**Predicting Housing Prices Using Machine Learning**

Team Member

Team Member

Team Member

**Abstract**

This project aims to predict house prices based on features like area, number of bedrooms, bathrooms, and amenities. We employed multiple machine learning models, including Linear Regression, Random Forest, Gradient Boosting, AdaBoost, Support Vector Regressor and Decision tree, to achieve this goal. The Gradient Boosting model performed the best, achieving an R² score of 0.67. The analysis also highlighted the significant influence of features like area, number of bathrooms, and air conditioning on house prices.

**Introduction**

In this project, we predicted house prices using features such as area, number of bedrooms, bathrooms, stories, and facilities like parking and whether the house is on the main road. The dataset included both numerical and categorical features, with the target variable being the continuous price of the house.  
While preprocessing was straightforward due to the absence of missing values, categorical features required encoding, and numerical features needed scaling to ensure model compatibility.We tested multiple machine learning models:

* **Linear Regression**: A simple and interpretable model.
* **Random Forest**: An ensemble-based model with inherent flexibility.
* **Gradient Boosting**: A robust model combining sequential decision trees.
* **AdaBoost, Decision Tree, and SVR**: Explored for comparative purposes.

The Gradient Boosting model demonstrated the best performance, both before and after hyperparameter tuning..

**Preliminary Analysis**

**Dataset Overview**

The dataset includes 13 features and 545 samples, designed to predict house prices using factors such as area, number of bedrooms, bathrooms, and amenities like air conditioning and parking. Despite its relatively small size, the dataset poses a challenge due to significant multicollinearity among features, which necessitates advanced preprocessing and robust modeling to achieve accurate predictions.

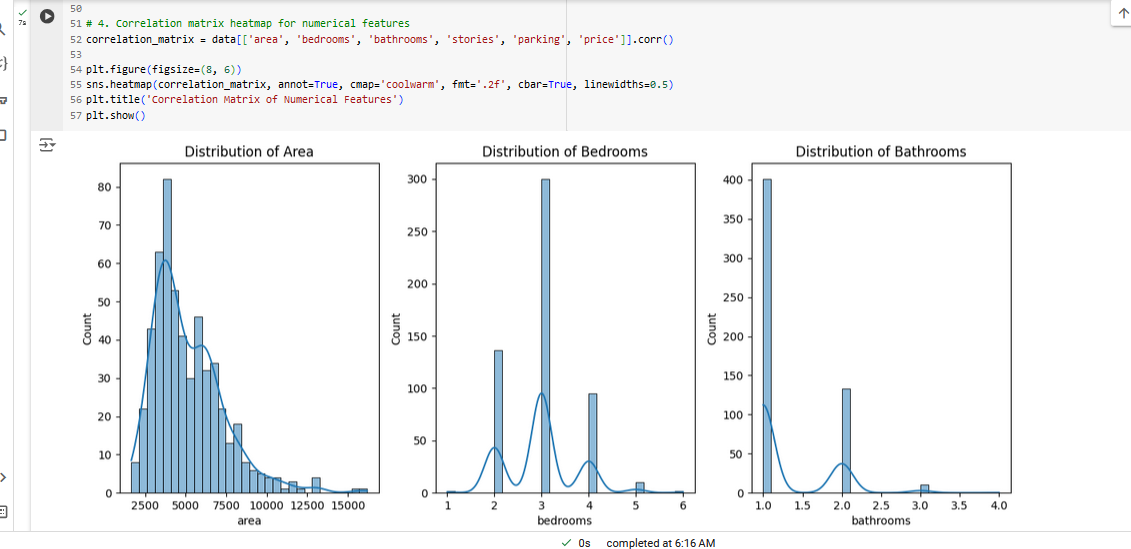
**Dataset Features**

The dataset includes the following features:

