Analyzing of Student Alcohol Consumption and Consequences

Aid Semić [a,1,*]

[a] International Burch University, Sarajevo, Bosnia and Herzegovina
[1] aid.semic@stu.ibu.edu.ba
* corresponding author

ARTICLE INFO

Article history
Received July 20, 2022
Revised August 7, 2023
Accepted May 22, 2022

Keywords
alcohol consumption
students
grades
prediction
EDA
XGBoost

ABSTRACT

Student alcohol consumption is very interesting topic today. It is known that average number of students who consume alcohol today is rapidly increasing. Here is idea to analyze how it affect their studying and grades. To predict the grades of students we have to take a lot of aspect in consideration. This research uses a range of different parameters and machine learning algorithms, such as Exploratory Data Analysis and XGBoost. By using this methodologies and algorithms I have to admit that I didn't develop the best model for prediction.

1. Introduction

Underage and harmful college drinking are serious public health issues that have a devastating impact on students' lives on campuses around the world. College drinking has become a tradition that many students view as essential to their time in higher education. Some students already have drinking habits before they enroll in college, and the atmosphere there might make things worse. A US national survey found that nearly 33% of full-time college students between the ages of 18 and 22 who consumed alcohol in the previous months also engaged in binge drinking.

According to the most recent data from the US National Institute on Alcohol Abuse and Alcoholism (NIAAA), roughly 1,519 college students between the ages of 18 and 24 are thought to have perished as a result of alcohol-related unintentional injuries, including car accidents. According to the most recent NIAAA figures, over 696,000 students between the ages of 18 and 24 are allegedly assaulted by fellow students who have consumed alcohol. Researchers have verified a long-standing finding that 1 in 5 college women experience sexual assault during their time in college, despite the fact that calculating the incidence of alcohol-related sexual assaults is extremely difficult because sexual assault is often underreported. Most sexual assaults that occur in colleges involve alcohol or other drugs. Research on the connections between alcohol use and sexual assault among college students is still ongoing. In order to more accurately estimate the number of assaults due to alcohol, additional national survey data are required.

About one in four college students claim that drinking has caused them to miss class or fall behind on their assignments. According to a nationwide poll, college students who binge drink at least three times a week are nearly six times more likely than those who drink in moderation to perform poorly on an exam or assignment (40 % vs. 7%). Additionally, binge-drinking students had a fivefold higher likelihood of skipping a class (64 % vs. 12 %).

2. Related Work

This study investigated the association between two performance-related student metrics and drinking group norms. The multilevel hypotheses that (a) house alcohol climate is associated with student alcohol consumption, (b) student alcohol consumption is associated with student withdrawal behavior (i.e., missing class) and (c) that student alcohol consumption mediates the link between house alcohol climate and student withdrawal behavior are supported by data from 96 undergraduate students (mean age 22 years) living in 21 student houses. There was no correlation between student alcohol use and academic performance (i.e., average grades). The theory that home cohesion would mitigate the association between house alcohol climate and student alcohol consumption also received little empirical support. Future research implications are highlighted.

A sample of 106 employed college students was used to evaluate the intra-individual correlations between daily work pressures and alcohol use over a period of 14 days. Based on a theoretical framework for tension reduction, we hypothesized that exposure to workplace stressors would lead to increased alcohol use among employed college students, especially among men and those with higher daily expectations for the tension-relieving effects of alcohol. We discovered that hours worked were positively correlated with the quantity of beverages consumed after adjusting for the day of the week. Work-school conflict and workload were unrelated, but drinking was negatively correlated with both, especially when students strongly believed that alcohol had a calming effect. There was no proof that gender had any effect on how stressful job was. The findings show that employment throughout the academic year affects college students' drinking significantly and imply that the job context would be a suitable intervention setting to address the issue of student drinking.

