Predict Hospital Readmission rates of Diabetes patients

Bindu Jacob

Capstone project Report

SpringBoard Data Science Career track

1. Introduction

1.1 About Diabetes

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar.

34.2 million people, or 10.5% of the U.S. population, have diabetes. An estimated 26.8 million people - or 10.2% of the population - had diagnosed diabetes. Approximately 7.3 million people have diabetes but have not yet been diagnosed (2018). Diabetes was the seventh leading cause of death in the United States in 2017 based on the 83,564 death certificates in which diabetes was listed as the underlying cause of death.

1.2 Motivation

As a consequence of the increase in diabetes diagnoses, the hospital readmissions are also increasing. A hospital readmission is when a patient who is discharged from the hospital get readmitted within a period of time.

The Hospital Readmissions Reduction Program (HRRP) is a Medicare value-based purchasing program that encourages hospitals to improve communication and care coordination to better engage patients and caregivers in discharge plans and, in turn, reduce avoidable readmissions.

Centers for Medicare & Medicaid Services(CMS) has assigned hospital readmissions as one of the metric for hospital quality and penalizes hospitals for having too many Medicare patients readmitted within 30 days.

1.3 Who benefits?

Major benefits for this readmission predictive model are for hospitals and other health service providers where they can assess the patient health before discharge and take appropriate steps to mitigate readmissions and better the quality of patient care.

Patients also benefit as they receive high-quality care during their hospitalizations and their transition to the outpatient setting will likely have better outcomes, such as survival, functional ability, and quality of life.

Overall reducing the rate of hospital readmissions has great potential to help constrain health care costs and improve the quality of health care.

1.4 Problem Statement

Predict if a diabetes patient will be readmitted within 30 days after the patient was discharged from the hospital, based on the patient's medical history information.

2. Data Wrangling

2.1 Data Overview

The data is sourced from UCI Machine learning Repository, Diabetes 130-US hospitals for years 1999-2008. The raw data had two files:

- diabetic_data.csv - contains patient history information
- IDs_mapping.csv - file containing the descriptions to the IDs used in the main file

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It has 101766 observations and 50 attributes

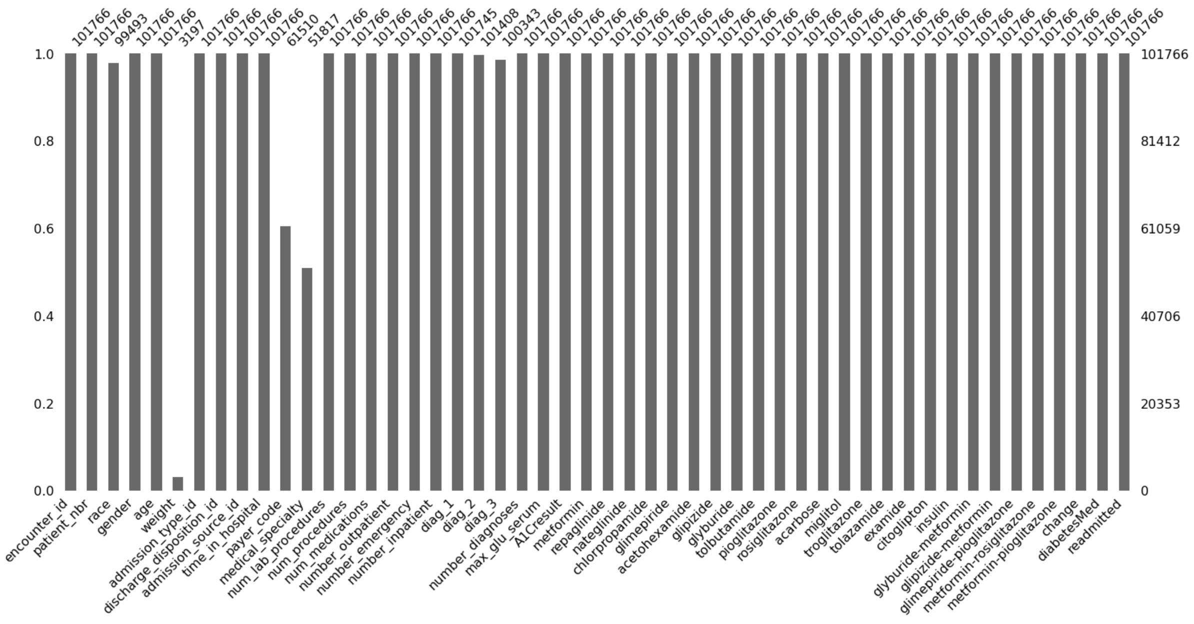
Each observation had the patient's medical history of the hospital encounter, demographic data, diagnoses, procedures and time in hospital etc

2.2 Data Cleaning

2.2.1 Missing values

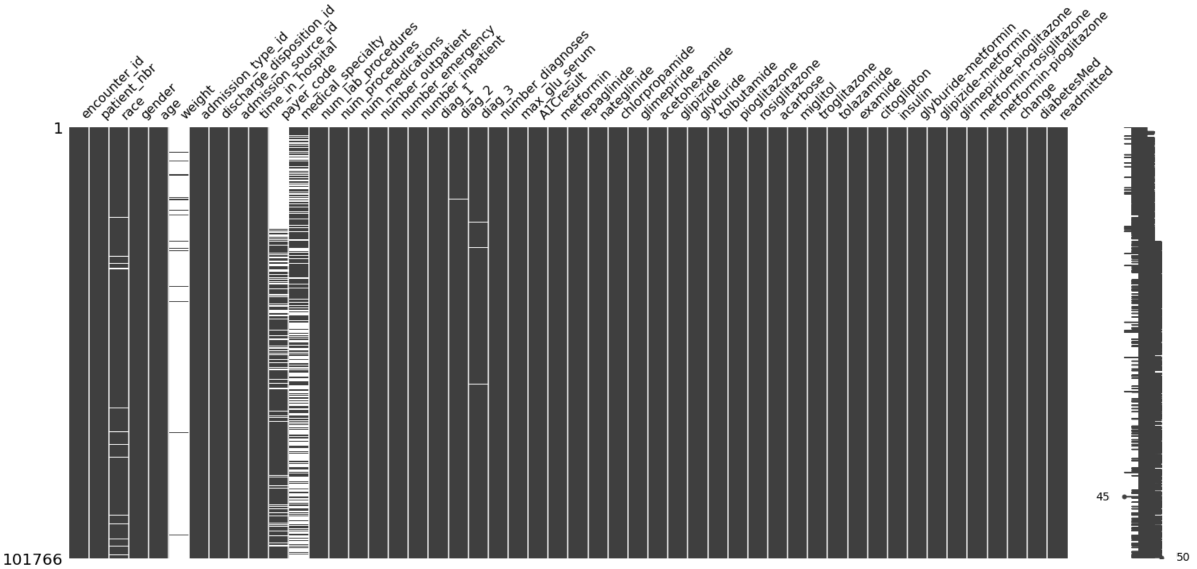
The dataset had many missing values represented as '?' which were stored as NaN values while loading the data to pandas dataframe

Visualization of the nullity of the dataset using missingno is shown below



From the figure, weight has the most missing values about 97% and so this column is removed. payer_code related to the patient's insurance information has also about 40% missing values and this is also removed as it is not relevant to the outcome.

Further investigations of other missing attributes is done using missingno's nullity matrix which shows the missing patterns. Here columns with no missing values are shown as fully black and columns with missing values are shown with white lines.



medical_specialty which is the specialty of the admitting physician is also missing about 50% of the values. Missing values of this column is tagged as 'missing'. race is also missing values and these are also tagged as 'missing'. diag_1, diag_2, diag_3 related to primary, secondary and tertiary diagnoses of the patients are also missing values. These missing values are removed as they are unique to a patient, and does not make sense to fill these with any values.

2.2.2 Duplicate data

There are atleast 20% of the patients with more than one encounter with the hospital. Duplicate records of these patients based on unique patient number are removed and only one record of the patient's first admission to the hospital is retained.

2.2.3 Formatting

admission_id (Id corresponding to the type of patient admission), discharge_disposition_id (Id for discharged details) and admission_source_id (Id for how the patient was referred to the hospital) are three type of Ids that had corresponding descriptions which was mapped from the supporting files to the dataframe so as to fully understand these data for further exploratory analysis.

2.2.4 Invalid values

gender has 'Female', 'Male' and three 'Unknown/Invalid' values. These three values were removed.

3. Exploratory Data Analysis

3.1 Investigation

After cleaning, the dataframe contained 70413 observations and 48 attributes. These attributes need to be explored further to understand more about the data.

3.1.1 Target Variable

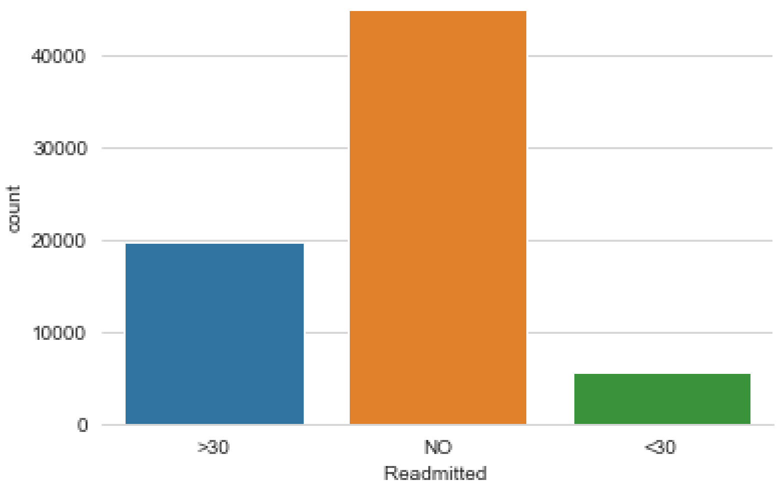
Since the goal of our model will be to predict whether a patient will get readmitted to the hospital. The variable of interest is 'readmitted' and has the following values:

'<30' - if the patient was readmitted in less than 30 days

'>30' - if the patient was readmitted in more than 30 days

'No' - if the patient was not readmitted

This problem wil be tackled as a multiclass classification



```
NO      0.640166
>30     0.280602
<30     0.079233
Name: readmitted, dtype: float64
```

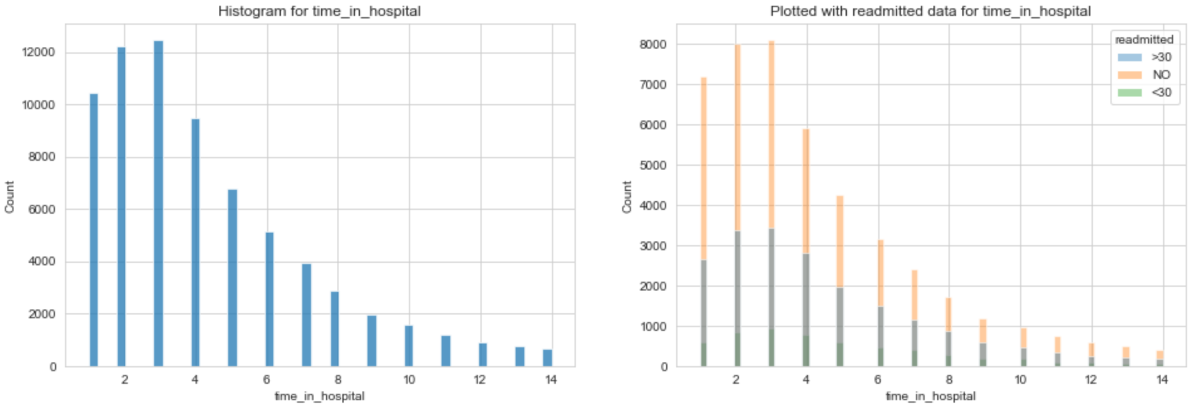
The majority of the data from this readmission distribution shows that people are not readmitted (64%) and about 28% are admitted after more than 30 days. Only around 8% are getting readmitted in less than 30 days.

3.1.2 Numerical Variables

There are two attributes enounter id, which refers to a unique id given to a patient for every visit to the hospital and patient number, which is also unique and refers to a single patient are not plotted. Plotting and analyzing rest of the numerical variables in the data.

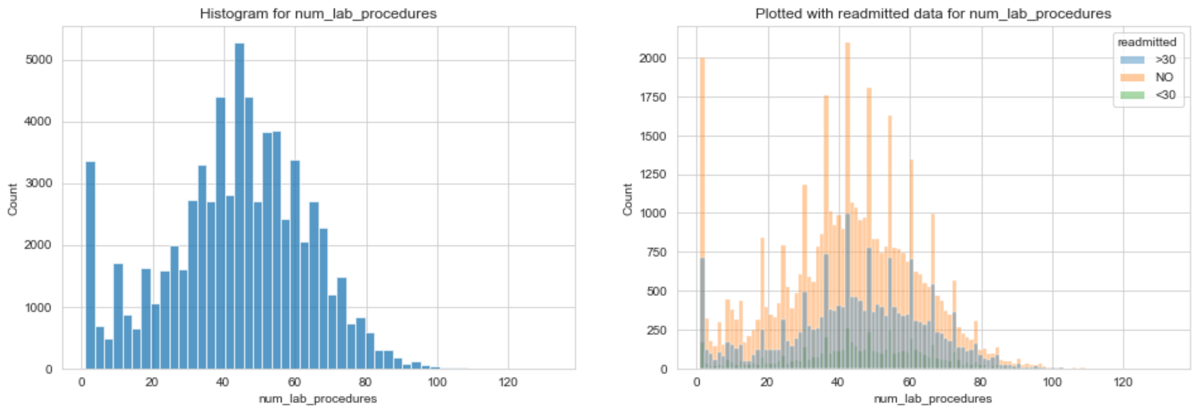
3.1.2.1 Time in Hospital

The chart below shows the time that the patient spent in the hospital from encounter to discharge and the rates of readmission of the patient. The data shows that majority of the patients spents between two to three days in the hospital. And also the distribution shows that higher readmission rates are more for people who stay less time at the hospital.



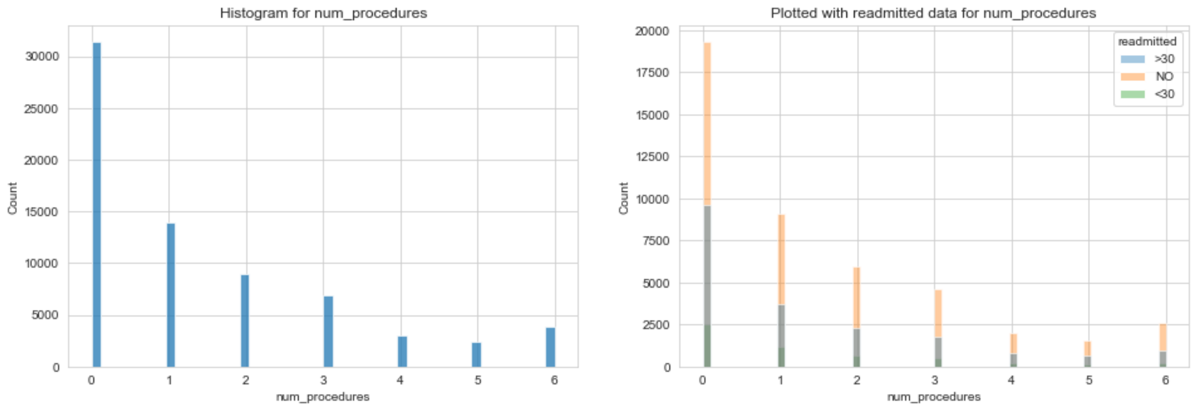
3.1.2.2 Number of lab procedures

The chart shows the number of lab tests performed while the patient was in the hospital and majority of the tests was between 40 to 50 tests. There are some outlier cases where more than 100 tests were performed and some patients had only one test.



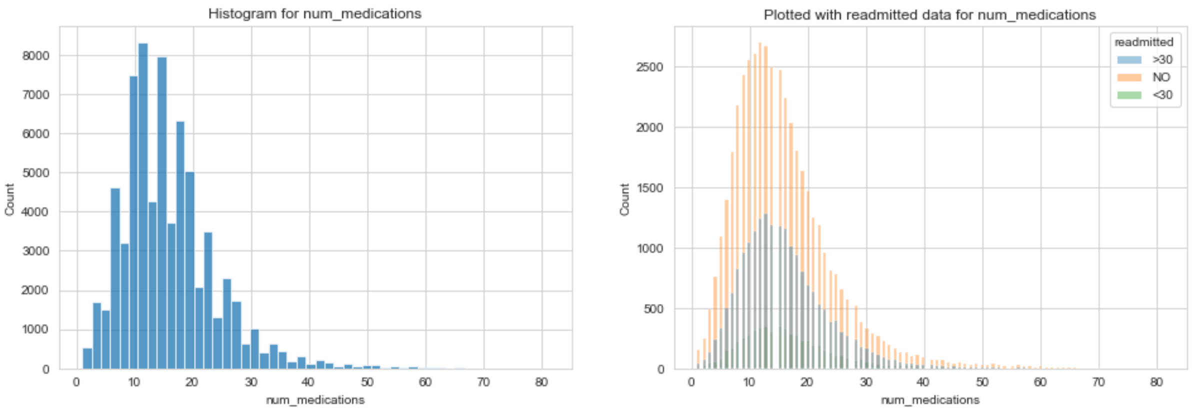
3.1.2.3 Number of procedures

Number of procedures (other than lab) performed while the patient was in the hospital is shown below. The distribution shows that more patients had zero procedure done with few having a maximum of six tests procedures.



