CWX Report - March 2024

Table of Contents

Introduction

This activity demonstrates how we will predict and build a classification model with the given hospitals diabetes dataset," diabetic_data.csv" and predict whether a patient will readmit to the hospital within 30 days.

Description of the given dataset

The given raw data has 101766 rows and 47 columns. This dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Each row concerns hospital records of patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days.

Using the given dataset, the goal is to determine the early readmission of the patient within 30 days of discharge. The data is not clean, so need to apply appropriate methods to clean the data. A suitable model with the given raw data needs to be built and appropriate performance metrics to evaluate the performance of your model.

Part 1

In the **first part** of the activity, we imported libraries and read the data file to build up a basic model which will further help us in predicting this dataset.

1. Data Munging (Data Cleaning and Transformation)

- Initiated with data cleaning where columns like “encounter_id” have been removed.
- Identified the missing columns and replacing the character ‘?’ with NaN, which was represented instead of missing values. Showing the summary of the missing values.
- Converted the response variable, ‘`readmitted`’ to binary feature, by replacing the cells which contains ‘<30’ to 1 and ‘>30’ and ‘NO’ to 0 respectively.
- Calculated the percentage of missing values and dropped the columns with >90% missing values.
- Dropped the columns which are near zero-variance columns.
- Dropped rows with null values and showing the resulting data frame shape and summary statistics of numerical columns.
- Identified and removed the outliers from the numerical columns of the data frame. Removed the zero variance columns post outliers are performed.
- Applied Min-Max scaling feature normalisation for normalising the numerical columns.

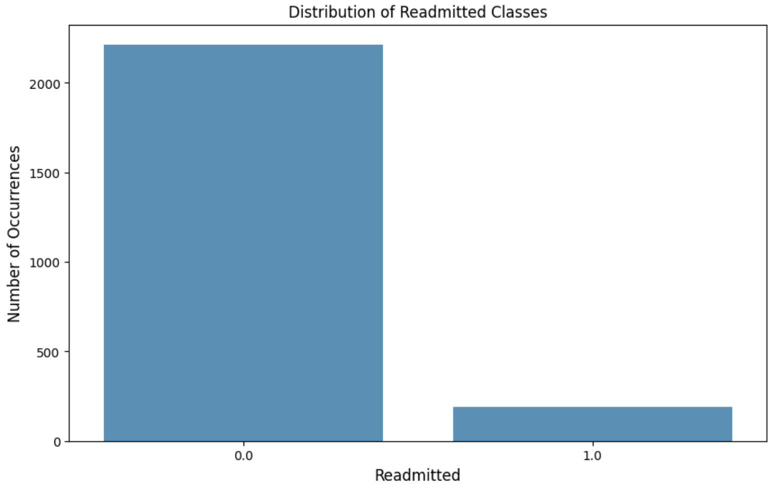
2. Data Visualization

- The aggressive cleaning resulted in a smaller dataset.
- Started data exploration, to show the relationships between the various features of the dataset.

a. Distribution of readmitted unique classes

- The below “figure 1”, shows distribution of unique classes in the column readmitted. It shows two unique categories ‘0’ and ‘1’ in readmitted column in ‘x-axis’ and number of occurrences in the ‘y-axis’. The bar charts shows that the number of patients readmitted in <30 days is very less compared to that of patients who are not readmitted or readmitted in >30 days.

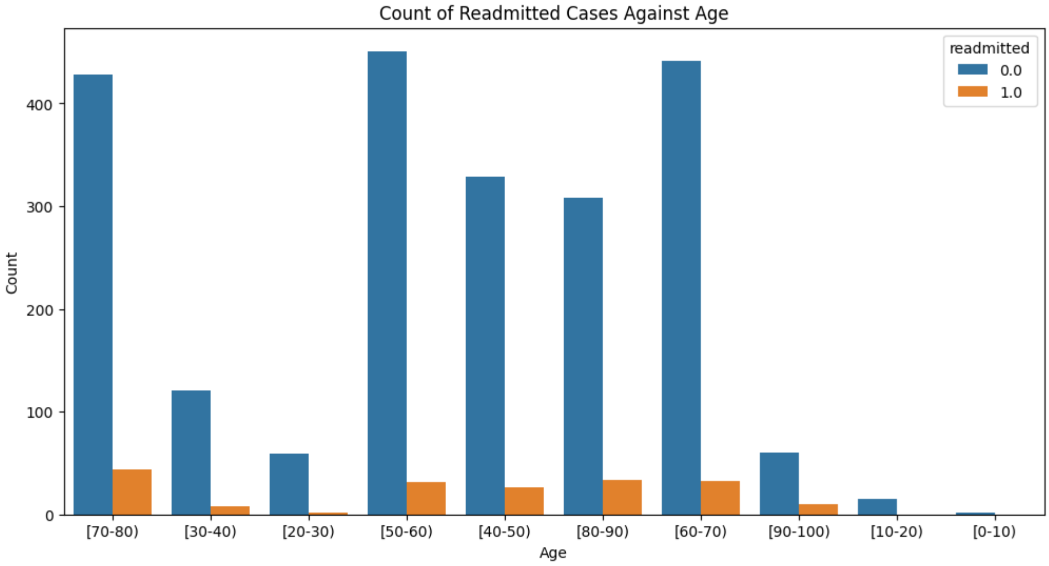
Figure1:



b. Count of readmitted cases against age

- ➢ The below “figure 2”, shows grouped bar charts, comparing the count of patients readmitted across different age groups. Most age groups show higher count of not being readmitted or readmitted >30 days, with the most prominent age group being [50-60].

