STATS 415: Data Mining - Dept. of Statistics, University of Michigan - Ann Arbor

Alcohol Consumption Classification for College Students

Tahmeed Tureen, Aidan Kearns & Anthony Zheng

Abstract

In recent times, students have begun practicing binge drinking as a social norm in college life. Some, more than others, consume a great deal of alcohol throughout their college career. In this data mining report, we attempt to construct a prediction model for heavy college drinkers from a given set of data regarding college students from two distinct universities in the country of Portugal. We wanted to construct a predictive model that we can use to classify a student as a heavy drinker based on their background and life choices. To assess our data set we explored statistical methods such as lasso, ridge regression, logistic regression and random forest and chose the optimal model to create our prediction model.

Data Set

Our data set is courtesy of the UCI Machine Learning website, which contains 32 variables regarding alcohol consumption and student background information from two distinct universities in Portugal. The number of observations range to a total of 1044 students and each attribute represents specific contextual information such as family size, class absences, and alcohol consumption. Fortunately, there is no missing data and the majority of the data is categorical with differing factor levels for different variables. Specifically, the data set consists of binary, numeric and nominal variables.

Initial Variables

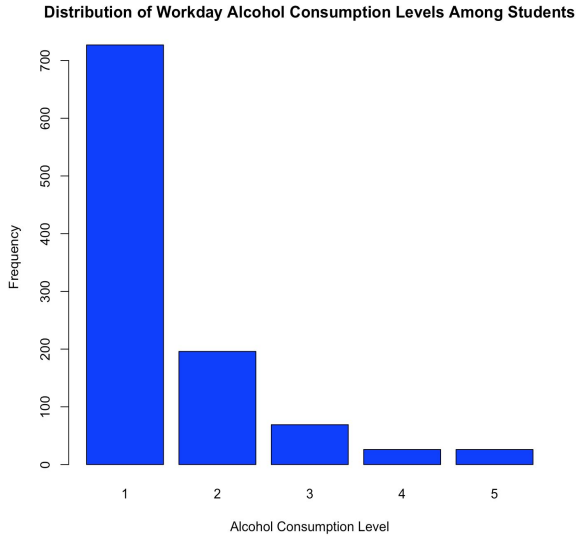
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

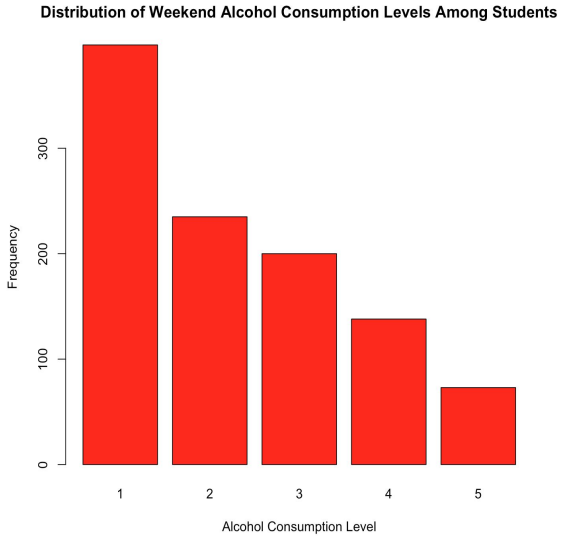
Variable Modification

Before performing our variable modifications, we observed the summary statistics and initial distributions of the Dalc (workday alcohol consumption) and Walc (weekend alcohol consumption) variables from our dataset. We decided to look at these variables because they are the only direct measures of alcohol consumption in our initial data set.

Table 1: Summary Statistics for Variables Walc & Dalc

Variable	Min	1st Quar.	Median	Mean	3rd Quar.	Max
Dalc	1.000	1.000	1.000	1.494	2.000	5.000
Walc	1.000	1.000	2.000	2.284	3.000	5.000

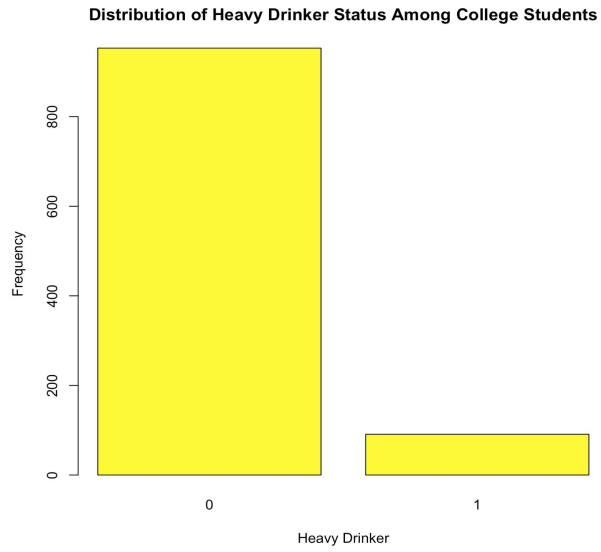




According to the bar graphs, we can see that the distribution of workday alcohol consumption is more heavily right-skewed than the distribution for weekend alcohol consumption. This shows that alcohol consumption levels are generally lower during the week than on the weekend. This is also supported by the summary statistics table, which show the mean alcohol consumption levels of Dalc and Walc as 1.494 and 2.284 respectively.

