Health Analytics: Logistic Regression to Guide Treatment for Diabetes Patients

Leyla Ahmadli
City, University of London
Department of Computer Science
leyla.ahmadli@city.ac.uk

Abstract - Diabetes, which affects millions of lives worldwide, is a chronic disease that requires constant monitoring and clinical supervision to ensure life satisfaction of affected patients. Multitude of tests must be carried out to effectively oversee the condition, different medications may be prescribed at various stages and when outpatient care is insufficient patients must be hospitalized promptly. In this work we develop logistic regression models to predict medication change and readmission into inpatient care from demographic and clinical visit data and medical test results. We report that the logistic regression model can successfully identify true positives of the test data with 77% accuracy and the model has highest sensitivity to the results of select medical tests. The results of this model can be used by clinicians to decide whether certain tests should be recommended to guide the diagnoses and medication prescriptions.

Keywords – diabetes, logistic regression, predictive analytics

I. Introduction

Diabetes is one of the common chronic diseases[1] that affect 34.2 million people in United States as of 2018[2]. There is not a single medical condition for diabetes, the main types of diabetes are type 1 and type 2 diabetes. Type 1 diabetes, also called as juvenile diabetes, occurs when the immune system attacks the insulin producing cells in the pancreas. Type 2 diabetes is often associated with lifestyle choices like body mass index or inactivity[3]. Various medications or combinations of them can be used to treat different forms and types of diabetes, and if the patients are not careful or if the treatment is not correctly selected, hospitalization and inpatient care may be required to improve the healthy life of the affected patients[4].

To reduce the workload required for diagnoses and to improve the reliability of diagnostics, nowadays statistics and data science are applied at an increasing pace[5]. The statistical methods can be used to give diagnoses based on the known risk factors observed from the previously diagnosed patients and these methods can be used to give helpful information about patients' health. In this work, we develop a logistic regression model to study the record of diabetes patients from United States to see how well the records of age, gender, history of medication, hospitalization, and test results can help forecast readmission requirement and necessity to alter the course of medication. From the result of logistic regression, we look at key attributes which explain requirements for medication change and readmission.

II. Analytical Questions

We are interested in seeing whether it is possible to predict the need for readmission or for medication change based on the data available to the hospitals. To reduce the potential of complications arising from diabetes, it is necessary to expeditiously detect whether the medications currently prescribed are helping the patient's treatment and change to a different course of medications if they are not. If the information available to the clinicians indicates a preference for hospitalization rather than outpatient treatment, doing so without delay may be important in saving and prolonging the patient's life.

If a model to predict a need for readmission or a need for medication change is successful, the clinicians can use the results of such a model trained on the population data as an aid to the patient data available to them when making recommendations for the continuation of the course of treatment. Alternatively, the model can be used as a preprocessing step to filter out those who do not require readmission or those who should continue the medications currently recommended.

Besides the success of the model, we are looking to find which features are most significant in helping to predict the need for medication change or readmission. Knowing the foremost contributing factors to the model can help guide the physicians e.g., in determining whether certain medical tests should be carried out more often than they currently are.

To state directly, the questions we are looking to answer in this research are

- Is it possible to predict the need for hospitalization or change of medication with a logistic regression model?
- How well does the model predict the target variable when applied to the test data?
- What are the most important explaining variables which should be included in such a model?

III. Data

The dataset was taken from UCI Machine Learning Repository, and it is publicly available data. The data has around 100000 records of patient from United States gathered between the years 1999 – 2008. It is an aggregate data contributed by physicians of different specializations from 130 different hospitals. The data has 50 columns, some of which are categorical, and some are numerical values. It is an unprocessed dataset and needs cleaning. The available attributes like age, gender and race are helpful at visualization how the target values, medication change, and readmission differ among demographic groups. There is a wealth of other information available in the chosen dataset as well. We have access to the results of 25 different tests done on the patients; given we are looking to explain the need for medication change, the results of tests are expected to be highly significant in giving information about altering the course of treatment. We also have access to the number of times different patients were hospitalized, how long they were hospitalized, how many medical procedures they had and whether they visited the hospitals for an emergency procedure, these variables are expected to be relevant to explain the need for readmission for an inpatient care.

