Predicting Students' Alcohol Consumption Using Gradient Boosting on a Weighted Alcohol Consumption Score

Introduction

Machine learning has taken over the space of predictive analytics in recent years and can be applied to virtually any subject domain. One major advantage of this is that new methodology and techniques are being invented and improved upon literally every day. The goal of this paper is to show how experimenting with some of these newer methods can improve predictability on a specified problem.

I don't think there's any dispute over the idea that student alcohol consumption is an issue and a hot topic at universities everywhere. During my college days, my university warned their students about the dangers of alcohol while at school. This was done by cramming thousands of students in the auditorium and delivering a lecture to an uninterested audience. My guess, this is not a very effective approach. To be effective, this type of promoting should be geared towards smaller groups, if not a 1 on 1 discussion. The problem here is that this would not be a feasible approach in dealing with an entire student body. So, what if the university could target those who might benefit the most from such a discussion? Well, that's where machine learning comes in.

Before completing my own study on this dataset, my intuitive hypothesis was that it would be entirely possible to predict a student's alcohol consumption using other characteristics about the student. To back me up on this notion, googling "student alcohol consumption" provides this image before all others (*1Dental.com* photo credit):

I am going to use characteristics about students that have nothing to do with alcohol consumption to predict whether the student might benefit from a 1 on 1 "intervention". To do this, I use a dataset found on kaggle that originated from a study completed by P. Cortex and A. Silva that was predicting secondary school student performance in Portugal back in 2008. The title of the study can be found in the **References** section of this paper. A major underlying assumption here is that findings from this dataset can be applied to other schools.

The dataset being used was obtained through a survey of math and Portuguese language students and contains social, gender and school-related information about the participating students. After reading the discussion board on kaggle, I determined I had 959 unique students to play with. The

full list of information provided on these students can be found in the **Data Exploration** section of this paper, but as a sneak peek, it contains information on where the student resides, the student's family relations, some of the student's extra-curricular activities and two questions that measure the student's alcohol consumption during the week and weekend.

The Problem

Predicting student alcohol consumption can have a broad understanding. For specificity purposes, I will explain how the alcohol consumption questions are measured first, and then how I plan to use them as a predictive outcome second.

There are two variables in the data that measure alcohol consumption, *Dalc* and *Walc*. *Dalc* measures workday alcohol consumption while *Walc* measures weekend alcohol consumption. Both variables are measured on a likert-scale from 1-5 where 1 indicates very low and 5 indicates very high. For simplicity, I combined these two variables into one using a weighted-average technique I learned while consulting on a project in grad school that examined a survey on week day and weekend exercise. This technique was shown to be an effective form of measuring a person's exercise (at least for that project) and so I decided to adopt the ideology behind it for my project. I call the combination of the two variables the "Weighted Average Alcohol Consumption Score" or WAACS for short. Here's how it works:

- Multiply each variable by their respective weight
    - $(5/7)$ for work day alcohol consumption
    - $(2/7)$ for weekend alcohol consumption
- Take the average of the two weighted scores

The reason work day alcohol consumption is weighted at $(5/7)$ and weekend alcohol consumption is weighted at $(2/7)$ is because there are typically 5 work days and 2 weekend days in the week. Also, for my purposes and in my opinion, it's important to have work day alcohol consumption weighted higher as it would likely be more detrimental to the well-being of the student – so it works out. The equation for WAACS is as follows:

$$WAACS = \frac{\left(\frac{5}{7}\right) * Week\ Day\ Alcohol\ Consumption + \left(\frac{2}{7}\right) * Weekend\ Alcohol\ Consumption}{2}$$

Since I am interested in identifying students who would benefit the most from some type of intervention or discussion I decided to break this problem out into two potential paths. For both paths, I defined the highest set of scores as individuals at or above the 3rd quartile of WAACS scores.

The first path is to predict WAACS and then look at the highest scores. For this path, I will explain how I used regression techniques to predict WAACS scores first, and then discuss those who are in the highest set of scores.

The second path is to first assign the highest scores, and then predict who is in the top set of scores. For this path, I will be using binary classification techniques to predict a binary indicator of whether someone is in the highest set of scores.

Measuring Performance

Measures for the predictions vary from each problem path and so are not comparably "apples to apples", but both are crucial for concluding how effective the predictions are.

