

EXAM 2

Total: 375 points

What to Submit

- You will submit 5 files. D2L will allow multiple submissions. Submit one file at a time. Details follow.
- One file. *File Name:* Exam 2_537_LastName_FirstName.docx or .pdf (no .zip). In this file, post screenshots for each part in sequence. If you cannot do a part, state that!
- **Q1.R** In this file submit the R-code only. (if .R extension is disallowed by D2L, try .zip or submit as .doc)
- **Q2.R** In this file submit the R-code only.
- **Q3.R** In this file submit the R-code only.
- **Q4.R** In this file submit the R-code only.
- Include comments in your code.
- Make sure your code is written nicely (i.e. adequate tabs, spaces, etc.)
- Failure to adhere to any of the above instructions will result in point-deductions.

Q1

100 points

File to Use: **ToyotaCorollaFinal.csv**

The file is an extension of the example you saw in the Primer posted on D2L. It includes the sales price and other information on the car, such as its age, odometer mileage, fuel type, horsepower, etc. There are more than 35 attributes in total, and more than 1400 records. Your goal is to predict the price of a used car from the given data and specifications. For this purpose, use 50% of the records for training, keep 30% for validation, and then use the remaining 20% of the records for testing.

For the outcome variable Price, use the following predictor variable set: Car_Age, Odometer, Fuel_Type, HP, Automatic, Doors, Qtr_Tax, Build_Guarantee, Guarantee_Months, Cold_Air, Cold_Air_Auto, CD_Player, Powered_Windows, Sports, and Towing.

- A. From the above-mentioned predictors, which (say 3 or 4) are most important for predicting the price of the car.
 - B. Now, the predictors that you deem most important, use them to assess the model performance (for predicting prices).
-

Q2**75 points**

File to Use: **LawnMowers.csv**

A marketing company is tasked by a lawn mower manufacturer to identify whether a given household will be a prospective owner or not. The given file lists a sample of households that the company collected classifying the households on the basis of Income (in \$1000s) and Lot Size (in 1000 sq. ft).

- A. What type of regression is best suited in this case, and why?
 - B. Create a colored scatter plot of Income vs. Lot Size and distinguish between owners and nonowners. Looking at the plot, owners have a higher avg. income, or nonowners?
 - C. What are the chances that a household with a \$50,000 income and a lot size of 25,000 sq. ft. is an owner?
-

Q3**150 points**

File to Use: **Airlines.csv**

Frequent flier data for almost 4000 participants is listed in the file. It includes the participants' mileage history, and how the miles were attained (e.g. using Credit card, etc.) or spent in the previous year. Our goal is to look for groups of passengers that may have some underlying commonalities for the purpose of targeted marketing.

- A. Apply hierarchical clustering on the given data (normalize the data first). How many clusters appear?
 - B. If you don't normalize, what happens then?
 - C. Cluster stability check: Remove 5% of the data randomly (i.e. take a random sample of 95% of the records), and repeat the analysis. Does the dendrogram look the same?
 - D. Use k-means clustering with the number of clusters that you found in part A. Are the results comparable/similar?
 - E. Which clusters would you target for offers, and what types of offers would you target to customers in that cluster?
-

Q4**50 points**

File to Use: **Groceries.csv**

The file contains information collected from a one-month operation of a real-world grocery store. It contains 9835 transactions (rows) and 169 unique items (columns) bought by the customers. Conduct a market basket analysis on the given dataset with a support of 0.005, and confidence of 0.15. Use transactions that have a minimum of three items in them.

What is the total number of rules you get?

Now sort the rules by lift, and state the top most rule in Normal English.

NOTE:

Due to the nature of questions, and choice of solutions, the chances of Your code matching another student's code are slim. Thus, please do NOT plagiarize!
