

  
Final report on

**DIABETIC PATIENT READMISSION**

**PREDICTION**

**Post Graduate Program in Data Science Engineering**

**Location: Bengaluru Batch: JAN2022**

**Submitted by:**

Atul Mathdhari

Partha Pratim Konch

Muhammed Harshad K M

Rahul Singha

Rahul Rao R

**Mentored by:**

Mr. Jatinder Bedi

**Table of Content**

|  |  |  |
| --- | --- | --- |
| Sl.No | Title | Page No |
| 1 | **Industry details** | 4 |
| 2 | **Abstract** | 4-5 |
| 3 | **Objectives** | 5 |
| 4 | **Understanding the Problem** | 5 |
| 5 | **Background Research** | 6 |
| 6 | **Application** | 7 |
| 7 | **Dataset and Domain** | 8 |
| 8 | **Data Pre-Processing**   * **8.1** Missing value and mismatched value * **8.2** Duplicate value and their treatment * **8.3** Remove Uninformative Features | 9-10 |
| 9 | **Project Details**   * **9.1** Project Statement * **9.2** Complexity Involved * **9.3** Project Outcome | 11 |
| 10 | **Exploratory Data Analysis**   * **10.1** Univariate Analysis * **10.2** Bivariate Analysis | 11-20 |
| 11 | **Feature Engineering**   * **11.1** Non-target features * **11.2** Target Feature | 21-23 |
| 12 | **Data Pre-Processing for model** | 24 |
| 13 | **Model building & parameters** | 25-26 |
| 14 | **Predictive modelling** | 27 |
| 15 | **Algorithms used** | 28-29 |
| 16 | **Results** | 29-30 |
| 17 | **Cross validation** | 30-31 |
| 18 | **Conclusion**   * **18.1** Important Features * **18.2** Business inference | 32 |
| 19 | **Limitation** | 33 |
| 20 | **Future Work** | 33 |
| 21 | **Closing Reflection** | 34 |

**1.Industry details**

* 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. 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.

### 3.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.

### 4.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.

### 5.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.

### 6.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 each category
    - For numerical variables there is statistical test carried out. All the data is made norm also 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 and needs.

The data is scaled and transformed based on the normality requirements, there is clear imbalance in the target variable. [The dataset is oversampled](https://arxiv.org/abs/1106.1813#%3A~%3Atext%3DUnder%2Dsampling%20of%20the%20majority%2Cclassifier%20to%20the%20minority%20class) [and then Logistic Regression,](https://arxiv.org/abs/1106.1813#%3A~%3Atext%3DUnder%2Dsampling%20of%20the%20majority%2Cclassifier%20to%20the%20minority%20class) [Random Forest Classifier, ADABOOST](https://arxiv.org/abs/1106.1813#%3A~%3Atext%3DUnder%2Dsampling%20of%20the%20majority%2Cclassifier%20to%20the%20minority%20class) [Classifier, XGBOOST Classifier models are built etc.](https://arxiv.org/abs/1106.1813#%3A~%3Atext%3DUnder%2Dsampling%20of%20the%20majority%2Cclassifier%20to%20the%20minority%20class)

### Dataset and Domain

#### 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

## **8.1 Missing value and mismatched value:**

## There are no standard missing values in the dataset like 0, nan etc.

## There are non-standard missing values present in the data like “?” for 5 features namely **“race”, “Gender”, “diag1”, “diag2”, “diag3”.**

## The missing value details and the treatment done are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | No of missing values | Percent | Type | Treatment |
| Race | 1948 | 2.72 | ? | Imputed it with “others” |
| Gender | 3 | 0.004 | Unknown/ Invalid | Imputed with mode ‘Female’ |
| diag1 | 1225 | 1.71 | ? | Imputed as “-1” category |
| Diag2 | 294 | 0.41 | ? | Imputed as “-1” category |
| Diag3 | 11 | 0.02 | ? | Imputed as “-1” category |

