
HOSPITAL
EXIT
HOSPITAL
ENTRANCE

Predict Hospital Readmission rates of Diabetes patients

Bindu Jacob

Capstone project

SpringBoard Data Science Career Track

What is diabetes?

- Diabetes mellitus, commonly known as diabetes, is a metabolic disease that causes high blood sugar.
- Occurs when the pancreas does not produce enough insulin
- Or when the body cannot effectively use the insulin it produces

Who has diabetes and why is it important?

- 34.2 million people, or 10.5% of the U.S. population, have diabetes.
    - An estimated 26.8 million people or 10.2% of the population - had diagnosed diabetes.
    - Approximately 7.3 million people have diabetes but have not yet been diagnosed (2018).
- High prevalence of diabetes makes it a common comorbid condition in hospitalized patients

Why we need to focus on 30-day Readmissions?

- It is a high-priority health care quality measure and target for cost reduction.
- The cost due to this is estimated to be close to $25 billion per year in the U.S. Patients with diabetes are frequently admitted to the hospital.
- Medicare (Insurance) evaluates this metric and penalizes the hospital.

For these reasons, reducing the rate of hospital readmissions has great potential to help constrain health care costs and improve the quality of health care.

What is the problem to solve?

Predict if a diabetes patient will be readmitted within 30 days after the patient was discharged from the hospital, based on the patient's medical history information.

Who might care?

- Government agencies
- Healthcare systems
- Hospitals
- Health Insurance companies

Dataset

- Source - UCI Machine learning Repository - Diabetes 130-US hospitals for years 1999-2008
- The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks.
- This dataset has 101766 observations and 50 attributes
- Each observation had the patient's medical history of the hospital encounter, demographic data, diagnoses, procedures and time in hospital etc

Data Wrangling

Missing data

- There are several attributes with missing data, and as they are unique to a patient, these records were removed
- weight attribute was dropped due to 97% missing values

Duplicate data

- Several patients had more than one encounter with the hospital and so only one encounter was retained

Mapping of IDs

- Several attributes had ID's and so had to remap these with the corresponding text

