Springboard Data Science Course
Data Science Project

‘Hospital Readmission Predictor’

Manju Nair (2019)

Contents

Introduction

Hospital readmission is a high-priority health care quality measure and target for cost reduction currently in healthcare domain. A hospital readmission is when a patient who has been discharged from the hospital, gets re-admitted again in less than 30 days.

Hospital readmission rates for certain conditions are now considered an indicator of hospital quality, and also affect the cost of care adversely. 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, is one such initiative of Centers for Medicare & Medicaid Services

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. Reducing readmission rates of diabetic patients has the potential to greatly reduce health care costs while simultaneously improving care.

My project 'Hospital Readmissions Predictor' identifies previous inpatients with the probability of getting readmitted, using machine learning.

Overview of Data Set

I used a publicly available dataset from UCI repository (link) containing diabetes patients data for 130 US hospitals spanning a decade (1999–2008) containing 101,766 observations over 10 years. The dataset has over 50 features including patient characteristics, conditions, tests and medications history.

Data Wrangling

Initial examination of the dataset shows that there are over 100,000 records and around 50 data columns of which a few are 'Id' columns and the rest are either numeric or categorical data.

1. Cleaning and Consolidating the Data

The column 'discharge_disposition_id' indicates what happened to patient post hospitalization. From the csv file, the ids "# 11, 13, 14, 19, 20, 21" are related to death or hospice, so those can be dropped. Significant number of categorical variables are present, and 'One Hot Encoding' was used to convert them into numerical variables.

2. Missing Values

A couple of columns like 'weight' and 'payer_code' where around 98% of rows were blank were dropped. 'Dropna' and 'fillna' were used to drop or fill blank columns as applicable to the scenario.

3. Outliers

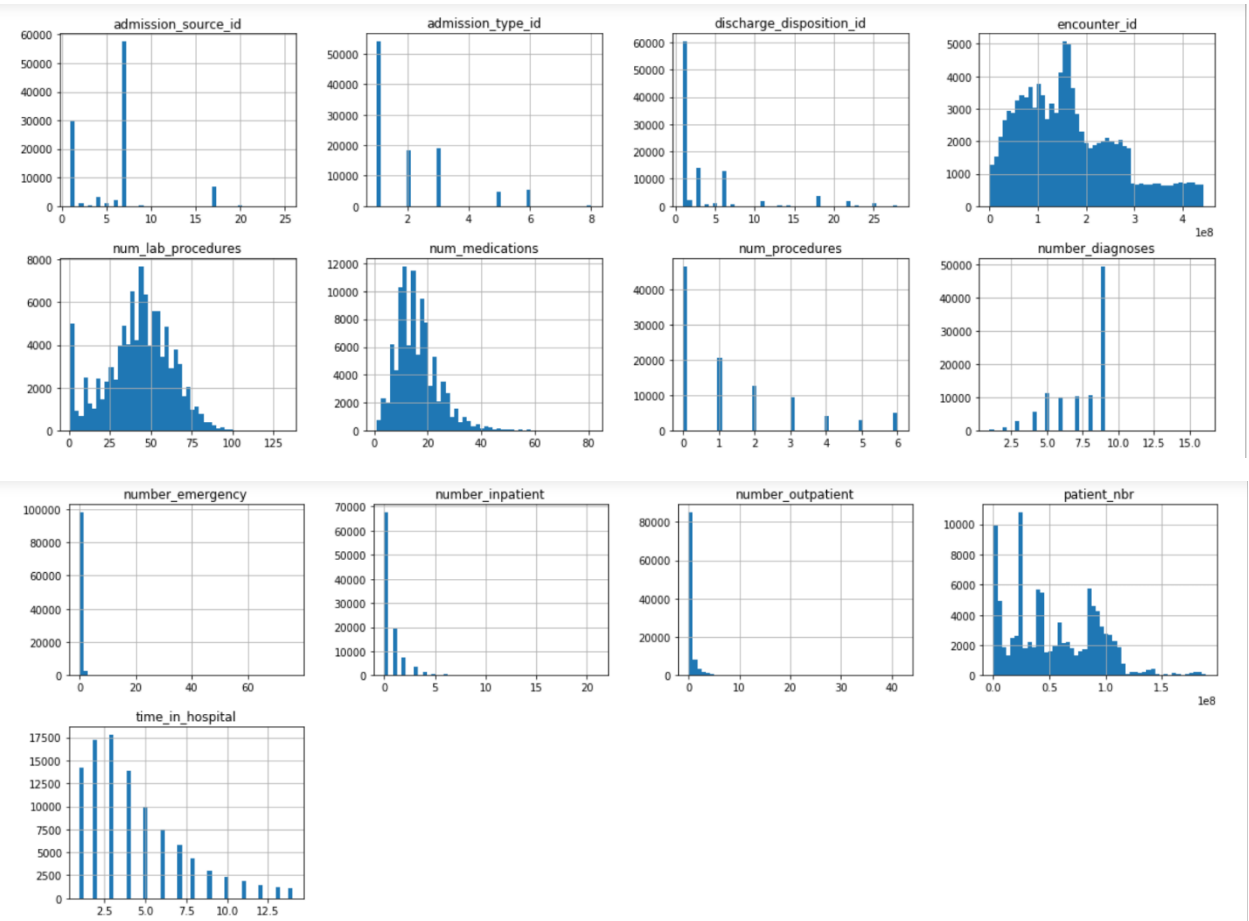
There were not much significant outliers to be worked upon

Exploratory Data Analysis

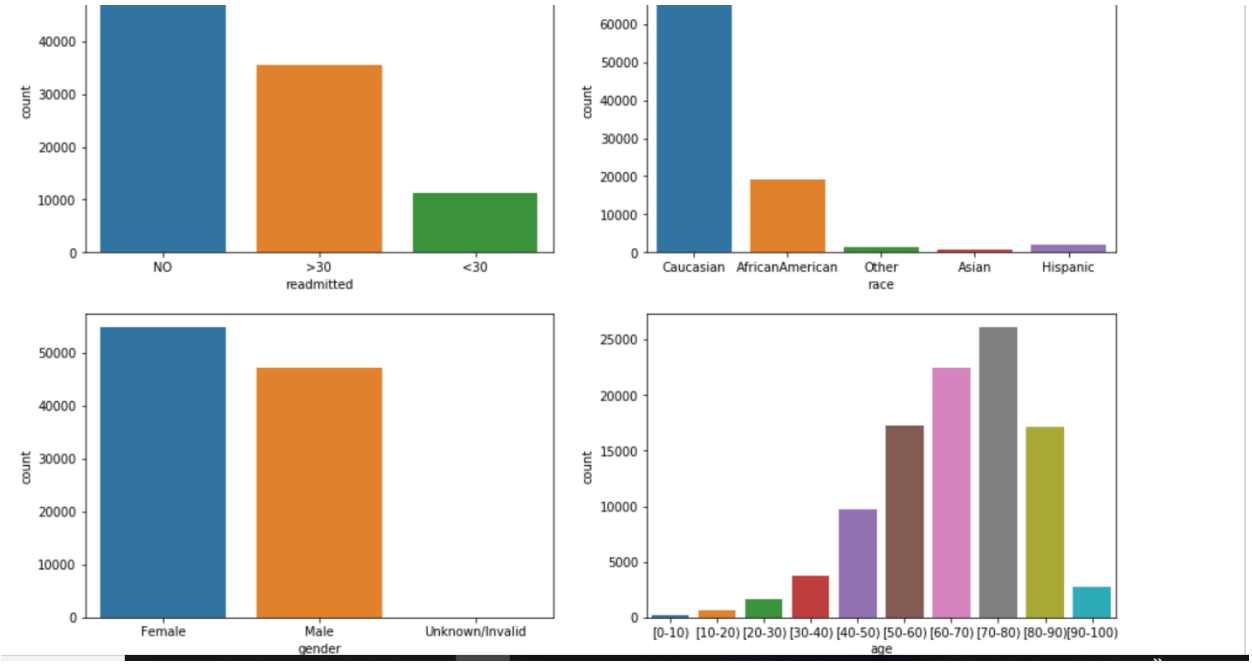
'Diabetic_data.csv' data file has a huge set of feature variables and doing a thorough EDA is key to finding out , which of this have the maximum impact on predicting which patients have the likelihood of getting readmitted under 30 days post discharge from hospital. The task is in finding out the list of key feature variables.

