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

UCI Machine Learning on Kaggle.com

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

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

school	
GP	349
MS	46

sex	
F	208
M	187

address	
R	88
U	307

famsize	
GT3	281
LE3	114

Pstatus	
A	41
T	354

schoolsup	
NO	344
YES	242

famsup	
NO	153
YES	242

paid	
NO	214
YES	181

Exploring the Data – Binary Responses 2

Counts of Binary Responses Continued

activities	
NO	194
YES	201

nursery	
NO	81
YES	314

higher	
NO	20
YES	375

internet	
NO	66
YES	329

romantic	
NO	263
YES	132

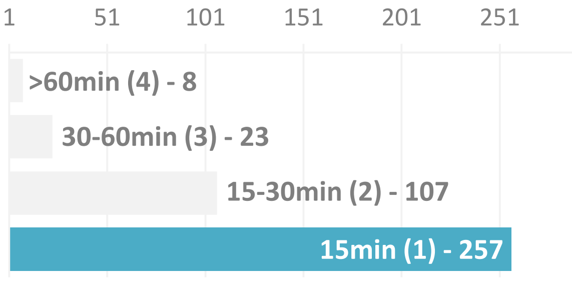
Dalc	
NO	351
YES	44

Walc	
NO	236
YES	159

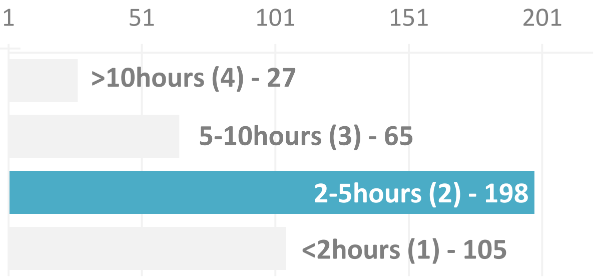
Exploring the Data – Likert-type Scale

Responses in a Rating Scale

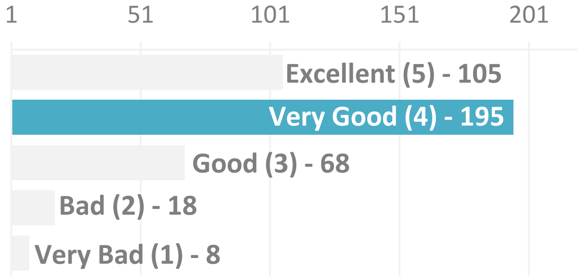
Travel Time



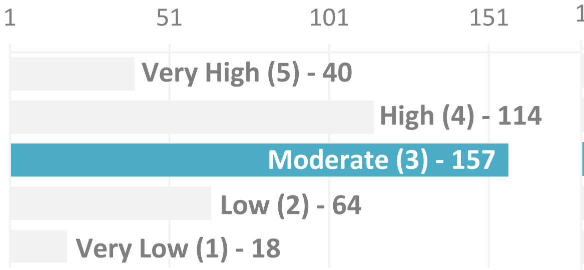
Study Time



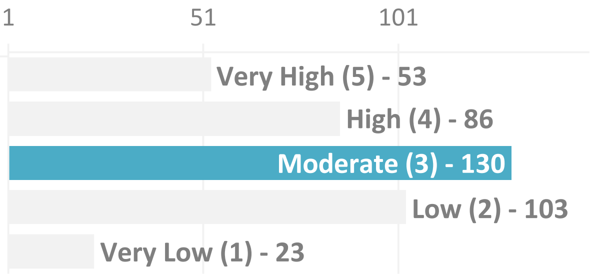
Quality of Family Relationships



