HOUSE PRICE PREDICTION WITH MACHINE LEARNING

STUDENT NAME:	AYOMIDE OTINWA
STUDENT NUMBER:	C00313536
COURSE NAME:	Master in Applied Artificial Intelligence
DEPARTMENT:	Department of Computing
COURSE CODE:	ZPRGC5201
WORD COUNT:	2800
SUPERVISOR:	Jamal Tauseef
DATE OF SUBMISSION:	7 December 2024

Abstract- Machine learning has significantly increased house price prediction by delivering more accurate and data-driven insights into real estate decision-making. This project focuses on **predicting house prices using machine learning** algorithms to make decisions in the housing industry. Five regression and an ensemble model were developed and evaluated in this project: Linear Regression, Decision Tree, Support Vector Regression (SVR), Random Forest, Gradient Boosting, and a Voting Ensemble. Models were evaluated using the Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 score. Random Forest outperforms all other models with an MSE of 382.85, R^2 score of 0.757, and an MAE of 12.33, which was good in identifying underlying patterns in data. In contrast, the SVR performed poorly with an R^2 value of 0.116 and the highest MAE being 25.511. The results show how effective ensemble models and tree-based algorithms are in predicting house prices.

Keywords- Machine Learning, House price prediction, Regression, model, Linear Regression, Decision Tree, Support Vector Regression, Random forest, Gradient boosting, Voting Ensemble, mean square error, Mean Absolute Error, R^2 score.

I. INTRODUCTION

The economy depends on the real estate market which has an impact on the national economic stability, business strategies, and individual financial decisions [1]. One of the major challenges in this is properly predicting house prices. Housing prices are usually influenced by several factors such as location, property size, area, and market values. Normal methods of price prediction usually struggle to handle the relationships between these variables.

The recent advancements in technology have made machine learning a powerful tool for predictive analytics with high levels of accuracy. Machine learning algorithms can examine large amounts of data, recognize hidden patterns, and produce useful insights to make the algorithms effective in house price predictions. This project predicts house prices using machine learning methods and lays a good foundation for data-driven decision-making in the real estate sector.

The main aim of this study is to determine the most effective machine learning model for house price prediction. By studying

results, this project not only shows the efficiency of machine learning techniques but also offers information about their benefit in the real estate industry. These findings illustrate the potential of machine learning in this field and how it can assist real estate professionals, buyers, and sellers in making well-informed decisions.

This report studies the performance of 6 different models and contains sections on methodology, data preparation, results, discussion of findings, conclusion, and future works.

II. RESEARCH QUESTIONS

A. How do different structural features impact the price of houses in the dataset?

B. What role do location attributes play in determining the price of residential properties?

III. PROBLEM STATEMENT

Analyzing how the design and whereabouts of homes affect their prices is crucial for making informed decisions in the housing market world. Understanding the impact of factors such as bedroom count and distance to water bodies can assist individuals like buyers and sellers in manipulating markets skillfully. The first research question delves into aspects vital for property assessment, While the second sheds light on location characteristics underscoring the role of position, in the real estate valuation process. This study adds value to real estate analysis by providing insights that can enhance market forecasts and investment tactics. Moreover, it maintains standards by utilizing accessible data sources ensuring transparency throughout the research process.

IV. RELATED WORK

Several studies have examined various machine learning techniques for predicting house prices, often leveraging datasets such as the Boston housing dataset. For instance, [2] developed a decision tree model and concluded that factors such as the number of houses, population quality, location, education, and crime rates significantly influence house prices. Similarly, [3] compared Linear Regression, Decision Tree, and Random Forest models on the Boston dataset. While their findings highlighted the relative strengths of tree-based models, the

study acknowledged that the dataset's small sample size restricted the generalizability of the results to broader contexts.

[4] employed a Random Forest model using the UCI Boston dataset and achieved a commendable error margin of ±5%. However, the study faced challenges due to the limited geographic scope and dataset size, which hampered its applicability to larger or more diverse markets. Other researchers, such as [5] and [6], explored more advanced algorithms like XGBoost and call tree regression, respectively. Chowhaan's work, in particular, demonstrated the capability of XGBoost to handle intricate data structures effectively. However, these studies also noted outliers and regional constraints that impacted the models' accuracy.

