Final Project

Due: Friday, May 3rd @ 11:59PM

Implementing a Model to Predict Hospital Readmission from Diabetes Diagnosis

Brianna L. Palmisano, Jianna J. Estevez & Julianna M. Lomonte

St. John's University, The Peter J. Tobin College of Business

BUA 611/3311: Machine Learning for Business

Professor Yanni Ping, PhD.

Spring 2024

Table of Contents

I. Introduction

This model utilizes several features of diabetes diagnosis to predict levels in patients' hospital readmissions. The target of this paper is to analyze whether these diabetes diagnosis variables hold association in forecasting hospital readmission using decision tree classifiers, logistic regression, and discriminant analysis algorithms.

With machine learning, we hope to alleviate hospital operation costs by choosing which combination of measures of diabetes diagnosis will result in readmission. The length of the dataset (a decade) will provide us with enough observations to accurately predict readmission results. Ultimately, the goal of this analysis is to help companies within the healthcare industry to better allocate resources to strategically reduce the immense costs of hospital readmissions currently compromising company efficiency.

II. About the Dataset

Link to data the source.

Descriptions of "Hospital Readmission" Variables

The dataset titled, *Predicting Hospital Readmissions* has been retrieved from the data science online platform *Kaggle*, containing historical data for 10 years of patient information. This data consists of both numerical and categorical variables, as well as calculated variables.

Numerical Variables…

"time_in_hospital" - days (from 1 to 14)
"n_procedures" - number of procedures performed during the hospital stay
"n_lab_procedures" - number of laboratory procedures performed during the hospital stay
"n_medications" - number of medications administered during the hospital stay
"n_outpatient" - number of outpatient visits in the year before a hospital stay
"n_inpatient" - number of inpatient visits in the year before the hospital stay
"n_emergency" - number of visits to the emergency room in the year before the hospital stay

Categorical Variables…

"age" - age bracket of the patient
"medical_specialty" - the specialty of the admitting physician

"diag_1" - primary diagnosis (Circulatory, Respiratory, Digestive, etc.)

"diag_2" - secondary diagnosis

"diag_3" - additional secondary diagnosis

"glucose_test" - whether the glucose serum came out as high (> 200), normal, or not performed

"A1Ctest" - whether the A1C level of the patient came out as high (> 7%), normal, or not performed

"change" - whether there was a change in the diabetes medication ('yes' or 'no')

"diabetes_med" - whether a diabetes medication was prescribed ('yes' or 'no')

"readmitted" - if the patient was readmitted at the hospital ('yes' or 'no')

