Capstone Project – Report

Prediction of Diabetes Treatment Type to reduce Readmission

Domain – Healthcare Analytics

Mentored by,

Ms. Akhila Gurre

Submitted by,
Group – 5

Aravind Rao
Advait Saraf
Rahul Verma
Saurajit Sahoo
Tanuj Verma

ACKNOWLEDGEMENT

We would like to thank our mentor **MS AKHILA NAGA SAI GURRE**, Data Science Consultant, SYNITI, Bangalore for providing her valuable guidance and suggestions for our Project work. We also thank her for the continuous encouragement and the interest shown towards us to complete our Project work.

We are extremely grateful to our all teaching and non-teaching staff members of **GREAT LEARNING**, who showed keen interest and inquired our developments.

We would like to express our Gratitude to all teaching and non-teaching staff members of **Great Lakes Institute of Management**, for providing their support and guidance for our project.

We greatly admire and acknowledge the constant support we received from our friends and team members for all the effort and hard work that they have put into completing this project.

Content

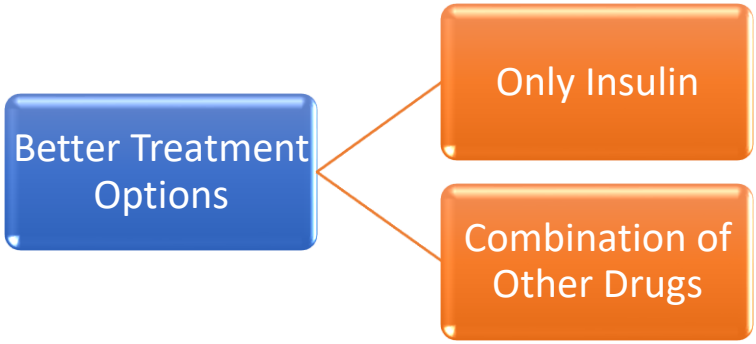
1. Abstract:

This Dataset relates to the Domain of HealthCare Analytics and is derived from the UCI's Machine Learning Repository. This dataset describes the records for Diabetes 130-US hospitals for years 1999-2008. This dataset consists of 1,01,766 instances of data & 50 features. It has 50 features which are explicitly divided into 13 numerical features & 37 categorical features. It also includes 23 categorical features for medications. This dataset can have colossal influence for discovering the factors blameworthy for higher treatments costs.

2. Problem Statement:

1. **What Would you achieve by this project?**

   This project will give insight into the better treatment option and readmission rate of the patient based on the different types of medication & insulin provided to the patient thereby providing a better solution to reduce the readmission rate. This is an academic project which focuses on understanding the problems caused due to readmission, which is impacting the efficiency of the hospitals and being a burden on diabetic patients in terms of treatment costs. The project aims to predict better treatment option based on the historic data that can help in reducing the readmission rate which could be either only insulin or only combination of medications.



2. **How would you help the business or clients?**

   The analysis of the data and different algorithms will optimise our results. Our focal point will be the parameters like Insulin, diabetic medicine, age, HbA1c result, diagnosis etc. to cure diabetes and to contour the readmission rate of the patients so that the cost of medications can be reduced with lesser readmissions. The parameters used in the data will navigate towards the quality of medications provided by the hospitals due to which readmission rate will be reduced by implementing some positive measures.

3. **Limitation of the project**

   Our project is a not a generalized model but by fair means, it helps to identify the best treatment option that can be provided to a patient and reducing the readmission rates.

3. Introduction:

Diabetes is dramatically increasing of increased obesity, sedentary lifestyle and aging populations. Interventions to improve diabetes outcomes can be directed at individuals with diabetes, health providers or the health systems. Hospital readmission is a compelling measure for the cost reduction. The diabetes is basically a misfortune for the patients which upshots in the cost of treatment and reveal the quality of treatment given to the patients. The reduction in the readmission rates act as a pivot to diminish the health care cost and significantly improve the quality of treatments. The government agencies and Health care systems are appraising the considerable amount of readmission rates to revamp the quality of treatments.

3.1 Need for the project:

This data elucidates that diabetes leads to other numerous complications like Kidney, Cardiovascular and eye diseases etc. HbA1c is measure of blood glucose levels. It is normal if less than 6% but becomes a pre-diabetic state when lies within 7%. If HbA1c is higher, the complications related to diabetes increases. There are several

medications provided to the diabetic patients and the non-diabetic patients as well. This survey manifests that how the health care industry impacts the treatment of diabetic patients and the medications provided to them. This analysis of large clinical database is undertaken to examine the historical patterns of diabetes care in patients with diabetes admitted to US hospital and to inform the future directions which may lead to improvements in patient safety. The readmission rate is the cornerstone in our project. It extricates the cost of treatments which is burdensome for patients due to repetitive readmissions.

3.2 Objective of the Project

The objective of the project is to develop a classification algorithm to predict a better treatment type to reduce the patient readmission within 30 days of discharge based on the different conditions. To predict the same, we have taken 47 variables like age, race, gender, different diagnosis, medications given, insulin etc into consideration. We will be predicting the impact of insulin on the readmission rate, insulin & other medication on the readmission rate & only medications other than insulin on the readmission rate.

3.3 Scope

The readmission rates are being contemplated to reduce the cost of treatments. Reducing readmission rates is the vital step to be achieved due to which quality of treatments will improve and treatment costs will be reduced to much extent.

3.4 Tools

- Programming Languages: Python
- Visualisation: Python & Tableau

4. Data Description

4.1 Description

The dataset consists of 1,01,766 instances of data & 50 features. The dataset is spread over 50 features including patient characteristics, conditions, tests and 23 medications. This is explicitly divided into 13 numerical features & 37 categorical features.

Feature Name	Description	Type
encounter_id	Unique identifier of an encounter	Numerical
patient_nbr	Unique identifier of a patient	Numerical
race	Values: Caucasian, Asian, African American, Hispanic, and other	Categorical
gender	Values: male, female, and unknown/invalid	Categorical
age	Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)	Categorical
weight	Weight in pounds.	Numerical
admission_type_id	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	Categorical
discharge_disposition_id	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	Categorical
admission_source_id	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	Categorical
time_in_hospital	Integer number of days between admission and discharge	Numerical
payer_code	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	Categorical
medical_specialty	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	Categorical
num_lab_procedures	Number of lab tests performed during the encounter	Numerical

num_procedures	Number of procedures (other than lab tests) performed during the encounter	Numerical
num_medications	Number of distinct generic names administered during the encounter	Numerical
number_outpatient	Number of outpatient visits of the patient in the year preceding the encounter	Numerical
number_emergency	Number of emergency visits of the patient in the year preceding the encounter	Numerical
number_inpatient	Number of inpatient visits of the patient in the year preceding the encounter	Numerical
diag_1	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	Categorical
diag_2	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	Categorical
diag_3	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	Categorical
number_diagnoses	Number of diagnoses entered to the system	Numerical
max_glu_serum	Indicates the range of the result or if the test was not taken. Values: ">200," ">300," "normal," and "none" if not measured	Categorical
A1Cresult	Indicates the range of the result or if the test was not taken. Values: ">8" if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured.	Categorical
change	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"	Categorical
diabetesMed	Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"	Categorical
readmitted	Days to inpatient readmission. Values: "<30" if the patient was readmitted in less than 30 days, ">30" if the patient was readmitted in more than 30 days, and "No" for no record of readmission.	Categorical

metformin, repaglinide, acetohexamide, pioglitazone, troglitazone, glyburide-metformin, glimepiride-pioglitazone, metformin-pioglitazone, nateglinide, glipizide, rosiglitazone, tolazamide, glipizide-metformin, metformin-rosiglitazone, chlorpropamide, glyburide, acarbose, examide, glimepiride, tolbutamide, miglitol, citoglipton, insulin	Values: "up" if the dosage was increased during the encounter, "down" if the dosage was decreased, "steady" if the dosage did not change, and "no" if the drug was not prescribed	Categorical

4.2 Source:

The domain is Healthcare Analytics and the dataset we have used is derived from the UCI's Machine Learning Repository "Diabetes 130-US hospitals for years 1999-2008 Data Set" which is available publicly. The dataset consists of 1,01,766 instances of data & 50 features.

Dataset Name:

"Diabetes 130-US hospitals for years 1999-2008 Data Set"

Variable Categorization

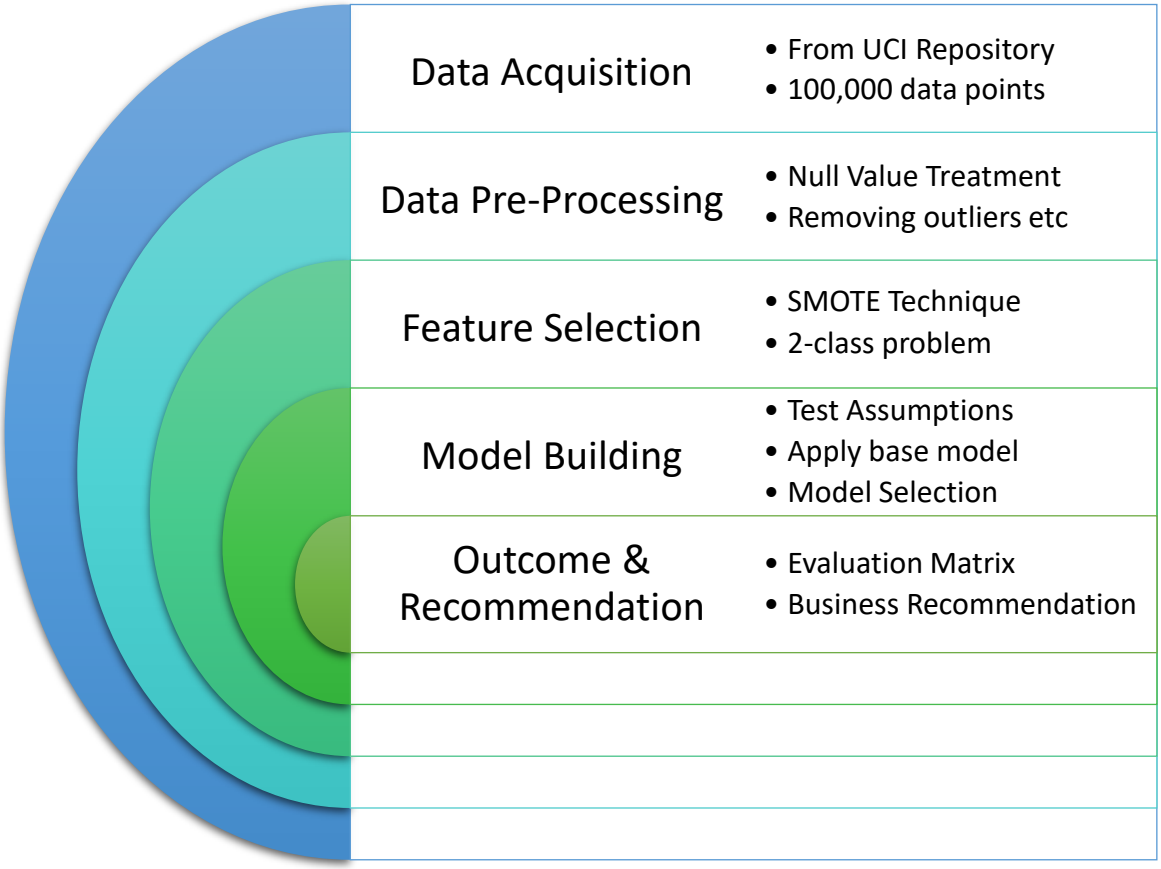
No. of Categorical Variables	37
No. of Numerical Variables	13