The results suggest that while machine learning models show promise in housing price prediction, their full potential remains untapped. Future research should prioritize using larger, more diverse datasets and developing techniques to manage data anomalies and outliers better. Addressing these limitations could enhance the accuracy, robustness, and real-world applicability of machine learning models in predicting house prices across varying markets and conditions.

V. METHODOLOGY AND IMPLEMENTATION

A. Data Collection

The first phase of the project was obtaining the house price dataset used for prediction purposes. The dataset was sourced from Kaggle and authored by Anmol Kumar and comprises 29,000 samples of data and 12 features. The data was imported into Python and stored as a DataFrame, referred to as "df."

	landlord	construction	Govt approved	rooms	property type	sqr ft	ready to move	resale	address	long	lat	price
0	Owner	0	0	2	BHK	1308.236407	1	1	Kshl Layout,Bangalore	12.969910	77.597960	55.0
1	Dealer	0	0	2	BHK	1275.000000	1	1	Vishweshwara Nagar,Mysore	12.274538	76.644605	51.0
2	Owner	0	0	2	BHK	933.159722	1	1	Jigani,Bangalore	12.778033	77.632191	43.0
3	Owner	0	1	2	BHK	929.921143	1	1	Sector-1 Vaishali,Ghaziabad	28.642300	77.344500	62.5
4	Dealer	1	0	2	BHK	999.009247	0	1	New Town,Kolkata	22.592200	88.484911	60.5

Fig. 1. Importing Dataset

The dataset used has 29,451 samples of data, 2 features, and each feature provides different information concerning the houses for price prediction. The description of each of the features in detail is as follows:

FEATURE NAME	DESCRIPTION
Landlord	Category marking who has listed the property
Construction	Under Construction or Not
Govt_Approved	Government-approved or Not
Rooms	Number of rooms
Property type	Type of Property
Sqr_ft	Square foot of the house
Ready_to_move	Category marking if the house is ready to move into
Resale	Category marking if the house is on resale or not
Address	The address of the property
Long	Longitude of the property
Lat	Latitude of the property
Price	The price of the property (target)

Table 1. Features

All of these features provide a wide range of information about the property, allowing for a thorough analysis to forecast house prices depending on a number of variables, including size, location, condition, and type of property. The goal of the models is to predict the "Price," by using the other feature values.

B. Data Preprocessing

1) Removing Duplicates

The first step in the data preprocessing stage involved identifying and removing duplicate data from the dataset to reduce redundancy. Duplicate data introduce bias to the data by over-representing certain samples. There were 401 duplicate records removed from the dataset leaving 29050 samples of data. This step was necessary to ensure that each sample in the dataset represents unique samples.

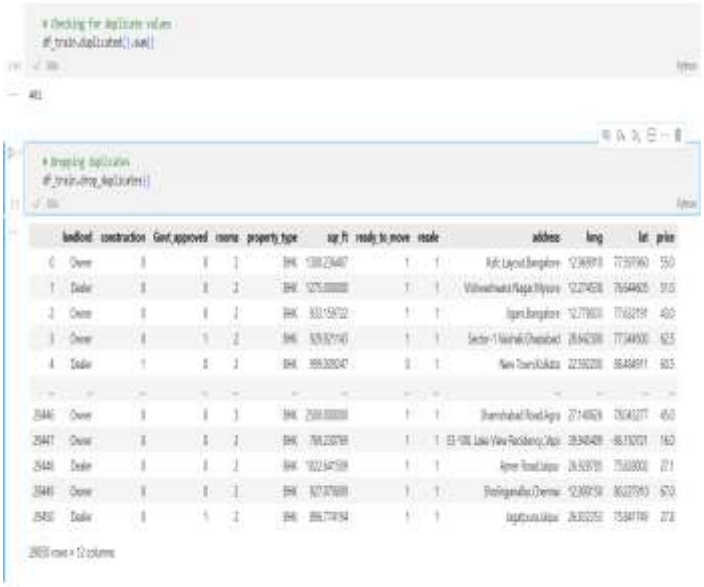
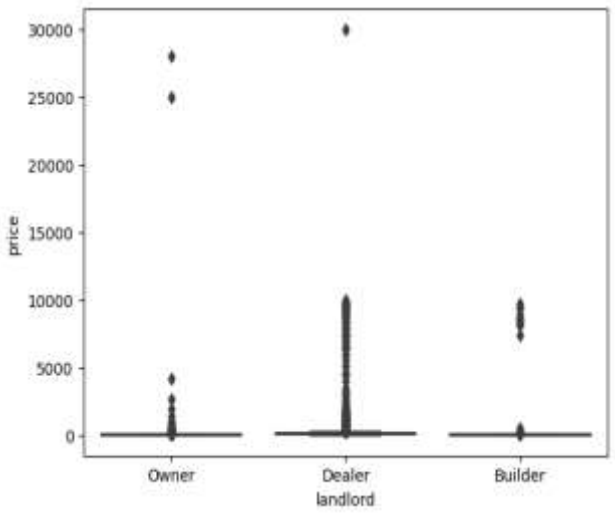