Calculated Variables…

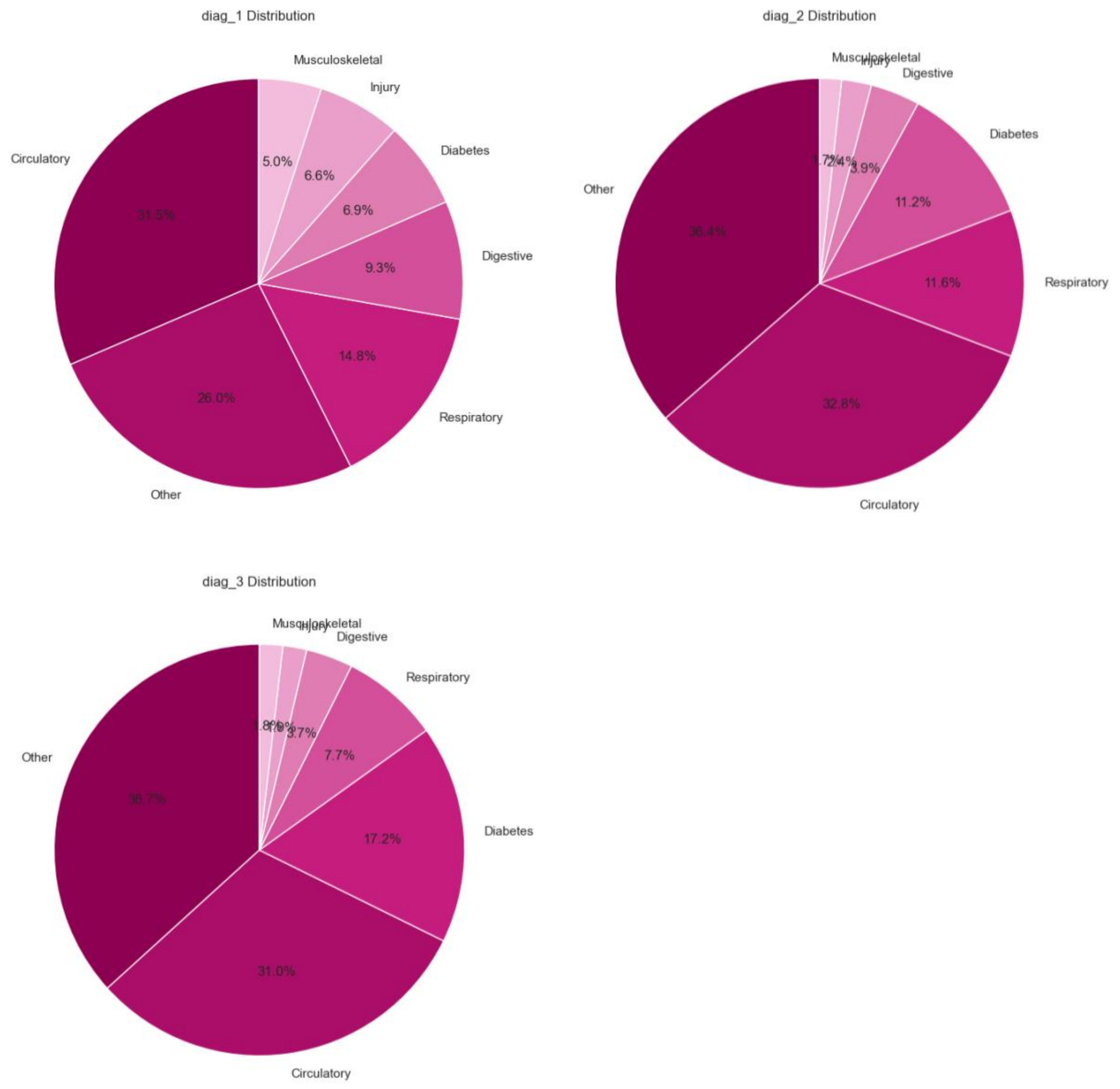
"average_time_in_hospital" - average length of hospital stay (days)

"average_#_of _procedures" - average number of procedures

Distribution of Diagnosis

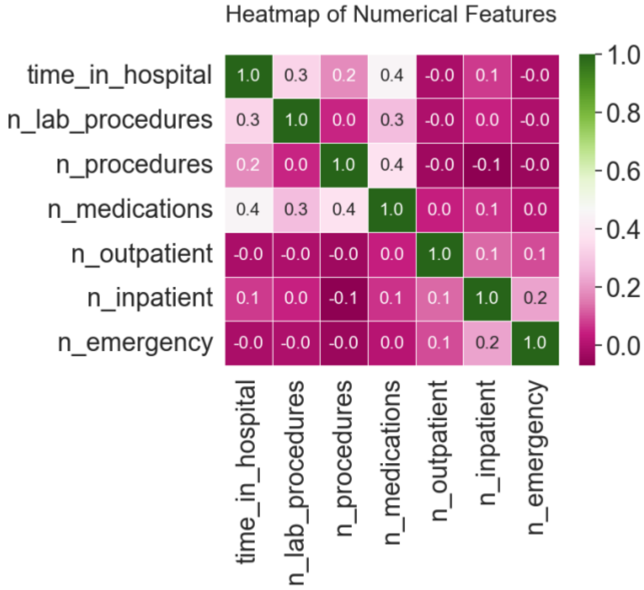
The pie charts below show the distribution around the diagnosis results and its subcategories (type of diagnosis) that are inclusive of all variables in the data. This allows us to visualize a pattern in high percentages of diagnosis that are “Circulatory” and “Other”. The decrease in percentage of Circulatory from diagnosis 1 to diagnosis 3 seems to migrate to Other. This insists that patients are having an initial circulatory diagnosis, later being diagnosed with something else by their third diagnosis.