5. Literature Survey

This data elucidates that diabetes leads to other numerous complications like Kidney, Cardiovascular and eye diseases etc. HbA1c is measure of blood glucose levels. It is normal if less than 6% but becomes a pre-diabetic state when lies within 7%. If HbA1c is higher, the complications related to diabetes increases. There are several medications provided to the diabetic patients and the non-diabetic patients as well. This survey manifests that how the health care industry impacts the treatment of diabetic patients and the medications provided to them. This analysis of large clinical database is undertaken to examine the historical patterns of diabetes care in patients with diabetes admitted to US hospital and to inform the future directions which may lead to improvements in patient safety. The readmission rate is the cornerstone in our project. It extricates the cost of treatments which is burdensome for patients due to repetitive readmissions.

The diabetes increased substantially due to individual risk-factors, environmental risk factors, global changes, prevalence of obesity, hypertension, inadequate medications and higher treatment costs etc. A population level increase in prevalence of diabetes may be attributable to wide range of potential factors. The changes in diagnostic criteria and decreasing mortality rates among individuals with diabetes contributed to the rise in prevalence of this condition. However, the government agencies and HealthCare systems adopted some measures like identifying the population and several parameters that can literally improve the scenarios where diabetes is increasing substantially. The readmission rates are being contemplated to reduce the cost of treatments. Reducing readmission rates is the vital step to be achieved due to which quality of treatments will improve and treatment costs will be reduced to much extent.

6. Solution Methodology



7. Data Pre-Processing

7.1 NULL Values Treatment

In the dataset, features like weight, payer_code, medical_speciality, diag_1, diag_2, diag_3 have '?' character which needs to be replaced by null values. This has been converted first. Once the character was replaced, we calculated the percentage of NULL values for each variable. Below are the details:

Sl. No.	Feature Name	% of Null Values
1	weight	96.86
2	medical_specialty	49.08
3	payer_code	39.56
4	race	2.23
5	diag_3	1.4
6	diag_2	0.35
7	diag_1	0.02

Table 1 - Null value percentage

For all the above features, as per their data types & % missing we followed different methods to impute the missing values which are mentioned below.

i. **Weight & Medical Speciality:**

The variables 'weight' & 'medical_speciality' has 96.86% & 49.08% null values respectively which exceeds the threshold of 40% so we decided to drop the columns for our further analysis.

ii. **Payer Code:**

Though the 'payer_code' has 39.56% null values, we decided to drop the variable considering it is not important predictor for our further analysis of detecting readmission rate.

iii. **Race:**

The variable race has 2.23% of null values. Considering that it is a categorical value & may be an important predictor, we have imputed the null values with KNN imputer & it was imputed with mode i.e. 'Caucasian' class.

iv. Diagnosis 1, Diagnosis 2, Diagnosis 3:

The variables 'diag_1', 'diag 2', 'diag 3' are having 0.02%, 0.35% & 1.4% null values respectively. Considering the number of instances to be less, we have dropped the respective rows having the null values.

7.2 Duplicate Entries

In the dataset, we observed that in the span of 10 years, there are people who have been readmitted multiple times. So, we decided to use only the 1st encounter per patient from the 'patient_nbr' column as it represents unique id for each patient.

After all the data pre-processing is over, we are left with 70,230 instances of data & 47 variables.

8. Data Exploration (EDA):

8.1 Univariate Analysis

a. Age

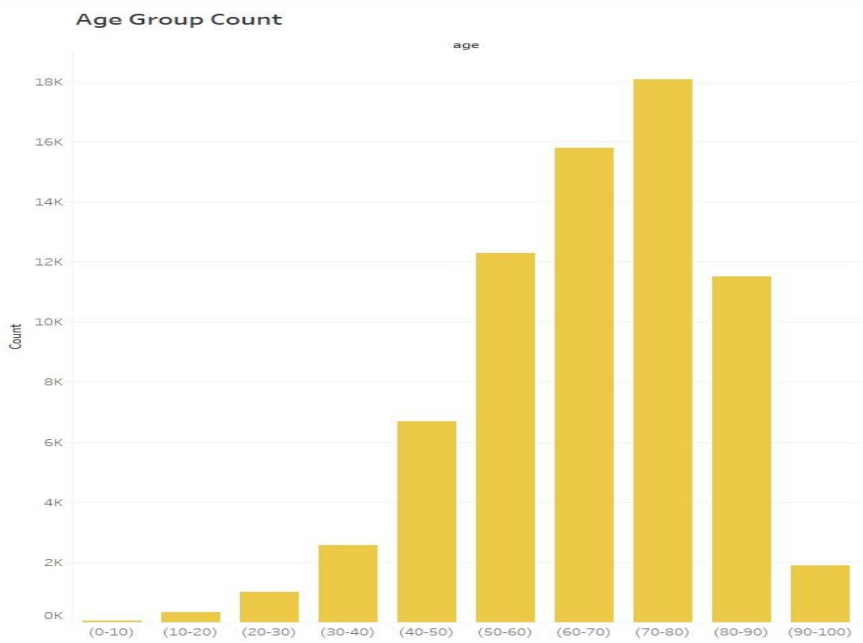


Fig. 1

1. From the count plot it is clear that the age group(70-80) is most impacted by Diabetics.
2. The graph shows that with increase in the age, the no of diabetes patients increases.

b. Diagnosis

a. Patients are diagnosed for primary, secondary and additional secondary diagnosis(*columns diag_1, diag_2, diag_3*)

b. Count plot given below shows the no of diagnosis for all the three given heads