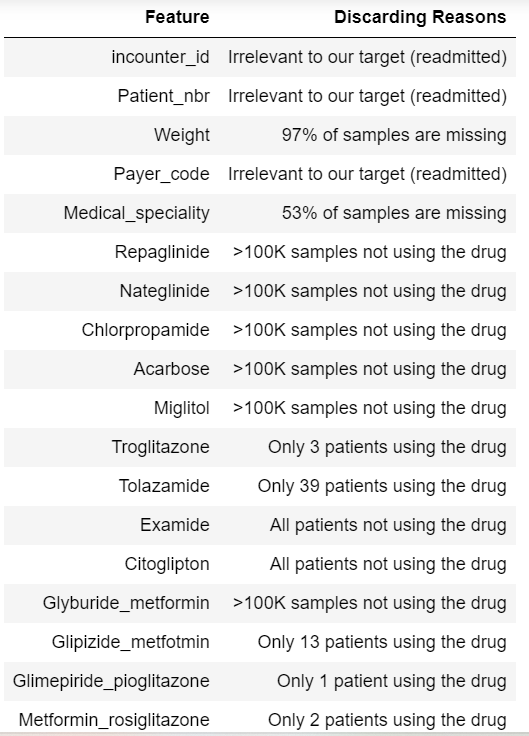
## **8.2 Duplicate value and their treatment:**

## There are many duplicate entries in the dataset.

## We have dropped / deleted the duplicate entries.

##### **8.3 Remove Uninformative Features:**

* There are 21 uninformative features in the dataset
* We have discarded the uninformative features
* The uninformative features are as follows: -



**After the data preprocessing our cleaned dataset contains 71518 entries and 29 features**

* 1. **Project Details-**

#### 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.

#### 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

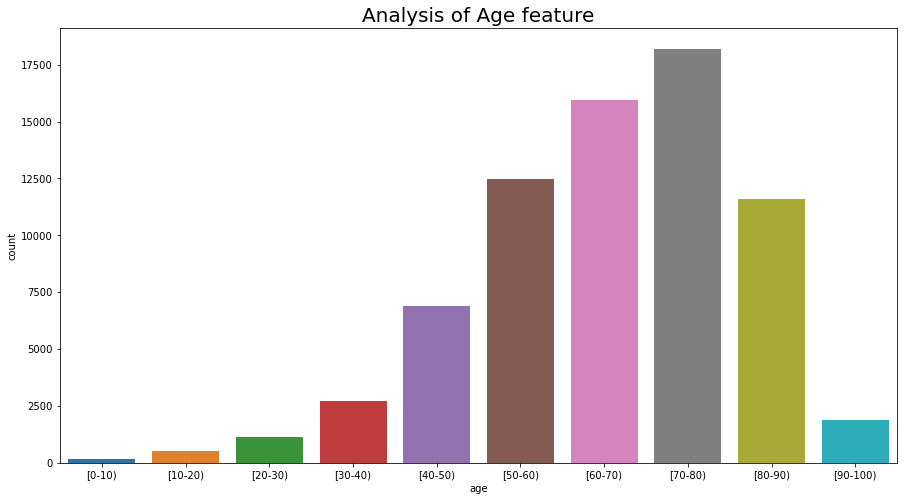
need to readmit but model say “no to readmit” then that will cause readmission cost to the hospital so, misclassification rate should be as low as possible**.**

