****

**PREDICTING READMISSION OF DIABETIC PATIENTS**

**GROUP-1**

**TEAM MEMBERS**

1. Veeresh Kanthi
2. Ditto Paul
3. Mangala Naikar
4. Manish Mohan
5. Shabaz Khan
6. Shammu Raj U

**MENTOR**

Jatinder Bedi

**Table of Contents**

[1. Industry Review 3](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398009)

[2. Dataset and Domain](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398010)

[2.1. Data Dictionary 4](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398011)

[2.2. Varible Categorisation 5](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398012)

[2.3. Preprocessing Details 11](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398013)

[2.3.1. Dropping of Redundant columns 11](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398015)

2.3.2. Changing the datatype of the variables 11

2.3.2. Missing values/Null value treatment 13

3[.Exploratory Data Analysis](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398024)

3[.1. Univariate Analysis 1](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398025)4

[3.2. Multivariate Analysis 1](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398030)9

4. Feature Engineering

[4.1. Encoding](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398019) 20

[4.2. Scaling](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398020) 20

[4.3. Hypothesis Testing](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398019) 21

5. Model Building and Evaluation

[5.1. Model Building and Evaluation without SMOTE](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398019) 21

[5.2. Model Building and Evaluation with SMOTE 2](file:///C:\Users\pauld\Documents\Adarsh%20Exams\Project\Capstone+project+Final+version%20(1)\Capstone%20project%20Final%20version\Capstone_Grp1_ONL_DEC21%20Project%20Final%20report.docx#_Toc38398020)4

**1. Industry Review**

It is increasingly recognized that the management of hyper glycemia in the hospitalized patient has a significant bearing on outcome, in terms of both morbidity and mortality.

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 readmission is a policy priority aimed at improving healthcare quality.

Machine learning helps in providing more accurate predictions than current practices. For this study, the original dataset underwent data preprocessing, exploratory data analysis steps to ensure the success of building a robust predictive model.

**2. Dataset and Domain**

**2.1 Data Dictionary**

Health Care Data Analytics is basically the analysis of data using different techniques to discover patterns and trends and make use of this information to take actionable steps and guide the decision-making process. The goal of healthcare domain is to assure the application's quality, dependability, performance, safety, and efficiency. The main use case of Health Data Analytics is managing and controlling the spread of diseases.

Diabetes is a wide spread chronic disease that is accompanied with irregularities of blood glucose levels due to problems related to insulin. Diabetes is increasing at an alarming rate all over the world. According to the CDC’s (Centre for Disease Control) National Diabetes Statistics report for 2020, the cases of diabetes have risen to an estimated **34.2 million in US alone**. Diabetes is affected by height, race, gender, age but a major reason is a sugar concentration.

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.

**2.2Variable categorization (count of numeric and categorical)**

Dataset Name: Diabetes-130-us-hospitals-for-years-1999-2008

Link:<https://www.kaggle.com/code/iabhishekofficial/prediction-on-hospital-readmission>

Shape of the Dataset: Rows - 101766, Columns – 50

|  |  |
| --- | --- |
| Independent variables | encounter\_id, patient\_nbr, race, gender, age, weight, admission\_type\_id, discharge\_disposition\_id, admission\_source\_id, time\_in\_hospital, payer\_code, medical\_specialty, num\_lab\_procedures, num\_procedures, num\_medications, number\_outpatient, number\_emergency, number\_inpatient, diag\_1, diag\_2, diag\_3, number\_diagnoses, max\_glu\_serum, A1Cresult, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, citoglipton, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, metformin-pioglitazone, change, diabetesMed |
| Target variable | readmitted |

Details about the independent and target variables are as follows:

Out of the total 49 independent variables 11 are numerical in type and rest are of categorical type.

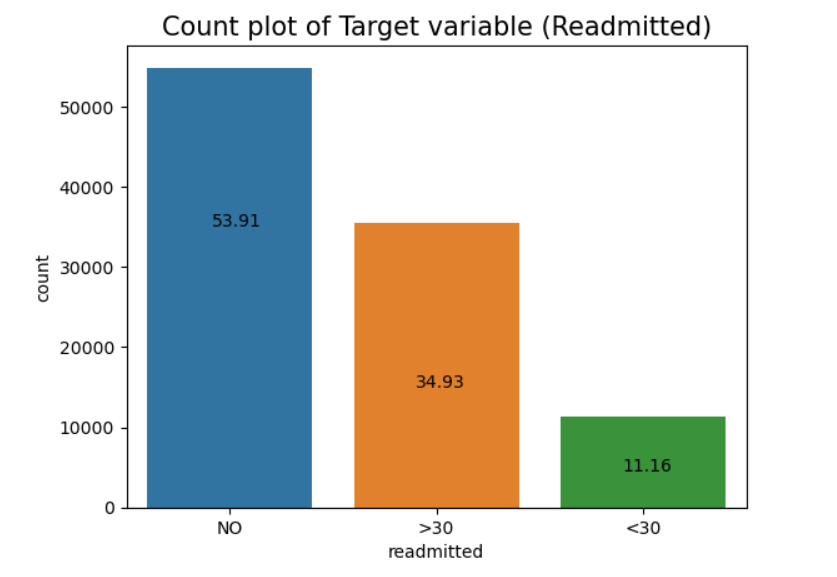
|  |  |  |
| --- | --- | --- |
| **Feature name** | **Type** | **Description** |
| Encounter ID | Numeric | Unique identifier of an encounter |
| Patient number | Numeric | Unique identifier of a patient |
| Weight | Numeric | Weight in pounds. |
| Time in hospital | Numeric | Integer number of days between admission and discharge. |
| Number of lab procedures | Numeric | Number of lab tests performed during the encounter. |
| Number of procedures | Numeric | Number of procedures (other than lab tests) performed during the encounter. |
| Number of medications | Numeric | Number of distinct generic names administered during the encounter. |
| Number of outpatient visits | Numeric | Number of outpatient visits of the patient in the year preceding the encounter. |
| Number of emergency visits | Numeric | Number of emergency visits of the patient in the year preceding the encounter. |
| Number of inpatient visits | Numeric | Number of inpatient visits of the patient in the year preceding the encounter. |
| Number of diagnoses | Numeric | Number of diagnoses entered to the system. |