Fig. 2. Dropping duplicate values

2) Visualizing and Removing Outliers

Outliers are values in the data which deviate from the rest of the data. These outliers negatively impacted the performance of the machine learning models by skewing them. Boxplots and Scatterplots are very useful visualizations for identifying outliers in data. To address this issue the dataset was checked for outliers and removed to make the data evenly distributed. To visualize the outliers, the price was plotted against the number of Square feet and landlord of the house using a box plot and scatterplot.



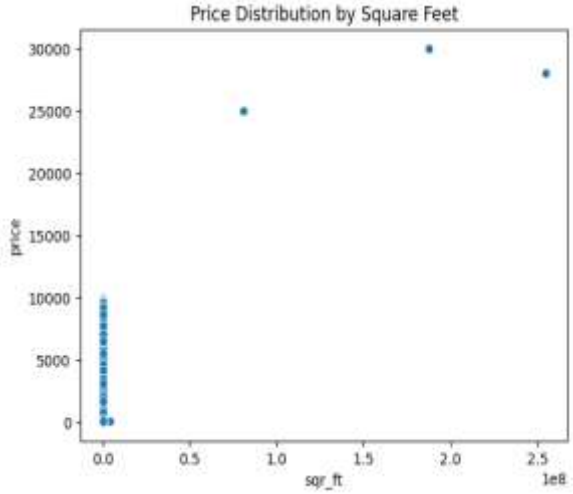


Fig. 3. Detecting Outliers

The Inter Quartile Range was used to find and remove the outliers in the data. The Price and Square foot column had values outside of the range and were removed using the interquartile formula, with Q1 and Q3 as the first and third quartiles respectively.

```
# Removed the outliers in the dataset for effective distribution

def remove_outliers_iqr(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]

# applying the function to the price column and sqr_ft column
df_train = remove_outliers_iqr(df_train, "price")
df_train = remove_outliers_iqr(df_train, "sqr_ft")
✓ 0.0s
```

Fig. 4. Removing Outliers

The values that were below the first quartile Q1 and above the third quartile Q3 were removed for both the Price and Square foot columns. In doing this, the dataset was filtered down to 25,575 samples of data which gave a more robust and high-quality dataset to work on.

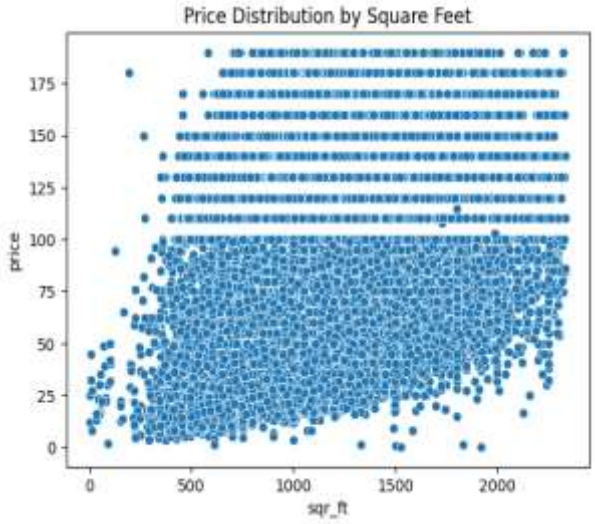


Fig. 5. Filtered Dataset from outliers

3) Feature Engineering

Feature engineering was done on the dataset to make it more normal and consistent for the models. Two techniques were carried out to improve the predictive capability of the models by designing and encoding new features:

Extracting city information from the address:

The address column contained full addresses for each property. To simplify it, the city name was extracted from the address with string manipulation. This involved splitting the data at the commas and selecting the last segment of the string to get the city name.

