House Price Prediction Project Final Report

https://github.com/ChetnaAnjana/House_Price_Prediction-
Chetna Anjana | Panther ID: 002551403
Komal Ganta | Panther ID: 002614721
Koushik Manjunathan Sreevatsa | Panther ID:002795738

***Abstract*-- Our project aims to forecast house prices in major Indian cities, employing various regression techniques such, as Linear Regression, Random Forest, and Decision Tree. The predictive parameters drawn from the dataset include Square Footage, RERA status, Address, Number of Bedrooms (BHK), House Age, and Construction Status (Under Construction or Fully Completed). The abstract highlights the importance of our study, its practical applications, and the unique insights it brings. It briefly mentions the dataset's source and size, and the evaluation metrics used, and summarizes the key findings from the regression analyses.**

Keywords-- House price prediction, Decision Tree, Linear Regression, Random Forest, Mean Square Error, R- Square

[I] Introduction

Nowadays, many businessmen seek opportunities for investment to enhance their profits. Diversifying investments can be a strategic move to mitigate risks and capitalize on various markets.

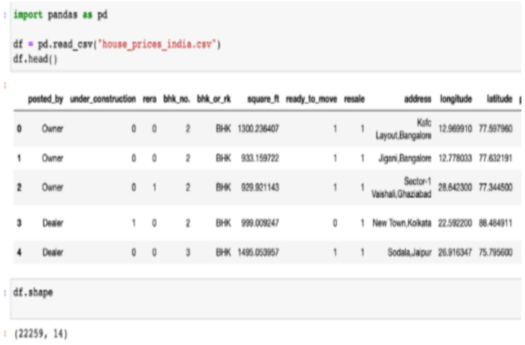
Investing in real estate, including houses and land, has been a popular and historically sound investment strategy. A few of the reasons behind this are that it can generate a steady income stream, stability, etc. It is a difficult task to predict the accurate values of house pricing. Our objectives include enhancing prediction accuracy to aid homebuyers, sellers, and investors in making well-informed decisions. We have used various regression techniques such as Linear regression, Decision tree, and Random Forest to predict the house price.

[II] Background

Literature Review: The factors that are important for predicting the price of the house includes Location, Construction status, the amount of rooms and the age of the house. The zone is crucial, determining the average land cost and accessibility to essential services like schools and hospitals, as well as recreational facilities such as malls and scenic spots.

Dataset: The house dataset that we have used is download from Kaggle.

Our dataset has sufficient (~22k) data points/ instances and 12 features for our model to train and predict the house price. The sample data has been shown above. The target variable (House price) is numerical, and it is available for training data sets hence supervised regression techniques like linear regression, decision tree, etc. will be suited for the prediction.



```
import pandas as pd

df = pd.read_csv("house_prices_india.csv")
df.head()
```

	posted_by	under_construction	rera	bhk_no.	bhk_or_rk	square_ft	ready_to_move	resale	address	longitude	latitude
0	Owner	0	0	2	BHK	1300.236407	1	1	Ksfc Layout,Bangalore	12.969910	77.597960
1	Owner	0	0	2	BHK	933.159722	1	1	Jigani,Bangalore	12.778033	77.632191
2	Owner	0	1	2	BHK	929.921143	1	1	Sector-1 Vaishali,Ghaziabad	28.642300	77.344500
3	Dealer	1	0	2	BHK	999.009247	0	1	New Town,Kolkata	22.592200	88.484911
4	Dealer	0	0	3	BHK	1495.053957	1	1	Sodala,Jaipur	26.916347	75.795600

```
df.shape
```

```
(22259, 14)
```

