Prediction on Diabetes Patient's Hospital Readmission using Machine learning

Presented By:

Sai Kiran Reddy Kotha (700746206)

Cherukuri Venkata Ramana (700742704)

Kanduri Pavan Kumar (700740975)

Chelledi Sushmitha (700700167)

Vamshi Juluri (700744970)

Bird’s-eye view of the project:

1. Business Problem: Description of the problem with respect to Business
2. Problem Statement: Description of the problem with respect to Machine Learning.
3. Source of the Data.
4. Existing Approaches: Some solutions to the problem.
5. My Improvements: What I tried to improve in my approach.
6. Exploratory Data Analysis: Analyzing the given Data
7. First Cut Approach: My take on the Problem.
8. Models Used: Algorithms that I have used.
9. Model Comparison: Comparison of Results.
10. Future Work.
11. References

INTRODUCTION

Diabetes, commonly known as diabetes, is a metabolic disease that causes high blood sugar. About 422 million people worldwide have diabetes, the majority living in low-and middle-income countries, and 1.6 million deaths are directly attributed to diabetes each year. Patients with diabetes may be at higher risk of readmission than those without diabetes. In a study of 4769 medical patients, diabetes was associated with a statistically significant 40 % increased risk of readmission within 90 days

1. Business Problem:

It is important to know if a patient will be readmitted to some hospital. The reason is that you can change the treatment, in order to avoid a readmission.

In this database, you have 3 different outputs:

No readmission;

Readmission in less than 30 days (this situation is not good, because maybe your treatment was not appropriate);

Readmission in more than 30 days (this one is not so good as well the last one however, the reason can be the state of the patient.

In this context, you can see different objective functions for the problem. You can try to figure out situations where the patient will not be readmitted, or if their are going to be readmitted in less than 30 days (because the problem can the treatment), etc... Make your choice and let's help them creating new approaches for the problem.

Content

The data set 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. 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 into 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.

Medications were administered during the encounter

The data contains such attributes as a patient number, race, gender, age, admission type, time in the hospital, the medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatients, inpatient, and emergency visits in the year before the hospitalization, etc.

2.Problem Statement

The Prediction on Diabetes Patient's Hospital Readmission deals with the prediction of patients who have been suffering from diabetes. Here we are going to predict if the patients suffering from diabetes are going to readmit within a month or not. The main intention in solving the problem is to reduce the number of patients being readmitted due to diabetes and here there is 40% more chance of patients suffering from diabetes to be readmitted. And the money spent on the readmission accounts to billions. So by focusing more on this type of problem we will be able to save money and as well as wealth because if the number of readmissions is more then it means that there is a fault with the treatment. Here the amount of money spent on diabetes readmission is more and the people suffering from diabetes are increasing day by day and there should be an effective way in which the factors which are more dependent on predicting if the patient is going to be readmitted can be more focused on this problem is an example of binary classification in machine learning this problem is also a supervised problem because the targets for train data is already given and the model will be learned on the labeled data. Step by step method of solving the problem

1. Explanatory data analysis should be done to get in-depth information about the data
2. clean and process the data so that we can feed the data into the machine learning model
3. Apply various algorithms/ models to solve the problem.
4. Optimize the models.
5. Compare the results.
6. Select an appropriate model/approach that fulfills all the requirements and gives the best score.
7. Predict the values using the selected model and submit the solution

This data is an extremely imbalanced data so use the area under curve **(AUC)** as primary metric and **f1** score as an evaluation metric

3. Source of Data:

Diabetes 130-US hospitals for years 1999-2008 Data Set

Abstract: This data has been prepared to analyze factors related to readmission as well as other outcomes pertaining to patients with diabetes.

Data Set Characteristics:	Multivariate	**Number of Instances:**	100000	**Area:**	Life
Attribute Characteristics:	Integer	**Number of Attributes:**	55	**Date Donated**	2014-05-03
Associated Tasks:	Classification, Clustering	**Missing Values?**	Yes	**Number of Web Hits:**	433048

Data Set Information:

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. 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, et

Source:

The data are submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grant UL1 TR00058 and a recipient of the CERNER data. John Clore (jclore **'@'** vcu.edu), Krzysztof J. Cios (kcios **'@'** vcu.edu), Jon DeShazo (jpdeshazo **'@'** vcu.edu), and Beata Strack (strackb **'@'** vcu.edu). This data is a de-identified abstract of the Health Facts database (Cerner Corporation, Kansas City, MO).

