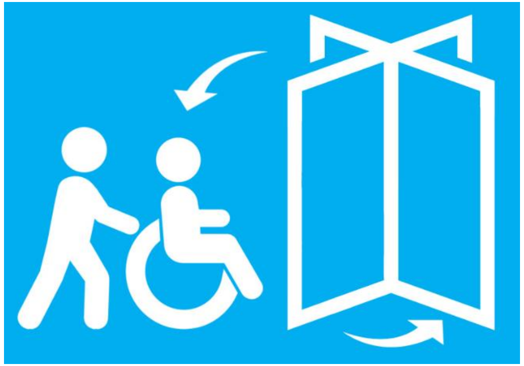
Diabetes Readmission Prediction



Sheik Basha
1760B37

Introduction

- Readmission Rate

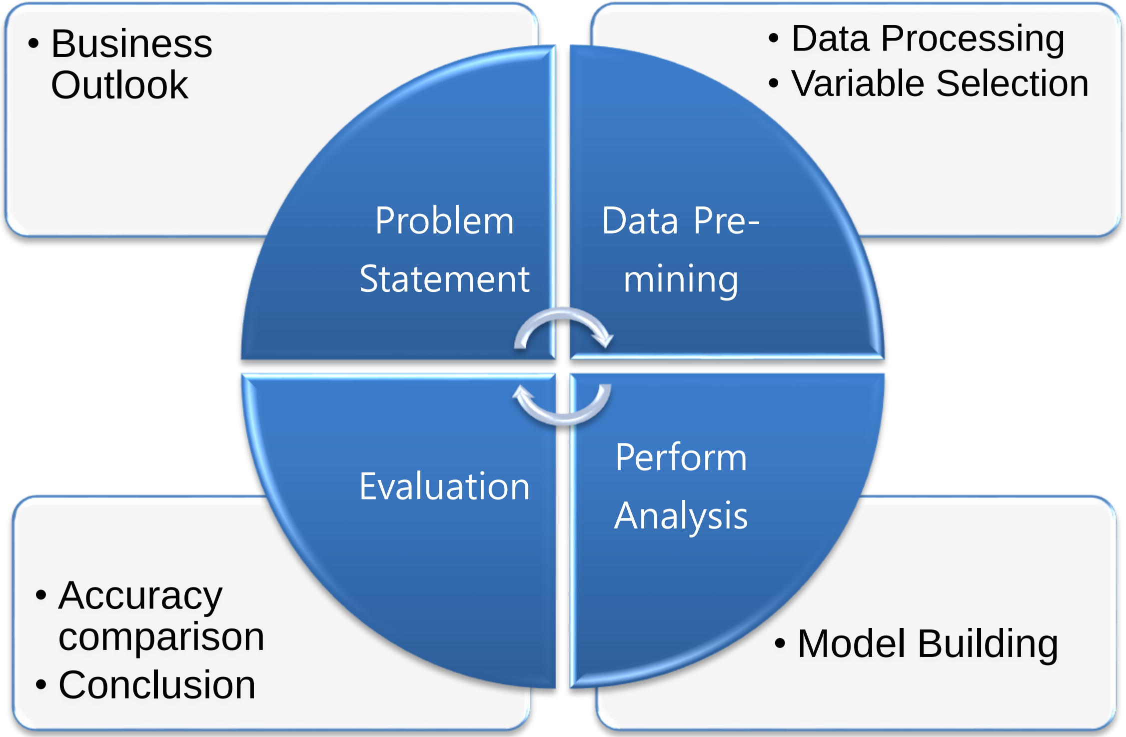
  A hospitalization that occurs within 30 days after a discharge.

- Why Is Readmission Important ?

  Reduce cost of care and medical disputes.

  Improve patients’ safety and health.

Methodology/ Process



Problem Statement

- Identify the major factors that contribute to hospital readmi ssions.

- **Goal :** To make effective prediction on readmissions which will enable hospitals to identify and target patients at the higher risk.

Data Pre-mining

Data Pre-processing

Feature Engineering

Variable Extraction

Data Pre-processing

- Dataset was given with 34650 records and 45 variables out of which one is target variable.
- Categorize Diagnosis into 18 groups.
- Re-categorize Age group.
- Drop variables which has more than 50% of null Values.

Feature Engineering

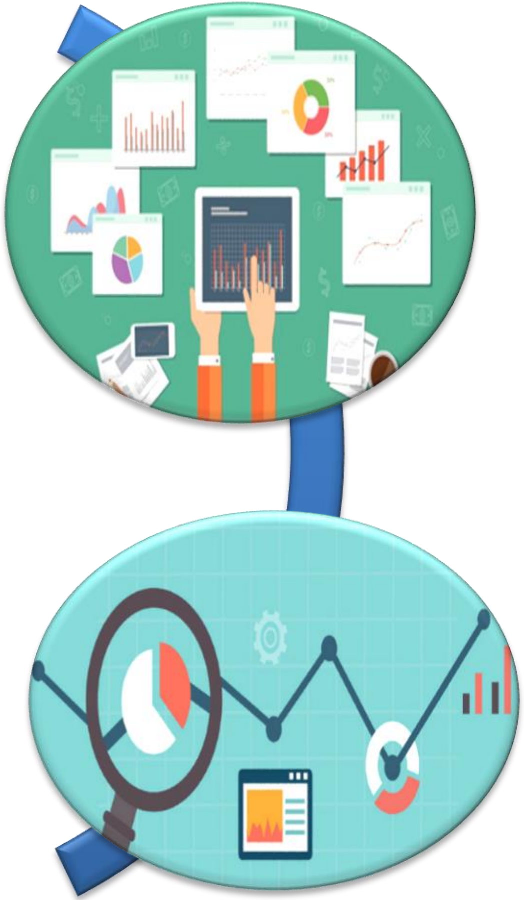
- A new Feature have been extracted based on the existing variables.
- Days _Spent : No. of days spent in hospital

