Boston Pricing Prediction

Cyrus Benjamin C. Canape
College of Information and Computing Science
University of Santo Tomas
Manila, Philippines
cyrusbenjamin.canape.cics@ust.edu.ph

***Abstract*— This report is a predictive analysis of the median value of owner-occupied homes in Boston using the Boston house-price dataset.**

I. Introduction

This paper is based on a previous study conducted by Harrison, D. and Rubinfield, D.L., *"Hedonic housing prices and the demand for clean air"* which explored the relationship between housing prices and various socio-economic and environmental factors. According to the Harrison and Rubinfield, house prices are dependent on various factors such as crime rates, number of rooms, property taxes rates, and other socio-economic and environmental conditions

The Boston house-price dataset, a widely used dataset for regression analysis containing 13 independent variables that influences property values will be used. The paper aims to analyze these relationships to enable individuals better market valuation and data-driven decision-making, and to find the median value of owner-occupied home in $1000's (MEDV).

Linear models and its variants such as ridge regression, lasso, regression, and elastic net regression is used for the analysis. Models are evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score to determine the predictive effectiveness. Correlation heatmaps, feature importance plots are used to aid in interpretation.

This study offers a real-life data analysis application in real estate. Providing better market valuation and data-driven decision-making for homebuyers and realtors.

II. Methodology

A. Data Preprocessing

To improve data quality, model performance, and reliability, the data will be preprocessed using the following steps:

- **Handling Missing Values:** No missing values were found in the dataset
- **Outliers Detection:** Boxplots was used to visualize and detect the outliers. The highest outliers were (B, CRIM, ZN) but no removal was performed
- **Feature Scaling:** StandardScaler was used to normalize numerical features for scale consistency.
- **Train-Test Split:** Sklearn was used to train and test datasets

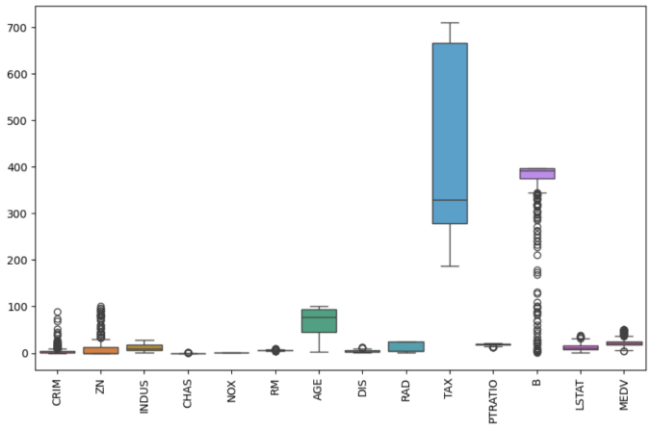

Figure II.A. Outliers Detection using Boxplots

B. Feature Selection & Engineering

- **Pair Plot:** Pairplot was used to visualize relationship between feature and target variable.
- **Correlation Matrix:** Heat map was used to visualize the correlation between feature and target variable

