Final report for Capstone Project-1

Title- Analyzing the factors behind the alcohol consumption and relation with the

final grades in school

By- Ayush Mittal To- Srdjan Santic

Colorado State University Springboard mentor

**Problem-**

Students who consume a lot of alcohol are known to have bad grades. Also, the reason for alcohol consumption by students is unknown in many cases. Is it just an addiction? Or there are some other factors which control the alcohol consumption. To find the answers, I am going to analyze the data set “Alcohol consumption and Average Grades.” I am trying to see if there is a significant relationship between the alcohol consumption and average grades. If yes, what type of relationship is that and what other factors or features contribute to it and How the grades of the student are affected based on the demographic and social features.

**The Client-**

The Client can be schools, US Department of Education, Child Welfare Society or any government and public organizations which play a vital role in the education of teenagers. After this study, the client would be able to make changes in the current system or would be able to advise to parents about which factors play a crucial role in grades and alcohol consumption of their child.

**Data-**

The Data is available on Kaggle. Here is the link-

https://www.kaggle.com/uciml/student-alcohol-consumption

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school-related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por).

**Work Done/Approach-**

1. Data Wrangling and cleaning-

Steps were taken in Data Wrangling process-

1. Read the CSV files into the two separate data frames named as ‘math\_df’ and ‘port\_df’.
2. Make a join between these two data frames and see how many students are common, the common data frame is named as the ‘common\_df.’
3. Now a correlation analysis is done for the mathematics and Portuguese final scores for the 382 common students. It is found that the correlation coefficient is – 0.48. This shows a strong relationship between the scores.
4. In the next step a graph is plotted to visualize the relationship, The first scatter plot is confusing and does not the convey the required information to the viewer.
5. To make the visualization clear, the slice of the data frame is plotted as a line plot. It can be easily inferred now that students with high mathematics score have high Portuguese scores and vice versa. The probability of this trend is almost ~50%.
6. To catch the outliers, it is checked if any final grade is greater than the 20. The final grades are on the scale of 1 to 20, so no grade should be greater than 20. There were no outliers in the code.
7. Similarly, both the data frames are checked for any null values.
8. Most of the features in the data set are categorical. They need to be converted into one hot encoding. Most of the features like ‘traveltime’, ‘studytime’, ‘famreal’ and ‘famtime’ are label encoded. These features represent the quantities like how much a student travels to reach school? How much is the average study time? How much average time is spent with the family? All these features with higher value represents the higher amount of these features, which make sense. Machine learning algorithms should give more weightage to these features if they have higher values.
9. However, on the other hand, the features like- ‘Mjob, ‘Fjob’, ‘reason’ and ‘guradian’ have categories which are independent of each other. Hence, it makes sense to one hot encoding for these features. It increases the number of features in the data set, but this is a necessary step to avoid misinterpretation by Machine Learning Algorithms.
10. After one hot encoding, the Data is clean and clear for further processing.

Link to the IPython notebook-

<https://github.com/ayusmittal/SpringBoard-Projects/blob/master/Capstone-Project-1/Data_Wrangling_CAP1.ipynb>

1. Data Storytelling-

The objective of this part is to transform the analysis into a simple story. Which is easy to understand by the user. Below are the findings in the Data Storytelling-

