Data Analytics and Predictions 2019

Project: Prediction of Mortality in SAH patients

Jaehwan Han, Seokhun Kim, Babak Soltanalizadeh

INTRODUCTION

Subarachnoid hemorrhage (SAH) is sudden bleeding into the subarachnoid space where the most common cause of spontaneous bleeding is a ruptured aneurysm. Symptoms include sudden, severe headache, usually with loss or impairment of consciousness. Approximately 17% of patients dies at the time of the bleeding and if they survive initially they can die after early rebleeding or have major complications. Based on the National FINRISK Study, there exists 437 cases of SAH where, 18% died before arriving to the hospital, 26.5% died in the first 30 days after the bleeding, and 5.6% died in the first year and after surviving in the first year, 65% of the 233 patients died over the follow-up period of approximately 9 years on average [1]. The number of patients of SAH is around 9 per 100,000 person-years.

Diagnosis is done through CT or MRI. CT scanning is a faster method, and since early diagnosis and faster initial treatment plays an important role in SAH patient management, it is usually the preferred method of diagnosis. Also, CT angiography has been used these days instead of catheter angiography for detecting aneurysms since it is a less invasive modality. Mortality of hospitalized patients has significantly declined over the past 3 decades. Many common therapies for SAH have created controversy, and various recent neuroprotective clinical trials have produced negative results.

By the way, the best practice for patients with SAH has been provided by the American Heart Association/American Stroke Association and the Neurocritical Care Society. The process is firstly controlling the bleeding and then treating the delayed cerebral ischemia secondary to cerebral vasospasm by induced hypertension and endovascular therapies where the main treatment for delayed cerebral ischemia in SAH patients are three common Vasopressors including dopamine, phenylephrine, and norepinephrine. Patients undergoing these three treatments are considered more severe cases and comprise almost 10% of SAH patients.

Datasets

In this project the Cerner Health Facts EMR dataset for SAH patients has been used which came from over 700 participating Cerner client hospitals and clinics in the United States.

The original database includes information on patient demographics, encounters, diagnoses, procedures, lab results, medication orders, vital signs, and other clinical observations however in this study we have used only the encounters, procedures, lab results, and medication orders data sets. The effective population queried included 4,838 total patients admitted to any hospital using Cerner as their EMR from the years 2000 to 2015 where we have been assigned to apply different statistical approaches in order to model the "Morality" in 2000 patients.

Data Pre-processing

The four data tables—encounter, medication, procedure, and lab—were merged using ENCOUNTER_ID as the key. Before merging, the medication and procedure tables were preprocessed as follows: (1) For each ENCOUNTER_ID, we counted the number (i.e., frequency) of each medication or procedure with the Group_by method, (2) The medication or the procedure (categorical) variable was one-hot-coded into dummy variables of unique number of medications or procedures, and each dummy variable was multiplied with the count column so that the value of the variables has count information, and (3) All the one-hot-coded variables were collapsed for each ENCOUNTER_ID, so that the data table has only 1 observation for each ENCOUNTER_ID. The lab data table was preprocessed in the same steps, except that we obtained mean, maximum, and minimum values of each lab procedure instead of frequency because the measured lab results were continuous.
After all the preprocessing, (1) the encounter and medication data tables were merged, (2) the procedure data table was merged further, and (3) the lab data table was merged finally. The final merged data table had 1370 observations.

Although we calculated mean, maximum, and minimum of lab results for each ENCOUNTER_ID, due to high multicollnearity between the three, we dropped maximum and minimum, and kept only mean values for analyses. Finally, the outcome variable 'Mortality' was converted into a factor variable for the following classification models.

Analyses using Machine Learning Models

1. Lasso for variable selection

The greater number of predictors (p = 2,284) than the number of observations (n=1,370) might give rise to problems such as no unique parameter estimates, under-identified model, and poor prediction. Therefore, we performed a Lasso including all the variables, in order to select important ones that do not shrink to 0 at the optimal shrinkage parameter lambda.

The result of Lasso with different lambdas on 10-fold cross-validation (CV) showed that the optimal lambda was 0.0093 and its CV mis-classification error rate was 13.21%. Among the

2,284 that were entered, 255 variables did not shrink to 0, and thus, were selected our predictor subset.

Additionally, we tried 1 more Lasso on the selected subset of 255 predictors, to check if there would be any other predictors to be dropped. However, the 2nd Lasso resulted in only 1 more variable reduction, and thus, we decided to stick to the 1st Lasso subset.

2. Logistic regression

As the first step of model fitting, we performed a Logistic regression with the 255 predictors, and also computed its 10-fold CV mis-classification error rate.

The training error rate was 1.53%, and the CV error rate was 12.96%, indicating that the linear classification boundary already works well.

3. Linear discriminant analysis

To test another linear classification boundary that works well for separated classes, we performed a Linear discriminant analysis with the 255 predictors, and also calculated its leave-one-out CV error rate.

