Name: Huijun An

UID: 904499500

BIOS 273 Final Project

Prediction of Diabetes Patient Readmission Classification

Introduction

Diabetes is a group of metabolic disorders characterized by a high blood sugar level over a prolonged period.[1] The reason for diabetes is that the patient's body is not producing enough insulin, or not responding to insulin properly. Common risk factors for diabetes include weight, inactivity, family history, race, age, high blood pressure, etc. There are about 425 million people suffering from diabetes worldwide (8.8% of adult population) as of 2017, and the rate is rising. The global cost for diabetes related expenditure was about $727 billion in 2017.[2] With the high prevalence rate of diabetes, the question of bringing proper care to every patient has arisen. In 2014, the research article "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records" by Beata Strack brought out the problem that few national assessments serving as baseline of change of diabetes care are present in the hospitalized patients (specifically non-ICU). The researchers doubted that this problem affects inpatient safety and inpatient care cost. To address the problem, they fitted a multivariate logistic regression model to prove the relationship between readmission and indicators of attention to diabetes care. A typical indicator, HbA1C measurements, which stands for the average blood glucose level for the last two to three months, was selected. The result of this research paper shows that the relationship between the probability of readmission and the HbA1C measurement depends on the primary diagnosis, and the data suggest that the greater attention to diabetes

reflected in HbA1C determination may improve patient outcomes and lower cost of inpatient care.[3]

While the paper focuses on using statistical model to find the relationship between target measurements and outcome and identify significant factors, the fact that patient readmission affects patient care quality and cost worth people's attention. If the readmission status of patient can be accurately predicted, potential patient who is going to be readmitted can be given more intensive and personalized care to improve their care quality and decrease their cost. In light of the fact that the dataset being used in this paper has not been used by any other researchers, the purpose of this project is to use multiple machine learning algorithms to find the best model for future patient readmission prediction with highest performance. Algorithms that are going to be used in this project includes Logistic Regression, Support Vector Machine (SVM), Classification and Regression Trees (CART), Random Forest, and XGBoost.

Methods & Results

A. Data Acquisition and Pre-processing

The data for this project was acquired from UCI Machine Learning Repository with the name "Diabetes 130-US hospitals for years 1999-2008 Data Set". The original dataset is in csv format. It contains 101,766 records and 50 attributes. The detailed descriptions of the attributes are listed in the original research paper by Strack. Note that in the original paper, in order to account for within subject correlation, the paper only used the first encounter data in each patient, resulting in a 30% reduction in the dataset. For this project, since the aim is the accuracy of classification, within subject correlation is not a concern. However, considering that patients with more records may suggest that the patient has been readmitted and has a higher chance of being readmitted,

this information was put in a new column called “Single record” in binary format indicating if the patient has only one record in the dataset.

There are three attributes in the dataset containing more than 50% missing values, weight, payer code, and medical specialty. The weight column contains as high as 97% missing value. Therefore, the entire column was eliminated for the integrity of the dataset. However, considering that weight is an important risk factor of diabetes, the reason for missing so many records worth research. It turns out that prior to HITECH legislation of the American Reinvestment and Recovery Act in 2009, hospitals and clinics were not required to capture the weight of patient in a structured format.[3] Since the dataset contains information in the time frame of 1999 to 2008, it makes sense for it to miss 97% of weight records. The attribute payer code contains 52% of missing values, indicating the type of payment used by patient. After logical reasoning of deeming it as irrelevant with patient readmission, this column was also deleted. The feature medical specialty, which indicates the type of admitting physician, has 53% of missing values. Since it was concluded as a significant covariate in the original paper, only the missing rows were excluded (which also lightened the scale of the data by half) and the column was kept in the dataset. Around 2 % of missing data were present in race and diagnosis, rows that are missing were deleted. The final count of records for this project is 50,727.

