HOSPITAL READMISSION WITHIN 30 DAYS PREDICTION

TASMIA KAYENAT

SPRINGBOARD - DATA SCIENCE CAREER TRACK

AGENDA

Problem Statement

Data Overview

Exploratory Data Analysis (EDA)

Preprocessing and Training

Modeling

Key Insights

Recommendation

Q&A

PROBLEM STATEMENT

Everyday we see thousands of patients getting readmitted to the hospitals even after getting the same care.

- Is age a factor here?
- Does race make a difference when it comes to hospital readmissions?
- Can we predict the pattern of readmissions based on the variables like age, race, gender?

My aim is to develop a machine learning model that predicts the likelihood of a patient being readmitted to the hospital within 30 days after discharge.

DATA OVERVIEW

THE RAW DATA WAS TAKEN FROM KAGGLE, WHICH IS THE DATA OF DIABETIC PATIENTS BEING ADMITTED TO 130 US HOSPITALS BETWEEN THE YEARS 1999 - 2008.

DATA OVERVIEW

Initial dataset: 101766 rows & 50 columns

Cleaned data: 101766 rows and 43 columns

Dropped and replaced missing values to avoid biased results

Checked for inconsistencies and saved the cleaned data

DATA OVERVIEW

MESSY DATA VS CLEANED DATA

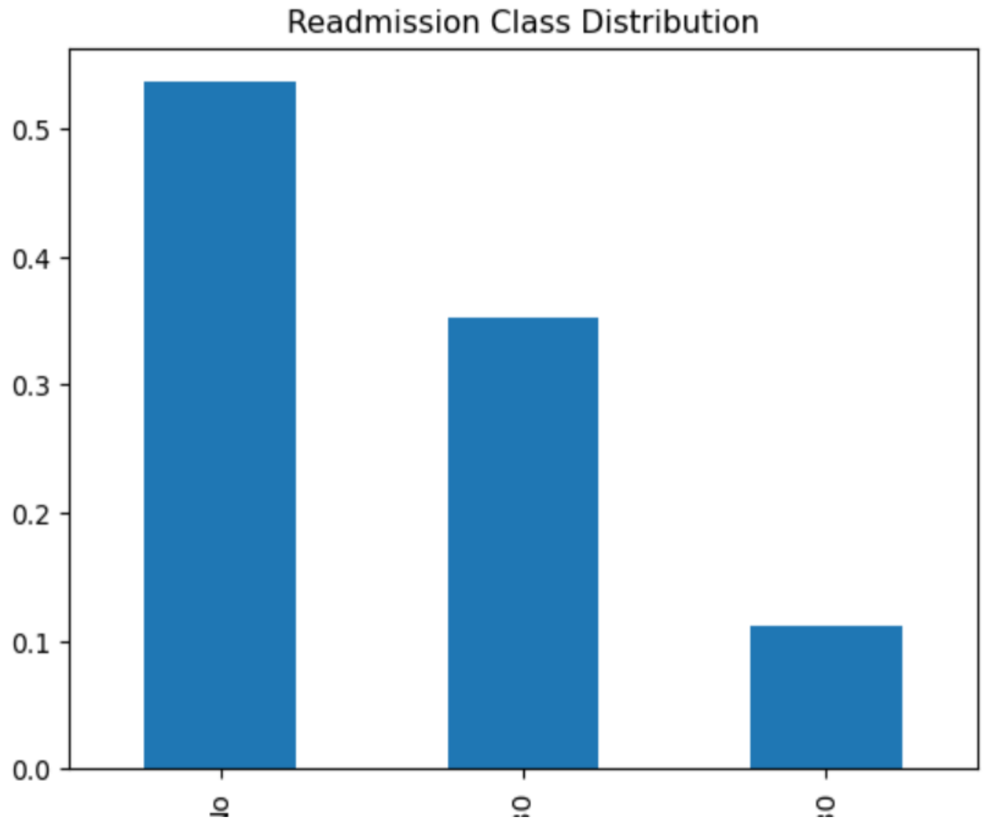
	encounter_id	patient_nbr	race	gender	age	weight	admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	...	citoglipton
0	2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1	...	No
1	149190	55629189	Caucasian	Female	[10-20)	?	1	1	7	3	...	No
2	64410	86047875	AfricanAmerican	Female	[20-30)	?	1	1	7	2	...	No
3	500364	82442376	Caucasian	Male	[30-40)	?	1	1	7	2	...	No
4	16680	42519267	Caucasian	Male	[40-50)	?	1	1	7	1	...	No