Variable Extraction

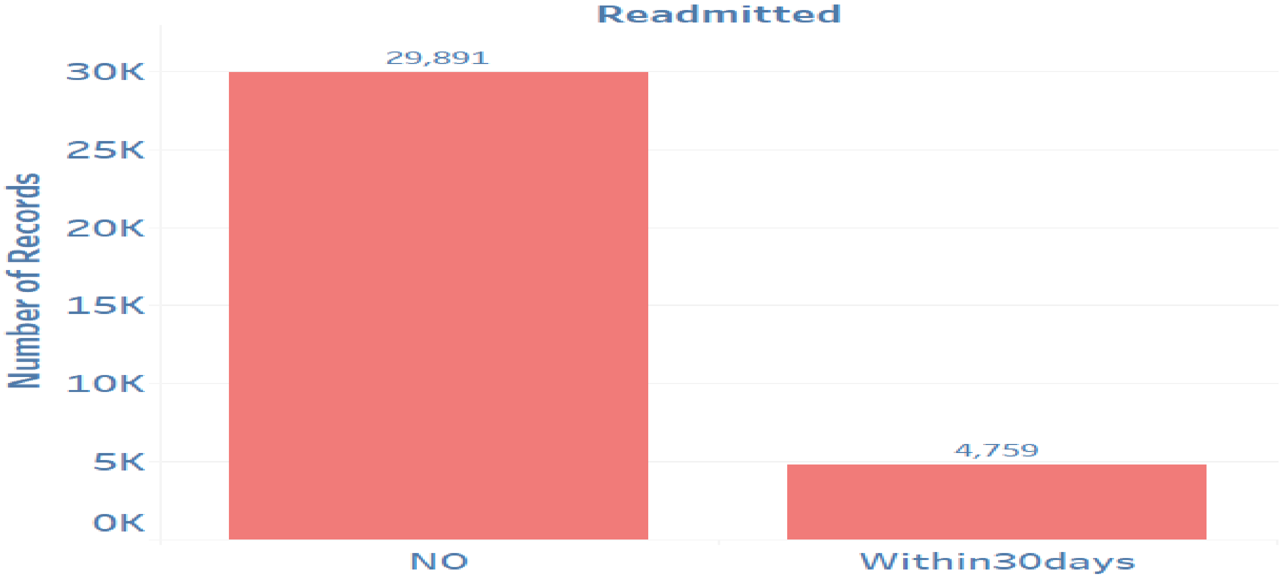
- ❖ By using the chisq test between the various extracted variables and the target variables, could able to find the most useful variables.

- ❖ Remove irrelevant variables (EX: patient ID)

