Report Details

Code Details:

- **File Name**: R00224350-Haseeb-ML-Asg2.ipynb (attached)
- **Developed in**: Jupyter Notebook
- **Language**: Python
- **Python Interpreter**: 3.10 (package with Anaconda)
- **Platform:** Google Colab (SaaS)

Word Document Report:

File Name: R00224350-Haseeb-Ahmed-ML-Asg2.pdf

Annexures:

List of Features - Boston Housing Dataset	**-**	**Annexure A**
Jupyter Notebook Details – Boston Housing Dataset	**-**	**Annexure B**
Glossary of Terms	**-**	**Annexure C**

Boston Housing - Real Estate Price Prediction Using Machine Learning

Haseeb Ahmed, Student of MSc AI, MTU, Cork, Ireland

haseeb.ahmed@mycit.ie

Abstract

In this project, the primary focus is on predicting house prices, a crucial consideration for both prospective buyers and real estate agencies. The multifaceted nature of this prediction involves various factors such as location, amenities, and market trends. Recognizing the challenges posed by the dynamic and occasionally exaggerated nature of housing prices, the author employs an advanced automated Machine Learning model.

The model incorporates a range of regression techniques, including Simple Linear Regression, utilizing the Boston Housing Dataset. The overarching goal is to enhance the accuracy of future house price predictions. To measure model accuracy, the study employs rigorous metrics such as R-Squared, Mean Square Error (MSE), and Cross-Validation.

In addition to predictive modeling, the research delves into understanding the dataset's internal dynamics. Correlation analysis using a heat map is conducted to identify attributes that significantly impact model predictions. Furthermore, the study addresses the issue of outliers in the dataset, recognizing their potential to distort accuracy. Through meticulous outlier removal, the research aims to improve overall model performance.

Key findings reveal that Random Forest Regression outperforms other regression techniques across various measuring metrics, while Gradient Boosting Regression exhibits suboptimal performance. This comprehensive exploration not only advances the understanding of machine learning applications in real estate prediction but also provides valuable insights for stakeholders navigating the complex landscape of housing market.

1 Introduction

Effectively estimating the value of real estate, whether for buyers, sellers, is a pivotal task in the dynamic and ever-changing Real-Estate market. The intricate nature of dealing in houses poses challenges in accurate property valuation, leading to issues of over- and under-validation. Traditional methods often fall short in addressing these concerns, emphasizing the need for advanced techniques.

1.1 Choice of Dataset

In response to the complexities of property evaluation, few studies have covered Boston Housing Dataset for various purposes[1][2]. This research project employs Machine Learning (ML), to analyze the Boston Housing Dataset comprehensively. This dataset has been chosen because, it encompasses multifaceted features, including size, area, and location, as well as more relevant factors such as inflation rates and the age of the property[3]. These variables collectively influence property prices, necessitating a sophisticated approach to modeling.

1.2 Research Goal

The goal of this study is to develop a machine learning model that can accurately predict the median home value for a given neighborhood based on these factors. The study delves into the supervised learning approach within ML, where the model is trained with inputs and their desired outputs, aiming to create a general rule for mapping inputs to outputs. Furthermore, the research explores unsupervised learning, where the system identifies hidden structures in the data, and reinforcement learning, where the program performs certain goals in a dynamic environment. In this context, ML algorithms are applied to the Boston Housing Dataset, creating diverse models to predict property prices. The study emphasizes the significance of AI and ML in addressing the challenges of property valuation, offering a good perspective on the intricate interplay of various features in determining housing prices.

3 Methodology

Following sequence of steps has been adopted to carry out this project as in figure 1:-