Recall that for **the first path**, I am predicting WAACS score. To do this, I utilize different regression techniques and assess the strength of my predictions and how they improve by looking at the mean squared error (MSE) and adjusted R-squared (adj. R^2). MSE measures the difference between the predicted WAACS score and the actual WAACS score. The Adj. R^2 measures the variance in the WAACS score that is explained by the model while considering the dimensions of the dataset.

Recall that for **the second path**, I am predicting whether a student is at or above the 3rd quartile of WAACS scores. To do this, I utilize different binary classification techniques and assess the strength of my predictions and how they improve by looking at accuracy, precision and recall. Because of the nature of my problem, I wanted to focus my attention on precision and recall and so thus I also measured the F1 score, which is the average of precision and recall. The precision is measuring the % of students who were at or above the 3rd quartile out of the ones who were predicted to be. The recall is measuring out of all the students who were at or above the 3rd quartile, the % of students who were accurately predicted. Since I am predicting students who would benefit from an intervention or discussion about alcohol, I wanted to focus on those who are at or above the 3rd quartile and thus the F1 score is a perfect measure to assess the model.

$$MSE = \frac{\sum_{i=0}^{n}(Actual\ WAACS - Predicted\ WAACS)^2}{n}$$

$$Adj.\ R2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

$$Accuracy = \frac{\#\ of\ Accurately\ Predicted\ Students}{Total\ \#\ of\ Students}$$

$$Precision = \frac{\#\ of\ Accurately\ Predicted\ Students\ in\ 3rd\ Quartile}{Total\ \#\ of\ Predicted\ Students\ in\ 3rd\ Quartile}$$

$$Recall = \frac{\#\ of\ Accurately\ Predicted\ Students\ in\ 3rd\ Quartile}{Total\ \#\ of\ Actual\ Students\ in\ 3rd\ Quartile}$$

$$F1\ Score = \frac{Precision + Recall}{2}$$

Data Exploration

As promised, the full list of variables (45 total after structuring) that were used in the predictions of WAACS scores in addition to some descriptive stats accompanying them.

- **School** - Gabriel Pereira or Mousinho da Silveira
- **Sex** – Male or Female
- **Age** – Student's Age (15 -22)
- **Address** – Urban or Rural
- **Famsize** – Family size (> or <= 3)
- **Pstatus** – parent's cohabitation; living together or apart
- **Medu** – Mother's Education; None to Higher Education
- **Fedu** – Father's Education; None to Higher Education
- **Mjob** – Mother's Job (4 categories and other)
- **Fjob** – Father's Job (4 categories and other)
- **Reason** – Reason to Choose School (3 categories and other)
- **Guardian** – Student's guardian (Mother, father or other)
- **Traveltime** – Home to School Travel time (4 numeric options)
- **Studytime** – Weekly Study Time (4 numeric options)
- **Failures** - # of Past Class Failures (numeric)
- **Schoolsup** – Extra Educational Support
- **Famsup** – Family Educational Support
- **Activities** – Extra-Curricular Activities
- **Nursery** – Attended Nursery School
- **Higher** – Wants a Higher Education
- **Internet** – Internet access at home
- **Romantic** – In a Romantic Relationship
- **Famrel** – Quality of Family Relationships
- **Freetime** – Free time after school (numeric 1 – very low to 5 – very high)
- **Goout** – Going out with friends (numeric 1 – very low to 5 – very high)
- **Health** – Current Health Status (numeric 1 – very bad to 5 – very good)
- **Absences** - # of School Absences (numeric 0 – 93)

Some data structuring was needed before any analysis could be performed. First, the data had to be merged as its source was 2 separate excel files – 1 for the math students and 1 for the Portuguese language students. Based on discussions on kaggle (ultimately suggested by Carlo Ventrella), there are 82 students who belong to both datasets and merging was to be done on all data set features with the exclusion of the **paid** variable (also excluded in the analysis). You can see the discussion here: https://www.kaggle.com/uciml/student-alcohol-consumption/discussion/26889

Additional structuring consisted of forcing each of the following binary variables as "dummy" variables where 1 indicates a yes and 0 indicates a no:

- Schoolsup
- Famsup
- Activities
- Nursery
- Higher
- Internet
- Romantic

The rest of the survey questions that were categorical in nature were one-hot encoded whereas the scalar questions were left alone.

The Walc and Dalc Variables → WAACS

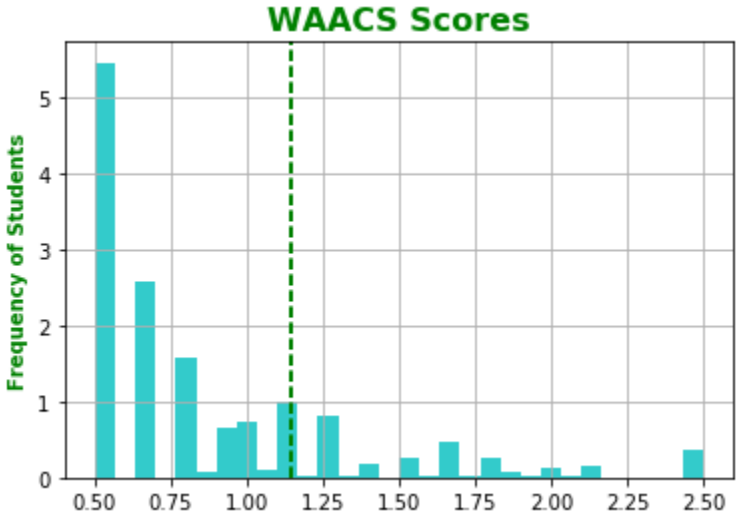
Below is a correlation matrix for the Dalc, Walc, WAACS and at or above the 3rd quartile for WAACS variables:

	Dalc	Walc	WAACS	WAACS ↑ 3rd Q.
Dalc	1.00	0.63	0.95	0.81
Walc	0.63	1.00	0.84	0.69
WAACS	0.95	0.84	1.00	0.84
WAACS ↑ 3rd Q.	0.81	0.69	0.84	1.00

The correlation coefficient between WAACS and Dalc is 0.95 and the correlation coefficient between WAACS and Walc is 0.84, while the correlation coefficient between Dalc and Walc was only 0.63. This suggests to me that the WAACS score did a good job at catching the variance of both Dalc and Walc in a single score. Some key points about these variables are:

- The median value for Dalc was 1.0
- The median value for Walc was 2.0
- The average WAACS score was 0.870624
- There are 254 students who are at or above the 3rd quartile for their WAACS scores
    - The 3rd quartile score was 1.142857

The following histogram shows the frequency of students for each WAACS score. All students at or above the 3rd quartile for WAACS scores are to the right of the dashed line:



The Predictor Variables

There were 45 total predictor variables that were considered in the analysis. Though each are interesting in their own way, I felt that the following table does a good job at showing what's going on with the student population:

Predictor Summary	All Students (n = 959)	Students Below 3rd Quartile (n = 714)	Students Above 3rd Quartile (n = 245)
AVG. Age	16.8	16.7	16.9
% Male	43.6%	36.4%	64.5%
% Female	65.4%	63.6%	35.5%
Median Travel Time	<15 min	<15 min	<15 min
Median Study Time	2 to 5 hours	2 to 5 hours	2 to 5 hours
Median Past Failures	0.0	0.0	0.0
% w/ Families > 3	70.6%	72.4%	65.3%
% w/ Internet Access	78.7%	77.5%	82.5%
AVG. Going Out	3.2	3.0	3.8
Quality of Family Relations	3.9	4.0	3.8
AVG. Current Health Status	3.6	3.5	3.7
AVG. Free Time After School	3.2	3.1	3.5
% w/ Separated Parents	11.9%	12.3%	10.6%
% Romantically Involved	35.8%	36.6%	33.5%
AVG. # of Absences	4.7	4.2	6.2
Median # of Absences	2.0	2.0	4.0
% Wanting a Higher Ed.	90.8%	92.3%	86.5%
% Chose School for Close to Home	24.3%	23.7%	26.1%
% Chose School for Course Preferance	41.5%	42.4%	38.8%
% Chose School for School's Rep.	23.6%	25.5%	18.0%
% Chose School for Other Reasons	10.6%	8.4%	17.1%
AVG. Score for Father's Education	5th to 9th grade	5th to 9th grade	5th to 9th grade
AVG. Score for Mother's Education	secondary ed.	5th to 9th grade	secondary ed.
% w/ Extra Educational Support	11.2%	11.6%	9.8%
% w/ Family Educational Support	61.3%	62.3%	58.4%
% Attended Nursery School	80.1%	81.4%	76.3%

An additional inference I want to note about these descriptive stats are that the following variables all have one thing in common – the % difference for students above the 3rd quartile compared to students below is **greater than 10%:**

- % Male
- Avg. Frequency of Going Out
- Avg. Free Time After School
- Avg. & Median # of School Absences
- % Who Chose the School Because It Was Close to Home & Because of Other Reason

Similarly, the following variables are with the theme that the % difference for students above the 3rd quartile compared to students below is **less than 10%**:

- % w/ Separated Parents
- % w/ Extra Educational Support
- % Who Chose School for Its Reputation

So based on these 8 variables, it seems that the students above the third quartile tend:

1. To be male
2. To go out w/ friends a lot
3. To have a lot of free time after school
4. To miss a good amount of school
5. To have chosen the school because it was close to home or because of other reasons
6. To have parents who are together
7. To not have extra educational support
8. To not have chosen the school for its reputation

Methodology

Because this problem has the 2 paths, there are 2 methods that were used. First, recall the two paths:

1. Predict WAACS and then assess the top scores (those at or above the 3rd quartile)
2. Predict those who are at or above the 3rd quartile of WAACS scores

The nature of the first path calls for a regression technique as the WAACS scores can take on scalar values between 0.5 and 2.5. The second path calls for a classification technique as students are either at or above the 3rd quartile, or they are not. One beautiful aspect of the newer branches of machine learning is that algorithms used for either of my paths are adapted for the other. For example, I will give an overview of two methods that I used to see improvements on my benchmark models (linear regression & logistic regression, paths 1 & 2 respectively). The two algorithms were examined for both problem types and they were ***random forest*** and ***gradient boosting***. The tool used to perform this analysis was python 2.7.

First, my benchmark models that I compared to – linear regression and logistic regression. The linear regression used for the benchmark modeled the relationship between all 45 predictor variables and the WAACS scores. This technique captured coefficients on how each variable acts as a linear equation to solve for WAACS scores. The logistic regression models a likelihood that a student is at or above the 3rd quartile. In this case, each of the predictor variables contribute to a probability. Both methods lack one main component of successful machine learning techniques – *randomness*. Random forest and gradient bosting both utilize this quality in their modeling, in addition to iterating, which allows the model to narrow in on the *importance* of each predictor variable as they are tested w/ subsets of the data.

Importance can be defined differently across techniques, but the main point that is common in each definition is that it is a measure of influence on the variance of the outcome given by one variable.

Not to be interpreted as a coefficient, importance gives insight on how the model utilizes its randomness. To be more specific, I will give a brief overview of random forest and gradient boosting methods:

Random forest is an ensemble of decision trees where each tree is built on a *random* subset of the data, and chooses the most important variable for splitting in each subset. This is done for n trees and then each tree "votes" for importance and thus variables are ranked based on the votes across all trees in the forest.

Gradient boosting methods use a loss function to minimize the error modeled in *random* subsets of the data in the same fashion of a random forest. This method, however, focuses its attention on outcomes that have the largest error or are incorrectly classified, making this technique theoretically more powerful. Each iteration has its own set of weights for the observations present allowing for large error or miss-classified observations to have higher consideration in the model.

For each path, improvements on the benchmark models are applied using both random forest and gradient boosting. In each case (i.e. regression and classification), gradient boosting outperformed random forest and is the model of my choice. Because of this, I will be focusing most of my attention on the results of this method.

As a means of testing the performance of the benchmark models and improved models, I split the dataset into a training dataset and a testing dataset with an 80/20 split. This resulted in the training dataset having 767 rows and the testing dataset having 192 rows. Each model will be built only considering the training data and they will be evaluated on the unseen testing data.

Analysis

Recall that for my benchmark models I used linear & logistic regression. The measures of performance I used to compare models are the following with the respective results of each benchmark model:

Path 1 - Regression

	Benchmark
MSE	0.14
Adjusted R^2	0.33

Path 2 - Classification

	Benchmark
Recall	0.38
Precision	0.58
F1-Score	0.46

All results recorded are based off the testing dataset.

Improvements

Each improvement model was carefully tuned using a grid search technique scoring on R^2 for regression and the F1-Score for classification. How this works is that the model is evaluated on

different sets of values for each parameter and the value that results with the best R^2 or F1-Score were used.

