Predicting Failure: Part III

Ashley Darnell, Ethan Hawk, Joshua Nibbe

DATA-151-A

Due Nov. 18th

Introduction

The goal of this paper is to outline the work and research done in order to understand what causes the number of failures students attain while in the educational system. Student failure is a major problem both in the United States and around the world. It is important for educators to understand why students fail, so that they may better aid their student's in the future. This is often a very complex problem because the lives of individual students can be very unique. There are many factors that can influence the grades of a student such as family, social life, educational background, health, and more. The data we are working with is more complicated than U.S. data because the legal drinking age in Portugal is only 16 years, therefore alcohol intake must be factored into the equation even at the highschool level. One can see how this model could become complex very quickly.

In a paper published by Springer, the authors go on to detail how they believe that the most influential theoretical explanations of student failure are the student's social and academic integration into the educational institution. After identifying what they believed to be the main causes of academic failure, the authors described their concerns in having imbalanced data as well as their process for creating a model for predicting failure in the future. Some of the classification models that they experimented with were k-nearest neighbor and the decision tree model. They also used a technique called genetic programming for classification. As a result of their efforts, they have proposed specific genetic programming models that they believe can be used to obtain accurate rules for predicting a student's academic performance (Màrquez-Vera). This article has proved to be of great value because of the similarities in their research and the results we are attempting to produce.

A reference article from the UCI Machine Learning Repository was helpful for comparison with the outline of this project. In the article, Paulo Cortez writes about the quality of education of Portuguese students and how it has improved in the last decades. Cortez's research involved collecting real-world data from student grades, demographic information, as well as social and school related data by using school reports and questionnaires. Cortez looked closely at the two core classes which were Mathematics and Portuguese in a binary/five-level classification and regression analysis, and four models which were Decision Trees, Random Forest, Neural Networks, and Support Vector Machines. He tested three input selections which were with and without previous grades. Cortez discusses that the results show, "that a good predictive accuracy can be achieved, provided that the first and/or second school period grades are available" (Cortez). Cortez concluded that as a direct outcome of this research, more efficient student prediction tools can be developed meanwhile improving the quality of education and enhancing school resource management. Cortez's research is of value because we can draw parallels between our prediction methods and his.

Another similar article was written by Rosa Maria de Castro and Dora Isabel Fialho Pereira who are faculty members of the Arts and Humanities at Universidade of Madeira in Funchal, Portugal. Castro and Pereira wrote that Portuguese schools have a high failure and drop out rates have spawned a number of initiatives that aim for their reduction. They formed a study that evaluates the relationship between internal working models with students, their perceptions of the quality of their relationships with teachers, and their academic performance using three measures which are: the "Inventory of Attachment in Childhood and Adolescence" (IACA) measure, the "Inventory of Parent and Peer Attachment" (IPPA) measure (which concerns that attachment to the teacher), and a socio-demographic questionnaire on a sample of 305 students from the 8th grade to regular education and the ACC (De Castro). The authors gathered results saying that students that are under the ACC program are less secure in their education than students who are in RE in all three measures. From this article, we can gather that certain educational tracks have a lifelong impact on students.

DATA & METHODS

In this project we used data from 649 students of a Portuguese secondary school. This data comes from the UCI Machine Learning Repository. The data consists of 33 variables with 17 categorical and 16 numerical. These variables include information about the students' home life, health, habits, parents, social life, study habits, and more. The data cleaning involved clearing out unnecessary colons and quotation marks so that it could be read in properly to Python. Several data-mining techniques were applied throughout the modeling process such as transformation of variables, data cleaning, and data splitting.

The data, available on the UCI Machine Learning Repository, includes information about students attending both math classes as well as Portuguese language classes. At this time in the project we have chosen to only use the data from the Portuguese language class in our models. We will be focusing mainly on recall score when evaluating our models, because what is most important in determining a good model in our case as it can predict the true positives (failures). In other words, we don’t want to undercount failures or create more false negatives (non-failures). The tool most frequently used in analysing our data was Google Colaboratory, however R Studio was also used to create linear regression models of the top predictors chosen by a random forest feature importance graph. The models did not yield great results so we chose not to pursue further analyses in R Studio.

To begin examining the dataset, we used Google Colaboratory to create several graphs including scatter plots, box plots, heatmaps (correlation shown in Figure (1)), and histograms. We also examined the descriptive statistics to look at each variable. From this exploratory analysis we noted that the dataset contained many categorical variables. We found interesting relationships between several variables including age and failures, this proved to be vital in further analysis due to the fact that age explained a lot of the variance in failure in many of our models. This relationship is shown in Figure (2) as a boxplot. We were also able to see that not many of our variables had a high correlation to the target variable failures, and that our data was imbalanced as many students had no failures. Failure was originally coded as a categorical variable with five categories. The students could have either zero, one, two, three, or four past class failures. Even though four was a possible category of the failures variable, there were no students that had a value of four. We transformed this variable into a binary variable with two categories, fail or no fail.It was then coded as a dummy variable. In order to begin modeling, most of the variables needed to be transformed into dummy variables. After examining the data, transforming the variables, and splitting the data into training and testing groups we began modeling with a decision tree.

RESULTS

When we began modeling, we were using all of the predictors available in the dataset, however this was a very complex model. The decision tree, using all of the predictors, gave us a recall of .846, while this is a good score we wanted to tweak the model to see if we could simplify it. We were able to simplify the model to use only seven predictors while maintaining a recall of .815 from a decision tree model through the process of guess and check as well as some intuition from the exploratory analysis. This was good, however we thought we could do better, so we modeled our data in a random forest. This model produced a recall score of .8. The confusion matrix of the random forest led us to believe that the data was imbalanced. To remedy this issue we first tried oversampling, which resulted in low scores. Using the SMOTE method along with a random forest we were able to get an average recall score of .83 and produce a feature importance graph, shown in Figure (3), which lists the variables in order of importance relative to ability to predict failure. Using the top 10 features, as suggested by the graph from the random forest, we attempted to create a linear regression model in R Studio. After backwards elimination, this resulted in a model with only 2 significant predictors that had a very low correlation.

Because our data contained many categorical variables, we decided a logistic regression model might be a better fit. We first ran chi-squared tests between every non-integer variable and the target (failures) that had to be transformed into a binary variable of fail or no fail. The chi-square test yielded significant p-values for the variables “reason”, “higher”, and “guardian”. A mosaic plot of “reason” and “failures” is shown in Figure (4). After transforming all variables into dummies, the logistic regression model was performed as well as a cross validation. This gave us a mean accuracy score of .83 for the logistic regression model. While the mean scores were lower, we were able to get a better confusion matrix by lowering the threshold at which the model categorizes each case as fail or no fail. A threshold of .2 allows for the most correct categorizations of failures, while keeping false categorizations of failures as low as possible. The resulting confusion matrix can be seen in Table (1). Finally we produced a ROC curve from the logistic regression that is shown in Figure (5) and an AUC score. The AUC score is a metric that allows us to see how well a model distinguishes between categories in a binary regression analysis. The AUC produced from the random forest was .855, while the AUC from the logistic regression was .883. The ROC curve for our random forest model can be seen in Figure (5).

Figure (1): Heatmap for Correlation

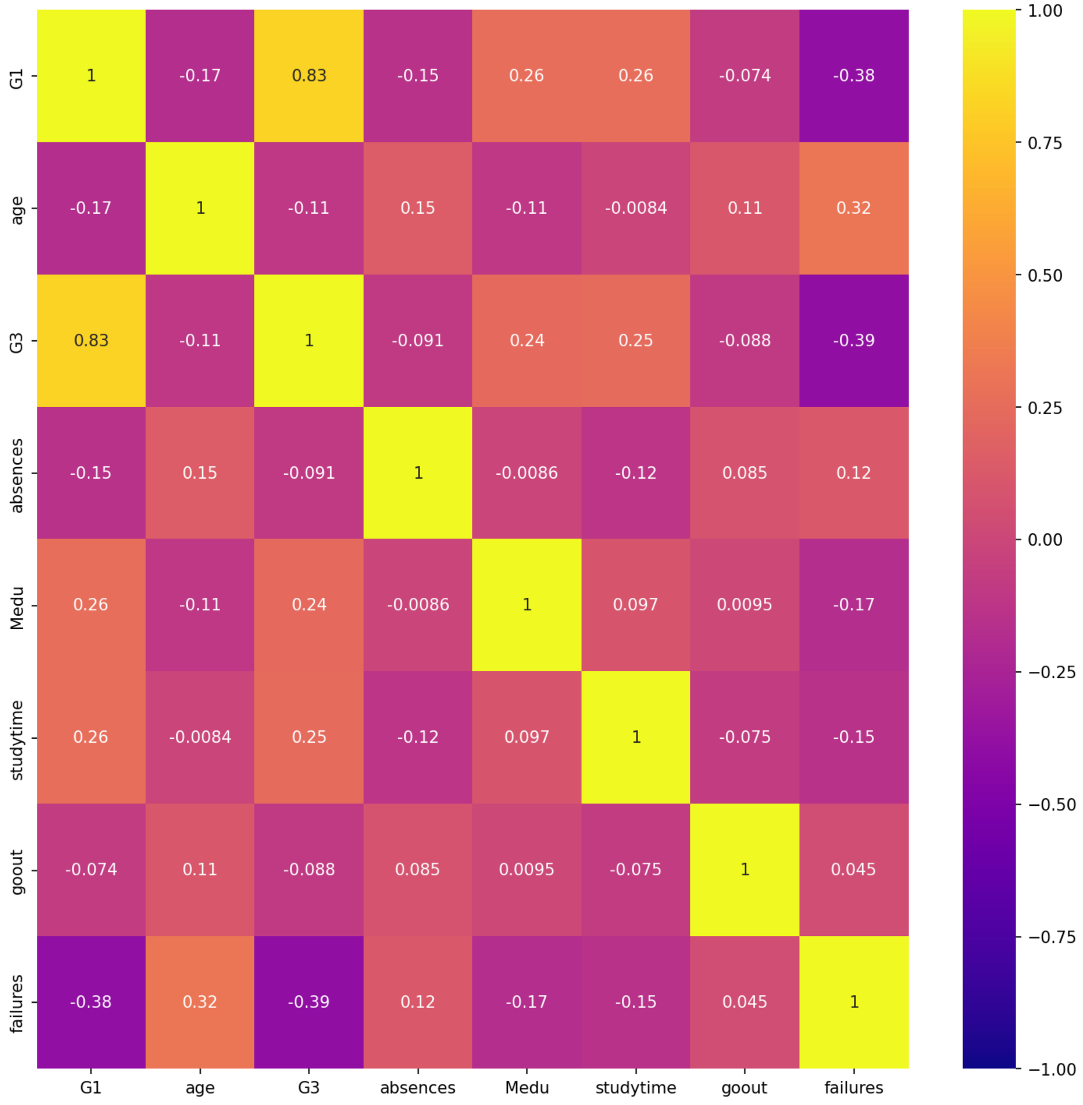


Figure (2): Failures vs. Age

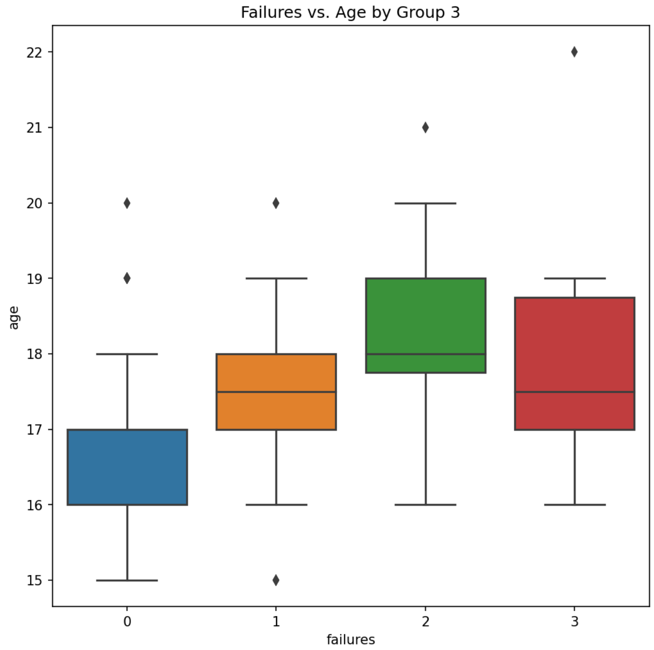


Figure (3): Random Forest Feature Importance