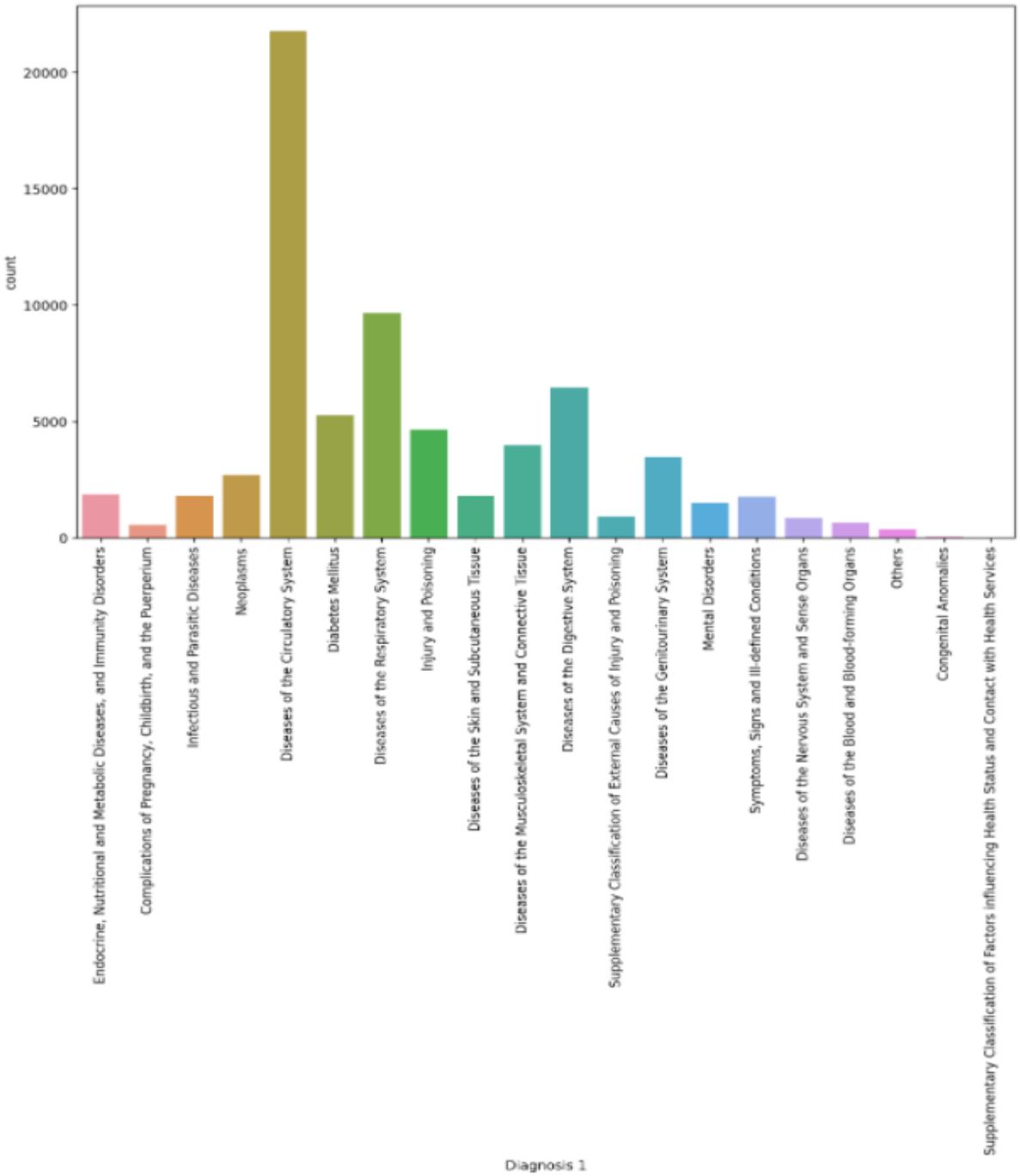


Fig. 2

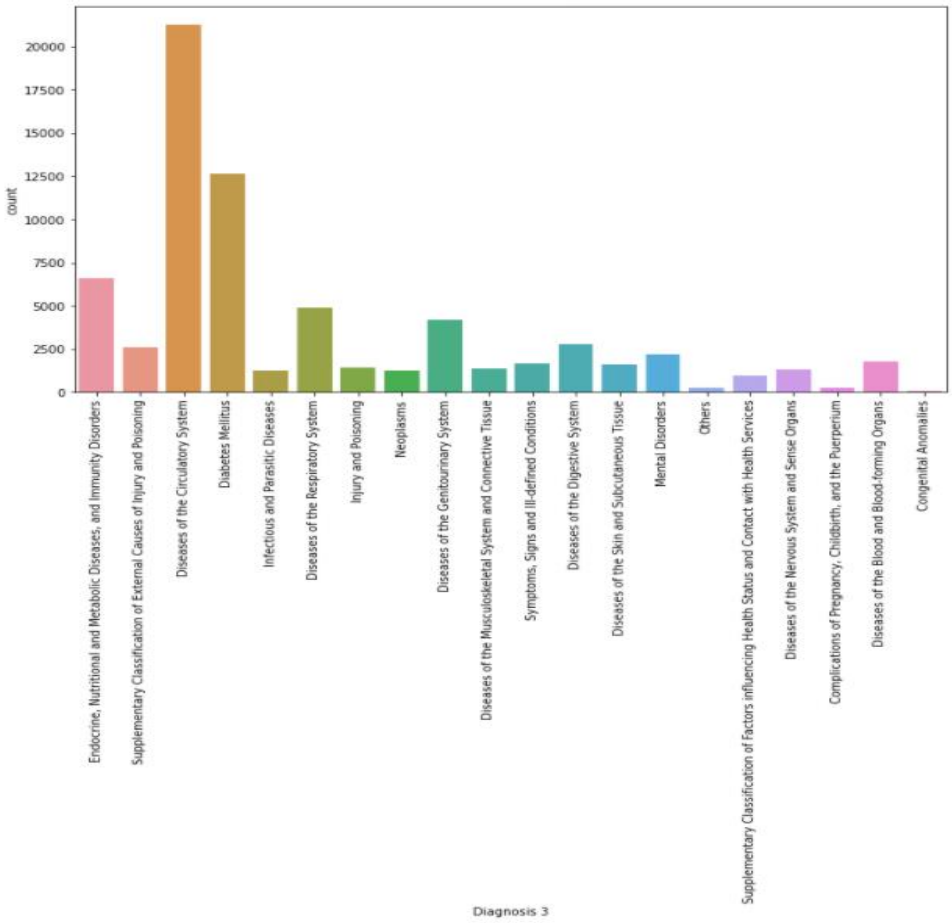


Fig. 3

a. It is seen that in every head, 'Diseases of Circulatory System' had maximum count suggesting that it is major factor influencing diabetics.
b. Also diabetics was least impacted by the diseases of Nervous System, Digestive system, Pregnancy, Sensory organs.

c. Readmitted

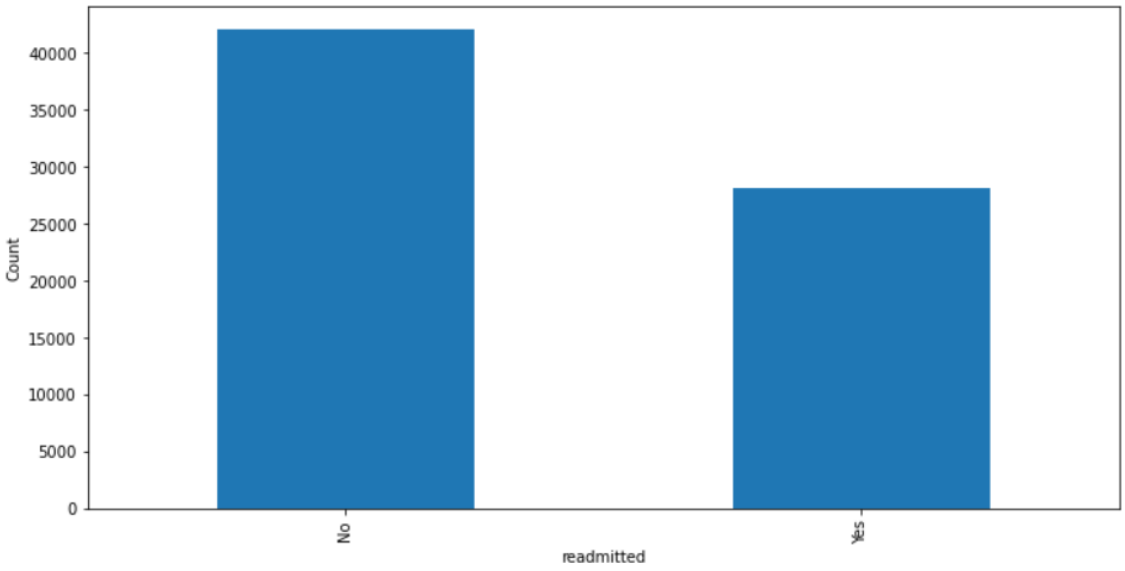


Fig. 4

1. There are two classes in count plot, 'Yes' denotes no of patients re-admitted, whereas 'No' denotes number of patients cured and didn't require any further treatment.
2. There is significant difference between the classes 'Yes' and 'No' and 'No' class shows maximum count suggesting that the prescribed treatment for treating diabetes is effective.

d. Number of Lab Procedures:

1. It indicates the number of lab tests performed during the encounter

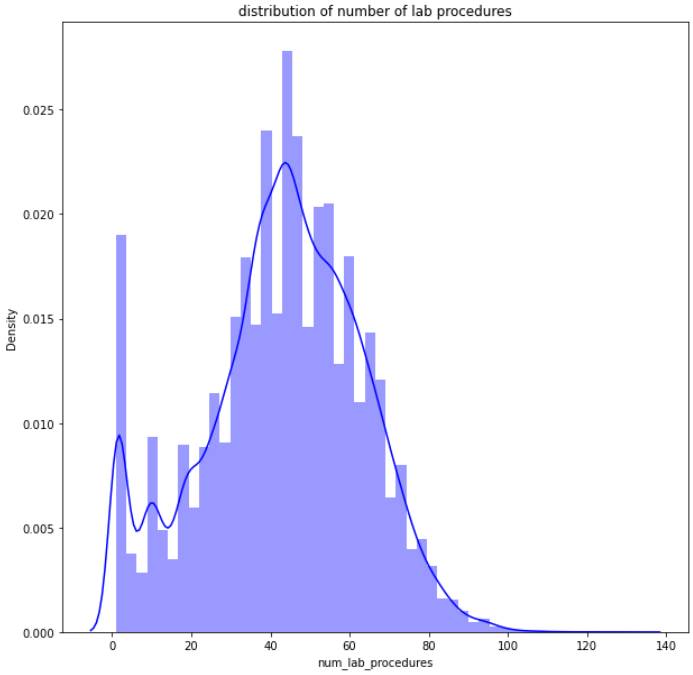


Fig. 5

2. From the distplot it is clear that a maximum of 100 lab tests were performed on a particular patient during the encounter.
3. The maximum no of patients were treated with 40-60 lab tests, while there were also considerable values for no lab tests.

e. Number of Medications

1. It is a number of different generic names administered during the encounter.

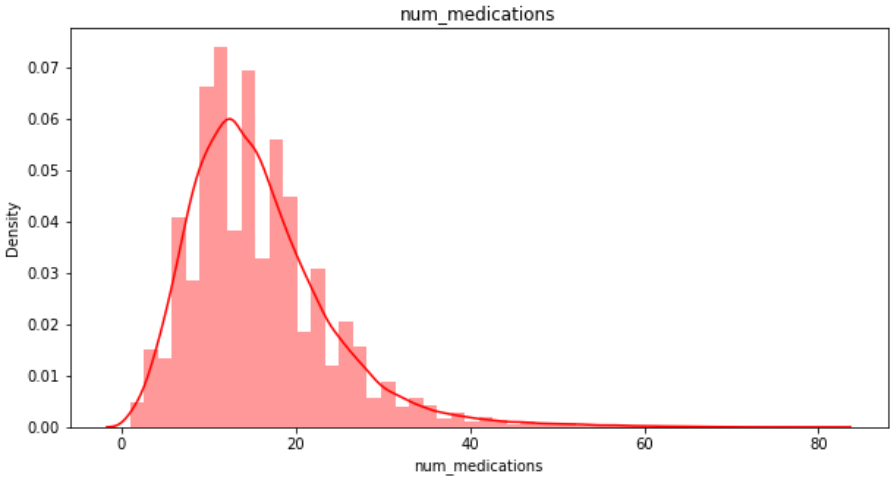


Fig. 6

2. The number of medications ranges from 0 to 40 where maximum no of patients suffering from diabetes were treated with 10-15 medications.

