Readmittance to Hospitals Among Patients with Diabetes

Richard Kim
Southern Methodist University

Joaquin Dominguez
Southern Methodist University

September 19, 2022

I. Introduction

Diabetes is a chronic disease that affects people of all ages across the globe. The disease is categorized into different forms of diabetes, referred to as Type I, Type II, and gestational diabetes, all of which involve the body's inability either to produce or effectively use insulin.[1] In the cases where patients require hospitalization, diabetes is associated with longer inpatient stays and higher rates of mortality and readmittance. Not only does this imply a clear impact on the lives and well-being of the patients, but research has shown an overall financial impact in medical expenses as a result of these factors.[2]

Exploring the causes of these factors may then be beneficial for both the patients and the hospitals, investigating potential preventive measures for those that are at risk of these factors. Machine learning techniques are useful tools for prediction of outcome measures or classification of independent observations. On the current topic of inpatient diabetes patients, we focus on predicting readmittance. By exploring the explanatory factors, we may gain insight into the key features that contribute to readmittance among diabetes patients.

We looked at over 40 variables to create a model that could classify whether a patient was readmitted after 30 days, readmitted within 30 days, or not readmitted. We used a logistic regression to predict readmittance of patients and to determine which features are most important in making this prediction.

II. Methods

i. Data Preprocessing

The diabetes dataset consisted of information on 101,766 patients with diabetes, 49 features, and the response variable, 'readmitted'. Categorical features that were binary in the data, like gender, or could be represented as numeric values, like age, were converted into ordinal variables. Any features that had a single value and provided no variability were dropped. Additional steps were taken for 8 features that contained missing values. For features 'gender', 'weight', 'diag_1', 'diag_2', and 'diag_3', patients with missing values were removed from the data because the missing values for each feature comprised less than 1% of the total data set. There were no duplicates in the data. The final training data frame comprised 80,192 rows and 47 columns, while the test data frame comprised 20,049 rows.

ii. Exploratory Data Analysis

ii.1 Correlation

Correlations allow us to form a picture of the patterns and dynamics between the variables. We first looked at correlation of features in the following forms:

- Correlation to target variable ('readmitted')

[1] Diabetes Overview. ADA., 2022

[2] Hussain, Alkharaiji, and Idris 2020

- Mutual correlation of non-target variables
- Correlation to variables with missing values for imputation

Identifying those variables that had a high correlation with the target variable provides a safeguard during feature selection. Since most of the variables in the data-set were categorical variables, correlation scoring methods varied. In examining correlation to the target variable, we found considerably weak relationships, with the highest scores being 0.23 and 0.13 for 'number_inpatient' and 'diag_1', respectively. This did not bode well for performance of classification models to come. Next, in addressing multi-collinearity, we looked at the correlation scores of all non-target variable pairs and set a threshold for high multi-collinearity as >=0.7, such that a variable-pair reflecting such score would be removed from the data-set. In this case, no variable-pairs contained a score higher than 0.7, with the highest being 0.64 for 'insulin' and 'changeYes0'.

Following, in the interest of addressing missing values via varying imputation methods, we wanted to check the highest correlative pairs in those variables that contained missing values ('race', 'payer_code', and 'medical_specialty') as a preliminary step to imputation via regression. In line with previously mentioned correlation trends, we found weak relationships between these variables and their highest-correlated-pair, with 'payer_code'-'encounter_id' as the highest pair with 0.55.

ii.2 Assumptions of Linear Regression

Binary/Multiclass
Response variable is multi-class, adapted for in model argument.

Independence
We may assume independence for all instances.

Multicollinearity
Multicollinearity was addressed in the correlation stage of the EDA; no multicollinearity found.

Linearity
In all variables, there exists a linear relationship between the independent variable, x, and the logit of the dependent variable, y.

Sample Size
With the training set consisting of >80,000 rows, sample size is sufficiently large.

iii. Imputation

There were 3 features 'race', 'payer_code', and 'medical_specialty' that had missing values. These missing values were imputed after splitting the full data into an 80/20 training and test data sets. All features with missing values were categorical. which were imputed using two methods.

The first method imputed values based on the mode of each category. The most common value for each categorical feature was used to replace all the missing values. While this is a simpler method of imputation, a more complex method may not be necessary depending on the data.

In the second method, we imputed the values using logistic regression, predicting the missing values based on the other variables. With this approach, we created a logistic regression model for each categorical feature with missing values. Each model included all other features as explanatory variables, except features with missing values and the response variable. This method incorporates other available information about each patient to impute the missing data.

iv. Modeling

Given the goal of this study to identify predictors of readmittance, it is crucial that the model is able to provide information on feature importance. To accomplish this, we perform a Yeo-Johnson power transformation to standardize all numerical variables prior to modeling.

All features in the final training data set were included in developing a classification model. A logistic regression model was trained to classify the readmittance status of diabetes patients. We utilized Scikit-Learn's *pipeline* class to set up the gridsearch workflow, logistic regression to train the model, gridsearchcv class to execute the search and refit the best performing model, and cross-validate class to ascertain the accuracy score from five-fold, shuffled internal cross validation. The response variable, *readmitted*, was a multiclass variable. Thus, our logistic regression model was adapted to account for this using the one versus one method.

The logistic regression model was trained for each method of imputation, mode and classification. We compare the performance metrics of the two models in our results.

III. Results

i. Accuracy, Precision, and Recall

The logistic regression model using mode imputation achieved an overall accuracy score of 58.63%, a precision score of 49.97%, and a recall score of 31.38%.

The accuracy score at nearly 60% is far above chance levels for a variable with three groups. The 49.97% precision score is also promising, as it implies that almost half of the positive predictions made by the model were correct. Unfortunately, the low recall score indicates that there are a lot of missed positive predictions. We can see from Table 1 that the model is having a hard time classifying readmittance within 30 days. This is likely in part due to the imbalance of data across groups. Most patients in the actual data were not readmitted. Consistent with that imbalance, the model frequently predicted that patients would not be readmitted, even when they were.

The next model using classification imputation achieved an accuracy score of 58.61%, a precision score of 49.69%, and a recall score of 31.83%.

Mode - Confusion Matrix			
	<30	>30	No
True <30	28	880	1305
True >30	34	2857	4100
True No	49	1926	8870

Table 1: *Confusion matrix of logistic regression performance after imputing missing values with the mode of each feature.*

While this model achieved a slightly higher recall, it also produced slightly worse accuracy and precision. However, the difference in scores are nearly negligible, with the largest difference at around 0.6%. The confusion matrix results after imputing with a logistic regression is shown in Table 2, where we see a very similar pattern to the results after imputing with the mode in Table 1. The minimal difference in performance metrics across the two models provides evidence that the choice in imputation has a relatively small impact on the logistic regression results.

Classification - Confusion Matrix			
	<30	>30	No
True <30	29	900	1284
True >30	42	2897	4052
True No	51	1976	8818

Table 2: *Confusion matrix of logistic regression performance with classification method for imputation.*

ii. Feature Importance

Feature importance was determined by ranking the variables by the absolute value of their coefficient values. By ranking with the absolute values, we focus on the magnitude of each feature's contribution rather than the directionality. Given that the data were standardized prior to modeling, the coefficients across our variables are on the same scale. This allows for direct comparisons, where we may interpret larger coefficients as indicative of greater contribution or importance. We summarize our findings on feature importance below.

The most important features identified by our logistic regression model using mode imputation were, in order:

- *glimepiride-pioglitazone*
- *changeYes0*
- *insulin*
- *glyburide-metformin*
- *miglitol*

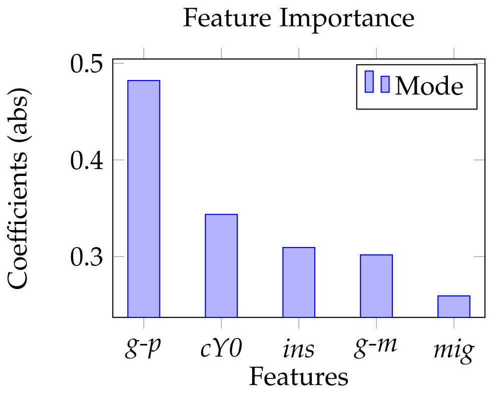

Figure 1: *Feature Importance (Mode)*

Similarly, the most important features identified by the other model using classification imputation were, in order:

- *glimepiride-pioglitazone*
- *changeYes0*
- *glyburide-metformin*
- *insulin*
- *metformin-pioglitazone*

Not only did both models identify 4 of the same features in their top 5, the first and second most important features were the same across both models. Figures 1 and 2 display the top 5 features from the mode imputation and classification imputation models, respectively. Both figures compare feature importance by visualizing the magnitude of the coefficients.

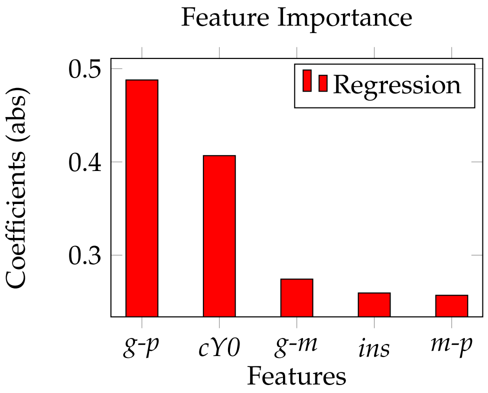

Figure 2: *Feature Importance (Regression)*

IV. Conclusion

The logistic regression model using the mode for imputation, as opposed to applying a classification method for imputation, produced a higher accuracy and precision but a lower recall, all by a very small margin. Given that imputing with the mode is simpler and computationally less expensive, we proceed with results from the model that uses the mode.

Based on the results from this model, we conclude that the most important features in predicting readmittance were *glimepiride-pioglitazone*, *insulin*, *changeYes0*, *glyburide-metformin*, and *miglitol*. Above all other features, these variables had the highest predictive power for readmittance status.

Most of the features selected above are medications, meaning risk of readmittance may be dependent upon what medications the patient is taking. Our results also show that knowing whether the patient changed medications has comparably high predictive power.

In this study, we aimed to predict whether patients will be readmitted and, if so, if they will be readmitted within or after 30 days. Unfortunately, our final model had relatively low recall, which is arguably the most important metric for the goal of this study. A low recall score suggests that this model would fail to identify a lot of patients who end up being readmitted later on. Used in practice, this model may still be helpful to predict readmittance, allowing hospitals to take preventive measures for some of their patients' wellbeing and reduce medical expenditure. However, wanting performance metrics, especially in recall, suggest that further modeling is warranted.

i. Code

Please refer to the attached Python notebook.

References

[Diabetes Overview. ADA., 2022] Diabetes Overview. ADA. (2022). Retrieved September 18, 2022, from https://diabetes.org/diabetes

[Hussain, Alkharaiji, and Idris 2020] Hussain, Z., Alkharaiji, M., and Idris, I. (2020). Evaluating the effect of inpatient diabetes education on length of stay, readmission rates and mortality rates: A systematic review. *British Journal of Diabetes*, 20(2), 96–103. https://doi.org/10.15277/bjd.2020.256