

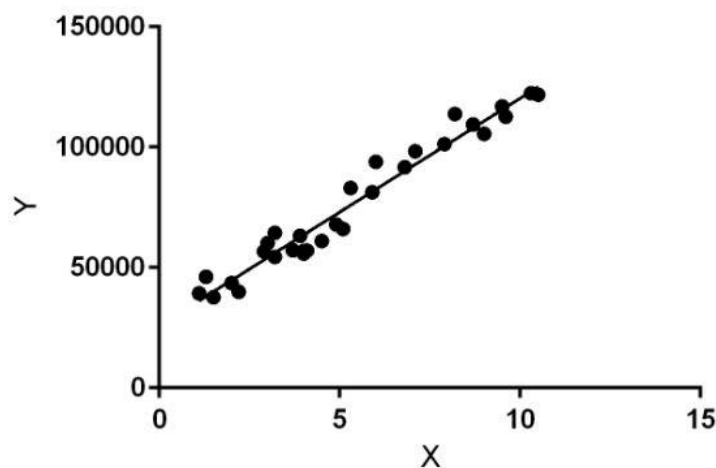
Experiment No. 1
Analyze the Boston Housing dataset and apply appropriate Regression Technique
Date of Performance:
Date of Submission:

Aim: Analyze the Boston Housing dataset and apply appropriate Regression Technique.

Objective: Ability to perform various feature engineering tasks, apply linear regression on the given dataset and minimise the error.

Theory:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Dataset:

The Boston Housing Dataset

The Boston Housing Dataset is derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA. The following describes the dataset columns:

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 centres

RAD - index of accessibility to radial highways

TAX - full-value property-tax rate per \$10,000

PTRATIO - pupil-teacher ratio by town

B - $1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town

LSTAT - % lower status of the population

MEDV - Median value of owner-occupied homes in \$1000's

Code:

Conclusion:

1. What are features have been chosen to develop the model? Justify the features chosen to estimate the price of a house.
 - A linear regression model was developed to estimate house prices using the 'LSTAT' (percentage of the lower status of the population) feature. This choice is justified by the intuition that areas with a higher percentage of lower-status individuals might have lowerpriced housing due to economic disparities.
 - By focusing on the 'LSTAT' feature alone, the model simplifies the prediction process, which is important for illustration purposes. However, it's important to note that real-world housing price prediction models often consider multiple features like crime rate ('CRIM'), average number of rooms ('RM'), and accessibility to employment centers ('DIS') among others, as these collectively offer a more comprehensive understanding of housing dynamics.
 - While this simplified model captures a basic trend, a more robust model would involve incorporating a broader set of features to enhance predictive accuracy

2. Comment on the Mean Squared Error calculated.

- The Mean Squared Error (MSE) evaluates linear regression's performance by measuring the squared difference between actual and predicted house prices.
- Calculated MSE values for training and testing datasets reveal model accuracy and generalizability. Comparing MSE helps detect overfitting or underfitting. Reported MSE values serve as a reference for assessing predictive performance, specifically the 'LSTAT' feature's effectiveness.
- Consider data scale and characteristics when interpreting lower MSE values. Compare with complex models or algorithms for a deeper understanding of predictive accuracy