Capstone Project 1- Predicting Alcohol Use of College Students

Final Report

Marisa Mitchell

Introduction

This project investigates the alcohol consumption of college students. Roughly 60 percent of college students ages 18-22 drank alcohol in the past month and of those students nearly 2 out of 3 engaged in binge drinking. Binge drinking can lead to a variety of harmful consequences for these college students such as death, assault, sexual assault, and academic problems. College level administrators and working groups may be interested in minimizing these consequences for students so that they can provide interventions to help these students. The aim of this project is to develop a way to predict which students are at risk of high alcohol use and thus abuse so that they can make informed decisions about interventions they provide to these students.

Dataset

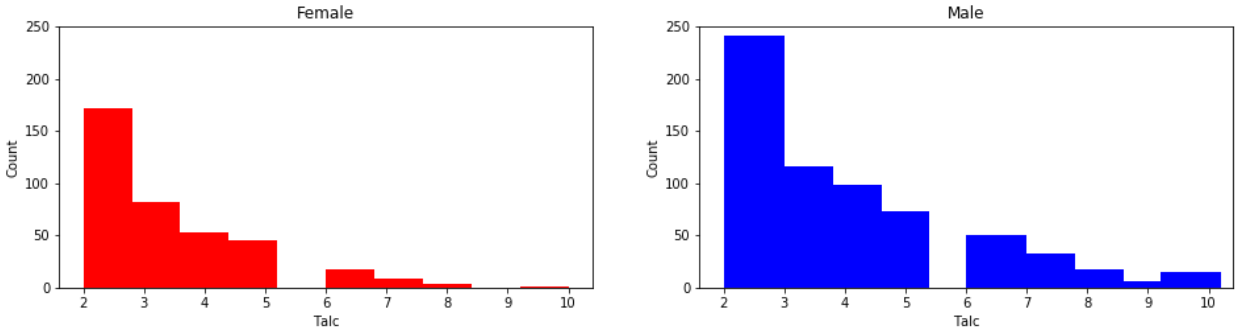
The dataset used in this project was the UCI Machine Learning Student Alcohol Consumption dataset located on Kaggle (https://www.kaggle.com/uciml/student-alcohol-consumption/data). This dataset contains 33 variables related to 649 students who were registered in a math Portuguese language course at one college. The variables contained in this dataset relate to the student's sex and age as well as information about their family (e.g., family size, parent's education level, parent's job), school related variables (e.g, travel time to school, studytime, and past course failures), student's social life (e.g., if they are in a romantic relationships, time going out with friends, and free time), alcohol consumption (both workday and weekday), and academic information (e.g., absences from school and grades from first period, 2nd period and final grade in the Portuguese class). The workday and weekend alcohol consumption variable are numeric ratings from 1(very low) to 5 (very high). This dataset did not have any missing values for any of the variables. I created a new composite variable of total alcohol consumption by adding the value of the workday and weekend variables together. This new variable for total alcohol consumption was then binned into three groups with low total alcohol use category representing students with total alcohol use equal to or below 3, medium use representing students with a total alcohol use between 4 and

6, and high alcohol representing those with total alcohol values of 7 or higher. Additionally I created dummy variables for several of the categorical variables in the dataset such as sex, address (urban or rural), family size, mother's education, and father's job. This data configuration was used for several of the models that I used to predict level of total alcohol use. I created a second configuration of the data where several string variables (e.g. School, sex, address, family size) were converted to numerical values. This second configuration of data was used for the tree-based models I evaluated.

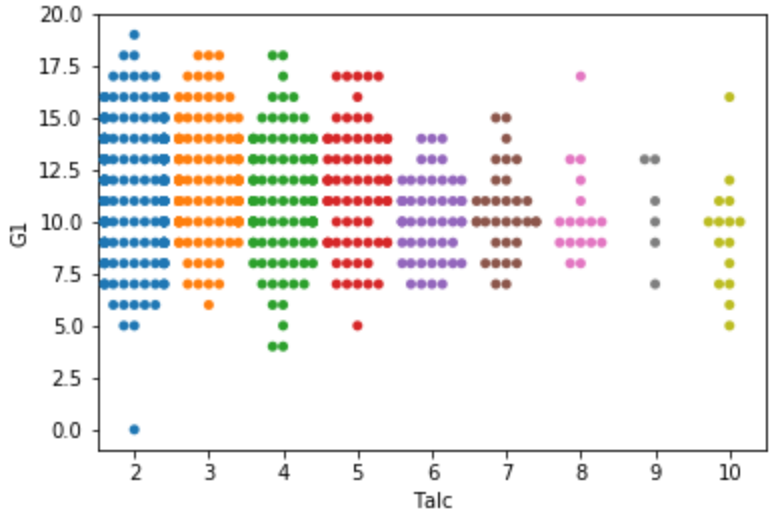
Exploratory Analysis

The findings of this report are given without the complete Python code. The Python code in which the findings are based can be found on Github.

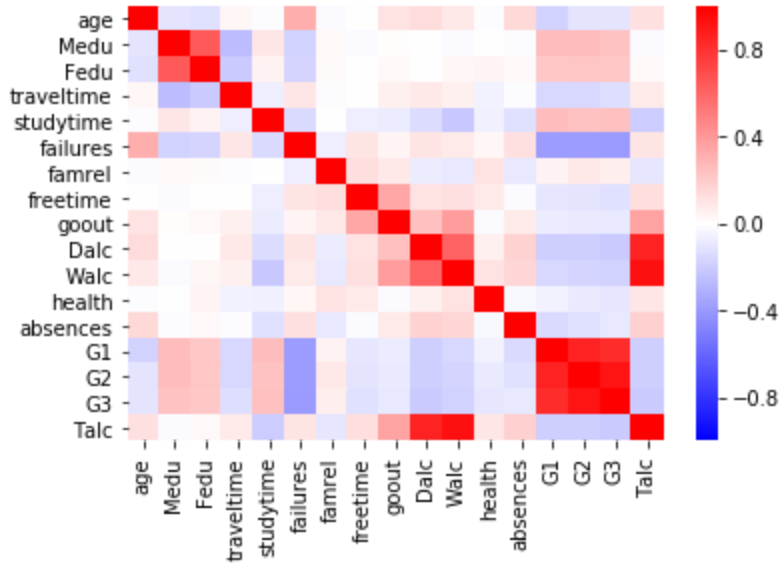
Based on descriptive statistics, exploratory data analysis, and correlations male students have a higher mean total alcohol consumption that female students.



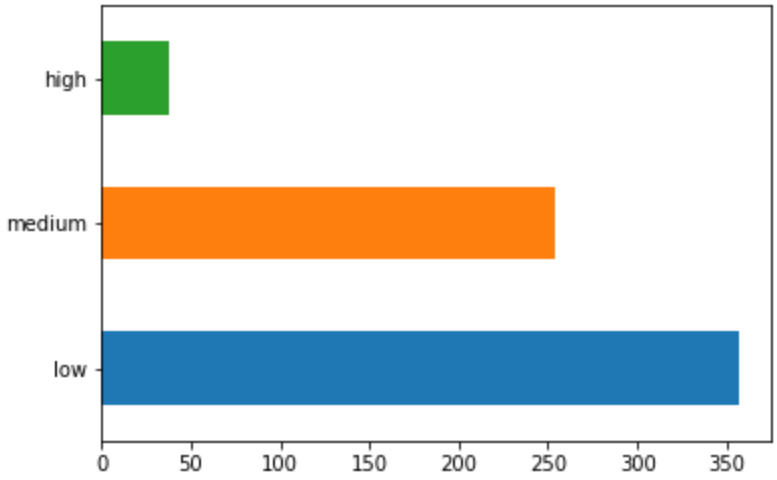
The average total alcohol consumption was 3.78 for all students regardless of sex. The average age of students in the dataset was 16.74. There was a moderate and positive correlation between workday and weekend alcohol consumption (pearson r= .62) which was statistically significant. The amount of time students go out with friends was weakly and positively correlated with the total alcohol consumption of students (person r=.36). Previous class failures had a weak and negative correlation to grades in the course (pearson r's ranged from -.38 to -.39). Workday and weekend alcohol consumption had strong and positive correlations to total alcohol consumption (.86 and .93 respectively).