In order to constrain the size of this project, many attributes with way too many categories were combined into several large categories. For patient diagnosis, originally it was coded in ICD9 codes, and it was recoded in a categorical format based on ICD9 broad definitions as “Circulatory”, “Respiratory”, “Digestive”, “Diabetes”, “Injury”, “Musculoskeletal”, “Genitourinary”, “Neoplasms”, and “Other”. Medical specialty was recoded in broader format as done in the research paper as “Emergency/Trauma”, “Family/General Practice”, “Internal

Medicine", "Surgery", and "Other". Admission type and source were recoded in a binary format as 1 being emergency, 0 being other. Discharge disposition was recoded in a binary format as 1 being home, 0 being other. There exist 24 medications indicators regarding diabetes treatment. For simplicity of this project, for medications excluding insulin, they were together combined into three binary categorical variables as having medication or not, having increasing medication level or not, and having decreasing medication level or not. One-hot encoding technique was used for all nominal categorical variables. For age, it was originally coded in categorical format as 0-10, 10-20, …, 90-100 and it was transformed into continuous format of 0, 10, …, 100. The univariate analyses for covariates and outcome variable are summarized in Table 1. The outcome variable has a balanced result of 55.5% readmitted to 45.5% not readmitted.

Table 1. Summary Statistics

i. Categorical Features

Features	**Record Number (%)**
Gender	
Female	27,442 (54.1%)
Male	23,285 (45.9%)
Race	
White	37,333 (73.6%)
Black	10,942 (21.6%)
Other	2,452 (4.8%)
Admission source	
Emergency	25,207 (49.7%)
Other	25,520 (50.3%)
Admission type	
Emergency	20,339 (40.1%)
Other	30,388 (59.9%)
Discharge Disposition	
Home	32,191 (63.5%)
Other	18,536 (36.5%)
HbA1C Test	
Yes	8,792 (17.3%)
No	41,935 (82.7%)
HbA1C Test result high	
Yes	6,607 (13.0%)

No	44,120 (87.0%)
Glucose Serum Test	
Yes	2,896 (5.7%)
No	47,831 (94.3%)
Glucose Serum Test result high	
Yes	1,388 (2.7%)
No	49,339 (97.3 %)
Insulin Usage	
Yes	27,606 (54.4%)
No	23,121 (45.6%)
Insulin Increase	
Yes	5,702 (11.2%)
No	45,025 (88.8%)
Insulin Decrease	
Yes	6,450 (12.7%)
No	44,277 (87.3%)
Other Medication Usage	
Yes	24,052 (47.4%)
No	26,675 (52.6%)
Other Medication Increase	
Yes	1,769 (3.5%)
No	48,958 (96.5%)
Other Medication Decrease	
Yes	1,007 (2.0%)
No	49,720 (98.0%)
Single Record	
Yes	27,958 (55.1%)
No	22,769 (44.9%)
First Diagnosis	
Circulatory	15,351 (3.0%)
Respiratory	6,680 (13.2%)
Digestive	4,443 (8.8%)
Injury	3,451 (6.8%)
Diabetes	4,478 (8.8%)
Musculoskeletal	3,089 (6.1%)
Genitourinary	2,302 (4.5%)
Other	10,933 (21.6%)
Second Diagnosis	
Circulatory	16,288 (3.2%)
Respiratory	5,017 (9.9%)
Digestive	2,032 (4.0%)
Injury	1,206 (2.4%)
Diabetes	7,149 (14.1%)
Musculoskeletal	952 (1.9%)
Genitourinary	3,437 (6.8%)

Other	14,646 (28.9%)
Third Diagnosis	
Circulatory	15,205 (30.0%)
Respiratory	3,493 (6.9%)
Digestive	1,895 (3.7%)
Injury	930 (1.8%)
Diabetes	9,071 (18.9%)
Musculoskeletal	1,039 (2.0%)
Genitourinary	2,914 (5.7%)
Other	16,180 (3.2%)
Medical Specialty	
Emergency/Trauma	7,540 (14.9%)
Family/General Practice	7,271 (14.3%)
Internal Medicine	23,925 (47.2%)
Surgery	7,742 (15.3%)
Other	4,249 (8.4%)
Diabetes Medication Prescribed	
Yes	39,540 (77.9%)
No	11,187 (22.1%)
Change in Diabetes Medication	
Yes	23,742 (46.8%)
No	26,985(53.2%)
Readmitted	
Yes	28,132 (55.5%)
No	22,595 (44.5%)

ii. Continuous Features

Features	**Mean ± standard deviation**
Age	70.2 ± 1.6
Time in Hospital	4.4 ± 3.0
Number of Lab Procedures	42.9 ± 19.6
Number of Other Procedures	1.4 ± 1.7
Number of Medications	15.8 ± 8.3
Number of Diagnoses	7.1 ± 4.0
Number of Emergency Visit	0.21 ± 1.2
Number of Inpatient Visit	0.64 ± 1.7
Number of Outpatient Visit	0.25 ± 0.94