After data cleaning final dataset contained 70413 observations and 48 attributes

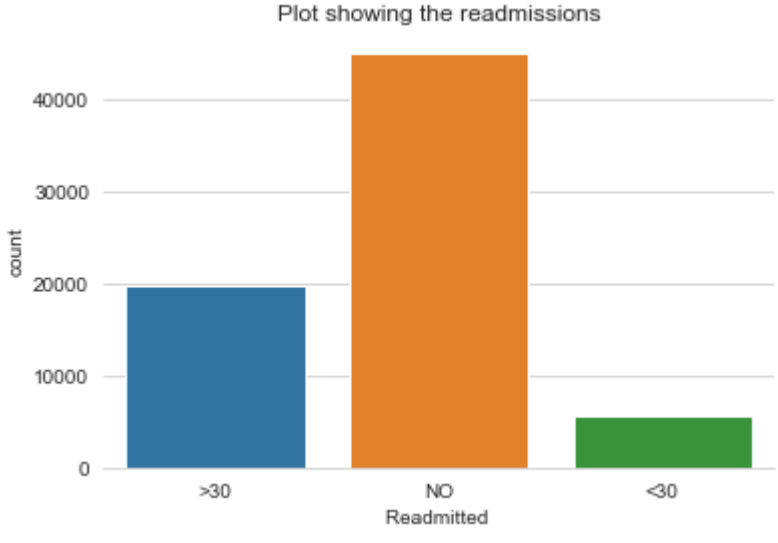
Exploratory Data Analysis

Target

- readmitted
- Multiclass classification

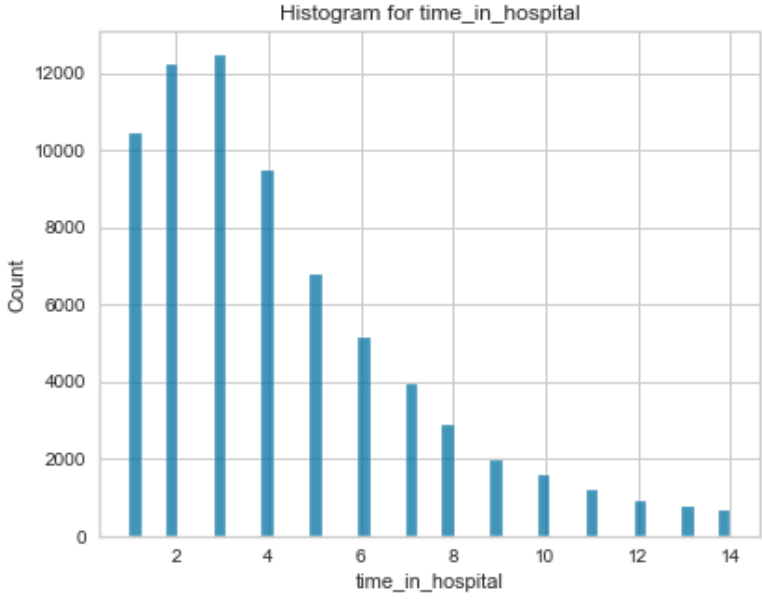
Features

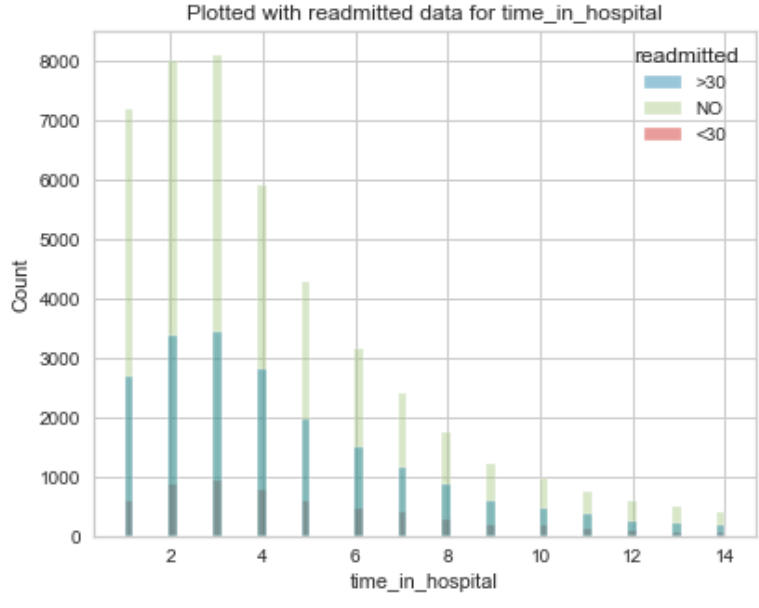
- patient medical history : time in hospital, number of diagnoses, number of medications, number of procedures, prior visits
- patient demographic data : race, gender, age
- medications given to the patients
- ICD9-codes diagnoses

