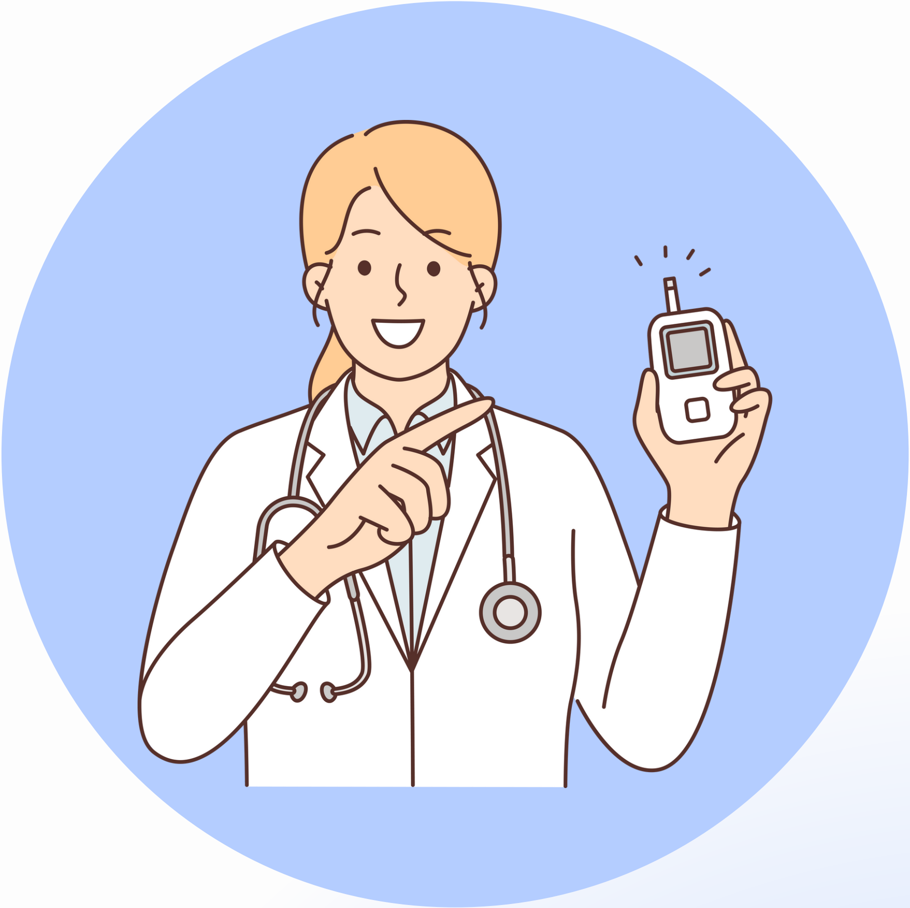
Addressing Early Hospital Readmissions

A Data-Driven Approach to Managing Diabetes

Presented by Louis Cockenpot, Foucauld Estignard and Hector Mell Mariolle

Overview

Introduction



Diabetes is a chronic health condition characterized by **elevated levels of blood glucose** resulting from the body's inability to produce or effectively utilize insulin.

Globally, diabetes is a major health concern affecting millions. According to the International Diabetes Federation, approximately **463 million** adults (20-79 years) were living with diabetes in 2019, and this number is expected to rise to **700 million** by 2045, with a significant proportion of them undiagnosed.

The hypothesis is as follows: some patients **do not receive the necessary care**, which would anticipate their return to hospital

Reducing early hospital readmissions is a policy priority aimed at improving healthcare quality.

Introduction

To answer this major concern, we studied the data from 130 US hospitals over 10 years (1999-2008).

More specifically, we focused on patients diagnosed with diabetes, who underwent laboratory, medications and stayed up to 14 days.

Our goal is to understand what are the main factors that lead to early readmission, to help hospitals organizing and directing their efforts towards more effectively managing diabetes.

Will the patient be back in the hospital within 30 days?