The parameters that were tuned for the random forest model were:

- The number of trees in the forest
- The number of features to consider when looking for the best split
- The maximum depth of the tree
- The minimum number of samples required to split an internal node
- The minimum number of samples required to be at a leaf node

The parameters that were tuned for the gradient boosting model were:

- The number of boosting stages to perform
- Maximum depth of the individual estimators
- The minimum number of samples required to split an internal node
- The minimum number of samples required to be at a leaf node

Each model used a fixed learning rate of 0.1. The following results for the regression path indicate the performance of the gradient boosting method beat random forest, and allow for a major improvement on the benchmark models when assessing the adjusted R^2:

	Benchmark	Random Forest Tuned	Gradient Boosting Tuned
MSE	0.14	0.15	0.15
Adjusted R^2	0.33	0.75	0.90

Similarly, the results of the classification models indicate that gradient boosting out-performs random forest when assessing recall, precision and the F1-Score:

	Benchmark	Random Forest Tuned	Gradient Boosting Tuned
Recall	0.36	0.49	0.71
Precision	0.65	0.69	0.85
F1-Score	0.47	0.57	0.77

The adjusted R^2 for the gradient boosting model is 2.6 times the benchmark, where I'm now able to infer that nearly 90% of the variance in the WAACS scores is explained by the gradient boosting model. The F1-Score for the classification is 1.6 times the benchmark, a large jump from 0.47 to 0.77. Both models achieve a very reasonable result that gives me confidence in the improvement and possibility of using such models at scale. I will compare the regression and classification models and give my input on which is the better model in the **Regression or Classification?** section.

Now that I have established the best method for each path, I want to dive a little deeper inside the models by looking at the importance of the predictors for the gradient boosting models for each

path. Though the regression and classification paths are predicting 2 different metrics, there are a lot of overlapping inferences that can be made by assessing the importance rankings of each. For example, if I wanted to look at the top 10 most important variables for each path, I notice that 9 out of 10 of the variables are the same.

The importances can be seen visually in the following plots:

During the exploratory phase of this project, I noticed that the students with the higher WAACS scores tended to have certain characteristics. The importance plots assured me that many of these characteristics had a valid influence in WAACS score. These characteristics included the average frequency of going out, average free time after school, the number of school absences and choosing the school because it was close to home, all of which ended up in the top set of important variables.

Regression or Classification?

So far I've been examining 2 paths for the problem of identifying students who may benefit the most from an intervention or discussion on the dangers of alcohol consumption at school. So, which path is the one that will be the most effective? I propose that path 1 (i.e. regression on the WAACS scores) is the most effective solution and here's why:

Predictions based off path 1 yield estimated WAACS scores. I already know what the recall, precision and F1 scores are for the classification method and so to determine whether the regression does a better job, I needed to do a conversion that would allow me to compare "apples to apples".

To do this, I assigned all the estimated WAACS scores a 1 if it was at or above the 3rd quartile for all WAACS scores in the original dataset. I then calculated recall, precision and the F1 score for these newly classified predictions, and received the following results:

	Logistic Regression Benchmark	Gradient Boosting Classification	Gradient Boosting Classified Regression Predictions
Recall	0.36	0.71	0.87
Precision	0.65	0.85	0.96
F1-Score	0.47	0.77	0.91

Implementing this process on the predicted WAACS score resulted in a 23.1% increase in recall, a 13.2% increase in precision, and overall, a 18.4% increase in the F1-Score.

Discussion and Future Work

Student alcohol consumption is an issue at universities everywhere. I'm sure methods for dealing with this issue vary from school to school, however, the effectiveness can be debatable. After thoroughly analyzing the dataset shared on kaggle, I propose an approach for using information about the students having nothing to do with alcohol to identify which students may benefit the most from a small group, perhaps 1 on 1 "intervention".

Using gradient boosting to predict a weighted average alcohol consumption score (WAACS), I was able to achieve a F1-Score of 0.91 on a subset of the data that was not used in the building of the model. This score represents identifying students who are at or above the 3rd quartile of WAACS scores. Ultimately, my approach utilizes components of both regression and classification, and assesses measures accordingly.

I'm aware that the population of students used in this study is not very large, however, what I am proposing is that these findings should be enough evidence to utilize a method such as this to identify students who may benefit from an "intervention". I believe that the power of machine

learning methods can be utilized in virtually any domain, and my hope is that this study proves its use in dealing with a problem at a potentially very large scale.

There are new algorithms and techniques coming out every day. For this problem, I found that gradient boosting proved to be successful. An improvement on this method has come out quite recently called *xgboost* and I would be very interested in seeing how well it works with this very same problem.

An important aspect of this problem is the limitation to the size of the dataset. If I had access to a larger dataset with the same survey results, I may find that some survey questions are more important than what my study has shown. Hopefully the idea catches on and more data becomes open sourced so that a refresh on this analysis can be possible.

References

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.

Aarshay Jain. May 28, 2016. Complete Guide to Parameter Tuning in Gradient Boosting (GBM) in Python. Aanalyticsvidhya. April, 2017. https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/