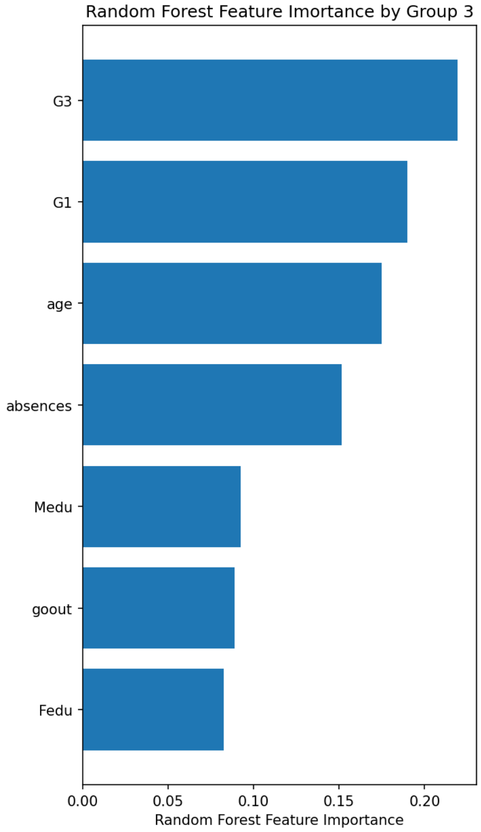


Figure (4): Mosaic Plot for Reason and Failure

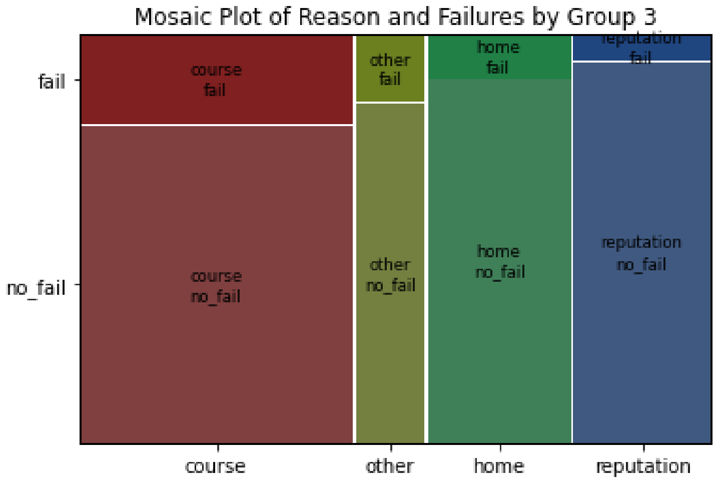
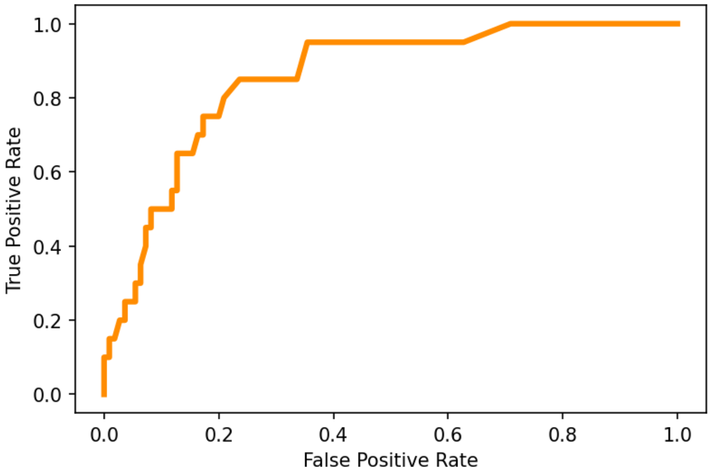


Figure (5): ROC curve for Random Forest



Table(1): Confusion Matrix for Logistic Regression

	0	1
0	141	24
1	10	20

DISCUSSION

The results of this project were both surprising and expected. The outcome of the decision tree and random forest were good as we expected, however, there were surprising details that came about from testing and exploratory analysis. For example, we found that absences, and age were important numerical predictors while reason was a significant categorical predictor. Out of all of our variables, we did not predict that these would be among the top. We also did not understand how the imbalanced data would affect the models. Our target variable, failure, was imbalanced because more students had a value of zero than any other category. As a result of this we had to try several balancing methods. By using the SMOTE method, we were able to increase the recall score slightly as well as get a balanced dataset to create more models. It was essential that we transform the target variable into a binary variable where zero was equivalent to no failures while one, two, and three were equivalent to failure. We used this newly transformed variable to do a logistic regression model.

CONCLUSION

Throughout the modeling process, we cleaned, split, transformed, and balanced the data. We modeled the data using decision trees, random forests, and logistic regression. The random forest model with SMOTE balancing produced the best average recall score. By using a subset of the data we were able to get the best recall score while keeping the model as simple as possible. The best model we found predicts failure with first quarter grades, third quarter grades, absences, mother and father's education, and how often the student goes out. This means that when we observe these factors of a students social and academic life and look at each factor as a decision sampled over again multiple times, you could get a reasonable prediction for a student's academic performance. When we compare the results of the random forest to the logistic regression, it is just a matter of what is most important. The logistic regression produces a lower recall score, yet it is able to classify target variable better, as we saw with the AUC score and ROC curves.

While the predictors chosen may impact the number of failures a student obtains during a school year, each student should be viewed as an individual when it comes to receiving an education. This model should not be used to determine a student's worth or make decisions for the student based on predictions of more failures in the future. An ethical application of this analysis would be to aid teachers in understanding what causes failures so that they can make changes where they are able to. As discussed in previous sections, there are so many aspects of a students life that affect their academic performance and it is a dynamic and ever changing process. The point of our study was simply to find the most common and impactful elements that make a difference in student performance.

REFERENCES

Cortez, Paulo, and Alice Silva. "USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE." *UCI Machine Learning Repository: Student Performance Data Set*, 2008, https://archive.ics.uci.edu/ml/datasets/student+performance.

De Castro, Rosa Maria, and Dora Isabel Fialho Pereira. "Education and Attachment: Guidelines to Prevent School Failure." MDPI, Multidisciplinary Digital Publishing Institute, 20 Feb. 2019, https://www.mdpi.com/2414-4088/3/1/10/htm.

Márquez-Vera, Carlos, et al. "Predicting Student Failure at School Using Genetic Programming and Different Data Mining Approaches with High Dimensional and Imbalanced Data." *Applied Intelligence*, Springer US, 26 Aug. 2012, https://link.springer.com/article/10.1007/s10489-012-0374-8.