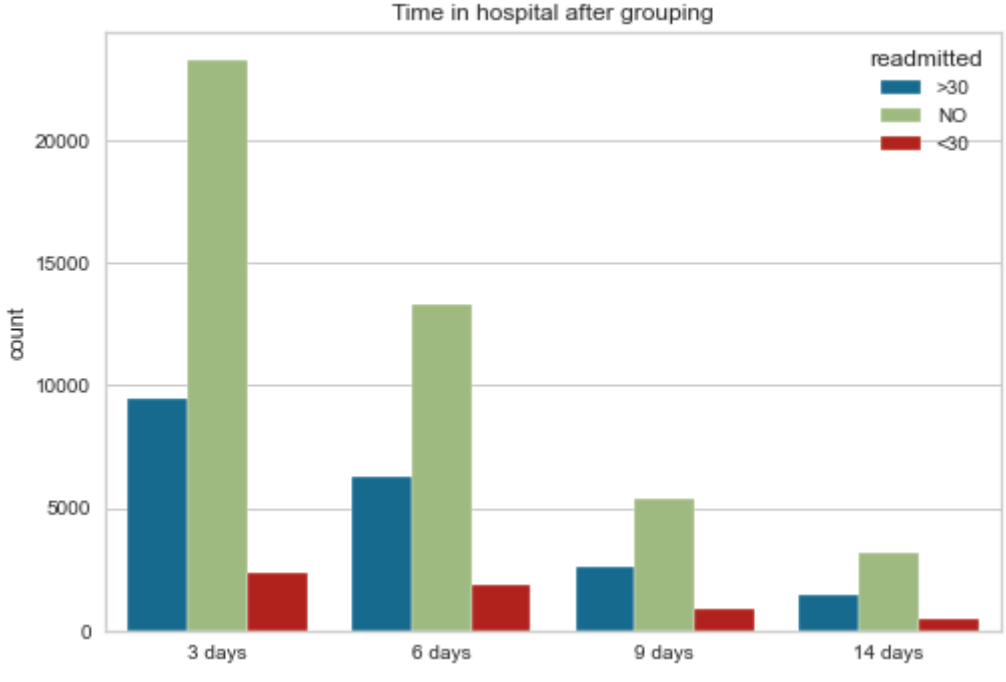

Plot showing the readmissions
40000
30000
20000
10000
0
count
>30
NO
<30
Readmitted

Numerical Data

- After plotting and evaluating the various numerical data, we could see a clear pattern for binning these data to categories.
- Majority of patients stayed in the hospital for 3 days
- All the numerical data were binned similar to time in hospital attribute


Histogram for time_in_hospital
Count
time_in_hospital
0
2000
4000
6000
8000
10000
12000
2
4
6
8
10
12
14


Plotted with readmitted data for time_in_hospital
readmitted
>30
NO
<30
Count
time_in_hospital
0
1000
2000
3000
4000
5000
6000
7000
8000
2
4
6
8
10
12
14


Time in hospital after grouping
readmitted
>30
NO
<30
count
20000
15000
10000
5000
0
3 days
6 days
9 days
14 days

