Examining the Student Alcohol Consumption

Project Report

Mine Your Business

STA3920 Data Mining | Fall 2018

Group Members

Haonan Ou
Kaijun He
Min Zheng
Sandy Ramos
Soohun Han

1. Dataset used and the problem addressed in this analysis

The dataset we chose is about the student alcohol consumption. The dataset was shared by UCI Machine Learning on Kaggle.com and it is from a survey conducted in Europe, about student background and alcohol consumption.

The main point we want to learn from this data is:

How are the students' personal background (family size, age, gender, etc.), if any, correlated to their alcohol consumption

2. Exploratory Data Analysis

The dimension of the dataset was 395 records and 33 variables pertaining to different personal backgrounds such as age, gender, family size, and so on about each student as well as the rate of alcohol consumption.

Upon examining the structure of the data using structure function, we found out that 16 of the 33 variables were integer variables.

Using the summary function, we gained the five number summary for continuous variables; counts of binary responses; counts for the Likert scale; and nominal variables.

- As for the continuous variable summary, there was an outlier for the 'absences' variable.
- For binary responses, the NO was nearly 8 times more than YES for weekday alcohol consumption whereas NO response was only 1.5 times more than YES for the weekend alcohol consumption
- Some responses in the form of a rating scale

We noticed that the response variables were recorded as integers and because of the categorical nature of our response variable, we converted the continuous variables to categorical variables using as.factor() function.

Also, our response variable, the alcohol consumption rate, was separated into two – weekday alcohol consumption and weekend alcohol consumption with a rating scale of 1 to 5.

- So we had to convert our response variable into a binomial variable.
- We set a consumption rate less than or equal to 2 as "NO" and greater than 2 as "YES" for alcohol consumption.
- After that, we converted these two variables as categorical ones like before.

For simpler analysis, we converted these responses to a binary type, where we consider the consumption rate less than or equal to 2 as 'NO' and greater than to as 'YES' for alcohol consumption. At the same time, we made two variable contrasts or dummy variables where 0 takes 'NO' and 1 represents 'YES'.

3. Reducing Variables

Our primary assumption of possible predictors that could have some correlation with a higher alcohol consumption rate was as follows:

Sex: the student's gender
Pstatus: the parent's cohabitation status – whether the parents are living together or apart
Romantic: the student's relationship status
Absences: the number of school absences
Failures: the number of past class failures
Famrel: the quality of a family relationship

However, to get a more meaningful set of predictors, we decided to use the stepwise regression.

We first created two variables

- One, a null variable containing only the intercept.
- And a full variable containing all the variables from the dataset and regress them in a glm function with the family option set to "binomial"

Then we set up a variable for the step function where we set the vector, the scope, and the direction to "both".

- We repeated this process for the 2nd response variable which is 'weekend alcohol consumption' or Walc.

Since we have two response variables, we repeated this process twice. For the first response variable, weekday alcohol consumption, the starting AIC was 278.04. By the end of the regression, the AIC was reduced down to 226.76 with the following predictors:

- sex; goout; school; absences; traveltime; activities; higher; reason; famsize; nursery

For the second response variable, weekend alcohol consumption, the starting AIC was 534.48 and the final AIC was reduced down to 437.07 with the following variables:

- goout; sex; fjob; absences; famrel; paid; traveltime; address; famsize; nursery; activities

In result, only 50% of the predictors we assumed in the beginning was also included in the best subset selection - and they were, sex, absences, a family relationship; but parent's cohabitation status, romantic, and failure was not included in the result.

We had the largest odds ratio for sexM predictor where, holding all other predictors fixed, the one-unit increase in SexM, increased the odds of weekend alcohol consumption by a factor of 8. In another words, we expect to see about 8 times more in the odds of weekend alcohol consumption for male students.

4. Developing models; comparing model performances; selecting models

Since our response variables are categorical in nature, we selected logistic regression and decision tree models as our data mining methods. Since our dataset is small, we needed to employ a resampling method for our dataset.

We initially used only the bootstrap as our resampling method but we noticed that our dataset was highly unbalanced. We use the ROSE package to resolve this issue. By applying the ROSE package, we can balance the data. After the dataset is balanced, we apply it to the bootstrap to create balanced sample datasets.