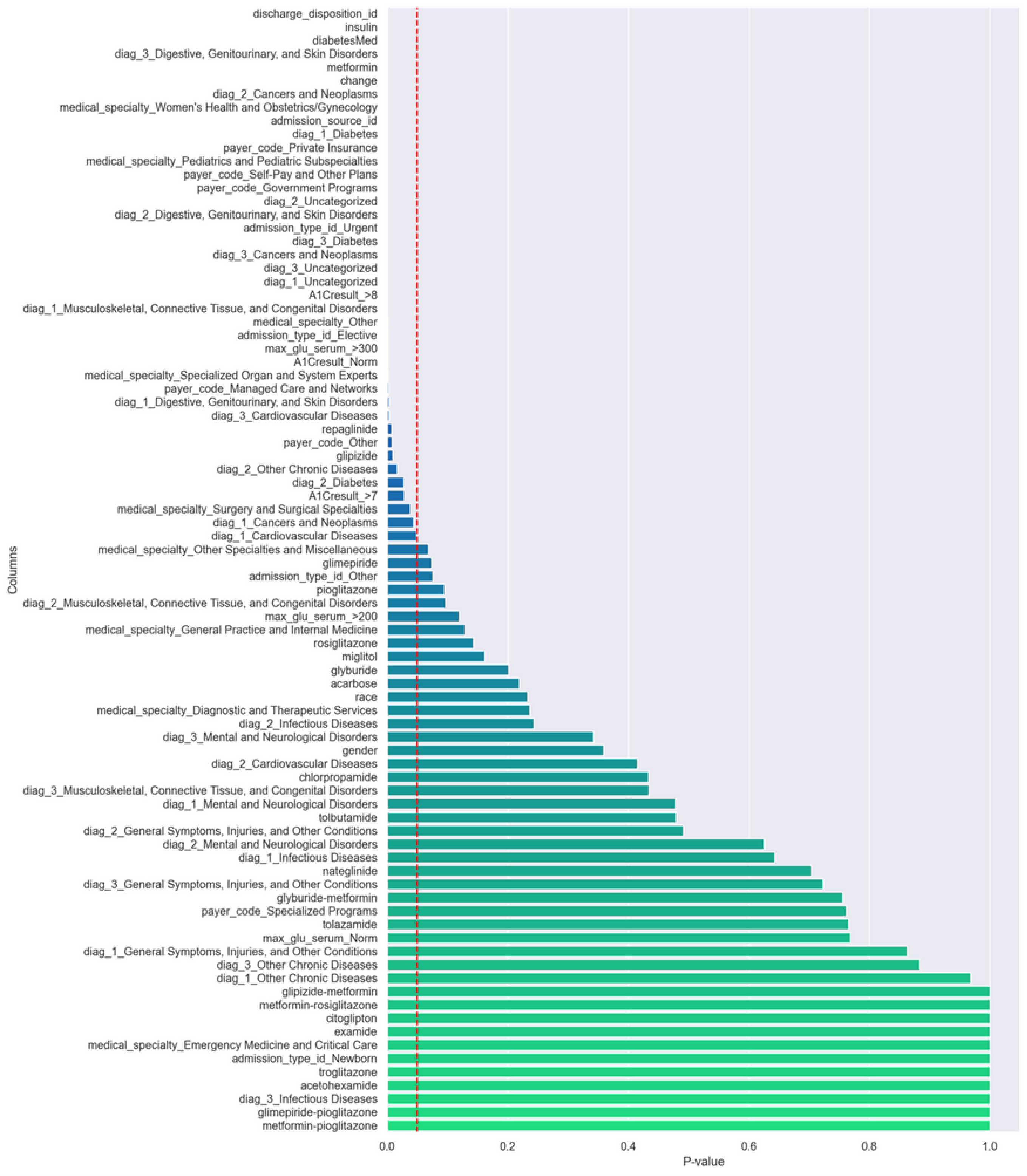
Dataset Exploration

Gives information about the **medical condition of the patient** such as: his weight, his age, his gender... And also his medical record: number of visits during the past year

We also have information about his **hospital admission** and **hospital stay details**. We can check the admission type, the medical specialty, the time in hospital, all the diagnostics and even more.

In the dataset we also have access to the **treatment** and **diabetes medications** that the patient received during his stay at the hospital.

Taking the best from our data



- In order to maximize the efficiency of our solution, we focused on selecting the most relevant features, determined by using statistical tests, such as the **Chi squared.**

- To mesure the **statistical association** between each variable and the target variable, we computed their **p-value**, which indicates the probability of wrongly concluding that an association exists.

- We only kept the columns that have a p-value **under 5%** (red line) to predict the readmission of a patient with our models. This ensures us that the variables are **statistically related.**

A solution built along with medical specialists*

- To ensure the **efficiency** of our solution, we **grouped the data** into broader categories that we carefully choosed, in order to maintain the information carried by our data.

- This is illustrated by the way we handled the diagnosis codes. These codes can take more than 800 distinct values, corresponding to the International Classification of Diseases (ICD-9).

- This enables us to apply a **One Hot Encoding** without increasing too much the dimensionality of our data

- One Hot Encoding is particularly relevant in our case, as there is **no ordinal link** between the values.