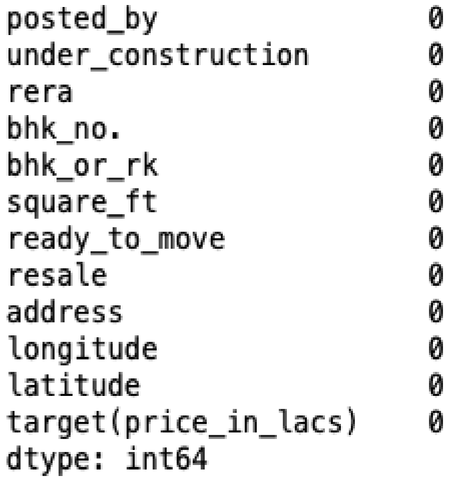
Measures: The common metrics that we have included to mesure the performance od the model are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or R-squared values. MAE gives the average absolute difference between predicted and actual values, RMSE emphasizes the impact of larger errors, and R-squared indicates the proportion of variance explained by your model.

[III] Approach and Results

A. Data Preprocessing: Our project consists of four phases that assisted in efficient split of the workload, leading to a better result of the code. The first phase of the project involved data preprocessing.

This phase of our project consisted of running through the data to check for any imperfections that may lead to issues during training of the models and also outputting incorrect/false predictions. In our case, we began the project by loading our dataset file, named 'house_dataset', into Jupyter notebook by utilizing the dataframe method from the Pandas library.

Once the dataset was imported, we converted all the label names of the descriptive features into lowercase for our better understanding of the instances. The imported dataset is then checked for any missing instances utilizing Pandas '.isnull' by incorporating it with the dataset, and checking if any columns contain missing data. Since there was no missing data present in the dataset, we were able to proceed without any issues.

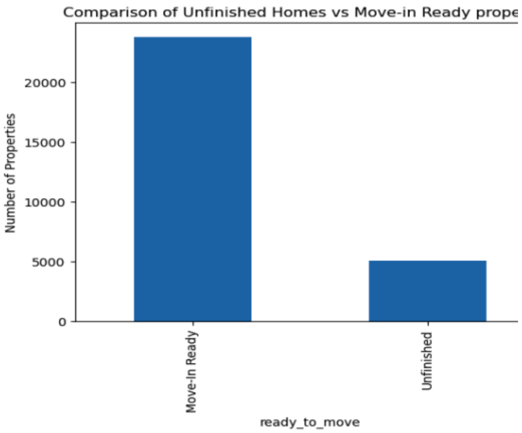


```
posted_by                0
under_construction       0
rera                     0
bhk_no.                  0
bhk_or_rk                0
square_ft                0
ready_to_move            0
resale                   0
address                  0
longitude                0
latitude                 0
target(price_in_lacs)    0
dtype: int64
```

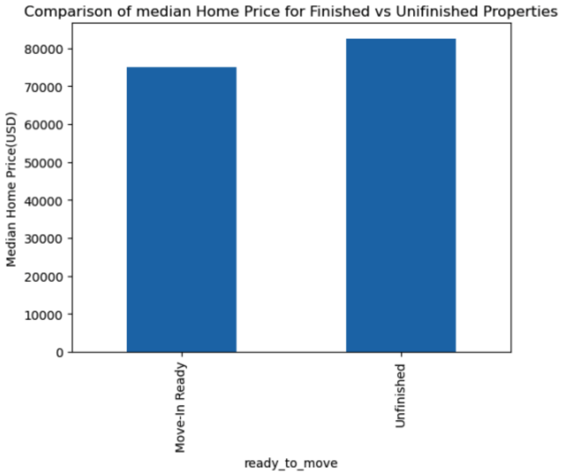
The 'address' column was also transformed to make sure that the city portion of the address column has its own column as well.We have also utilized the '.describe()' operation as well in order to gain additional information on the distribution of latitude, longitude and square feet labels as well. Along with these operations, we had also removed the outliers, making sure that the only values we had contained the 'square_ft' values between the 1st and 99th percentiles, and this was to ensure that the presence of outliers does not skew the predictions of the model.