This study intends to evaluate the associations between tobacco and alcohol use, physical activity (PA), and demographic factors among students in Romania. There were 253 participants in this study, students at Transilvania University in Brasov (112 men and 141 women). The findings demonstrated that moderate cigarette use and harmful alcohol use were both highly prevalent across age, gender, study year, and PA level groups. Despite the fact that there was no significant association between alcohol usage and tobacco use, it was determined that there was a negative correlation between PA level and cigarette use. Additionally, there was a negative significant association between alcohol usage among students and age, year of study, and PA level. In the end, the model used to predict alcohol and cigarette use had heterogeneous coefficients.

Significant progress has been achieved in defining and forecasting youth alcohol intake using cognitive and developmental techniques. The subjective expected utility (SEU) theory, the theories of reasoned action and planned behavior, and the alcohol-related outcome expectancy theory are only a few of the decision-making theories that are examined in the current review. Also discussed is the developmental literature on how peers and parents influence teenage alcohol usage. The theory of planned behavior and alcohol-related outcome expectancy theory are combined in the proposed model, with changes based on results from the developmental literature. Discussion is had about the implications for future study.

3. Results and Discussion

3.1 Dataset

The datasets that we have used in this work is taken from kaggle. The author of this dataset is Ruslan Sikhamov. The dataset contains data on 395 students.33 features are available. They are mostly category variables. While some variables appear numerical, they are actually categorical. For instance, Medu represents the mother's education (numerical values: 0 for no education, 1 for primary education (4th grade), 2 for secondary education (5th to 9th grade), or 3 for further education). There are just 5 actual numerical characteristics: age, absences, grades, and so on (G1, G2, G3).

There are no missing values. Each row in this dataset represent a students with his information.

To obtain a better idea of the dataset we have,We have to look at the goals (grades), some fundamental features (age, sex, etc.) and the level of alcohol consumption. We'll look at a few characteristics (which, in my biased opinion, can affect the target and alcohol consumption) and their impact on the target and alcohol consumption.

3.2 Expolatory Data Analysis

Data scientists use exploratory data analysis (EDA), which frequently makes use of data visualization techniques, to examine and analyze data sets and summarize their key properties. It makes it simpler for data scientists to find patterns, identify anomalies, test hypotheses, or verify assumptions by determining how to modify data sources to achieve the answers they need. EDA's major goal is to encourage data analysis before making any assumptions. It can assist in finding glaring errors, better understanding data patterns, spotting outliers or unusual occurrences, and discovering intriguing relationships between the variables. To make sure the findings they create are reliable and relevant to any desired business objectives and goals, data scientists can employ exploratory analysis. EDA aids stakeholders by assuring them that they are posing the proper questions. Standard deviations, categorical variables, and confidence intervals are all topics that EDA may aid with. EDA's features can then be used for more complex data analysis or modeling, including machine learning, when it is finished and conclusions have been formed from it.

3.3 XGBoost

The Gradient Boosting framework is used by the open-source software library XGBoost to develop optimal distributed gradient boosting machine learning algorithms. Extreme Gradient Boosting (XGBoost) is a distributed, scalable gradient-boosted decision tree (GBDT) machine learning framework. The top machine learning library for regression, classification, and ranking issues, it offers parallel tree boosting. Comprehensing the machine learning ideas and techniques that supervised machine learning, decision trees, ensemble learning, and gradient boosting are built upon is essential to understanding XGBoost. In supervised machine learning, a model is trained using algorithms to discover patterns in a dataset of features and labels, and the model is then used to predict the labels on the features of a new dataset. XGBoost is a scalable and extremely accurate gradient boosting solution that pushes the limits of computing power for boosted tree algorithms. It was created primarily to enhance the performance and computational speed of machine learning models. Trees are constructed using XGBoost in parallel as opposed to GBDT's sequential method. It employs a level-wise approach, scanning over gradient values and assessing the quality of splits at each potential split in the training set using these partial sums. In the past few years, XGBoost has significantly increased its popularity by assisting individuals and teams in winning almost all Kaggle structured data competitions. Companies and researchers submit data for these competitions, and then statisticians and data miners compete to create the best models for forecasting and characterizing the data.

