SPRINGBOARD DATA SCIENCE COURSE

CAPSTONE PROJECT 1

Predicting the likelihood of hospital readmission for Diabetes Patients

Abhishek Sukhadia

12th April, 2020

Contents

1. Introduction 1

2. Data Acquisition and Cleaning 2

3. Data Wrangling and Outlier Analysis 3

3.1 Data Wrangling 3

3.2 Outlier Analysis 4

4. Exploratory Data Analysis 6

4.1. Hospital Readmission Rates 6

4.2. Visit Number 7

4.3. Patient Medications 8

4.4. Patient Characteristics 9

4.5. Patient Admission and Discharge Information 11

4.6. Patient Medical Diagnosis 12

4.7. Patient Encounters Details 14

4.8. Frequentist and Inferential Statistics 15

1. Introduction

As per Wikipedia, hospital readmission refers to an "episode when a patient who had been discharged from a hospital is admitted again within a specified time interval." These readmission rates are used as a metric to measure quality of health services in any country. The most common time frames used are 30-day, 90-day or 1 year.

Hospital Readmission rates were made part of (Patient Protection and Affordable Act, 2010 in the United States. Under this act, a hospital gets penalized if they have higher than expected readmission rates. Thus it has become an important measure for hospitals to monitor and reduce their readmissions. It has become an important indicator of hospital quality and affects the cost of care adversely. Being able to determine factors that lead to higher readmissions, and correspondingly being able to predict which patients will get readmitted can help hospitals save millions of dollars while improving quality of care.

The most important stakeholders are Government Healthcare regulatory authorities, Hospitals and Patients. Center for Medicare & Medicaid Services (CMS), a federal agency within the United States Department of Health and Human Service, has established a Hospital Readmissions Reduction Program. This project sees direct application to such programs run by government agencies. Hospitals can directly use the results of this project to identify factors that are causing high readmission rates and thus concrete action to reduce the same. This would help them save penalties and thus improve the quality of care for the patients. Currently CMS has included 6 diseases in this program and they are regularly adding new disease conditions to the list. I have taken diabetes as the disease to predict readmissions even though it is not part of the CMS list.

In US, diabetes affects nearly 34.2 million Americans - just over 1 in 10 have diabetes (National Diabetes Statistics Report, 2020) and this number is expected to grow substantially every year. It's the fifth leading cause of death in America, more than breast cancer and AIDs combined. The American Diabetes Association released new research on March 22, 2018 estimating the total costs of diagnosed diabetes have risen to $327 billion in 2017 from $245 billion in 2012, when the cost was last examined. For the cost categories analysed, care for people with diagnosed diabetes accounts for 1 in 4 health care dollars in the U.S., and more than half of that expenditure is directly attributable to diabetes. This proves that there is huge burden both on patients and hospitals on the growing costs and readmissions contribute a significant portion of this burden. Reducing readmission rates of diabetic patients has the potential to greatly reduce health care costs while simultaneously improving care.

2. Data Acquisition and Cleaning

I have used data available on Kaggle which was in turn sourced from UCI repository which contains de-identified diabetes patient encounter data for 130 US hospitals (1999–2008) containing 101,766 observations over 10 years. The dataset has over 50 features including patient characteristics, conditions, tests and 23 medications.

Each row is a single encounter of a patient with details given below.

Patient Characteristics:

- Encounter ID, Patient Number, Race, Gender, Age, Weight

Patient Hospital Admission and Discharge Details:

- Admission Type, Discharge Disposition, Admission Source, Time in Hospital, Payer Code, Medical Specialty

Patient Encounter Details:

- Number of lab procedures, procedures other than lab tests, medications, outpatient visits, inpatient visits, emergency visits

Patient Medical Diagnosis Details:

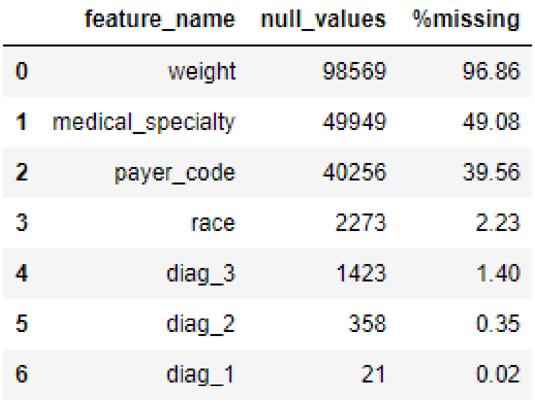
- Primary, secondary and additional secondary Diagnosis, Number of diagnoses, A1C result, Glucose serum test result,

Patient Medications:

- change of medications, diabetes medications, 24 features of drug medications (generic names)

Details about each field can be found in the research article "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records".

The above data was downloaded as a CSV and then imported in python. More details of this can be found in this **IPython Notebook**. Post importing, data was checked for missing values which were present as "?" in the data set. Table 1 is the summary of the missing values.



	feature_name	null_values	%missing
0	weight	98569	96.86
1	medical_specialty	49949	49.08
2	payer_code	40256	39.56
3	race	2273	2.23
4	diag_3	1423	1.40
5	diag_2	358	0.35
6	diag_1	21	0.02

Table 1: Features with missing values

Missing values were handled differently for each feature.

- **Weight** have ~97% missing values. There is poor interpretability of the missing values, so it is best to drop this column
- **medical_speciality** and **payer_code** have 40-50% missing values. I have decided to drop it but there are ways to deal with it. Alternatively, if these columns are not dropped then, it is best to create a separate category of 'missing'.
- **race, diag_3, diag_2, diag_1** have only <=2% missing values, so missing rows were dropped for diagnosis and replaced with mode of the column values for race.
- A missing first diagnosis while data have diagnosis second and third, is also a bad data. But second and third diagnosis can give us some idea about the readmission. So we can either keep this missing values under "missing" category or just ignore the rows. I am going with first one.
- In **gender**, there are values like "Unknown/Invalid" which are missing. These columns are also dropped

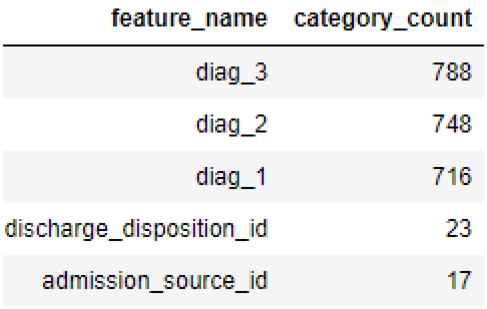
Apart from removing missing values, data required further cleaning based on domain knowledge of readmissions.

We ignored patients who died post hospital admission since they have zero probability of readmission and thus would bias the data. I also dropped two columns where all records have same values (examide and citoglipton) since it cannot provide any information about readmission. More details about the data cleaning process can be found in this **IPython Notebook**. The cleanest data is ready for data wrangling.

3. Data Wrangling and Outlier Analysis

3.1 Data Wrangling

This data sets contains categorical data with number of categories ranging from 2 to 788. The features with high categories can increase the dimensionality of the data and make is sparse. So it was decided to reduce the number of categories through grouping. Table 2 shows the features identified for grouping of categories.



feature_name	category_count
diag_3	788
diag_2	748
diag_1	716
discharge_disposition_id	23
admission_source_id	17

Table 2: Features with high category count

Diagnosis columns (diag_1, diag_2 and diag_3) contain numbers that are ICD9 Codes describing specific diseases. They have been clubbed into larger categories in accordance with ICD9 to reduce the categories.

For discharge disposition, admission type and admission source, ~90% of data falls under 3-4 categories (Figure 1). Thus we can club the remaining categories with the existing category or under others. This would help reduce the noise from the data. Attempt was made to group items based on similar distribution of target variable and same feature characteristics like “urgent and Emergency”, “Not Applicable, Not Available, Missing” etc (Figure 2). More Details of the grouping is available in this **IPython Notebook**.