**Categorical**

|  |  |  |
| --- | --- | --- |
| **Feature name** | **Type** | **Description** |
| Race | Nominal | Values: Caucasian, Asian, African American, Hispanic, and other |
| Gender | Nominal | Values: male, female, and unknown/invalid |
| Age | Nominal | Grouped in 10-year intervals: [0, 10), [10, 20), ..., [90, 100) |
| Admission type | Nominal | Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available. |
| Discharge disposition | Nominal | Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available. |
| Admission source | Nominal | Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital. |
| Payer code | Nominal | Integer identifier corresponding to 17 distinct values, for example, Blue Cross\Blue, Shield, Medicare, and self-pay. |
| Medical specialty | Nominal | Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct, values, for example, cardiology, internal medicine, family\general practice, and surgeon. |
| Diagnosis 1 | Nominal | The primary diagnosis (coded as first three digits of ICD9); 848 distinct values. |
| Diagnosis 2 | Nominal | Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values. |
| Diagnosis 3 | Nominal | Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values. |
| Glucose serum test result | Nominal | Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured. |
| A1c test result | Nominal | 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. |
| Change of medications | Nominal | Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”. |
| Diabetes medications | Nominal | Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”. |
| 23 features for medications | Nominal | For the generic names: metformin, repaglinide, nateglinide, chlorpropamide,glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. 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. |

**Target variable**

|  |  |  |
| --- | --- | --- |
| Readmitted | Nominal | 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. |

****

**Observation from Target variable:**

1. “NO”: No readmission

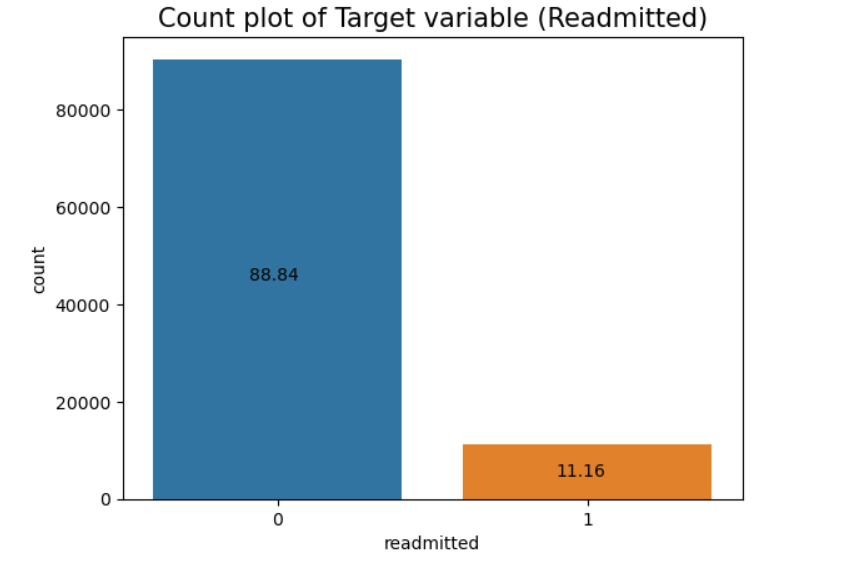
2. “<30”: It indicates Readmission < 30 days which is a very bad situation and highly correlated with the quality of care from the health center.

3. “>30”: It indicates Readmission > 30 days which is also bad but has less correlation with the quality of care from the health center and might be due to other external factors.

There are 3 classes in the target variable and hence we classify the problem into two based on our problem statement:

“**0**”: Not Readmitted = (NO or >30 days)

“**1**”: Readmitted = (< 30 days)



From the above count plot, it is clear the data is imbalanced.

**2.3 Preprocessing Data Analysis**

Data preprocessing, a component of [data preparation](https://searchbusinessanalytics.techtarget.com/definition/data-preparation), describes any type of processing performed on [raw data](https://searchdatamanagement.techtarget.com/definition/raw-data). It is a step that involves transforming raw data so that issues owing to the incompleteness, inconsistency, and/or lack of appropriate representation of trends are resolved to arrive at a dataset that is in an understandable format.

**2.3.1 Dropping of Redundant columns**

1) As we observed the columns ‘encounter\_id’ and ‘patient\_id’ have a greater number of unique values so, these columns can be dropped.

encounter\_id: 101766

patient\_nbr: 71518

2) Also, the columns “examide” and “citoglipton” have only one unique value and hence these also does not have an impact on our target variable. So, these columns can be dropped too.

