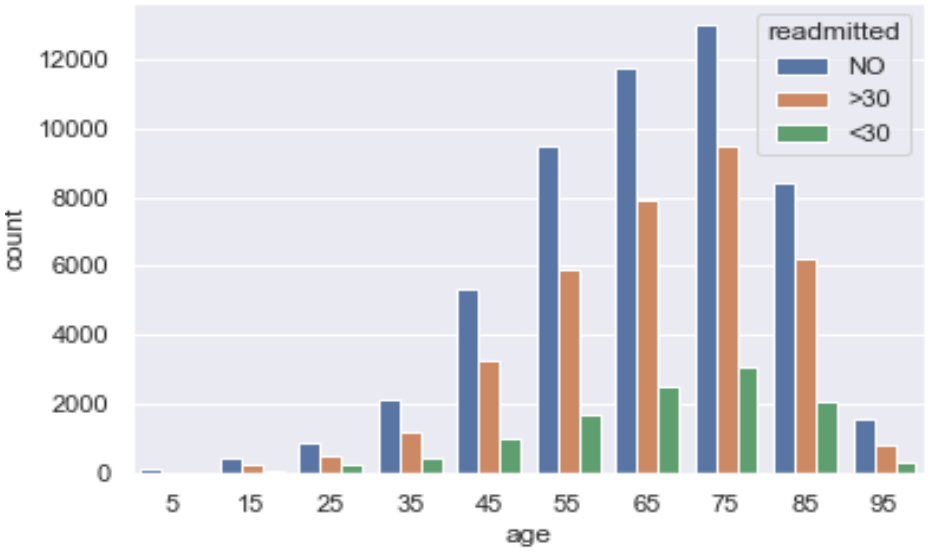
### 10.1.2 Bivariate Analysis

❖ **Gender Vs Readmitted:**



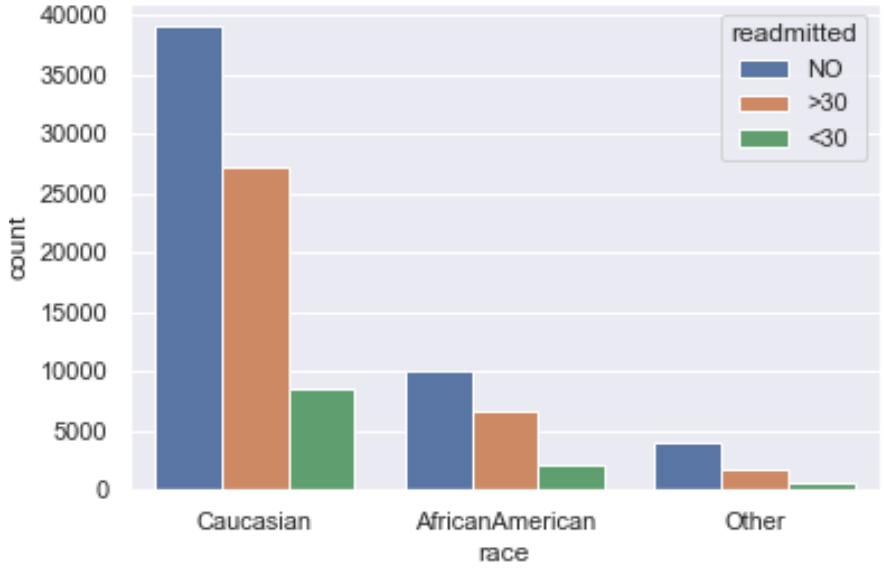

- We can see from the above plot, the **proportion of males readmitted is almost equal to females**, although females are little more prone to be readmitted than males.

## ❖ Age vs Readmitted:



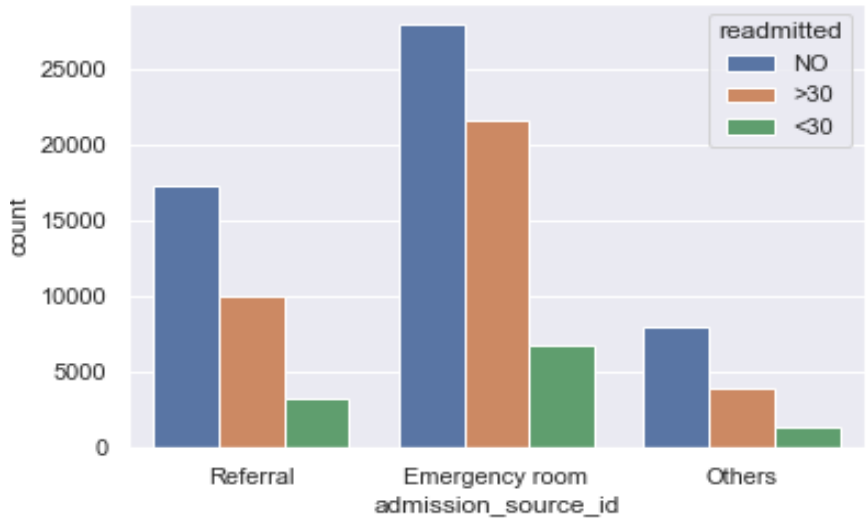

- From the above plot and analysis over age, we can conclude that though we have our **mostly affected population to be elderly,** but we **cannot underrate** the **mid-age generation (i.e., 20-40)** because it also associates considerable ratio of affected cases.

## ❖ Race vs Readmitted:



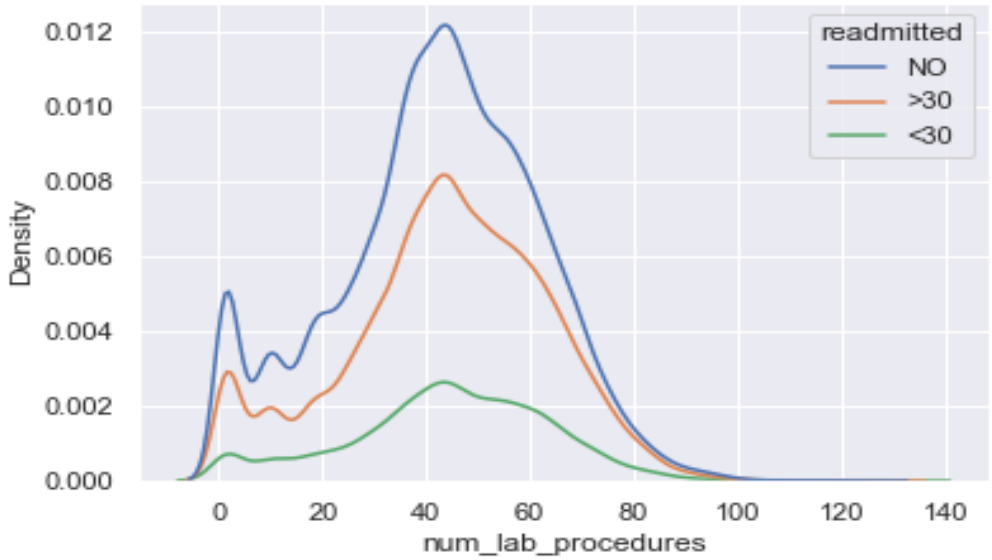

- The above analysis indicates that although we have our **most of the population to be "Caucasians"**, but the **readmission chances of rest of the races** are almost **similar**.