Our response variable of interest is whether or not a student is a heavy drinker. Obviously, the term “heavy drinker” is very ambiguous and is perceived differently by many people. To eliminate this ambiguity, we observed the summary statistics table for the variables Dalc

(workday alcohol consumption) and Walc (weekend alcohol consumption). Based on the table, we noted that the 3rd Quartile for both Dalc and Walc are 2.000 and 3.000 respectively. We chose to create a new variable "heavydrinker" that is assigned a value of 1 if a student has a Dalc rating of at least 2.000 **and** a Walc rating of at least 3.000; a value of 0 is assigned otherwise for the "non heavy drinkers". We defined it in this manner to assign a student with the label "heavy drinker" based off of the statistics of the students in the data set. This helps eliminate the ambiguity of the term in the context of this study. Utilizing our definition of "heavy drinker", we found that 91 of the 1044 students were heavy drinkers in our set, which can be seen in the bar graph below. This newly created variable is our response variable of interest.



Before proceeding to data analysis, we opted to remove the initial variables Walc and Dalc from our data set because they were both used to construct our response variable. We do this because intuitively we note that Walc and Dalc will have high predictive power for "heavydrinker" because "heavydrinker" was created from those two variables. This means that keeping Walc and Dalc in our analysis will strip predictive power from other variables, which may cause annoyances as well as collinearity issues in our predictive model.

Data Exploration

After creating the new variable "heavydrinker", we separated the data into two different sets for further data exploration. The first set includes only the students with the "heavydrinker" positive response, while the other set only includes non "heavydrinker" students. We performed this split to compare the summary statistics for the two types of students and their associated variables. The first noticeable difference is the male to female ratio. The overall data set is 26.1% male (272 of 1044). The set of non heavy drinkers is very similar in terms of percentages, 25.6% male (244 of 953). However, the "heavydrinker" set is dominated by males at 85.7% (78 of 91). This is an interesting statistic to note considering

that the majority of the full data set is female at 73.94% (772 of 1044). Another noticeable difference we observed was in the “reason” variable (reason why the student attends their given university). These results are summarized in the following table.

Table 2: Distribution for Reason Predictor

	Full Set	**Heavydrinker Set**	**Non-Heavydrinker Set**
Course	430	30	400
Home	258	25	233
Other	108	20	88
Reputation	248	16	232

Again, the full set and “non-heavydrinker” set have similar distributions, but the “heavydrinker” set has a higher rate of “other” (22.0% compared to 10.3% and 9.2% respectively) and a lower rate of course preference (33.0% compared to 41.2% and 42.0%). There is also a noticeable difference in the “study time” variable. The full dataset has a median of 2.0 and mean of 1.97, the “non-heavydrinker” set has a median of 2.0 and mean of 2.002, while the “heavydrinker” set has a median of 1.00 and mean of 1.673 for “study time”. We also note that there is a gap in the nursery variable (whether or not the student attended nursery school). 80.0% of the total students attended nursery school, 80.8% of “non-heavydrinker” also attended, however, only 60.0% of “heavydrinkers” attended. “Heavydrinker” students tend to have more freetime with a median of 4.0 and mean of 3.626, “non-heavydrinker” students have a median of 3.0 and mean of 3.161 and the full data set has a median of 3.0 and mean of 3.201. “Goout”, the variable that measures the amount a student goes out with friends, is similar to freetime in percentages. “Heavydrinker” set has a median of 4.0 and mean of 3.956, “non-heavydrinker” has median 3.0 and mean of 3.08, the full dataset has a median of 3.0 and mean of 3.156. The “heavydrinker” set has a generally higher number of absences than the other sets, with median 6.0 and mean 6.484. Meanwhile, “non-heavydrinker” has median 2.0 and mean 4.239; it should be noted that there is an outlier value of 75 in this set. The full set has a median of 2.0 and mean of 4.435 for number of absences. Finally, “heavydrinker” has lower

grades than “non-heavydrinker” in all three academic periods: (median 10 and mean 10.2 for G1, median 10 mean 10.51 for G2, median 11 mean 10.63 for G3) for “heavydrinkers” versus (median 11 mean 11.31 for G1, median 11 mean 11.32 for G2, and median 12 mean 11.41 for G3) for “non-heavydrinkers.”