**2.3.2 Changing the datatype of the variables**

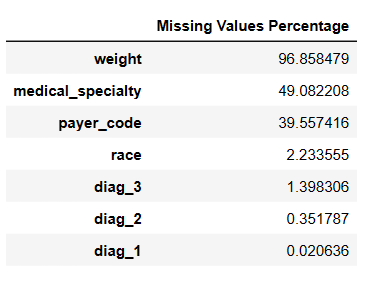
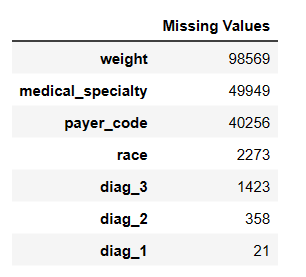
The variables “admission\_type\_id”, “discharge\_disposition\_id” and “admission\_source\_id” are categorical even though they have numerical values. There are 8 unique values for “admission\_type\_id” and we grouped few of the variables together and made it to 3 unique values. We did the same for “discharge\_disposition\_id” and “admission\_source\_id”. The 28 unique values of “discharge\_disposition\_id” were converted into 3 unique values and similarly 26 unique values of “admission\_source\_id” were converted to 3 unique values. This grouping helps to reduce the complexity in these variables.

**2.3.3 Missing values/Null value treatment**

Missing data is defined as the values or data that is not stored (or not present) for some variables in the given dataset.

In the dataset, blank shows the missing values. Usually in datasets, missing values are represented by Nan. In the dataset taken for analysis, missing value are represented by **‘?’.**

Below are the number of missing values and percentage of missing data.



The rate of missing values for “weight”, “medical specialty” and “payer\_code” are so high. So, either these columns can be dropped or can be imputed.

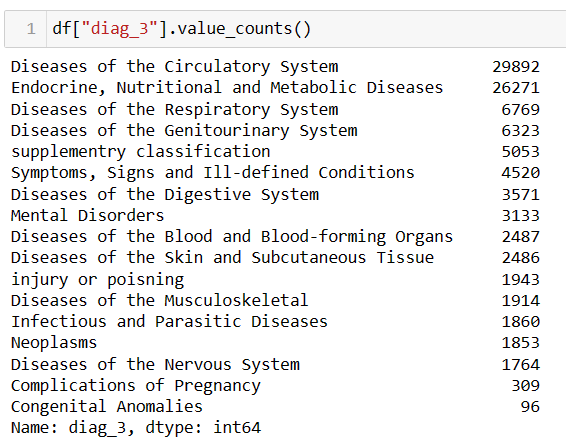
“payer\_code” can be removed since it doesn’t have any significant impact on our target variable. The two approaches taken were:

1. Drop the columns weight and medical\_specialty
2. Impute the columns with mode of the data

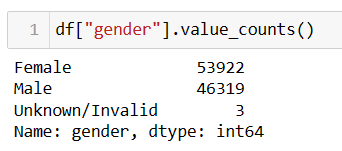
We were getting similar results by using two approaches and hence we decided to drop all the three columns.

The variable race is having 2.23% of the missing data. We replace those values with “Other” since we do not know which race does that category belongs to.

In case of the columns “diag\_1”, “diag\_2” and “diag\_3”, we are going to remove the rows which has null values. Then we did replace the numerical values with corresponding categorical values for better understanding. Below is the value counts of column “diag\_3” after we replace it with categorical values.



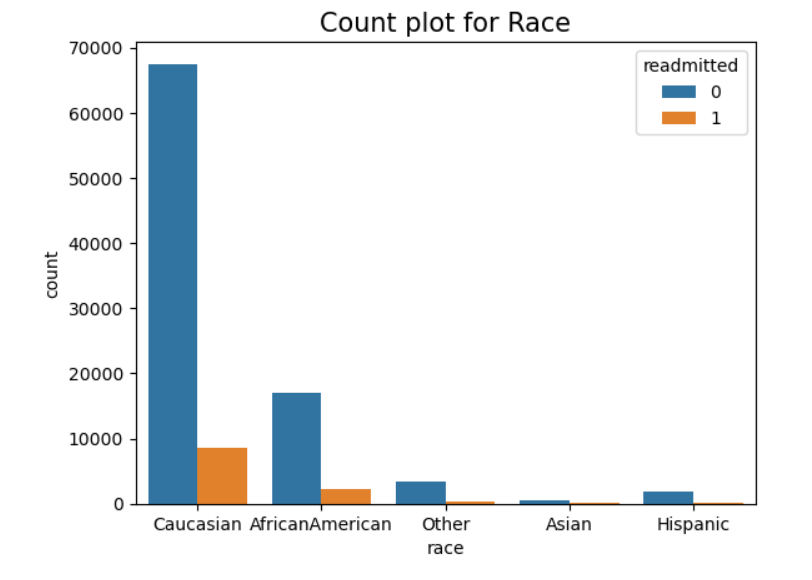
In case of column “Gender”, there were 3 unknown values, we removed those rows as well.