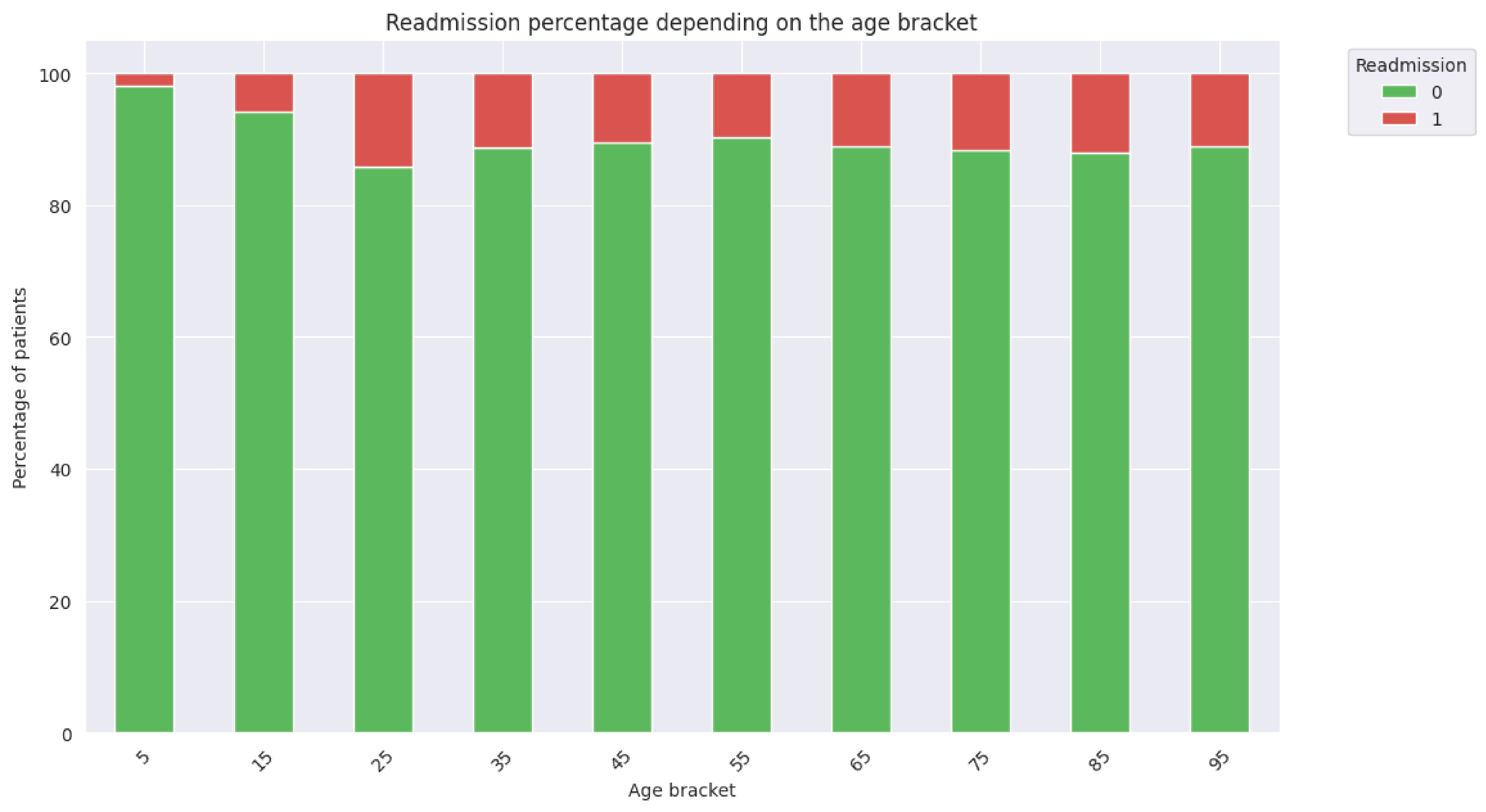
- We used a similar approach to encode the data regarding the medical specility, the admission type, the discharged category...

ICD-9 Code Range	Category
Starts with '250'	Diabetes
1-139	Infectious Diseases
140-239	Cancers and Neoplasms
390-459	Cardiovascular Diseases
240-279, 280-289, 460-519	Other Chronic Diseases
290-319, 320-389	Mental and Neurological Disorders
520-579, 580-629, 680-709	Digestive, Genitourinary, and Skin Disorders
710-739, 740-759, 760-779	Musculoskeletal, Connective Tissue, and Congenital Disorders
780-799, 800-999, Starts with 'V' or 'E'	General Symptoms, Injuries, and Other Conditions
Others	Uncategorized

