Capstone Project in Analytics:

Hospital Readmission Prediction System

Submitted By:

Rakhee Moolchandani

Presented To:

Dr.Amin Gharipour

Kent State University

December-10-2021

Table of Contents

Executive Summary

Diabetes is a medical condition that is caused due to insufficient production and secretion of insulin from the pancreas in case of Type-1 diabetes and defective response of insulin Type-2 diabetes. Diabetes is one of the most prevalent medical conditions in people today.
Hospital readmission for diabetic patients is a major concern in the United States. Over $250 million dollars was spent on treatment of readmitted diabetic inpatients in 2011 alone. Diabetes is chronic and does not have any specific cure.
Hospital readmission rates for certain conditions are now considered an indicator of hospital quality and affect the cost of care adversely. Hospital readmissions of diabetic patients are expensive as hospitals face penalties if their readmission rate is higher than expected and reflects the inadequacies in health care system. For these reasons, it is important for the hospitals to improve focus on reducing readmission rates. Identify the key factors that influence readmission for diabetes and to predict the probability of patient readmission.
This project aims to train a classification model that can predict the chances of a diabetic patient to be readmitted to the hospital within 30 days of discharge.

The dataset was originally used in the research. The data are submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grant UL1 TR00058 and a recipient of the CERNER data. John Clore (jclore '@' vcu.edu), Krzysztof J. Cios (kcios '@' vcu.edu), Jon DeShazo (jpdeshazo '@' vcu.edu), and Beata Strack (strackb '@' vcu.edu). This data is a de-identified abstract of the Health Facts database (Cerner Corporation, Kansas City, MO).

This report presents an analysis of the dataset with the help of different analytical and statistical methods such as feature engineering and data visualizations, handling missing values, identifying the relationship between different attributes as well as determining important attributes of the dataset.

Since the business problem is a classification problem of determining whether the patient will be readmitted or not, five simple modeling strategies were utilized i.e., Naïve Bayes, K-Nearest Neighbor, Logistic Regression, Decision tree, Random Forest, and XG Boost modeling methods were built and evaluated their performance based on Accuracy, Precision, Sensitivity, specificity, AUC, and ROC plot.

These modeling methods also helped in understanding the key factors involved in the process of patient readmission and their evaluation suggests that the best way forward would be to use the XG Boost model which provides better prediction results on the given dataset. This model will help the domain experts to summarize and describe the business rules giving intuition for the data towards the decision process. Therefore, developing better patient assessment.

Overview

Diabetes Mellitus (DM) is a chronic disease where the blood has high sugar level. It can occur when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin, it produces (WHO). Diabetes is a progressive disease that can lead to a significant number of health complications and profoundly reduce the quality of life. While many diabetic patients manage the health complication with diet and exercise, some require medications to control blood glucose level. As published by a research article named "The relationship between diabetes mellitus and 30-day readmission rates", it is estimated that 9.3% of the population in the United States have diabetes mellitus (DM), 28% of which are undiagnosed. In recent years, government agencies and healthcare systems have increasingly focused on 30-day readmission rates to determine the complexity of their patient populations and to improve quality. Thirty-day readmission rates for hospitalized patients with DM are reported to be between 14.4 and 22.7%, much higher than the rate for all hospitalized patients (8.5–13.5%).

The main purpose of this Project is to facilitate healthcare institutions in predicting readmission of a diabetic patient by allowing the model to learn the relation among features and their importance in determining whether the patient will be readmitted or not. This helps the hospitals in providing the best inpatient treatment and improve the cost efficiency of healthcare centers. At the same time, it is important to identify the key factors responsible for the readmission of a diabetic patient. Hospital readmission is a crucial healthcare quality measure that helps in determining the level of quality of care that a hospital offers to a patient and has proven to be immensely expensive. The results are encouraging with patients having changes in medication while admitted having a high chance of getting readmitted. Identifying Prospective patients for readmission could help the hospital systems in improving their inpatient care, thereby saving them from unnecessary expenditures. Diabetes is one of the non- communicable diseases that are on the rise with massive urbanization and a drastic change of lifestyle in many countries. It is expected to turn into the seventh most prevalent mortality factor by 2030 and millions of deaths could be prevented each year through better analytics. When assessing the quality of care delivered by a health center, readmission is the metric of choice. It measures the number of patients that need to come back to the hospital after their initial discharge. The readmission can be classified into three broad categories such as unavoidable, planned, and unplanned. The unavoidable readmission that is highly predictable mostly due to the nature of the pathology or patient's condition (i.e. cancer phase IV, metastasis). Secondly in the planned readmission which is directly prescribed by the healthcare professional to the patient (i.e. check-up, transfusion). Lastly, the unplanned is defined as readmission that shouldn't have happened given the practitioner's diagnosis and could have been avoided if proper care was given to the patient post-discharge. Unavoidable and planned readmissions already are highly anticipated. However, predicting unplanned readmission is of prime interest due to its inherent uncertainty.

In our research, 10 years (1999-2008) of clinical care data from 130 hospitals and 70,000 diabetic inpatients in the US was used for data mining and analytics for two objectives:

1. This project aims to predict whether a diabetes patient will be readmitted within 30 days of discharge. Different machine learning models and ensemble methods are applied to train this data.

2. To extract critical risk factors that correlates with readmission of diabetic patients.

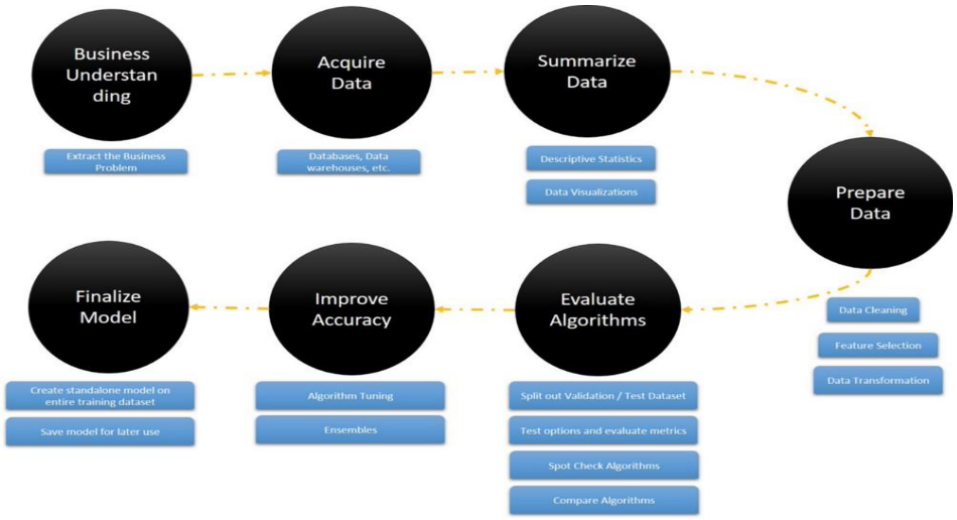
Statement of the Problem

Hospital readmission is an indicator of the quality of care and is a driver for the increasing cost of healthcare. Like other chronic diseases, Diabetes is associated with a higher risk of hospital readmission. In this project, I will evaluate several machine learning approaches to predict the probability of hospital re-admissions for diabetic patients. The data set used for this study contains more than 100,000 diabetic patient data and 55 variables including length of stay, insulin, and in-patient visits from hospitals in the United States. We leverage several pre-processing techniques and investigate the performance of the various models. The significant variables contributing to the analysis are the number of in-patients, length of stay, number of medications, number of diagnoses, and age. The results demonstrate the viability of the techniques in providing a better understanding of factors influencing hospital re- admission.

To identify the factors that lead to the high readmission rate of diabetic patients within 30 days post discharge and correspondingly to predict the high-risk diabetic-patients who are most likely to get readmitted within 30 days so that the quality of care can be improved along with improved patient's experience, health of the population and reduce costs by lowering readmission rates. Also, to identify the medicines that are the most effective in treating diabetes.

Scope of the Project

The beneficiaries of this project are twofold, the patient themselves who will benefit in terms of disease management, overall health, and early detection. The health service providers will gain, they will have a better understanding of the data where action can be taken to reduce early readmissions associated with the patient diagnosis. Early detection and treatment are essential in order to provide better treatment to patients and potentially saving lives and reducing readmitted patients' treatment healthcare costs.



Business Analytics Project Lifecycle

Dataset Exploration

Dataset:

For our analysis, we have picked up a publicly available dataset from UCI Machine Learning repository2. It covers data on diabetes patients across U.S. hospitals during a 10-year period from 1999 to 2008. There are 101,766 unique hospital admissions in the dataset from approximately 70,000 unique patients. The dataset is spread over 50 features including patient characteristics, conditions, tests and 23 medications.

