City University of New York

School of Professional Studies

**Hospital Readmissions: Prediction and Management of Readmission Risk (30 days risk model) for Diabetic Patients**

Nicholas Schettini & Niteen Kumar

Data 698 - Master Research Project

Professor Arthur O’Connor

# 

# 

# 

# 

# 

# 

# Table of Contents

Contents

[Table of Contents 2](#_Toc3841316)

[Introduction 3](#_Toc3841317)

[Background 3](#_Toc3841318)

[Data Source 4](#_Toc3841319)

[Hypothesis 4](#_Toc3841320)

[Approach and Methodologies 4](#_Toc3841321)

[Prior Research 7](#_Toc3841322)

[Assessment 25](#_Toc3841323)

[Tools 25](#_Toc3841324)

[References 26](#_Toc3841325)

# Introduction

Hospital readmissions of diabetic patients are one of the most common problems that in healthcare today. Hospital readmission rates are also used as an indicator of the quality of health care services.

A readmission can be defined in multiple ways, including:

* All-cause unplanned readmissions that happen within 30 days of discharge from the index (i.e., initial) admission.
* Patients who are readmitted to the same hospital, or another applicable acute care hospital for any reason.
* Readmissions to any applicable acute care hospital are counted, no matter what the principal diagnosis was.

Our study is focused on patients with diabetic background who keep returning to acute care hospitals as unplanned readmissions within 30 days.

According to the national healthcare data, readmission is also one of the main reason why our nation’s healthcare cost is so high. In 2012[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4439931/), the Affordable Care Act (ACA) established the Hospital Readmission Reduction Program (HRRP) as a way to incentivize hospitals to develop policies and procedures to reduce the number of readmissions per year within a hospital institution. Hospitals are penalized when patients are discharged and readmitted within 30 days. Readmissions are associated with negative patient and financial outcomes. [2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4439931/)

In this research project, we will build a model to predict readmission scenarios to minimize 30 days readmission risk. This model will have following benefits and advantages:

* Model can be used to find which patients have higher probability of readmission
* Model can help to understand different attributes that impact the outcome (readmissions)
* It can also help to bring down overall healthcare cost will also improve overall quality of healthcare service

# Background

Diabetes is one of the most prevalent medical conditions in population health. Human body uses a hormone called insulin to help control blood sugar. Diabetes is mainly caused due to insufficient production and secretion of insulin from the pancreas.

There are 2 types of diabetes:

* Type 1: The body does not produce enough insulin. This form of diabetes affects children and young adults.
* Type 2: The body produces insulin but does not use it properly to control blood sugar. Most people diagnosed with diabetes have this form of the disease.

More than 700,000 New Yorkers have diabetes, and nearly a third do not know they have it. Many people with diabetes have no symptoms. For others, symptoms may develop slowly over time or are so mild, they go unnoticed. Without proper care, diabetes can cause serious health problems like heart disease, blindness, kidney failure, and amputations of the legs or feet.

# Data Source

The UCI Machine Learning Repository: Diabetes 130-US Hospitals for years 1999-2008 Data Set contains 100000 instances of data, with 55 attributes. The dataset represents 10 years of clinical data at 130 US hospitals. The data contains attributes such as patient’s race, gender, age, time in hospital, number of lab tests performed, number of inpatient, outpatient, and emergency visits for prior year, etc. [3](https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008)

# Hypothesis

**Null Hypothesis (H0**): Patients demographics (age, gender, race) and medical indicators (medications, lab procedures, number of visits (inpatient and outpatient)) with after being discharged from the hospital do not have significant hospital readmissions within 30 days.

**Alternative Hypothesis (HA):** A significant number of patients readmitted to acute care hospital after being discharged within 30 days based on demographics (age, gender, race) and medical indicators (medications, lab procedures, and the number of visits (inpatient and outpatient)

# Approach and Methodologies

The below steps outline the approach for the research project:

1. Data Collection
2. Data Preparation and Data Wrangling
3. EDA (Exploratory Data Analysis)
4. Model building
5. Model evaluation and Model Selection
6. Data visualization and Dashboard

**Data Collection**: The data will be collected from the UCI web link. The project dataset is hosted on github and used directly for research purpose. In future, if needed, we also intend to use appropriate database (SQL/ NoSQL) in order to store clean and use case specific datasets in different tables that can be used for data visualization purpose.

**Data Preparation and Data Wrangling**: The dataset characteristics is multivariate and has both numeric and nominal data types. The dataset also has missing values which need to be handled using appropriate data wrangling techniques such as data impute or drop the attributes that no statistically significant. The data preprocessing is an important part of data analysis and really helps to identify features present in the dataset that can have significant impact on the model outcome.

Below table shows features available in the dataset, data type, description and missing values percentile. There are three features weight, payer code, medical specialty that have high percentage of missing values. So, these features can be safely ignored or dropped for model building models as they can't be treated as significant predictors. Also, race feature has 2% missing values whereas diagnosis code 1, diagnosis code 2 and diagnosis code 3 have less than 1% missing values.

| Feature name | Type | Description and values | % missing |
| --- | --- | --- | --- |
| Encounter ID | Numeric | Unique identifier of an encounter | 0% |
| Patient number | Numeric | Unique identifier of a patient | 0% |
| Race | Nominal | Values: Caucasian, Asian, African American, Hispanic, and other | 2% |
| Gender | Nominal | Values: male, female, and unknown/invalid | 0% |
| Age | Nominal | Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100) | 0% |
| Weight | Numeric | Weight in pounds. | 97% |
| Admission type | Nominal | Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available | 0% |
| Discharge disposition | Nominal | Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available | 0% |
| Admission source | Nominal | Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital | 0% |
| Time in hospital | Numeric | Integer number of days between admission and discharge | 0% |
| Payer code | Nominal | Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay | 40% |
| Medical specialty | Nominal | Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon | 49% |
| Number of lab procedures | Numeric | Number of lab tests performed during the encounter | 0% |
| Number of procedures | Numeric | Number of procedures (other than lab tests) performed during the encounter | 0% |
| Number of medications | Numeric | Number of distinct generic names administered during the encounter | 0% |
| Number of outpatient visits | Numeric | Number of outpatient visits of the patient in the year preceding the encounter | 0% |
| Number of emergency visits | Numeric | Number of emergency visits of the patient in the year preceding the encounter | 0% |
| Number of inpatient visits | Numeric | Number of inpatient visits of the patient in the year preceding the encounter | 0% |
| Diagnosis 1 | Nominal | The primary diagnosis (coded as first three digits of ICD9); 848 distinct values | 0.02% |
| Diagnosis 2 | Nominal | Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values | 0.35% |
| Diagnosis 3 | Nominal | Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values | 1.4% |
| Number of diagnoses | Numeric | Number of diagnoses entered to the system | 0% |
| Glucose serum test result | Nominal | Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured | 0% |
| A1c test result | Nominal | 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. | 0% |
| Change of medications | Nominal | Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change” | 0% |
| Diabetes medications | Nominal | Indicates if there was any diabetic medication prescribed. Values: “yes” and “no” | 0% |
| 24 features for medications | Nominal | 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 | 0% |
| Readmitted | Nominal | Days to inpatient readmission. Values: “<30” 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. | 0% |

The diagnosis codes present in the dataset are ICD-9 codes. ICD is important because it provides a common language for reporting and monitoring diseases. This allows the world to compare and share data in a consistent and standard way – between hospitals, regions and countries and over periods of time. It facilitates the collection and storage of data for analysis and evidence-based decision-making. On July 31, 2014 the U.S. Department of Health and Human Services (HHS) issued a rule finalizing October 1, 2015 as the new compliance date for health care providers, health plans, and health care clearinghouses to transition to ICD-10-CM/PCS. Since our dataset is before 2015, we will use ICD 9 specifications for research purpose.