After obtaining balanced sample datasets, we use these datasets to fit our model.

Logistic Regression
For the logistic regression, we used 4 different resampled datasets to train our model.

We use the glm() function containing the predictors obtained from stepwise regression. We set data equal to a balanced dataset; set subset equal to a training dataset created from the balanced dataset; and family equal to binomial.

We use subset option in glm() to fit a regression using only the observations corresponding to the subset, and set the family equal to binomial since we have a binary outcome.

After that, we use predict() function to evaluate the performance of the model using the test dataset. It calculates the predicted probabilities of alcohol consumption, and we set the type equal to "response" which will give us the predicted probability.

Classification Decision Tree
For the classification tree analysis, we use balanced dataset created from using ROSE and bootstrap as before.

We use the tree() function where we regress our response variable on predictors we obtained from stepwise regression. We set the subset equal to the balanced sample dataset.

Then we performed cross validation with K=10 to obtain the optimal tree size.

5. Model evaluation. Use figures/tables to show model performance

We used bootstrap to check the accuracy for the logistic models and the decision tree due to the limited sample size. The high imbalance in the response variable called for the use of a balanced training dataset to fit the model. We have produced several confusion matrices with balanced training datasets below.

Before proceeding, here are the test results on the original dataset:

Weekday (Dalc)

glm.pred5	test.truevalue	
	NO	YES
NO	347	32
YES	4	12

accuracy = 0.9088608

Weekend (Walc)

glm.pred5	test.truevalue	
	NO	YES
NO	192	54
YES	44	105

accuracy = 0.7518987

Accuracy Using Balanced Training Dataset on Weekday Alcohol Consumption

glm.pred6	test.truevalue2	
	NO	YES
NO	271	83
YES	80	268

accuracy = 0.7678063

glm.pred7	test.truevalue2	
	NO	YES
NO	258	72
YES	93	279

accuracy = 0.7649573

glm.pred8	test.truevalue2	
	NO	YES
NO	270	79
YES	81	272

accuracy = 0.7720798

glm.pred9	test.truevalue2	
	NO	YES
NO	276	104
YES	75	247

accuracy = 0.7720798

For weekday alcohol consumption, the mean accuracy was 0.7692308

Accuracy Using Balanced Training Dataset on Weekend Alcohol Consumption

glm.pred16	test.truevalue4	
	NO	YES
NO	114	39
YES	45	120

accuracy = 0.7358491

glm.pred17	test.truevalue4	
	NO	YES
NO	117	38
YES	42	121

accuracy = 0.7484277

glm.pred18	test.truevalue4	
	NO	YES
NO	110	42
YES	49	117

accuracy = 0.7138365

glm.pred19	test.truevalue4	
	NO	YES
NO	104	31
YES	55	128

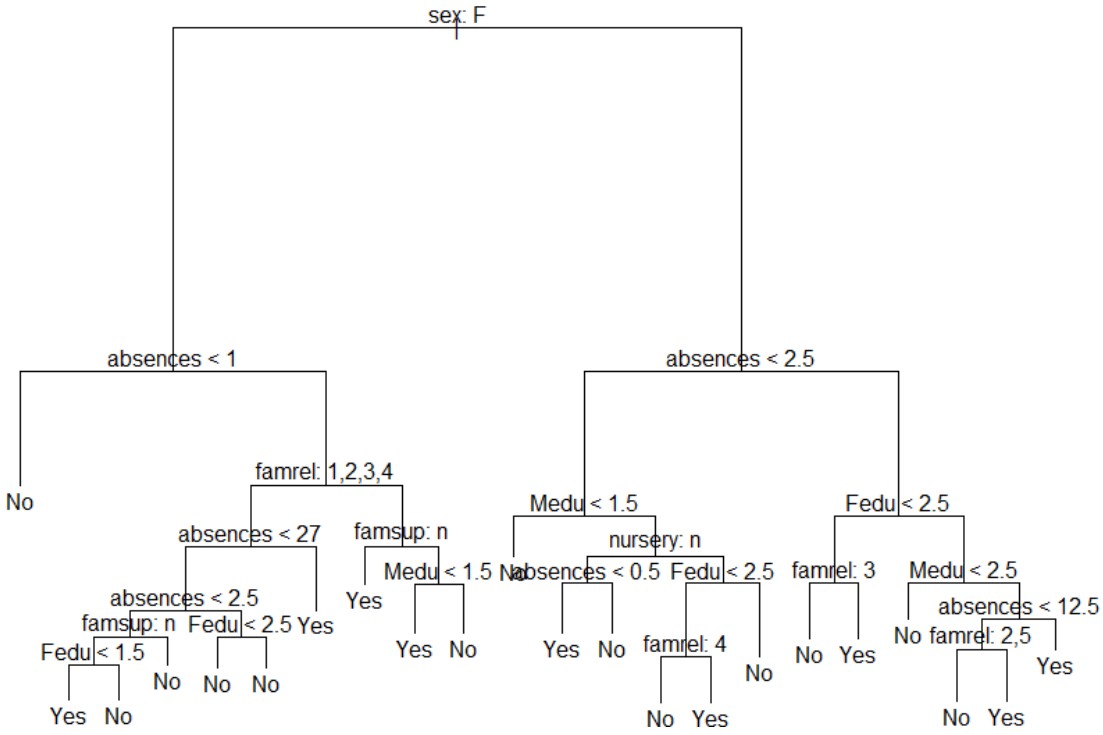
accuracy = 0.7295597

For weekday alcohol consumption, the mean accuracy was 0.7319182

Next, we move on to decision tree model. As before, we balance our dataset using the ROSE package and use bootstrap to produce resampled datasets. The results are as follows.

Decision Tree Results for Weekday Alcohol Consumption

Pruned Tree Model



Pruned Tree Model Results

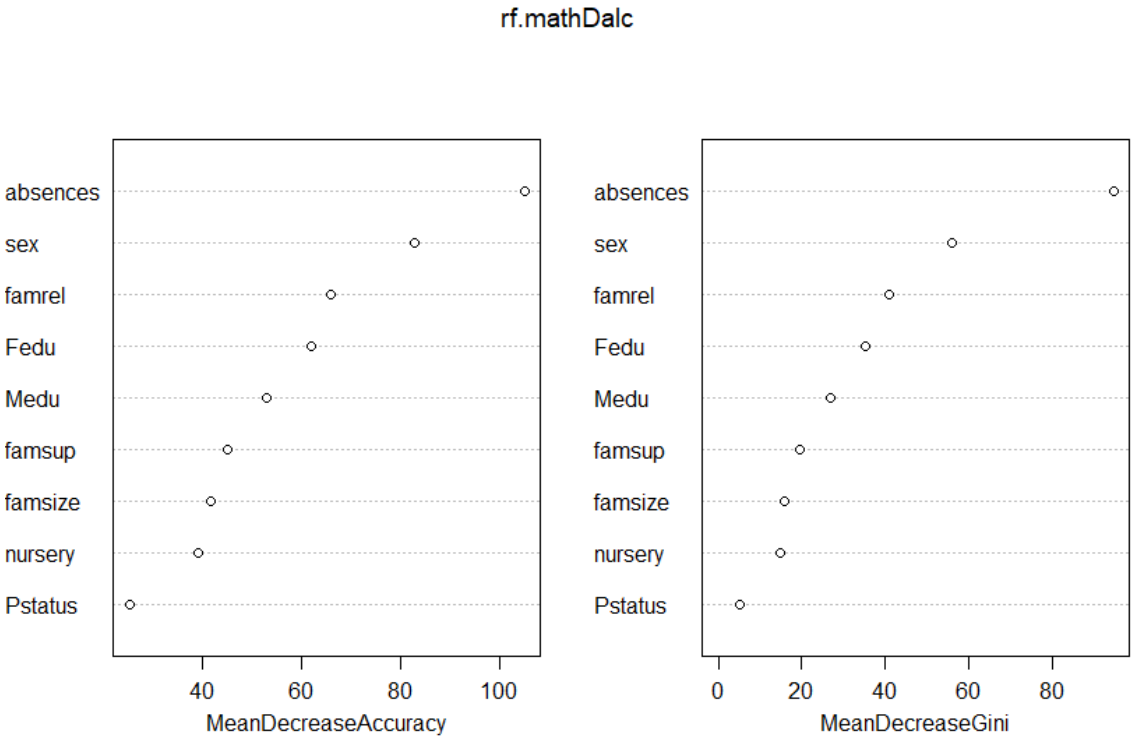
prunetree.pred	Dalc.test		
		NO	**YES**
	NO	86	13
	YES	39	111