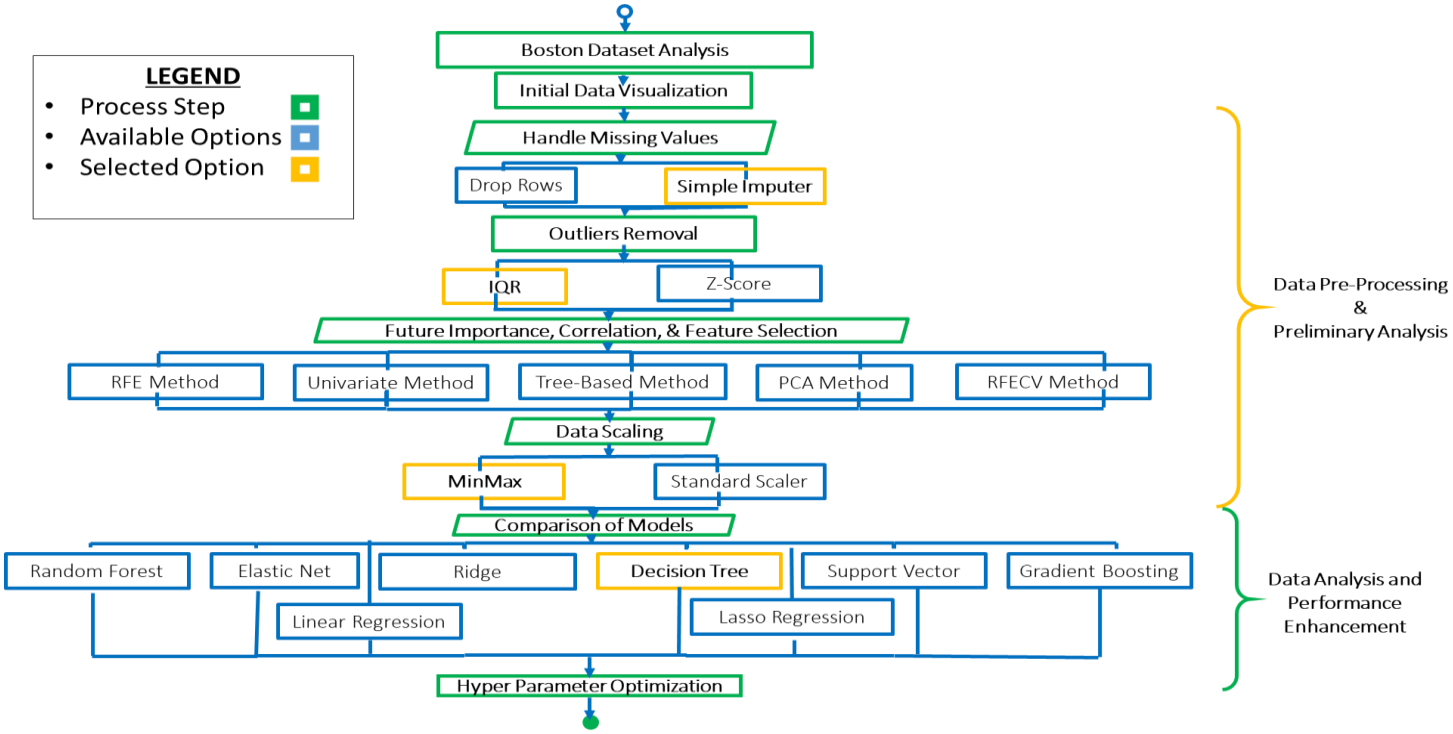

Figure 1: Boston Housing Problem - Solution Workflow

2 Data Exploration

2.1 Preliminary Dataset Analysis (Annexure A & B)

An initial EDA revealed a 'Range Index' (rows) of 511 entries (0 to 510) and 14 feature features constituting the dataset as described at Annex A. Out of 14 total features, the datatype of 11 features is 'float' and of three features (CHAS, RAD and TAX) are 'integer'. The RM feature (index 5) has lesser (506) non-zero entries which means that few entries (5) are NaN. Furthermore, no non-numeric type of feature has been observed. Owing to the small size of dataset, its memory usage is merely 56.0 KB. The dataset has been visualized in tabulated form (Jupyter Notebook (JN) Section 1.1 'Data Visualization'). Histogram plotting analysis (JN 1.2) shows a range from 0 to a little more than 700 along x-axis and 0-500 along y-axis. This graph shows that the region between 460 to 650. The distribution shows that there may be outliers in the data.

2.2 Data Pre-Processing

As in above mentioned analysis, this Boston housing dataset needed some preprocessing steps before proceeding to in-dept analysis. Therefore, a number of steps have been taken toward data pre-processing as given below:-

2.3.1 Firstly, Since, our dataset has regression problem and there are continuous values in the dataset, so, data balancing is not being carried out. (JN 2.1).

2.3.2 Since, the dataset has no non-numeric/ categorical values and a regression problem, thus, no feature encoding is required.

2.3.3 **Handling the Missing Values in Dataset**

Analysis of the dataset revealed 5 missing values in one of the descriptive features named 'RM' i.e. average number of rooms per dwelling. Two methods have been used to address this issue; In first one, all 5 rows containing null values have been removed (JN 2.3.1) resulting in reduction of row down to 506 from a total of 511. But since, data is important even in those rows having missing values. Thus, in order to maintain the sanctity of the data, another method has been used (JN 2.3.2) to fill data in such missing places. Therefore, this modified dataset has been used in further steps.

2.3.4 **Outliers Detection and Removal**

Removing outliers from a dataset can provide several benefits for data analysis and model building. Outliers are extreme values that deviate significantly from the rest of the data. They can distort the overall distribution of the data and lead to inaccurate insights and results. These outliers have been chalked out using boxplot analysis (JN 2.4.1) and were accounted for using Q1 and Q3 ranges of IQR (Interquartile Range) Method (JN 2.4.2) The model testing after IQR resulted in improvement of *Mean Square Error (MSE)* by 62.94 from 71.22 to 8.28. However, 'Z-Score' method was also tested for better analysis (JN 2.4.3) which resulted into a decrease in MSE by 50.74 i.e. from 71.22 to 20.48. Thus, the data processed through IQR was chosen for further processing in order to make the model robust and improve its performance.

2.3.5 Feature selection

Since there are several methods available for feature selection, they need to be tried for comparison purposes. Therefore, a method has been crafted for code reusability and model evaluation. It trains various models by taking X, y and model parameters and returns MSE for that model (JN 2.5.1). Following models have been pitched for this task:-

a. Univariate Feature Selection (JN 2.5.2)
b. PCA Feature Selection (JN 2.5.3)
c. RFE Feature Selection (JN 2.5.4)

d. Tree-Based Feature Selection (JN 2.5.5)

e. RFECV Feature Selection (JN 2.5.6)

Comparison has been carried out using a bar graph (JN 2.5.7) which reflects that RFECV has performed the best among all with MSE of 12.388 while the PCA has got the maximum MSE of 20.813 as shown below:-

