Data Analysis Report: Predicting and Reducing Hospital Readmissions

Project: Data Science

Professor:
Dirk VALKENBORG

Eleftherios Kokkinis

August 24, 2024

Contents

1 Introduction

1.1 Background and Motivation

Hospital readmissions are a significant challenge in the healthcare industry, impacting both the sustainability of healthcare systems and patient well-being. Unplanned readmissions not only increase healthcare costs but also place additional strain on already burdened healthcare systems. The challenge of reducing these readmissions aligns with the United Nations Sustainable Development Goals, particularly:

- **SDG 3: Good Health and Well-being:** Aims to ensure healthy lives and promote well-being for all at all ages. Reducing unnecessary readmissions contributes directly to improving patient outcomes and optimizing healthcare resources.
- **SDG 9: Industry, Innovation, and Infrastructure:** Encourages innovation in healthcare systems, including predictive analytics to enhance patient care and resource management.

By predicting the likelihood of hospital readmissions, healthcare providers can take proactive measures to prevent them, thereby contributing to the sustainability of the healthcare system and improving patient care.

1.2 Framing the Problem

The primary focus of this project is to develop a predictive model that can identify patients at high risk of being readmitted to the hospital. By accurately predicting readmissions, healthcare providers can implement targeted interventions, such as enhanced follow-up care, personalized treatment plans, or social support, to reduce the risk of readmission. This not only enhances patient outcomes but also reduces the financial and operational burden on healthcare facilities.

2 Objectives

2.1 Specific Objectives

- **Objective 1:** Develop a predictive model to identify patients at high risk of readmission.
- **Objective 2:** Provide actionable insights and recommendations for healthcare providers to reduce the risk of readmission.

Reducing hospital readmissions contributes to the sustainability of healthcare systems by optimizing resource allocation, reducing unnecessary healthcare costs, and improving patient outcomes.The focus on predictive modeling also supports evidence-based decision-making in healthcare.

2.2 How I adhere to a SMART Work Plan

- **Specific:** Predict hospital readmissions using patient data and provide recommendations to reduce readmission rates.
- **Measurable:** Success will be measured by the accuracy, precision, recall, and F1-score of the predictive model.
- **Achievable:** Utilize publicly available datasets from Kaggle, focusing on patient demographics and treatment history. Due to time constraints, the model will be a first iteration, with potential for refinement in future work.
- **Relevant:** The project addresses a critical issue in healthcare sustainability and aligns with the SDGs.
- **Time-bound:** I will complete the project within a **single day - 23 August - 24 August**, focusing on data collection, model development, evaluation, and documentation.

3 Technical Description

3.1 Hospital Readmissions Dataset 1

time_in_hospital	num_lab_procedures	num_procedures	num_medications	number_outpatient	metformin-rosiglitazone_No	metformin-pioglitazone_No	change_No	diabetesMed_Yes	readmitted
14	41	0	11	0	True	True	True	True	0
2	30	0	12	0	True	True	False	True	1
5	66	0	22	1	True	True	True	True	1
3	63	0	8	0	True	True	True	True	1
5	40	0	6	0	True	True	True	False	0

Table 1: Dataset 1 (not all columns are shown for the sake of space)

- **Number of Observations:** 25,000
- **Number of Features:** 65
- **Missing Values:** None reported in the dataset; however, many features are binary encoded, which could imply missing data was encoded as separate categories

3.1.1 Features/Variables

- **Time in Hospital:** Number of days a patient was hospitalized.
- **Num Lab Procedures:** Number of lab tests conducted during the hospital stay.
- **Num Procedures:** Number of procedures (excluding lab tests) performed.
- **Num Medications:** Total number of unique medications administered.
- **Number Outpatient:** Number of outpatient visits in the year before the hospital admission.
- **Number Emergency:** Number of emergency visits in the year before the hospital admission.
- **Number Inpatient:** Number of inpatient visits in the year before the hospital admission.
- **Number Diagnoses:** Number of diagnoses recorded during the hospital stay.
- **Race:** Binary features indicating race categories
- **Gender:** Binary features indicating gender categories
- **Age:** Binary features indicating age ranges
- **Payer Code:** Binary features indicating the type of insurance payer

