DNSC6315: Machine Learning II (Group 3)

Final Project Report

Price Prediction in Real Estate: A Comparative Analysis of Machine Learning Models

By Members of Casa Crunchers:

- Abdul Haleem Abdul Salam
- Sai Nityamani Sahith Matsa
- Kason Richard
- Rameen Ridah
- Zhenqi Zheng

Introduction

In the ever-evolving realm of real estate, where the pulse of the market beats with constant flux, we, the analysts from Casa Crunchers, are avidly exploring innovative methodologies to accurately forecast house prices. Armed with the arsenal of advanced data science techniques, analysts meticulously sift through a plethora of independent variables—from the quintessential metrics like the number of bedrooms, bathrooms, and square footage to the nuances of location and amenities. Among the array of methodologies at their disposal, linear regression, random forest regression, gradient boosting regression, ridge regression, support vector regression, bagging, and boosting stand out as indispensable tools, each offering distinct advantages in tackling multicollinearity and feature selection challenges. Through relentless model training and validation, Casa Crunchers refine their predictive algorithms to grasp the subtle intricacies of the housing market, empowering them to furnish clients with precise estimates tailored to their unique

needs and preferences. By harnessing the power of data-driven insights, these companies not only streamline the buying and selling process but also cultivate trust and transparency within the real estate ecosystem, thereby enriching the experience for both buyers and sellers alike.

Business Understanding:

Problem: Accurate prediction of housing prices is crucial for both buyers and sellers in the real estate market. It helps sellers set competitive prices and aids buyers in making informed purchasing decisions.

Managerial Decisions: Our model will assist real estate agencies and property developers in estimating house prices based on various attributes, facilitating effective pricing strategies and negotiations.

Dataset Overview:

The dataset utilized for this project was sourced from Kaggle, a leading platform for data science competitions and datasets. It comprises a comprehensive collection of housing-related features, including but not limited to the number of bedrooms, bathrooms, area size, location, and various amenities. The dataset provides a rich and diverse source of information, allowing for thorough analysis and model training to predict house prices accurately.

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
1	12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
2	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
3	12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
4	11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished

Data Preparation

The dataset was first loaded into a Pandas DataFrame from a CSV file named 'Housing.csv'. An initial exploration of the dataset revealed that it contains 545 observations with 13 columns. There were no missing values in the dataset, and no duplicate rows were found. The data types of the columns were appropriately identified and handled. Categorical variables were encoded using label encoding to convert them into numerical format for modeling.

```
df.shape
```

```
(545, 13)
```

```
df.isnull().sum()
```

```
price             0
area              0
bedrooms          0
bathrooms         0
stories           0
mainroad          0
guestroom         0
basement          0
hotwaterheating   0
airconditioning   0
parking           0
prefarea          0
furnishingstatus  0
dtype: int64
```

Exploratory Data Analysis

Exploratory Data Analysis was conducted to understand the relationships between various features and the target variable (housing prices). Scatter plots, count plots, and box plots were used to visualize the relationships and identify any potential patterns or trends in the data. For example, scatter plots(Figure 1.1) were used to visualize the relationship between area and price, while count plots(figure 1.2 and 1.3) were used to explore the distribution of categorical variables like air conditioning and hot water heating. Box plots(figure 1.4) analyzed key property characteristics such as price, square footage, number of bedrooms, number of bathrooms, number of floors, and parking spaces; we were able to gain insight into major trends and anomalies in the market.