Data Insights

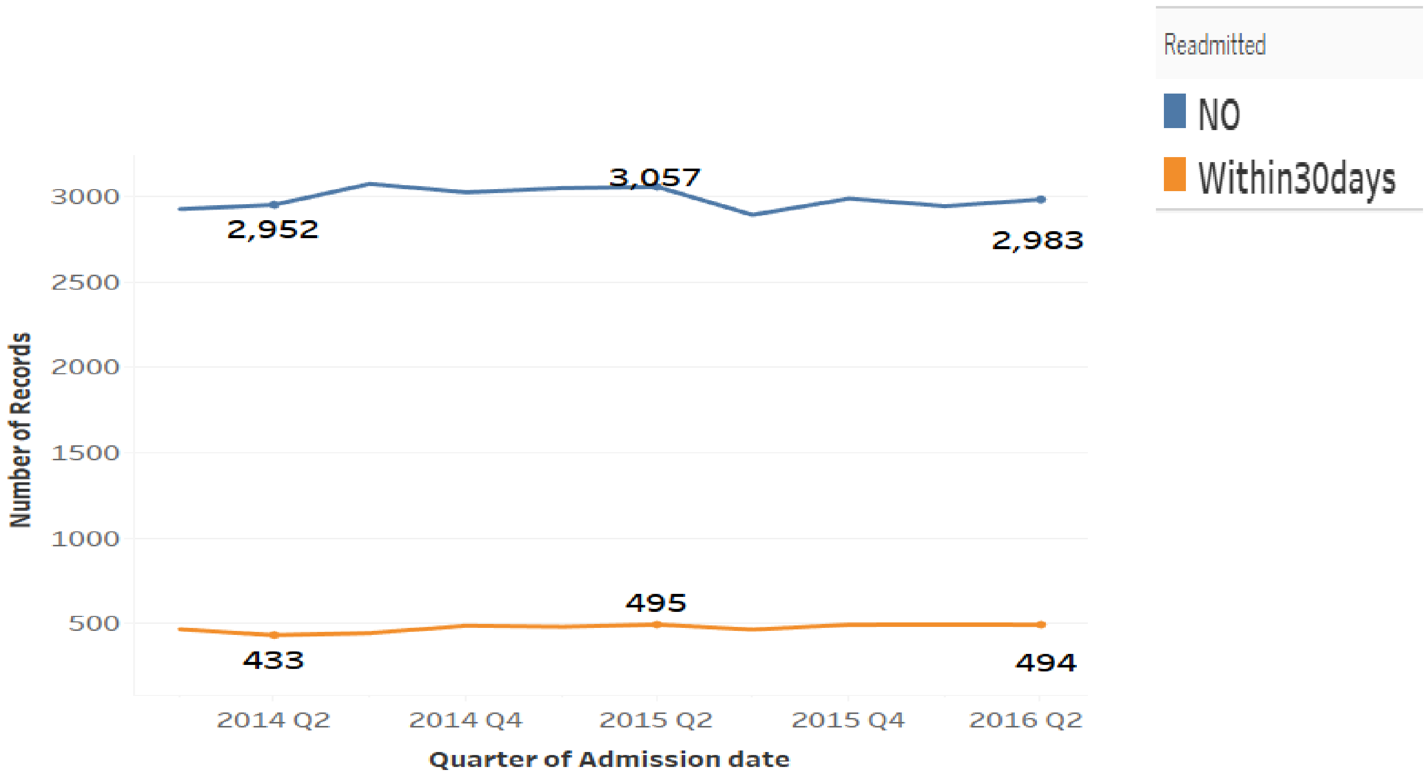


Target Variable Data Distribution

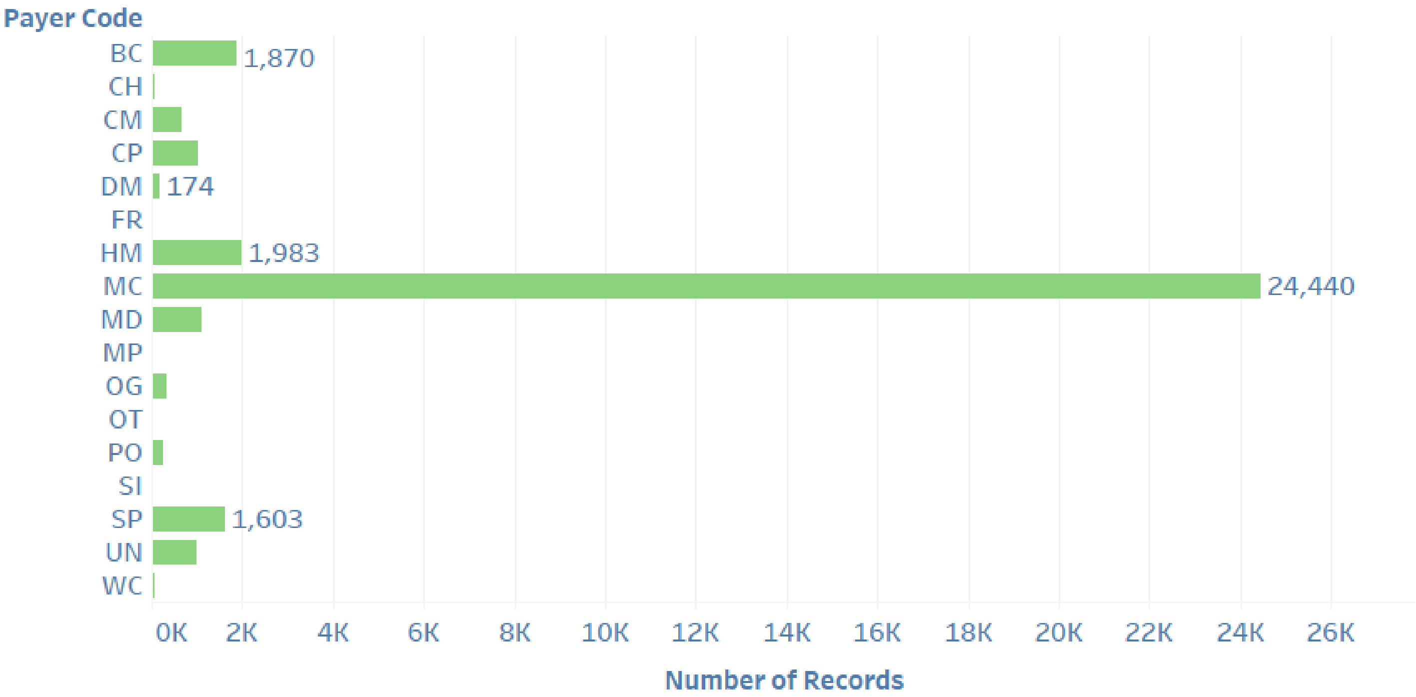


- Smote has been used to oversample the minority class data.