Dataset Characteristics:

1) All encounters are hospital admissions.
2) Only diabetic encounters are included (at least one of the three primary diagnosis was diabetes)
3) The patient stayed in the hospital for between 1 and 14 days
4) Laboratory tests were performed on the patient
5) Some form of medication was given to the patient during the stay at the hospital

Dataset Description:

Range Index: 101766 entries, Data columns (total 50 columns):

- **Encounter ID** Unique identifier of an encounter
- **Patient number** Unique identifier of a patient
- **Race Values:** Caucasian, Asian, African American, Hispanic, and other
- **Gender Values:** male, female, and unknown/invalid
- **Age** Grouped in 10-year intervals: 0, 10), 10, 20), ..., 90, 100)
- **Weight:** Weight in pounds
- **Admission type** Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
- **Discharge disposition** Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
- **Admission source** Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
- **Time in hospital** Integer number of days between admission and discharge
- **Payer code** Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay Medical
- **Medical specialty** Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
- **Number of lab procedures** Number of lab tests performed during the encounter

- **Number of procedures** Numeric Number of procedures (other than lab tests) performed during the encounter
- **Number of medications** Number of distinct generic names administered during the encounter
- **Number of outpatient visits** Number of outpatient visits of the patient in the year preceding the encounter
- **Number of emergency visits** Number of emergency visits of the patient in the year preceding the encounter
- **Number of inpatient visits** Number of inpatient visits of the patient in the year preceding the encounter
- **Diagnosis 1** The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
- **Diagnosis 2** Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
- **Diagnosis 3** Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
- **Number of diagnoses** Number of diagnoses entered to the system 0%
- **Glucose serum test result:** Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
- **A1c test result** Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.
- **Change of medications** Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”
- **Diabetes medications** Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”
- 24 features for medications for the generic names**: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin- pioglitazone**, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed
- **Readmitted** Days to inpatient readmission. Values: “0” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission

Exploratory Data Analysis

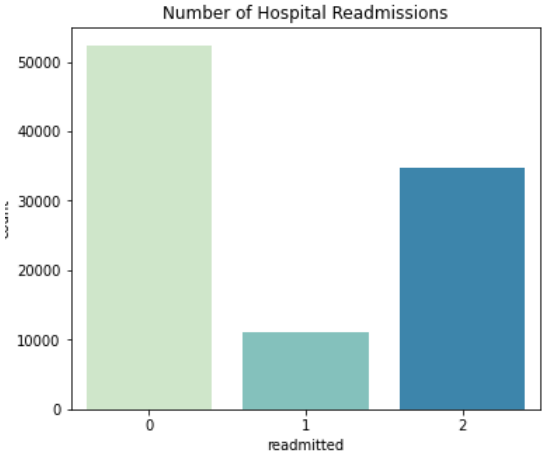
Exploratory Data Analysis (**EDA**) is an approach to analyzing datasets to summarize their main characteristics. We will apply this approach to understand data, get some context regarding it, understand the variables and the relationships between them, and formulate hypotheses that could be useful when building predictive models.

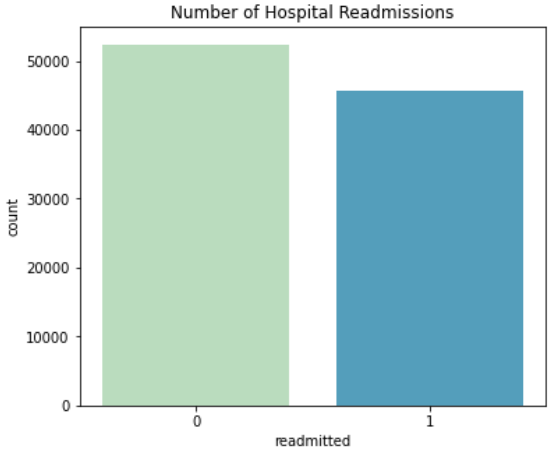
Target Attribute:

Target content changed to 1-0

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:

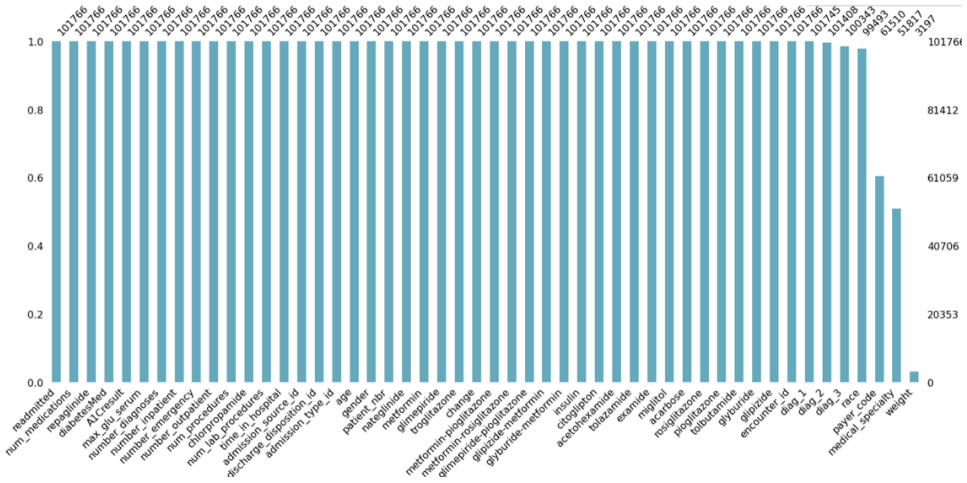
NO and >30: 0
<30: 1





Missing Values:

The dataset did have some missing values (NAs). Attributes race, payer_code, medical_speciality, weight has missing values.



Per our results above, let's take a closer look at all variables that have missing/null values. Since "weight" had the greatest number of missing values, let's start with that variable first.

Approximately 97% or more than 9,500 of the values denoted under the weight variable are missing or null values. Typically, in cases of missing values the next step is to determine which method to best impute the missing values but seeing as over 95% of the values are missing for a data set with over 100,000 observations, it makes much more sense to simply drop the column all together.
With almost 40% of its data missing, I also decided to drop payer_code variable.
For the missing null values, I researched to see if they indicated a different type of specialist to those listed above or it simply meant that there was no data available for that observation. I found out it was the former. I decided to drop medical_speciality variable seeing that almost 50% of the data was not available to begin with. Also trying to impute these missing values would introduce more inaccuracy since I'd simply have to make speculations for the missing data.
Approximately 2% or less than 2,300 of the values denoted under race are missing or null values. As dropping these won't result in a great impact to the overall dataset, that's the step I'll take.
As each of the diagnoses variables have missing or null values that are less that 2% of the remaining observations, dropping these observations won't result in a great impact to the overall dataset.

Summary:

- Dropped the following columns:
    - weight
    - payer_code
    - medical_speciality
- Dropped rows with nan/missing values for the following columns:
    - race

- diag_1
- diag_2
- diag_3

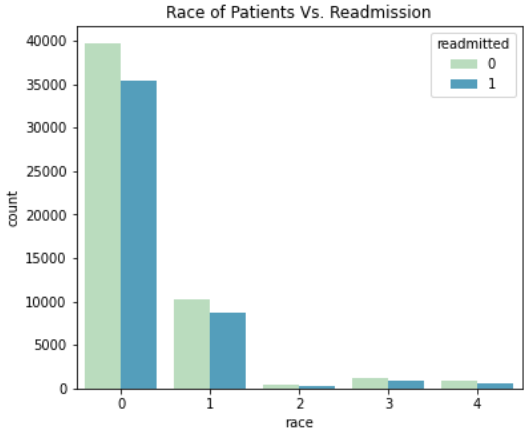
We are left with 98,053 observations or 96% of our original data set to continue cleaning.

Exploring Categorical values in the dataset:

Readmission Rate based on:

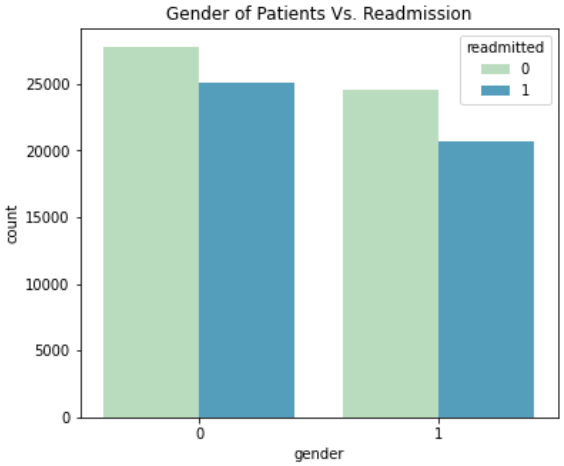
Race:
The data collected showcases that 76% of patients within the data set are Caucasian.

