An Evaluation of Data Mining Methods Applied to Adverse Events for Clinical Trials

Xiao-Song Zhong, UMBC, Paul Schuette, Scott Komo, Andrejus Parfionovas, FDA/CDER/OTS/OB

Disclaimer: The views and opinions in this poster are those of the authors, and do not represent policy or guidance of the US FDA.

Introduction

Data mining consists of novel ways to find unexpected relationships and to summarize the large observational data sets; Data Mining can include interaction effects.

- There are few papers on the application of Data Mining methods to AE analyses for clinical trials.
- There are few papers that make comparisons among different data mining software packages.

Project Goals

- Determine the strength and weaknesses of the two software packages under consideration.
- Identify unexpected AEs.
- Identify the AEs associated with higher risk groups.
- Identify potential classification and machine learning algorithms for AE analyses of clinical data.

The Basic Data Mining Process

- Data Preparation is critical and more time was spent organizing, cleaning, and generating indicator variables for AEs than any other steps.
- Splitting the data. Data was randomly split into a training data set (30% of total) and a testing data set (70%) of total.
- Explore: Searching for anticipated relationships, unanticipated trends and anomalies in order to gain understanding and ideas.
- Modify: Creating, selecting, and transforming the variables to focus the model selection process.
- Model:Using the analytical tools to search for a combination of the data that reliably predicts a desired outcome.
- Assess: Comparing the models using appropriate metrics to determine which appears to be best.

Data Source

- Submission underwent FDA review and was approved.
- Number of subjects was relatively large. Phase 3 clinical trial data came from multiple trials, multiple regions.
- Primary data sources were analysis data in ADaM-like files.
- Analysis Variables: Age, Race, Geographic Region, Sex, Treatment Arm, Cancer Population.

Data Mining Algorithms used

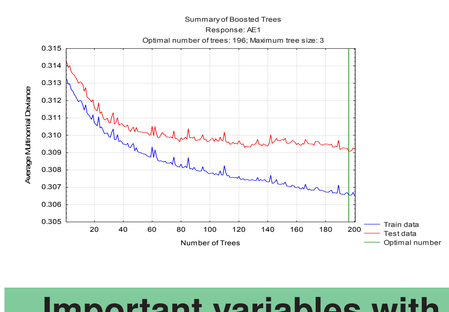
- SVM (Support Vector Machines)
- CRT (Classification and Regression Tree)
- MARS (Multivariate Adaptive Regression Splines)
- Boosted trees
- Exhaustive CHAID
- Best-Subset and Stepwise General Discriminant Analysis ANCOVA
- SANN (Select Automated neural network)
- Random Forests

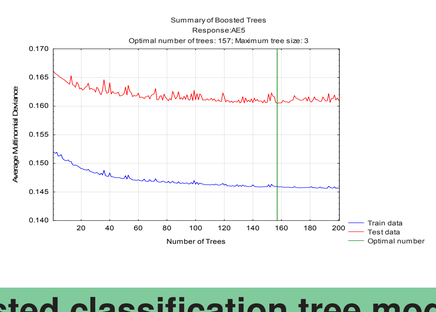
Results
Misclassification error rate from STATISTICA:

The table shows misclassification error rate from two AEs of nine considered using an EDA tool.

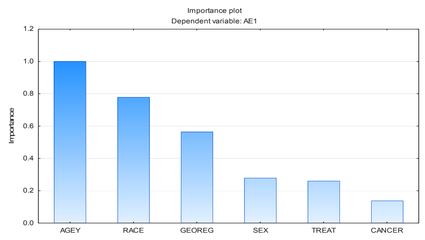
Model(AE1)	Training error %	Testing error %	Model(AE5)	Training error%	Testing error %
Neural network	3.87	3.34	Neural network	0.6	0.7
SVM	3.87	3.34	SVM	0.6	0.7
Boosted tree	3.87	3.34	Boosted tree	0.6	0.7
Random forest	27.63	30.17	Random forest	4.89	5.3
Class&RT	47.35	50.09	Class&RT	38.87	42.35

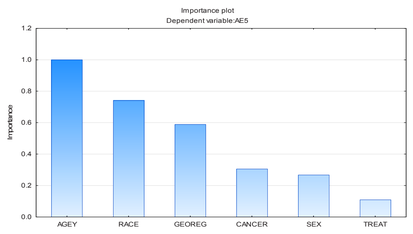
Risk estimates summary of boosted classification trees model from STATISTICA:





Important variables with Boosted classification tree models in STATISTICA:

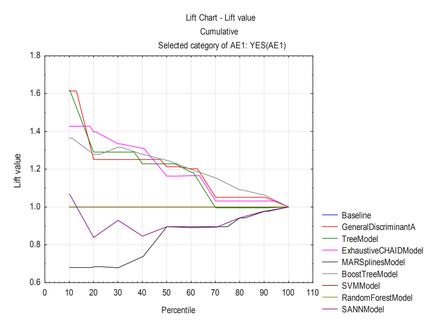


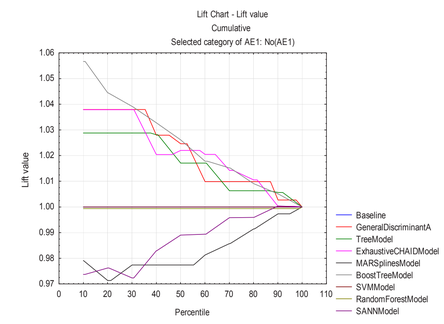


Age,race and Geographic Region are the most important predictors for the AE5, which is consistent with other AEs.