accuracy = 0.7911647

Bagging & RandomForest Results

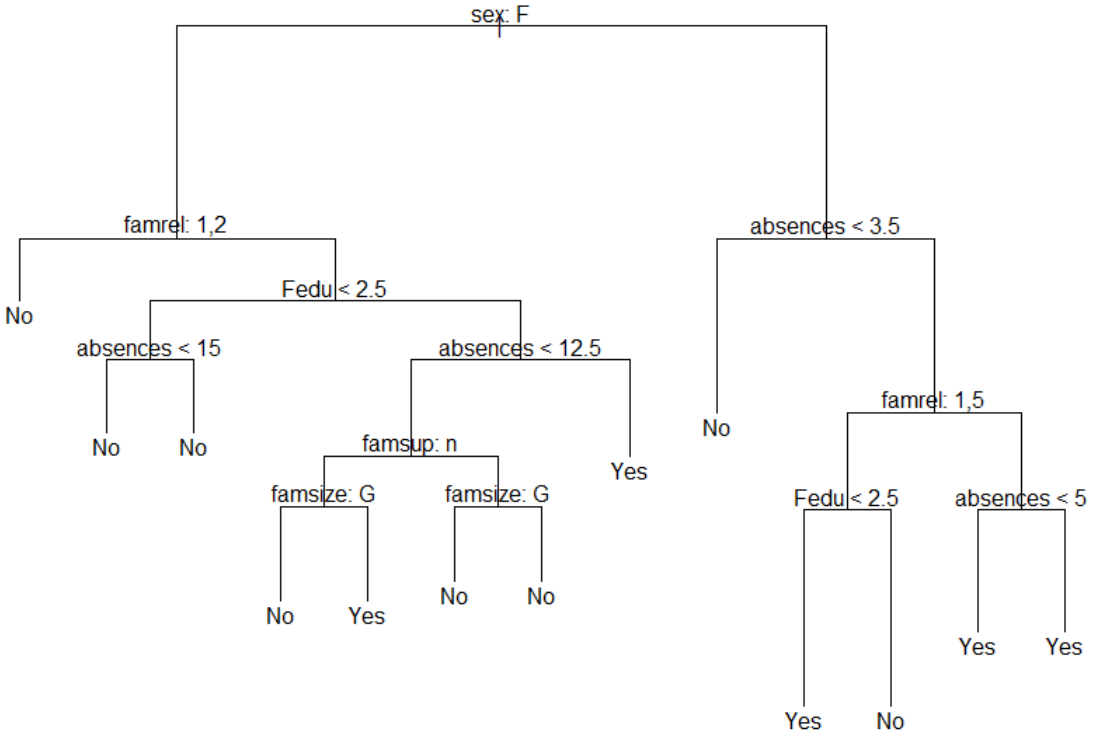
Bagging	RandomForest
0.875502	0.8795181

Verifying the Importance of Predictors: Accuracy and Gini Plot



Decision Tree Results for Weekend Alcohol Consumption

Pruned Tree Model



Pruned Tree Model Results

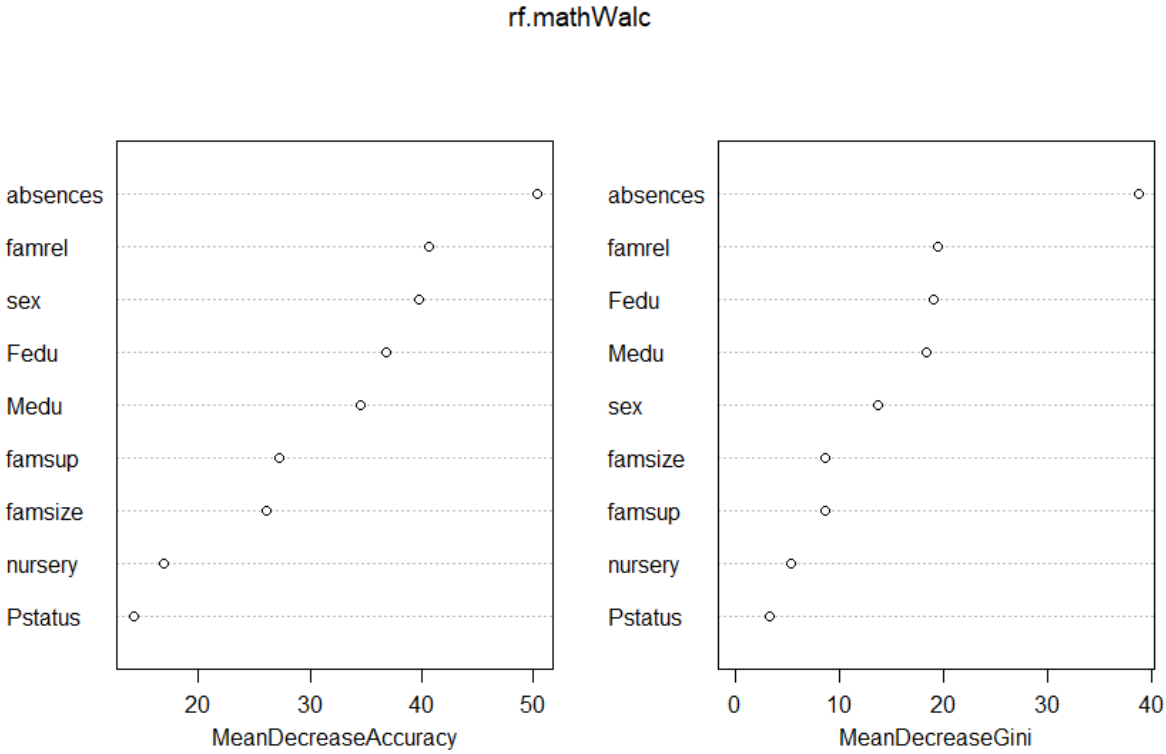
prunetree.pred	Walc.test		
		NO	**YES**
	NO	43	43
	YES	10	23

accuracy = 0.5546218

Bagging & RandomForest Results

Bagging	RandomForest
0.5378151	0.5546218

Verifying the Importance of Predictors: Accuracy and Gini Plot



6. State the final model and clearly show your findings. Use figures, tables, or formatted outputs

It is clear that the logistic model showed a better overall performance and outputs. Below is the summary of accuracy produced by each model.

	Logistic Regression	Tree	Bagging	RandomForest
Accuracy (Weekday)	0.7692308	0.7108434	0.8875502	0. 8795181
Accuracy (Weekend)	0.7319182	0.5546218	0.5378151	0.5546218

7. Conclusion; implications to the target audience

(Intercept)	math$sexM	math$goout2	math$goout3
7.293839e-04	8.607440e+00	1.597163e+00	2.281399e+00
math$goout4	**math$goout5**	**math$schoolMS**	**math$absences**
5.775550e+00	1.382077e+01	5.165561e+00	1.070722e+00
math$traveltime2	**math$traveltime3**	**math$traveltime4**	**math$activitiesyes**
2.670773e-01	1.509010e+00	7.757424e+00	4.403500e-01
math$higheryes	**math$reasonhome**	**math$reasonother**	**math$reasonreputation**
8.517728e+00	1.534271e+00	5.529093e+00	1.409036e+00
math$famsizeLE3	**math$nurseryyes**		
2.292881e+00	4.808093e-01		

Above is the odds ratio of various predictors. Disregarding some of the uncontrollable factors like a student's gender, some of the important predictors of higher alcohol consumption rate based on the odds ratio are as follows:

higher yes: wants to take higher education (yes)
traveltime 4: home to school travel time (> 1hour)
goout 4: going out with friends (4 = high)
reasonother: the reason to choose this school (other <than 'close to home'; 'school reputation'; or 'course preference')
schoolMS: student's school (MS = Mousinho da Silveira High School)
nursery yes: attended nursery school (yes)
activities yes: extra-curricular activities (yes)

Many of the predictors with high odds ratio overlaps with the predictors chosen during the stepwise regression. However, some factors such as father's job, and the quality of family relationship was not included in the list.

From this observation, we can reflect some characteristics of a person who we consider as 'outgoing' correlated with higher alcohol consumption. The more you go out with friends, the more activities you are involved, the more social interaction there is, there was a higher correlation with increased alcohol consumption.