	race	gender	age	admission_type_id	discharge_disposition_id	time_in_hospital	medical_specialty	num_lab_procedures	num_procedures
0	Caucasian	Female	5	6	25	1	Pediatrics-Endocrinology	41	0
1	Caucasian	Female	15	1	1	3	0	59	0
2	AfricanAmerican	Female	25	1	1	2	0	11	5
3	Caucasian	Male	35	1	1	2	0	44	1
4	Caucasian	Male	45	1	1	1	0	51	0

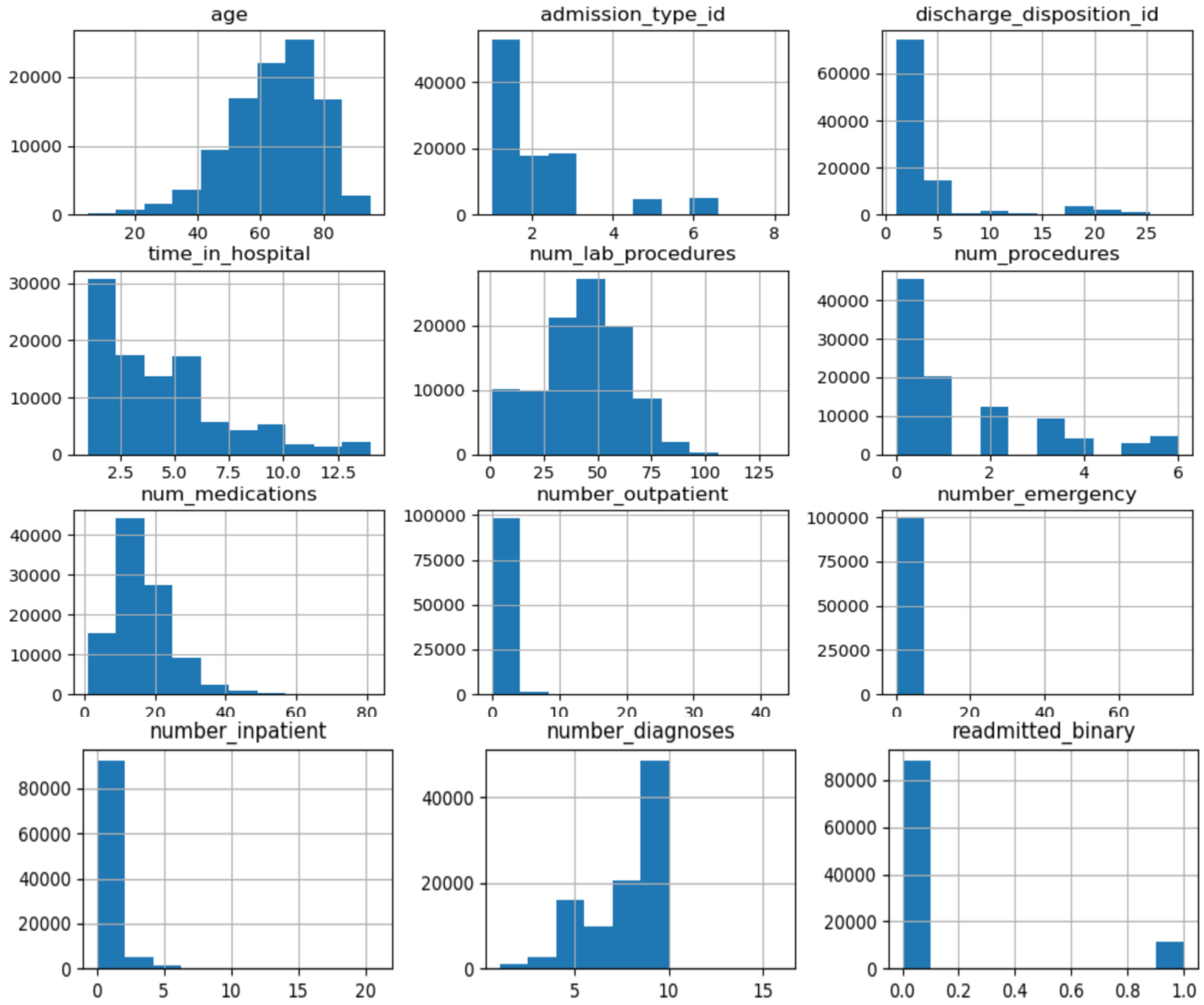
EXPLORATORY DATA ANALYSIS (EDA)