- **Medical Specialty:** Binary features indicating the medical specialty of the attending physician
- **Comorbidities:** Binary features indicating the presence of comorbid conditions
- **Treatment Details:** Binary features indicating whether certain treatments or medications were used

3.1.2 Target Variable

- **Readmission Status:** The target variable indicating whether a patient was readmitted (binary: 1 for readmitted, 0 for not readmitted).

3.2 Hospital Readmissions Dataset 2

encounter_id	patient_nbr	race	gender	age	metformin-rosiglitazone	metformin-pioglitazone	change	diabetesMed	readmitted
2278392	8222157	Caucasian	Female	[0-10)	No	No	No	No	NO
149190	55629189	Caucasian	Female	[10-20)	No	No	Ch	Yes	¿30
64410	86047875	AfricanAmerican	Female	[20-30)	No	No	No	Yes	NO
500364	82442376	Caucasian	Male	[30-40)	No	No	Ch	Yes	NO
16680	42519267	Caucasian	Male	[40-50)	No	No	Ch	Yes	NO

Table 2: Dataset 2 (not all columns shown)

3.2.1 Overview

- **Number of Observations:** 101,766
- **Number of Features:** 50
- **Missing Values:** Some features have missing values marked with ?.

3.2.2 Features/Variables

- **Race:** Categorical feature indicating race
- **Gender:** Categorical feature indicating gender
- **Age:** Categorical feature indicating age ranges
- **Admission Type:** Categorical feature indicating the type of admission
- **Discharge Disposition:** Categorical feature indicating the patient's discharge status (e.g., discharged to home, transferred to another facility).
- **Admission Source:** Categorical feature indicating the source of admission
- **Time in Hospital:** Number of days a patient was hospitalized.
- **Num Lab Procedures:** Number of lab tests conducted during the hospital stay.
- **Num Procedures:** Number of procedures performed during the hospital stay.
- **Num Medications:** Total number of medications administered.
- **Change in Medication:** Binary feature indicating whether the patient's medication was changed during the hospital stay.
- **Diabetes Medications:** Binary feature indicating whether diabetes medications were administered.
- **Readmitted:** The target variable indicating whether a patient was readmitted (within 30 days, more than 30 days, or not readmitted).

3.2.3 Target Variable

- **Readmitted:** The target variable is multi-class (indicating whether a patient was readmitted within 30 days, more than 30 days, or not at all). Since I am interested in a binary classification, this can be reduced to a binary variable indicating readmission in general (whether it is within 30 days or not).

3.3 Prediction Outcome

The primary outcome of this analysis is a binary classification that predicts whether a patient will be readmitted to the hospital. The goal is to maximize the model's accuracy while ensuring that it can be practically implemented in a healthcare setting, with a particular focus on minimizing false negatives (i.e., patients who are incorrectly predicted not to be readmitted).

4 Limitations and Constraints

Given the constraint of time, this project will prioritize the development of a working model over the exploration of complex, fine-tuned methodologies. Key limitations include:

- **Data Quality and Coverage:** The analysis will rely on publicly available datasets, which may have limitations in terms of data quality, completeness, and representativeness.
- **Model Complexity:** Due to time constraints, the model will likely be a basic machine learning model with potential for more sophisticated techniques to be explored in future work.
- **Generalizability:** The model will be tested on the available datasets and may not generalize to other healthcare settings.

5 Project Housekeeping

5.1 Reproducibility

To ensure reproducibility, the entire project workflow, including data collection, preprocessing, model development, and evaluation, will be documented using a Jupyter Notebook. All code, data processing steps, and model parameters will be clearly annotated, and the notebook will be shared alongside the final report.