|  |  |
| --- | --- |
| ICD Codes | Description |
| 001-139 | Infectious And Parasitic Diseases |
| 140-239 | Neoplasms |
| 240-279 | Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders |
| 280-289 | Diseases Of The Blood And Blood-Forming Organs |
| 290-319 | Mental Disorders |
| 320-389 | Diseases Of The Nervous System And Sense Organs |
| 390-459 | Diseases Of The Circulatory System |
| 460-519 | Diseases Of The Respiratory System |
| 520-579 | Diseases Of The Digestive System |
| 580-629 | Diseases Of The Genitourinary System |
| 630-679 | Complications Of Pregnancy, Childbirth, And The Puerperium |
| 680-709 | Diseases Of The Skin And Subcutaneous Tissue |
| 710-739 | Diseases Of The Musculoskeletal System And Connective Tissue |
| 740-759 | Congenital Anomalies |
| 760-779 | Certain Conditions Originating In The Perinatal Period |
| 780-799 | Symptoms, Signs, And Ill-Defined Conditions |
| 800-999 | Injury And Poisoning |
| V01-V91 | Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services |
| E000-E999 | Supplementary Classification Of External Causes Of Injury And Poisoning |

# Prior Research

Jamei, M., Nisnevich, A., Wetchler, E., Sudat, S., & Liu, E. (2017, July 14). Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5510858/>

The study attempts to predict the all-cause risk of 30-day hospital readmissions using neural networks.  The study samples electronic health record data from 20 hospitals across a health-network called Sutter Health, located in Northern California. The researches extracted about 324k patient rows from the organization without sensitive patient data.   The researchers looked at variables such as discharge time, insurance, race, LACE Score, previous emergency visits, inpatient visits, age, admission type, admission source, and so on.   The study used data from in-patient only visits and excludes outpatient visits.  The researches also mapped the patient data using Google’s Geocoding API to determine coordinates of each patient home, using 2010 census data.

The study mentions multiple models were used in their initial investigation, including logistic regression, random forests, and neural networks – with neural networks heavily outperforming the others. The type of neural net that was the best-performing was a two-layer network, with a dense hidden layer and dropout nodes between other layers to prevent overfitting.  The model was initially trained on all features from the dataset (1667 features), eventually being tuned to use a subset of the features (top 100), which showed a reasonable amount of near-optimal performance.

Shameer, K., Johnson, K. W., Yahi, A., Miotto, R., Li, L. I., Ricks, D., . . . Dudley, J. T. (2016). PREDICTIVE MODELING OF HOSPITAL READMISSION RATES USING ELECTRONIC MEDICAL RECORD-WIDE MACHINE LEARNING: A CASE-STUDY USING MOUNT SINAI HEART FAILURE COHORT. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5362124/>

The study mentions a successful use of a neural network model to predict the risk of patients being readmitted within 30 days of discharge.

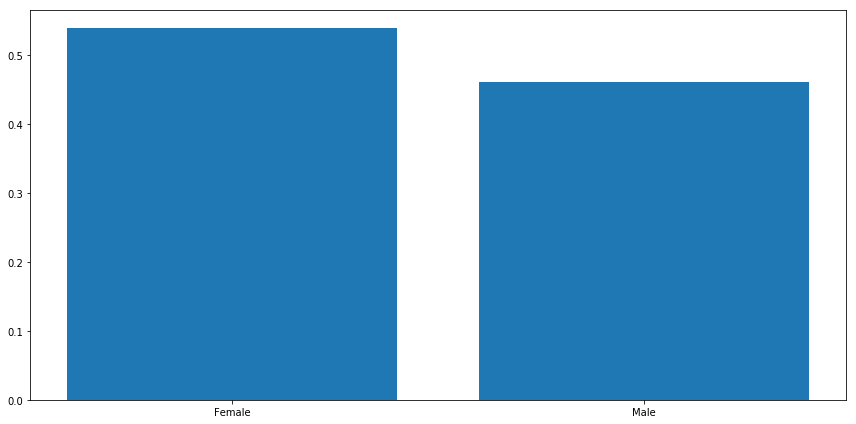
The study attempts to reduce the number of readmissions of patients with heart disease. Like the previous studies investigated, and our own dataset, the research team used HIPPA and PHI adherence during their study, as well as machine learning techniques. The study consisted of 1,068 patients admitted tom the Mount Sinai Heart system during the year of 2014. Each patient had a principal diagnosis of heart failure. While the study being used in our investigation focuses on patients with diabetes, the techniques used in this study will be useful.

Machine learning models were used, including Naïve Bayes – using a 70/30 split of the dataset, using python libraries such as scikit-learn. The team breaks the classification down to patients that were “RA” or “Readmitted” or “NonRA” – “Not readmitted”. In our investigations, we’re performing a similar approach by mapping readmit patients to 1 (readmit) and 0 (not readmit).

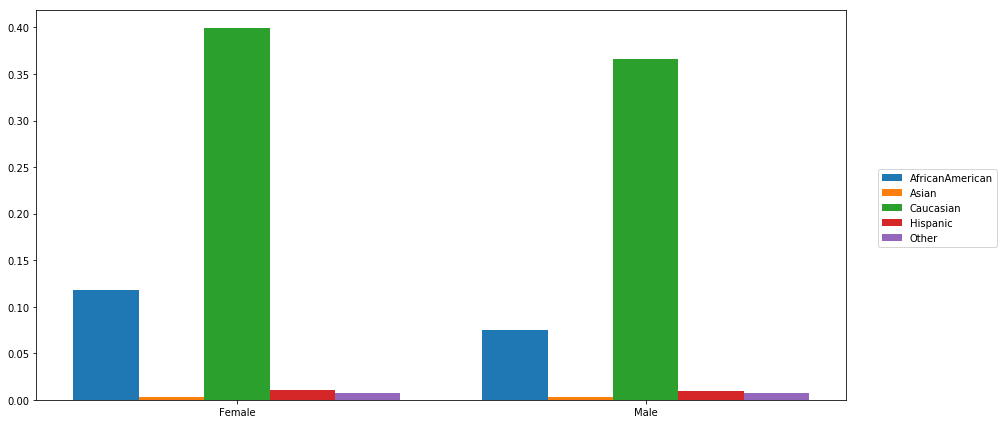
The results from the study mention that their approach provides better accuracy than previous literature on the topic – using only one year or admission data. The article mentions that to improve accuracy, more years of data should be used in the model.

**EDA (Exploratory Data Analysis)**: Once the data is cleaned, we’ll use EDA techniques to gain more insight into underlying data. We will use different plots and charts to view and find statistically significant attributes such as:

Patient Demographics: The below plot shows that number of female patients are more than male patients. Female population is 54% of the total population and male population is 46%. There are 3 patients for which gender information is unknown.



