Upload a .pdf titled *yourlastname_daii_assignment.pdf* with all of your code, figures, and summaries of responses to the project to Blackboard by Friday, August 7th at 7:45pm.

Part 1: Definition (20 Points)

Define the following terms and where appropriate compare and contrast with the alternative method.

- **a)** The Trade-Off Between Prediction Accuracy and Model Interpretability
- **b)** Supervised Versus Unsupervised Learning
- **c)** The Bias-Variance Trade-Off
- **d)** Linear Regression versus K-Nearest Neighbors
- **e)** Logistic Regression versus LDA versus QDA

Part 2: Regression (40 Points)

The table below displays catalog-spending data for the first few of 200 randomly selected individuals from a very large (over 20,000 households) data base.[1] The variable of particular interest is catalog spending as measured by the Spending Ratio (`SpendRat`). All of the catalog variables are represented by indicator variables; either the consumer bought and the variable is coded as 1 or the consumer didn't buy and the variable is coded as 0. The other variables can be viewed as indexes for measuring assets, liquidity, and spending.

```
'catalog.csv': 200 obs. of 21 variables:
SpendRat : num 11.8 16.8 11.4 31.3 1.9 ...
Age : int 0 35 46 41 46 46 46 56 48 54 ...
LenRes : int 2 3 9 2 7 15 16 31 8 8 ...
Income : int 3 5 5 2 9 5 4 6 5 5 ...
TotAsset : int 122 195 123 117 493 138 162 117 119 50 ...
SecAssets : int 27 36 24 25 105 27 25 27 23 10 ...
ShortLiq : int 225 220 200 222 310 340 230 300 250 200 ...
LongLiq : int 422 420 420 419 500 450 430 440 430 420 ...
WlthIdx : int 286 430 290 279 520 440 360 400 360 230 ...
SpendVol : int 503 690 600 543 680 440 690 500 610 660 ...
SpenVel : int 285 570 280 308 100 50 180 10 0 0 ...
CollGifts : int 1 0 1 1 0 0 1 1 1 0 ...
BricMortar : int 0 1 0 0 1 1 0 1 0 1 ...
MarthaHome : int 0 1 0 0 1 1 0 1 1 0 ...
SunAds : int 1 0 1 1 0 0 1 0 0 0 ...
ThemeColl : int 0 0 1 1 0 0 0 1 1 0 ...
CustDec : int 1 1 1 0 1 1 0 1 1 0 ...
RetailKids : int 1 1 1 0 0 0 0 1 0 0 ...
TeenWr : int 1 0 0 0 0 0 0 1 0 1 ...
Carlovers : int 0 0 0 0 0 1 0 1 0 0 ...
CountryColl: int 1 0 1 1 0 0 1 0 1 0 ...
```

[1] I thank David Cameron for providing the random sample of 200 observations from a large catalog spending

Data Cleaning

The goal of this section is to explore the data set and get it ready for analysis. There are no missing values in the data set, but there are some incorrect entries that must be identified and removed before completing the analysis. `Age` can be regarded as quantitative, and any value less than 18 is invalid. Length of residence (`LenRes`) is a value ranging from zero to someone's age. `LenRes` should not be higher than `Age`. `Income` is coded as an ordinal value, ranging from 1 to 12, it's left to you to decide if it should be treated as continuous or categorical. You should create a simple 1-2 paragraph summary of this section. Be sure to fully explain the reasoning behind transforming any columns and removing any rows. Simply saying that, "Campbell told me to" is not sufficient. Justify why it makes sense not to include any rows whose age is less than 18 or why we shouldn't use rows in which length of residence is larger than age.

Basic Summary

Provide a basic summary of the cleaned data set. Include a table of univariate statistics to summarize each variable. Choose meaningful summary statistics for each type of variable. You should also include a basic summary of the catalog spending (`SpendRat`) including an appropriate graphical display.

Modeling

We are interested in developing a model to predict spending ratio. Find a regression model for predicting the amount of money that consumers will spend on catalog shopping, as measured by spending ratio. Your goal is to identify the best model you can. In your write-up be sure to justify your choice of model, discuss any transformation you make to the variables, discuss your model fit, and discuss the effect of the significant predictors using both hypothesis tests and confidence intervals. Remember to check the conditions for inference as you evaluate your models. The data set is much too small to split into training and test data sets, so use cross validation in all your models

- **a)** Fit a linear model using least squares on the training set, and report the CV error obtained.
- **b)** Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the CV error obtained.
- **c)** Fit a lasso model on the training set, with λ chosen by cross-validation. Report the CV error obtained, along with the number of non-zero coefficient estimates.
- **d)** Fit a PCR model on the training set, with M chosen by cross-validation. Report the CV error obtained, along with the value of M selected by cross-validation.
- **e)** Fit a PLS model on the training set, with M chosen by cross-validation. Report the CV error obtained, along with the value of M selected by cross-validation.
- **f)** Find a regression method that we didn't discuss in class and fit it to the training set, report the CV error.
- **g)** Comment on the results obtained. How accurately can we predict the spending ratio? Is there much difference among the CV errors resulting from these five approaches?

Part 3: Classification (40 Points)

In this problem, you will develop a model to predict whether income exceeds $50K/yr based on census data.[2]

- **a)** Use the code in Blackboard to create the adult data set.
- **b)** Explore the data graphically in order to investigate the association between income and the other features. Which of the other features seem most likely to be useful in predicting income? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.
- **c)** Split the data into an 80% training set and a 20% test set. Set the seed at 1303.
- **d)** Perform LDA on the training data in order to predict income using the variables that seemed most associated with income in (b). What is the test error of the model obtained?
- **e)** Perform QDA on the training data in order to predict income using the variables that seemed most associated with income in (b). What is the test error of the model obtained?
- **f)** Perform logistic regression on the training data in order to predict income using the variables that seemed most associated with income in (b). What is the test error of the model obtained?
- **g)** Perform KNN on the training data, with several values of K, in order to predict income. Use only the variables that seemed most associated with income in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?
- **h)** Perform SVM on the training data, choose which kernel is best in order to predict income. What are the test errors for your best model?
- **i)** Fit a random forest model to the training data.
- **j)** Choose which model predicts income the best and justify your choice.

[2] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.