According to my swarmplot there are many more students who have a low total alcohol consumption than those who have higher levels of total alcohol consumption.



The above heatmap shows the correlations between the variables in the dataset. Grades from first period, second period, and final grades were all strongly and positively correlated with pearson r values ranging from .82 to .92. Weekday and weekend alcohol use were moderately and positively correlated with a pearson r value of .62.

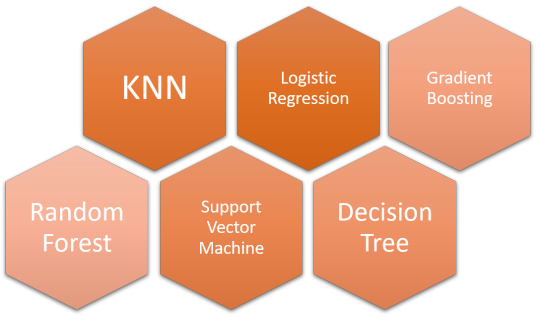


The above bar chart represents the number of students who were in the high, medium, and low total alcohol use categories once the data was binned. There was a relatively low representation of students in the high total alcohol use group (38) in comparison to those in the medium(254) and low (357) groups.

Analytical Method

The objective of this project was to produce an algorithm that would correctly classify students in one of 3 groups (high, medium, and low total alcohol use). I selected these six classification algorithms to model with my data; K nearest neighbors (KNN), logistic regression, gradient boosting, random forest, support vector machine (SVM), and decision tree. After training and testing these algorithms I choose the top 3 performing algorithms to do additional hyper-parameter tuning. After conducting hyper-parameter tuning on the top 3 models I analyzed the performance metrics in order to draw conclusions on which model would be the best to use.

As stated above I used two different configurations of data in this analysis. For the KNN, logistic regression, and SVM models I used the dataset in which categorical variables were dummy coded. The tree based models (decision tree, random forest, and gradient boosting) used the configuration of data which involved converting the categorical string variables to numerical variables.



Model Results and Evaluation

Initial Results

All models were run prior to hyper-parameter tuning using default settings. Results are shown in the below table. I used test data accuracy scores to select the highest performing models. This score represents the percent of cases that the model predicted the correct label. In this case the number of students that were correctly labeled with their actual total alcohol use group. The model with the highest accuracy score was SVM with an accuracy of 0.641. The logistic regression model had the next highest accuracy score 0.636. The third highest performing model was the gradient boosting model with an accuracy score of 0.621. Next was the random forest model which had an accuracy score of 0.615. The second worst performing model was the KNN model which had an accuracy score of 0.564. The decision tree model was the worst performing model with an accuracy score of 0.503. This means that the top three performing models in which I choose to do further hyper-parameter tuning on were the SVM, logistic regression, and gradient boosting models which were able to correctly predict 64.1%, 63.6% and 62.1% of students' total alcohol use categories respectively.

Model	Accuracy Score (Before Tuning)
KNN	0.564
Logistic Regression	0.636
SVM	0.641
Decision Tree	0.503
Random Forest	0.615
Gradient Boosting	0.621

Model Tuning Results

In order to tune the hyper-parameters of the three highest performing models I used the grid search function in the scikit-learn package. For the SVM model I selected the kernel and C (the penalty for misclassification) hyper-parameters to tune on. The kernel options I chose were the linear and the radial basis function (rbf) kernels. The C values I tuned on were 1, 10, 100, and 1,000. For the logistic regression model I chose to tune on the penalty (type of regularization used) and C hyper-parameters. I used the l1 and l2 penalties and C values of 1, 10, 100, and 1,000 for the hyper-parameter values. In the gradient boosting model I tuned on the n_estimators (number of sequential trees to be modeled) hyper-parameter with the values of 50, 100, 200, 300, and 500.

After tuning the logistic regression model had the highest accuracy score with a value of 0.646. The SVM model had the second highest accuracy score of 0.641 after tuning. Finally, the gradient boosting model had the lowest accuracy score of 0.631.