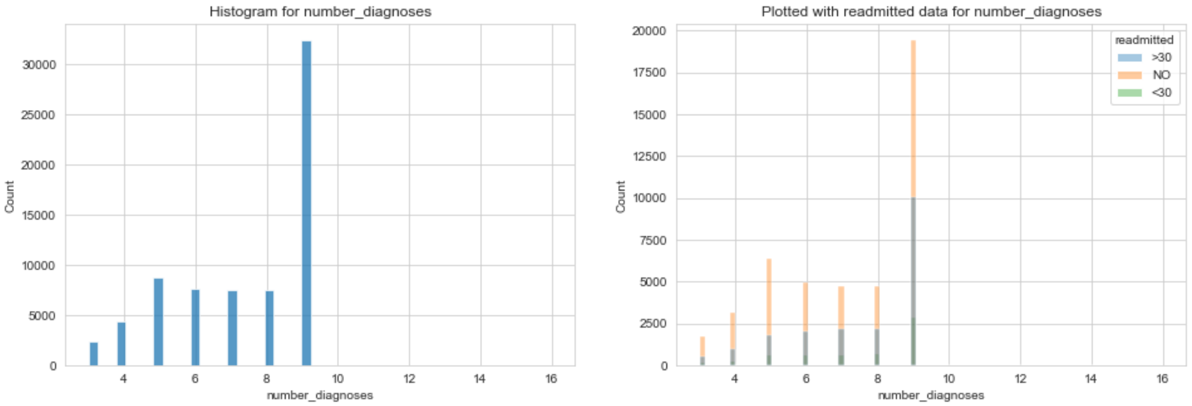
3.1.2.4 Number of medications

The total number of medications that was given to the patient during the hospital encounter is plotted below and most patients are given between 10 to 20 medications. Some outliers here show that more than 50 medications were also given. This distribution shows an equal distributions of medications given for all the classes of readmissions.



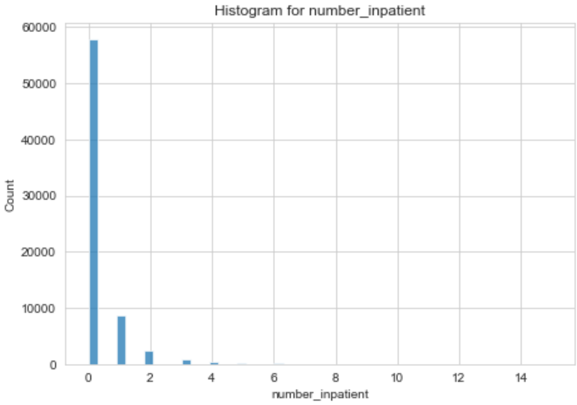
3.1.2.5 Number of diagnoses

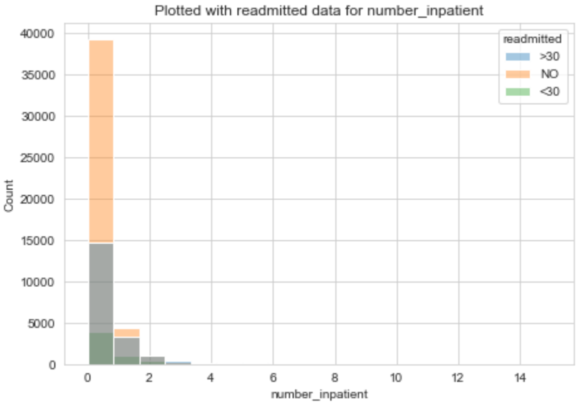
A large number of patients had been diagnosed with atleast 9 ailments. We can clearly see in the distribution that many patients who had less than 30 day readmission had also been diagnosed with atleast 9 illness.

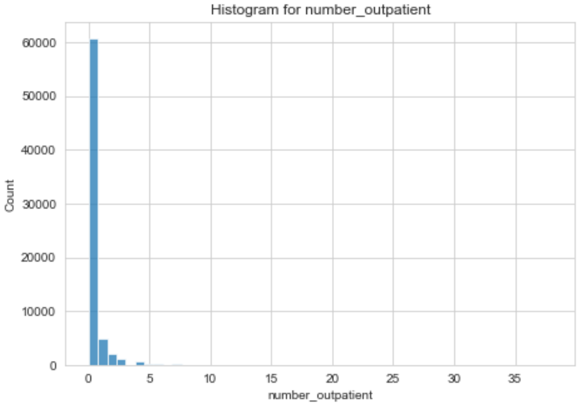


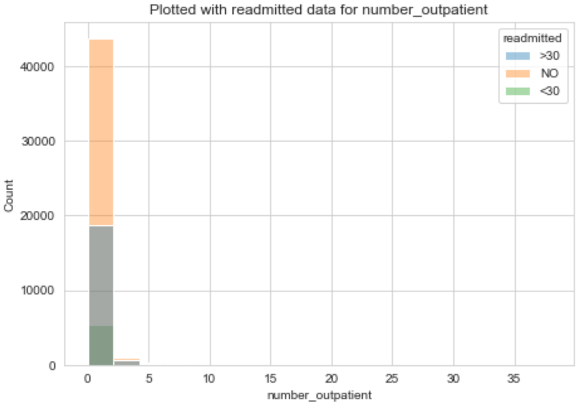
3.1.2.6 Type of prior visits

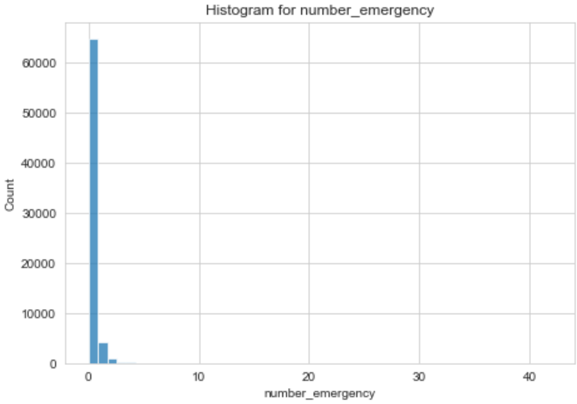
The inpatient, outpatient, and emergency visits that the patients made prior to the hospital encounter is plotted below. Some of the patients had prior inpatient visits to the hospital, majority shows that there was no prior visits. Since we had removed duplicate records (hospital encounter) for a patient, these variables' values may not be accurately captured. The raw data from UCI did not have the dates for the patient hospital encounters and so we cannot assess if these values are from the latest record of a patient.

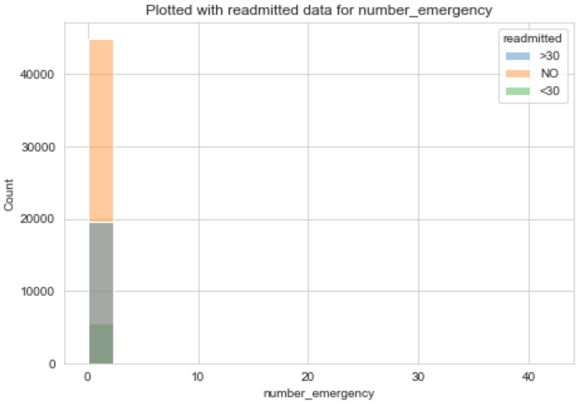

Histogram for number_inpatient


Plotted with readmitted data for number_inpatient


Histogram for number_outpatient


Plotted with readmitted data for number_outpatient


Histogram for number_emergency

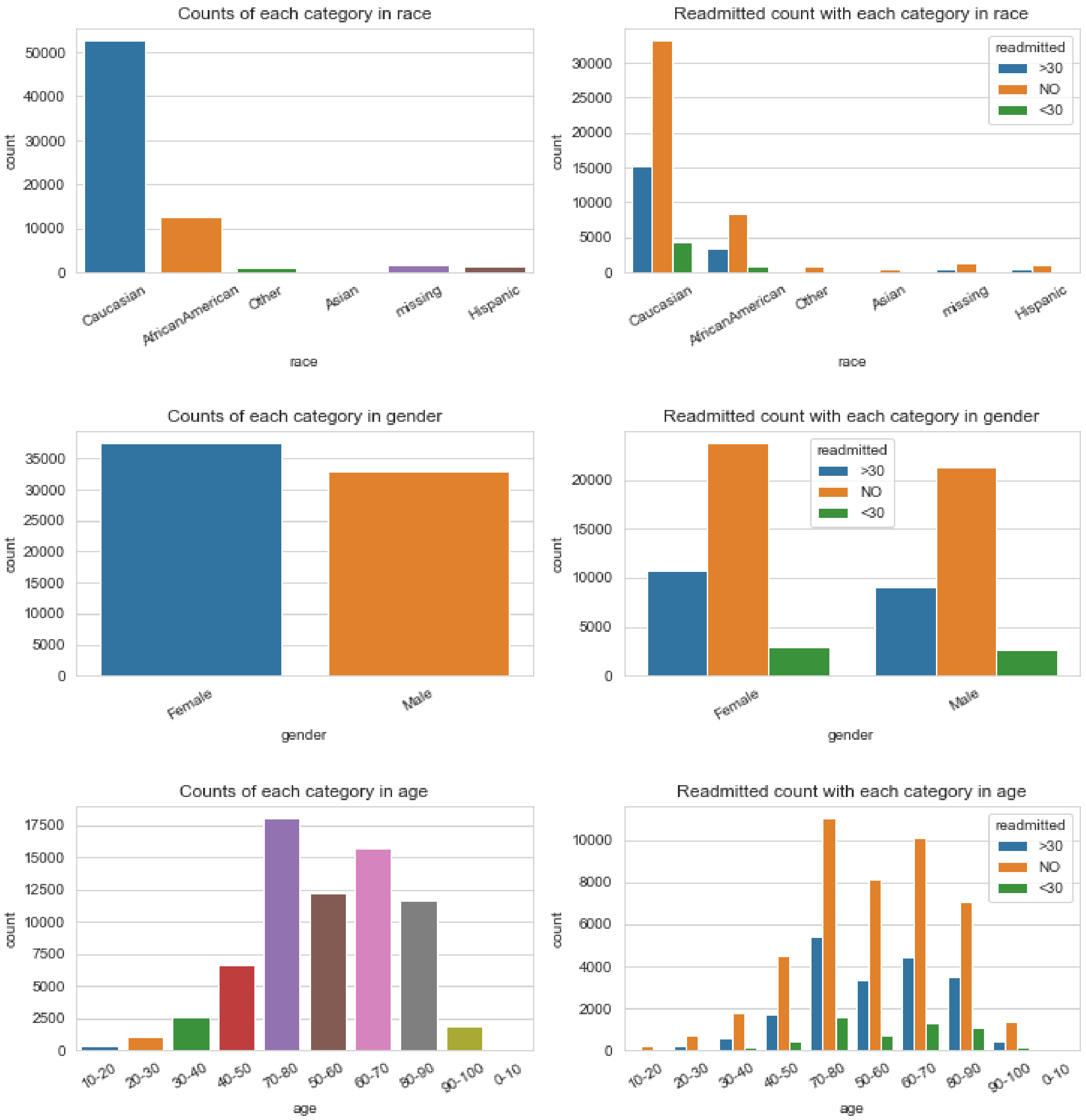

Plotted with readmitted data for number_emergency

3.1.3 Categorical Variables

Most of the attributes are categorical in this dataset. We can plot each sets of attributes to find out more about our data

3.1.3.1 Patient Demographics

Plotting patient's race, gender and age attributes shows us that at least 75% of the patients are caucasian with female slightly outnumbering males. From the readmission distribution we can see that half of the patients are not readmitted. Most of the patients that are hospitalized are of ages between 60 to 80 years of age and readmission rates are also high among these age groups.

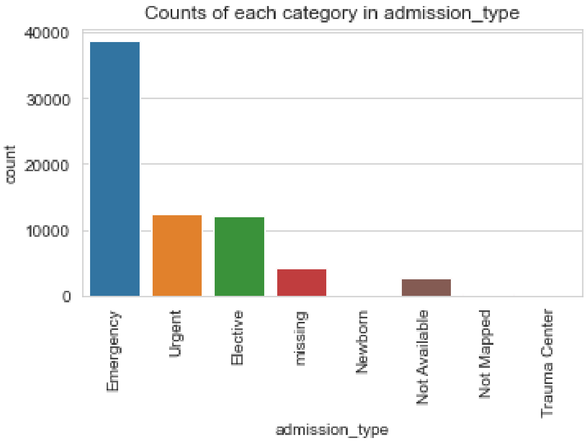


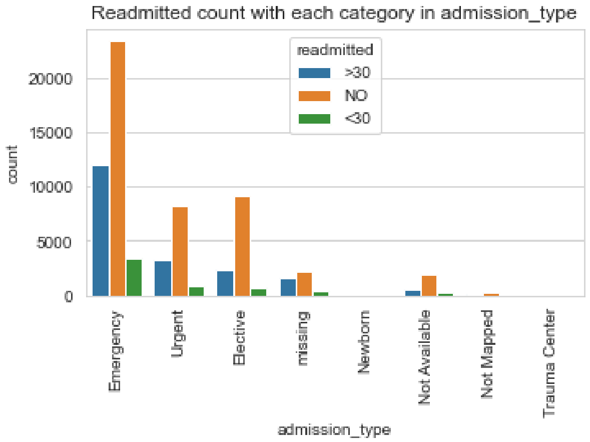
3.1.3.2 Variables admission_type, discharge_disposition, admission_source

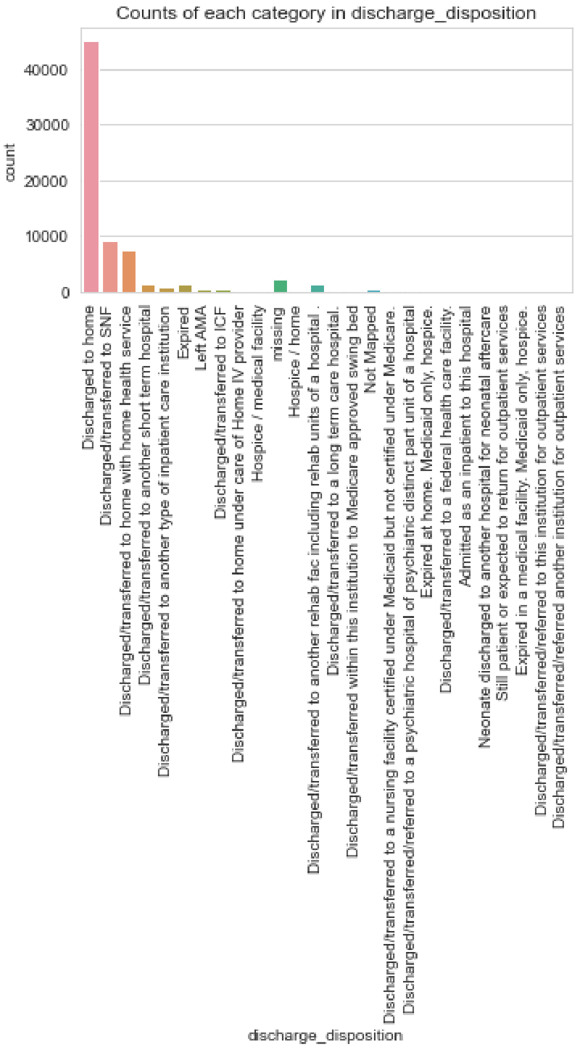
The admission type chart shows us that most of the patients were admitted under emergency and readmission rates also shows that these patients were also readmitted less than 30 days. Newborn, Notmapped and Trauma center has very less values to show up in the chart

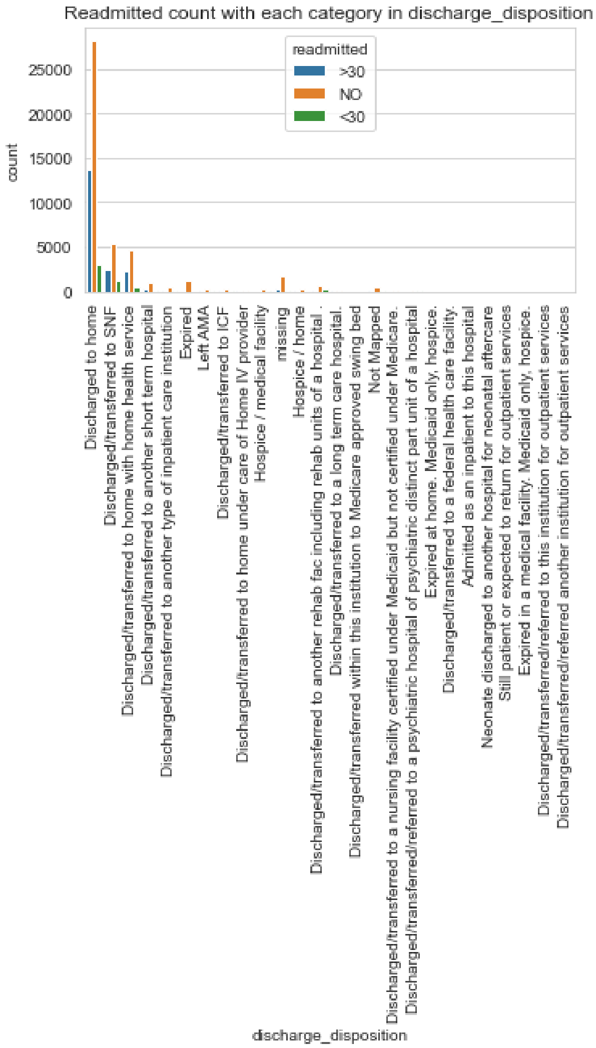
Discharge details shows that most of the patients were discharged to their home and few to a skilled nursing facilty.