IV. Analysis

A. Data preparation

Dataset has 50 columns, however not all of them carry valuable information for our analysis. The identification columns, columns with unclear information and columns with medical specialty are removed. The 'weight' column is also dropped because more than 96% of data in this column is missing. For the 'race' column missing values are replaced with 'Others' category as this category is already present in 'race' column. The entries with unknown values of gender attribute are removed from dataset; there are only 3 rows that had missing gender information and filtering them out does not affect the sample size significantly.

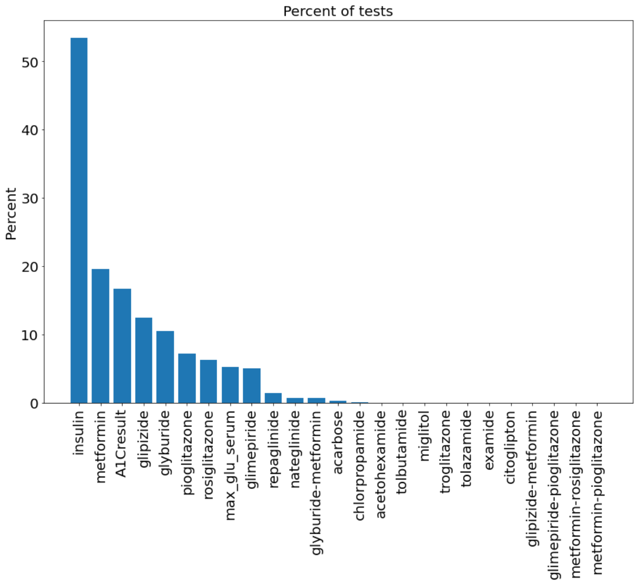

Figure 1. Percentage of tests with results

In figure 1 we see percentages of population for whom various tests were recommended. There are some tests which were never carried out, and some tests were recommended on a very small subset of the presenting patients. On the other hand, some tests were carried out on greater than 5% of the patients and insulin test is recommended for more than 50% of the patients. To select the test results as predictor variables in our models, we select tests which are recommended for at least 5% of the presenting patients. The rest of the test result columns are dropped from the model. The remaining test result variables have high variance which should help improve the performance of the model.

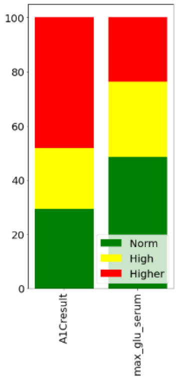

Figure 2.

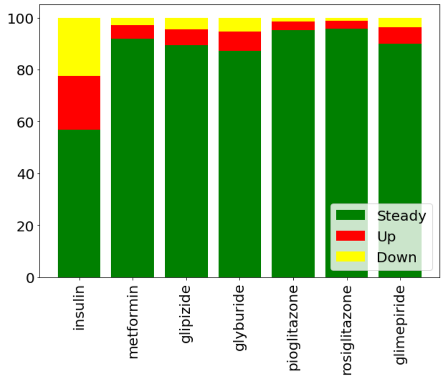

Figure 3.

From the stacked bar charts in figure 2 and 3 we see results of tests which have been carried out on at least 5% of the patients in the study. Both figures have ordered results. In the raw dataset, each of these tests are categorized to have 3 outcomes. We divide the test columns into two groups; in the first group of tests the desired medical outcome (norm) is the lowest value, while in the second group of tests the desired medical outcome (steady state) is between the other two categories.

B. Data derivation

Because our dataset contains categorical values as seen in figure 2 and 3, we replace them with numeric values. Categories like No, Steady, Up, Down are ordered so they can be replaced with numbers indicating their level. We also add an extra binary column per each test column which describes whether the test has been applied to the patient. Next, gender, change of medication and taking diabetes medication columns are replaced with binary values. Readmitted column contains 3 distinct categories (as NO, <30, >30), for our logistic regression we transform this column to binary values, 1 indicating if the patient was readmitted and 0 indicating not readmitted. Values of age attribute were presented as interval of 10 years for each record, and it was object rather than numeric value, we replace it with average of each 10-year interval.

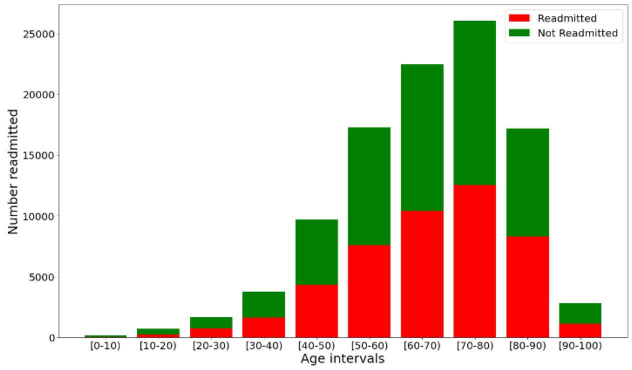

Figure 4. Number of readmissions by age

Figure 4 describes readmission requirements of patients by their age intervals. We see that higher age groups constitute a greater fraction of diabetes patients. However, the likelihood of readmission does not show a clear difference for different age groups.

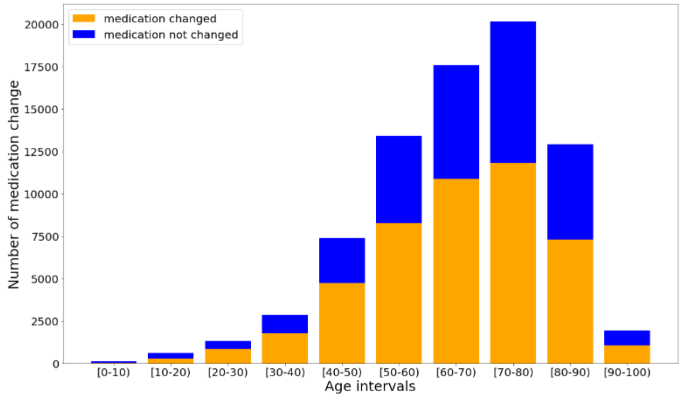

Figure 5. Number of medication change by age

Figure 5 depicts medication change for patients in different age groups. For this plot, only the patients already prescribed medication treatment are selected. There is no clear medication change with age, but we see that slightly more than half of the patients were recommended different medications.

From figure 6 it is interpreted as Caucasians constitute the highest fraction of patients diagnosed with diabetes which can be skewed since Caucasians are also the largest racial group in the United States. Since the race cannot be transformed into a single ordered variable unambiguously, this attribute is not likely to help our analysis. So, we drop this feature as well.

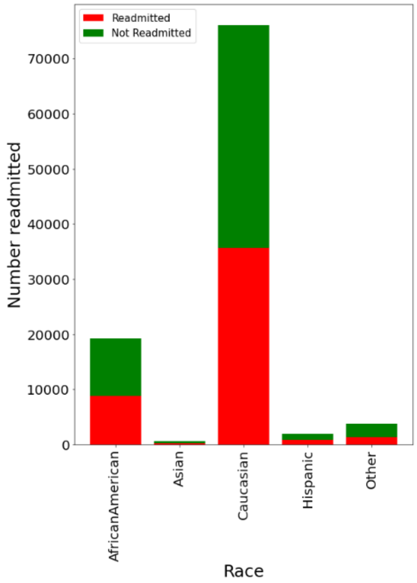

Figure 6. Number of readmissions by race

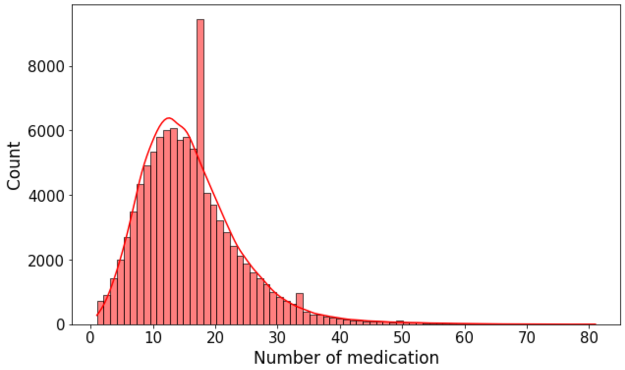

Figure 7. Histogram for number of medications

The feature ‘number of medications’ as presented in figure 7 has a left skewed distribution, which is characteristic of distributions like log normal distribution. To reduce the skew for this column, we originally tried to transform this column by taking the logarithm of values. However, when running the model with both transformed and untransformed ‘number of medications’ feature, we found that the model with untransformed ‘number of medications’ feature had higher accuracy and f1 score; hence we report the model results where this variable is untransformed.

C. Construction of model

We apply normalization to our data set to make all features vary between the range [0,1] which will make it easier for our model to better calculate relative influence of features when predicting the target values.

For this analysis logistic regression is used since our targets are binary classes. We first look at correlations between our features as can be seen in figure 8. Using this heatmap we remove features that are highly correlated before passing them into logistic regression. From the result of logistic regression, we have a post processing step where the features which are not statistically significant (defined as p value greater than 0.05) are dropped from the model and the regression is run a second time. This is to ensure that the model's performance is entirely due to the significant predictors.

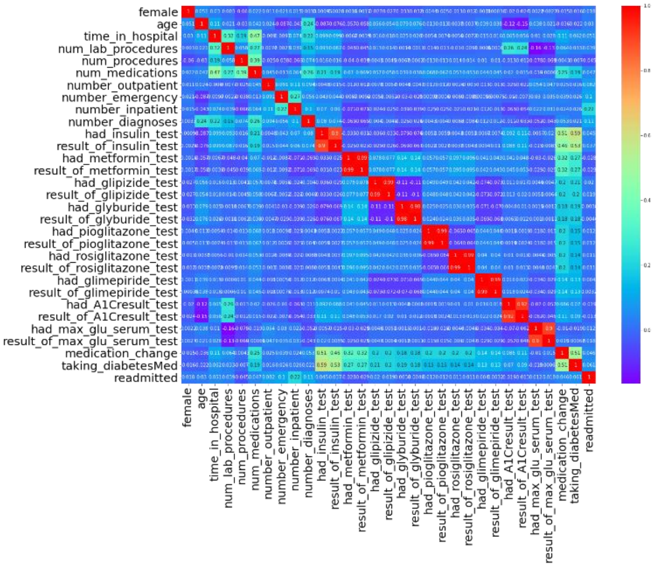

Figure 8. Heatmap of correlations between features

We also attempt another method where to reduce multicollinearity, PCA (Principal Component Analysis) is applied to all features before manually discarding highly correlated features. PCA method itself finds the principal components with highest variances which are uncorrelated by construction, this method both reduces the multicollinearity while maintaining the variance, hence explaining power, of the predictors.

Logistic regression	accuracy
Medication change	77%
Medication change PCA	75%
Readmitted	61%
Readmitted PCA	56%

Table 1.

D. *Validation of results*

Table 1 illustrates accuracy of models for unseen testing set. From the table we see that applying PCA to dataset does not perform as well as manually removing some of the highly correlated features, for both models. Compared to the manual selection of features, the drop in accuracy for the model with PCA was 2% lower when predicting medication change and 5% lower when predicting readmission. This can be because PCA takes only account of variance in features rather than their influence on each other.

PCA was surprisingly better at predicting minority class than logistic model without PCA, however, it was not as good at predicting the majority class.

V. Findings, Reflections, and Future Work

A. *Findings and reflections*

Figures 9 and 10 depict the ROC (Receiver Operating Characteristic) curves for the logistic regression models built to explain the needs for readmission and medication change. The ROC curves are obtained from benchmarking the explained variable in the test set against their prediction from the model and give an indication of how well the model can distinguish between true and false positives. The larger the area under the curve, the better the model is at distinguishing.