4. Existing Approaches:

As the healthcare system moves toward value-based care, CMS has created many programs to improve the quality of care of patients. One of these programs is called the Hospital Readmission Reduction Program (HRRP), which reduces reimbursement to hospitals with above average readmissions. For those hospitals which are currently penalized under this program, one solution is to create interventions to provide additional assistance to patients with increased risk of readmission. But how do we identify these patients? We can use predictive modeling from data science to help prioritize patients.

One patient population that is at increased risk of hospitalization and readmission is that of diabetes. Diabetes is a medical condition that affects approximately 1 in 10 patients in the United States. According to Ostling et al, patients with diabetes have almost double the chance of being hospitalized than the general population (Ostling et al 2017).

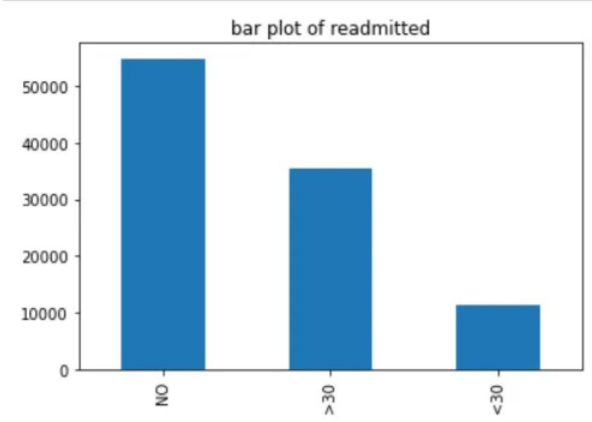
Through this project, we created a machine learning model that is able to predict the patients with diabetes with highest risk of being readmitted within 30 days. The best model was a gradient boosting classifier with optimized hyperparameters. The model was able to catch 58% of the readmissions and is about 1.5 times better than just randomly picking patients. Overall, I believe many healthcare data scientists are working on predictive models for hospital readmission

5. My Improvements:

1. I used AUC as a primary metric and f1 score to measure the metrics since data is extremely imbalanced
2. I used k best features to select the best features out of the dataset
3. To manage the imbalance instead of using smote I used upsampling of the test data
4. Here since it is an extremely imbalanced data I have tried various combinations of upsampling, downsampling with different ratios and found out that the undersampling performed the best
5. here weighted average is taken for getting the predictions of the data

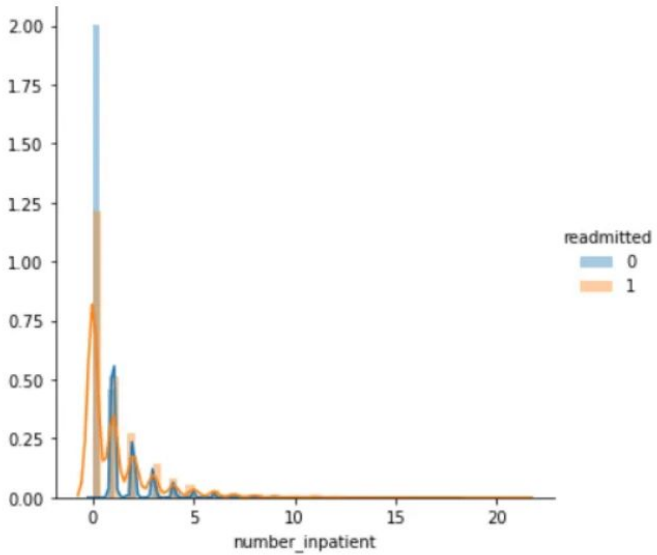
6. Exploratory Data Analysis:

1. This dataset consists set of variables, each representing a feature in the patients who are admitted into the hospital
2. The ground truth is labeled ‘readmitted’ and represents if the patient is readmitted within 30 days or not
3. ‘diabetes.csv’ file consisting of information about patients
4. The number of points in the data: 1010766
5. Total number of features 50
6. Total number of categorical features 37
7. Total number of numerical features 13
8. First looking at the demographics of the data it is recorded in united states of America
9. Bar plot of the output label shows that less number of patients are readmitted clearly stating that it’s a imbalanced data here the >30 also come under NO category as we are predicting readmission under 30 days



10. Bar plot of age shows more number of people aged between 50–80 are being admitted into the hospital

11. By looking at the relation between number of inpatient and the readmission rate if the number of inpatient visits are less there is less chance of readmission



7. First Cut Solution:

1.Check for the missing values to be dropped and unnecessary features from the dataset.
2.Encode all the categorical features using One-Hot Encoding.
3.After that combine all the features. using hstack
4.Apply feature engineering based on certain experiments and experiment with different features to get the best accuracy
5.Since it is a imbalanced data use undersampling,oversampling or pipeline
6.Apply baseline model which is simple like logistic regression and go to complex model
7.Evaluate all the models and select the best model that gives the best score.

8. Models Used:

1.Logistic regression: The first model that is used to get a benchmark score is Logistic Regression. It is selected as the Baseline Model. Since we have 50 features and it is a Classification task, Logistic Regression will give us a decent benchmark/baseline score to work upon.

```
# here using cross validation and gridsearch to determine the best hyperparameter
logit = LogisticRegression()
param_grid = {'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000] }
clf = GridSearchCV(logit, param_grid,scoring='roc_auc',cv=5,n_jobs=-1)
clf.fit(X_train_fs,y_train)
# selecting the best model and printing the confusion matrix
```

2.Decision tree:

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

```
parameters = {'max_depth': [1, 5, 10, 50], 'min_samples_split': [5, 10, 100, 500]}
dtree = DecisionTreeClassifier()
clf = GridSearchCV(dtree,parameters,scoring='roc_auc',cv=5,n_jobs=-1)
clf.fit(X_train_fs,y_train)
    # selecting the best model and printing the confusion matrix
dtree =clf.best_estimator_
dtree.fit(X_train_fs,y_train)
```

3. Random Forest:

Random forest are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

```
rm = RandomForestClassifier()
params={'n_estimators':[5,10,25,50,100,300,500],'n_estimators': [10, 25], 'max_features': [5, 10]
 'max_depth': [10, 50,75,100, None], 'bootstrap': [True, False]}
model_rf=GridSearchCV(rm,param_grid=params,cv=5,scoring='roc_auc',n_jobs=-1,verbose=1)
model_rf.fit(X_train_fs, y_train)
```

4. Weighted predictions of both decision tree and random forest:

Here I took the predictions from both the decision tree and random forest and added the probability of both of them and took the arg max to get the output here calibrated classifier is used to get the correct probability of prediction

9.Model Comparison:

Vectorizer	Model	f1-score	TEST AUC
ONE HOT ENCODING	LOGISTIC REGRESSION	0.27	0.62
ONE HOT ENCODING	DECISION TREE	0.27	0.62
ONE HOT ENCODING	RANDOM FOREST	0.27	0.62
ONE HOT ENCODING	COMBINED DECISION TREE RANDOM FOREST	0.27	0.63

Here the primary metric is area under curve and f1-score is primary metric because here the positive to negative ratio is 10:1 here we can't take accuracy as primary metric because if it is overfitted then we can get 90% accuracy so area under curve and f1-score should be taken into consideration

Here this case study is also done using deep learning check it in my github in the later section

10. Future Work:

1.In my solution I used upsampling with one hot encoding and response coding and label encoder which performed the same we can try using different encodings to check the change in performance.

2.Further and better hyperparameter training can improve the results.

11.References:

1.https://towardsdatascience.com/predicting-hospital-readmission-for-patients-with-diabetes-using-scikit-learn-a2e359b15f0

2.https://www.kaggle.com/iabhishekofficial/prediction-on-hospital-readmission

3.https://medium.com/@akshay.43279

4.https://www.kaggle.com/brandao/diabetes

5.https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

6.https://www.appliedaicourse.com/