STAT 636 - Group Final Project
Section 700

Matt Byrom matthew.james.byrom@tamu.edu
Raechel Dillehay raedillehay@tamu.edu
Haley Friel haleymfriel@tamu.edu

LASSO Penalized Logistic Regression & Boosted Tree Classification Methods for Secondary Education Student Data

INTRODUCTION:
For the purposes of the STAT 636 final project, a data set made available through Creative Commons licensing was used to present the differences between a LASSO penalized logistic regression and boosted classification trees.

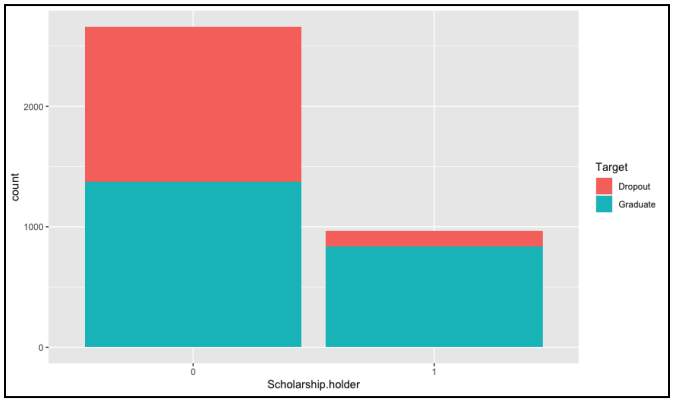
DATA DESCRIPTION:
The data set, titled *Predicting Student Dropout and Academic Success*, was created by researchers in the SATDAP program in Portugal. The data set includes 36 academic, demographic, and socio-economic variables for 4,424 students in which each row of data is a single student observation. The data set includes binary, ordinal, nominal, and continuous features. There are three possible categorical outcomes for each student: 1) dropout, 2) enrolled, or 3) graduated.

Complete information about the data set can be found online at the UC Irvine Machine Learning Repository. Data url: (https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success).

Unless otherwise noted, RStudio version 2023.09.1+494 was used for all statistical analyses and the codebase can be found on GitHub: https://github.com/mattByrom-tamu/STAT636_project.git.

COHORT OF STUDY:
For this analysis, the 794 observations of enrolled students were excluded in the model-building process. This changed the outcome variable to a binary result of graduated (coded as 1) or dropout (coded as 0). Of the remaining 3,630 students, 61% were graduates and 39% dropped out. Many of the academic metrics are specific to the Portugal academic system. In addition, the demographics of the 3,630 students lack diversity. Nearly all students in the cohort of study were single and of Portugal ethnicity. Students were predominantly daytime students with a previous secondary education qualification and 18% of them were pursuing a nursing degree.

EXPLORATORY ANALYSIS:
There were no missing values in the dataset. Variables were classified as categorical or continuous and summary statistics were calculated for the entire dataset to look at distributions. Summary statistics were also calculated for each of the target groups to identify variables that had significant differences between target groups. Correlations were calculated and visualized with a heatmap for evaluation. From the correlation matrix, it was clear that there is correlation amongst the various academic predictors such as first and second semester grades, enrollment status, and credit completion. It should be noted that Age was treated as a continuous variable. Categorical variables were further explored using bar plots. To easily visualize variables with differences between target groups. The chart (right) is an example of one of the significant variables in the



later models. It can be seen that a small percentage of people dropout if they have a scholarship. This analysis of the categorical variables provided overall information about the characteristics of the cohort as described in the section titled *Cohort of Study*.

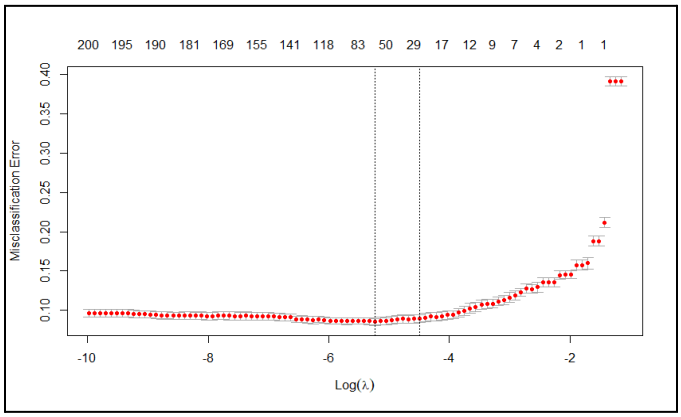
GOAL:

The main objective of this analysis was to explore two supervised learning methods and compare them using the same data set. By doing this, it was possible to evaluate the advantages of each. This paper reports the findings of a LASSO penalized logistic regression and a boosted classification tree model.

Secondly, the team was interested in assessing the influence of the 36 academic, demographic and socio-economic variables on student success (defined as graduated or dropped out). Ultimately, since the outcome is a categorical, binary, value, the goal was to find the best model to use for classification of students as either graduates or drop outs.