3.4 Proposed Method

For the experimental part of this work we have done the following. Since we are working with a lot of parameters I have decided to procces and format this data using EDA which we have mentioned earlier, this data was in csv format and I decided by using EDA to create the dataframe format of this data which will be used to make predictions.

The majority of students are between the ages of 15 and 18. 22 is the oldest student age group. In this dataset, there are a bit more females than males.

Now I'll look at the Dalc functionality. It represents a typical workday's alcohol consumption (from 1 - very low to 5 - very high).

Table 1. Dalc representation

1	0.70
2	0.19
3	0.07
4	0.02
5	0.02

We might deduce from this that the majority of students (70%) do not consume alcohol during the week (1 in Dalc denotes very low consumption). It's odd that there isn't a 0 option for "no alcohol at all." Because it's impractical to expect every student to consume alcohol, I'll treat "very low consumption" as "no alcohol" in this study). During the week, 19% drink a little more than nothing, and 2,5% drink a lot.

Another important feature is Walc - weekend alcohol consumption.

Table 2. Walc representation

1	0.38
2	0.22
3	0.20
4	0.13
5	0.07

The situation has entirely changed at this point.

On weekends, it appears that students begin to drink. It's also fascinating to watch how many kids who don't consume alcohol during the week do so on weekends and it is more than 45%.

Table 3. Dalc-Walc representation

1	0.54
2	0.24
3	0.15
4	0.05
5	0.01

The majority of kids either do not have many absences or do not have any at all. However, there are a few pupils who had more than 30 absences.

Boys certainly consume more alcohol than girls, which is to be expected. In addition, both groups drink more on weekends than on weekdays. Alcohol use on weekends rises with age, peaks at 17 years old, drops a little, and then rises dramatically again from 20-22 years old. However, as we saw in the last section, we only have a few pupils over the age of 19, therefore this data does not apply to them.

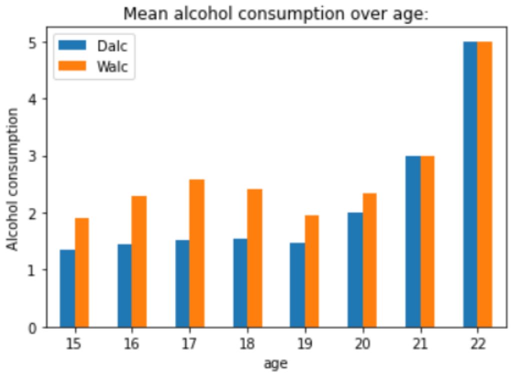

Figure 1. Alcohol consumption per age

Feature importances

Understanding which elements have an impact on the target and which don't is crucial, and it's helpful to gauge their influence in some way. According to our knowledge and experience, we approached the previous sentence stochastically by simply randomly taking into account a number of features that might have an impact on the target. However, it would be difficult to accurately take into account all 30 of our characteristics. There is, however, a deliberate and rational approach to go about doing it. Only the most crucial aspects will be carefully taken into account after we analyze the importance of each feature. Estimating feature importances can be done in a variety of ways, including through correlation analysis and machine learning algorithms.

Finding linear correlation between numerical variables is what the correlation analysis entails. Unfortunately, since we only have two numerical variables (age and absenteeism), applying the correlation analysis to our situation would be useless.

For the categorical characteristics' encoding there will only be one Hot Encoding method employed. We can now construct the random forest regression model after encoding the features and the target. Let's look at the top 20 traits (including dummy variables, there will actually be fewer) that affect the target the most. The next method is based on the RandomForestRegressor property. The process is straightforward: first fit the RandomForestRegressor, then call the attribute. In this section, we'll employ feature matrix and dependent variable vector. Because the RandomForestRegressor has so many hyperparameters to tweak, we must identify the best ones. GridSearchCV is going to be used.

