Predicting hospital readmission of diabetic patients using ensemble learning

Amey Bhole, s3411427
Martin Dijkhuizen, s1992031
Ilyas Aaqaoui, s3502317
Tos Sambo, s2385813

University of Groningen
Faculty of Science and Engineering

Abstract

Diabetes is a disorder in which there are high blood sugar levels. Therefore, the assessment of diabetes patients during hospitalization is of great importance. The goal of this research is to predict the probability that a diabetic patient is readmitted within 30 days of hospitalization. Predicting readmission could help to reduce cost of care, medical dispute and improve patients health and safety. The dataset used for this research was submitted by the Center for Clinical and Translational Research, Virginia Commonwealth University, and is available on the University of California Irvine Machine Learning Repository. In this paper, different bagging based ensemble models were created using logistic regression, naive bayes, random forest, K-nearest neighbours and extreme gradient boosting. As a result, the ensemble with logistic regression, extreme gradient boosting, random forest, and naive bayes provides the most accurate results with an AUC of 61.45%. Therefore, it is recommended to further improve this ensemble learning algorithm in order to improve the accuracy and AUC.

1 Introduction

Diabetes is a group of metabolic disorders in which there are high blood sugar levels. More than 29 million American adults have diabetes and another 86 million have prediabetes. In 2012, diabetes cost $245 billion in the United States [3]. The assessment of diabetes patients during hospitalization is therefore of great importance. The main goal of this research is to predict the probability that a diabetic patient is readmitted within 30 days of hospitalization. Readmissions are especially serious as they might indicate a failure of the health system to provide adequate support to the patient and are extremely costly to the system. Predicting readmission could help to reduce cost of care, medical dispute and improve patients health and safety.

The dataset used for this research was submitted by the Center for Clinical and Translational Research, Virginia Commonwealth University, and is available on the University of California Irvine Machine Learning Repository. The data set contains data of clinical care at 130 hospitals for over a period of 10 years (1999-2008). Variables can be differentiated in different categories: Admission and discharge details, patient medical history, patient admission details (number of diagnoses and lab procedures, time spent in hospital), Clinical results, Medication details and controlling demographic variables such as race and gender.

In this paper, different bagging based ensemble models were created using logistic regression, naive bayes, random forest and extreme gradient boosting. Ultimately, the ensemble with logistic regression, extreme gradient boosting, random forest, and naive bayes provides the most accurate results with (AUROC curve [95 percent CI]) of (0.6145 [0.6135, 0.6154]).

2 Data Preparation

2.1 Data Description

The dataset used initially has 101,766 observations and 50 variables. The features and corresponding descriptions from the original dataset are presented in Table 6. However, before starting the training, the dataset needed some preprocessing and cleaning, since most variables were categorical and a creation of dummy variables was needed.

2.2 Data Preprocessing

In terms of observations, this dataset contained multiple inpatient visits for some of the patients - multiple visits for the same parson - and hence the observations could not be considered as statistically independent, which is an assumption of the logistic regression model. Subsequently, we considered only the first encounter for each patient as the primary admission and determined whether or not they were readmitted within 30 days. This reduced the size of the data set to 69,990 observations with 50 variables. In addition, features with more than 50 percent missing data such as encounter id, patient number, weight and payers code were removed from the data set. Techniques could be used to generate values for the missing data, but this was deemed unpractical when the proportion missing is too large.

There are 8 numerical variables without any missing values. The diagnosis features (diag1, diag2, diag3) had icd9 codes, hence codes according to the International Statistical Classification of Diseases and Related Health Problems; a list of 6-character alphanumeric codes to describe diagnoses. These distinct codes were reduced down to 10 different categories for diag1, diag2 and diag3. Lastly, all the categorical variables with missing data were treated as another category.

To give an overview of the data, statistics on some of the features in the dataset are given in table 2.

Table 1: Descriptive statistics

Statistic	Mean	St. Dev.	Min	Max
Time_in_hospital	4.273	2.934	1	14
Num_procedures	1.425	1.757	0	6
Num_medications	15.665	8.287	1	81
Male	0.468	0.499	0	1
Age [0-30]	0.026	0.159	0	1
Age [30-60]	0.312	0.464	0	1
Age [60-100]	0.662	0.473	0	1
Insuline	0.510	0.500	0	1
Diabetes	0.082	0.275	0	1
Readmitted	0.090	0.286	0	1

Total number of observations, $N = 69,990$

It is also important to note that the target variable corresponding to readmission, had initially three categories: readmission within 30 days($\leq$ 30), readmission after 30 days ($>$ 30), and no readmission (NO). Since we are interested in readmission within a short period, we converted it to a dummy variable which is 1 when readmission within 30 days occured, and 0 otherwise. This resulted in an unbalanced dataset with approximately 10% positive cases and 90% negatives.

3 Methods

For an accurate prediction of hospital readmittion of diabetic patients within 30 days, we will construct an ensemble method. Ensembling is a technique of combining multiple classification algorithms, also called base-learners. The final output of the ensemble method is then calculated by using plurality voting on the predictions of these base-learners. That is, the final prediction equals the class having the maximum number of votes. In our context, we will consider the following four different base-learners; the Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), and the Extreme Gradient Boosting (EGB) algorithms. We will experiment with different combinations of algorithms and pick the one that generates the best AUC (Area Under Curve).