- It is all about visualization!

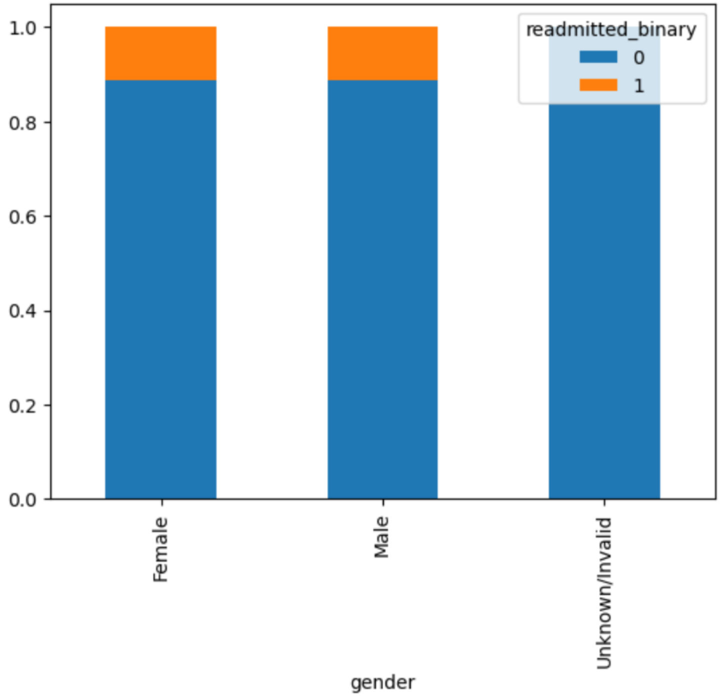
Target Variable:

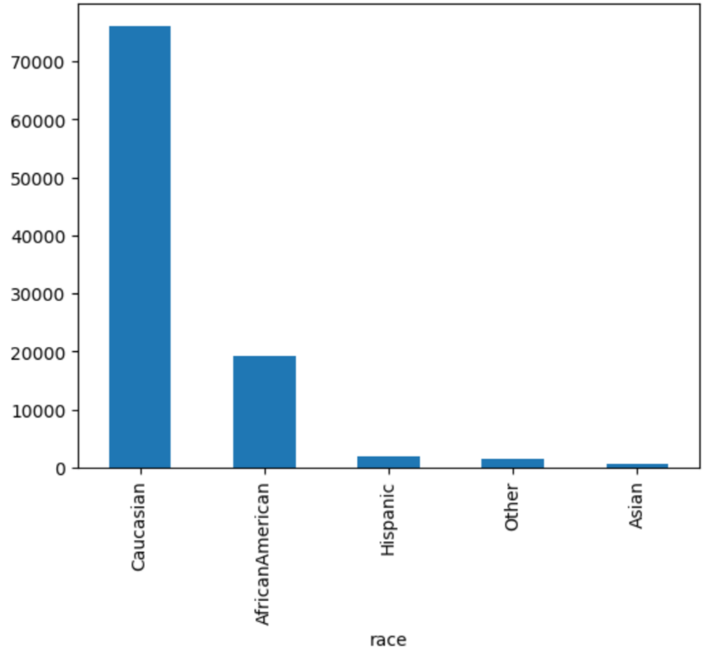


NUMERIC VARIABLES:



CATEGORICAL VARIABLES:





PREPROCESSING AND TRAINING

- One Hot Encoding
- Target Distribution
- Feature Scaling
- Distribution of Scaled Features
- Train – Test Split

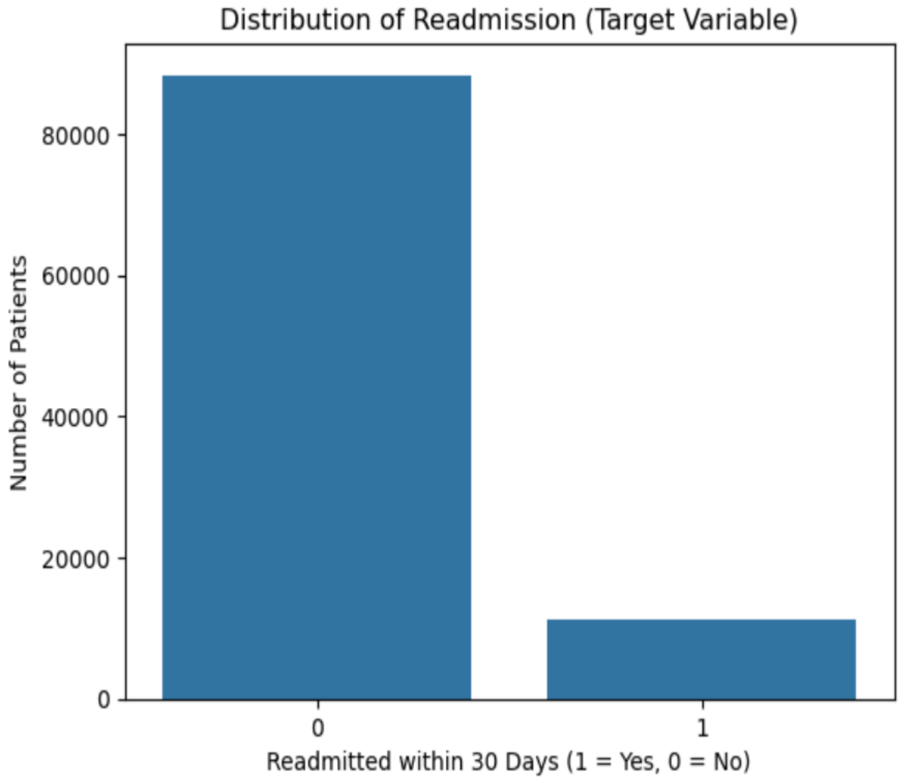
- One Hot Encoding:

A lot of machine learning algorithms cannot handle categorical variables directly. Hence, one-hot encoding was used to convert categorical columns like 'race' , 'gender', and 'medical specialty' into numerical indicators.

PREPROCESSING AND TRAINING

- Target Distribution:

To understand the balance between the readmitted and non-readmitted patients, I will be plotting the distribution of the 'readmitted binary' target variable.



PREPROCESSING AND TRAINING

- Feature Scaling:

To make sure that all the numeric features contribute equally to the model, standardization using 'StandardScaler' was applied. This is particularly helpful for algorithms like SVM, KNN, and Logistic Regression. But first the feature distribution was checked, then numeric columns were selected and then finally scaling was applied.

- Distribution of Scaled Features:

To visualize the distributions to confirm they are centered around 0 with unit variance.

- Train –Test Split:

The preprocessed data was split into training and testing subsets using an 80-20 ratio to train models on one part of the data and evaluate performance on unseen data.