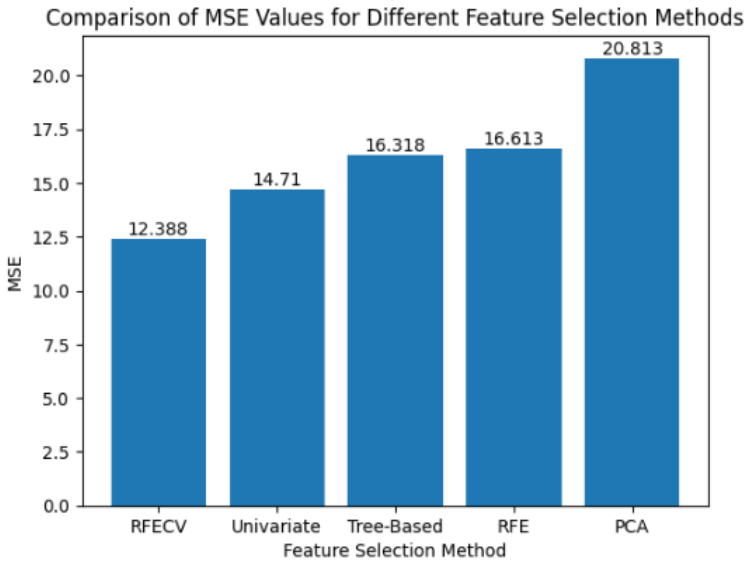

Figure 2

2.3.6 **Data Scaling**

Scaling ensures that all features are treated equally during the learning process. Without scaling, features with larger magnitudes may disproportionately influence the model thereby could lead to bias. In order to ascertain a better method for this purpose, a comparison of two methods has been carried out (JN 2.6) i.e. MinMax Scaler Method and the Standard Scaler Method. Both methods scaled down the number of rows to 469 from 511. The comparison graphs of both methods with trend lines display a better shape than the original graph of the original data. The MinMax Scaler more proportionately distributed the data thereby making smooth trendlines and shape similar to the original data distribution (figure 3). While the Standard Scaler, converged the data more towards the Q2 or central portion of the distribution (figure 4) and thus was not used.:-