- 0: Caucasian
- 1: African American
- 2: Asian
- 3: Hispanic
- 4: Other

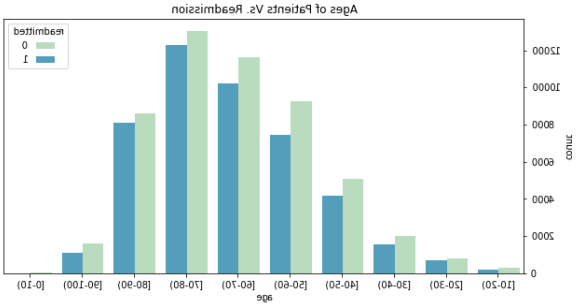


Gender:

- 0: Female
- 1: Male



Age:



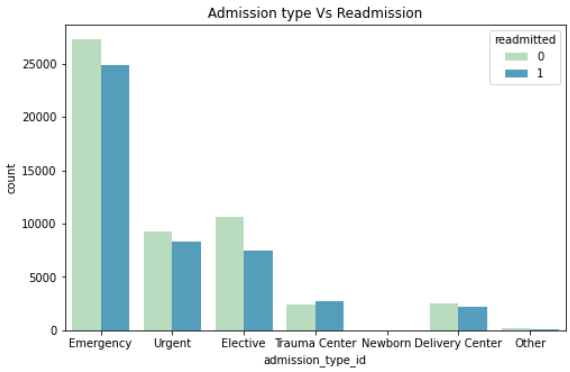
- In all age groups the number of readmissions never exceeded the number of non-readmissions.
- The age group with the highest readmissions overall were those between 70-80 years of age.

Admission Type:

Integer identifier corresponding to distinct values:

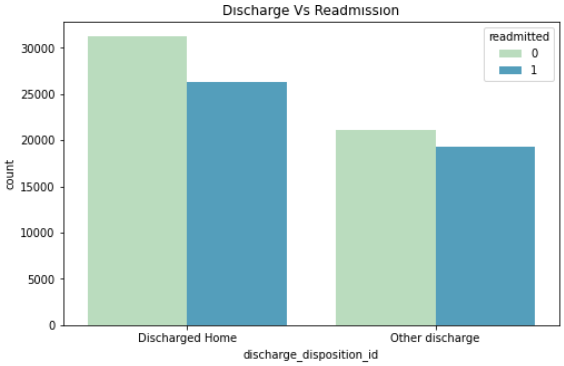
1. Emergency
2. Urgent
3. Elective
4. Newborn
5. Delivery
6. Trauma Center
7. Other

Reduced the initial 9 distinct values to just 6.



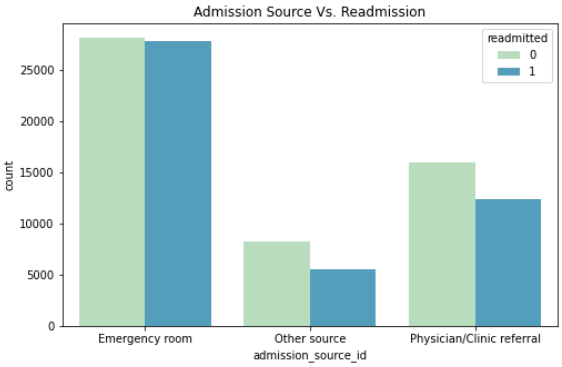
Discharge Disposition:

Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available. Reduced the initial 29 distinct values to just 2, simply quantifying whether a patient was discharged to their home or not.



Admission Source:

Indicates the origination of a patient's admission, for example, physician referral, emergency room, and transfer from a hospital. Based on the value counts above, I decided to reduce the distinct values to simply capture three instances, a physician referral, emergency room referral, or other.

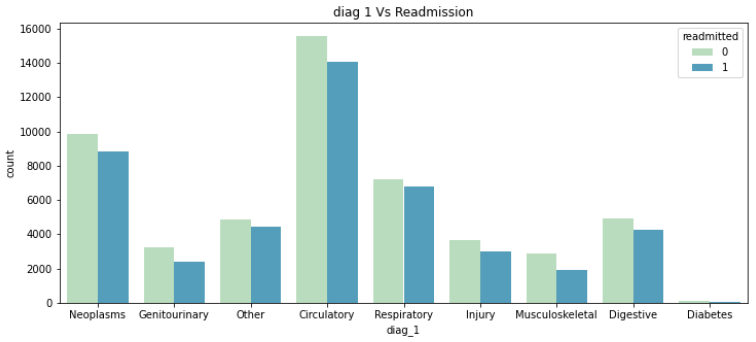


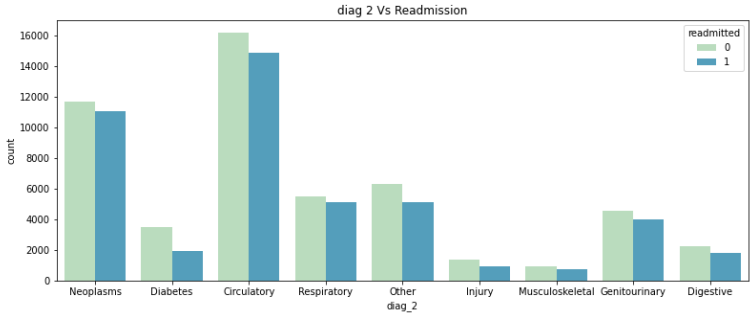
Diagnoses Attributes:

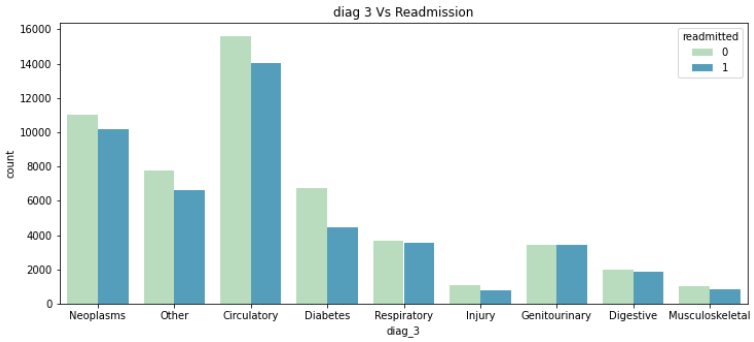
- 'diag_1'- primary diagnoses
- 'diag_2'- secondary diagnosis
- 'diag_3'- additional diagnosis

Each of these indicated a particular level of diagnosis for a given patient. However each had 848-923 distinct values coded as the first three digits of the International Classification of Diseases (ICD- 9). Based on research papers used for my analysis I decided to do what previous researchers and analysts had done and consolidate these into 9 major categories:

1. Circulatory
2. Respiratory
3. Digestive
4. Diabetes
5. Injury
6. Musculoskeletal
7. Genitourinary
8. Neoplasms
9. Others.



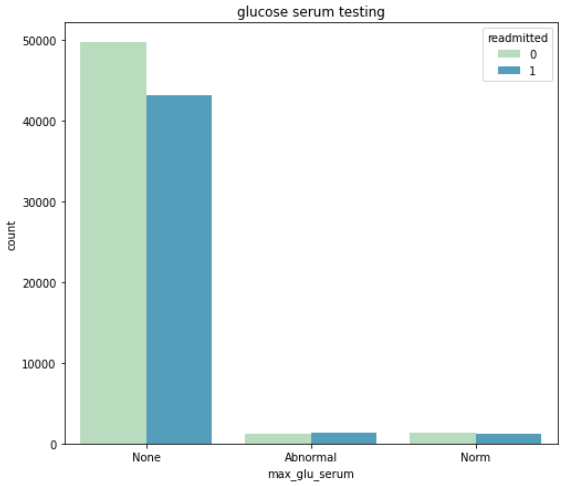




For each level of diagnoses data, it's noted that diabetes is always one of the least diagnosed diseases.

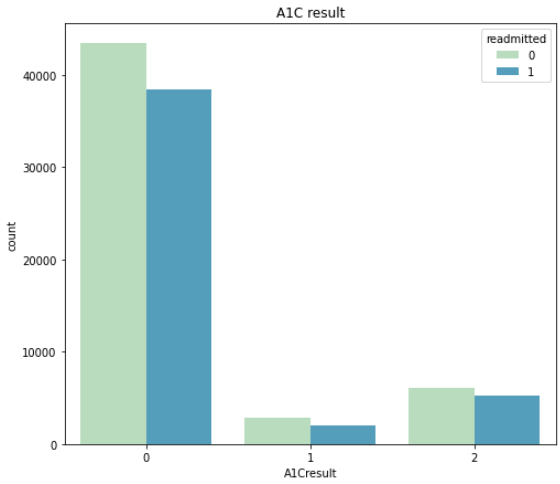
Glucose Serum Test Result:

The glucose serum test measures a person's blood sugar before and after they drink a liquid that contains glucose. When measured a blood sugar level of 200mg/dL or higher indicates a person has diabetes. I grouped our instances of ">200" and ">300" together as abnormal and dummified all the observations.



