Machine Learning – Simple Linear Regression Assignment

Name: Nour Mohamed Hussein Mahmoud Mohamed Kamaly

ID: 20191700701

Department: Scientific Computing

Section: 5

This assignment is composed of 2 parts:

1. 6 simple univariable linear regression models on preprocessed data
2. 6 simple univariable linear regression models on unprocessed data

Approaches to preprocess the data:

1. Outlier detection using interquartile range on the label (“house price of unit area “): outliers are considered noise to the data, and they don’t describe the normal distribution of the sample, by removing them, it will make the predictions better (function implemented from scratch).
2. Feature scaling on (“latitude, longitude, distance to nearest MRT station”): as these features contain values that are on a different scale that the label, I transformed them to another range (0->1) to boost my prediction and be on a scale like the labels (using MinMaxScaler from sci-kit learn).
3. Splitting the transaction date into year and others and taking only the year part by converting it to a string and then slicing it.

Models and their mean squared error (on the processed data):

Feature	Mean Squared Error
Transaction Date	162.64848775463088
House Age	153.0559919741399
Distance to nearest MRT station	82.6432129308194
Num of convenience stores	102.94708718321607
Latitude	109.46061446401903
Longitude	112.62355306955502

Conclusion: the best variable for this task is distance to nearest MRT station.

Models and their mean squared error (on the unprocessed data):

Feature	Mean Squared Error
Transaction Date	184.68931783705574
House Age	176.50047403131393
Distance to nearest MRT station	100.88574959799587
Num of convenience stores	124.47199212769486
Latitude	129.56861389100305
Longitude	134.11606939001436

Conclusion: the best variable for this task is distance to nearest MRT station.