Figure 1.1

```
Text(0.5, 1.0, 'Hot water heating')
```

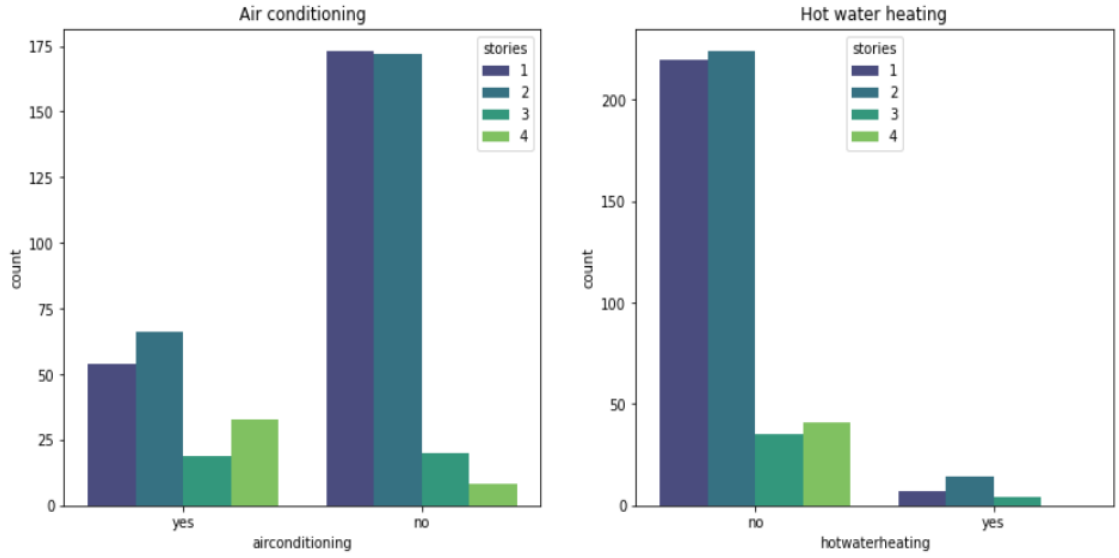


Figure 1.2

<AxesSubplot:xlabel='basement'>

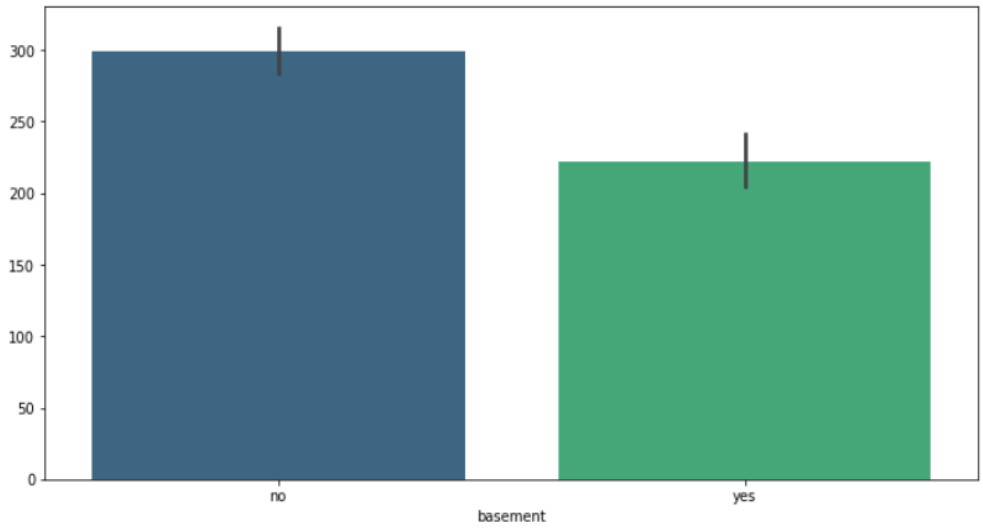


Figure 1.3

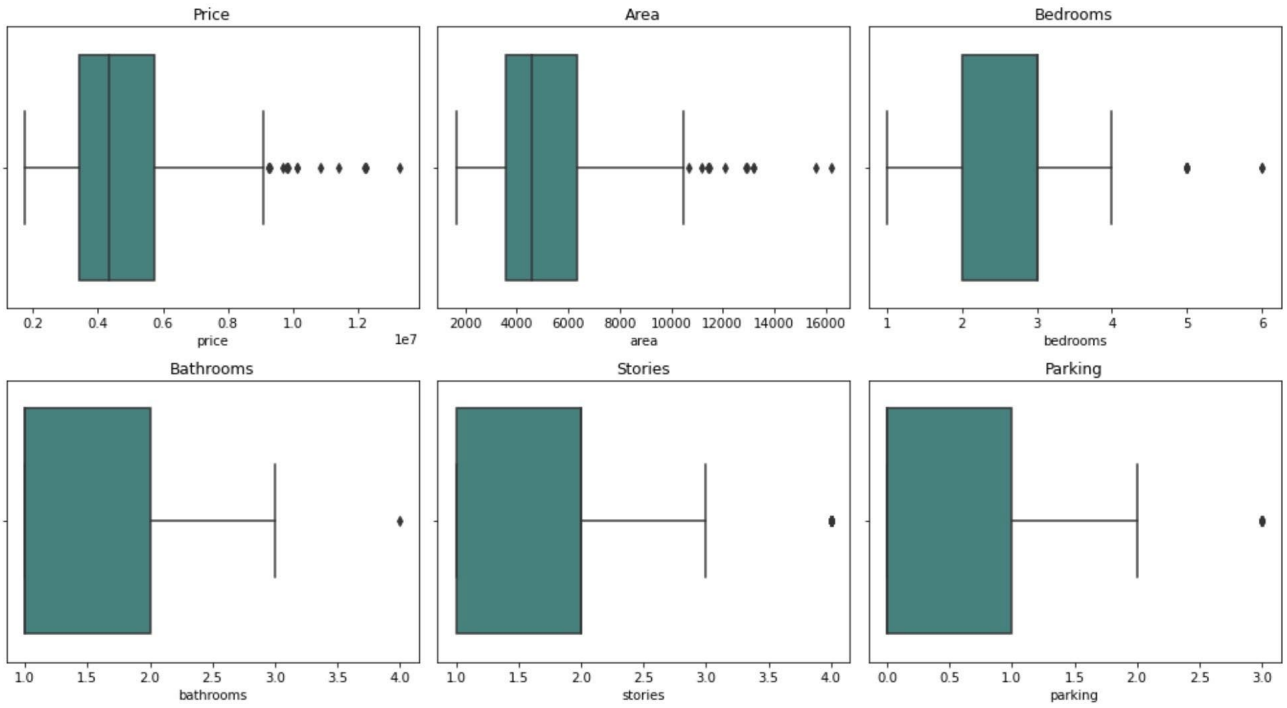


Figure 1.4

Model Building and Evaluation

Several regression models were trained and evaluated using the prepared dataset. The models included:

1. **Linear Regression**
2. **Random Forest Regression**
3. **Gradient Boosting Regression**
4. **Ridge Regression**
5. **Support Vector Regression**
6. **Bagging**
7. **Boosting**