**3. Exploratory Data Analysis (EDA)**

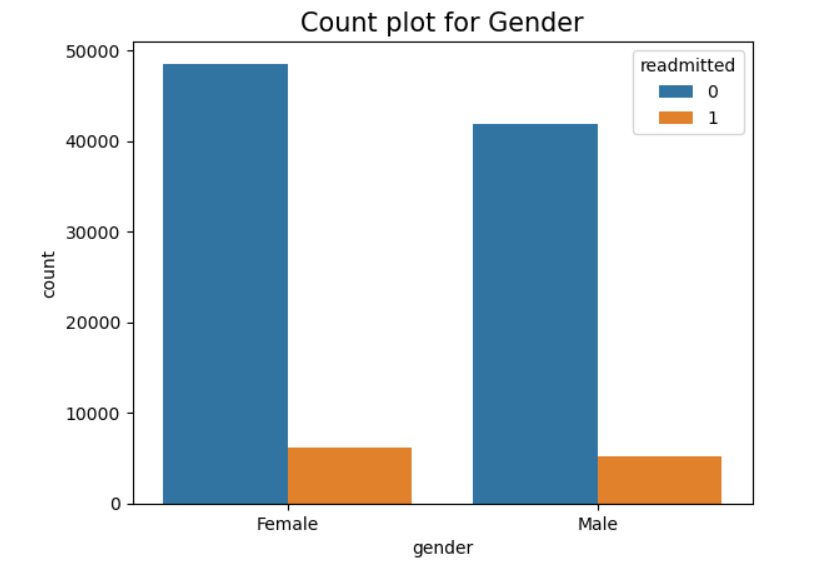
The relationship between different variables and the target column are as follows:

**3.1 Univariate Analysis**

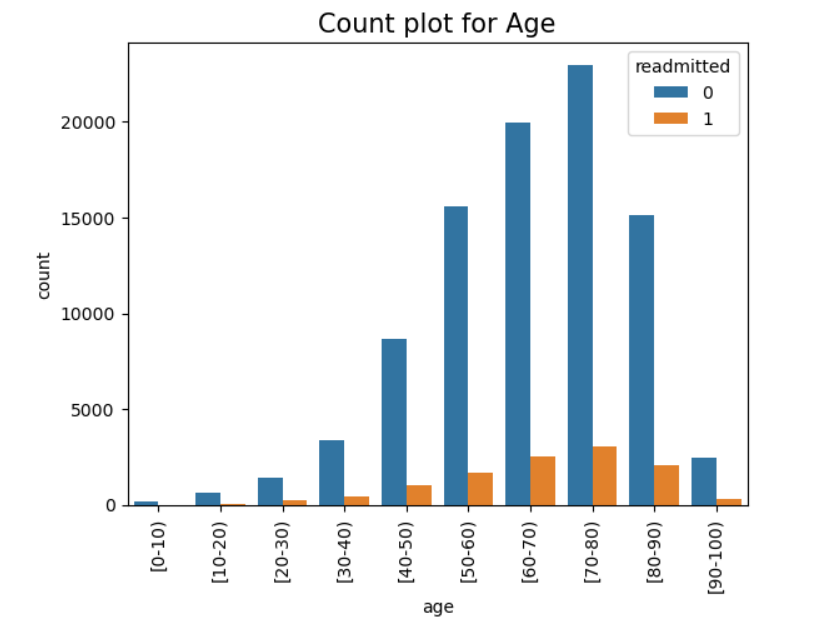


Caucasian has a high chance of readmission followed by African American and Other races.

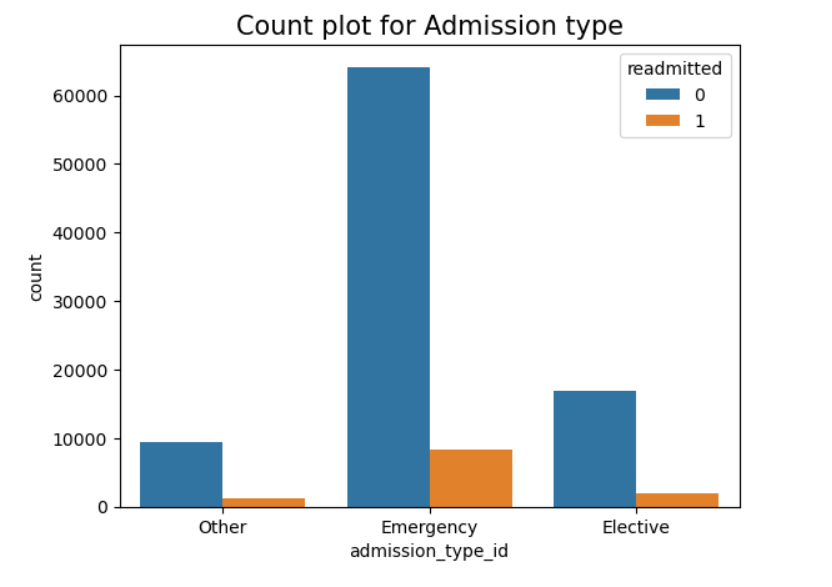
Gender does not have a significant impact on whether a patient gets readmitted or not.



