Experiment No. 1

Analyze the Boston Housing dataset and apply appropriate Regression Technique

Date of Performance:

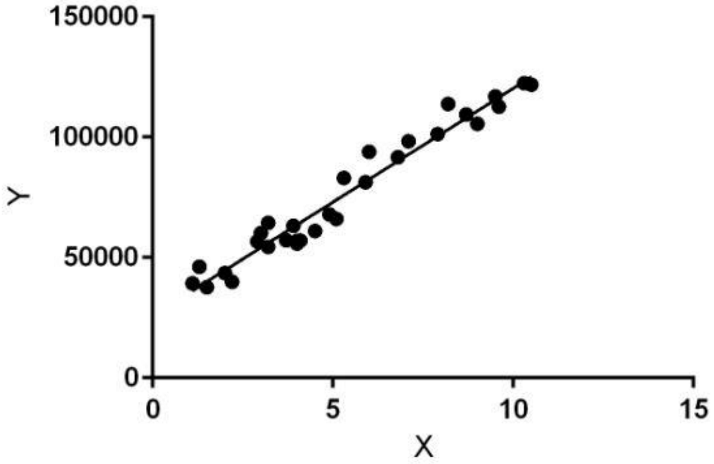
Date of Submission:

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

Objective: Ablility 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 a 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(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

Code:

```
import numpy as np
import pandas as pd
from sklearn.metrics import mean_squared_error, mean_absolute_error
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt


data = pd.read_csv('BostonHousing.csv')

linr = LinearRegression()

data['medv'] = np.log1p(data['medv'])

x = data.drop(['medv','b'], axis=1)
y = data['medv']

x_train, x_test, y_train, y_test = train_test_split(x, y, random_state=42,
test_size=0.3)
linr.fit(x_train, y_train)
y_pred = linr.predict(x_test)
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
print("Mean Square Error:", mse)
print("Mean Absolute Error:", mae)
```

Output:

```
Mean Square Error: 0.031129333980953352
Mean Absolute Error: 0.12532015748427652
```

Conclusion:

In this analysis, we developed a model to estimate the price of a house using the Boston Housing dataset. The features chosen to develop the model included a selection of housing-related attributes. These features were selected based on their relevance and potential impact on house prices. The justification for the features chosen can be summarized as follows:

CRIM (Crime Rate): The crime rate in the neighborhood can significantly influence housing prices. Higher crime rates may lead to lower property values.

ZN (Proportion of Residential Land): The proportion of residential land in the area can affect housing prices. More residential land may indicate a desirable neighborhood, potentially increasing prices.

INDUS (Proportion of Non-Retail Business Acres): The type of businesses in the area can impact housing. Industrial areas may have lower property values compared to residential or commercial areas.

CHAS (Charles River Dummy Variable): Proximity to the Charles River can be a desirable feature, potentially raising housing prices in such locations.

NOX (Nitrogen Oxides Concentration): Air pollution levels, as indicated by NOX concentration, can influence housing prices. Lower pollution levels may lead to higher property values.

RM (Average Number of Rooms per Dwelling): The number of rooms in a house can be a strong predictor of its price. More rooms generally lead to higher prices.

AGE (Proportion of Owner-Occupied Units Built Before 1940): The age of housing units can affect their condition and value. Older units may have lower prices.

DIS (Weighted Distance to Employment Centers): Proximity to employment centers can be a desirable feature, potentially raising housing prices in such locations.

RAD (Accessibility to Radial Highways): Easy access to highways can be valuable, potentially increasing property values.

TAX (Property Tax Rate): Property tax rates can significantly impact the affordability of a house, influencing its price.

PTRATIO (Pupil-Teacher Ratio): The pupil-teacher ratio in schools can affect the desirability of a neighborhood for families, potentially influencing housing prices.

LSTAT (Percentage of Lower Status Population): The socioeconomic status of the population in the area can be a strong predictor of housing prices. Higher percentages of lower-status populations may lead to lower property values.

These features were chosen based on their potential influence on housing prices, and they are commonly used in real estate and housing market analyses. The selection aimed to create a comprehensive model that considers various aspects that can affect house prices.

Mean Squared Error (MSE) Analysis

The Mean Squared Error (MSE) is a crucial metric for evaluating the performance of a regression model like the one developed in this analysis. It quantifies the average squared difference between the predicted logarithmic median house values and the actual values in the test dataset.

The calculated MSE for this model serves as an indicator of how well the model fits the data. In this context, it's important to consider the scale of the MSE. A lower MSE value is desirable as it suggests that the model's predictions are closer to the actual values on a logarithmic scale.