Prediction of readmission rate of patients with diabetes

Quanquan Zhang
Department of Mechanical and
Aerospace Engineering
University of California, Irvine
CA, Irvine
quanquaz@uci.edu

Xiaofang Zhang
Department of Electrical
Engineering and Computer Science
University of California, Irvine
CA, Irvine
xiaofaz7@uci.edu

Yiran Luo
Department of Electrical
Engineering and Computer Science
University of California, Irvine
CA, Irvine
yiranl27@uci.edu

ABSTRACT

Diabetes is a serious and prevalent disease all over the world. There are various factors contributing to diabetes. In this project, by analyzing the datasets named Diabetes 130-US hospitals for years 1999-2008 Data Set, we are going to figure out which factors make critical contribution and suitable machine learning models to make prediction of readmitted rate of patients with diabetes. In this paper, we investigate performance of two machine learning modes in predicting readmitted rate. First, we do the data exploration. Then, we tune the model's hyperparameter and try to maximize the performance of our model with readmitted rate prediction. Experiments results show our models achieve over 80% prediction accuracy.

KEYWORDS

Machine learning, Decision Tree, Neural Network

1 Introduction

Diabetes was the nation's seventh-leading cause of death in 2019, accounting for 87,647 deaths annually [1]. Those with diabetes are twice as likely to have heart disease or a stroke than those without diabetes. Also, high rates of 30-day readmission are both costly for the hospital and detrimental for the patient, and diabetes remain one of the greatest risk factors for increased 30-day readmissions. Therefore, it is important to figure out what factors make most contribution and make accuracy prediction of readmission rate.

Our dataset [2] 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 medications, diabetic medications, number of outpatients, inpatient, and emergency visits in the year before the hospitalization, etc.

2 Data exploration and Featuring Engineering

2.1 Data exploration

First, to intuitively understand the dataset as a whole, we split data set into two parts based on whether patients readmitted within 30 days. As shown in Figure 1, the number of readmitted patients is less than the number of not admitted. Then, we will have an intuitive impression about the relationship between single feature and the readmission rate from the graph. As shown in Figure 2, Caucasian race has more readmission rates. As for age, patients with age 60 and 70 have a higher readmitted rate. And it seems that there is no relationship between gender and readmission. If number of medications between 10 and 20, there is more chances that patient will admit again. As shown in Figure 3, patients stay in hospital between 2 and 4 days have a higher readmitted rate. Readmission rate looks similar whether the change of medication occurs or not. But patients get diabetes medication readmitted often. As for the max glue serum test, patients without it get a higher readmission rate. As shown in Figure 4, it is obvious that patients who don't take a1ctest get higher readmission rate. There is no relationship between the number of lab procedures and readmission rate. And patients who are admitted as emergency or discharged to home have higher readmission rate.