A1c Test Results:

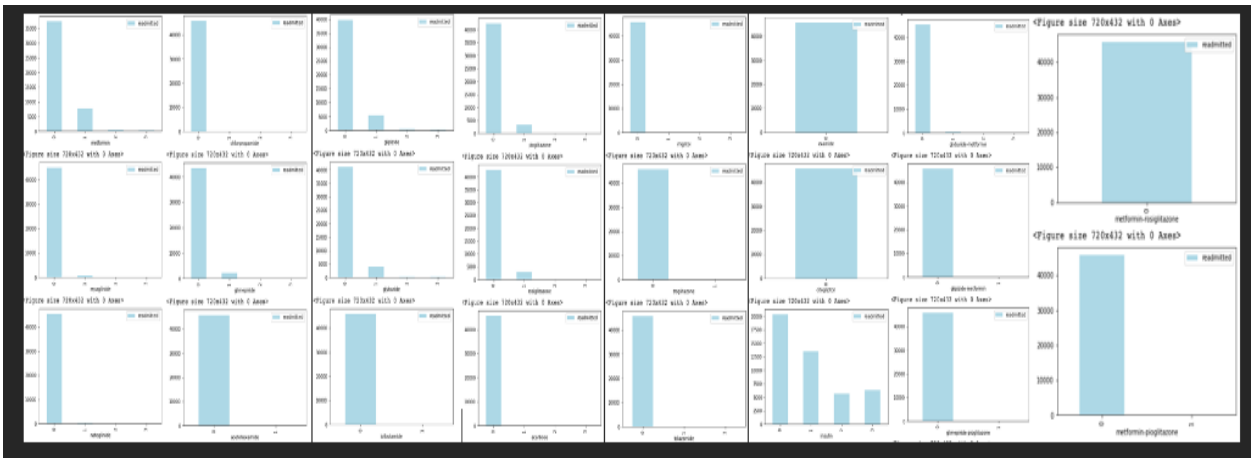
The A1C test measures your average blood sugar level over the past 2 or 3 months. An A1C below 5.7% is normal, between 5.7 and 6.4% indicates you have prediabetes, and 6.5% or higher indicates you have diabetes. I grouped our instances of of ">7" and ">8" together as abnormal and dummified all the observations.



Administered Medications:

Variable indicates whether the drug was prescribed or there was a change in the dosage.

- 0: "No" if the drug was not prescribed
- 1: “steady” if the dosage did not change
- 2: “up” if the dosage was increased during the encounter
- 3: “down” if the dosage was decreased

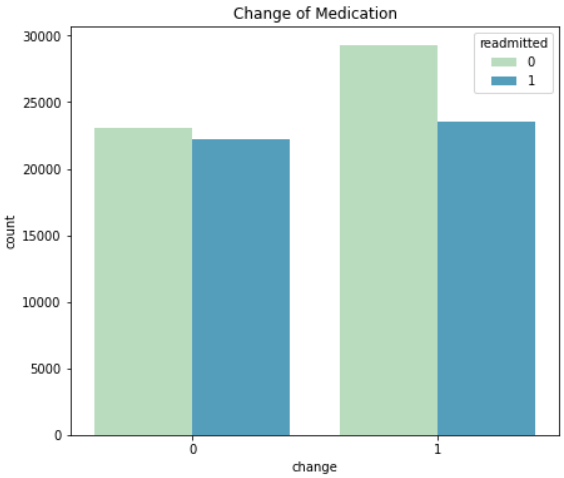


Based on the graphs above, I determined that the following medications were not administered at all, value of "0", I decided to drop these since they had no impact on our overall outcomes.

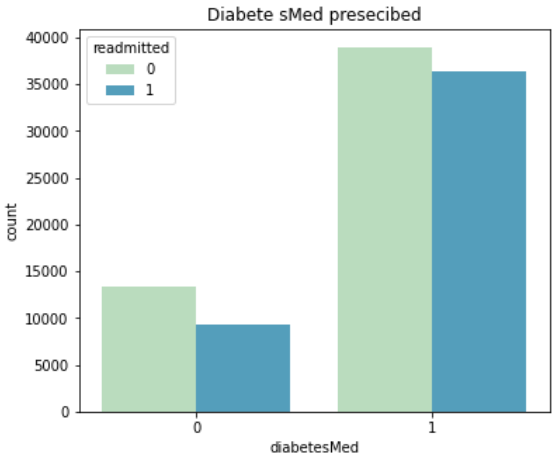
- chlorpropamide
- acetohexamide
- troglitazone
- tolbutamide
- acarbose

- miglitol
- troglitazone
- tolazamide
- examide
- citoglipton
- glipizide-metformin
- glimepiride-pioglitazone
- metformin-rosiglitazone
- metformin-pioglitazone

Change of Medication:

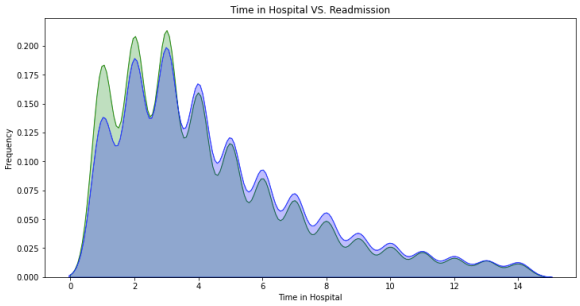


Diabetes Medication Prescribed:



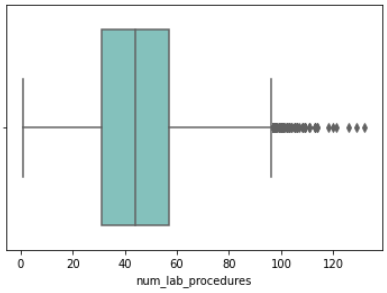
Exploring Non-Categorical values in the dataset:

Time:
On average patients spent 4.4 days in the hospital. Shortest being 1 day and the longest being a total of 14 days.

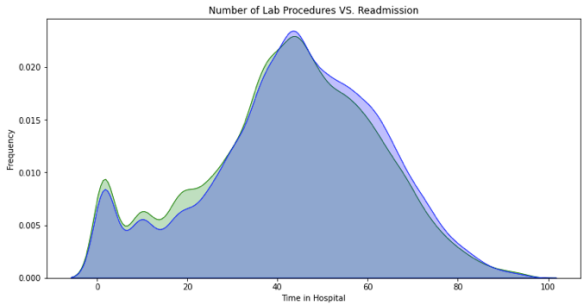


Number of Lab Procedures:

There seems to be a greater percentage of 0 lab tests performed during initial admission. On average 43 lab tests were performed during the patients' initial admission with a max of 132 performed, for a lay person that seems like a lot of procedures to perform on one patient. A look at the outliers indicates just that with the boxplot displayed below. It looks like anything past 96 procedures isn't the norm.



142 observations lie outside the threshold. I've decided to drop these outliers to reduce as much bias as I possibly can.

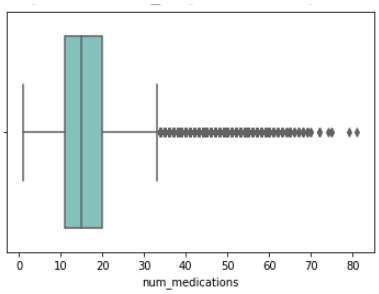


Number of Procedures:

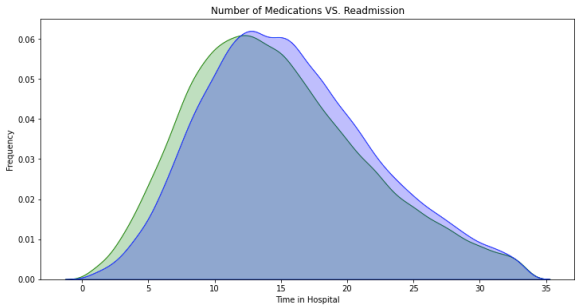
On average 1.3 non-lab test procedures were performed during the patient's initial admission.

Number of Medications:

On average 16 distinct generic types of medication were administered to the patient during their initial admission. A maximum of 81 distinct types of medication were administered which I deem to be a large amount for one patient and possibly not the norm, since that a big jump from the average. A look at the outliers indicates just that with the boxplot displayed below. It looks like anything past 33 distinct types of medications isn't the falls out of the upper limit.



3262 observations lie outside the threshold. I've decided to drop these outliers to reduce as much bias as I possibly can.



Number of Outpatient Visits:

The number of outpatient visits of the patient in the year preceding the encounter ranged from zero visits all the way up to 42 visits.

Number of Emergency Visits:

The number of emergency visits of the patient in the year preceding the encounter ranged from zero visits all the way up to 76 visits.