Figure 1: Diagnosis Distribution



The heatmap in *Figure 2* is a matrix of the numerical features, showing the correlation that the number of medications prescribed is most correlated with the features number of procedures, number of lab procedures and the length of hospital stay.

Figure 2: Heatmap of Numerical Variables



III. Cleansing, Pre-Processing & Transformation Overview

Dropping Bad & Missing Data

The column for medical specialty observations was dropped for this analysis, due to there being over 90% of the data missing for this variable. We decided this would be a poor variable for our model and that it would be best to exclude this feature. Once this column was eliminated, there were few missing points in the diagnosis 3 column because most were diagnosed in diagnosis' 1 or 2.

Even though this data needed to be removed from the data frames being analyses, the immense amount of data at hand is still more than enough to run these predictions. A table displaying the included features with the number of missing values within those features was created to confirm that there are no other features to be concerned of in terms of missing data.

Creating New Features

Two new variables were created to calculate the averages for this model. The first variable calculates the average number of days patients were admitted, and the second variable adds the average number of procedures for each row of data; "average_time_in_hospital" and "average_#_of _procedures". A variable was created for hospital stay to have the same average for each row of data, as this average is the same across the entire data set. Though, the variable created for average number of procedures adds the average number of procedures for each row of data, this includes the total number of procedures and divides it by two for each row. In contrast to average hospital stay, average number of procedures is different for each row, taking into consideration the two types of procedures that are being totaled.

Encoding Categorical Variables

To be includable as input features in this research, the necessary categorical variables were first encoded using both regular and one-hot encoding to create binary vectors. The categorical variables that are regularly encoded are both the glucose and A1C tests, as well as the readmitted variable (y, or the dependent variable). Whereas variables patient age, diagnosis 1, diagnosis 2, diagnosis 3, if there were changes in medication, and if diabetes medication was prescribed, have all been encoded using one-hot encoding. All other variables did not need to be encoded to fit our model.

The categorical variables encoded using regular encoding require that their data is given a hierarchy in terms of how important that result is to the objective of our model, such as diagnostic test results and readmissions. Having a high glucose or A1C test is most important for our model, having a normal glucose serum or A1C level is the next important, and not having either test performed is least important. Here, data where patients were readmitted is more relevant than those who were not, assigning a hierarchy to this data (yes, before no), regardless of this being a binary vector.

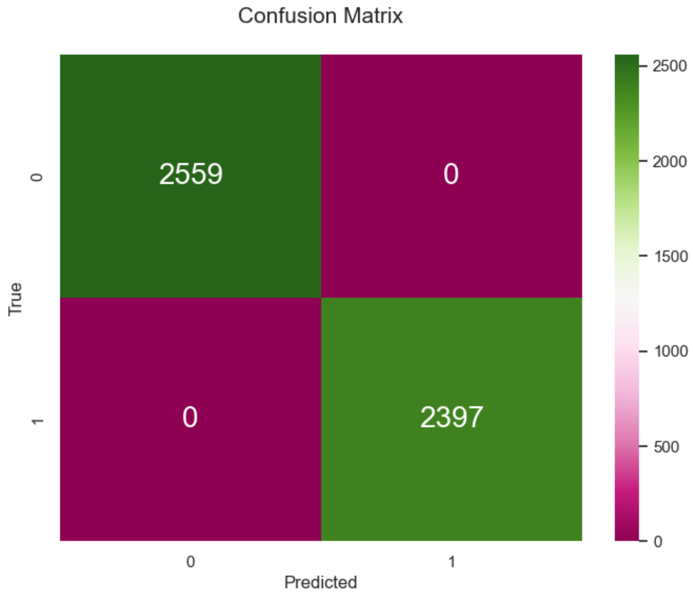
In one-hot encoding, each categorical variable is given a column, and the data elements are either hot or cold (0 or 1), a binary vector; a single variable is one and the rest are zero. This helps us avoid imposing ordinal relationships on categorical variables when using it in our model evaluation. Rather, where the categorical variables don't have a meaningful numerical relationship (like a yes/no relationship), they are given one.

IV. Technology & Analysis

Logistic Regression Classification

In this project, we focused on classification for the binary dependent variable “readmitted”. The first attempt at using a classification model to make predictions in our dependent variable started with a logistic regression classification method.

Figure 3: Logistic Regression Classification; Confusion Matrix



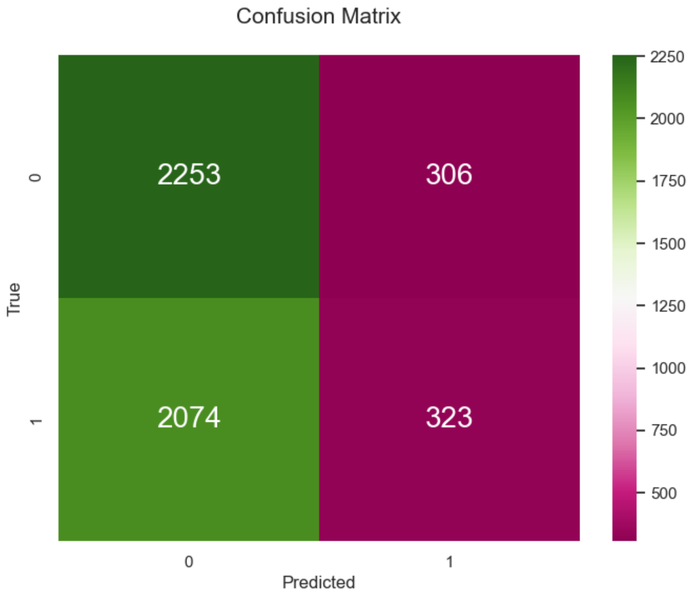
The confusion matrix in *Figure 3* visualizes that the true 0 and predicted 1 suggest that the data may be overfitted for the model or that there is a class imbalance here. This is an initial assumption of bias in this model.

After splitting the data into a training set and testing set, the model produced a perfect score (at 1.00) for all validations; accuracy, precision, recall, and F1 score. These results insist that this model is overfitted to the data, giving nearly perfect results when using the testing set against the training set. This overfitting can be due to the immense amount of data and variables that are similar, possessing multicollinearity. The results from this method incline that further classification is needed to develop a more valid model of prediction.

Decision Tree Classification

Due to overfitting of the data, a decision tree classifier was chosen to fine tune the model. As previously mentioned, the overfitting is believed to be because of how many variables are in the data set. By using decision tree classification, the top five features were selected; “diag_2_Diabetes”, “diag_2_Circulatory”, “diag_1_Respiratory”, “diag_1_Other”, and “time_in_hospital”. From here, we split the data again based on these top five features and ran a logistic regression. From our results, this seemed to fix the issue of overfitting the data, as seen in *Figure 4’s* confusion matrix.

Figure 4: Decision Tree Classification; Confusion Matrix



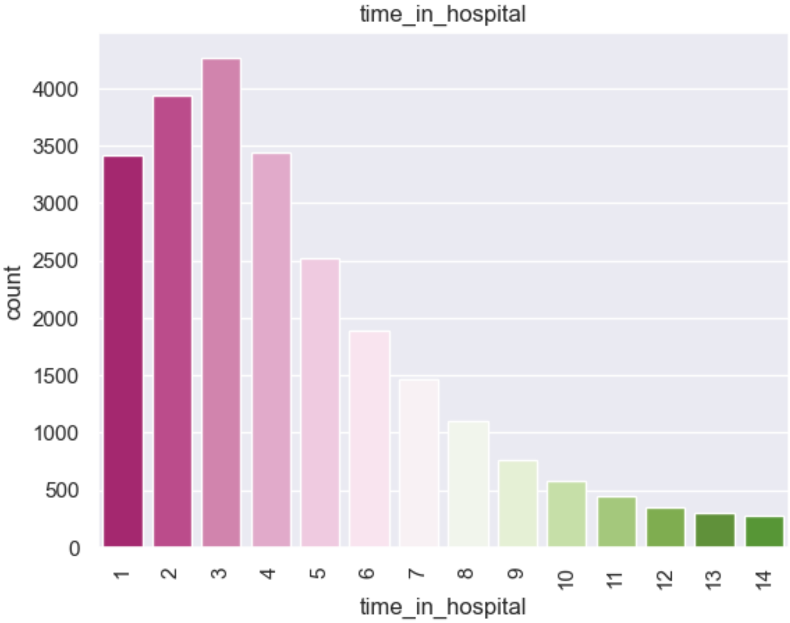
Consequently, the decision tree classification model produced a low accuracy score (at .52) and lower precision score (at .41), and an even lower recall score (at .13), insisting a lack of sensitivity and failing to recognize a large amount of true positive instances in the data. The F1 score (at 1.00) insists an imbalance in the data because of how contradictory this was to the recall score.