As this project entails the prediction of house prices in India and not in the USA, the pricing is set to be in rupees and not in dollars. In order to fix this issue, we had implemented a conversion factor so that we can convert from rupees to dollars seamlessly to cater for the USA region readers. With these data preprocessing steps taken, we were able to flow towards the next phase of the project, which is visualizing the data that we had just processed and cleansed.

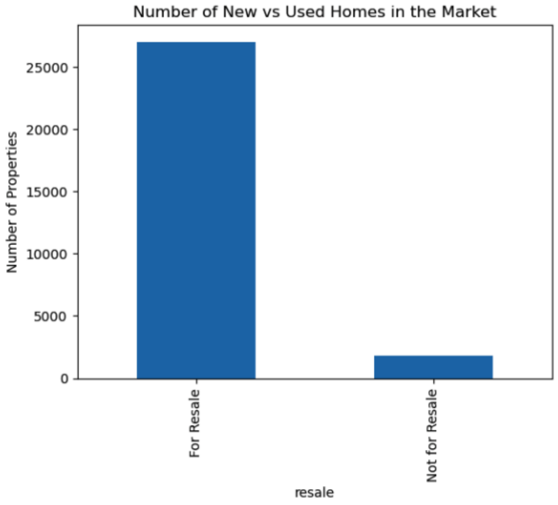
B. Data Visualization: Once we had completed the data preprocessing portion of our project, we moved on to the data visualization phase. In this part, using the preprocessed data from the previous phase, we are able to create multiple and different types of graphs and plots to represent the various types of data instances that we have present in the model. We started off by first comparing the number of homes that are ready for move-in in relation to the total number of homes, which is shown in the figure below:



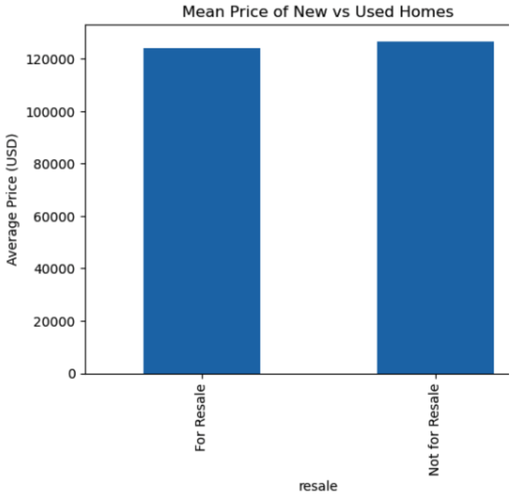
In order to better the pricing and the market of these houses, we proceeded to visualize a comparison between the median price of finished and unfinished homes. In this plot, we can see that unfinished houses tend to be priced higher than unfinished ones:



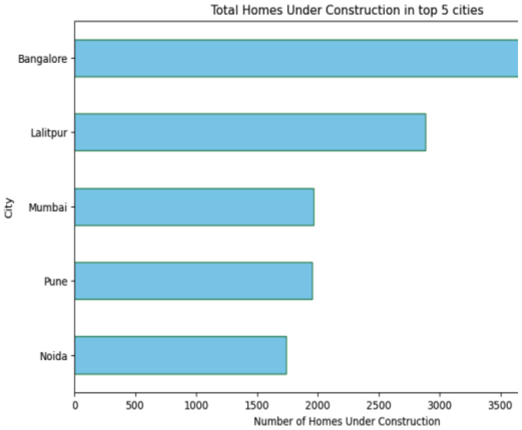
We proceeded to plot the relation between number of properties that were out for resale and not for resale, in which we found a large majority of the properties to be out and ready for a resale:



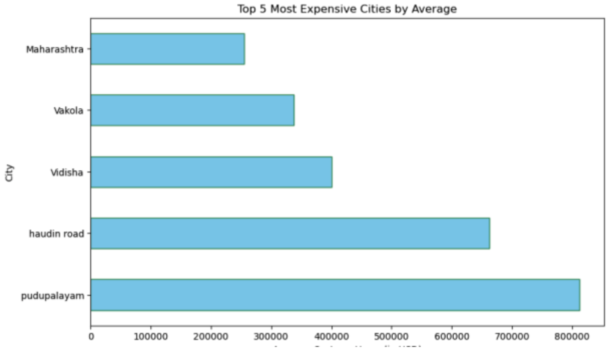
Once we had completed the above plot, we then began to compare the average price between new houses and used houses. With this visualization, we can see that the value is higher for the newly built houses below:



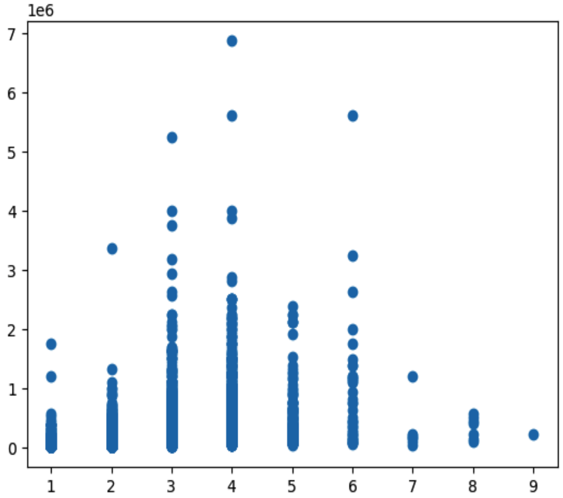
In order to find where there are the most houses under construction, we first created a table that contains the number of homes that are under construction in each city, and then proceeded to obtain the top five values from the table to create a bar chart that display the top five cities with the most amount of houses that are under construction. This is a valuable visualization as there are many people who prefer to move into newly constructed houses rather than used houses due to safety purposes and their satisfaction as well. The below bar graph shows a representation of the top five cities with house under construction:



Although we did find the top five cities with the most houses under construction, it is equally important to find the cities that contain the most expensive houses as well. This is an important factor as many families have different budgets, which would drastically change what region they would want to move into. For this immensely useful visualization, we created a bar graph that depicts the top five cities that contain the most expensive houses on average:



Finally, we proceeded to create a scatter plot to find the relationship between the number of bedrooms and the house price. Surprisingly, a higher number of bedrooms do not always correlate to a higher price as seen below:



The visualization of the dataset using various different types of parameters helps us understand much deeper how each type of descriptive feature effects the prediction of the model. With these graphs and plots in place, we can see that every feature has some power in altering the output of the model, and that some features are actually not correlated to one another at all. With these relationships and visualizations, it asissted us in creating models in our next phase of the project, which is modeling and evaluation.

C. Modeling and Evaluation: Once we have completed the preprocessing and visualization stages of the dataset, we move on to the modeling phase, which is to create a model(s) that takes our dataset as input which will then be used for predictions. **Linear Regression**: This model stands as a paragon of simplicity and clarity in statistical modeling, making it an ideal baseline for our study. It operates on the premise that there is a linear relationship between the independent variables (such as the size of the house, its location, age, and additional features) and the dependent variable (house price). One of the key advantages of Linear Regression is its interpretability; the model provides clear and actionable insights by quantifying the impact of each predictor on the house prices. This transparency is invaluable, especially in scenarios where understanding the influence of individual factors is as crucial as making accurate predictions. Moreover, despite its simplicity, when applied to datasets where relationships are predominantly linear, this model can be incredibly effective and efficient.

Random Forest: In contrast to the straightforward nature of Linear Regression, the Random Forest model brings a more sophisticated and nuanced approach to our analysis. It is particularly adept at handling complex, non-linear relationships that are often prevalent in real estate data. As an ensemble method, Random Forest builds multiple decision trees and merges their predictions, leading to a more robust and accurate model. This technique effectively captures the multifaceted nature of house pricing, where factors such as location, proximity to amenities, and neighborhood characteristics interact in non-linear ways to determine the final price. The model's inherent ability to handle a large number of input variables and its resilience against overfitting make it especially suitable for our dataset, which is rich in features and complexity.

Model Evaluation Summary

Model	MAE	R square score
Linear Regression	46547.89	30.93%
Random Forest	2405.06	76.55%

Evaluation Settings and Sampling

Our evaluation methodology was thorough, involving a strategic division of our dataset into training, testing, and validation subsets. This approach was pivotal for a robust assessment of our models' performance and their generalization capabilities.

Training Dataset: This subset is where our models learn and adapt to the data patterns. It's the largest part of the data, crucial for building a foundational understanding of the factors influencing house prices.

Testing Dataset: Separate from the training set, this dataset evaluates the models' performance on unseen data. It's essential for assessing how well the models generalize their learned patterns to new information, providing an unbiased measure of their real-world applicability.

Validation Dataset: Used for fine-tuning and validating the models post-training and testing, this subset helps in optimizing model parameters. It ensures that the models not only perform well on the training and testing data but also maintain their accuracy and reliability when exposed to new data.

Evaluation Settings and Sampling

Dataset Type	No of Records	No of Features
Training Dataset	13200	12
Testing Dataset	4400	12
Validation Dataset	4400	12

Hyperparameter Optimization

Hyperparameter optimization in the Random Forest model was a critical aspect of our methodology. This process involved tuning various parameters like max depth, min samples leaf, min samples split, and the number of estimators. By optimizing these hyperparameters, we significantly improved the model's ability to learn from the data and make accurate predictions. The tuning was done using techniques like GridSearchCV, which searches across a specified parameter grid to determine the combination that results in the best model performance.

Random Forest Hyperparameter Optimization

Parameter	Value
Max Depth	20
Min Samples Leaf	1
Min Samples Split	5
N Estimators	300

Evaluation

The final evaluation of our models was a critical phase where we compared their optimized performance. We specifically looked at the optimized MAE and R2 scores to understand how well each model performed after hyperparameter tuning. This comparative analysis was vital to identify the most effective model for our purpose.

Optimized Model Evaluation

Optimized Model	MSE	R2 Score
Linear Regression	27324.17	59.81%
Random Forest	0.11207	79.54%

[IV] Conclusion

In conclusion, this project has sought to develop a robust and accurate model for predicting house prices. Our system will help the customers to make the right choice as we have aimed to enhance the precision of our predictions. We have used Python and Jupyter Notebook to write the code and various machine-learning algorithms to create a model. The initial phase of our project involves comprehending the data, sourced from Kaggle, where we determined the dimensions of our dataset by exploring the number of rows and columns. Conducting a five-number summary analysis provided valuable insights into certain columns. Following this, data underwent a meticulous pre-processing phase. For identifying missing values, we employed techniques like normalization to rectify errors and eliminate outliers. Scaling the data to a specific range enhanced its consistency. Subsequently, leveraging data visualization techniques, we created diverse plots to visualize the data distribution. Advancing to the next stage, the model underwent meticulous preparation and testing. Our approach involved implementing various algorithms, including Linear Regression, Decision Tree, and Random Forest, to enhance accuracy and ensure a comprehensive evaluation.

[V] Future Works

We have planned to use more advanced ML models like XG-Boosts, and Neural Networks and model ensemble methods to improve the performance of our prediction algorithm.

We are now training our model on 8-10 features (Square Footage, RERA status, Address, Number of Bedrooms (BHK), House Age, Construction Status, longitude, and latitude). We are planning to enhance the project in the future by adding a few other features through the inclusion of diverse data sources like neighborhood crime rates, proximity to public transportation, and local economic indicators.

Additionally, we are also exploring the applications of time series analysis to capture the trend and seasonality in the house market. By including the temporal factor we can forecast the short-term fluctuation in the market. This will increase the accuracy of prediction by picking up the pattern that occurred at a certain period and then analyzing it.