# ‘Hospital Readmission Predictor’

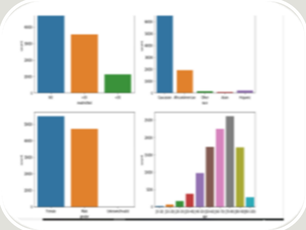
CAPSTONE PROJECT : MANJU NAIR

# Table of Contents

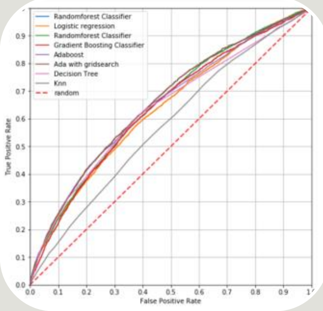
- Summary
- Context
- Datasets And Data Wrangling
- Exploratory Data Analysis – Findings
- Model Selection And Evaluation
- Results and Conclusion
- Future Work

# Summary

‘Hospital Readmissions Predictor’ uses the power of machine learning to identify a hospital’s previous inpatients who have more probability of getting readmitted within 30 days of discharge. This helps the hospitals to create interventions to provide additional assistance to patients with increased risk of readmission, and avoid getting penalized .



Finding : Features like number of medicines taken by the patient, number of procedures undergone, number, time spent in hospital , age etc. are important features that decide on the likelihood of a patient getting readmitted



Result:

Predicts with an accuracy of 67%, the patients with the likelihood of getting readmitted within 30 days of discharge from a hospital on unseen data.

Future Work : To further improve the model, some more of the categorical variables can be explored and added to feature engineering variable list to see how they influence the target variable.

Hyperparameter Tuning : Optimize the hyperparameters of the models to test the improvement in performance

# Context

- The Center for Medicare & Medicaid Services,(CMS) which is part of the Department of Health and Human Services (HHS) has created Hospital Readmission Reduction Program (HRRP),  to improve the quality of patient care .
- One way this is  implemented is  by reducing reimbursement to hospitals with above average readmissions as penalization. This in turn enforces the hospitals under this program to take steps to create interventions to provide additional assistance to patients with increased risk of readmission
- Diabetes is not yet included in the penalty measures, but data indicates that American hospitals spent over $41 billion per annum on diabetic patients who got readmitted within 30 days of discharge
- My project ‘Hospital Readmissions Predictor’   identifies previous inpatients who have more probability of getting readmitted within 30 days of discharge, using machine learning.

**Source : CMS Homepage

# Datasets and Data Wrangling

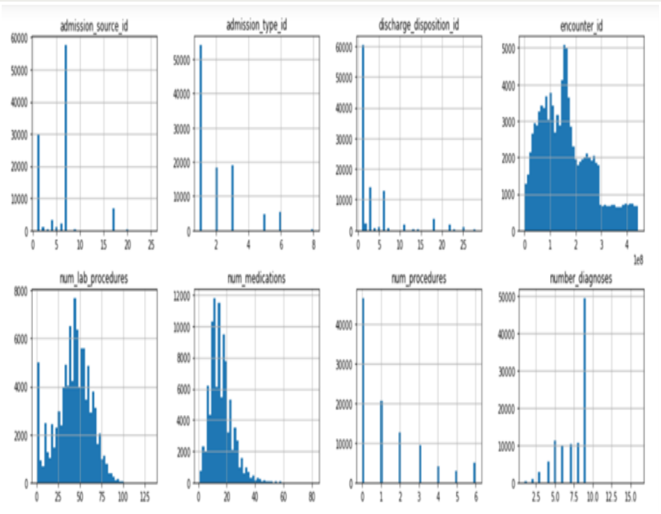
## Data sets

- ‘diabetic_data’ .csv data which has all the hospitalization related details of diabetic patients in hospital between 1998 and 2008 containing above 101,766 observations records with 50 variables
- ID’s mapping file which has the description of the various IDs used in the diabetic_data.csv

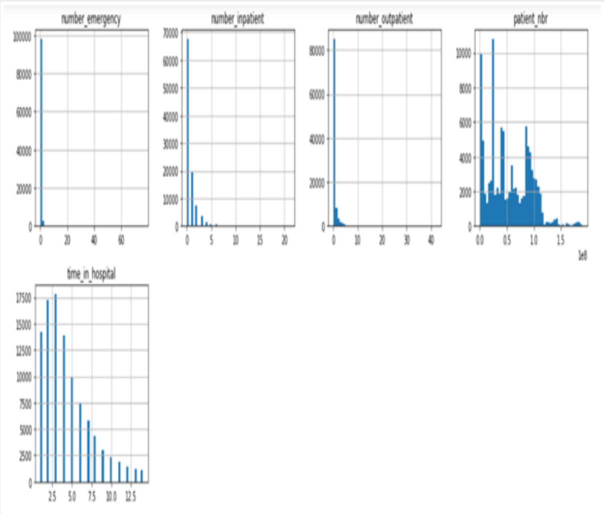
## Data Wrangling

- Some unnecessary attributes with a lot of missing values were dropped and a new columns introduced as target variable
- For categorical variables, One Hot Encoding was used to conver to numerical variables. For columns with a long list of categorical values, ‘others’ column was used to club barely used values
- Post feature engineering completion, there are 132 variables, of which 123 are categorical and 9 are numeric.