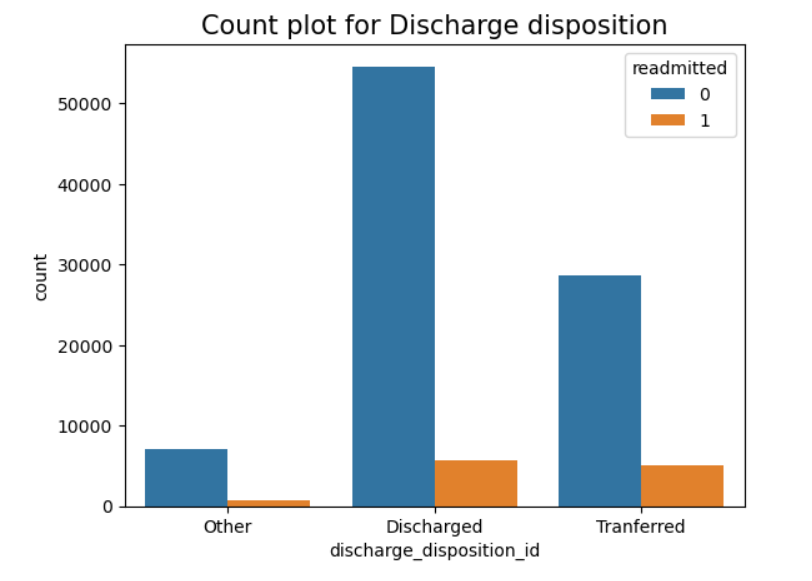
Age group [70 – 80) has higher chance of getting readmission followed by ages [60 – 70) and [80 – 90)



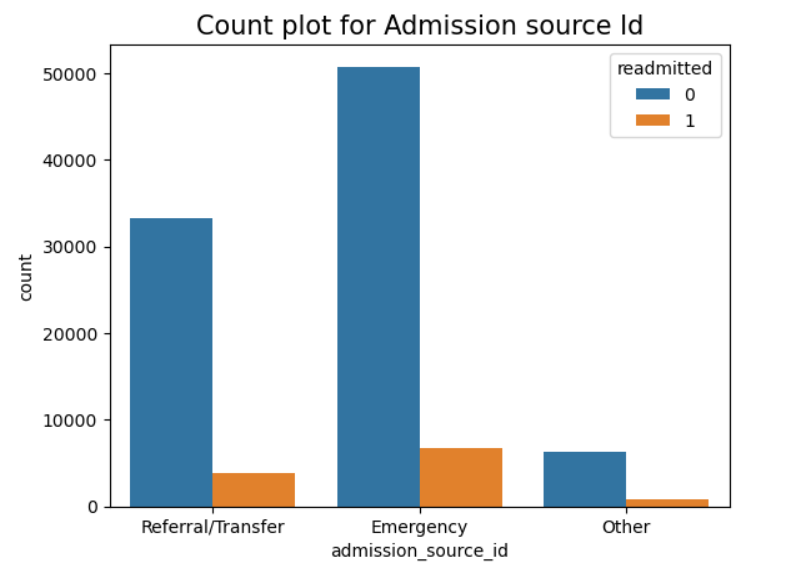
Patients who are admitted as Emergency has a higher chance of readmission than Elective and Others.



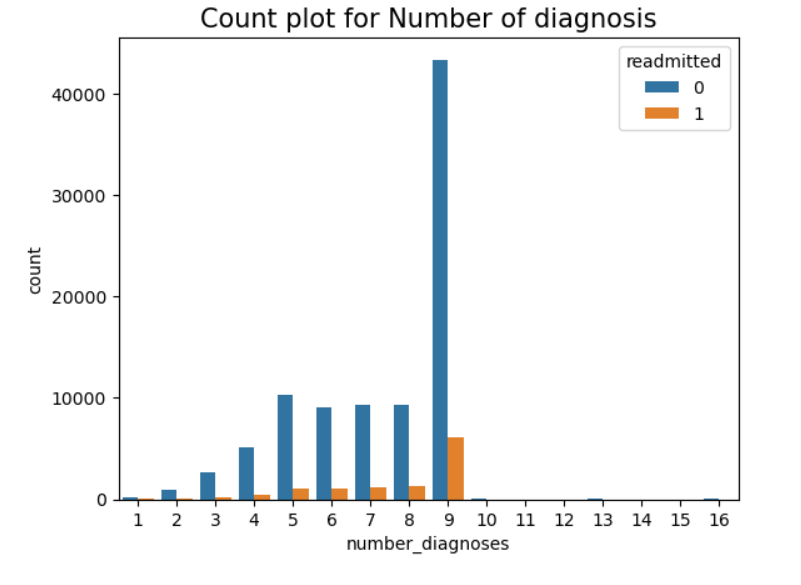
Patients who are getting Transferred or Discharges has a higher chance of readmission than other categories.



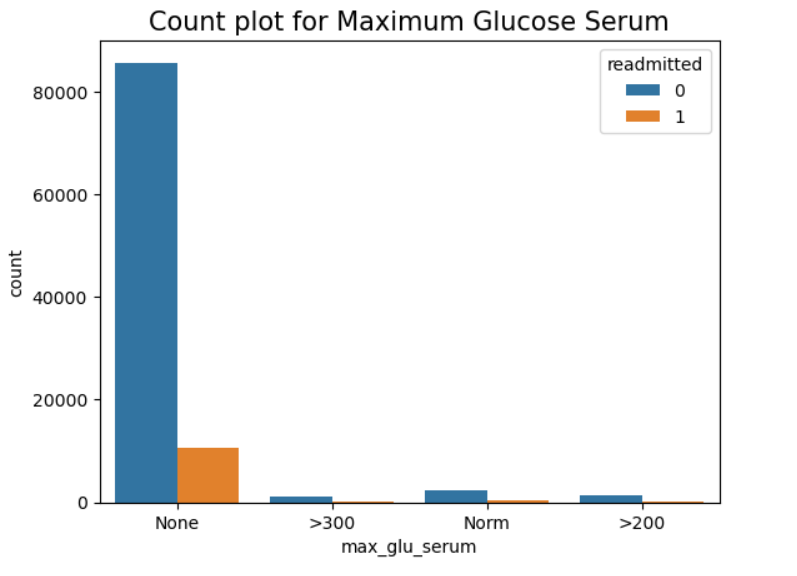
Patients having admission source id as Emergency has a higher chance of readmission than the others.



