# Predicting the “Re-admission possibility of a patient into the hospital”

## Abstract:

Hospital readmissions are expensive and reflect the inadequacies in healthcare system. Early identification of patients facing a high risk of readmission can enable healthcare providers to to conduct additional investigations and possibly prevent future readmissions. This not only improves the quality of care but also reduces the medical expenses on readmission.

Machine learning methods have been leveraged on public health data to build a system for identifying diabetic patients facing a high risk of future readmission. Number of laboratory tests, discharge disposition and admission type were identified as strong predictors of readmission. These insights can help healthcare providers to improve inpatient diabetic care.

## Introduction:

A dataset containing medical records of 34650 diagnosed with diabetes, in train dataset. There are 44 Independent features in the initial dataset, one dependent feature. 20 feature out of 44 independent features related to drugs. The distribution of class variable is 13.7% for readmitted within 30 days and 86.3% for Not admitted.

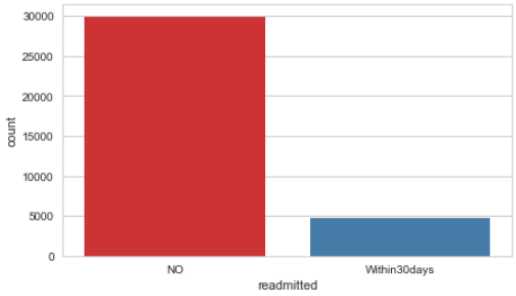



*Figure: Class Imbalance*

## Methodology:

Starting from the scrach, we first do Hypothesis generation, Data Exploration, preprocessing, understand feature importance, modelling, interpretation and Analysis. We could also understand on what futher developments can be done for the analysis.

## Data Exploration:

We start with missing values. Few columns in dataset has many missing values. Here are the columns and their missing values percentage.

- Weight -> 96%
- Medical Specialty -> 47%
- Payer Code -> 42.5%

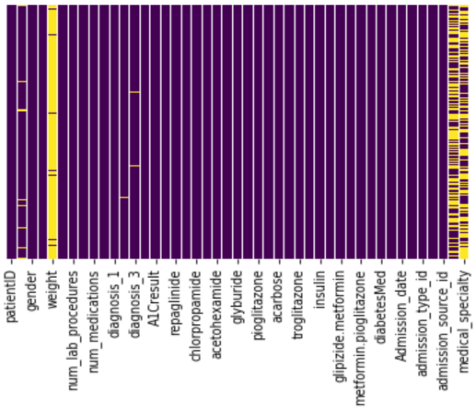



*Figure: Missing values Visualization*

Now, let us examine the pair-plot of the Numeric variables in our dataset.

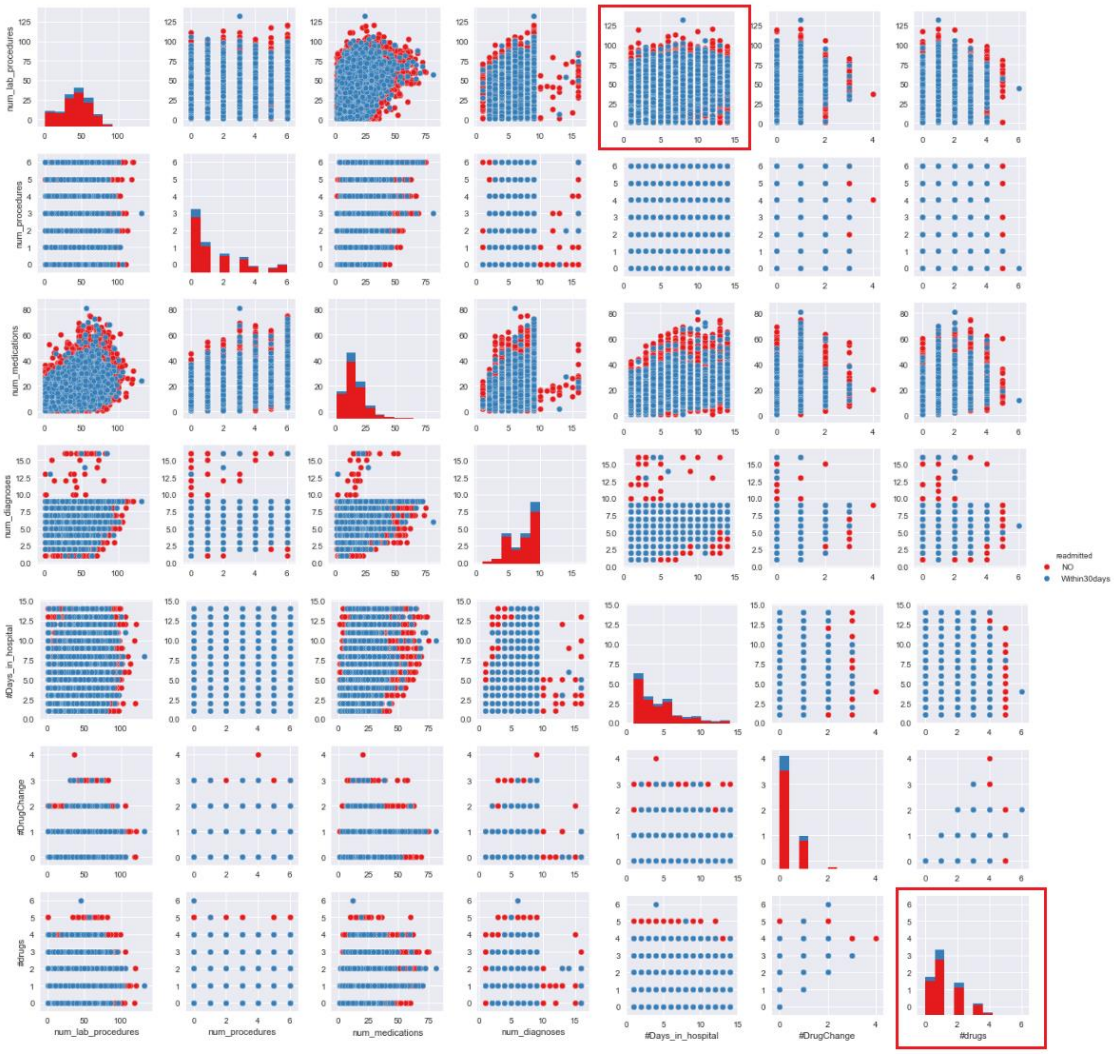



*Figure: Pair-Plot*

From pair-plot, patients who are many drugs are likely to get readmitted.

Here are few other visualizations of our data.

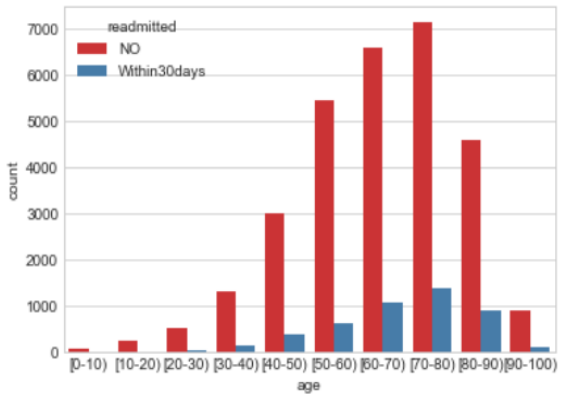



*Figure: Age Vs Readmission*

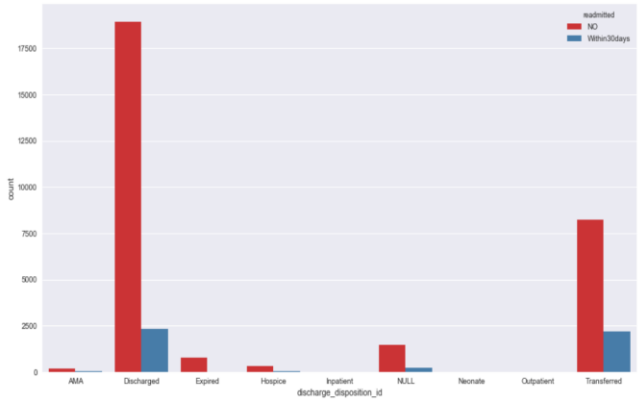



*Figure: Discharge_dispositio_id Vs Readmission*

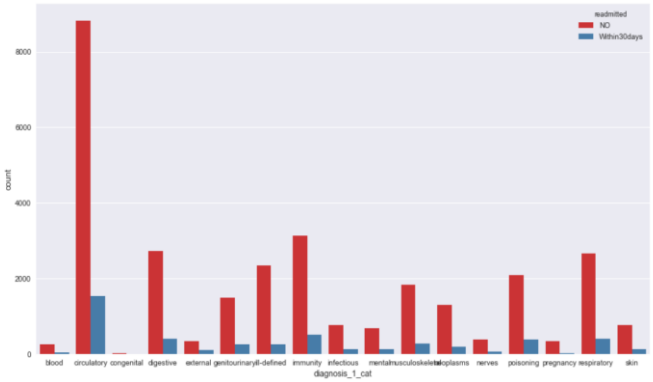



*Figure: Dignosis_1_cat Vs Readmission*

From these plots, it is clear that

- Patient numbers dramatically increase with age, and reaches a peak in the range of 70-80 years old.
- Patients in discharge_decomposition category, 'Discharged' and 'Transfered' could get readmitted.
- Patients in diagnosis_1 category, 'Circulatory' and 'Immunity' could get readmitted.

## Data Preprocessing

Preprocessing steps are given below

- To manage the level discrepancies, Train and test sets are merged together and preprocessed.
- Weight, Payer Code and Medical Speciality has high missing value. Atleast, around 50% of data in those columns is missing.
- Days difference between Discharged and Admission dates is considered Days in Hospital.
- Diagnosis_1, Diagnosis_2, Diagnosis_3 column codes are grouped into meaningful categories as per ICD-9 codes.
- Levels of age, discharge_disposition_id, admission_type_id and admission_source_id are condensed to meaningful levels.
- IDs and Dates are dropped.

## Feature Engineering

Feature Engineering steps done are as follows.

#Drugs: Is derived out of drugs used on patient during treatment.

#DrugChange: Is the drug dosage change, i.e. number of Ups or Downs of drugs used on patient.

## Handling Class Imbalance

- SMOTE

  Synthetic Minority Oversampling Technique is a statistical technique for increasing the number of cases in the dataset in a balanced way. The module works by generating new instances from existing minority cases that you supply as input.

- Manual Up-sampling by duplicating records of Minority Class.

  As a consequence of applying SMOTE, new dataset is generated where the columns names are no longer retained. As a result, explicability of feature importance after building the model is lost.

  To overcome this issue, manually minority class records are duplicated to match majority class records.

(Minority class records are duplicated 6 times to match majority class records)

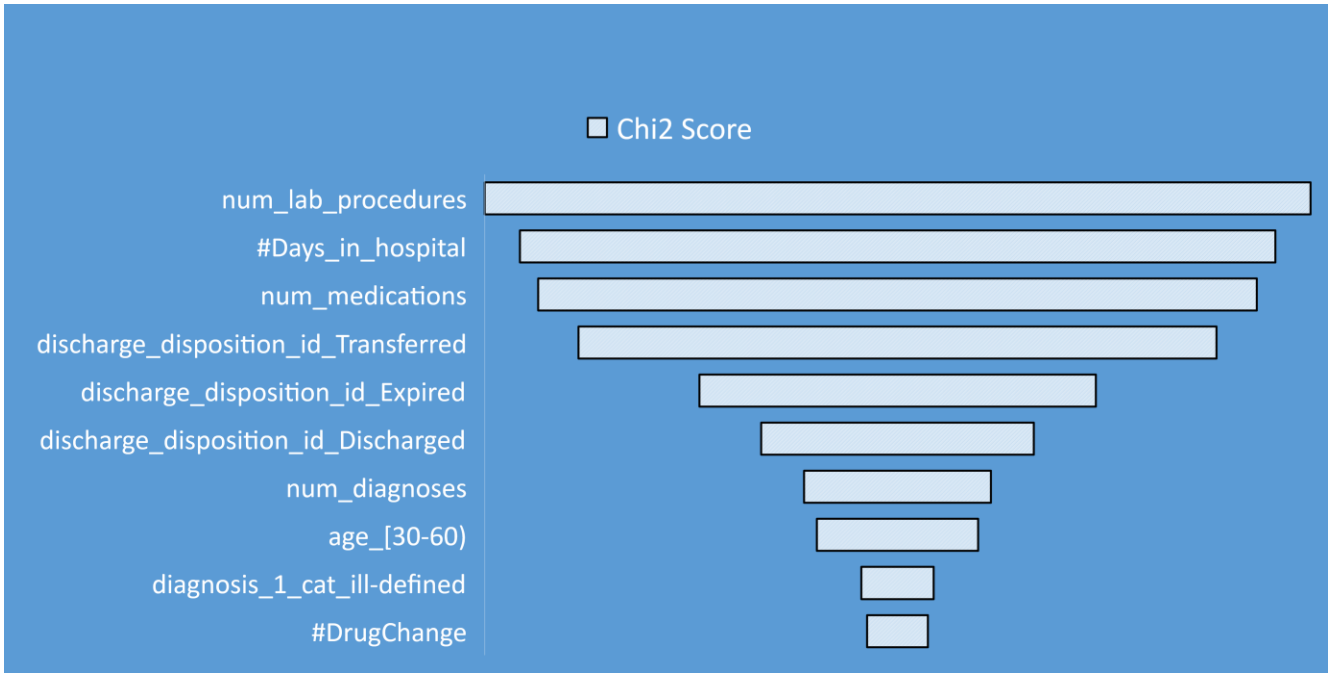
NO: 29,891

Within30days: 4,759

Manual Up-sampling: 4,759*7= 33,313

## Feature Importance – ChiSquare Test

- As there are many features in our dataset, it is important to understand significant features which influence the target variable.
- Top 10 features with Chi-Square scores are arranged in descending order.




## Modeling

- The choice of models is governed primarily by aim to understand the most important factors, along with their relative effects on readmission.
- Therefore, models that have little or no interpretability (neural networks, support vector machines, nearest neighbors, etc.) are given less precedence(at this stage).
- Models Implemented:
    - Logistic Regression
    - Decision tree
    - Random Forest

- Naïve Bayes
- Gradient Boosting Classifier

## Logistic Regression

- The model gave accuracy is 58% and recall is 70% on final test data.
- Recursive Feature Elimination (RFE) is used to understand important features.
- Top 5 features as per RFE ranking are
  - 'discharge_disposition_id_Expired'
  - 'medical_specialty_Otolaryngology'
  - 'medical_specialty_Pediatrics-CriticalCare'
  - 'medical_specialty_Pediatrics-Endocrinology'
  - 'medical_specialty_Proctology‘.
- Classification Report on Validation set:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| NO | 0.61 | 0.54 | 0.57 | 8858 |
| Within30days | 0.63 | 0.70 | 0.66 | 10104 |
| avg / total | 0.62 | 0.62 | 0.62 | 18962 |

## Decision Tree

- Parameter tuning is done based on grid search.
- Hyperparameters used for decision tree are criterion='gini', max_depth=90, max_features=40, min_samples_split=2
- Decision tree gave a recall of 35% on final test set.
- Top 5 important attributes of decision tree are
  - Num_lab_procedures
  - Num_medications
  - #Days_in_hospital
  - Num_diagnosis
  - Num_procedures
- Classification Report on Validation set:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| NO | 1.00 | 0.82 | 0.90 | 8858 |
| Within30days | 0.86 | 1.00 | 0.93 | 10104 |

## Random Forest

- Hyperparameters used for random forest are criterion='gini', max_depth=90, max_features=40, min_samples_split=2, n_estimators=10
- Random Forest gave a recall of 43% on final test set.
- Top 5 important attributes of decision tree are
    - Num_lab_procedures
    - Num_medications
    - #Days_in_hospital
    - Num_diagnosis
    - Discharge_disposition_id_Transfered
- Classification Report on Validation set:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| NO | 1.00 | 0.94 | 0.97 | 8858 |
| Within30days | 0.95 | 1.00 | 0.97 | 10104 |
| avg / total | 0.97 | 0.97 | 0.97 | 18962 |

## Naïve Bayes

- Naive Bayes algorithm is a probabilistic model for classification. It assumes that given the class features are statistically independent of each other.
- Naïve Bayes is used to improve recall.
- Up-sampling with SMOTE, diagnosis_3_cat dropped with Naïve Bayes gave best recall of 75%.
- Classification Report on Validation set:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| NO | 0.59 | 0.58 | 0.58 | 8858 |
| Within30days | 0.63 | 0.64 | 0.64 | 10104 |
| avg / total | 0.61 | 0.61 | 0.61 | 18962 |

## Gradient Boosting Classifier

- Gradient boosting involves loss function to be optimized and allows weak learners to make predictions. This is an additive model to add weak learners to minimize the loss function.
- Parameters used to train this model are criterion='mse', learning_rate=0.15, max_depth=3.
- Up-sampling with SMOTE, with Gradient Boosting Classifier gave recall of 65%.
- Classification Report on Validation set:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| NO | 0.64 | 0.53 | 0.58 | 8858 |
| Within30days | 0.64 | 0.73 | 0.68 | 10104 |
| avg / total | 0.64 | 0.64 | 0.64 | 18962 |

## Model Comparison

(Recall with respect to Readmission within 30 days)

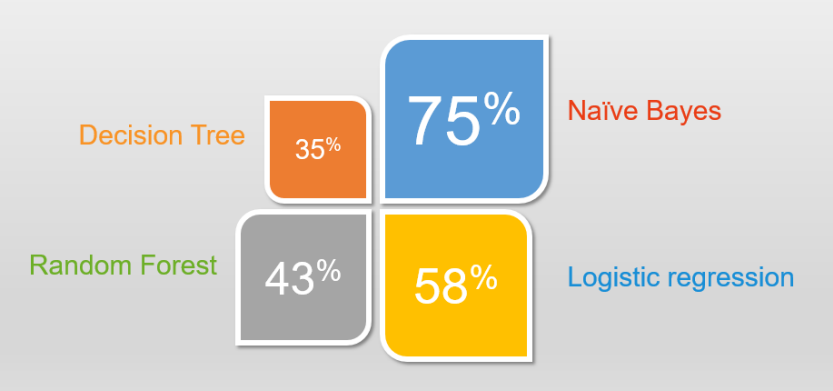
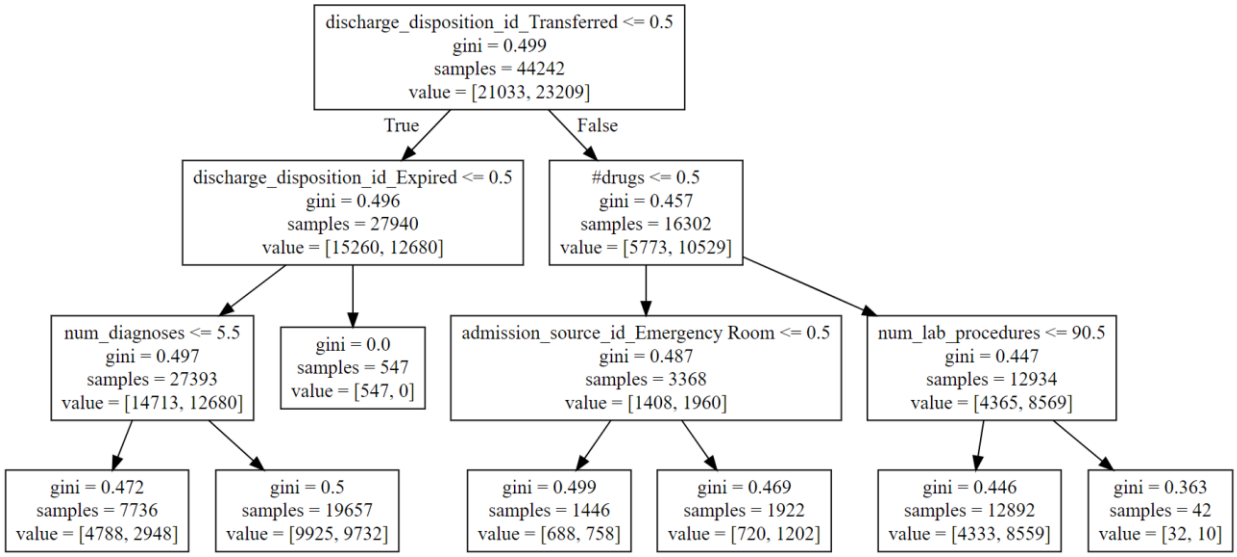



*Figure: Model Comparison*

## Analysis

(To understand the rules a decision tree, a small tree with depth 3 is created)



## Further Improvements

- This work uncovers the features that are critical in identifying high risk of readmission.
- Other Machine learning algorithms can also be used to predict readmission.
- Some critical data which seems to missing in the original dataset can be collected for better recall prediction.

## References:

[1]https://www.hindawi.com/journals/bmri/2014/781670/tab3/

[2]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3996476/

[3] https://www.coursera.org/learn/competitive-data-science/lecture/wzi5a/hyperparameter-tuning-ii