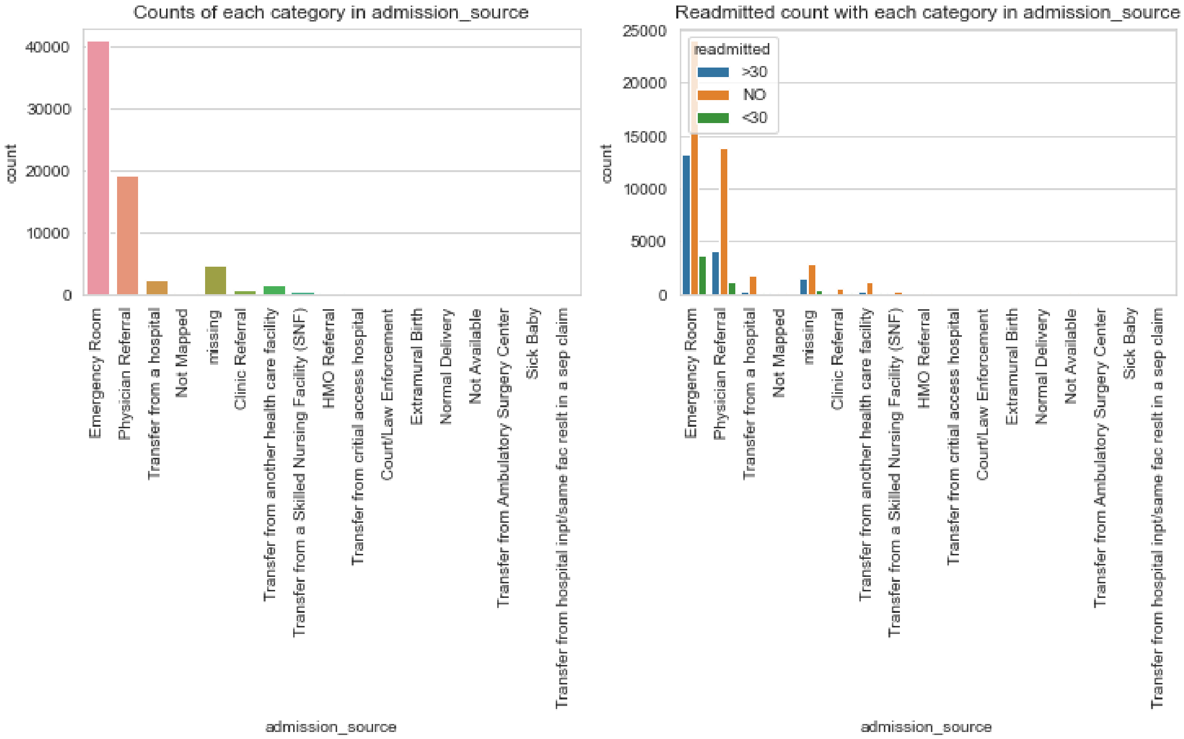
Admission source tells us that majority of the patients were admitted from Emergency room and from Physician referral. Readmission rates for less that 30 days are also higher among these patients.


Counts of each category in admission_type


Readmitted count with each category in admission_type


Counts of each category in discharge_disposition


Readmitted count with each category in discharge_disposition

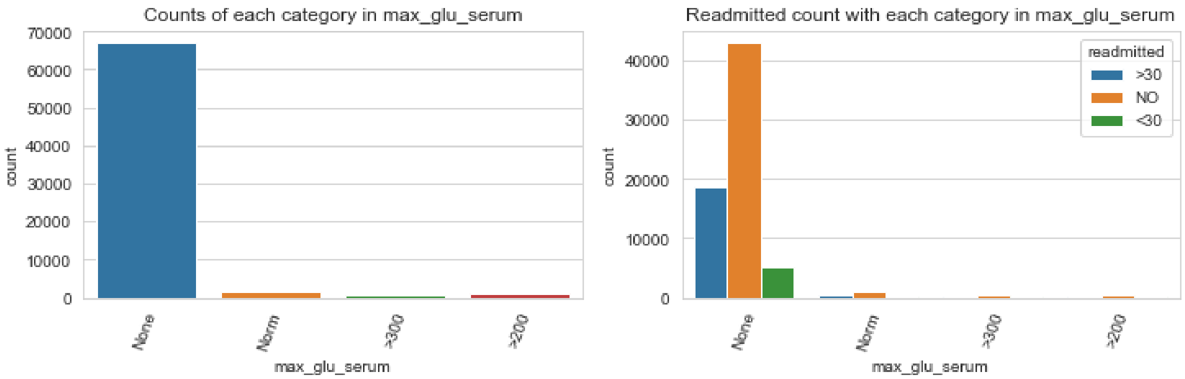


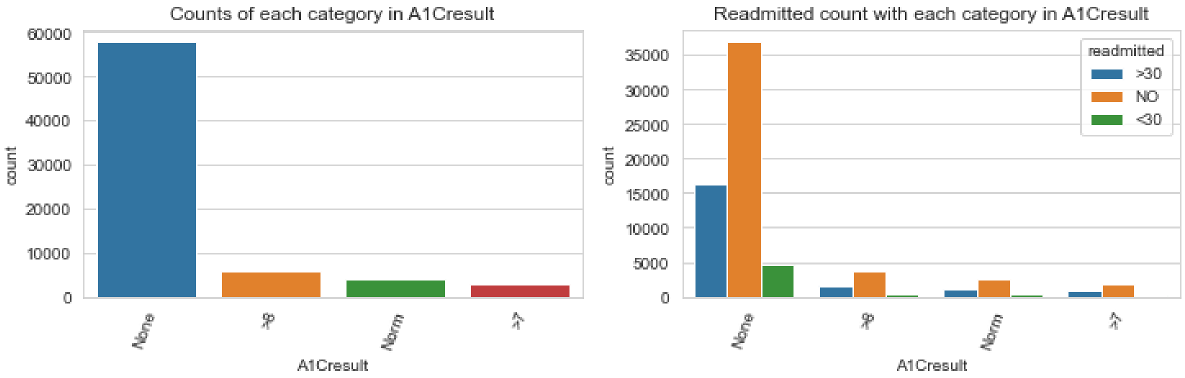
3.1.3.3 Test Results

Two types of tests are used to measure the blood glucose levels and determine if a patient is diabetic or not.

First test is Glucose serum test (max_glu_serum) which is the simplest and most direct single test available to measure blood glucose level in mg/dL. Almost all of the patients were not administered this test (None value). 200 mg/dL or above result is considered diabetic. Only very few patients were given this test.

A1C test is a blood test that provides information about average levels of blood glucose over the past 3 months and is used to diagnose whether the patient has diabetes. Majority of the patients were not administered this test. Any value greater than 7 are considered diabetic.



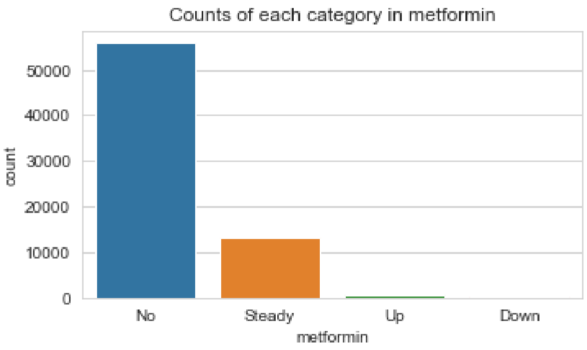


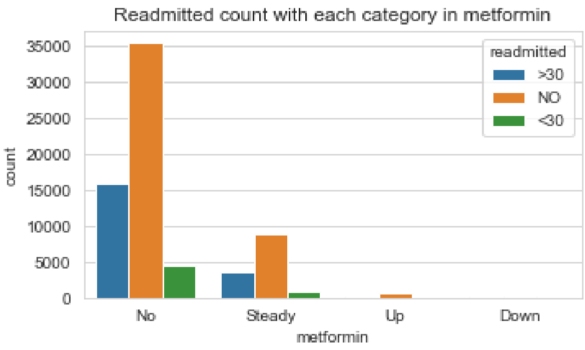
3.1.3.4 Medications Prescribed

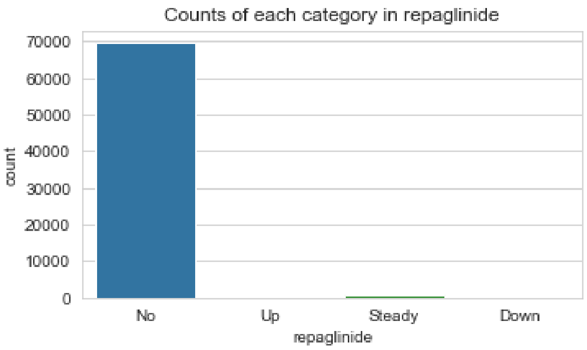
There are 23 different types of medications' information record for each patient. Values 'up' if the dosage for the drug was increased during the encounter, 'down', if the dosage was decreased, 'steady', if the dosage was not changed and 'no', if the drug was not prescribed.

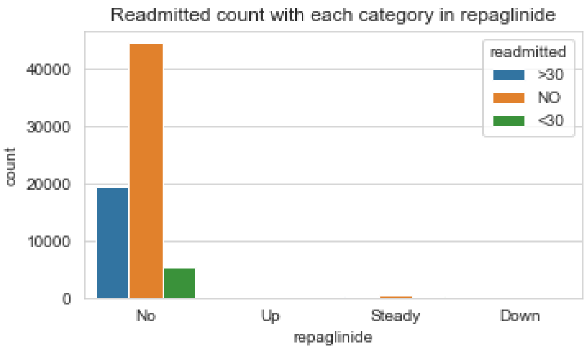
Insulin and metformin were two drugs mainly administered to the patient during the encounter. Here domain knowledge about the types of drugs would be needed to further feature engineer these drug information. For this project we will take into account all these medications.

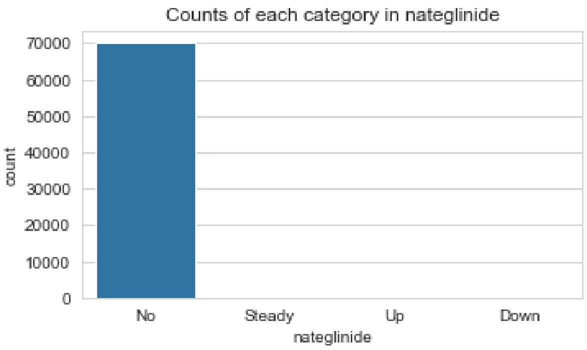
About 75% of the patients were administered some form of diabetes medications during the encounter.