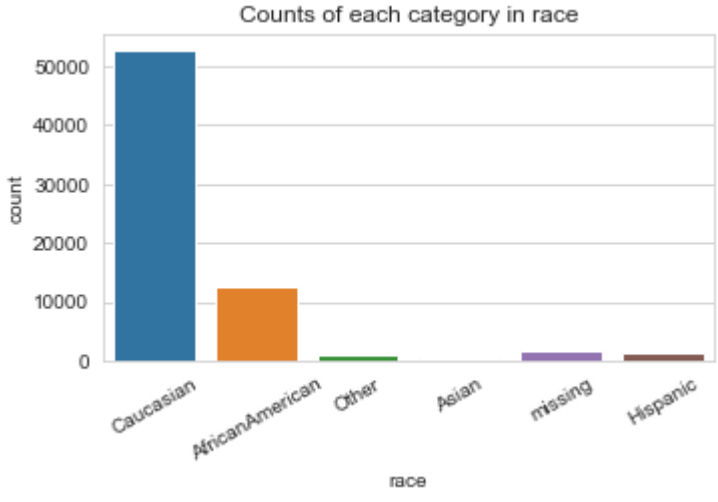
Categorical Data

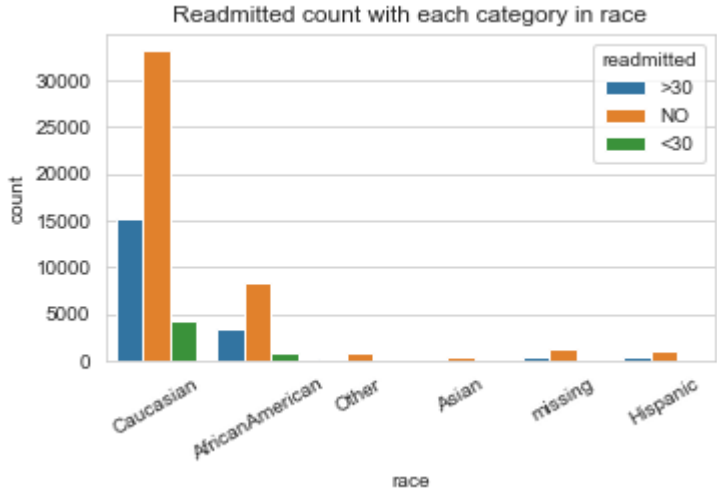
Patient Demographic data

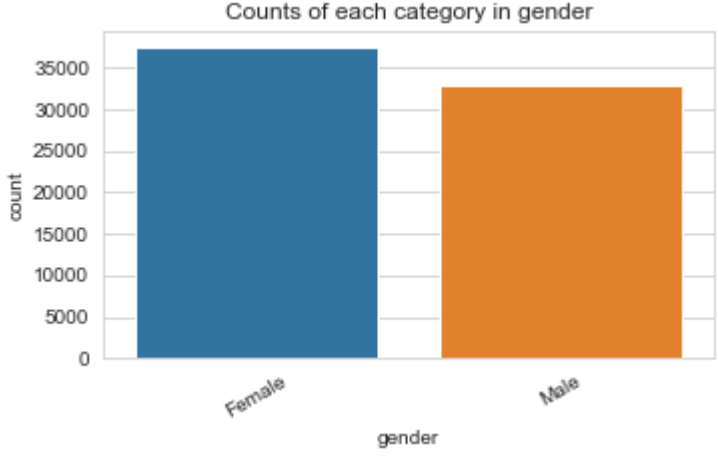
From the plotted graph we can see that majority of the patients are caucasian, with females slightly outnumbering males and most of the patients' ages range from 60 to 80 years.

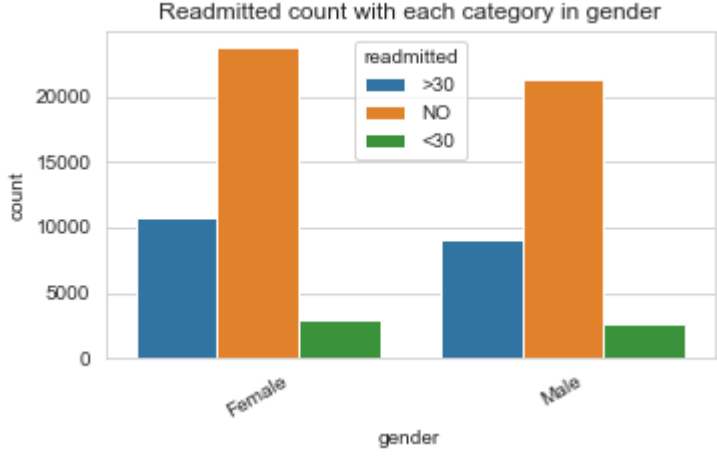
Other features

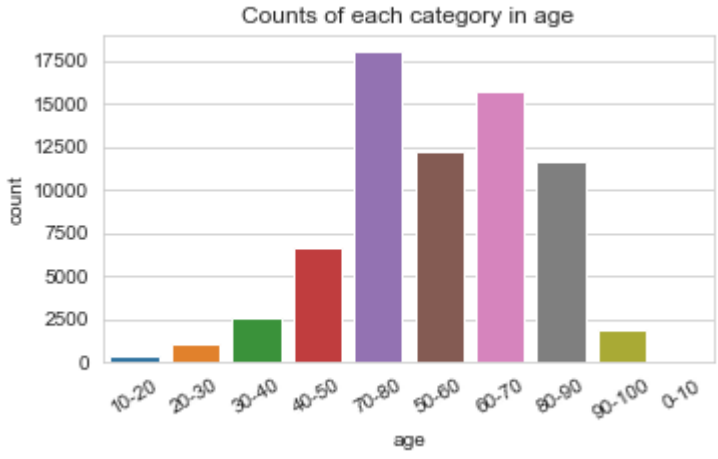
All other features where plotted and analyzed and features with high cardinality were re-mapped to bring down the variances in the data

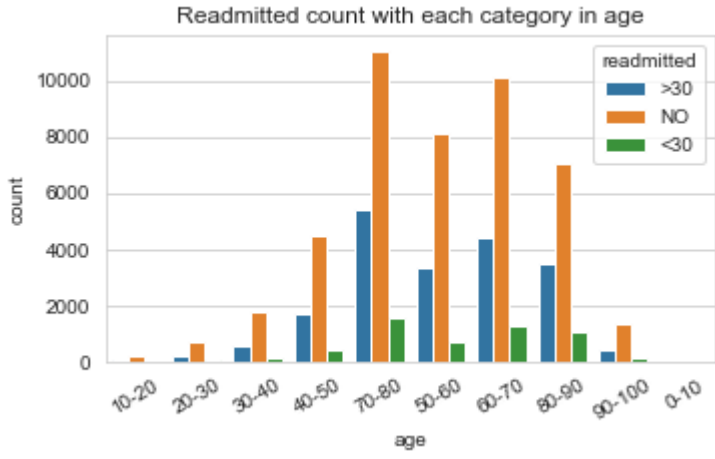

Counts of each category in race
count
race
Caucasian
AfricanAmerican
Other
Asian
missing
Hispanic