The training error rate was 3.36%, but the CV error rate was 49.20%, possibly due to multi-collinearity of predictors, violation of equal variance-covariance matrix between the 2 classes, and the not-clear-separation between the 2 classes. That is, the result implies that even though the classification boundary could be close to linearity, the 2 classes may not be separable, and the distributions of predictors may be far from normality.

4. Ridge logistic regression

Although the performance of logistic regression was quite good, we still have 255 predictors for 1370 observations, which could have produced an overfitting of the logistic model. Thus, to reduce the variance of a many-parameter linear model by giving a slightly higher bias, we performed a Ridge logistic regression with different shrinkage parameter lambdas on 10-fold CV.

The result showed that the optimal lambda was 0.0175, and the CV error rate was 7.52% (i.e., CV accuracy ~ 92.5%), which improved that of logistic regression quite a bit. Additionally, the training error rate was 2.55%, demonstrating that the Ridge model is more biased (i.e., higher training error rate) but has a less variance (i.e., smaller CV error rate) than the logistic regression model without a shrinkage.

5. Random forest

Although the linear models worked quite well, we also decided to fit nonparametric tree-based models to check if they provide a better prediction performance. We did not try a single tree model with pruning process and a bagging model due to their possibility of overfitting.

Instead of those two, we performed a Random forest model with different predictor subset sizes for each split (mtry) and different tree numbers for bootstrap resampling (ntrees, to check which combination produces the best OOB error rate, which corresponds to CV error rate.

The best CV error rate for OOB (12.63%) was achieved for the combination of mtry of 90 and ntree of 500. Moreover, the training error was 0%, indicating that the model was in fact overfitted. However, note that in spite of overfitting, the CV error was rather good, thanks to the Random forest's averaging method across many trees, which gave rise to a lower variance.

6. Gradient boosting

Also, we performed a tree-based Gradient boosting to test if this slow learning method works well for the data. We used combinations of shirinkage parameters of [0.001, 0.005, 0.1, 0.2, 1], interaction depths of [1, 2, 4], and ntrees of [200, 400, 600, 800, 1000].

The best CV error rate (11.46%) was achieved for the combination of shrinkage of 0.1, interaction depth of 1, and ntree of 600. Additionally, the training error rate was 6.50%. This Gradient boosting method did not overfit the data as shown by a relatively higher training error rate, and its estimate of test error rate (i.e., CV error rate) was better than that of random forest. However, their performance is not quite better than logistic regression (12.96%), which could be actually a little overfitted with 255 predictors, and quite worse than Ridge logistic (7.52%), which prevented overfitting to some extent by giving bias.

These findings that nonparametric tree-based models with their own methods of reducing overfitting did not perform better than a shrunken linear model, reinforces the idea that the true classification boundary is close to linearity.

7. Support vector classifier (SVC)

Further, we performed a linear support vector classfier with different costs of [0.01, 0.1, 0.5, 1, 10, 100]. The best CV error rate (10.81%) was achieved with the cost of 0.1. Additionally, the training error was 2.04%.

The estimate of test error rate (i.e., CV error rate) of linear SVC is better than that of logistic (12.96%) because it has a higher bias and lower variance due to the low value of cost (0.1). The participation of many support vectors (277 out of 1370) due to this low cost resulted in a little higher training error and quite smaller CV error rate than logistic regression.

However, this result is still worse than that of Ridge logistic regression, which could give an even more bias than this model and reduce quite more variance eventually.

8. Support vector machine (SVM using Radial kernel)

Finally, we performed a nonlinear radial SVM. As expected from the previous findings, this nonlinear classifier did not work very well for the data. The best CV error rate (40.36%) was achieved for sigma (gamma) of 0.01 and cost of 10, and the training error was 0.07%.

This model was also overfitted as random forest, but produces a way higher CV-error than random forest due to its inherent high nonlinearity; the random forest is a nonparametric model with high flexibility, but it can also be quite well fitted to linearity, owing to its averaging method.

Conclusion

The results from the series of machine learning models presented that a Ridge regression model produced the best CV misclassification rate for the data.

Since CV error rate is a reliable estimator of test error, we can conclude that the Ridge regression model with the selected 255 predictors will be the best model (of all that we have tried) that can predict a new test data with the best accuracy.

The better performance of linear models, especially with shrinkage being applied (Ridge is the best and linear SVC is the next), and the finding that the complicated non-parametric methods (random forest and boosting) produced only slightly better results than the non-shrunken linear model (logistic) and explicitly worse results than the shrunken linear models, indicate that the true classification boundary for mortality is quite close to linearity.

In the future, we may apply a new method, if any, for further variable selection because the current 255 variables are still a lot, and therefore, undermine the interpretability of models. If we do not have a new method, we may rely on the (sorted) variance importance metric from random forest or gradient boosting to further select more important variables (e.g., 50 or 100), and may apply Ridge regression or other shrinkage methods on them, hoping to get a better prediction accuracy. However, since this method is quite arbitrary, we did not pursue it further at this point.