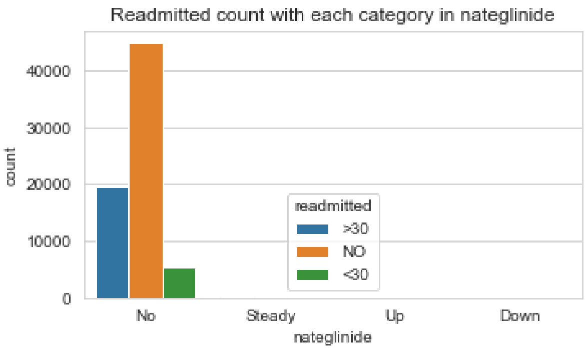

Counts of each category in metformin

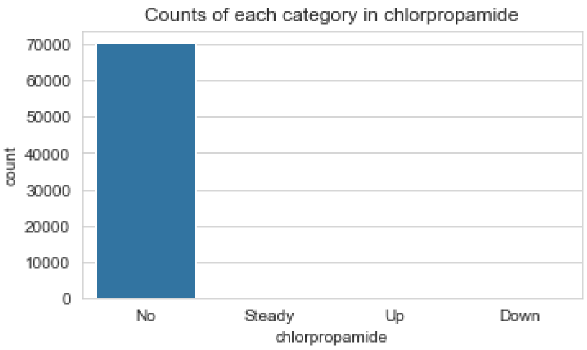

Readmitted count with each category in metformin

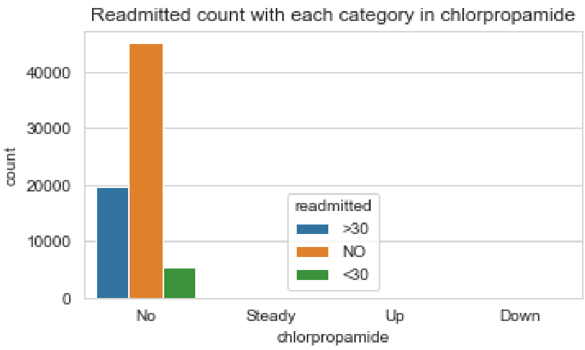

Counts of each category in repaglinide

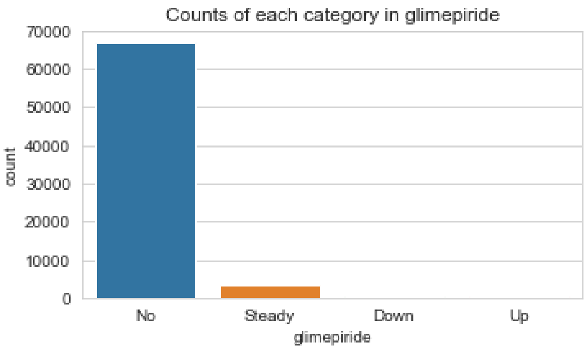

Readmitted count with each category in repaglinide

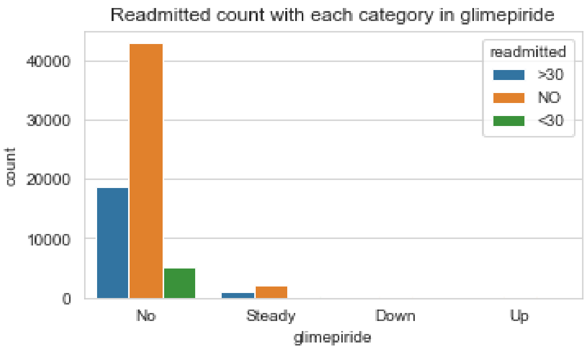

Counts of each category in nateglinide

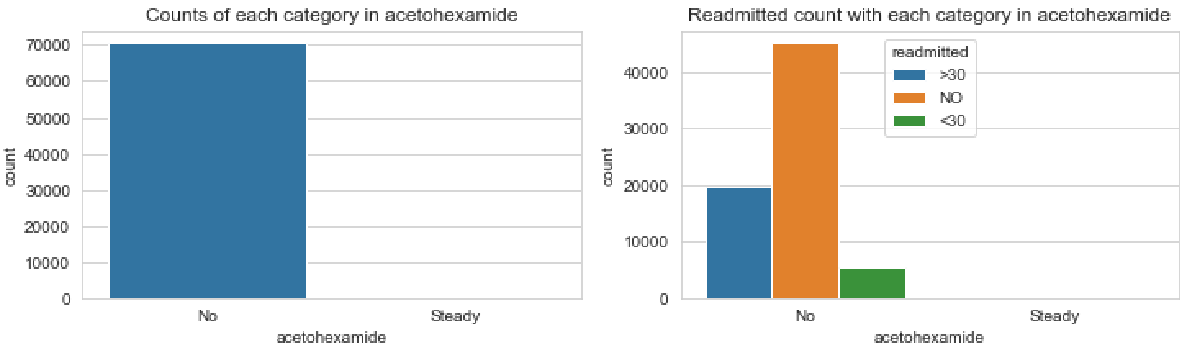

Readmitted count with each category in nateglinide

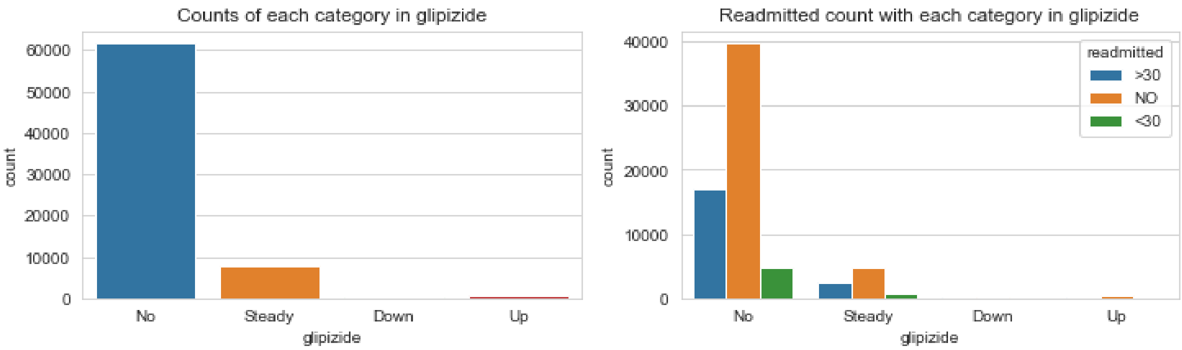

Counts of each category in chlorpropamide

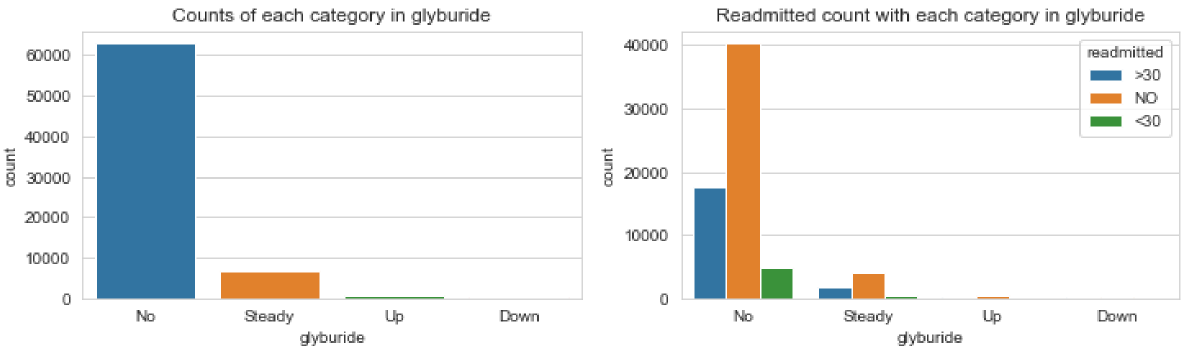

Readmitted count with each category in chlorpropamide

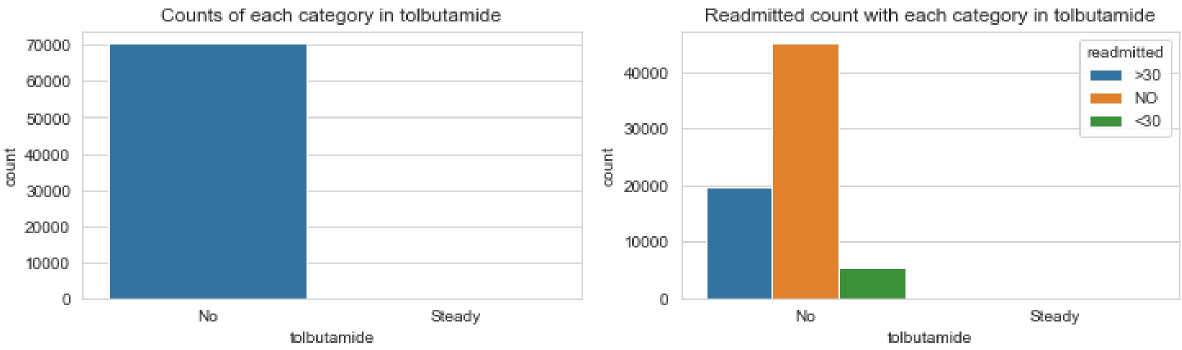

Counts of each category in glimepiride

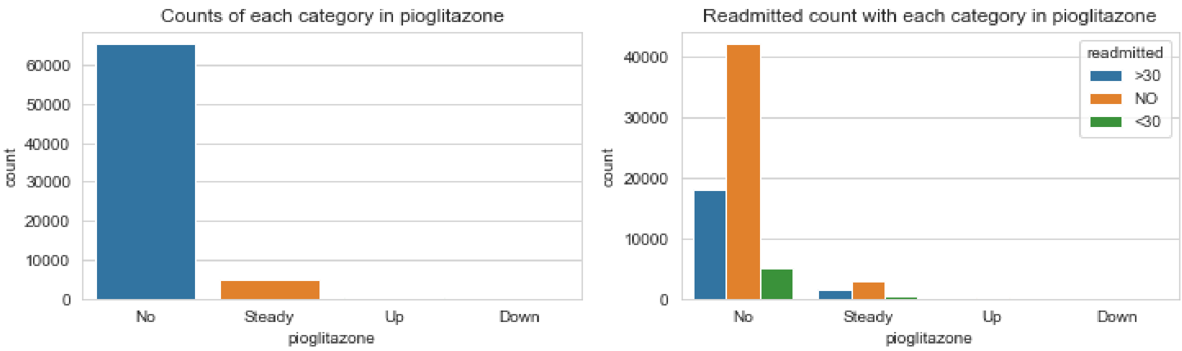

Readmitted count with each category in glimepiride

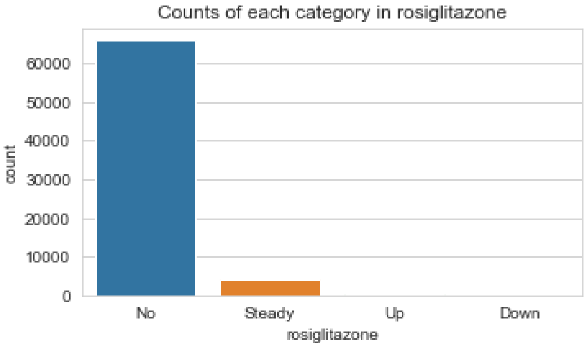

Counts of each category in acetohexamide
count
0
10000
20000
30000
40000
50000
60000
70000
No
Steady
acetohexamide
Readmitted count with each category in acetohexamide
count
0
10000
20000
30000
40000
readmitted
>30
NO
<30
No
Steady
acetohexamide

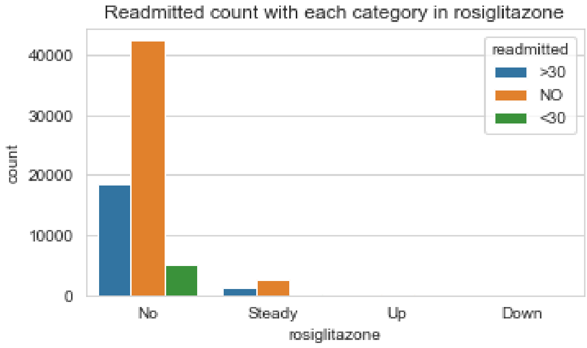

Counts of each category in glipizide
count
0
10000
20000
30000
40000
50000
60000
No
Steady
Down
Up
glipizide
Readmitted count with each category in glipizide
count
0
10000
20000
30000
40000
readmitted
>30
NO
<30
No
Steady
Down
Up
glipizide

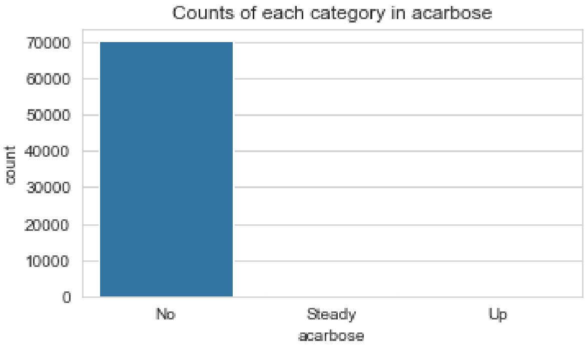

Counts of each category in glyburide
count
0
10000
20000
30000
40000
50000
60000
No
Steady
Up
Down
glyburide
Readmitted count with each category in glyburide
count
0
10000
20000
30000
40000
readmitted
>30
NO
<30
No
Steady
Up
Down
glyburide

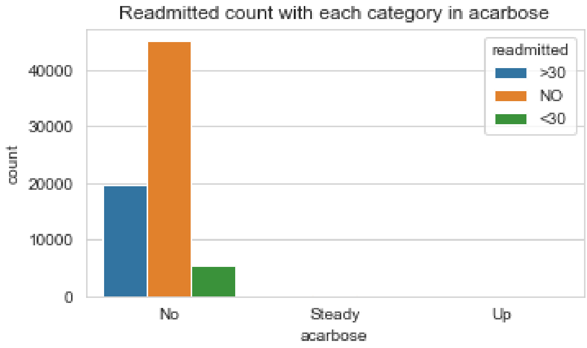

Counts of each category in tolbutamide
count
0
10000
20000
30000
40000
50000
60000
70000
No
Steady
tolbutamide
Readmitted count with each category in tolbutamide
count
0
10000
20000
30000
40000
readmitted
>30
NO
<30
No
Steady
tolbutamide

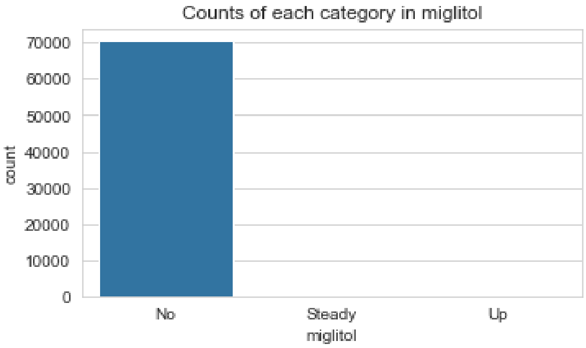

Counts of each category in pioglitazone
count
0
10000
20000
30000
40000
50000
60000
No
Steady
Up
Down
pioglitazone
Readmitted count with each category in pioglitazone
count
0
10000
20000
30000
40000
readmitted
>30
NO
<30
No
Steady
Up
Down
pioglitazone

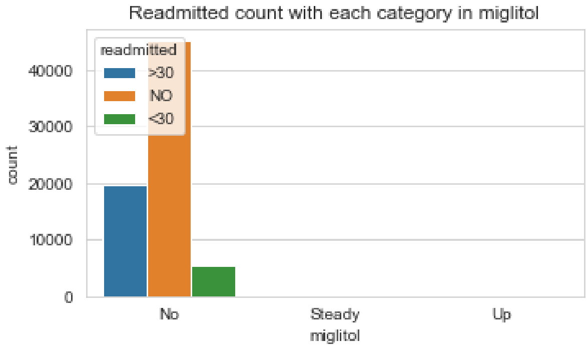

Counts of each category in rosiglitazone

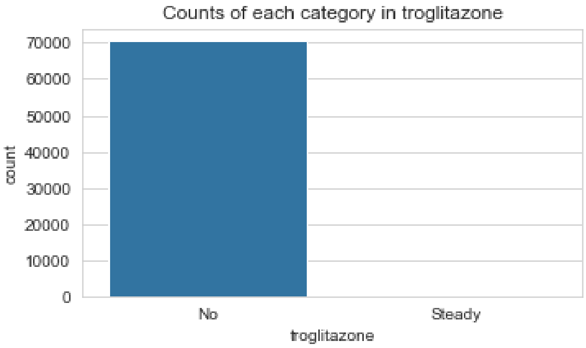

Readmitted count with each category in rosiglitazone

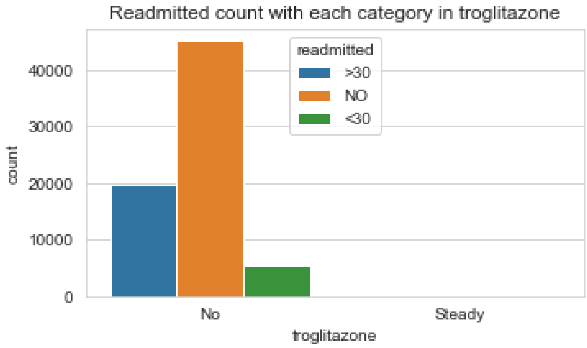

Counts of each category in acarbose

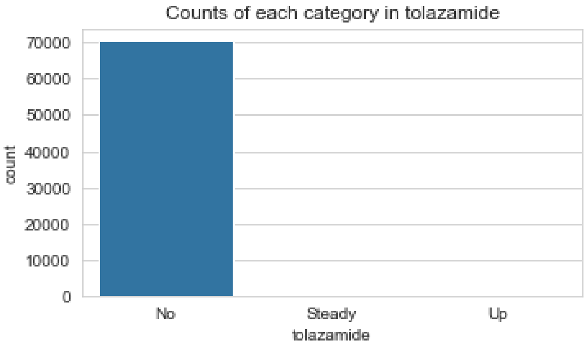

Readmitted count with each category in acarbose

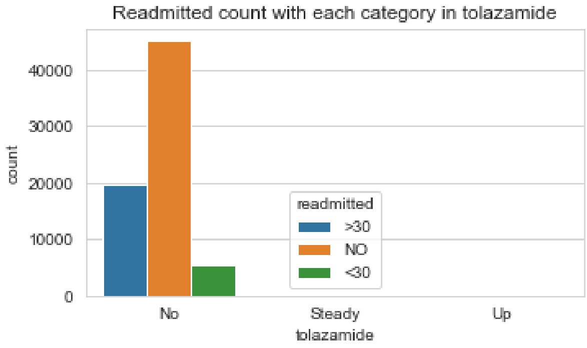

Counts of each category in miglitol

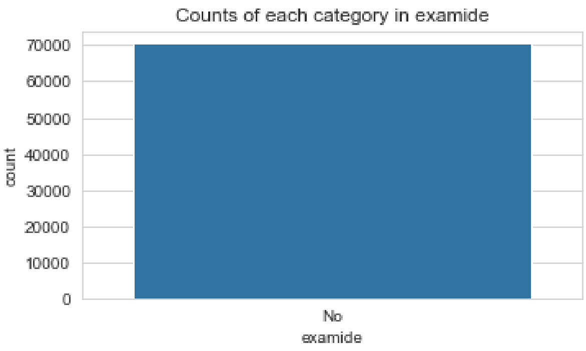

Readmitted count with each category in miglitol

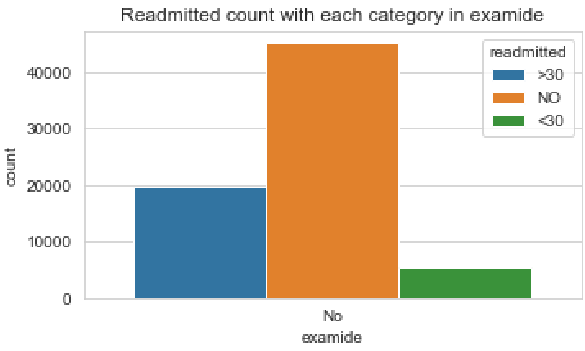

Counts of each category in troglitazone

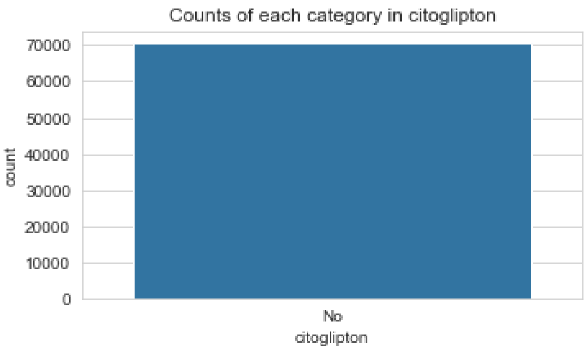

Readmitted count with each category in troglitazone

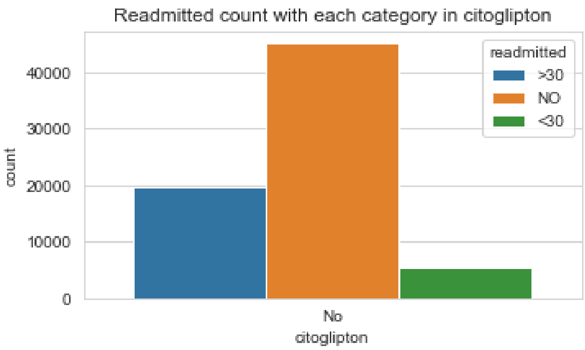

Counts of each category in tolazamide

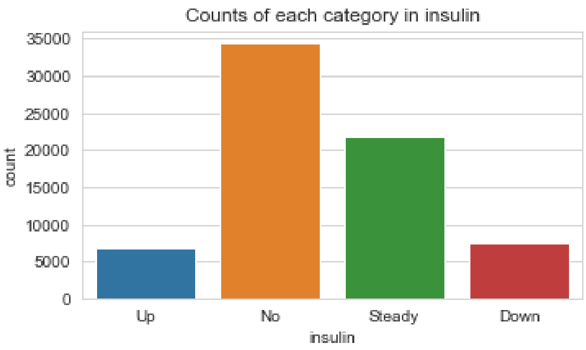

Readmitted count with each category in tolazamide

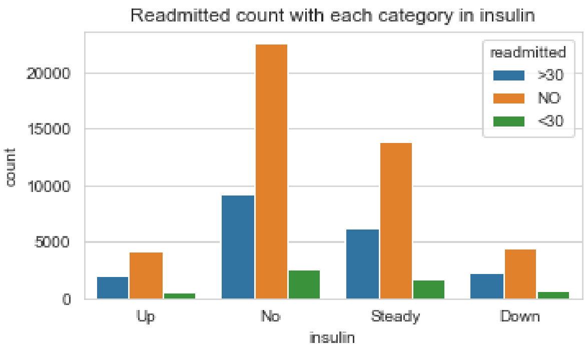

Counts of each category in examide
count
70000
60000
50000
40000
30000
20000
10000
0
No
examide

