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 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:

```
num_lab_procedures num_procedures  num_medications num_diagnoses
Min.   :  1.00     Min.   :0.000   Min.   : 1.00   Min.   : 1.000
1st Qu.: 31.00     1st Qu.:0.000   1st Qu.:10.00   1st Qu.: 5.000
Median : 44.00     Median :1.000   Median :14.00   Median : 8.000
Mean   : 42.65     Mean   :1.453   Mean   :15.58   Mean   : 7.123
3rd Qu.: 57.00     3rd Qu.:2.000   3rd Qu.:20.00   3rd Qu.: 9.000
Max.   :132.00     Max.   :6.000   Max.   :81.00   Max.   :16.000
```

Summary (categorical features)

```
   patientID                 race           gender          age              readmitted
 PT11101:    1   ?              :    0   Female:18222   [70-80):8532   NO          :29891
 PT11102:    1   AfricanAmerican: 6334   Male  :16428   [60-70):7677   Within30days: 4759
 PT11103:    1   Asian          :  271                  [50-60):6098
 PT11104:    1   Caucasian      :26641                  [80-90):5486
 PT11105:    1   Hispanic       :  786                  [40-50):3409
 PT11106:    1   Other          :  618                  [30-40):1468
 (Other):34644                                          (Other):1980
  AdmissionID       Admission_date      Discharge_date    admission_type_id discharge_disposition_id
 ADM10251:     1   2014-11-02:   56   2015-01-28:   58   1:19510            1      :22606
 ADM10252:     1   2015-10-07:   56   2014-09-21:   56   2: 6334            3      : 4262
 ADM10253:     1   2014-04-28:   55   2014-09-28:   56   3: 7112            6      : 3634
 ADM10254:     1   2014-11-22:   55   2016-01-12:   56   4:    5            22     :  765
 ADM10255:     1   2015-05-18:   55   2016-06-26:   56   5: 1513            11     :  763
 ADM10256:     1   2015-11-09:   55   2015-10-11:   55   7:   16            2      :  756
 (Other) :34644    (Other)   :34318   (Other)   :34313   8:  160            (Other): 1864
```

```
 admission_source_id  diagnosis_1      diagnosis_2      diagnosis_3     max_glu_serum  A1Cresult
 7      :20180       414    : 2550   250    : 2834   250    : 5363   >200:  449     >7  : 1418
 1      :10934       428    : 1661   276    : 2180   401    : 3422   >300:  316     >8  : 3042
 4      : 1467       410    : 1468   428    : 1887   276    : 1615   None:33059     None:28242
 6      : 1067       786    : 1440   427    : 1684   428    : 1253   Norm:  826     Norm: 1948
 2      :  510       486    : 1111   401    : 1619   427    : 1251
 5      :  285       715    :  992   599    : 1079   414    : 1242
 (Other):  207       (Other):25428   (Other):23367   (Other):20504
  metformin       repaglinide     nateglinide     chlorpropamide  glimepiride      acetohexamide
 Down  :  218    Down  :   12    Down  :    4    Down  :    1    Down  :   69     No:34650
 No    :27260    No    :34232    No    :34426    No    :34618    No    :32830
 Steady: 6753    Steady:  372    Steady:  211    Steady:   30    Steady: 1639
 Up    :  419    Up    :   34    Up    :    9    Up    :    1    Up    :  112


  glipizide       glyburide       tolbutamide     pioglitazone    rosiglitazone    acarbose
 Down  :  158    Down  :  183    No    :34642    Down  :   26    Down  :   42     No    :34565
 No    :30481    No    :30922    Steady:    8    No    :32197    No    :32503     Steady:   80
 Steady: 3735    Steady: 3245                    Steady: 2343    Steady: 2037     Up    :    5
 Up    :  276    Up    :  300                    Up    :   84    Up    :   68
```

Data Pre-mining

Applied three types of methods here:

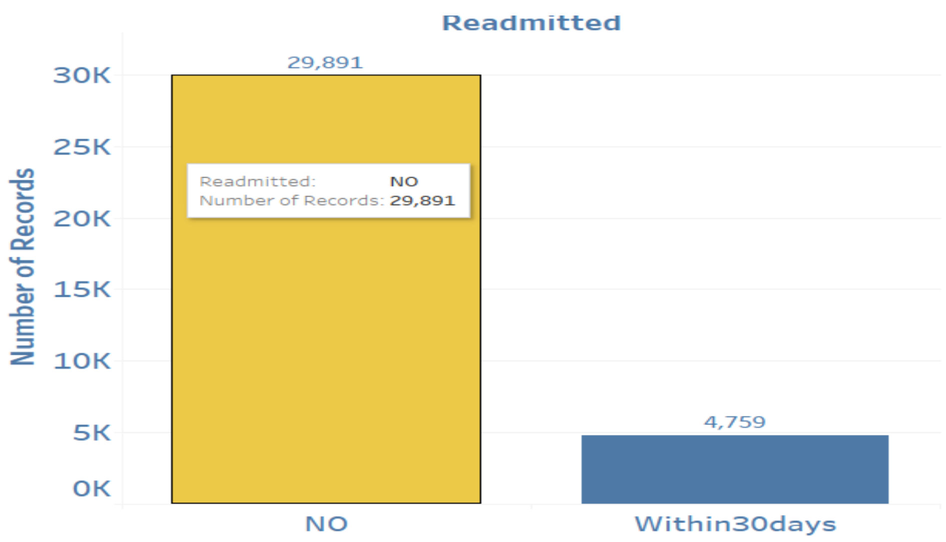
- Cleaning tasks such as dropping bad data, dealing with missing values.
    - The columns namely weight, payer code, medical specialty have been dropped as it contains more than 80% of null values.
    - Replaced “?” with “NA” and performed Central Imputation on the dataset.
- Modification of existing 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.