Readmitted count with each category in race
readmitted
>30
NO
<30
count
race
Caucasian
AfricanAmerican
Other
Asian
missing
Hispanic


Counts of each category in gender
count
gender
Female
Male


Readmitted count with each category in gender
readmitted
>30
NO
<30
count
gender
Female
Male


Counts of each category in age
count
age
0
2500
5000
7500
10000
12500
15000
17500
10-20
20-30
30-40
40-50
70-80
50-60
60-70
80-90
90-100
0-10


Readmitted count with each category in age
readmitted
>30
NO
<30
count
age
0
2000
4000
6000
8000
10000
10-20
20-30
30-40
40-50
70-80
50-60
60-70
80-90
90-100
0-10

Models

Data modelling was done using Pycaret

Top 3 models with best outcomes are:

1. CatBoost Classifier
2. Light Gradient Boosting Machine (Light GBM)
3. Extreme Gradient Boosting (XGBoost)

Tuning models

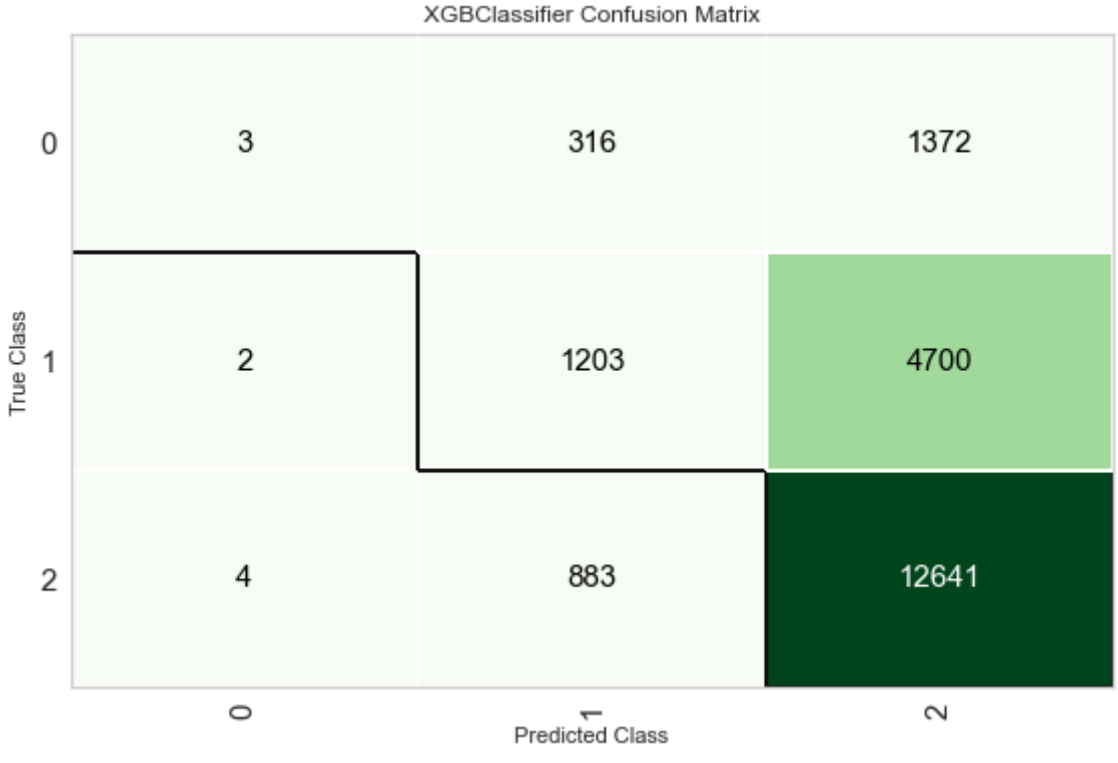
After tuning all the chosen models, XGBoost gave a slightly better performance