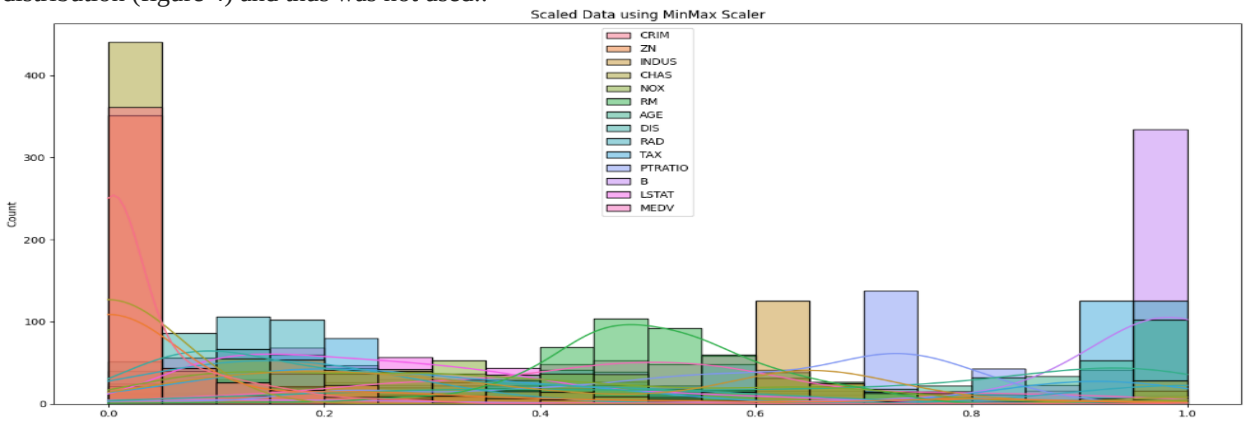

Figure 3

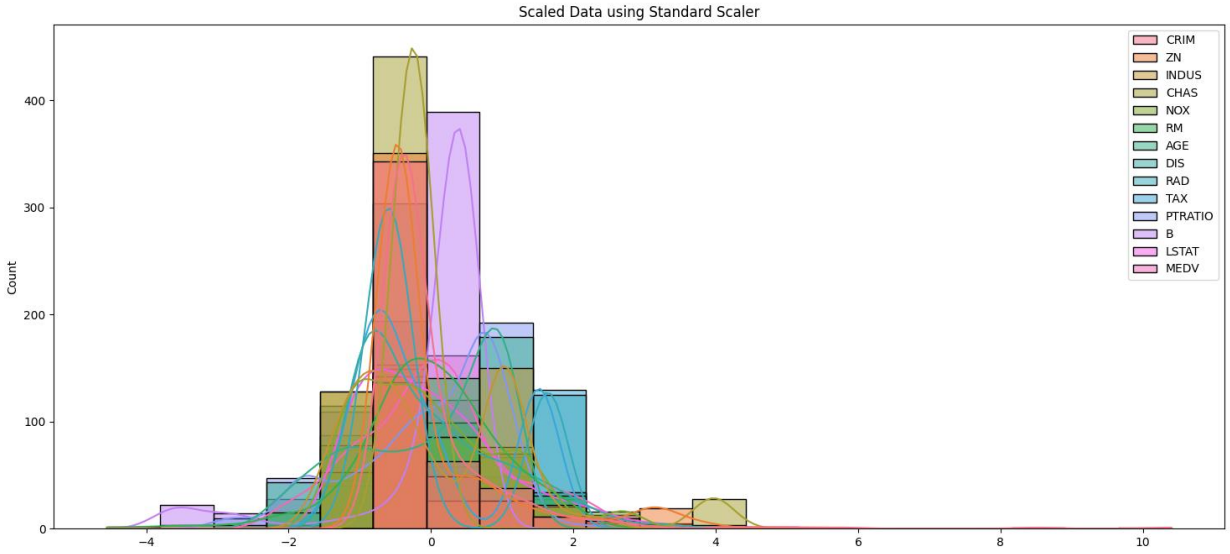

Figure 4: Standard Scaler Method

3.1 Feature Importance (JN 2.5.0)

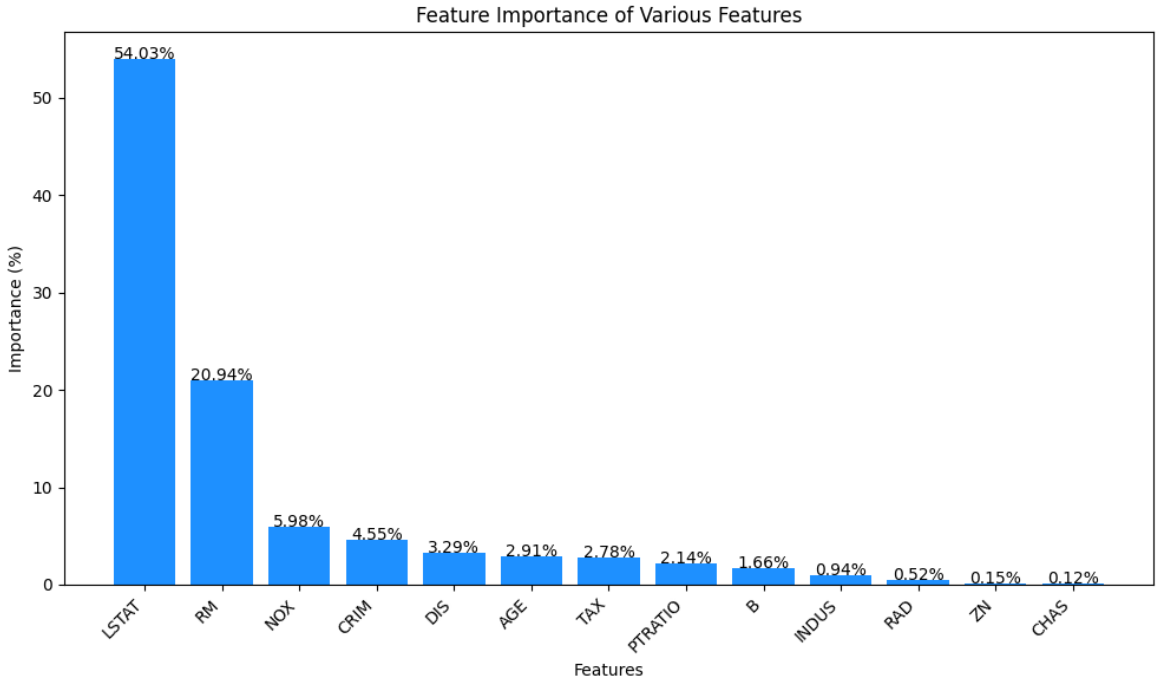

Figure 5

These were calculated and converted into percentage for better assimilation by incorporating the cross-validation object of optimized Random Forest model as in figure 5. This graph shows the most and the least influential features. Clearly, LSTAT is exceptionally influential as compared to the rest. RM is slightly influential. CRIM and NOX stand

below 10% although slightly above 5%. While rest of them are not much influencing the target feature MEDV, since, they are below 5%. While, CHAS, ZN and PAD are even below 1%; thereby being almost non-influential.

While the feature correlation matrix is showing Pearson correlation coefficients between various features is shown in figure 6.

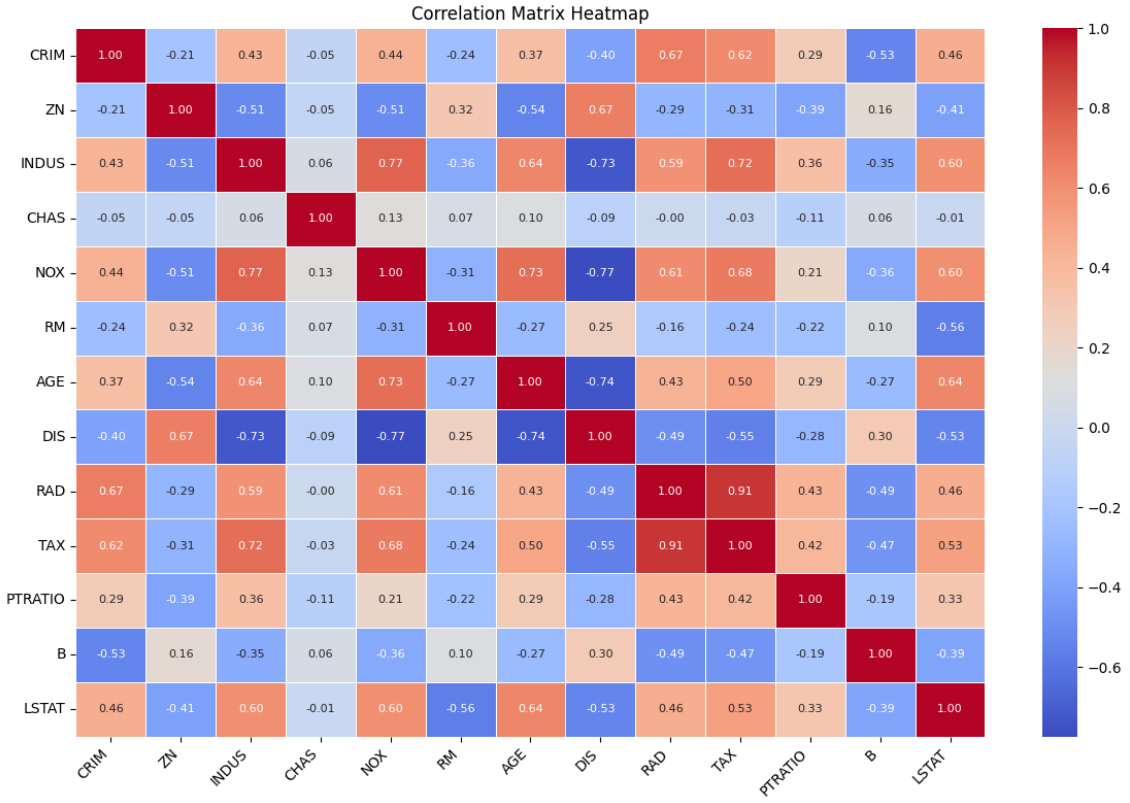

Figure 5: Feature Correlation

Notable observations include strong positive correlations between 'INDUS' (proportion of non-retail business acres per town) and 'NOX' (nitric oxides concentration), which is 0.77, and between 'RAD' (index of accessibility to radial highways) and 'TAX' (full-value property tax rate per $10,000), which is 0.91. There is also a strong negative correlation between 'DIS' (weighted distances to five Boston employment centers) and 'NOX' (-0.77).

These strong correlations indicate significant relationships between certain features. For example, the high correlation between 'INDUS' and 'NOX' suggests that areas with more industrial activity tend to have higher pollution levels. Similarly, the strong correlation between 'RAD' and 'TAX' implies that properties with better access to radial highways are associated with higher property taxes. Conversely, the negative correlation between 'DIS' and 'NOX' indicates that areas further from employment centers tend to have lower pollution levels. This understanding of feature relationships is crucial for interpretation in the context of predicting housing prices.

3.2 Enhanced Visualization of Most Important Relationships (JN 2.5.0)

Scatter plot analysis has been carried out for top four features in figure 7 above (according to their feature importance). The plot between LSTAT and MEDV shows an inverse correlation showing that the areas with a higher percentage of lower-status population may be associated with lower housing prices. The plot between RIM and MEDV shows a weak positive correlation which means that as RIM increases, MEDV tends to increase. However, there is a lot of scattering in the data which suggests that the age of housing units is one of many factors that affects median housing values. The scatter plot between NOX and MEDV shows a strong negative correlation suggesting that neighborhoods

with higher levels of air pollution tend to have lower median housing values. The scatter plot between CRIM and MEDV shows a strong negative correlation with most of the data in an exponential form. This shows that as crime rate increases, MEDV exponentially decreases, thereby tend to lower median housing values.