f. No. of Procedures

1. They are a number of tests performed other than lab tests during the encounter

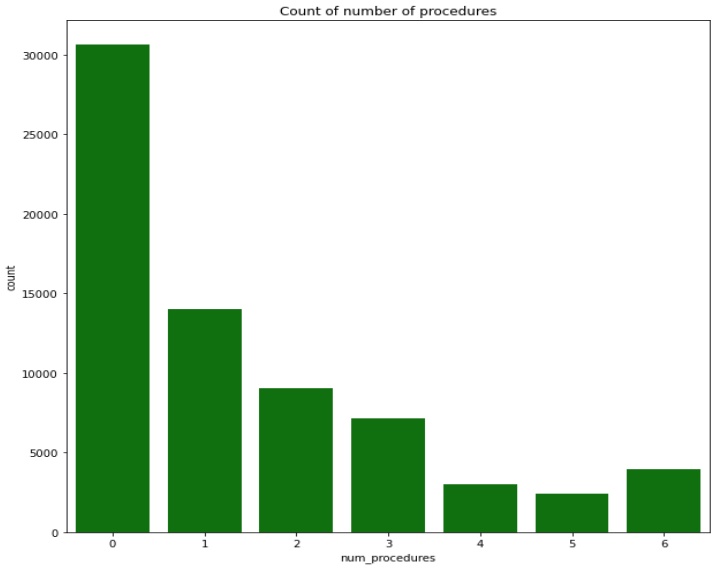


Fig. 7

2. The count plot suggests that there were few additional tests performed other than lab tests.

g. Admission Type ID

1. It contains 9 distinct integer values where the values denotes below categories:
   - 'Emergency'
   - 'Urgent'
   - 'Elective'
   - 'Newborn'
   - 'Not Available'
   - 'NULL'
   - 'Trauma Center'
   - 'Not Mapped'
   - 'Not Available'.

2. Count plot shows distribution of each of the listed categories

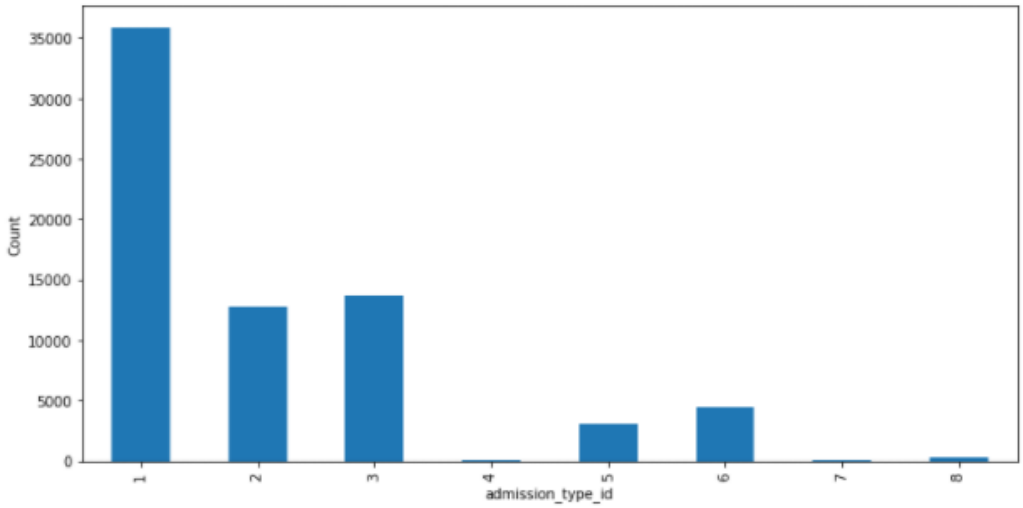


Fig. 8

3. Maximum patients suffering from diabetics were 'Emergency', while the categories 'Elective' and 'Urgent' have also shown significant values.

8.2 Bi-variate Analysis

a. Race vs Diabetes Patients:

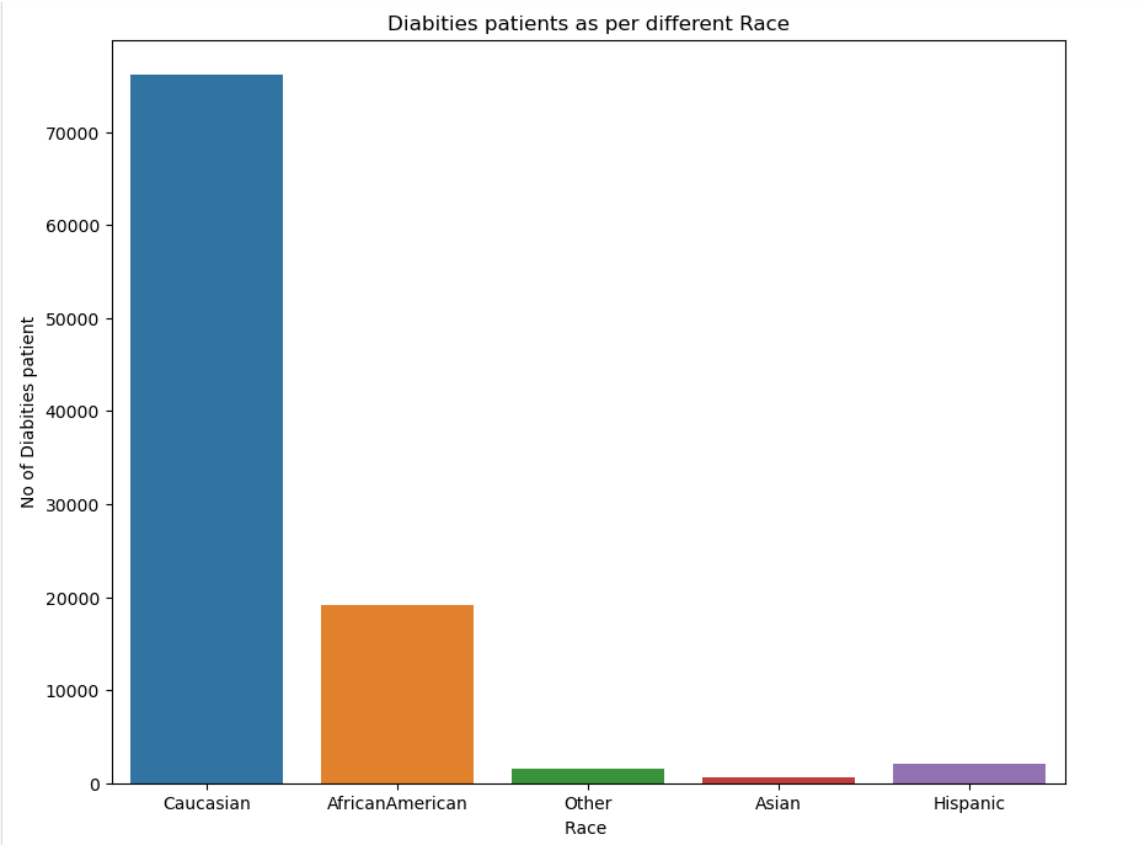


Fig. 9

From the count plot it can be deduced that the 'Asian' race has the least count which means that out of the races considered for analysis it is healthiest.

On the other hand, 'Caucasian' race has the highest number indicating that it has impacted this race the most