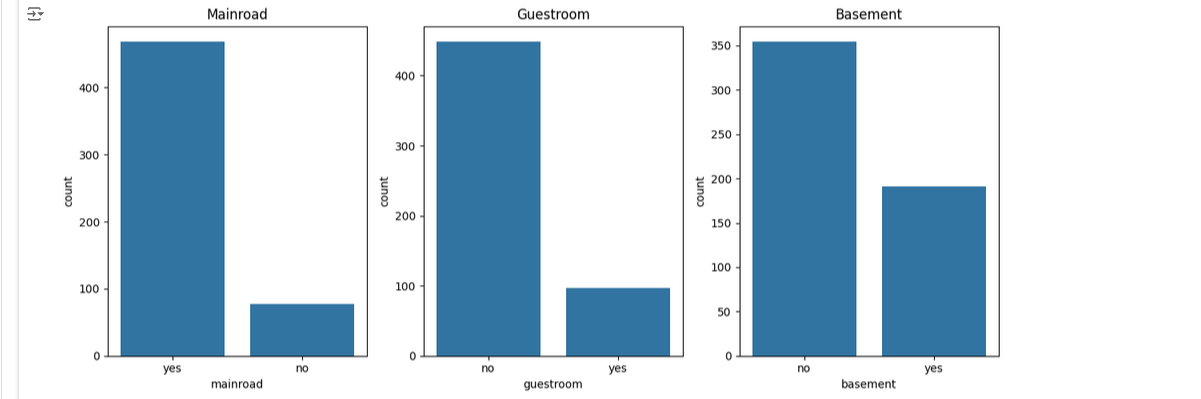
|  |  |  |
| --- | --- | --- |
| **Feature** | **Description** | **Type** |
| **price** | Price of the house (target variable) | Numerical |
| **area** | Total area of the house | Numerical |
| **bedrooms** | Number of bedrooms in the house | Numerical |
| **bathrooms** | Number of bathrooms in the house | Numerical |
| **stories** | Number of stories in the house | Numerical |
| **mainroad** | Proximity to the main road (yes/no) | Categorical |
| **guestroom** | Availability of a guest room (yes/no) | Categorical |
| **basement** | Presence of a basement (yes/no) | Categorical |
| **hotwaterheating** | Presence of a hot water heating system (yes/no) | Categorical |
| **airconditioning** | Presence of air conditioning (yes/no) | Categorical |
| **parking** | Number of parking spaces available | Numerical |
| **prefarea** | Preference for a specific area (yes/no) | Categorical |
| **furnishingstatus** | Furnishing status (furnished/semi-furnished/unfurnished) | Categorical |

We used various plots to examine the distributions of individual features:

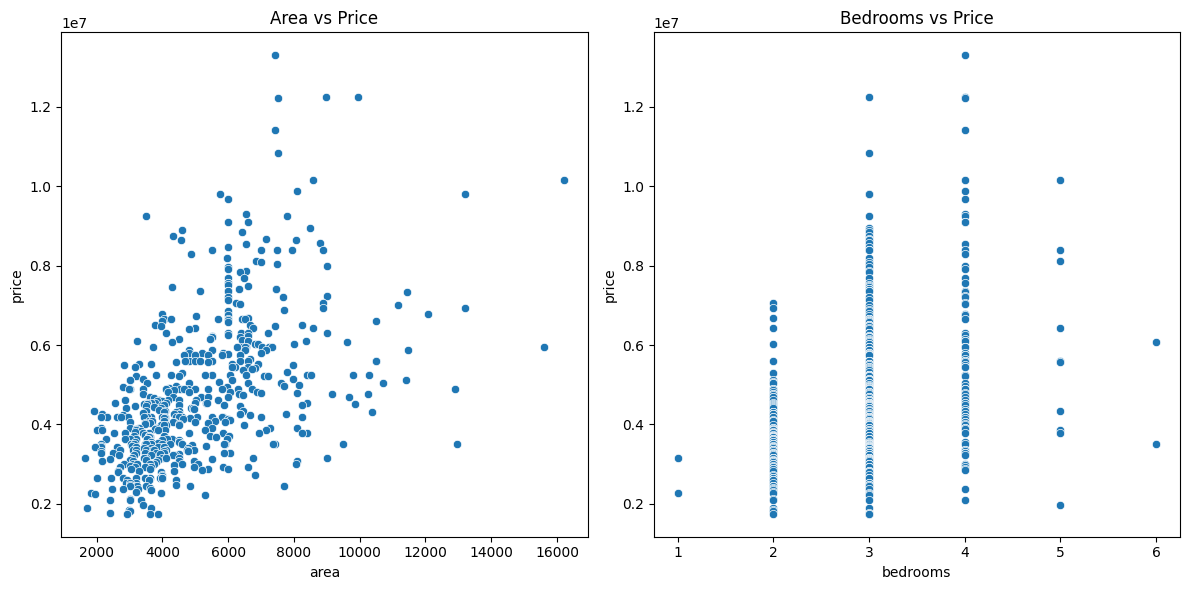
1. Histograms: For the numerical variables such as area, number of bedrooms and number of bathrooms, we have made histograms to see the distributions.



