WHAT IS THE EFFECT OF ALCOHOL CONSUMPTION ON A STUDENT'S FINAL CLASS GRADE?

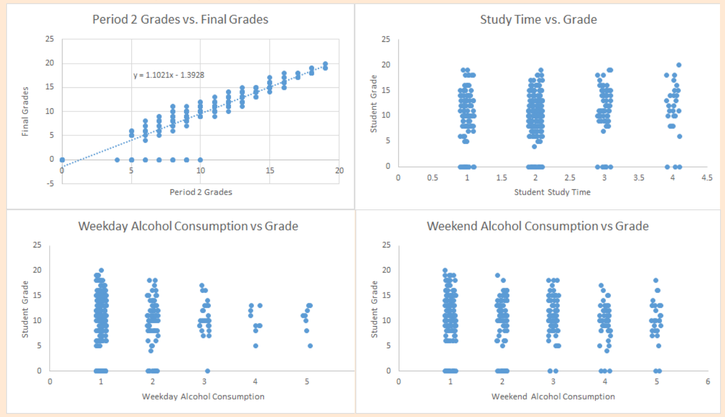
Anna Boonyanit (annaboon@stanford.edu), Carly Davenport (carly17@stanford.edu), Erik Van (evan32@stanford.edu)

OVERVIEW

Our project seeks to discover the relationship between alcohol consumption and academic performance contextualized with other potential influences. We aim to better understand how impactful alcohol consumption is on a student's final grade in relation to different socioeconomic and behavioral factors.

DATA & INITIAL OBSERVATIONS

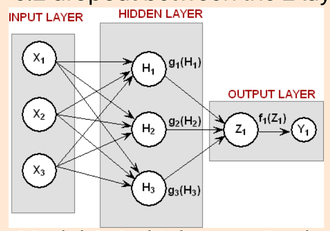
- Utilized a dataset from Kaggle that recorded the final grades of 395 students and their social information
- The inputs were the socioeconomic (i.e. parent education) and behavioral (i.e. alcohol consumption) factors, which was formatted as either categorical (yes/no) or numerical (Rating between 1-5)
- The numerical data was standardized and categorical data was one-hot-encoded. For example, the variable "Mother job" was divided such that each job was a different factor, (i.e. 'Mjob_at_home') and has a value of either 0 or 1.
- The output is the prediction of the student grade based on the following format based on Portugal's grading system:
  - 16-20 = A, 14-15 = B, 10-13= C, 0-9 = F.



- Based on preprocessing data analysis, one key finding is that a student's ability to pass a class is strongly dependent on study time and whether or not students pay for a class. However, individual grade scores are dependent on weekday alcohol consumption and parent education.
- Most students claim a low weekday alcohol consumption so the wide range of grades is concentrated in these low ratings
  - Weekday consumption is weak indicator for grades.
- Weekend alcohol has a wider distribution of grades and there is a very slight downward trend in the highest student grade as the weekend alcohol consumption rating goes up
  - Weekend consumption is a stronger indicator for grades

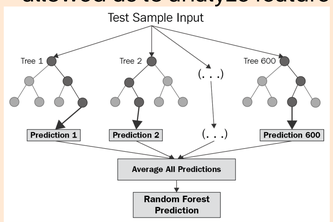
IMPLEMENTATION

- Utilized two classification models
- Model 1: TensorFlow's Keras to implement a two layer neural net
- First layer had 100 nodes, used relu activation function
- Second layer had 4 nodes (4 outputs), used softmax function
- 0.2 dropout between the 2 layers



Two-layer neural net diagram

- Model 2: Scikit-learn to implement a random forest, which allowed us to analyze feature importance.



Random forest diagram

- We also tried implementing the neural network with the 15 most important features (according to the random forest) only

CHALLENGES

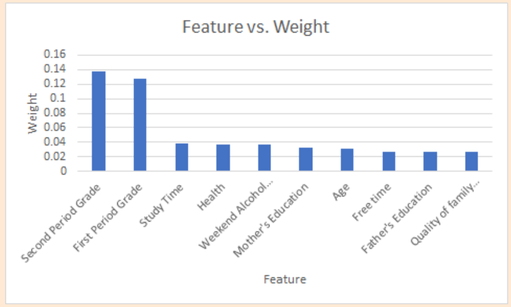
- Difficult to trace feature importance using neural network
- Small dataset makes it hard to get an accurate neural net, can't draw too many conclusions
- We attempted to solve these issues using random forest to determine the most important features and then only include those as inputs into the neural net to narrow the number of features.
  - The results improved slightly from the neural net, as we were over-fitting the model less. However, the random forest did not improve.

NEXT STEPS

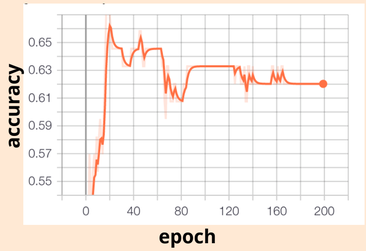
- Gather more data to run more comprehensive testing, since our dataset was rather small and therefore couldn't use larger neural net
- Try more models to see which features they consider most important and use those on the neural network
- Try to create a tool where you could input student information and it could return how likely they are to achieve a certain grade

RESULTS AND ANALYSIS

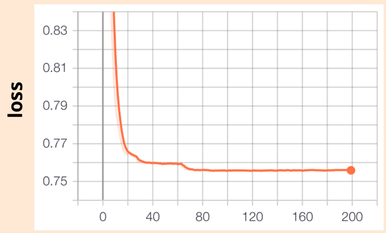
Method	Test set accuracy
Neural network w/ our data split and 30 epochs	60.1%
Neural network w/ Sklearn's split and 30 epochs	74.2%
Neural network w/ Sklearn's split And only top 15 features and 52 epochs	78.5%
Neural network w/ our data split And only top 15 features and 15 epochs	60.7%
Random forest w/ our data split	68.4%
Random forest w/ Sklearn's split	84.8%
Random forest w/ Sklearn's split And only 15 features	81.0%



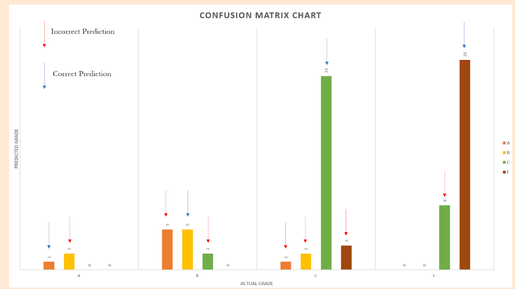
- We found that alcohol wasn't a very strong predictor of grades overall as the weights in the random forest demonstrated. However, weekend alcohol consumption still was a top 5 feature being weighed, so we can conclude that alcohol consumption has some influence on grades.
- The following TensorBoard graphs and confusion matrix graph represent a run with a neural network with Sklearn's data split and all features (this run achieved 70.8% accuracy)
- Test Accuracy - led us to utilize early stopping to reduce overfitting



- Training Loss



- Training Accuracy



- All mis-predictions were only by one letter grade except one (i.e. only once was the difference between true and predicted 2-letter-grades)
- There is a much higher rate of incorrect predictions for grades A and B (66.6% and 58.3% respectively) than C and D (20.0% and 23.5% respectively). This is likely because there is much less training data for A's and B's.
- 14 over predictions, 6 under predictions, suggesting model tends to over-predict