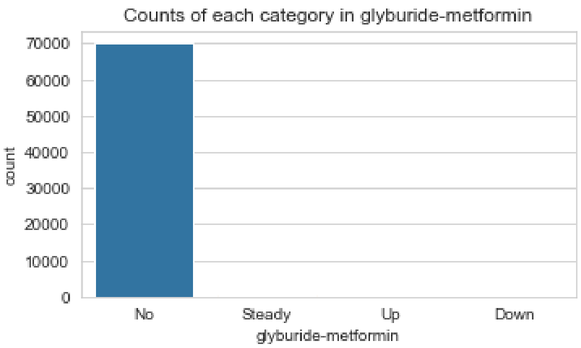

Readmitted count with each category in examide
count
40000
30000
20000
10000
0
readmitted
>30
NO
<30
No
examide

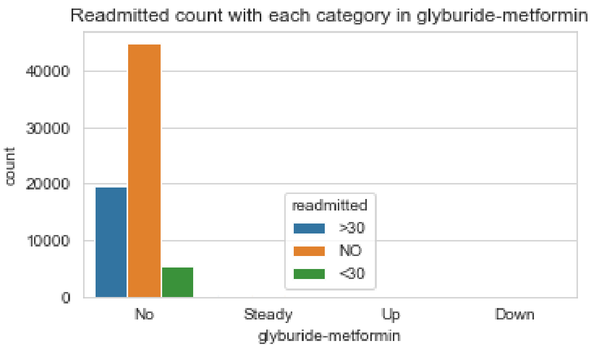

Counts of each category in citoglipton
count
70000
60000
50000
40000
30000
20000
10000
0
No
citoglipton

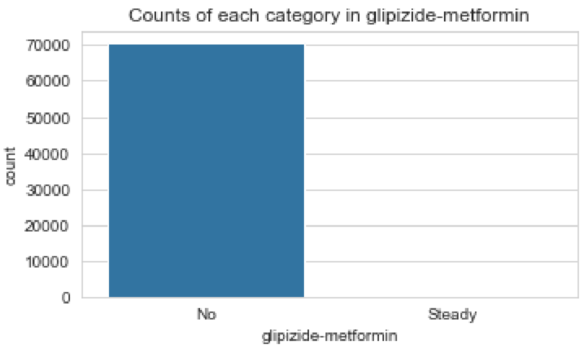

Readmitted count with each category in citoglipton
count
40000
30000
20000
10000
0
readmitted
>30
NO
<30
No
citoglipton

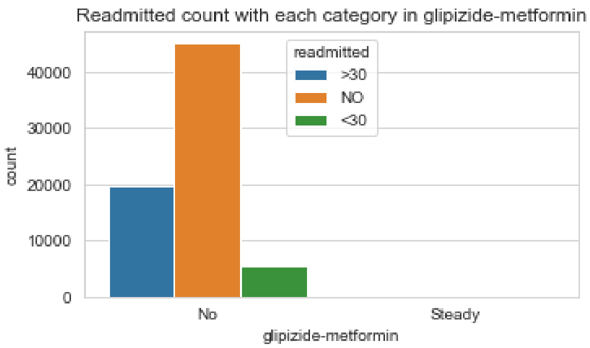

Counts of each category in insulin
count
35000
30000
25000
20000
15000
10000
5000
0
Up
No
Steady
Down
insulin


Readmitted count with each category in insulin
count
20000
15000
10000
5000
0
readmitted
>30
NO
<30
Up
No
Steady
Down
insulin

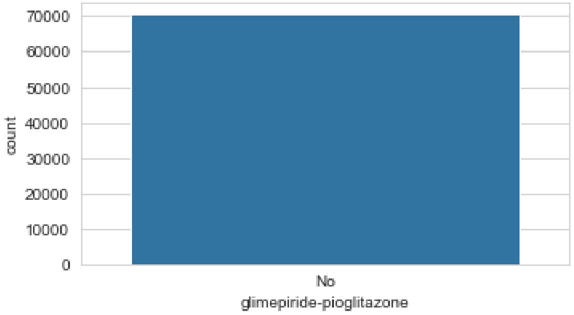

Counts of each category in glyburide-metformin
count
70000
60000
50000
40000
30000
20000
10000
0
No
Steady
Up
Down
glyburide-metformin


Readmitted count with each category in glyburide-metformin
count
40000
30000
20000
10000
0
readmitted
>30
NO
<30
No
Steady
Up
Down
glyburide-metformin

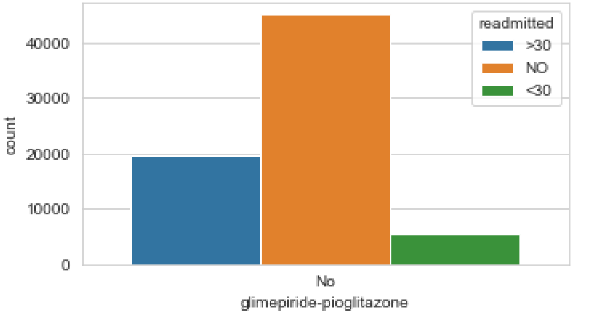

Counts of each category in glipizide-metformin
count
70000
60000
50000
40000
30000
20000
10000
0
No
Steady
glipizide-metformin


Readmitted count with each category in glipizide-metformin
count
40000
30000
20000
10000
0
readmitted
>30
NO
<30
No
Steady
glipizide-metformin

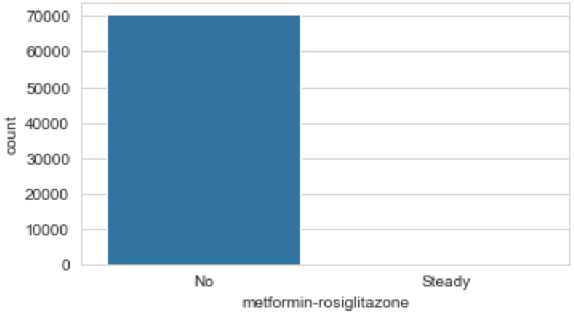
Counts of each category in glimepiride-pioglitazone



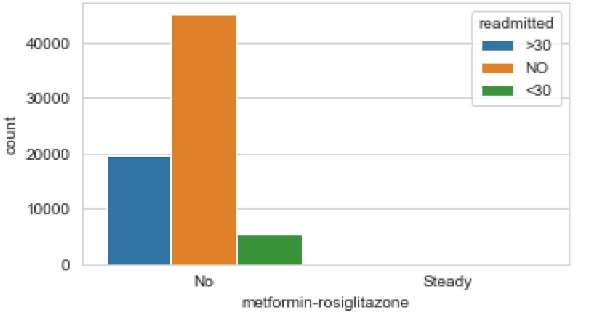
Readmitted count with each category in glimepiride-pioglitazone



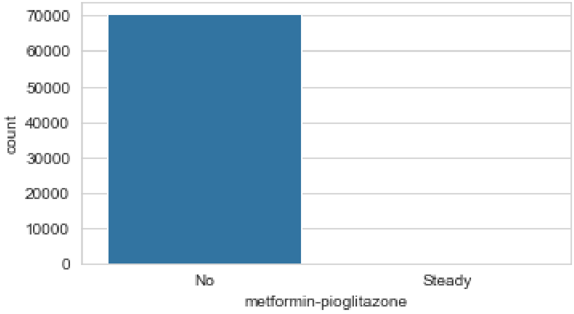
Counts of each category in metformin-rosiglitazone



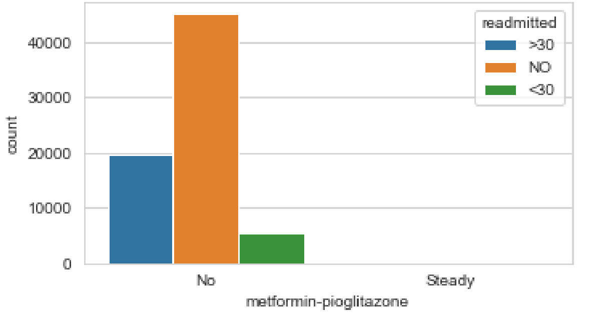
Readmitted count with each category in metformin-rosiglitazone



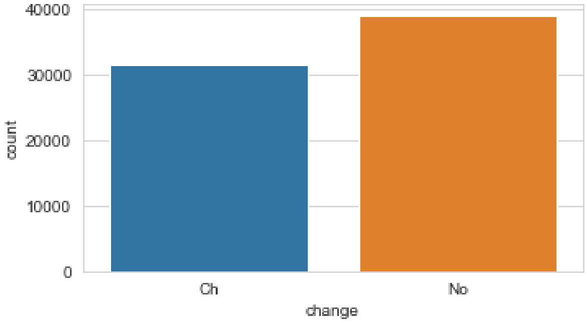
Counts of each category in metformin-pioglitazone



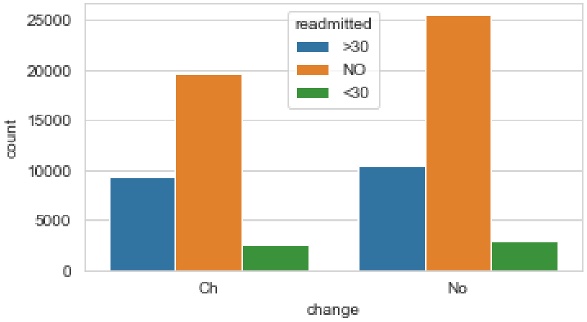
Readmitted count with each category in metformin-pioglitazone



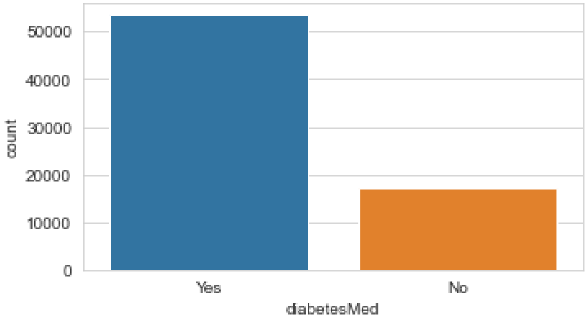
Counts of each category in change



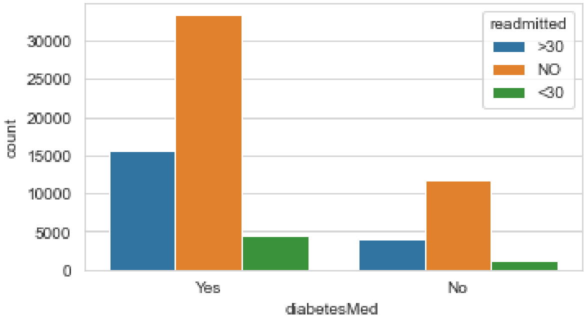
Readmitted count with each category in change



Counts of each category in diabetesMed



Readmitted count with each category in diabetesMed



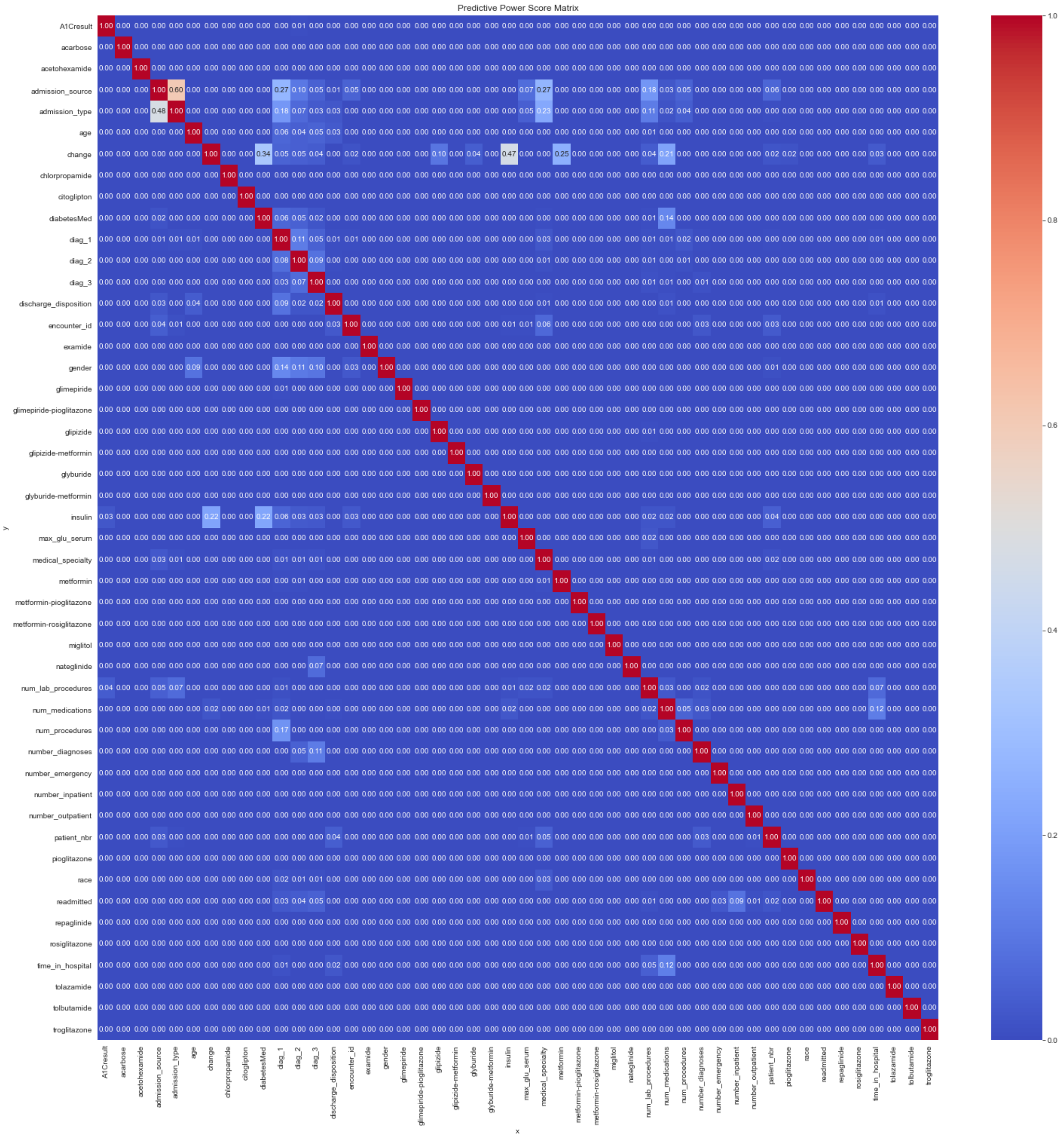
3.1.3.5 High Cardinality variables

High cardinality variables like : medical specialty which is the specialty of the admitting physician, has about 70 values and these needs to be feature engineered

diag_1 (refers to primary diagnose), diag_2 (refers to secondary diagnose) and diag_3 (refers to tertiary diagnose) has over 700 values and these also needs to be feature engineered using domain knowlege. These attributes contain the ICD9-codes for each type of diagnoses.

3.2 Predictive power score matrix

A Predictive power score matrix is used here to analyze the correlation between the variables and assess if there are any good predictors for the target variable and also to eliminate attributes that just add random noise or also to eliminate attributes that can be predicted by other attributes because they don’t add new information.



The following can be analyzed from the predictive power score matrix:

- we can see that the admission_source and admission_type are correlated, so we will only consider admission_type and drop admission_source from our model.
- change and insulin are correlated, we will keep insulin as this is the most common drug given to patients during the encounter

3.3 Feature Selection and Engineering

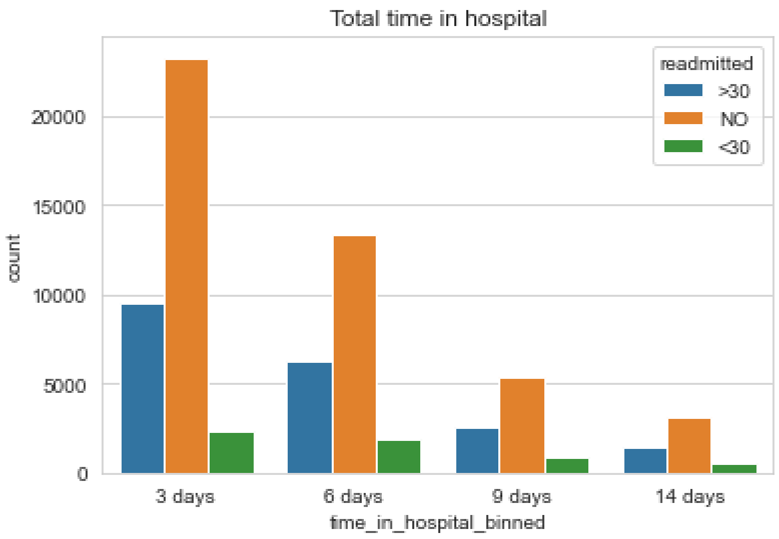
The following attributes are further selected and engineered to reduce their cardinality.