#### 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.

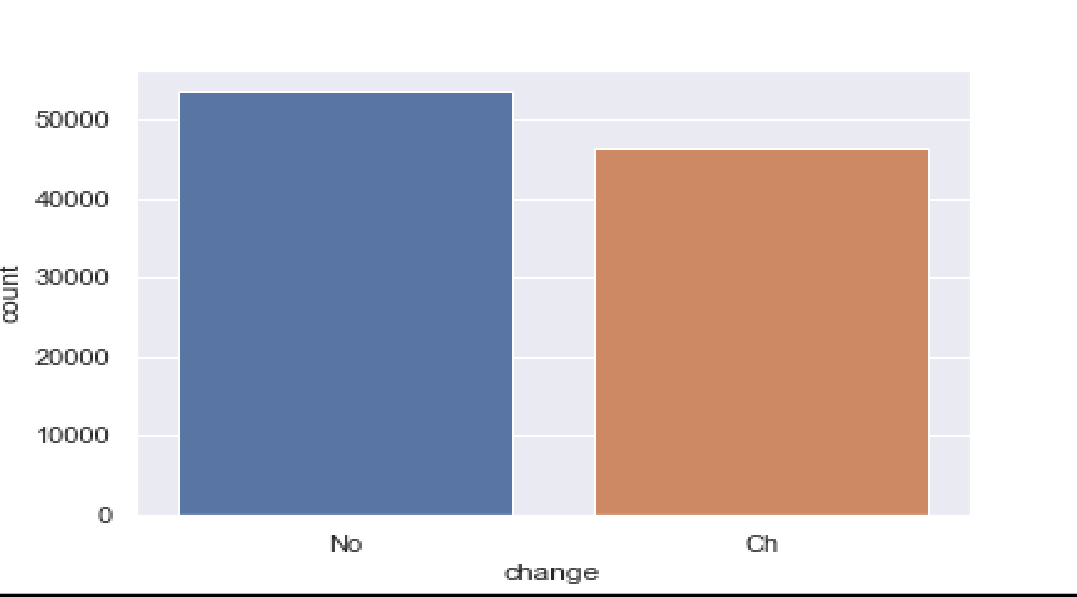
#### Exploratory Data Analysis

* 1. **Univariate Analysis**
* **Age**



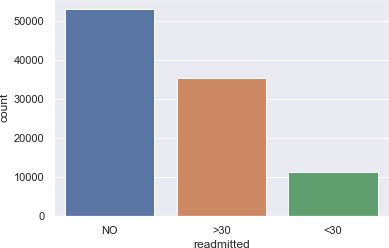
* Most of the patients falls in the age range 50-90 in our dataset.

#### Changes In Medications



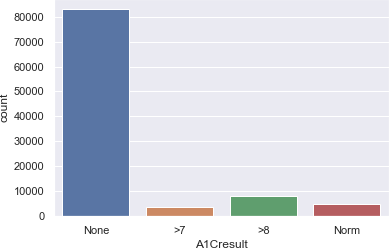
* + Graph shows that around **46% of the patients** were observed with change in their diabetic routine medications.

#### Readmitted column

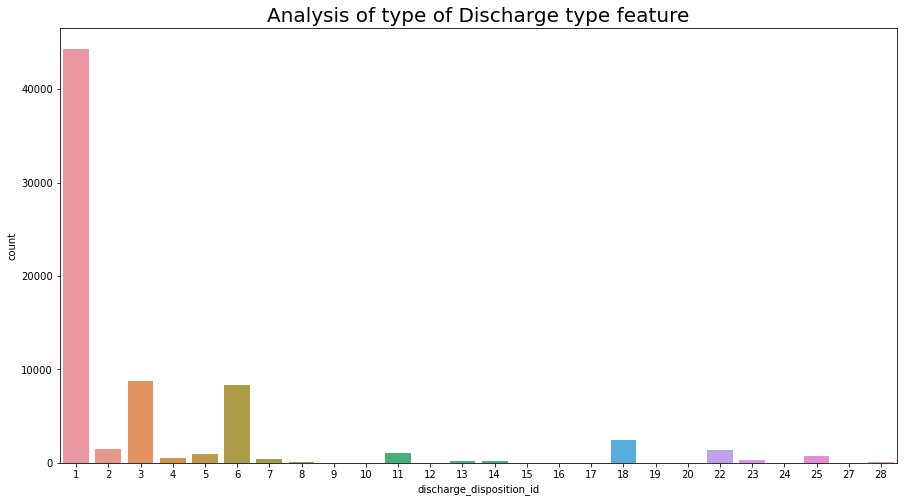


* + From our data only **11% of the patients** were readmitted within **30 days** and **35% after 30 days**. Also, there is **53% of population** which remained **unaffected of readmission** due to diabetic conditions.