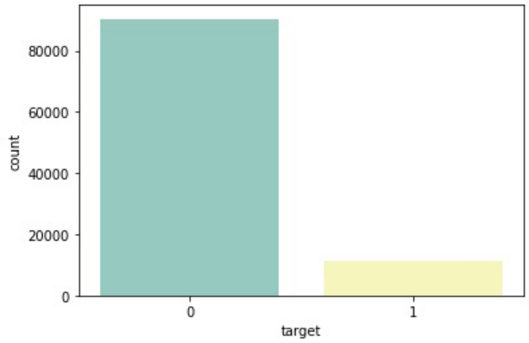

Figure 1: Distribution of readmission

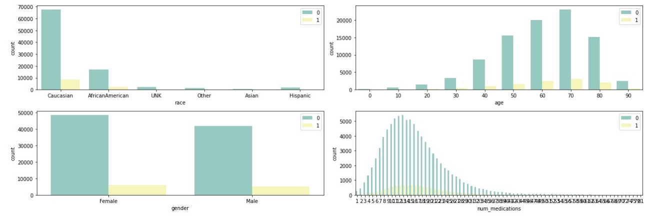

Figure 2

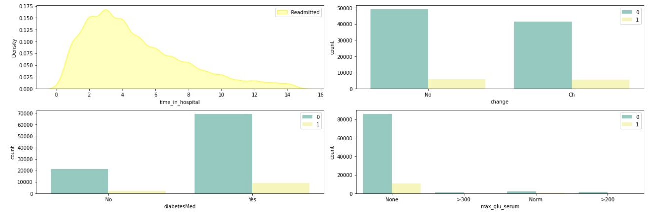

Figure 3

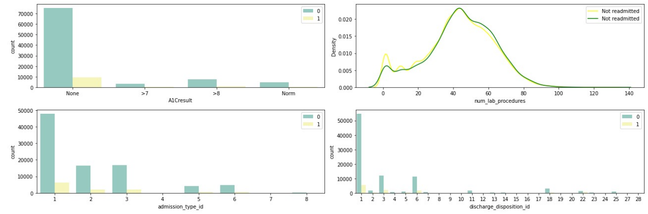

Figure 4

2.2 Feature fix and missing data imputation

After looking through the raw dataset, we found that there are many missing and duplicate data. In addition, there are many categorical variables. First, we remove duplicate data based on the feature of the patient number because each patient has only one number. After drop off duplication, there are 71518 samples in each feature. Next, we removed the encounter id, patient id and payer code because these features are not relevant to the target class. Based on data visualization, we intuitively found that in many features of drugs, almost all data (>95%) are 'Yes' or 'No', such as repaglinide, nateglinide, chlorpropamide. This shows that the sample is extremely unbalanced. Especially for tree models, the information gain of these features is almost zero, which is not helpful for classification, so we delete these features. After counting missing values, it is obvious that the proportion of missing data for the weight was 96%, and the proportion of missing data for the medical_specialty feature was 49%, so we removed these two features. For missing values of other features, we need to fill in some data based on specific methods. For the missing values of race, we use 'other' to fill in because it is a categorical variable and has a relatively large proportion of missing values. For diag_1, diag_2 and diag_3, we use the mode for imputation because it is a numerical variable and the missing proportion is small.

2.3 Feature transform

In this section, we convert categorical variables into numerical variables. The raw data for age is a range value and we use the median instead of each range. For example [10, 20) maps to 15. For the features of diag_1, diag_2 and diag_3, the data are mapped to categorical variables according to ICD-9 (https://en.wikipedia.org/wiki/List_of_ICD-9_codes). Then we convert all categorical data to numerical data using one-hot encoding. At first, we divided the patient's class into two groups '<30, >30'and 'No'. But after reviewing Strack's paper, they grouped'>30'and 'No' into one category, '<30' as another category. This is because the project only cares about early readmission. Based on data visualization, we found that the class of '<30' accounted for a small proportion, which would lead to an imbalanced training dataset. So, in the subsequent model training, we use the SMOTE algorithm to solve this problem.

3 Models and performance

In this project, we take the random forests model and the neural networks model. Random forests model has many advantages. It has extremely high accuracy and can effectively operate on high-dimensional big data. It introduces a bagging algorithm and reduces the effect of overfitting. The advantage of neural networks lies in its strong robustness and fault tolerance, and it can fully approximate any complex nonlinear relationship.

3.1 Random Forest

The original dataset is tabular form, we first thought of using a random forest model. For the training dataset, we use SMOTE to address the class imbalance problem. The four best hyperparameters of the model were obtained by cross-validated grid search and using ROC-AUC as the scoring criterion. There are 200 estimators, the max depth is 6, min_samples_split is 100, and min-samples-leaf is 10. The accuracy of the model is 0.891, the auc is 0.527, and the recall is 0.065.

3.2 Neural Networks

The next model we chose was Multi-layer Perceptron Neural Networks. We also used SMOTE on the training set to balance the labels. Cross-validated grid search was utilized to find the optimal parameters. We used ROC-AUC as the scoring criterion instead of accuracy because we think ROC-AUC is a better and more comprehensive yardstick in the binary classification problem like this. Eventually the best model was a 3-layer neural network with 20, 20 and 15 nodes in each layer respectively. The activation function was logistic and the regularization parameter 'alpha' was set to 0.34. This model gave us an ROC-AUC of 0.579 and an accuracy of 0.802 and a recall of 0.083 on the test set.

