Predicting Failure: Part IV

Ashley Darnell, Ethan Hawk, Joshua Nibbe

DATA-151-A

Due Dec. 16th

Introduction

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 that will be examined in this paper is more complicated than U.S. data because the legal drinking age in Portugal is only 16 years old, therefore alcohol intake must be factored into the equation even at the high school level. One can see how this model could become complex very quickly. 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 Portuguese educational system.

The work of Marquez-Vera, is relevant in the discussion of predicting failure and influential factors of education. He talks about how he believes that the most likely explanations of student failure are the student’s social and academic integration into the educational institution. Marquez-Vera and co-authors (2012) described their concerns in having imbalanced data as well as their process for creating a model for predicting failure in the future. They 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.

Another pertinent author that can be noted in the field of education as well as data science is Paulo Cortez. In his 2008 article he 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.

Another similar article was written by Rosa Maria de Castro and Dora Isabel Fialho Pereira. Castro and Pereira wrote that Portuguese schools have a high failure and drop out rates that 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.

Many data scientists have attempted to predict what influences failure of high school students. This means that there is a lot of theory to begin with and many different paths that could be taken. From the work of Marquez-Vera, it is known that a student's social life is a driving element in educational success. Cortez's journal aids in understanding what a typical Portuguese education system is like, therefore making it easier to interpret the variables that come from the Portuguese data set as well as any results that might be obtained. Castro and Pereira give insight into how the quality of teacher relationships may have a high influence on student failure. The purpose of this project will be to combine all of this knowledge to get an even more accurate and holistic model to predict what causes student failure.

DATA & METHODS

In this project data was used from 649 students of a Portuguese secondary school which are publicly available from the UCI Machine Learning Repository. This data was collected over one academic year. While there is data available that includes information about students attending both math classes as well as Portuguese language classes, at this time in the project it was decided to only use the data from the Portuguese language class in the models. The dataset chosen 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 target variable, *failures,* was transformed into a binary to help with imbalanced data. The entire list of variables can be found in (Table 1) .The original dataset’s format was modified and saved for use throughout the project. This modified dataset was split into training and testing data with 20% reserved for testing and the rest used to train the model.

Table 1: Variable List

*Variables*

*School - The school a student attends ( ‘GP’ Gabriel Pereira or ‘MS’ Mousinho da Silveira)*

*Sex - Student’s sex ( ‘F’ Female or ‘M’ male)*

*Age - Student’s age (15-22)*

*Address - Student’s home address type ( ‘U’ urban or ‘R’ rural)*

*Famsize - Family Size ( ‘LE3’ Less than or equal to 3 or ‘GT3’ Greater than 3)*

*Pstatus - Parent’s combination status (‘T’ Living together or ‘A’ Apart)*

*Medu- Mother’s education (0: none, 1: primary education, 2: 5th-9th grade, 3: secondary education, or 4 higher education)*

*Fedu - Father’s education (0: none, 1: primary education, 2: 5th-9th grade, 3: secondary education, or 4 higher education)*

*Mjob - Mother’s job (‘teacher’, health’,’services’,’at\_home’, or ‘other’)*

*Fjob - Father’s job (‘teacher’, health’,’services’,’at\_home’, or ‘other’)*

*Reason - reason to choose this school (‘home’, ‘reputation’, ‘course’, or ‘other)*

*Guardian - Student’s guardian (‘mother’, ‘father’, or, ‘other’)*

*Traveltime - Home to school travel time ( 1: <15 min, 2: 15-30 min, 3: 30-60 min, 4: >60 min)*

*Studytime - Weekly study time ( 1: <2 hours, 2: 2-5 hours, 3: 5-10 hours, 4: >10 hours)*

***Failures*** *-* ***number of past class failures ( 0,1,2, or 3)\*\*target variable\*\****

*Schoolsup - Extra educational support ( ‘yes’ or ‘no’)*

*Famsup - Family educational support (‘yes’ or ‘no’)*

*Paid - Extra paid classes within the course (‘yes’ or ‘no’)*

*Activities - Extracurricular activities (‘yes’ or ‘no’)*

*Nursery - Attend nursery school (‘yes’ or ‘no’)*

*Higher - Wants to take higher education (‘yes’ or ‘no’)*

*Internet - Internet access at home (‘yes’ or ‘no’)*

*Romantic - With a romantic relationship (‘yes’ or ‘no’)*

*Famrel - Quality of family relationships ( 1-very low to 5 - very high)*

*Freetime - Free time afterschool (1-very low to 5 - very high)*

*Goout - Going out with friends ( 1-very low to 5 - very high)*

*Dalc - Workday alcohol consumption ( 1-very low to 5 - very high)*

*Walc - Weekend alcohol consumption ( 1-very low to 5 - very high)*

*Health - Current health status ( 1-very bad to 5 - very good)*

*Absences - Number of school absences ( 0-93)*

*G1 - first period grade (0-20)*

*G2 - second period grade (0-20)*

*G3 - final grade (0-20)*

To evaluate the performance of models, average cross validation recall score will be used. This is because predicting the rate of failure is the goal of the research and this recall score will demonstrate how well a model is classifying failure or non-failure. To begin examining the dataset, an exploratory analysis was done to create several graphs including scatter plots, box plots, heatmaps for correlation, and histograms. This initial analysis was useful for understanding the spread and shape of the data as well as which variables would be the best fit in a features vector for modeling. To begin modeling it was necessary to decide which variables would be included as predictors. In order to find these predictors, a heatmap (Figure 1) was examined to find any variables that were highly correlated with each other.