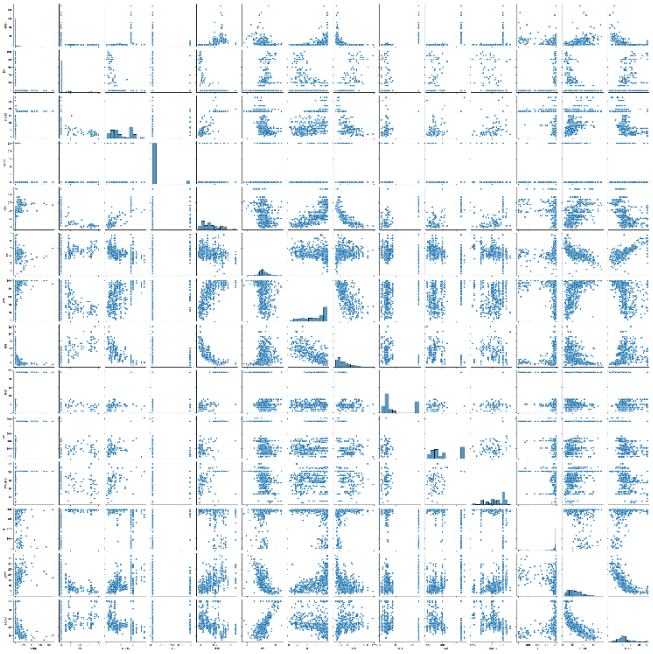

Figure II.B.1. Pair Plot

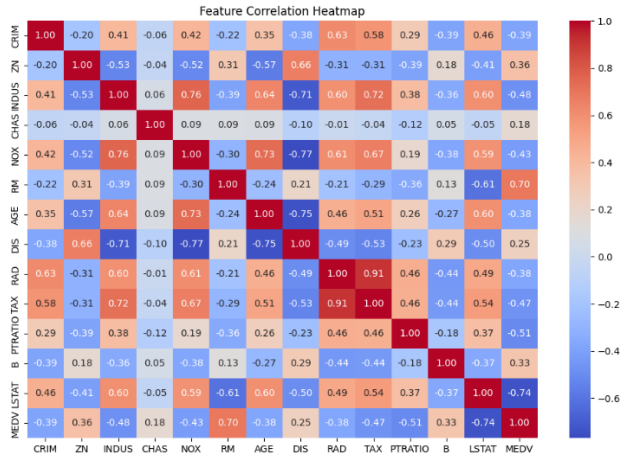

Figure II.B.2. Heatmap with all variables

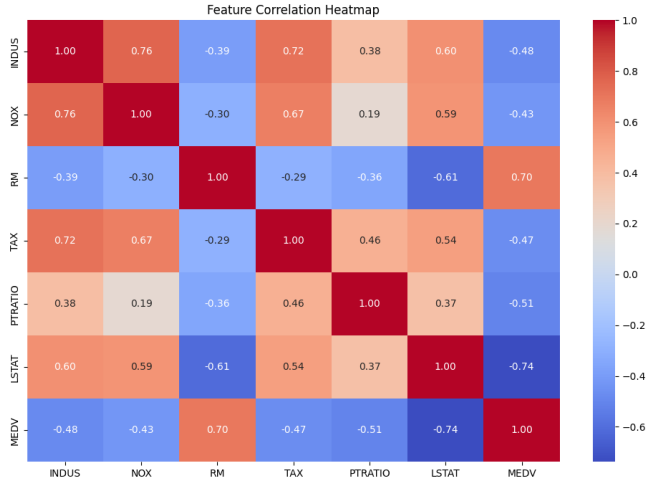

Figure II.B.3. Heatmap with above 0.4 and below -0.4 correlation

C. Model Implementation

Four regression models; linear regression, ridge regression, lasso regression, and elastic net regression were implemented:

- **Linear Regression:** Serves as the baseline model.
- **Ridge Regression:** Reduces overfitting
- **Lasso Regression:** Removes unimportant features.
- **Elastic Net Regression:** Balances feature selection and multicollinearity.

III. Experiments

The experiment was conducted by doing a hyperparameter tuning for the ridge regression, lasso regression, and elastic net regression.

A. Hyperparameter Tuning

Hyperparameter tuning is used to optimize the model parameters to improve performance and find its best settings for the machine learning model. The alpha range that was used is *"alpha": [0.1, 1.0, 10.0, 100.0],* which covers a wide range of possible regularization strength. This was applied by using the GridSearchCV and RandomSearchCV and was used for ridge regression, lasso regression, and elastic net regression.

- **GridSearchCV:** Tested all possible combination of hyperparameters and performed cross-validation by splitting the training data into 5 parts. This was used for ridge regression and elastic net regression.
- **RandomizedSearchCV:** Randomly selects hyperparameter values. This was used for lasso regression.

Hyperparameter Tuning		
Regression Model	***Best Alpha***	***Best l1 Ratio***
Ridge Regression	0.1	None
Lasso Regression	0.1	None
Elastic Net Regression	0.1	0.1

Table III.1. Hyperparameter Tuning Results

After performing GridSearchCV and RandomizedSearchCV, the best alpha for all three regression is 0.1 which causes the model to be like linear regression. This indicates that the model prefers low regularization meaning complex models work better.

IV. Results

A. Model Evaluation

Upon evaluation, the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score were calculated.

Model Evaluation			
Regression Model	***MSE***	***RMSE***	***R^2***
Linear Regression	24.291119474	4.92860218	0.668759493
Ridge Regression	24.301025500	4.92960703	0.668624412
Lasso Regression	25.155593753	5.01553524 1022056	0.656971280
Elastic Net Regression	24.439742316	4.94365677	0.666732830

Table IV.A.1. Model Evaluation

In *Figure A.2. Comparison of Regression Models*, it shows that the lasso regression has the highest MSE, followed by the elastic net regression, ridge regression, and linear regression, respectively. Based on the obtained values, linear regression exhibits the most optimal performance as it provides the lowest errors indicated by the MSE and RMSE, and a better model fit indicated by the R^2.