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6573	0.6729	0.3816	0.5785	0.5860	0.1451	0.1822
1	0.6547	0.6748	0.3784	0.5937	0.5810	0.1364	0.1730
2	0.6588	0.6721	0.3807	0.6059	0.5843	0.1427	0.1836
3	0.6573	0.6666	0.3812	0.6287	0.5844	0.1426	0.1809
4	0.6478	0.6656	0.3712	0.5855	0.5719	0.1168	0.1493
5	0.6462	0.6643	0.3695	0.5731	0.5696	0.1116	0.1435
6	0.6622	0.6789	0.3866	0.6060	0.5915	0.1574	0.1979
7	0.6541	0.6713	0.3780	0.5709	0.5816	0.1372	0.1722
8	0.6543	0.6697	0.3767	0.5722	0.5796	0.1333	0.1705
9	0.6514	0.6758	0.3737	0.6455	0.5748	0.1222	0.1585
Mean	0.6544	0.6712	0.3778	0.5960	0.5805	0.1345	0.1712
SD	0.0047	0.0045	0.0049	0.0242	0.0064	0.0133	0.0158

Confusion Matrix

Encoded labels are : '<30': 0, '>30': 1, 'NO': 2

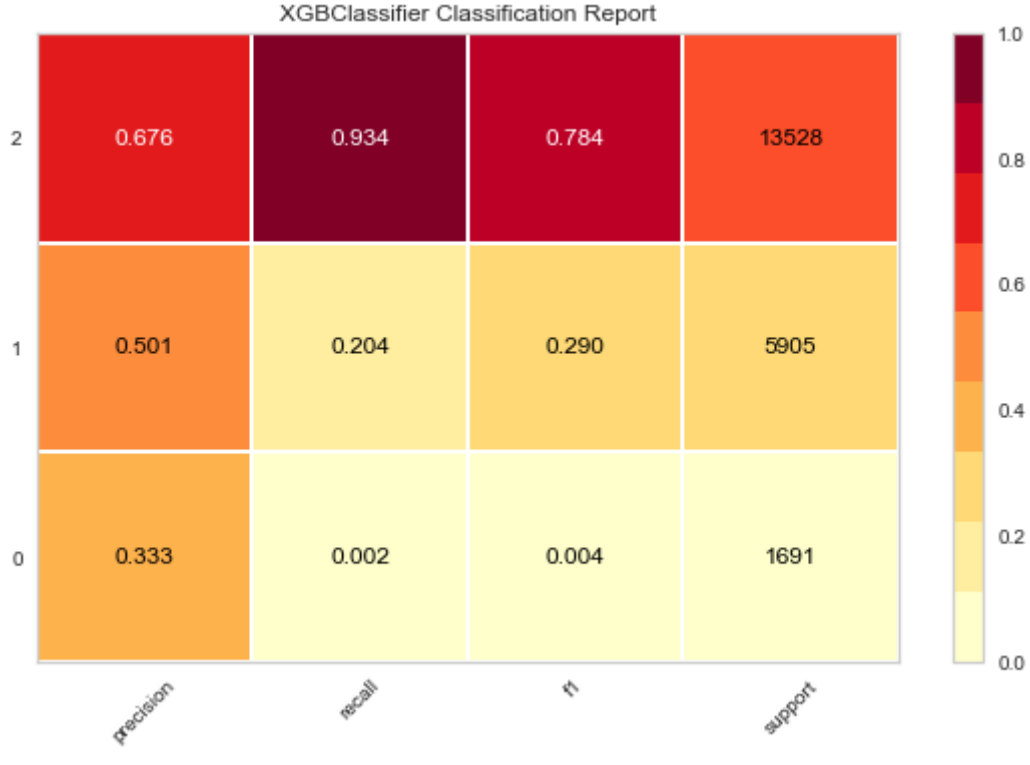
- We can see from this matrix that the class imbalance of data has influenced the results.
- For less than 30 days false positives outweighs the false negatives