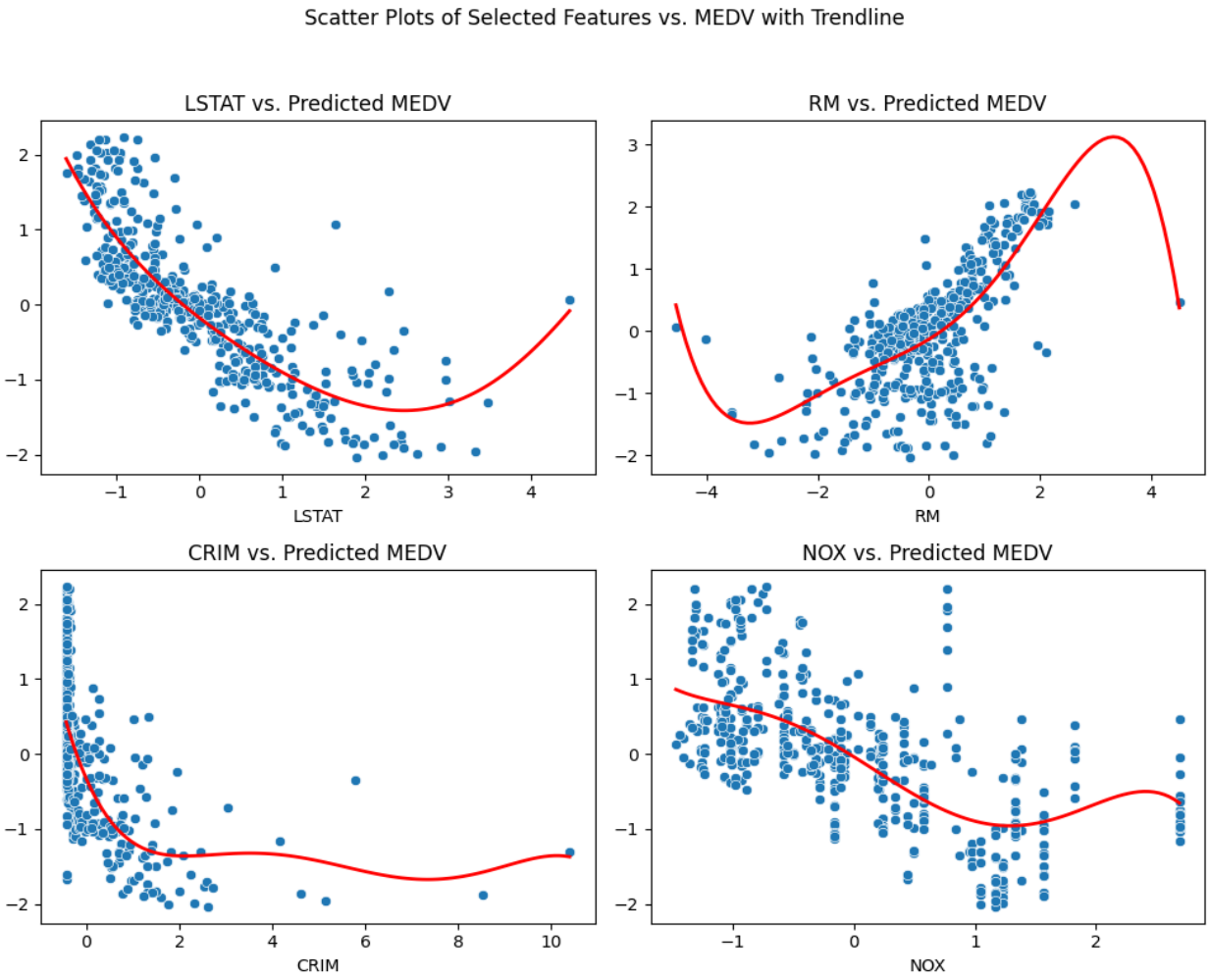

Figure 6

In figure 8, the plot DIS and MEDV shows a positive correlation. MEDV values increase until reaching a threshold when the line becomes flat but then again tend to slightly increase. The AGE and MEDV plot shows decrease in MEDV at a lower pace w.r.t AGE. So is the case with TAX and the MEDV plot. In PTRATIO plot, the wave pattern shows uneven trend with an overall lowering of prices w.r.t increase in PTRATIO. While, the rest of features including B, INDUS, RAD, ZN and CHAS have feature importance below 2%. Thus, they might not play an important role. While, the rest of features including B, INDUS, RAD, ZN and CHAS have feature importance below 2%. Thus, they might not play an important role.

