Boston House Price Prediction 10.11.2022

Bibekananda Sahoo
Intern at Skillvertex
Minor Project Group 8
September Batch 2022

Overview

In this minor project, we are going to do implementing a salable model for predicting the house price prediction using some of the regression techniques based on some of the features in the dataset which is called Boston House Price Prediction.

About the Datasets

This dataset contains 13 columns including our target column. Below are the complete description of all those columns:-

- CRIM - per capita crime rate by town
- ZN - the proportion of residential land zoned for lots over 25,000 sq. ft.
- INDUS - the 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 - the average number of rooms per dwelling
- AGE - the 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 dollar
- 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
- MEDV - Median value of owner-occupied homes in $1000's

Where the column “MEDV” is our target column. We have a total of 506 rows.

- All the attributes are having Integer/float data type
- But the attribute “CHAS” is actually a dummy attribute so while doing EDA and Model building we have to consider this column as a dummy column.

Data Cleaning

- We do a null value count for all the attributes and there is no null value present in the dataset.
- Also, check the Data type of all columns and that all those attributes are in the right data type.

Exploratory Data Analysis

1. **Some attributes which seem linearly correlated to the target column MEDV are**
- LSTAT - % lower status of the population
- RM - the average number of rooms per dwelling
2. **Where RM is positively correlated and LSTAT is negatively correlated**
3. **Attribute 'Chas' is highly imbalanced**
- The mean house price is a little high in the case of Charles River
4. **Some highly correlated attributes are**
- Attribute 'RAD' and 'TAX' has positively correlated with a value of 0.91
- Attribute 'RAD' and 'CRIM' are also positively correlated with a value of 0.81
- we have to drop the "RAD" column due to the multicollinearity issue
5. **And the dependent variable which is 'MEDV' is positively correlated with 'RM' with a value of 0.70 and negatively correlated with 'ISTAT' with a value of -0.73 its also negatively correlated with "CRIM", "INDUS", "NOX", "AGE", "RAD", "TAX" and "PTRATIO"**

Hyper-Parameter Tuning For choosing the Best Model

These are the regression model which we use for Model Building

- **Linear Regression**
- **Random Forest Regressor**
- **KNeighborsRegressor**
- **Support Vector Machines**
- **Decision Tree**
- **XGBRegressor**

	Model	Best_Score	Best_Parameter
3	RandomForest	0.853203	{'criterion': 'squared_error', 'n_estimators':...
5	XGBRegressor	0.853151	{}
2	SVM	0.807856	{'C': 1.0, 'gamma': 'scale', 'kernel': 'rbf'}
4	KNN	0.758585	{}
1	Decision_Tree	0.693277	{'criterion': 'squared_error'}
0	Linear_Regression	0.658515	{}

- **As we can clearly interpret that Random Forest and XGBRegressor are having highest R-squared**
- **So we go with the Random Forest algorithm for our prediction.**

Finding Best attributes

- **Using RFE(Recursive Feature Elimination) for finding the top 5 features**
- **These are the top 5 features according to RFE**
- **“CHAS”, “RM”, “DIS”, “PTRATIO”, “ISTAT”**

Evaluation Parameter

- These are the model parameters after we run the model in train data
    1. **R-Squared: 0.9774547904402869**
    2. **Adjusted R^2: 0.9771251821133905**
    3. **MSE: 0.022545209559713113**
    4. **RMSE: 0.1501506229081755**
- We can clearly see that the difference between R-Squared and Adjusted R-Squared was so low which means there are no model complexity issues.
- There may be some multicollinearity issues present so we also do a VIF check

	Features	VIF
4	lstat	1.97
1	rm	1.66
2	dis	1.35
3	ptratio	1.19
0	chas	1.03

- **We can clearly see that the VIF value is less than 2 for all the features so which means there are no multicollinearity issues.**

Residual analysis and validating the assumptions

- **Error terms are normally distributed with a mean of approximately zero.**
- **By observing the Q-Q plot we can clearly see that most of the points are in the line so residuals are normally distributed.**
- **Error terms have constant variance because the points are randomly scattered; there is no such pattern.**

FINAL Inference

- **The top 5 crucial features are**
- **CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)**
- **RM - the average number of rooms per dwelling**
- **DIS - weighted distances to five Boston employment centers**
- **PTRATIO - pupil-teacher ratio by town**
- **LSTAT - % lower status of the population**

- **where PTRATIO and LSTAT are negatively impacting the house price and CHAS, RM, and DIS are positively impacting.**