Final Project Data Memo

Vardan Martirosyan

2022-09-30

Overview of Dataset

- This data contains the final grades of college students who were enrolled in a math or portugese class in secondary school. It should be noted that while the description of the dataset says it contains both the final grades for the math class and the portugese class, the dataset on the website only has the final grades and information for the math class. As a result, for my project, I plan on only using the math grades dataset. In addition, it contains many attribute variables about each student themselves, such as information about their sex, age, how many classes they've failed in the past, etc.
- I will be obtaining this dataset from kaggle.com. I will download it from the website itself. The link to this dataset can be found here: https://www.kaggle.com/datasets/janiobachmann/math-students
- There are 395 observations in this dataset, and 32 predictors.
- I will be working with binary, numeric, and nominal/categorical variables.
- There is no missing data, but on the Kaggle website, it says that the attributes from 'schoolsup' to 'romantic' are mismatched. I downloaded the CSV file and looked at it myself, and it doesn't look like those predictors are mismatched from my point of view. However, if it does end up being an issue, then I'll choose to not include these features/predictors in my analysis.

Overview of Research Questions

- The variable I am most interested in predicting is the final grade of a student at the end of a math class. If there's time, other variables I might be interested in predicting are the first period grades and second period grades. I have two primary questions. First, how accurately can we predict a student's grade with the given predictors? Second, what types of techniques, regression, and classification methods can lead to the best results for the prediction of a student's grade?
- The primary response/outcome variable I am interested in is G3, which is the final grade of a student, and is a numeric variable (that ranges from 0 to 20).
- I think that the first question will best be answered with a regression approach. However, I am not certain about this. As a result, another goal for this project is to implement the variety of regression and classification techniques we learn during class into my code to answer this question. That is, I plan on creating many different models to answer the first question. These models include linear regression, logistic regression, decision trees, K-means, etc. By doing this, I hope to determine which model would be most effective for answering the first question.
- I think that the predictors that will be especially useful are 'G1' and 'G2', which represent a student's first period grade and second period grade, respectively. Another predictor that will be useful is 'failures', which represents the number of past class failures. I think that these predictors may be

useful for a few reasons. One, 'G1' and 'G2' will be useful, since they are directly related to a student's final grade, 'G3'. Second, I think that 'failures' might also be a useful predictor, since it can let us know if a student has had any troubles academically, and may be more prone to getting a poor grade in a class.

- We are asked to describe if the goal of our model is descriptive, predictive, inferential, or a combination, and to explain our reasoning. As stated before, I plan on making many models to answer my primary question of determining a student's final grade based on their predictors. As my main goal is trying to predict a student's grade based off of predictors, all of my models will be predictive. There may be a use for a descriptive model in terms of building a student profile, but this would be used to then help make a prediction model for the first question I am trying to answer.

Proposed Project Timeline

Below is my proposed Project Timeline. I plan on having my dataset loaded, and beginning my exploratory data analysis, at the beginning of Week 2.

Week 2: Begin and Finish Exploratory Data Analysis

Week 3: Implement a Linear Regression Model

Week 4: Implement a Logistic Regression Model

Week 5: Test out Resampling Methods/Adding Discriminant Analysis to the Models

Week 6: Test Model Selection/Predictor Selection into the Linear and Logistic Regression Models

Week 7: Implement Decision/Classification Tree Models

Week 8: Implement K-Means and Clustering Classification Models

Weeks 9-10: Wrap up the Model Implementation, Gather Information About All of the Results, Start/Finish Writing Up Report.

Any Questions/Concerns

- There are no major problems/difficult aspects of the project that I anticipate. The main concern I have is as follows. The mismatched data might end up being an issue, and I may have to get rid of all of those predictor variables, which would be unfortunate, as they seem like very interesting predictor variables and may be useful to the creation of the model.
- I have no specific questions for the professor or the instructional team.