Predicting Diabetic Patients Hospital Readmission Rates

Aaron Onserio

Daniel Ekale

Emily Njue

Robert Mbau

Yussuf Hersi

Jimcollins Wamae

Edna Wanjiku

Outline

- Overview
- Problem Statement
- Objectives
- Data Understanding
- Feature Selection
- Modeling
- Model Evaluation
- Deployment
- Conclusion
- Recommendations

Overview

Introduction

The project aims to predict diabetic patients readmission rates using features available in the "Diabetes 130-US hospitals for years 1999-2008" dataset.

Problem Statement

- Discover the power of predictive modeling in healthcare. Our cutting-edge project aims to develop a remarkable solution that accurately estimates the likelihood of readmission for diabetic patients.

- By unraveling the factors associated with readmission, we empower healthcare providers to intervene proactively, delivering timely care and slashing readmission rates.

Objectives

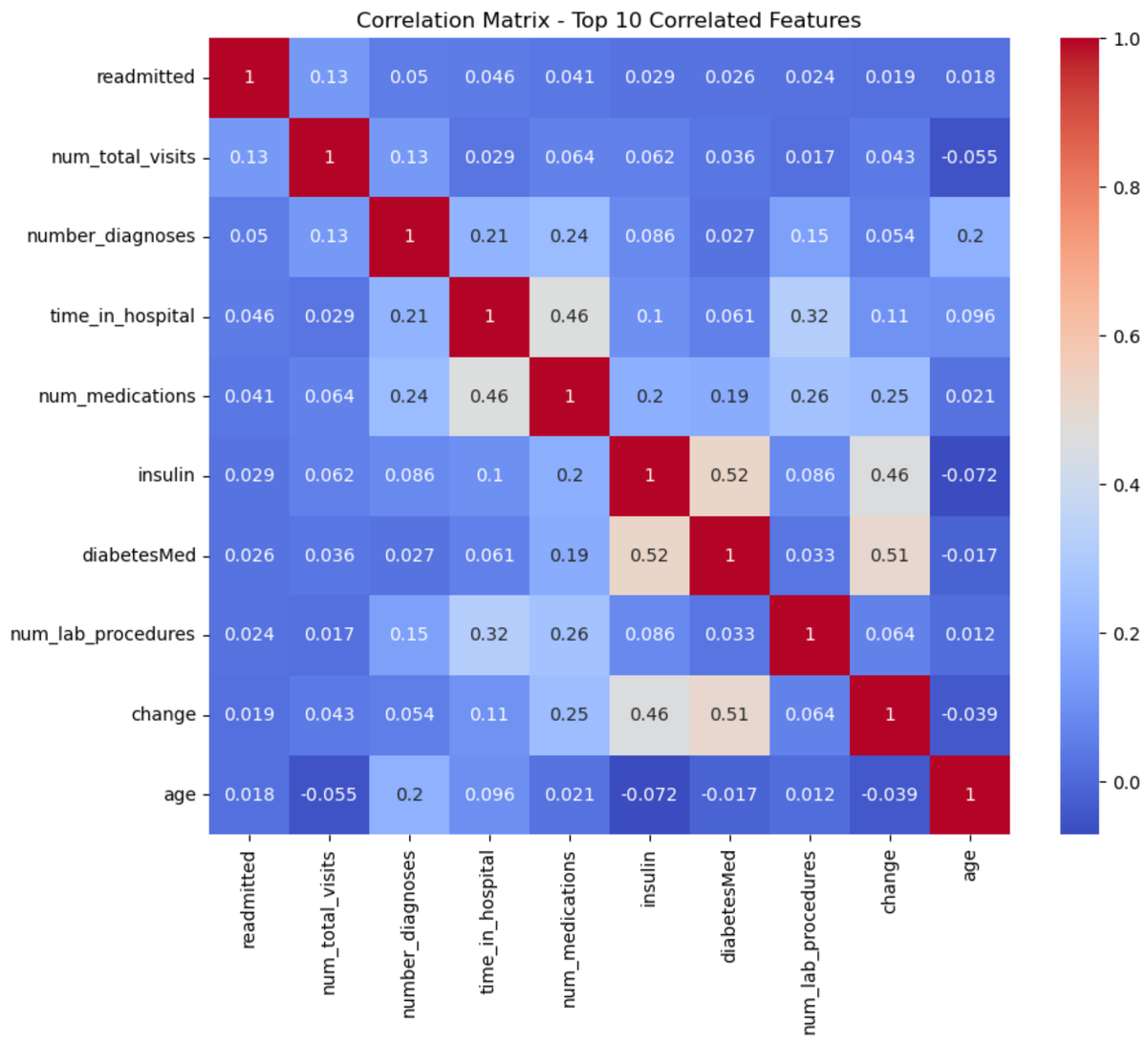
- Conduct comprehensive exploratory analysis to uncover insights specific to diabetic patients.
- Meticulously preprocess the dataset, handling missing values and selecting influential features.
- Develop a robust predictive model using advanced algorithms for accurate readmission estimation.
- Uncover factors influencing readmission rates among diabetics, providing actionable insights for healthcare providers.

Data Understanding

- Data was acquired from UC Irvine(https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008)
- It contains information from 130 US hospitals over a time period from 1998 to 2008.
- In additon, it has over 40 features and over 100,000 instances.
- The dataset also contains quite a number of missing values which is denoted by '?'.

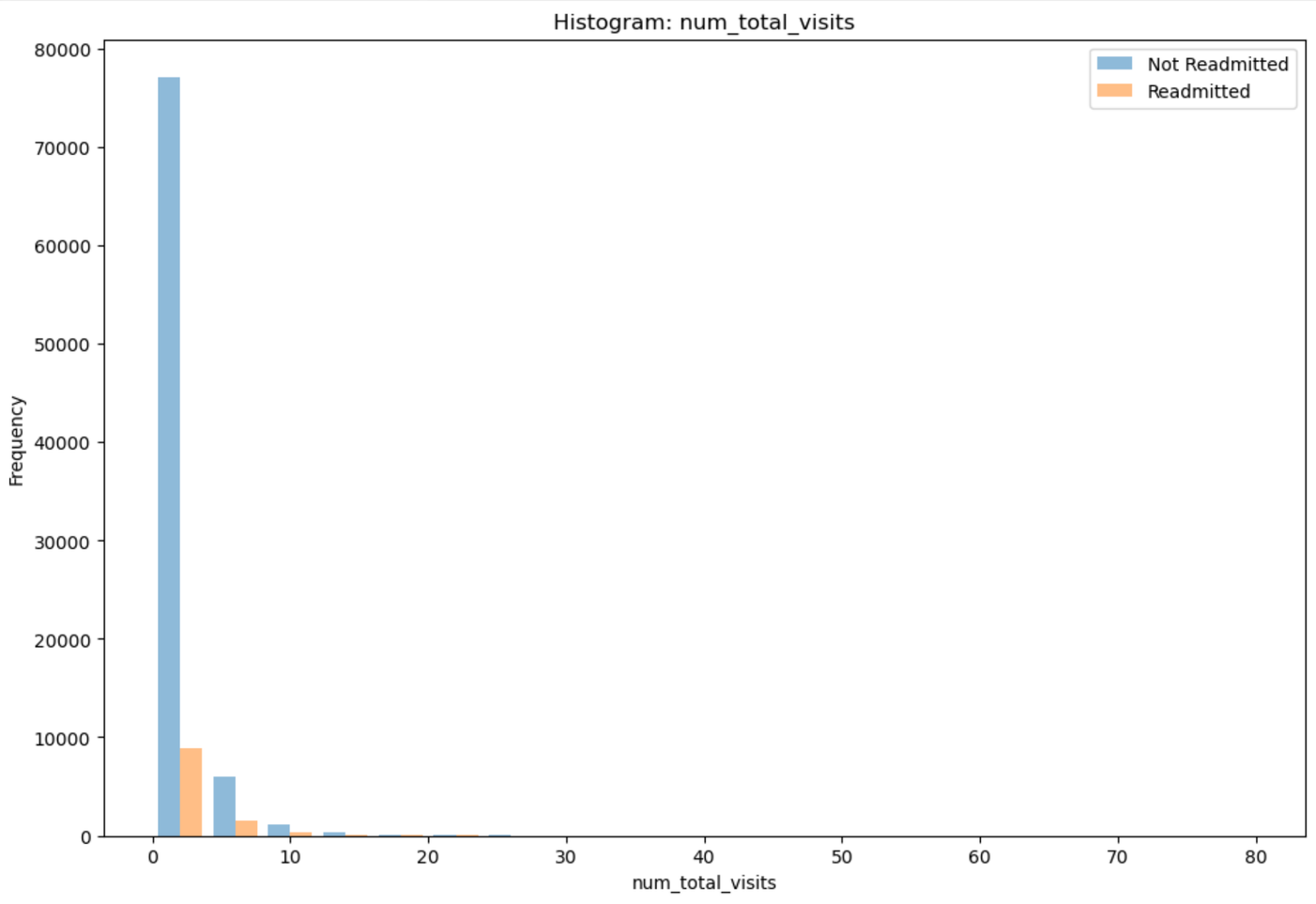
Exploratory Data Analysis (EDA)

- We began EDA by examining the correlation matrix to identify the top ten features that exhibit a strong correlation with the target column, 'readmitted.'