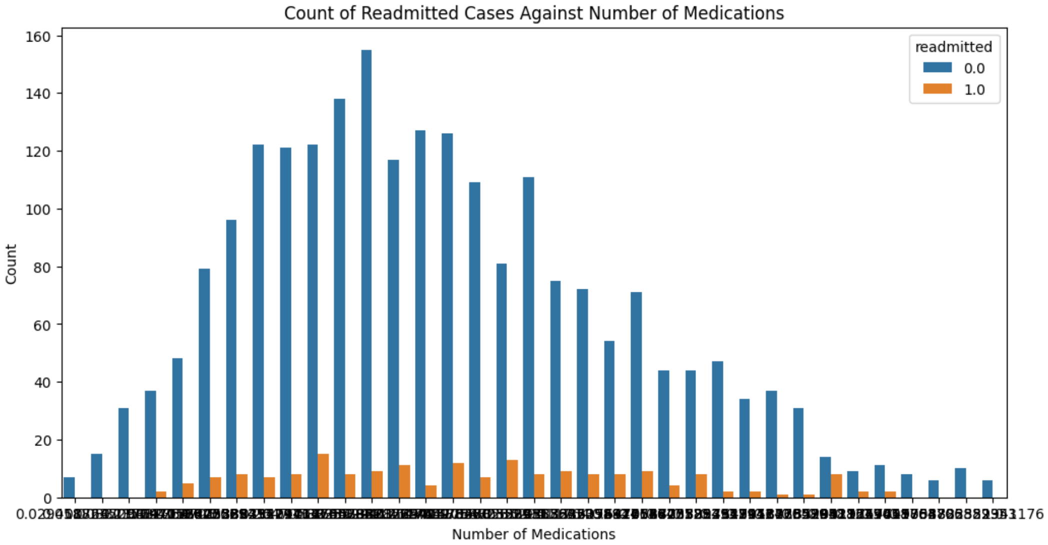
Figure 2:



c. Count of readmitted cases against number of medications

- ➢ The below “figure 3”, shows grouped bar charts, comparing the count of patients readmitted against number of medications. Most number of medications show higher count of not being readmitted or readmitted >30 days

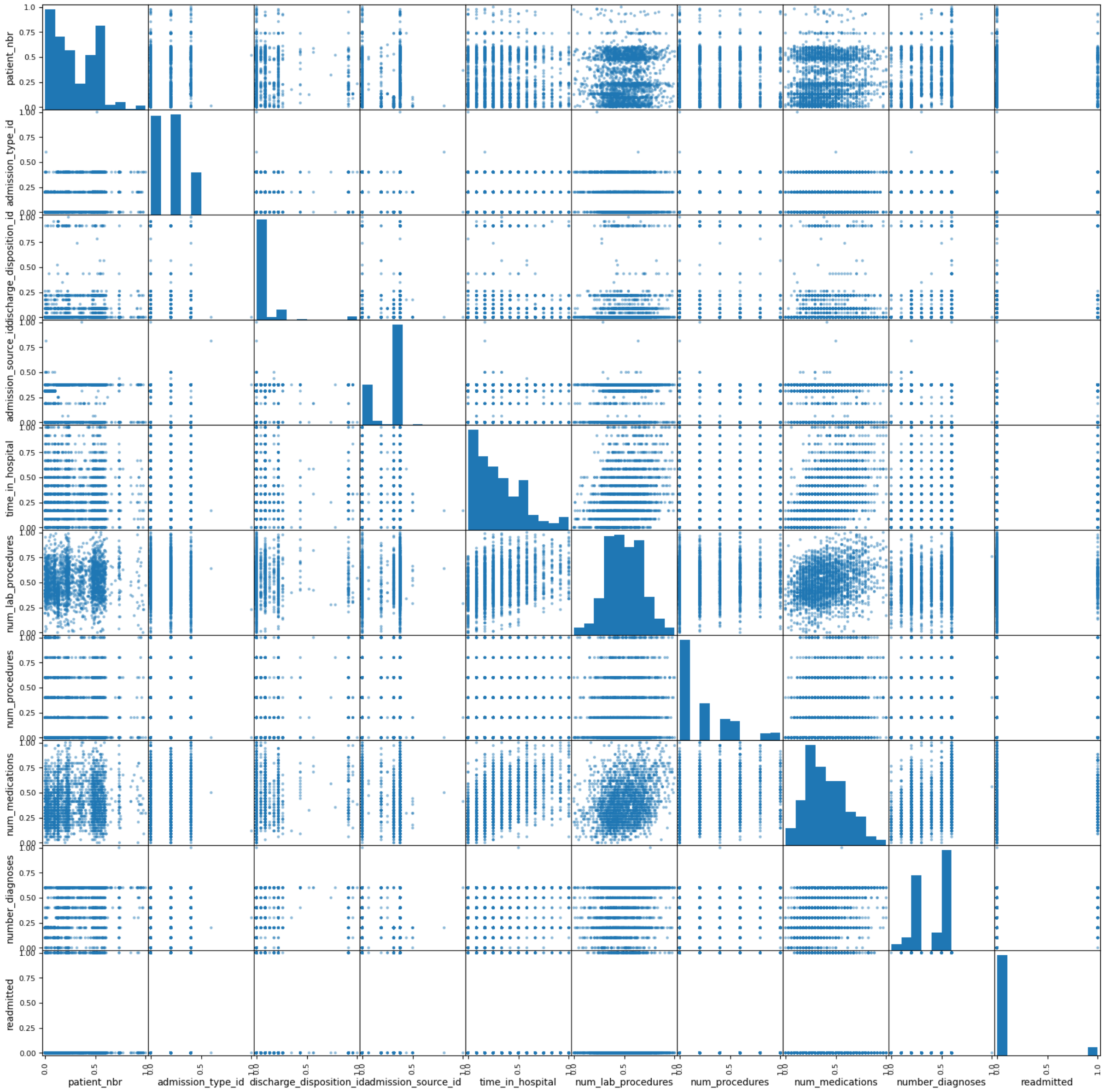
Figure 3:



d. Scatter Matrix

- The below “figure 4”, shows a scatter plot matrix for each pair of numerical variables in the provided Data set. Each columns in the matrix depicts the relationship between two variables, one along the x-axis and y-axis. The diagonal of the matrix displays a histogram of each variable that gives information on the distribution of each variable and can aid in the identification of outliers or unusual values.

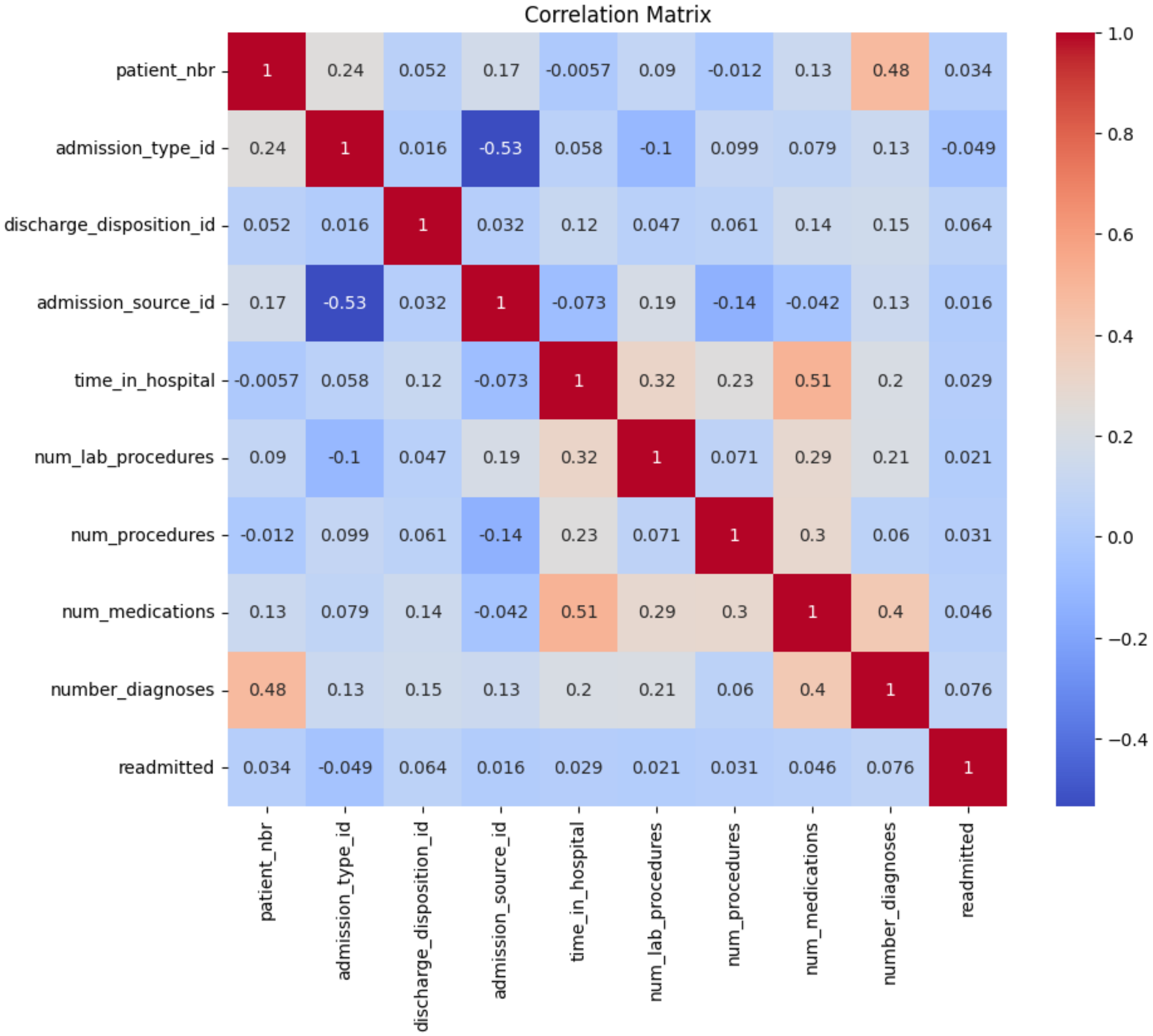
Figure 4:



e. Correlation Matrix

- The below “figure 5”, shows a heatmap, which provides a rapid summary of the relationship between distinct Data Frame characteristics. The values range from -1 to 1, with 1 being a perfect positive correlation, -1 representing a perfect negative correlation, and 0 representing no connection. Here the colour intensity shows the degree of the correlation, with deeper hues indicating a stronger correlation. A positive correlation indicates that as one variable grows, so does the other. Whereas a negative correlation suggests that as one variable increases, so does the other.

Figure 5:



f. Distribution of each numerical columns against readmission

- The below figures from “figure 6” to “figure 15”, shows distribution of each numerical column against readmission, which helps in identifying the correlations, patterns that are crucial for feature selection.

Figure 6:

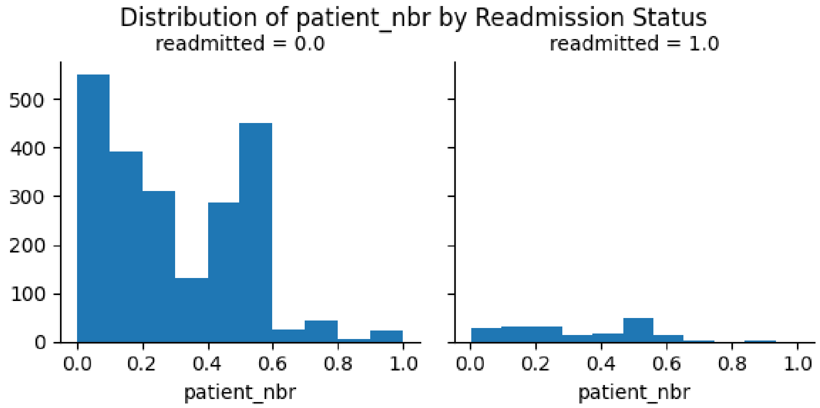


Figure 7:

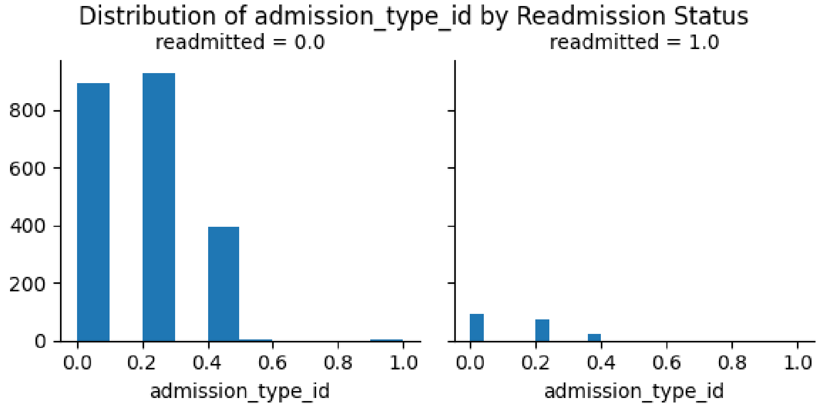


Figure 8:

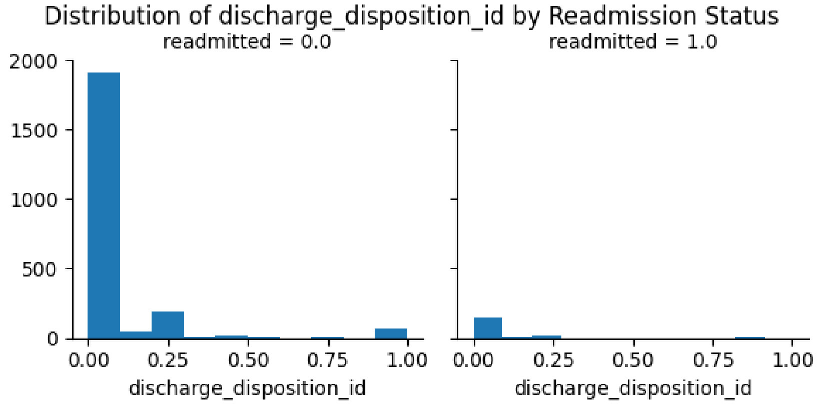


Figure 9:

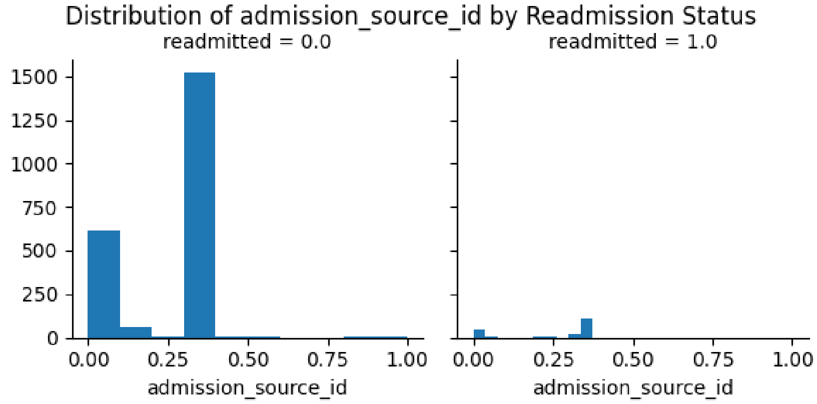


Figure 10:

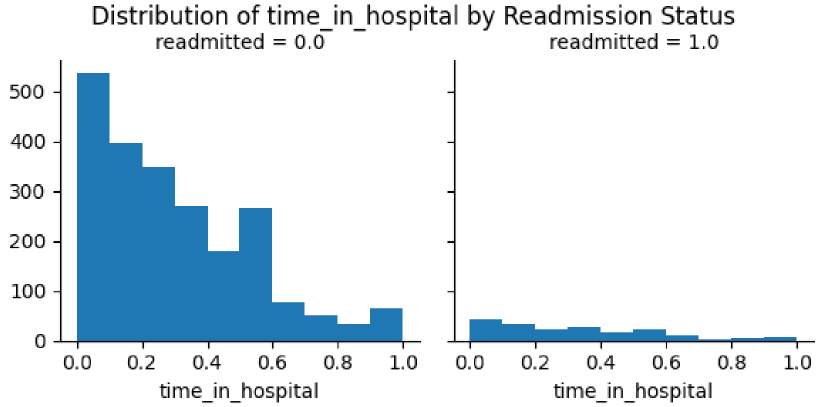


Figure 11:

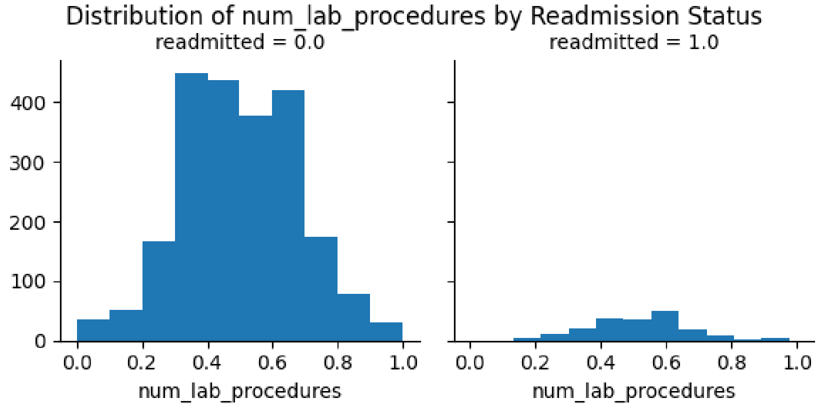


Figure 12:

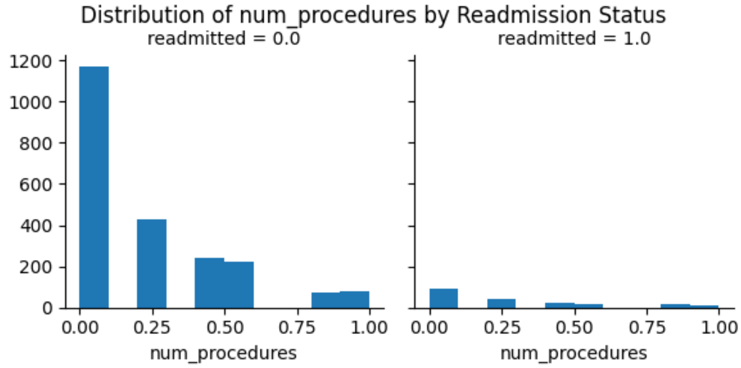


Figure 13:

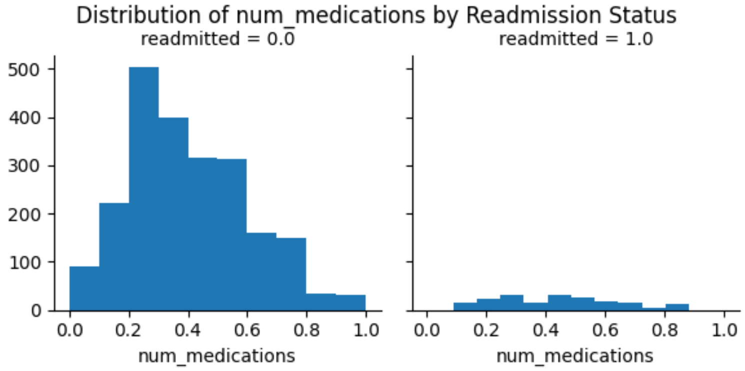


Figure 14:

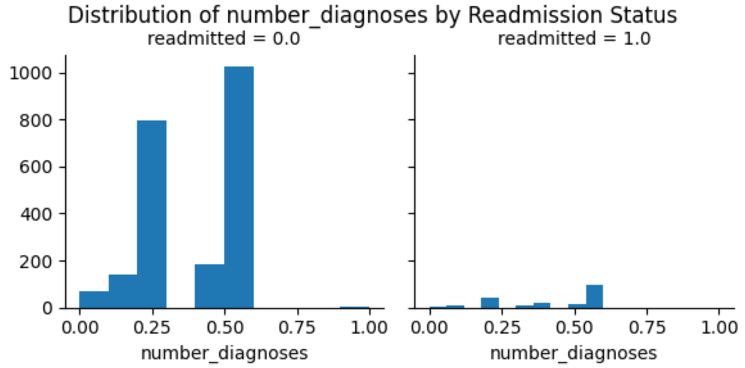
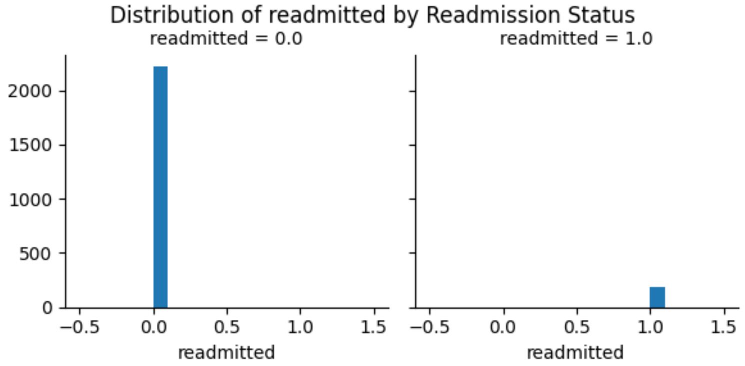


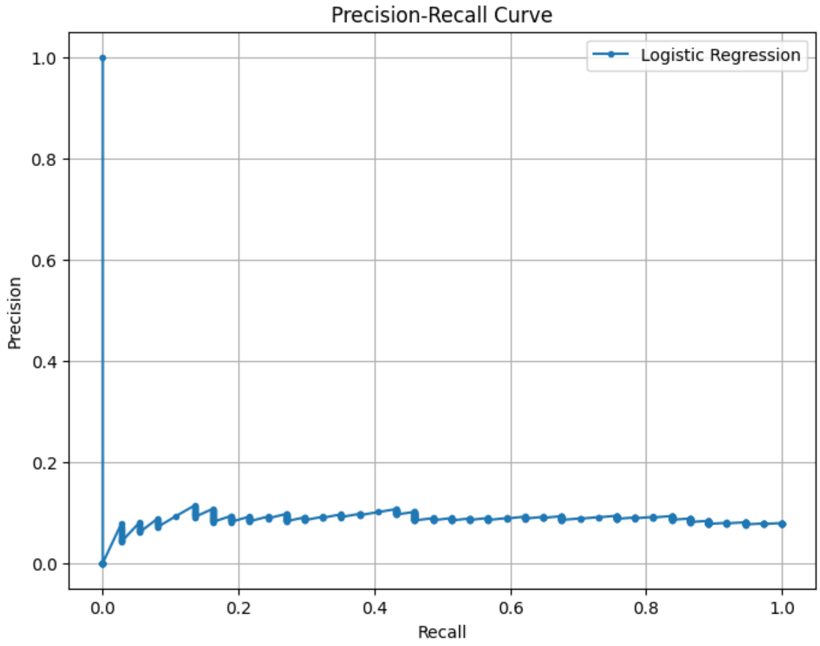
Figure 15:



3. Model Building

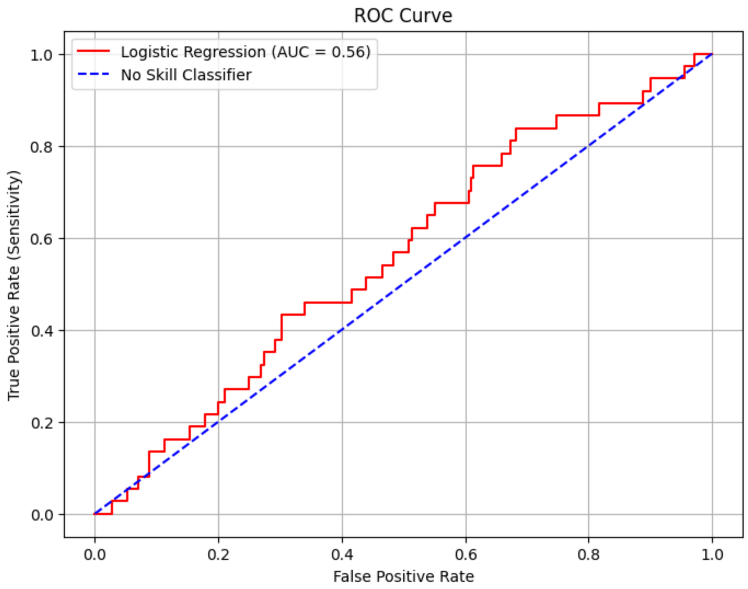
- Selected the predictors which will have impact in predicting readmission.
- Selected the predictor columns data and readmitted columns data for estimating the feature sections. Further by using this appropriate features to predict the readmission.
- Defined X and Y for the model and further split them accordingly to train and test.
- Defined a linear model with Logistic Regression, to train the model. Further got the intercept and coefficient of the model as below:
  - Linear model intercept: [-3.11332666]
  - Linear model Coefficient: [[0.17341569 -1.34521779 0.76699102 0.1836179 0.04644122 0.10542888, 0.51960027 0.42517802 0.7717646]}
- Showing the model scores for train and test data as below:
  - Score against training data: 0.921436004162331
  - Score against test data: 0.9230769230769231
- Evaluated the model with various performance metrics like, Mean hits, Accuracy score, cross validation mean as below:
  - Mean hits: 0.9230769230769231
  - Accuracy score: 0.9230769230769231
  - Cross-validation mean scores: 0.9217652143845088
- Further evaluated the model with Precision recall curve (Figure 16) and ROC curve (Figure 17) along with AUC as below:

Figure 16:



- AUC = 0.5608108108108107

Figure 17:



- Calculate the mean squared error of the model as below:
  - Mean Squared error: 0.07823553890969621
- To summarize the model shows high scores in cross-validation mean and for accuracy in predicting outcomes, with a score above 0.92 on both training and testing data. But the ability to distinguish between classes could be better, as indicated by AUC of 0.56.

Part 2

1. Improved Model

Reconsidering the entire data frame again after removing the ‘encounter_id’ column, replacing the character ‘?’ with NaN and converting the readmitted column to binary values.

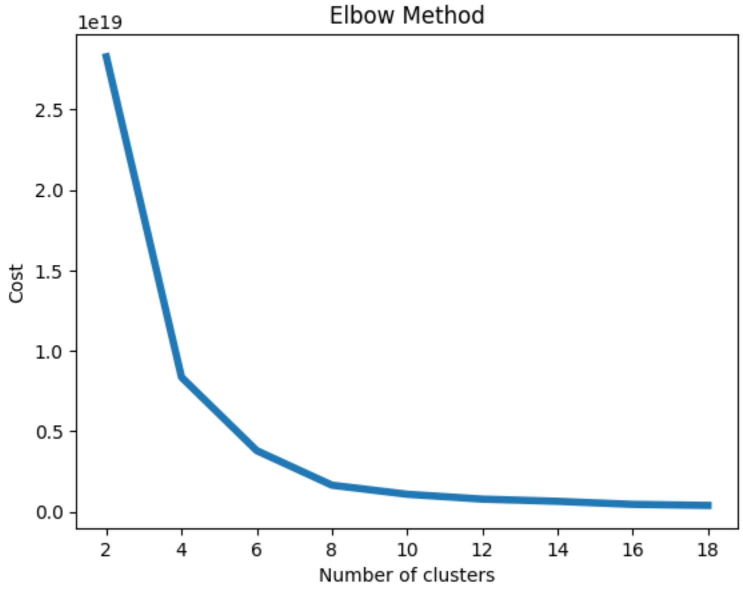
a. Data Cleaning

- Split the columns list based on their data types as indicated in glossary as, Id, categorical and numerical.
- Instead of dropping the empty values, applying imputation to the columns based on categorical and numerical data. Further showing the summary of the missing values in dataset.
- Identified and removed outliers from the numerical columns in the dataset and further removing the columns with zero variance.
- Applied min-max scaled feature normalization on numerical columns, to normalize the numerical values, further replacing those numerical values in the dataset accordingly.

