**Case Study 2: Modelling Readmission of Diabetic Patients**

Dylan Scott

Jobin Joseph

Taylor Bonar

Satvik Ajmera

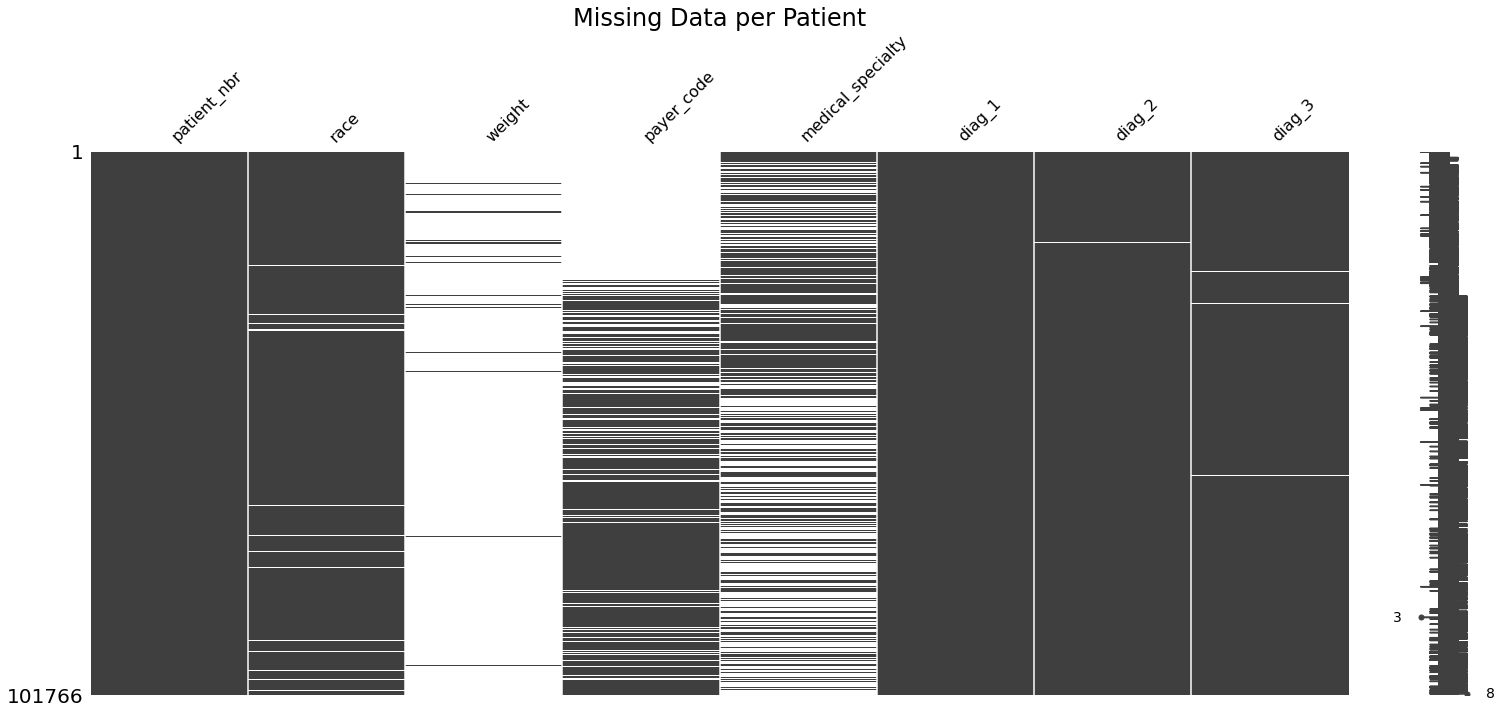
# 1 Introduction

The study is tasked to investigate whether diabetic patients would be readmitted to a hospital within the next 30 days (about 4 and a half weeks) based on their clinical records using Multinomial Logistic Regression. Identifying diabetic patients with a history of poor hyperglycemia management is an important aspect for outcome in terms of patient health, morbidities, and mortality. Prediction of whether a patient would result in readmission for hyperglycemia could be helpful in educating a patient on glycemic index (GI) in combination to additional insulin treatments to prevent readmission. The dataset utilized was split into two types of information: diabetic\_data.csv for anonymized patient records and ID\_mappings.csv for metadata context.

# 2 Methods

We wanted to try two different methods of data imputation and compare the results between Method A and Method B. Below, we have explained differences in how we have preprocessed the data and treated the missing values. Both methods have similarities, but crucial differences we will explain in this section.