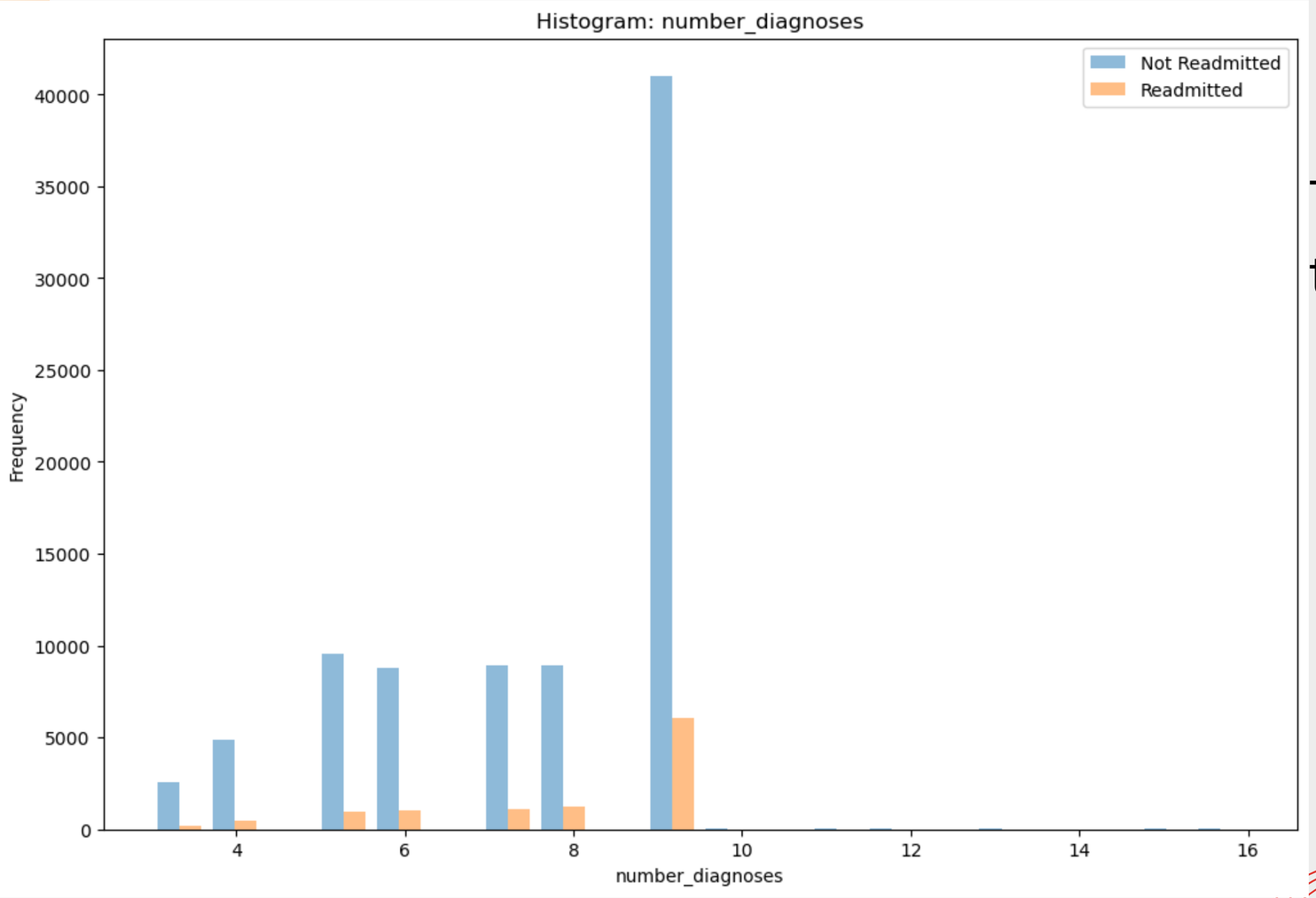


Exploratory Data Analysis (EDA)

- In our correlation analysis, we identified the top five features most correlated with 'readmitted.'
- These features include : **the number of total visits, number of diagnoses, time spent in the hospital, and number of medications.**

