Examining the Student Alcohol Consumption

Mine Your Business

Contents

Addressing Questions Exploratory Data Analysis Reducing Variables Bootstrapping Logistic Regression Classification Tree Conclusion

Data Source & Questions

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

UCI Machine Learning on Kaggle.com

Exploring the Data

dim(math) 395 Records, 33 Variables

Used structure function str(math)

16 int variables

16/33

Exploring the Data – Summary Statistics

Five Number Summary of Continuous Variables

**age Medu Fedu failures absences**

**Min** 15 0 0 0 0

**1Q** 16 2 2 0 0

**Median** 17 3 2 0 4

**Mean** 16.7 2.75 2.52 0.33 5.71

**3Q** 18 4 3 0 8

**Max** 22 4 4 3 75

Exploring the Data – Binary Responses

Counts of Binary Responses

**address school**

**sex**

**famsize**

**GP** 349

**F** 208

**R** 88

**GT3** 281

**MS** 46

**M** 187

**U** 307

**LE3** 114

**Pstatus**

**A** 41

**T** 354

**schoolsup**

**famsup**

**paid**

**NO** 344

**NO** 153

**NO** 214

**YES** 242

**YES** 242

**YES** 181

Exploring the Data – Binary Responses 2

Counts of Binary Responses Continued

**higher activities**

**nursery**

**internet**

**NO** 194

**NO** 81

**NO** 20

**NO** 66

**YES** 201

**YES** 314

**YES** 375

**YES** 329

**romantic**

**NO** 263

**YES** 132

**Dalc**

**Walc**

**NO** 351

**NO** 236

**YES** 44

**YES** 159

Exploring the Data – Likert-type Scale

Responses in a Rating Scale

**Travel Time** 1 51 101 151 201 251

**15-30min (2) - 107**

**15min (1) - 257**

**Study Time**

1 51 101 151 201

**<2hours (1) - 105**

**Quality of Family Relationships**

1 51 101 151 201

**>60min (4) - 8**

**>10hours (4) - 27**

**Excellent (5) - 105**

**30-60min (3) - 23**

**Very Good (4) - 195 Good (3) - 68 Bad (2) - 18**

**Very Bad (1) - 8**

1 51 101 151

**5-10hours (3) - 65**

**2-5hours (2) - 198**

**Free Time**

**Going out with Friends**

**Current Health Status**

1 51 101

**Very High (5) - 53**

**Very Good (5) - 146 High (4) - 86**

**Good (4) - 66**

**Very Low (1) - 18**

**Moderate (3) - 130**

**Low (2) - 103 Very Low (1) - 23**

1 51 101

**Very High (5) - 40**

**High (4) - 114 Moderate (3) - 157**

**Bad (2) - 45**

**Moderate (3) - 91**

**Low (2) - 64**

**Very Bad (1) - 47**

Exploring the Data – Nominal Response

Nominal Response Variables

**Mother's Father's Job**

**Job**

0 50 100 150

0 50 100 150

**at\_home**

**at\_home**

**Health**

**Health Teacher**

**Teacher Services**

**Services Other**

**Other**

1 101 201 301 **Student's Guardian**

**Reason to Choose this School**

1 51 101 151 201

**Mother**

**Father**

**Other**

**Course Preference**

**Close to Home**

**School Reputation**

**Other**

Converting Data – Categorical Variable

famrel, health, Mjob, Fjob, traveltime, studytime, freetime, goout, G1, G2, G3

math$famrel=as.factor(math$famrel) math$health=as.factor(math$health) math$Mjob=as.factor(math$Mjob) math$Fjob=as.factor(math$Fjob) math$traveltime=as.factor(math$traveltime) math$studytime=as.factor(math$studytime) math$freetime=as.factor(math$freetime) math$goout=as.factor(math$goout) math$G1=as.factor(math$G1) math$G2=as.factor(math$G2) math$G3=as.factor(math$G3)

Converting Data – Binomial Variable

Convert integer response to binomial “YES” and “NO”

math$Dalc[math$Dalc>2]="Yes" math$Dalc[math$Dalc<=2]="No" math$Walc[math$Walc>2]="Yes" math$Walc[math$Walc<=2]="No" math$Dalc=as.factor(math$Dalc) math$Walc=as.factor(math$Walc)

Converting Data – Variables Contrasts

Two Variable Contrasts

contrasts(math$Dalc) contrasts(math$Walc)

**Dalc YES**

**NO** 0

**YES** 1

**Walc YES**

**NO** 0

**YES** 1

Choosing Predictors – Assumption

sex: the student’s gender

Pstatus: the parent’s cohabitation status

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

Choosing Predictors – Stepwise Selection

Stepwise Regression

null = glm(math$Dalc ~ 1, family="binomial",data = math) full = glm(math$Dalc ~ math$school+math$sex+math$age+math$address+math$famsize+math$P status+math$Medu+math$Fedu+math$Mjob+math$Fjob+math$reason+m ath$guardian+math$traveltime+math$studytime+math$failures+math$sch oolsup+math$famsup+math$paid+math$activities+math$nursery+math$hi gher+math$internet+math$romantic+math$famrel+math$freetime+math$ goout+math$health+math$absences+math$G1+math$G2+math$G3, family="binomial", data = math) step.reg = step(null, scope=list(lower=null, upper=full),direction = 'both') summary(step.reg)

Choosing Predictors – Result

Dalc AIC: 278.04 > 226.76

sex, goout (going out with friends), school, absences, traveltime, activities, higher (wants to take higher education), reason (reason to choose school), famsize, nuersery

Walc AIC: 534.48 > 441.27

goout, Fjob (father’s job), sex, absences, famrel, nursery, paid (extra paid classes within the course subject), traveltime, address (urban/rural), activities

Choosing Predictors – Result Cont.

50% of our assumption

**sex, absences, famrel, Pstatus, romantic, failure**

The Largest Odds Ratio – Male Students

Model Selection – Bootstrapping

Limited data size

Sampling with replacement

4 resampled datasets from bootstrap

set.seed(14568) train.dalc1=sample(nrow(math), 395 , replace=TRUE)

set.seed(23258) train.dalc2=sample(nrow(math), 395 , replace=TRUE)

set.seed(36585) train.dalc3=sample(nrow(math), 395 , replace=TRUE)

set.seed(45823) train.dalc4=sample(nrow(math), 395 , replace=TRUE)

Mean accuracy

Logistic Regression

Using 4 resampled datasets from bootstrap

Apply logistic model on each dataset

Logistic Regression – Implementation

The R Code

The Prediction

Logistic Regression – Results

table(glm.pred1, test.truevalue)

**test.truevalue**

**glm.pred1 NO YES**

**NO** 334 24

**YES** 17 20

Classification Tree

Definition

Purpose

Classification Tree – Predictors

**famsize**

**LE3: <=3; GT3: >3**

**nursery YES; NO Pstatus T: living together; A: living apart**

**famrel Numeric:**

**from 1 (Very Bad) – 5 (Excellent)**

**Medu**

**0: None; 1: Primary Edu. (4th grade) 2: 5th-9th grade; 3: Secondary Edu. 4: Higher Education**

**Fedu Same as Medu**

**famsup YES; NO absences Count**

Classification Tree – Result (Weekday)

Tree Diagram for the Weekday Alcohol Consumption

treeplot.pruned tree with optimal size 9

Classification Tree – Result (Weekend)

Tree Diagram for the Weekend Alcohol Consumption

treeplot.pruned tree with optimal size 8

Classification Tree – Confusion Matrix

Confusion Matrix

mean(prunetree.pred==Dalc.test) [1] 0.8810127

**Dalc.test**

**Prunetree.pred NO YES**

**NO** 392 22

**YES** 29 15

mean(prunetree.pred==Walc.test) [1] 0.685544

**Dalc.test**

**Prunetree.pred NO YES**

**NO** 182 54

**YES** 67 92

Classification Tree – Interpretation (Weekday)

Interpretation (Weekday)

**Important Variables**

• absences

• Fedu

• Medu

• sex

• famrel

• nursery

Classification Tree – Interpretation (Weekend)

Interpretation (Weekend)

**Important Variables**

• Absences

• Fedu

• Sex

• Famsup

• Medu

• famrel

Classification Tree – Evaluating the Tree Model

**Tree Bagging Random Forest**

**Accuracy (Weekday) 0.8810127 0.8101266 0.8101266**

**Accuracy (Weekend) 0.685544 0.5949367 0.6202532**

Classification Tree – Key Findings

Evidence ***“There is evidence suggesting that individuals who are children of alcoholics have a higher probability of becoming alcoholic or problem drinkers as a result of their unstable childhood family systems.”***

***– Professor Engs, Ruth C, Indiana University studying Family Background of Alcohol Abuse and Its Relationship to Alcohol Consumption among Students***

Conclusion

**Logistic Regression > Decision Tree**