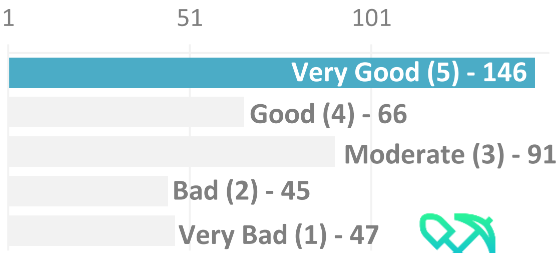
Free Time



Going out with Friends



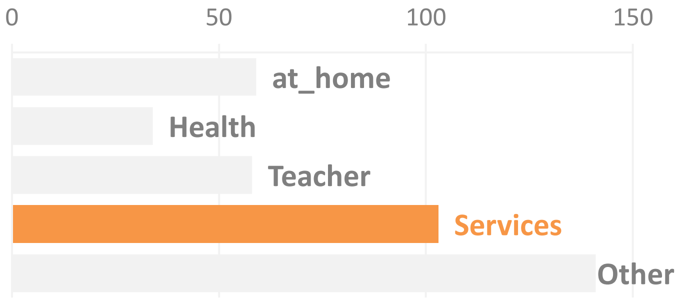
Current Health Status



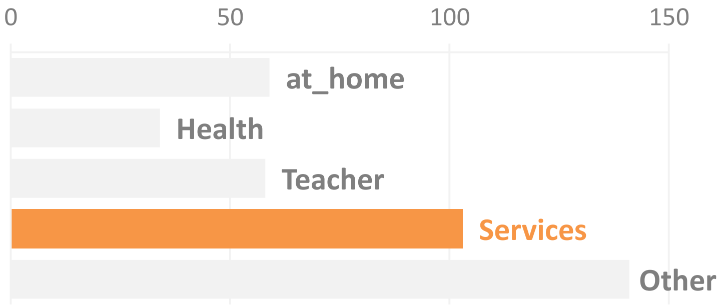
Exploring the Data – Nominal Response

Nominal Response Variables

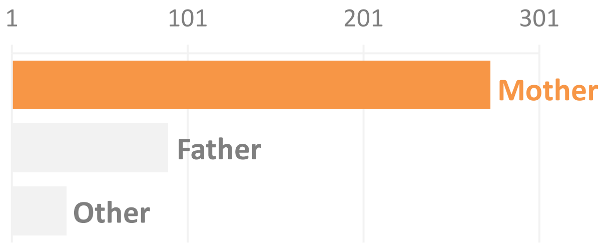
Father's Job



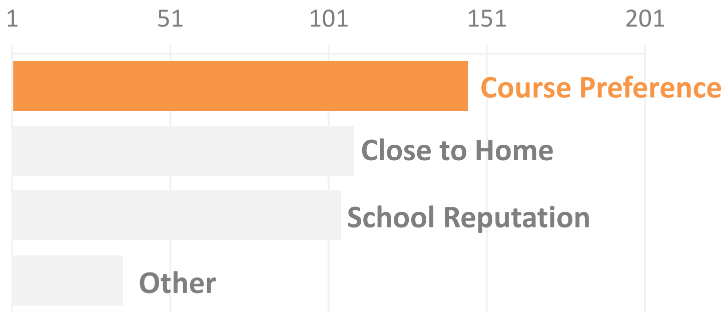
Mother's Job



Student's Guardian



Reason to Choose this School



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$famsize+math$Pstatus+math$Medu+math$Fedu+math$Mjob+math$Fjob+math$reason+math$guardian+math$traveltime+math$studytime+math$failures+math$schoolsup+math$famsup+math$paid+math$activities+math$nursery+math$higher+math$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)
```

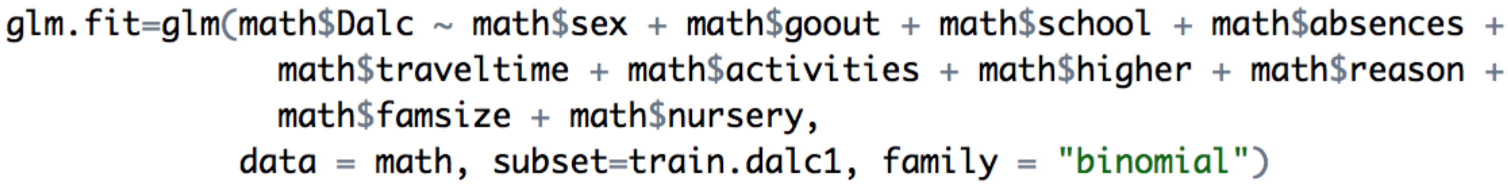
- Mean accuracy

Logistic Regression

- Using 4 resampled datasets from bootstrap
- Apply logistic model on each dataset

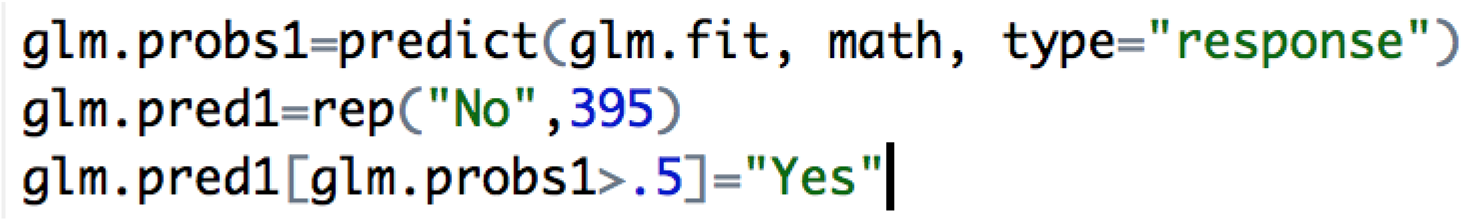
Logistic Regression – Implementation

The R Code