Let's take a look at the 20 features, and highlight the most essential ones by combining dummy variables into single features.

The most important features based on the Random Forest:

- absences - number of school absences.
- schoolsup - extra educational support.
- failures - number of past class failures.
- age.
- Medu - mother's education.
- Fjob - father's job (whether father is teacher or not).
- sex.
- freetime - free time after school.
- studytime - weekly study time.
- famsup - family educational support.
- Fedu - father's education.
- Mjob - mother's job.
- reason - reason to choose this school.
- higher - wants to take higher education.

4. Results and Discussion

First step to get final result was to perform Modeling. To estimate the model performance in the future we have to create a test set. In our case, it will contain 20% of all data.

The goal value must be distributed similarly in the test and training sets. Change the random seed in train test split if this requirement is not met. Similar distributions can be found. We can infer that the split went off without a hitch.

The major objective of this section is to create a model that can forecast a student's grade. Three steps are involved:

- Setting up the model.

- Analyzing the test set performance of the model.
- Evaluating the model's suitability (comparison with a constant model).

We'll be utilizing the XGBRegressor from the xgboost package for training. The decision is based on the fact that, on table data, gradient boosting typically yields the best results. GridSearchCV will be used to fine-tune the model's hyperparameters.

For evaluating we will use RMSE (root mean squared error).

Table 4. RMSE results

Set	RMSE
Training set	3.2080592352679003
Test set	3.188232469843065

The test set's performance is little lower than the training set's, and it was not something what is expected. Although it is appropriate, the performance on the test set is a little bit worse than on the training set. We can infer that overfitting did not occur. If a model performs better than a constant model, whose output is always the same (for instance, the mean or median value), it can be said to be sufficient or meaningful. RMSE will be used for comparison. On the test, our model had an RMSE of 3.19. The result of RMSE is 3.129226832177046.

The constant model has a marginally superior performance. We must concede, however, that I was unable to develop a reasonable model for grade prediction.

After the whole work, it is very simple to discuss these results. The preprocessing, encoding, and whole process of EDA were very successful. The process of implementation of algorithms was successful too, but the results were a bit disappointing. Our constant model score is much superior, and it obviously means that our prediction model is not good. Related work that I analyzed had better results, and I probably missed something in my work. In the future, it is possible to try other different algorithms on this dataset and try to get better results.

5. Conclusion

This paper describes how EDA and XGBoost regressir may be used to set up and carry out data analysis on a selected dataset. On this dataset, EDA was demonstrated along with the implementation of XGBoost regressor to obtain accuracy scores of this models. Implementation of EDA process was fully successuful. For better results of prediction it is necessary to conduct more research to determine whether the XGBoost regresor can be enhanced to get better results or it is necessary to try some other algorithm. We got more superior performacne score on constatn model and that means that this is not good prediction model. Furthermore, future research may use other training, testing, and accuracy algorithms to get better picture of this topic.

References

[1] National Institute on Alcohol Abuse and Alcoholism, College drinking

[2] Group Alcohol Climate, Alcohol Consumption, and Student Performance, Jennifer Carson, Julian Barling, and Nick Turner

[3] College student employment and drinking: A daily study of work stressors, alcohol expectancies, and alcohol consumption. Butler, A. B., Dodge, K. D., & Faurote, E. J. (2010).

[4] Predicting Tobacco and Alcohol Consumption Based on Physical Activity Level and Demographic Characteristics in Romanian Students. Georgian Badicu, Seyed Hojjat Zamani Sani, Zahra Fathirezaie

[5] Rational decision perspectives on alcohol consumption by youth: Revising the theory of planned behavior. Tara L.Kuther

[6] Statistics Knowledge Portal, Exploratory Data Analysis

[7] https://www.nvidia.com/en-us/glossary/data-science/xgboost/