MODELING

Logistic Regression Results:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	17724
1	1.00	1.00	1.00	2175
accuracy			1.00	19899
macro avg	1.00	1.00	1.00	19899
weighted avg	1.00	1.00	1.00	19899

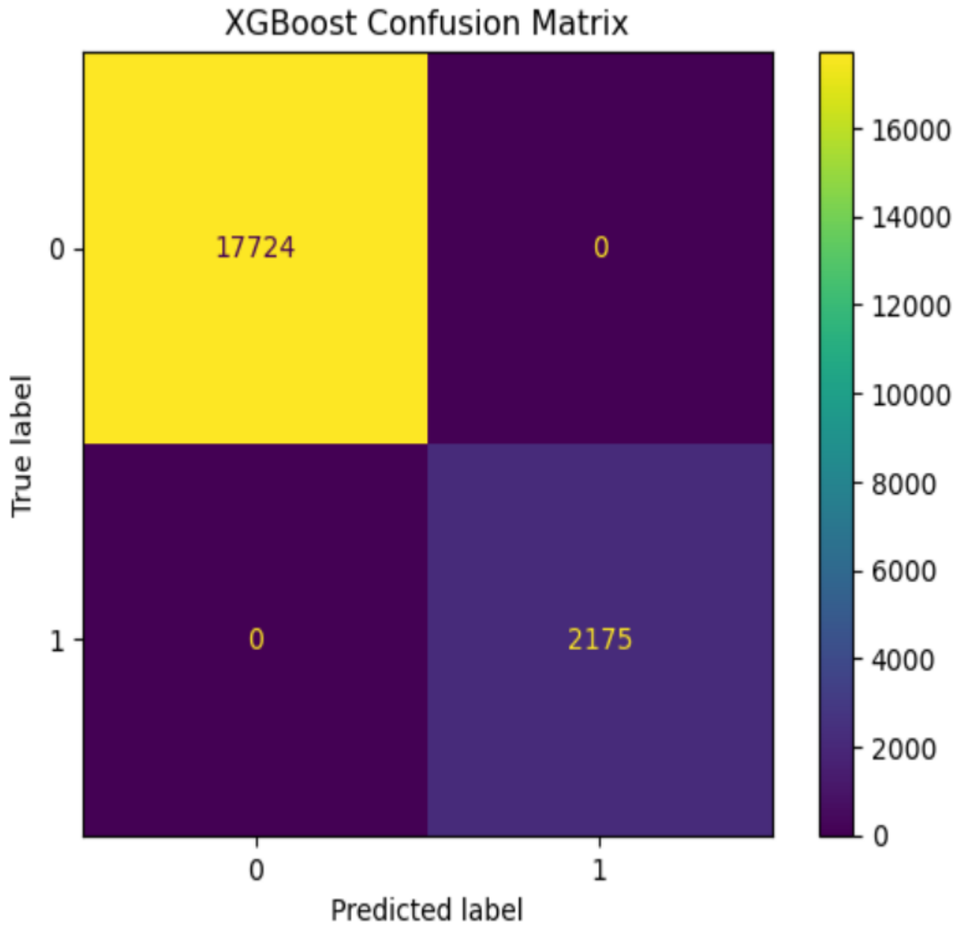
Random Forest Results:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	17724
1	1.00	0.99	0.99	2175
accuracy			1.00	19899
macro avg	1.00	0.99	1.00	19899
weighted avg	1.00	1.00	1.00	19899

