Predicting Stroke Recurrence

Robert P. Yerex, Ph.D

October, 2014

UVA Medical Center Accountable Care Organization

**Objectives**. To develop a predictive model using statistical learning predict recurrence of strokes among patients previously hospitalized for one or more strokes.

**Data Sources.** The Copenhagen Stroke Study data, which includes longitudinal patient information from 652 individuals hospitalized with 24 hours of stroke onset. The first study to make use of the data from this cohort investigated the influence of age on stroke outcome [3]. A ten year follow-up study was conducted to elucidate the relationship between total serum cholesterol and stroke severity [4]. The follow-up study augmented the original data set with information on time to subsequent stroke and/or all-cause mortality.

**Statistical Learning / Predictive Model.** A Random Survival Forest (RSF) algorithm was used to predict each patient’s hazard function post discharge.

**Model Performance.** The final version of the model achieved a Boot632plusErr rate of 16%

# Background

Application of statistical learning techniques to the development of predictive algorithms applied to patient outcome data involves a specialized form of time to event modeling due to both right and left censoring of the data. Analytical modeling of this type of data has, for several decades, used some form of *Survival Analysis* technique [1]. These techniques were developed with an eye towards interpretation and analysis of the effect of various static and time dependent covariates included in the model on the time to the event of interest. Both parametric and non-parametric approaches have been used successfully. The non-parametric techniques, while the easiest to apply, are generally the weakest in terms of interpretation strength and are considered to be overly conservative in most cases. Highly complex parametric, multi-variate, time dependent covariates models have been developed that are much more useful for interpretation. Any of these approaches can be used to build predictive models. However, as with many techniques, a desire for greater interpretability greatly sacrifices predictive accuracy [2]. When the goal is the development of the most accurate predictions possible, regardless of causal interpretability, there are a number of techniques available that can be effectively applied. What follows is an example of applying one such technique, *Random Survival Forests* to a data set from the *Copenhagen Stroke Study.*

# Data

The Copenhagen Stroke Study data includes longitudinal patient information from 652 individuals hospitalized with 24 hours of stroke onset. The first study to make use of the data from this cohort investigated the influence of age on stroke outcome [3]. A ten year follow-up study was conducted to elucidate the relationship between total serum cholesterol and stroke severity [4]. The follow-up study augmented the original data set with information on time to subsequent stroke and/or all-cause mortality.

Table 1: Sample of Data

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | sex | hypTen | ihd | prevStroke | othDisease | alcohol | diabetes | smoke | atrialFib | hemor | strokeScore | cholest | time | status |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 80 | female | no | no | no | yes | no | no | yes | no | no | 55 | 5.2 | 1369 | 1 |
| 74 | male | no | yes | yes | no | yes | no | no | no | no | 54 | 5.6 | 1657 | 1 |
| 71 | male | no | no | no | no | yes | no | no | no | no | 58 | 6.9 | 3670 | 0 |
| 65 | female | yes | no | no | no | no | no | no | no | no | 44 | 5.4 | 4262 | 0 |
| 69 | male | yes | no | yes | no | no | no | yes | no | no | 54 | 4.7 | 1157 | 1 |
| 72 | female | yes | yes | no | no | yes | no | yes | no | no | 58 | 5.7 | 4245 | 0 |
| 76 | female | yes | no | yes | yes | no | no | yes | no | no | 53 | 7 | 4256 | 0 |
| 80 | male | yes | no | no | no | yes | no | no | no | no | 54 | 5.9 | 400 | 1 |
| 72 | female | no | no | no | no | no | no | no | no | no | 50 | 5.1 | 2986 | 1 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

To illustrate the nature of the attributes that are candidates for feature selection, Table 1 presents the attributes for a random sample of observations. The nature of the features to be used as covariates has an important impact on algorithm selection and an initial investigation into their distribution and cross-correlations is the first step in building a predictive model.

Figures 1 and 2 present the density plots for *age* and *total serum* cholesterol within the study population respectively. The third continuously distributed feature is strokeScore, or *Stroke Score*, which is based on the Scandinavian Stroke Score, a subjectively assigned score/index with three sub-indices [5]. The somewhat odd nature of this combined index is shown in. Figure 1. Subjective, multipart scales, often show this type of distributional behavior. If treated as a continuous variable, such a feature can be increase the overall “noise” level of the model. An easy fix is to collapse the buckets into ones that show smoother distributional behavior.



Figure 1: Distribution of Age



Figure 2: Distribution of Total Serum Cholesterol



Figure 3: Distribution of Stroke Score (strokeScore)

# Modeling

The initial model included all the 13 features shown in Table 1. The target is event occurrence, either another stroke, or all-cause mortality, at a given point in time. Essentially this can be thought of as a classification model across a discrete time scale. Three algorithms were tested for best fit; Cox Regression with Backward Elimination (BE), a Random Survival Forest, and a Conditional Forest.