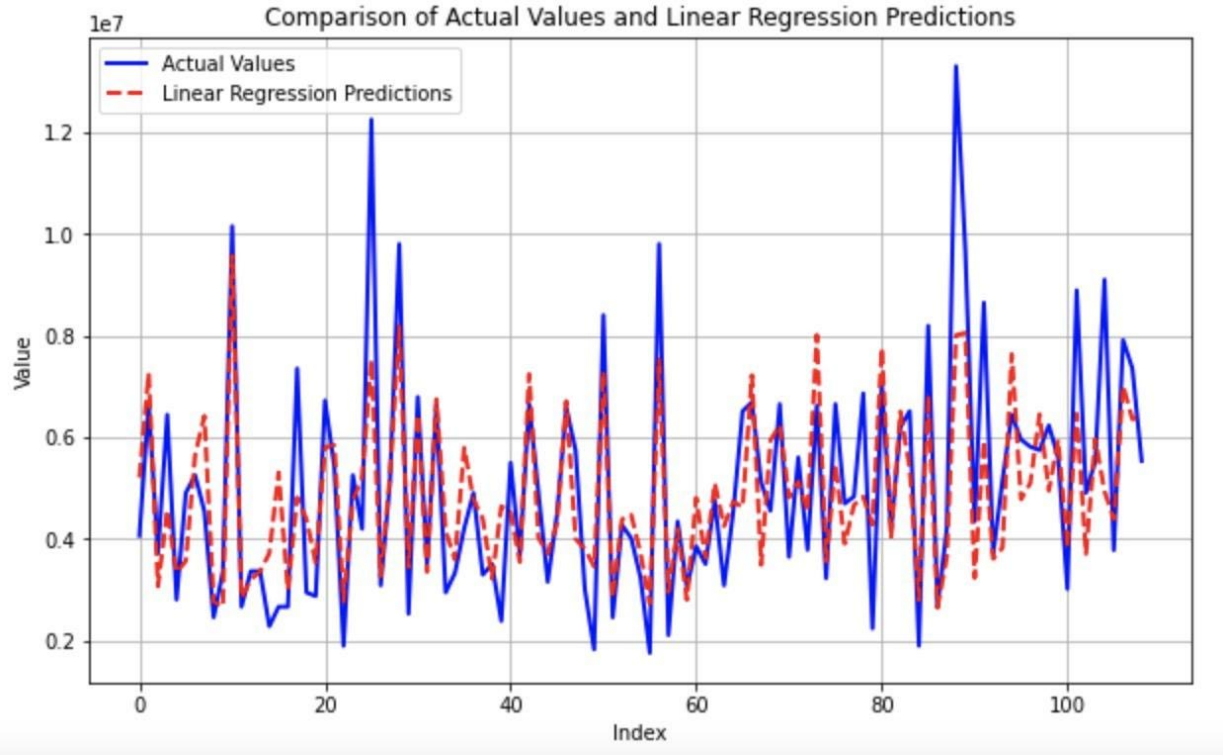
The dataset was split into training and testing sets, and feature scaling was performed where necessary. Each model was trained using the training set and evaluated using the testing set. Evaluation metrics such as mean squared error (MSE) and R-squared (R^2) were used to assess the performance of each model. Additionally, predictions from each model were compared to the actual housing prices to visualize their performance.

Linear regression

In this context of housing price prediction, linear regression attempts to establish a linear relationship between the independent variables (features) and the dependent variable (price). The model was instantiated using Linear Regression() class from scikit-learn and was trained on the training data using the fit() method. After running this model, we calculated the Mean Squared Error which came out to be 1771751116594.0342 and the R-Squared which was 0.6495. The Linear Regression model shows some predictive capability with a moderate R-Squared, there is still room for improvement, as indicated by the relatively high MSE. Further analysis and potentially more modeling techniques could enhance the model's performance.

The prediction plot below visually compares actual values with predictions made by a linear regression model, illustrating how closely the predictions align with the real data across an index of data points. The linear regression prediction shows the model generally follows the overall

trend of the actual data. There are noticeable discrepancies, especially at points where the actual values peak or drop sharply. These deviations suggest that the linear regression model, despite capturing the general pattern, struggles with extreme values and might not fully account for possible non-linear relationships within the data.