b. Age Vs. Readmitted

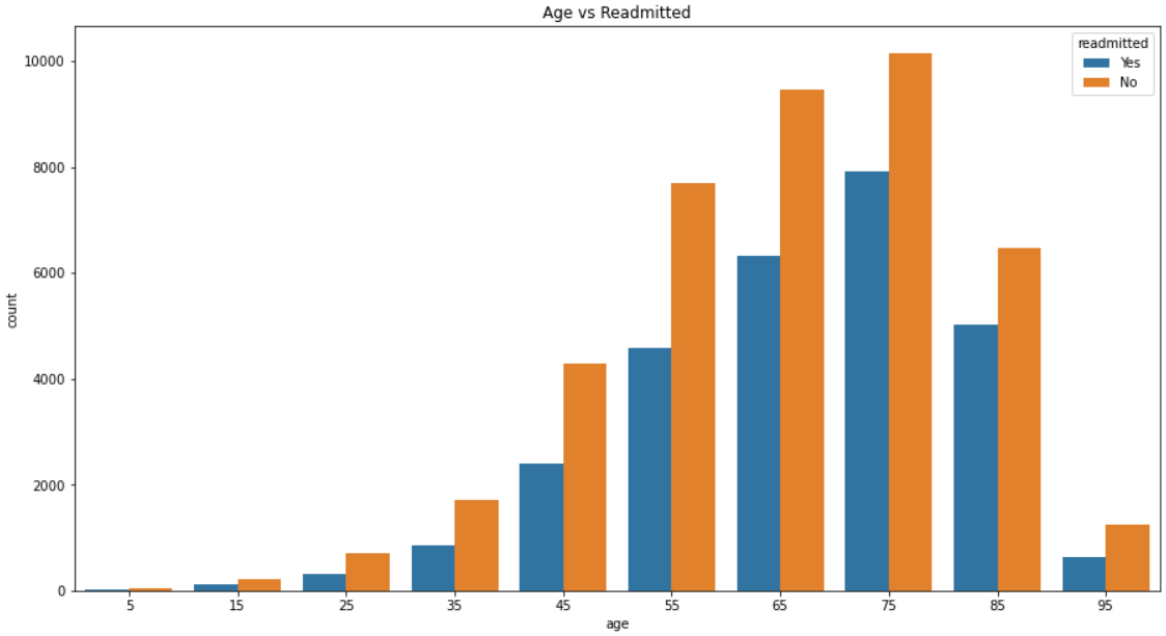


Fig. 10

1. Age influenced readmission rate and it was majorly observed with patients in age group 55-85 years.

c. Other Drugs

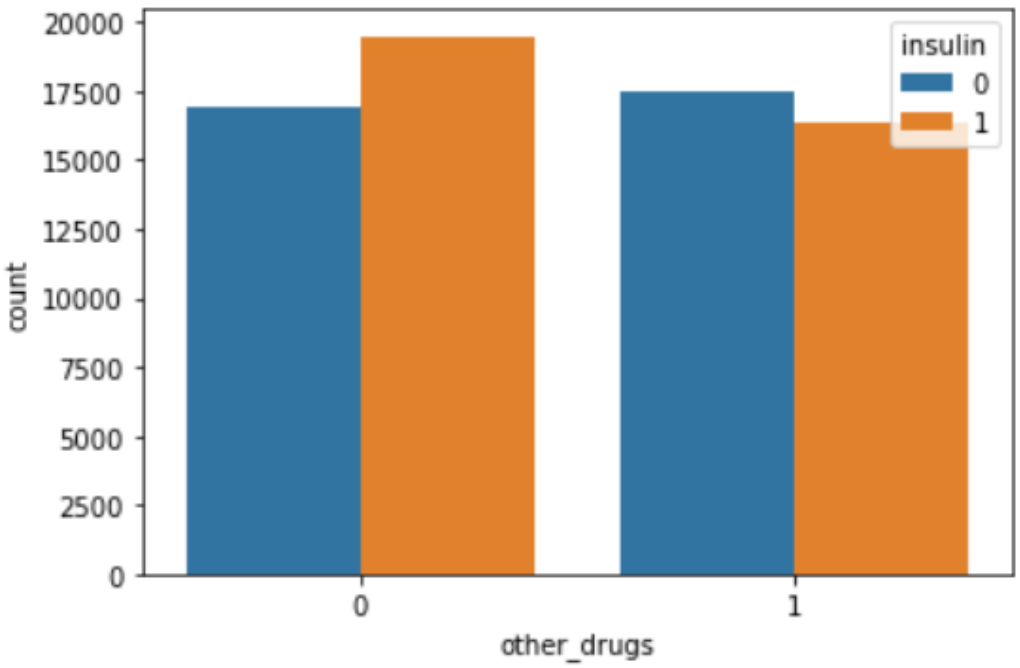


Fig. 11

1. From above plot we can infer how the medications were prescribed, whether one or combination of drugs and insulin were given together, or were they given separately, or no medication was given
2. Approximately 17000 patients were not given any medication

c. Treatment type (solo-insulin and combination of insulin and other drugs):

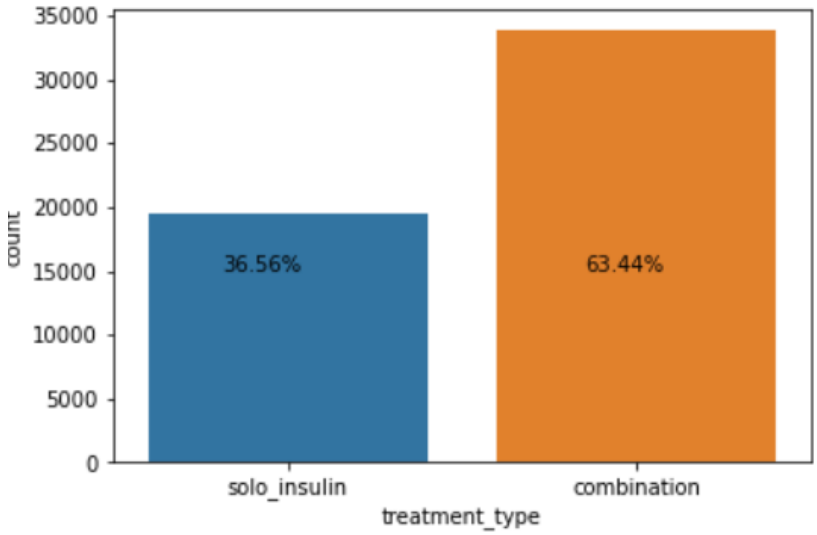


Fig. 12

1. Majority patients(about 63%) are treated with a combination of insulin and other drugs

8.3 Multi-variate Analysis

Multi- collinearity in the dataset:

1. To check multi-collinearity in the dataset, a heat map was plotted from sea born Library package.
2. The heat map shows the extent of relationship between various variables (value of co-relationship coefficient 1 or -1 shows strong co-relation while the value of 0 indicates no – correlation).

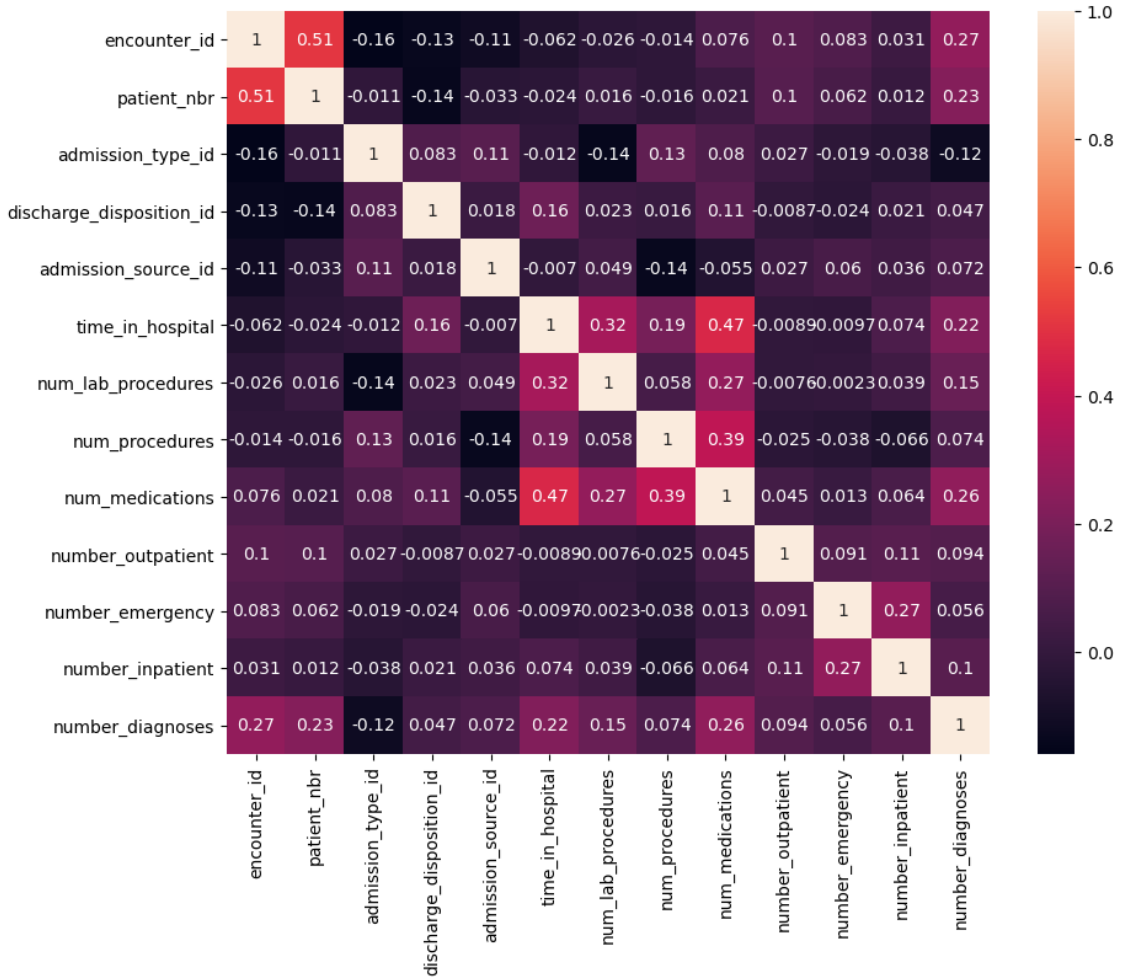


Fig.13

3. From the heat map it is clear that there is no multi-collinearity in the dataset.

Presence of Outliers and it's Treatment:

We plot a boxplot to check presence of outliers in the data set

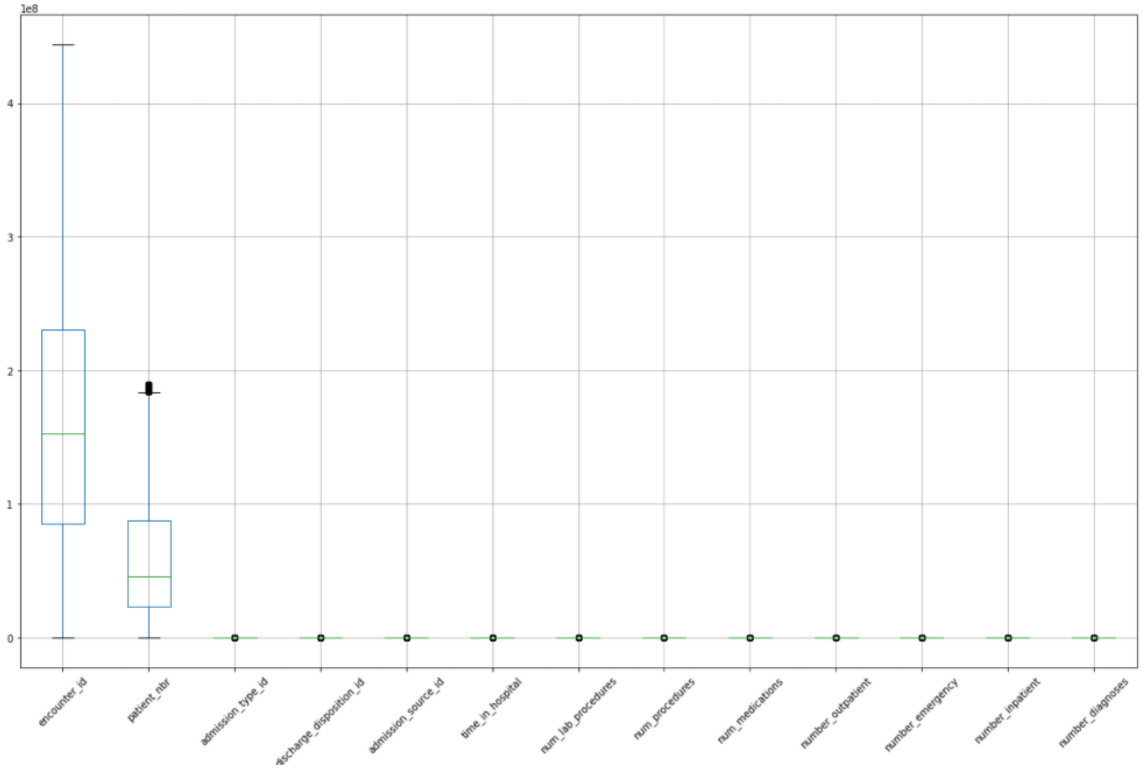


Fig. 14

1. From the box-plot it is clear that the variable 'patient_nbr' contains significant outliers.
2. The outliers can be treated using IQR(Inter Quartile Range) method
3. Steps performed in IQR:
   i. Find first quartile(Q1)
   ii. Find third quartile(Q3)
   iii. IQR = Q3-Q1
   iv. Define lower normal data range with lower limit as (Q1- 1.5*IQR) and upper limit as (Q3+ 1.5*IQR)
   v. Any data point outside this range is considered as outlier and should be removed for further analysis

