BUSINESS OVERVIEW

Management of diabetes in hospitalized patients has a significant bearing on outcome, in terms of both morbidity and mortality. This project aims at predicting how likely a patient will be readmitted after discharge based on a number of features. Stakeholders are majorly interested in reducing the readmission rates in order to preserve and allocate hospital resources more efficiently. The recommendations are meant to inform future directions which might lead to improvements in future patient safety. This information may prove valuable in the development of strategies by caregivers to reduce readmission rates and costs for the care of individuals with diabetes.

The problem statement is:

How can diabetic patients’ health records help to predict patient readmission rates?

DATA UNDERSTANDING

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes and 101766 instances.

Information was extracted from the database for encounters that satisfied the following criteria

(1) It is an inpatient encounter (a hospital admission).

(2) It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis

(3) The length of stay was at least 1 day and at most 14 days.

(4) Laboratory tests were performed during the encounter.

(5) Medications were administered during the encounter

The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization. The target column was selected to be the readmitted column and the other columns were the predictors. The data contains no missing values or duplicates.

DATA PREPROCESSING

Using domain knowledge a number of columns were selected from the dataset columns. The criteria for selecting the columns was based on which columns seemed more applicable to the problem statement. Data cleaning involved removing missing values that were encoded as relevant values, splitting the data into train and test sets, performing ordinal encoding on the binary columns and performing one hot encoding on the other categorical columns.

MODELING

The metrics that were used to select the best model are the recall score, accuracy score and the cross-validation scores.

The reasons for selecting these metrics include:

1. This project aims at focusing on recall as the main metric since prioritizing on reducing false negatives, ensures that patients who are likely to be readmitted are correctly identified by the model.A high recall implies that the model will not incorrectly predict that a patient will not be readmitted when they actually will. By maximizing the recall, we are reducing the chance of missing potentially important cases. Preventing missed readmissions is an important factor in healthcare as they indicate inadequate care.

2. The overall correctness of the model's prediction was determined by the accuracy metric. Since class imbalances were removed the accuracy scores in the models are reliable values.

3. Cross validation accuracy scores for the training and tests sets were used to indicate if the model is overfitting. Models which were not generalizable were not selected.

RESULTS

From the above analysis the best model which can predict whether or not a patient will be readmitted before or after 30 days following discharge is the Logistic Regression Classifier model generated using pipelines. The reason for this is because it has the highest recall value and a pretty high accuracy as well compared to the other models.

With regards to computational complexity the KNeighbors Classifier and the Random Forest Classifier took the most time to be processed especially after incorporating pipelines.

The Logistic Regression Classifier Model had the following results:

a. Recall Score 0.4268829236483637

b. Accuracy Score 0.5033082214215526

c. Cross Validation Score (-0.4520046256610761, -0.4503294864706085)

The Logistic Regression Classifier was selected as the best model because it had the best overall metric scores and its computational complexity was low.

CONCLUSION

While carrying out the above analysis limitations were encountered; the dataset was large therefore it was prone to noise which ultimately affected the performance of the models and the dataset has a nonrandomized study design. Nonetheless the application of Logitsic Regression Classifier in the prediction of readmission is crucial in formulating strategies that deal with patient care, resource allocation and medical intervention. Hospitalization is a unique opportunity for providers to influence change to patient’s health outcome trajectories.