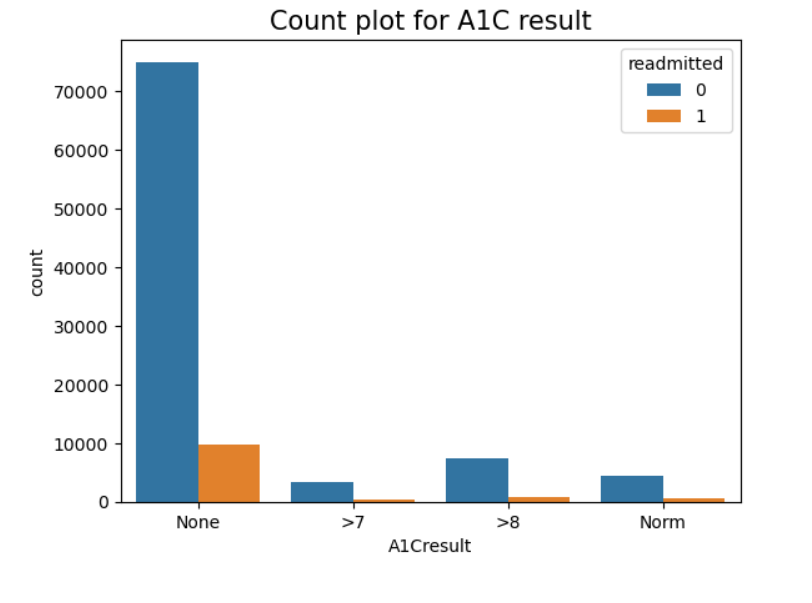
Patients who are diagnosed with 9 diseases have higher chance of readmission.



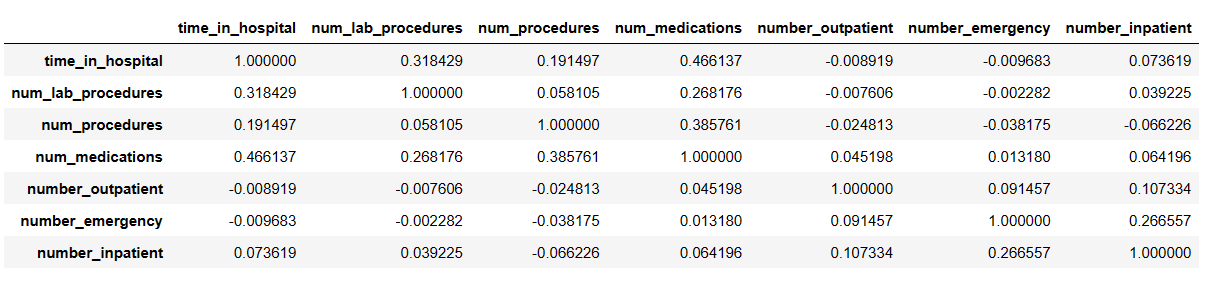
Patients who have not taken Max glucose serum have a higher chance of readmission.

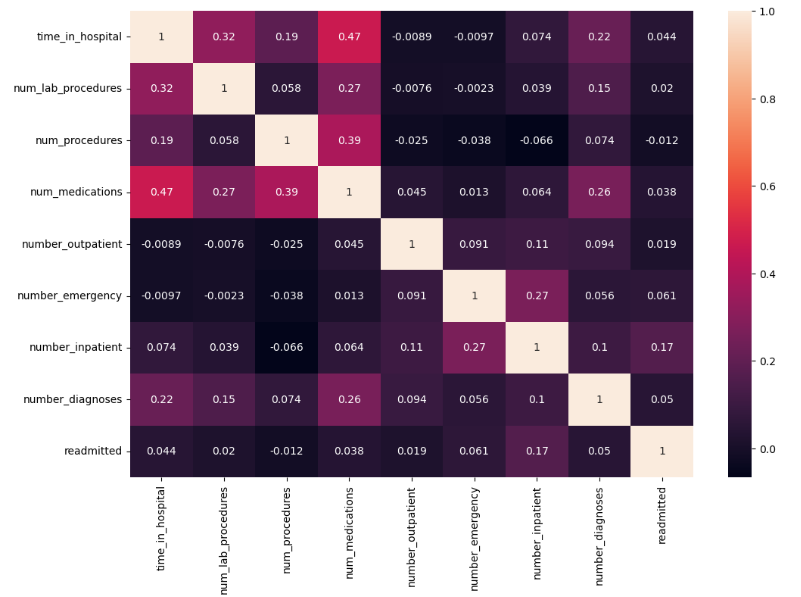


Patients who have not taken A1C Result have a higher chance of readmission.



**3.2 Multivariate Analysis**





The features with comparatively high correlation:

Time\_in\_hospital – num\_medications : 0.47

Time\_in\_hospital – num\_lab\_procedures : 0.32

Num\_procedures – num\_medications : 0.39

**4. Feature Engineering**

**4.1 Encoding**

We have used different encoding techniques since there are different types of variables used.