Model	Accuracy Score (After Tuning)
Logistic Regression	0.646
SVM	0.641
Gradient Boosting	0.631

To further evaluate these models I examined the confusion matrix and classification report for each of these models using the best estimators. Below you will find tables of the confusion matrix and classification report for each of these models.

Logistic Regression Confusion Matrix and Classification Report

	high	low	medium
high	1	2	7
low	1	92	23
medium	1	35	33

Group	Precision	Recall	f1-score	support
avg/total	0.63	0.65	0.63	195
High	0.33	0.10	0.15	10
Low	0.71	0.79	0.75	116
Medium	0.52	0.48	0.50	69

SVM Confusion Matrix and Classification Report

	high	low	medium
high	0	2	8
low	0	98	18
medium	0	42	27

Group	Precision	Recall	f1-score	support
avg/total	0.59	0.64	0.61	195
High	0.00	0.00	0.00	10
Low	0.69	0.84	0.76	116
Medium	0.51	0.39	0.44	69

Gradient Boosting Confusion Matrix and Classification Report

	high	low	medium
high	3	2	5
low	0	87	29
medium	2	34	33

Group	Precision	Recall	f1-score	support
avg/total	0.63	0.63	0.63	195
High	0.60	0.30	0.40	10
Low	0.71	0.75	0.73	116
Medium	0.49	0.48	0.49	69

Although the logistic regression model had the highest accuracy of any of the models the recall for each group illustrates that it was rather poor at classifying the students in the high alcohol use group with a recall of only 0.10, meaning that only 10% of the students who are actually in the high alcohol group were identified as such by the model. Further the logistic regression model was much better at correctly identifying the students in the low alcohol group with a recall score of 0.79.

The SVM model which had the second highest total accuracy score was actually found to be the worst at classifying the students in the high alcohol group with a recall score of 0.00. This model did not predict any of the students who were in the high alcohol group, instead identifying most of them as the medium alcohol group. Again, this model had the highest recall for the low group with a score of 0.84.

Despite the gradient boosting model having the lowest accuracy score it was actually the best at identifying the students in the high alcohol group with a recall score of 0.30. Similarly to

the other models it was also the best at predicting the students in the low alcohol group with a recall score of 0.75.

Conclusions

The purpose of this project was to create a model which is able to predict the level of college students' alcohol use so that college level administrators can use this prediction to be proactive and target students with interventions that would prevent potential harmful effects of alcohol abuse. Overall the logistic regression model had the highest levels of accuracy in classifying students' alcohol level after hyper-parameter tuning. However, this model performed rather poorly in correctly identifying the students in the high alcohol use group. The model which had the highest rate of correct predictions for the high alcohol group was the gradient boosting model.

In fact, all models had a difficult time predicting the students in the high alcohol group with much more accuracy in predicting those in the low alcohol group. This may not be a problem if the school wants to target students in the medium and high alcohol groups but would be a potential limitation if the school chooses to only target those identified as in the high alcohol group.

My recommendation to college level administrators using these models would be to use the gradient boosting model if they choose to only provide interventions to those students in the high alcohol group. This may be the most cost effective method as it would mean providing interventions to the least amount of students. If administrators have the means to target a larger group of students I would suggest that they use the logistic regression model. This model would work relatively well as most of the students in the high alcohol use group would be incorrectly identified as medium alcohol use but would still receive interventions.

Next Steps

The largest challenge in creating a model which accurately predicts the students in the high alcohol group is the small number of students in the dataset which were actually in the high alcohol group. There were only 38 students in the high alcohol use group in the dataset in comparison to the 254 and 357 students in the medium and the low groups, respectively. This created an unbalanced dataset which made the analysis difficult. To correct for the unbalanced dataset the following next steps could be taken to improve the models:

- Collect more data
- Apply resampling techniques such as oversampling or undersampling
- Generate synthetic samples
- Use a penalized model

In addition to next steps to deal with the unbalanced data future work on this project could involve an investigation into the most important features. This would help college administrators be able to further understand which potential factors are associated with high alcohol use and in turn help them to develop interventions and supports targeted at those factors.