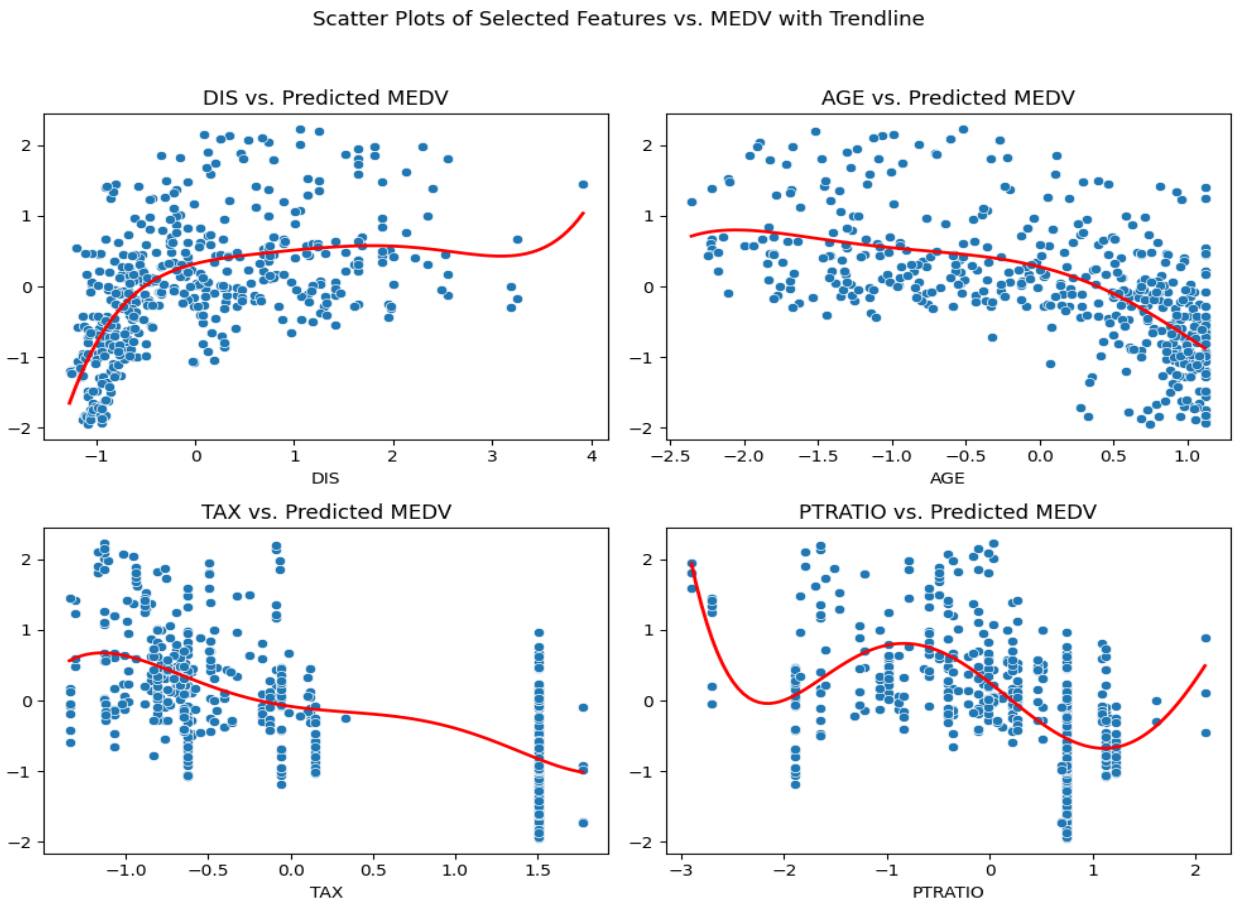

Figure 7

3.3 Deductions from Feature Importance and Scatter Plots:

The analysis of the Boston Housing Dataset yielded significant findings, shedding light on essential factors influencing housing price predictions in the Boston area. The analysis identified few crucial features strongly influencing housing prices as in the study carried out [4]. Notably, LSTAT and RM are more impactful than the rest. As high as about 50% of the lower socio-economic income status (LSTAT) value indicates that a larger proportion of the population in that area has a lower economic status that is making the model sensitive and largely dependent on this factor. While the average number of rooms (RM) could also drive real estate related business with having a one-fifth feature importance and a linear progressive plot. Thus, these could drive policy-making and formulation of business models for various businesses including real estate, public and private properties, urban and town planning, and not the least; the educational campaigns much stronger than the others. The pollution (Nitric oxides concentration - NOX) factor causes lower quality of life by increasing health issues[5] and thus, affect prices of housing. But it has got about 6% share in feature importance and a lower recession slope, thus causes a limited impact. A 4.5% proportion of the factor of area with more crime (CRIM) shows a weak impact on the housing prices (MEDV) despite the fact that the prices are inversely correlated and are little impacted by that. The ‘distances to five Boston employment centers’ factor has a limited impact on housing prices owing to low importance (3.3%) despite a positive correlation which falls flat on after a certain threshold near zero. The owner-occupied vintage houses also have about 3% importance with a negative correlation, thus affects little to prices. The property-tax factor with a negative correlation and an importance of 2.8% doesn’t formidably impact either. The pupil-teacher ratio by town impacts a little more than 2% with a wave pattern

thereby may be studied in conjunction with other factor(s). However, its overall impact is sliding downwards. The black population ratio, non-retail business acres, the Charles River vicinity, accessibility to radial highways, Proportion of residential land, and non-retail business acres did play a little to no role in this regard despite sounding relevant factors.

4.1 Performance Comparison of Regression Models (JN 3)

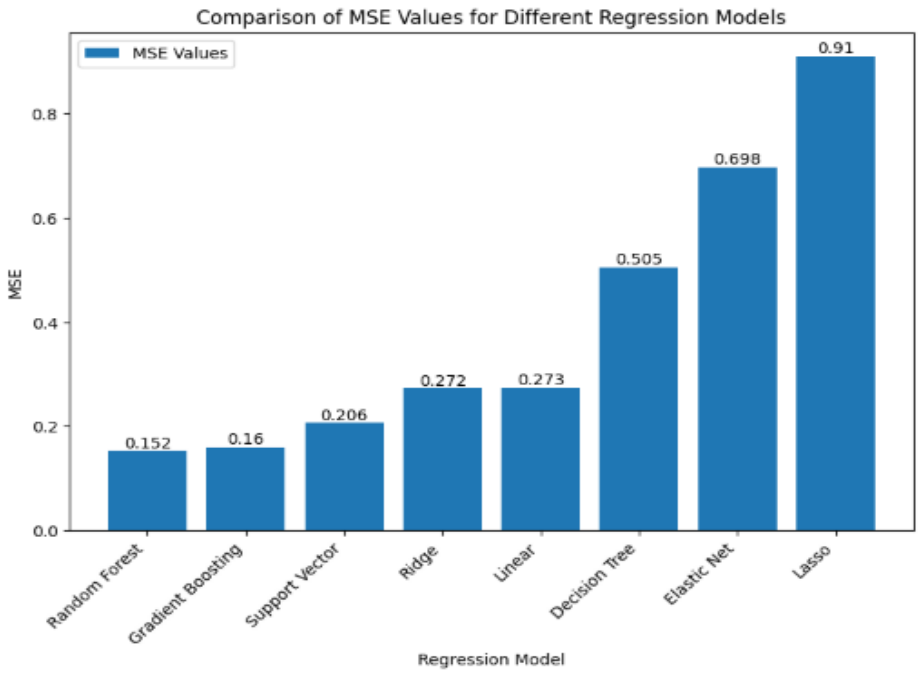

Figure 8

As many as 8 regression models have been pitched to test their performance. (figure 9). In order to make a comparison, algorithms such as Lasso, Elastic Net, and Ridge regression were also employed. Ridge and Lasso models are used for regression problems, which means they are designed to predict continuous outcomes. Lasso, short for Least Absolute Shrinkage and Selection Operator, is a statistical formula used in ML whose main purpose is the feature selection and regularization of data models. Elastic Net Regression is a powerful algorithm that combines the features of both Lasso and Ridge Regression. It is a regularized regression technique that is used to deal with the problems of multicollinearity and overfitting, which are common in high-dimensional datasets. In order to measure performance, Mean Squared Error (MSE) unit was used. While it is worth mentioning that the accuracy, precision, recall, and F1 score are not typically used to measure the performance of *regression models*. These metrics are designed for evaluating classification tasks, where the model predicts discrete class labels (e.g., spam or not spam).

The above comparison suggests that the Random Forest Regression model outperformed all others with a close competitor of Gradient Boosting Regression model. The former achieved the highest accuracy by 0.152 MSE. While the latter achieved 0.16 MSE, thereby, making a marginal difference. The third is Support Vector Regressor model. Thus, the top model (Random Forest) has been shortlisted for further investigation.