First Analysis

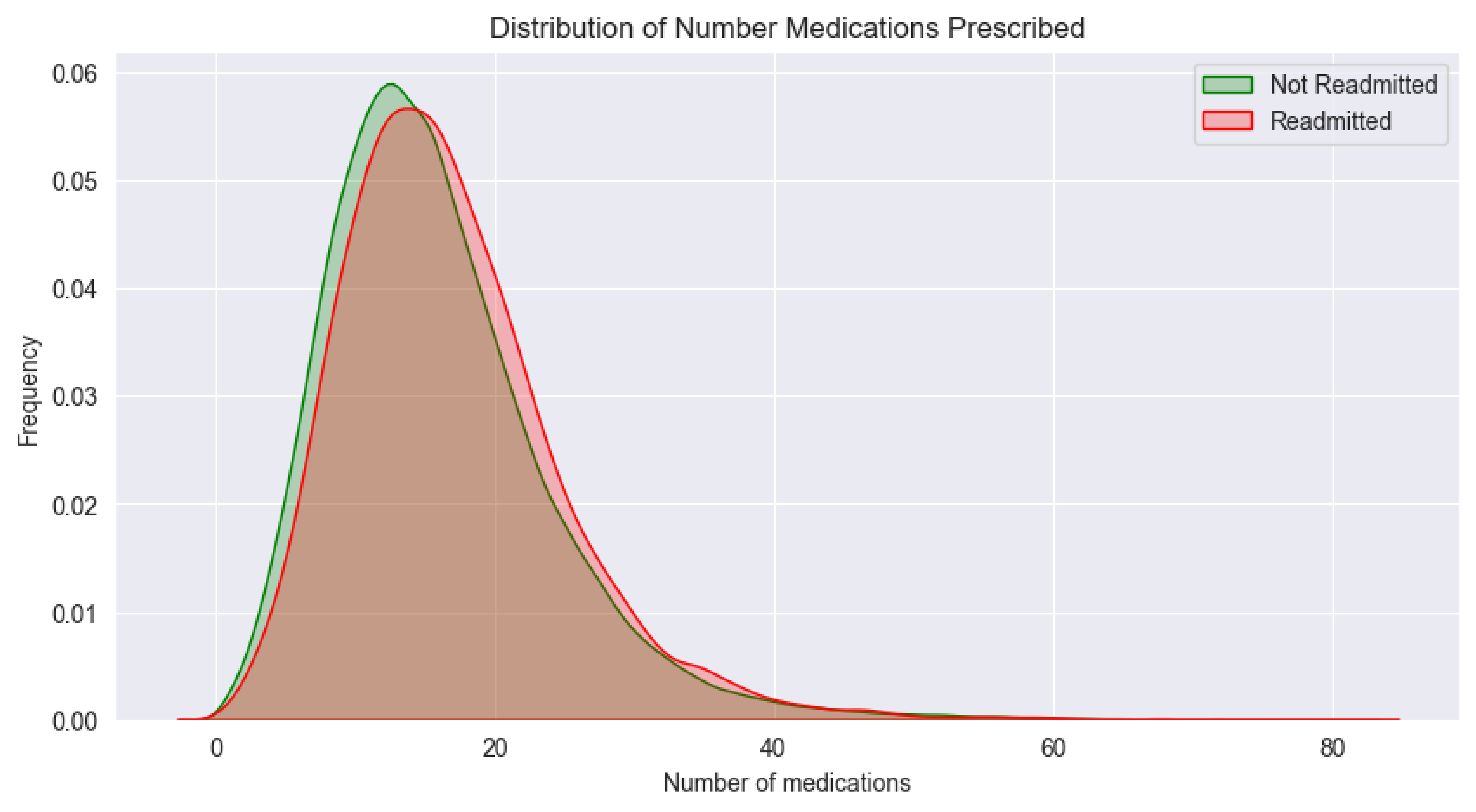
1. **Readmission rate by age**

We noticed that young patients tended to be less readmitted within the next month, maybe because of a better follow-up for children. However, we should nuance this as the dataset is not balanced at all between the different age brackets and children are much less represented.