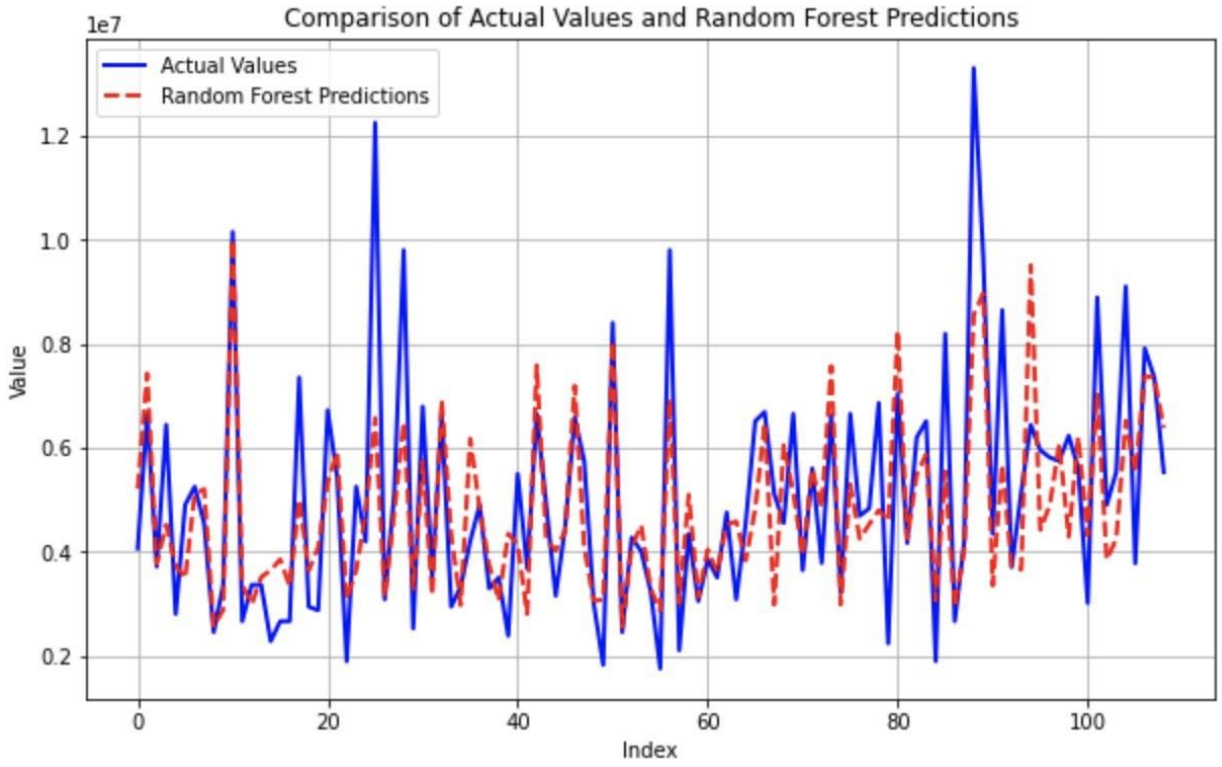


Random Forest Regression

Random Forest is an ensemble learning technique that builds multiple decision trees during training and merges their predictions to improve accuracy and reduce overfitting. With housing price prediction, Random Forest Regression can capture complex relationships between features and target variables more effectively compared to a single decision tree. After training this model, the Mean Squared Error came out to be 1963538216518.6526 and the R-Squared was 0.6115. The R-Squared again indicates that the Random Forest Regression model demonstrates moderate predictive capability and there is still room for improvement indicated by the high MSE.

For the Random forest predictions below, it can be seen that this graph highlights the ability of the Random Forest to approximate the actual data points across the range, showing that the

predictions often closely follow the actual trends but with some deviations, particularly at higher values or more volatile sections of the data.