❖ **Admission Source Vs Readmitted:**



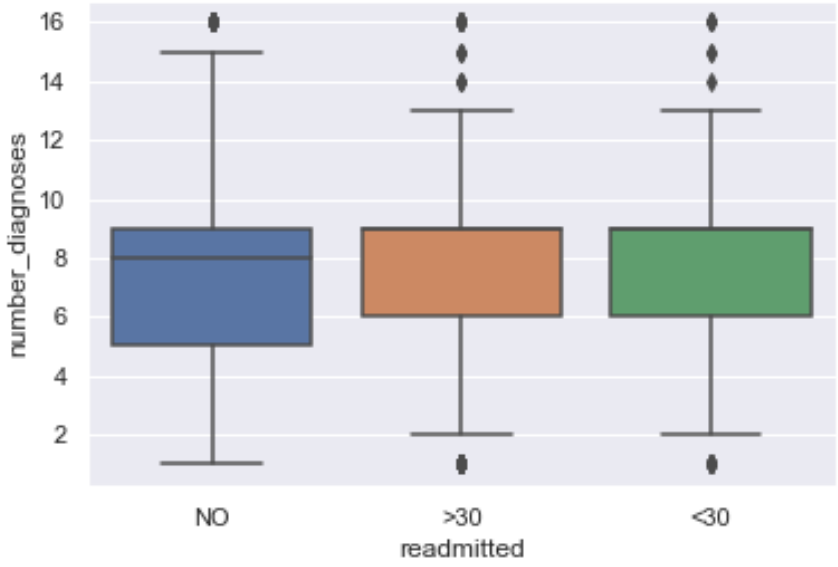

- Although we have data **imbalance among admission_source_id** categories, the chances of **readmission** are almost **close to each other**.

❖ **Number of Lab Procedures Vs Readmitted:**



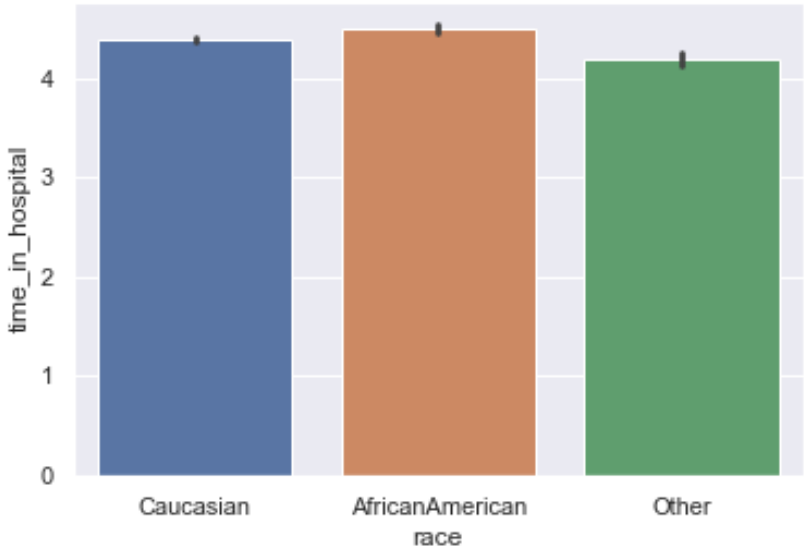

- The plot above shows the **variance** among number of **lab procedures** with respect to readmission, we see an insightful distribution of number of lab procedures among readmission of patient. **Irrespective of the number of lab procedures, there are considerable chances of readmission of patient.**

❖ **Number of Diagnoses Vs Readmission:**




- From Boxplots above we can infer that although the distribution of number of diagnoses is not similar, the median for all categories in readmission is almost similar. Thus, we can conclude that **readmission** is **independent** of the number of **diagnoses**.

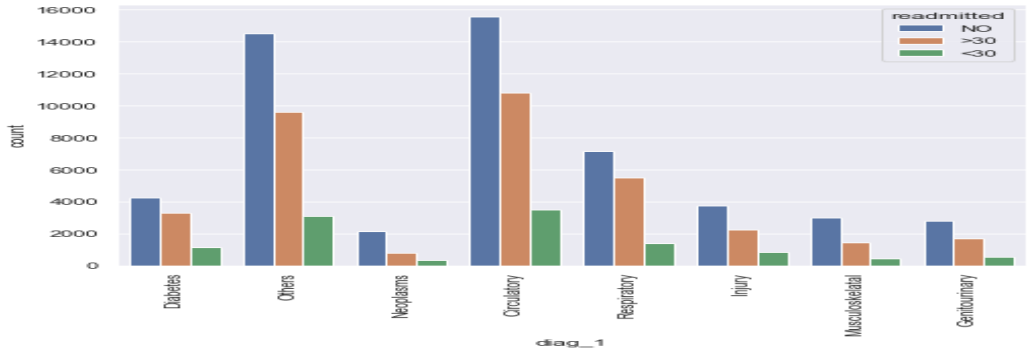
❖ **Race Vs Time in Hospital:**



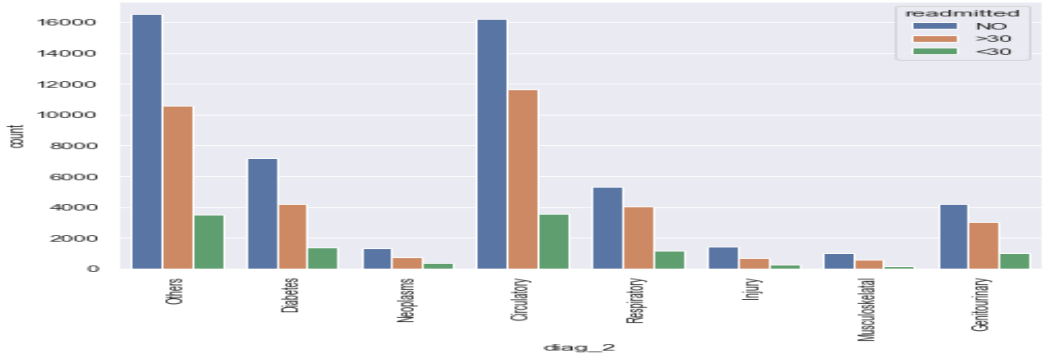

- Irrespective of the race, the time spent by patients is almost similar. But **African Americans** are more likely to spend more time than Caucasian followed by other race.