5.2 Documentation of Work

The project will include detailed documentation covering the following areas:

- **Data Provenance:** Clear documentation of data sources, including URLs, data acquisition methods, and any preprocessing steps applied.
- **Model Development:** A step-by-step explanation of the model development process, including feature selection, model training, and evaluation metrics.
- **Results Interpretation:** Discussion of model performance and potential implications for healthcare practice.

5.3 Data Provenance

All datasets used in this project will be sourced from Kaggle, and any modifications or transformations applied to the data will be documented in detail. The data will be stored in a structured format, and any derived features or labels will be reproducible using the provided code.

6 Pre-processing and Final dataset

The two datasets represented similar information but in a quite different structure. I had to ensure consistency across both datasets before concatenation. Dataset 1 presented almost all information if a binary TRUE or FALSE fashion. TRUE in most cases meant that something was NOT present. Several columns were merged together to match the structure of Dataset 2. Common columns were kept while others had to be carefully renamed before concatenation since they conveyed the same kind of information. Information that did not exist in both datasets was discarded.

All kinds of treatments in dataset 1 were coded in TRUE or FALSE while dataset 2 often included other categories. TRUE meant that something was not present so it was coded to 0 and FALSE to 1. For dataset 2 the case was the opposite and 'Yes' was coded to 1 while 'No' to 0.
The case was similar for the target variable which for dataset 2 had 3 categories, 'no', within 30 days and after 30 days. These were coded to 0 and 1 (0 for 'no' and 1 for all others).

race	gender	age	time_in_hospital	payer_code	metformin-rosiglitazone	metformin-pioglitazone	change	diabetesMed	readmitted
Caucasian	0	[50-60)	14	SP	0	0	0	0	0
Caucasian	1	[50-60)	2	SP	0	0	1	0	1
Caucasian	1	[80-90)	5	MC	0	0	0	0	1
Caucasian	1	[50-60)	3	?	0	0	0	0	1
Caucasian	1	[80-90)	5	?	0	0	0	1	0

Table 3: Sample Data from Concatenated Dataset

The final dataset contains 126,766 rows and 41 columns.

7 Short Exploratory Data Analysis

No significant class imbalance exists.

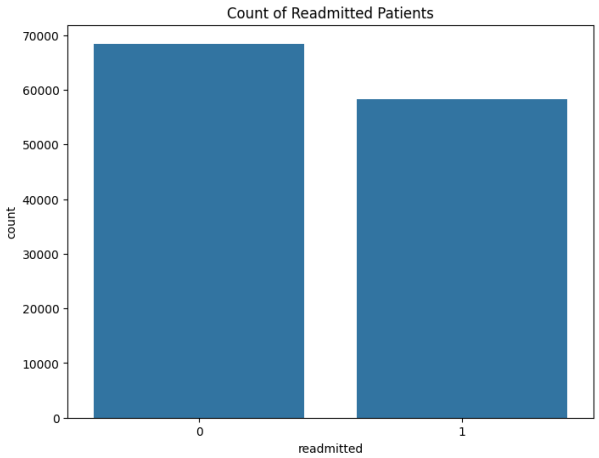

Figure 1: Counts of readmitted and not-readmitted. The observations of patients that were not-readmitted are slightly more. Overall no great class imbalance appears to exist

Other interesting observations:

- Patients who had more emergency visits in the past years than non-readmitted
- Readmitted patients also had more inpatient visits
- And also more outpatient visits than the non-readmitted ones