* Target Encoding = 'diag\_1', 'diag\_2','diag\_3'
* Label Encoding = 'age', 'race', 'weight','admission\_type\_id', 'discharge\_disposition\_id', 'admission\_source\_id','medical\_specialty', 'max\_glu\_serum', 'A1Cresult', 'metformin', 'repaglinide', 'nateglinide', 'chlorpropamide', 'glimepiride', 'glipizide', 'glyburide', 'pioglitazone', 'rosiglitazone', 'acarbose', 'miglitol', 'tolazamide', 'insulin', 'glyburide-metformin'
* Get\_dummies Encoding = 'gender', 'acetohexamide', 'tolbutamide',

'troglitazone','glipizide-metformin', 'glimepiride-pioglitazone',

'metformin-rosiglitazone', 'metformin-pioglitazone', 'change', 'diabetesMed'

**4.2 Scaling**

The rest of the numerical features are scaled as well.

The scaling is done using StandardScaler for the rest of the numerical variables.

Numerical variables which are scaled are:'time\_in\_hospital', 'num\_lab\_procedures','num\_procedures','num\_medications','number\_outpatient', 'number\_emergency','number\_inpatient' and 'number\_diagnoses'

**4.3 Hypothesis Testing**

**5. Model Building and Evaluation**

**5.1 Model Building and Evaluation without SMOTE**

Since the target variable is categorical, our problem is a classification problem.

We tried to build the following models:

1. Logistic Regression
2. KNN
3. Decision Tree
4. Random Forest
5. Gradient Descend
6. ADA Boost

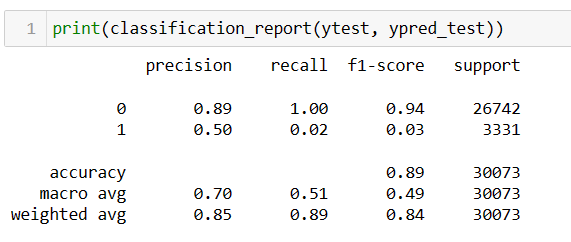
Evaluation metrics are used to measure the performance of difference models.

Different evaluation metrics used are:

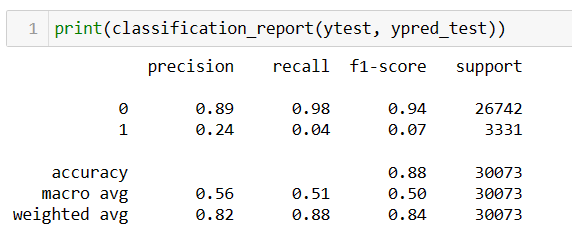
1. Accuracy
2. Recall
3. Precision
4. F1 Score
5. Classification report

Let us look at results of various models

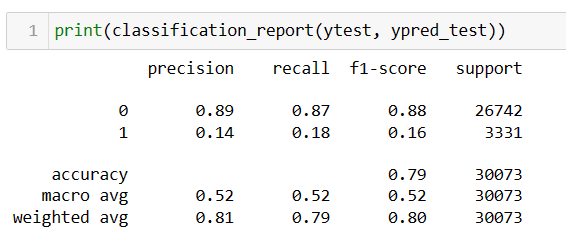
1. **Logistic Regression**

****

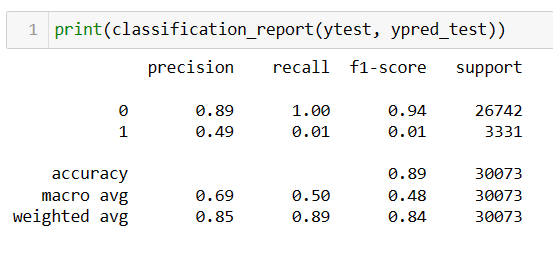
1. **KNN**

****

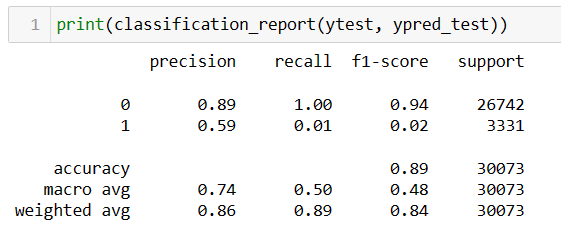
1. **Decision Tree Classifier**



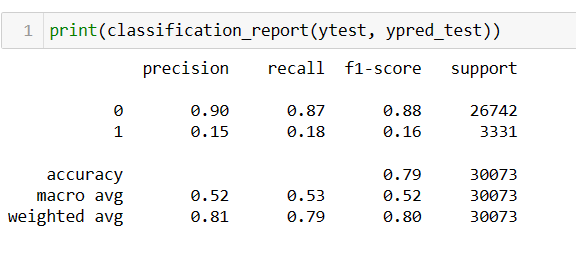
1. **Random Forest Classifier**

****

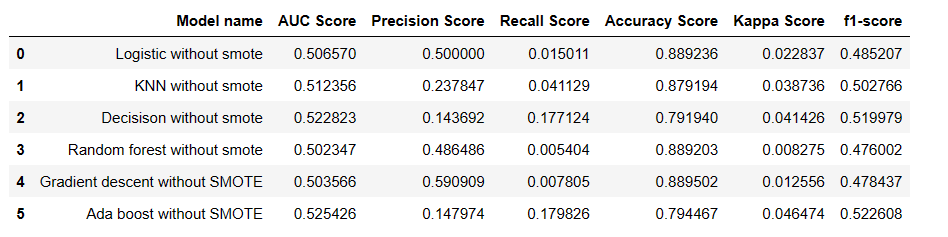
1. **Gradient Descend**

****

1. **ADA Boost**

****

Now let us see the evaluation metrics of all the models in one data frame.