Exploratory data analysis is done to get a basic sense about the data. Since there are numerous variables, finding out any correlation will help in reducing dimensionality of data.

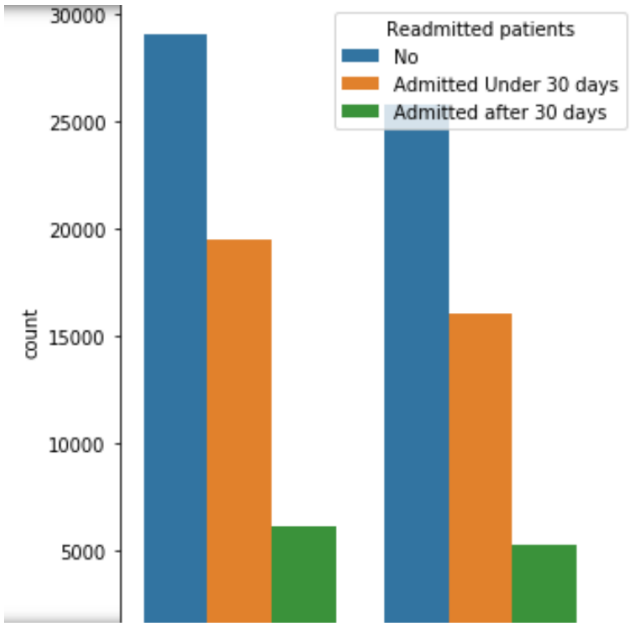
Below given visualization will help understand the dataset better.



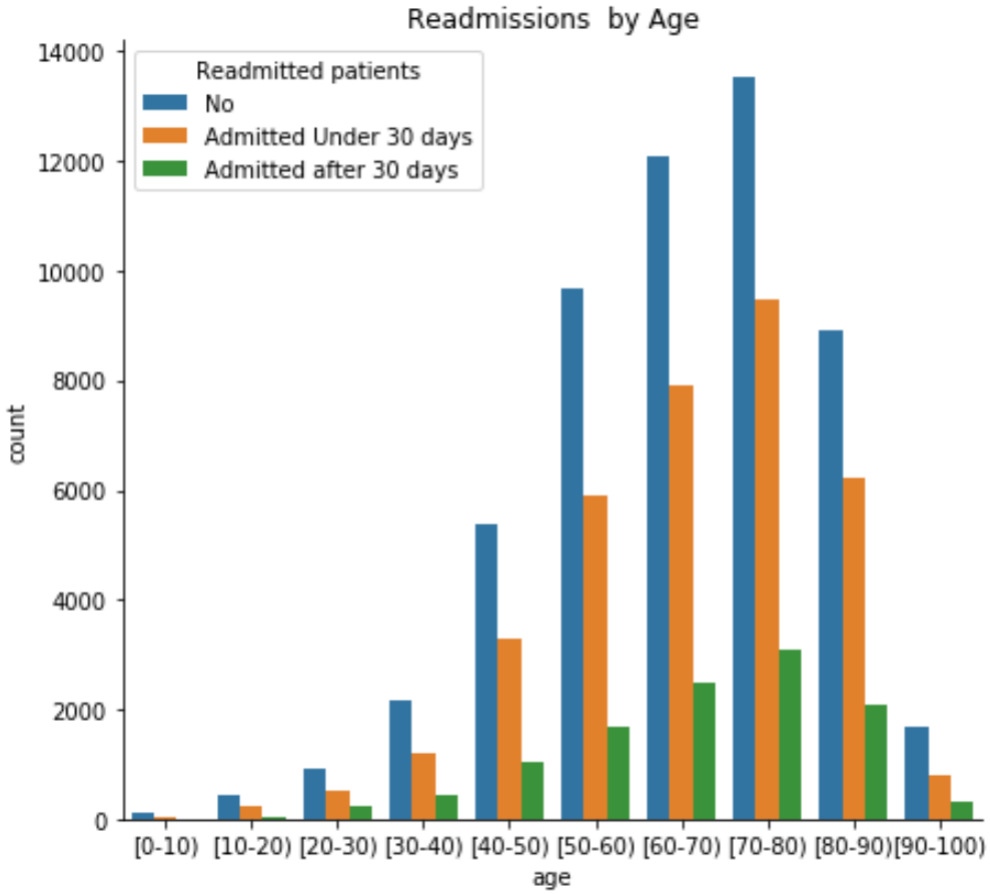
It is interesting to note that the number of lab procedures and number of medications taken, follow a bell curve in the population getting admitted to the hospital. It would be worth the effort to check for influence.