3.3.1 Grouping/Binning Numerical variables

Adaptive binning, where we use the data distribution to transform numerical variables into categorical variables taking also the target classes into account. Here the data distribution itself decide our bin ranges.

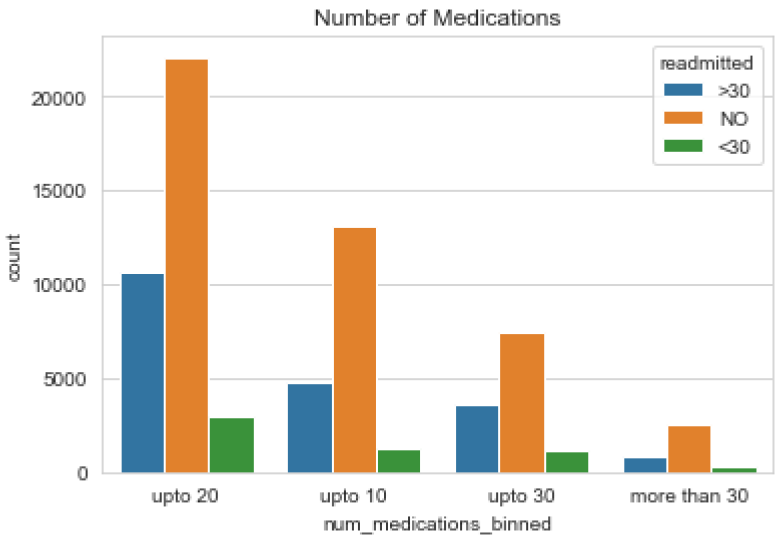
3.3.1.1 Time in Hospital

Based on the histogram analysis of this variable, we could group this into 3 days, 6 days, 9 days and 14 days time in hospital to account for skewed data



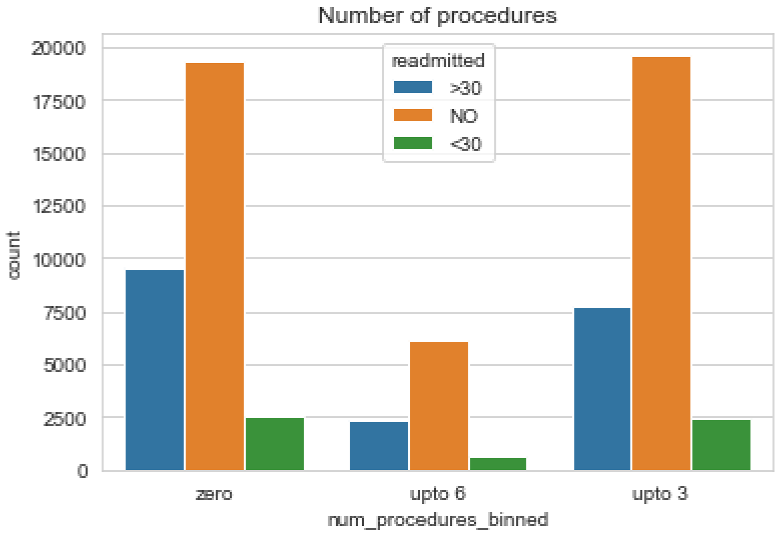
3.3.1.2 Number of medications

Since the original distribution is skewed to the right we can categorize the groups as taking up to 10, 20, 30 and from 30 upto the max number of medications. The new chart is showed below.



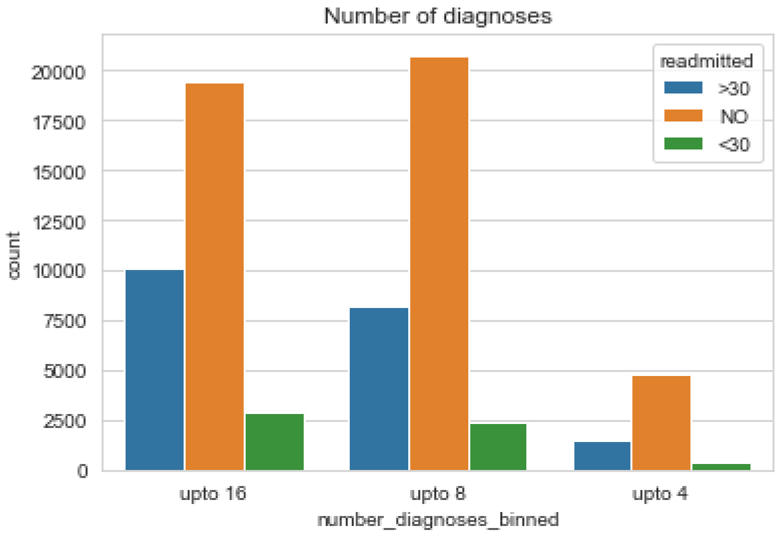
3.3.1.3 Number of procedures

The original chart shows that patients are given 0 to 6 maximum number of procedures. This can be grouped as 0, upto 3 and upto 6 procedures after analyzing the chart to capture the data distribution



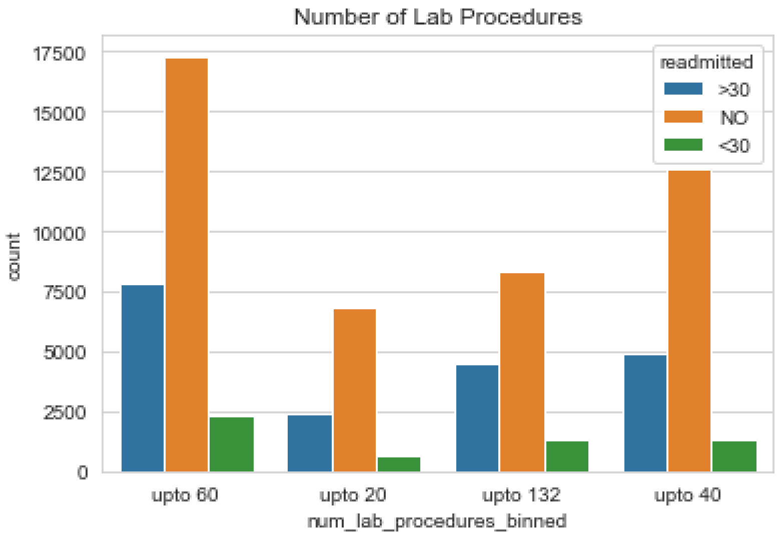
3.3.1.4 Number of diagnoses

Since the majority of the patients got 9 diagnoses and the distribution shows a minimum 3 disgnoses and maximum of up to 16 diagnoses, the bins are split as upto 4, upto 8 and upto 16 diagnoses.



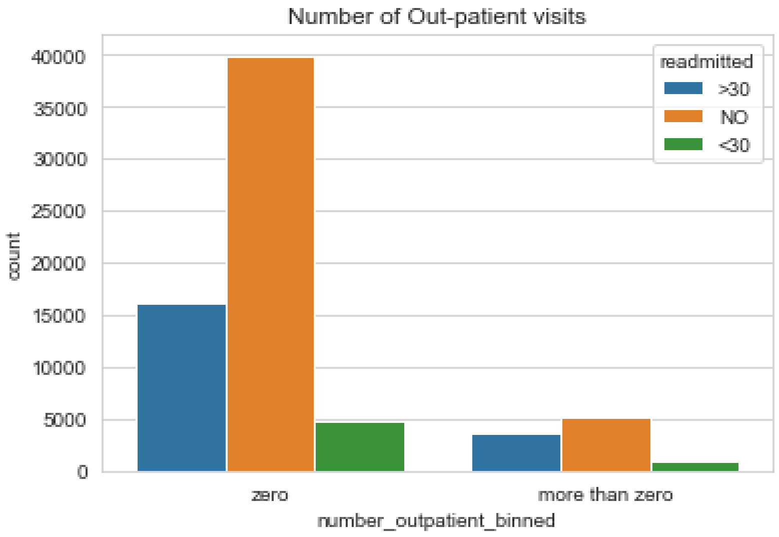
3.3.1.5 Number of lab procedures

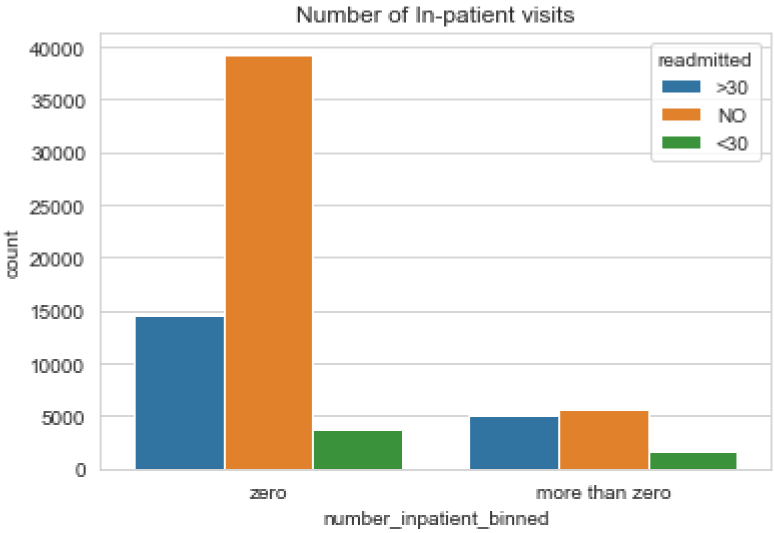
From the distribution the groups are formed into 4 ranges - upto 20, 40, 60 and 132 (maxmium) bins. The new chart is shown below

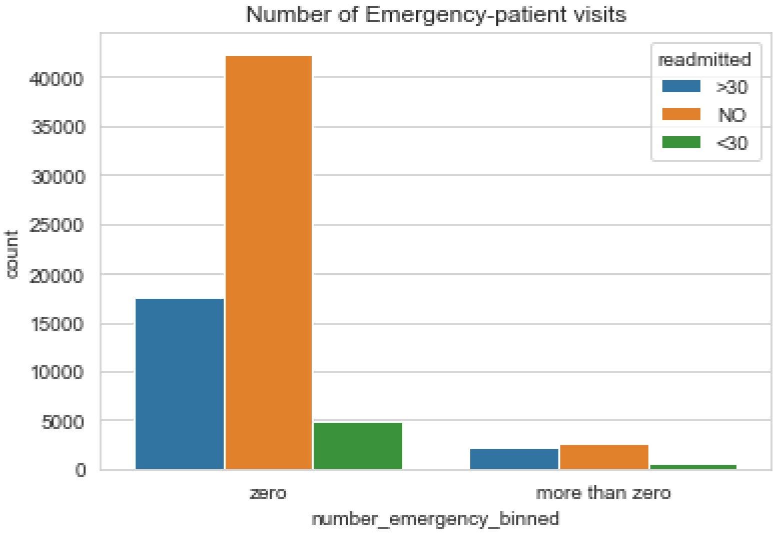


3.3.1.5 Number of outpatient, inpatient and emergency visits

From the plotted distribution, majority of the patients has zero prior in-patient, out-patient and emergency visits. So the bins were split into zero and more than zero groups







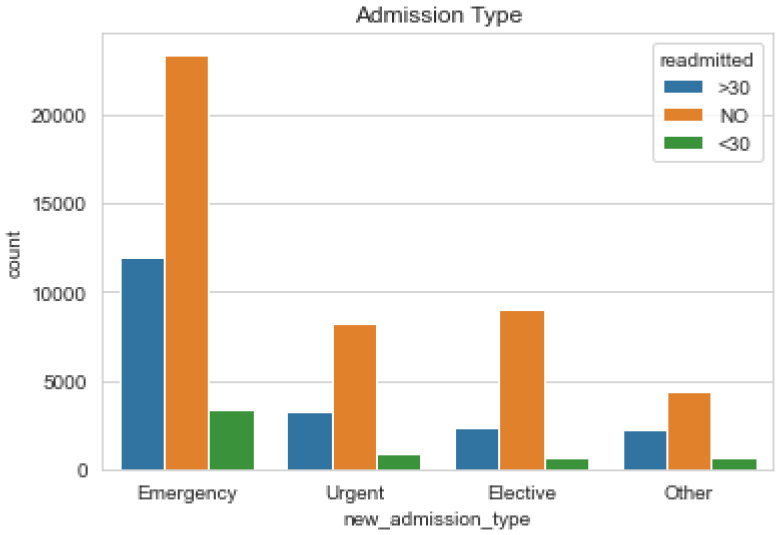
3.3.1 Binning Categorical variables

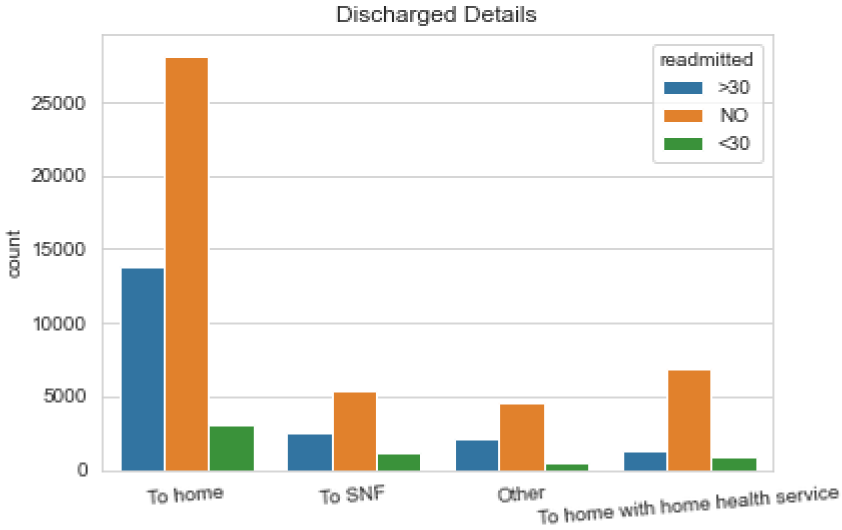
Categorical variables that contain rare or sparse values are grouped together to reduce the cardinality. Domain knowledge is also used here to bring down high cardinality variables to few relevant ones.

3.3.1.1 Variables - Admission type, discharge disposition

Admission type values - missing, Not Available, Not Mapped, Trauma Center and Newborn all these have been move to 'Other' group as they make up only about 8% of the observations.

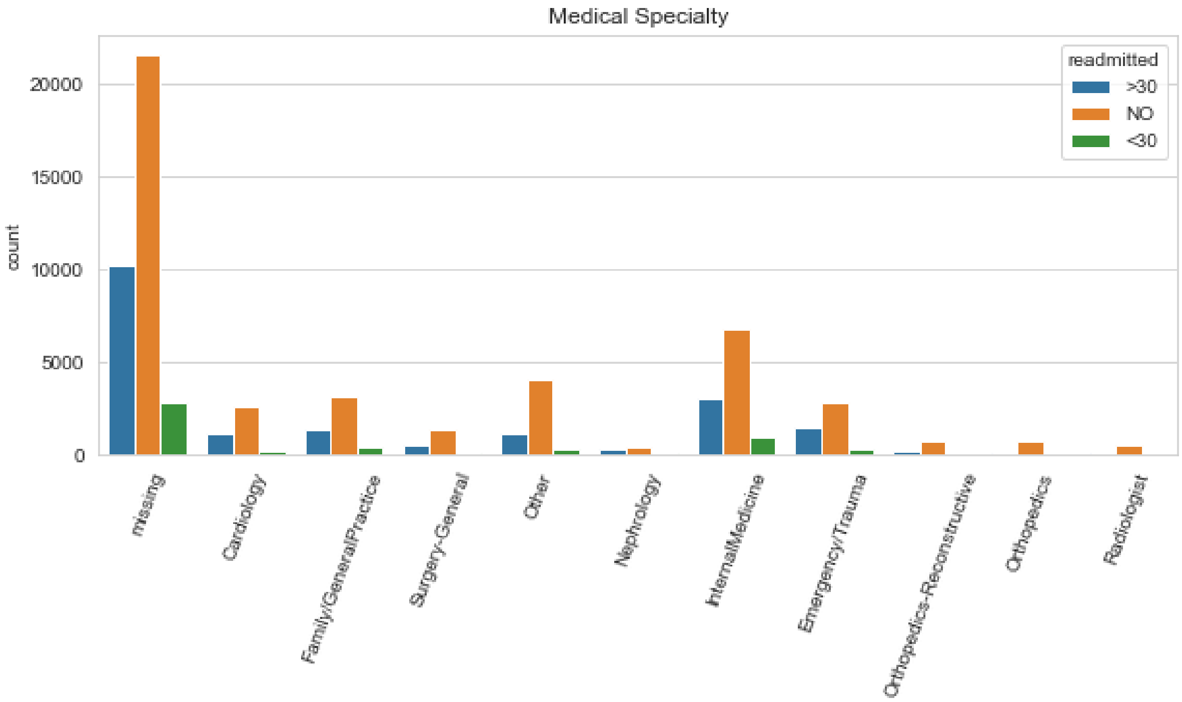
Discharge details have 26 unique values with rare labels with less than 1% observations. These were moved to 'Other' category and the 3 top categories are selected.