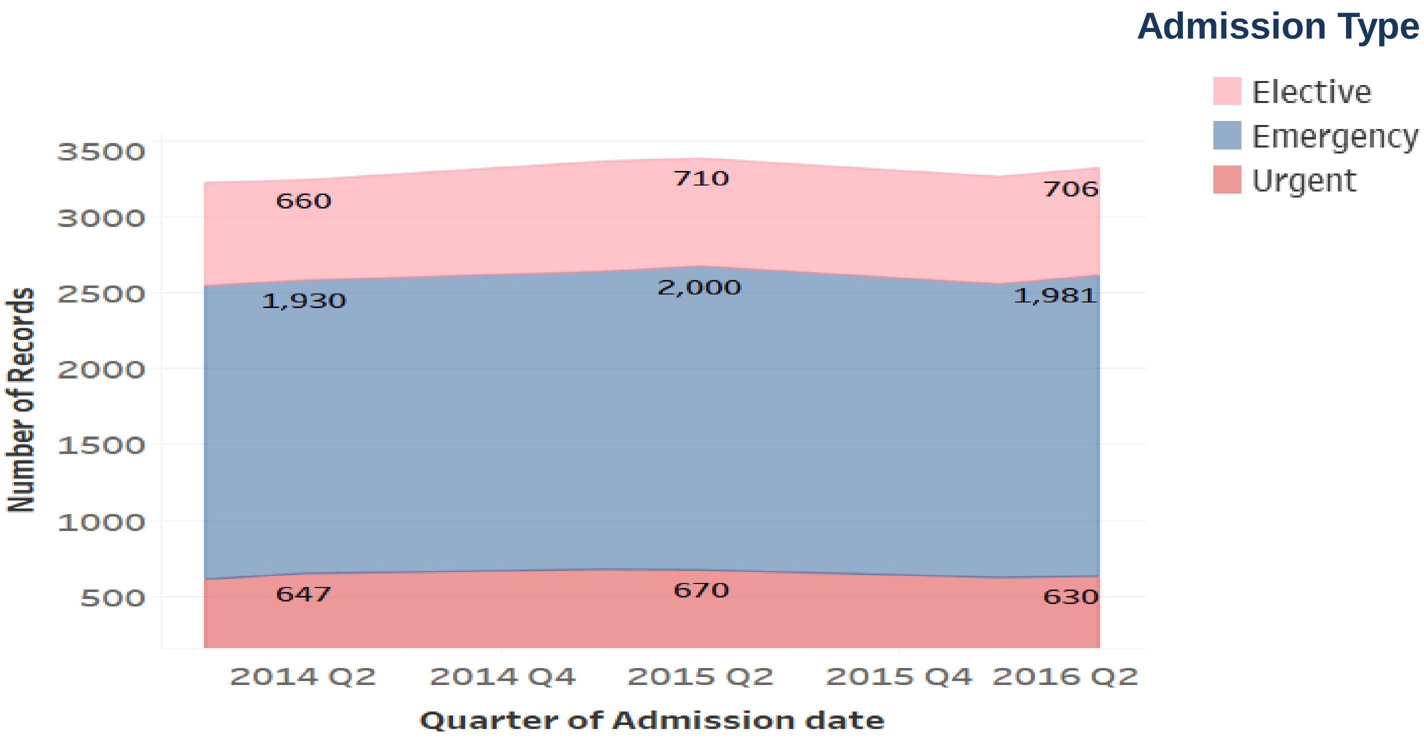
No. of Patient Records Trend



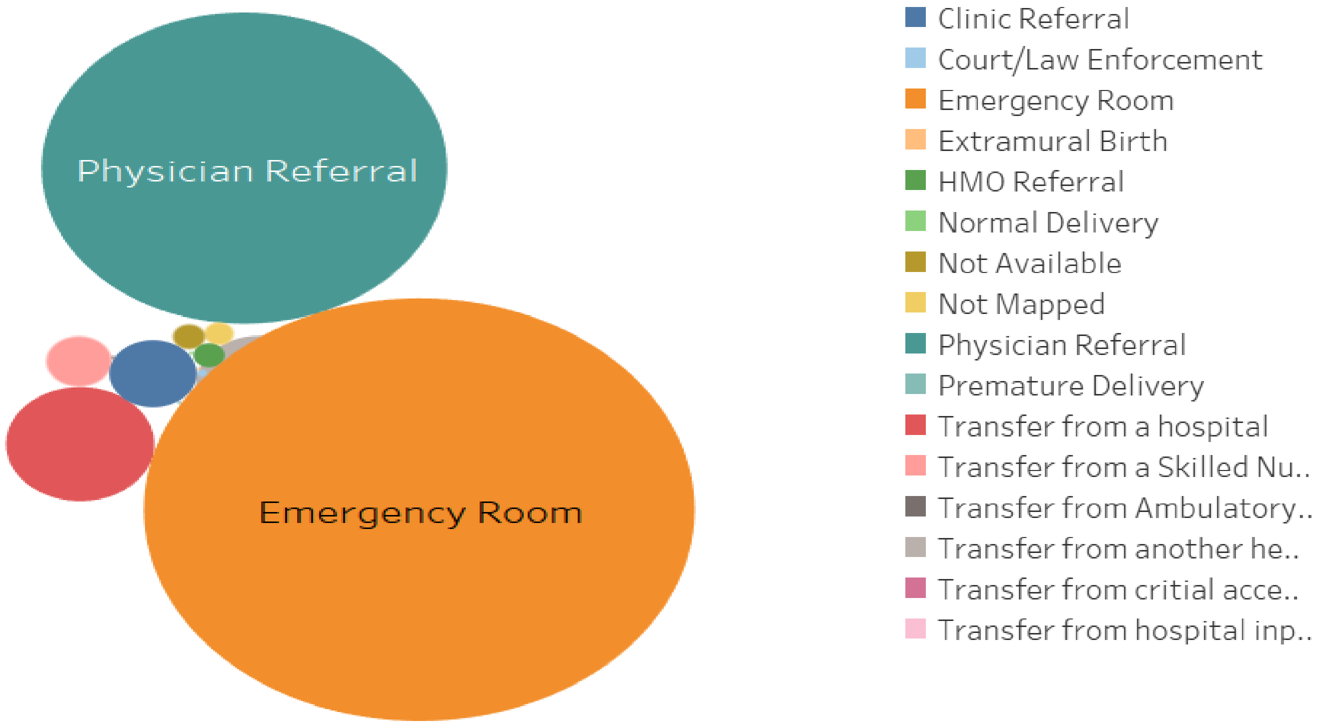
No. of Records Significantly varying with Payer