**5.2 Model Building and Evaluation with SMOTE**

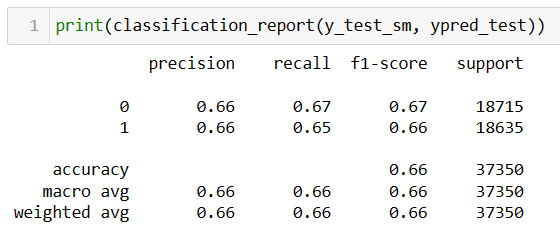
SMOTE is an over sampling technique we apply to deal with imbalanced data.

We tried to build the following models:

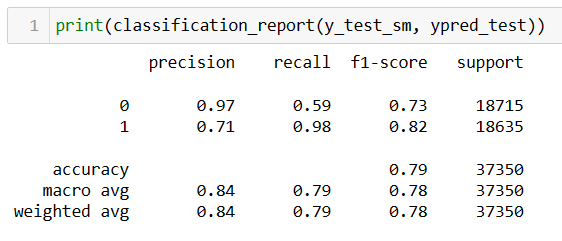
1. Logistic Regression
2. KNN
3. Decision Tree
4. Random Forest
5. Gradient Descend
6. ADA Boost

Results with the SMOTE are as follows:

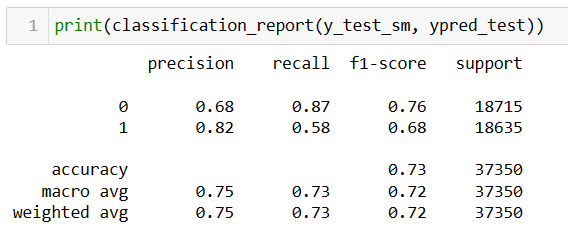
1. **Logistic Regression**

****

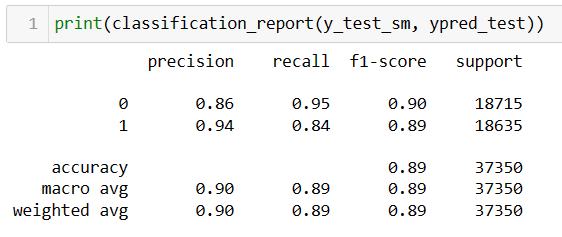
1. **KNN**

****

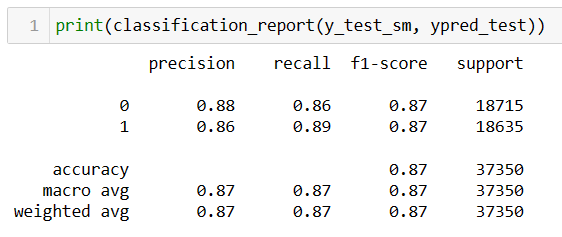
1. **Decision Tree**

****

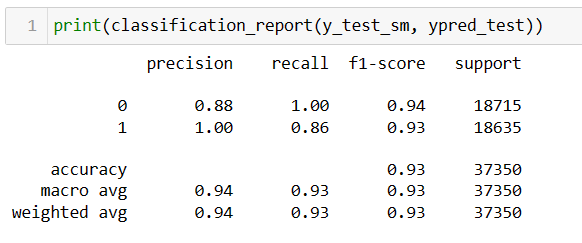
1. **Random Forest**

****

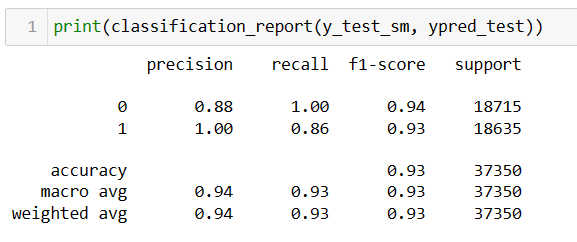
1. **ADA Boost**

****

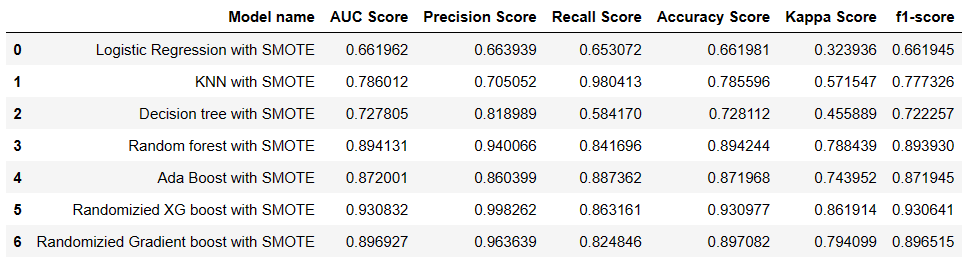
1. **Gradient Boost**

****

1. **XG Boost**

****

Now let us see the evaluation metrics of all the models in one data frame.



From the above results, XG Boost gives the best results with f1 score of 0.93. With out SMOTE, maximum f1 score was 0.52 for ADA boost.