#### A1Cresult column

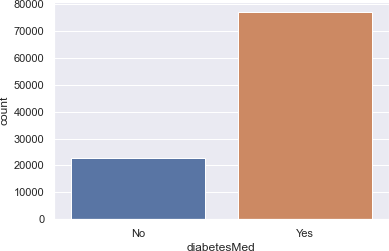


* + A1C test results denotes the average blood sugar level of the patient. Here, it is distributed in ranges where **above 6.5 denotes** that the **patient is diabetic**. It also indicates that around **12% of samples from our dataset are diabetic**. Apart from that, we have majority of unmeasured values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

****

* Discharge disposition type feature explain the type through which the patient was discharge. 1 explain the that most people got discharge to home.

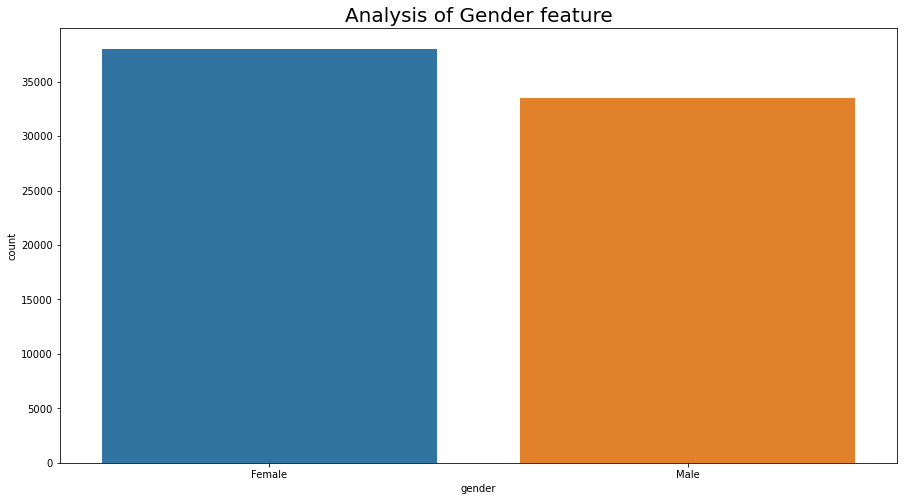
#### Diabetes Med



* + We can observe from plot that around 77% of samples from our data take

diabetes medicines.

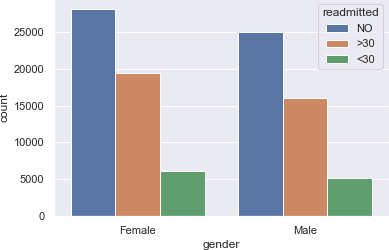
* **Gender:**

****

* + The number of patients is of gender female, also there are some missing values is in the dataset.

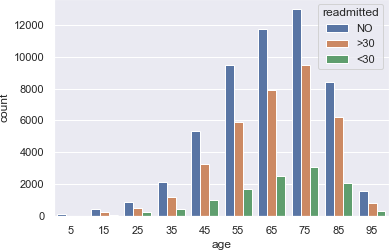
### 10.2 Bivariate Analysis

#### Gender Vs Readmitted:



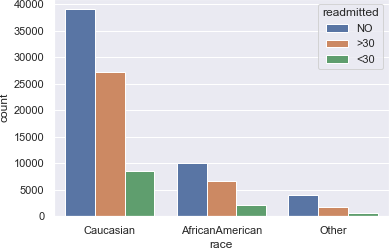
* + We can see from the above plot, the **proportion of males readmitted is almost equal to females**, although females are little more prone to be readmitted than males.

#### Age vs Readmitted:



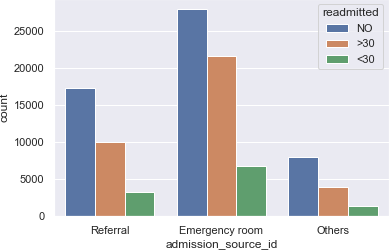
* + From the above plot and analysis over age, we can conclude that though we have our **mostly affected population to be elderly,** but we **cannot underrate** the **mid-age generation (i.e., 20-40)** because it also associates considerable ratio of affected cases.

#### Race vs Readmitted:



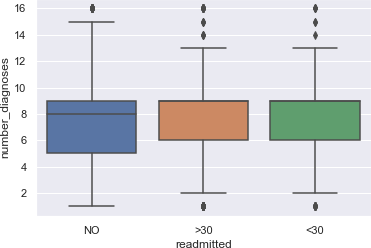
* + The above analysis indicates that although we have our **most of the population to be "Caucasians"**, but the **readmission chances of rest of the races** are almost **similar**.

#### Admission Source Vs Readmitted:



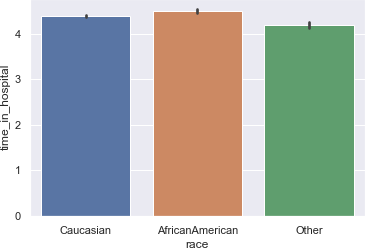
* + Although we have data **imbalance among admission\_source\_id** categories, the chances of **readmission** are almost **close to each other**.

#### Number of Diagnoses Vs Readmission:

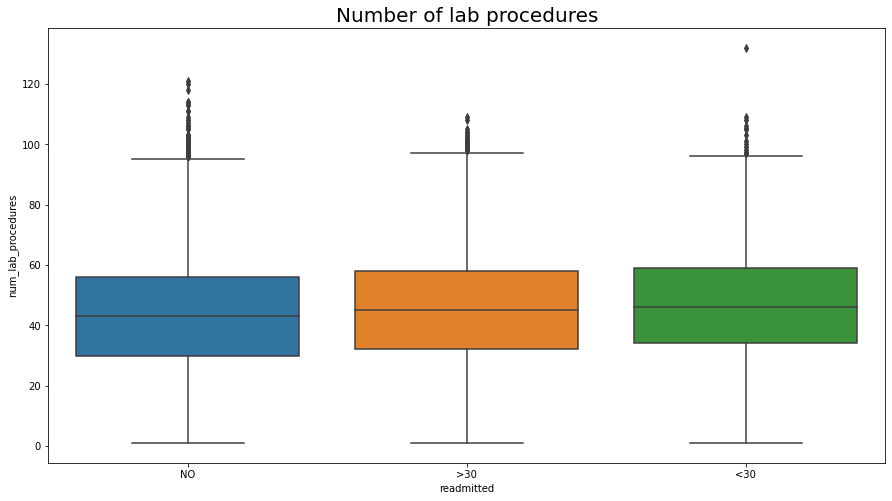


* + From Boxplots above we can infer that although the distribution of number of diagnoses is not similar, the median for all categories in readmission is almost similar. Thus, we can conclude that **readmission** is **independent** of the number of **diagnoses**.

#### Race Vs Time in Hospital:



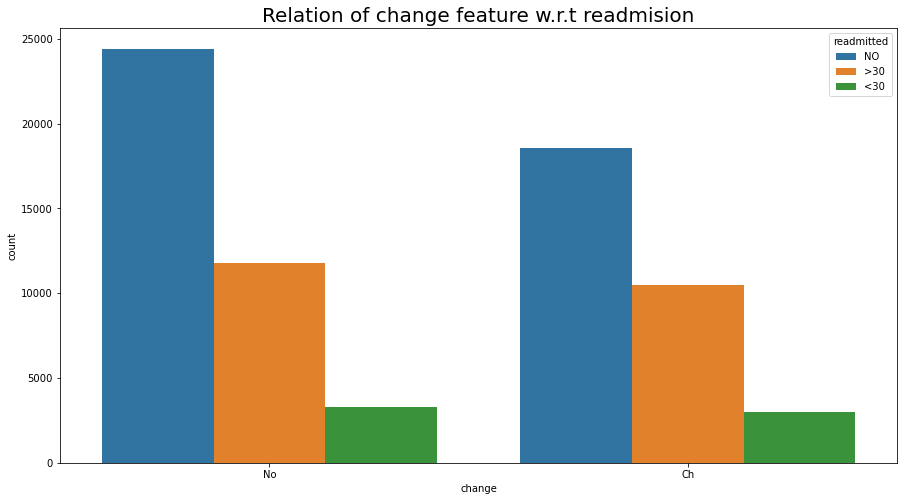
* + Irrespective of the race, the time spent by patients is almost similar. But **African Americans** are more likely to spend more time than Caucasian followed by other race.
* **Number\_of\_lab\_procedured vs readmitted:**