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.



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



All the 3 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 trends can be drawn, no significant enough information is seen to proceed with the modeling. As next step, feature importance can be used to order the feature variables according to their significance, which would help the prediction.

Feature Engineering

In the feature engineering section, the list of variables are closely analyzed to decide which ones have to be retained for the predictive model. There were a significant number of categorical variables. They were converted to numerical variables using the 'One Hot Encoding' method.

The field ' medical_specialty' has a large range of around 75 categorical values'. Since One Hot encoding creates one column corresponding to each unique value, this would result in the creation of extra 75 columns, contributing to data sparsity. So I decided to keep the top 20 values for 'medical_specialty' and consolidate the remaining values to a single column 'others'.

The column age is given as age range. I convert it to numeric form of the corresponding median value for each range.

The target variable 'readmission_label' is created and assigned value of '1' or '0' depending on whether the patient was readmitted under 30 days or not as stored in the variable' readmitted'.

Post feature engineering completion, there are 132 variables, of which 123 are categorical and 9 are numeric.

Splitting Data into Training and Testing Samples

The data set was split into training and validation sets. The idea behind splitting the data is to measure how well the model would work on unseen data. In this project, I split into 70% train, 30% validation data.

Handling Class Imbalance

At this point, I would like to pause and ponder about class imbalance of the dataset. 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 .In class imbalance, the model gets trained with more records of dominant type(That being' 'no-readmission' class here), resulting in the model predicting incorrectly for non-dominant type records as there were insufficient number of them in the training set.

With a greater imbalanced ratio, the decision function favor the class with the larger number of samples, usually referred as the majority class.T his results in inaccuracy in prediction.

In such cases, the data is balanced through

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 used to balance the data

For negating the class imbalance, I did both oversampling and undersampling to balance the data and then trained the Random Forest Classifier on both the resultant balanced data set. It can be observed that with oversampling, although the accuracy is 88%, the True Negative prediction is very less, '110'. In the case of training data set balanced using undersampling, although the accuracy is only 62%, the TN prediction is '1991'. Since that's the non -dominant class, accuracy in predicting non-dominant class correctly would be taken as criteria for the preferred balanced data set, so I would be proceeding with testing further algorithms on balanced dataset obtained by undersampling.