In order to assess how the results of the statistical analysis will generalize to an independent data set, we randomly split the data into a training set and a testing set. The training set is used to fit the models, whereas the testing set is used to assess model validation. Furthermore, each base-learner is trained over a slightly different dataset, sampled from the training set with replacement. This is also known as bagging. As is common practice, the training set consists of 80 percent of the total sample and the remaining 20 percent is appointed to the testing set.

3.1 Evaluation

In order to evaluate the models performances, the whole procedure of partitioning the data set is repeated n times (where n is large). Collecting all the output will provide understanding in the dispersion and distribution of the validation measures and allows us to create, for example, confidence intervals.

One way to evaluate the performance of our ensemble and the individual base-learners is to determine the predictive accuracy. The predictive accuracy could be used as a statistical measure of how many observations in the testing set are correctly classified. This would amount to counting the true positives and true negatives compared to the total, including the false positives and false negatives. Nevertheless, the predictive accuracy must be approached with caution, because the predictive accuracy may only reflect the underlying class distribution of the dependent variable.

Therefore, we also consider another measure for overall model performance which is based on the ROC-Curve [4]. In the ROC graph the true positive rate (the sensitivity) is plotted against the false positive rate (equal to 1-specificity, where specificity is the true negative rate) for a threshold which varies from 0 to 1. A good model will realize a high true positive rate whereas the false positive rate remains small. Thus, for a good model the ROC-curve will rise steeply close to the origin, and flatten at a value near the maximum of 1. On the contrary, for a poor model the ROC-curve will lie adjacent to the diagonal where the true positive rate equals the false positive rate which implies that the model makes random predictions. Subsequently, the Area Under the ROC-Curve (AUC) is a well-defined measure for overall model performance [5]. Here, good models achieve an AUC approximating 1, while poor models will have an AUC near 0.5.

In order to get more insight in the actual predictions we will compute the confusion matrix. This matrix consist of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN)

For this paper, all of the statistical analyses were conducted in R on a computer running Windows.

3.2 Base learners

3.2.1 Logistic Regression

We apply a binominal logistic regression, since it allows prediction of a binary dependent variable based on the analysis of the independent variables.

The Akaike Information Criterion (AIC) provides a method for assessing the quality of a model through comparison of related models [2]. For more than one similar candidate models (where all of the variables of the simpler model occur in the more complex models), the model which corresponds with the smallest AIC should be selected.

For this algorithm we compare various models and eventually select 41 variables consisting of patient demographics, medical history and admission details as our regressors to predict hospital readmittion within 30 days.

3.2.2 Random Forest

A random forest is an emsemble learning method itself. As can be seen in figure 1, a random forest is a series of decision trees. Using plurality voting to find the class which appeared most often, the output of the random forest is the then the mode of the classes.

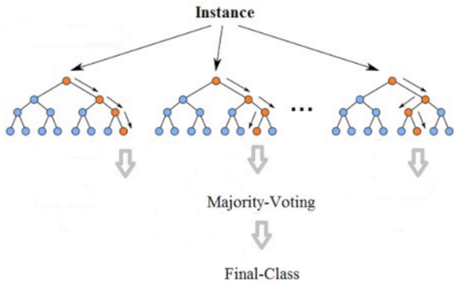

Figure 1: An example of a random forest

In addition, a variable importance test can be produced. However, since a random forest is not negatively effected by including features of low importance, reducing the number of features does not increase the accuracy of the random forest. Furthermore, seperate decision trees are prone to overfitting their training sets. Individually they have a low bias but usually a high variance. Using a random forest, this overfitting and high variance is reduced which provides the random forest with a much better accuracy in general. Obviously, the number of decision trees affects this result. For this purpose the accuracy was plotted against the number of trees used. As can be seen in figure **??**, using 10 or more trees provides the best results. Using more trees greatly increases computation time while hardly increasing the accuracy.

3.2.3 Naive bayes

Naive Bayes is a classification technique which represents a supervised learning method as well as a statistical method for classification. It assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes.

It can solve diagnostic and predictive problems. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The data set consisted of 9 numerical variables and 34 categorical variables. The model was trained with numerical and categorical variables as well as with all the variables converted to categorical variables. The Naive Bayes model was tuned during the training using the laplace smoothing from 0 to 3. Different variables were used obtained from variable importance method in random forest but the AUC went down when the variables were removed from the model. The most accurate model used all the variables in the data set where the numerical variables were converted to categorical variables and had lapace smoothing as 1.

3.2.4 Extreme Gradient Boosting

Extreme gradient boosting is a supervised learning technique based on principles of gradient boosting. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. XGBoost uses a more regularized model formalization to control over-fitting, which gives it better performance. Extreme gradient model was trained using the 10 fold cross-validation and was tuned using the parameters booster as gbtree , eta = 0.03, maximum depth as 5 and objective as binary logistic regression.