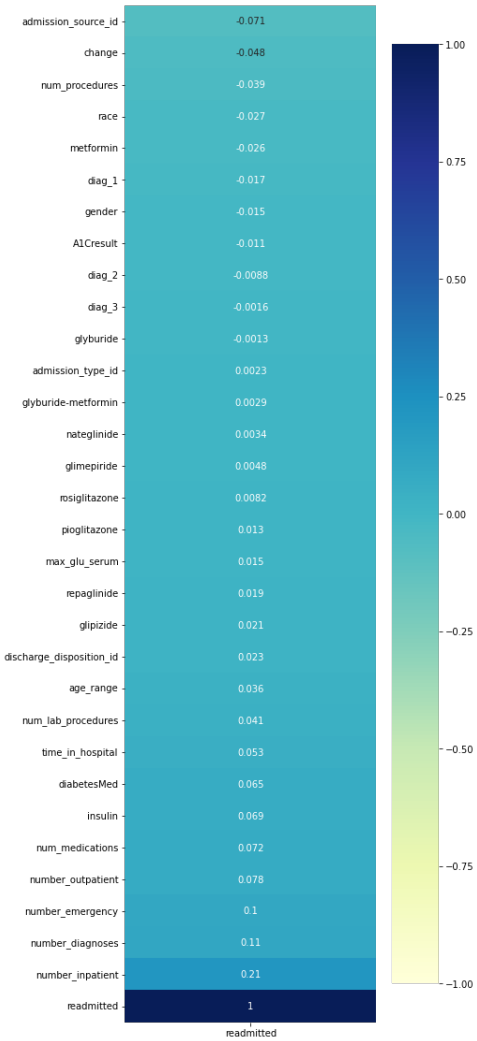
Number of Inpatient Visits:

The number of inpatient visits of the patient in the year preceding the encounter ranged from zero visits all the way up to 21 visits.

Multi-Collinearity:

Multicollinearity generally occurs when there are high correlations between two or more predictor variables

We can identify which variables are affected by multicollinearity and the strength of the correlation. Heatmap helps us to visualize the correlation between variables. It takes a rectangular data grid as input and then assigns a color intensity to each data cell based on the data value of the cell. This is the great way to get visual clues about the data.



These correlations aren't very strong (>0.5), the variables with the strongest correlations to our target variable of readmission, are the number of impatient visits, number of diagnoses, and number of emergency visit in the preceding year.

After all our cleaning, we are left with 32 features, 1 target feature, and 94,620 observations.

Modeling Strategy

There are hundreds of machine learning algorithms available to data scientists, and new ones are created every day. The correct algorithm for a given machine learning problem is the prerequisite for a good model that can then become a good business tool.
All models have some predictive error, given the statistical noise in the data, the incompleteness of the data sample, and the limitations of each different model type. Therefore, the notion of a perfect or best model is not useful. Instead, we must seek a model that is "good enough.

Since it is a Classification problem, following a supervised learning approach therefore the preference was given to simple and easier classification models which can help in understanding the data and decision process. These models can not only achieve high performance but also, provide interpretability helping few summarizing and descriptive rules giving intuition about the data that can directly help domain experts understand the decision process for classification tasks.

Analysis and evaluation from different models help in providing not only better prediction results but with better interpretability and explanation to the results. Below is the list of models:

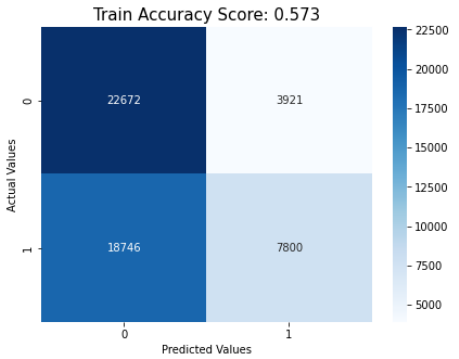
1. Naive Bayes
2. KNN Classification
3. Logistic Regression
4. Decision Trees
5. Random Forest
6. XG Boost

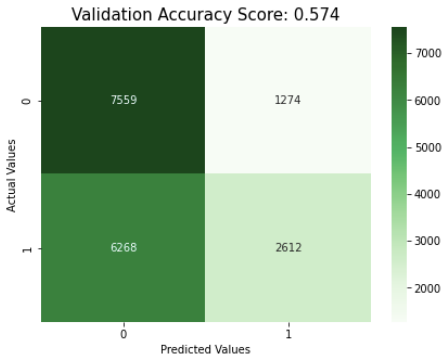
Naive Bayes

Naïve Bayes is a great model if the data is not complex, and the task is relatively simple. The Naïve component relies on independent assumptions and looks for the likelihood point from a data set that will exhibit similar features to a new random variable. It is a **high bias/low variance classifier** and is helpful when working with a limited amount of data. Due to the simple nature of Naïve Bayes, **it does not tend to overfit data**, enabling the ability to train data quickly.

```
Naive Bayes Classifier
Training:
AUC:0.637
accuracy:0.573
recall:0.294
precision:0.665
fscore:0.408
specificity:0.853

Validation:
AUC:0.637
accuracy:0.574
recall:0.294
precision:0.672
fscore:0.409
specificity:0.856
```





Observations:

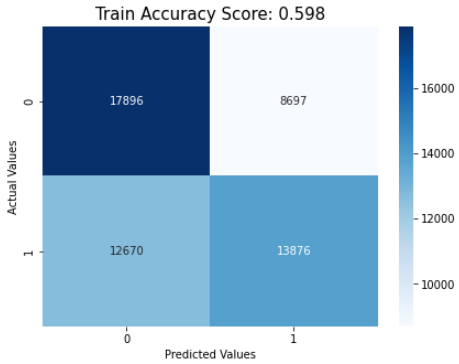
- The Naive Bayes model gives and Accuracy of 57.4%, recall of 29.4% Specificity of 85.6%.
- The AUC for Naive Bayes model is 63.7%.
- Based on confusion matrix, the specificity is higher than recall, which means the models is performing better in classifying the people who will be readmitted.

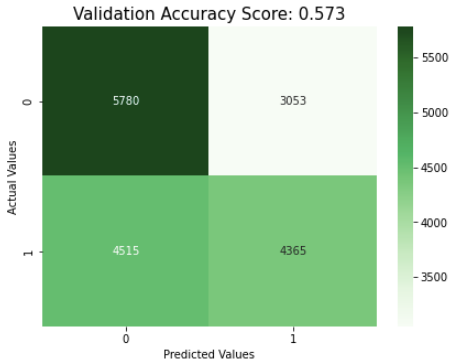
KNN

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in proximity. In other words, similar things are near to each other. The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with mathematics of calculating the distance between points on a graph.

```
k Nearest Neighbours
Training:
AUC:0.637
accuracy:0.598
recall:0.523
precision:0.615
fscore:0.565
specificity:0.625

Validation:
AUC:0.606
accuracy:0.573
recall:0.492
precision:0.588
fscore:0.536
specificity:0.604
```





Observations:

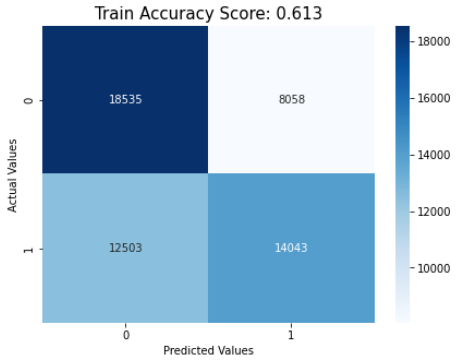
- The KNN model gives and Accuracy of 57.3%, recall of 49.2% Specificity of 60.4%.
- The AUC for Naive Bayes model is 60.6%.
- Based on confusion matrix, the specificity is higher than recall, which means the models is performing better in classifying the people who will be readmitted.

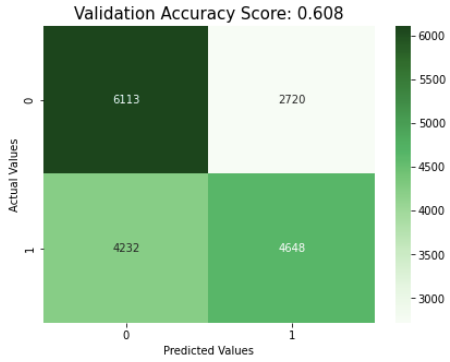
Logistic Regression Model

Logistic Regression utilizes the power of regression to do classification. One of the main reasons for the model's success is **its power of explain ability** i.e. calling out the contribution of individual predictors, quantitatively.
Unlike regression which uses Least Squares, the model uses Maximum Likelihood to fit a sigmoid curve on the target variable distribution.

```
Logistic Regression
Training:
AUC:0.658
accuracy:0.613
recall:0.529
precision:0.635
fscore:0.577
specificity:0.697

Validation:
AUC:0.658
accuracy:0.608
recall:0.523
precision:0.631
fscore:0.572
specificity:0.692
```





Observations:

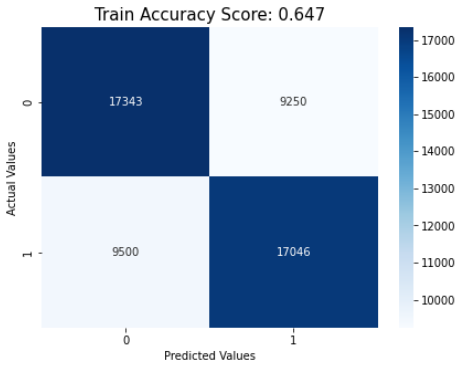
- The Logistic Regression model gives and Accuracy of 60.8%, recall of 52.3% Specificity of 69.2%.
- The AUC for Logistic Regression model is 65.8%. It performed better than the Naive Bayes model.
- Based on the confusion matrix, the specificity is higher than sensitivity, which means the model has performed better to classify the people who will be readmitted.

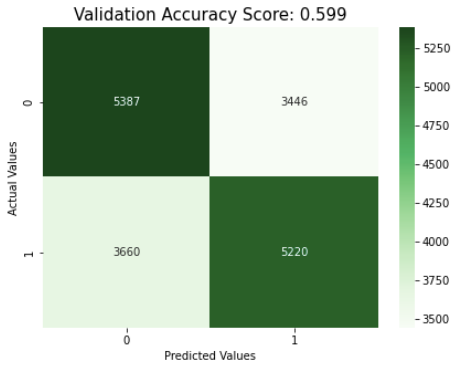
Decision Tree

To see how a decision tree predicts a response, we need to **follow the decisions in that tree from the beginning node down to a leaf node** (each split in the tree is called a leaf) which contains the response. Classification trees give **nominal responses, such as true or false**. Decision trees are relatively easy to follow, and we can **see a full representation of the path** taken from the beginning (root) to the leaf.

```
Decision Tree
Training:
AUC:0.711
accuracy:0.647
recall:0.642
precision:0.648
fscore:0.645
specificity:0.649

Validation:
AUC:0.632
accuracy:0.599
recall:0.588
precision:0.602
fscore:0.595
specificity:0.606
```





Observations:

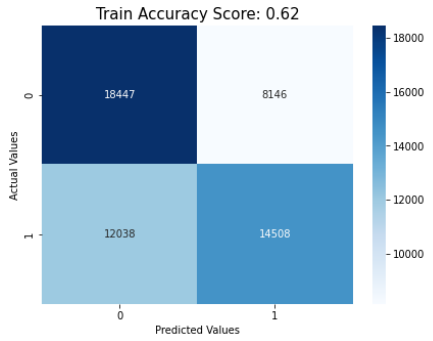
- The Decision Tree model gives and Accuracy of 59.9%, recall of 58.8% Specificity of 60.6% .
- The AUC for Decision Tree model is 59.9%. The Decision Tree model did not outperform all the models executed before Naive Bayes, KNN, Logistic Regression.
- From the confusion matrix, we can see that specificity is slightly higher in comparison to recall, which means the models has performed better in classifying the people who will be readmitted. The derivation of a Decision Tree is Random Forest Regression, which is where the same algorithm is applied multiple times – effectively a team of decision trees offering an average of different predictions.

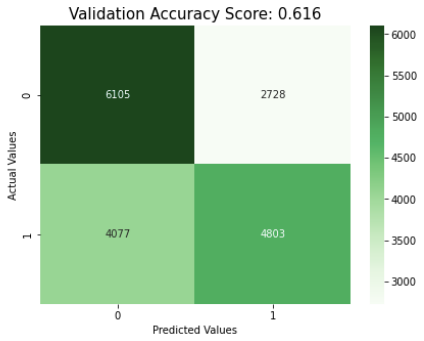
Random Forest

A Random Forest is a **reliable ensemble of multiple Decision Trees**. It is more popular for classification problems than regression applications. It builds individual trees via bagging and split using fewer features. This results in a diverse forest of uncorrelated trees exhibit reduced variance; therefore, is **more robust towards change in data and carries its prediction accuracy to new data.**

```
Random Forest
Training:
AUC:0.672
accuracy:0.620
recall:0.547
precision:0.640
fscore:0.590
specificity:0.694

Validation:
AUC:0.663
accuracy:0.616
recall:0.541
precision:0.638
fscore:0.585
specificity:0.691
```





Observations:

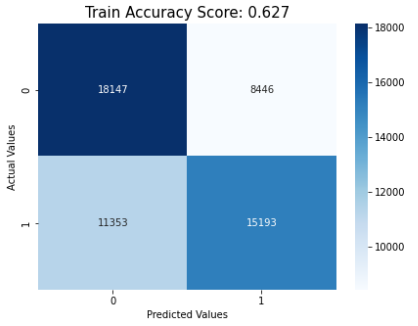
- The Random Forest model gives and Accuracy of 61.6%, recall of 54.1% Specificity of 69.1%.
- The AUC for Random Forest model is 66.3%. Sensitivity is higher than recall, which means that model has performed better in classifying the people who will readmitted.
- The Random Forest model has outperformed all the models executed before Naive Bayes, Logistic Regression, KNN, Decision Tree.

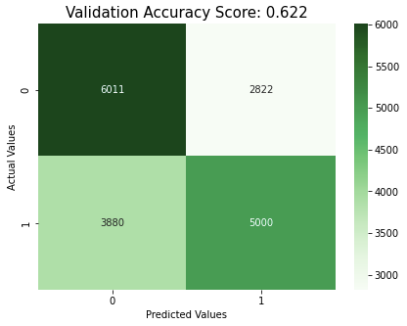
XG Boost

XG Boost stands for extreme Gradient Boosting. XG Boost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. XG Boost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

```
XGBOOST
Training:
AUC:0.681
accuracy:0.627
recall:0.572
precision:0.643
fscore:0.605
specificity:0.682

Validation:
AUC:0.671
accuracy:0.622
recall:0.563
precision:0.639
fscore:0.599
specificity:0.681
```





Observations:

- The XG Boost model gives and Accuracy of 62.2%, recall of 56.3% Specificity of 68.1%.
- The AUC for XG Boost model is 67.1%. Sensitivity is higher than recall, which means that model has performed better in classifying the people who will readmitted.
- The XG Boost model has outperformed all the models executed before including Random Forest.

Evaluation of Classification Model's Performance

Evaluation of all classification models included a range of different metrics with their advantages and shortcomings.

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

Accuracy = correct prediction / all predictions

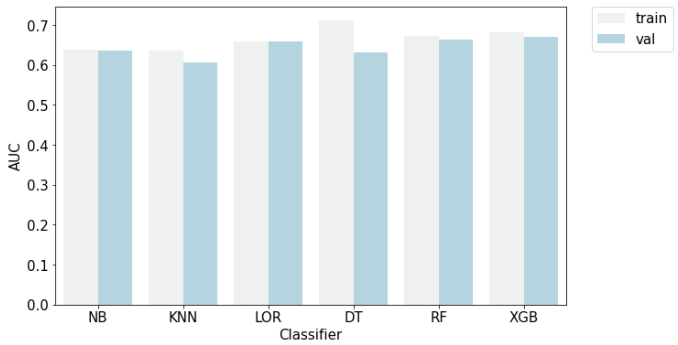
Precision is defined as the fraction of relevant examples (true positives) among all of the examples which were predicted to belong in a certain class.

Sensitivity and Specificity

Sensitivity, also known as the true positive rate (TPR), is the same as recall. Hence, it measures the proportion of positive class that is correctly predicted as positive.

Specificity is like sensitivity but focused on negative class. It measures the proportion of negative class that is correctly predicted as negative.

	classifier	data_set	auc	accuracy	recall	precision	fscore	specificity
0	NB	train	0.637392	0.573439	0.293830	0.665472	0.407662	0.852555
1	NB	val	0.636596	0.574211	0.294144	0.672156	0.409212	0.855768
2	KNN	train	0.636595	0.597904	0.522715	0.614717	0.564995	0.625089
3	KNN	val	0.606483	0.572743	0.491554	0.588434	0.535649	0.604098
4	LOR	train	0.658480	0.613071	0.529006	0.635401	0.577343	0.696988
5	LOR	val	0.658296	0.607520	0.523423	0.630836	0.572132	0.692064
6	DT	train	0.711016	0.647152	0.642131	0.648235	0.645169	0.648629
7	DT	val	0.632333	0.598826	0.587838	0.602354	0.595007	0.605796
8	RF	train	0.671753	0.620166	0.546523	0.640417	0.589756	0.693679
9	RF	val	0.663065	0.615819	0.540878	0.637764	0.585339	0.691158
10	XGB	train	0.680553	0.627411	0.572327	0.642709	0.605480	0.682398
11	XGB	val	0.670900	0.621634	0.563063	0.639223	0.598731	0.680516

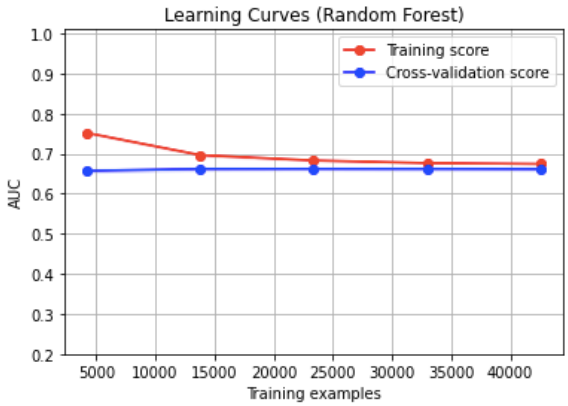


AUC is an abbreviation for the area under the curve. It is used in **classification analysis** to determine which of the used models predicts the classes best. If the classifier is **good AUC will**

be close to 1. AUC is a better measure of classifier performance than accuracy because it does not bias on size of test or evaluation data. Accuracy is always biased on the size of test data.

