CIS 3920

Term Project

Students Performance On Alcohol

What are we trying to accomplish for this project?

This project aims to find clearer data on alcohol consumption and determine which factors and features help determine who is likely to drink alcohol. This also looks at effects whether directly or indirectly on students' studies and the degree to which they are affected.

Project Description/Abstract

Alcohol consumption upon its quantity greatly affects the frontal cortex leading to poor decision making and concentration. Even though poor academic performance can be attributed to many other factors such as habits, sleep and nutrition, our hypothesis aims to know how bad alcohol affects students and whether or not a group of students are affected more than others. We'll use a dataset from a study done on students to further create linear regression models, random forests, apriori models and thus find any correlations that might be present.

Introduction

Good academic performance consists of various factors including hours dedicated to studying and future planning eventually leading to good grades. This excellence at school is due to many factors such as family, free time, previous school performance, interest, extracurricular activities, and friends. Each of these factors affect students in multiple different ways including how prone they are to drink alcohol. Factors such as family size, parent's level of education, parents' jobs also affect students' performance given that since there is weak or no guidance/ supervision on these students studying might be less likely. The data type being used is cross sectional data given that data was collected from students individually. The data used is on Portuguese students from the school Gabriel Pereira and Mouzinho da Silveira. Consumption of alcohol on these students was possible since Portugal's drinking age is 18. The following variables such as age, gender, family status, level of education will be compared to one another.

Dataset being used and description of the dataset

We will be utilizing a data set that has gathered information from 382 Portuguese students in secondary school. The study attempts to predict students' performance in school. This dataset includes information about the students gender, social life, their families, and academic habits. The data set tracks the grades of the students in two subjects, math and Portuguese. The distribution of students based on gender is 53% female and 47% male. The data was collected from two different schools, 88% of the students surveyed came from the same school.

Literature Survey

Past work on this dataset tries to predict what factors affect student achievement with a primary focus on alcohol consumption. A report from Cortez and Silva tries to see how alcohol has affected these students' grades through use of Data Mining. Cortez and Silva utilize classification and regression to understand this real world raw data. Many classification algorithms are used to see which combination of variables have a significant impact on school performance.

Project Methodology

Random Forest:

Random forest is used to determine which set of features would have most likely influenced a student to delve into drinking as a student. This method is used in RStudio. To determine how to divide the group into who is most likely, students were divided and grouped based on how much they drank during the weekend (Walc). Students who drank at least 3 times in the weekend were considered severe drinkers while students who drank less than three were considered soft drinkers. To see what features contribute, we trained the model (cross-validation) where 50% go through training and 50% go through testing. During this process, the algorithm will decide which 4 features best determine which student is most likely to drink alcohol. The use of a confusion matrix will be used to determine how accurate the classification of the 2 variables are when pairing with the chosen 4 variables, by determining how times something was misclassified and unclassified. An ROC curve will visualize accurately the classification for the random forest and AUC (area under curve) will give a numerical number of how accurate the random forest is.

Packages used for random forest:
tidyverse, caret, grid, GGally, gridextra, rpart, corrplot, randomforest, ROCR

Regression:

Given the number of different variables for this specific dataset it made sense to use multi linear regression. This is much more useful than simple linear regression because it allows

multiple explanatory variables and their relationship with the response variable. For this methodology we used R in Jupyter Notebook and the function linear regression equation which is represented as lm. First, a response variable will be tested with different predictor variables and based on their p-value and r-square we will know whether or not they will be useful. In addition to this in order to find out whether or not the model follows a normality distribution, a histogram will be used to check whether the model's distribution is normal , right-skewed or left-skewed. The reason for this is to check if there is any linearity within the variables. Having done so, we will make we will change the response variables and predictor variables until we find the desired

P-value and R -squared. In the data we will use up to four predictor variables but no less than two given that the point is to see the response variables in relation to multiple variables. Furthermore, to gain a more complete understanding of our model we will use summary(model_name) to find out things such as the estimate, standard error, p-value and R^2. We will also use code to find out the relationship between the residual and the fitted data to find any linearity and ultimately we will graph in hopes that such visualization might become helpful for the interpretation of the data.

Association Rules:

Association Rules are used to analyze data and find hidden relationships within datasets. It highlights items that frequently appear in the data together in the form of an *if-then* rule. The implication of the association rule algorithm is that it provides item sets (X) that were found in the dataset and gives you an item (Y) that is likely to be included in that item set as well. We then use metrics such as support, confidence, and lift to help determine if relationships are relevant or not. Support is an indication of how frequently a set of items appear together. Confidence is an indication of how often the support-rule has been found to be true.
Lift is a measure of association using both support and confidence.

