Timothy Brathwaite
Kevin Johnson

May 12th, 2014
Final Project CS 289A

Introduction:

Numerous authors have written about either the potential or realized benefits of combining classification and regression trees (CART) and logistic regression models. As early as 1998, Steinberg and Cardell noted that CART and logistic regression models have complementary strengths: CART is adept at representing complex structures in one's data, while logistic regression models can parsimoniously represent linear-in-parameter relationships that hold across one's entire dataset. It was also noted by Steinberg and Cardell, that early methods of combining logistic regressions and CART models were not terribly successful because they attempted to run logistic regressions within the terminal leaf nodes of the decision tree, where the variability within the feature vectors and the size of the available observations had been greatly reduced.

By 2000, researchers such as Kuhnert, Do, and McClure were using decision trees to aid in feature selection for logistic regression analysis of motor vehicle injury data. In 2006, researchers began to successfully use decision trees to augment logistic regression models in the medical realm by building trees specifically to increase the fit of the logistic regression and then incorporating the leaf nodes of the tree in to the model as dummy variables (Su). The technique of placing leaf-node dummy variables in the logistic regression models slowly entered the transportation field in the context of mode choice, with a trio of papers published between 2007 and 2011 (Timmermans and Arentze, 2007; Kim, 2009; Kim and Kim, 2011), but it remains seldom used among transportation researchers as of today. Meanwhile, the decision tree dummy variables + logistic regression combination has continued to be adopted by researchers in various fields for tasks such as bankruptcy prediction (Brezigar-Masten and Masten, 2012).

In this paper, we apply and compare the prediction results of the hybrid CART-Logistic Regression model described by Steinberg and Cardell to the traditional methods of using a single decision tree or using a stand-alone logistic regression. The classifiers will be compared on the basis of their total prediction accuracy, their sensitivity, their specificity, their precision, and their precision versus recall curves. Additionally, the learning rates of these classifiers will be compared.

Data:

The data for this project from the 2011 and 2012 National Automotive Sampling System (NASS) General Estimates System (GES). The NASS GES database is a national, stratified sample of police reported crashes involving at least one motor vehicle. For each crash, the data is separated into multiple files such as the "accident" file, the vehicle file, the person file, the "parkwork" file, etc. These files contain complementary pieces of information, at varying levels of specificity, and they can be linked by means of unique identifiers that are present in each file.

For this project, the many crashes were organized at the level of the person file, so there was one feature vector for each person. The particular people of interest in this project were cyclists who were injured in a crash, and who suffered either an evident and non-incapacitating injury, or cyclists who suffered either a evident and incapacitating injury or a fatal injury. As a result, each feature vector contains information about the injured cyclist, about the vehicle which struck the injured cyclists, about the occupants of the vehicle which struck the cyclist, and about the general circumstances in which the crash took place.

Notable features of the dataset include the fact that of the 331 variables present in the dataset, 11 of them were measured on a nominal scale. Additionally, the dataset is unbalanced, with approximately 74% of the data belonging to "Class 0", the people who suffered evident and non-incapacitating injuries. The remaining approximately 26% of observations correspond to people who suffered incapacitating or fatal injuries.

Goal:

The goal of this project is to take the NASS GES data, and by using three different classification algorithms, compare the results of predicting the probability of a cyclist being in an incapacitating or fatal injury, given that the cyclist is injured in a crash at all. In terms of risk decomposition within the field of traffic safety, we want to model the probability of serious injury or fatality given injury, $P(S \cup F | I)$ where "S" is the event that a cyclist suffers a severe injury (i.e. an incapacitating injury), "F" is the event that a cyclist suffers a fatal injury, and "I"is the event that a cyclist suffers an injury of any severity from evident and incapacitating up to a fatal injury.

The three algorithms to be compared are CART, logistic regression, and the hybrid technique of Steinberg and Cardell which combines CART and logistic regression models.

Methodology:

In growing the decision trees, maximum information gain will be used as the splitting criteria, and the trees will be grown to full depth before being pruned according to "reduced error pruning." Due to the large number of potentially important variables, feature selection for the logistic regression model will be performed via a modified version of the "purposeful selection" algorithm described by Hosmer, Lemeshow, and Sturdivant (2013) in Applied Logistic Regression. First, all of the variables will be tried one at a time in a univariate logistic regression, and all the variables with p-values of less than 0.2 will be retained. Once all variables have been entered into a univariate logistic regression, all of the variables selected in the first step will be simultaneously entered into a large multivariate logistic regression. One at a time, if any variables have a p-value of greater than 0.05, the variable with the highest p-value will be removed and then the model will be re-fit with the remaining variables. This will continue until the model only contains variables with significant p-values. There are further steps to the purposeful selection algorithm, but they were curtailed due to time constraints.