## ❖ Diagnoses Vs Readmission:

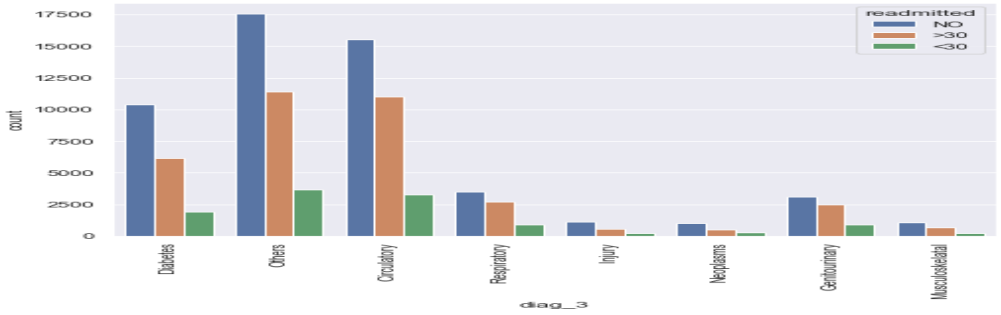
- **Diagnosis 1 Vs Readmission:**




- **Diagnosis 2 Vs Readmission:**



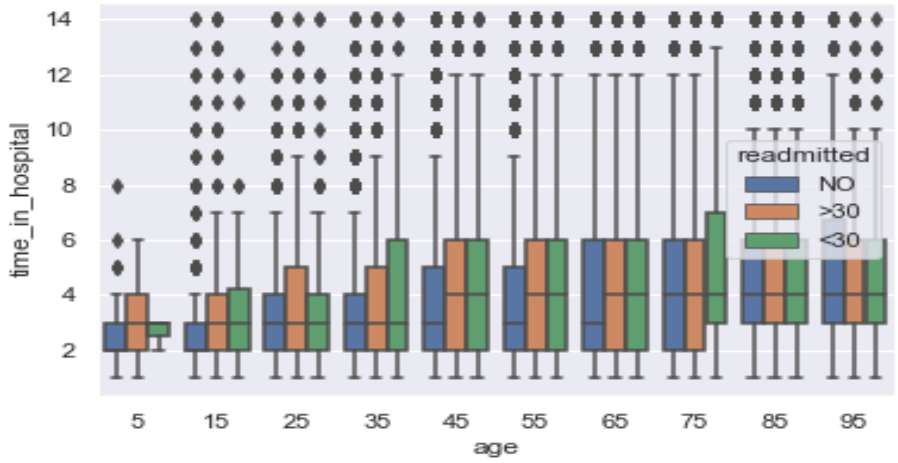

- **Diagnosis 3 Vs Readmission:**




- Above plots and analysis indicates that, when compared in all three diagnoses the patients diagnosed with **Circulatory and Respiratory** diseases are **more likely** to get **readmitted** as compared to other.

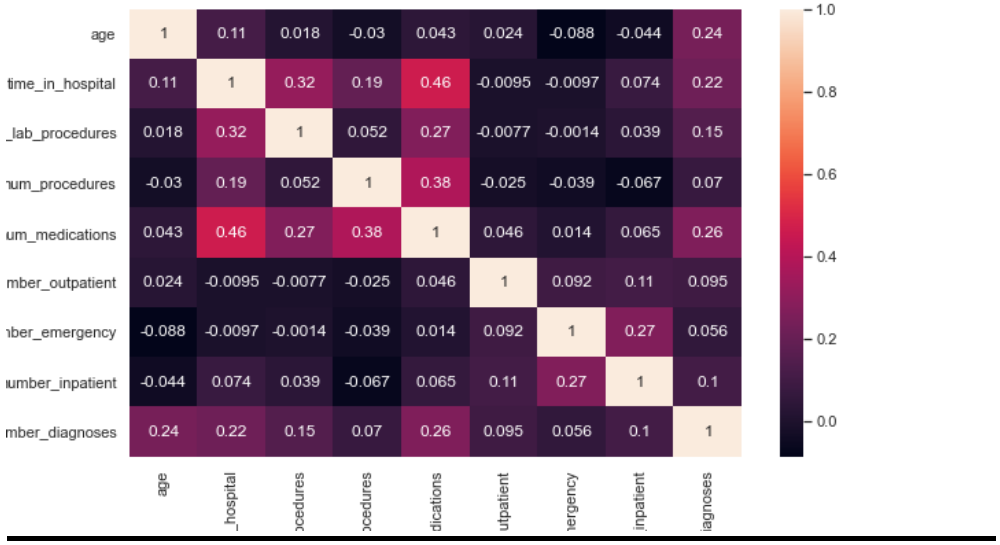
### 10.1.3 Multivariate Analysis

❖ **Age Vs Time in Hospital Vs Readmitted:**




- From the above plot we can see that as the **age of the patients increases**, the **time they spent in the hospital also increases** up to a point, along with chances of getting readmitted.

## 10.2 CORRELATION MATRIX




From the above **Heatmap** we can say that there is no strong **Correlation** among the independent features.

## 11.0 Base Model Building:

The choice of the right model performance measures is highly critical since the dataset is a highly imbalanced dataset and the conversion rate. Model accuracy alone may not be enough to evaluate a model. Hence the following model performance measures have been used to evaluate the models, based on the confusion matrix built for the predictions on the training and test dataset:

| | Negative (Predicted) | Positive (Predicted) |
|---|---|---|
| Negative (Observed) | True Negative (TN) | False positive (FP) |
| Positive (Observed) | False negative (FN) | True positive (TP) |

### ❖ Sensitivity or recall:

Sensitivity (Recall or True positive rate) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall or true positive rate (TPR).

### ❖ Precision:

Precision (Positive predictive value) is calculated as the number of correct positive predictions divided by the total number of positive predictions.

### ❖F1-Weighted:

F1 score is an overall measure of a model's accuracy that combines precision and recall A good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0.

### 11.1.1 Predictive Modelling:

- Primary solution of the project is to rightly predict, whether the patient will get readmitted within 30 days or after 30 days or will not get readmitted at all. So, we are trying to solve our problem by building a multiclass classification model.

- There are different types of classification algorithms for modelling multiclass classification problems. So, we are using two different algorithms to build our base models.

- Performance of multiclass classification models can be evaluated using different metrics. F1-score, precision and recall are three such metrics used here to evaluate the base models.

- The target feature 'readmitted 'contains 3 categories which are NO, '>30'and '<30'. These categories are encoded using the labels 0, 1 and 2 respectively.
- As the scaling methods affect model performance significantly, so we are exploring various scaling methods for the base models. Based on the final evaluation metrics, the performance of different algorithms with respect to different scalars could be analyzed.

- Here we are building our base model and checking its performance with two different changes in our dataset:

1. By label encoding the medication columns, which have sub-categories like up, steady, down and no.
2. By one-hot encoding the sub-categories in medication columns using pandas get dummies

- The algorithms that are used for classification are:

1. Logistic Regression.
2. Random Forest.

### 11.1.2 LOGISTIC REGRESSION

- Logistic regression has become an important tool in the discipline of machine learning. It is one of the most used machine learning algorithms for classification problems. The purpose of logistic regression is to estimate the probabilities of events, including determining a relationship between features and the probabilities of outcomes. Logistic Regression is one of the base models for solving our multiclass classification problem.

### 11.1.3 RANDOM FOREST

- Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For our multiclass classification problem, we are using a random forest classifier as one of the base models.