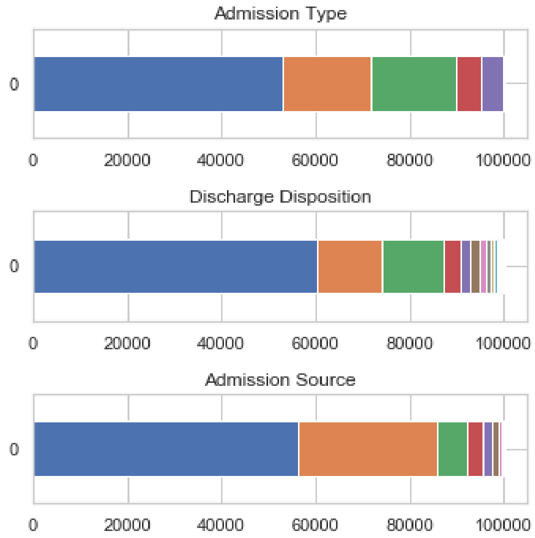


Figure 1: Pre-Grouping Hospital Admission Features

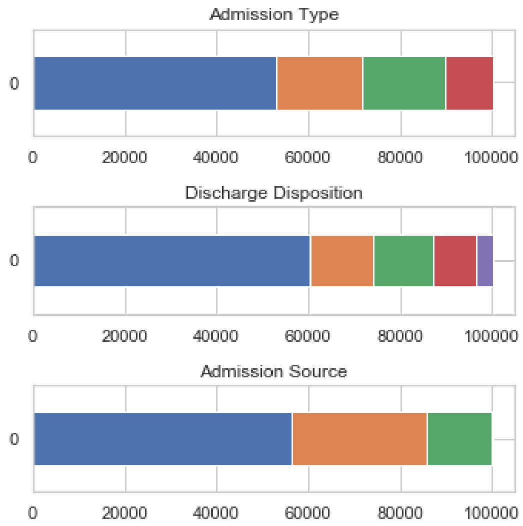


Figure 2: Post-Grouping Hospital Admission Features

By looking at the individual patient’s id, it was observed that many of the patients have had multiple inpatient visits post readmission. This implies there are non-unique encounters for patient. This means, those observations are not statistically independent. The patient next visit data might get affected based on the patient’s previous visits. Since it violates one of the assumptions of logistic regression, I have created a new feature column to deal with this multiple inpatient visits. The new column would be the ‘visit_num’ defined as the visit number for that patient i.e. whether it is his/her first visit, second visit, etc. The rank is determined by the ascending order of the encounter_id. The groups made were 1, 2, 3, 4-30 and > 30. Post this grouping patient and encounter index were dropped. More Details of the grouping is available in this **IPython Notebook**.

3.2 Outlier Analysis

The numerical variables in the data set are all discrete count data. Figure 3 shows the distribution of all discrete variables in the dataset. All the variables shows long tailed distribution skewed towards either right or left. This points out that there are outliers that need to be removed. E.g. A person having number of medications administered during encounter as 80 is very highly unlikely and using that data would only skew our results. Similar logic applies to other variables. So with cut-off value as 3-sigma (99.7%), outliers observations were removed from the data. Figure 4 shows the distribution of variables post outlier analysis. This cleanest data is then ready for exploratory analysis.

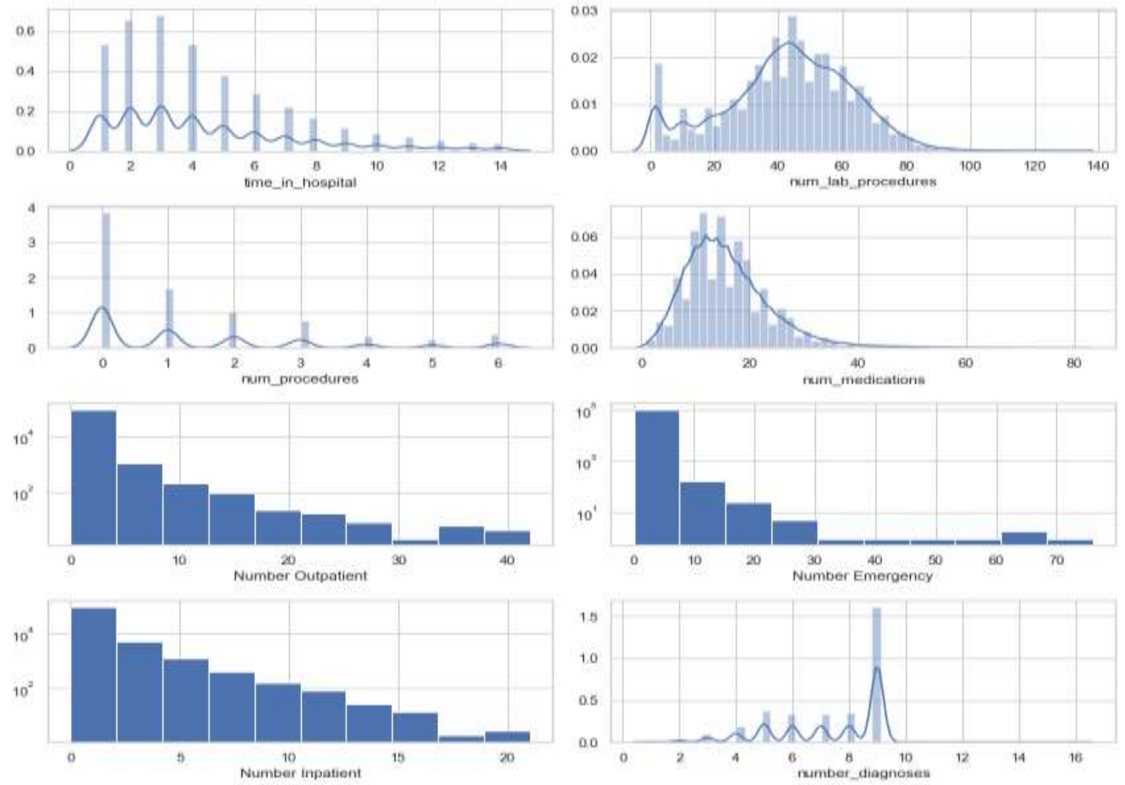


Figure 3: Original Distribution of Discrete Numeric Variables

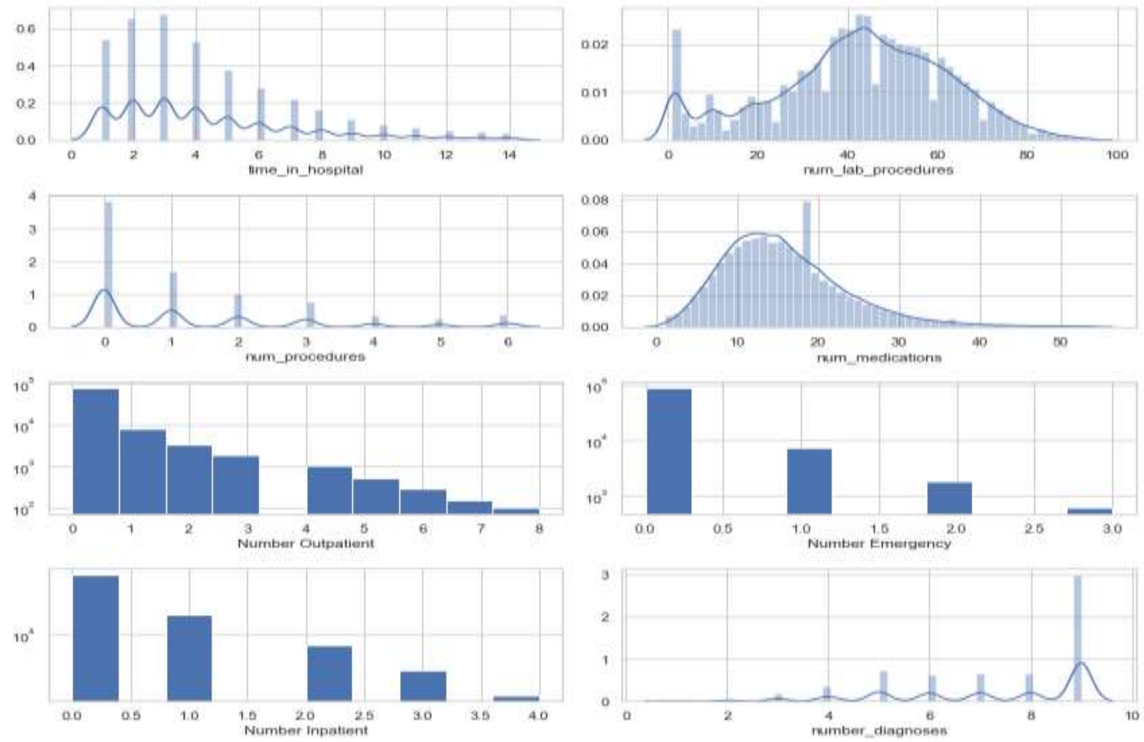


Figure 4: Post Outlier Analysis Distribution of Variables

4. Exploratory Data Analysis

In the cleaned data, there are now 96551 patient records, 44 features/columns and one target variable. We will go through most of the features in the dataset to explore the relation with the readmission rate.