Statistical significance of variables:

i. Numeric variables:

	encounter_id	patient_nbr	admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	num_lab_procedures	num_procedures
count	1.017660e+05	1.017660e+05	101766.000000	101766.000000	101766.000000	101766.000000	101766.000000	101766.000000
mean	1.652016e+08	5.433040e+07	2.024006	3.715642	5.754437	4.395987	43.095641	1.339730
std	1.026403e+08	3.869636e+07	1.445403	5.280166	4.064081	2.985108	19.674362	1.705807
min	1.252200e+04	1.350000e+02	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
25%	8.496119e+07	2.341322e+07	1.000000	1.000000	1.000000	2.000000	31.000000	0.000000
50%	1.523890e+08	4.550514e+07	1.000000	1.000000	7.000000	4.000000	44.000000	1.000000
75%	2.302709e+08	8.754595e+07	3.000000	4.000000	7.000000	6.000000	57.000000	2.000000
max	4.438672e+08	1.895026e+08	8.000000	28.000000	25.000000	14.000000	132.000000	6.000000

Fig. 15

1. From the count values in statistical summary, it can be concluded that there are some patients admitted to hospital for multiple times since it exceeds the no of unique entries in the dataset

ii. Categorical variables:

	race	gender	age	weight	payer_code	medical_specialty	diag_1	diag_2	diag_3	max_glu_serum	A1Cresult	metformin	repaglinide
count	101766	101766	101766	101766	101766	101766	101766	101766	101766	101766	101766	101766	101766
unique	6	3	10	10	18	73	717	749	790	4	4	4	4
top	Caucasian	Female	[70-80)	?	?	?	428	276	250	None	None	No	No
freq	76099	54708	26068	98569	40256	49949	6862	6752	11555	96420	84748	81778	100227

Fig. 16

1. The statistical summary of categorical variables provides information on the count of each variables, unique entries, max occurring field and it's corresponding frequency.

Class imbalance and it's treatment:

1. There were 2 classes considered in the target variable for this classification problem depending upon the treatment performed.
2. Only Insulin or combination other_drugs (Insulin and other drugs)
3. The distribution of the variables is as below:
   Only Insulin: 19,487
   Other_drugs: 33,818
4. From the distribution of classes, it is clear that the classes were imbalanced. So, we applied SMOTE technique to handle the imbalances.

9. Machine Learning

9.1 Base Model:

Logistic Regression

- It is a predictive algorithm using independent variables to predict the dependent variable. It is like Linear Regression, but with a difference that the dependent variable should be categorical variable.
- Independent variables can be numeric or categorical variables, but the dependent variable will always be categorical.
- Logistic regression is a statistical model that uses Logistic function to model the conditional probability.

Representation of Logistic Regression

y = e^ (b0 + b1*x) / (1 + e^ (b0 + b1*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

Train- Test and Split:

We divide the dataset into train and test in the ratio of 70:30, train the model on training dataset and validate results using test datasets.

Performance Metrics:

Many learning algorithms have been proposed. It is often valuable to assess the efficacy of an algorithm. In many cases, such assessment is relative, that is, evaluating which of several alternative algorithms is best suited to a specific application.

People even end up creating metrics that suit the application. In this article, we will see some of the most common metrics in a classification setting of a problem.

The most commonly used Performance metrics for classification problem are as follows,

- Accuracy
- Confusion Matrix
- Precision, Recall, and F1 score
- ROC AUC
- Log-loss

Accuracy

Accuracy is the simple ratio between the number of correctly classified points to the total number of points.

To calculate accuracy, scikit-learn provides a utility function.

```
from sklearn.metrics import accuracy_score          #predicted y values
y_pred = [0, 2, 1, 3]                               #actual y values
y_true = [0, 1, 2, 3]accuracy_score(y_true, y_pred)
0.5
```

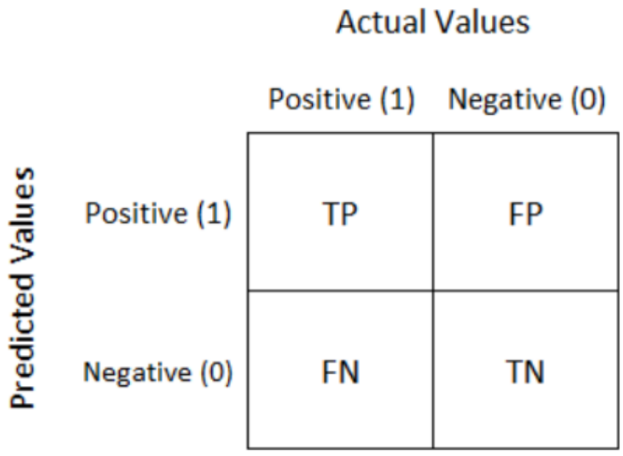
Accuracy is simple to calculate but has its own disadvantages.

Limitations of accuracy

- If the data set is highly imbalanced, and the model classifies all the data points as the majority class data points, the accuracy will be high. This makes accuracy not a reliable performance metric for imbalanced data.
- From accuracy, the probability of the predictions of the model can be derived. So from accuracy, we can not measure how good the predictions of the model are.

Confusion Matrix

Confusion Matrix is a summary of predicted results in specific table layout that allows visualization of the performance measure of the machine learning model for a binary classification problem (2 classes) or multi-class classification problem (more than 2 classes)



Confusion matrix of a binary classification

- TP means **True Positive**. It can be interpreted as the model predicted positive class and it is True.
- FP means **False Positive**. It can be interpreted as the model predicted positive class but it is False.
- FN means **False Negative**. It can be interpreted as the model predicted negative class but it is False.
- TN means **True Negative**. It can be interpreted as the model predicted negative class and it is True.

For a sensible model, the principal diagonal element values will be high and the off-diagonal element values will be below i.e., TP, TN will be high.

To get an appropriate example in a real-world problem, consider a diagnostic test that seeks to determine whether a person has a certain disease. A false positive in this case occurs when the person tests positive but does not actually have the disease. A false negative, on the other hand, occurs when the person tests negative, suggesting they are healthy when they actually do have the disease.

For a multi-class classification problem, with 'c' class labels, the confusion matrix will be a (c*c) matrix.

To calculate confusion matrix, sklearn provides a utility function

```
from sklearn.metrics import confusion_matrix
y_true = [2, 0, 2, 2, 0, 1]
y_pred = [0, 0, 2, 2, 0, 2]
confusion_matrix(y_true, y_pred)
array([[2, 0, 0],
       [0, 0, 1],
       [1, 0, 2]])
```

Advantages of a confusion matrix:

- The confusion matrix provides detailed results of the classification.
- Derivates of the confusion matrix are widely used.
- Visual inspection of results can be enhanced by using a heat map.

Precision, Recall, and F-1 Score

Precision is the fraction of the correctly classified instances from the total classified instances. **Recall** is the fraction of the correctly classified instances from the total classified instances. Precision and recall are given as follows,

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

Mathematical formula of Precision and Recall using the confusion matrix

For example, consider that a search query results in 30 pages, out of which 20 are relevant. And the results fail to display 40 other relevant results. So the precision is 20/30 and recall is 20/60.

Precision helps us understand how useful the results are. Recall helps us understand how complete the results are.

But to reduce the checking of pockets twice, the F1 score is used. F1 score is the harmonic mean of precision and recall. It is given as,

$$F1\ score = \frac{2*Precision*Recall}{Precision+Recall}$$

When to use the F1 Score:

- The F-score is often used in the field of information retrieval for measuring search, document classification, and query classification performance.
- The F-score has been widely used in the natural language processing literature, such as the evaluation of named entity recognition and word segmentation

Log Loss

Logarithmic loss (or log loss) measures the performance of a classification model where the prediction is a probability value between 0 and 1. Log loss increases as the predicted probability diverge from the actual label. Log loss is a widely used metric for Kaggle competitions.

$$\text{log-loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1 - y_i) \log(1 - p_i).$$

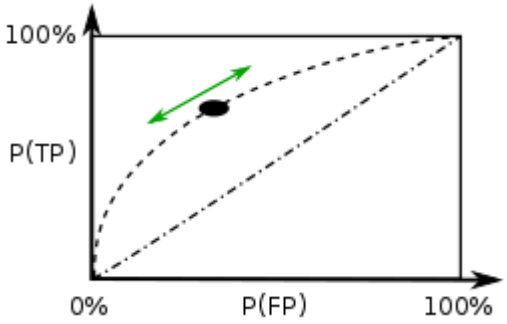
Here 'N' is the total number of data points in the data set, yi is the actual value of y and pi is the probability of y belonging to the positive class. Lower the log-loss value, better are the predictions of the model.

To calculate log-loss, scikit-learn provides a utility function.

from sklearn.metrics import log_losslog_loss(y_true, y_pred)

ROC AUC

A Receiver Operating Characteristic curve or **ROC curve** is created by plotting the True Positive (TP) against the False Positive (FP) at various threshold settings. The ROC curve is generated by plotting the cumulative distribution function of the True Positive in the y-axis versus the cumulative distribution function of the False Positive on the x-axis.



The dashed curved line is the ROC Curve. The area under the ROC curve (ROC AUC) is the single-valued metric used for evaluating the performance.