Upon investigation, we discovered there were primarily 3 problematic fields that contained missing data: weight, medical\_specialty and payer\_code. (Refer to Figure 1)



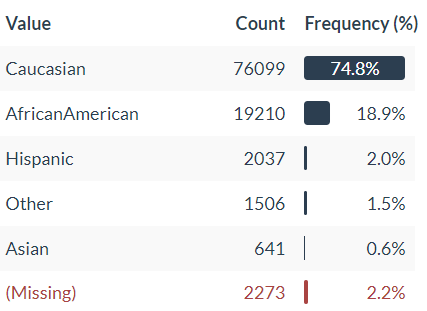
*Figure 1: Missing Data Aggregation. The top three columns with missing data were* weight, medical\_specialty, and payer\_code.

As you can see, 97% of the weight data is missing. Due to this large volume of missing data, we cannot impute or substitute values into our dataset for our model, so we moved forward of removing this specific field. We dropped the *encounter\_id* and *patient\_nbr* features for Method A and Method B because these are used for unique identification.

# Method A

In method A, we found 2-3 fields marked unknown or missing for less than 1% of the total dataset: gender and race. Due to these missing fields only affecting less than 1% of the dataset’s population, the team has decided to move forward with imputing the dataset using the mode for race. As seen in Figure 2, Caucasian made up 74% of our dataset. We imputed the missing race to be Caucasian.

**Race Distribution of Diabetic Patients**



*Figure 2: Racial Distribution of Diabetic Patients in Dataset*

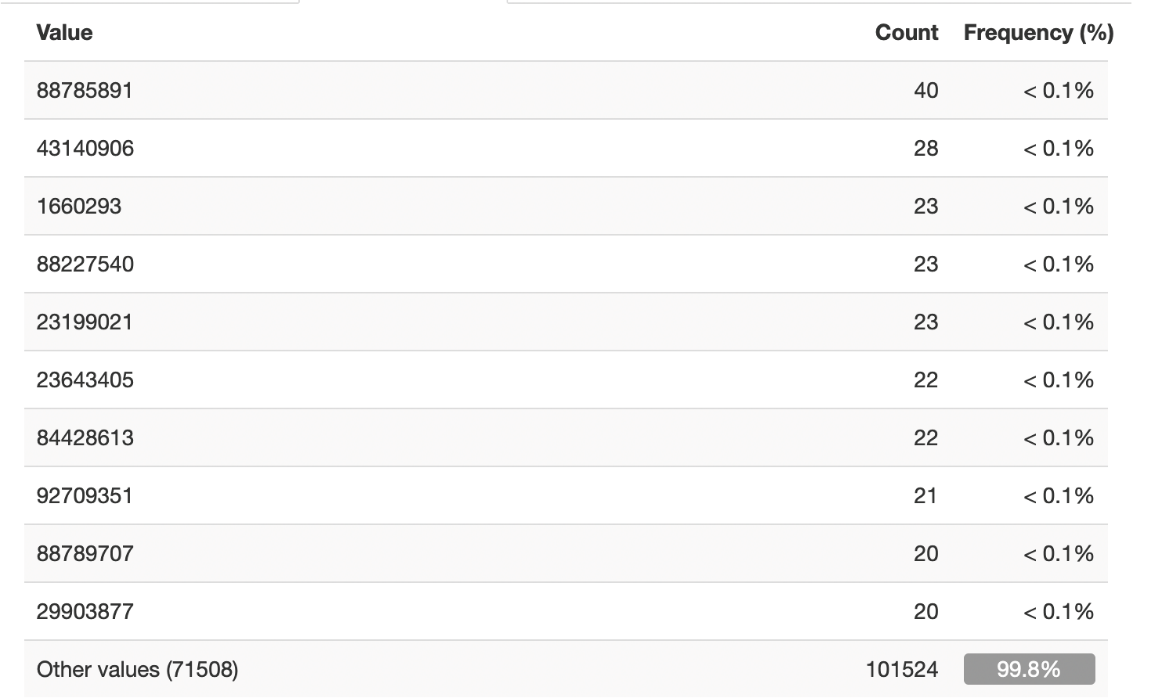
The discharge\_disposition is a field that contains information regarding the discharge outcome. We used the meta data to drop categories 11,13,14,19,20 and 21 since they were related to either death or hospice. This was done due to the unlikelihood of readmittance given one of these discharge dispositions.