```
glm.fit=glm(math$Dalc ~ math$sex + math$goout + math$school + math$absences +
              math$traveltime + math$activities + math$higher + math$reason +
              math$famsize + math$nursery,
            data = math, subset=train.dalc1, family = "binomial")
```

The Prediction



```
glm.probs1=predict(glm.fit, math, type="response")
glm.pred1=rep("No",395)
glm.pred1[glm.probs1>.5]="Yes"
```

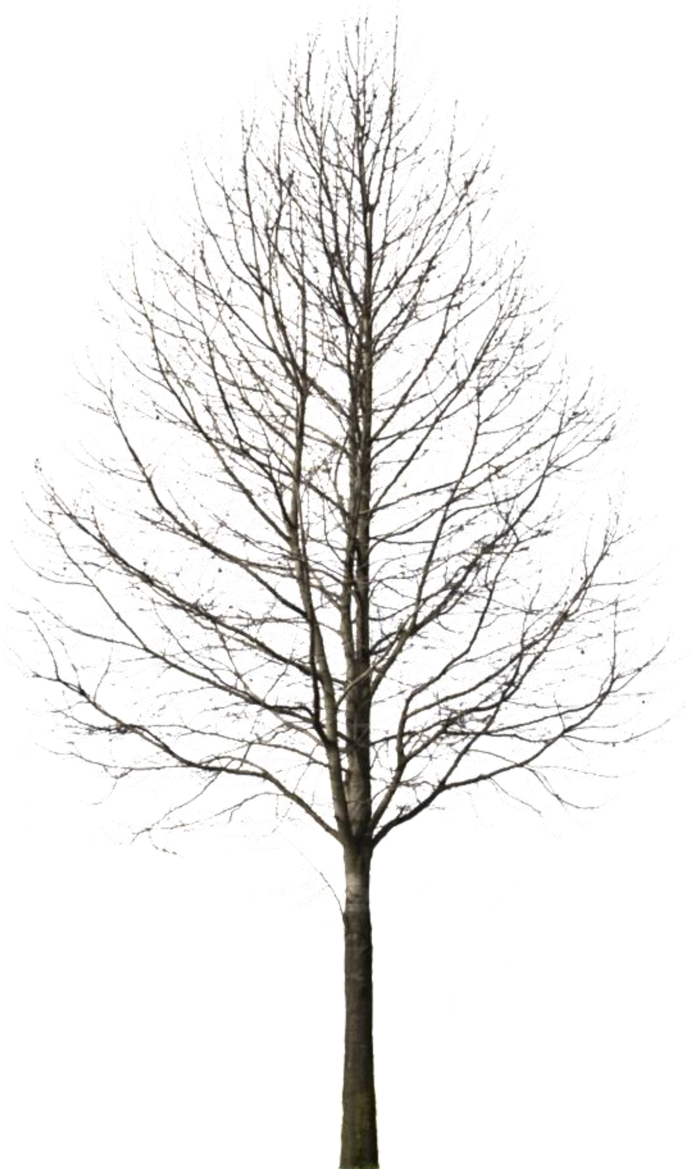
Logistic Regression – Results