## 11.2 BASE MODELS WITH LABEL ENCODED MEDICATION COLUMNS

| | LR_f1_score | LR_recall | LR_precision | RF_f1_score | RF_recall | RF_precision |
|---|---|---|---|---|---|---|
| Standard_scaler | 0.510350 | 0.572291 | 0.542071 | 0.515982 | 0.561913 | 0.524696 |
| Min_Max_scaler | 0.507234 | 0.571451 | 0.546245 | 0.516708 | 0.562253 | 0.525985 |
| Max_abs_scaler | 0.508489 | 0.571051 | 0.540465 | 0.515841 | 0.561513 | 0.527135 |
| Robust_scaler | 0.520828 | 0.573531 | 0.527816 | 0.515645 | 0.560843 | 0.525195 |
| Power_transformer | 0.511064 | 0.572541 | 0.541900 | 0.517743 | 0.563482 | 0.529857 |

**Figure 1: Model evaluation metrics for different algorithms with label encoded columns**

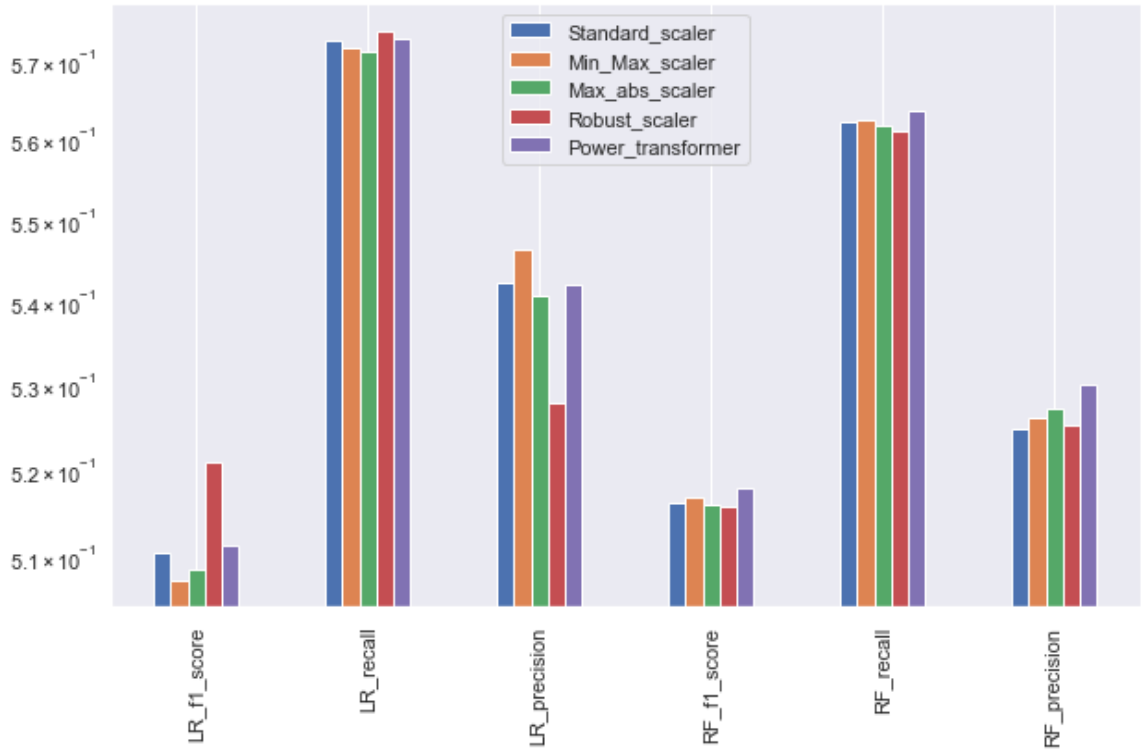



**Figure 2: Graphical representation of evaluation of different models with label encoded columns**

The highlighted metrics in figure 1 shows that logistic regression is performing better when compared to random forest classifier. We can also observe that logistic regression model is giving better performance on this dataset when scaled using Robust scaler and Min-Max scaler.

## 11.3 BASE MODELS WITH ONE HOT ENCODED MEDICATION COLUMNS

| | LR_f1_score | LR_recall | LR_precision | RF_f1_score | RF_recall | RF_precision |
|---|---|---|---|---|---|---|
| Standard_scaler | 0.509391 | 0.572301 | 0.544780 | 0.517983 | 0.561503 | 0.526505 |
| Min_Max_scaler | 0.507359 | 0.571011 | 0.541710 | 0.517892 | 0.561443 | 0.524311 |
| Max_abs_scaler | 0.507408 | 0.570742 | 0.541680 | 0.517759 | 0.561253 | 0.523062 |
| Robust_scaler | 0.520229 | 0.573011 | 0.537645 | 0.516454 | 0.560203 | 0.523091 |
| Power_transformer | 0.509379 | 0.572301 | 0.544423 | 0.516309 | 0.560153 | 0.524196 |

Figure 3: Model evaluation metrics for different algorithms with one hot encoded column

## BASE MODEL - With Transformation

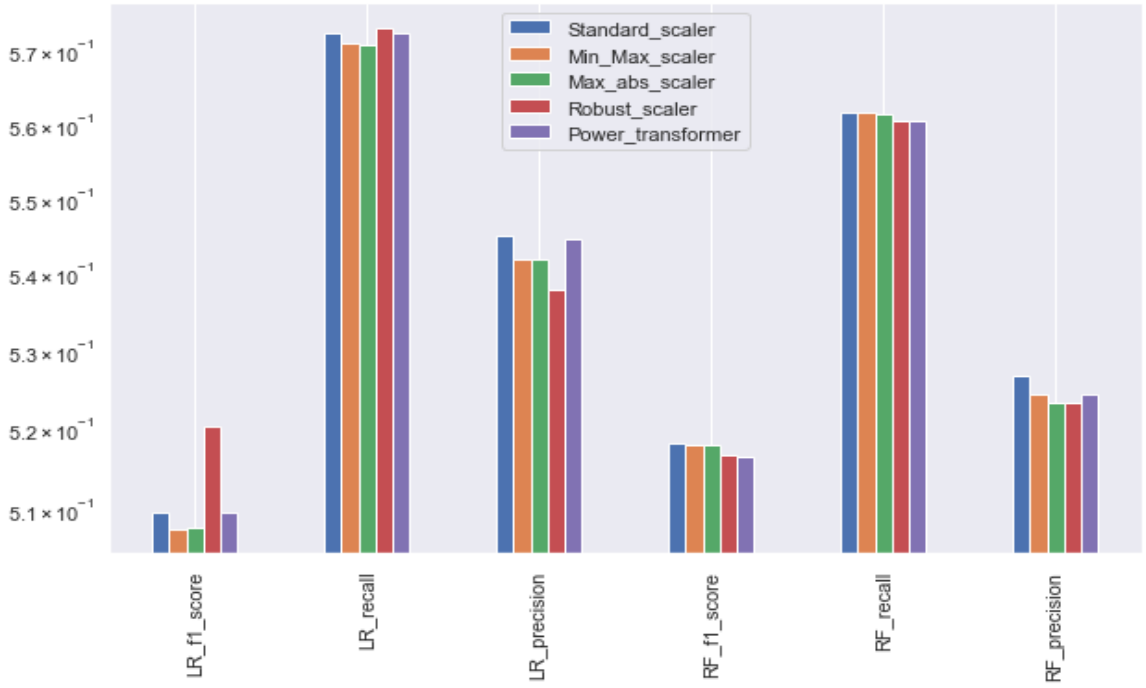



Figure 4: Graphical representation of evaluation of different models with one hot encoded column

Comparing figure 1 and figure 3, we can see that the model built using label encoded medication columns and one hot encoded medication columns are giving almost similar results. With one hot encoded column also logistic regression is performing better when compared to random forest classifier.