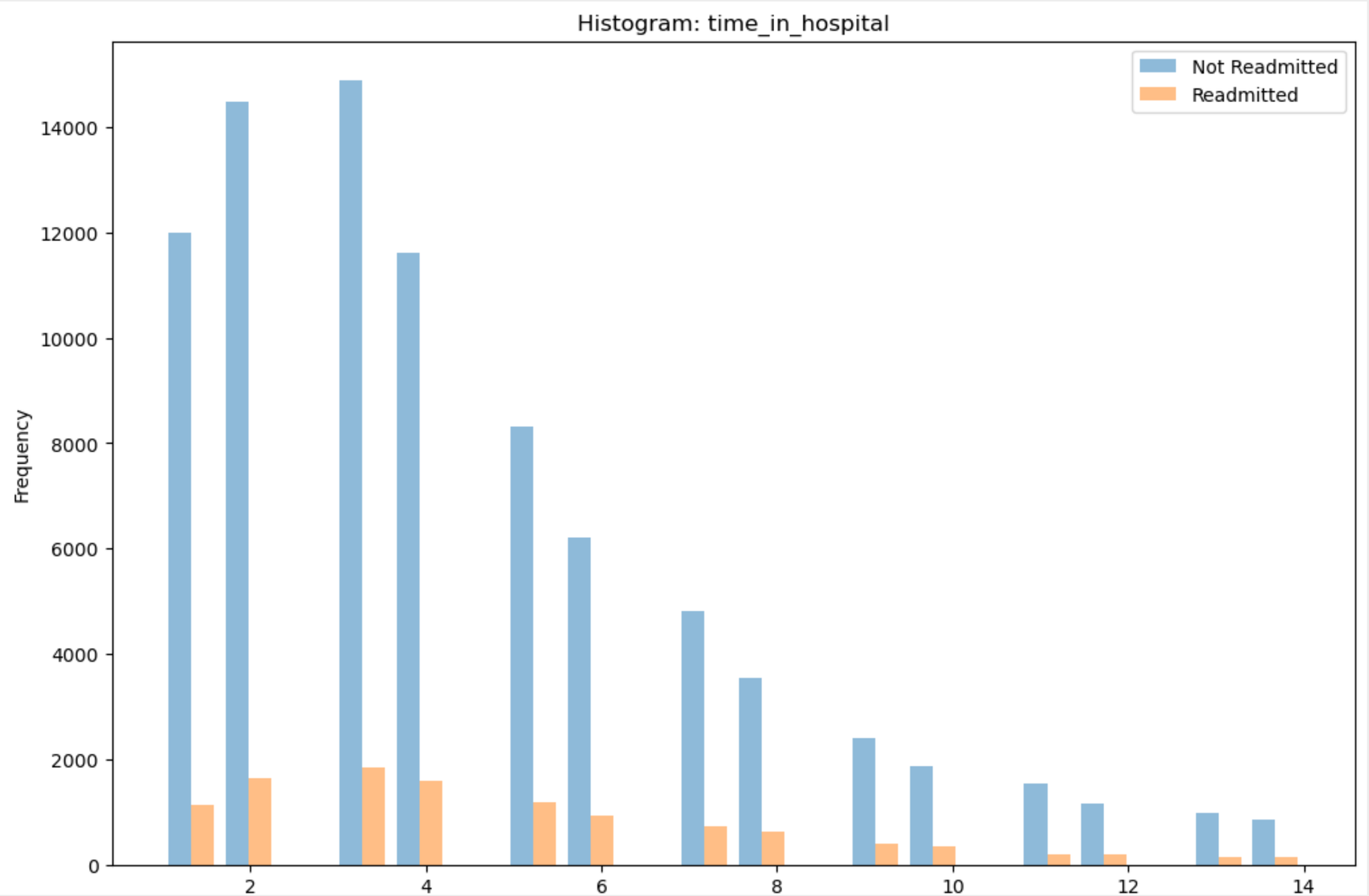


Exploratory Data Analysis (EDA)



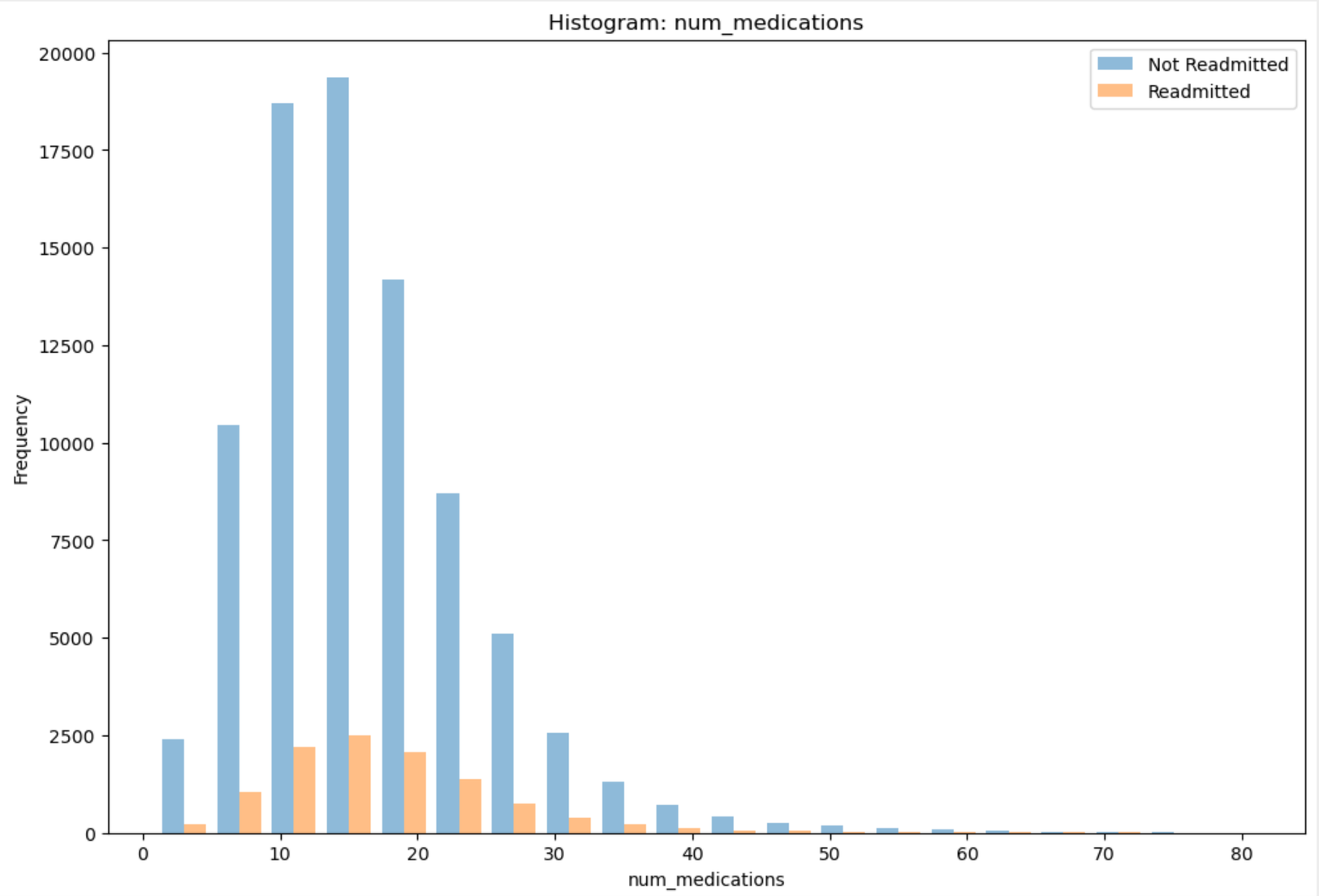
These were the number of diagnoses that were done to the patients.