```
table(glm.pred1, test.truevalue)
```

		test.truevalue	
glm.pred1		NO	YES
	NO	334	24
	YES	17	20

Classification Tree

- Definition
- Purpose

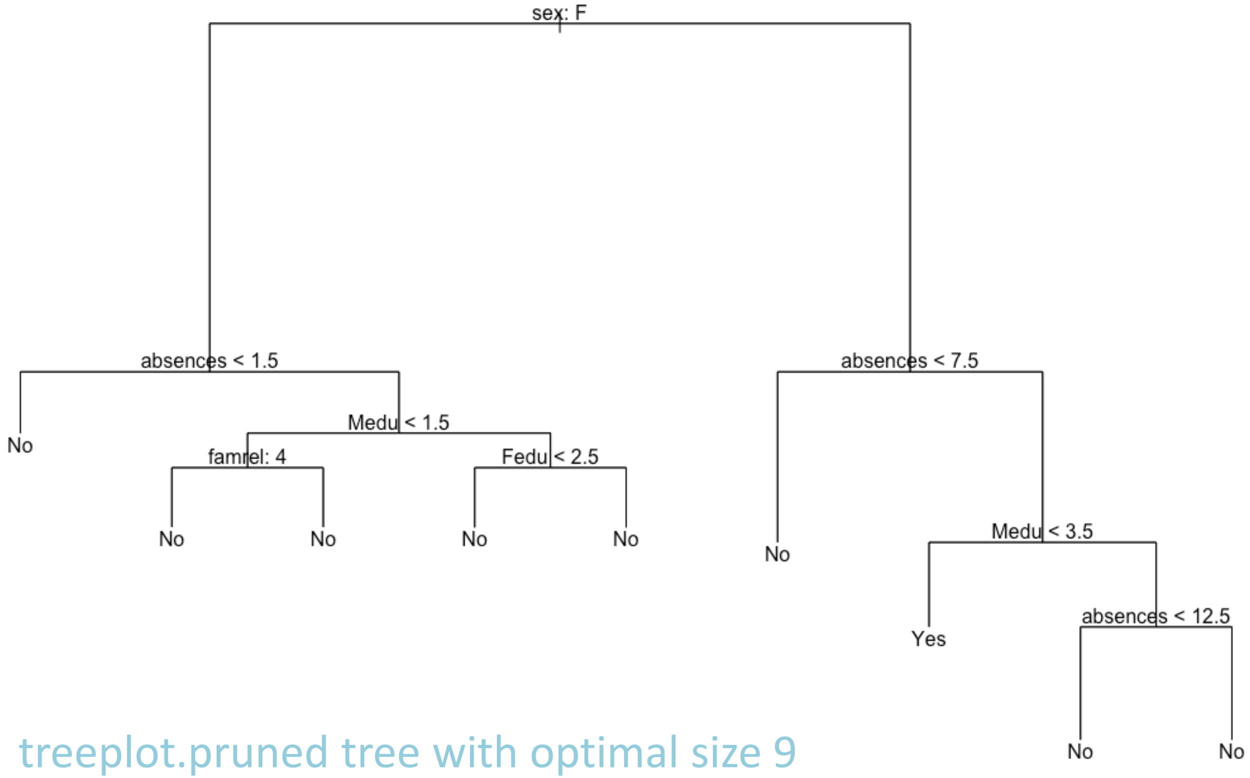


Classification Tree – Predictors

famsize	LE3: <=3; GT3: >3
Pstatus	T: living together; A: living apart