XGBClassifier Confusion Matrix
True Class
0
1
2
3
316
1372
2
1203
4700
4
883
12641
0
1
2
Predicted Class

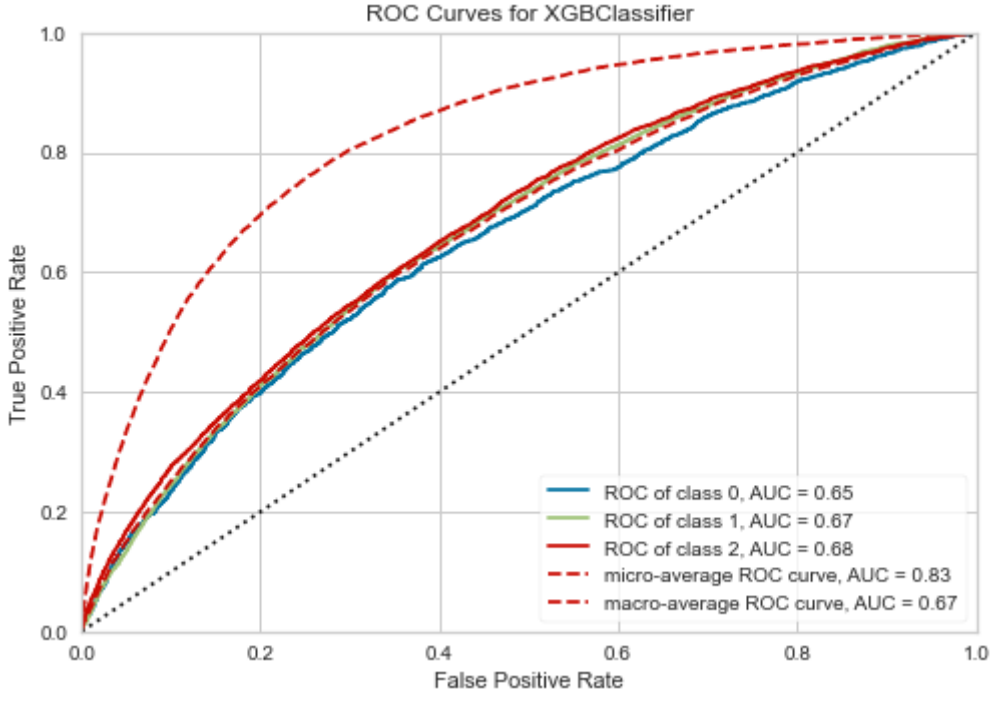
Classification Report

Encoded labels are : '<30': 0, '>30': 1, 'NO': 2

- The model does a better job of predicting that the patient will not be readmitted
- The model does not predict well if the patient will be not be readmitted for less than 30 days