Exploratory Data Analysis (EDA)



This graph represent the number of times in which a patient spent in the hospital

Exploratory Data Analysis (EDA)



This graph represent the total number of medications a patient was given

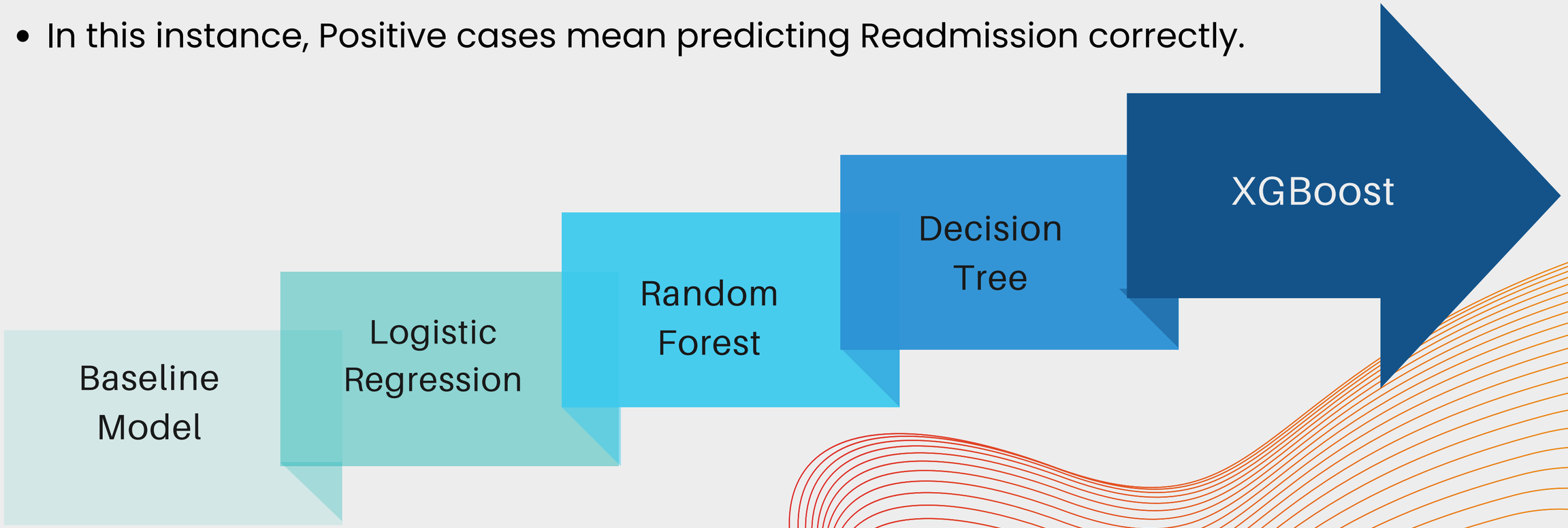
Feature Selection

- During the feature engineering and selection, we found 23 features to be important for modelling.
- We used Chi Square and Anova, and managed to extract the significant features.
- Based on our analysis, we found the following significant features (columns) with their corresponding p-values, indicating their level of significance in relation to 'readmitted'.
- These features provide valuable insights into the factors influencing readmission.

	columns	p_value	significance
40	num_total_visits	0.000000e+00	Significant
39	readmitted	0.000000e+00	Significant
5	discharge_disposition_id	8.396779e-169	Significant
14	number_diagnoses	7.163232e-55	Significant
7	time_in_hospital	2.968908e-46	Significant
32	insulin	4.101414e-42	Significant
10	num_medications	5.062578e-36	Significant
11	diag_1	1.158059e-29	Significant
13	diag_3	7.972029e-24	Significant
12	diag_2	2.487150e-18	Significant
38	diabetesMed	1.282343e-15	Significant
17	metformin	5.128988e-15	Significant
8	num_lab_procedures	2.436319e-13	Significant
6	admission_source_id	4.992654e-09	Significant
37	change	9.474549e-09	Significant
4	age	1.974873e-08	Significant
16	A1Cresult	9.411177e-08	Significant
15	max_glu_serum	8.854120e-05	Significant
9	num_procedures	5.982499e-04	Significant
23	glipizide	6.188992e-03	Significant
18	repaglinide	9.280168e-03	Significant
21	glimepiride	3.349191e-02	Significant
26	pioglitazone	4.230423e-02	Significant
27	rosiglitazone	5.755043e-02	Insignificant
29	miglitol	7.680071e-02	Insignificant

Modeling

- We created multiple models to elicite the best success metric.
- Recall was the metric used for the project because it measures how well the model will predict positive cases.
- In this instance, Positive cases mean predicting Readmission correctly.



Baseline Model

Accuracy: 0.882

	precision	recall	f1-score	support
0	0.88	1.00	0.94	21102
1	0.00	0.00	0.00	2811
accuracy			0.88	23913
macro avg	0.44	0.50	0.47	23913
weighted avg	0.78	0.88	0.83	23913

- This model was used to evaluate the perfomance on the training and testing data.
- It had a Recall of 100% on the No Readmission and 0% on the Readmission.
- The poor results on predicting the Readmission was due to class imbalance where the No Readmission was about 88% of the training data.

Logistic Regression:
Balanced data

```
Training score: 0.6042060193930235
Test Score: 0.609400579272132


Classification Report

_______________
              precision    recall  f1-score

           0       0.60      0.65      0.63
           1       0.61      0.57      0.59

    accuracy                           0.61
   macro avg       0.61      0.61      0.61
weighted avg       0.61      0.61      0.61
```

- This model improved Recall upto 61% which is a very big improvement from the previous models.
- However, this is still not a good result due to the magnitude of the outcome we are dealing with.

Random Forest

	precision	recall	f1-score
0	0.89	1.00	0.94
1	0.99	0.88	0.93
accuracy			0.94
macro avg	0.94	0.94	0.94
weighted avg	0.94	0.94	0.94

- The success metric improved to 94% for this model.
- This exceeded the recall threshold that was set.

Decision Tree

```
Training score: 0.9868089661251731
Test Score: 0.8603450447046972


Classification Report

_________________
                precision    recall  f1-score

           0       0.86      0.86      0.86
           1       0.86      0.86      0.86

    accuracy                           0.86
   macro avg       0.86      0.86      0.86
weighted avg       0.86      0.86      0.86
```

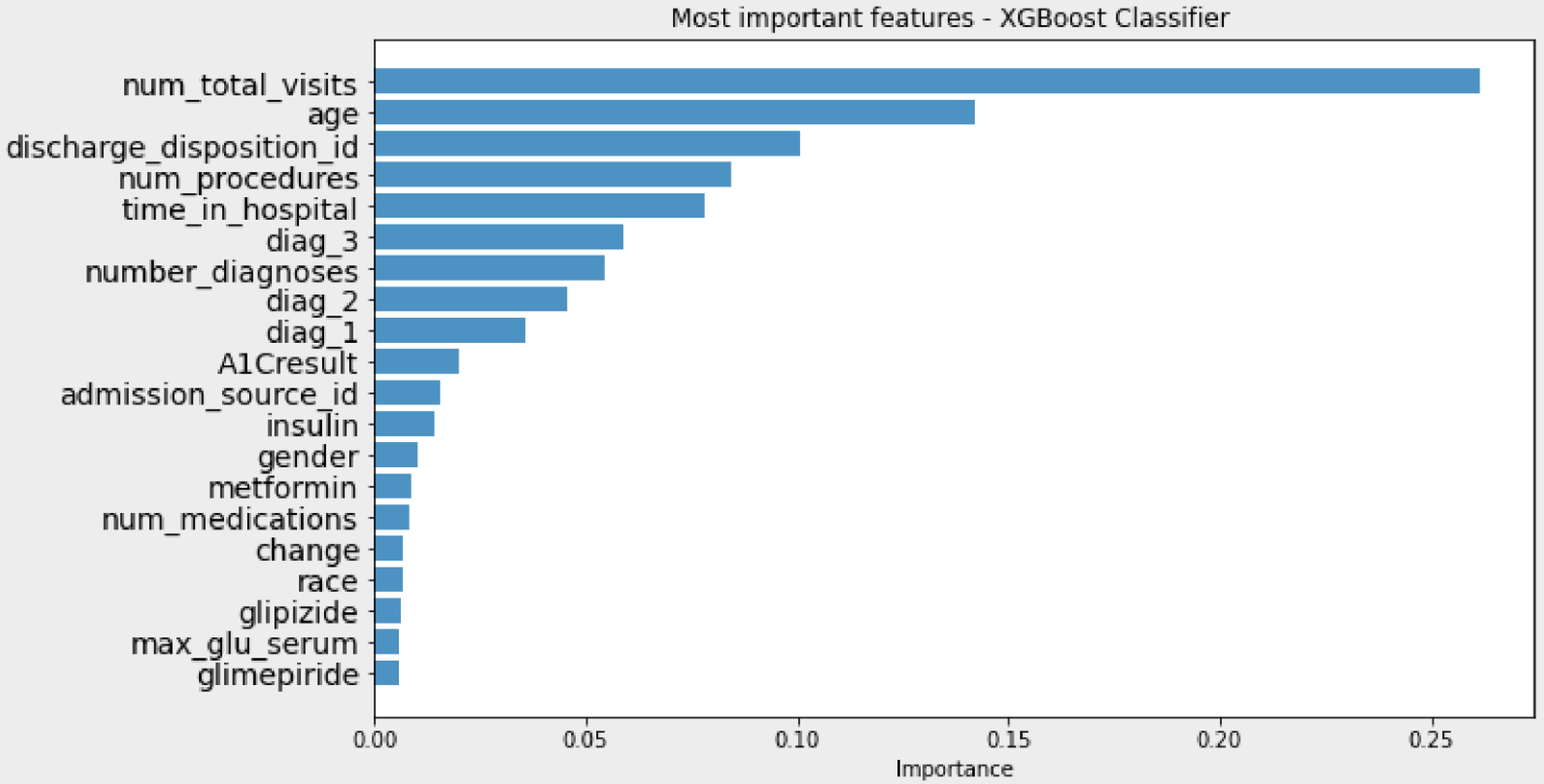
- Although it achieved the set target, there was a better model already.
- The model was able to predict 86% accurately.
- The model was overfitting slightly on the training data.

XGBoost

- The model achieved a Recall of 93%.
- Although there was a model with a higher recall, this was the best model because it was able to generalise to all the data.

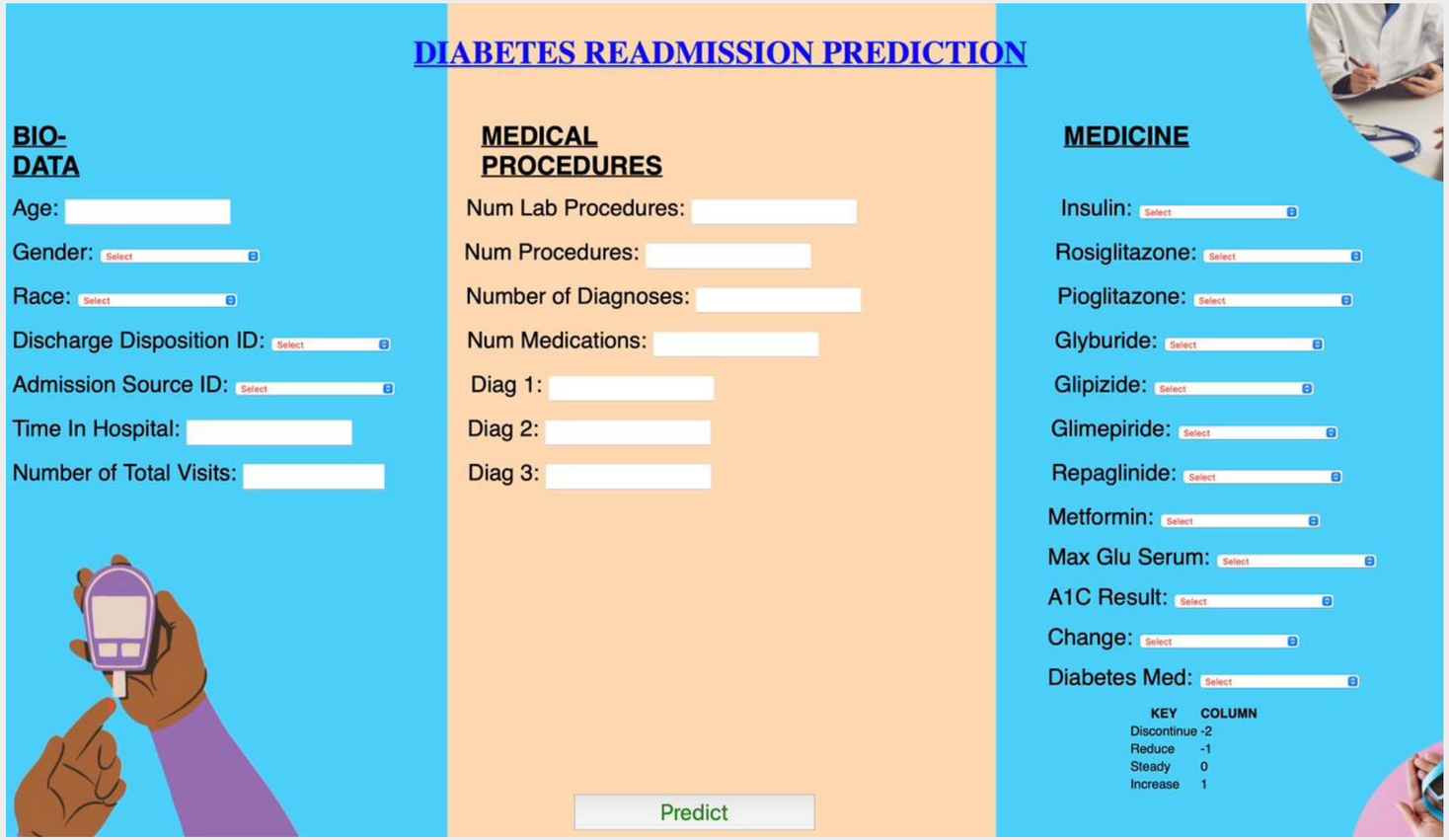
```
Training score: 0.9384523359778365
Test Score: 0.9338559375393527


Classification Report

_________________
                precision    recall  f1-score

            0       0.89      1.00      0.94
            1       1.00      0.87      0.93

     accuracy                           0.93
    macro avg       0.94      0.93      0.93
weighted avg       0.94      0.93      0.93
```