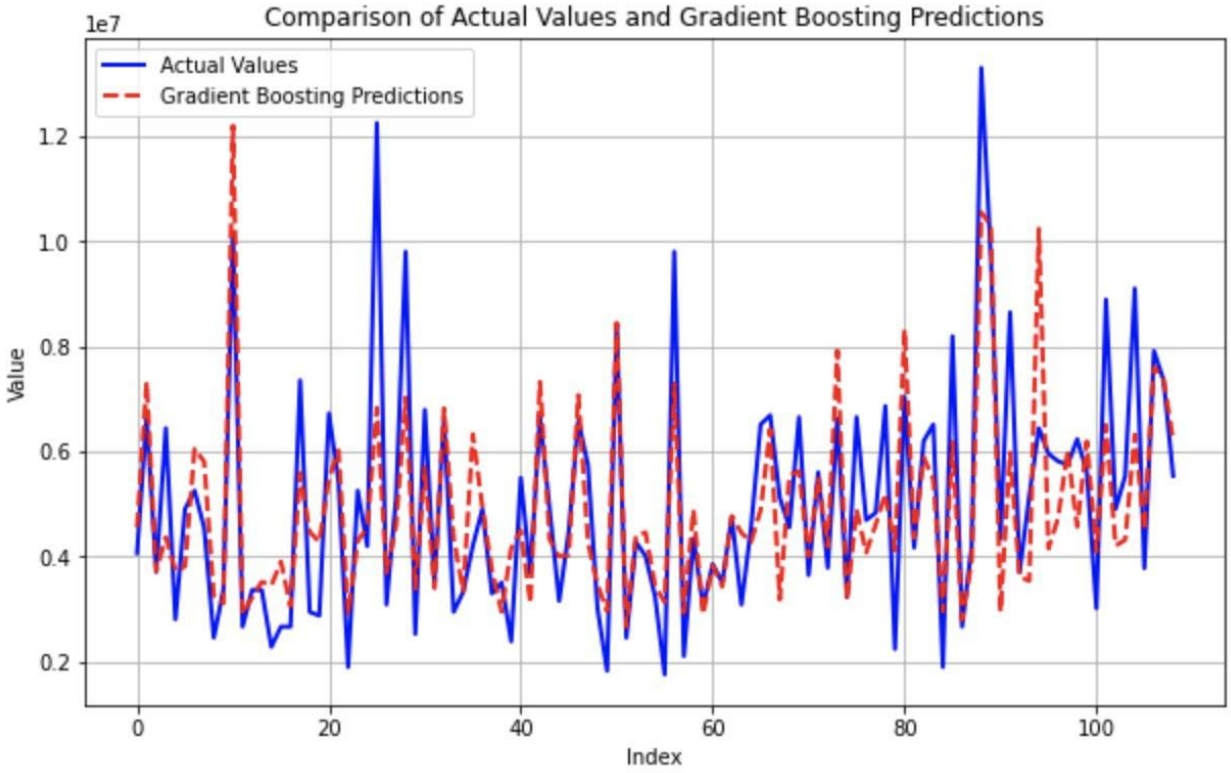


Gradient Boosting Regression

Gradient Boosting Regression operates uniquely when applied to predicting housing prices. It builds upon errors from previous predictions, optimizing its approach through an iterative process. Unlike simpler models that analyze the entire dataset uniformly, Gradient Boosting leverages a series of decision trees-each designed to improve upon the shortcomings of its predecessors. The procedure is initialized in scikit-learn with the GradientBoostingRegressor class, which is trained using a sequence of residuals to refine its accuracy progressively. Once the model is trained, it's evaluated using metrics such as Mean Squared Error (MSE) and R-squared. For example, with an MSE of approximately 1694870370248.4102 and an R-squared of 0.6647, Gradient Boosting demonstrates a robust ability to fit the data better than other models. While the MSE indicates there's still a little gap between the predicted and actual values, suggesting room for further refinement, Gradient Boosting modular approach allows for adjustments in learning

parameters to better handle the dataset's complexities. This adaptability makes it highly effective in real estate market analysis, where the influences on property prices can be intricate and interdependent.

When comparing actual values with Gradient Boosting model predictions, shown by the solid blue and dashed red lines, respectively. The Gradient Boosting predictions generally follow the actual data trends closely, although some discrepancies appear at points of rapid value change. This indicates that the model captures the overall patterns well but may struggle with sudden fluctuations in the data.

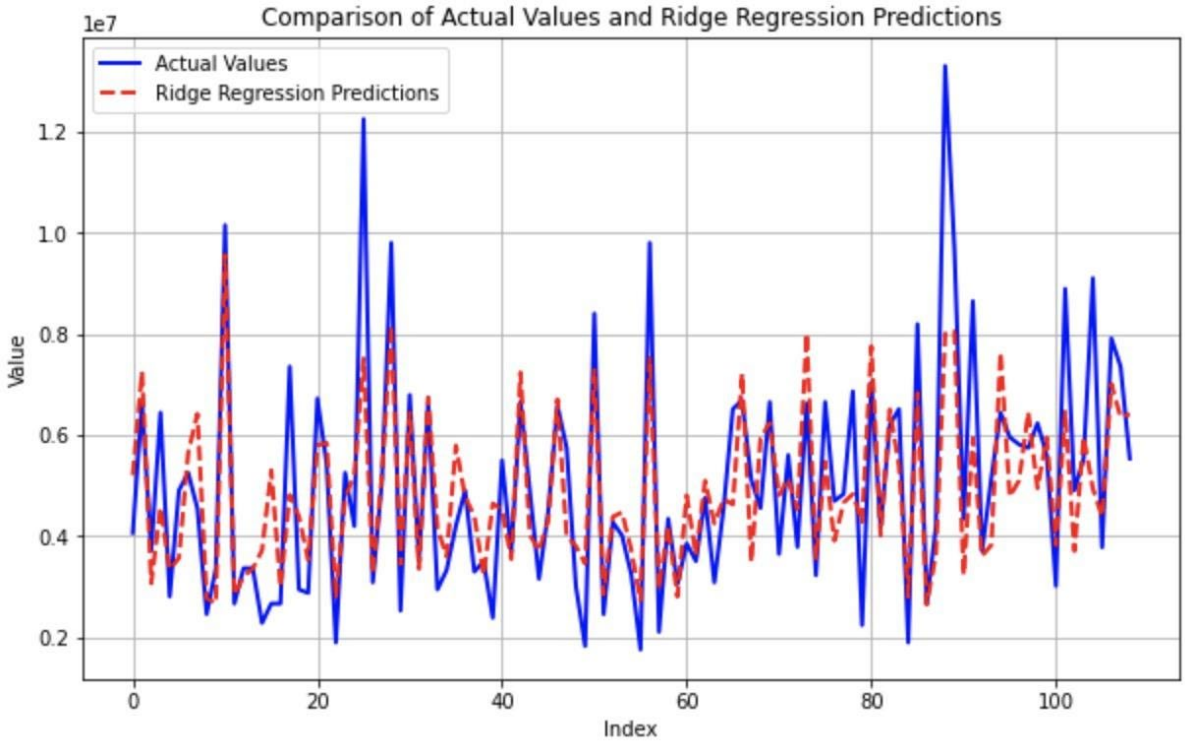


Ridge Regression

Ridge Regression is a linear regression model that includes regularization, which helps prevent overfitting by penalizing large coefficients. It is particularly useful when multicollinearity exists among the predictor variables. The Mean Squared Error ended up being 1772333186531.0132 and the R-Squared was 0.6494. For Ridge Regression, the R-Squared

indicates a relatively good predictive performance. However, the MSE is still relatively high. Overall, the model provides good insight for the house price predictions, but there may be room for more improvement.

When displaying a comparison between actual values and predictions from Ridge Regression model. The visualization shows that the Ridge Regression predictions closely track the actual data points, though with some discrepancies, particularly at the peaks. This indicates that the model generally captures the overall trend of the data but may struggle with precision at higher value points.

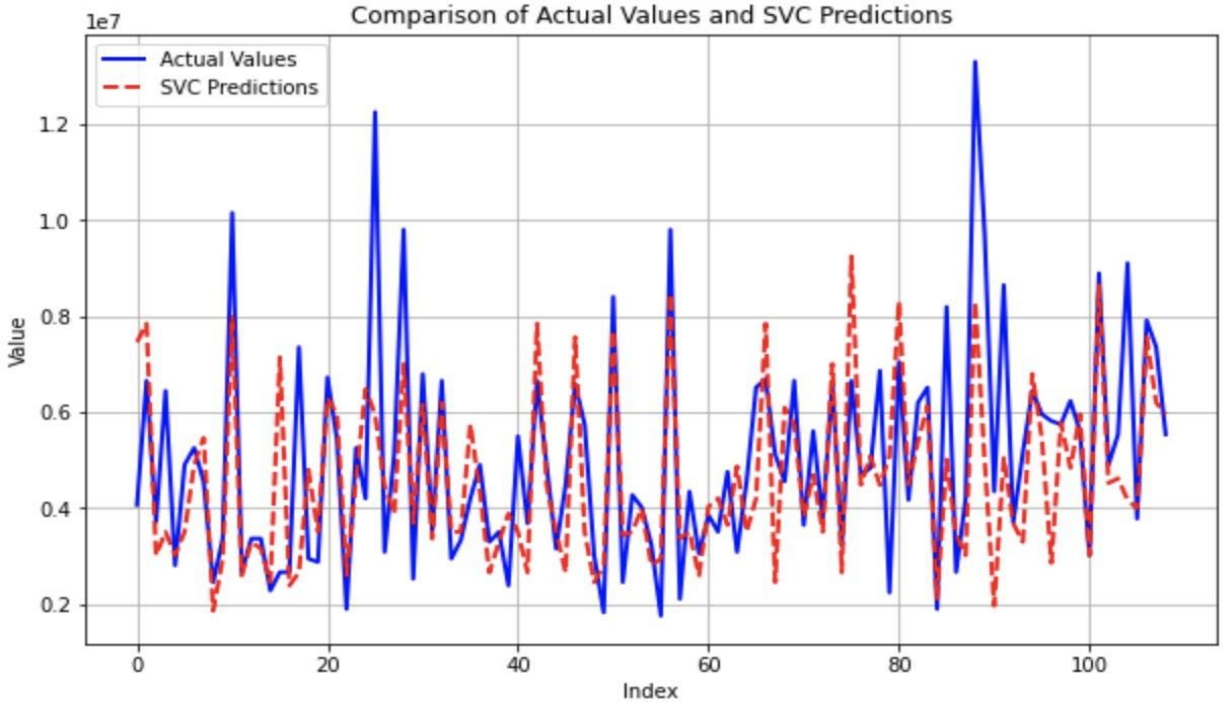


Support Vector Regression

Support Vector Regression (SVR) is a supervised learning algorithm used for regression tasks that constructs a hyperplane or set of hyperplanes in a high-dimensional space to predict the continuous target variable. It is based on the same principles as Support Vector Machines (SVM) for classification but changed for regression problems. After running the model, the calculation came out to be 3065089133027.523 for the Mean Squared Error and 0.3936 for the R-Squared. The MSE for this model is high which indicates a large discrepancy between the actual and predicted values. The R-Squared for this model is also low which indicates a smaller proportion

of the variation of the dependent variable explained by the independent variables. Both values suggest that the model may not perform well in capturing the variation in the target variable and the model may not be effectively capturing the underlying patterns in the data. This is an obvious indication that the Support Vector Regression model more than likely is not the best choice for predicting housing prices in this dataset.

The visualization below(The comparison of accrual values and SVC predictions) reveals that the SVC predictions closely follow the fluctuations of the actual data, though with notable deviations, especially in capturing the peaks. This indicates that while the model aligns well with the general trends, it may have challenges with precision during more extreme changes in the data.