b. Kmeans Clustering

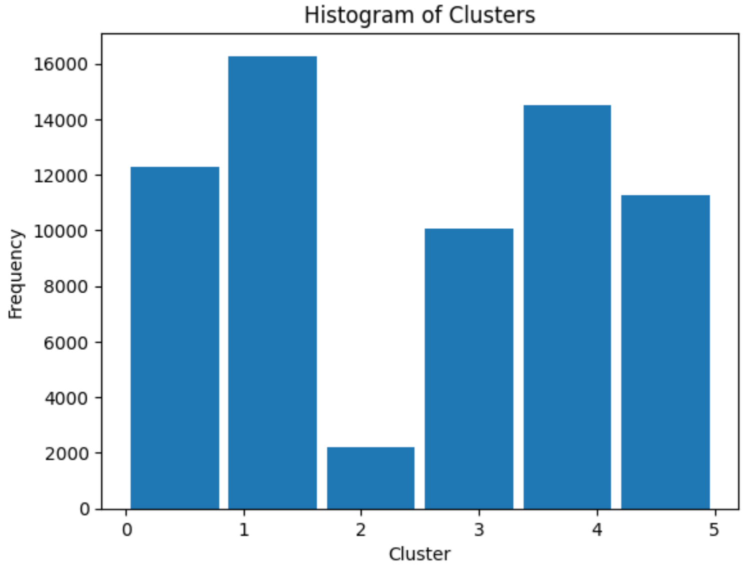
- Applied elbow method to determine the optimal number of clusters, based on the point of inflection or "elbow" in the Elbow method plot. The number of clusters corresponding to this point is often chosen as the optimal number of clusters for k-means clustering as shown in figure 18.

Figure 18:



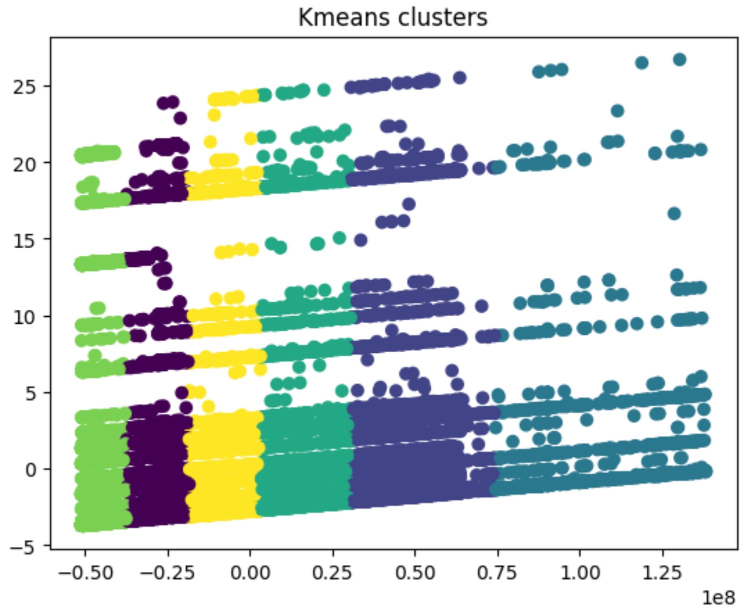
- ➢ Applied K-Means clustering, which is a well-known clustering method that is used to group related data points into a predefined number of clusters. The purpose of K-Means is to divide a given collection of data points into k clusters (where k is a pre-specified integer) in such a manner that the within-cluster sum of squares is minimized. With the K-Means technique, we have taken 6 clusters and centroids of Kmeans for this given Diabetic dataset.
- ➢ The frequency of each cluster in the data set is shown by the histogram (Figure 19). We can see the distribution of data points among the different clusters by visualizing the histogram.

Figure 19:



- ➢ Further Plotting Kmeans clusters, to allow us to see the clusters in two dimensions and see how well the K-means algorithm grouped the data points into distinct clusters in below figure

Figure 20:



- Trained each clusters with the linear model with Logistic Regression, and evaluated each cluster model and their respective precision recall curve and ROC curve is plotted accordingly.

Conclusion

Based on the above elbow method, we have used 6 clustering for our model and its respective values as shown below in a table format,

Cluster	X_Feature	Y_Feature	Training Score	Testing Score	Mean Hits	Accuracy Score	Cross-validation score	AUC
Cluster0	(12293, 11)	(12293,)	0.917	0.915	0.915	0.915	0.916	0.52
Cluster1	(16271, 11)	(16271,)	0.901	0.899	0.899	0.899	0.900	0.50
Cluster2	(2198, 11)	(2198,)	0.921	0.929	0.929	0.929	0.923	0.39
Cluster3	(10048, 11)	(10048,)	0.903	0.913	0.913	0.913	0.905	0.49
Cluster4	(14500, 11)	(14500,)	0.902	0.911	0.911	0.911	0.904	0.48
Cluster5	(11282, 11)	(11282,)	0.893	0.891	0.891	0.891	0.892	0.47

Overall, based on the result we executed all 6 clusters and the values are captured in a table as in above. Cluster 2, shows the best performance with the highest testing score (0.929), mean hits (0.929), accuracy score (0.929), and cross-validation score (0.923), even though its AUC value is the lowest (0.39). It suggest that while the model predicts accurately for the given data, it may not generalize well for class separation.

References

[1] Beata Strack, Jonathan DeShazo, Chris Gennings, Juan Olmo, Sebastian Ventura, Krzysztof Cios, John Clore. "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Record"2014. [Online]. Available: http://www.hindawi.com/journals/bmri/2014/781670/.