3.3.1.2 Variable - medical speciality

This variable has 71 unique values and so top 10 categories of medical speciality are taken and the rest are moved to 'Other'



3.3.1.3 Variables - diag_1, diag_2, diag_3

These variables have high cardinality of over 700 values since they contain the ICD-9 codes,which represent the diagnosis/reason a procedure is done. We will use the domain knowledge here to create fewer categories.

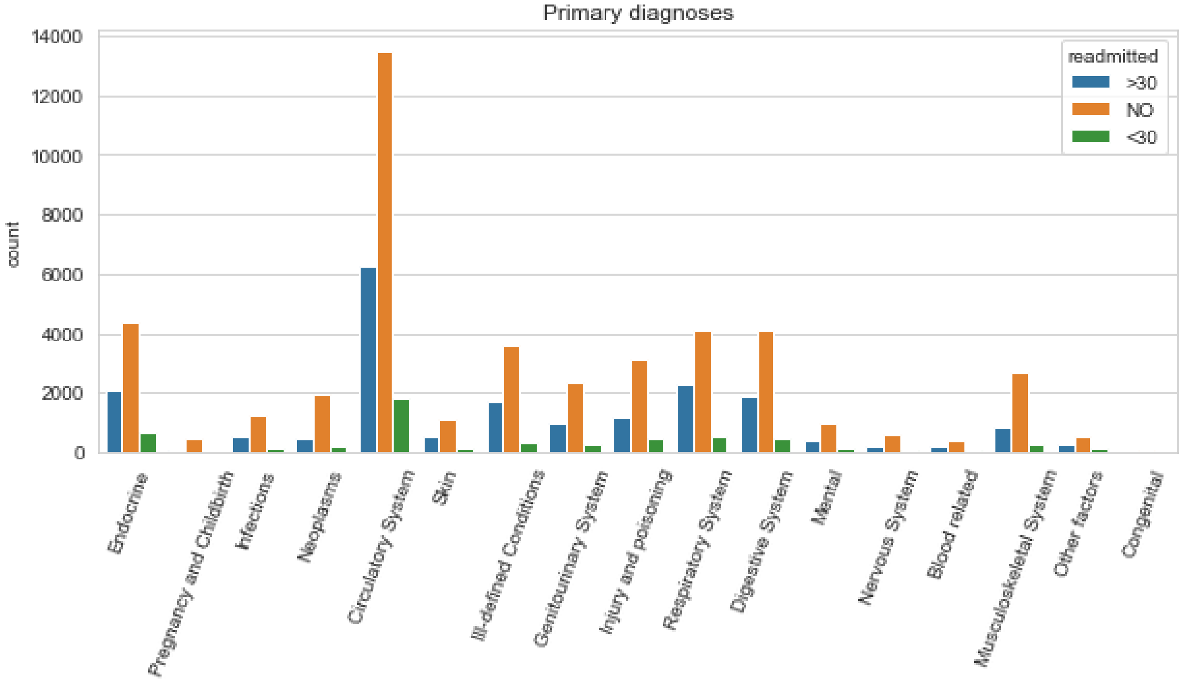
The values are mapped to the ICD-9 range codes as shown below:
001-139 Infectious And Parasitic Diseases
140-239 Neoplasms
240-279 Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders
280-289 Diseases Of The Blood And Blood-forming Organs
290-319 Mental, Behavioral And Neurodevelopmental Disorders
320-389 Diseases Of The Nervous System And Sense Organs
390-459 Diseases Of The Circulatory System
460-519 Diseases Of The Respiratory System

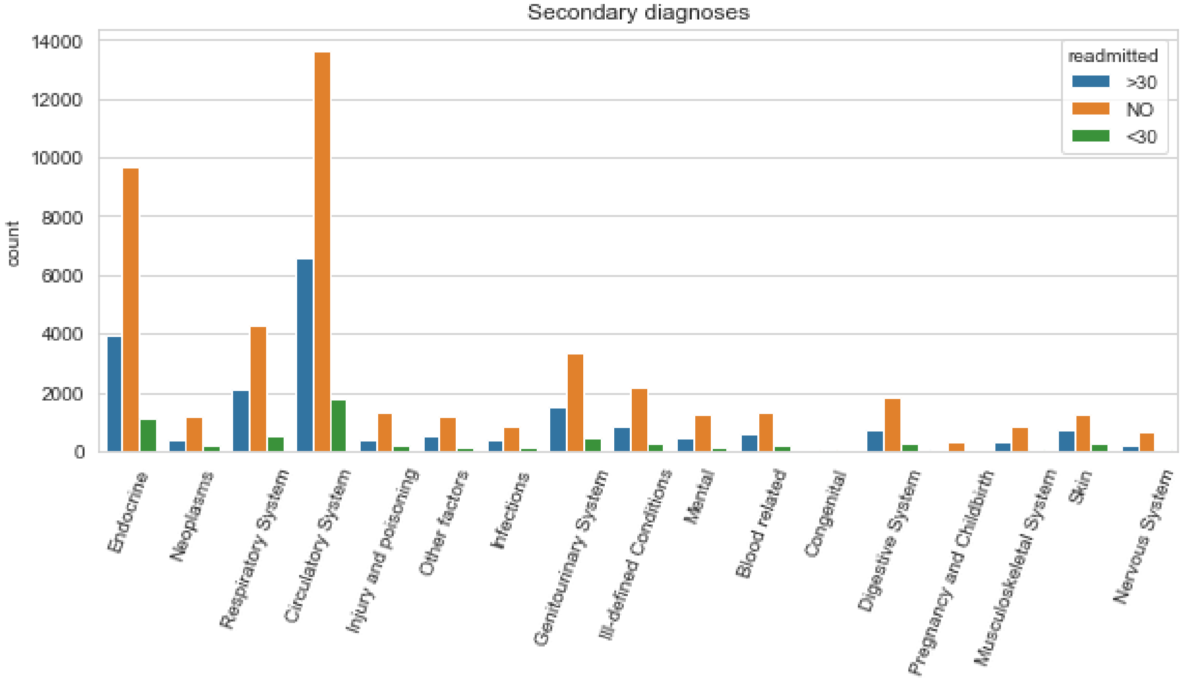
520-579 Diseases Of The Digestive System
580-629 Diseases Of The Genitourinary System
630-679 Complications Of Pregnancy, Childbirth, And The Puerperium
680-709 Diseases Of The Skin And Subcutaneous Tissue
710-739 Diseases Of The Musculoskeletal System And Connective Tissue

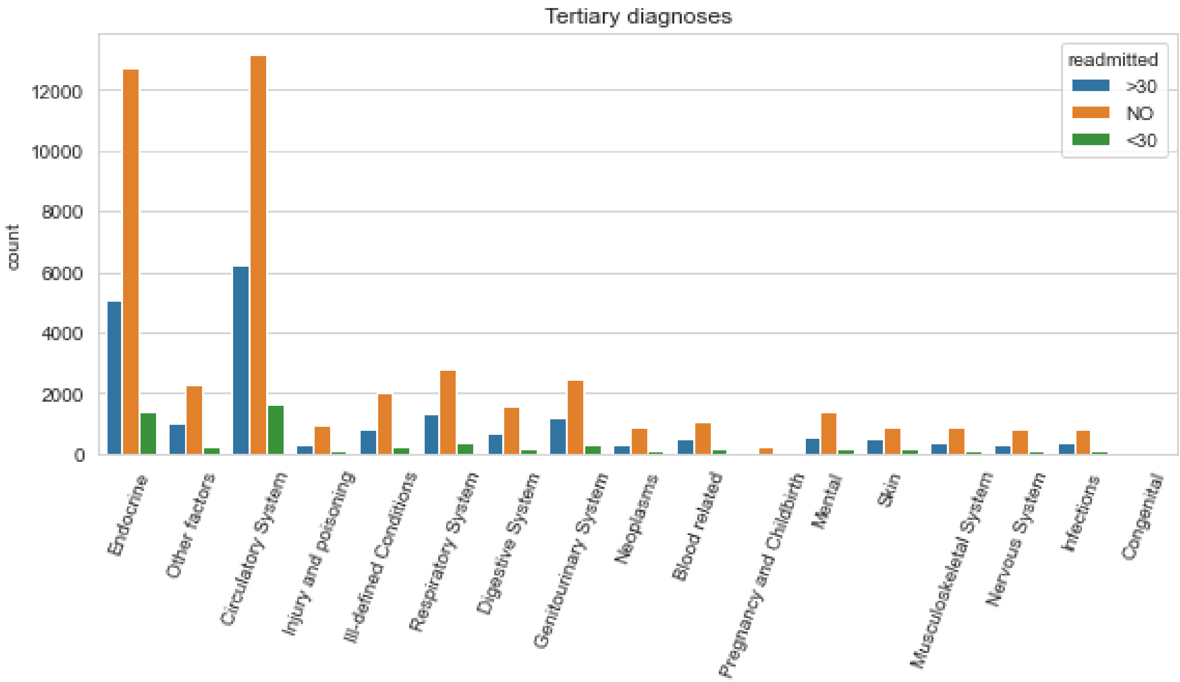
740-759 Congenital Anomalies
760-779 Certain Conditions Originating In The Perinatal Period
780-799 Symptoms, Signs, And Ill-defined Conditions
800-999 Injury And Poisoning

V01-v91 Supplementary Classification Of Factors Influencing Health Status
E000-e999 Supplementary Classification Of External Causes Of Injury And Poisoning

Circulatory related diseases score high in primary, secondary and tertiary diagnoses. Endocrine related diseases score next in these diagnoses







3 Data Modeling

For data modeling, PyCaret is used to compare, train, tune and evaluate the machine learning model.

3.1 PyCaret

Setting up the Environment in PyCaret

3.1.1 Initialize Setup

Below are the various parameters that are setup for performing this classification task

	Description	Value
0	session_id	123
1	Target	readmitted
2	Target Type	Multiclass
3	Label Encoded	<30: 0, >30: 1, NO: 2
4	Original Data	(70413, 44)
5	Missing Values	False
6	Numeric Features	0
7	Categorical Features	43
8	Ordinal Features	False
9	High Cardinality Features	False
10	High Cardinality Method	None
11	Transformed Train Set	(49289, 174)
12	Transformed Test Set	(21124, 174)
13	Shuffle Train-Test	True
14	Stratify Train-Test	False
15	Fold Generator	StratifiedKFold
16	Fold Number	10
17	CPU Jobs	-1
18	Use GPU	False
19	Log Experiment	False
20	Experiment Name	clf-default-name
21	USI	b6fc
22	Imputation Type	simple
23	Iterative Imputation Iteration	None
24	Numeric Imputer	mean
25	Iterative Imputation Numeric Model	None
26	Categorical Imputer	constant
27	Iterative Imputation Categorical Model	None
28	Unknown Categoricals Handling	least_frequent
29	Normalize	False
30	Normalize Method	None
31	Transformation	False
32	Transformation Method	None
33	PCA	False
34	PCA Method	None

	Description	Value
35	PCA Components	None
36	Ignore Low Variance	False
37	Combine Rare Levels	False
38	Rare Level Threshold	None
39	Numeric Binning	False
40	Remove Outliers	False
41	Outliers Threshold	None
42	Remove Multicollinearity	False
43	Multicollinearity Threshold	None
44	Clustering	False
45	Clustering Iteration	None
46	Polynomial Features	False
47	Polynomial Degree	None
48	Trignometry Features	False
49	Polynomial Threshold	None
50	Group Features	False
51	Feature Selection	False
52	Features Selection Threshold	None
53	Feature Interaction	False
54	Feature Ratio	False
55	Interaction Threshold	None
56	Fix Imbalance	False
57	Fix Imbalance Method	SMOTE

3.1.2 Compare all the models

This function trains all the models in the model library using default hyperparameters and evaluates performance metrics using cross-validation. The function produces a data frame with the performance statistics for each model and highlights the metrics for the best performing model

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.6536	0.6678	0.3687	0.5961	0.5677	0.1107	0.1562	7.7240
lightgbm	Light Gradient Boosting Machine	0.6535	0.6726	0.3754	0.5947	0.5775	0.1288	0.1662	0.4300
catboost	CatBoost Classifier	0.6531	0.6713	0.3820	0.6005	0.5852	0.1436	0.1749	6.2410
lr	Logistic Regression	0.6517	0.6579	0.3701	0.5900	0.5703	0.1147	0.1545	7.4630
ridge	Ridge Classifier	0.6515	0.0000	0.3667	0.5658	0.5653	0.1055	0.1485	0.1090
lda	Linear Discriminant Analysis	0.6514	0.6576	0.3714	0.5921	0.5715	0.1166	0.1553	0.5020
xgboost	Extreme Gradient Boosting	0.6510	0.6662	0.3822	0.5964	0.5846	0.1416	0.1708	11.5610
ada	Ada Boost Classifier	0.6498	0.6474	0.3669	0.5748	0.5658	0.1055	0.1452	0.6760
rf	Random Forest Classifier	0.6477	0.6526	0.3690	0.5759	0.5687	0.1094	0.1442	2.4080
svm	SVM - Linear Kernel	0.6454	0.0000	0.3472	0.5651	0.5292	0.0454	0.0906	0.7540
et	Extra Trees Classifier	0.6438	0.6462	0.3729	0.5721	0.5737	0.1179	0.1450	3.8730
qda	Quadratic Discriminant Analysis	0.6220	0.5008	0.3341	0.4964	0.5076	0.0023	0.0053	0.3490
knn	K Neighbors Classifier	0.5805	0.5699	0.3674	0.5361	0.5533	0.0845	0.0868	27.3290
dt	Decision Tree Classifier	0.5263	0.5396	0.3692	0.5316	0.5288	0.0720	0.0721	0.3340
nb	Naive Bayes	0.0927	0.5743	0.3371	0.5904	0.0412	0.0033	0.0198	0.0870

Gradient Boosting Classifier, Light Gradient Boosting Machine (Light GBM) and CatBoost Classifer returned the best model in terms of Accuracy. Extreme Gradient Boosting(XGBoost) has the best recall metric.

So will be taking Light GBM, CatBoost and XGBoost for further assessment in terms of creation and tuning of the models

3.1.3 Create the models

This function returns a table with k-fold cross validated performance metrics along with the trained model object. Evaluation metrics used for Classification are: Accuracy, AUC, Recall, Precision, F1, Kappa, MCC

3.1.3.1 Light GBM

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6559	0.6776	0.3791	0.6141	0.5823	0.1378	0.1756
1	0.6529	0.6737	0.3750	0.5682	0.5777	0.1297	0.1657
2	0.6616	0.6734	0.3801	0.6089	0.5838	0.1442	0.1903
3	0.6579	0.6734	0.3805	0.6292	0.5834	0.1413	0.1814
4	0.6460	0.6664	0.3695	0.6069	0.5691	0.1113	0.1428
5	0.6474	0.6696	0.3695	0.6382	0.5691	0.1101	0.1440
6	0.6581	0.6765	0.3778	0.5795	0.5812	0.1368	0.1794
7	0.6517	0.6697	0.3754	0.5670	0.5780	0.1293	0.1634
8	0.6549	0.6717	0.3766	0.5723	0.5794	0.1334	0.1715
9	0.6485	0.6737	0.3703	0.5623	0.5708	0.1137	0.1483
Mean	0.6535	0.6726	0.3754	0.5947	0.5775	0.1288	0.1662
SD	0.0049	0.0032	0.0041	0.0266	0.0055	0.0120	0.0157

3.1.3.2 CatBoost

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6529	0.6708	0.3831	0.5854	0.5871	0.1469	0.1766
1	0.6529	0.6767	0.3816	0.5912	0.5852	0.1451	0.1755
2	0.6586	0.6728	0.3859	0.6183	0.5903	0.1545	0.1897
3	0.6470	0.6710	0.3781	0.5913	0.5795	0.1306	0.1580
4	0.6504	0.6662	0.3804	0.6239	0.5827	0.1386	0.1678
5	0.6504	0.6650	0.3780	0.5977	0.5801	0.1325	0.1638
6	0.6579	0.6750	0.3857	0.6032	0.5895	0.1520	0.1876
7	0.6525	0.6689	0.3837	0.5973	0.5871	0.1469	0.1760
8	0.6531	0.6689	0.3798	0.5960	0.5830	0.1400	0.1726
9	0.6552	0.6778	0.3835	0.6008	0.5877	0.1488	0.1811
Mean	0.6531	0.6713	0.3820	0.6005	0.5852	0.1436	0.1749
SD	0.0033	0.0040	0.0027	0.0114	0.0036	0.0075	0.0094

3.1.3.3 XGBoost

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6541	0.6708	0.3864	0.5948	0.5901	0.1525	0.1821
1	0.6539	0.6707	0.3836	0.5980	0.5874	0.1485	0.1791
2	0.6549	0.6696	0.3825	0.5948	0.5856	0.1449	0.1786
3	0.6496	0.6625	0.3857	0.6167	0.5863	0.1434	0.1703
4	0.6452	0.6607	0.3791	0.5980	0.5790	0.1288	0.1547
5	0.6482	0.6605	0.3763	0.5792	0.5782	0.1285	0.1582
6	0.6549	0.6701	0.3894	0.6058	0.5915	0.1552	0.1852
7	0.6480	0.6631	0.3777	0.5646	0.5814	0.1362	0.1633
8	0.6492	0.6585	0.3793	0.6043	0.5815	0.1355	0.1643
9	0.6516	0.6750	0.3817	0.6078	0.5848	0.1425	0.1719
Mean	0.6510	0.6662	0.3822	0.5964	0.5846	0.1416	0.1708
SD	0.0032	0.0054	0.0040	0.0142	0.0043	0.0088	0.0099

3.1.4 Hyperparameter Tuning

After training the model, we can optimize even further with the tuning of hyperparameters. This function automatically tunes the hyperparameters of a model using Random Grid Search on a pre-defined search space. Here, tuning Light GBM, CatBoost and XGBoost to evaluate the best performing model.