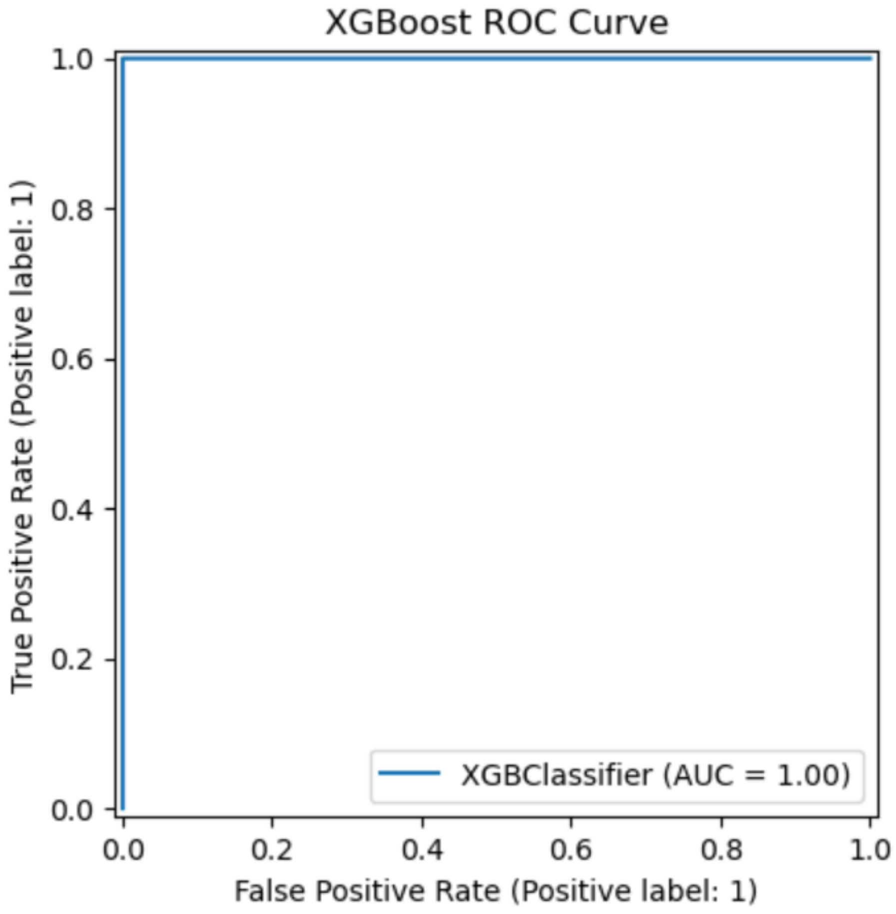
XGBoost Results:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	17724
1	1.00	1.00	1.00	2175
accuracy			1.00	19899
macro avg	1.00	1.00	1.00	19899
weighted avg	1.00	1.00	1.00	19899

MODELING

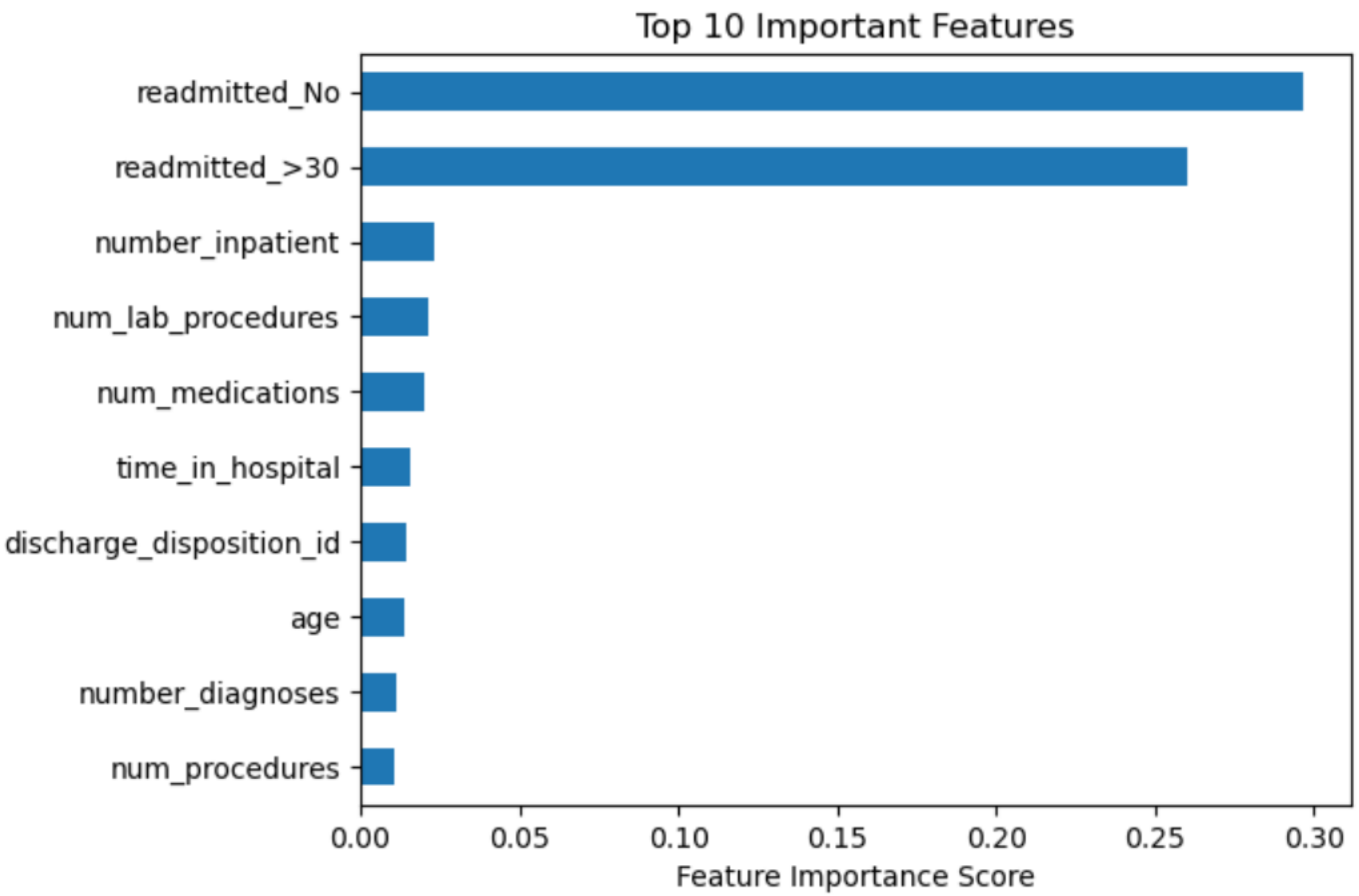


XGBoost Confusion Matrix
17724
0
0
2175
True label
Predicted label
0
1
16000
14000
12000
10000
8000
6000
4000
2000