# Exploratory Data Analysis – Findings (1/3)



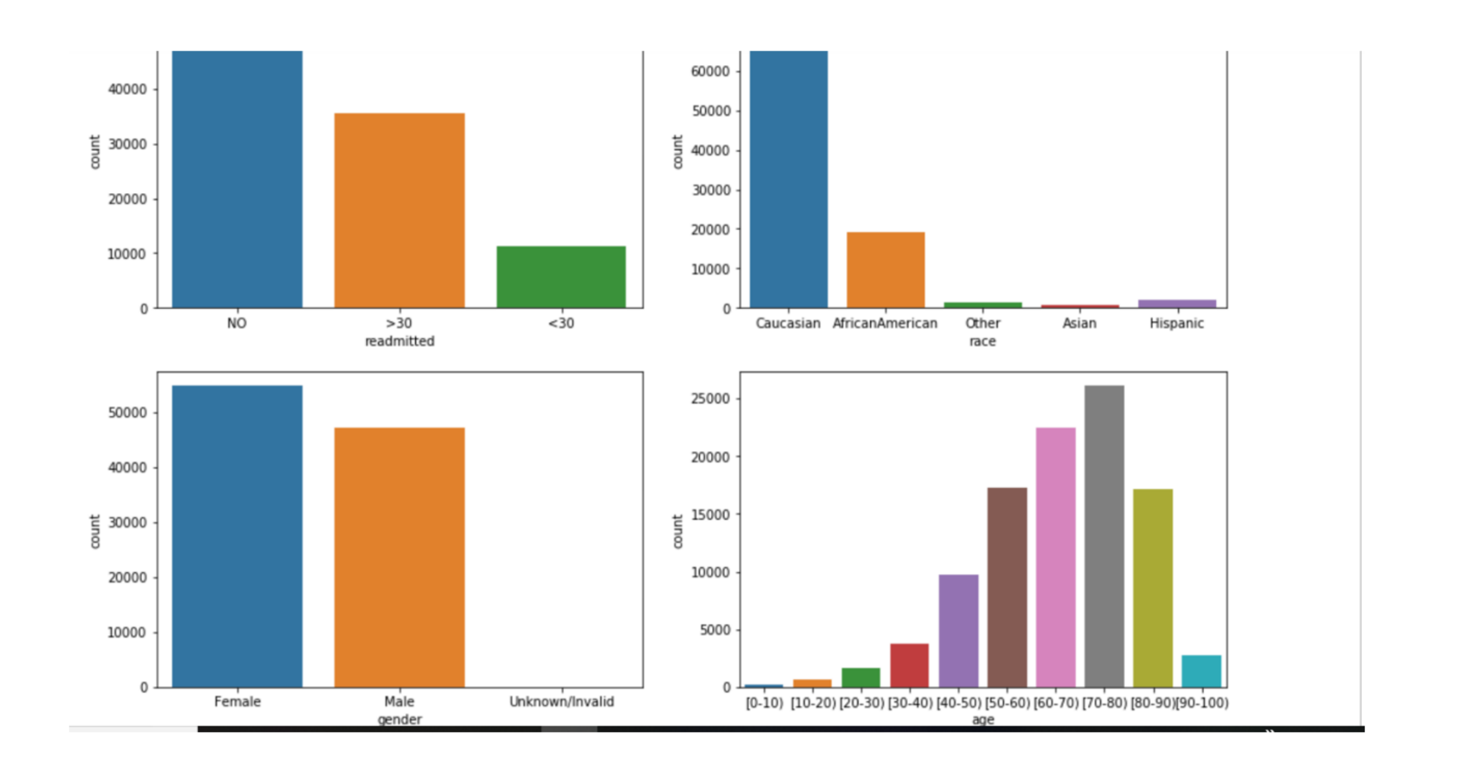
Distribution of fields that are numeric



Distribution of fields that are numeric

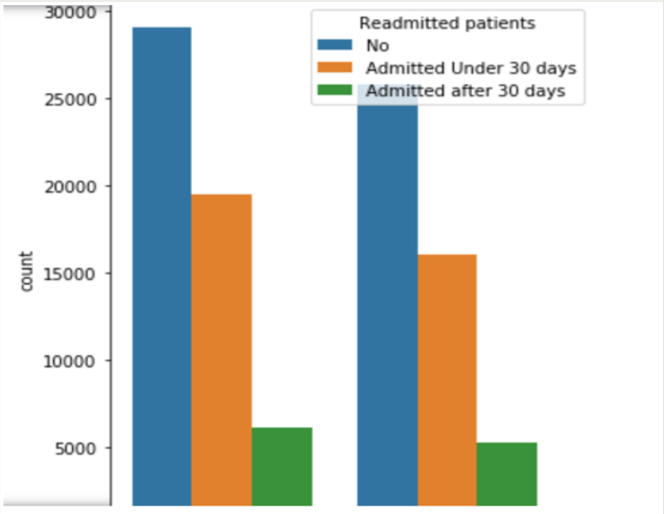
- It is interesting to note that the number of lab procedures and number of medications taken, follow a bell curve in the population getting admitted to the hospital. It would be worth the effort to check for influence.

# Exploratory Data Analysis – Findings (2/3)

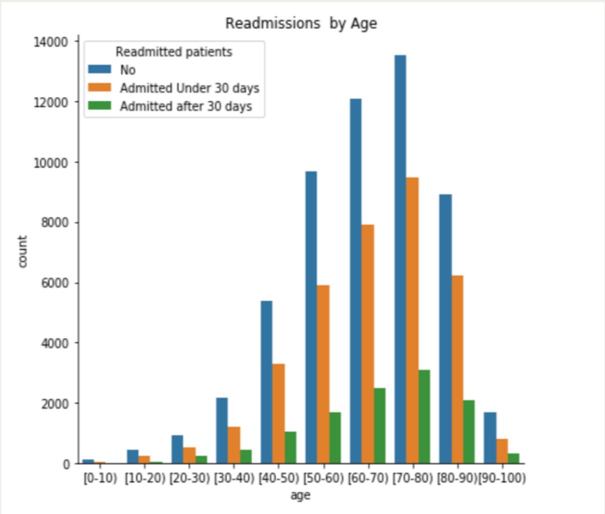


Readmission based on 'race' does not indicate that a specific race has more chances of readmission, as it could also be just reflective of ethnic distribution of population in that city. It is interesting to note that female population has more cases of re-admission. As expected, age pattern follows a bell curve.

# Exploratory Data Analysis – Findings (3/3)



It can be noted that readmission categorization follows similar pattern across both the genders.



The3 categories 'Admitted Under 30 days', 'Admitted after 30 days' and ' No readmissions' follow a bell curve peaking at 70-80s

- From the exploratory data analysis and statistical inferences drawn, though some patterns can be drawn, no significant enough information is seen to aid the prediction model.

# Model Selection and Evaluation

## Splitting Data into Training and Testing Samples

The data set is split into training and validation sets with the idea of measuring how well the model would work on unseen data. In this project, I split into 70% train, 30% validation data.

## Handling Class Imbalance

A quick glimpse at the number of readmission records and no readmission records indicate that the dataset is heavily skewed with ' no readmissions' at 88.61% and ' readmissions' at 11.38%.

This is called **class imbalance** . With a greater imbalanced ratio, the decision function favor the class with the larger number of samples, usually referred as the majority class. This results in inaccuracy in prediction.