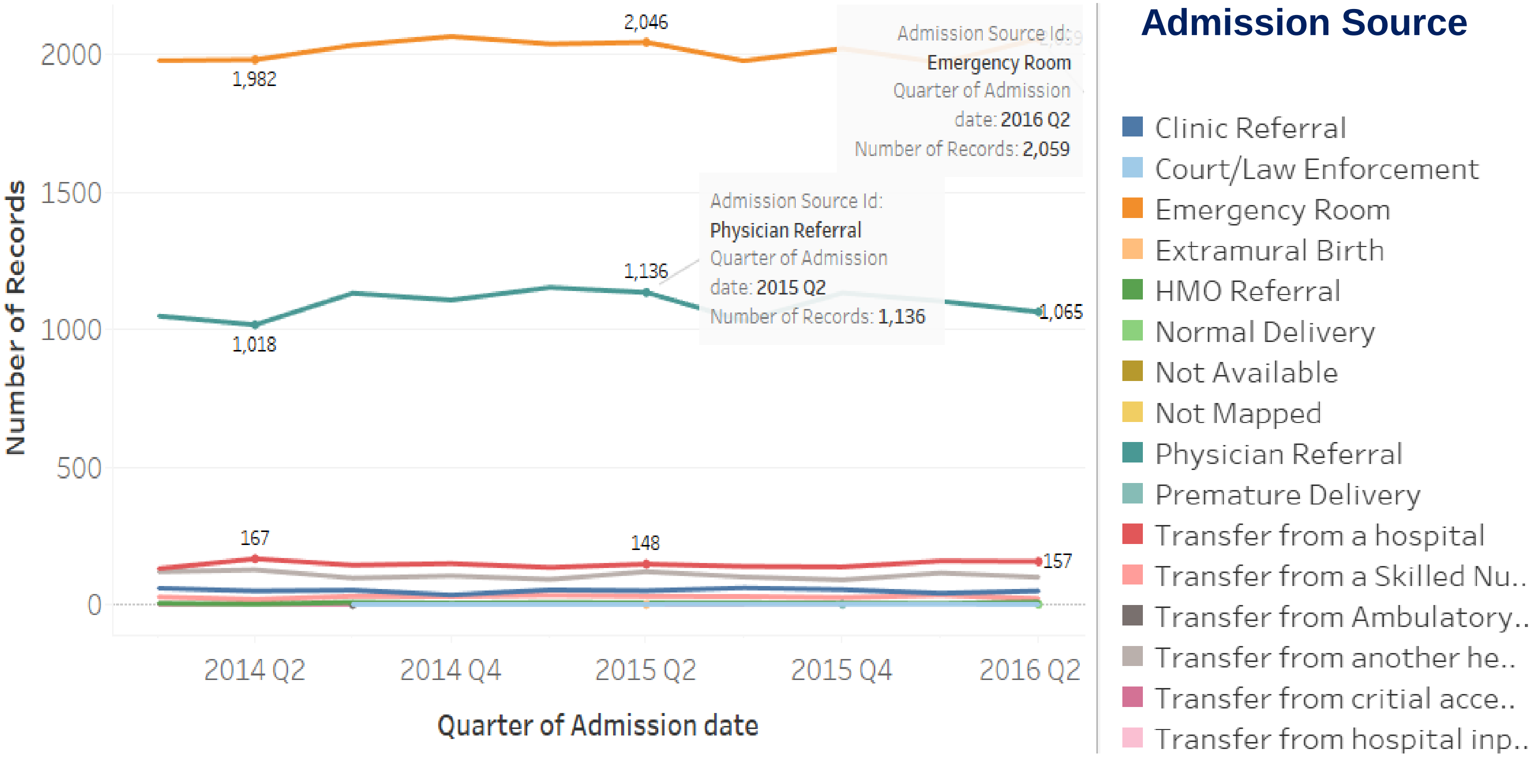
Admission Type Trend



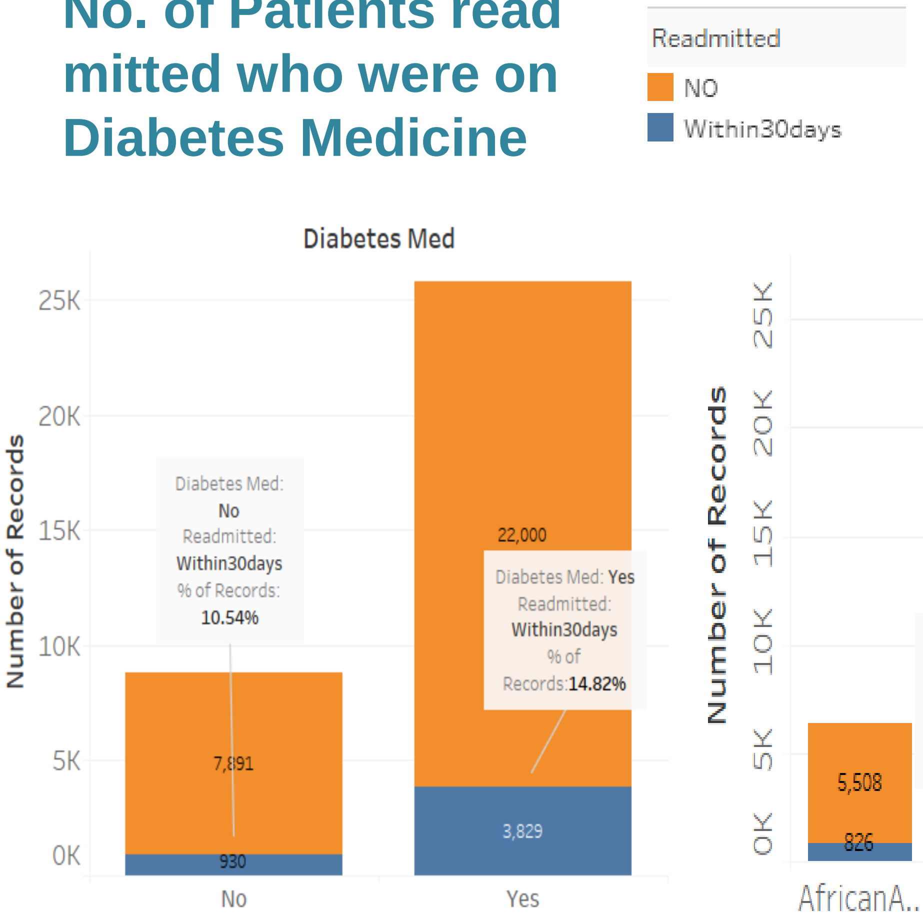
Admission Sources



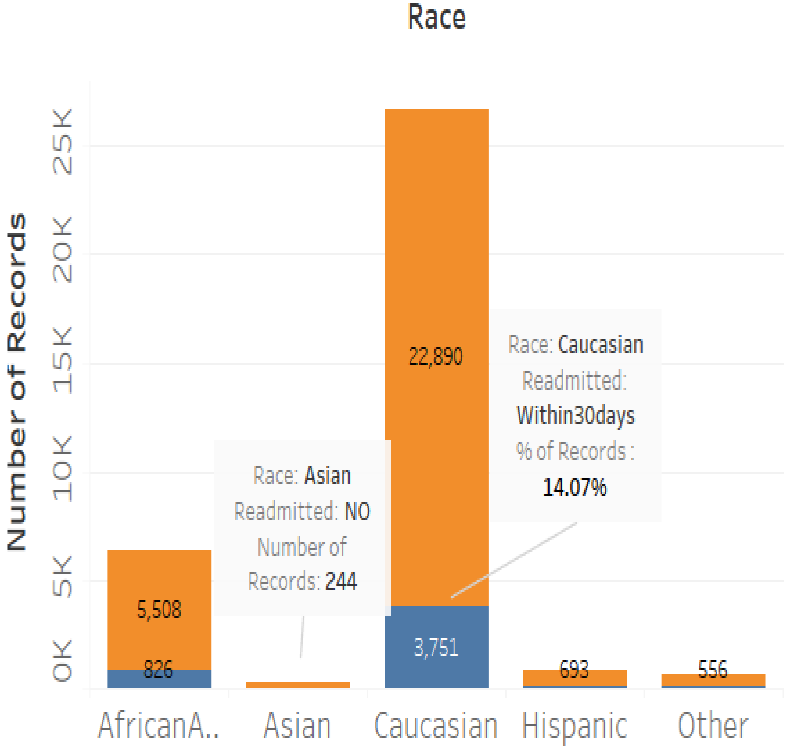
Admission Source Trend



No. of Patients readmitted who were on Diabetes Medicine



No. of Patients readmitted Vs Race

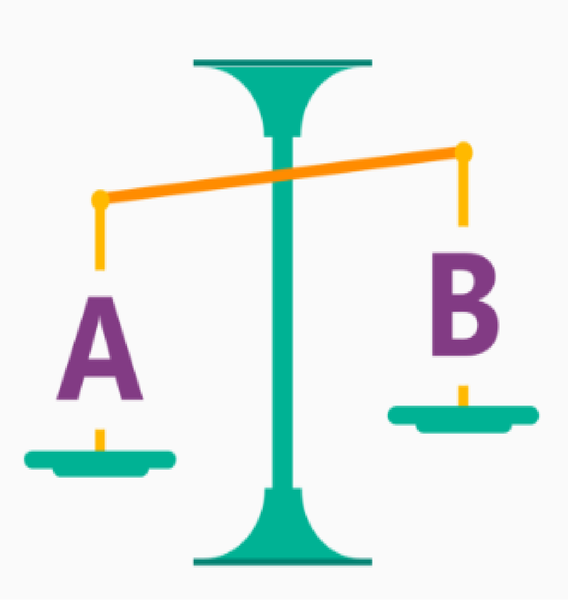


Model Building

Model Building Methods

Classification

- Decision Tree
- Bagged CART
- Logistic Regression



Decision Trees

- Easy to interpret.
- Useful in Data exploration.
- Less data cleaning required.

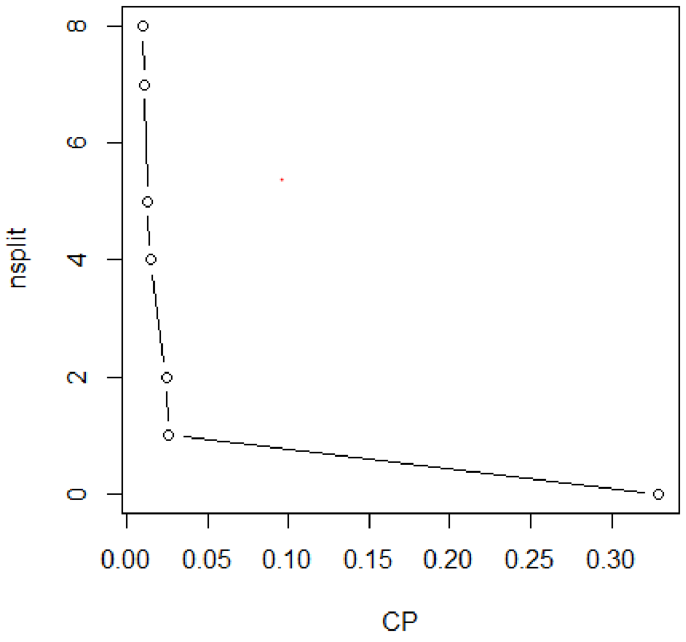
Model Performance:

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

Decision Tree..

Variables actually used in tree construction:

- A1Cresult
- discharge_disposition_id
- max_glu_serum
- num_diagnoses
- num_lab_procedures
- race



Bagged CART

- ❖ Reduce the variance of our predictions.
- ❖ There are various implementations of bagging models Random forest is one of them.

Model Performance using “treebag” :

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

Random Forest

- Handle large data set with higher dimensionality.
- Identify most significant variables i.e. the model outputs **Importance of variable.**
- The sub-trees are learned so that the resulting predictions from all of the subtrees have less correlation.