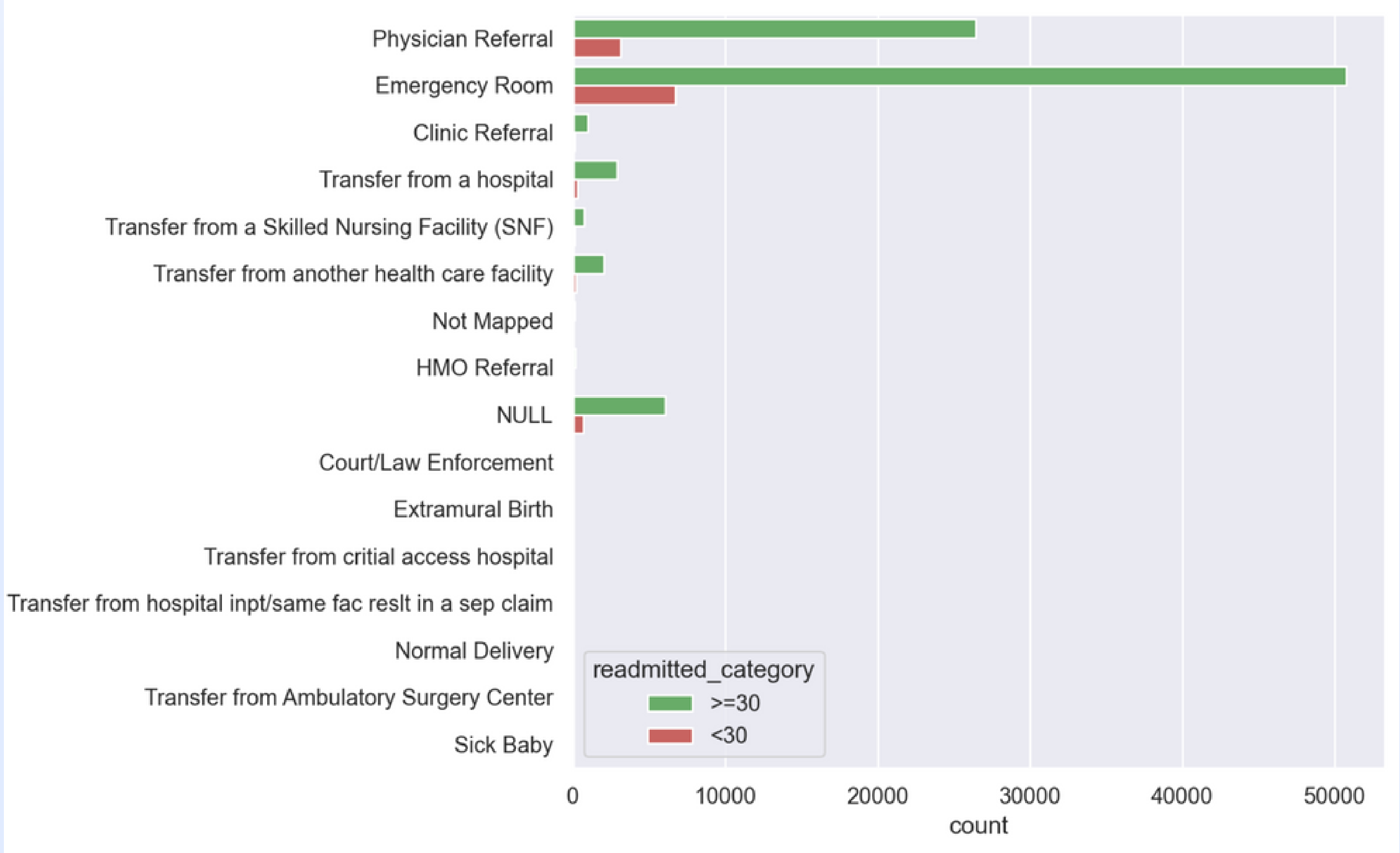


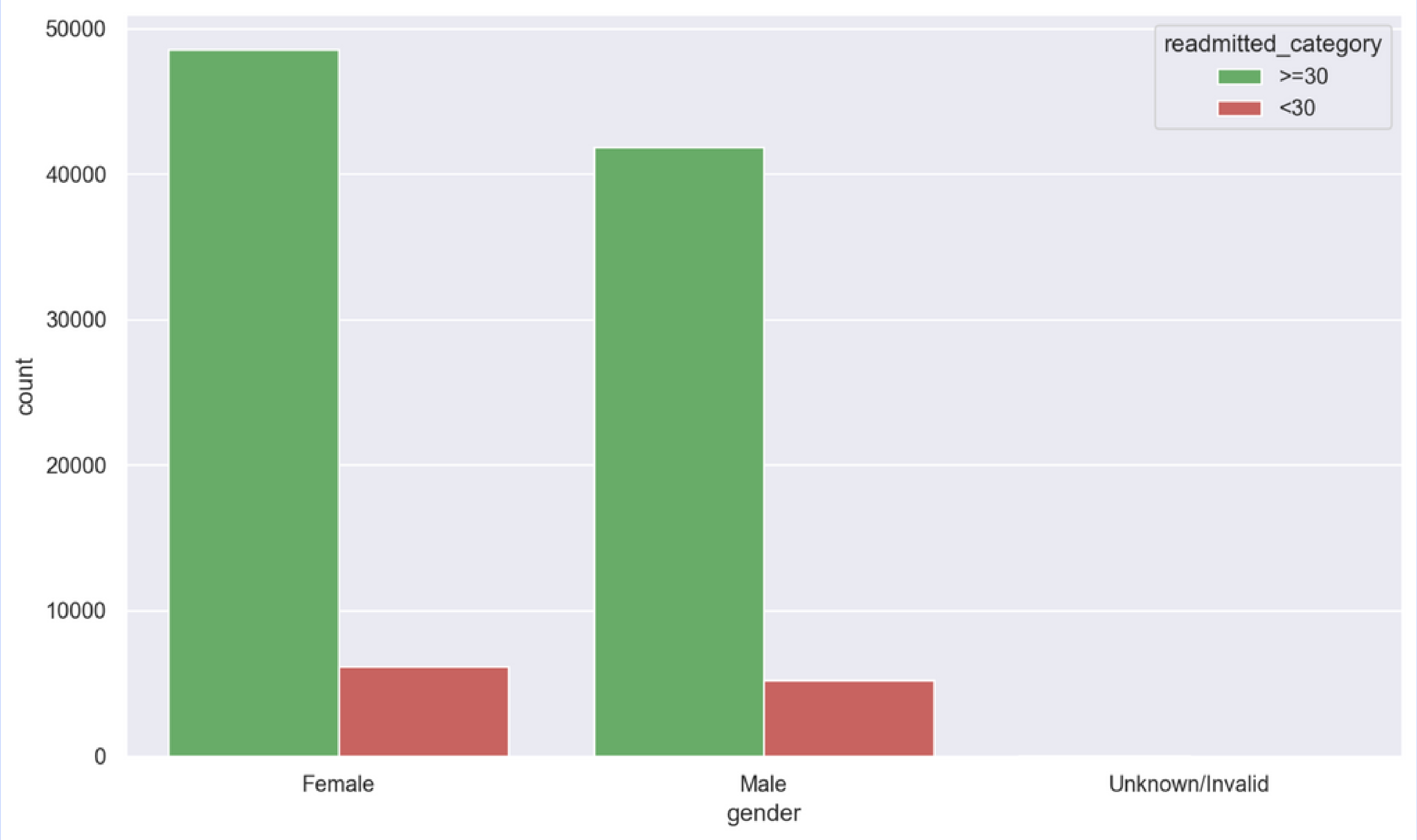
2. Readmission rate according to the number of medications

We observe that the number of medications taken does not seem to decrease the likelihood of being readmitted within the next month. While this may initially seem counterintuitive, it's understandable, considering that patients with more severe conditions are likely to require more medications and simultaneously have a higher risk of readmission.



3. Distribution of admission type and gender





Most of the patient suffering from diabetes come to the hospital in an emergency state making the diabete management even more critical. This often leads to an excessive anticipation of their discharge from hospital and, consequently, increases early readmission.

On the other hand, gender does not seem to affect patients' anticipated hospital readmission rate.

Performance metrics

AUC Score

AUC is an abreviation for Area Under the Curve. It is a commonly used metrics in Machine Learning. It evaluates the performance of binary classification models

Precision

Precision is a measure of how many of the positively predicted instances are actually true positives.

Recall

Recall measures the proportion of actual positive instances that are correctly identified by the model. It focuses on the ability of the model to find true positive.

F1 Score

The F1 score is a single metric that combines precision and recall into a balanced measure of a model's performance.

Our Models

We focused on tree based models because they are less inclined to assume linear relationships between variables.

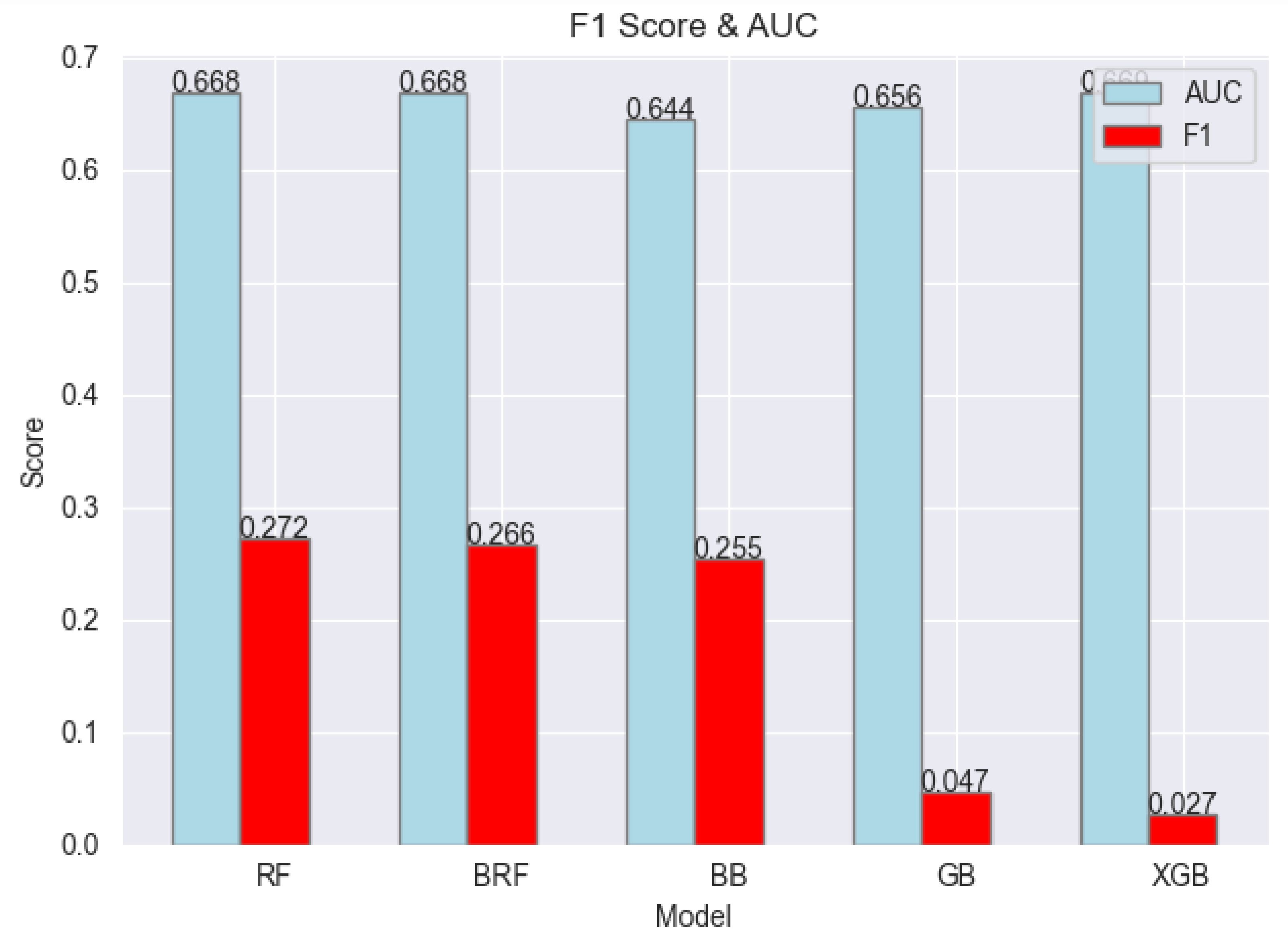
- **Random Forest Classifier** : is an ensemble learning method that constructs multiple decision trees during training and outputs the mode (classification) or mean prediction (regression) of the individual trees

- **Balanced Random Forest**

- **Balanced Bagging Classifier** : is an ensemble method similar to Random Forest, but it is specifically designed to address imbalanced datasets.

- **Gradient Boosting** : is an ensemble technique where multiple weak learners (usually decision trees) are trained sequentially.