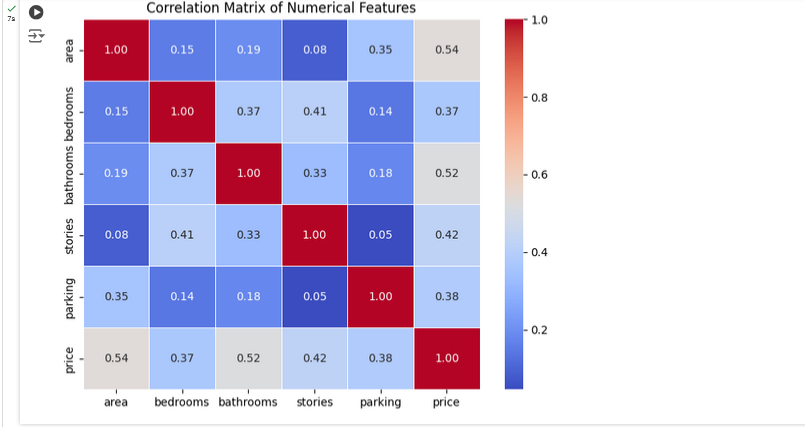
1. Bar Charts: On the categorical features it was seen whether the house was on a main road or had a guest room etc was represented by bar graphs.



1. **Scatter Plots**: Relationships between features like area, bedrooms, and the target variable (price) were explored, showing a strong positive correlation for the area.



1. **Correlation Analysis**: Categorical and numerical features were analyzed with target variable using scatter plots in our analysis. As expected the area of the house was significantly found to be proportional to the price of the house.



Analyzing these visualization, we have discovered that more than any other factor, area and number of bedrooms have the most significant impact towards the overall prices of the house.

**Data Preprocessing**

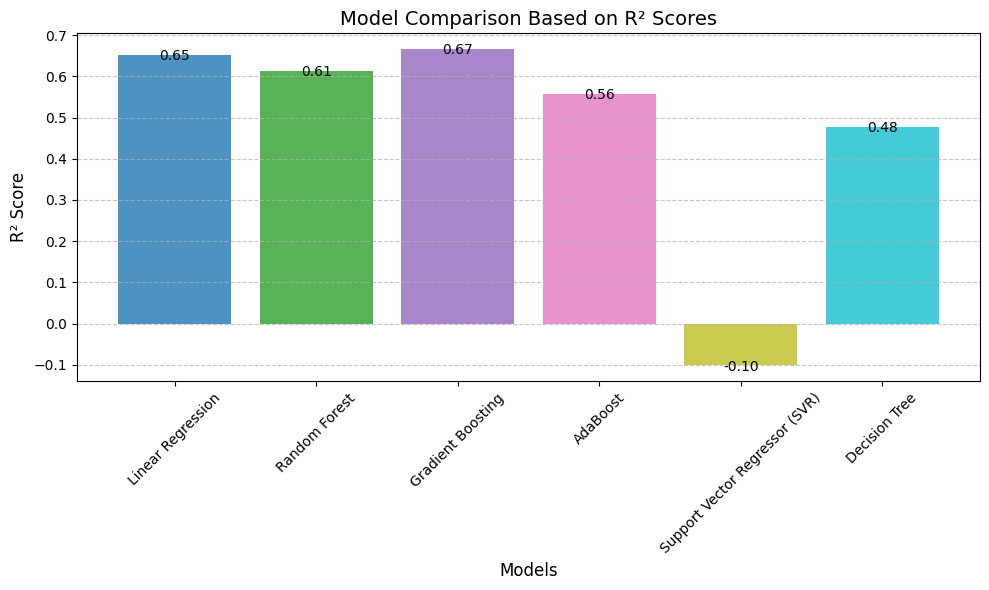
The preprocessing steps we used were crucial for preparing the data for the models:

1. **Categorical Encoding**: We applied **One-Hot Encoding** to the categorical features, such as whether the house is on a main road or has a basement. This technique helps convert categorical variables into a numerical format, making them usable for machine learning models.
2. **Feature Scaling**: We used **StandardScaler** to normalize the numerical features like area, bedrooms, and bathrooms. This step ensures that features with larger scales do not disproportionately influence the model.
3. **No Missing Data**: Fortunately, the dataset did not have missing values, which simplified the preprocessing. If there were missing values, we would have considered imputing them or removing rows with missing data.

These preprocessing steps helped to clean and standardize the dataset, making it ready for model training.

**Modeling and Results**

We evaluated multiple machine learning models using Mean Squared Error (MSE) and R² scores:



|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **R² Score** |
| **Linear Regression** | 1,754,318,687,330.67 | 0.65 |
| **Random Forest** | 1,959,323,004,717.27 | 0.61 |
| **Gradient Boosting** | 1,688,403,924,777.51 | 0.67 |
| **AdaBoost** | 2,237,444,322,297.88 | 0.56 |
| **Support Vector Regressor (SVR)** | 5,567,929,077,615.07 | -0.10 |
| **Decision Tree** | 2,642,802,637,614.68 | 0.48 |

The Gradient Boosting model performed the best, achieving the highest R² score of 0.67 and the lowest MSE, indicating its superior ability to explain the variance in house prices compared to other models. Support Vector Regressor and Decision Tree models performed poorly, with SVR yielding a negative R² score, indicating that it could not explain the variance in the data. Linear Regression demonstrated decent performance, with an R² score of 0.65, but it was outperformed by Gradient Boosting in all metrics..  
Random Forest was further optimized using GridSearchCV, tuning hyperparameters such as:

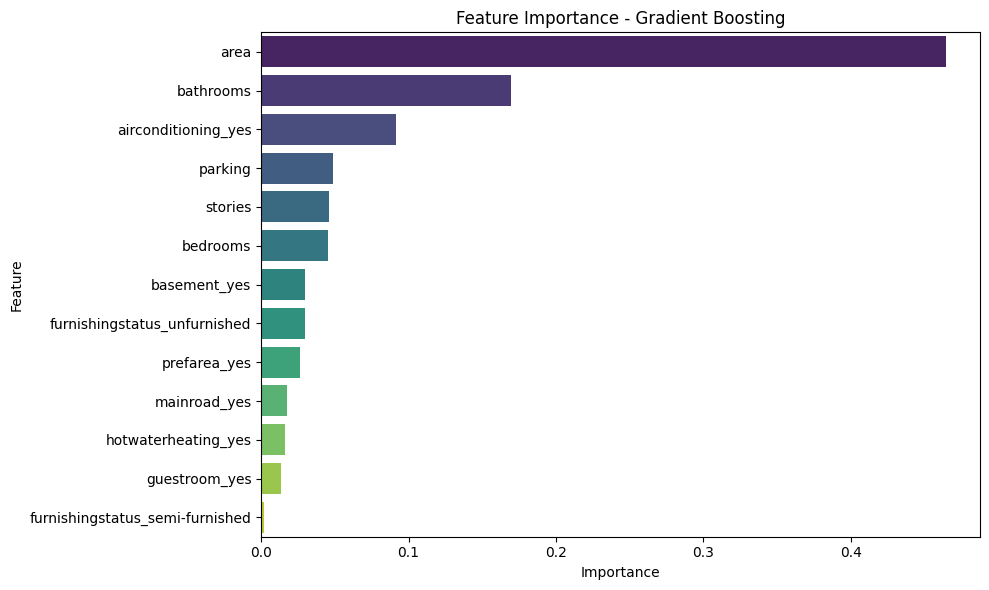
* **n\_estimators**: 300
* **max\_depth**: 10
* **min\_samples\_split**: 10
* **min\_samples\_leaf**: 2