All variables that had an absolute correlation of at least .11 with *failures* were included in the features vector. The threshold of .11 was made because any higher value would cause nearly every variable to fall out of the model. This is due to the fact that there was very low correlation among each variable and *failures.* Once the features vector was created, the models could be made from them. The first models that were created were a decision tree, random forest, and logistic regression. Then, these models were all replicated using SMOTE oversampling.

RESULTS

The model that performed the best out of the six that were attempted was the random forest with SMOTE oversampling. The results from each model performed can be seen in a table below (Table 2). Based upon the average recall score with cross validation, the random forest with SMOTE oversampling far outperforms all other models. This model does a good job of classifying as well, as shown by the ROC curve (Figure 2) and confusion matrix (Table 3).

Table (2) : Results from Modeling

|  | Recall Score | Avg. Recall from Cross Validation | Weighted Avg. Recall | ROC score |
| --- | --- | --- | --- | --- |
| Decision Tree | .76 | .41 | .81 | .79 |
| Decision Tree w/ SMOTE | .82 | .30 |  | .86 |
| Random Forest | .78 | .31 | .88 | .85 |
| **Random Forest w/ SMOTE** | **.72** | **.94** | **.86** | **.86** |
| Logistic Regression | .85 | NA | .86 | .87 |
| Logistic Regression w/ SMOTE | .61 | .86 |  | .82 |

Figure (1): Heatmap for Correlation (Predictors Only)

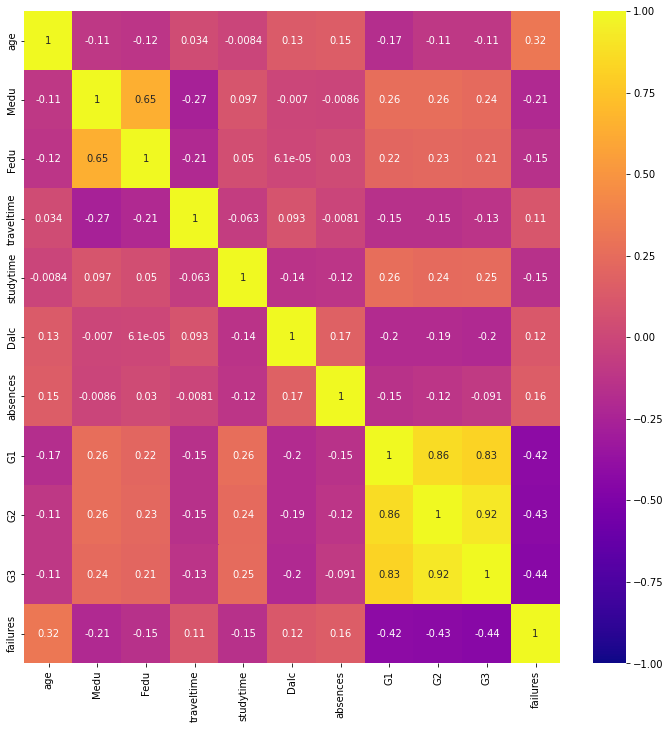
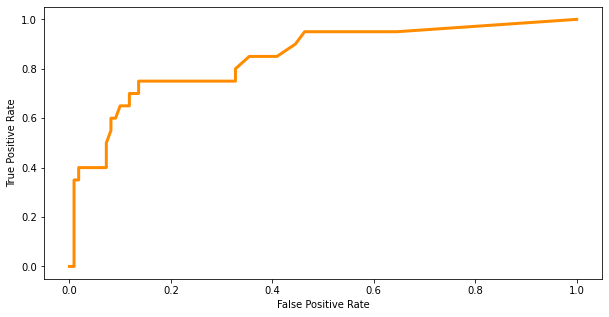


Figure (2): ROC curve for Random Forest w SMOTE



Table(3): Confusion Matrix for Random Forest w SMOTE

|  | 0 | 1 |
| --- | --- | --- |
| 0 | 84 | 26 |
| 1 | 5 | 15 |

DISCUSSION

The results of this project were mostly as expected. The outcome of the decision tree and random forest were good as we expected, however, there were surprising details that came about from modeling and exploratory analysis. For example, absences and age were important numerical predictors. 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 there were several balancing methods that had to be done. By using the SMOTE method, it was able to increase the recall score slightly as well as get a balanced dataset to create more models. It was essential to transform the target variable into a binary variable where zero was equivalent to no failures while one, two, and three were equivalent to failure. The newly transformed variable was used in all of the models.

It is important to note that there is multicollinearity among the predictors used to create these models; *G1, G2,* and, *G3.* This can be seen in the heatmap of correlation between predictors (Figure 1). There were attempts to remove this, however, time constraints did not allow a completion of the models that contained the correction. Due to this multicollinearity, the results of the models (Table 1) should be taken into account with caution.

Other researchers who have worked on similar projects, as mentioned in previous sections, obtained similar metrics as seen in this report. The Education and Attachment article tended to care more about the psychological aspects of a student's performance by thinking about the relationship between the teacher and students. The authors who wrote the Grading in Portuguese Secondary School Physical Education article were interested in how physical education affects the overall performance in school. The authors focused on average grades according to the Physical Education quartile.

CONCLUSION

In this study, it was found that age, mother’s education, father’s education, travel time to school, study time, daily alcohol consumption, absences, and grades are the best predictors of student failure. These factors, along with a random forest model with SMOTE oversampling, can be used for several academic purposes. Future uses of this model should include helping students to avoid academic failure by understanding how their outside influences are affecting them. This could allow teachers to get more insight into a student’s life and how they are dealing with more than just academic struggles. Future research could include more variables like learning styles, learning disabilities, number of siblings, and more.

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.

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>.

R, Vishal. “Feature Selection - Correlation and P-Value.” *Medium*, Towards Data Science, 23 Feb. 2020, https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf.