

School of Computer Science Engineering and Technology

Course- BTech
Course Code- CSEL301
Year- 2022
Date- 05-09-2022

Type- Core
Course Name-AIML
Semester- Odd
Batch- 5th Sem

Lab Assignment No. 3.1_1

Exp. No.	Name	CO-1	CO-2	CO-3
3.1_1	Simple Linear regression	✓	✓	

Objective: To implement Simple Linear regression model using scikit-learn library.

About Dataset: The market historical data set of real estate valuation are collected from Sindian Dist., New Taipei City, Taiwan.

Data Set Characteristics:	Multivariate	Number of Instances:	414	Area:	Business
Attribute Characteristics:	Integer, Real	Number of Attributes:	7	Date Donated	2018-08-18
Associated Tasks:	Regression	Missing Values?	N/A	Number of Web Hits:	158853

Attribute Information:

X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)

X2=the house age (unit: year)

X3=the distance to the nearest MRT station (unit: meter)

X4=the number of convenience stores in the living circle on foot (integer)

X5=the geographic coordinate, latitude. (unit: degree)

X6=the geographic coordinate, longitude. (unit: degree)

Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Download the dataset available on:

(<https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set>)

1. Load dataset into Pandas Data Frame (5)
2. Display the first 5 rows (5)
3. Remove the columns X2 "The house age" (5)
4. Check whether data contains missing value or not. if require, pre-process the data. (5)
5. Read and store the features "X2=the house age" of data in X and output variable in Y "house price of unit area". (5)

6. Split the dataset into train and test in the following ratio (Hint: Use `train_test_split` class, use Splitting ratio 80:20) (10)
7. Create Linear Regression Models on the splitting criterion as mentioned above (Hint: Use `sklearn.linear_model.LinearRegression` class) (10)
8. Find out the linear regression coefficients (i.e., m and c) (10)
9. Write the equation of SLR (10)
10. Perform the prediction on the test dataset. (10)
11. Check the performance of the model on test dataset by Calculating the 'Mean Squared Error' (MSE) and R2-Score (Hint: `sklearn.metrics.mean_squared_error` function) (10)
12. Plot the regression line for test dataset (i.e., Y_{pred} vs Y_{actual}) (10)
(Hint: Use scatter plot and line plot of Matplotlib Library)
13. Train the model against different features like X_3 vs Y , X_4 vs Y , X_5 vs Y and X_6 vs Y and Identify the most desirable feature for the dependent variable Y (20)

Suggested Platform: Jupyter Notebook/Google Colab Notebook

Packages: `numPy`, `Pandas`, `sklearn`, `matplotlib.pyplot`