Model Selection Learning Curve

In the case of random forest, we can see the training and validation scores are similar but they both have low scores. This is called high bias and is a sign of underfitting.



Depending on learning curve, there are a few strategies you can employ to improve your models

High Bias:

- Add new features
- Increase model complexity
- Reduce regularization
- Change model architecture

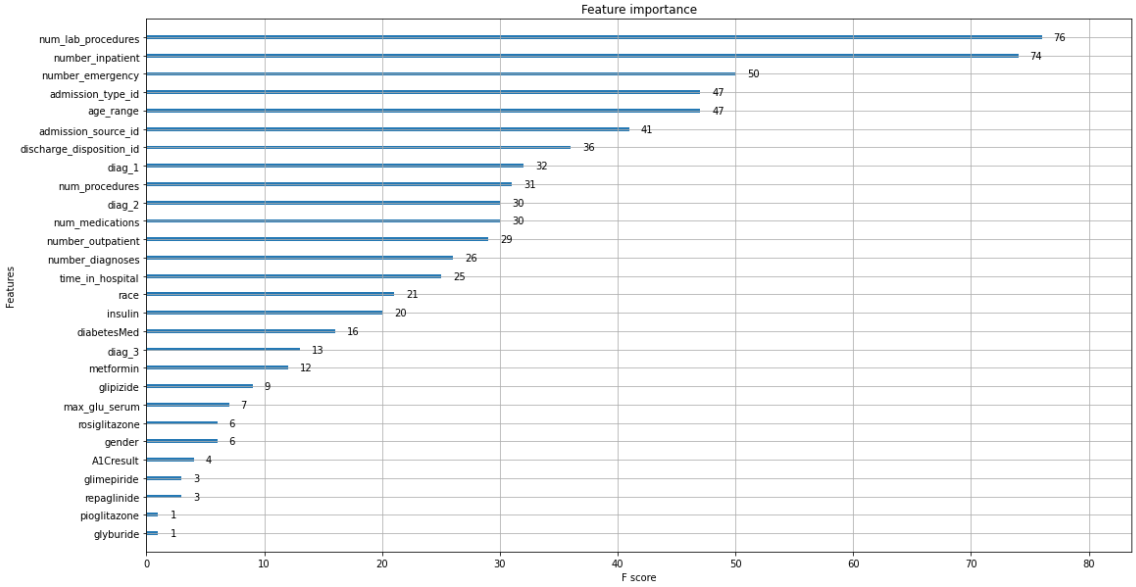
High Variance:

- Add more samples
- Add regularization
- Reduce number of features
- Decrease model complexity
- Add better features
- Change model architecture

Feature Importance

Feature selection is for filtering irrelevant or redundant features from your dataset.

Several techniques were tested to perform feature selection including statistical significance tests, forward selection, backward elimination, and feature importance. Based on the collective results of these techniques, below are the set of most important features selected for model building.




```
best_features.index
```

```
Index(['number_inpatient', 'number_diagnoses', 'diabetesMed',
       'number_outpatient', 'number_emergency', 'admission_source_id',
       'max_glu_serum', 'time_in_hospital', 'age_range', 'num_medications',
       'admission_type_id', 'discharge_disposition_id', 'metformin',
       'num_procedures', 'num_lab_procedures', 'insulin', 'diag_1', 'diag_2',
       'glipizide', 'gender', 'race', 'repaglinide', 'A1Cresult', 'diag_3',
       'glyburide'],
      dtype='object')
```

Hyperparameter Tunning

Hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. Hyperparameter settings could have a big impact on the prediction accuracy of the trained model. Optimal hyperparameter settings often differ for different datasets and different models. Therefore, they should be tuned for each dataset and model. Since the training process doesn't set the hyperparameters, there needs to be a meta process that tunes the hyperparameters.

Hyperparameter Tuning for the best 2 models in our project:

Random Forest Model:

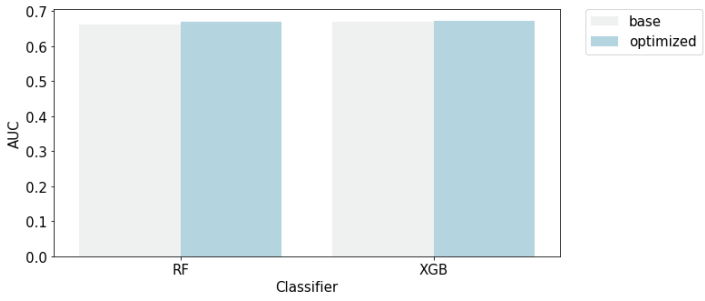
```
Baseline Random Forest
Training AUC:0.672
Validation AUC:0.663
Optimized Random Forest
Training AUC:0.714
Validation AUC:0.669
```

XG Boost Model:

```
Baseline XGB
Training AUC:0.681
Validation AUC:0.671
Optimized XGB
Training AUC:0.707
Validation AUC:0.673
```

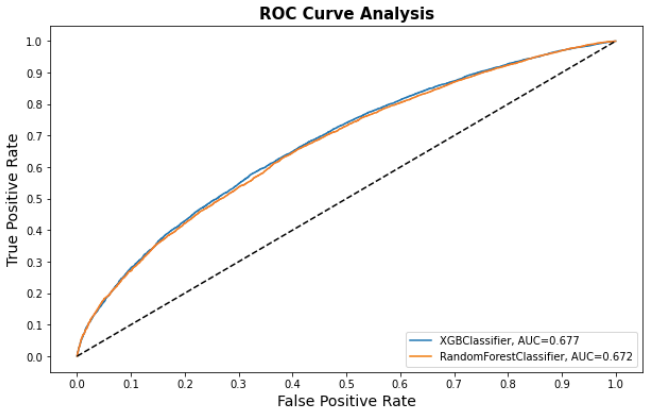
Hyperparameter Tunning Results:

	classifier	data_set	auc
0	RF	base	0.663065
1	RF	optimized	0.669049
2	XGB	base	0.670900
3	XGB	optimized	0.672530



ROC – AUC Comparison of Models:

Classifiers that give **curves closer** to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



After evaluating the models based on performance metrics i.e. Accuracy, Sensitivity, Specificity, AUC and ROC, XG Boost model has performed the best in of every single metric.

XG Boost produced an accuracy of 62.88% with an AUC of 67.73%

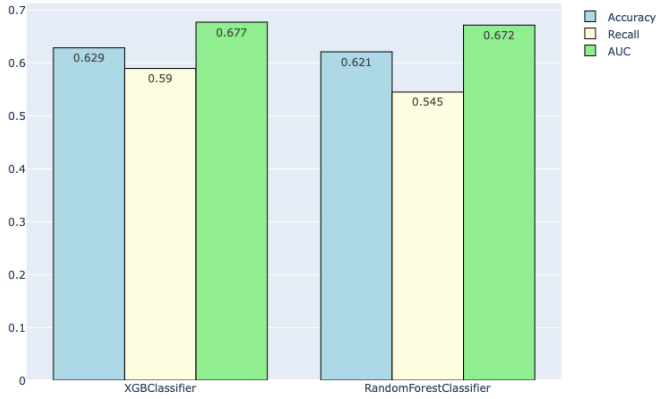
Therefore, we can choose XG Boost model for the dataset to predict the behavior of the drivers based on different scenarios and can use the variable importance plot to identify the important variables.

