ADIA Summercamp SUAD/SCAI

Hackathon : Housing Price

During the course you have studied several methods for exploratory data analysis, data visualization tools and data modelling techniques (simple linear regression).

Data is like Pandora's box that has precious mysteries to reveal when treated in the right way. It demands patient analysis before we take it to the next step.

This exercise demonstrates how we can extract valuable information from real world data namely Housing price data.

The problem demands identifying patterns related to housing price, visualization of different variables and understanding the relationship between exploratory and output variables.

Reference: HousingPrice.csv

-The output should be a Jupyter Notebook, with calculations done in Python with a presentation

Reading the data:

1.Load the dataset in Jupyter Notebook

2.Find the primary information (numbers of observations, numbers of variables, columns names, column information)

3.Check how many unique observations are there for each variable

4. Summarize the data frame statistically (mean, median...), what did you notice?

Cleaning the data:

5. Verify whether there are "NA" values and remove them if there are any.

6. Verify whether there are "NULL" values and remove them if there are any.

7. Remove the row where the price is 0

8. Remove the column that you cannot use (categorical or text ...)

Analysing the data:

10. Analyze the relationship between the variables:

-Correlation Matrix

11. Scale your data (we do not want some feature to be voted as “more important” due to scale differences. 10m = 10000mm, but the algorithm is not aware of meters and millimetres)

12. Calculate covariance matrix of the scaled data

13.Compute the eigen decomposition of the covariance Matrice (eigen values and eigen vectors)

Forecasting the data:

The variable that we plan to forecast is “Price”

14. Split the data into “Train” and “Test” data using the `train_test_split` function.

Your data should be scaled (provided you answered question 14).

If not, check the help of StandardScaler function and use it.

15. Import linear regression, and r2 score functions from the Sklearn module.

16. Fit the model to the train data, then predict the price from the test sample.

What can you conclude about linear regression’s performance?

Important Guidelines:

1. Group has to prepare one Presentation with 5-10 Slides answering all the questions asked above (maximum 10 mins).

2. Slides should include the followings (1-2 Slide(s) for each): Project Title and Student Names, Problem Statement, Data Description and Source, Answers to the Questions asked (2-4 slides), Name of the Methods used, What did you learn from this.

3. Present the project in front of a Jury followed by Question and Answer Session. Evaluation will be based on your work, presentation and answers to questions raised by the Jury.