Predicting Readmissions of Diabetes Patients

GROUP - 3

FINAL REPORT

**Mentored by:**

MRS. **Anjana Agrawal**

**Submitted by:**

Shekhar

Narendran  
Pratibha

ABSTRACT: Healthcare is a highly evolving domain. In such an evolving area, new trends appear daily and older standards become out of fashion very fast. The paper presents a methodology for seizing knowledge from rich information sources of the healthcare domain and processing it in order to improve the domain understanding for organizations and individuals. To achieve that, we suggest the development of technologies for systematically & automatically updating, dynamic knowledge maps of healthcare domain. These maps will act as advanced business intelligence support for market, products & technology watch and will represent a novel knowledge management practice (technology evolution assessment). The methodology may lead to a knowledge management platform for the intelligent monitoring, planning, forecasting of technological evolution in highly dynamic domains. Doctors, researchers and healthcare organizations that utilise the platform will achieve state-of-the-art awareness in research and development and will increase competitiveness in their domain. The dataset is to obtain a measurement of HbA1c for patients with diabetics is a useful predictor of readmission rates which may prove valuable in the development of strategies to reduce readmission rates and costs for the care of individuals with diabetes

1. INTRODUCTION:

As the healthcare system moves toward value-based care, CMS has created many programs to improve the quality of care of patients. One of these programs is called the Hospital Readmission Reduction Program ([HRRP](https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction-Program.html)), which reduces reimbursement to hospitals with above average readmissions. For those hospitals which are currently penalized under this program, one solution is to create interventions to provide additional assistance to patients with increased risk of readmission. But how do we identify these patients? We can use predictive modeling from data science to help prioritize patients.

One patient population that is at increased risk of hospitalization and readmission is that of diabetes. Diabetes is a medical condition that affects approximately 1 in 10 patients in the United States. According to Ostling et al, patients with diabetes have almost double the chance of being hospitalized than the general population ([Ostling et al 2017](https://clindiabetesendo.biomedcentral.com/articles/10.1186/s40842-016-0040-x" \t "_blank)). Therefore, in this article, I will focus on predicting hospital readmission for patients with diabetes.

In this project I will demonstrate how to build a model predicting readmission in Python using the following steps

* data exploration
* feature engineering
* building training/validation/test samples
* model selection
* model evaluation

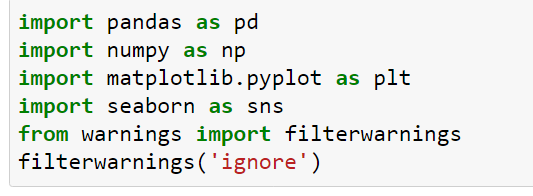
**Project Definition**

Predict if a patient with diabetes will be readmitted to the hospital within 30 days.

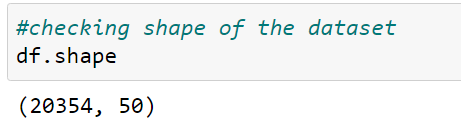
# Data Exploration

The data that is used in this project originally comes from the UCI machine learning repository ([link](https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008)). The data consists of over 100000 hospital admissions from patients with diabetes from 130 US hospitals between 1999–2008.

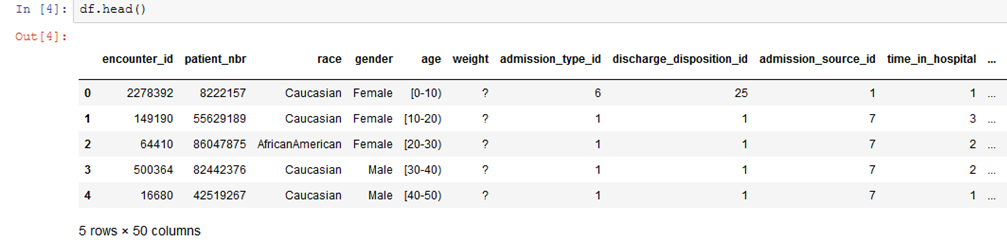
In this project, we will utilize Python to build the predictive model. Let’s begin by loading the data and exploring some of the columns.





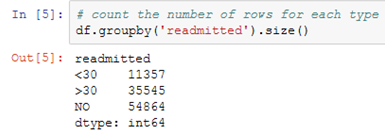


From briefly, looking through the data columns, we can see there are some identification columns, some numerical columns, and some categorical (free-text) columns. These columns will be described in more detail below.

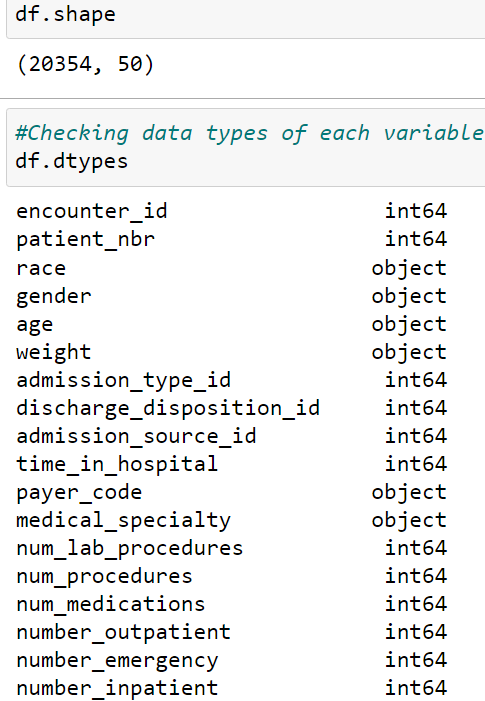


There is some missing data that are represented with a question mark (?). We will deal with this in the feature engineering section.

The most important column here is readmitted, which tells us if a patient was hospitalized within 30 days, greater than 30 days or not readmitted.

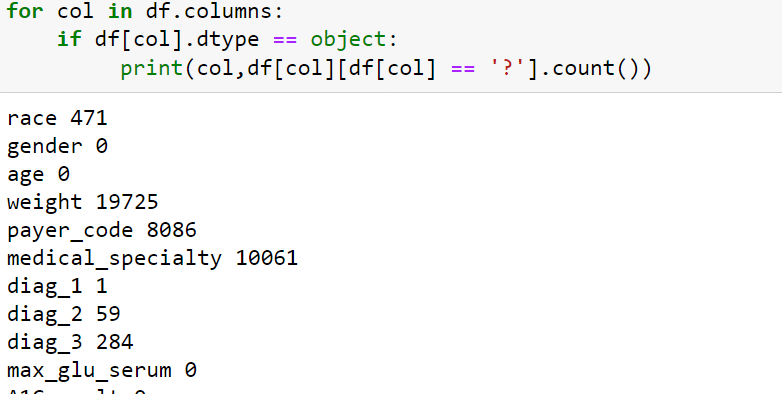


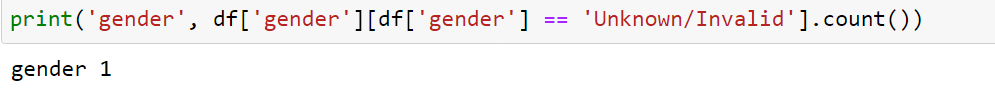
Checking data types of each variable



-Checking for missing values in dataset

-In the dataset missing values are represented as '?' sign





#Dealing with Missing Values

#Variable weight contains approximate 98% of the missing values so there is no significance in filling those missing values

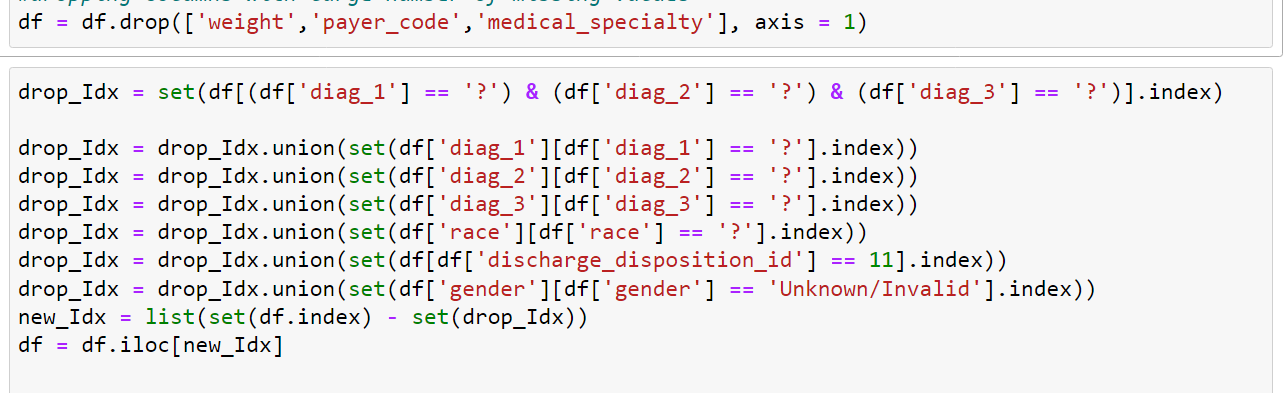
#so we decided to drop these variables.

#Variable Payer code and medical specialty contains approximate 40% missing values

#so we also dropped these variables. Variables race, diag\_1, diag\_2, diag\_3 and gender

#contains very less missing values as compared to other attributes which we dropped

#so for these attributes we also decided to drop those where missing values contains.



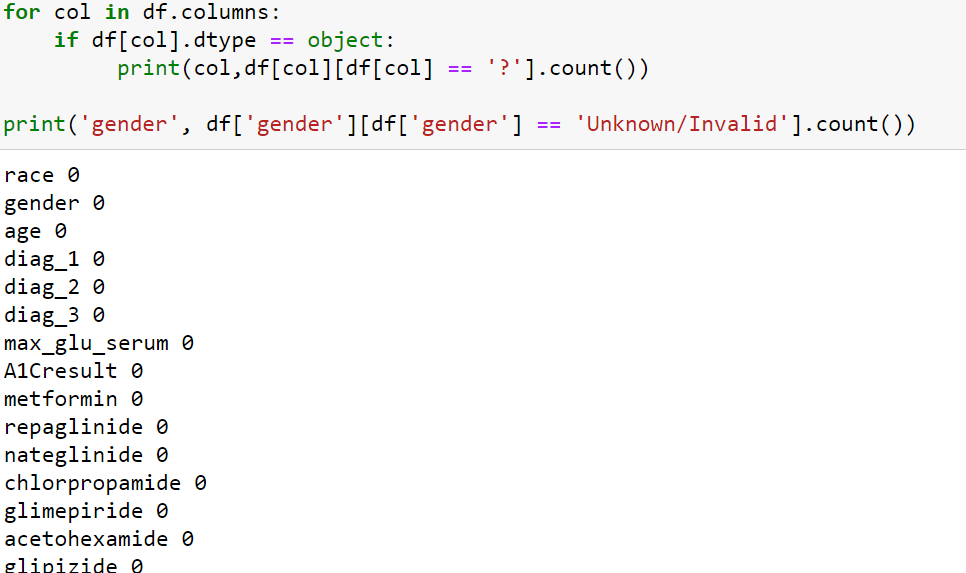
#variables (drugs named citoglipton and examide), all records have the same value.

#So essentially these cannot provide any interpretive or discriminatory information

#for predicting readmission so we decided to drop these two variables



#Checking for missing values in the data



###Feature engineering

#This is highly subjective, and partly depends on a knowledge of health care services, and making sense of the potential relationships between features.

#There are perhaps thousands of ways to try here. We tried some

#Service utilization: The data contains variables for number of inpatient (admissions), emergency room visits and outpatient visits for a given patient in the previous one year.

#These are (crude) measures of how much hospital/clinic services a person has used in the past year. We added these three to create a new variable called service utilization (see figure below).

#The idea was to see which version gives us better results.

#Granted, we did not apply any special weighting to the three ingredients of service utilization but we wanted to try something simple at this stage.



#Number of medication changes:

#The dataset contains 23 features for 23 drugs (or combos) which indicate for each of these, whether a change in that medication was made or not during the current hospital stay of patient.

#Medication change for diabetics upon admission has been shown by previous research to be associated with lower readmission rates.

#We decided to count how many changes were made in total for each patient, and declared that a new feature.

# re-encoding admission type, discharge type and admission source into fewer categories

#The reasoning here was to both simplify the model and possibly discover a relationship with number of changes regardless of which drug was changed.

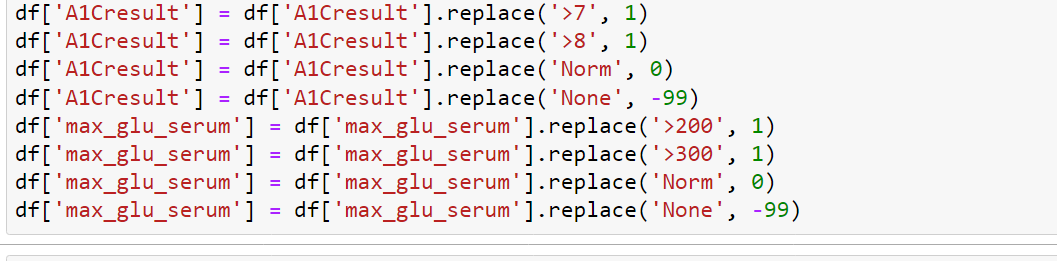
###Encoding some variables:

#The original dataset used string values for gender, race, medication change, and each of the 23 drugs used.

#To better fit those variables into our model, we interpret the variables to numeric binary variables to reflect their nature.

#For example, we encoded the “ medication change ” feature from “No” (no change) and “Ch” (changed) into 0 and 1.

##We also reduced both A1C test result and Glucose serum test result into categories of Normal, Abnormal and Not tested.



#Dealing with age: There are different ways to deal with this.

#The dataset only gives us age as 10 year categories,

#so we don’t know the exact age of each patient.

#The previous study on this dataset used age categories as nominal variables,

#but we wanted to be able to see the effect of increasing age on readmission,

#even if in a crude way. To do that, we assume that age of the patient on average lies

#at the midpoint of the age category. For example, if the patient’s age category is 20–30 years,

#then we assume the age = 25 years. So we converted age categories to midpoints,

#resulting in a numeric variable:

#Collapsing of Multiple Encounters for same patient Some patients in the dataset had more

#than one encounter. We could not count them as independent encounters because

#that bias the results towards those patients who had multiple encounters.

#Thus we tried multiple techniques to collapse and consolidate multiple encounters for same patient such as:

#Considering more than 2 readmissions across multiple encounters as readmission for collapsed record.

#Considering average stay at hospital across multiple encounters.

#Considering the percentage of the medication changes across multiple encounters

#Considering the total number of the encounters to replace the encounter unique ID

#Considering the combination of diagnoses across multiple encounters as a list However, taking the features such as “diagnosis”,

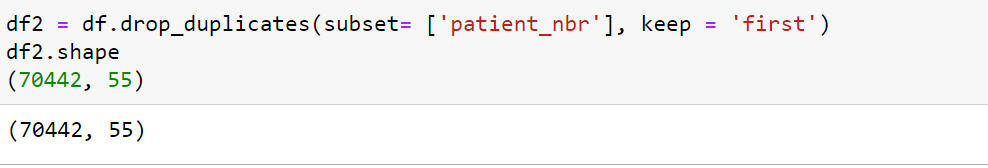
#for instance, we did not find it not meaningful to combine multiple categorical values into

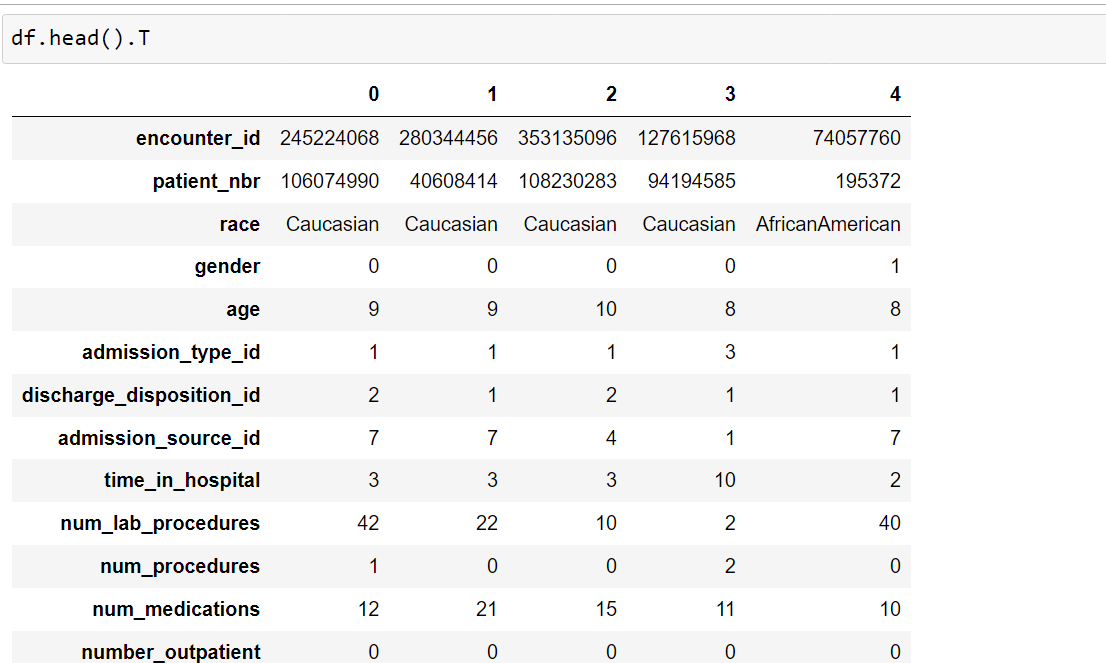
#an array for building data model.

#We then considered first encounter and last encounter separately as possible representations

#of multiple encounters. However, last encounters gave extremely imbalanced data for readmissions

#(96/4 Readmissions vs No Readmissions) and thus, we decided to use first encounters of patients with multiple encounters.





#Encoding the outcome variable:

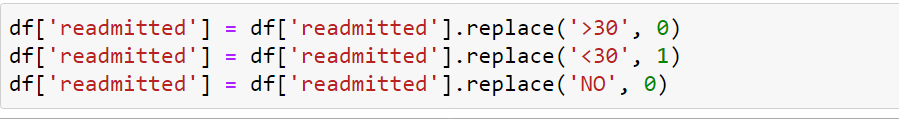
#The outcome we are looking at is whether the patient gets

#readmitted to the hospital within 30 days or not.

#The variable actually has < 30, > 30 and No Readmission categories.

#To reduce our problem to a binary classification,

#we combined the readmission after 30 days and no readmission into a single category:



# Categorization of diagnoses:

# The dataset contained up to three diagnoses for a given patient (primary, secondary and additional). However, each of

# these had 700–900 unique ICD codes and it is extremely difficult to include them in the model and interpret meaningfully.

# Therefore, we collapsed these diagnosis codes into 9 disease categories in an almost similar fashion to that done in the

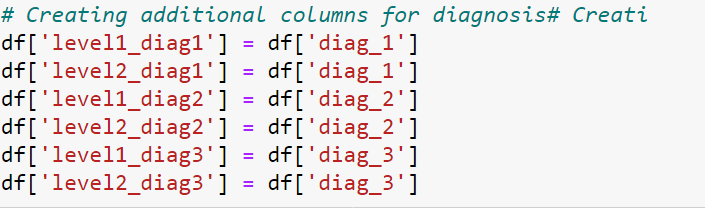
# original publication using this dataset. These 9 categories include Circulatory, Respiratory, Digestive, Diabetes,

# Injury, Musculoskeletal, Genitourinary, Neoplasms, and Others. Although we did this for primary, secondary and additional

# diagnoses, we eventually decided to use only the primary diagnosis in our model. Doing this in python was slightly

# cumbersome because, well, we are mapping the disease codes to certain category names. Below code should demonstrate this

# easily.



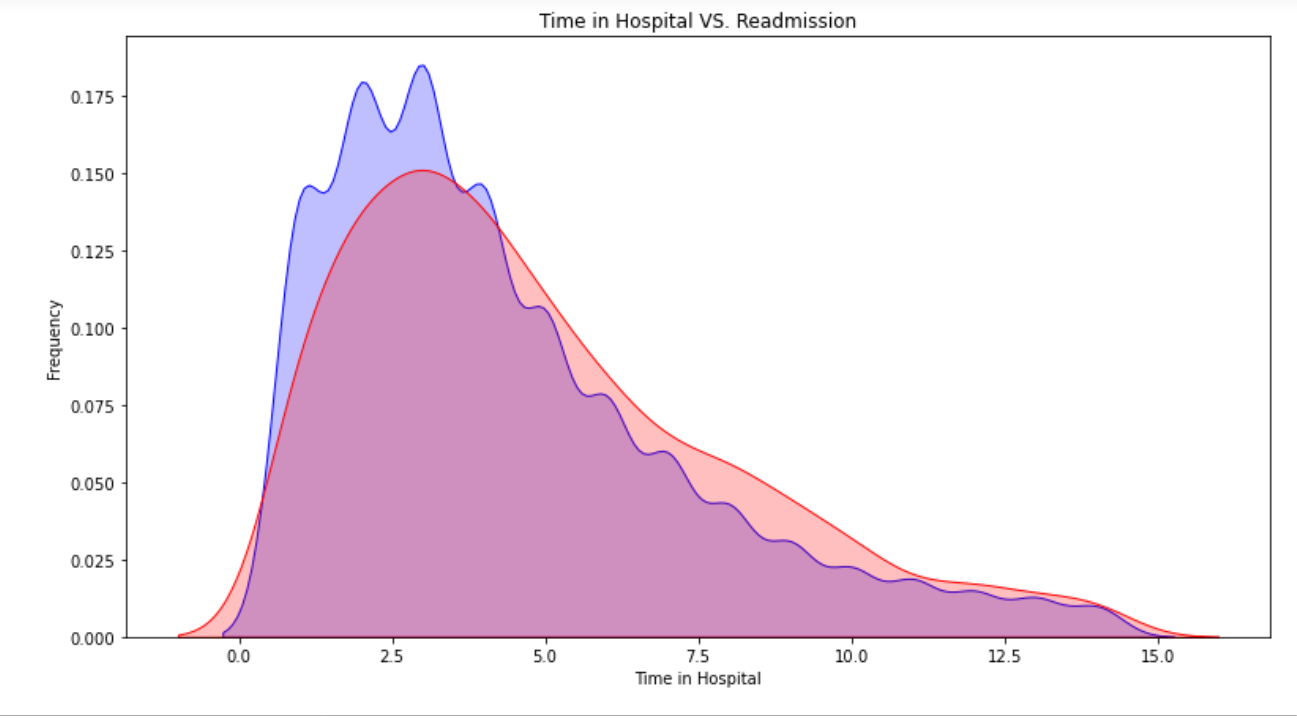
# Data Visulation

# #Our target variable is imbalance.

# #Number of readmitted patient are quite less as compared to Not readmitted.

# 

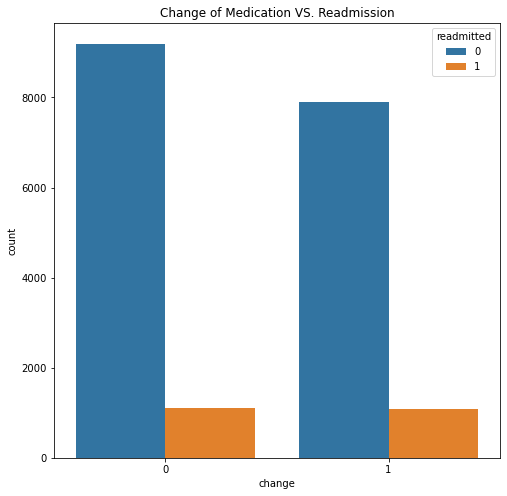
###Time in Hospital and Readmission



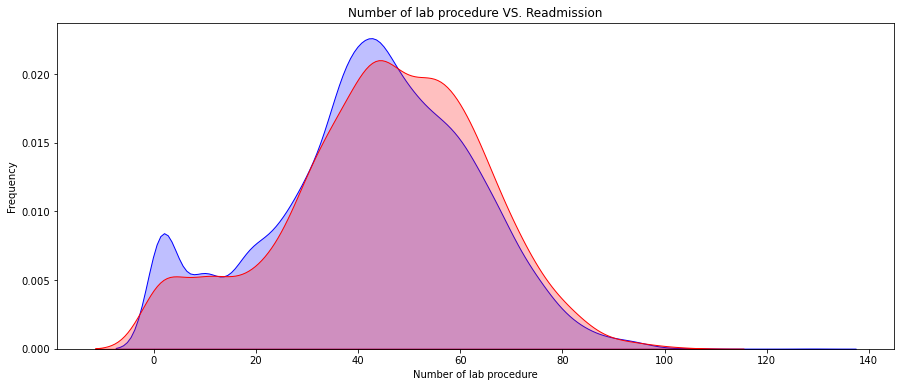
----#Change of Medication and Readmission

#Change = 1

#No Change = 0



##Number of lab procedure and Readmission



We have shown only few diagrams.

# Pre-Modeling Data Preprocessing

# 

# 

# 

# #Number of medication used:

# #Another possibly related factor could be the total number of medications used by the patient

# #(which may indicate severity of their condition and/or the intensity of care).

# #So we created another feature by counting the medications used during the encounter

# #(keys variable in code below is continued from above):

# 

# Modeling

# 

# Logistic Regression

# 

# 

# ##Since our target variable is having class imbalance problem, So will use SMOTE technique to resolve it

# 

# 

# Decision Tree

# 

# 

# 

# 

# 

# ### in reference with the above chart we can infer that few parameterd like (time\_in\_hospital,num\_medications,

# #num\_procedures,number\_inpatient\_log1p) are most important predictor variable that play a vital role while predicting the

# #the target variable.

# Random Forest

# 

# 

# 

# 

# Model Comparision

# 

# Conclusion

# Through this project, we created a machine learning model that is able to predict the patients with diabetes with highest risk of being readmitted within 30 days. The best model was a gradient boosting classifier with optimized hyperparameters. The model was able to catch 58% of the readmissions and is about 1.5 times better than just randomly picking patients