Group 5: Matthew Conway, Trent Hollandsworth, Valerie Davis

STOR 390 Final Project: Summary

**What we have done so far**

* The two data sets, one from a Math class and the other from a Portuguese Class, are imported and merged. Observations with a 0 as the final grade (G3) were dropped, as the other variables did not support this outcome. We assume there were other factors that led to the 0 final scores (e.g. the student did not take the assessment). We also removed the observation with the sole final grade of 1 as an outlier.
* In the EDA, examined trends between gender, age, alcohol consumption, parent’s education level and job status, amount of study time, number of failures and absences, and extra help given.

**What we have found**

* Gender
  + Males and Females perform roughly the same. G3 scores are roughly distributed the same.
  + Males drink more heavily. Both genders increase alcohol consumption on weekends.
  + Males are more likely to be in a romantic relationship.
* Age
  + Age against G3 scores shows a slight downward trend (i.e. test scores drop as age increases).
  + Alcohol consumption increases with age. Drinking for all ages increases over weekends.
* Drinking
  + More drinking does not show clear trend of decreased health.
  + Positive relationship between consumption of alcohol and going out.
* Parent education/job
  + Positive relationship between G3 scores and mother and father education levels. Slightly more positive relationship with mother’s education levels.
  + Interest for pursuing higher education increases with the higher levels of mother’s education level.
  + Positive relationship with parent’s education level and amount of educational support.
* Study time
  + Increased study time has positive relationship with test scores.
  + Individuals in a romantic relationship show more time devoted to studying.
* Failures and absences
  + Increased failures and absences have negative relationship on grade.
  + Absences do not show relationship with bad health or difficulty getting to school (not absent due to health reasons or travel restrictions).
* Extra help
  + Access to internet shows 1 point higher scores on tests.
  + Taking extra paid classes shows negative trend in test scores. This could be due to increased work load or the need to take extra classes due to failure first time.

**What we have left to do**

* Variable selection
  + A guided version using what we have learned from EDA
  + An automatic version (probably using LASSO)
* Further model creation
  + Ideally sticking to linear regression or classification using the variables via the selection process.
  + However the original paper utilizing this data found that tree-based methods were most effective, so we may try to include a tree-based model instead.