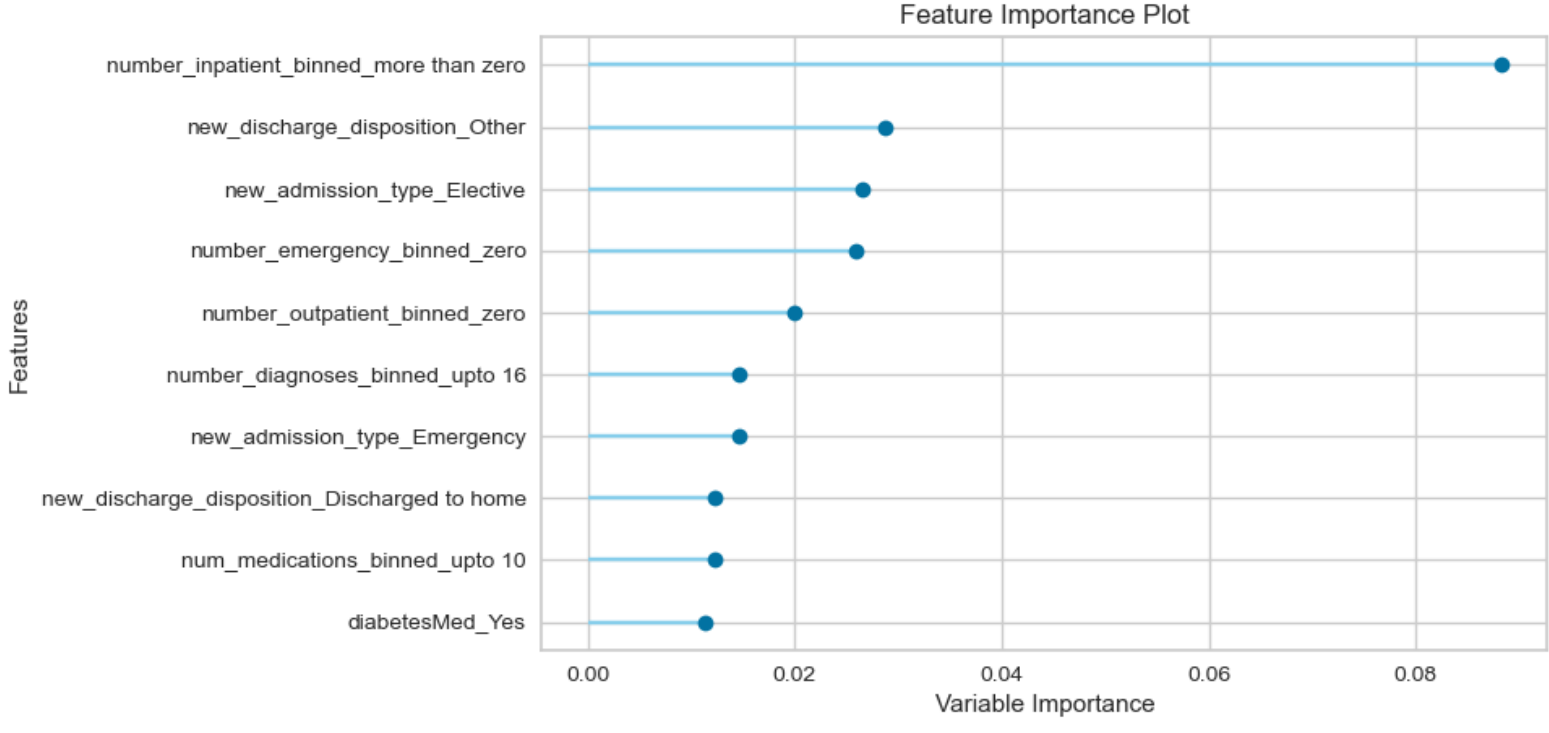

XGBClassifier Classification Report
2
1
0
0.676
0.934
0.784
13528
0.501
0.204
0.290
5905
0.333
0.002
0.004
1691
precision
recall
f1
support
1.0
0.8
0.6
0.4
0.2
0.0

Plot AUC


ROC Curves for XGBClassifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0, AUC = 0.65
ROC of class 1, AUC = 0.67
ROC of class 2, AUC = 0.68
micro-average ROC curve, AUC = 0.83
macro-average ROC curve, AUC = 0.67

Feature Importance

- One of the main feature of importance was the prior number of inpatient visits by the patient
- Patient discharge details, type of admission and prior number of emergency visits by the patient are also other importance features


Feature Importance Plot
number_inpatient_binned_more than zero
new_discharge_disposition_Other
new_admission_type_Elective
number_emergency_binned_zero
number_outpatient_binned_zero
number_diagnoses_binned_upto 16
new_admission_type_Emergency
new_discharge_disposition_Discharged to home
num_medications_binned_upto 10
diabetesMed_Yes
Features
0.00
0.02
0.04
0.06
0.08
Variable Importance

Prediction using XGBoost on the test data

- Accuracy is slightly better than on the trained data, and this shows that the model is resilient to overfitting

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Extreme Gradient Boosting	0.6555	0.6744	0.3800	0.5993	0.5834	0.1406	0.1763

Take aways from this model and future works

- prior inpatient or emergency patient visits to the hospital is a good indicator of whether the patient will be readmitted or not
- Hospitals can take steps to better their inpatient and emergency care
- Data was unbalanced and so the model is biased towards the more common class, 'NO'

- Further feature reduction techniques could be used to get a better model
- Could use techniques like SMOTE to solve data imbalance issue

Acknowledgement

SpringBoard mentor Raghunandan Patthar for his guidance and support