* + Number of lab procedures some outliers in all classes
  + Most of number of lab procedures are in range of 30 to 55 for all classes
  + Number lab procedures for classes other than class 0(no readmission) is slightly

higher

* + Their mean and median are almost similar in all classes.
* **Medication\_change vs readmitted**



* + The patient who are not getting admitted are the one for which the medication change is not occur are more than the patients which have medication change.
  + patients who are getting readmitted in more than 30 days, are somewhat same in number for both the category of change feature, that means the weather change occur in medication happen or not does not affect the readmission ratio of the patients. The number of people who are getting readmitted in less than 30 days are somewhat same for both the class of change feature, that indicates that this is not the factor affecting the readmission ratio. need further analysis for more appropriate result.

**11.0: Feature engineering:**

**11.1: Non-Target Features:**

# Discharge Disposition ID:

Discharge Disposition ID corresponding to [11 or 13 or 14 or 19 or 20 or 21] indicates patient has expired so there is no chance that it will readmit again so we will remove these records.

Discharge Disposition ID has lots of distinct values using domain knowledge we will convert them into small number of categories. They are as follows-

**cat1 = [6, 8, 9, 13]   
cat2 = [3, 4, 5, 14, 22, 23, 24]   
cat10 = [12, 15, 16, 17]   
cat11 = [19, 20, 21]  
cat18 = [25, 26]**

# Admission Type ID:

Similarly, Admission Type ID can be converted into small number of categories -

**cat1 = [2, 7]   
cat5 = [6, 8]**

# Glucose Serum test:

# A blood glucose test is used to find out if your blood sugar levels are in the healthy range. It is often used to help diagnose and monitor diabetes.

# We have categorized the Glucose Serum test as follows-

|  |  |  |
| --- | --- | --- |
| none | 0 |  |
| norm | 1 |  |
| >200 | 2 |  |
| >300 | 3 |  |

# A1C test:

# An A1C test is a blood test that reflects your average blood glucose levels over the past 3 months We have categorized the Glucose Serum test as follows-

|  |  |
| --- | --- |
| None | 0 |
| Norm | 1 |
| >7 | 2 |
| >8 | 3 |

# Diagnosis feature: The diag\_1, diag\_2, diag\_3 are categorical features but they have. lots unique values if we apply one hot encoding directly it will result in. lots of feature values that will eventually invoke "curse of dimensionality" problem.

# We will use the domain knowledge provided in the description of data to create fewer values.

# Race Feature: Imputing the nan value with other because may be the person is not comfortable with disclosing their race or maybe they couldn’t find their race in the column.

# Gender Feature:

# We have imputed the male as 1 and female as 0

**Drugs Features:**

We still have 7 features, each one represents the change in the patient's dosage of a specific drug, during hospital encounter. Those 7 drugs are the following:

* metformin
* glimepiride
* glipizide
* glyburide
* pioglitazone
* rosiglitazone
* insulin

We have categorized the drugs as follows-

# No= 0

# Steady=2

# Down=1

# Up=3

# Change Feature: We have categorized the medication changed done for the patients as follows

# No change=0

# Change=1

# DiabetesMed Feature:

# We have categorised the patients with diabetes medication as follows: -

# Yes=1

# No=0

**11.2 Target Feature:**

**Readmitted Feature:**

Redmitted column has values like '>30' that is patient readmitted after 30 days and 'NO' that is patient not readmitted and '<30' that is patient readmitted before 30 days

So, we have replaced

* ‘'NO' ---------------------0
* ‘>30‘& ‘<30' -----------1

# 12 Data Processing for Model:

# First changing the categorical columns into the numerical through one-hot encoding.

# Making the dataset into 2 different databases named as x and y for train test split.

# X is the cleaned dataset after excluding target variable.

# Y is the cleaned dataset’s having only the target variable (“readmitted”)

# We have done a train test split of 70-30 ratio respectively.

# 

# 13 Model Building & parameters:

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:



### 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).

### 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.

**ROC\_AUC:**

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

**Accuracy Score:**

Machine learning model accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data.