For the hybrid CART-Logistic Regression model, a decision tree will be built as above. Then, for leaf nodes where the training observations were of more than one choice, the leaf nodes will be added to the logistic regression as dummy variables. Then, the modified purposeful selection algorithm (described above) will be repeated with two differences. First, the node-dummy variables will be forced to stay in the model while statistically insignificant variables are removed, and second, the intercept will be suppressed for model identifiability purposes.

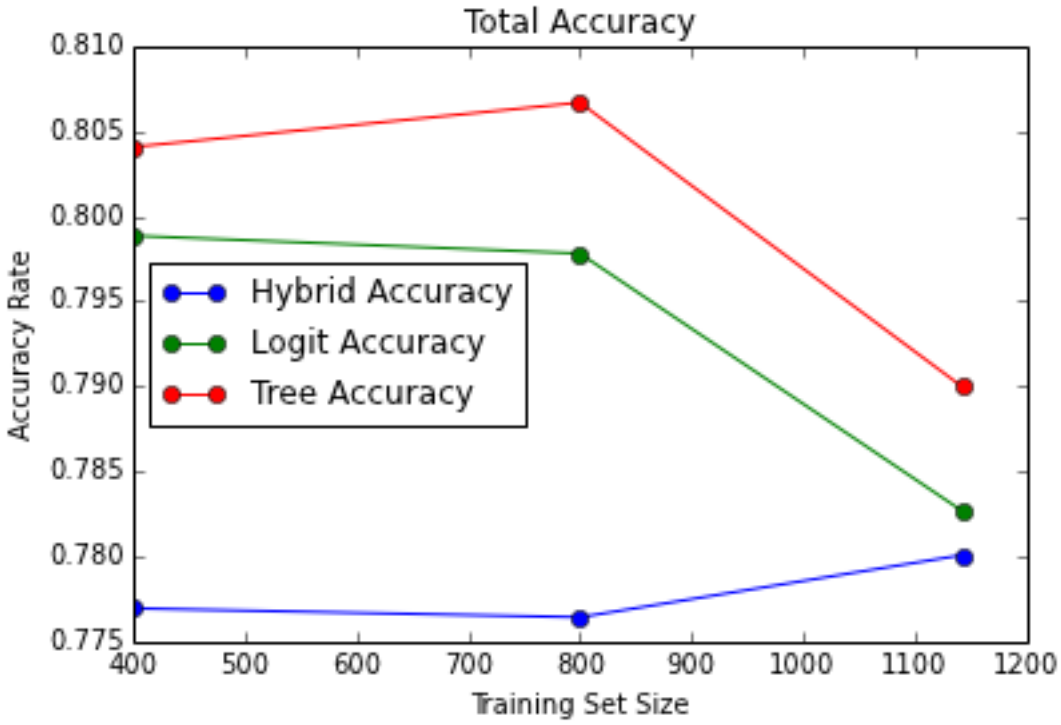
Twenty percent, or 285 observations, of the 1,427 2011 NASS GES observations was randomly removed to act as a pruning set for the decision tree. The remaining eighty percent or 1,142 data points were used to train both the decision tree, the pure logistic regression, and the hybrid-logistic regression. The test set used to evaluate the predictive performance of the three algorithms was the 2012 NASS GES data. Classification was performed using a threshold of 0.5. Observations for which the classifier assigned a probability of 0.5 or greater for the probability of being in an incapacitating or fatal accident were predicted as belonging to Class 1. Conversely, observations that were assigned a probability of less than 0.5 for suffering a severe injury were predicted as belonging to Class 0.

For the hybrid CART-logistic regression model, the classification was only slightly different. If the observation from 2012 was sorted by the tree into a leaf which for which the training data was all of one class, then the 2012 observation was assigned a probability of 1 for being in the same class as the

training data. If the 2012 observation was sorted into a leaf which contained training examples of both classes, then the hybrid CART-logistic regression model was used to classify the observation.