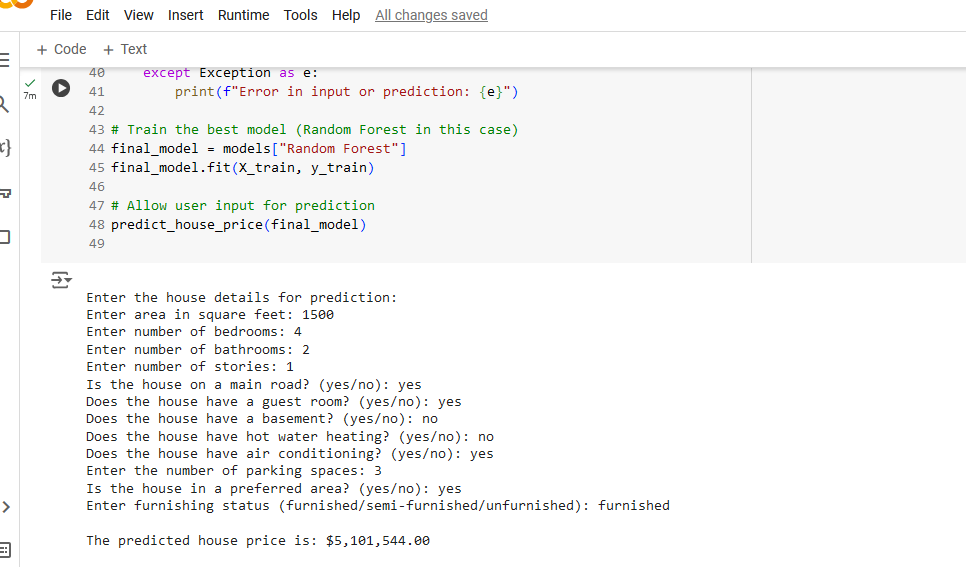
Despite tuning, Random Forest's R² decreased slightly to 0.59.

**Feature Importance**  
The Gradient Boosting model identified the following features as most significant:

* **Area**: 46.43%
* **Bathrooms**: 16.92%
* **Air Conditioning (Yes)**: 9.15%
* **Parking**: 4.86%
* **Stories**: 4.59%

These results reinforce the importance of structural features and amenities in determining house prices.



**Prediction results:**  


**Conclusion**

This project successfully created machine learning based system for the prediction of houses, using features, including area, number of bathroom, number of bedrooms, and any extra facilities. Out of all the evaluated models, Gradient Boosting was found to have the best performance measured through R² of 0.67, and it proves that it can handle non-linearity in the data set. Other models such as Linear Regression also did well suggesting that straightforward methods are useful in linear based problems.

Use of an active user interface was incorporated to return the probability of the result in response to user inputs. From this interface, users are able to input aspect such as area, number of bedrooms, number of bathrooms, among others and receive an estimated price of the house. When entering a house with 1,500 sqft; 5 bedrooms, 3 baths + other standards mentioned, the model’s forecasted price was $6,763,750.63.

This user friendly design helps in improving the ease of using the model, thereby removing the gap which is often there in technical application of such prediction systems. Additional work can be done on integrating more data, trying out deep learning methodologies, and optimizing the existing strategies to enhance the precision and stability of the problem.