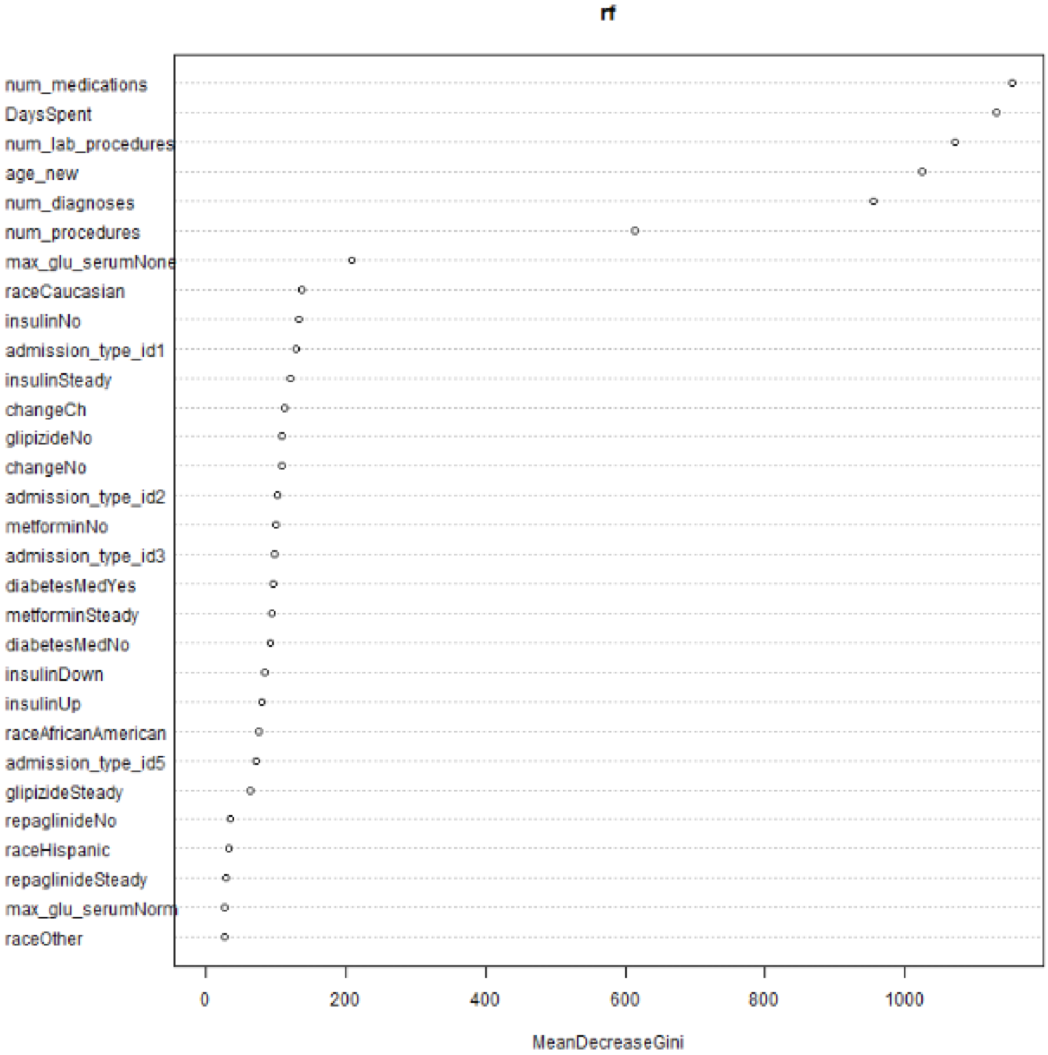
Random Forest ..

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

Model Performance :

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%

Important Variable Plot



Logistic Regression

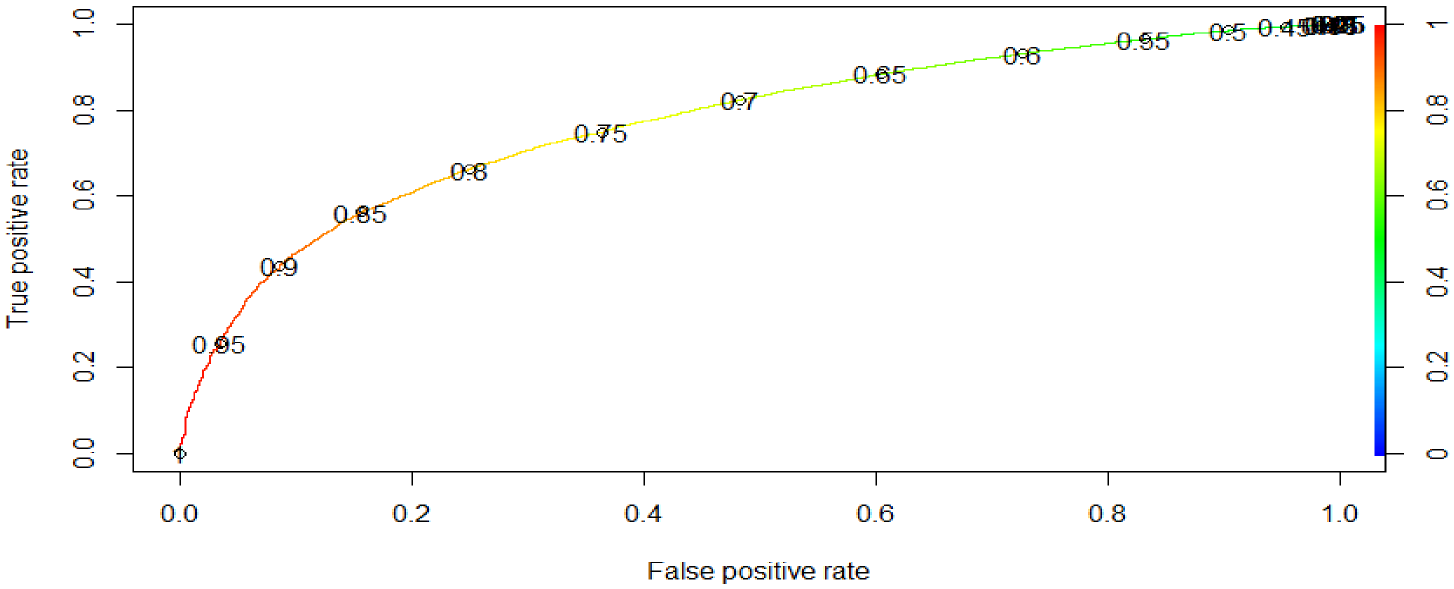
- Have probability for reference.
- The threshold value determines whether the probability value should be assigned to True or False.