Model	Accuracy	Recall	AUC
Random Forest	0.891	0.065	0.517
Neural Networks	0.802	0.083	0.579

Table 1 Model Performance

4 Conclusion and thoughts

Based on Table 1, we choose the neural network model because its AUC and Recall are relatively high. Even though the Accuracy is slightly lower compared to the random forest model, in this project we mainly compare the AUC. AUC can reflect the performance of the classifier, that is, the ability to identify the positive class. Obviously, the classes in the dataset are imbalanced, and the number of negative class samples (>30 & No) is much larger than the number of positive class samples (<30). Specifically, the positive class in the project only accounts for 10%, and many positive samples are judged to be negative after training. Negative samples are almost all judged as negative class. Therefore, the reason for the high accuracy is that the number of negative samples is large, and the prediction results of negative samples are accurate, thus covering up the fact that the positive class is inaccurate. We can clearly see from the recall of the two models that the positive class prediction is very poor. This also proves our analysis.

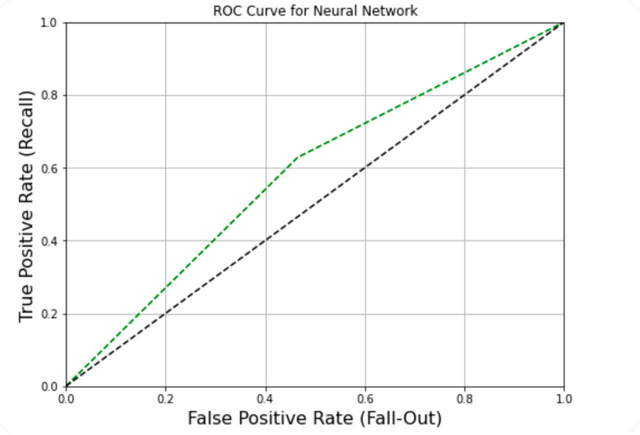

Figure 5 Neural Network ROC-AUC Curve

We think the model is underfitting because of low accuracy and there’s still plenty of room for improvement, but our model is limited by the AUC scoring criteria. In addition, in the process of parameter grid search, all parameters are defined by us. The parameters combination after search can only be the best in the selected parameters, which does not mean that they are the global optimal solution. So, if we want to study if the selected parameters are overfitting or underfitting, we can judge by drawing a learning curve in the future. The learning curve is to compute the accuracy of the training dataset and validation dataset with different dataset sizes. By looking at the performance of the model on the data, it can be determined if the model has high variance or high bias, and if the training set increase can reduce the overfitting. In Figure 6, as the number of samples increases, if the error of both the training set and the validation set is high, the model is underfitting. If the errors all reach a low value, the model performs best. If the training set error is low and the test set error is high, the model is overfitting. Our neural network model is underfitting. We expect that the model can be improved in the following ways: 1. Multi-classification to address sample imbalance problem 2. Introducing weights into the model 3. Dimensionality reduction of dataset

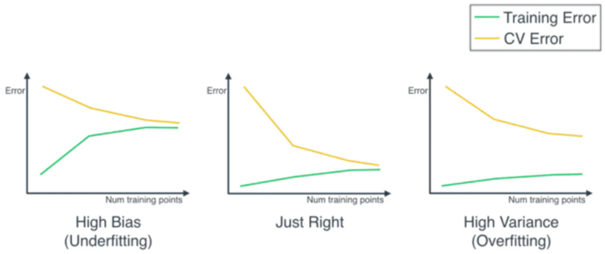

Figure 6 Learning Curve

4 Division of task

We have roughly split the project into three parts as follows:

Xiaofang Zhang: Data Exploration and Visualization

Quanquan Zhang: Feature Engineering and Random Forest model

Yiran Luo: Feature Engineering and Neural Network model

Model parameters tuning, result analysis and report writeup were done by everyone.

REFERENCES

[1] https://www.americashealthrankings.org/explore/annual/measure/Diabetes/state/ALL

[2] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.