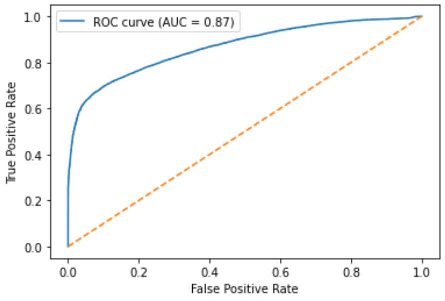

Figure 9. ROC curve of medication change

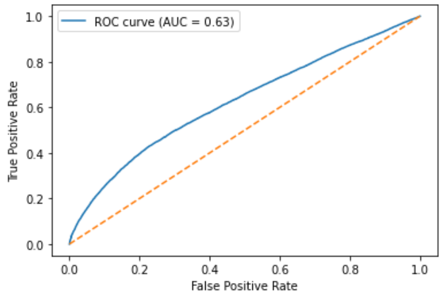

Figure 10. ROC curve of readmission

We clearly see that the model for medication change has better predictive power than that for the need for readmission. From the result of the model for readmission, the target variable has highest sensitivity to features like number of outpatients, emergency, and inpatient. We think that the readmission model can be improved if more data like general demographic, mortality, income data as well as body mass index (BMI) is included.

The medication change model showed that the features which were both statistically significant in the model and which the model is most sensitive to are attributes such as number of medications, having insulin test, results of metformin, glipizide, glyburide, pioglitazone, rosiglitazone, glimepiride tests and age. Compared to the other tests results of which are statistically significant to the model, we found that keeping the binary variable for insulin test which indicates whether the test is done or not is better for model operation than the outcome of the insulin test. This implies that if an insulin test is already recommended for the patient in the near past, recommending other tests and looking at their results is better for determining the treatment plan than relying on the result of the insulin test itself. This model would be useful for clinicians to decide which new tests to recommend for their patients if they are considering changing the course of medication. Helping the clinicians narrow down the potential tests to recommend for their patients will both expedite the diagnosis process and help utilize the testing resources in a more efficient way.

B. Future work

We see that the data currently available in this dataset is not sufficient to develop a logistic regression model to predict readmission. We would like to include data that would include an overall population of different age groups and genders. Looking at the pattern of how the fractions of diabetes patients in overall population change, as well as including the mortality data to account for the lower number of diabetes cases at the highest age groups is expected to improve the quality of the model. We also think that BMI and income will help improve the quality of the model; the latter can be a proxy as to whether the patient can financially afford an inpatient treatment.

The dataset in this analysis was recorded for the 1999-2008 period. The analysis should be performed again on newer data, since 2009 first to see how whether the model quality persists with the new data. Additionally, with greater use of technology and digital tools, the newer data is expected to have more information (greater number of potentially helpful columns) and lower probability of large gaps in the rows and columns of the dataset.

WORD COUNTS

Section Name	Word counts	Maximum word limit
Abstract	140	150
Introduction	254	300
Analytical questions	294	300
Data (Materials)	203	300
Analysis	978	1000
Findings, reflections, and further work	524	600

[1] J. Silva, E. Souza, A. G. Echazú Böschemeier, C. Costa, H. S. Bezerra, & E. Feitosa, "Diagnosis of diabetes mellitus and living with a chronic condition: participatory study," BMC *public health*, 2018. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5989432/

[2] Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2020. Available: https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

[3] "Symptoms & Causes of Diabetes". Available: https://www.niddk.nih.gov/health-information/diabetes/overview/symptoms-causes

[4] J. J. Marín-Peñalver, I. Martín-Timón, C. Sevillano-Collantes, & F. J. Del Cañizo-Gómez, "Update on the treatment of type 2 diabetes mellitus," *World journal of diabetes* vol. 7,17 (2016). Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5027002/

[5] W. Raghupathi, & V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health information science and systems, 2*, 3, (2014). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/