1. The distribution of the Maths and portuguese score is normal.
2. It was observed that in both the classes the number of females are higher than the males.
3. Then the percentage of students are calculated who drink none alcohol to very high alcohol on the daily and weekly basis. All these analysis is shown in a pie chart so the user can understand things easily.
4. Calculations are done to see, how many students drink alcohol illegally. It is found that approximately 24% of student’s drink alcohol illegally in both the maths and the Potuguese class.
5. It is tried to see what could be the reason of underage drinking? Like what factors contribute most in illegal drinking of students Vs. the factors of students who do not drink alcohol illegally.
6. It can be observed that there is no significant difference in the average of Pstatus (Parent Status). Therefore, the parent status whether they are living together or separate does not contribute to underage drinking. However, if you look at the 'goout' feature, which represents the time spent with friends outside, it is high for students who drink illegally, i.e., 3.23 as compared to the students who do not drink, i.e., 2.7.
7. Also, Number of absence is greater for the student who drinks illegally, i.e., 5.51 as compared to nondrinking students, i.e., 3.4.
8. Last, underage students who do not drink alcohol has strong family relationship average of 4.07 as compared to the drinking students, i.e., 3.75
9. Initially it is observed that there is not much difference in grades of the students (Maths class) who drink alcohol and who do not. But detailed analysis showed that, for the Portuguese class.  Students who consume less alcohol has better grades as compared to the students who consumes more alcohol.
10. So after detailed analysis it is found that in Maths class, boys who drink less alcohol have good grades as compared to the boys drink high amount of alcohol on a weekly basis. However, opposite results are found for girls, girls who drink high amount of alcohol have better grades as compared to the girls who do not drink girls.
11. After diving deep into data, it is found that the number of samples (girls) who drink high alcohol are very few, only 6 to 7. Out of which three have exceptionally high scores greater than 15. This justifies that these girls might be brilliant or exceptional students who know when to party or when to study, or maybe they are outliers who gave wrong feedback on the drinking habit’s.

Link to the IPython notebook is-

<https://github.com/ayusmittal/SpringBoard-Projects/blob/master/Capstone-Project-1/Data_storytelling_CAP1.ipynb>

1. Data Statistics-
2. Pearson Correlation analysis is done for different features and the final grades. All shown in a colorful heat map. Findings are below-

For the Maths class-

It is observed that there is a lot of negative correlation between different variables. However, the first term and mid term grades have a high positive correlation with the final grades.   
a) Higher the mother and father education, higher the score of students.

b) Number of failures are negatively correlated with the student final grade, it means that

higher the number of failures, lower the score of a student.

c) Going out with friends frequently and large alcohol consumptions are negatively correlated with the student final grades.

For the Portuguese class-

It is observed that there is a lot of negative correlation between different variables. However, the first term and mid term grades have a high positive correlation with the final grades.   
a) Higher the mother and father education, higher the score of students.

b) Number of failures are negatively correlated with the student final grade, it means that higher the number of failures, lower the score of a student.

c) Going out with friends frequently and large alcohol consumptions are negatively correlated with the student final grades.

d) Higher study time results in better grades for the students in Portuguese.

e) Having a good family relationship also contribute in the overall studies of a student and leads to a better final score.

With the help of averages, other analysis is-

1. It is clearly observed students with an aim of higher education are serious and get better grades in final Maths test.
2. Having a romantic relationship can bring your final score down a little bit.
3. Taking extra classes in Maths help students to get good finals scores. however, there is no guarantee that extra classes will help a student to get good final grade in Portuguese.
4. Participation in extracurricular activities does not affect your final grades.
5. No significant difference in scores for students who attended nursery and who do not attend nursery school
6. No significant relation between parent’s jobs and the child grades.
7. It is clear from above analysis that extra educational support is given to weak students and still their grade is low as compared to students who do not get an extra educational support.
8. In this whole notebook, different relationships have been discussed concerning the final grades of a student.

Link to the IPython Notebook is –

https://github.com/ayusmittal/SpringBoard-Projects/blob/master/Capstone-Project-1/Statistics\_CAP1.ipynb

Machine Learning models-

In this section of the report, three different Machine Learning models have been implemented. The different models implemented are-

1. Linear Regression
2. Support Vector Regression
3. Ensemble Learning Model (Using Decision Tree regressor as a base model)

All the above mentioned models are trained on both the Maths and Portuguese class data sets and results are summarized in this report. Hyper parameters are changed to optimize the model.