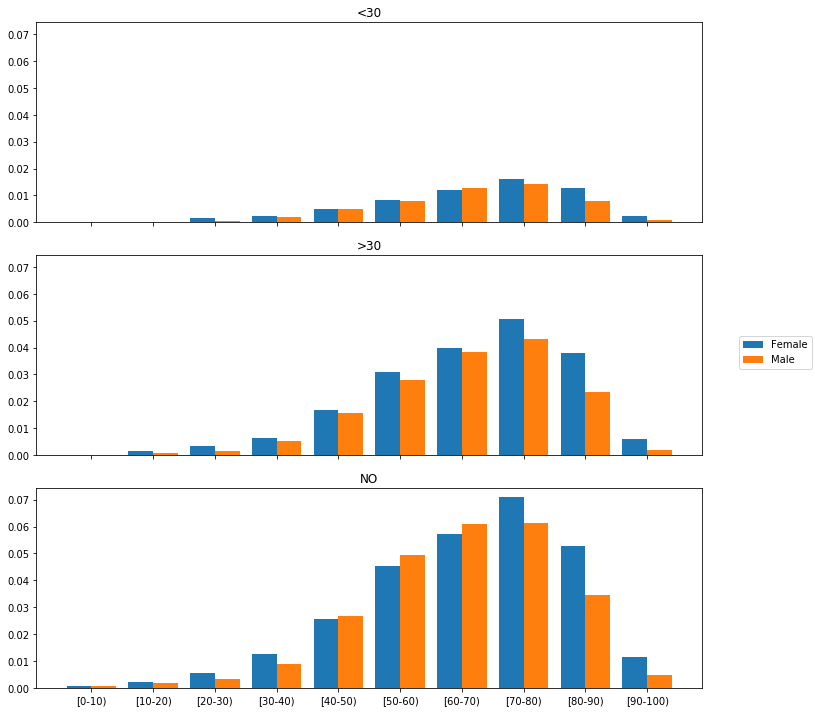
The data distribution for race and gender features indicate that both male and female population for caucasian patients are greater than 35%. The number of female patients appear to be high in all races.



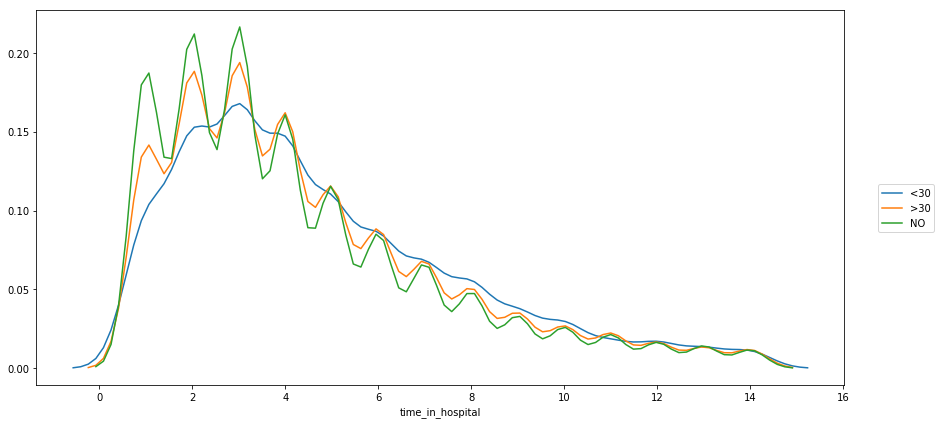
From the above plot it can be inferred that the percentage of caucausian is higher at this location. Below is the tabular representation of the race and population percentile present in the dataset. African americans make up 19% of overall patient population. Hispanic and asian population is also pretty low.

|  |  |
| --- | --- |
| Race | Population % |
| Caucasian | 75 % |
| African American | 19% |
| Hispanic | 2% |
| Asian | < 1% |
| Other | 1.5% |

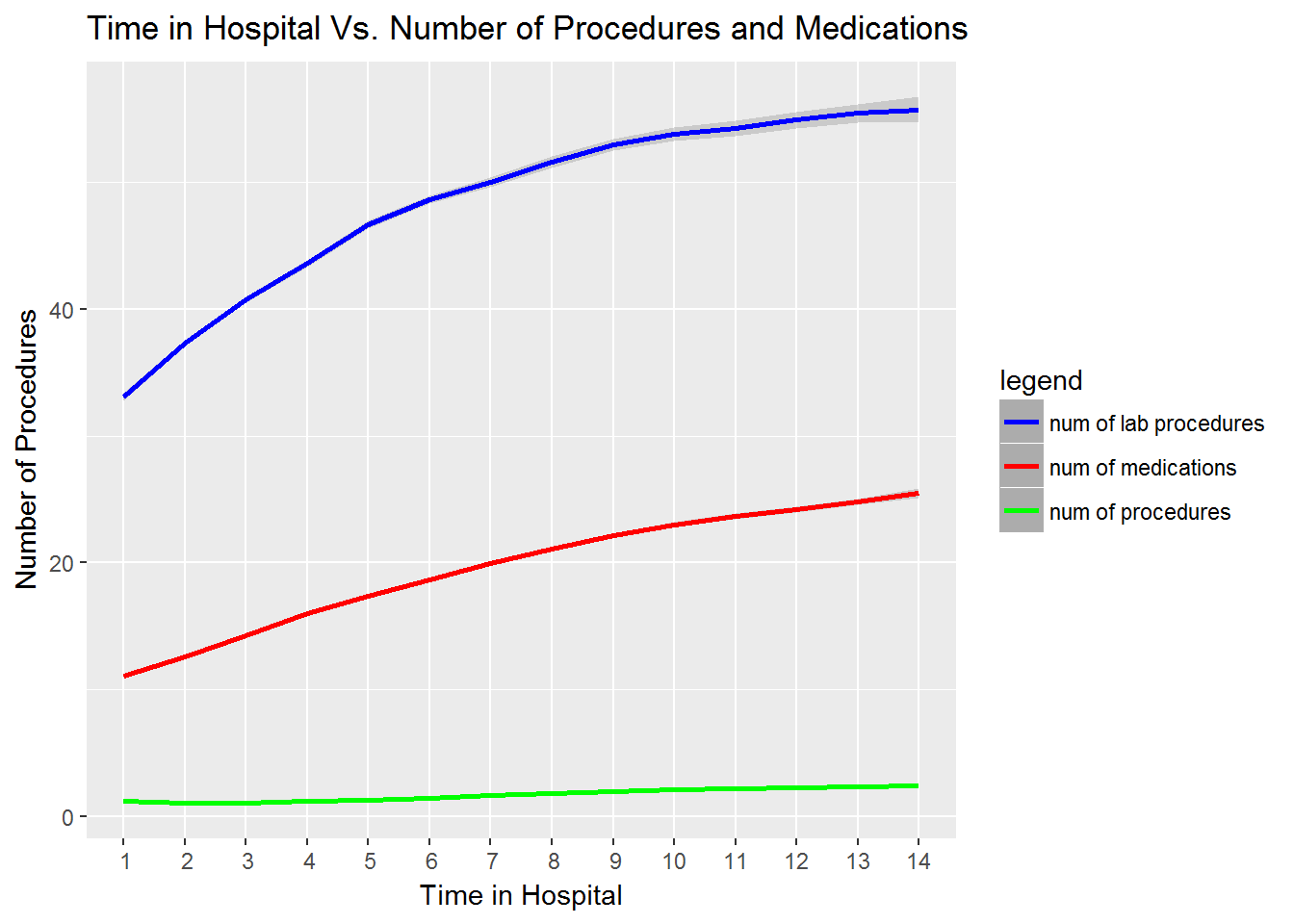
We also tried to find the corelation between race, gender, age and how these features relate to the reponse variable ‘readmitted’. It appears that number of patients are lowest who got readmitted in less than 30 days. Readmission rate appears to be highest for both male and female for more than 30 days period. There are singnificant number of patients who do not return to the hospital. The readmission data for the age group between 40 to 90 appears to have normal distribution and peaking for the age group 70-80 for both more than 30 days and less than 30 days period.