3.1.4.1 Light GBM

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6579	0.6704	0.3774	0.5791	0.5805	0.1348	0.1778
1	0.6537	0.6683	0.3731	0.5685	0.5749	0.1246	0.1639
2	0.6579	0.6650	0.3742	0.5787	0.5760	0.1272	0.1741
3	0.6575	0.6714	0.3747	0.5797	0.5765	0.1269	0.1730
4	0.6468	0.6610	0.3672	0.5553	0.5671	0.1081	0.1417
5	0.6506	0.6674	0.3681	0.5647	0.5676	0.1089	0.1488
6	0.6606	0.6767	0.3778	0.5833	0.5809	0.1373	0.1847
7	0.6533	0.6687	0.3725	0.5690	0.5739	0.1218	0.1615
8	0.6531	0.6680	0.3742	0.5696	0.5760	0.1257	0.1637
9	0.6512	0.6699	0.3699	0.5666	0.5701	0.1133	0.1526
Mean	0.6543	0.6687	0.3729	0.5715	0.5743	0.1229	0.1642
SD	0.0040	0.0039	0.0034	0.0082	0.0045	0.0095	0.0129

3.1.4.2 CatBoost

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6567	0.6691	0.3707	0.5783	0.5708	0.1170	0.1664
1	0.6539	0.6671	0.3695	0.5675	0.5698	0.1160	0.1595
2	0.6590	0.6669	0.3716	0.5824	0.5719	0.1206	0.1738
3	0.6581	0.6681	0.3722	0.5811	0.5727	0.1212	0.1719
4	0.6458	0.6609	0.3626	0.5529	0.5601	0.0943	0.1302
5	0.6482	0.6626	0.3623	0.5591	0.5588	0.0924	0.1335
6	0.6588	0.6753	0.3722	0.5828	0.5727	0.1214	0.1736
7	0.6506	0.6666	0.3673	0.5641	0.5664	0.1070	0.1477
8	0.6506	0.6666	0.3664	0.5639	0.5649	0.1049	0.1467
9	0.6520	0.6700	0.3649	0.5692	0.5622	0.0997	0.1464
Mean	0.6534	0.6673	0.3680	0.5701	0.5670	0.1094	0.1550
SD	0.0044	0.0037	0.0036	0.0100	0.0051	0.0107	0.0155

3.1.4.3 XGBoost

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6573	0.6729	0.3816	0.5785	0.5860	0.1451	0.1822
1	0.6547	0.6748	0.3784	0.5937	0.5810	0.1364	0.1730
2	0.6588	0.6721	0.3807	0.6059	0.5843	0.1427	0.1836
3	0.6573	0.6666	0.3812	0.6287	0.5844	0.1426	0.1809
4	0.6478	0.6656	0.3712	0.5855	0.5719	0.1168	0.1493
5	0.6462	0.6643	0.3695	0.5731	0.5696	0.1116	0.1435
6	0.6622	0.6789	0.3866	0.6060	0.5915	0.1574	0.1979
7	0.6541	0.6713	0.3780	0.5709	0.5816	0.1372	0.1722
8	0.6543	0.6697	0.3767	0.5722	0.5796	0.1333	0.1705
9	0.6514	0.6758	0.3737	0.6455	0.5748	0.1222	0.1585
Mean	0.6544	0.6712	0.3778	0.5960	0.5805	0.1345	0.1712
SD	0.0047	0.0045	0.0049	0.0242	0.0064	0.0133	0.0158

Tuned XGBoost returned the best results, so using this model for further analysis.

3.1.4 Model Analysis

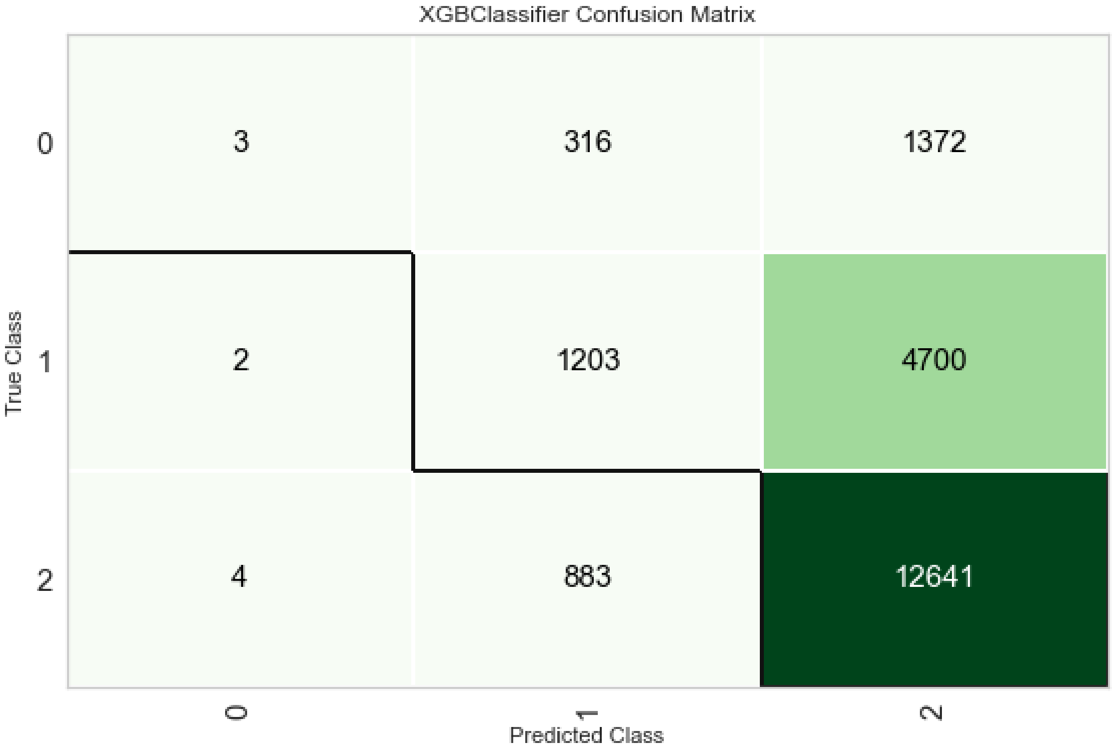
Performance evaluation and diagnostics of a trained machine learning model, tuned XGBoost is explored here.

3.1.4.1 Confusion Matrix

For the target class the Label Encoded is as follows:

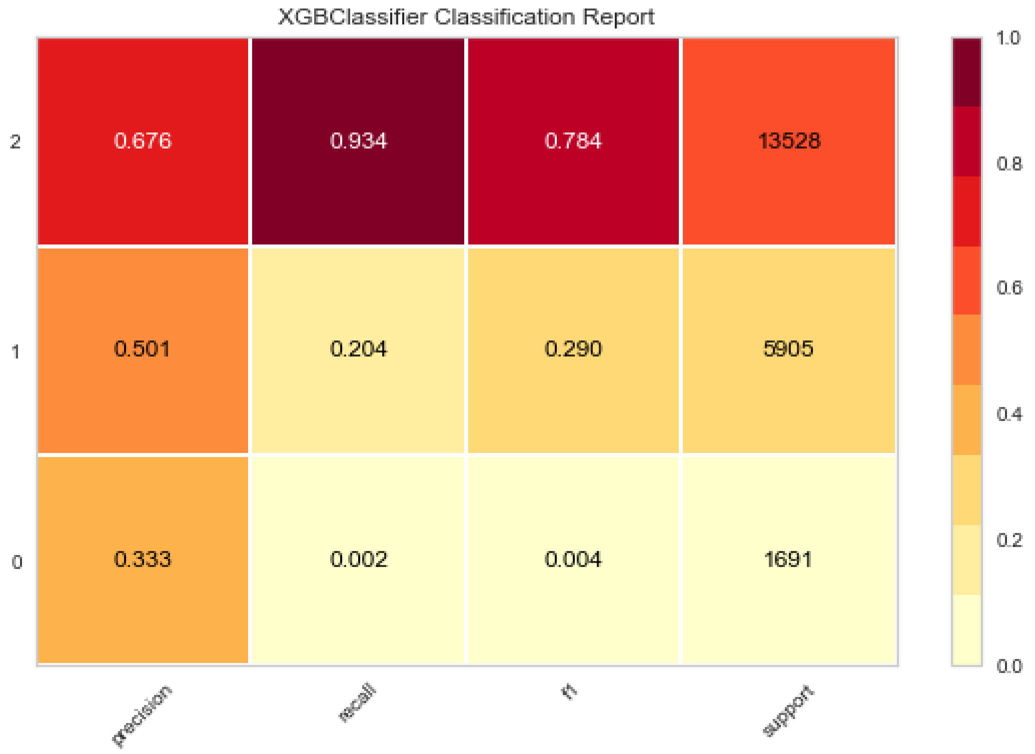
- '<30': 0
- '>30': 1
- 'NO': 2

The model is predicting correctly the maximum number of true positives for 'NO' label. It is correctly predicting that more patients that do not require readmission. Since data related to class label '<30' is sparse we are getting much less information from this model

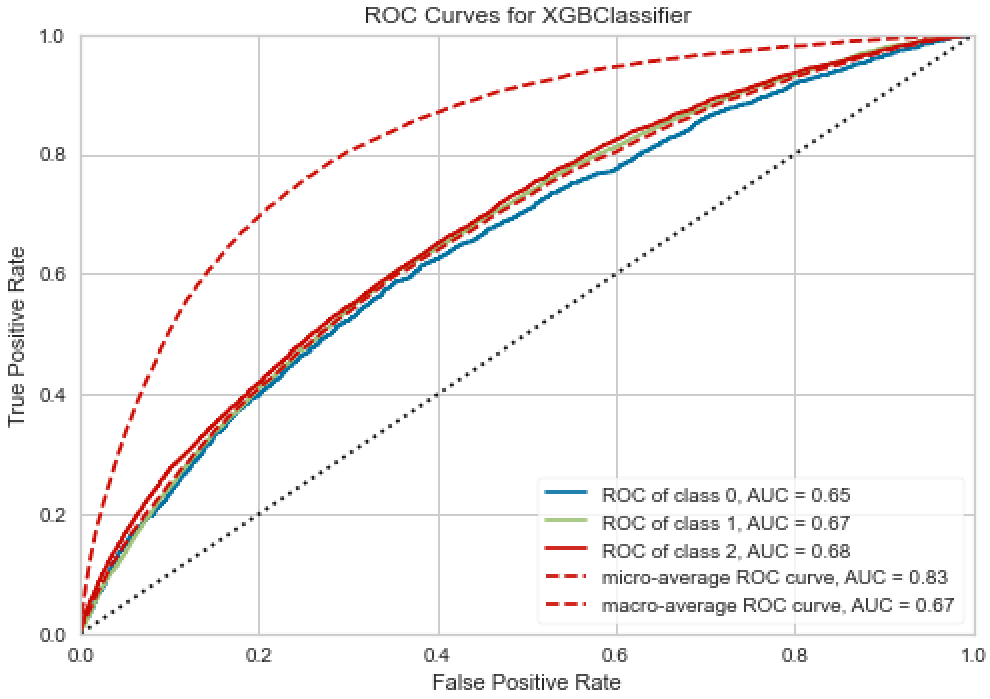


3.1.4.2 Classification Report

The prediction model for class label 'NO' performs better than class label '<30'.


XGBClassifier Classification Report
2
1
0
0.676
0.934
0.784
13528
0.501
0.204
0.290
5905
0.333
0.002
0.004
1691
precision
recall
f1
support
1.0
0.8
0.6
0.4
0.2
0.0

3.1.4.3 AUC

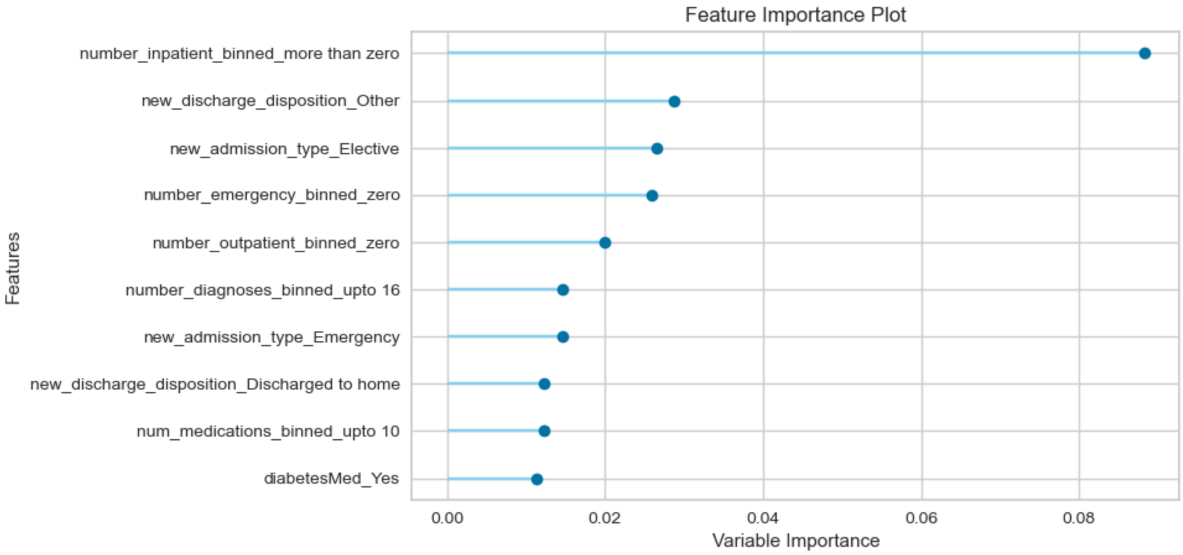

ROC Curves for XGBClassifier
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0, AUC = 0.65
ROC of class 1, AUC = 0.67
ROC of class 2, AUC = 0.68
micro-average ROC curve, AUC = 0.83
macro-average ROC curve, AUC = 0.67

3.1.4.4 Feature Importance

The feature that is predicted to be most important is the number of inpatient visits prior to the hospital encounter. So if the patient was admitted as inpatient in the past this increases their chance of being readmitted.

Discharged details , admission type and number of emergency visits prior to the hostial encounter also are shown as important features.

These information can be used by health service related authorities to better their care giving program. They can improve their inpatient and emergency care protocols to keep the readmission rates to a minimum.



3.1.5 Model Prediction

Predicting the model using the test data yielded better results with tuned version of the XGBoost classifer whch was 0.6544. This shows our model is not constrained by overfitting.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Extreme Gradient Boosting	0.6555	0.6744	0.3800	0.5993	0.5834	0.1406	0.1763

4. Conclusions and Future work

- From the model analyzes we can see that the data is unbalanced, 68% of the target variable is for 'NO', 28% for '>30' and 8% for '<30'. We can improve in this area by collecting more data and also by using techniques like undersampling, oversampling, SMOTE or combinations of these techniques to have a more balanced data.
- The data did not contain any dates related to the patient hospital encounter, so we cannot assess that we are using the latest health record for the patient. About 20% of the patients had multiple encounters and we used only the first record of the patient.
- Model performance can be improved with less features. So further EDA and with the help of domain knowlege can again attempt to reduce redundant features
- Can explore binary classification, where the target variable, readmitted can be 'YES' or 'NO' to answer whether a patient might be readmitted for less than 30 days
- Can leverage the indepth exploratory data analysis of this data to extrapolate various patterns with multiple features.

5. References

- https://www.who.int/news-room/fact-sheets/detail/diabetes (https://www.who.int/news-room/fact-sheets/detail/diabetes)
- https://www.diabetesresearch.org/diabetes-statistics (https://www.diabetesresearch.org/diabetes-statistics)
- https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program (https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program)