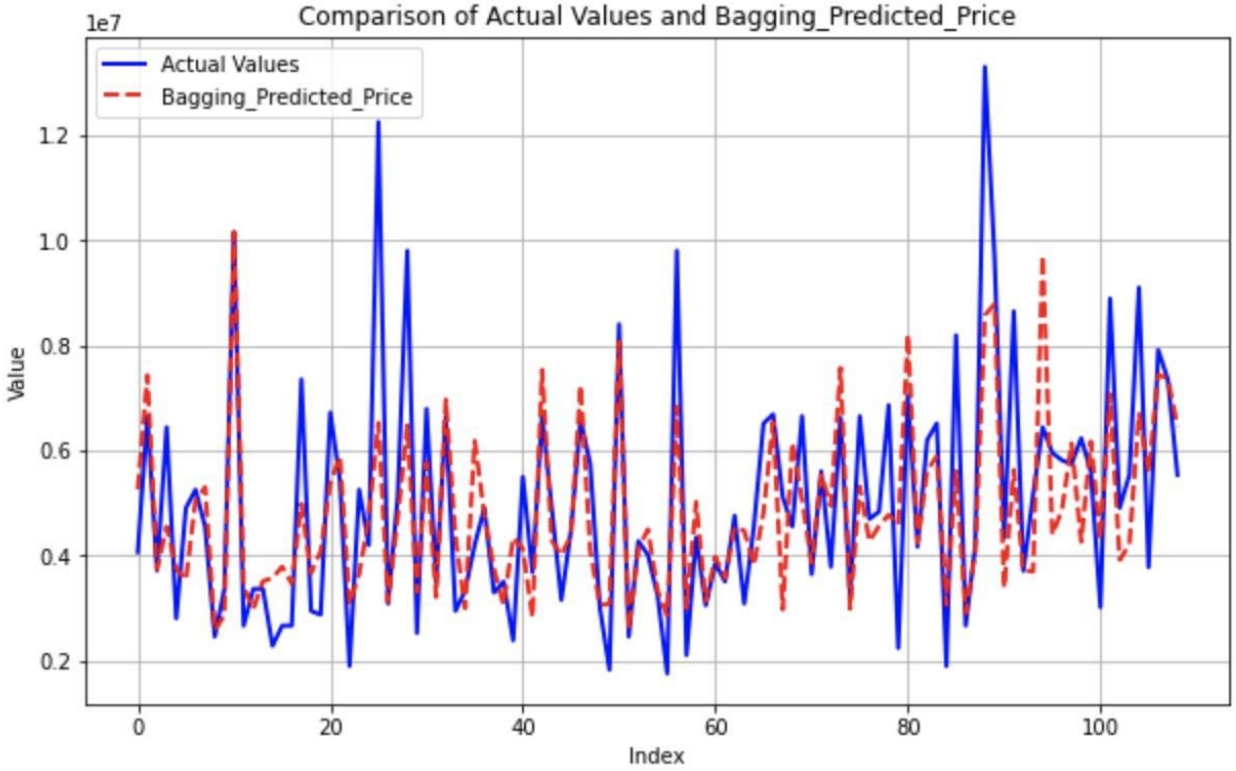


Bagging

Bagging offers a distinct approach to predicting housing prices by utilizing the power of ensemble learning, which stands in contrast to both linear regression and boosting methods. This technique involves creating multiple models like decision trees. Each trained on a slightly different subset of the training data. These subsets are generated by randomly sampling the original dataset with replacement, ensuring that each model has a unique perspective on the data, thereby reducing variance and improving model robustness. When we output the model, it amalgamates the insights from various independent models to form a consensus output. This method effectively diminishes the likelihood of overfitting, which is often a drawback of more complex, non-linear models that

attempt to fit every detail of the training data. The bagging method showcased a mean squared error (MSE) of approximately 1977244607051.165 and an R-squared value of 0.6088. These metrics suggest that while bagging provides a robust predictive capability with better error handling than some other models, it still needs room for optimization to enhance accuracy and model fit.

The Bagging model predictions shown below closely follow the actual data trends, demonstrating a good fit, although there are occasional discrepancies, particularly at data points with sharp peaks or sudden changes. The model seems effective in capturing the general pattern of the data, but it may not fully capture the extremities or the most volatile fluctuations.The Bagging model predictions closely follow the actual data trends, demonstrating a good fit, although there are occasional discrepancies, particularly at data points with sharp peaks or sudden changes. The model seems effective in capturing the general pattern of the data, but it may not fully capture the extremities or the most volatile fluctuations.

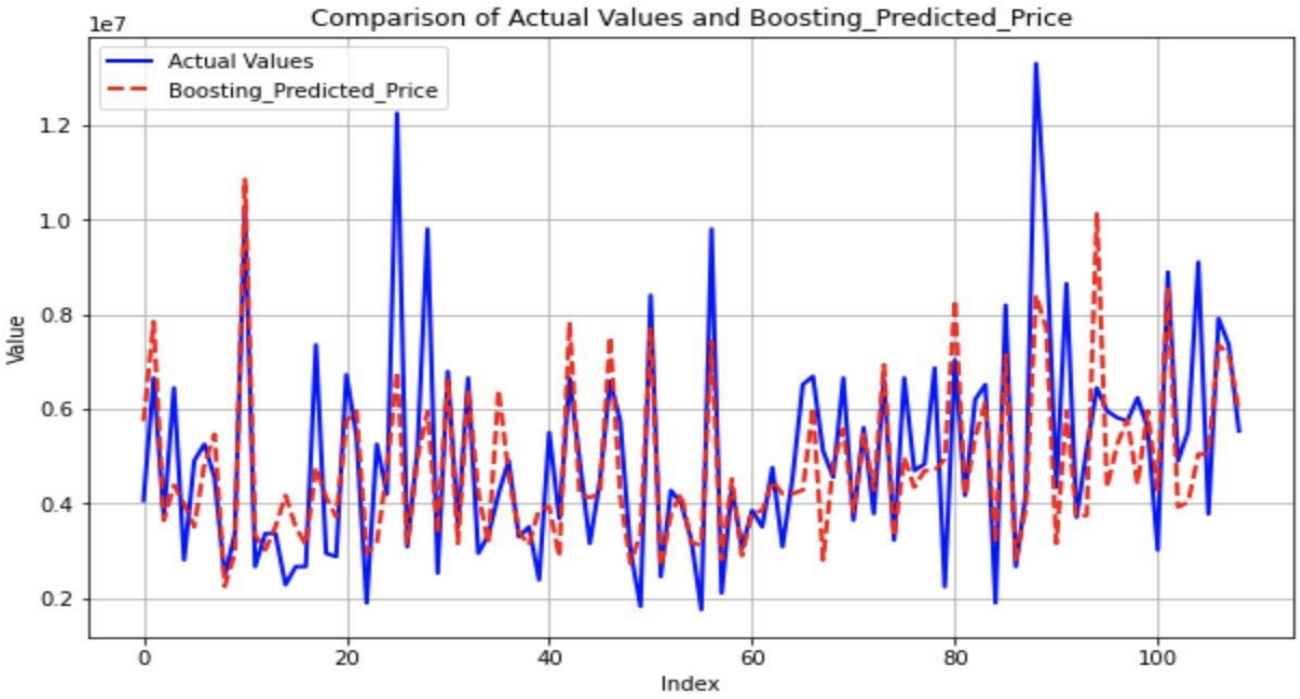


Boosting

Boosting is an ensemble learning technique that combines multiple weak learners sequentially to create a strong learner. AdaBoostRegressor, which is a boosting algorithm, is used

for regression tasks. After running the model, the Mean Squared Error came out to be 2077443429238.3057 and the R-Squared came out to be 0.5889. This model shows a high MSE and a lower R-Squared which suggests that the model may not perform well in capturing the variation in the target variable and the model may not be effectively capturing the underlying patterns in the data. This is an indication that the Boosting model is more than likely not the best choice for predicting housing prices in this dataset.

From the comparison between actual value and prediction, the predictions closely track the actual data's movements, demonstrating the model's capability to follow the overall trends, albeit with some minor deviations at peaks and through more volatile sections. This visualization suggests that the Gradient Boosting model is generally effective but may occasionally miss the mark on exact peaks or rapid changes in the dataset.



Results and Discussion

The results of the analysis revealed that the Gradient Boosting Regression model performed the best among all models, achieving the lowest Mean Squared Error and highest R-Squared score. This indicates that the Gradient Boosting model was most effective in capturing

the relationship between the features and the target variable, resulting in more accurate predictions of housing prices. The Bagging Regression and Random Forest Regression models also performed well but were slightly less accurate compared to Gradient Boosting.

Hyperparameter and Cross Validation

We employed hyperparameter tuning using GridSearchCV along with cross-validation to find the best combination of hyperparameters. The optimized model was then evaluated using various metrics on the test set. The best hyperparameters for the Gradient Boosting Regressor were found to be learning rate at 0.05, number of estimators at 100, maximum depth at 5, minimum samples per leaf at 4. The best R-squared score achieved during cross-validation was 0.6307, indicating a good fit of the model to the training data. The model was evaluated on the test set using an R-squared of 0.6552 and a Mean Squared Error of 1742594144026.4756. Finally, we made predictions on new data using the optimized model.

Prediction Price

In examining the original price list differentiated by area, it's evident that there are discrepancies between the actual housing prices and those predicted by the Gradient Boosting model. Specifically, in the first area (as per the first row of the dataset), the model forecasts a house price of $10,679,539 compared to the actual price of $13,300,000. Similarly, for the second area, the model's prediction of $11,142,564 falls short of the actual price of $12,250,000. In the third area, the predicted price is significantly lower at $6,825,461, whereas the actual price is $12,250,000. The trend continues in the fourth area, with a predicted price of $11,290,257 versus an actual price of $12,215,000. Lastly, in the fifth area, the model predicts $9,124,701, again underestimating the real price of $11,410,000. This pattern indicates that despite utilizing consistent features across the dataset, the model consistently estimated lower prices for the houses in all five scenarios.

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	1	0	0	0	1	2	1	0
1	12250000	8960	4	4	4	1	0	0	0	1	3	0	0
2	12250000	9960	3	2	2	1	0	1	0	0	2	1	1
3	12215000	7500	4	2	2	1	0	1	0	1	3	1	0
4	11410000	7420	4	1	2	1	1	1	0	1	2	0	0

Conclusion

In conclusion, this analysis illuminates potential discrepancies between the pricing provided in the original dataset and the estimates generated by our model. While the original data may have overestimated house prices, our model's more conservative predictions could reflect a more accurate valuation of the properties, given the features used. This underscores the importance of meticulous data analysis and careful model training to ensure precise and reliable predictions in practical applications. The effectiveness of machine learning algorithms in predicting housing prices based on various attributes was clearly demonstrated in this project. The Gradient Boosting Regression model stood out as the best-performing model, closely followed by Bagging Regression and Random Forest Regression. These models prove to be invaluable tools for real estate professionals, homeowners, and policymakers in making well-informed decisions regarding housing prices. Continued refinement and optimization of these models could potentially enhance their accuracy and utility even further.

Contributions:

All team members will work collaboratively on all aspects of the project, including implementing machine learning models, evaluating model performance, conducting data preprocessing and exploratory data analysis, and writing the final project report and PPT.