Thierry Isaac N'kouka

Mid-Sem Report

This report will give a description of the insights obtained from the Boston house prices dataset and will explain how the following were solved:

- Clean data by identifying if there are any missing values, outliers or need for standardization.
- Carry out basic exploratory analysis by plotting bar plots, histograms, pie charts etc depending on your preference and data type.
- Split the data into train and test by using 25% of the data for testing and the remaining data for training the model.
- Evaluate the performance of your model using MSE.

First of all, the Boston house prices dataset being used in this project is made of fourteen attributes:

-CRIM per capita crime rate by town
- ZN proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS proportion of non-retail business acres per town
- CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX nitric oxides concentration (parts per 10 million)
- RM average number of rooms per dwelling
- AGE proportion of owner-occupied units built prior to 1940
- DIS weighted distances to five Boston employment centers
- RAD index of accessibility to radial highways
- TAX full-value property-tax rate per $10,000
- PTRATIO pupil-teacher ratio by town
- B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT % lower status of the population
- price

And five hundred and six instances or rows. (refer to the code for more details on the this).
The average price of houses is 22.532806, the maximum and the minimum are 50.000000 and 5.000000 respectively.

Machine learning process requires, to clean the data before building any model. The first thing we are going to do with this data set is to find null values, and as stated in the data set characteristics ": Missing Attribute Values: None", nevertheless this must still be verified. For that purpose a function called "isnull()" was used to determine the null values in the data set and as you will see, there was zero null value found in every series.
After that we need to find outliers, which are values that fall out of the range, this happens when the standard deviation is higher, for example a standard deviation equal to 3 will implicate a spread of the data in other words some data points will be out of the actual range.

For that we need to find the Z_score of every value of every series to determine outliers. Remember, the higher the std deviation, the higher the Z_score of values will be and the more the data points will be spread out. After detecting and dropping the row with outliers, we end up with 415 rows.
Now that our data set has been cleaned, let's look at least one of the attributes CRIM :
The Per Capita Crime Rate By Town, is very low in the first three hundred towns, in fact it varie s between 0 and 5; and of all the towns in the data set, these would be the best place to live as saf ety is one of the most important factors a tenant looks for.

In the end we have split the dataset into train and test by using 25% of the data for testing and the remaining data for training the model.
After model evaluation, here are the coefficients obtained:

	Coefficient
CRIM	-1.795714e-01
ZN	2.871761e-03
INDUS	5.283073e-02
CHAS	1.198069e-13
NOX	-1.182039e+01
RM	5.252177e+00
AGE	-2.618816e-02
DIS	-1.237648e+00
RAD	2.660713e-01
TAX	-1.366262e-02
PTRATIO	-9.455442e-01
B	4.553445e-03
LSTAT	-4.533907e-01

After the model built, the following mean scared error was obtained: 16.785950616140024