The higher the AUC, the better the performance of the model at distinguishing between the classes.

In general, an AUC of 0.5 suggests no discrimination, a value between 0.5–0.7 is acceptable and anything above 0.7 is good-to-go-model

When to use ROC:

- ROC curves are widely used to compare and evaluate different classification algorithms.
- ROC curve is widely used when the dataset is imbalanced.
- ROC curves are also used in verification of forecasts in meteorology

9.2. Pros and Cons:

Pros of Logistic Regression:

i) Logistic Regression performs well when the dataset is linearly separable.

ii) Logistic regression is less prone to over-fitting.

iii) Logistic Regression not only gives a measure of how relevant a predictor is, but also its direction of association. iv) Logistic regression is easier to implement, interpret and very efficient to train.

Cons of Logistic Regression:

i) Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables

ii) If the number of observations are lesser than the number of features, Logistic Regression should not be used, otherwise it may lead to overfit

iii) Logistic Regression can only be used to predict discrete functions.

Therefore, the dependent variable of Logistic Regression is restricted to the discrete number set.

To improve our prediction capability, we fit our train dataset with different models and compare their performances.

Reselecting our Base Model :

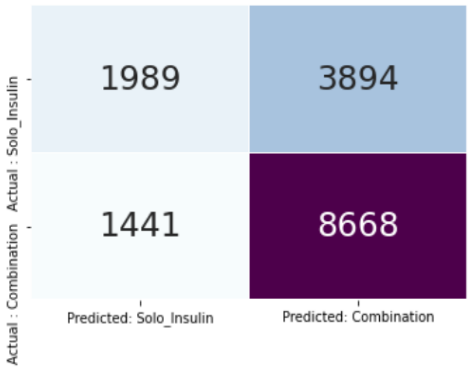
MODEL	TRAIN	TEST	PRECISION	RECALL	F1
Logistic Regression	0.67	0.67	0.69	0.86	0.76
Decision Tree	0.68	0.67	0.65	0.67	0.64
Random Forest	0.99	0.67	0.68	0.69	0.67
KNN	0.74	0.61	0.58	0.61	0.59
Naïve Bayes	0.65	0.65	0.63	0.65	0.61

Base Model Results:

The base Logistic Regression model gives the following prediction:

	precision	recall	f1-score	support
0	0.58	0.34	0.43	5883
1	0.69	0.86	0.76	10109
accuracy			0.67	15992
macro avg	0.63	0.60	0.60	15992
weighted avg	0.65	0.67	0.64	15992

```
Accuracy score for the logistic model is: 66.64
F1 score for the model is: 64.05
```



Base model's precision, recall and F1-score are not so appealing so further techniques such as SMOTE, feature selection and feature tools are applied to improve the metrics of the predictive model.

F1 Score is 64.05%. We try to improve this score using SMOTE, an oversampling technique, to deal with class imbalance.

9.3 Ensemble Techniques:

Ensemble Technique combines several individual predictive models to come up with the final predictive model for better accuracy of the model.

1. Bagging
2. Boosting

Bagging and Boosting are ensemble techniques that reduce bias and variance of a model. It is a way to avoid overfitting and underfitting in Machine Learning models.

Bagging:

Bagging is a powerful ensemble method that helps to reduce variance, and by extension, prevent overfitting. Random forest is the most popular technique in the bagging methods category. It is used for classification as well as regression problems. Random forest is a combination of decisions to identify and locate the data point, inappropriate class. It selects a set of features, only those can decide best split at each node of the decision tree. Some random subsets are generated from the original dataset. Only a random set of features are considered to decide the best split, at each node in the decision tree. Each subset is fitted on the decision tree model. The final prediction is nothing but the average of the predictions from all decision trees.

Boosting:

Boosting is a general ensemble method that creates a strong classifier from a number of weak classifiers. This is done by building a model from the training data, then creating a second model that attempts to correct the errors from the first model.

AdaBoost:

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique that is used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the

weights are re-assigned to each instance, with higher weights to incorrectly classified instances. Boosting is used to reduce bias as well as the variance for supervised learning.

Gradient Boosting:

Gradient boosting classifier is a set of machine learning algorithms that include several weaker models to combine them into a strong big one with highly predictive output. Models of a kind are popular due to their ability to classify datasets effectively.

XGBoost:

XGBoost stands for extreme Gradient Boosting. It became popular in the recent days and is dominating applied machine learning and Kaggle competitions for structured data because of its scalability. XGBoost is an extension to gradient boosted decision trees (GBM) and specially designed to improve speed and performance.

LightGBM:

LightGBM is a fast, distributed, high performance gradient boosting framework based on decision tree algorithms, used for ranking, classification and many other machine learning tasks. It can be used in classification, regression, and many more machine learning tasks. This algorithm grows leaf wise and chooses the maximum delta value to grow.

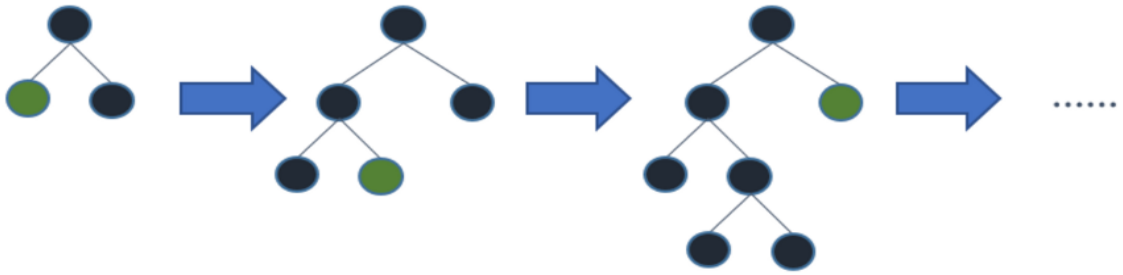
LGBM Classifier:

Light GBM is a gradient boosting framework that uses tree based learning algorithm.

Light GBM grows tree vertically while other algorithm grows trees horizontally meaning that Light GBM grows tree **leaf-wise** while other algorithm grows level-wise. It will choose the leaf with max delta loss to grow. When growing the same leaf, Leaf-wise algorithm can reduce more loss than a level-wise algorithm.

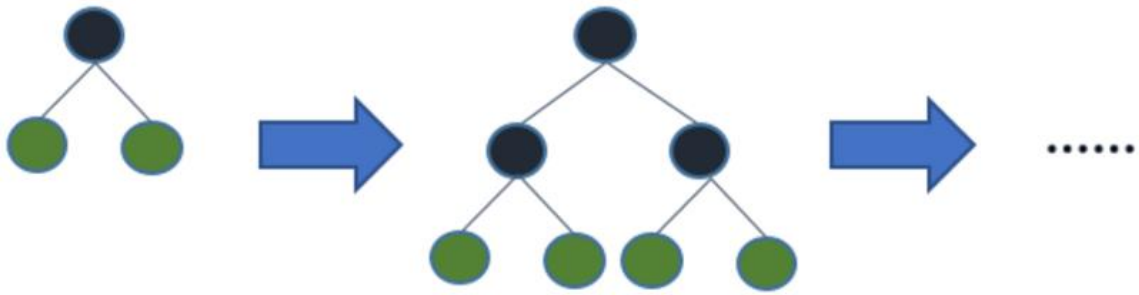
Below diagrams explain the implementation of LightGBM and other boosting algorithms.

Working of LGBM



Leaf-wise tree growth

Working of other tree-based algorithms

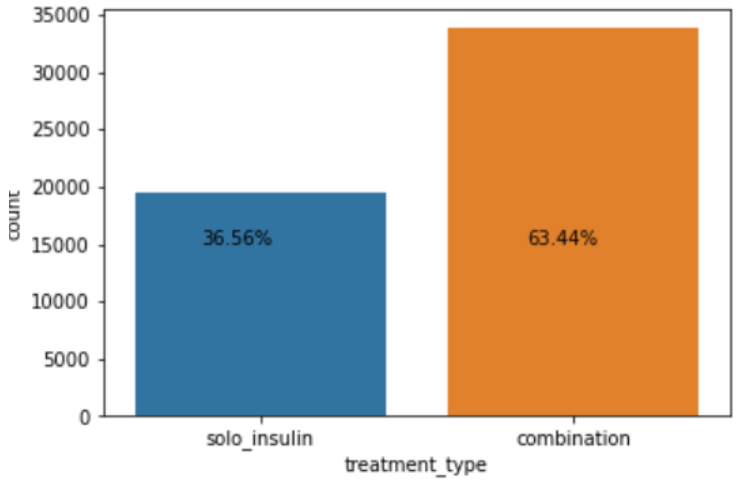


Level-wise tree growth

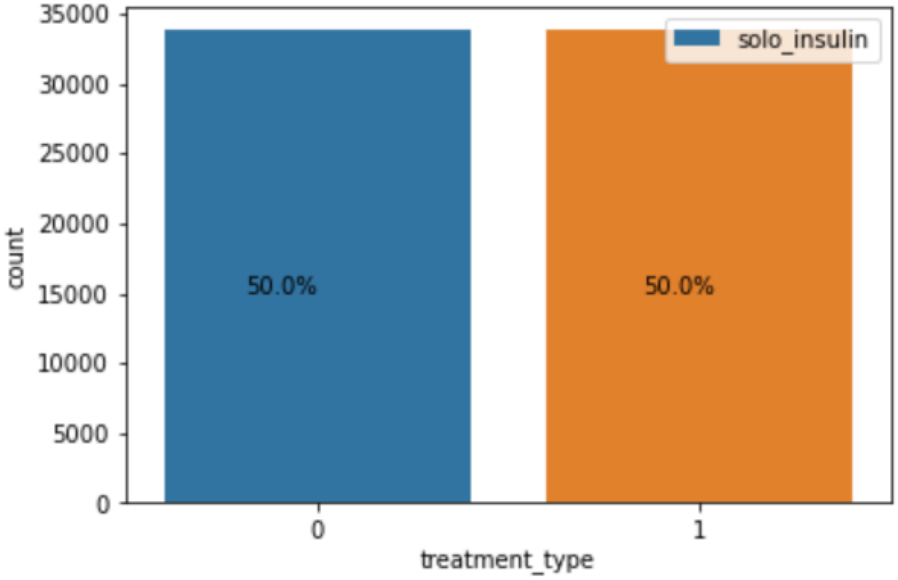
9.4 Synthetic Minority Oversampling Technique (SMOTE):

This two-class dataset is imbalanced (63% vs 37%). As a result, there is a possibility that the model built might be biased towards to the majority and over-represented
class. After applying Synthetic Minority Oversampling Technique (SMOTE) to over sample the minority class, we obtain good F1-score & Recall score with selected features.