Model Evaluation



- The model was able to generalise to new data as seen from the test scores.
- The high recall means that the model can accurately predict 93% of the positive cases correctly.

Deployment



- Our project had many significant features and they all are included in the interface.
- The features have been sub-divided into relevant classes for ease of access.

Conclusion

- The baseline model performed poorly, with an accuracy of 0.88 and low recall for the positive class.
- Logistic Regression with imbalanced data had very low recall and precision.
- Logistic Regression with balanced data improved the recall to 0.05 but is still not satisfactory.
- Random Forest achieved the highest success metric with a recall of 0.94 and an accuracy of 0.94.
- Decision Tree had high accuracy but slightly lower recall compared to other models.
- XGBoost performed well with a recall of 0.93 and high accuracy.

Recommendation

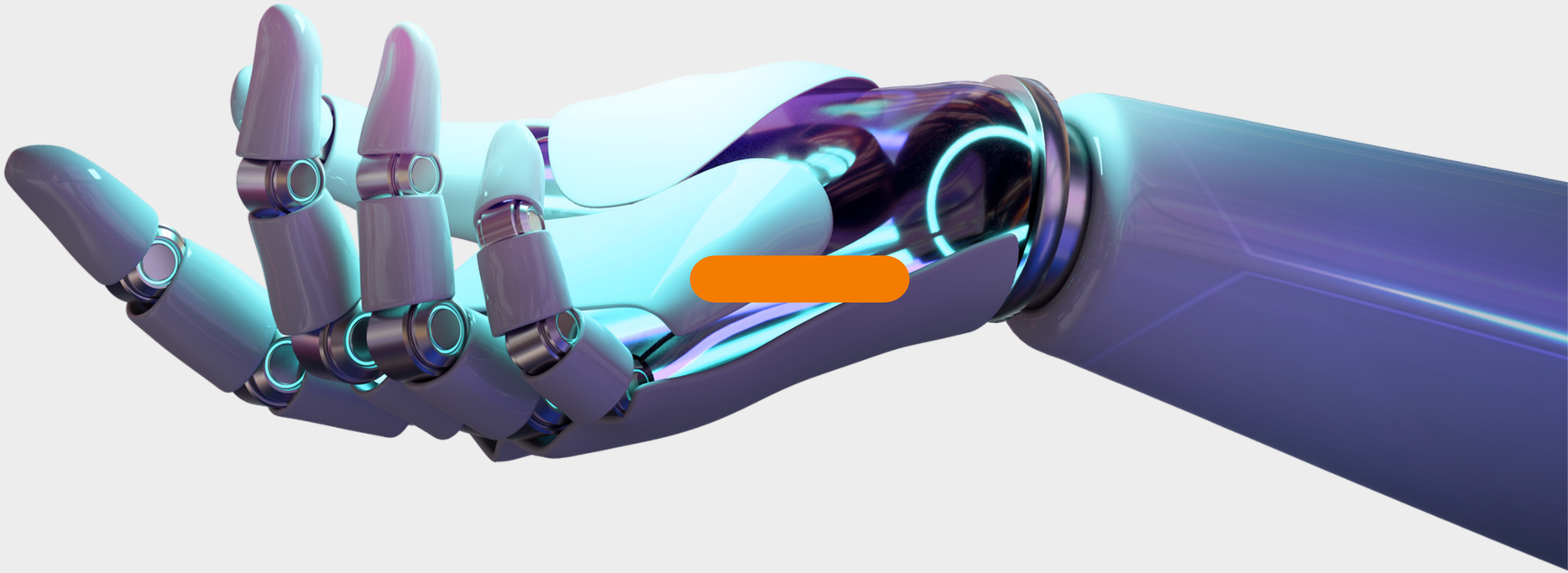
- XGBoost is recommended due to their higher recall and accuracy compared to other models.
- Consider further optimizing the models to improve the recall for the positive class.
- Explore ensemble methods to combine the strengths of multiple models and improve overall performance.

Next Steps

- Gather more data to increase the positive class instances.
- Get more recent data.

Q and A

ASK AWAY!



Thank You