Summary results of different algorithms for AE1

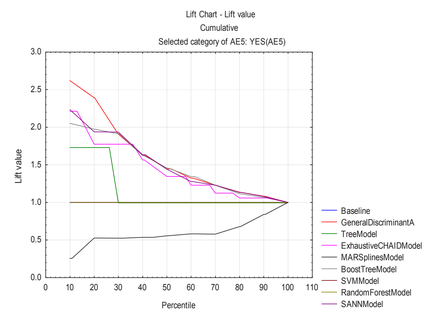
Compute overlaid lift charts from all algorithms

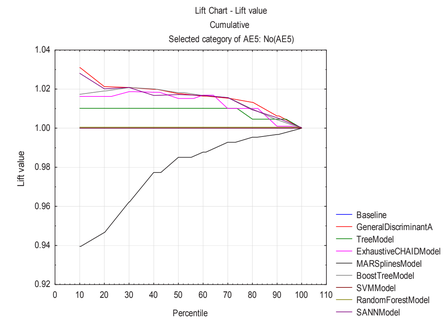




Summary results of different algorithms for AE5

Compute overlaid lift charts from all algorithms

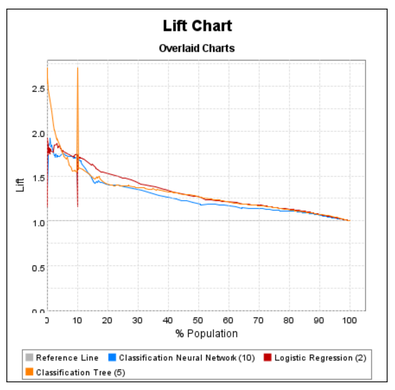




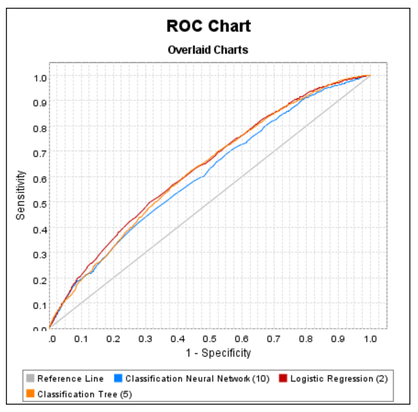
A Boosted classification tree appears to be the best model for both AEs.

Classifation of TIBCO SpotFire Miner

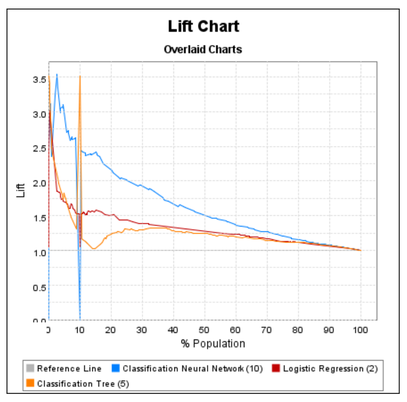
AE1



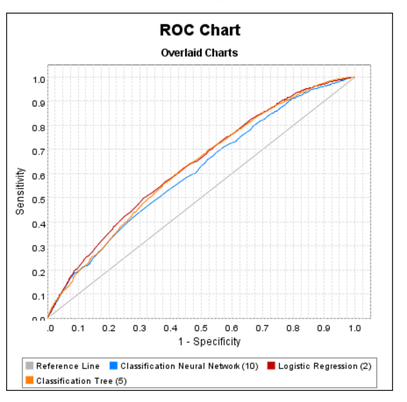
AE1



AE5



AE5



No one algorithm performed uniformly well using Spotfire Miner, possibly due to the limited number of algorithms used.

Conclusions.

- One algorithm (Boosted Trees) performed well for all AEs under consideration. This method was available with STATISTICA, but not with Spotfire Miner.
- There are more options for STATISTICA but more data cleaning was required.
- Data mining methods can be helpful in identifying covariates related to adverse events.
- Predictive modeling and data mining have the potential to identify groups at high risk for specific AEs. The regulatory impact of predictive modeling is not clear at this time.

References

- Austin PC. A comparison of classification and regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Statistics in Medicine. 2007;26:2937-2957.
- Austin PC, Le DS. Boosted classification trees result in minor to modest improvement in the accuracy in classifying cardiovascular outcomes compared to conventional classification trees. Am.J Cardivas Dis. 2011; 1(1); 1-15
- Freund Y, Schapire R. Machine Learning: Proceedings of the Thirteenth International Conference. San Francisco, CA: Morgan Kauffman; 1996. Experiments with a new boosting algorithm; pp. 148-156.
- Nisbet R, Elder, J,Miner Gary. Statistical Analysis and Data Mining Application. (ACADEMIC PRESS)

CDER Office of Biostatistics

Mail: Paul.Schuette@fda.hhs.gov

Supported by ORISE