Final Model Chosen:

As we can see from the above results, Random Forest and XG Boost are still giving good results after hyperparameter tuning. But the training and prediction time of Random Forest is far more than that of XG Boost. So, Random Forest is not scalable and not suitable for production because it will increase costs and time while giving similar results as XG Boost which takes only 2-3 seconds to train on this large dataset thus saving costs and time while giving the best results.

```
test_result
```

classifiers	accuracy	recall	auc
XGBClassifier	0.628831	0.589770	0.677323
RandomForestClassifier	0.621327	0.545054	0.671795



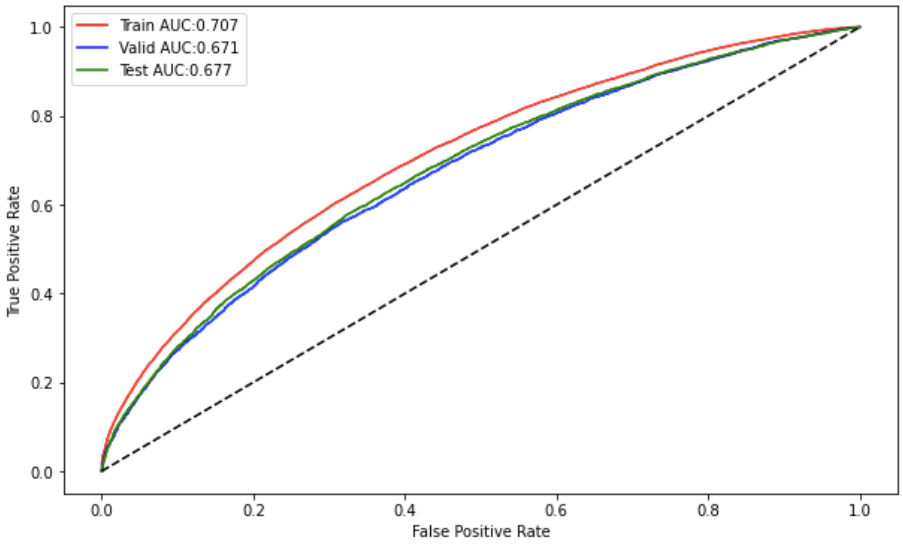
Predicting Results

Below are the test results predicted using the chosen XG Boost Model.

```
Training:
AUC:0.707
accuracy:0.647
recall:0.612
precision:0.658
fscore:0.634
specificity:0.682

Validation:
AUC:0.671
accuracy:0.622
recall:0.563
precision:0.639
fscore:0.599
specificity:0.681

Test:
AUC:0.677
accuracy:0.629
recall:0.590
precision:0.606
fscore:0.598
specificity:0.663
```



Conclusion

30-day hospital readmission of diabetes patients is of prime importance for health centers and is found

very stressful due to the current models limit in term of performance and generalizability. To cope with this challenge, this project implemented a comprehensive pre-processing framework to improve the initial data quality, hence empowering the model's efficiency. The suggested pre-processing framework included comprehensive data cleaning, data reduction and transformation aiming at better optimizing and selecting prominent features for 30-day unplanned readmission among diabetes patients

We Performed Explanatory Data Analysis and applied extensive feature engineering, feature selection and on Diabetes patient's hospital readmission data. We used Naïve Bayes, KNN, Logistic Regression, Decision Tree, Random Forest, and XG Boost classifiers to predict the readmission rate. All our algorithms are evaluated using the area-under-the-curve (AUC) & confusion Matrix, which is equivalent to the c-statistic in the binary classification scenario. Our results showed that number of inpatient and number of diagnoses are the two most critical factors in readmission prediction. These results provided valuable suggestions to inpatients monitoring policy that may reduce short-term readmission and public healthcare cost in the future. The overall comparison of methods based on their accuracy sensitivity and specificity was generated as the output with XG Boost Classifier as the best model with 63% accuracy.

Based on the classes that models produced, diabetic patients who are more likely to be readmitted are either women, or Caucasians, or outpatients, or those who undergo less rigorous lab procedures, treatment procedures, or those who receive less medication, and are thus discharged without proper improvements or administration of insulin despite having been tested positive for HbA1c.

Diabetic patients who do not undergo vigorous lab assessments, diagnosis, medications are more likely to be readmitted when discharged without improvements and without receiving insulin administration, especially if they are women, Caucasians, or both.

Insights

- This project aims to train a classification model that can predict if a user will accept a coupon given his/her answers to some survey questions based on different driving conditions.
- It presents an examination of the dataset with the help of different insightful and statistical techniques such as feature engineering and data visualizations, handling missing values, identifying the relationship between different attributes as well as determining significant characteristics of the dataset.
- With the help of visuals (bar charts), we were able to analyze the patients' readmission accuracy based on different scenarios. The visuals were able to provide an accessible way to see and understand trends and patterns in the data. This understanding can be used to tell a story, drive decisions, and create predictive
- Feature importance is a basic approach to interpreting the model. Interpretability leads to reliability. Ensuring that small changes in the input do not lead to large changes in the prediction.
- Understanding of the business problem is a must to come up with right solution. Data analysis helped in understanding of relationship between different attributes and how they can impact the decisions of the drivers towards coupon acceptance.
- We were successfully able to build models with better accuracy and precision by cleaning the data, identifying relationships between different attributes, and removing non-contributing elements.
- Modeling methods also helped in understanding the key factors involved in the process of patients' hospital readmission and their evaluation suggests that the best way forward would be to use the XG Boost model which provides better prediction results on the given dataset. The model also provides a list of important variables, domain experts can rely on it to summarize and describe the business rules giving intuition for the data towards the decision process. Therefore, developing patients' hospital readmission for different categories of patients and under different situations.

Business Recommendations

This project was designed to help hospitals decrease readmission rates for diabetic patients. With the proposed model, hospitals can target patients in high-risk percentiles. Not only are these patients at higher risk of being readmitted, but the model precision is also considerably better for the higher percentiles which means hospitals can efficiently use their resources to reduce readmission rates by administering medication that are highly effective in treating diabetes like insulin, metformin, glipizide, and glyburide. Medical facilities can take precautionary measures with these patients during their initial admission by making A1C and maximum glucose serum test compulsory and providing the treatment accordingly. A follow-up visits to check their progress should also be schedule at the time of discharge.

This allows hospitals to provide a better quality of healthcare to their patients and reduce the readmission rates. This reduction can help hospitals avoid penalties that are incurred for high readmission rates, leading to reduction in health expenditures for hundreds if not thousands of dollars per diabetes patient, while simultaneously improving health outcomes, and saving lives.

To summarize, these are the recommendations proposed to the business client:

- Medical facilities can take precautionary measures with patients during their initial admission by making A1C and maximum glucose serum test compulsory and providing the treatment accordingly reason being 80-90% of the readmitted patients had not gone under these tests.
- A follow-up with the discharged patients should be one to keep a track of their health and to counsel them time-to-time.
- High-risk patients' current medicines' regime should be re-evaluated, and the most effective medicines should be considered.
- Most Effective Medicines as per the findings are Insulin, Metformin, Glipizide. These medicines are coming out to be statistically significant, coming out to be quite important with respect to different machine learning models employed, are most widely prescribed, and are associated with low risk of readmission if given to the patient.
- The annual plans, financials and infrastructure / inventory of the hospital should be planned accordingly by considering the predicted readmissions.
- Hospitals must provide extra attention and care to high-risk patients.

Future Scope

Geographical factor can also be included as a feature like whether a patient is from urban or a rural region. This would allow determining if the readmission factors differ based on patient geographical location or if similar traits are observed nationwide. In addition, this would strengthen both urban and rural models while assessing the importance of age categorization.

This research study has only targeted patients with diabetes. Readmission prediction model needs to be generated for other key health conditions also such as heart disease, kidney disease etc. in Indian Healthcare system. In the future studies, planned and unplanned (emergency) readmissions needs to be considered.

Various other key features in the medical records, like family history (to find hereditary information), emotional status (depression), socioeconomic status, and lifestyle habits (exercise), smoking status and season of readmission need to be collected and analyzed. It will be interesting to perform a more exhaustive exploration of additional features in the dataset and study their relevance towards predicting the risk of readmission.

Living with diabetes is challenging and distressful. Diabetic patient's condition cannot be understood only from his medical charts. There is a need to collect and analyze both subjective and objective patient information to fully understand the occurrence of readmission of patients with diabetes. Subjective data can be captured by interviewing patients or by conducting surveys which will enrich the depth of patient information. The conversation between doctor and patient can also be collected and analyzed which could help to extract important features corresponding to patient's willpower and attitude by text-mining techniques. This information might improve the intelligent models to identify patients at high risk of readmission.

Given more time, we could run stacked models and neural networks to see if the AUC and recall is improved or not. We can also dive deeper into the most important features of the models to see which categories of the features are affecting the classification. We can also try to change the classification threshold for some models to see if they improve the performance, especially reducing the false positives to improve the precision score.

References:

- 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.

  Website: https://www.hindawi.com/journals/bmri/2014/781670/

- The Hospital Readmissions Reduction (HRR) Program, Center for Medicare and Medicaid Services.

  Website: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction- Program

- Diabetes 130-US hospitals for years 1999-2008 Data Set provided by Centre for Clinical and Translational Research, Virginia Commonwealth University.

  Website: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

- https://towardsdatascience.com/predicting-hospital-readmission-for-patients-with-diabetes-using-scikit-learn-a2e359b15f0