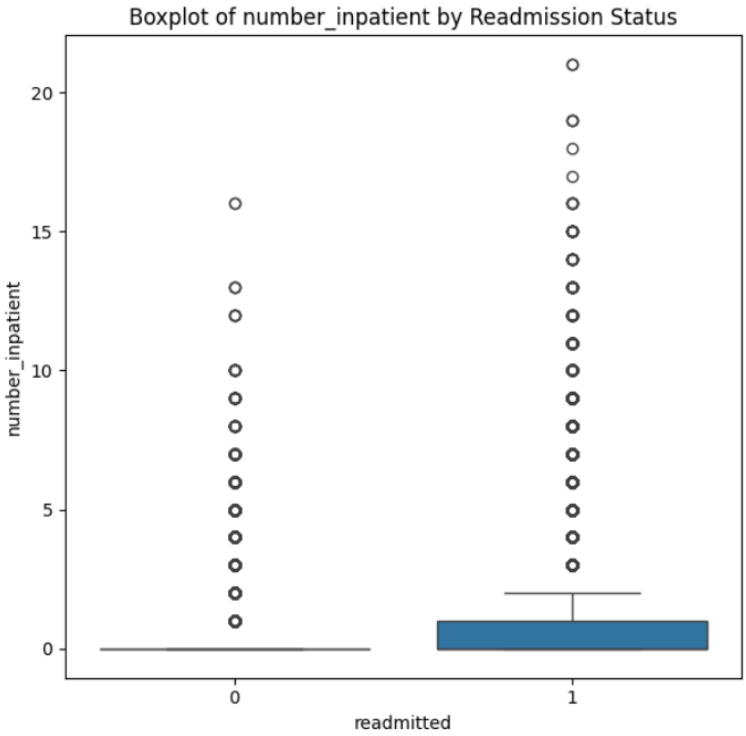

Figure 2: Distribution of admitted and re admitted over inpatient visits

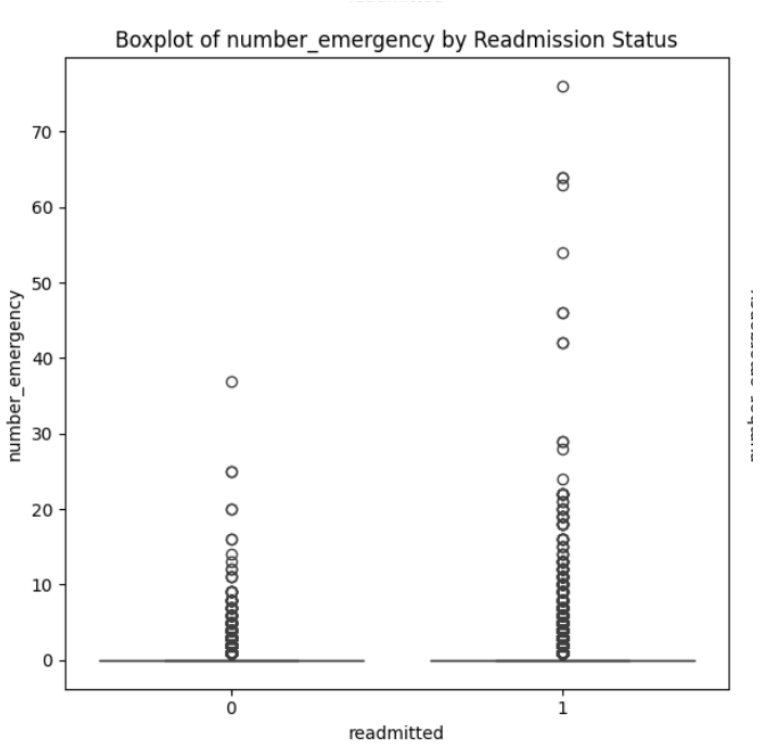

Figure 3: Distribution of admitted and re admitted over emergency visits

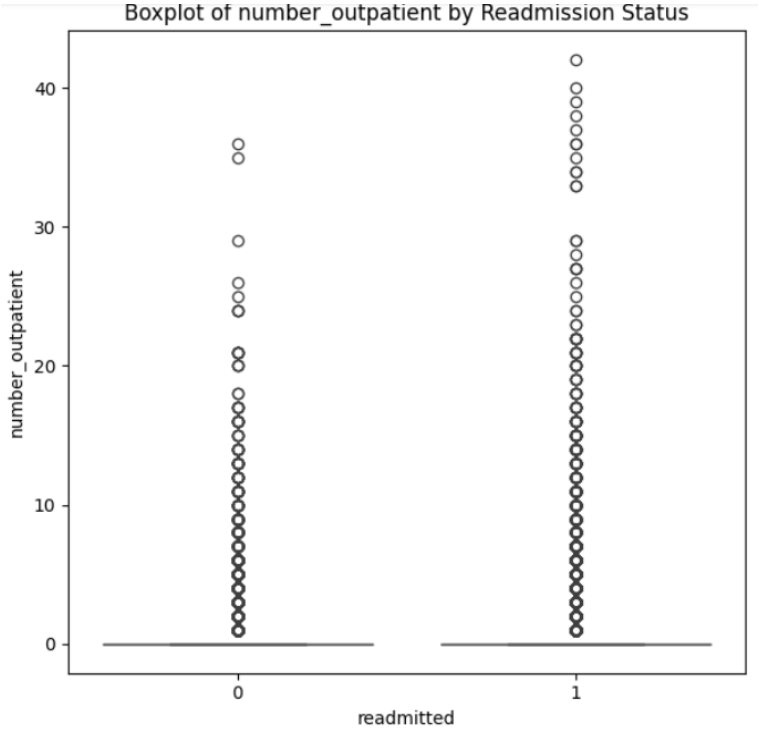

Figure 4: Distribution of admitted and re admitted over outpatient visits

8 Analysis of Modelling Results

Predictive modelling was performed using the random forest algorithm.Technical details on the implementation are provided in the Jupyter notebook. Feature selection was carried out and significant improvement in the model's performance was achieved.

The model was evaluated on the following scores:

Metric	**Value**
Random Forest Accuracy	0.7183