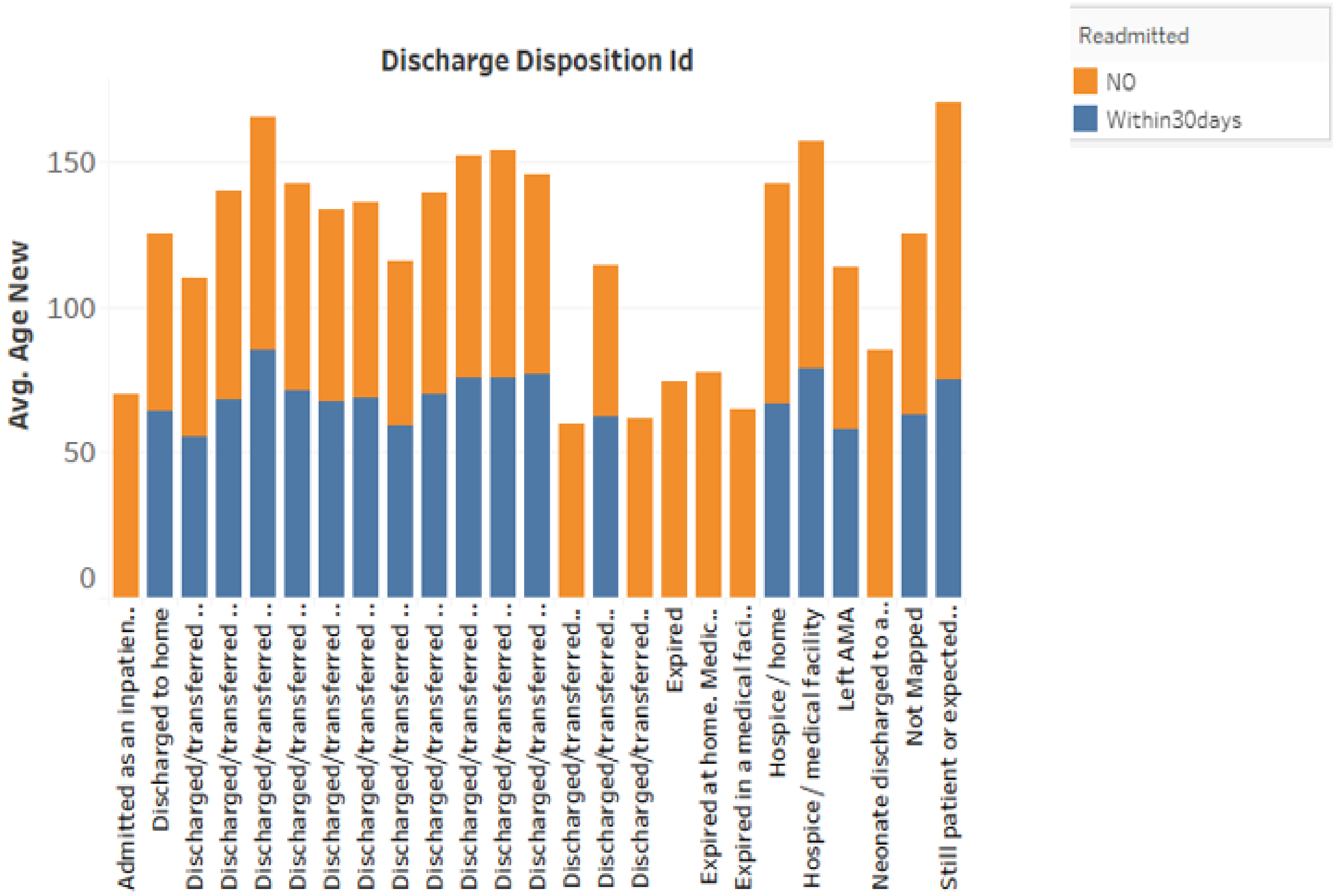


Data Balancing

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.

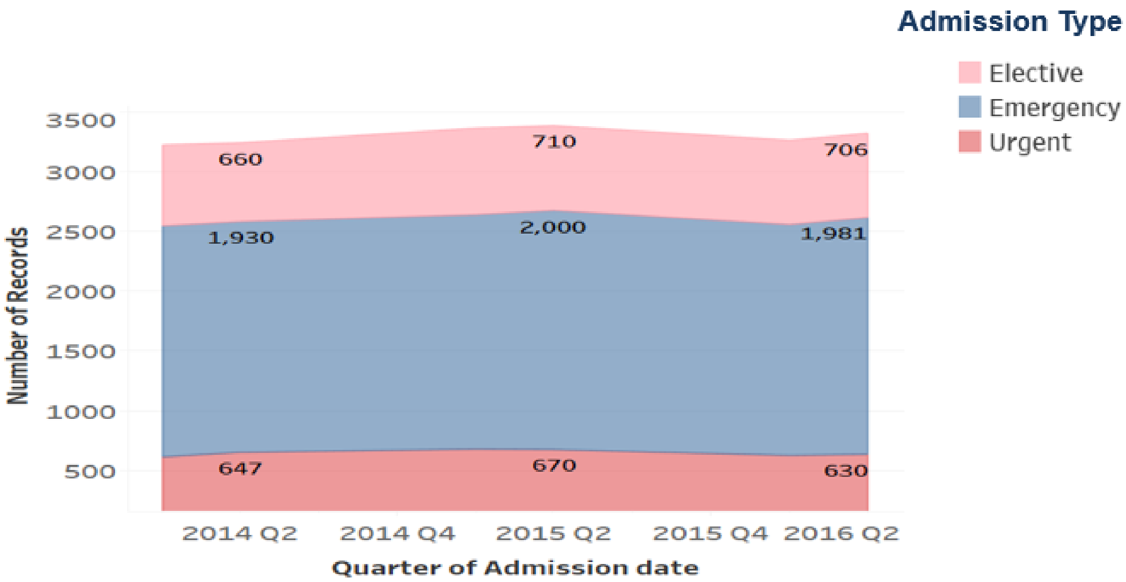
Discharge Disposition

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:

perc. under	perc. over	Recall	Accuracy
0	300	16.23%	74.42%

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:

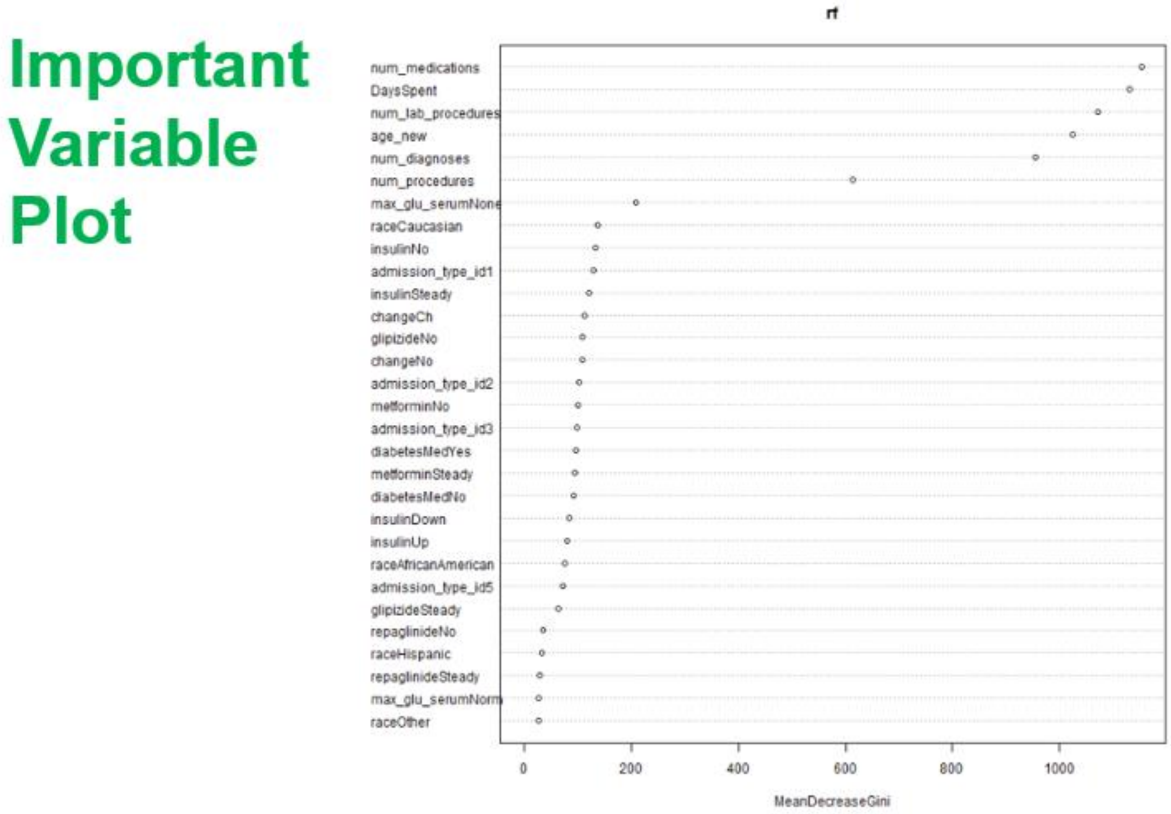
perc. under	perc. over	Recall	Accuracy
50	400	35.56%	68.05%
100	500	14.40%	76.69%

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:

perc. under	perc. over	Recall	Accuracy
0	300	9.81%	78.09%
50	400	48.41%	62.10%
50	500	61.57%	55.19%
28	500	65.59%	51.67%

Logistic Regression

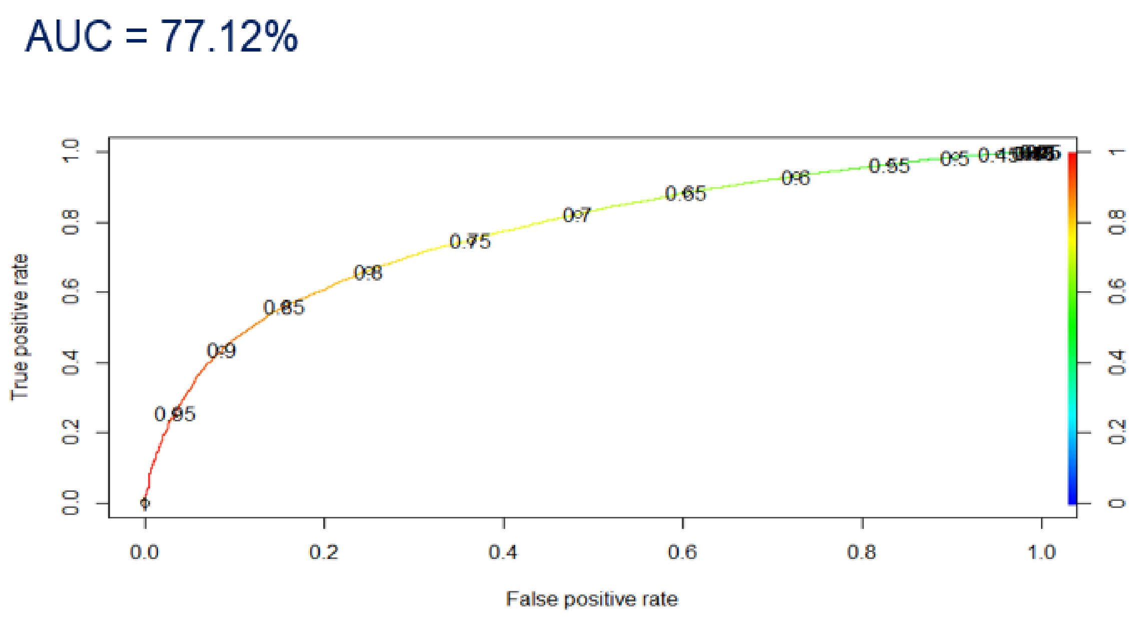
Logistic Regression is used to predict a binary outcome given a set of independent variables and can help us understand the relative impact and statistical significance of each factor on the probability of readmission.

Below is the model performance result:

perc. under	perc. over	Threshold Level	Recall	Accuracy
28	550	0.7	66.87%	52.78%
		0.68	69.87%	51.07%
		0.6775	70.61%	50.58%
33	450	0.66	68.98%	50.89%

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).

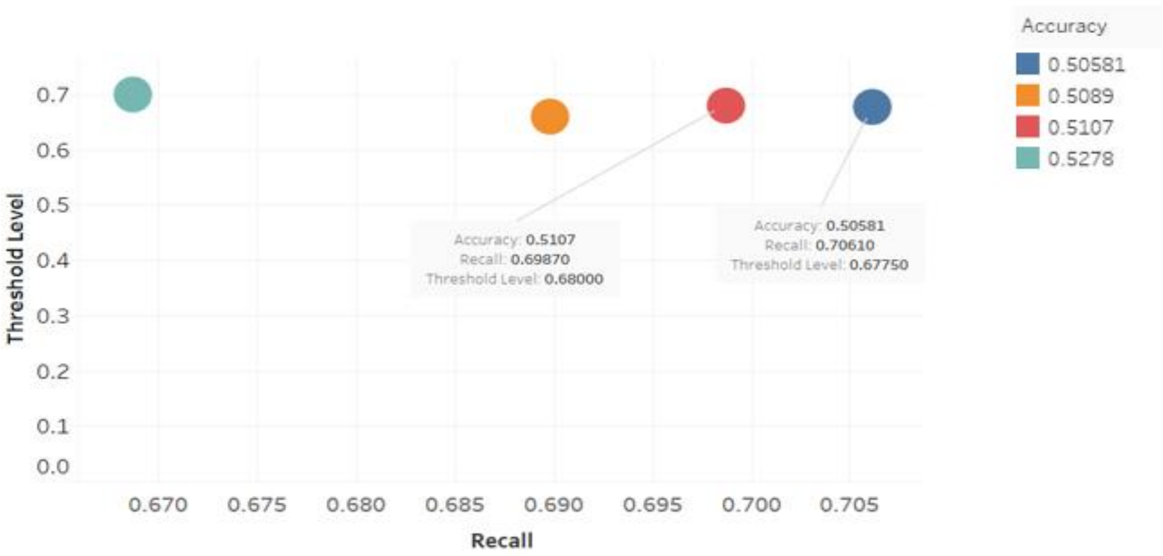
The area under curve (AUC), referred to as index of accuracy is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model.



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.