Class balance is achieved by

**Subsampling /Undersampling:** where the records of dominant type are removed, to balance the data, or

**Oversampling** : where more of the non-dominant type records are added to balance the data

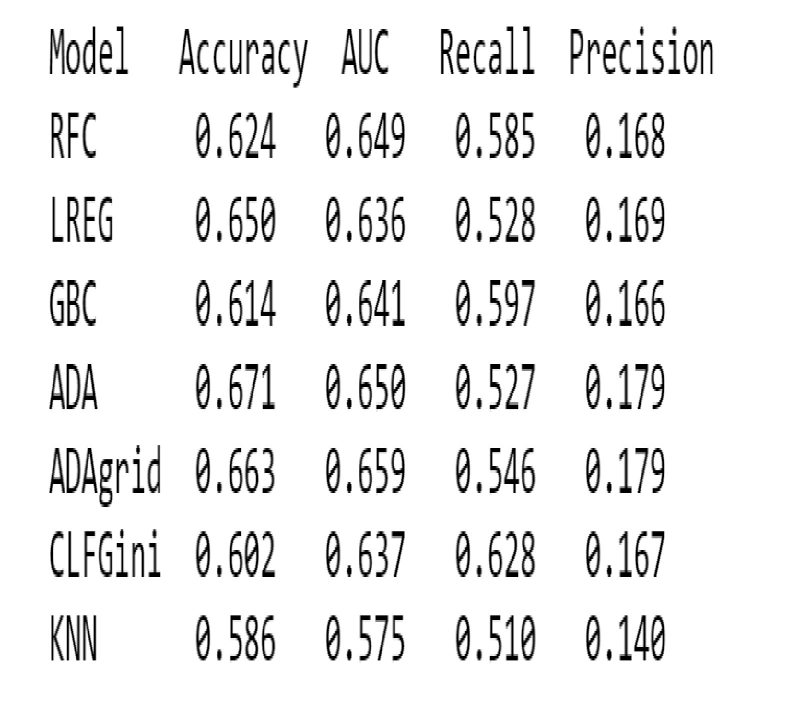
# Model Selection and Evaluation

Data set  is trained on the following algorithms and analyzed the predictive power by comparing their accuracy, auc, precision, recall and roc curves.

1. Random Forest Classifier
2. Logistic Regression
3. Gradient Boosting Classifier
4. AdaBoost Classifier
5. AdaBoost with GridSearchCV
6. Decision Tree
7. K Nearest Neighbor(knn)

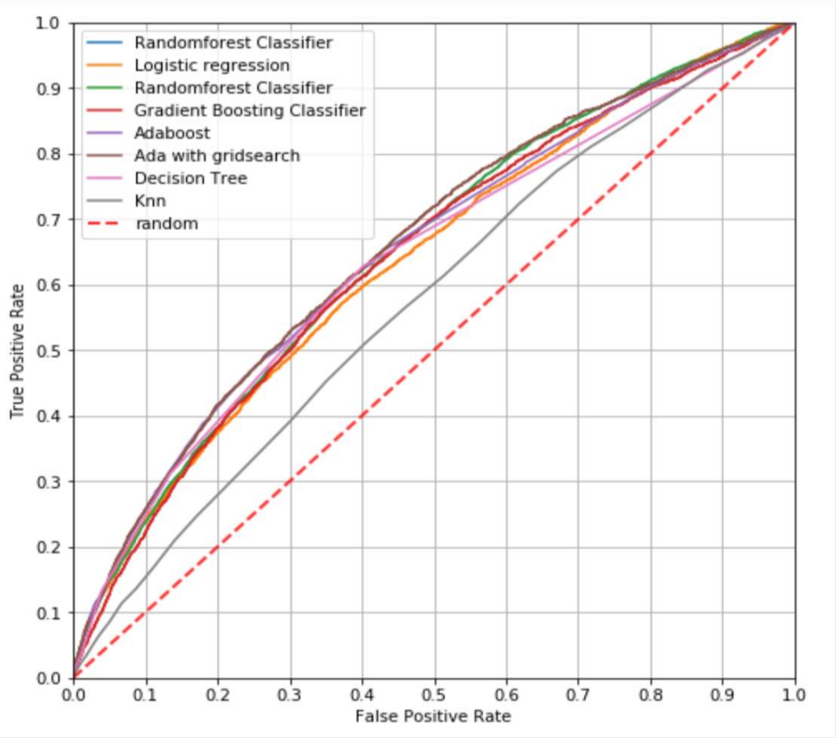
# Results and Conclusions

Results of the evaluation using the various models are given in the table below:



| Model | Accuracy | AUC | Recall | Precision |
|---|---|---|---|---|
| RFC | 0.624 | 0.649 | 0.585 | 0.168 |
| LREG | 0.650 | 0.636 | 0.528 | 0.169 |
| GBC | 0.614 | 0.641 | 0.597 | 0.166 |
| ADA | 0.671 | 0.650 | 0.527 | 0.179 |
| ADAgrid | 0.663 | 0.659 | 0.546 | 0.179 |
| CLFGini | 0.602 | 0.637 | 0.628 | 0.167 |
| KNN | 0.586 | 0.575 | 0.510 | 0.140 |

ROC curve comparison is given below



# Results and Conclusions

1) Looking at the tabular column, ADA Boosting Classifier with GridSearchCV has the best value for AUC making it the model of choice for prediction in this project

2) However, accuracy of all the models ranges between 58-67%. Further, applying more pre-processing techniques might help improve the accuracy.

3) With this ML model, it is possible to predict the probability of a patient with diabetes getting readmitted within 30 days , with 66% accuracy and AUC of 0.67.

# Future Work

1. To further improve the model, some more of the categorical variables can be explored and added to feature engineering variable list to see how they influence the target variable.
   - One such variable would be diag1, diag2, diag3—they are categorical and have a lot of values.
2. Hyperparameter Tuning: In the current model, the default values for hyperparameters were used for modeling. Future work can include tuning the hyperparameters to see whether the models can be improved.

# Thank You

## Acknowledgments

Srdjan Santic