COMP 237 - Online lab assignment “Linear Regression”

Due date: End of week #6

Purpose:
The purpose of this Lab assignment is to:

1. To get hands-on experience of applying supervised machine Learning namely the Linear regression algorithm to solve a business problem.
2. To get hands-on experience with data exploration and pre-processing for machine learning problems.

Pre-requisite to carrying out the assignment:

1. Download from the course shell the following comma separated file: Ecom Expense.csv. This file contains information of a typical commerce website users activities.

   A brief description of the column names of the dataset is, as follows:

   - Transaction ID: Transaction ID for the transaction
   - Age: Age of the customer
   - Items: Number of items in the shopping cart (purchased)
   - Monthly Income: Monthly disposable income of the customer
   - Transaction Time: Total time spent on the website during the transaction
   - Record: How many times the customer has shopped with the website in the past
   - Gender: Gender of the customer
   - City Tier: Tier of the city (3 options)
   - Total Spend: Total amount spent in the transaction

2. Go through and watch all “Linear & Logistic” lecture and lab tutorials related to modules # 5 & 6 to understand the concepts and the presented code.

General Instructions:
Be sure to read the following general instructions carefully:

1. This assignment must be completed individually by all the students.
2. Only provide the requested screenshots and make sure to have a complete screenshot, partial screenshots will not earn any marks.
3. You will have to provide a **demonstration video for your solution** and upload the video together with the solution on **eCentennial** through the assignment link. See the **video recording instructions** at the end of this document.
4. In your 5-minute demonstration video you should explain your solution clearly, going over the main code blocks and the purpose of each module/class/method also demoing the execution of exercises #1 & 2. Youtube links and links to google drive or any other media are not acceptable, the actual recording must be submitted.
5. Any submission without an accompanying video will lose 70% of the grade.

6. In your analysis report make sure you provide an introduction and clearly state the facts and findings. Any submission missing Analysis report will lose 70%.

Submission:

There are three elements to be submitted for this assignment in one zipped folder (All subject to grading as per the rubric for this assignment):

1. For each exercise that requires code, please create a project folder and include all project python scripts/modules and screenshot of output, as needed. Name all python scripts your firstname_linear.py. Name the folder "Exercise#X_firstname", where X is the exercise number and firstname is your first name. (In total 2 folders for this assignment).
2. For all questions that require written or graphic response create one "Word document" and indicate the exercise number and then state your response. Name the document "Written_responses_firstname", where firstname is your firstname. (In total one word or pdf document).
3. All submissions need to be accompanied with a recorded demonstration video not to exceed 5 minutes in length

Create one zipped folder containing all of the above, name it Linear_firstname where firstname is your firstname.

Assignment - exercises:

1. **Exercise #1 :** Sampling and noise (20%)

   Requirements:

   a. Using numpy sample 100 numbers from a uniform distribution and store it into variable x. Make sure that the sample contains numbers greater than zero and numbers less than zero within the uniform distribution interval. For more info checkout: https://numpy.org/doc/stable/reference/random/generated/numpy.random.uniform.html
   b. Set the seed to be the last two digits of your student number.
   c. Generate y data using x data you generated in point above according to the following relationship y = 12x -4.
   d. Using matplotlib generate a scatter plot of x and y. Set alpha the transparency to 0.5. Make sure to give the plot an appropriate title and label the axis's. For more info checkout: https://matplotlib.org/3.3.2/api/_as_gen/matplotlib.pyplot.scatter.html#matplotlib.pyplot.scatter
   e. Add(injecting noise) to the y data using the from the normal (Gaussian) distribution (i.e. y=12x-4 + noise). For more information checkout: https://numpy.org/doc/stable/reference/random/generated/numpy.random.normal.html
   f. Reproduce the plot in point d above and change the title.
   g. Compare the two plots and point out the difference in your analysis report, try to explain why.

2. **Exercise # 2:** commerce website predictions (80%)

   Requirements:

   a. Get the data :
      i. Load the "Ecom Expense.csv" data into a data frame, name the dataframe ecom_exp_firstname , where firstname is your firstname.
   b. Initial Exploration:
      i. Display (print) the first 3 records.
      ii. Display (print) the shape of the dataframe.
      iii. Display (print) the column names.
      iv. Display (print) the types of columns.
      v. Display (print) the missing values per column. (You will have to write code to illustrate the column name and the number of missing values per column. If there are no missing values print 0.
   c. Data transformation:
      i. Using "Get dummies" transform all the categorical variables in your dataframe into numeric values.
      ii. Attach the newly created variables to your dataframe and drop the original columns.
      iii. Remove the original categorical variables columns. Use pandas drop method and select the correct argument values. For more info checkout : https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop.html
      iv. In the same manner drop the Transaction ID column.
      v. Write a function that accepts a dataframe as an argument and normalizes all the data points in the dataframe. Use pandas .min() and .max().