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

### 14 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

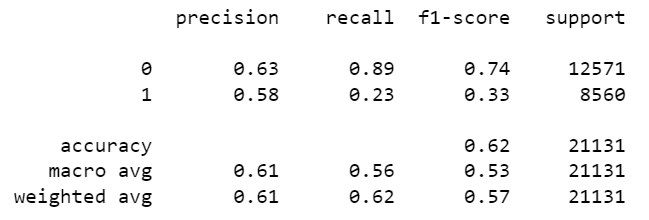
* The algorithms that are used for classification are:
* Logistic Regression.
* Grid Search for Decision tree.
* XG Boost classifier
* Gradient boost Classifier
* Random Forest.
* Ada boost with Random Forest
* Light GBM Classifier

**15 Algorithm used**

* **Logistic regression:** It 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.
* **Decision tree:** is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.
* **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.
* **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.
* **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
* **Light GBM Classifier:** Gradient-Based One-Side Sampling (GOSS) is a method that leverages the fact that there is no native weight for data instance in GBDT. Since data instances with different gradients play different roles in the computation of information gain, the instances with larger gradients will contribute more to the information gain. Thus, in order to retain the accuracy of the information, GOSS keeps the instances with large gradients and randomly drops the instances with small gradients.

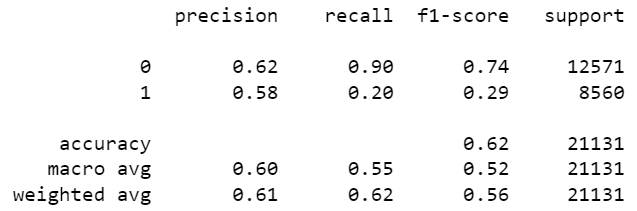
**16 Results:**

* **Logistic Regression:**

****

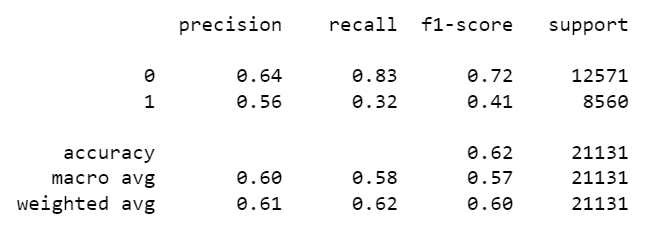
ROC\_AUC Score: 0.5583052002941052

* **Grid Search for Decision tree:**

****

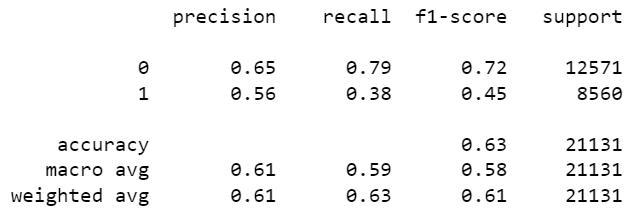
ROC\_AUC Score: 0.550364667009145

* **Random Forest Classifier:**

****

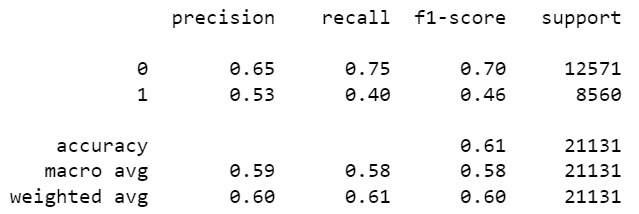
ROC\_AUC Score: 0.5750170155014843

* **AdaBoost with Random ForestClassifier:**

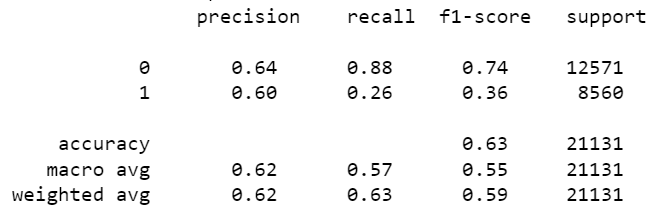


ROC\_AUC Score: 0.5876106936897488

* **XGBoost Classifier:**

