STA3920 Project Ideas

0. Merge the two files

1. First question: How does students’ personal background (family size, age, gender, etc.), if any, correlated to their alcohol consumption?

-Use alcohol consumption as a response

-Predictors to use

-use forward selection (find optimal predictors)

-use bootstrap (data size is limited; find out mean squared error)

-sex, ptatus, romantic, absence, failures, famrel (confirm if it’s categorical or quantitative)

-two multiple regression model (one for weekday alcohol consumption, one for weekend

alcohol consumption)

-logistic regression

-Clustering(possible categorical variables)

2. 2nd question: How does alcohol consumption, if any, affects students’ academic performance?

-Use GPA as the response, alcohol consumption as a predictor

3. Presentation

1. Intro (question/exploratory analysis)
2. Reducing variables (how we determine variables)
3. Logistic regression
4. Decision Tree
5. Bootstrap
6. Conclusion

1. Introduction

-Introduce the group

-Go over the table of contents

-Address the questions & cover exploratory 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 about student

background and alcohol consumption conducted in Europe

-What we want to know from this data set is:

-How are the students’ personal background (family size, age, gender, etc.), if

any, correlated to their alcohol consumption?

-To give a little bit of overview of this dataset (Exploratory Analysis)

-The data were obtained in a survey of high school students enrolled in math

class

-The data contains 395 records with 33 explanatory variables related to personal

backgrounds such as age, gender, family size, and so on about each student as

well as the rate of alcohol consumption

-We used the structure function (str()) in R to examine the structure (chart)

-we learned that 10 of the variables were numerical variables 23 of them

are categorical

-Here are the summary statistics for the continuous variables (chart/explain)

-And this is the summary statistics for the categorical variables (chart/explain)

-In order to proceed with the data analysis, we converted the continuous

variables to categorical ones using the as.factor() function (picture)

-At the same time, we made two variable contrasts or dummy variables

2. Reducing Variables (choosing predictors)

-Our primary choice of possible predictors of higher alcohol consumption rate was:

1. Sex: the student’s gender

2. Pstatus: the parent cohabitation status; whether the parents are living together

or apart

3. Romantic: whether the student is in a relationship or not

4. Absence: the number of school absences

5. Failures: the number of past class failures

6. Famrel: the quality of family relationship scaled to 1-5 (chart)

-However, since it was based on our intuition, and because there were many more

predictors in the dataset, we decided to use the best subset selection method in R which

will allow us to reduce the variables to a more meaningful ones

-Using the stepwise regression (step()) function, we set the test to F and direction to both

as the option we obtained the following results: (chart/explain)

3. Logistic Regression

4. Classification Decision Tree

* **Definition:** [Classification tree](https://en.wikipedia.org/wiki/Classification_tree) analysis is when the predicted outcome is the class to   
  which the data belongs.
* **Purpose:** To predict the Daily/Weekly Alcohol Consumption associated with a variety of familty’s condition.
* **Predictor:** 
  + Famsize(family size) -- LE3: Less or equal to 3; GT3: greater than 3
  + Pstatus(Parent's cohabitation status) -- T: living together; A: living apart
  + Medu(Mother’s education) -- 0 - none; 1 - primary education (4th grade);2 - 5th to 9th grade; 3 - secondary education; 4 - higher education)
  + Fedu(Father’s education) -- Same as Medu
  + Famsup(Family educational support) -- Yes; No
  + Nursery(Attended nursery school) -- Yes; No
  + Famrel(Quality of Family relationships) numeric: from 1 - very bad to 5 - excellent)
  + Absences(Number of school absences)
* Result & Interpretation
  + Showing the decision trees for both Daily/Weekly Alcohol Consumption(MSE)
  + R code consist of training and testing dataset, applied library *rpart* for displaying decision trees
    - Features and each function
* Summary:
  + Consuming alcohol does harm to students physically and mentally.
  + Discussing the crucial factors as result of family’s background could be more likely to cause students consume alcohol.

5. Bootstrap

Why bootstrap:

-limited amounts of data

-Estimate error rate

-Train model on the entire data set

-We have decided to use bootstrap because of the limited data size we had. By

implementing bootstrap, we were able to obtain a larger dataset and therefore make an improved inference from the data

Definition:

-the size of the original data set is n, sample it with replacement n times.

-use as training data

Our data:

-the bootstrap approach on our sample containing n = 395 observations.

-each bootstrap data set

contains n = 395 observations, sampled with replacement from the original data set.

Result

We apply variables into logistic regression model 4 times with different bootstrap data set.

The average correctly predicts 90% of Dalc status. But only 64% of Walc status.

6. Conclusion