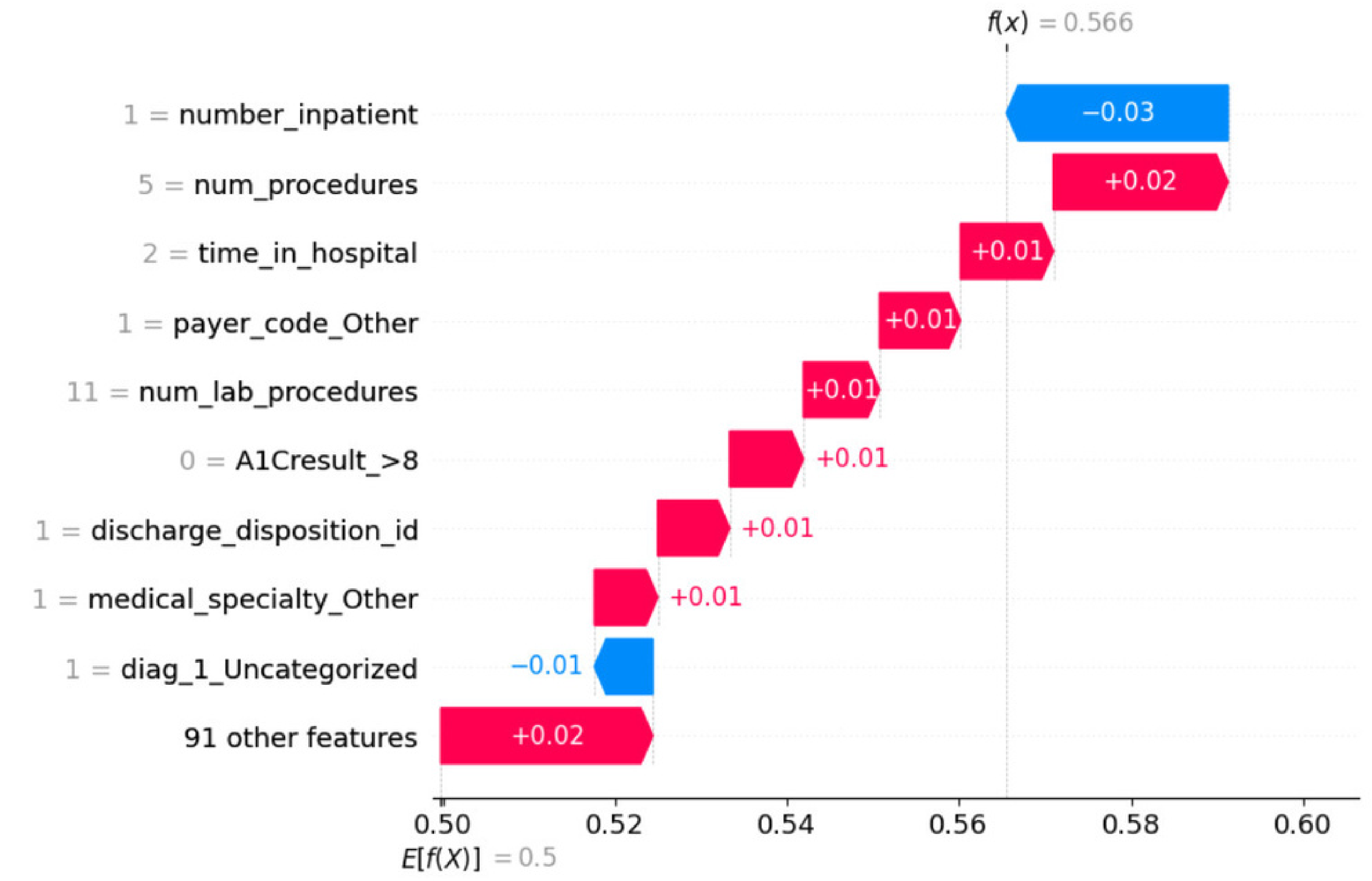
- **Xtreme Gradient Boosting**



We have chosen Random Forest with the class balancing parameter, which is the most efficient in our case.

Explainability

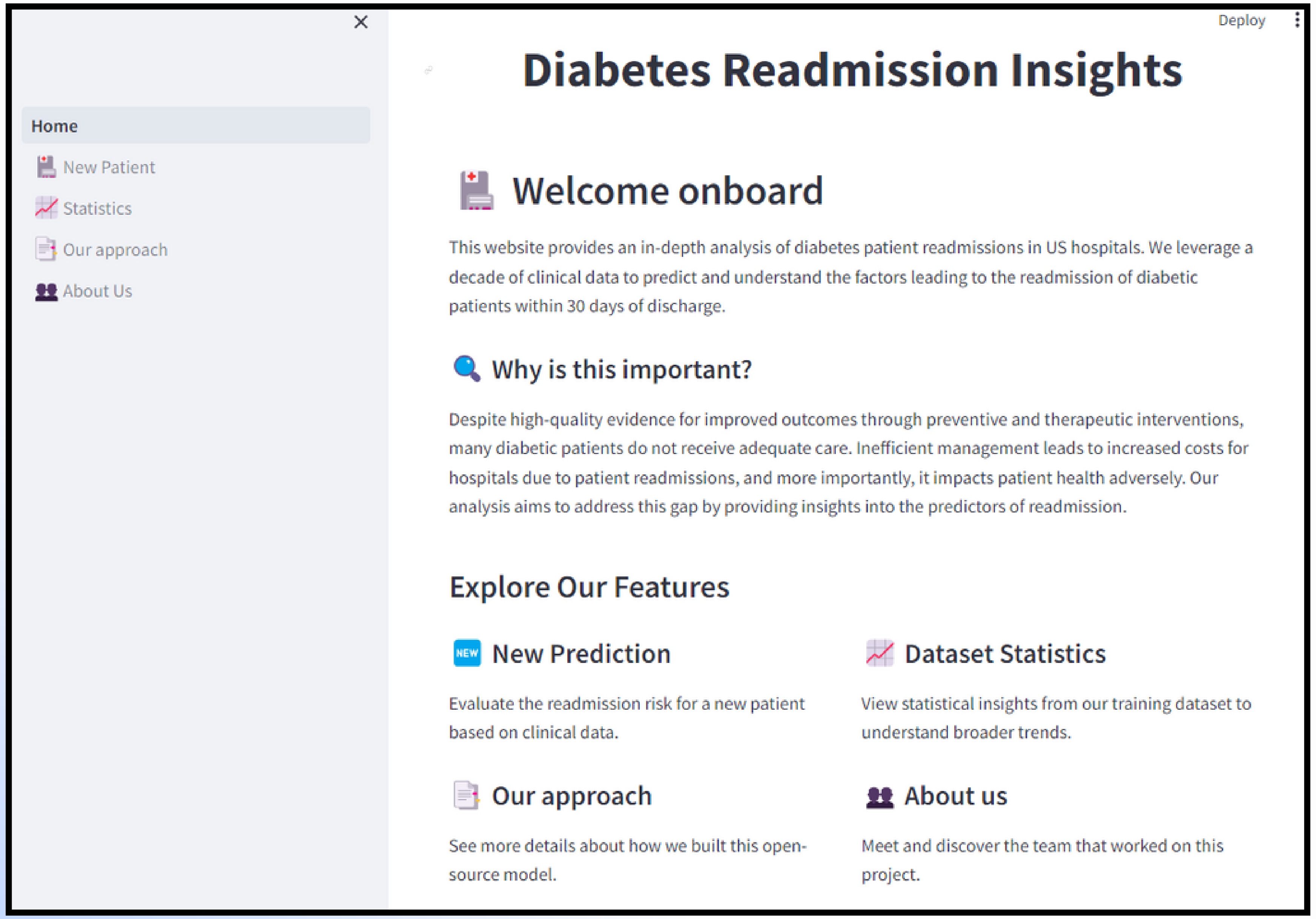
- One of the main strengths of our solution, beyond its accuracy, is that it enables you to **understand why a given patient is likely to be readmitted or not**
- This is done by using **SHAP**, a Python library that tells you which feature impacted positively or negatively the prediction
- In our specific use of SHAP we understood what are the variables that made a patient come back to the hospital.
- We understood that: num_impatients (number of visits during the previous year) had a major impact on the readmission of the patient



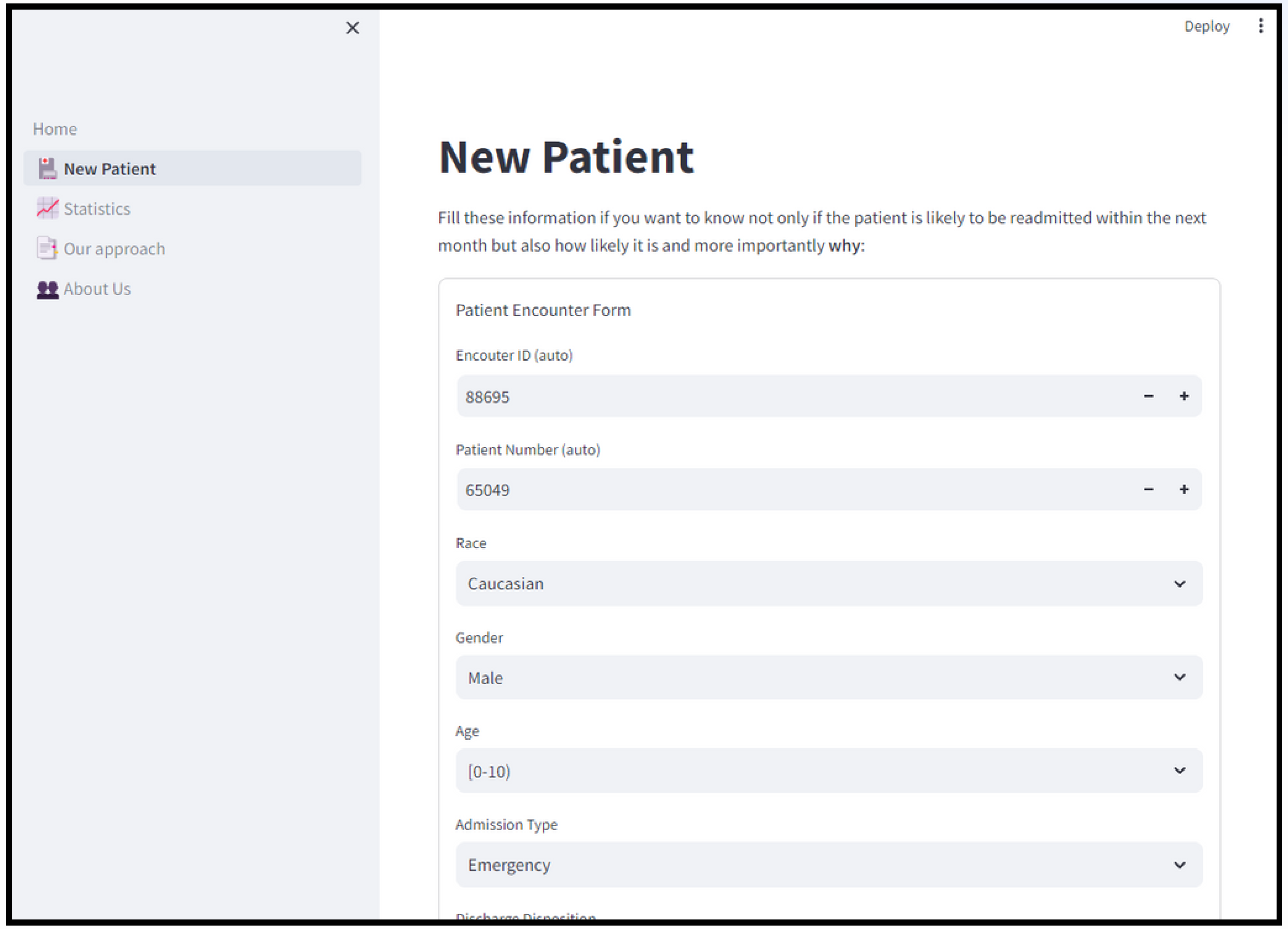
Example of a patient that is likely to be readmitted

Our Website

Moreover, our solution has been built to be easy to use. You directly try it on our website.



Home Page



New Patient

Conclusion

We managed to highlight the factors that played a major role in the readmission of a patient, through an accurate, easy-to-use and transparent solution.

We hope that our work will help hospitals to identify better where they should direct their efforts in order to ensure a more effective diabetes management.