Understanding Data Mining

Lesson 2: Using R for Data Mining Name____________________________

I. Many times, a situation depends on numerous factors. Someone's credit score, for example, isn't just determined by a single answer to a question on a loan application. List some of the things you think a credit company asks for on a credit application.

____

____

____

Obviously, some pieces of information you listed above are more relevant than others in determining whether or not someone should get a loan. As a result, we need a process that can analyze these large amounts of data in order to determine which factors are the BEST predictors of someone's "credit-worthiness." One such process is called ***data mining***.

II. Search the internet for five facts about data mining. Record them below:

1. ____

2. ____

3. ____

4. ____

5. ____

III. We can use R to experiment with the data mining process. Let's first try out the process on a data set that is already in R.

1. Type **`attitude`** to see the data set we are going to use.

2. Type **`help(attitude)`** to see the details of the data set. It is always important to understand the source of data when trying to make predictions and draw conclusions!

3. List the variables presented in this data set below, and then circle the dependent variable:

____

4. Now return to your R console and load the data using the command **`data(attitude).`**

5. Let's find a linear model to relate the variables that impact a person's overall rating. Our command will be: **`att<-lm(attitude$rating ~ attitude$complaints +attitude$privileges +attitude$learning + attitude$raises + attitude$critical + attitude$advance).`**

Record the equation below:

____

6. We can use an analysis of variance to draw initial conclusions about which variables are the most influential in a person's overall rating. To access this, type the command **`anova(att).`**

7. The variables with the highest F-values have the highest degree of influence on a person's overall rating. List the variables in order from the highest to the lowest F-values:

____

8. While F-values help us identify the overall relevance of each variable, it doesn't allow us to draw conclusions about which variables may have no measurable effect on a person's overall rating. As a result, statisticians often times rely on variable selection processes to identify the variables that should be considered. We will try three methods of variable selection:

-*forward selection*: start with the first variable and keep adding variables to the prediction equation

-*backward selection*: start with all variables and eliminate variables from the prediction equation one at a time

-*stepwise selection*: like forward selection, except some variables may be deleted after they have been added

9. Let's try forward selection first. The command is **`attfor<-step(lm(attitude$rating~1, data=data.frame(attitude)), scope=list(lower=~1, upper=~attitude$complaints + attitude$privileges + attitude$learning + attitude$raises + attitude$critical +attitude$advance), direction = “forward”).`**

10. Look at the final step and record the relevant variables as identified by forward selection.

____

11. Now, let's try backward selection. The command is **`attback<-step(att, direction = “backward”)`**. Look at the final step and record the relevant variables according to backward selection.

____

12. Finally, let's try stepwise selection. The command is **`attstep<-step(att, direction = “both”).`** As before, record the relevant variables according to stepwise selection.

____

13. Our conclusion is that the factors that should be considered when determining someone's overall rating are:

____

14. How could our conclusion be helpful to people studying the results of the survey? ________

____

IV. Use the internet to research three uses of data mining. Make sure to explain how data mining is used in each of the three cases:

1. ____

____

2. ____

____

3. ____

____