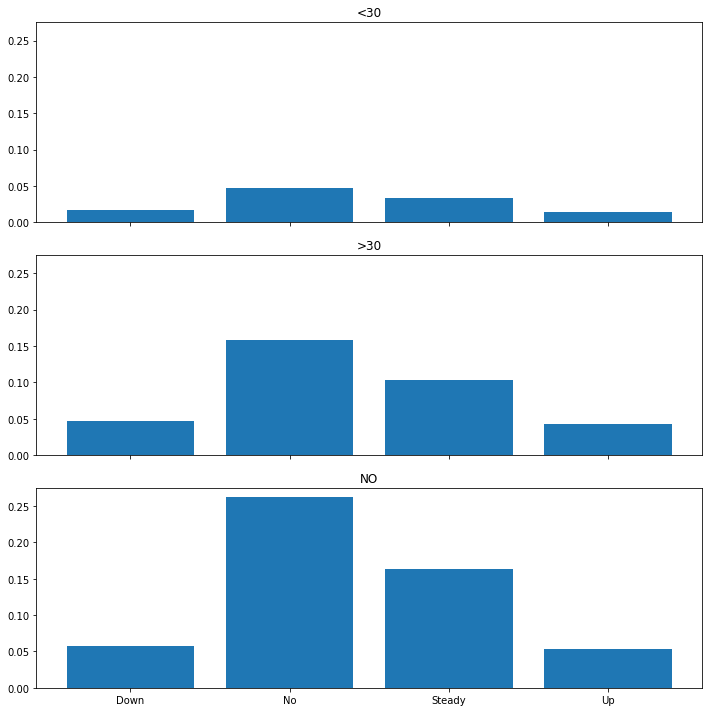
It also appears that patients who are discharged within 2 to 5 days are most likely to be readmitted. The plot is right skewed and the number of patients who do not come back also appear to be released within 2 to 5 days. So, the length of stay appears to be one of the predictor for the readmission but there are definitely other predictors available in the dataset which have larger impact on the readmission outcome.



Both features number of lab procedures and number of medications have upward trend w.r.t. length of stay in the hospital. We can also notice that number of procedures remain low as the patient stays longer in hospital. So, both number of lab procedures and number of medications can play significant role for readmission cases.

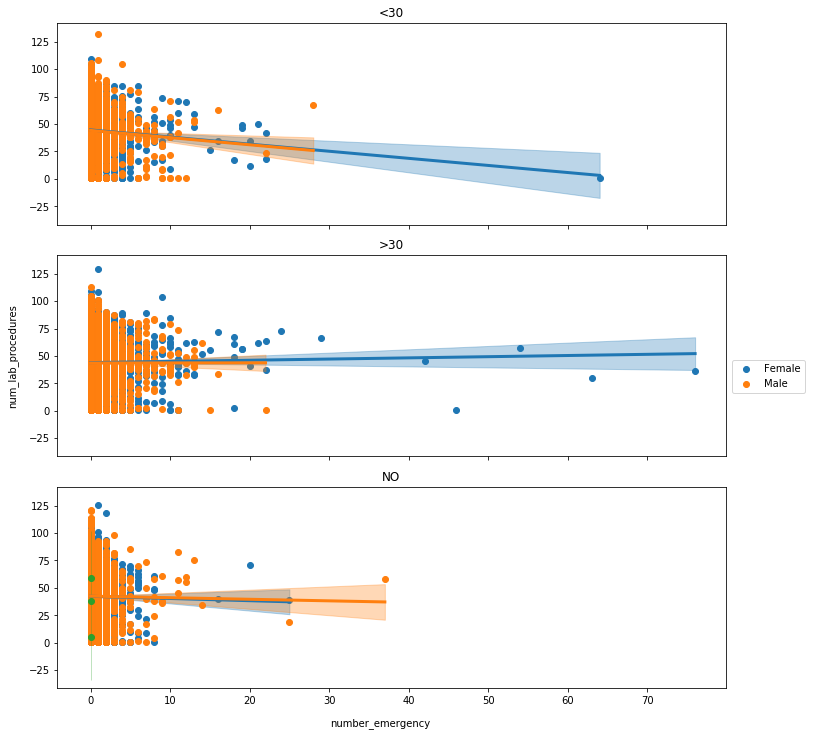


**Bar Plot** - Med – Insulin Vs readmitted



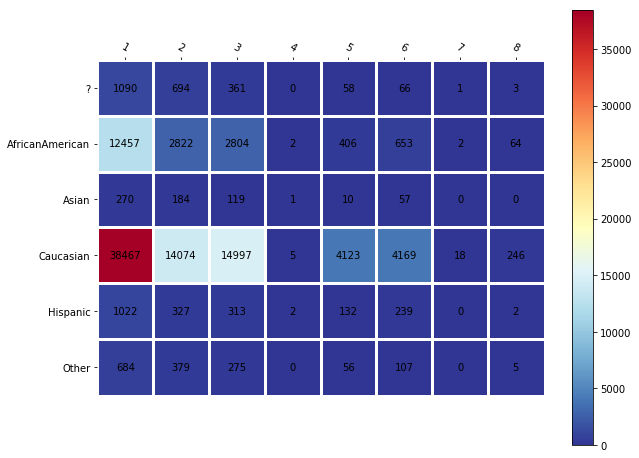
The above graph shows the various insulin levels for patients that were admitted <30 days, > 30days, and not readmitted. In the patients readmitted <30 days, the data shows an almost flatline of the different insulin levels, with “No” as being slightly higher than the others. In patients readmitted >30 days, the data shows “No” being the highest recorded, followed by steady. In patients not readmitted/ “No”, “No” appears as the highest recorded insulin, followed by steady.

**Joint plot** - Regression fit for – number of emergency , number of lab procdures, gender and readmitted response label.

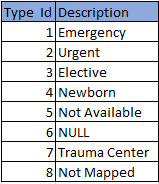


The above graphs show the number of emergency visits vs. the number of lab procedures by gender, with a regression line. Overall, the number of lab procedures seems to occur within the first few visits to the emergency department. One potential reason of this could be that certain procedures have a limit of when they can be performed – MRI, CT, X-RAYs. These exams expose the patient to radiation, so exams of these types must be regulated.

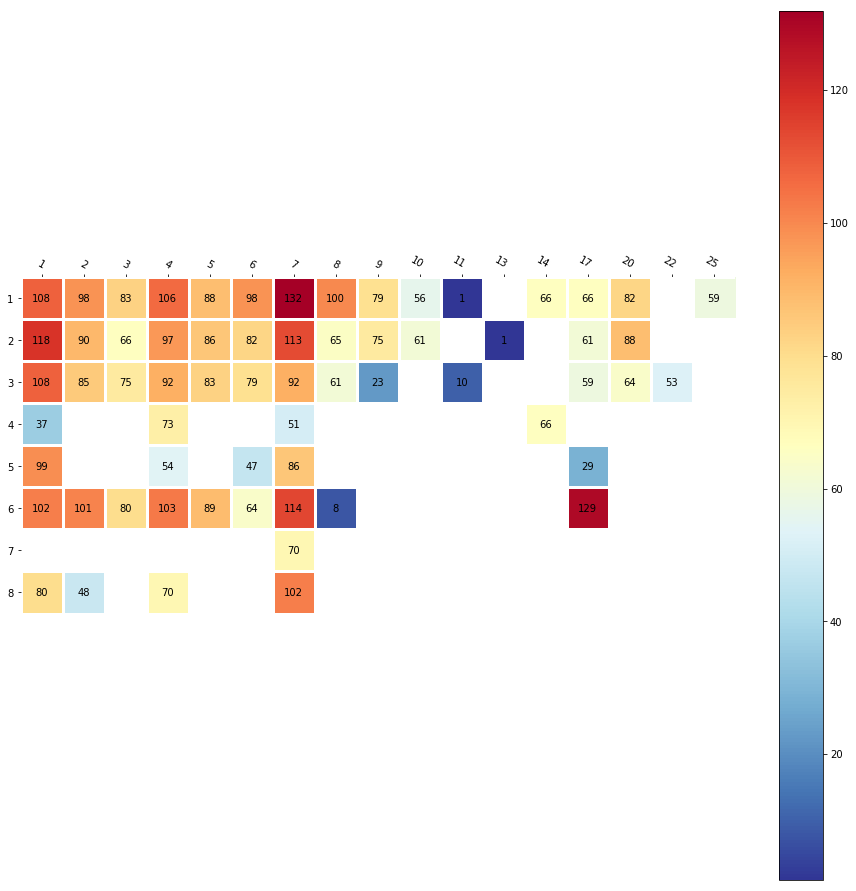
**Heat map** – admission source id and race



The above heatmap shows the number of admission source ID vs race. The higher the count of admission source, the redder they appear in the heatmap. The less the admission source, the bluer they appear. According to this heatmap, the reddest square is for Caucasian with an admission source ID of 1 – which corresponds to Emergency visits. The below table shows more information on admission source mapping.



**Heat map** –Admission Type vs. Admission Source



The above graph shows the admission type vs admission source. The redder the square, the more occurrences. The bluer the box, the less occurrences. The below tables show a legend for the different column and row names (numbers).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Description | ID | Description | ID | Description |
| 1 | Physician Referral | **11** | Normal Delivery | **21** | Unknown/Invalid |
| 2 | Clinic Referral | **12** | Premature Delivery | **22** | Transfer from hospital inpt/same fac reslt in a sep claim |
| 3 | HMO Referral | **13** | Sick Baby | **23** | Born inside this hospital |
| 4 | Transfer from a hospital | **14** | Extramural Birth | **24** | Born outside this hospital |
| 5 | Transfer from a Skilled Nursing Facility (SNF) | **15** | Not Available | **25** | Transfer from Ambulatory Surgery Center |
| 6 | Transfer from another health care facility | **16** | NULL | **26** | Transfer from Hospice |
| 7 | Emergency Room | **17** | Transfer From Another Home Health Agency |  |  |
| 8 | Court/Law Enforcement | **18** | Readmission to Same Home Health Agency |  |  |
| 9 | Not Available | **19** | Not Mapped |  |  |
| 10 | Transfer from critial access hospital | **20** | Normal Delivery |  |  |

**Model Building**: In this project, we will attempt to classify and display patients that are at risk of being readmitted to a hospital unit after discharge. Different modeling techniques will be used in our readmission predictions. Model results will be evaluated and compared to each other in order to select the most accurate technique.

Based upon the initial understanding and preliminary analysis of dataset, we are intending to evaluate below models but not limited to:

**Data Preprocessing:** In order to apply machine learning algorithms, we need to preprocess the data so that it can properly fit into the model.

* We already have dropped features: weight, medical specialty and payer code
* The primary diagnosis codes diag1, diag2 and diag3 have very small number of missing values, so we are dropping the records where are all 3 diagnosis codes are missing
* Also, we are primarily interested in readmissions and the significant features impacting the outcome of the model, so it makes sense to drop patients who have expired and that can be determined by feature discharge\_disposition\_id
* One of the goals is to focus on 30 days readmission risk model, so patients that are readmitted after 30 days can be treated as 'NO' in response 'readmitted'
* Map diagnosis codes according to ICD 9 specifications
* There are various features available in the dataset. We would be considering following features for model and normalizing them to fit into the model. The idea is to start with the simplest set of features and then focus on creating composite features and other feature engineering techniques to enhance the model
  + time\_in\_hospital
  + num\_lab\_procedures
  + num\_procedures
  + num\_medications
  + number\_outpatient
  + number\_emergency
  + number\_inpatient
  + number\_diagnoses

**Modeling Techniques:** One of the main goals is to understand how a particular set of features perform with different models. The model accuracy is important, but we would also focus on interpretability of the model.

We randomized (to avoid any selection bias) and split the pre-processed clean data into Training and Test Data set, in an 80:20 ratio, which allowed us to train our models on 80% of the data and use the other 20% to assess the performance of our models. Moreover, we used multiple fold cross-validation on our training data to ensure that our models were both scalable and robust and were not over-fitting on the training data. We also ensured that the models were trained and evaluated on the same data to ensure performance comparison across models for the exact same data being trained and evaluated upon.

We intend to evaluate below models:

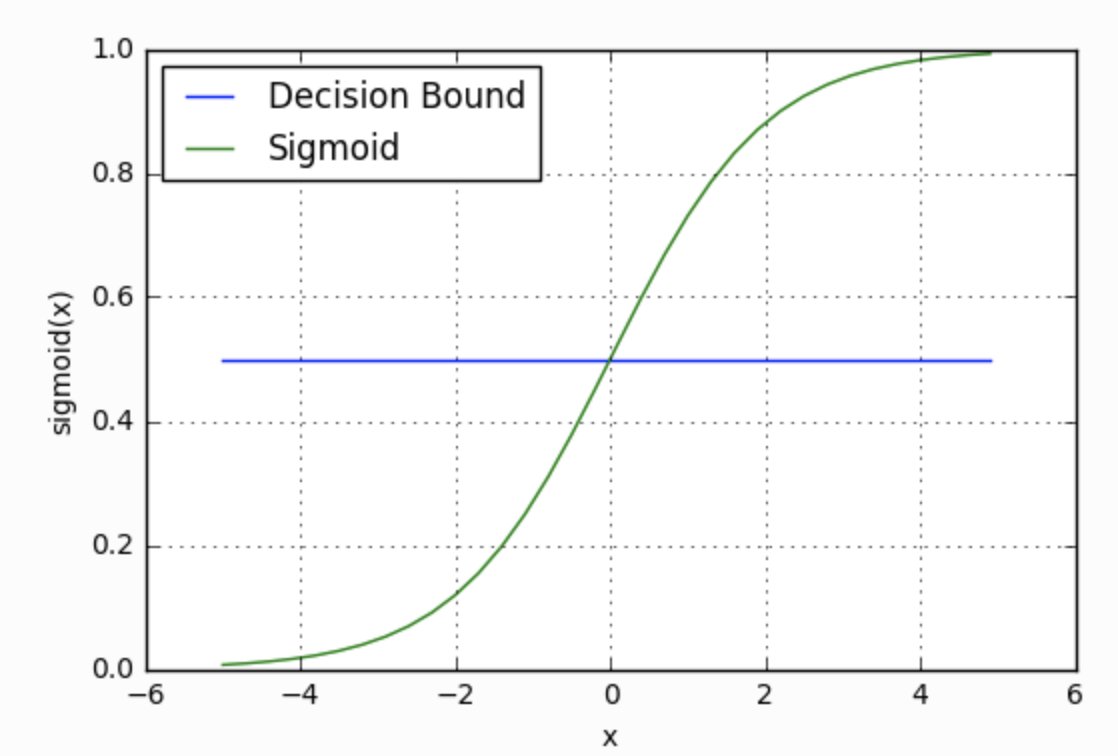
* Logistic Regression
* Decision Trees
* Random Forest
* KNN (K Nearest Neighbors)
* Gradient boost Classification
* Neural networks

**Logistic regression** is widely used classification technique used to assign observations to discrete cases. [6](https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html) Logistic regression provides observations such as: pass/fail, hot/cold, small/large, yes/no. The model works by selecting what's known as a Decision boundary - a type of benchmark or selection criteria that classifies your observations. For example, if we set the decision boundary to:

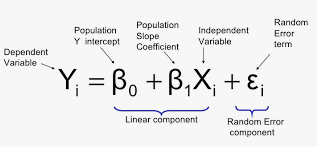
P >/ 0.5, class = yes

P < 0.5, class = no

A resulting value above or below the boundary line will determine the classification.



[fig. 2](https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html)



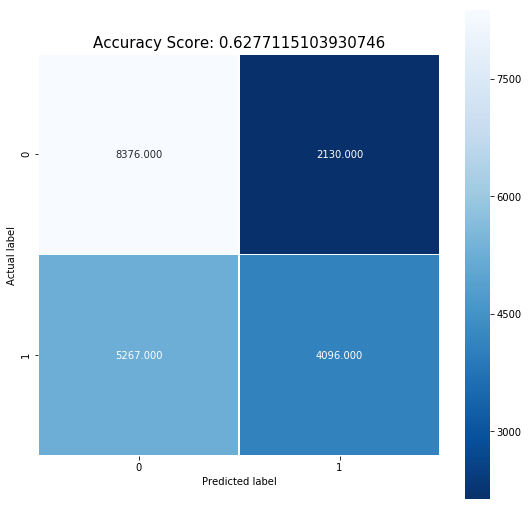
The logistic regression model from sklearn model is used. We chose high tolerance value as stopping criteria and small c value for stronger regularization.

Below is the summary for classification:

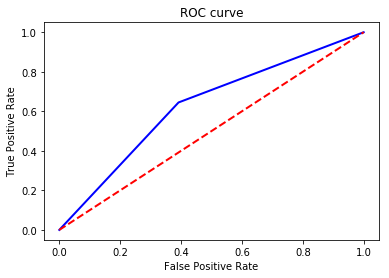
**Approach 1.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | f1-score | Support |
| 0 | 0.61 | 0.80 | 0.69 | 10506 |
| 1 | 0.66 | 0.44 | 0.53 | 9363 |
| Avg./ total | 0.63 | 0.63 | 0.61 | 19869 |

The accuracy score of the model is 62%



Also, below is the ROC curve which shows false positive rate and true positive rate



True Positive Rate (TPR): This is the percentage of patients predicted as readmitted. It’s the number of TP divided by the number of purchased items (TP + FN).

False Positive Rate (FPR): This is the percentage of patients predicted as not readmitted. It’s the number of FP divided by the number of not purchased items (FP + TN)

Two accuracy metrics are as follows:

* Precision: This is the percentage of patients that have been predicted as readmitted. It’s the number of FP divided by the total number of positives (TP + FP).
* Recall: This is the percentage of patients that have been not predicted as readmitted. It’s the number of TP divided by the total number of purchases (TP + FN).

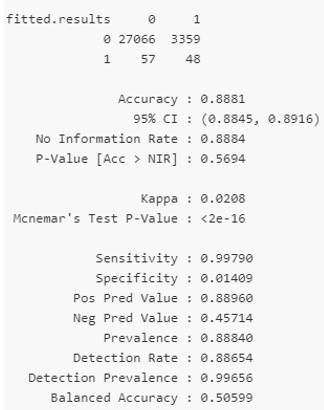
**Approach 2.**

After extracting significant values from running a logistic model with all factors, some of the more significant values were extracted to be used in the model, as shown below:

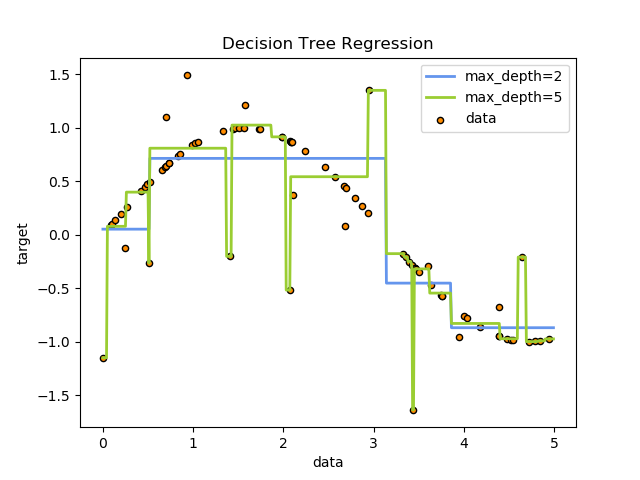
* Age
* Gender
* Race
* Admission\_source\_id
* Insulin
* Time in Hospital
* Change
* Num lab procedures
* Num inpatient
* Num outpatient
* Num Diagnoses
* Num Medications
* Num Emergency

The data was then mapped to add a readmission indicator to a new column. This allows the model to attempt to predict the outcome of our response variable. 

After creating the new column ‘readmit’, the data was randomly broken up into a 70/30 split. This allows for the evaluation of the performance of the model using one data set. This create a train/test split. After running the model, the logistic regression output confusion matrix is below:



**Decision Trees** are one of the most commonly machine learning models. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. One advantage of tree-based methods is that they have no assumptions about the structure of the data and are able to pick up non-linear effects if given sufficient tree depth. We can fit decision trees using the following code. For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.

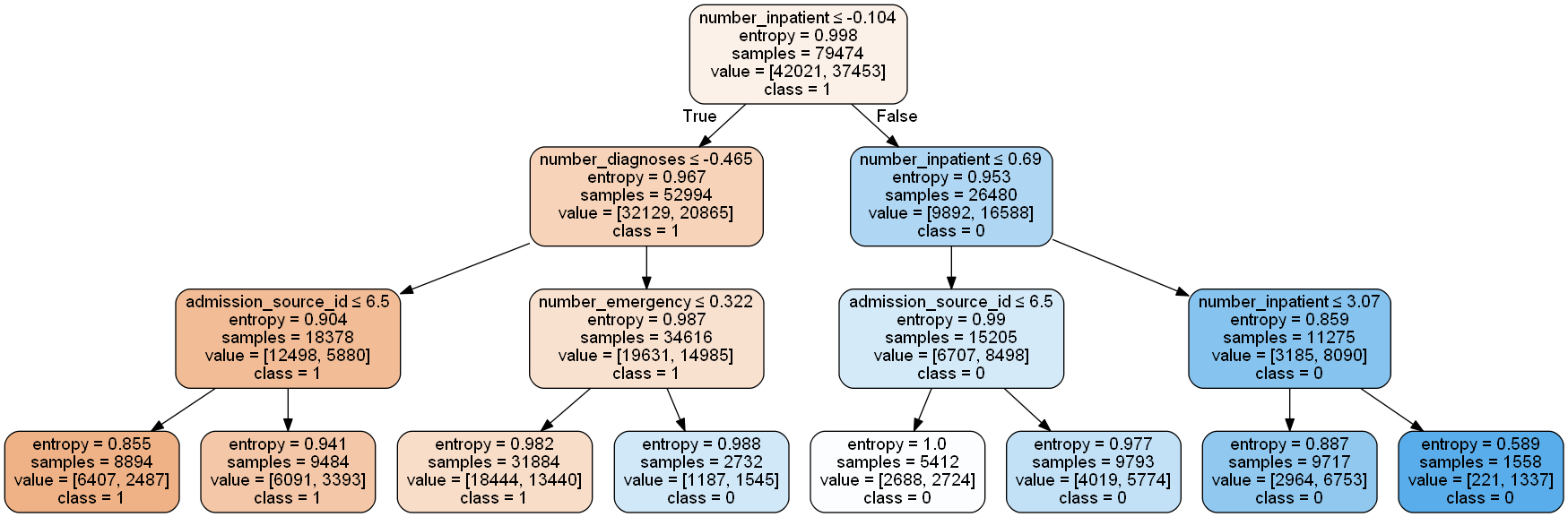


[fig. 3](https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html)

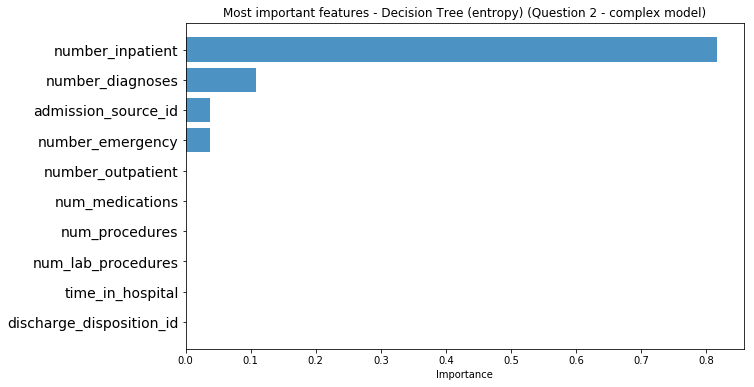
In the decision tree approach, we used 'entropy' as criterion and used 3 as max depth of the tree. The minimum sample leaf was 5 and random state parameter was set to 100.

The accuracy of the decision tree model is 62.35 %

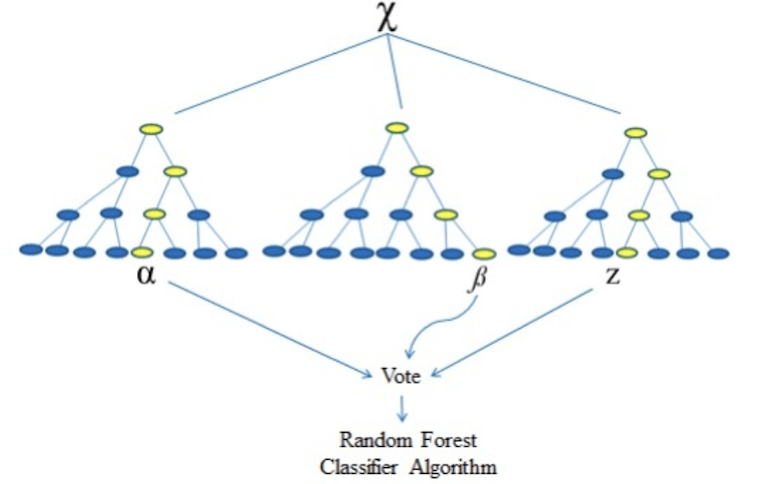
The below decision tree diagram shows the decision tree of depth 3.



Also, below are the most important features of decision tree (entropy). number of inpatients has the largest impact on the outcome of the model. Surprisingly, number of lab procedures and length of stay in the hospital do not appear to be significant in decision tree of depth 3.

****

**Random forests** are categorized as a supervised learning algorithm, used to make predictions in both classification and regression problems. [4](https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd) A random forest works by creating multiple Decision Trees and merging the results together - this creates a more accurate prediction. [5](https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd)

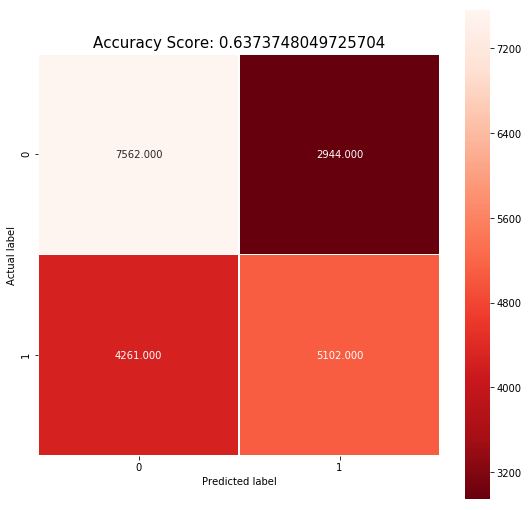


[fig. 4](https://www.rxdatascience.com/blog/machine-learning-for-pharma-using-random-forest)

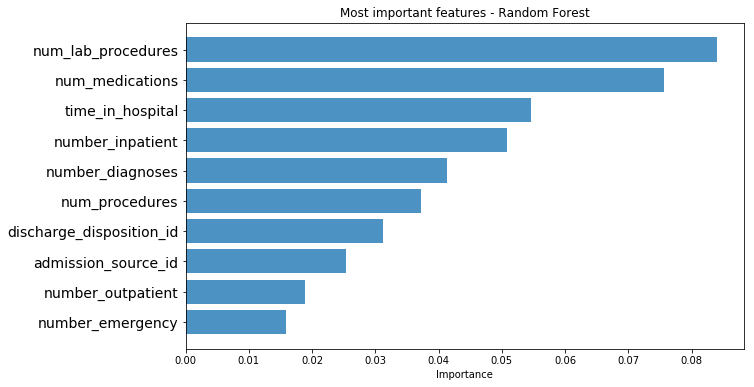
Below is the summary for Random forest classifier:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | f1-score | Support |
| 0 | 0.41 | 0.72 | 0.68 | 10506 |
| 1 | 0.63 | 0.54 | 0.59 | 9363 |
| Avg./ total | 0.64 | 0.64 | 0.63 | 19869 |

Below is the confusion matrix of random forest with accuracy score of 63.73%

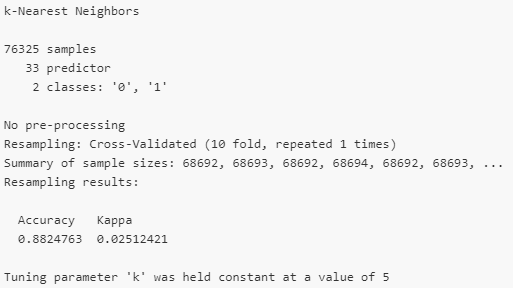
****

Most important features of random forest is as shown below. The number of lab procedures appear to have largest impact on the outcome followed by number of medications and length of stay in hospital.

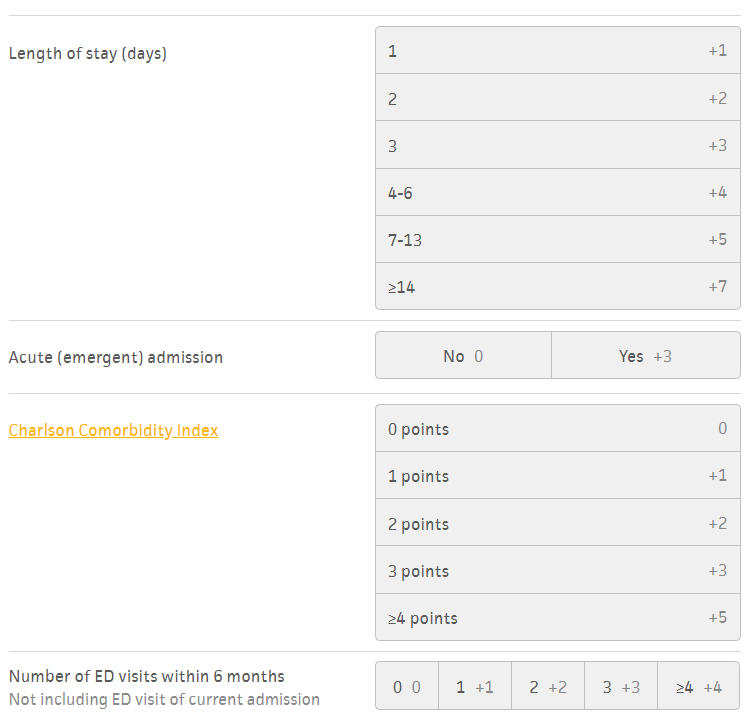


**KNN** is one of the simplest machine learning models for datasets for which we need to perform classification and clustering techniques. The given dataset has hundred thousand data points and for a given sample point, the model looks at the k closest data points and determines the probability by counting the number of positive labels divided by K. This model is easy to implement and understand but comes at the disadvantage of being sensitivity to K and takes a long time to evaluate if the number of trained samples is large. Below is one the example of KNN machine learning model.

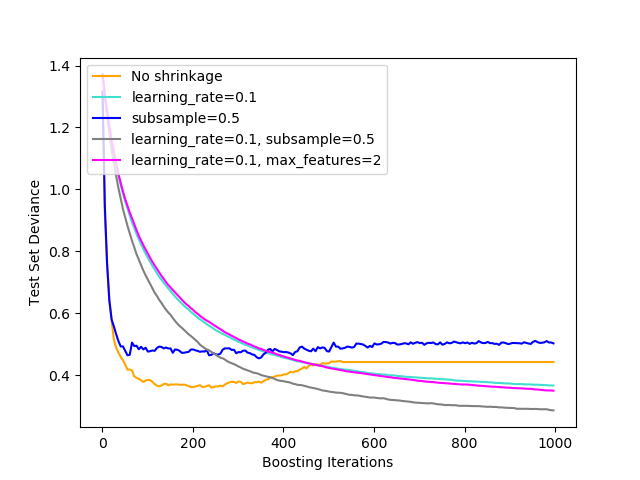
The data is first manipulated using a library called “mlr”. This allows easy creation of dummy variables. Dummy variables allow categorical variables to be inputted into the model. Due to performance issues on the local PC – the number of repeats and tune length had to be adjusted to 1. The below chart shows the KNN output.



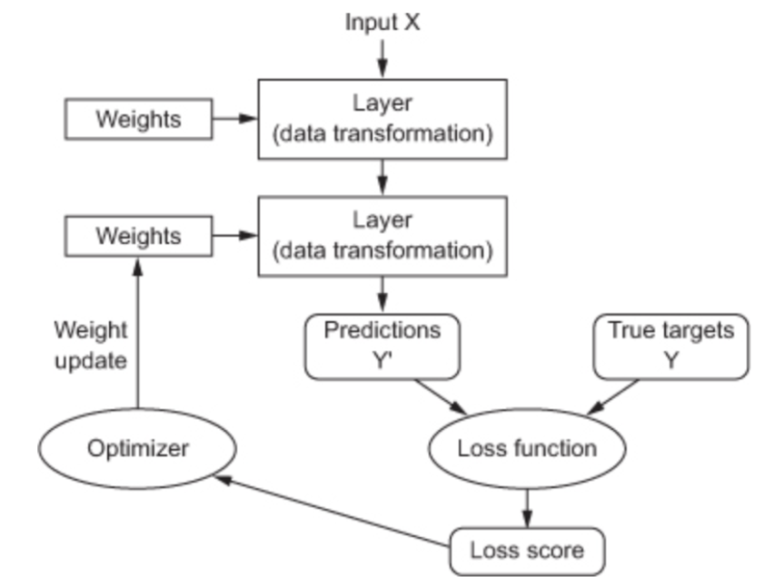
**Future Steps**

* We will continue to evaluate other important features of the dataset such as medications provide, change in medications and a1c test level
* LACE score evaluation for better predictions
* Other models, as listed below, need to be evaluated and compared

**Gradient boosting classifier** is another approach to improving decision trees is using a technique called boosting. In this method, you create a bunch of shallow trees that try to improve on the errors of the previously trained trees. One model that uses this technique paired with a gradient descent algorithm (to control the learning rate) is known as gradient boosting classifier. To fit gradient boosting classifier, we can use the following code.

****

**Neural networks** are a set of algorithms used in both unlabeled and labeled classification and regression problems. Neural networks are made up of multiple layers, called nodes. The more layers, the more depth the model has. Each layer contains a weight - a bunch of numbers. These weights are assigned random values at first, which the model then implements a series of random transformation. To measure the accuracy of the model, a loss score is calculated based on the distance the observed value is from the true target. After each iteration (training loop), the weights are slightly modified, which decreases the loss score. [7](https://livebook.manning.com/#!/book/deep-learning-with-python/chapter-1/55)



[Fig 6](https://livebook.manning.com/#!/book/deep-learning-with-python/chapter-1/55)

**Model Evaluation and Model Selection:** After trying several models we need to come up with baseline model and depending on the learning curve, below strategies will be considered in order to improve the model performance:

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

**Data Visualization and Data Dashboard**: Once the model is selected, we will end to end pipeline. The model will be accessed through an interactive user friendly UI/UX. The model information will be displayed on a dashboard, created in either R (Shiny) or Python (Flask/Dash).

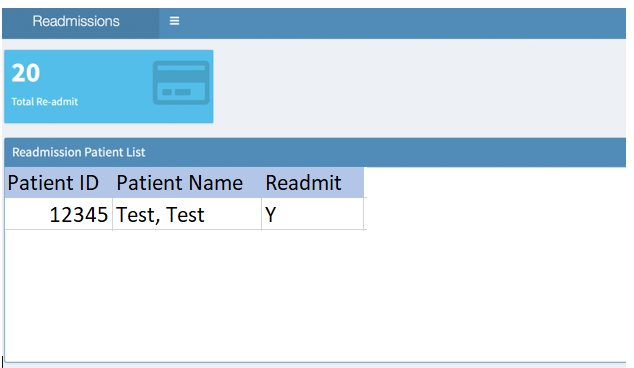


Fig: Sample dashboard created in Shiny

# Assessment

To assess the effectiveness of the predictive readmission model, we will judge the model on different success criteria, such as: the tools ability to predict if a patient is likely to be readmitted to a hospital within 30 days.

# Tools

The machine learning methodologies considered to help create a predictive model for this use case include: Random Forest [8](https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd), Logistic Regression [9](https://www.statisticssolutions.com/what-is-logistic-regression/), KNN, GBC, and neural networks from Kears [10](https://keras.io/), using both Python and R; and their various libraries (numpy/pandas, tidyverse). We will have to consider our machines processing power, and the amount of data points available (especially for use of a neural network).

# References

* <https://www.nychealthandhospitals.org/services/diabetes-care/>
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4439931/>
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4439931/>
* <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>
* <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>
* <https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html>
* <https://livebook.manning.com/#!/book/deep-learning-with-python/chapter-1/55>
* <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>
* <https://www.statisticssolutions.com/what-is-logistic-regression/>
* <https://keras.io/>
* <https://www.rxdatascience.com/blog/machine-learning-for-pharma-using-random-forest>
* <https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html>
* <https://livebook.manning.com/#!/book/deep-learning-with-python/chapter-1/55>
* <https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html>