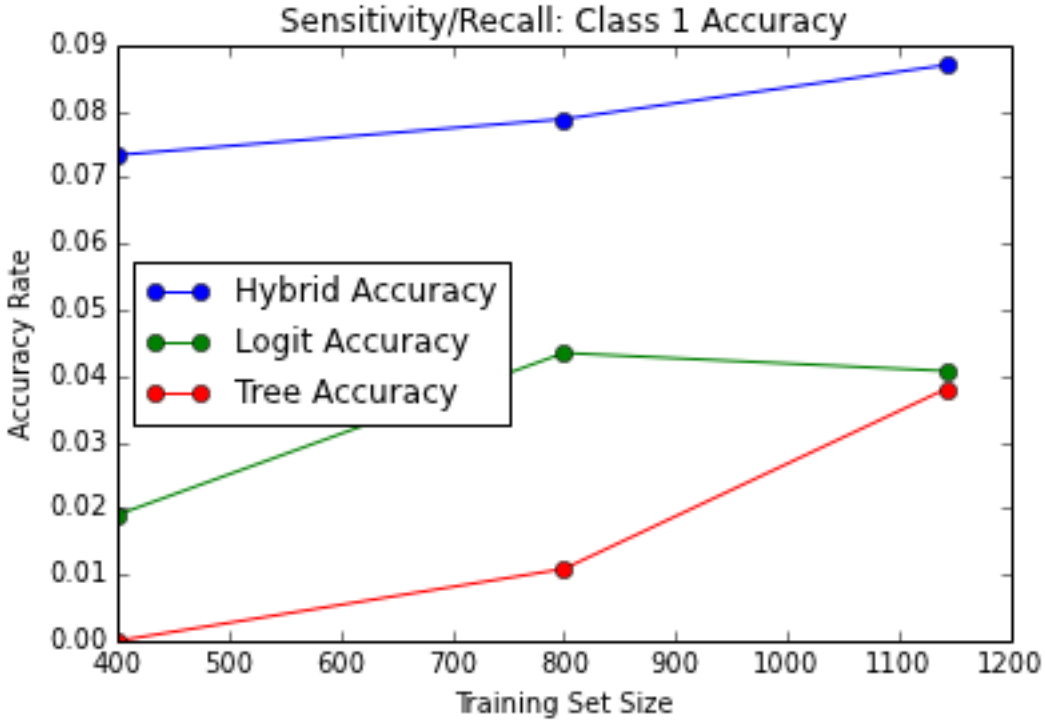
Results:

For sample sizes of 400, 800, and 1142, the total accuracy, sensitivity, specificity, precision, and precision versus recall graphs are shown below based on predictions of the 2012 data.

