**Diabetic Readmission Prediction**

**Abstract**

Hospital readmission is a high-priority health care quality measure and target for cost reduction, particularly within 30 days of discharge (30-day readmission, aka early readmission). The burden of diabetes among hospitalized patients, however, is substantial, growing, and costly, and readmissions contribute a significant portion of this burden. Reducing readmission rates among patients with diabetes has the potential to greatly reduce health care costs while simultaneously improving care.

**Introduction**

A hospital readmission is when a patient who is discharged from the hospital, gets re-admitted again within a certain period of time. Hospital readmission rates for certain conditions are now considered an indicator of hospital quality, and also affect the cost of care adversely. For this reason, Centres for Medicare & Medicaid Services established the [Hospital Readmissions Reduction Program](https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html) which aims to improve quality of care for patients and reduce healthcare spending by applying payment penalties to hospitals that have more than expected readmission rates for certain conditions.

**Problem Statement**

 A leading hospital in the US is suddenly seeing increase in the patient readmission in less than 30 days. This is serious concern for the hospital as it may indicate insufficient treatment or diagnosis when the patient was admitted first and later released under clean bill of health. Not only the image of hospital as healthcare provider is compromised, this is also increased cost to the entire Medicare ecosystem in form of increased insurance claims.

**Aim**

Being able to determine factors that lead to higher readmission in such patients, and correspondingly being able to predict which patients will get readmitted can help hospitals save millions of dollars while improving quality of care.

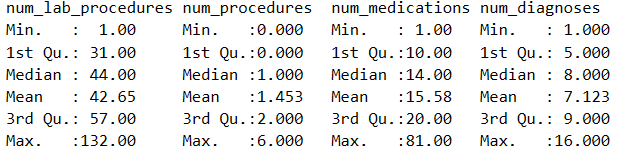
The objective is: Classify the patients treated by this hospital into two primary categories:

* **Readmitted within 30 days**
* **Not readmitted**

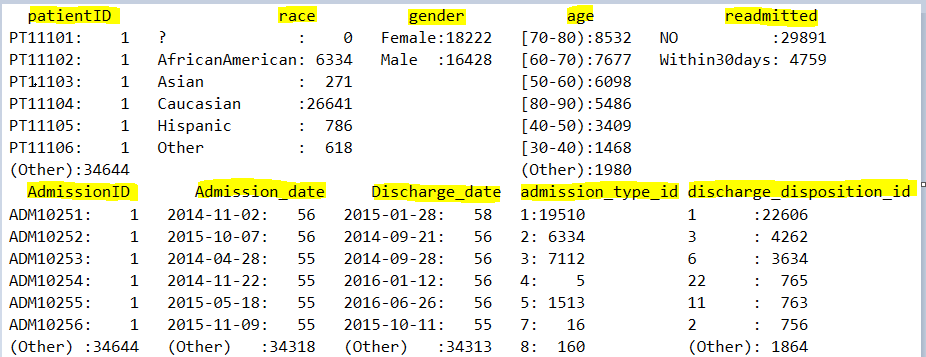
**Dataset**

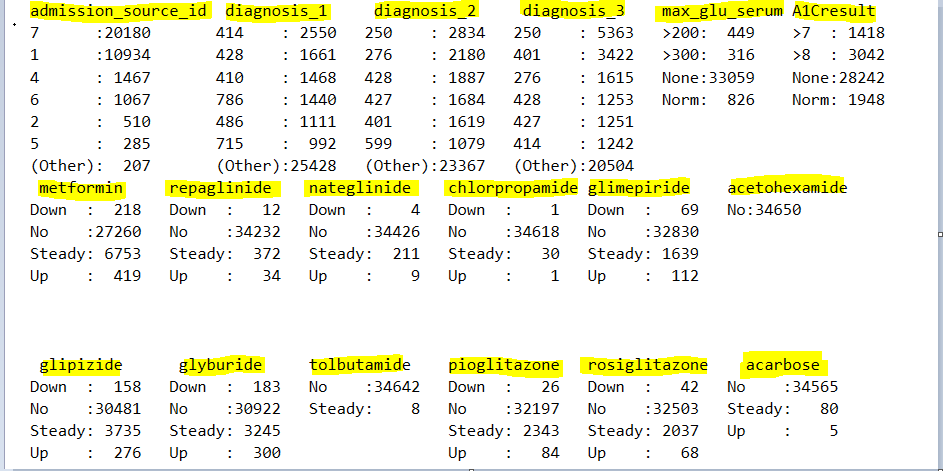
The dataset has over 34650 records and 45 features including patient characteristics, conditions, tests and medications.

Below are the summary of numerical features:



**Summary (categorical features)**





**Data Pre-mining**

Applied three types of methods here:

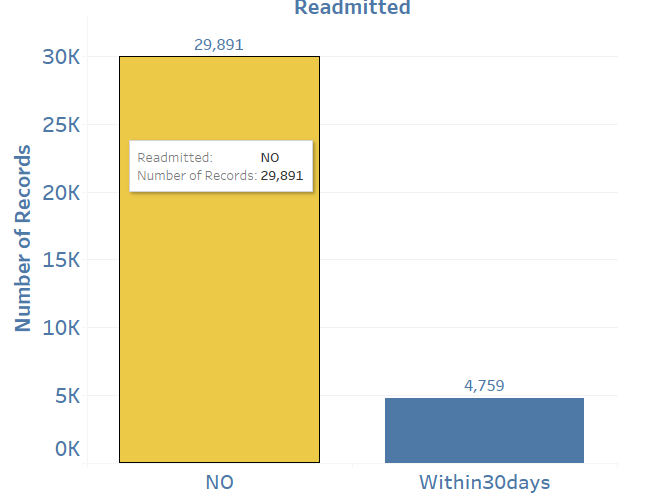
* Cleaning tasks such as dropping bad data, dealing with missing values.
* The columns namely weight, payer code, medical specialty have been dropped as it contains more than 80% of null values.
* Replaced “?” with “NA” and performed Central Imputation on the dataset.
* Modification of existing features
* The dataset contained up to three diagnoses for a given patient (primary, secondary and additional). However, each of these had several unique ICD codes and it is extremely difficult to include them in the model and interpret meaningfully. Therefore, these diagnosis codes have been collapsed into different disease categories in an almost similar way. These categories include Circulatory, Respiratory, Digestive, Diabetes, Injury, Musculoskeletal, Genitourinary, Neoplasms, Others etc.
* Re-categorize age group into discrete form.
* Creation or derivation of new features, usually from existing ones.
* A new Feature have been extracted based on the existing variables.

Day’s \_Spent: No. of days spent in hospital

By using the chisq test between the various extracted variables and the target variables, could able to find the most useful variables and removed irrelevant variables (EX: patient ID).

**Data Analysis**

The given dataset has an imbalance target data distribution with 4,759 records of “Within30days” and “29,981” records of “No” i.e. not been readmitted.

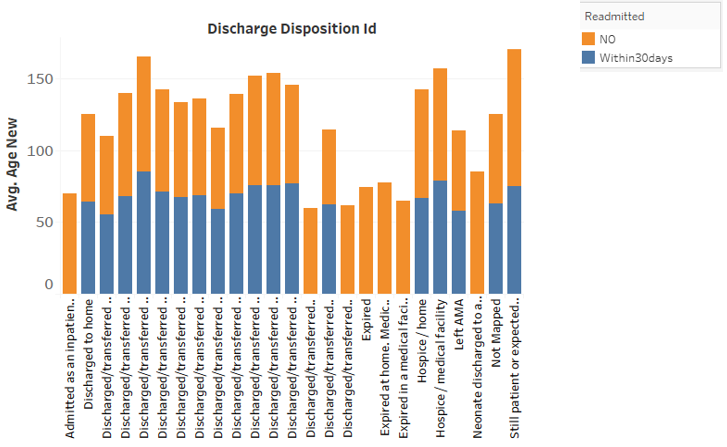


**Data Balancing**

Data was highly imbalanced with respect to readmissions (**only 13% records for 30-day readmissions**), leading to high accuracy. Moreover, the high accuracy could be attributed not to the generalizability of our model to diverse patient records but to the baseline accuracy of 90%: predicting that no patient would be readmitted. This was evident from the poor precision and recall of our model in predicting patient readmissions. We used synthetic minority over-sampling technique (**SMOTE**) to oversample our underrepresented class of readmissions.

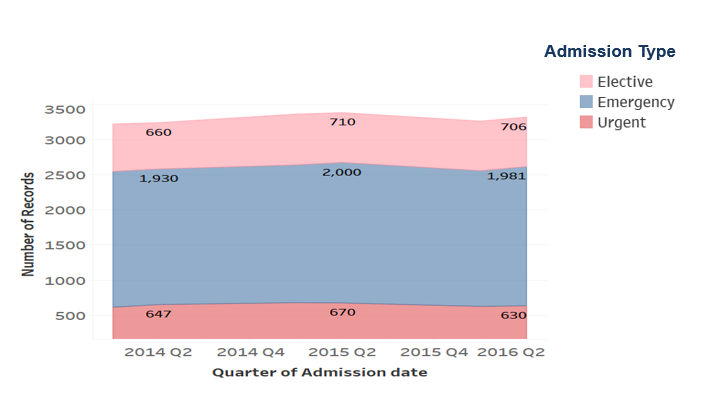
**Discharge Disposition**

As observed below, few discharge disposition have zero readmissions. Since we are trying to predict readmissions, those patients who died during this hospital admission, have zero probability of readmission.



**Admission Type Trend**

The below Trend Chart shows that, more number of patients are getting admitted under “Emergency” on comparison with other types.



**Model Building**

The choice of models is governed primarily by our aim to understand the most important factors, along with their relative effects on medication change and readmission. Thus, while model accuracy is important, model interpretability in order to devise corrective measures is a key criterion for the model selection. The models that have been implemented include:

**Decision Trees**:

Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes. By iteratively and hierarchically observing the level of certainty of predicting whether someone would be readmitted or not, we find the relative importance of different factors.

Variables actually used in tree construction:

 A1Cresult

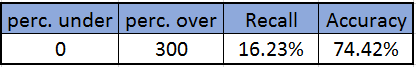
 discharge\_disposition\_id

 max\_glu\_serum

 num\_diagnoses

 num\_lab\_procedures

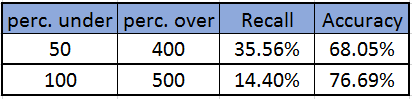
 race

Below is the model performance result: 

**Bagged CART**

Bagging is a technique used to reduce the variance of our predictions by combining the result of multiple classifiers modelled on different sub-samples of the same data set. There are various implementations of bagging models. Random forest is one of them

Below is the model performance result:

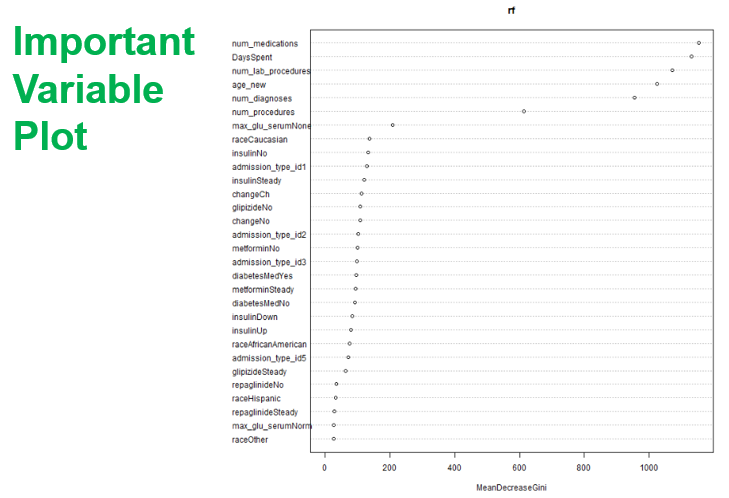


**Random Forest**

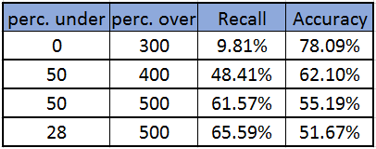
Random forests are made up of randomly grown trees. This method works in a way that instead of relying on a single decision tree, we try many different trees with randomly assigned subsets of features. The final prediction is then calculated by voting across predictions made by all the trees in the forest.

By considering more than one decision tree and then doing a majority voting, random forests helped in being more robust predictive representations than trees.

Below Plot shows important variables that has majorly influenced the accuracy of the model:



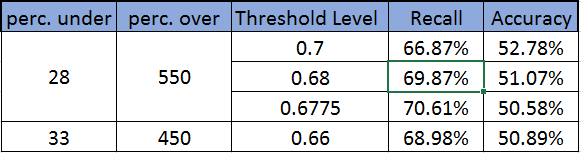
Below is the model performance result:



**Logistic Regression**

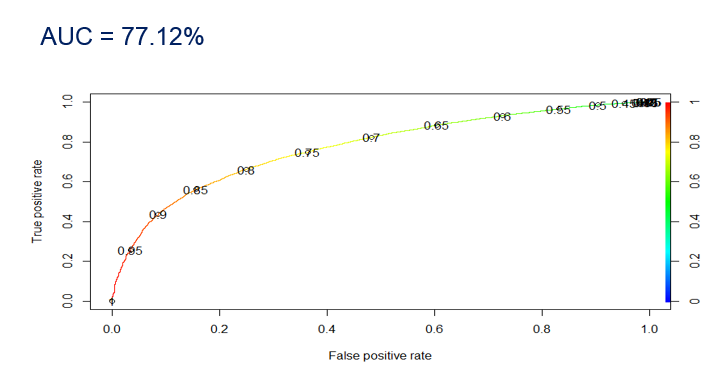
Logistic Regression is used to predict a binary outcome given a set of independent variables and can help us understand the relative impact and statistical significance of each factor on the probability of readmission.

Below is the model performance result:



Below diagram shows the ROC curve which summarizes the model’s performance by evaluating the trade-offs between true positive rate (sensitivity) and false positive rate (1- specificity).

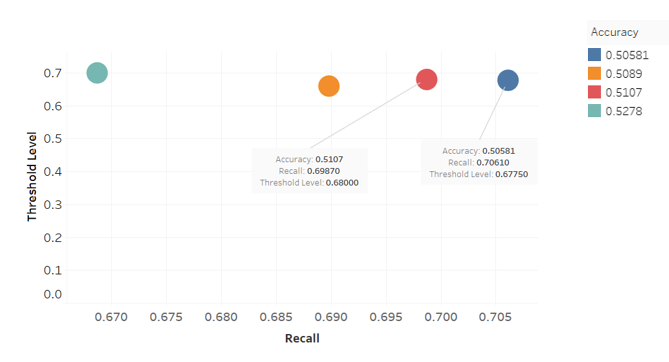
The area under curve (AUC), referred to as index of accuracy is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model.



Threshold Vs Recall & Accuracy

Logistic Regression predicts the probability of occurrence of an event by fitting data to a logit function. The threshold value determines whether the probability value should be assigned to True or False.

Below image shows, accuracy and recall at different threshold level and as observed 0.6775 threshold level, gives out the good recall.



**Comparing model performance**

The below results shows different model and their respective recall and accuracies. Model has been experimented on various smoted dataset and results have been changed significantly.

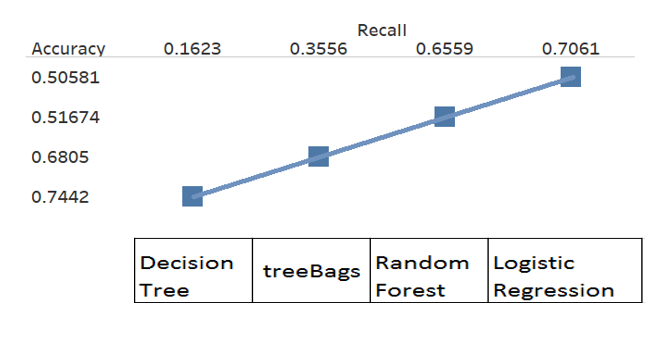


Image representing track from initial to final result.

**Limitations**

The dataset at hand provides some really useful information. However, a key thing to understand here is that the quality of predictions depend not only on the volume of data available, but on variety as well. We are limited by the information at hand, which is a comprehensive but not an exhaustive account of all the factors that may affect hospital readmissions. Besides other factors mentioned above, there may be many other factors depending on situation that could be affecting readmissions.

**Conclusion**

* Data pre-mining is of upmost importance in improving the model accuracy.
* The readmission groups are related to admission source, admission type, discharge disposition and number of inpatient visits.
* Instead of tracking all attributes, hospitals are suggested to focus on number of patient’s inpatient visits, admission source, admission type, discharge disposition.
* Hospitals are advised to concern not only inpatient treatment but also continuing care after discharge.