Daniel Corral

DSC530

Final Project paper

What external factors affect final math scores? Is it the obvious ones such as missing class, or drinking too much alcohol? Or, do parental factors or prior history come into play? These are the questions that I set out to answer while doing my data analysis.

First, I obtained a dataset of students in secondary institutions. My first mistake was selecting a dataset with only 396 cases. I should have been looking in the thousands, but this was my first analysis. I did like the number of potential external factors that were part of the dataset.

The main variables that I focused on were final scores and the items that may impact them. Those items were mother and father’s education, absences, gender, family size, and alcohol consumption during the week and on weekends.

I looked at average scores which were from 0-20 which I am not sure how those convert to the standard 100-point scale but no additional information was offered. The average scores were lower with lower parental education, high weekday/weekend alcohol use, and high family size. So far it seemed to be lining up with my hypothesis. Histograms showed a visual representation of the spread of scores for the variables.

I then used correlation and scatterplots to show the relationship between the different variables and final scores. This is where my surprise came in. Only one variable that I checked had more than a weak relationship with final score. That was prior failures and even that was only a weak relationship. Everything else using different methods of correlation only showed very weak correlation.

I have concluded that due to my data size the correlation numbers are not reliable. I don’t feel that this dataset accurately reflects large populations. I learned more from the average values and the correlation positivity/negativity than I did from the pure correlation numbers. Those showed that with higher parental education and higher study time scores tended to go up. With higher alcohol use regardless of weekday or weekend, higher age, and higher prior failures, the score went down. Those are all assumptions that I had prior to running the analysis. However, if my null hypothesis was that those factors did not have a strong impact on grades then I have failed to prove it wrong.

Reference

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.