Model Performance :

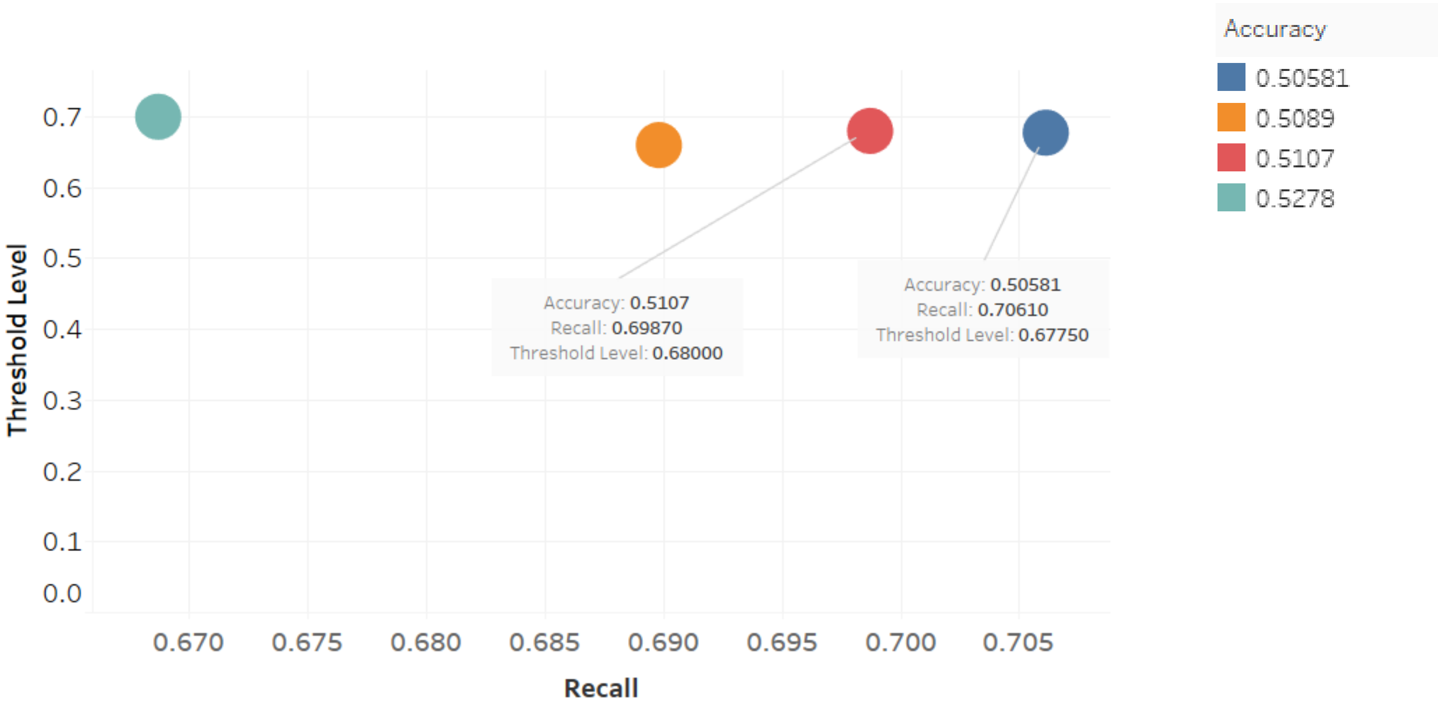
perc. under	perc. over	Threshold Level	Recall	Accuracy
28	550	0.7	66.87%	52.78%
		0.68	69.87%	51.07%
		0.6775	70.61%	50.58%
33	450	0.66	68.98%	50.89%

ROC Curve

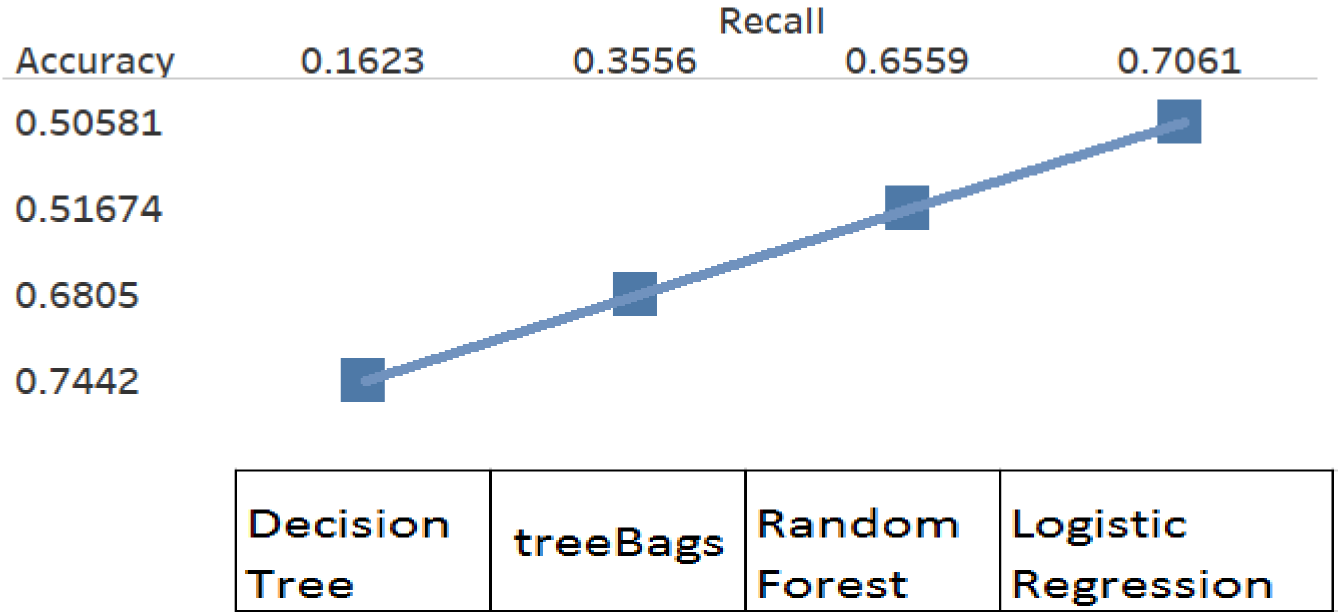
AUC = 77.12%



Threshold Level Vs Recall & Accuracy



Model Performance



Conclusion

- 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.

Q&A

Thanks