4 Results

First, it must be mentioned that as only 10% of the patients in our data set are readmitted within 30 days, a constant prediction of no readmission within 30 days is bound to be correct 90 percent of the time. However, this predictive accuracy is non-informative and useless. This problem is known as the accuracy paradox. Consequently, we will only consider the AUC as a valid measure to evaluate the models.

Testing with different combinations of algorithms, only one landed the best results. it includes logistic regressiong, Naive Bayes, Random Forest and Extreme Gradient boosting. As mentioned before, cross-validation is conducted with $n = 300$ iterations. From all these iterations, the resulting mean and the confidence interval of the AUC for our ensemble and base-learners are presented in Table (2). Moreover, the mean and confidence interval of the confusion matrix of the ensemble is shown in Table (3).

Table 2: Average AUC for the ensemble and the base-learners

	Ensemble	Logit	EGB	RF	NB
AUC	0.6145	0.6046	0.6073	0.5389	0.6474
95% Conf	(0.6135, 0.6154)	(0.6023, 0.6069)	(0.6065, 0.6082)	(0.5381, 0.5397)	(0.6468, 0.6479)

Note: The average AUC is based on $n = 300$ iterations

Table 3: Confusion matrix for the ensemble

	TP	FP	TN	FN
Ensemble	463	2910	9831	795
95% Conf	(452, 473)	(2843, 2978)	(9763, 9898)	(784, 805)

Note: based on $n = 300$ iterations

Note that the confidence intervals are calculated under the assumption that the AUCs and the elements of the confusion matrix follow the Normal distribution. Based on graphs visualizing the goodness-of-fit of the Normal distribution this assumptions are valid.

From Table (2) one can see that the Naive Bayes algorithm outperforms the other base-learners and even the ensemble itself. However, this difference is small compared to the ensemble. Moreover, the prediction ability of the logistic regression and Extreme Gradient Boosting are somewhat equal. The Random Forest performs the worst with an AUC of 0.5389.

Regarding the confusion matrix, Table (3) shows that, on average, the ensemble was able to correctly predict 463 patients out of the 13998 as being readmitted to the hospital within 30 days. In addition, on average 9831 diabetic patients who were not readmitted are accurately predicted. However, the ensemble wrongly predicts on average 2910 patients as being readmitted, whereas 795 patients are wrongly predicted as not being readmitted within 30 days.

In some other combinations of algorithms, we included K Nearest neighbours, but the AUC did not improve, table (4) and (5).

Table 4: Confusion matrix for other ensembles used

	TP	FP	TN	FN
All 5 algorithms	83	268	12445	1170
Logit, NB and knn	84	302	12432	1180
Logit, NB and RF	73	254	12486	11185

Table 5: Average AUC for tested ensembles and the base-learners

	Logit	NB	RF	Knn	EGB
All 5 algorithms	0.605	0.6322	0.543	0.5007	0.608
Logit, NB and Knn	0.538	0.631	-	0.502	-
Logit, NB and RF	0.582	0.635	0.534	-	-

5 Conclusion and Discussion

The Naive Bayes slightly outperformed some more complex methods. This should not come as a surprise, as Ashari et. al (2013)[1] showcased that Naive Bayes's good performance is caused by the zero-one loss function used in classification.

Eventhough the results obtained are above the base accuracy, these values are still low. Therefore, it is recommended to further improve this ensemble learning algorithm in order to improve the accuracy and AUC. The difficulty to obtain excellent accuracy is due to multiple reasons. First, most of the variables in the dataset were categorical with many categories or were largely missing values, the necessity to transform them into dummy variables or to completely remove some of them can result in the loss of information and interpretability. Secondly, transforming the readadmission variable produced an unbalanced data, where only 10% of the cases are positive (for readmission within 30 days). One way to address this is by oversampling from observations where readmission occured within 30 days and undersampling from observations where it did not. This technique is called Synthetic Minority Oversampling Technique (SMOTE). However, it did not improve the AUC. Another way is to assign weights to each class in the majority vote, in a way that favours the underrepresented class. Lastly, the presence of experts in the field diabetes could have helped in making a numerical representation of categorical variables in a more relevant way than creating dummy variables.

References

[1] A Min T Ahmad A, Iman P. Performance comparison between naïve bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. *International Journal of Advanced Computer Science and Applications*, 4(11), 2013.

[2] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[3] American Diabetes Association. Economic costs of diabetes in the u.s. in 2012. *Diabetic Care*, 2013.

[4] Hjalmar R. Bouma Fred Geus Anne H. Epema Jose Castela Forte, Marco A. Wiering. Predicting long-term mortality with first week post-operative data after coronary artery bypass grafting using machine learning models. *Journal of Machine Learning Reaserch*, 68, 2017.

[5] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

6 Appendix

Table 6: List of features and their descriptions in the initial dataset

Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	1%
Gender	Nominal	Values: male, female	0%
Age	Nominal	Grouped in 3 intervals: [0, 30), [30, 60) and [60, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	47%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	1%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	1%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
24 medication variables	Nominal	Medication taken during hospitalization	0%