Below is the summary of the models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear Regression (Maths) |  | Portuguese |  |
|  | ABS | MSE | ABS | MSE |
| With all Features | 3.39 | 4.19 | 2.15 | 2.86 |
| With all Features, but removing students with zero score in finals | 2.32 | 2.77 | 2.02 | 2.52 |
| Taking limited features into consideration | 2.4 | 2.7 | 2.09 | 2.55 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Support Vector Regression (Maths) |  |  | Portuguese |  |
|  | ABS | MSE |  | ABS | MSE |
| With all features RBF | 3.33 | 4.16 |  |  |  |
| With all features Linear | 3.39 | 4.23 |  |  |  |
| With all features Poly | 4.99 | 7.12 |  |  |  |
| With Optimum C and Gamma for RBF C = 16 and gamma = 0.1 | 3.19 | 3.85 |  |  |  |
| With Limited features and Optimum Values |  |  |  |  |  |
| RBF | 2.29 | 2.82 |  | 2.1 | 2.55 |
| Linear | 2.23 | 2.76 |  | 2.11 | 2.55 |
| Poly | 3.22 | 4.18 |  | 2.2 | 2.99 |

As observed from the previous results. The ML model always perform better if the limited features are fed as input. Additionally, limited features have a direct correlation with the final grade (either positive or negative relationship). So in this section only limited features are considered while training the data.

Based on the different simulations and the changing parameters (like depth and parameters) for the Ensemble learning technique, below is a brief summary for the best results.

Decision Tree regressor is used as a baseline model.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Ensemble Learning (Maths) |  |  |
|  | ABS |  |  |
| n =100, d =20 | 2.35 |  |  |
| n =200, d =50 | 2.28 |  |  |
| n =300, d =30 | 2.25 |  |  |
| n =400, d =40 | 2.28 |  |  |
| n =500, d =30 | 2.28 |  |  |
| n =600, d = 20 | 2.26 |  |  |
|  |  |  |  |
|  | Ensemble Learning (Portuguese) |  |  |
|  | ABS |  |  |
| n =100, d =5 | 2.165 |  |  |
| n =200, d =20 | 2.18 |  |  |
| n =300, d =50 | 2.16 |  |  |
| n =400, d =40 | 2.17 |  |  |
| n =500, d = 5 | 2.14 |  |  |
| n =600. d = 50 | 2.18 |  |  |

**Conclusion-**

1) Linear Regression has the best results.

2) Using Limited features like-

a) Mother and Father education

b) Number of Failures in past

c) Hanging out time with friends

d) Alcohol consumption

e) Taking study time and family relationship features

Improves the accuracy of the Linear Regression model.

3) The best results for a Support Vector Regression comes from using a RBF kernel with C = 16 and gamma = 0.1

4) Different decision tree regressor were simulated to find the optimum depth of a tree. 20 and 5 Comes out to be the best depth for the tree.

5) Plugging the best depth into an Ensemble learning technique, it is observed that different models give different performance when different depth are plugged in.

6) The best performance is at a depth of 30 with 300 estimators in the Ensemble learning model (maths) and for Portuguese (n = 500, and depth = 5).

Finally, it can be concluded that-

1) Mother and Father education contributes most to the final grades of a student.

2) Alcohol Consumption is directly related to the strength of a Family relationship. A student with a strong family relationship has a less probability of drinking less Alcohol.

3) Also, drinking more alcohol can bring down your grades. However, there are few exceptions where hive volume of drinking alcohol does not bring the grades down. They are students who believes in "Work Hard and Party hard".

4) Higher study time results in better grades for the students in Portuguese.

5) Number of failures are negatively correlated with the student final grade, it means that higher the number of failures, lower the score of a student.

6) In last, it is found in this study that, 24 % of student's drink Alcohol illegally, Government shall issue a mandate to the Liquor shops, asking them to strictly verify the Id's of customers who are coming in store to buy Alcohol.

Code can be found at –

<https://github.com/ayusmittal/SpringBoard-Projects/blob/master/Capstone-Project-1/Final_Machine_Learning_CAP1.ipynb>