Upon investigating the patient\_nbr we noticed some of the patients had multiple visits recorded. This led us to try a few different methods to help satisfy the assumption of independence. Given the nature of a hospital we thought that the last visit would be most important for modelling. Therefore, we dropped all duplicate patient visits before the last visit. Though this did improve our model accuracy at 74% it gave us a false sense of accuracy because removing all visits besides the last visit for a patient made our data very imbalanced. With the “No readmittance” category making up 74% of our training data. This becomes an issue in the deployment of a model as it will not perform well with a different dataset. Therefore, we did not remove any visits from customers and assumed that each visit was independent of another.

For later analysis many of the categorical variables were converted to numbers. This will help with identifying feature importance later. A dictionary is provided in the code section of this paper but, race, gender and age were all converted to numerical values. Next, we looked at one hot encoding the data so we could start our modeling. We used sklearn’s OneHotEncoder and StandardScaler to one hot encode and scale our data.

Since we were comparing two different methods here, we wanted to note the distribution of our categories. We had 51488, 11207, 35130 for no readmission, less than 30 days and over 30 days respectively, meaning our test data will still be skewed to no readmission.

**Number of Encounters Per Patient**



*Figure 3: Number of Encounters Per Patient ID*

# Method B

In Method B, we wanted our model to meet the independence assumption of logistic regression. However, in Method B, we kept only the first instance of a visit for each patient rather than the last visit. Thus, our training set went from 101,000 observations to 71,518 observations where each observation is a unique patient’s first visit. From there, we joined all the *admission\_type\_id,* *discharge\_disposition\_id*, *admission\_source\_id* with their respective names and removed the NaNs from these id’s because they were defined as null in the *ID\_mappings csv.* This resulted in the removal of 12,794 observations that contain at least one NaN value in *admission\_type\_id*, *discharge\_disposition\_id*, and *admission\_source\_id*.

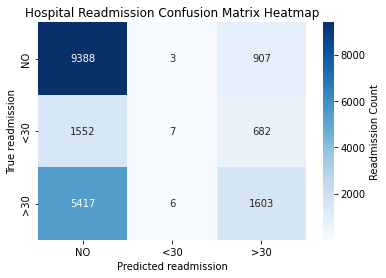
From there, weight and payer code were removed since there would be no way to impute these columns. Like the paper, we removed the levels from discharge\_disposition\_id that were related to *hospice* and *expired*, because this would bias our model's predictions for readmittance. For *medical specialty* and *race*, we decided to impute the NaNs as "Missing" because it makes sense to learn these features in the model.

In the end, we reduced the data to 60,672 observations with 45 input features. The response variable was recoded as an integer for each class between 1 to 3 ("NO":1,">30":3,"<30":2). For preprocessing, we used sklearn's pipeline to standard scale and one-hot-encode the categorical variables. We used sklearn’s stratified shuffled split where there were 5 splits and test size are 10%, because this cross-validation method allows for training and testing on balanced classes. Lastly the method for scoring is based on accuracy over the folds. The classes consist of 36,176 observations for NO, 19,116 observations for >30 and 5,380 observations for <30.

For Method A and Method B, we built and tested logistic regression models with L1/L2 regularization and without regularization for model evaluation. For variable importance we see that soft-max logistic regression results in each feature having a weight per class. We ranked feature importance for Method A and B, by taking the summation of the absolute value of the weights for each of classes and sorting them by this.

# 3 Results

For method A we used GridSearchCV from Sklearn to help hyper tune our parameters as well as compare different penalties of L1 and L2. Our overall best model had a C value of 0.001 and a penalty of L2. This model had an accuracy score of 56.5%. This model also included 5-fold cross validation with a test size of 0.20. Fig 3 shows where the misclassifications were in our model. As we can see the model was heavily influenced by no readmittance which was expected due to the distribution of the test data. We do not believe that our model’s performance was impacted heavily by the imputation of race.



*Figure 4: Confusion Matrix on Method A classification*

**Variable importance – For Method A**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Features** | **NO** | **<30** | **>30** | **Overall Importance (ABS)** |
| Unknown Race | 0.102636 | -0.01852 | -0.08411 | 0.205273 |
| insulin\_Steady | 0.0654 | -0.01959 | -0.04581 | 0.130799 |
| A1Cresult\_Norm | 0.047504 | -0.01507 | -0.03243 | 0.095007 |
| Other Race | 0.046789 | -0.0075 | -0.03929 | 0.093579 |
| num\_procedures | 0.04501 | -0.01483 | -0.03018 | 0.09002 |

