Predicting Hospital Readmission Rates of Diabetic Patients

Yi Nei Charlene Lau

University of Rochester

Table of Contents

Introduction:

Diabetes is a chronic condition affecting more than millions of people in the United States. Patients diagnosed with diabetes have a high chance of being readmitted into hospital. HbA1c is also known as glycated hemoglobin and is widely used as a measurement for average blood sugar levels over a significant period of time. A higher HbA1c indicates a higher risk of diabetes-related health problems. In the given dataset, it is shown in the A1Cresult column. Diabetic patients usually have HbA1c at over 6.5%. The HbA1c is only one measure that could have an affect on hospital readmission rates. Patient identity and hospital logistics also contain data which may be useful for readmission rate analysis. The readmissions of the patients come with cost concerns for hospital administration. Improved predictions of whether patients will be readmitted into hospital will provide the opportunity to reduce cost burdens for hospitals.

Related Work:

The research paper titled "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records" was published by BioMed Reasearch International in 2014. The dataset used was from the UCI Machine Learning Repository and is available online. The data represents 10 years (from 1999-2008) of data captured across 130 hospitals in the United States. The logistic regression model was fitted for the data. It was concluded that the decision for HbA1c to be measured for a patient was a crucial factor for readmission rates.

Methodology:

The goal of the project is to predict the impact on readmission rates for patients within 30 days. The aim is to predict binary class labels after constructing a model based on the training set in order to classify new data and its class labels: readmitted within 30 days or not readmitted within 30 days. The data used for this project, like the research paper, includes 101766 records with 51 attributes.

To do preprocessing of the data, the data will be cleaned and features with a high percentage of missing values will be removed. More than 50% of missing data will be considered a high percentage. The percentage of missing data in the dataset is as follows: Weight (97%), Payer code (40%), Medical specialty (49%). Features with dominant categories will be dealt with as they cause the class-imbalance problem which is commonly seen for healthcare datasets where a large proportion of results are negative with only a few positives.

Feature selection will be performed on the data to select the top features for fitting the classification models. Four classifier models will be fitted to the data. These include Logistic regression, Random Forest, AdaBoost and Support Vector Machines (with the linear and RBF kernel). The performance of the models will be compared for the best accuracy.

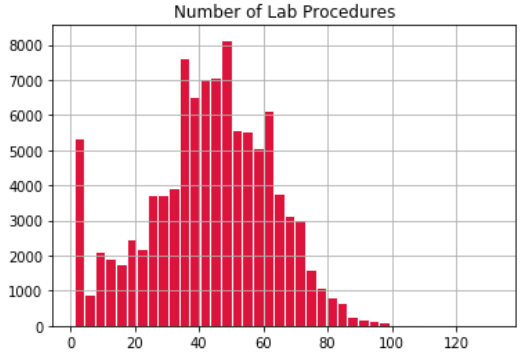
Experiment:

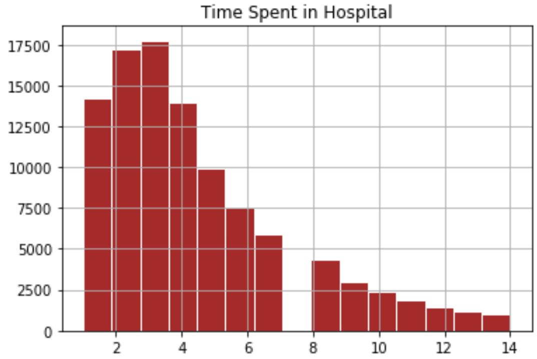
1) Data Preprocessing:

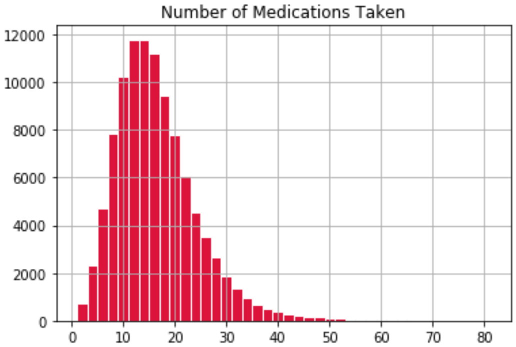
The dataset consists of 51 attributes which can be split into three general groups: Patient Information, Hospital Logistics and Medication Status. Both nominal and numerical attributes are present in the database. Nominal attributes include Race, Gender, Age, Admission type and Discharge disposition. Numeric data includes Number of procedures and Number of inpatient visits. The attribute Readmitted was set as the response variable. It consisted of three outcomes: <30, >30 and NO which were regrouped into two groups: <30 and not <30, resulting in a binary classification problem.

The percentage of missing data in the dataset is as follows: Weight (97%), Payer code (40%), Medical specialty (49%). The decision made was to remove these features. Features that were used as identifying codes, EncounterID and PatientNbr were removed. Features that were highly imbalanced were removed. These included the medications and diagnoses. After data cleaning, 18 features remained.

2) Data Visualization:



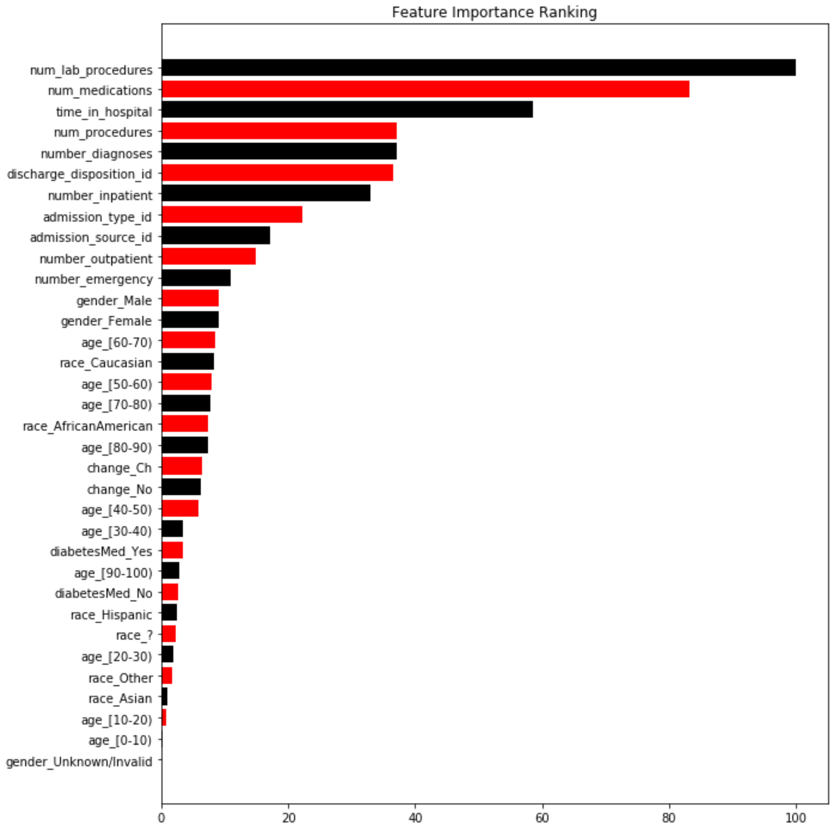




The frequency distributions of numerical attributes Number of Lab Procedures, Time Spent in Hospital and Number of medications taken were plotted for a distribution of the data. There is no trivial distribution that can be identified.

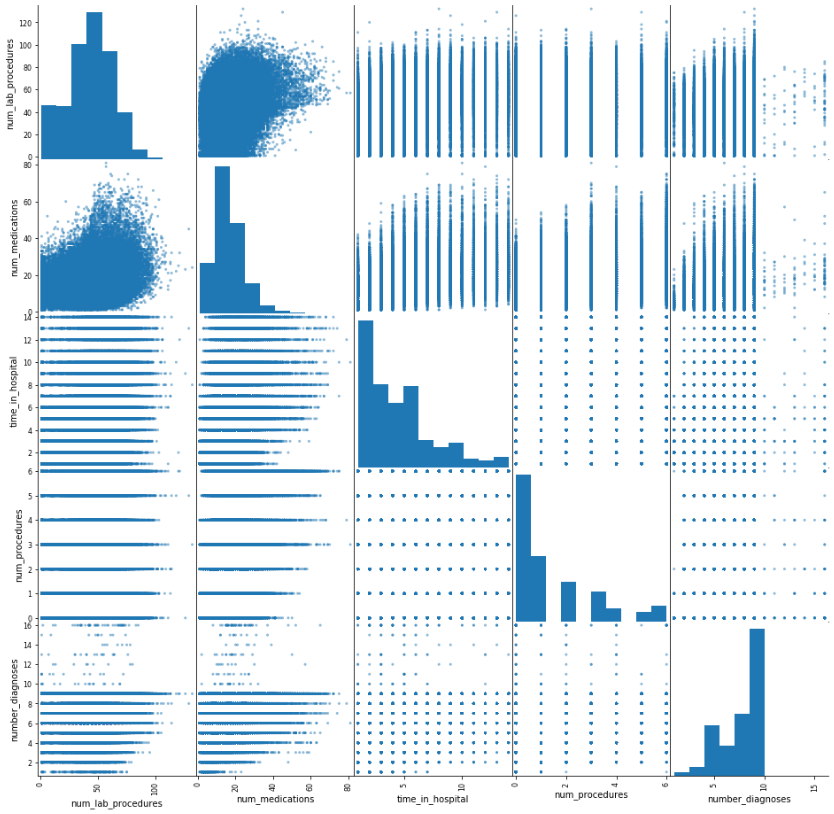
3) Feature selection:

Feature ranking was performed on the remaining 18 variables while nominal categories such as Age, Gender and Race were converted into separate features. The result was 34 attributes. The Random Forest Classifier was used for feature ranking and features were ranked based on relative importance.



Summary of Most Important Features:
Top feature: Number of Lab Procedures
Top 2 features: Number of Lab Procedures, Number of Medications
Top 3 features: Number of Lab Procedures, Number of Medications, Time in Hospital
Top 4 features: Number of Lab Procedures, Number of Medications, Time in Hospital, Number of procedures
Top 5 features: Number of Lab Procedures, Number of Medications, Time in Hospital, Number of procedures, Number of Diagnoses

The scatterplot matrix was plotted to look for multicollinearity between the top five features. There was no evidence of collinearity and the features are independent of one another.



Results:

Table showing Test Accuracies for Top feature combinations:

	Top feature	Top 2 features	Top 3 features	Top 4 features	Top 5 features
Logistic Regression	0.887	0.887	0.887	0.887	0.885
Random Forest	0.887	0.884	0.868	0.857	0.871
AdaBoost	0.887	0.887	0.887	0.887	0.885
Support Vector Machines	0.887	0.887	0.886	0.886	0.886

The highest test accuracy was 88.7% which was achieved by fitting the models on the most important feature: Number of lab procedures. The SVM used was with the linear kernel with C=1. All four classifiers returned the same accuracy for the Top feature. The accuracy decreased slightly with the addition of more features for all models except Random Forest. The Random Forest model did not show a general trend with the addition of more features.

Classification report for Top 5 features:

Classifier	Precision	Recall	F1 score	Test Accuracy
Logistic Regression	0.79	0.89	0.84	0.885
Random Forest	0.81	0.87	0.83	0.871
AdaBoost	0.79	0.89	0.84	0.885
Support Vector Machines				
-with linear kernel	0.79	0.89	0.83	0.886
-with RBF kernel	0.80	0.89	0.83	0.886

Using the classification report for the top 5 features, it can be seen that the SVM performs slightly better than the other models. The performance for logistic regression and AdaBoost were the same, and Random Forest performed slightly worse than the rest. The precision and recall are highest for SVM with RBF kernel with precision=0.8 and recall=0.89. Note that the training accuracy of the SVM with the RBF kernel was 0.971 which was extremely high. The model also took the longest to fit and had highest computation time.

Conclusion:

The overall results seemed to classify the readmission rates of diabetic patients with an accuracy of up to 88.6%. The prediction for which patients will be readmitted within 30 days or over 30 days shows good accuracy. The most important features in determining the prediction included Number of Lab Procedures, Number of Medications, Time in Hospital, Number of procedures and Number of Diagnoses. These factors were ranked as the features with most importance. All five of the most important features were hospital logistic factors which shows that paying attention to medical logistics improve the efficiency of predicting patient readmission, and thereby reducing resource wastage and financial implications.

Bibliography:

Diabetes.co.uk. 2018. Guide to HbA1c. https://www.diabetes.co.uk/what-is-hba1c.html Accessed: 23 March 2018

McDonald, J. 2014. Multiple Logistic Regression. www.biostathandbook.com/mulitple logistic.html Accessed: 23 March 2018

Strack, B et al. 2014. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*. Accessed: 23 March 2018