Model Selection and Evaluation

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

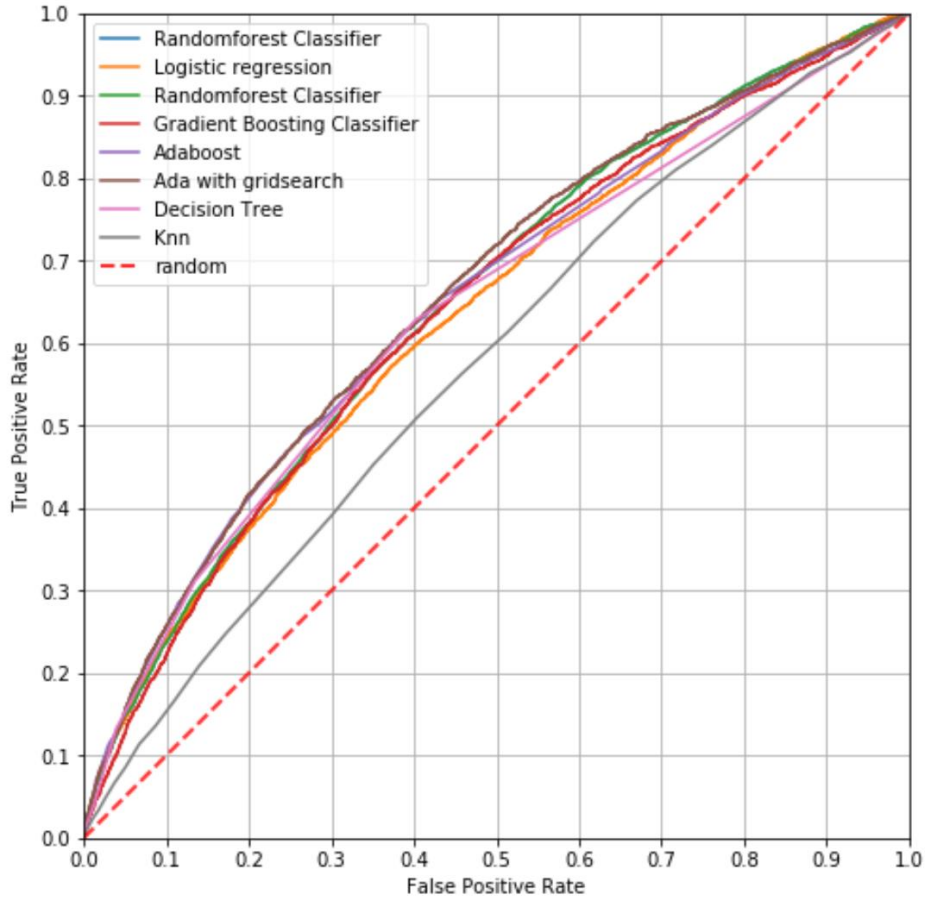
1. Random Forest Classifier
2. Logistic Regression
3. Gradient Boosting Classifier
4. AdaBoost Classifier
5. AdaBoost with GridSearchCV
6. Decision Tree
7. K Nearest Neighbor(knn)

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:



1) Looking at the tabular column, ADA Boosting Classifier with GridSearchCV has the best value for AUC and would be 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. In future version, this would be added.
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.

Appendix

List of Ids

admission_type_id	description
1	Emergency
2	Urgent
3	Elective
4	Newborn
5	Not Available
6	NULL
7	Trauma Center
8	Not Mapped
discharge_disposition_id	description
1	Discharged to home
2	Discharged/transferred to another short term hospital
3	Discharged/transferred to SNF
4	Discharged/transferred to ICF
5	Discharged/transferred to another type of inpatient care institution
6	Discharged/transferred to home with home health service
7	Left AMA
8	Discharged/transferred to home under care of Home IV provider
9	Admitted as an inpatient to this hospital
10	Neonate discharged to another hospital for neonatal aftercare
11	Expired
12	Still patient or expected to return for outpatient services
13	Hospice / home
14	Hospice / medical facility
15	Discharged/transferred within this institution to Medicare approved swing bed
16	Discharged/transferred/referred another institution for outpatient services
17	Discharged/transferred/referred to this institution for outpatient services
18	NULL
19	Expired at home. Medicaid only, hospice.
20	Expired in a medical facility. Medicaid only, hospice.
21	Expired, place unknown. Medicaid only, hospice.
22	Discharged/transferred to another rehab fac including rehab units of a hospital .
23	Discharged/transferred to a long term care hospital.
24	Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.

25	Not Mapped
26	Unknown/Invalid
30	Discharged/transferred to another Type of Health Care Institution not Defined Elsewhere
27	Discharged/transferred to a federal health care facility.
28	Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
29	Discharged/transferred to a Critical Access Hospital (CAH).
admission_source_id	description
1	Physician Referral
2	Clinic Referral
3	HMO Referral
4	Transfer from a hospital
5	Transfer from a Skilled Nursing Facility (SNF)
6	Transfer from another health care facility
7	Emergency Room
8	Court/Law Enforcement
9	Not Available
10	Transfer from critial access hospital
11	Normal Delivery
12	Premature Delivery
13	Sick Baby
14	Extramural Birth
15	Not Available
17	NULL
18	Transfer From Another Home Health Agency
19	Readmission to Same Home Health Agency
20	Not Mapped
21	Unknown/Invalid
22	Transfer from hospital inpt/same fac reslt in a sep claim
23	Born inside this hospital
24	Born outside this hospital
25	Transfer from Ambulatory Surgery Center
26	Transfer from Hospice