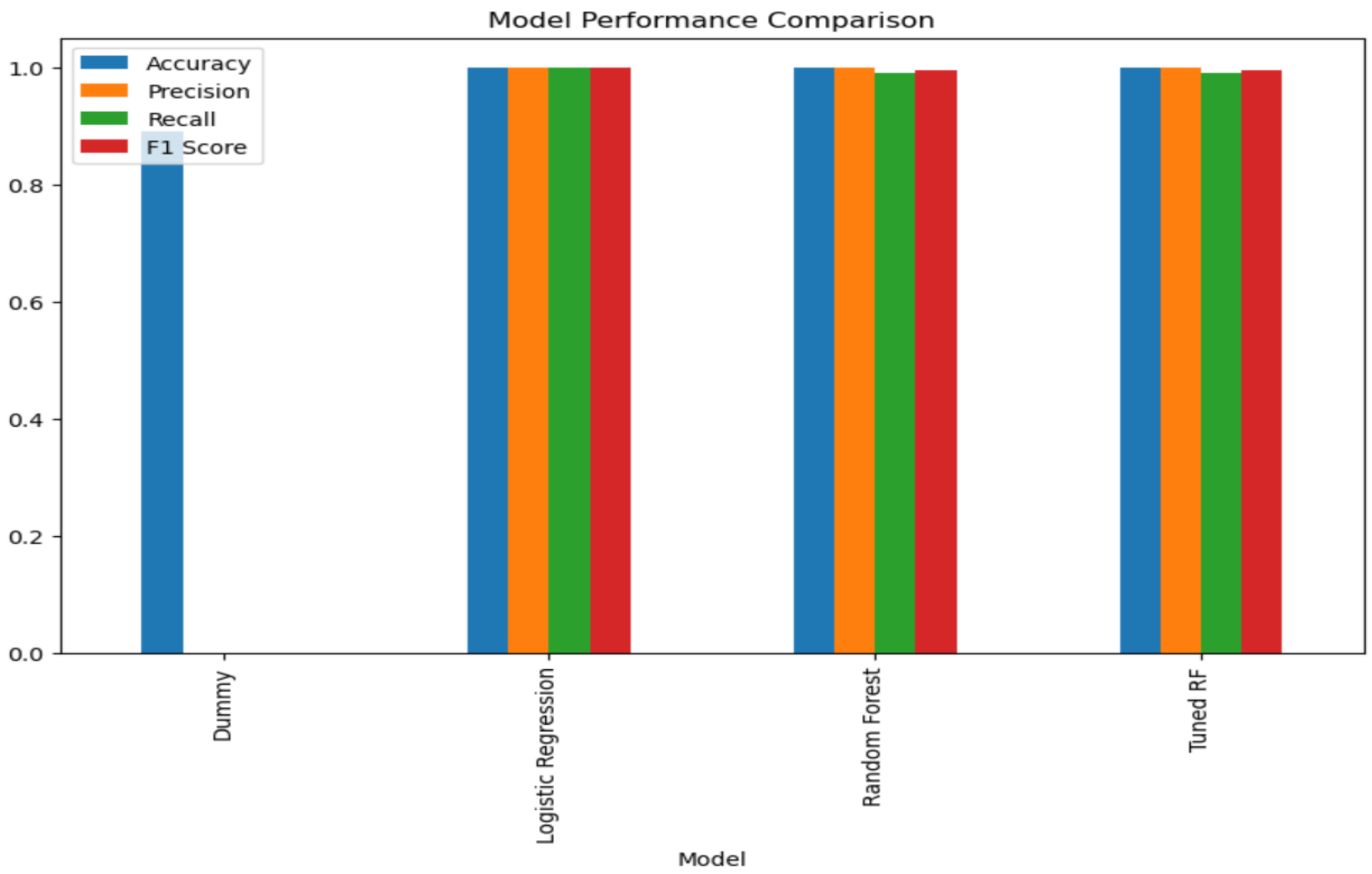
XGBoost ROC Curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
XGBClassifier (AUC = 1.00)
0.0
0.2
0.4
0.6
0.8
1.0

MODELING



Top 10 Important Features
readmitted_No
readmitted_>30
number_inpatient
num_lab_procedures
num_medications
time_in_hospital
discharge_disposition_id
age
number_diagnoses
num_procedures
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Feature Importance Score

MODELING



Model Performance Comparison
Accuracy
Precision
Recall
F1 Score
1.0
0.8
0.6
0.4
0.2
0.0
Dummy
Logistic Regression
Random Forest
Tuned RF
Model

CONCLUSION

- Despite thorough preprocessing and model tuning, the predictive performance across these models had poor outcome.
- AUC Scores: Ranged from approximately 0.47 to 0.53, indicating limited discriminative ability.
- Accuracy: While some models achieved higher accuracy, this was misleading due to class imbalance, with models predominantly predicting the majority class.
- Confusion Matrices: Revealed that models struggled to correctly identify readmitted patients, often misclassifying them as non-readmitted. Several factors contributed to the models' limited performance.
- Insufficient Feature Set: The dataset lacked comprehensive clinical details such as specific procedures, medication types, and vital signs, which are crucial for accurate predictions.
- Class Imbalance: A disproportionate number of non-readmitted patients led to models biased towards predicting the majority class.
- Model Limitations: Standard machine learning models may not capture the complex patterns associated with hospital readmissions without richer data.

THANK YOU

- Tasmia Kayenat
- aashatasmia1999@gmail.com