* - G1: First Period Grade, G2: Second Period Grade G3: Third Period Grade

Classification

The goal of our project is to find the most accurate model for predicting whether a student classifies as a “heavydrinker”. To choose the optimal model for our prediction we observed the classification errors for the following statistical methods: ridge regression, lasso regression, logistic regression and random forest. We are interested in the method with the lowest classification error. Furthermore, many of the predictor variables in the data set are categorical variables (26 of 31) with differing number of factor levels. Thus, we used the model.matrix() function in the statistical software R to split up the categorical variables and create dummy variables; the new matrix has 40 variables. We then randomly split this data into a training and test set of sizes 730 and 314 respectively for our model analysis.

For lasso and ridge regression, we fit models using cross validation to determine the best tuning parameter (lambda). We repeated this for both the original training data and for the data with two way interactions on every predictor combination.

Table 3: Best Lambdas & Their Classification Error Rates

	Lasso	**Lasso With Interaction**	**Ridge**	**Ridge With Interaction**
Best Lambda	1.439624e-02	0.094718751	0.094717627	1.993739
Classification Error	8.917197%	10.82803%	9.55414%	9.55414%

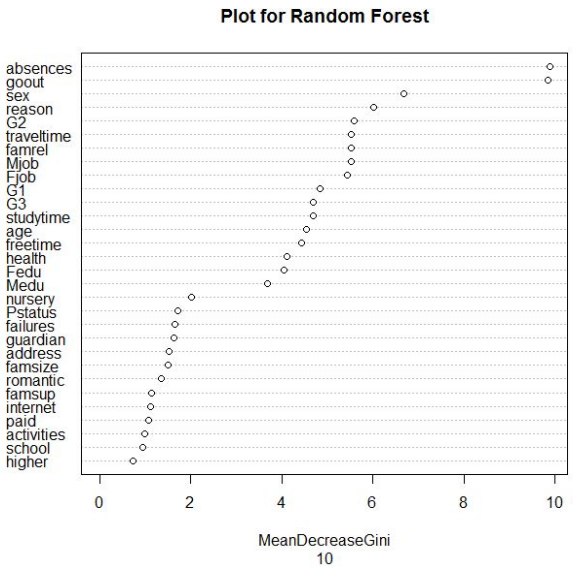
Next, we fit a logistic regression model, whose classification error rate was found to be 8.045977%. This is lower than that of lasso and ridge. On the next page, there is a table comparison of the logistic classification of “heavydrinker” with the true values (value of 1 representing heavydrinker):

Table 4: Classification Table for Logistic Reg. Model

Logistic Class.	True Value	Frequency
0	0	935
1	0	18
0	1	66
1	1	25

Finally, we applied random forest, which had the lowest classification error rate amongst all of our models. We created several models using different m values (number of variables considered at each split) of 10, 15, 20, and 25. The least accurate of the four random forest models, m = 25, had a classification error rate of 7.6%, which is already lower than that of the other supervised learning methods. The optimal random forest model was for m = 10, which had the lower error rate of 6.687898%. Below are the random forest output, classification table, and graph of Mean Decrease Gini of the explanatory variables for m = 10 variables at each split:

```
  predict_tree test_x Freq
1            0      0  284
2            1      0    1
3            0      1   20
4            1      1    9
[1] 0.06687898
                    0          1 MeanDecreaseAccuracy MeanDecreaseGini
school       2.659517  1.4581564             2.915625        0.9447083
sex         16.239464 24.2092324            23.652427        6.6715184
age          5.458913  6.4575170             7.500575        4.5291588
address      4.874405  3.7052873             6.051420        1.5245934
famsize      2.598809  5.9439382             4.553959        1.4902755
Pstatus      5.773860  4.5951144             6.771770        1.7180675
Medu         7.826163  3.4153617             8.116787        3.6839392
Fedu         7.003787  7.3102509             9.535690        4.0459591
Mjob         8.770014  7.4005627            10.630929        5.5153270
Fjob         9.441058 11.4987912            13.153927        5.4311524
reason       9.934039 11.4810395            14.485362        6.0116150
guardian     4.137857  3.2837868             5.162736        1.6246358
traveltime   5.650332 11.6922959            10.045085        5.5214082
studytime    9.786397 11.4504211            13.762196        4.6774350
failures     6.470697 -2.0222467             5.210046        1.6465655
schoolsup    3.257672  3.7221606             4.425073        0.6617947
famsup       2.988832  4.4775804             4.680564        1.1266094
paid        -2.430944  1.0414825            -1.665281        1.0746245
activities   2.954874  3.0021286             3.616652        0.9759215
nursery      6.693464 10.1835020            10.280862        2.0236714
higher       2.251348  0.6591480             2.368871        0.7234429
internet     6.095413  3.0199486             6.967766        1.1220733
romantic     5.508442  3.2948929             6.296075        1.3397800
famrel       6.050015 10.0868282             9.698433        5.5172851
freetime     6.375647  7.0851469             8.479794        4.4391223
goout       11.729143 18.8855514            18.740942        9.8544190
health       6.606876  5.1555122             8.329199        4.1105599
absences     7.062871  5.5818480             8.920628        9.8803566
G1           7.988178  2.4460923             8.568541        4.8481613
G2           8.577000  0.8366994             8.765114        5.5905390
G3           9.338514 -3.3622993             9.019723        4.6941568
```



