Final Project Memo

Lauren Huang

2022-09-23

An overview of your dataset

1. What does it include?

The dataset I chose is on student performance in 2 secondary education Portuguese schools. There are many predictor variables, such as school, sex, age, family size, mother's education, father's education, mother's job, father's job, travel time, study time, failures, and more. The goal of training a model on this data will be to accurately predict a student's G3 (final grade) based on their (G1) 1st and (G2) 2nd period grades.

2. Where and how will you be obtaining it? Include the link and source.

The data is in csv format and comes from kaggle. The data is separated into 2 files, student performance in Math and Portuguese classes. https://www.kaggle.com/datasets/whenamancodes/student-performance?select=Portuguese.csv

3. About how many observations? How many predictors?

The Math data file has about 390-400 observations, and the Portuguese data file has about 650 observations. Both files have the same 30 predictors.

4. What types of variables will you be working with?

I'll be working with independent variables(predictors) like school, age, travel time, study time, etc. and a dependent variable(s), which is G1 (first period grade), G2 (second period grade) and G3 (final grade).

5. Is there any missing data? About how much? Do you have an idea for how to handle it?

There doesn't seem to be any missing data.

An overview of your research question(s)

6. What variable(s) are you interested in predicting? What question(s) are you interested in answering?

I am interested predicting G1 (first period grades) and G2 (second period grades), and G3 (students' final grades). This is because G3 grades are highly correlated to/dependent on G1 and G2 grades. In other words, It would be much harder to predict G3 without knowing G1 and G2 first. I am interested in questions like: What is the relationship between a mother or father's education and the student's final grade? What is the

relationship between a mother or father's job and the student's final grade? What is the relationship between the distance between school and home and student's final grade? Ultimately, by asking these questions I'm trying to understand the relationship between each predictor and the students' grades.

7. Name your response/outcome variable(s) and briefly describe it/them.

I expect my response/outcome variable(s) to be G1, G2, and G3(students' grades). First I'll need to train my model on assigning students a grade by using my predictor variables(school, age, family size, study time, failures, etc.). I'll need to create predictions on G1 and G2 first, then use those results to predict G3. (As G3 is highly correlated/dependent on G1 and G2)

8. Will these questions be best answered with a classification or regression approach?

I think for my data set it is a bit tricky to tell between regression or classification, but I would choose to use the classification approach. This is because the students' grades are on a scale of 0 to 20 (See G1, G2, G3). The student grades are quantitative values, but they can only get one grade from many on the scale (the grading scale is not continuous, there's no data with grades of 0.25, 0.5, 0.75, etc.).

9. Which predictors do you think will be especially useful?

I think predictors like school, travel time, study time, failures will be most useful because they can tell us the direct impact on the outcome. Other predictors like mother's education, father's education, mother's job, father's job may help reveal the importance/value of parental guidance or familial emphasis on education, but because these are qualitative traits it is harder to tell their direct impact on student performance.

10. Is the goal of your model descriptive, predictive, inferential, or a combination? Explain.

I think the goal of my model is predictive. I am using existing student data and information to make a prediction about what a student's grade will be.

Your proposed project timeline

1. When do you plan on having your data set loaded, beginning your exploratory data analysis, etc? Provide a general timeline for the rest of the quarter.

I have downloaded my data set, and upon approval of the data memo, plan to load my data set within the next 1-2 weeks. After that, I'll train and test my model (est. 1-3 weeks, may take longer?), and lastly I'll explore my data (1-3 weeks). I'll use the last couple of weeks before finals to make adjustments + clean up final details before turning the project in.

Any questions or concerns

1. Are there any problems or difficult aspects of the project you anticipate?

As I was studying my data set, I felt confused on how I should approach my response variable(s). My first thought was to pick G3 (final grade) as the response variable, since the data set indicated that all predictors pointed toward that target attribute. Meanwhile, it is also mentioned that G3 is "strongly correlated" with G1 and G2 (first, second period grades), and it would be hard to know G3 without G1 and G2. One method I'm considering: I would need to first train my model to predict G1 and G2, then use those results to train my model to predict G3.

2. Any specific questions you have for me/the instructional team?

Do you think it would a good idea to merge the 2 data files? (to get more observations) or just pick one (either Math or Portuguese)? Also, if the instructional team has any advice regarding my thoughts on the possible difficulty I mentioned above, please let me know! I would greatly appreciate your help.