Descriptive Statistics

The only significant information that can be drawn from our descriptive statistics of the top five features of importance can be seen around the variable “time in hospital”. For hospital stay, the standard deviation (at 3.004) is close to the mean (at 4.465, but less than), suggesting little dispersion in the data and high levels of precision. All four other variables contrast these results, with a higher standard deviation than mean.

This statistical analysis insists that on average 25% of patients stay for two days, 50% stay for four day, and 75% stay for 6 days (considering a maximum hospital stay of 14 days and a minimum of one, objectively). When looking at hospital stay compared to other features of importance via the bar chart in *Figure 5*, we can visualize that more people are leaving after at least a 3-day stay than the latter.

Figure 5: Bar chart of Patients vs. Length of Hospital Stay

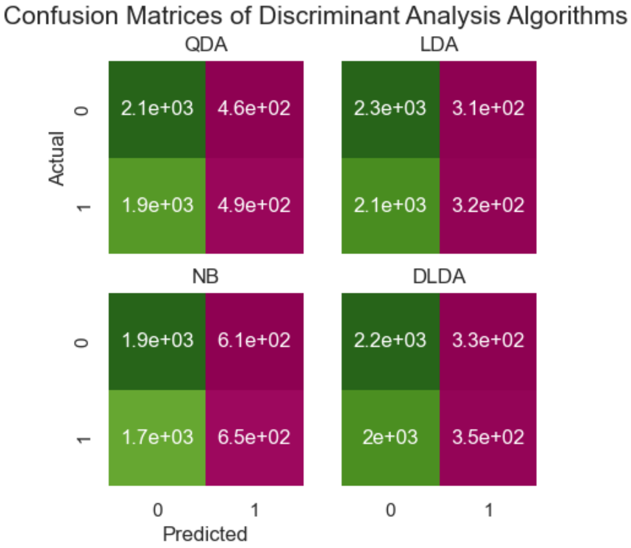


These statistical results lead us to predict that the number of days you spend in the hospital has an impact or relationship with readmission rate. The other independent variables also increase in correlation overall, due to this relationship. As the time variable increases, our other features of importance such as number procedures, lab procedures and medications objectively increase as well. The bar chart shows that most of the patients spend about 2-4 days in the hospital for the course of their treatments. We can expect that those who stay longer most likely receive more treatments or need more recovery time, either giving more time for healing, or possible further complications.

Discriminant Analysis Algorithm

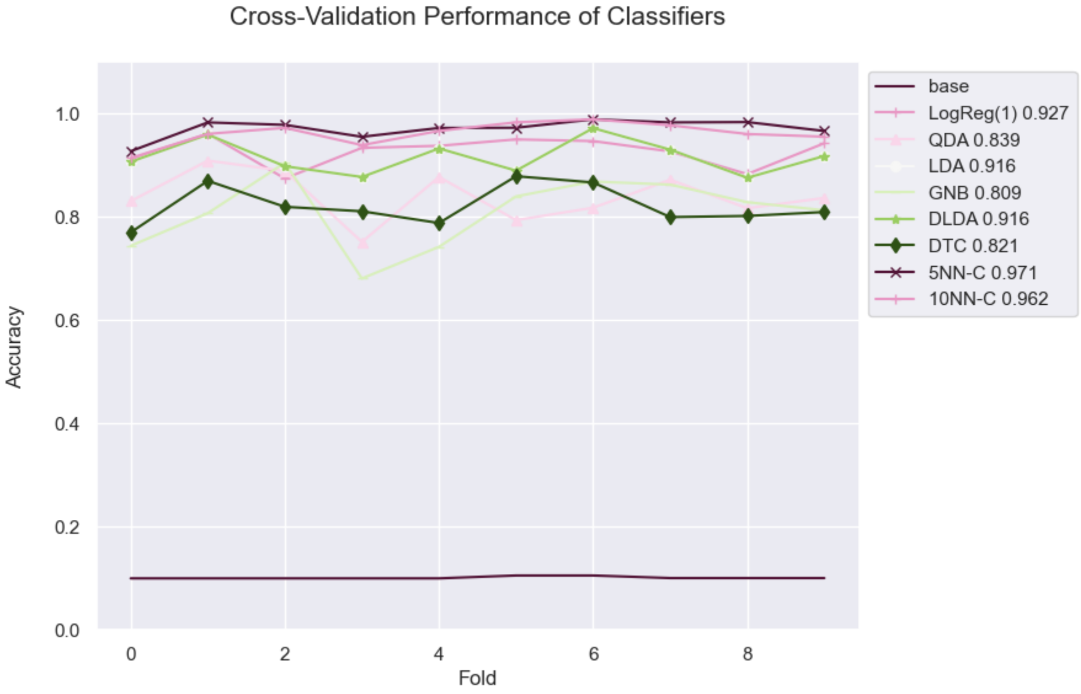
A discriminant analysis algorithm was run once results improved, attempting to incorporate classification methods towards the objective of this predictive model. The discriminant analysis consisted of quadratic, linear, and diagonal linear discriminant analysis as well as naive bayes algorithms.

Figure 6: Discriminant Analysis Algorithm Confusion Matrix



The computed results include an increase in both precision and recall scores. There was a greater improvement in precision, or an improvement in the model's ability to recognize instances predicted to be positive that were actually positive cases (true positives) in the data.

Figure 7: Cross-validation Curve; Performance of Classifiers



After building these methods, the model runs a 10-fold cross-validation (CV) across these classification methods based on its evaluation method, accuracy. When looking at *Figure 7*, comparing logistic regression to discriminant analysis; logistic regression had a higher average CV score (at .927) compared to both the QDA score (at .850) and LDA score (at .916).

VI. Conclusions

Although the initial attempt at using classification via logistic regression to assess risk of hospital readmission in diabetes patients was unsuccessful due to overfitting, the further attempts via using decision tree classification and the discriminant analysis algorithm have given more promising results as far as model reliability. Using the cross-validation technique we can conclude that the final models are more valuable in terms of real-world application towards this discipline of research.

This predictive approach to identifying at-risk patients is a personalized healthcare solution that integrates machine learning algorithms with patient care delivery. In future development of this intersection in disciplines, more research and refinement should be applied to this and similar models, in efforts of improving model accuracy, interpretability, as well as scalability of data. Our original model with all the variables, although accurate and precise, the model was most likely very overfitted. When running additional models, one can see that there would need to have further revisions that can continue to improve the quality and accuracy of the model.

Business Insights

The statistical research models in this paper aim to produce insight for companies within the healthcare industry by forecasting the likelihood of hospital readmissions in diabetes patients. These predictions can be used to make data driven decisions that influence patient outcomes by improving their ability to identify patients that are at a higher risk of readmission, as well as reduce the financial impact of hospital readmissions. The factors influencing hospital readmissions in this model include the duration of hospital stay, the frequency of medical procedures, and the diagnostic test results.

Companies within the healthcare industry can use this model to make predictions that help officials to make decisions that increase the efficiency of operations within their facilities. This may include risk driven efforts towards adjusting treatment mapping for patient readmission risk to decrease readmission rates. These efforts can improve overall operational efficiency of treatment by decreasing the length of hospital stays, or the financial ramifications behind hospital discharge.

VII. Bibliography

Sources

DD. “Predicting Hospital Readmissions.” *Kaggle*, 10 Mar. 2023, www.kaggle.com/datasets/dubradave/hospital-readmissions?resource=download.