10

As we observe from the graph, the variables with a high Mean Decrease Gini, or the variables that were most important in partitioning the data are split into three groups (visually trivial). The highest Mean Decrease Gini is found in “absences” and “goout” which have values of 9.88 and 9.85 respectively. The following important variables on the graph were “sex” and “reason” at 6.67 and 6.011 respectively. The random forest model did extremely well in classifying students as “non-heavydrinker” (284 out of 285 correctly classifed) but not as well at classifying “heavydrinker” (9 out of 29 correctly classified). Overall, the classification error rate of 6.67% was the lowest of any model. In our data exploration, we noted that “goout” and “absence” were both higher for the “heavydrinker” split data. “Heavydrinker” was also heavily male (85%) and the distribution of “Heavydrinker” students’ reasons for attending their given school was different than the distributions of the other datasets. We noticed in particular that the distribution of “sex” had a large discrepancy between students classified as “heavydrinker” and those classified as “non-heavy drinker”. This was reflected in the Random Forest model since “sex” had a very high Mean Decrease Accuracy (23.652427), 5 points higher than the next highest variable in the model. We see that the four variables that we mentioned were important are variables we initially thought to be important in predicting “heavydrinker.”

Conclusion

When attempting to classify students as “heavydrinker”, we conclude that random forest considering 10 variables at each split is the best model in terms of overall classification accuracy. This model’s top variables in terms of Mean Decrease Gini were “absence”, “goout”, “sex”, and “reason” which we thoroughly mentioned in the Classification section. Overall, no model that we fit performed well when attempting to classify students as “heavydrinker.” Below is a table of the classification error rate of predicting a student as a “heavydrinker” for the optimal model in each supervised learning method:

Table 5: Classification Errors for Predicting Heavy Drinker as Positive

	Lasso	**Lasso with Interaction**	**Ridge**	**Ridge with Interaction**	**Logistic Reg.**	**Random Forest**
Class. Error	96.55%	100%	100%	100%	72.53%	68.97%

As we can see, ridge and lasso performed extremely poorly when predicting a student as a "heavydrinker". If a model were to predict that **every** student is "non-heavydrinker," it would have a misclassification error rate of 9.24% which appears to be accurate, but in reality it is not. This is because of the small number of "heavydrinker" instances in the test set (29 of 314), so even if the model misclassified every "heavydrinker" as "non-heavydrinker" the error rate will appear low . In classifying a positive "heavydrinker", random forest once again performs the best. Although no model has an accuracy rate of over 50% when attempting to classify students as positive for "heavydrinker," random forest is still clearly the optimal model for predicting students classified as "heavydrinker" based on error rate. However, none of our models have strong predictive power when classifying a student as a "heavydrinker".

Limitations

There were several limitations in the dataset that we used. First of all, there were likely many collinear variables. Since the majority of the variables were all categorical, it was not possible to measure the collinearity and note its specific effects. This collinearity issue **may** have caused ridge and lasso to both perform poorly. In addition, in terms of real-world application of our findings, there is a contextual limitation because our data set is from universities in Portugal. The students that were surveyed were Portuguese and likely have different customs in terms of college drinking. The data came from two different universities in Portugal that were not evenly distributed (772 from one school 272 from the other). Dalc and Walc, the variables we used to create our response variable "heavydrinker," were scores of 1 to 5 found from a survey. We have no prior knowledge of how these scores were calculated, and again, there could be a difference in what an American and Portuguese person considers a level 5 weekday or weekend drinker.

Resources Used

1) James Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. N.p.: Springer Texts in Statistics, n.d. Print
2) Fabio Pagnotta, Hossain Mohammad Amran. *Using Data Mining To Predict Secondary School Student Alcohol Consumption.* Department of Computer Science, University of Camerino. (The Data Set)