## 12.0 Feature Extraction:

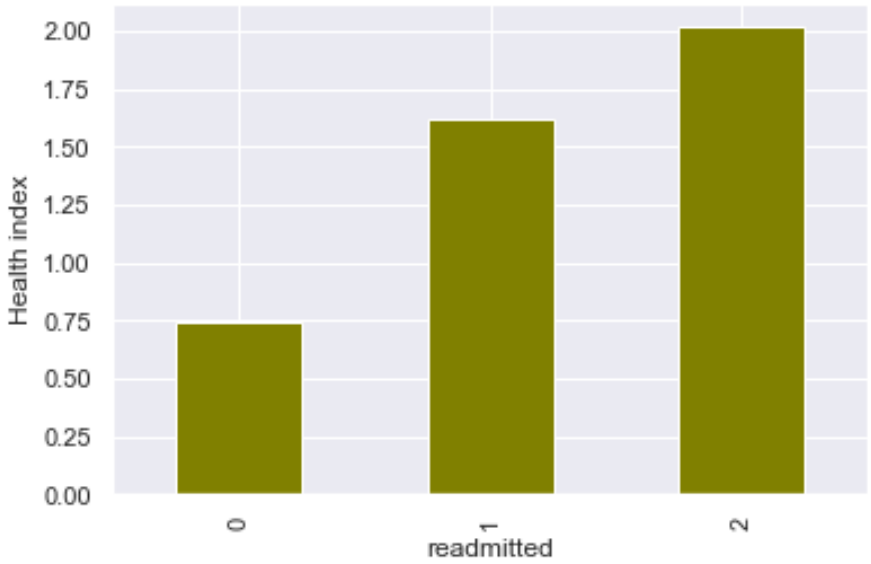
According to the domain knowledge and analyzing the features, also considering the correlation among them, we created few new features as follows:

### 12.1.1 Health Index:

If the frequency of person's visit to the hospital is high, then we can think of that person to be less healthy and less healthy patient tends to readmit quickly let's create health index variable. Higher the health index lessen the chance that person will readmit (indirectly proportional)

Health index = (1 / (number_emergency + number_inpatient + number_outpatient))

**Here, 0= No Readmission, 1= Readmission after 30 Days, 2=Readmission within 30 days**

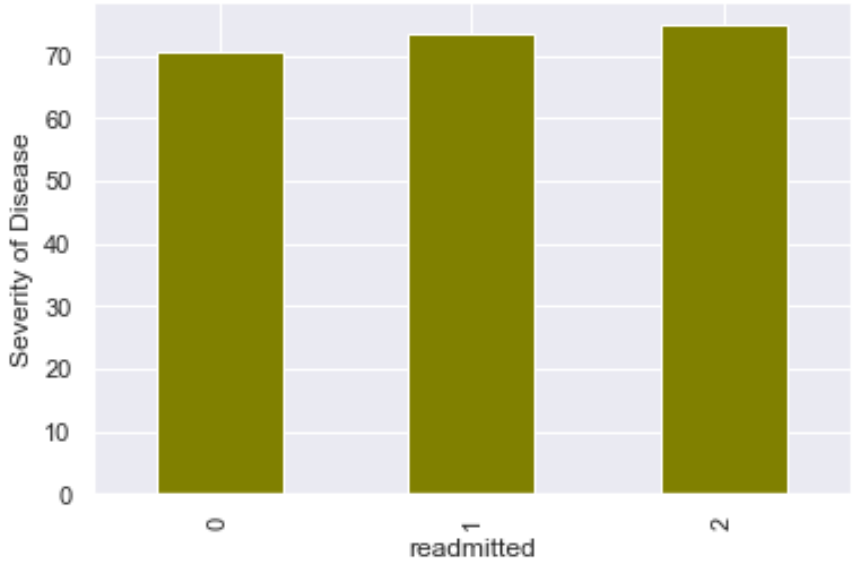


### 12.1.2 Severity of Disease:

Severity of disease is high if patient is spending lots of time in hospital and going through number of complicated test so, lets create severity of disease as one of the features. To get probabilistic interpretation lets divide it by total values.

severity_of_disease = (time_in_hospital + num_procedures + num_medications + num_lab_procedures + number_of_diagnoses)

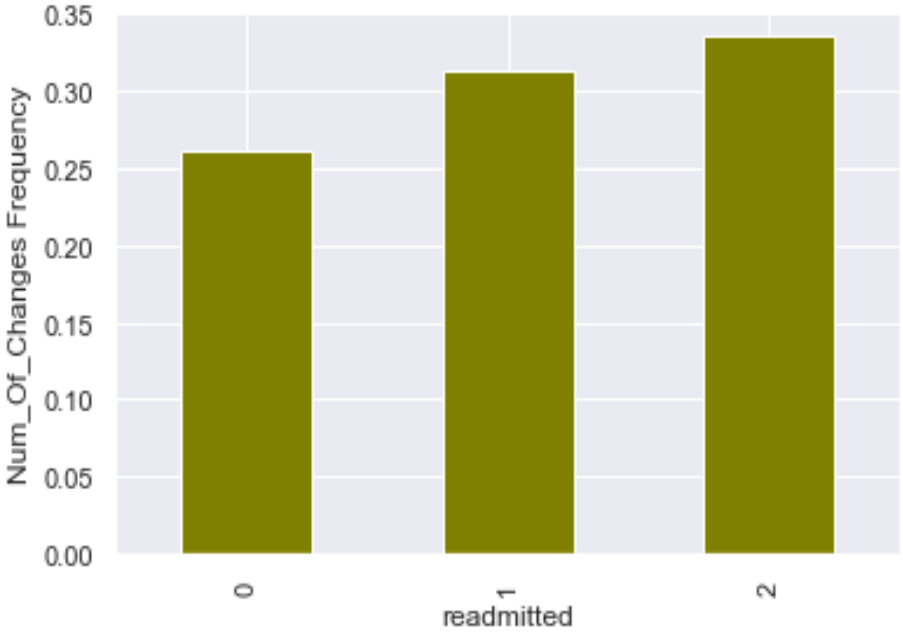
**Here, 0= No Readmission, 1= Readmission after 30 Days, 2=Readmission within 30 days**




### 12.1.3 Number of Changes:

Research has found that the patient which keep going through changes(up/down) in proportion of medications tends to readmit, so we have engineered new variable called as ‘number_of_changes.’. This captures number of medications whose proportion have changed. We calculated it for each patient.

**Here, 0= No Readmission, 1= Readmission after 30 Days, 2=Readmission within 30 days**




# 13.0 Feature Selection:

There are several ways in which feature engineering can be done, in this project there are 2 ofthem used. First is based on Multicollinearity among the independent variables. Second is based on the p value results from the statistical tests. Let us go through both in detail to make sure which features have been selected for the final model.

## 13.1 Selecting Significant Features:

### 13.1.1 Categorical Features (Using Chi-Squared Test)

Chi- Squared test is used to check whether two categorical features have some relationship or not. Here we want features to have some relation with class label if some features have no relation with class label it makes sense to remove them. Chi-Squared internally performs hypothesis testing and give us the p-value, using this p-value we can remove some features which are not important we kept significance level/alpha as 0.4 that is if there is more than 60% chance that feature is important then only, we will keep that feature in our final feature set.

### 13.1.2 Significant Categorical Features:

1. race
2. Gender

3. admission_type_id
4. discharge_disposition_id
5. admission_source_id
6. diag_1
7. diag_2
8. diag_3
9. max_glu_serum
10. A1Cresult
11. Change
12. diabetesMed
13. readmitted

### 13.1.3 Numerical Features (Using ANNOVA test)

- Let's use Chi2- Contingency Test to check whether numeric features and readmitted column are dependent or independent if some features are found to be independent on readmitted, we will simply remove them.

- We found that correlation is always close to zero but, ANNOVA test doesn't capture the non-linear relationships so, rather than using correlation coefficient we will use p-value to get "rejected features list" Here hypothesis testing is done assuming null hypothesis as variables are independent. Let's set significance level/alpha as 0.4 now we can remove features with p-value > alpha by declaring them to be statistically independent of class label "readmitted".

### 13.1.4 Significant Numerical Features:

1. Age
2. time_in_hospital
3. num_lab_procedures
4. num_procedures
5. num_medications
6. number_outpatient
7. number_emergency
8. number_inpatient

9.number_diagnoses

10.Metformin

11.Repaglinide

12.Glipizide

13.Glyburide

14.Acarbose

15.Insulin

16.Health_index

17.severity_of_disease

18.number_of_changes

## 13.2 **Feature Selection Using Recursive Feature Elimination**:

- RFE is a wrapper-type feature selection algorithm. This means that a different machine learning algorithm is given and used in the core of the method, is wrapped by RFE, and used to help select features. This contrasts with filter-based feature selections that score each feature and select those features with the largest (or smallest) score.
- This is achieved by fitting the given machine learning algorithm used in the core of the model, ranking features by importance, discarding the least important features, and re-fitting the model. This process is repeated until a specified number of features remains.

We have performed both the methods for Feature Selection, the Recursive feature selection (RFE) as well as the statistical approach. By doing so, we got the following results:

| | LR_SC | RF_SC | LR_RFE | RF_RFE |
|---|---|---|---|---|
| f1_weighted | 0.497682 | 0.521708 | 0.501997 | 0.522278 |
| recall_weighted | 0.567252 | 0.568592 | 0.569982 | 0.568232 |
| precision_weighted | 0.536662 | 0.540914 | 0.543449 | 0.537562 |

As the result we have obtained, the test results for both statistical and RFE methods are almost similar, thus for easing the computational method and timing, we have chosen the statistical approach for selecting the most significant features for the model.

## 14.0 Summary of EDA and Feature Engineering:

- From the EDA we get to know that the features like number of lab procedures, Diabetes Medication, admission are important in our task. Combination of features like “change + DiabetesMed”, “age + time_in_hospital”, “age + number_impatient” and “change + admission_source_id” can give us lots of information which is helpful in the given task.

- We have tried model-based imputation to fill the missing values but found out that the features with very a smaller number of distinct values should be imputed using model-based approach so we used model based imputation to fill missing values of race feature only.

- In feature engineering for columns like change, insulin, etc we have No, Up, Steady values but after experiments we found that it is good idea to convert them into numbers.

- First, we kept the categorical values intact but with this approach we were getting very low AUC so, we used domain knowledge to reduce the distinct categories of categorical features and that improved the accuracy.

**NOTE:** At start we had all the features included and most the features were not important to the given task. So, we used ANNOVA test and Chi-Squared test to remove the unnecessary features and this improved our model's accuracy.

As the result we have obtained, the test results for both statistical and RFE methods are almost similar, thus for easing the computational method and timing, we have chosen the statistical approach for selecting the most significant features for the model.

## 15.0 Data Sampling:

After successfully selecting the significant features and we perform different model building on them. Further analyzing the results obtained from different models.
Since there is a limitation with us of bulk dataset which makes the computational method and timing more complex, we have chosen to sample the data from the whole population data.
For our multiclass target variable and bulk dataset, it is desirable to split the dataset into train and test sets in a way that preserves the same proportions of examples in each class as observed in the original dataset. This is called a stratified train-test split.

### 15.1 Tests to verify sample target is representation of population target:

**Ho: proportion of target in population is equal to proportion of target in sample**

**H1: proportion of target in population not equal to proportion of target in sample**

- Performing two sample proportion z-test, and Chi-square test for goodness of fit.
- From proportion test it is evident that we accept null hypothesis i.e., sample is representation of population
- From chi-square test as the p-value is greater than 0.05, it is evident that we fail to reject null hypothesis i.e., sample is representation of population.

## 16.0 Model Building: (Without Tuning)

### 16.1.1 SVC C-Support Vector Classification:

The implementation is based on library svm. The fit time scales at least quadratically with the number of samples and may be impractical beyond tens of thousands of samples. For large datasets consider using **LinearSVC** or **SGDClassifier** instead, possibly after a **Nystroem** transformer.

### 16.1.2 K-Nearest Neighbour Algorithm:

KNN works by **finding the distances between a query and all the examples** in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

### 16.1.3 AdaBoost Classifier:

An AdaBoost classifier is **a meta-estimator that begins by fitting a classifier on the original dataset** and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

### 16.1.4 Random Forest:

A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

### 16.1.5 XGBoost Classifier:

XGBoost provides **a wrapper class** to allow models to be treated like classifiers or regressors in the scikit-learn framework. ... The XGBoost model for classification is called XGBClassifier. We can create and fit it to our training dataset. Models are fit using the scikit-learn API and the model. fit () function.

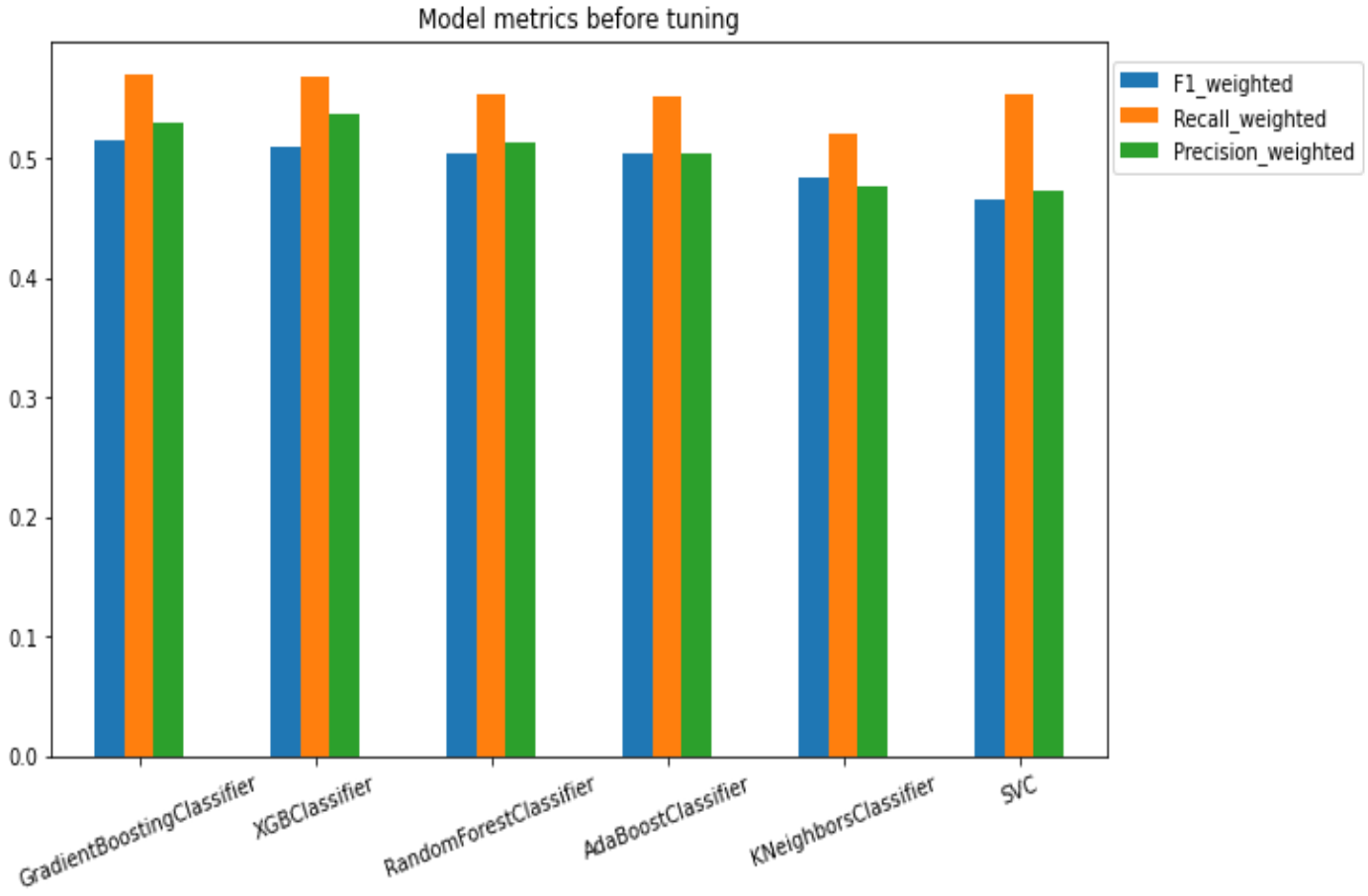
### 16.1.6 Gradient Boosting Classifier:

Gradient boosting classifiers are **a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model**. Decision trees are usually used when doing gradient boosting.

## 16.2.1 Results:

| | f1-score | recall | precision |
|---|---|---|---|
| **SVC** | 0.465434 | 0.552717 | 0.473132 |
| **KNeighborsClassifier** | 0.483617 | 0.519872 | 0.476770 |
| **AdaBoostClassifier** | 0.503777 | 0.552067 | 0.503803 |
| **RandomForestClassifier** | 0.504736 | 0.553367 | 0.512412 |
| **XGBClassifier** | 0.509982 | 0.568215 | 0.536538 |
| **GradientBoostingClassifier** | 0.514369 | 0.569615 | 0.529057 |

## 16.2.2 Graphical Representation of Different model's output:



## 17.0 Model Building: (With Hyperparameter Tuning)

### 17.1 Results after tuning all models:

1. **Random Forest Classifier:**

   f1_weighted: 0.5070576964659778

   recall_weighted: 0.5596161928556241

   precision_weighted: 0.5154207676167762

2. **AdaBoost Classifier:**

   f1_weighted: 0.5114114985914617

   recall_weighted: 0.5612160303802446

   precision_weighted: 0.5150099955834049

3. **XGBoost Classifier:**

   f1_weighted: 0.5210665826137267

   recall_weighted: 0.5667148481208213

   precision_weighted: 0.5298765942874173

4. **Gradient Boost Classifier:**

   f1_weighted: 0.5226612053062879

   recall_weighted: 0.5684146881438171

   precision_weighted: 0.5272921300578394

5. **Stacking Classifier:**

   f1_weighted: 0.5131383139770068
   recall_weighted: 0.5578161103942403
   precision_weighted: 0.5129917063598038

### 17.2 Results:

| | F1_weighted | Recall_weighted | Precision_weighted |
|---|---|---|---|
| GradientBoostingClassifier | 0.522661 | 0.568415 | 0.527292 |
| XGBClassifier | 0.521067 | 0.566715 | 0.529877 |
| StackingClassifier | 0.513138 | 0.557816 | 0.512992 |
| AdaBoostClassifier | 0.511411 | 0.561216 | 0.515010 |
| RandomForestClassifier | 0.507058 | 0.559616 | 0.515421 |

The one more limitation here we face is the imbalance we have within the data itself. The categories are imperfectly distributed within the dataset in biased proportions. Thus, we overcome that problem by oversampling the data by using the SMOTE technique.

SMOTE is an **oversampling technique that generates synthetic samples from the minority class**. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier.

Thereby performing different classifier models to obtain desired results.

