This homework is meant to give you practice with machine learning algorithms and implementing them using R. Machine learning is currently a hot field being used by both academics, including economists, for research and industry professionals so, aside from it being required for a grade, it is of your own personal interest to take this homework seriously.

To receive full credit for this homework, your D2L submission should be two files and a link to your GitHub repository:

1. Your solutions for Problem 1 in the file "Lastname_Firstname_HW3.pdf"
2. Your R code used to solve Problem 1 in the file "Lastname_Firstname_HW3.R"
    - Please use the R script template "ECON_418-518_HW3_R_Template.R" to format your solutions for these problems.
    - You must comment each line of code you write!
3. A link to your GitHub repository when you submit the homework on D2L (see Problem 2).

This homework is worth 100 points (10%) of your grade.

Problem 1 (80 Points)

Please download the data "ECON_418-518_HW3_Data.csv" on D2L. A description of this data set can be found at `https://archive.ics.uci.edu/dataset/2/adult` (clickable link). We will do some preliminary data cleaning and analysis. Then, we will apply a variety of machine learning algorithms to try to predict whether an individual's income is over 50K or not. We will then determine which model does the best job at doing so out of a wide variety of possible models.

Set your working directory to the folder you stored the data set in. Load the data set into R as a data table using R's read.csv() function. At the beginning of your R script be sure to set your seed to 418518 using the command "set.seed(418518)."

Please follow these instructions and answer these questions:

(i) (10 Points) Drop the columns "fnlwgt", "occupation", "relationship", "capital-gain", "capital-loss", and "educational-num" from the data table.

(ii) (15 Points) Please clean the data by following the steps below. This will involve using R's "ifelse()" function and the logical "or" operator.

(a) Convert the "income" column to a binary indicator where if an observation has an income value of ">50K", then change that value to a 1 and 0 otherwise.

(b) Convert the “race” column to a binary indicator where if an observation has a race value of “White”, then change that value to a 1 and 0 otherwise.

(c) Convert the “gender” column to a binary indicator where if an observation has a gender value of “Male”, then change that value to a 1 and 0 otherwise.

(d) Convert the “workclass” column to a binary indicator where if an observation has a workclass value of “Private”, then change that value to a 1 and 0 otherwise.

(e) Convert the “native_country” column to a binary indicator where if an observation has a native_country value of “United-States”, then change that value to a 1 and 0 otherwise.

(f) Convert the “marital_status” column to a binary indicator where if an observation has a marital_status value of “Married-civ-spouse”, then change that value to a 1 and 0 otherwise.

(g) Convert the “education” column to a binary indicator where if an observation has an education value of “Bachelors”, “Masters”, or “Doctorate”, then change that value to a 1 and 0 otherwise.

(h) Create an “age_sq” variable that is the aged squared.

(i) Standardize the age, age squared, and hours per week variables.

(iii) (10 Points) Please do some data analysis by answering the following questions:

(a) What is the proportion of individuals with income greater than $50,000.00 in the data set?

(b) What is the proportion of individuals in the private sector in the data set?

(c) What is the proportion of married individuals in the data set?

(d) What is the proportion of females in the data set?

(e) What is the total number of observations with a value for any column in the data set? i.e., how many NAs are in the data set?

(f) Convert the “income“ variable to a factor data type so “R“ knows it is discrete.

(iv) (10 Points) Split the original data set into training and testing sets using a 70-30 split by following these steps:

(a) Use R’s “floor” function combined with “nrow(dt) * 0.70” to find the last training set observation.

(b) Create the training data table to be the first row until the last training set observation.

(c) Create the testing data to be the observation after the last training set observation all the way to the end of the original data table.

(v) (15 Points) Now, we will estimate lasso and ridge regression models using "income" as the outcome variable and all other covariates as the explanatory variables.

(a) What is the point of using lasso and ridge regression? What is the difference between them both? What are their pros and cons?

(b) Estimate a lasso regression model using R's "train()" function contained in the "caret" package. For each model, use 10-fold cross-validation to find the best value of shrinkage parameter λ out of 50 possible evenly spaced values from 10^5 to 10^{-2}.

(c) What value of λ gives the highest classification accuracy? What is this classification accuracy?

(d) Which variables have coefficient estimates that are approximately zero when estimating a lasso regression model with this value of λ. If you defined the "train(...)" command to be "cv", then you can find these coefficients by using the "coef(cv$finalModel, s = cv$bestTune$lambda)" command.

(e) Estimate lasso and ridge regression models using only the non-zero coefficient estimate variables you obtained in the prior step. Once again, use a grid of 50 evenly spaced values of λ from 10^5 to 10^{-2} and 10-fold cross-validation via caret's "train()" function. Which model has the best classification accuracy rate?

(vi) (15 Points) Lastly, we will estimate a random forest model using "income" as the outcome variable and all other covariates as the explanatory variables.

(a) What is the process of bagging and how does the random forest model use bagging?

(b) Evaluate three random forest models using the "randomForest" package and the "caret" package's "train()" function. The first random forest model should include one-hundred trees, the second should have two-hundred trees, and the third should have three-hundred trees. Each forest should be estimated using splits of two, five, and nine random possible features. Use caret's "train()" function with 5-fold cross-validation. This process may take a few minutes to run on your computer so give it time. It took 3.75 minutes on mine.

(c) Which model gives the highest classification accuracy?

(d) How does this model compare to the best model you obtained in Part (v)?

(e) Create a confusion matrix using caret's 'confusionMatrix()' function. Make predictions using the entire training data table to generate the confusion matrix. What is the number of false positives? How about the number of false negatives? Does this suggest any issue with our model? (Think about the fraction of individuals with income greater than fifty-thousand in the data set.)

(vii) (5 Points) Evaluate the best model out of every estimated model on the testing data. What is the classification accuracy of this model when evaluated on the testing set? What does this classification accuracy represent?

Problem 2 (20 Points)

GitHub is a platform for storing code, tracking changes, and collaborating with others. Often times, employers will ask you to provide links such as LinkedIn when submitting a job application. If you have any desire to obtain a job working with data, you can provide a link to your GitHub account that has your data analysis/coding projects. It can be excellent way for you to show off the projects you are working on and increase the probability you get hired!

For this problem you will:

1. Create a GitHub account at https://github.com/.
2. Create a new repository (repo) named ML_Income_Classifier_Using_UCI_Adult_Data_Set.
3. Upload your R code from Problem 2 to the repo.
4. Create a README.md file describing the project and the R file.

Please attach a link to your GitHub repo when you submit this homework on D2L.