The target variable for this project is called "readmitted" which contains three values:

- "<30" : Person gets readmitted within 30 days (0)
- ">30": Person gets readmitted after 30 days (1)
- "NO": Person doesn't get admitted (2)

4.1. Hospital Readmission Rates

The distribution of the target variable in **Figure 5** shows that our dataset is highly imbalanced. The distribution of "<30" (class of interest) is just 10.7% compared to remaining two classes. This imbalance needs to be kept in mind while applying our machine learning algorithms. It would also require special treatment at many places.

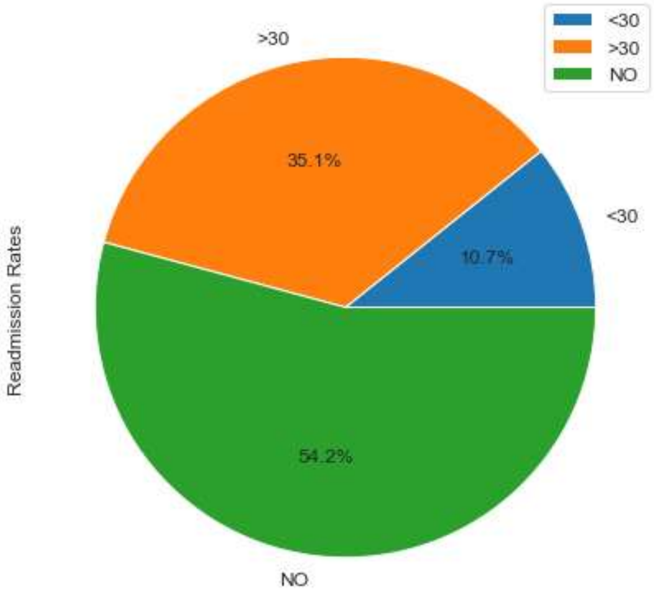


Figure 5: Distribution of Target Variable – Readmitted

We define the readmission rates as:

$$\text{Readmission rates} = \frac{\text{A person getting readmission within} < 30 \text{ days}}{\text{Total Number of Patient for a given scenario}}$$

Readmission would depend on lot of factors like type of medicine prescribed, patient medical conditions, patient age, race, results of laboratory tests, time in hospital, whether patient is visiting for the first time or multiple times, etc. In the following subsections, we will go through many of the above mentioned factors and explore the distribution of our readmission (<30 class). As per Section 2, there are broadly 5 categories of information that are embedded in most of the fields.

1. Patient Medications
2. Patient Characteristics
3. Patient Hospital Admission and Discharge info
4. Patient Medical Diagnosis
5. Patient Encounter Details

Before we deep dive into above categories, we will look at one of the important feature that would affect our readmission rates.

4.2. Visit Number

A person admitting for the first time, his/her visit number would be 1. Post first readmission, if he gets readmitted again or visit the hospital, his/her visit number will then be 2. Thus visit number is the number of visit to hospital made by the patient post first readmission. This feature was created as per Section 3.1. ~30% of my data has visit number as >1.

On comparing visit number with readmission rates, we see that as the visits of the patient increases to the hospital, his/her patient readmission chances also increases (Figure 6). Common sense would say that after a particular number of visits, readmission rate will be zero either because patient gets died or gets cured or gets transferred to other hospitals. During our data cleaning and outlier analysis, we have removed all such cases. So that would not be visible in our trend.

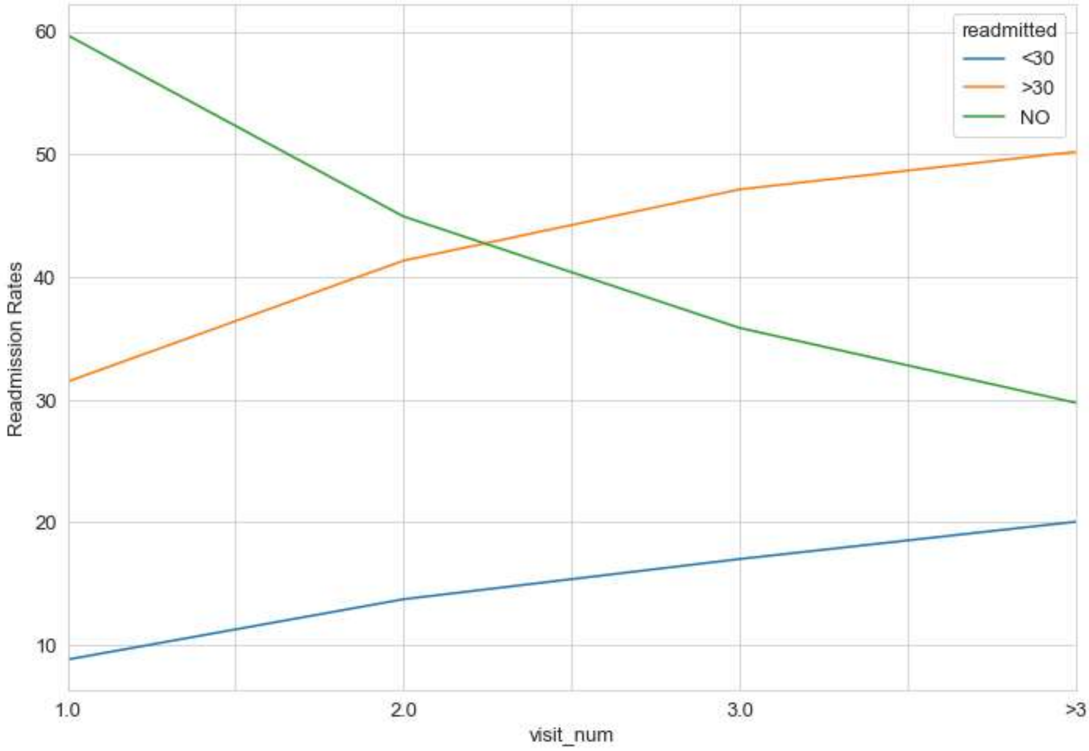


Figure 6: Readmission rates of different visit numbers

4.3. Patient Medications

Medications (21 Features)

For the 21 features, exploratory analysis in Figure 7 shows that there were few medicines not prescribed to 99.9% of the patient. This observation told us that even if we remove those features, it would not affect my target variable distribution. Thus we removed 9 medicines from the cleaned data set and final dataset will have 35 features. More Details of which medicines were removed is available in this **IPython Notebook**.

