Gabriel Solomon Holland

Data Science – Conclusions

Skip to bottom for final conclusion

**Initial Question**: Do students who drink more and miss classes tend to fail more classes or get lower grades?

**Immediate correlations:**

Chart, bar chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

This project has shocked me in many ways because the reality I’m seeing is that it is incredibly difficult to predict success because of the sheer number of factors. Simply looking at correlations – there are almost none. Alcohol consumption, which is a survey-based scale 1-5, had almost no correlation with how many absences there were.

I see a trend across dozens of charts showing that there are no correlations amongst two features which at initial glance completely disproved my hypothesis. This lack of correlations was supported by my correlation matrix, nearly every feature had a sub 20% correlation.

Chart, treemap chart

Description automatically generated

**Advanced Conclusions**

Since two features didn’t give any results, I went to machine learning. Machine learning started showing some significantly better answers. My target was number of failures since the target was only a range of 0-15 (roughly) as compared to the trimester grades which was 0-100. Using that target I made several sets of features to test with.

* Set x = Weekday alcohol consumption and number of absences
* Set w = age, mother’s education, father’s education, studytime, and number of times they went out
* Set v = all the features above plus weekend alcohol consumption and how much freetime they had.

Set x and set v were the ones that I was really focused on. The most interesting result was that across all types of learning, across both x and v, machines could only get 85.38% exactly. There was frequently no difference in accuracy between x and v, it seemed almost like the machine was ignoring all the extra features that v offered. Based on two features predicting one y rather than 1:1 features to y as seen prior to machine learning, we were able to jump from almost nothing to an 85% accurate prediction. 85.38 was the exact number for:

* Support vector machine, set v and x, for linear, RBF, and sigmoid kernels.
* Forest ensemble, set v and x, using soft voting, logreg, svm, and random forest machines.
* Multi-layer perceptron, set v, set x, AND set w.

Other close to 85% results were:

* Decision tree classifier
  + Set x 83.08% accuracy
  + Set w 84.62% accuracy
  + Set v 82.31% accuracy
* Forest Ensemble
  + Set x,w,v with 82.31% accuracy
  + Used:
    - Soft voting
    - Logreg
    - Svm
    - Random forest with 15 predictors
  + Yes the same forest ensemble got 85.38% on other runs

We see two numbers appear multiple times: 85.38% and 82.31% which both appear multiple times across multiple runs which likely means that they are important and might be the actual correlations between these features, the problem is it is hard to see how they’re deciding this. In order to get a better understanding of how they were classifying, I made a decision tree visualizations. Note that these are HUGE pictures so you can zoom in and if Windows permits, it will come into focus. I will also attach some of these to the assignment itself. The color is based on the gini impurity, red is highest, then orange, yellow/beige, green, blue, purple. Upon closer examination I realized a trend.

Set X: Graphical user interface

Description automatically generated

The reason they get 80% accuracy is because they guess 0-2 failures *every single time* and 0-2 appear roughly, shocker, *80% of the time*. It has high accuracy because it always guesses based on the sheer probability of someone NOT failing, not because it understands what’s going on. That would explain the ~10% accuracies on G1-G3, there isn’t an easy probability to guess.

**Final conclusions**

No, you can’t really predict whether a student will fail based on our dataset, even with set v (all our features), our machine learning algorithms hit an 85% accuracy – only because it always guesses the answer that appears 85% of the time. There simply aren’t enough correlations to, using any of our algorithms, accurately predict if a student will fail. We also can’t predict a student’s grades.

As proven by our dozens of graphs and correlation matrix, simple 1:1 correlations don’t exist. Our highest performing machine learning algorithms are abusing the statistical frequency of a specific answer.

Despite my initial question being disproved quite spectacularly, there were a lot of interesting results. Parental levels of education had a 65% correlation, one of the highest across all my data. A logical other result is that student’s trimester grades are closely correlated, students tend to get similar scores across trimesters. Another logical result was weekday vs weekend alcohol consumption, which makes quite a bit of sense. One that I expected more from was number of failures, which was so decisively not correlated with anything that it almost seems wrong, the most impactful cause of failures was, with 32%, the student’s *age*. Dalc (weekday alcohol consumption) also had almost no impact on anything, it impacted a student’s free time (obviously) but really nothing else. Looking at all these conclusions one must start considering the flaws in a survey-based dataset, many of these features were a scale 1-5 based on students self-reporting. That level of inconsistency in the data could single handedly explain why my results were so incredibly lackluster. Students, especially underage students, might be uncomfortable with admitting how much they drink. Other students might exaggerate how much they study. The only features that are verifiable by the school that collected the data were the student’s grades, absences, and failures. I think on the simple fact of data quality the drawn conclusions are unreliable.