MAS 651 HW 2 (due 11:59 pm on Feb 2nd, 2022)

Reading: Please read the Python examples I posted on Blackboard before working on the hw problems.

Problem 1 (30 points) Analyzing Advertising Data

(Hint: You may find useful Python commands from Logistic_example2.ipynt and Xgboostexampleb.ipynt)

This data set contains the following features:

'Daily Time Spent on Site': consumer time on site in minutes

'Age': cutomer age in years

'Area Income': Avg. Income of geographical area of consumer

'Daily Internet Usage': Avg. minutes a day consumer is on the internet

'Ad Topic Line': Headline of the advertisement

'City': City of consumer

'Male': Whether or not consumer was male

'Country': Country of consumer

'Timestamp': Time at which consumer clicked on Ad or closed window

'Clicked on Ad': 0 or 1 indicated clicking on Ad

ad_data = pd.read_csv('https://raw.githubusercontent.com/wangx346/MAS651/main/advertising.csv')

(1) Print the shape of the DataFrame using data.shape
(2) Define your outcome Y to be the variable: "Clicked on Ad". Print out how many zeros are in Y and how many ones are in Y. Use sns.countplot to make a count plot for Y.
(3) Make a bar chart to show how the frequency of clicking on the ad depends on gender. Make a histogram of age.

(4) Create a training data set and a test data set such that the test data set contains randomly 20% of the data set. Set the random seed to be 40. Read the Xgboostexampleb.ipynt example for binary classification. Implement Xgboost for classification on the training data.

(5) Evaluate the algorithm of Xgboost classifier on the test data. Report the accuracy, confusion matrix. Plot the ROC curve.

(6) Could you find better choices of parameters using GridSearchCV?

Problem 2 (20 points): Analyzing Boston housing data with Lasso

(Hint: You may find useful Python commands from regression2.ipynt and Xgboostexamplea.ipynt)

Analyze the Boston housing data set in the Xgboostexamplea.ipynt example.

(1) Create a training data set and a test data set such that the test data set contains randomly 20% of the data set. Set the random seed to be 40. Implement Lasso regression with tuning parameter set to be 1 on the training data. Report the estimated regression coefficients.
(2) Evaluate the above fitted Lasso model on the test data and report the RMSE.
(3) Select the tuning parameter for Lasso using a 10-fold cross-validation. What is your selected tuning parameter value?
(4) Use the chosen tuning parameter to fit the final model on the training data. Report the estimated coefficients. Evaluate the final model on the test data and report the RMSE.