*Figure 5: Top 5 Feature importance for Method A*

Looking at Method A, one could see that there are varied factors that have high importance for this model. The highest factor in the first model was unknown race, this indicates that the data points with empty race columns usually lead to no readmittance. This makes sense as the patient did not come back to the hospital, so it was not possible to fill in the race part of the information. The second was insulin. This makes sense as insulin is one of the primary medications used in controlling diabetes. It is normally a hormone that is produced by the pancreas in controlling glucose levels in the blood. Next in importance was the A1 C result, this is important, as having a normal level is indicative of blood sugar being in the proper range within the last few months. Glycated hemoglobin is a form of hemoglobin that is chemically linked to sugar. A normal A1C level is below 5.7% on average. Other races were also a variable that had an influence on readmittance, this is due to not enough data on the other races which could affect if the treatment is specific to them. Finally, the number of procedures was especially important, as the more procedures that one has the more likely one will be back to complications and or overall deterioration of one's health.

**Method B – Model Performance**

|  |  |  |
| --- | --- | --- |
| L1/L2 Regularization OR No Regularization | Best C – Inverse Regularization Parameter | Mean Accuracy Over Folds |
| L1 | 0.1 | 61.32% |
| L2 | 0.1 | 61.41% |
| N/A | N/A | 60.73% |

*Figure 6: Method B Performance for different penalties*

For Method B, we ran a grid search with L1 and L2 regularization to find the best mean accuracy over the folds. The values of C that were used for testing were 0.001, 0.01, 0.1. We found that the best performing logistic regression model was with L2 regularization and C = 0.1 resulting in the mean accuracy of 61.41%.

**Variable Importance – For Method B**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Features** | **NO** | **<30** | **>30** | **Overall Importance (ABS)** |
| diag\_1\_V58 | -0.5461 | 0.8284 | -0.2823 | 1.6567 |
| discharge\_disposition\_id\_Discharged/transferred to a long term care hospital. | 0.4198 | -0.6438 | 0.2240 | 1.2875 |
| discharge\_disposition\_id\_Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital | -0.2703 | 0.6190 | -0.3487 | 1.2381 |
| discharge\_disposition\_id\_Discharged/transferred within this institution to Medicare approved swing bed | -0.4772 | 0.5534 | -0.0762 | 1.1068 |
| discharge\_disposition\_id\_Discharged/transferred to another rehab fac including rehab units of a hospital. | -0.3158 | 0.5257 | -0.2099 | 1.0514 |

*Figure 7: Top 5 Feature importance for Method* ***B***

In Method B we can see that the discharge ID played a significant role in our model. 4 of the 5 highest (farthest from 0) feature importance's were regarding the discharge disposition ID. Knowing the discharge disposition can go a long way in helping to predict whether someone will return to the hospital and within what timeframe. We can see from Figure 7 that someone with a discharge disposition ID of transferred to long term care hospital has a strong positive importance on them not returning to the hospital while it also has a strong negative importance on them returning with 30 days. This is valuable information to have when modeling and the two polar ends of these feature importance helps us identify which discharge disposition IDs to look at when making a prediction. The other IDs show the opposite story as they have negative importance for no readmission but positive importance for returning in 30 days. These may be more severe circumstances for the patients and may indicate they will need to be readmitted soon to the hospital.

# 4 Conclusion

When we compare Method A to Method B, we see they both performed similarly with method B resulting in a better accuracy score of 61.32% opposed method A’s 56.5% accuracy. The optional model for these two methods were also different where method A favored a lower C value at 0.001 and a penalty of L2 whereas method B favored a higher C value of 0.1 and the same L2 penalty. When it comes to deploying a model and giving our recondition, we found that method B would perform better from a production standpoint. Many of the features deemed important to method B revolved around the discharge reason such as long-term care, psychiatric hospital care and transferred within this institution to Medicare approved swing bed were among the highest (farthest from 0) features we saw. This means that knowing discharge code will help our model categorize whether someone will return to the hospital or not and how soon.

# Appendix Code

\

import pandas as pd

# Read initial csv and fill missing values appropriately

diabetes\_df = pd.read\_csv("../dataset\_diabetes/diabetic\_data.csv", header=0, na\_values="?", low\_memory=False)

# Generates Missing Matrix on Relevant Fields (i.e., Figure 1)

import missingno

import matplotlib.pyplot as plt

missingno.matrix(diabetes\_df[["patient\_nbr","race", "weight", "payer\_code","medical\_specialty", "diag\_1","diag\_2","diag\_3"]])

font1 = {'color':'black','size':24}

plt.title("Missing Data per Patient", *fontdict*=font1)

\

Method A file:



Method B file:

