Step 1:

Loading BOSTON dataset.

- LON and LAT are the longitude and latitude of the center of the census tract.

- MEDV is the median value of owner-occupied homes, measured in thousands of dollars.

- CRIM is the per capita crime rate.

- ZN is related to how much of the land is zoned for large residential properties.

- INDUS is the proportion of the area used for industry.

- CHAS is 1 if a census tract is next to the Charles River else 0

- NOX is the concentration of nitrous oxides in the air, a measure of air pollution.

- RM is the average number of rooms per dwelling.

- AGE is the proportion of owner-occupied units built before 1940.

- DIS is a measure of how far the tract is from centres of employment in Boston.

- RAD is a measure of closeness to important highways.

- TAX is the property tax per $10,000 of value.

- PTRATIO is the pupil to teacher ratio by town

Step 2:

Basic EDA.

Step 3:

Stats with output/target MEDV variable.

Using LON and LAT for plotting Geo-map

Step 4: Outliers.

Robustness of Outliers

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| RMSE Value | | Outliers | | | | | | Without Outliers | | | | | |
| Seed(35) | | Seed(45) | | Seed(52) | | Seed(35) | | Seed(45) | | Seed(52) | |
| Train | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test |
| Size Split | 70:30 | 4.65 | 5.27 | 4.52 | 6.12 | 4.86 | 3.99 | 4.35 | 5.63 | 4.57 | 5.64 | 4.63 | 4.96 |
| 75:25 | 4.70 | 4.97 | 4.29 | 6.64 | 4.79 | 4.08 | 4.27 | 6.03 | 4.68 | 3.84 | 5.14 | 4.38 |
| 80:20 | 4.60 | 4.30 | 4.42 | 5.98 | 4.68 | 4.34 | 4.21 | 5.37 | 4.39 | 5.15 | 4.51 | 5.48 |

Step 5: Plotting and modelling requirements.

|  |  |
| --- | --- |
|  |  |

Step 6: Conclusion.

We can observe CART model is better than Linear Regression model both by visually and by RMSE.