## 18.0 Model Building: (With Hyperparameter Tuning & SMOTE)

- Target Variable distribution without SMOTE
  0: 3155 1: 2118 2: 728

- Target Variable distribution with SMOTE
  0: 3155 1: 3155 2: 3155

### 18.1 Result after Random Forest Classifier:

f1_weighted: 0.6873419929568708

recall_weighted: 0.6903328050713154

precision_weighted: 0.6952459184452985

### 18.2 Result after Gradient Boosting Classifier:

f1_weighted: 0.7014247367018527

recall_weighted: 0.701743264659271

precision_weighted: 0.7091445629178832

### 18.3 Result after XGBoost Classifier:

f1_weighted: 0.6759726747155105

recall_weighted: 0.6768092974115162

precision_weighted: 0.6801709054411543

### 18.4 Result after Stacking Classifier:

f1_weighted: 0.685217329039336

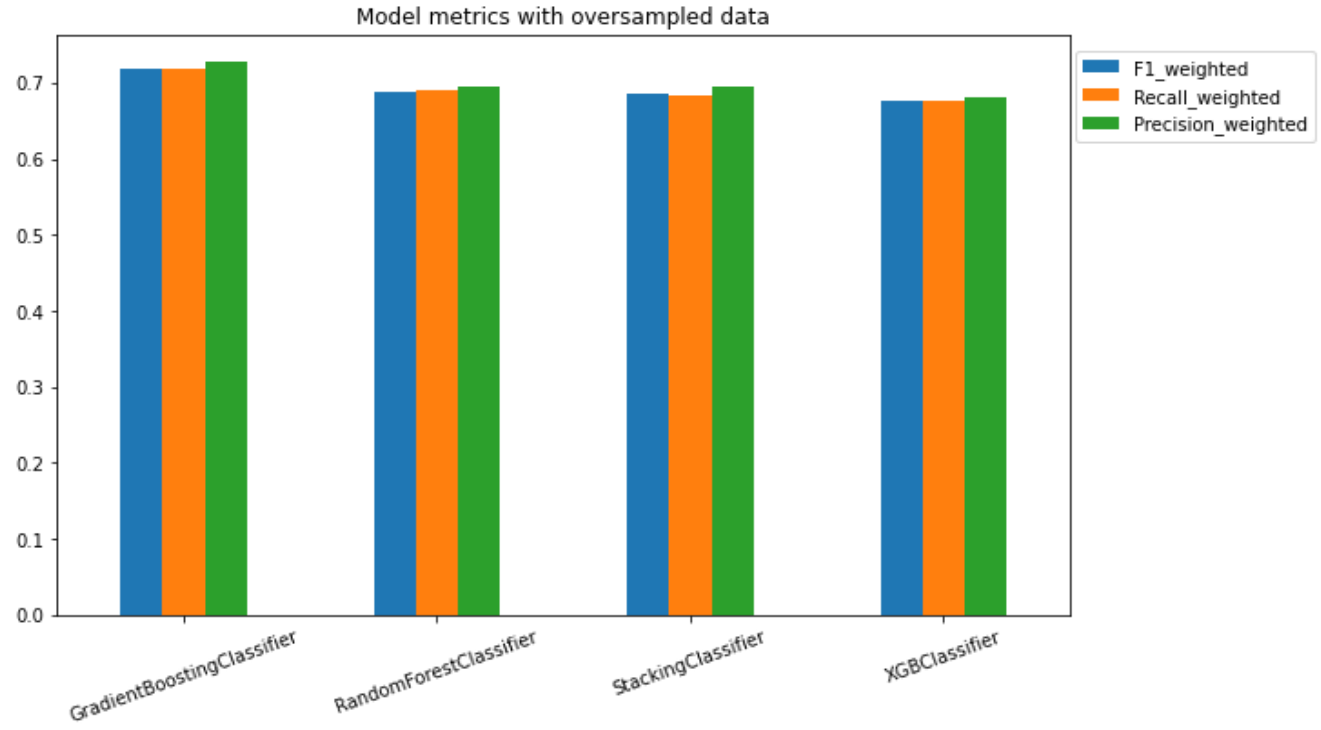
recall_weighted: 0.6834653988378235

precision_weighted: 0.6948057858564288

### 18.5 Result Metrics:

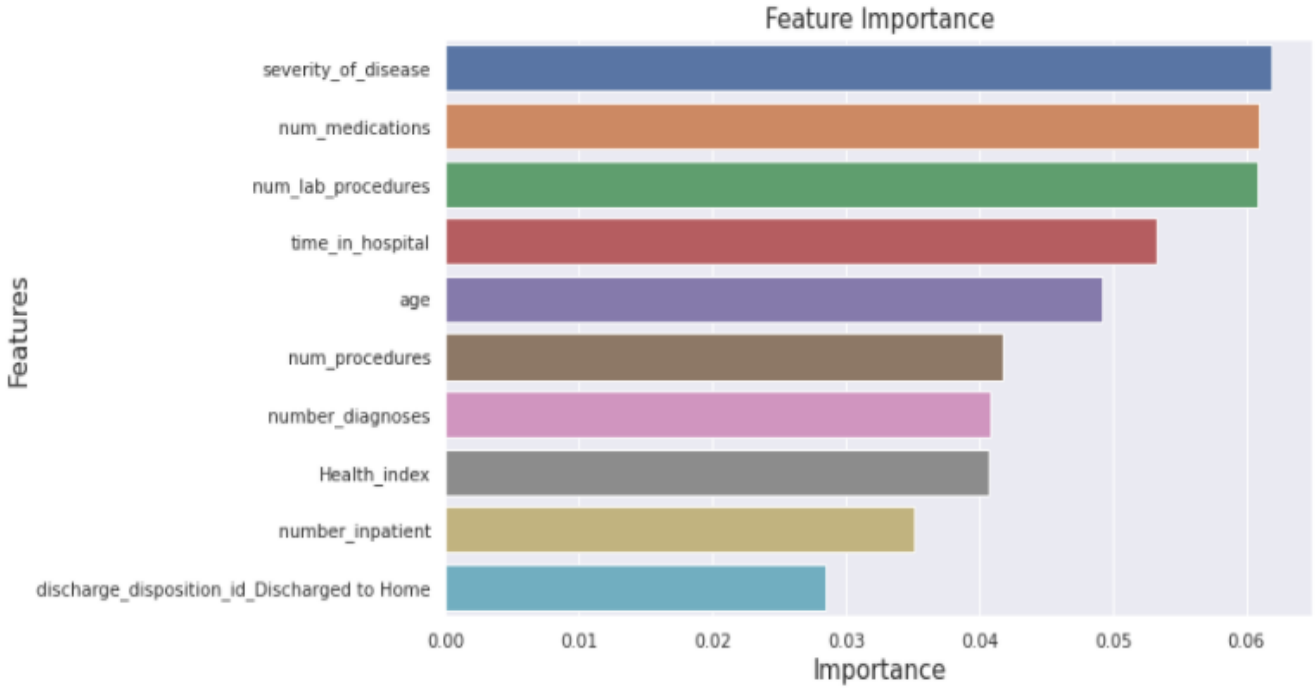
| | F1_weighted | Recall_weighted | Precision_weighted |
|---|---|---|---|
| GradientBoostingClassifier | 0.719425 | 0.718119 | 0.727532 |
| RandomForestClassifier | 0.687342 | 0.690333 | 0.695246 |
| StackingClassifier | 0.685217 | 0.683465 | 0.694806 |
| XGBClassifier | 0.675973 | 0.676809 | 0.680171 |

## 18.6 Graphical Representation of Different model's output:




# 19.0 Conclusions:

## 19.1 Feature Importance




From the above plot we can see that the features that were important in our model for prediction are severity_of_disease, num_medication,

num_lab_procedures, time_in_hospital and age.

Severity of Disease: Severity of disease is high if patient is spending lots of time in hospital and going through number of complicated tests.

Number of Medication: Number of distinct generic names administered during encounter.

Number of Lab Procedures: Number of Lab tests performed during the encounter.

Time in Hospital: Integer number of days between admission and discharge.

Age: Age groups of different patients.

## 19.2 Business Inference

Since we are dealing with sensitive medical data, it is important for us to look for both precision and recall. Hence we are interested in F1_score, which is harmonic mean of precision and recall.

In our best performing model(ada boost), we have got

- Precision of 71% which indicates that out of total predicted positives 71% are true positives
- Recall of 70% which indicates that out of total actual positives 70% are predicted are true positives
- F1_score, which is harmonic mean of precision and recall is around 70 percent

## 19.3 Limitations

Despite tuning the parameters model performance did not change drastically because of the class imbalance of target variable. We overcome this problem by oversampling through SMOTE

However in real life situation SMOTE will be the tool of last resort, so before that we need to ask client for more balanced data

Since we had rows more that 1 lakh and columns more than 80, it was computationally difficult to tune parameters for the models.

We overcome the above challenge by taking sample of dataset, that represents the population

## 19.4 Closing Reflection

There was application of statistics, feature selection and feature extraction during processing the dataset. Different modelling techniques were explored. Our base model which was a linear model and non linear model. Model tuning was performed on every model.Learned to implement SMOTE to overcome class imbalance.

A lot of domain knowledge was acquired in the field of healthcare.Parameter related to the treatment of diabetic patients were learnt. These are some of the major takeaways we got from this project.