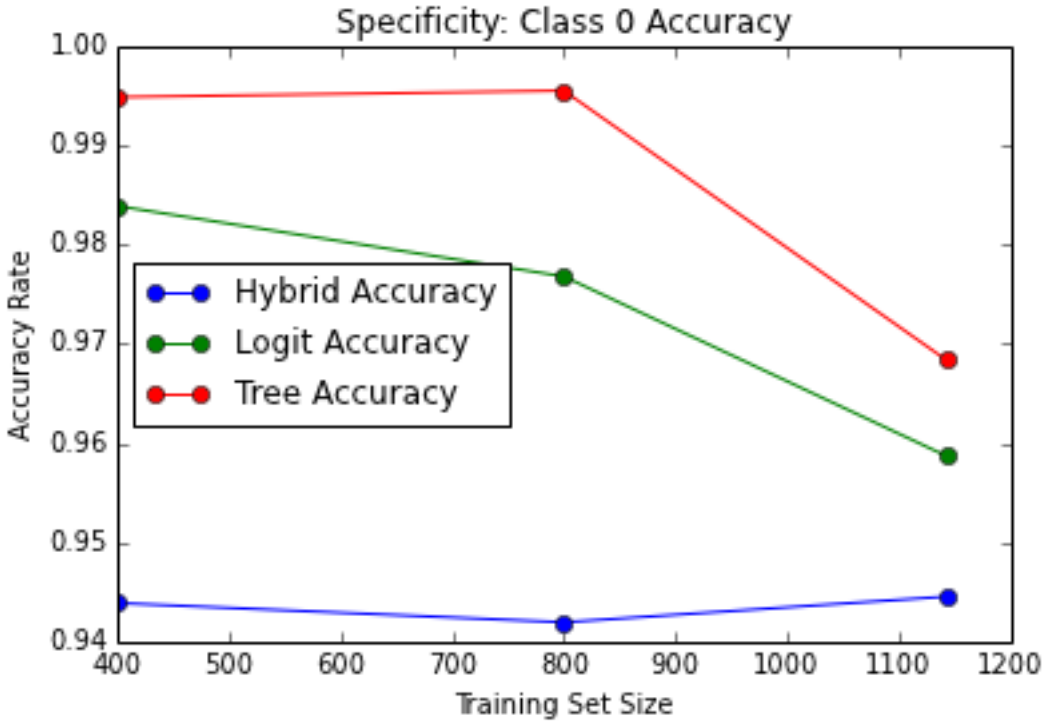


From the total accuracy chart we can see that in terms of overall predictive accuracy, the decision tree always performs best, the logisitic regression (i.e. the logit model) always performs second best, and the hybrid CART-logistic regression has the worst overall predictive accuracy.

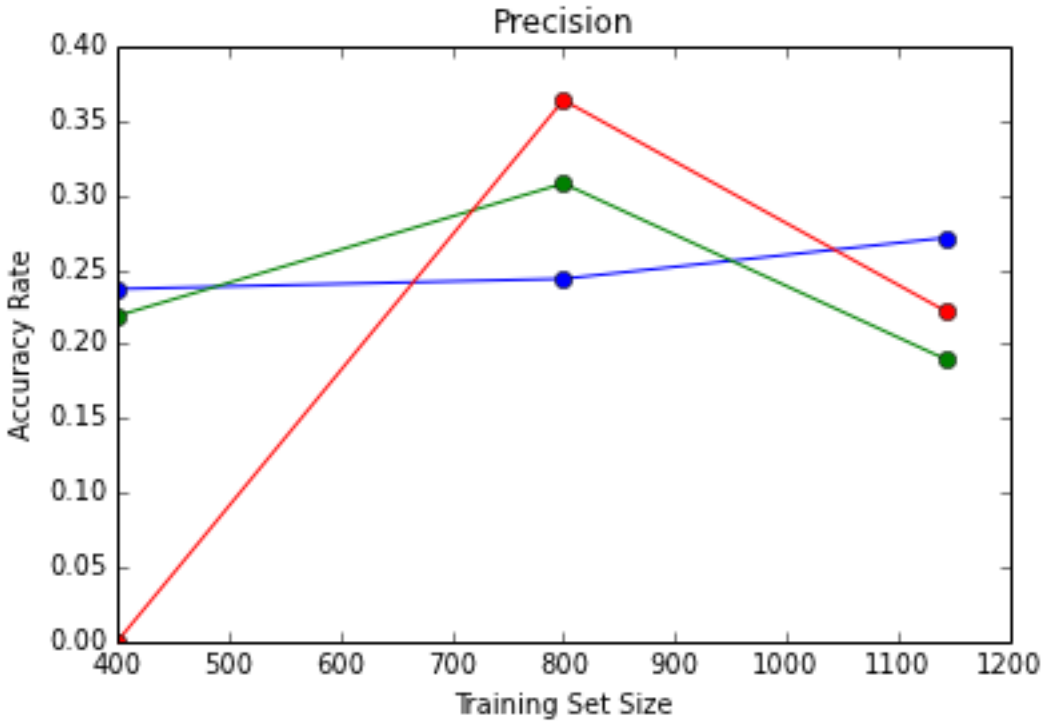
In terms of sensitivity however (see figure below), the story is exactly reversed with the hybrid CART-logistic regression performing the best, the pure logistic regression performing second best, and the decision tree performing the worst. The explanation for these results is that the imbalanced nature of the dataset means that it is always better to predict the majority class, which in our case is Class 0 where only minor injuries are incurred. When a classifier makes a number of predictions of observations being Class 1, it is easy to be wrong and thus lower one's overall accuracy. This is a classic case of classifiers tending to predict the majority class when trained on imbalanced data. The hybrid CART-logistic regression counteracts this tendency.