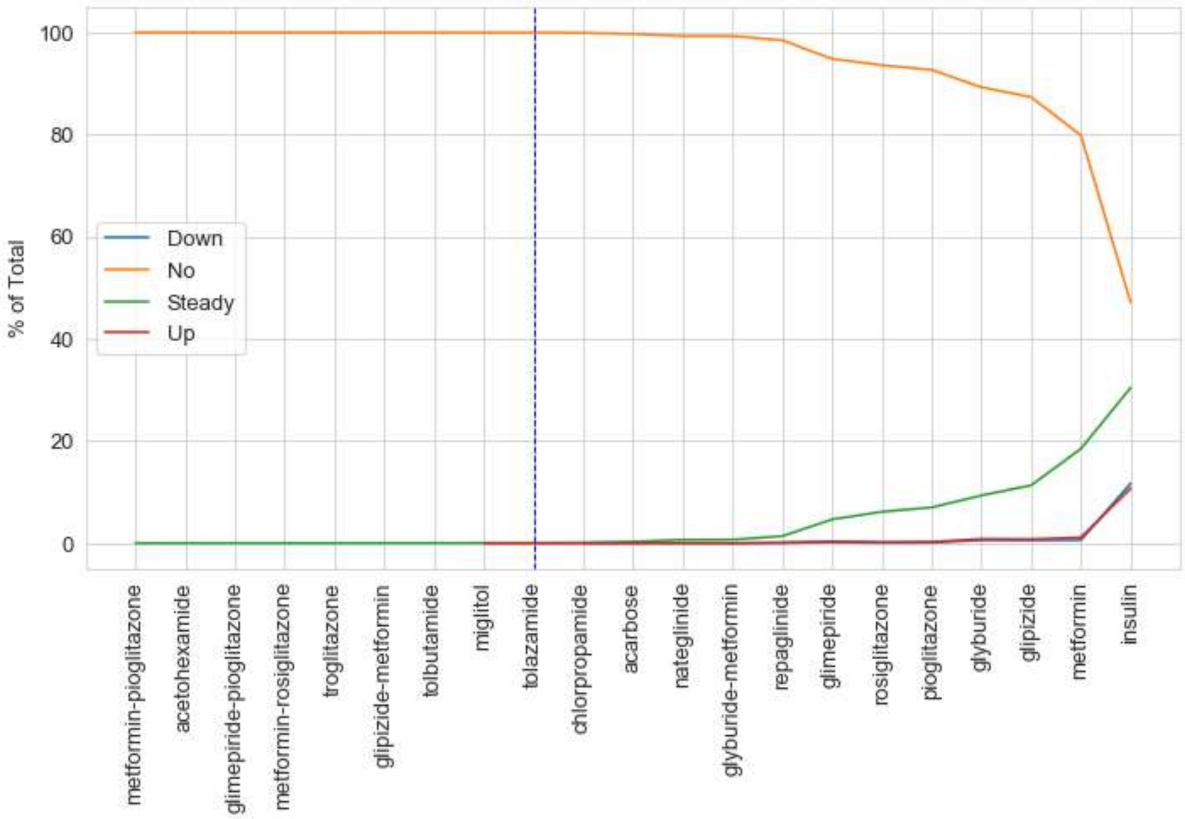


Figure 7: Medications and their prescription changes

Change of Medications

Patients whose medications were changes have higher chance of getting readmitted (Figure 8). Possible reasons may be wrong medicines prescribed, experimenting with medicines, wrong dosage of medicines prescribed, and person's immunity to the medicines.

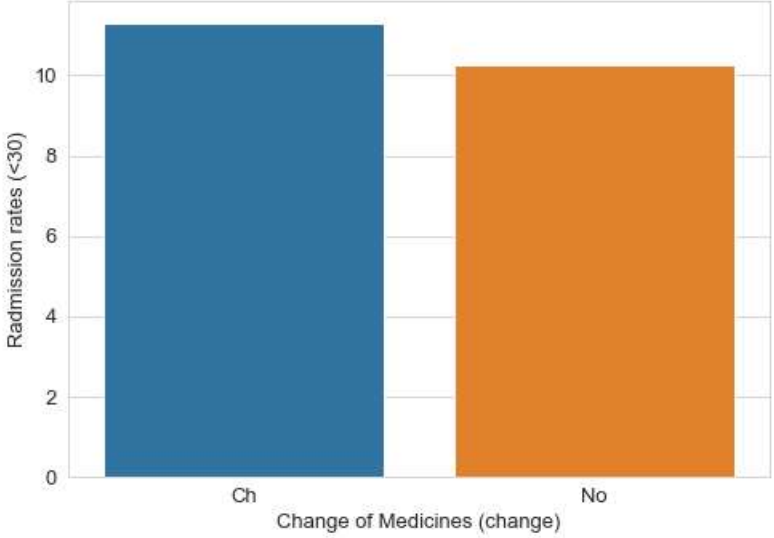


Figure 8: Change of Medicines affecting Readmission rates

4.4. Patient Characteristics

Race, Age and Gender

~77% of the patients are from Caucasian Race. This is as expected because the data set is from United States. Almost 90% of the patients more than 40 years of age, this implies that diabetes occurs in patient above the age of 40 years (Figure 9). The ratio of female and male is approximately uniform.

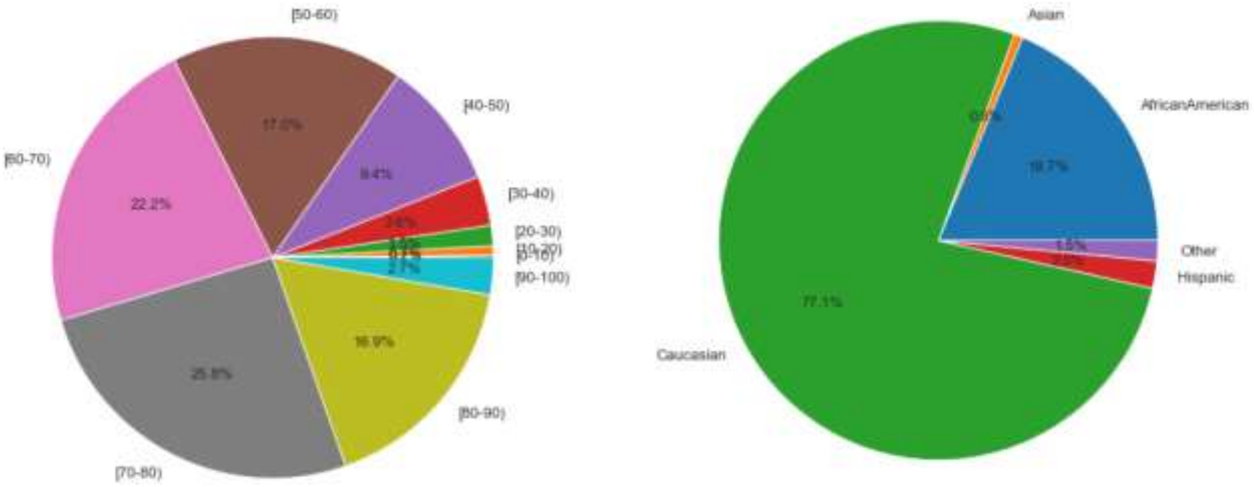


Figure 9: Age and Race Distribution of Patients

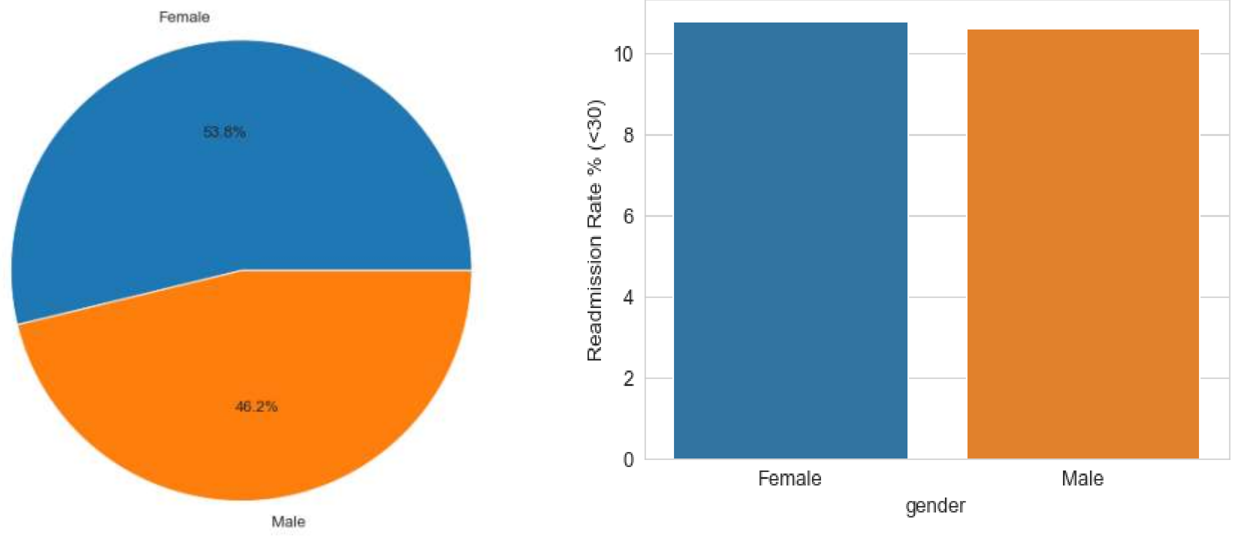


Figure 10: Gender Distribution of Patients and its relation with readmission rates

In Figure 10, we observe that gender doesn't play any major role in determining readmission rates. On exploring relation of race with readmission we found that, Caucasians and AfricanAmerican are at highest risk of being readmitted. With the exception of 20-30 age category, we see in Figure 11 that as age increases, chances of readmission also increases. Thus old people are more likely to get readmitted.

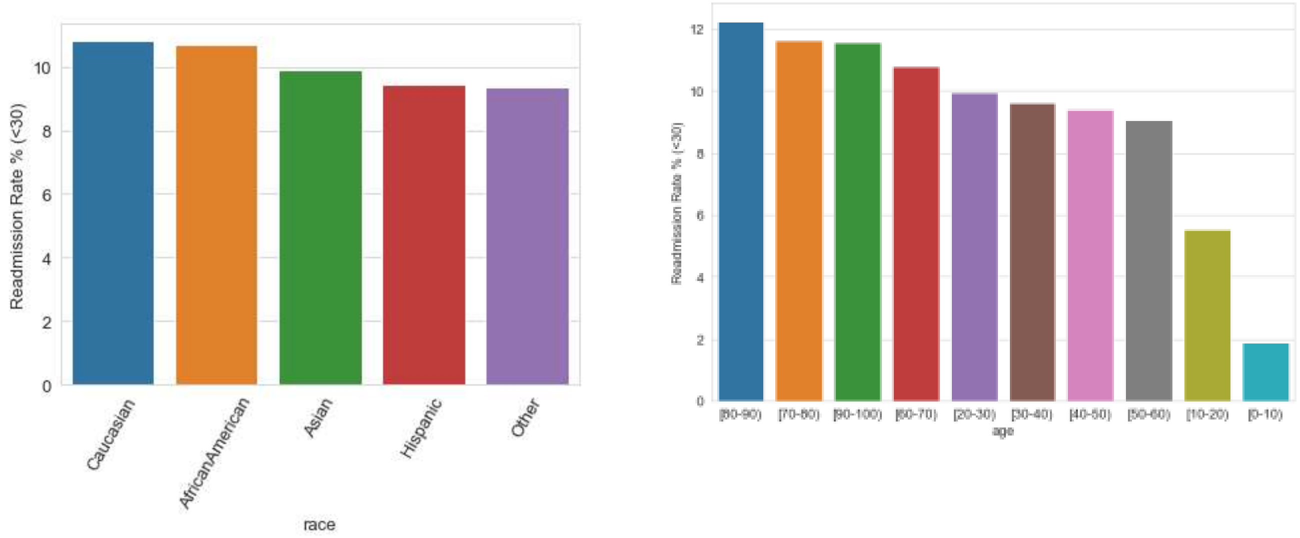


Figure 11: Readmission rates for age and race features

4.5. Patient Admission and Discharge Information

This information consists of reason for admission, admission source and how was the patient discharged. All the above three features affect the readmission rate because they may contain indirect information regarding the patient conditions at the time of admission and discharge. For example, a person coming for elective surgery would have a lesser chance of getting readmitted than a person coming as emergency. Similarly, a person getting discharged to home mean fully recovered compared to discharge to other facilities where he/she may not have recovered fully.

Exploratory analysis suggested that an Emergency Patient has high chance of getting readmitted compared to other reasons. In Figure 12, 7 and 1 represents emergency. ID mappings are available in the dataset. For discharges, a person getting discharged to home/hospice (Id = 1) have less chance of readmission compared to other categories (Figure 13).

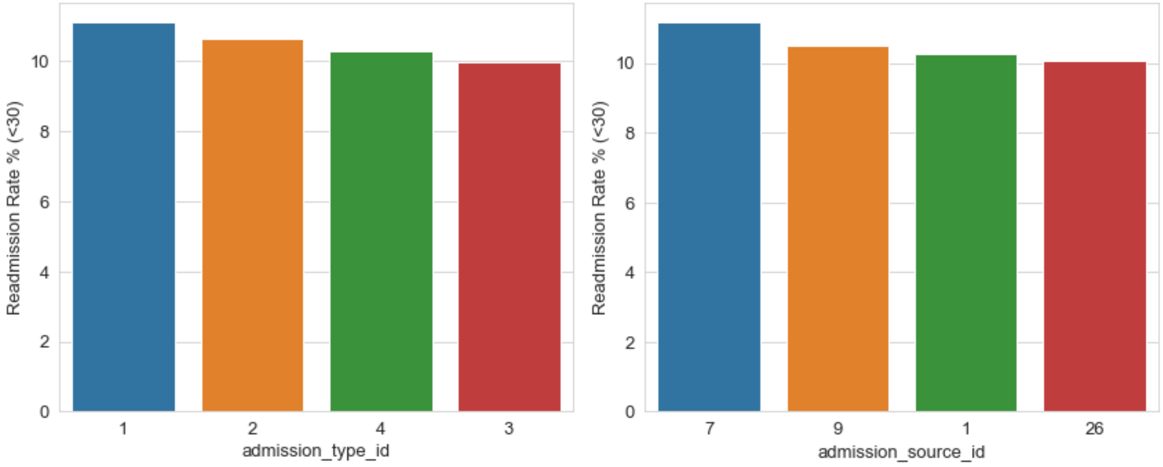


Figure 12: Admission Type and Source Readmission Rate

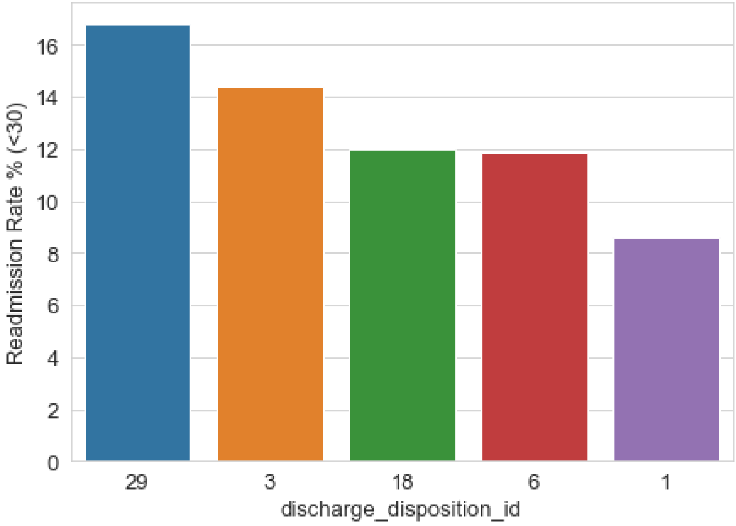


Figure 13: Readmission Rates for Discharge Disposition

4.6. Patient Medical Diagnosis

Glucose Serum and A1C Tests

Since we are dealing with diabetes patients, we will first look at the Glucose serum and A1C test results. Both these tests are performed to diagnose diabetes. A person with diabetes will show values of >200 in glucose test and >6% in A1c results. As both the value increases, intensity of diabetes also increases. Figure 14 and Figure 15 shows that diabetes patients whose lab results are >300 in glucose serum test and no A1C test was performed, have higher chance of readmission.

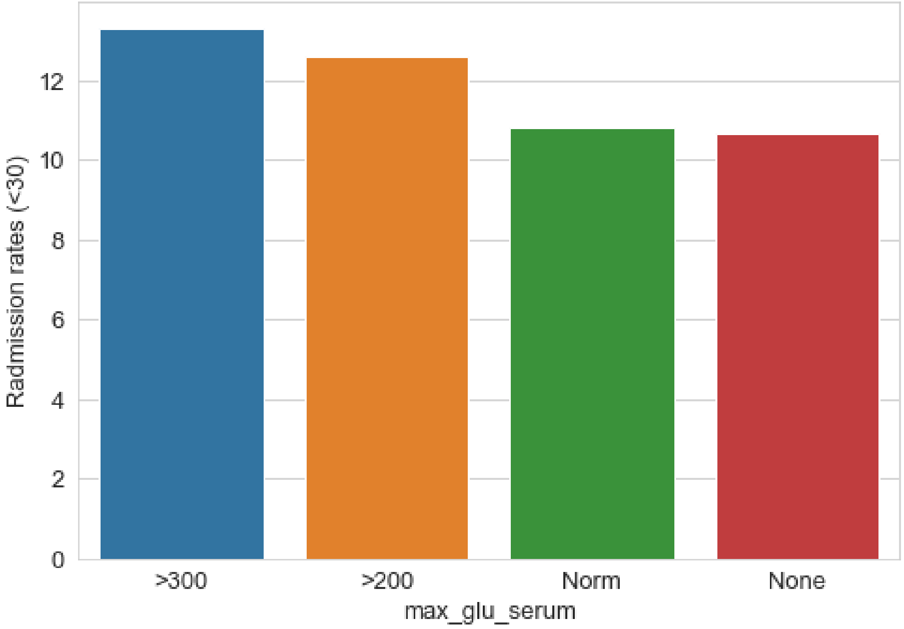


Figure 14: Glucose Serum Test Readmission Rates

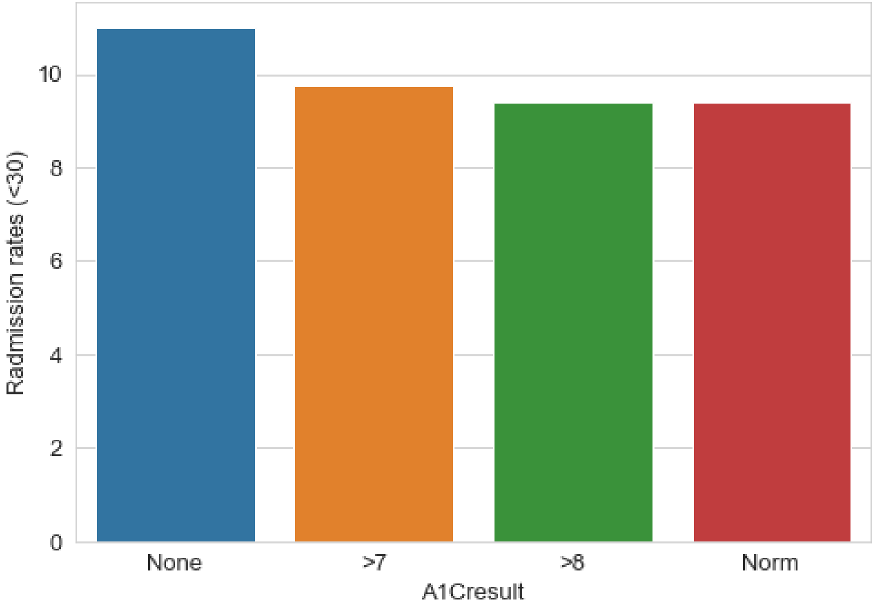


Figure 15: A1C Result Test Readmission Rates

Diagnosis

There are three diagnosis columns; primary, secondary and additional secondary diagnosis. This features represents the medical diseases of the patient i.e. whether patient has diabetes, cancer, etc. Just exploring the distribution of diagnosed diseases among the sample data, we found that Circulatory is the major diagnosis and present among 30% of the patients. Diabetes represents only ~8% among the patients (Figure 16).

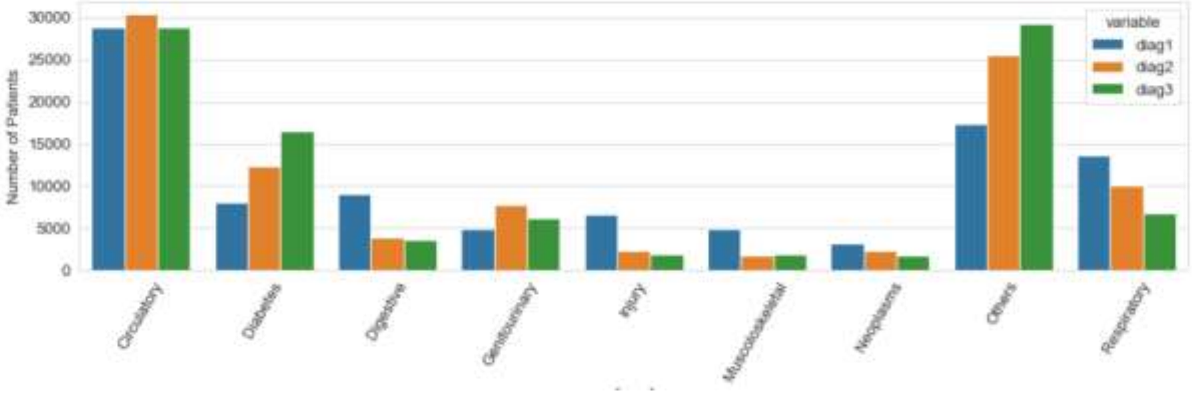


Figure 16: Types of Diagnoses among Patients

Figure 17 shows that a patient having primary diagnosis as diabetes have high readmission rates compared to other diseases. But when we see secondary and additional secondary diagnosis, diabetes is pushed towards the end and Neoplasms (Cancer Related) and Genitourinary (related to kidney) gets pushed up towards high readmission rates. This implies a particular combination of primary, secondary and additional secondary would affect the readmission rates. Our machine learning model will be further able to help us identify those combinations.

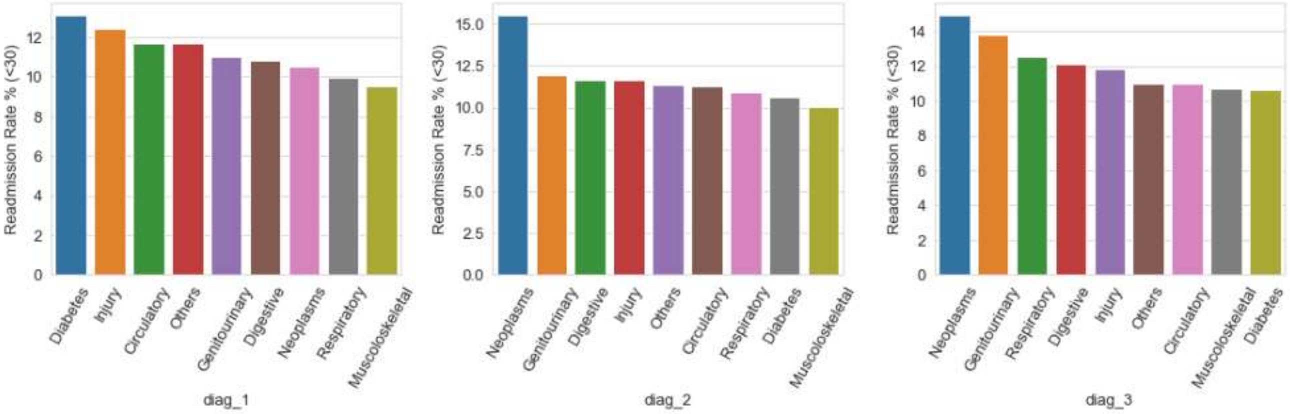


Figure 17: Readmission Rates across different diagnosis types

diag_1 – Primary Diagnosis

diag_2 – Secondary Diagnosis

diag_3 – Additional Secondary Diagnosis

4.7. Patient Encounters Details

This group contains following features:

- Number of lab procedures performed during the encounter
- Number of procedures other than lab tests during the encounter
- Number of distinct medications administered during the encounter
- Number of outpatient and Inpatient visits in the year preceding the encounter
- Number of emergency visits in the year preceding the encounter
- Number of diagnoses entered in system

All of the above features are discrete numeric variables i.e. count variables. The distribution of these variables is shown in Figure 4. Except for diagnoses number, all variables are skewed towards right suggesting they follow lognormal distribution. Number of diagnoses is skewed towards left. To explore relation between above features with readmission rate, we did a simplification. We combined '>30' and 'NO' in to one class No-Readmission (0) and '<30' as class Readmission (1). This would help us better visualize the relation. Figure 18 shows that except for 'number of procedures other than lab tests', all features have different distribution for readmission and no-readmission. The distribution for readmission is shifted towards right. We explore in next section whether this variation is statistically significant or not for each of the features.

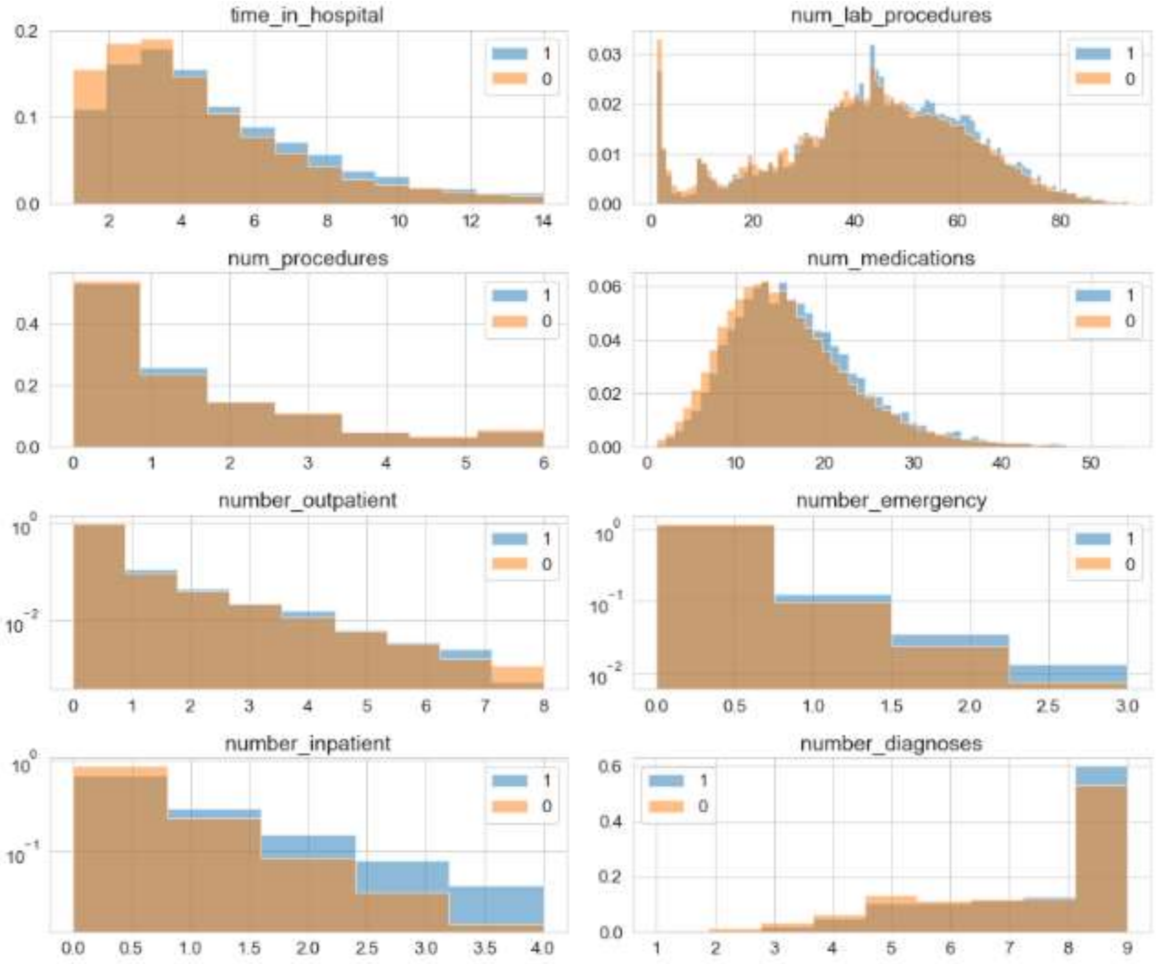
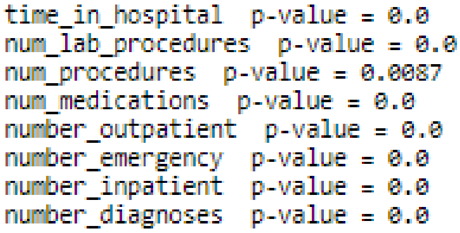


Figure 18: Distribution of patient encounter features between readmission and no-readmission

4.8. Frequentist and Inferential Statistics

In this section we will see which count variables have statistically significant relation with readmission rates. The Null hypothesis was that the mean of the feature samples for readmission and non-readmission are equal. The method used to test this hypothesis was Bootstrapped Hypothesis Test with difference of mean as test statistic and 5% significance level. P-value and confidence interval was calculated for all the features of Patient encounter group. More details is available in this **IPython notebook**. Below are the results of the test.



```
time_in_hospital  p-value = 0.0
num_lab_procedures  p-value = 0.0
num_procedures  p-value = 0.0087
num_medications  p-value = 0.0
number_outpatient  p-value = 0.0
number_emergency  p-value = 0.0
number_inpatient  p-value = 0.0
number_diagnoses  p-value = 0.0
```

It shows that all variables are statistically significant i.e. Null Hypothesis was rejected. This proves that distribution of patient encounter features are definitely different for readmitted patients and non-readmitted patients. Figure 19 shows the results graphically with confidence interval.

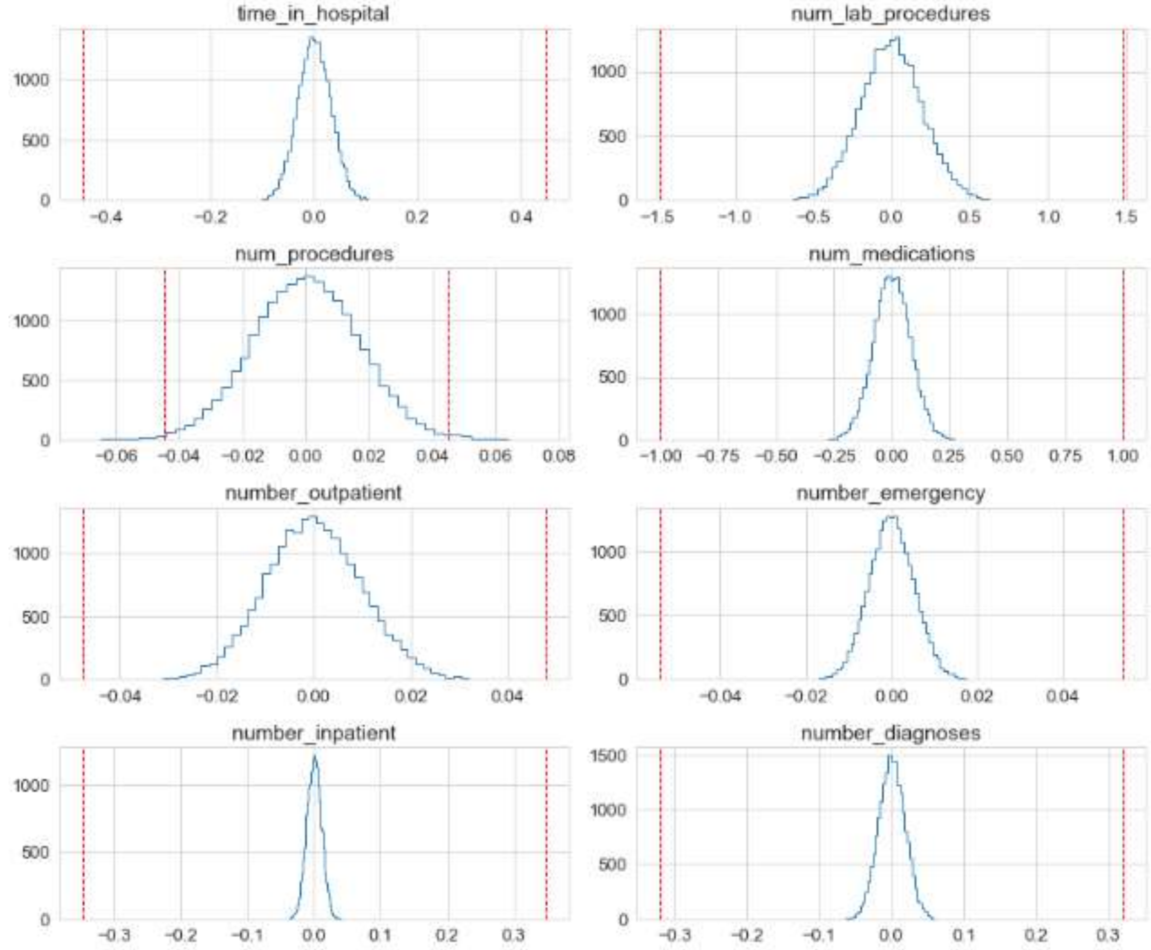
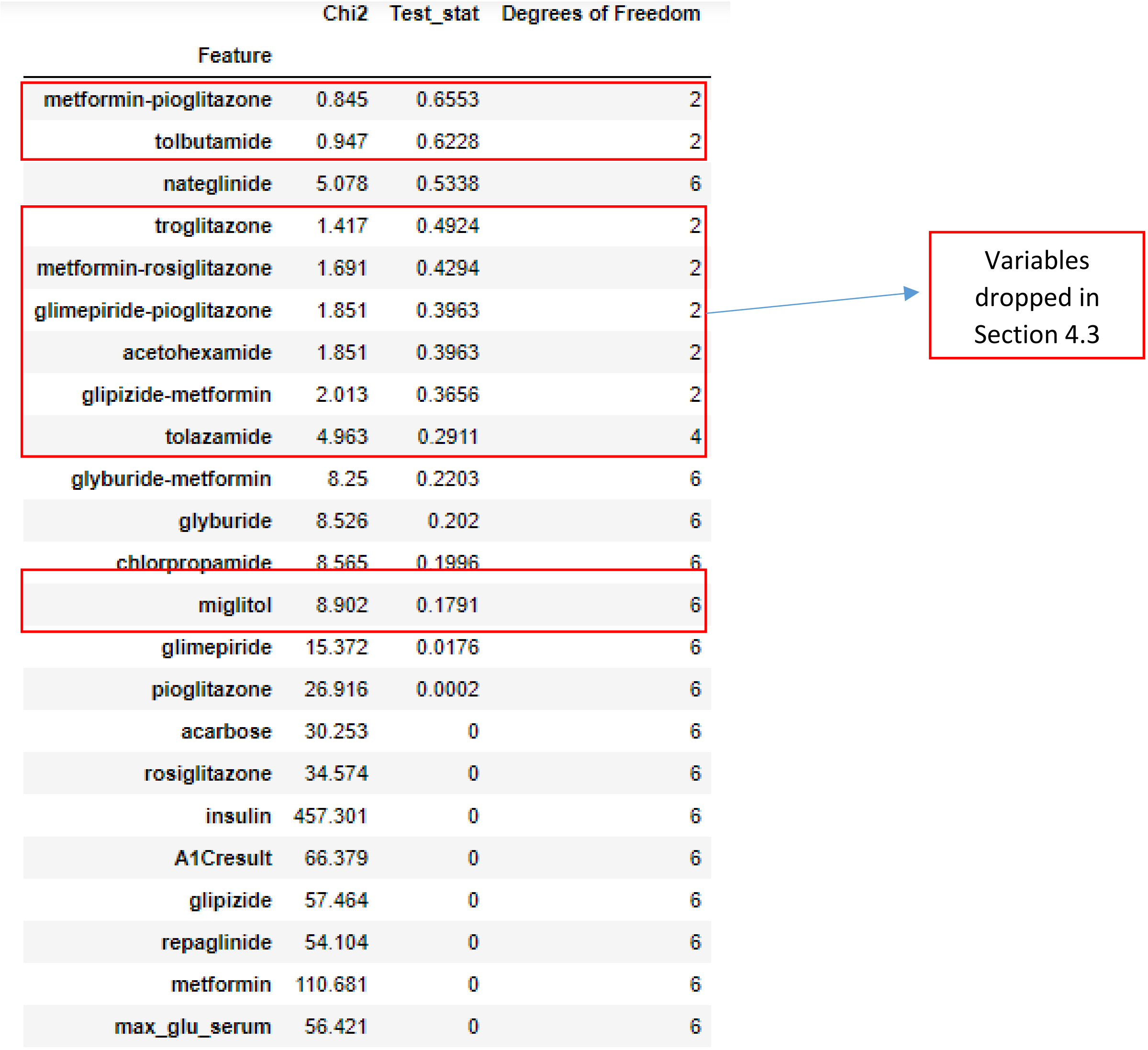


Figure 19: Results of Bootstrapped Hypothesis Tests

I also made use of Chi-square Independence Test to verify whether that columns of medicines that were dropped were insignificant and not related to Readmission rates. Chi-Square Test was used, since I was comparing two categorical variables i.e. medicine features and readmitted. The Null Hypothesis was that medicine features are independent of Readmission rates i.e. there is no relation between the two.

Below Table 3 shows that the 9 variables removed in Section 4.3 are all insignificant (Test_stat >0.05) and they passed our Null Hypothesis. This test also helped us identify more features that are not relevant. We can choose to remove those features too.


Variables dropped in Section 4.3

	Chi2	Test_stat	Degrees of Freedom
Feature			
metformin-pioglitazone	0.845	0.6553	2
tolbutamide	0.947	0.6228	2
nateglinide	5.078	0.5338	6
troglitazone	1.417	0.4924	2
metformin-rosiglitazone	1.691	0.4294	2
glimepiride-pioglitazone	1.851	0.3963	2
acetohexamide	1.851	0.3963	2
glipizide-metformin	2.013	0.3656	2
tolazamide	4.963	0.2911	4
glyburide-metformin	8.25	0.2203	6
glyburide	8.526	0.202	6
chlorpropamide	8.565	0.1996	6
miglitol	8.902	0.1791	6
glimepiride	15.372	0.0176	6
pioglitazone	26.916	0.0002	6
acarbose	30.253	0	6
rosiglitazone	34.574	0	6
insulin	457.301	0	6
A1Cresult	66.379	0	6
glipizide	57.464	0	6
repaglinide	54.104	0	6
metformin	110.681	0	6
max_glu_serum	56.421	0	6

Table 3: Results of Chi-Square Independence Test on Medicine Features

4.9. Multi-Collinearity

Common sense would say that a person spending more time in hospital is highly likely to have got high number of lab procedures performed and been prescribed large number of medicines (Figure 20). Similarly, 'change' i.e. change in medications column derives itself from the 24 features of medicine. Thus they two will be highly related. Multi-Collinearity was also observed between variables of Patient admission and discharge. More details about their proof is given in this **IPython Notebook**. This means that multi-collinearity is expected between various independent variables in the given dataset.

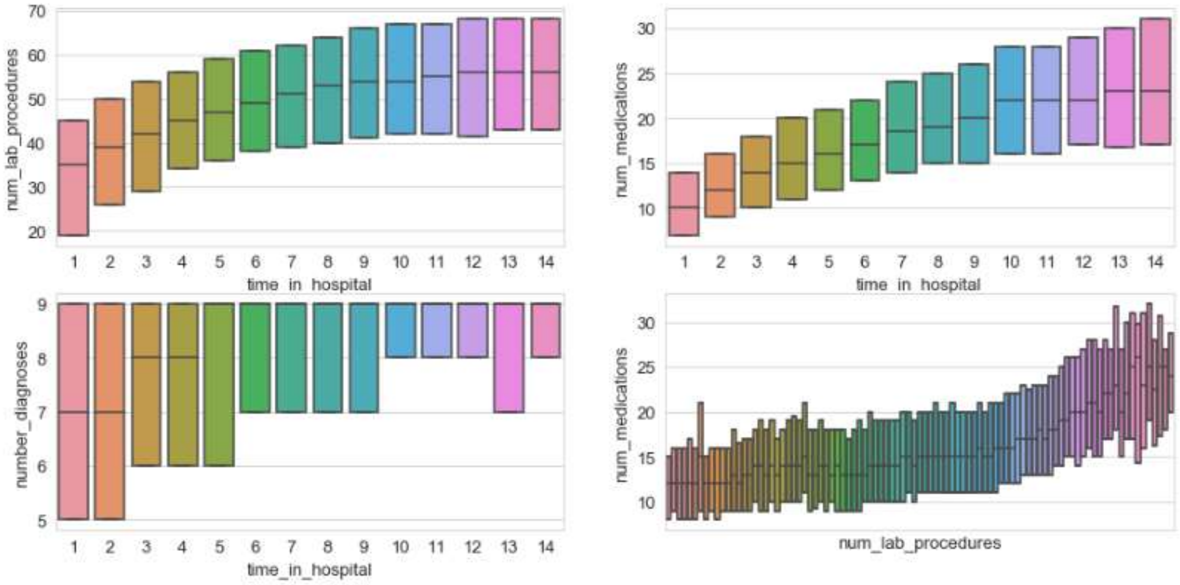


Figure 20: Multi-collinearity among features of Patient encounter group