We then used the apriori algorithm to reduce the amount of item sets to analyze. This is done by eliminating item sets that are not frequent, therefore eliminating all of its subsets. For this analysis we adjusted the minimum support and minimum confidence levels of the algorithm to help find relationships in the data.

Packages used for apriori algorithm:
arules

Analysis & Results-Anticipated Task and Technique

Multi-linear Regression Analysis- For this algorithm we tried to test for multiple variables that would give us an insight on whether or not there is a relationship between alcohol consumption and performance at school. We used the following formula in R: **lm ([target variable] ~ [predictor variables], data = [data source])**. At first we wanted to know if grades(G1,G2,G3)

changed when tested on predictor variables such as daytime alcohol consumption(Dalc) , weekly alcohol consumption(Walc) and freetime. After the formula was executed in R we printed a summary of the results with **summary(model)** . In this case we looked for the p-value and multiple R-squared to find out any correlations, however we found that the p-value in this example was 0.6649 and therefore there was not a statistically significant relationship with the response variables in place and therefore invalid. Furthermore, the multiple R-squared was very small : 0.004017 indicating that the fit of the regression is not perfect and not suitable. The T-value was also small meaning that there is a high probability that it happened by chance -0.611, -0.438, 0.468 respectively. In addition, our Estimate also known as regression coefficients were- 0.2105, -0.1032 and 0.1108 for our Dalc, Walc, and freetime predictor variables respectively. These indicate that the first two variables have opposite associations indicating that the better grades students get it is more likely to some small degree due to a decrease in alcohol consumption during the weekdays and weekends. However, since our p values are larger for each predictive variable it probably means there is a different factor creating these changes.

Primary Model

```
[8]: primary_model=lm(G3~Dalc+Walc+freetime, data=dataset)
     summary(primary_model)

     Call:
     lm(formula = G3 ~ Dalc + Walc + freetime, data = dataset)

     Residuals:
          Min       1Q   Median       3Q      Max
     -10.8451  -1.9412   0.3764   3.3651   9.4872

     Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
     (Intercept)  10.6050     0.8422  12.592   <2e-16 ***
     Dalc         -0.2105     0.3446  -0.611    0.542
     Walc         -0.1032     0.2356  -0.438    0.662
     freetime      0.1108     0.2368   0.468    0.640
     ---
     Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Residual standard error: 4.59 on 391 degrees of freedom
     Multiple R-squared:  0.004017,  Adjusted R-squared:  -0.003625
     F-statistic: 0.5256 on 3 and 391 DF,  p-value: 0.6649
```

During our analysis with linear regression we tested different response variables such as G1,G2,G3, studytime, indicating grades of students. Ideally we expected G3 and study time to have some negative correlation with alcohol consumption as in the higher the grade the lower the alcohol consumption, however, p-values and multiple R-squared were not sufficient to infer any possible calculation besides the one mentioned above. On the other hand, when we tested our multiple linear regression model with weekly alcohol consumption (Walc) as our response variable and failures(failures), daytime drinking(Dalc), absences (absences), and going out(goout) as our predictor variables our Multiple R- Squared was an astonishing : 0.4897. In other words 48.97% of the variation in weekly alcohol consumption can be explained by the failures, daytime drinking, and absences. To further support this idea the overall p-value for this model was: < 2.2e-16. However, when we looked at the data further we found out most of the multiple R squared value was due to daytime alcohol consumption and going out (p-value: <

2.2e-16, t-value: 14.942),(p-value: 1.33e-11, t-value: 6.973) respectively. This suggests that for students to engage in alcohol they must have had some previous experience with the substance and are able to go out in order to do so.

Best Model So Far

In [176]:
```
# 48.97 % of our predictive variables explains
# all the variation in the response variable around its mean.
best_model= lm(Walc~failures+Dalc+absences+goout,data=dataset)
summary(best_model)
```