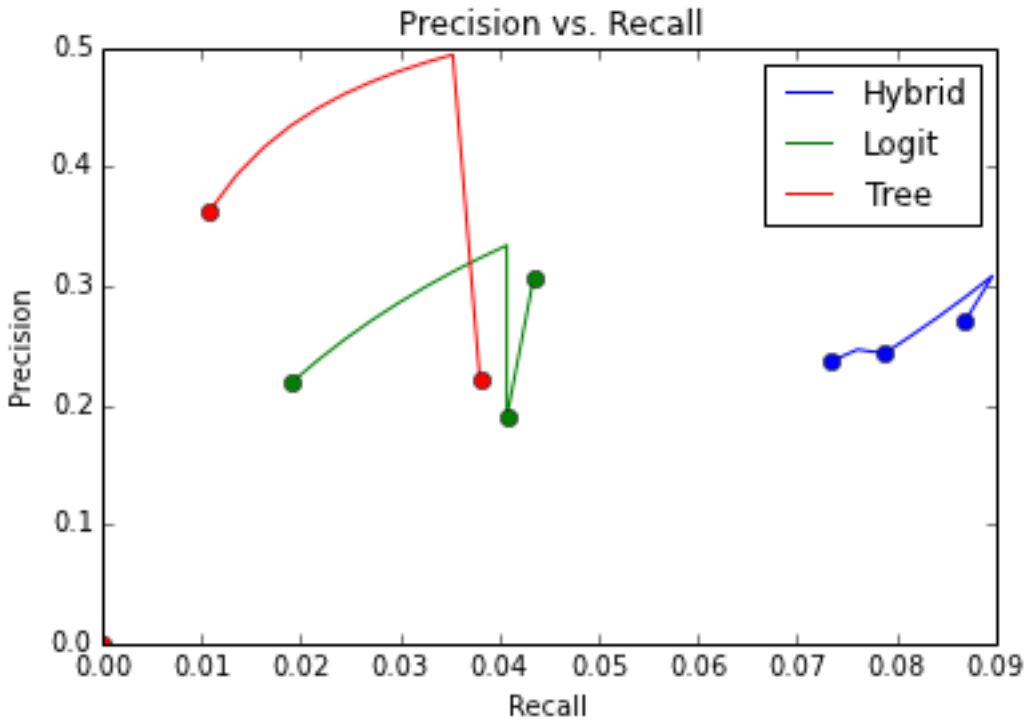
As per the story above, since the dataset is imbalanced in favor of the negative condition (having only a minor injury), the sensitivity graph, below, closely tracks the graph of overall accuracy.



As for precision, this metric measures the percentage of predicted Class 1 assignments that were actually correct. It is computed as the number of true positives divided by the sum of true positives plus false positives. In the following graph, we can see that for the hybrid CART-logistic model, precision pretty much increases steadily as the size of the training set increases. The precision of the logistic regression and decision tree vary widely though, in what response to what I hypothesize to be random effects of the sample that was drawn. The hybrid CART-logistic regression seems rather robust to exactly what data points are used to train it.



Lastly, a precision versus recall graph is shown below. The interpolation between the points within a given curve, whether it be for the decision tree, the logistic regression, or the hybrid model, were performed according to the recommendations of Davis and Goadrich (2006, p. 6). Given that the desired place for a curve to lie is in the upper right hand corner, it can be seen that no curve 'dominates' the other. While the hybrid CART-logistic regression provides the greatest recall, and greatest precision at two of the three sample sizes, no curve provides greater precision and recall at all sample sizes.



Conclusion:

In the end, the choice of a decision tree, a logistic regression, or a hybrid CART-logistic regression classifier for traffic safety modeling of cyclist injuries will necessitate a trade-off between overall predictive accuracy versus sensitivity or recall and to some extent precision as well. In this specific field however, the hybrid CART-logistic regression model seems to be the most appropriate model. Because the cost associated with incorrect predictions of severe or fatal injuries is much worse than the cost associated with an incorrect prediction of someone who suffers a minor injury, the sensitivity/recall should be the criteria most heavily weighed when judging the performance of a traffic safety classifier.

Future Work:

In this analysis a number of options were left un-investigated which may affect the findings. First, there are many different tree-building algorithms and only one was used in this work. Additionally, there are multiple ways to prune a tree, which may lead to differing results. Another beneficial extension would be to run multiple model building efforts at each training set size so that the overall performance of the algorithms can be seen without the influence of a particular random draw from the available training data. Here, only one model building effort was performed at each training set size. Lastly, this exercise was carried out on an imbalanced dataset. The results may differ if the data set was oversampled/undersampled to correct the imbalance or if different cost functions were used in each algorithm to counteract the tendency of the algorithms to emphasize correct predictions of the majority class.