METHOD: LASSO Penalized Logistic Regression

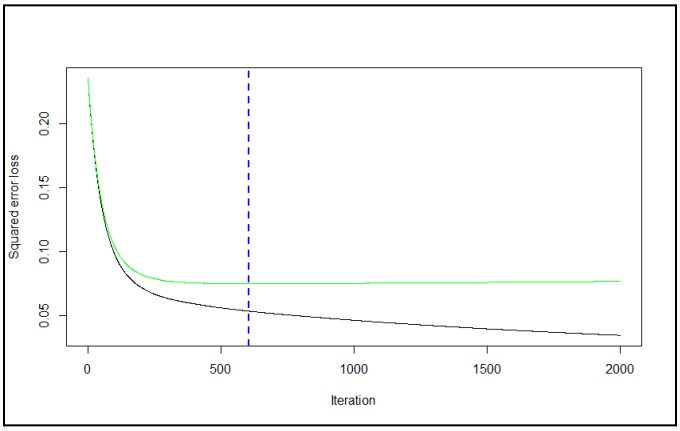
The LASSO penalized logistic regression method (referred to as just LASSO moving forward) was chosen because it allows for variable selection. This penalty is applied to the logistic regression model and shrinks non-significant variables to zero. This creates a more sparse model that is easier to interpret. The graph (right) depicts how each of the 36 variables were shrunk and eventually removed as lambda was increased.



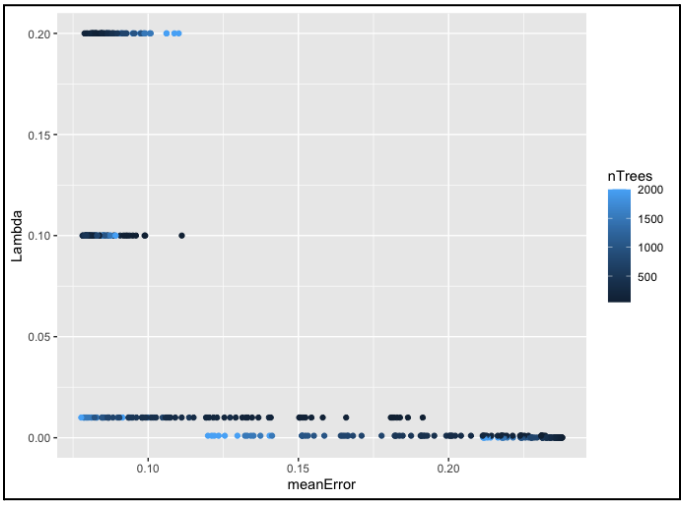
In order to perform the LASSO regression, a tuning parameter was found using tenfold cross-validation (CV). Two tuning parameter values were considered: 1) the minimized value, and 2) the value that creates the most regularized model such that the cross-validation error is within one standard error of the minimum (1se). The 1se value was used since it creates models with the number of variables equal to or less than the model found using the first option. The plot (right top) shows the difference of these two parameters with the 1se parameter value being the right dashed line. This ultimately makes for a simpler model that may be easily understood by a potential client.

METHOD: Boosted Classification Tree

In this method, trees are grown sequentially with each new tree using information from the previously grown tree in order to refine the model. Different combinations of tree depth, shrinkage (lambda), and tree count were used to find an optional model. The graph (right) shows the number of trees with the minimum classification error. An advantage of Boosted



classification trees is the resistance to overfitting, as shown by the green curve increasing very slowly after the minimum error indicated by the blue line.



Similarly to LASSO, fivefold cross-validation was used to determine tuning parameters and models by minimizing an estimated misclassification error. The plot (right) shows the misclassification error for different lambdas and number of trees.You can see that for the top three lambda, you could fit models with similar errors. However, you can see that for the large lambda (.20) the large tree counts lead to overfitting, while for the smallest lambdas, the low number of trees lead to underfitting.

COMPARISON & FINDINGS:

LASSO and boosted tree report variable significance in different ways (odds relationship vs. relative influence). Additionally, LASSO treats categorical variables separately and reports significance for each category within a particular variable whereas boosted tree reports influence based on the category as a whole.

Despite these differences, it was still possible to see that both methods resulted in similar results. The variables of *Second Semester Approved Credits, Age at Time of Enrollment, Course* (what might be called academic plan in American Universities), and *Tuition/Fees Up to Date* were significant in classifying students as graduates or drop outs in both models. The second-term academic life of a student seemed most significant in all parts of the analysis: exploration, LASSO model, and boosted tree model. It is clear that a student's second-term success (or lack thereof) really correlates with whether that student will graduate or drop out.

The boosted classification trees made better predictions than the LASSO model. The classification error for the classification trees was 8.06%, while the LASSO's classification error was 8.65%. However, it was decided that the LASSO penalized regression was a better model. First, it provides a sparse model. The simplicity makes it easier to write, interpret and share results of the model. Second, the LASSO penalized log regression model provides log odds estimates. These estimates are important because they give us a level of likelihood of a classification, rather than a binary yes/no result. This can be of value to schools so that they can identify students at different levels of risk and support them as needed. For example, a student about to drop out of college may require a different level of support than someone teetering on the edge between graduating and dropping out.

Suggestions moving forward would be to remove collinear variables as found in the *Exploratory Analysis* section. Any future classification attempts using this model should take care to remember the cohort of study. These models will likely perform poorly on a cohort that is more diverse.