         Below the formula for normalization:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

      vi. Call the new function and pass as an argument your transformed dataframe. By now all your data is numeric ☺
      vii. Display (print) the first two records.
      viii. Use pandas.hist to generate a plot showing all the variables histograms. Set the figure size to 9 inches by 10 inches. For more info, checkout : https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.hist.html
      ix. Use pandas.plotting.scattermatrix to generate a plot illustrating the relationships between : 'Age ','Monthly Income','Transaction Time','Total Spend'. Set alpha to 0.4, figure size to 13 inches by 15 inches .For more info checkout:

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.plotting.scatter_matrix.html

d. Build a model
   i. Assume a linear relationship between the output variable (label or target) Total Spend and the predictor variables (features): Monthly Income, Transaction Time, and both sets of dummy variables you created earlier in point "c" above.
   ii. Use "train test split" from sklearn to split your data into 65 % for training and 35% for testing.
   iii. Set the seed to be the last two digits of your student number.
   iv. Store the training data in a dataframe named as follows: x_train_firstname for the features (predictors) and the training labels y_train_firstname. Store the test data as follows: x_test_firstname and y_test_firstname.
   v. Using sklearn fit a linear regression model to the training data.
   vi. Display (print) the coefficients (i.e. the weights of the model).
   vii. Display (print) the model score (i.e. the R^2 of the model).
   viii. Repeat the steps from (i) but add the feature 'Record' to the list of predictors (features).
   ix. Display (print) the coefficients (i.e. the weight of the model).
   x. Display (print) the model score (i.e. the $R^2$of the model).
   xi. Compare the two models results and write a thorough analysis in your analysis report.

Write in your own words an analysis of each of the steps you carried out and the results you obtained in the "Written_responses_firstname" document. Be through and make sure not to miss anything.

------ End of Exercises ------

Rubric

Evaluation criteria	**Not acceptable**	**Below Average**	**Average**	**Competent**	**Excellent**
	0% - 24%	**25%-49%**	**50-69%**	**70%-83%**	**84%-100%**
Requirements in execises	Missing all requirements required	Some requirements are implemented.	Majority of requirements are implemented but some are malfunctioning.	Majority of requirements implemented.	All requirements are implemented Correctly.
Code Documentation	No comments explaining code.	Minor comments are implemented.	Some code is correctly commented.	Majority of code is correctly commented.	All code is correctly commented.
Design	No adherence to object design principles.	Minor adherence to object design principles.	Some object oriented and modulus design principles are adhered to.	Majority of Object oriented and modulus design principles are adhered to.	Object oriented and modulus design principles are adhered to.
Testing & Evaluation	No evidence of testing and evaluation of the requirements.	Minor evaluation and testing efforts.	Some of the requirements have been tested & evaluated.	Majority of requirements are tested & evaluated.	Realistic evaluation and testing, comparing the solution to the requirements.
Written analysis content	Missed all the key ideas; very shallow.	Shows some thinking and reasoning but most ideas are underdeveloped.	Indicates thinking and reasoning applied with original thought on a few ideas.	Indicates original thinking and develops ideas with sufficient and firm evidence.	Indicates synthesis of ideas, in-depth analysis and evidences original thought and support for the topic.
Written analysis report format and organization	Writing lacks logical organization. It shows no coherence and ideas lack unity. Serious errors. No transitions. Format is very messy.	Writing lacks logical organization. It shows some coherence but ideas lack unity. Serious errors. Format needs attention, some major errors.	Writing is coherent and logically organized. Some points remain misplaced. Format is neat but has some assembly errors.	Writing is coherent and logically organized with transitions used between ideas and paragraphs to create coherence. Overall unity of ideas is present. Format is neat and correctly assembled.	Writing shows high degree of attention to logic and reasoning of all points. Unity clearly leads the reader to the conclusion. Format is neat and correctly assembled with professional look.
Demonstration Video	Very weak no mention of the code changes. Execution of code not demonstrated.	Some parts of the code changes presented. Execution of code partially demonstrated.	All code changes presented but without explanation why. Code demonstrated.	All code changes presented with explanation, exceeding time limit. Code demonstrated.	A comprehensive view of all code changes presented with explanation, within time limit. Code demonstrated.

Demonstration Video Recording

Please record a short video (max 4-5 minutes) to explain/demonstrate your assignment solution. You may use the Windows 10 Game bar to do the recording:

1. Press the Windows key + G at the same time to open the Game Bar dialog.
2. Check the "Yes, this is a game" checkbox to load the Game Bar.
3. Click on the Start Recording button (or Win + Alt + R) to begin capturing the video.
4. Stop the recording by clicking on the red recording bar that will be on the top right of the program window.

(If it disappears on you, press Win + G again to bring the Game Bar back.)

You'll find your recorded video (MP4 file), under the Videos folder in a subfolder called Captures.

Or you can use any other video recording package freely available.