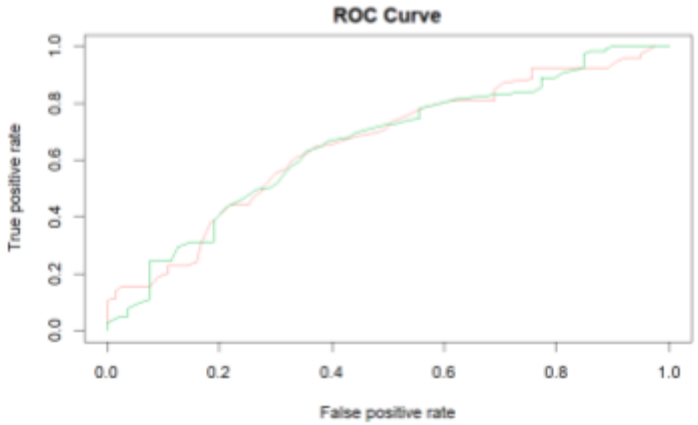
```
Call:
lm(formula = Walc ~ failures + Dalc + absences + goout, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0431 -0.7928 -0.1645  0.5607  3.1689

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.060893   0.146922   0.414    0.679
failures    0.049033   0.063580   0.771    0.441
Dalc        0.819602   0.054854  14.942  < 2e-16 ***
absences    0.009560   0.005865   1.630    0.104
goout       0.304113   0.043612   6.973 1.33e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9247 on 390 degrees of freedom
Multiple R-squared:  0.4897,    Adjusted R-squared:  0.4845
F-statistic: 93.56 on 4 and 390 DF,  p-value: < 2.2e-16
```

Random Forest: The random forest has shown that these features: goot (going outside with friends), absences, G3 (final grade), and age help determine which students have been drinking alcohol during the weekends. However, the AUC (0.657) has determined these features are not as accurate to predict which student is most likely to drink since the AUC is not close to one. This leads onto the idea that not all four features can determine who is most likely to drink. This means that either none of the features can accurately predict who is most likely to drink on the weekend or that one or few do not contribute to determining which students are more likely to drink. This is further supported by our validation model showing that the features used that rate importance of these features are goot as 100, absences as 75.01, G3 as 69.81 and age as 43.28.



ROC curve of classifying severe and heavy drinkers.

```
Confusion matrix:
       Severe Soft class.error
Severe     37   43   0.5375000
Soft       21   96   0.1794872
```

Confusion Matrix of severe and soft drinkers

	Overall <dbl>
goout	100.00000
absences	75.01565
G3	69.81343
age	43.28277
freetime	39.66630
studytime	39.35059
health	38.06215
Medu	36.25902
famrel	36.02768
Fedu	32.60410

Model showing importance of variables when classifying severe and soft drinkers

Association Rules Analysis:

The Association Rules and Apriori algorithm helped us find a few rules and item sets that we deemed to be interesting. We sorted our finds by lift.

```
     lhs                    rhs            support confidence  coverage     lift count
[1]  {sex=M,
      schoolsup=no,
      goout=[4,5]}       => {Walc=[3,5]} 0.1367089  0.8181818 0.1670886 2.032590    54
[2]  {sex=M,
      schoolsup=no,
      goout=[4,5],
      Dalc=[1,5]}        => {Walc=[3,5]} 0.1367089  0.8181818 0.1670886 2.032590    54
```

The first rule we want to highlight indicates the high association of being a male, with no school support, and goes out frequently with high weekend alcohol consumption. The second rule is interesting because even when you add weekday alcohol to the lhs items the support, confidence, and lift stay the same, indicating a high association.

```
[6]  {age=[17,22],
      goout=[4,5],
      Dalc=[1,5]}        => {G3=[0,10)} 0.1139241  0.5357143 0.2126582 1.627747    45
```

The next rule we want to highlight tells us that typically older students (ages 17-22) who go out frequently during the week and drink alcohol during the week have a high association with doing poorly on their final grades.

```
[47] {failures=[0,3],
      schoolsup=no,
      goout=[4,5]}       => {Walc=[3,5]} 0.2025316  0.6666667 0.3037975 1.656184    80
```

This rule highlights a few factors that are associated with high weekend alcohol consumption. We discovered that if a student has failed 0-3 classes, has no school support, and goes out frequently they are likely to consume high levels of alcohol on the weekends.

Conclusion

This project tries to show the severity of alcohol consumption on students and what factors may have contributed to determining which student drinks. This project uses three methods: multi-linear regression, random forest, association analysis to help determine what features may have students drink. It is commonly found in this project that students who do frequently go outside with their friends do drink alcohol. This tells that students who tend to be socially active may also drink as well. For future studies, we would hope to see other factors like if a student's circumstance, like impoverished families, have an effect on students to drink alcohol.

References

https://www.kirenz.com/post/2020-05-14-r-association-rule-mining/

https://www.kaggle.com/datasets/uciml/student-alcohol-consumption?resource=download

Multiple Linear Regression | A Quick Guide (Examples) (scribbr.com)

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.

Appendix