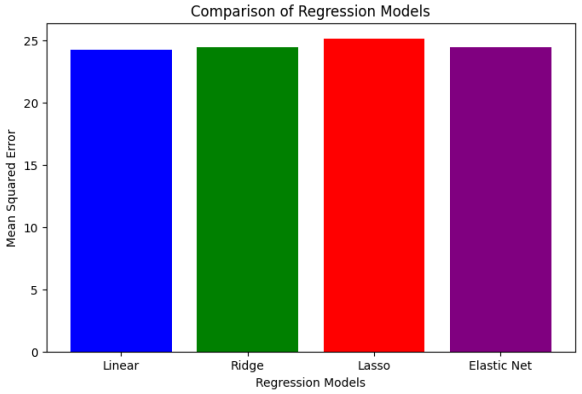

Figure IV.A.2. Comparison of Regression Models

B. *Model Plots*

In *Figure B.1. to Figure B.2.,* the regression models are shown through a plot. the x-axis as the actual price and the y-axis as the predicted price.

There is a minimal difference between the figures since the gap between the obtained MSE values are minimal as well. Upon observation, the regression model with the lowest MSE value, *Figure B.1. Linear Regression* has a noticeable difference with the regression model with the highest MSE value, *Figure B.3. Lasso Regression.*

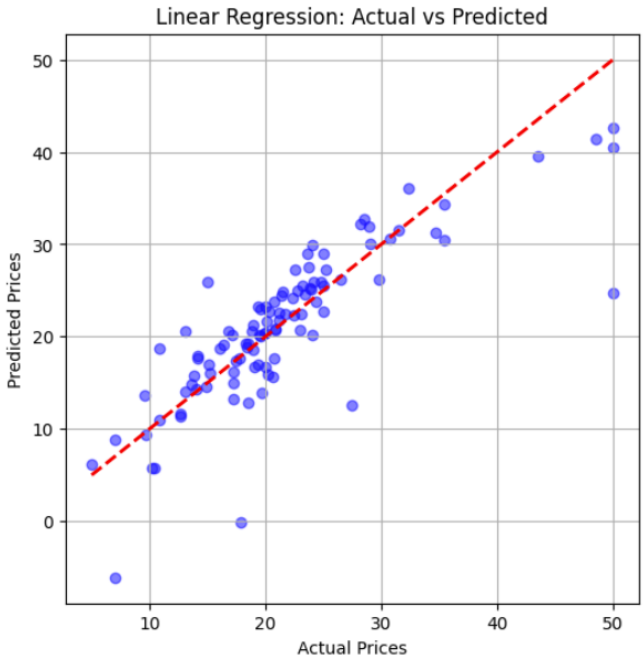

Figure IV.B.1. Linear Regression

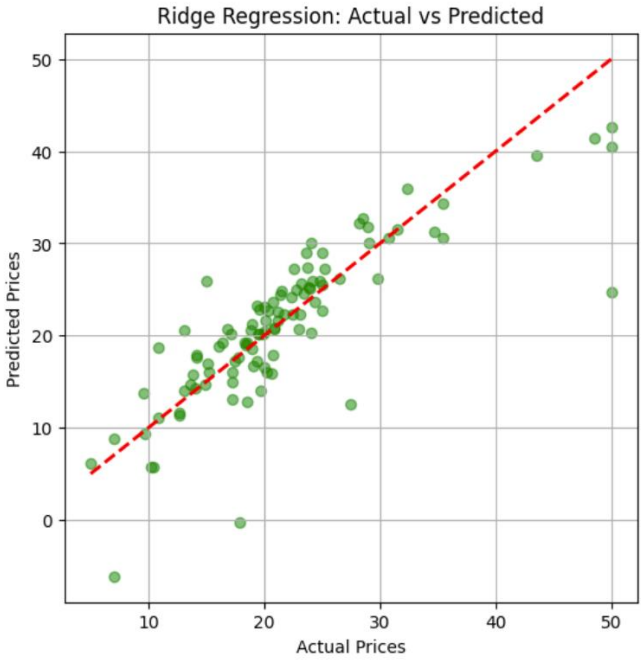

Figure IV.B.2. Ridge Regression

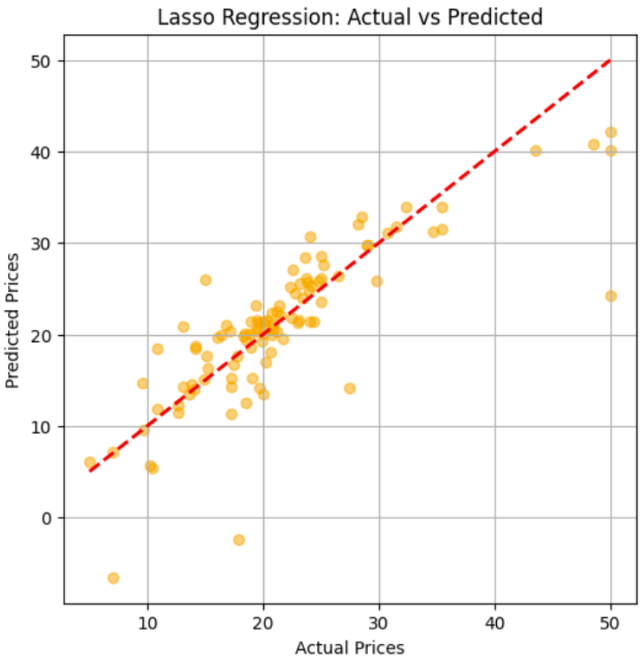

Figure IV.B.3. Lasso Regression

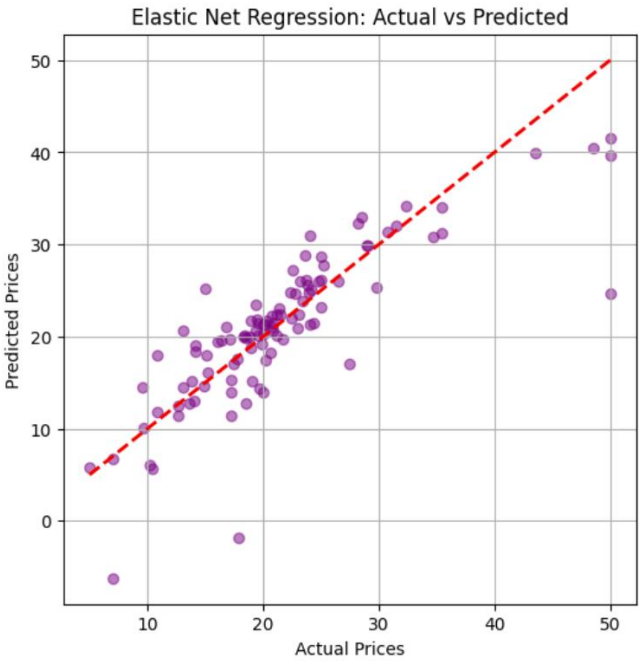

Figure IV.B.4. Elastic Net Regression

V. Conclusion

The Boston house-price dataset has gone through data preprocessing however it remained unchanged as it was already of high quality and reliability. However, after feature selection, only the variables *INDIUS, NOX, RM, TAX, PTRATIO, LSTAT, and MEDV* remained as it has the highest correlation *(above 0.4 or below -0.4)* with our target variable (MEDV).

Hyperparameter is then used to find the *alpha value* to be used for our regression model – ridge regression, lasso regression, and elastic net regression. The alpha range used was *[0.1, 1.0, 10.0, 100.0],* for a wide range of possible regularization. Upon calculation with the use of GridSearchCV and RandomSearchCV, the most suitable

alpha value for the three regression models is *0.1.* This means that all regression models retain more flexibility and complexity.

After identifying each regression model's alpha value, it is then used to compute the value of the *MSE, RMSE, and* R^2 of each model. As shown in *Table IV.A.1.,* the linear regression model, has the lowest MSE value of *24.291119474,* which is followed by the ridge regression model, lasso regression model, and elastic net regression model which has an MSE value of *24.301025500, 25.155593753,* and *24.439742316,* respectively.

To conclude, linear regression model is the best model to use for the prediction of MEDV. Based on its Root Mean Squared Error (RMSE) of 4.9286, the predicted median value of owner-occupied homes (MEDV) deviates by approximately $4928.60 from the actual values.

REFERENCES

[1] J. H. Friedman, "Hedonic housing prices and the demand for clean air," *Journal of Environmental Economics and Management*, Vol. 5, Issue 1., pp. 81-102, 1978. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/0095069678900062.

[2] "ML | Boston Housing Kaggle Challenge with Linear Regression," *GeeksforGeeks*, 2023. [Online]. Available: https://www.geeksforgeeks.org/ml-boston-housing-kaggle-challenge-with-linear-regression/.