In order to rectify this problem, we can do the following techniques to sample the data: 1. Up Sampling 2. Down Sampling 3. Up Sampling or Down Sampling using SMOTE



Even though smote gives the lesser f1 score we proceed with it since the class imbalance is treated here. So we proceed with the new train data.
After applying the SMOTE technique:



9.5 Tuning the Hyper Parameters:

A hyper parameter is a parameter that is set before the learning process begins. It is given by the user. We use the GridSearchCV technique to select the hyper parameters that enable our model to provide the best results. We give a range of different hyperparamater values, from which the GridSearchCV algorithm selects the optimum hyper parameters. GridSearchCV tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using

the Cross-Validation method. We then pass these hyper parameters to our base model, i.e.; we re-fit our base model before performing any further optimization techniques.

We fit the base model with the tuned parameters and we notice an increase in performance so we proceed with this base model with the tuned hyper parameters. Now the new F1 score is 69%

MODEL RESULTS AFTER HYPER PARAMETER TUNING AND SMOTE:

MODEL	TRAIN	TEST	PRECISION	RECALL	F1
Logistic Regression	0.69	0.69	0.69	0.69	0.69
Decision Tree	0.69	0.68	0.68	0.68	0.68
Random Forest	0.69	0.69	0.69	0.69	0.69
Ada Boost	0.68	0.69	0.69	0.69	0.69
Gradient Boost	0.68	0.68	0.68	0.68	0.68
XG Boost	0.75	0.72	0.73	0.73	0.73
LGBM	0.75	0.73	0.73	0.73	0.73

Feature Selection:

Recursive Feature Elimination (RFE):

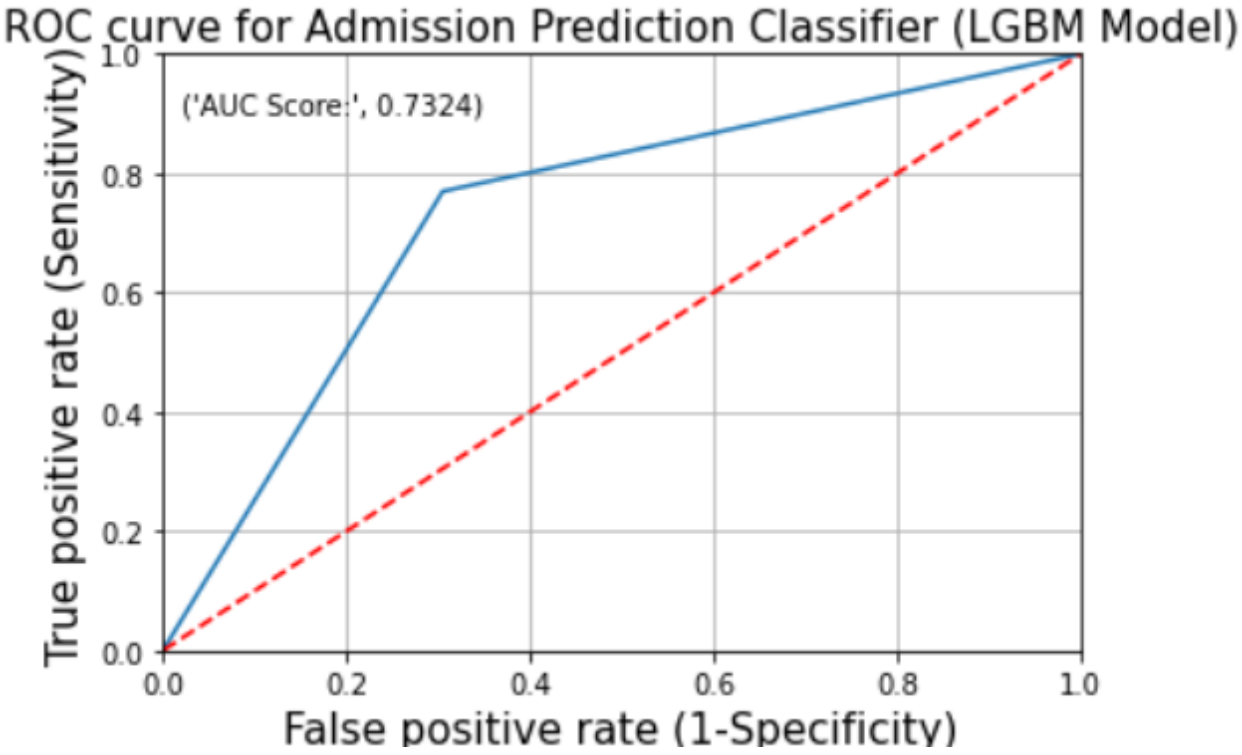
After the application of RFE, the original features were reduced from 58 to 21Features.

Index(['gender', 'age', 'admission_type_id', 'discharge_disposition_id', 'admission_source_id', 'num_lab_procedures', 'num_procedures', 'number_outpatient', 'number_emergency', 'number_inpatient', 'number_diagnoses', 'max_glu_serum', 'A1Cresult', 'change', 'readmitted', 'diag_1_LE', 'diag_3_LE', 'race__Asian', 'race__Caucasian', 'race__Hispanic', 'race__Other'], dtype='object')

These are the significant features in the model but when we built a model only with these significant features our model performance decreases. So we select our final model with tuned parametrs and all the features.

9.6 FINAL MODEL:

	precision	recall	f1-score	support
0	0.76	0.70	0.72	10257
1	0.71	0.77	0.74	10034
accuracy			0.73	20291
macro avg	0.73	0.73	0.73	20291
weighted avg	0.73	0.73	0.73	20291



It is observed that LGBM Classifier along with the features yielded the best results of all the other predictive models.

10. Business Suggestions

A business insight combines data and analysis to find meaning in and increase understanding of a situation, resulting in some competitive advantage for a business. Simply performing exploratory data analysis, building models and deriving insights won't be of any help if we are not able to leverage these insights into business solutions.

Here, we discuss some of the business solutions that will help improve the efficacy of diabetes treatment by providing patients the right sets of treatment required to treat diabetes.

Diabetes is an endemic disease and at the same time often ignored disease by the patients in the remote regions with least access to diabetes. Even if somehow they manage to reach the hospital in nearby metropolis, the skyrocketing costs of diagnosis and treatment makes it difficult for them to afford and they often refrain from treatment eventually.

Not only that, the ones living in metropolis are also risk due to the food lifestyle and their routine and it often comes late in realization that the person has a diabetes.

Moreover, there are a myriad number of treatments available to treat diabetes and often it gets late until the patient gets cured.

So, it gets important to know the right sets of treatment required for patient in treating diabetes given their medical history.

This machine learning model aims at suggesting the right combination of treatment to be taken by patients given their medical condition.

The classification model considers the eclectic data of the patients treated by diabetes across the globe. Parameters that considered include Age, Religion, Race, No of medications, treatment type (solo insulin or insulin and combination of 23 drugs).

This will ensure that the hospital overheads of the patients are reduced, and they receive the right treatment to cure them from diabetes

11. Project Outcome

There are different metrics that help in quantify performance of the model such as accuracy, recall, precision, F1 score. For our model, Accuracy score is the most accurate metric to judge the performance.

Before explaining the reason why, we will talk about the TP (True Positive) and True Negative (TN) predictions made by the model. With respect to our dataset, TP implies that the model has correctly labeled a patient who have prescribed combination of medicines as medicine. TN implies that the actual value of patient who have given solo insulin.

TN and TP both conditions may lead to reduce the Readmission rate to the hospital as and subsequent saving of the governments revenue and can reduce the overloads on the Hospitals and can cutdown the consumption of resources. Both are good to achieve the lower readmission rate or to control down the readmission rate. While precision gives weightage to FP which means that it predicted solo insulin prescribed to patient wrongly that can cause readmission rate, recall gives weightage to FN which means it predicted wrong that combination of medicines prescribed to patient which may also cause poor health of patient and may cause readmission. F1 score is the harmonic mean of precision and recall, thus giving weightage to both the conditions will also cause high readmission rate. So that we used accuracy as an evaluation metrics for our model.

Hence, Accuracy score is the best evaluation metrics for our model. We have obtained score of 75.10% & 73.23% on train and test set respectively. An AUC Score of 0.73 is closer to 1, which indicates that our model is able to distinguish between the two classes quite correctly.

Accuracy score is given as follows: (TN + TP) / (TN+FP+FN+TP)

The train score of our model is 75.10%, whereas the test score is 73.23%. The model is performing well. There is no case of overfitting and underfitting.

12. Conclusion

Our Project will evolve the classification algorithms to foresee better treatment types to turn down the readmissions for patients based on certain conditional medications like Insulin or both Insulin and medications. Since, the problem can't be disregarded due to escalated treatments costs. It becomes hinderance for patients to get their treatments done on time. Our project would simply tweak the standards of treatments. It will provide some insights for better treatment options and readmission rates of the patients that can help the patients with. This is an academic project which focuses on understanding the problems caused due to readmission, which is impacting the efficiency of the hospitals and also being a burden on diabetic patients in terms of treatment costs.

The project aims to predict better treatment option based on the historic data that can help in reducing the readmission rate which could be either only insulin or only combination of medications or both insulin & medications combined.

13. References

1. http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008
2. https://www.hindawi.com/journals/bmri/2014/781670/tab1/
3. https://towardsdatascience.com/performance-metrics-for-classification-machine-learning-problems-97e7e774a007
4. https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc