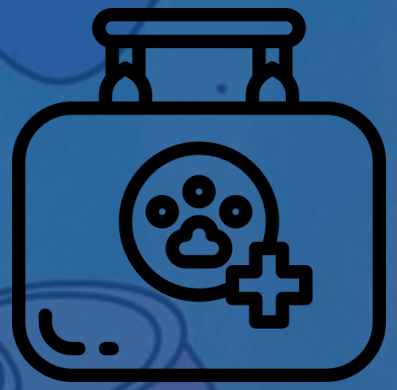


Diabetic Patient Hospital Readmission Prediction

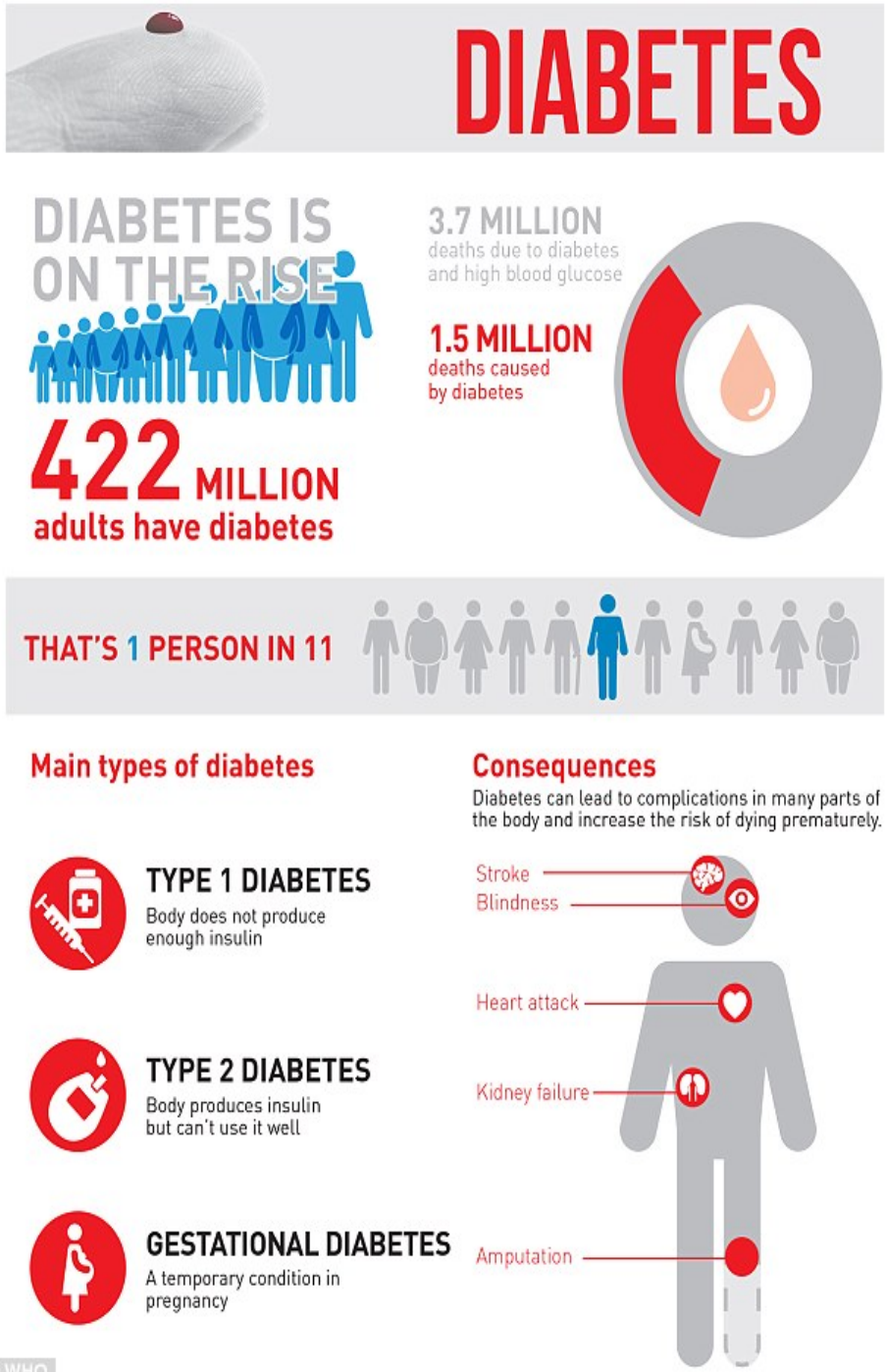
Wanxin Ye

M.S in Data Science at Brown University

2019.10.21

Github:

https://github.com/leavetina321/ML_diabetes_readmission.git



Introduction

Why we care ?

Diabetes problem has affected a great population in the US. According to the report released by WHO, one people in 11 is diagnosed as having diabetes. Meanwhile, people who is diagnosed as diabetes are more likely to be readmitted to hospital stated in the report by Ostling et al 2017.

Hospital readmission has also arose some discussions since it will add more financial burden to patients. To tackle with the excessive readmission problem, hospitals are requested to improve the qualities of the health care services, one of which concerns with reducing readmission rate. Hospital Readmission Reduction Program (HRRP) is created to reduce reimbursement to hospitals with excessive readmissions.

In order to help with reduce hospital readmission rate, this machine learning project will target at predicting whether the patient will be readmitted within 30 days.

About the Dataset

SOURCE

UCI machine learning repository.
This dataset contained clinical care at 130 US hospitals and integrated delivery networks from 1999 to 2008.

SIZE

There were 101,766 observations (71,518 unique patient Id).

Target Variables

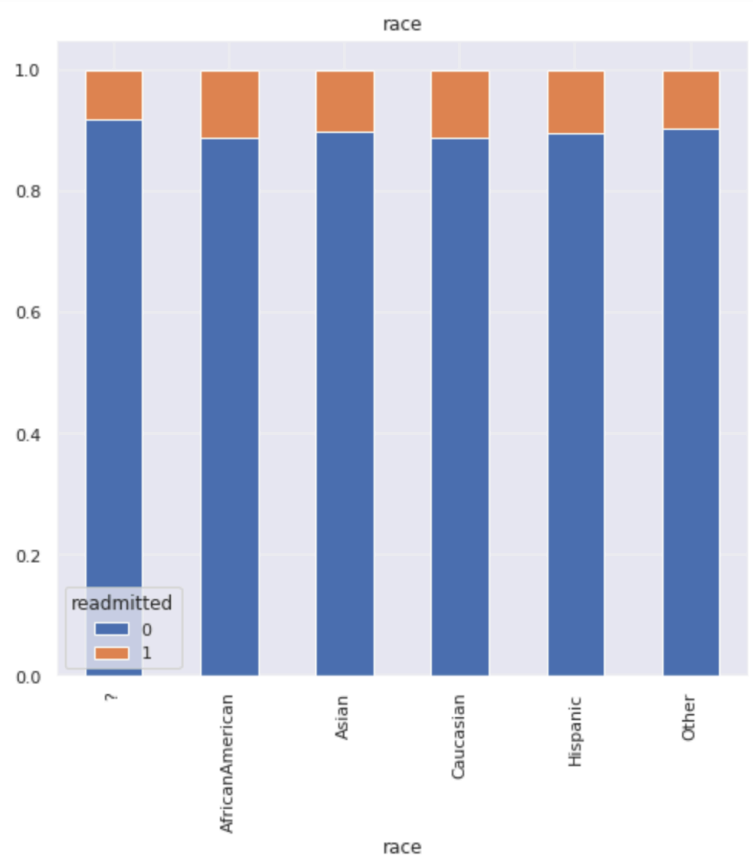
Class 0 : 88%

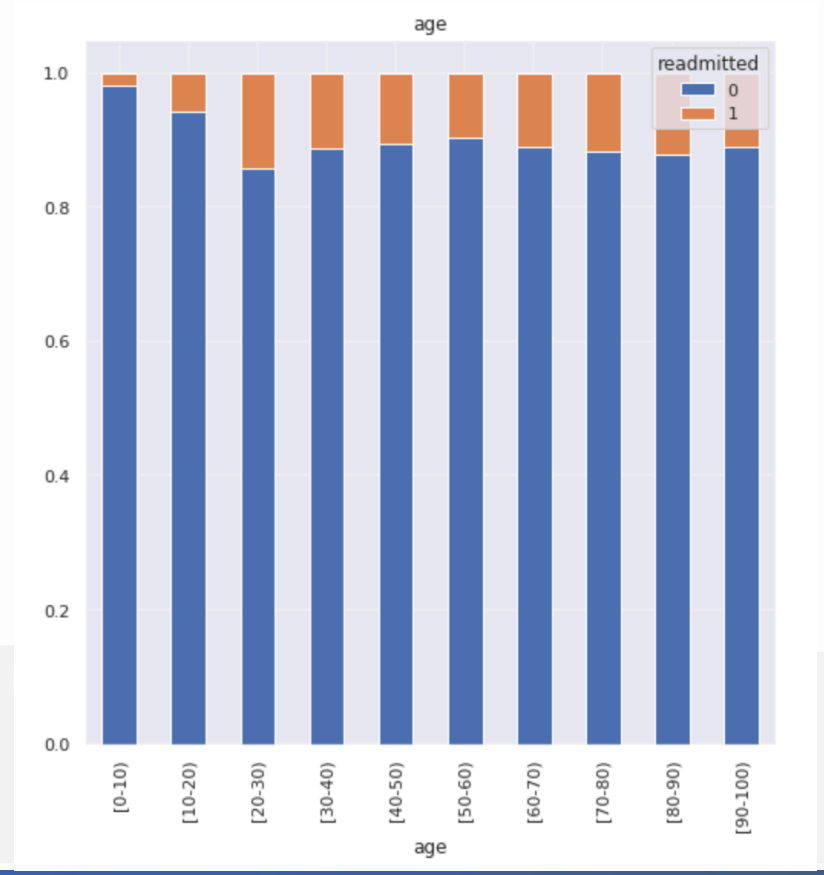
Class 1: 12%

FEATURE

It included 50 features (13 numerical and 37 categorical features)
The features include demographical features such as age and race, medical history such as number of lab procedures, number of medications, prescription such as 'chlorpropamide', 'glimepiride'

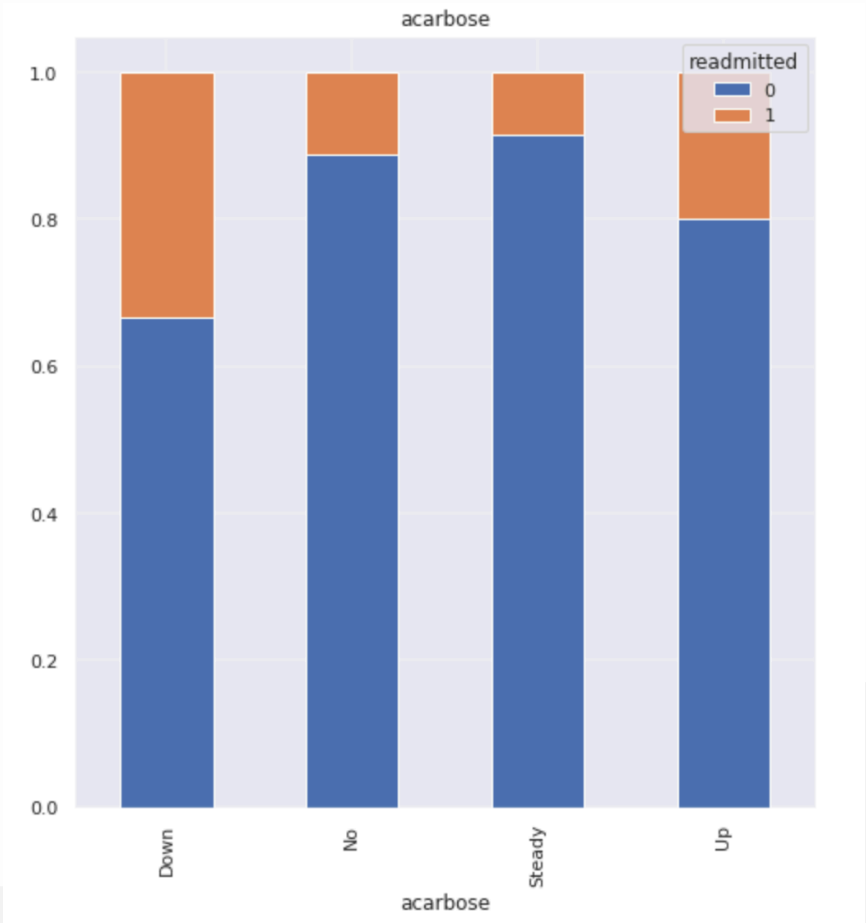
Exploratory Data Analysis

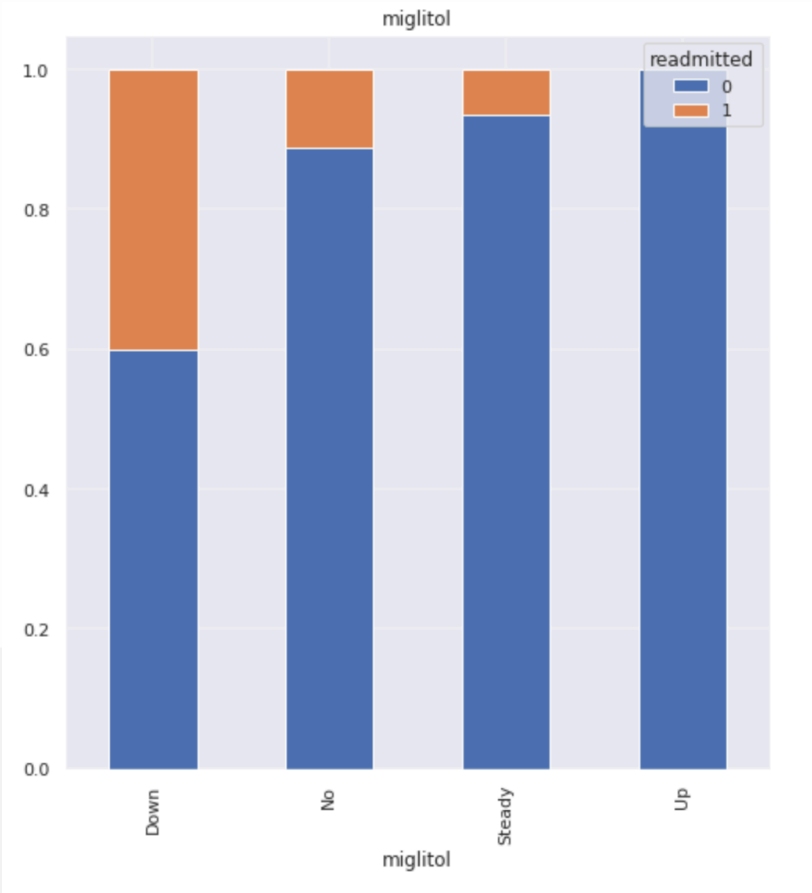




Race seems to be irrelevant as to whether the patient would be readmitted. Surprisingly, patient's readmission rate starts to go up after 20s.

Exploratory Data Analysis

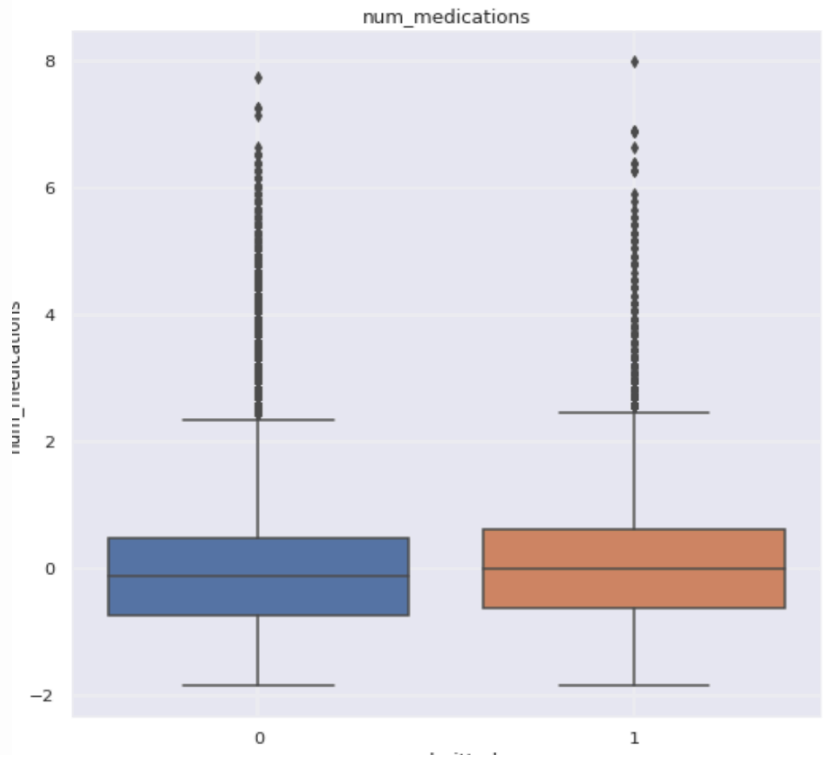


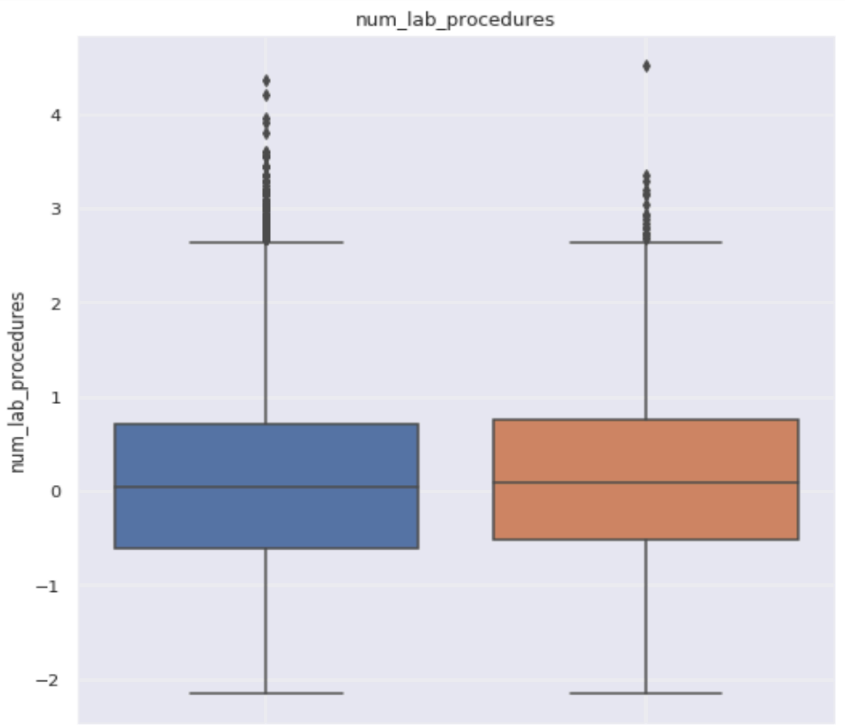


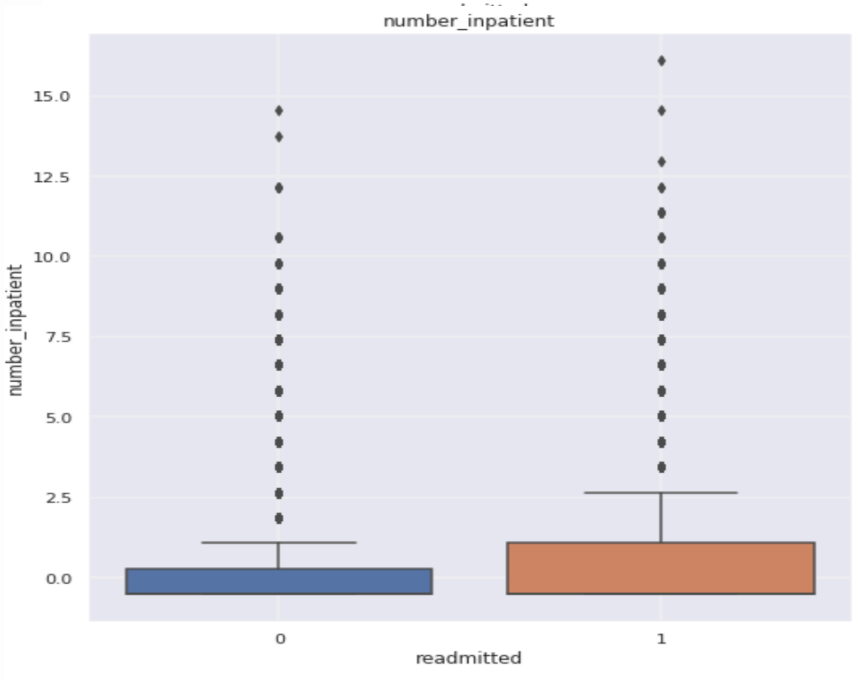
The level of medication doses also indicates some probability of readmission. For example, the lower the level of acarbose and miglitol doses, the higher chance the patient will be readmitted within 30 days

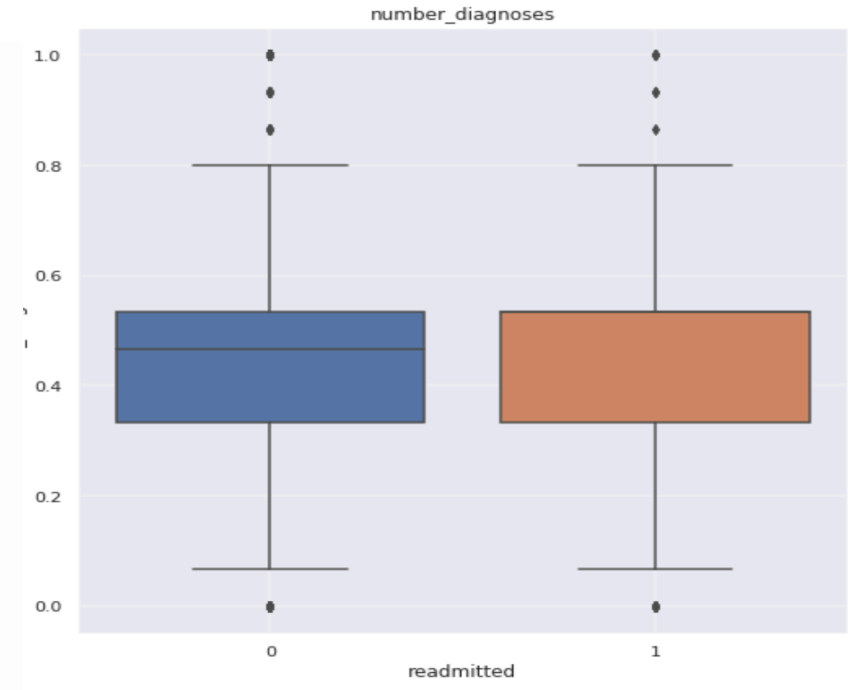
Exploratory Data Analysis

No major difference between number of medication, number of lab procedures, number of inpatient and number of diagnoses

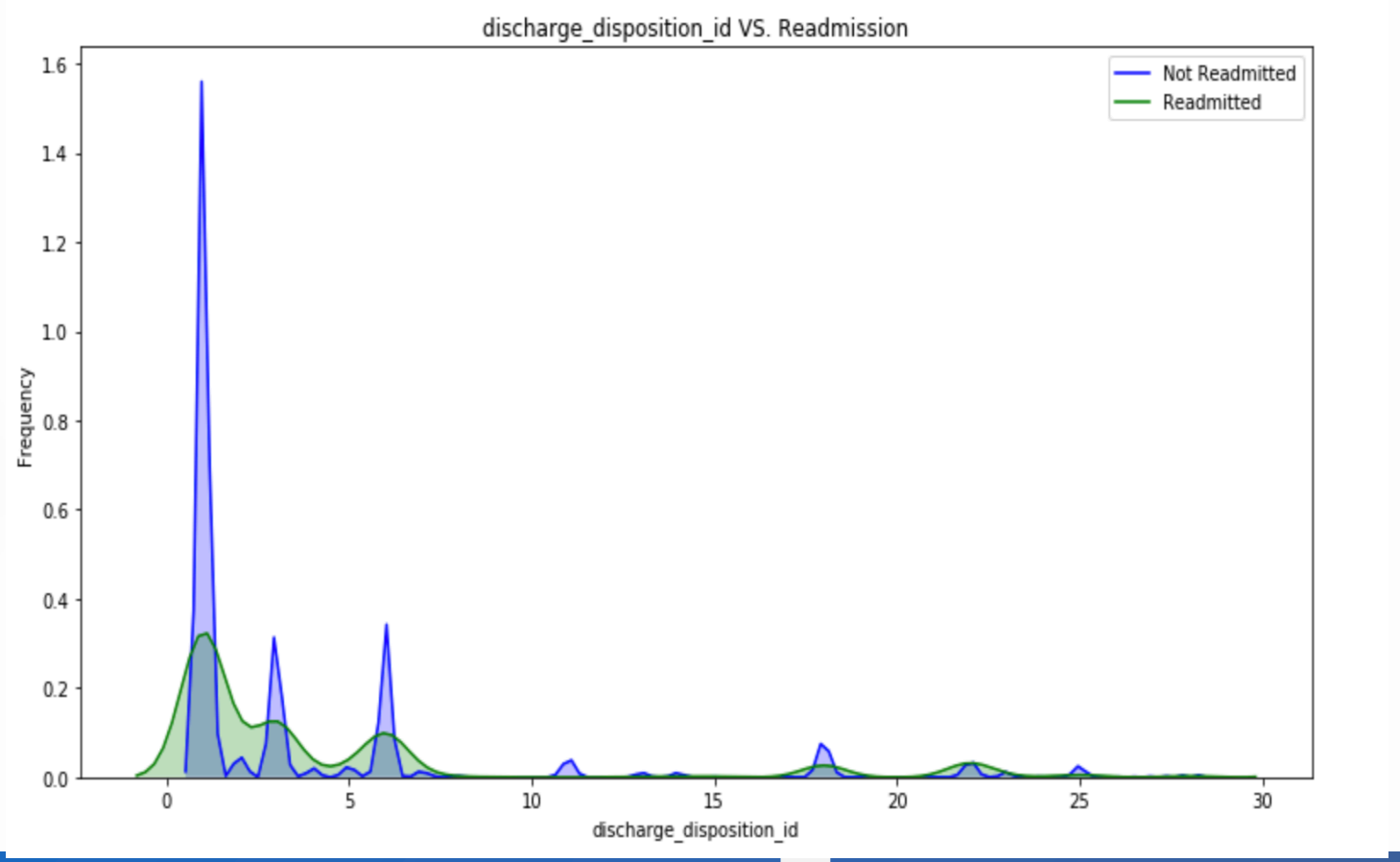






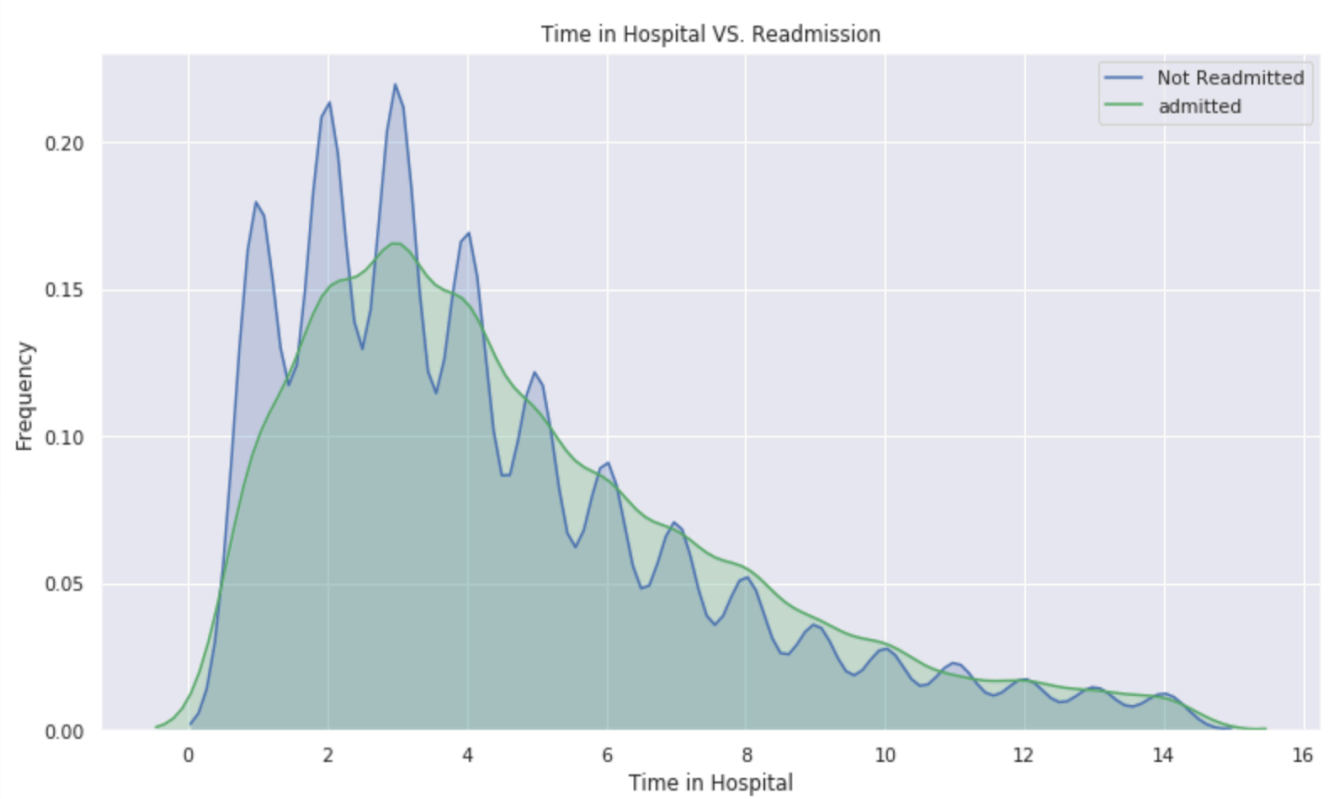


Exploratory Data Analysis



There's a clear distinction between not readmitted and readmitted patient in terms of discharge_disposition_id. We believe that discharge_disposition_id plays a significant role in the prediction

Exploratory Data Analysis



The trend shows the longer the time spent in hospital, more frequently the patient will be readmitted to hospital

Model and methods

Data process pipelines:
The Data Process pipeline includes the function to transform certain column values based on medical description , the function to handle missing values and the function to clean the data that's not the first readmission to the hospital.

Machine Learning Pipelines：
The pipeline will do the train test split and kfold cross validation. For most the case, we split the data into 80% for training and 20% for testing. After splitting the training set and test set, we apply one-hot encoder, standard scaler to the transform the data sets and input the transformed dataset to the model which we tried to tune the hyperparameters

Logistic Regression(L1) : Tune hyperparameters : {alpha ： np.logspace(-5,5,num = 20)}
After getting the best alpha based on the model that returned the best f1 score when we loop through the alpha values between np.logspace(-5,5) , we use the new model to predict the test dataset. Having the new model, we decided to use different random state to get the best outcome. Since it's a classification problem and it's an imbalanced dataset, I used f1 score, precision, recall score and auc_roc as the metrics to evaluate the results.

Random Forest: Tune hyperparameters: grid_para_forest = [{ "n_estimators": (300,500), "max_depth":(30,50),
"min_samples_leaf":(1,5), "min_samples_split":(2,7), "random_state": [42]}]
Before applying the rando forest model, pipeline will go through the same process to do the train test split. Since we have more than 2 hyperparameters to tune, I used grid search to help me find the best hyperparameters for my models and also do the cross validation. Likewise, f1 score, precision, recall score and auc_roc are the evaluation metrics.

Model and methods

Data process pipelines:
The Data Process pipeline includes the function to transform certain column values based on medical description , the function to handle missing values and the function to clean the data that’s not the first readmission to the hospital.

Machine Learning Pipelines：
The pipeline will do the train test split and kfold cross validation. For most the case, we split the data into 80% for training and 20% for testing. After splitting the training set and test set, we apply one-hot encoder, standard scaler to the transform the data sets and input the transformed dataset to the model which we tried to tune the hyperparameters

XgBoost: Tune hyperparameters : "eta": (0.001,0.4), "gamma":(0,30), "max_depth":(2,100)}
We also use the grid search to find the optimal hyperparameters for the Xgboost model. What makes a bit difference is that when we try to select the best hyperparameters, we use scoring=auc_roc as the evaluation. F1_score, precision score and recall score are also applied to evaluate the model performance.

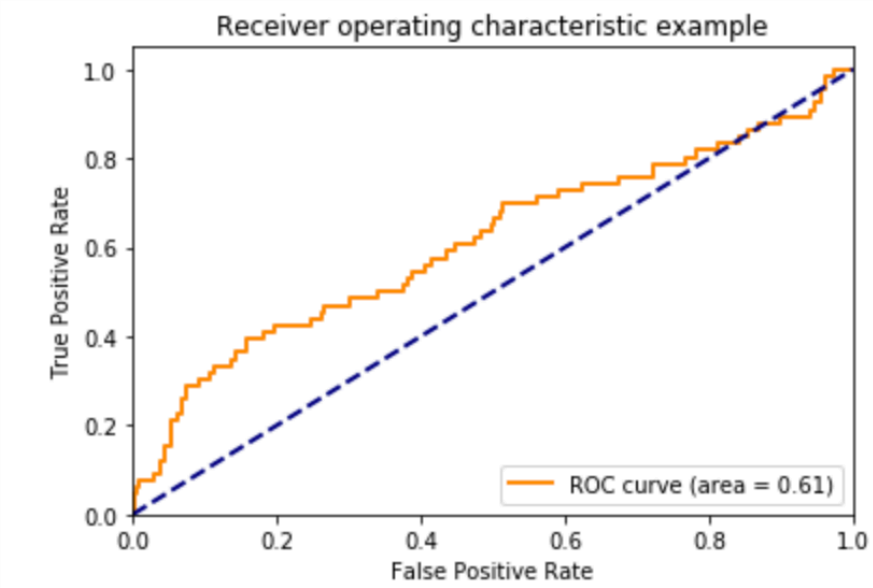
Beside model selection and model fitting, we also tried to get the model importance from both random forest and xgboost model.

Results Comparision

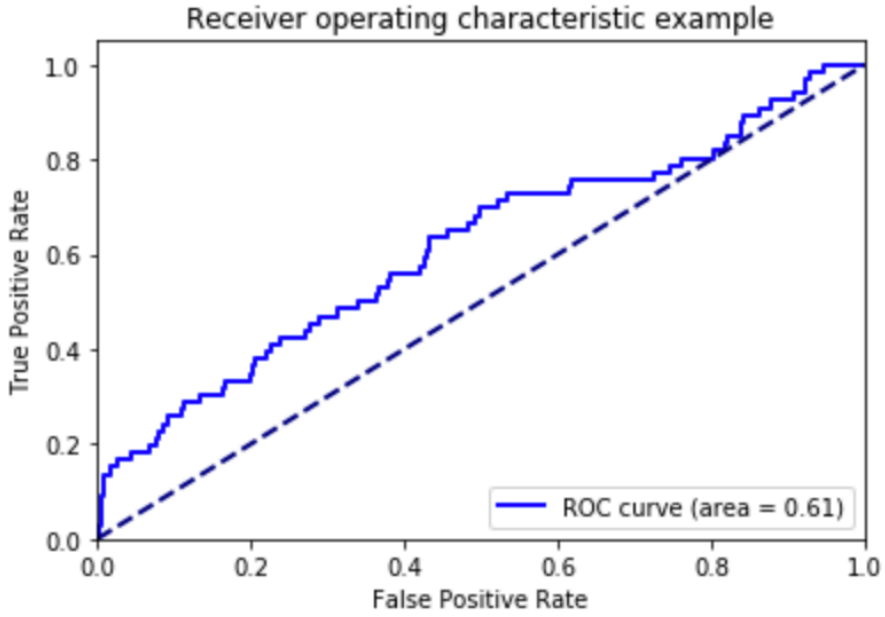
	class	Precision	Recall	F1_score
Logistic	0	0.91	0.99	0.95
Logistic	1	0.17	0.01	0.02
Random Forest	0	0.9248	1	0.9609
Random Forest	1	0	0	0
Xgboost	0	0.92588	1	0.9615
Xgboost	1	1	0.0151	0.02985

The baseline model is 0.9. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations . Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes . F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account We can see from the result that Xgboost performs the best in terms of precision, recall and f1_score. It is expected since Xgboost will always try to fit the difference from the last tree.

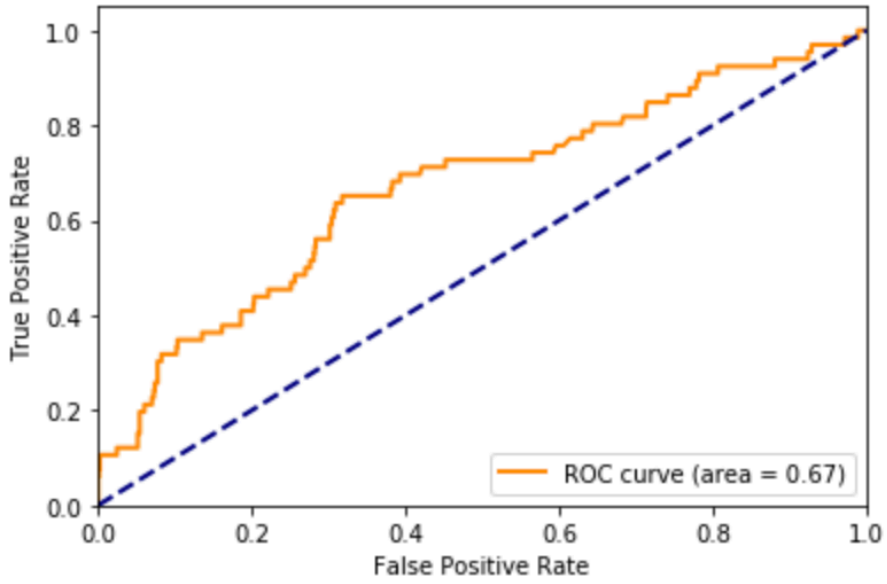
Results



Logistic regression ROC



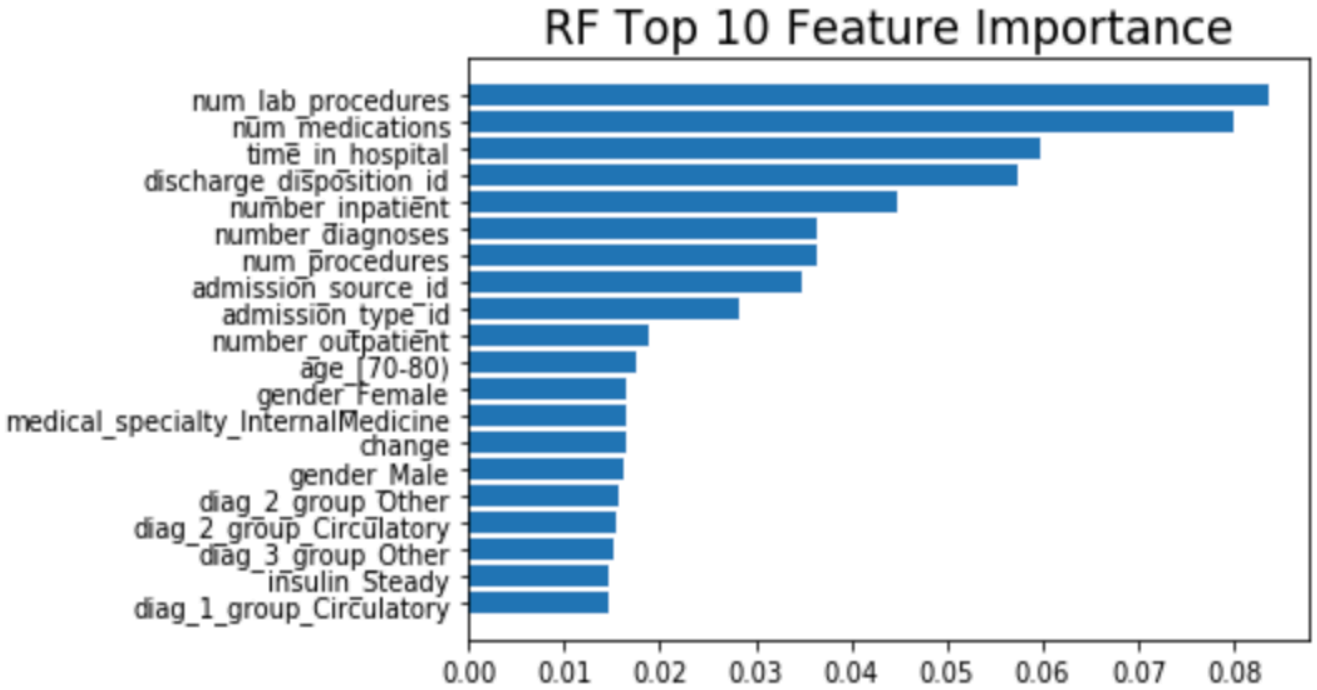
Random Forest ROC

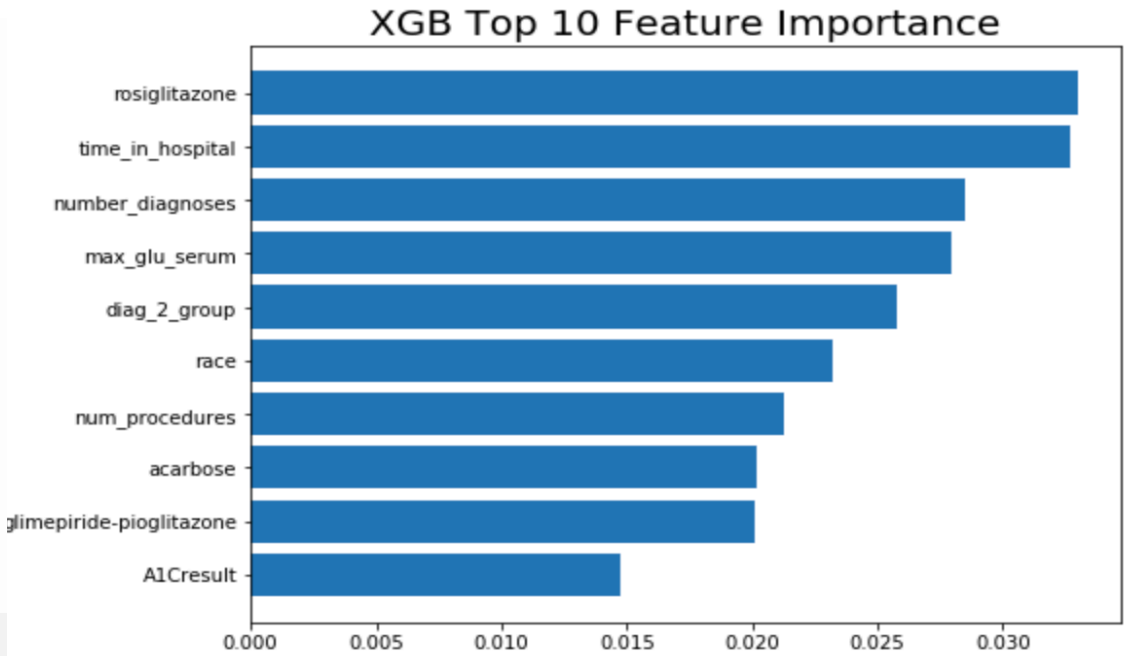


Xgboost ROC

The ROC are for both logistic regression and random forest are both 0.61 and for Xgboost is 0.67. This ROC curve also shows the model performance based on the True Positive rate and False Positive rate.

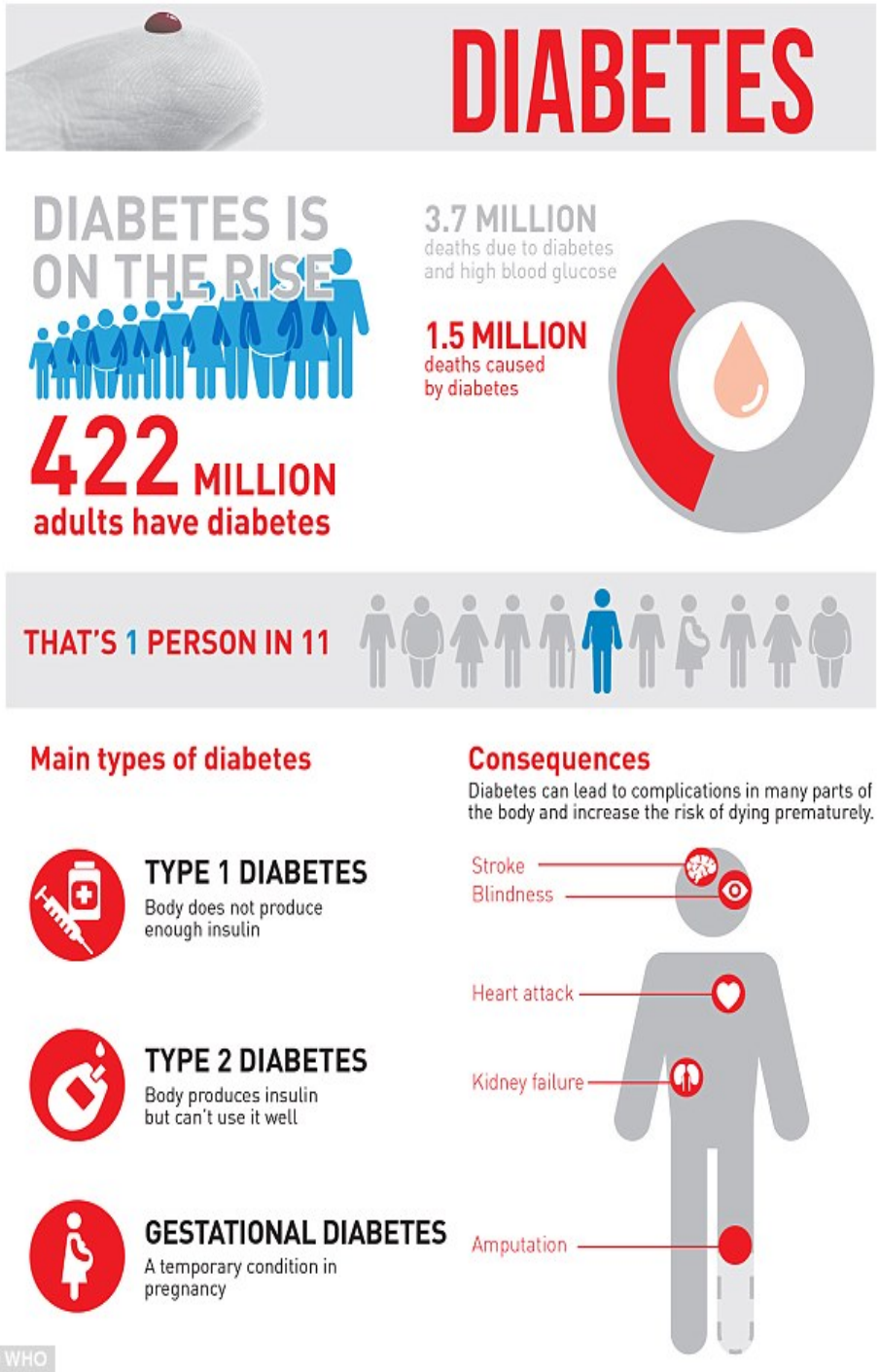
Results





Comparing the feature importance returned by both random forest and Xgboost model, we can see some of the variables play a significant role in our model performance. For example, number of procedures, time in hospital, number of diagnoses and the diagnosis 2 provided by the doctors.

The number of procedures indicate how many test has been run on the patient when the patient returns to the hospital. More procedures also means the situation of the patient is getting severer, which indicate a higher chance to be admitted to the hospital again. Time in hospital represents the days between admission and discharge. Longer of the stay suggests the worse the patient's condition and this also implies higher chance to get admitted.



Outlook

Health problem has been a hard issue to study but it also has a great significance if we can use machine learning to shad light for future medical research. For the future model performance, I would probably try to do the over sampling techniques to tackle with the imbalanced dataset. This could probably avoid some unintended bias. Furthermore, due to time constraint and computing power constraint, the model was not able to get the best of the hyperparameter. If possible, I will also try to do more feature engineering to select the variables which could explain the most variance of the model.

Reference

Data sources: UCI Diabetes 130-US hospitals for years 1999-2008
Data Set https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#

Publication :Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

Data Preprocessing

Originally:
101766 rows, 50 columns

After preprocessing:
101766 rows,
158 columns

Variables	Action	Reason
Weight/ examide/citoglipton	drop	Weight has 97% missingness and examide/citoglipton only have 1 value
diag_1/diag_2/diag_3	transform values	Transform the diagnosis data to string according to diagnosis table
Age/'metformin', 'repaglinide', 'nateglinide', 'chlorpropamide', …….'metformin-pioglitazone'	Apply Ordinal Encoder	Age might be a factor that shows the probability of getting diabetes/ The medication indicates patient's current severity of diabetes
num_procedures/numberdiagnoses	Apply minmax Scaler	Number of procedures and number of diagnoses both fall in certain range
time_in_hospital,num_lab_procedures,num_medications,number_outpatient,number_emergency,number_inpatient	Apply Standard Scaler	Standard these continuous variables in order to prevent large variance
Readmitted	Apply Label Encoder	If patient were readmitted within 30 days then '1' else '0'
'race', 'gender','medical_specialty','diag_1_group','diag_2_group','diag_3_group','max_glu_serum','A1C result'	Apply Onehot Encoder	The variable dimensions are necessary to be treated as single features