****ROC\_AUC\_Score:0.5780413001813252

* **Gradient Boost Classifier:**

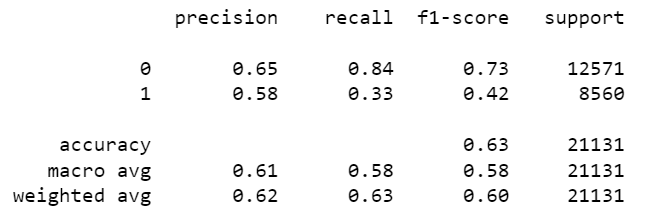


ROC\_AUC Score: 0.5721495735995248

**17 Cross Validation**:

* From the results that we obtain, we have seen that all models are overfitting.
* So, to control and check the overfitting we have done the Cross validation on the data.
* After checking for all the models, we have seen that the best performing model for the dataset comes out to be Light GBM classifier and the results are as follows: -

**Light GBM Classifier: -**

****

ROC\_AUC Score: 0.5837595913157192

Best parameters in the Light GBM are as follows: -

**Lamba 1 =1.5**

**Lambda 2 = 1.5**

**Minimum data in leaf = 30**

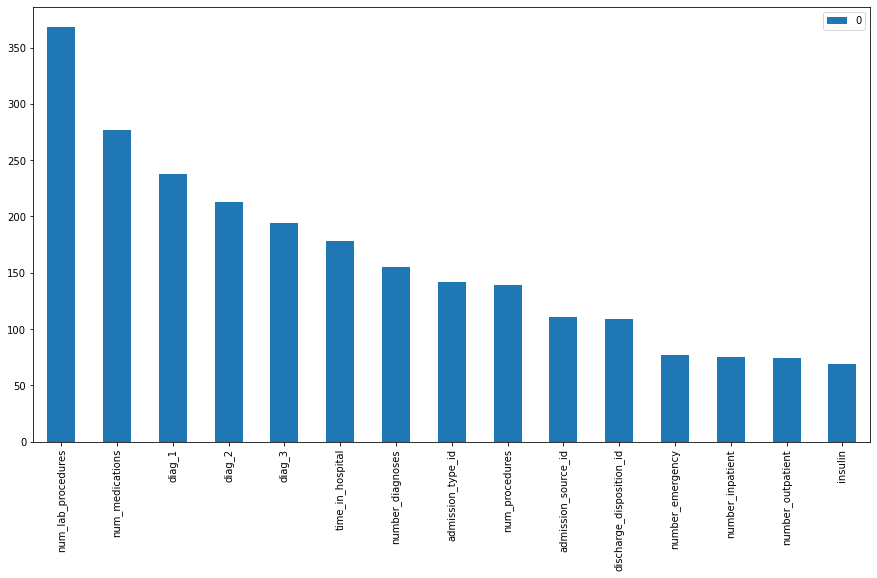
**Number of leaves = 31**

**Reg alpha = 0.1**

**18 Conclusion:**

* **18.1 Important Features:**

After performing grid search on for the best params for boosting and performing Cross-validation, we have found out the important features for the prediction are mentioned in the below figure:



# 18.2 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 (Light GBM Classifier), we have got

Precision of 58% which indicates that out of total predicted positive 58% are true positives

Recall of 63% which indicates that out of total actual positives , 63% are predicted are true positives

F1\_score, which is harmonic mean of precision and recall is around 42 percent

**19 Limitation:**

# After doing grid search and Cross validation also still the data is imbalance

# Since our data have less data so it is challenging to get good scores.

# 

# 20 Future work:

* Since the data is imbalance, we can come to better insights using deep learning.
* Deployment of the project.

# 21\_Closing Reflection:

There was application 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.

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.