## Prediction Accuracy

There are a number of methods to estimating prediction accuracy for continuous or categorical outcome targets [6]. Survival predictions do not easily fit into any of the typical approaches as the accuracy of a particular model depends on the outcome day against which it is being tested. Rather than simply predicting a single point in time outcome, or a number of days to a particular event, the interest is in describing a predicted *survival curve*. From such a curve, single day predictions can be extracted. One approach to estimating the accuracy of prediction when the output is a survival curve is to use *Prediction Error Curves* [7]. This approach involves calculating accuracy for each time point within the target time span using the *Brier Score* [8] and then integrating the derived empirical function over the time span desired, which Morgensen et. al. refer to as the *Integrated Brier Score*.

Table 2: Accuracy Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AppErr | BootCvErr | NoInfErr | Boot632plusErr |
| Null Model | 0.20 | 0.20 |  |  |
| Cox Regression | 0.16 | 0.23 |  |  |
| Random Survival Forest | 0.08 | 0.18 | 0.26 | 0.16 |
| Conditional Forest | 0.15 | 0.18 | 0.22 | 0.17 |
|  |  |  |  |  |

Table 2. presents the IBS values for each of the models tested, as well as null model (Kaplan Meier with no covariates). For the Null and Cox regression models, there is a closed form solution to estimating the Brier Score for each point in time.



Figure 4: Prediction Error Curves

For both the Random Survival Forest and Conditional Forest models, cross validation is accomplished using a technique known as *Boot 632 + Error*. An in depth description of this approach to cross validation can be found in [9].

## Variable Importance

The impact of each feature on the accuracy of the over-all model can be computed in a number of ways. The approach chosen here is known as *VIMP;* an in depth description of which can be found in [10].

Table 3: Variable Importance

|  |  |
| --- | --- |
| Feature | VIMP |
| Age | 0.039 |
| strokeScore | 0.026 |
| atrialFib | 0.005 |
| sex | 0.004 |
| cholest | 0.003 |
| alcohol | 0.003 |
| hypTen | 0.002 |
| smoke | 0.002 |
| prevStroke | 0.000 |
| hemor | -0.001 |
| diabetes | -0.002 |
| ihd | -0.003 |
| othDisease | -0.004 |

The VIMP measures for the 13 initial features are shown in Table 3.

## Model Tuning

The process of iteratively tuning a model involves a number of steps, each leading to an improved model in terms of accuracy of prediction. For this example, the next step was to 1) settle on an algorithm (Random Survival Forest), and 2) reduce the feature set to the most parsimonious without losing significant accuracy.

Having removed features; hemor, diabetes, ihd, and othDisease, the model was run again with the comparative accuracy results shown in Table 4. With the curves presented in Figure 5.

Table 4: Accuracy of Reduced vs. Full Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AppErr | BootCvErr | NoInfErr | Boot632plusErr |
| Full Model | 0.08 | 0.18 | 0.26 | 0.16 |
| Reduced Model | 0.08 | 0.17 | 0.26 | 0.15 |



Figure 5: Prediction Error Curves

# Conclusions

*Random Survival Forests* provide a powerful tool for developing predictive models of time to event patient outcomes where there is right censoring, while the *Brier Score*, and *Integrated Brier Scores*, calculated within a *632+ Bootstrap* protocol can be effectively use to calculate and compare expected prediction error rates across time.

# Citations

[1] M. J. Crowder, *Multivariate Survival Analysis and Competing Risks*, 1 edition. Boca Raton: Chapman and Hall/CRC, 2012.

[2] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, “Random Survival Forests,” *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, Sep. 2008.

[3] H. Nakayama, H. S. Jørgensen, H. O. Raaschou, and T. S. Olsen, “The influence of age on stroke outcome. The Copenhagen Stroke Study.,” *Stroke*, vol. 25, no. 4, pp. 808–813, Apr. 1994.

[4] T. S. Olsen, R. H. B. Christensen, L. P. Kammersgaard, and K. K. Andersen, “Higher Total Serum Cholesterol Levels Are Associated With Less Severe Strokes and Lower All-Cause Mortality Ten-Year Follow-Up of Ischemic Strokes in the Copenhagen Stroke Study,” *Stroke*, vol. 38, no. 10, pp. 2646–2651, Oct. 2007.

[5] “Multicenter trial of hemodilution in ischemic stroke--background and study protocol. Scandinavian Stroke Study Group,” *Stroke J. Cereb. Circ.*, vol. 16, no. 5, pp. 885–890, Oct. 1985.

[6] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, 2013 edition. New York: Springer, 2013.

[7] U. B. Mogensen, H. Ishwaran, and T. A. Gerds, “Evaluating Random Forests for Survival Analysis Using Prediction Error Curves,” *http://www.jstatsoft.org/v50/i11/paper*, Sep. 2012.

[8] R. M. B. Young, “Decomposition of the Brier score for weighted forecast-verification pairs,” *Q. J. R. Meteorol. Soc.*, vol. 136, no. 650, pp. 1364–1370, 2010.

[9] B. Efron and R. Tibshirani, “Improvements on Cross-Validation: The 632+ Bootstrap Method,” *J. Am. Stat. Assoc.*, vol. 92, no. 438, pp. 548–560, Jun. 1997.

[10] H. Ishwaran, “Variable importance in binary regression trees and forests,” *Electron. J. Stat.*, vol. 1, pp. 519–537, 2007.