B. Splitting the dataset

The dataset was randomly splitted int to three different parts: 10% Validation set, 18% Test set, and 72% Training set. The validation set is for the purpose of training hyper-parameters. Due

to the efficiency of computer, it is impossible to train the machine with the whole set of data. The purpose of creating training and test set is to avoid overfitting, as it is necessary to make the proposed model to perform well on dataset that it has never seen (test data), which stands for the generalization ability.

C. Logistic Regression

Logistic regression is a statistical regression model using a logistic function to model a binary outcome variable, in this case readmitted or not. Consider a model with vector of predictors X and a binary response variable Y, which we denote p = P(Y=1), a linear relationship can be assumed between the predictors and the log-odds of the even Y equals 1. The mathematical form is as following: $log\frac{p}{1-p} = \beta^T X$. The model can predict the probability of a record equals 1. Although the model itself does not perform classification, it can be transformed to a classifier by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class.

Table 2. Logistic regression confusion matrix

	0 (predicted)	1 (predicted)
0 (actual)	2988	1079
1 (actual	1136	3928

D. SVM

A Support Vector Machine (SVM) is a classifier using a separating hyperplane. The algorithm tries to find an optimal hyperplane to categorize new records into the correct class. Though in many cases, the sets are not linearly separable in the space. To solve this issue, the idea of kernel machine is proposed, implicitly mapping the inputs into high-dimensional feature

spaces. An example is visually displayed in Figure 1. Considering the dimension of the dataset, linear SVM classification is probably not a good classifier. Hence, gaussian kernel is used for this project. This kernel is defined as $K(x, x') = \exp(-\frac{||x-x'||^2}{2\sigma^2})$. The hyper-parameters need to be tuned in this method are gamma and cost, where gamma parameter defines how far the influence of a single training example reaches, and cost adjusts how hard or soft the margin classification should be. After parameter tuning, the best hyper-parameters are determined to be gamma = 0.125 and cost = 1.

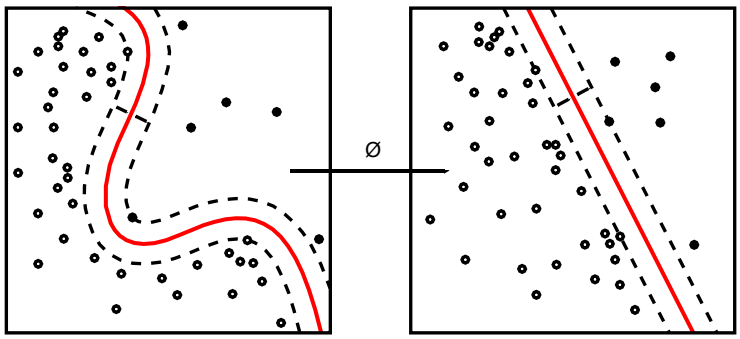

Figure 1. Kernel SVM Machine to Linear SVM Machine[4]

Table 3. SVM confusion matrix

	0 (predicted)	1 (predicted)
0 (actual)	2311	1756
1 (actual	983	4081

E. CART

CART is a machine learning method to produce a model predicting the value of a target variable based on several input variables. An interior node represents one of the input variables; there are edges to children for each of the possible values of that input variable; each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.[5] CART is actually an umbrella term. Many other ensemble

methods are based on CART with more than one decision tree constructed such as random forest and XGBoost. The fully-grown tree for this CART model is shown in Figure 2. Note that the later added variable *Single Record* contributes most in classification (Thanks to Dr. Ramirez), indicating that patients with more records are more likely to be readmitted.

Table 4. CART confusion matrix

	0 (predicted)	1 (predicted)
0 (actual)	3081	986
1 (actual	1158	3906

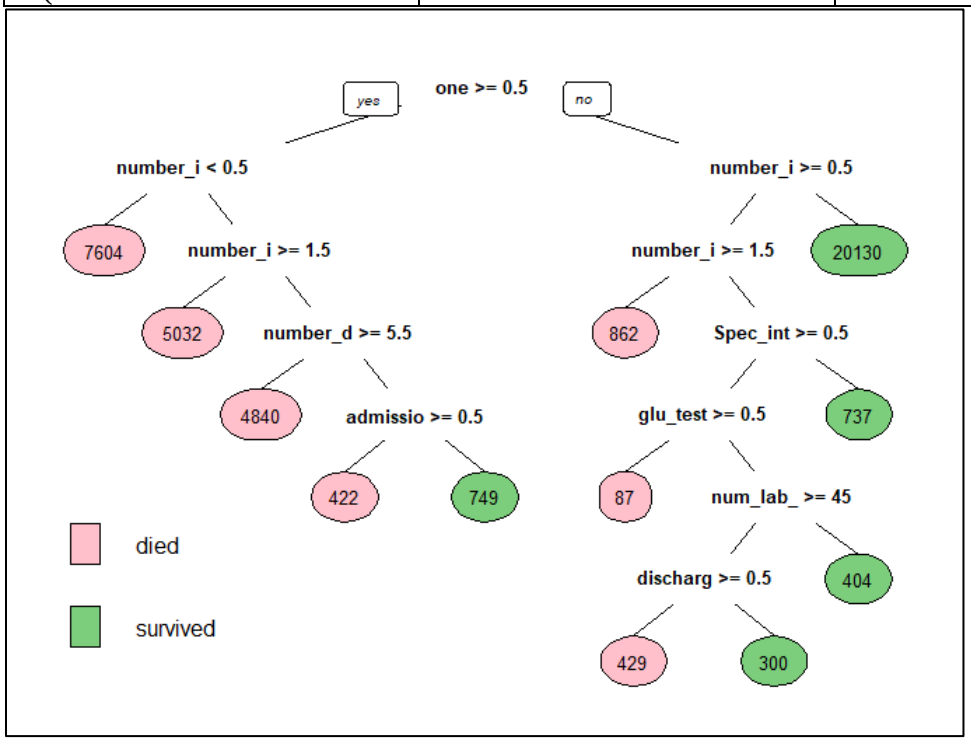

Figure 2. CART plot. Note that one stands for single record, *number_i* stands for number of inpatient visits, *number_d* stands for number of diagnoses, *admissio* stands for admission source, *Spec_int* stands for Internal medicine medical specialty, *glu_test* stands for glucose serum test, *num_lab* stands for number of lab procedures, and *discharg* stands for discharge disposition.

F. Random Forest

Random forest, as stated above, is an ensemble method based on CART. It consists of a large number of individual decision trees and uses the "crowd's wisdom". Each individual tree in the

random forest makes a classification and the most votes becomes the model's classification. The hyper-parameter, mtry, which stands for number of randomly selected predictors, were being tuned and its optimal value was 7. The tuning process is visualized in Figure 3. The hyper-parameter ntree, the number of branches will grow after each time split, was left with default value 500.

Table 5. Random forest confusion matrix

	0 (predicted)	1 (predicted)
0 (actual)	3041	1026
1 (actual	1095	3969

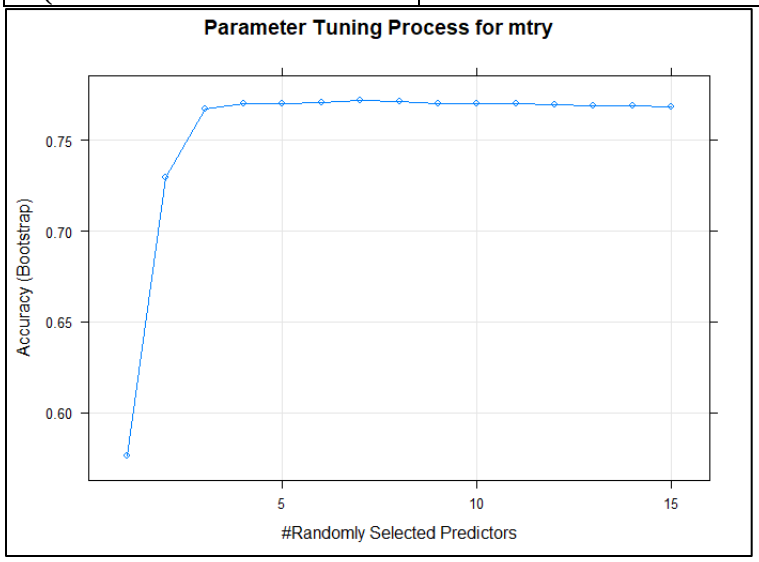

Figure 3. Accuracy vs. mtry for random forest

G. XGBoost

XGBoost is also an ensemble method based on CART. It uses the gradient boosting framework. The name XGBoost, stands for extreme gradient boosting. Boosting means building models from each "weak learner", i.e. each individual tree. Gradient means to minimize a loss function. The optimal hyper-parameters were nrounds (the maximum number of iterations) = 434, max_depth (the depth of the tree) = 5, eta (the learning rate) = 0.195, and lambda (L2

regularization on weight) = 0.144. Other hyper-parameters (min_child_weight, the minimum number of instances required in a child node, defaulted 1; subsample, the number of samples supplied to a tree, defaulted 1; colsample_bytree, the number of features supplied to a tree, defaulted 1) were left with default values.

Table 6. XGBoost confusion matrix

	0 (predicted)	1 (predicted)
0 (actual)	2909	1158
1 (actual	1008	4056

H. Summarized Result

The final results summarized in comparison of accuracies are listed in Table 2, where the definition of accuracy is number of correct predictions over total number of predictions.

Table 2. Summarized final results

Machine Learning Algorithm	**Accuracy**
Logistic regression	75.7%
SVM – Gaussian kernel	70.0%
CART	76.5%
Random forest	76.8%
XGBoost	76.3%

Discussion

The primary goal of this project is to use multiple machine learning algorithms to find the best model to predict patient's readmission. Based on the accuracy measure, we found out that Random Forest model has the best performance. The result of random forest being the best learner is not surprising. Considering the amount of data in this project (about 50k after data reduction), logistic regression and SVM tend to perform worse in such large datasets comparing

to random forest. Though it is unexpected that SVM is performing worse than logistic regression. Note that CART, random forest, and XGBoost are producing similar results and random forest is winning by 0.3%. Since random forest and XGBoost are ensemble methods from CART, it could be because of the original tree model is stable by itself. Therefore, its ensemble methods produce similar results. The reason for XGBoost being a little bit worse than the other two could be due to the fact that it is more sensitive to overfitting the noise in the dataset. In terms of machine learning time, SVM took the longest for the computer to run. Random forest took the second longest and XGBoost was the fastest. Hence, for final selection of actual use of the model to predict patients' readmission, considering that random forest and XGBoost has similar accuracy rate, it could be faster to just use XGBoost to save time. However, with the best learner random forest, the accuracy was only 76.6%, which means that 23.4% was misclassified. Limited by computer efficiency, extensive data reduction was performed to the original dataset. For example, the missing data in medical specialty was deleted to get rid of half of the data, while it could be a significant portion to readmission prediction. 24 medications' indicators were combined to reduce number of attributes, while some medications might indicate severe symptoms and lead to significance in predicting readmission. Moreover, given the nature of messiness of medical data, learners could be misled by the noises present which lead to overall low accuracy rate. Also, just as weight was not included in the dataset due to legislation, many other risk factors such as family history and blood pressure, that could lead to better prediction of readmission. Lastly, considering that the dataset is from 1999 to 2008, with ten years of medical development, the accuracy of selected learner on present data is doubtful. For future studies, a dataset with most recent data including more relevant features can be explored to make better predictions for new incoming patients' readmission.

Reference

1. "About diabetes". World Health Organization. Archived from the original on 31 March 2014. Retrieved 4 April 2014.
2. International Diabetes Federation (2017). IDF Diabetes Atlas, 8th edn. Brussels, Belgium: International Diabetes Federation.
3. 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.
4. Support-vector machine. (2019, Dec 1). https://en.wikipedia.org/wiki/Support-vector_machine.
5. Decision tree learning. (2019, Nov 14). https://en.wikipedia.org/wiki/Decision _tree_learning.