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
nursery	YES; NO
famrel	Numeric: from 1 (Very Bad) – 5 (Excellent)

| Classification Tree – Result (Weekday)

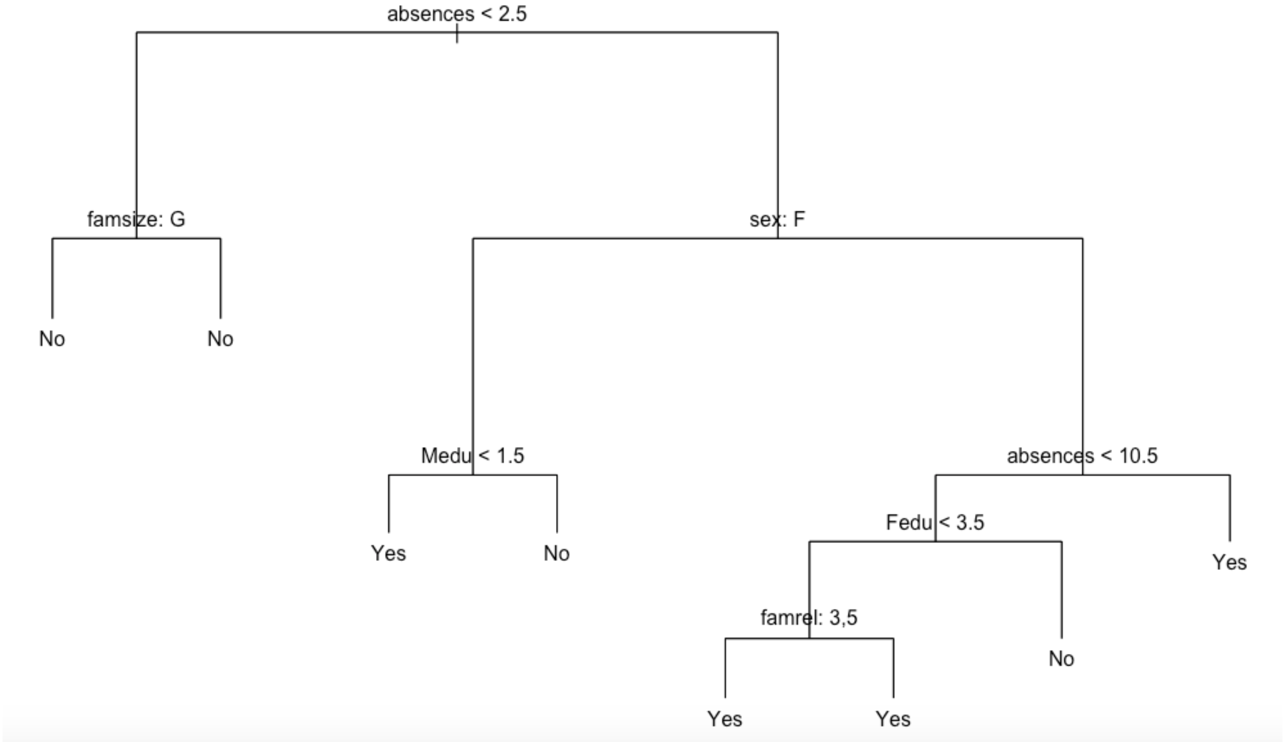
Tree Diagram for the Weekday Alcohol Consumption



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
```

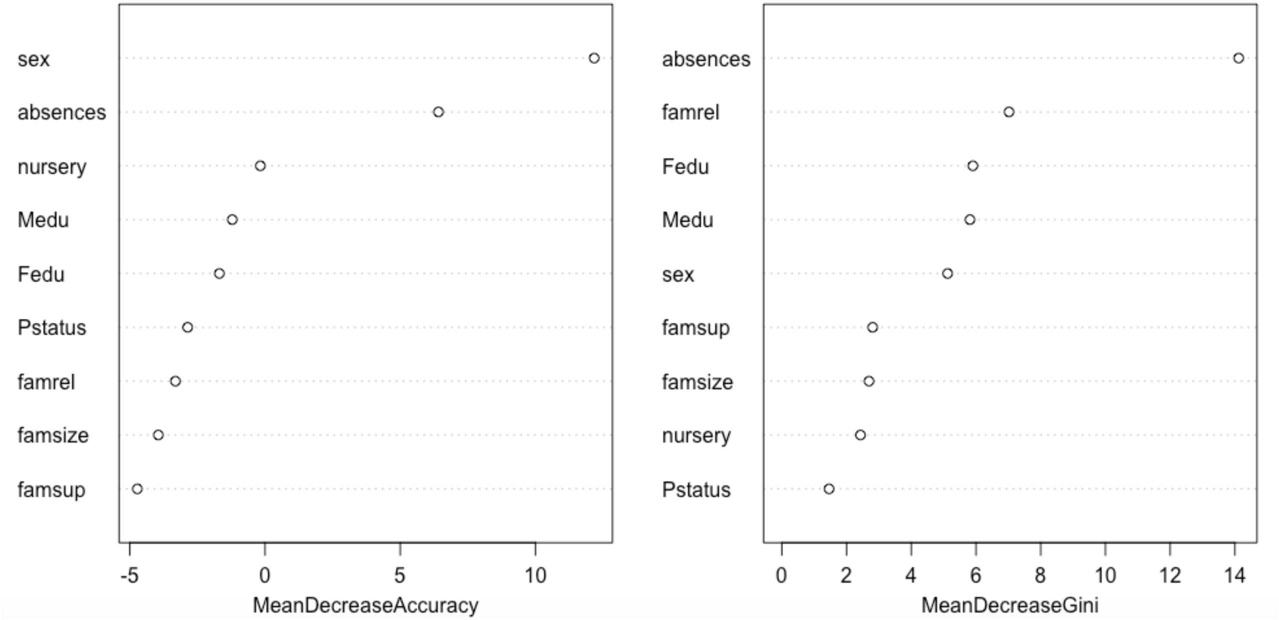
		Dalc.test	
Prunetree.pred		NO	YES
	NO	392	22
	YES	29	15

```
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)

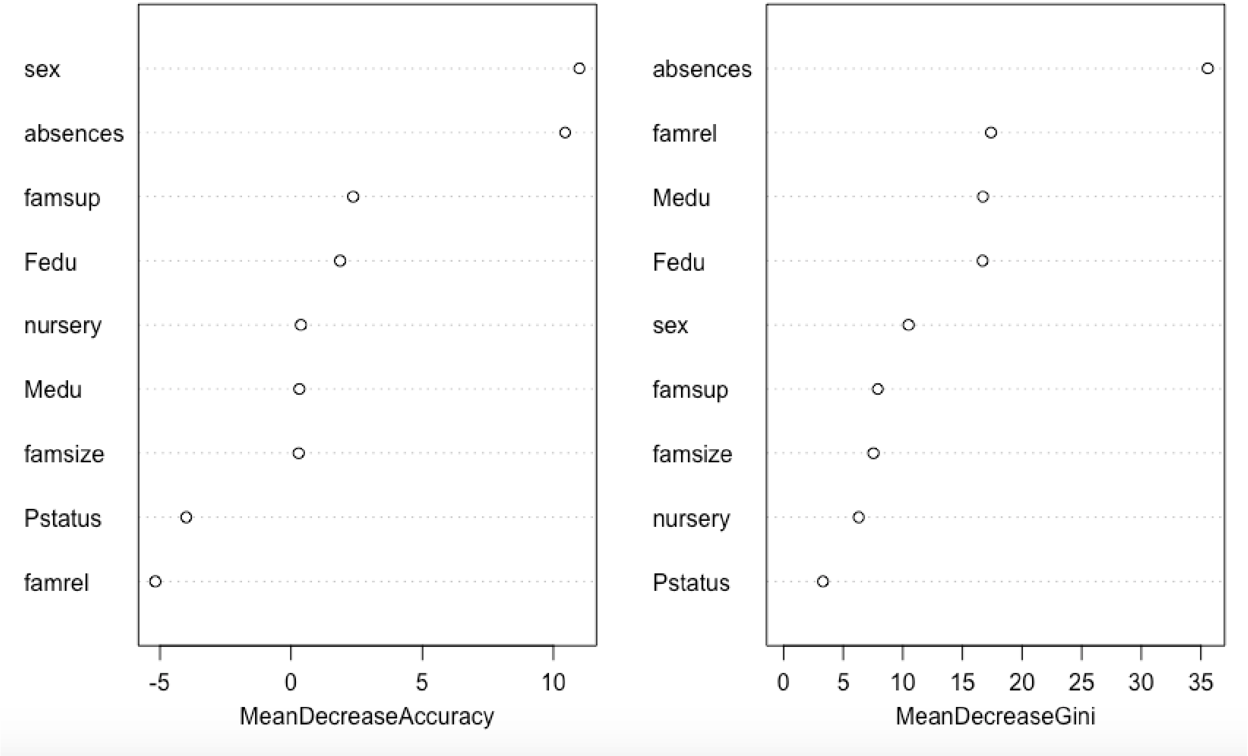


Important Variables

- 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