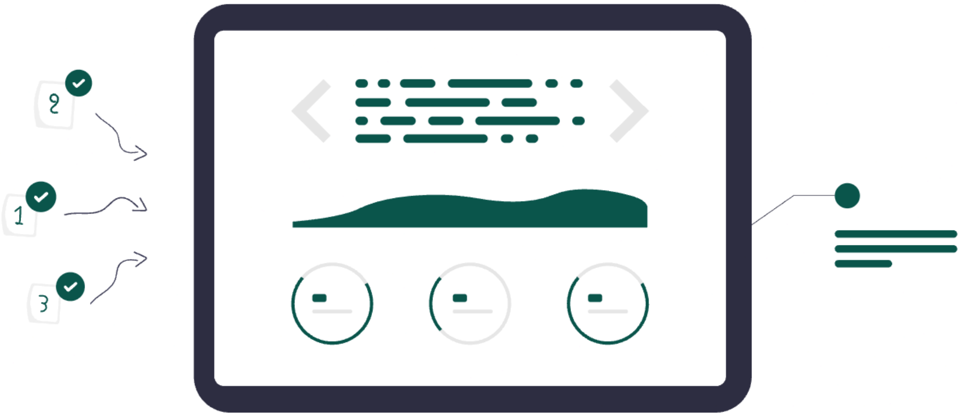
Assignment of CSE303

Exploratory Data Analysis on Boston Housing Dataset



Submitted By: Team 4

Aklhak Hossain 2022-3-60-057 https://akhlak.dev

Submitted To:

Dipayan Bhadra
Adjunct Faculty
Department of CSE
East West University

Overview

The Boston Housing dataset consists of data collected by the U.S. Census Service on housing in the Boston metropolitan area. It involves a number of features describing residential properties, socioeconomic indicators, and environmental variables within different neighborhoods.

The dataset is typically used to examine influences on housing prices.

There is one row for each neighborhood or housing district. The features have both numerical and categorical data, and these include:

- Crime rate per town
- Proportion of residential area zoned for large lots
- Mean number of rooms per unit
- Pupil-teacher ratio by town
- Access to highways
- Property tax rate
- Median home value (MEDV), the target variable

The data can be utilized for exploratory analysis, prediction modeling, and studying the association between house prices and location variables.

Data Preparation and Preprocessing

The data contains 14 columns and 506 rows. Most of the columns are numerical (float64 or int64), with the exception of the black column, which is object type and contains missing values.

- Missing values in the columns indus, nox, dis, ptratio, and black were found.
- Numerical missing values were imputed by mean method.
- The black categorical column was imputed with the most frequent value, followed by one-hot encoding.
- All features were standardized using StandardScaler for scaling values.
- Outliers were identified using Z-score technique and removed by removing data points with Z-score above 3.
- Final cleaned and normalized data was split into training (70%) and testing (30%) datasets for model training and evaluation

Machine Learning Models Applied

Four machine learning algorithms were implemented to predict the median house price (medv):

Linear Regression:

A method in statistics that models the relationship between the dependent variable and one or more independent variables by fitting a linear equation. It assumes a straight-line relationship and minimizes the sum of the squared deviations between predicted and actual values.

Decision Tree Regressor

A decision tree model that splits data into subsets based on feature values. It creates branches that constitute decision rules to predict continuous outputs. It can handle non-linear relationships but may overfit training data when not regularized.

K-Nearest Neighbors (KNN) Regressor

A form of instance-based learning that predicts the target value of a data point by taking the average of the target values of its nearest neighbors in the feature space. It is simple and effective but sensitive to the choice of neighbors and feature scaling.

Random Forest Regressor

An ensemble learning method that builds multiple decision trees from random data and feature subsets. It averages their predictions to improve generalization and avoid overfitting, typically achieving greater accuracy than an individual decision tree.

Model Training and Evaluation

All four models were trained on the training data and tested on the unseen test data. The Root Mean Squared Error (RMSE) measure was applied to assess the models, which is the average magnitude of the error in predictions.

The RMSE values I got were:

Model	**RMSE**
Linear Regression	5.3496
Decision Tree	3.5093
KNN (k=5)	5.3163
Random Forest	3.2425

The smaller the RMSE, the more accurate the model's prediction. Among all the models, Random Forest had the minimum RMSE and thus had the most accurate prediction of house prices.

Discussion on Suitable Attributes

To identify most influential features in house price estimation, correlation analysis was performed using heatmaps and pairplots.

Major findings:

- **Strong positive correlation** observed between rooms (`rm`) and house price (`medv`). Higher number of rooms generally indicate greater value.
- **Negative correlation** between percentage of population in lower status (`lstat`) and price, such that the areas with fewer deprived neighbors are more expensive.
- The other attributes such as crime rate (`crim`), nitrogen oxide level (`nox`), and highway accessibility (`dis`) also had a strong correlation, affecting the price.
- The categorical variable `chas` (Charles River dummy variable) did not play a strong effect after encoding.

These attributes contribute significantly to the predictive powers of the models. Other feature selection or dimension reduction techniques would further refine model performance by emphasizing the most important features.

Conclusion

This project employed four machine learning models for Boston house price prediction. Preprocessed data with care, imputation of missing values, scaling, and outlier removal preceded model training and evaluation on RMSE.

Random Forest gave the best prediction performance with the minimum RMSE (3.2425). Decision Tree gave good performance, and Linear Regression and KNN performed with higher errors.

The test showed that factors like socioeconomic status and number of rooms significantly affect the cost of homes. Proper data cleaning and feature scaling were crucial to the success of the model.

Hyperparameter tuning, feature engineering, and additional data sources can further improve accuracy in future projects.

Reference Links

Github Url:
https://github.com/Akhlak-Hossain-Jim/Learning-CSE-at-EWU/blob/main/Semester-8/CSE303/Lab/Section%206/Assignment%202