Random Forest AUC-ROC	0.8027

Table 4: Random Forest Model Performance Metrics

Although the model's performance is far from perfect, it still holds predictive value and potential for further enhancements exist.

The ROC curve is significantly better than random guessing which indicates a fair model with acceptable discriminatory ability.

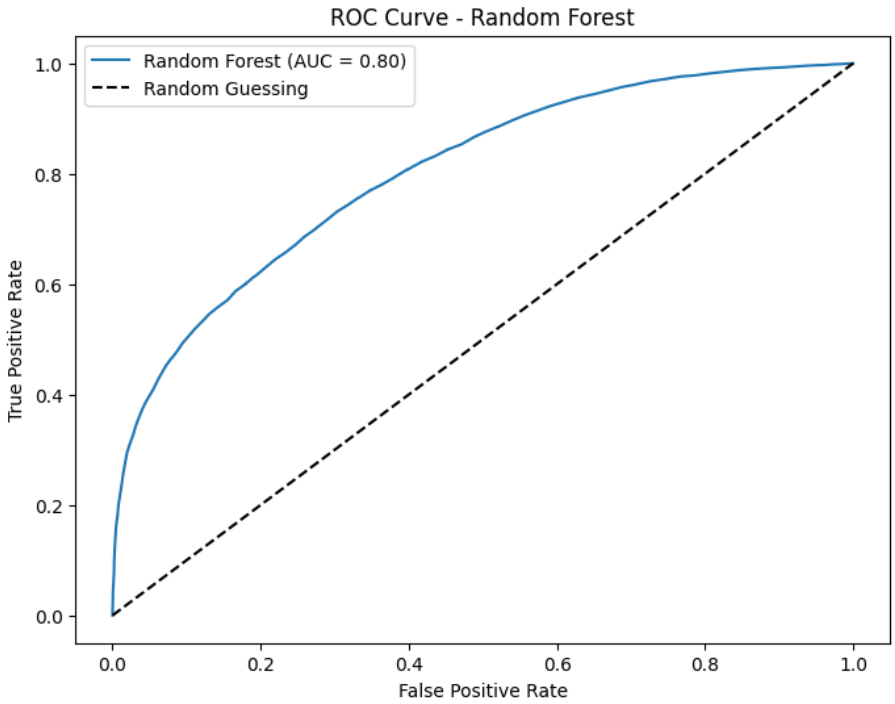

Figure 5: ROC curve

Below I can see the importance of each feature in the model's ability to correctly classify instances. The top 3 variables with predictive value were the number of medications that a patient received, the total procedures thay they went through during their stay in the hospital and the laboratory procedures they went through.