4.2 Hyper Parameter Optimization (HPO)

In order to find the best parameters, **Grid Search** has been incorporated with Random Forest Regression model i.e. the best performing model which resulted into MSE value of 0.133 (JN 4). Furthermore, the R-Squared Score has also

improved. Therefore, this model could be used for further analysis and cross-validation. The Table 1 shows the output comparison:-

METRIC	BEFORE OPTIMIZATION	AFTER OPTIMIZATION	IMPROVEMENT
MSE	0.15	0.105	30% decrease
R-SQUARED	0.84	0.90	6% increase

Table 1

4. Evaluation

4.1 Cross Validation

Random Forest model was applied to the evaluate the model using HPO parameters for optimized performance in the evaluation process (JN 5.1). It resulted into MSE 0.2 on the test set which shows that on average, the model's predictions are off by this amount. This value seems reasonable and suggests that the model is making reasonably accurate predictions on unseen data. The mean cross-validation score of 0.83 indicates good performance across different folds of the training data. It suggests that the model is performing well and consistently on different subsets of the training data. The low Standard Deviation of 0.04 suggests that the model's performance is relatively consistent across different folds, and there is not much variation in its predictions.

5. Conclusion and Future Work

In conclusion, the dominating factor impacting housing prices was the economic profile of that area, followed by the fact that how accommodative were those houses. Then comes the public health impacted by the environmental factor of air cleanliness, following by the public security assurance. While few other factors did a little play to almost negligible play.

5.1 Model Performance:

Evaluation metrics, including Mean Squared Error (MSE), scattered plots and bar charts demonstrated robust performance of the chosen machine learning models. The models effectively captured the variance in housing prices, providing reliable predictions and thus, brought actual role players to limelight which resulted into viable conclusions.

5.2 Limitations and Considerations:

The analysis highlighted potential limitations, such as data imbalances, presence of null values, requirement for removal of outliers and data scaling. These limitations underscore the importance of continuous refinement in future studies.

5.3 Implications:

The findings have several implications for real estate professionals, policymakers, and researchers. Understanding the key drivers of housing prices allows for more informed decision-making in property valuation, urban planning, and policy development.

5.4 Future Work

a. **Refinement of Models:**

- Future work could involve further refining machine learning models to capture complex non-linear relationships. Exploring advanced techniques like ensemble models or neural networks may enhance predictive capabilities.

b. **Incorporation of Additional Data:**

- Incorporating additional datasets, such as demographic trends, economic indicators, or recent property sales, may provide a more comprehensive understanding of housing market dynamics.

c. **Temporal Analysis:**

- Conducting a temporal analysis to account for changes over time would contribute to a more dynamic understanding of housing price trends, especially in response to economic fluctuations or policy changes.
- Features having least impact, may further be observed over a period of time. If they remain the same, they may be replaced with more relevant features.

d. **Localized Studies:**

- Zooming in on specific neighborhoods or regions within Boston for more localized studies could reveal detailed insights and cater to the unique characteristics of different areas.

6 References

[1] Sanyal, Saptarsi, Saroj Kumar Biswas, Dolly Das, Manomita Chakraborty, and Biswajit Purkayastha. 2022. "Boston House Price Prediction Using Regression Models." 2022 2nd International Conference on Intelligent Technologies (CONIT). IEEE. [Online]. Available: https://www.researchgate.net/publication/362812590_Boston_House_Price_Prediction_Using_Regression_Models.

[2] Ali, Ahmed. "Machine-learning analysis for Boston housing dataset." [Online]. Available: https://www.researchgate.net/publication/344196650_Machine-learning_analysis_for_Boston_housing_dataset

[3] Perera, P. (2023, January 4). The Boston Housing Dataset. Kaggle. Retrieved January 22, 2024, from https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset/notebook.

[4] Belsley, D., Kuh, E. and Welsch, R. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons.

[5] Harrison, D. and Rubinfeld, D. (1978). Hedonic Prices and the Demand for Clean Air. Journal of Environmental Economics and Management, 5(2), 81-102.

Annexure A

Features - Boston Housing Dataset

1. CRIM: Crime rate per capita by town
2. ZN: Proportion of residential land zoned for lots over 25,000 sq. ft.
3. INDUS: Proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (1 if tract bounds Charles River, 0 otherwise)
5. NOX: Nitric oxides concentration (parts per 10 million)
6. RM: Average number of rooms per dwelling
7. AGE: Proportion of owner-occupied units built prior to 1940
8. DIS: Weighted distances to five Boston employment centers
9. RAD: Index of accessibility to radial highways
10. TAX: Full-value property-tax rate per $10,000
11. PTRATIO: Pupil-teacher ratio by town
12. B: where Bk is the proportion of blacks by town
13. LSTAT: Percentage of lower status of the population
14. MEDV: Median value of owner-occupied homes in $1000's

Annexure B

Jupyter Notebook – Boston Housing Dataset Analysis

This annexure provides an overview of the Jupyter Notebook file used to conduct the analysis presented in this paper. The same is enclosed with this paper.

The Jupyter Notebook file includes the following steps:

1. Exploratory data analysis and preprocessing
2. Feature engineering
3. Machine learning model training and evaluation
4. Results Analysis

The Jupyter Notebook file also includes detailed explanations of the code and its outputs.

Annexure C

Glossary of Terms

Serial	Term / Abbreviation	Acronym
1.	**JN**	Jupyter Notebook
2.	**EDA**	Exploratory Data Analysis
3.	**MSE**	Mean Square Error
4.	**PCA**	Principal Component Analysis
5.	**RFE**	Recursive Feature Elimination
6.	**RFECV**	Recursive Feature Elimination with Cross-Validation
7.	**HPO**	Hyper Parameter Optimization