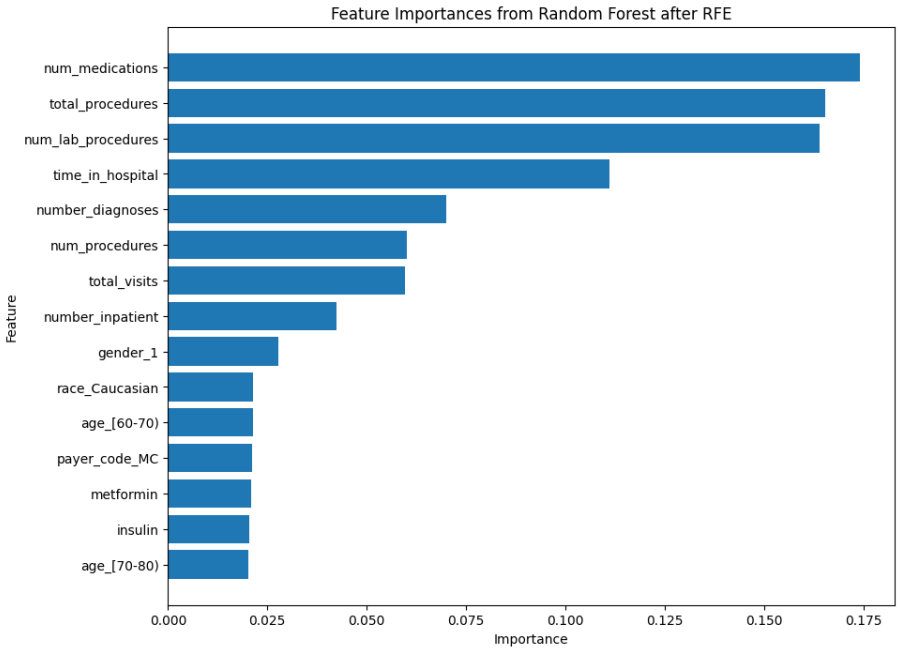

Figure 6: Feature Importance

9 Answering the problem's objectives

A predictive model was created that is able to classify re-admitted patients significantly better than random guessing. However, although I know which variables contributed the most to the correctness of these predictions, I cannot tell how each one of them affects the probability of re-admittance. For that purpose partial dependence plots and SHapley Additive exPlanations were used to help us understand how individual feature affect the outcome.

9.1 Partial Dependence Plots

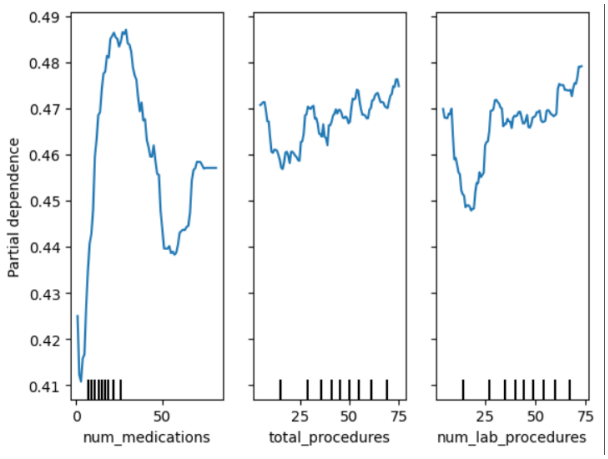

Figure 7: PDPs for the top 3 feature

For the sake of computational time, the top 3 features were selected for analysis. PDPs illustrate how the predicted outcome changes as the value of a feature. changes

In the case of the number of medications the plot shows a sharp initial increase in the predicted probability of re-admission as the number of medications increases from 0 to around 20. This suggests

that patients who are prescribed more medications have a higher likelihood of being re-admitted. The plot peaks around 20-30 medications, indicating the highest re-admission probability within this range. Beyond this peak, the probability starts to decline but then rises again slightly after 50 medications

The relationship between total_procedures and the predicted probability of re-admission shows small fluctuations with a slight upward overall trend.
The laboratory procedures is a part of the total procedures so we can see its effect now clearer. The plot shows an initial decrease in the probability of re-admission as the number of lab procedures increases up to around 30. Health care providers might want to look into the possibility that a moderate number of lab procedures are associated with a lower risk of re-admission, while an excessive number of lab procedures might be linked to more severe conditions, leading to a higher re-admission risk.

9.2 SHAP plots and values

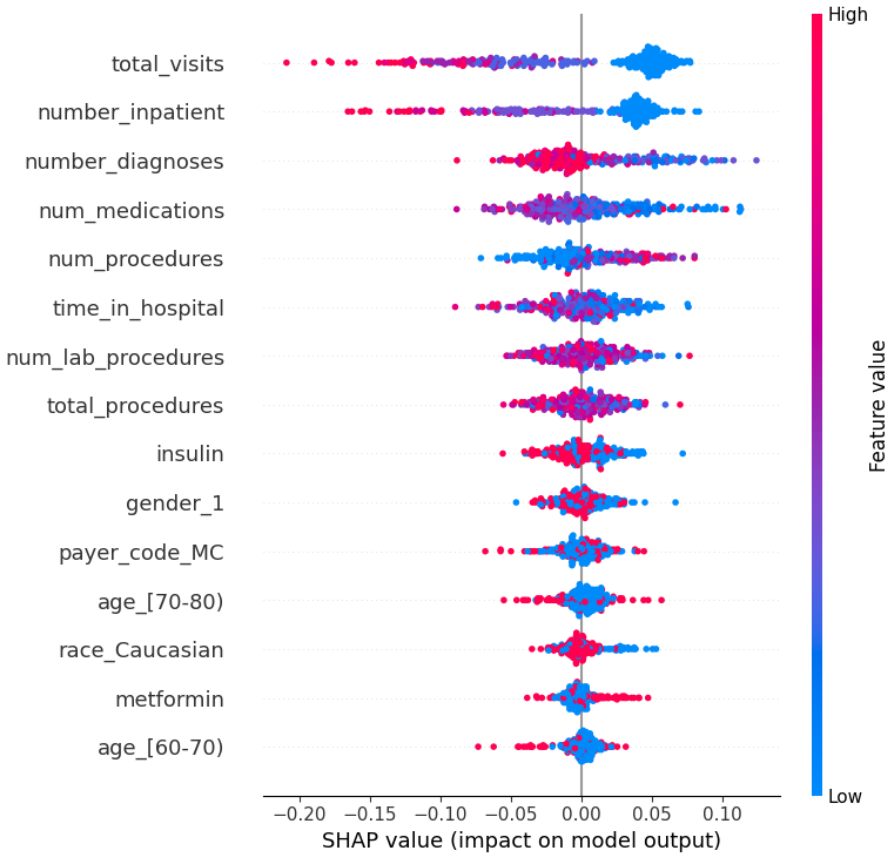

Figure 8: SHAP summary plot for a subset of 300 obs

9.2.1 Observations and Conclusions

The fact that low values of total_visits are predominantly on the right side of the plot with positive SHAP values suggests that having fewer visits is associated with an increased likelihood of re-admission.

That could imply that patients who have fewer total visits might not be monitored closely enough, leading to unresolved health issues and higher re-admission rates.

On the other hand, the red points on the left side indicate that higher values of total_visits are associated with a decrease in the probability of re-admission.

The case is opposite for the number of procedures. A lower number of procedures is associated with a smaller likelihood of readmission. While more procedures are linked to a higher likelihood. This could indicate patients with more complex health issues, whose condition might deteriorate again after their initial discharge.