Fig. 6 Feature Engineering

Encoding Categorical Variables:

The landlord, property type, and city information were converted to numerical format using one-hot encoding as seen in Fig. 6. This technique creates binary columns for all categories and makes it suitable for the machine learning model. To prevent

multicollinearity in the data, the first category for each feature was dropped using the drop_first parameter.

4) Feature Scaling

Feature scaling was applied to the data to normalize the independent features. This is important because models are sensitive to the scale of the input features. First, the independent variables (x) were separated from the target variable (y) to be able to scale the independent variables. Then the independent variables (x) were standardized using the Standard scaler from the sklearn library. Standardizing changes the data so that it has a mean of 0 and a standard deviation of 0. This makes sure that the features with larger variables do not overcome the features with smaller variables.

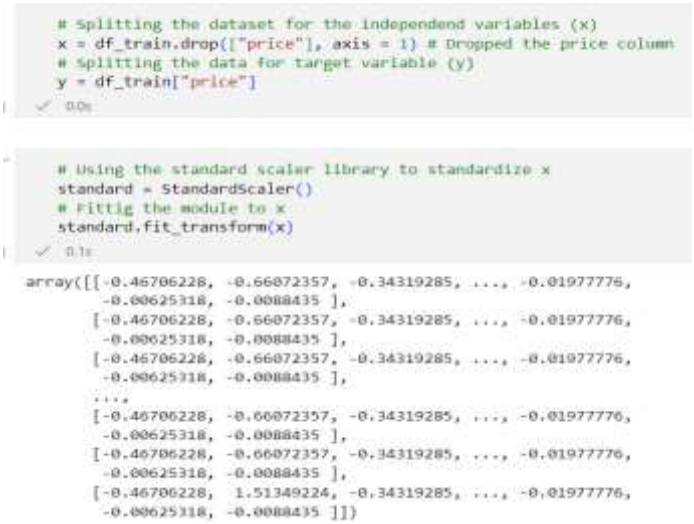


Fig. 7. Feature scaling

5) Data Splitting

The dataset was divided into a training and testing set using the train_test_split function from the sklearn library to estimate the performance of the machine learning algorithms. A split of 80% for training and 20% for testing was developed to let the models learn from the main dataset while still being evaluated on an unseen subset. The training set is where these models were trained, whereas the test set was kept for performance evaluation. A random_state was set to 42 so that the splits would provide reproducibility and consistency on multiple runs. This is an important step toward making the machine learning workflow robust and unbiased.

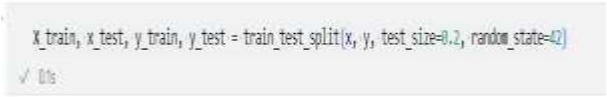


Fig. 8. Data Splitting

6) Model Training

Six machine learning models were initialized and trained with the dataset to predict house prices. Each of the models had specific characteristics that provided some benefits for the project, making sure there was diversity in the approaches. The models developed included linear regression,

decision trees, support vector regression, random forests, gradient boosting, and an ensemble model by voting.

Linear regression is the simplest and most interpretable algorithm initiated to check the assumption of linearity between the independent variables and the dependent ones. Thereafter, a decision tree regressor was introduced, as it naturally handles non-linearities in relationships due to segmentations of data into decision nodes based on feature values, making a tree-structured predictive model. Further, Support Vector Regression was applied to see if it could find the optimal hyperplane within a predefined margin, especially on data with complicated relationships.

There are also two ensemble models. The Random Forest Regressor which uses the idea of making the integral of multiple decision trees results in a better and more stable forecast. The Gradient Boosting Regressor builds decision trees sequentially, correcting the error that the previous decision tree made. Finally, the Voting Regressor combined the outcomes of all the models to come up with an average. This method leverages the strengths of individual models to provide robust and reliable predictions.

All the models were trained using the training dataset. This lets the models extract patterns from the data to allow them to predict house prices. By including these several models, this project allowed for comparing their performances and selection of the most effective model for predicting house prices.

Fig. 6. Model Training

7) Evaluation

Each of the models was evaluated using three evaluation metrics which were: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 Score. The MSE measures the difference between predicted and actual values. A lower value means a better result. MAE gives the average absolute deviation, hence presenting an interpretable measure of the prediction accuracy. It shows how far off the predictions are from the actual values. The R^2 Score shows how much a given model explains the variance in house prices. This score improves with values closer to 1.

Each of these methods was then implemented on the dataset, where the MSE, MAE, and R^2 of all these models were calculated. Further, a summarization of results into one table was made for easy comparisons. In this way, model performances were compared to find the best approach in the estimation of house prices.

VI. TECHNICAL SCENARIO

Smart Real Estate Investment Platform

A. Context

A real estate company wants to develop a smart platform that will help investors identify properties that are currently undervalued and want to make predictions about the future value of properties based on their size, location and other varying factors. This platform is designed to offer price estimations to enable them to make the right decisions.

B. Problem

Investors as well as home buyers have a hard time figuring out whether a certain property is overpriced or has the potential to increase in value in the future. These approaches involve the use of manual comparison of sales data and the market which is very tedious and inaccurate.

C. Proposed Solution

A machine learning-based house price prediction model can be implemented into the platform to deliver automated and precise price appraisals.

VII. RESULTS

After performing the evaluation for the various models, Random Forest had the best, with the minimum MSE of 382.85, the maximum R^2 at 0.757, having the least MAE of 12.33, hence being most accurate regarding house price prediction. The Voting

Ensemble model showed a good MSE of 463.57 and an R^2 of 0.706, therefore elaborating on the results regarding ensemble learning.

Decision Tree and Gradient Boosting showed similarly poor results with an MAE of about 15, though it was worse than a Random Forest. Linear Regression ran fairly well at MSE 615.58, though it was again outperformed by more complex models. Finally, the Support Vector Regression performed poorly with an MSE of 1392.82 and the highest MAE.

Model	**MSE**	**R^2**	**MAE**
Linear Regression	615.576222	0.609178	17.875300
Decision Tree	613.747910	0.610338	15.178658
Support Vector Regression	1392.819513	0.115715	25.512305
Random Forest	382.845882	0.756935	12.325959
Gradient Boosting	494.380900	0.686123	15.696498
Voting Ensemble	463.572452	0.705683	14.519205

Table 2. Results

In conclusion, as shown in table 2. Random Forest was the mode effective model for predicting house prices, while the voting ensemble model shows to be an alternative.

VIII. CONCLUSION AND FURTHER WORK

This study demonstrated the application of machine learning models to predict house prices, showing the potential of machine learning algorithms in addressing real-world scenarios. The results showed clear differences in the model performances, with Random Forests as the most accurate model for predicting house prices. The Ensemble Voting Regressor also proved to be effective.

Future works should incorporate larger and more diversified data to improve the findings' generalizability. Incorporating other variables like real estate trends would yield better predictability. More approaches such as neural networks and deep learning, which have yielded good results in nonlinear relations should be implemented.

Implementing all the factors mentioned above will extend the possibilities of machine learning models applied in real housing market forecasting.

REFERENCES

[1]. Zhao, C. and Liu, F. (2023) 'Impact of housing policies on the real estate market - Systematic literature review', *Heliyon*, 9(10), p. e20704. Available at: https://doi.org/10.1016/j.heliyon.2023.e20704. 2.10031728

[2]. Zhang, Z., 2021. Decision Trees for Objective House Price Prediction, in: 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). Presented at the 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), IEEE, Taiyuan, China, pp. 280–283.

[3]. Department of ICT, Comilla University, Cumilla , Bangladesh., Begum, A., Kheya, N.J., Department of ICT, Comilla University, Cumilla, Bangladesh., Rahman, Md.Z., Department of CSE, Britannia University, Cumilla, Bangladesh., 2022. Housing Price Prediction with Machine Learning. Int. J. Innov. Technol. Explor. Eng. 11, 42–46. https://doi.org/10.35940/ijitee.C9741.0111322

[4]. Adetunji, A.B., Akande, O.N., Ajala, F.A., Oyewo, O., Akande, Y.F., Oluwadara, G., 2022. House Price Prediction using Random Forest Machine Learning Technique. Procedia Comput. Sci. 199, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

[5]. Chowhaan, M.J., Nitish, D., Akash, G., Sreevidya, N., Shaik, S., 2023. Machine Learning Approach for House Price Prediction. Asian J. Res. Comput. Sci. 16, 54–61. https://doi.org/10.9734/ajrcos/2023/v16i2339

[6]. Gupta, A., Dargar, S.K., Dargar, A., 2022. House Prices Prediction Using Machine Learning Regression Models, in: 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC). Presented at the 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), IEEE, Tumkur, Karnataka, India, pp. 1–5. https://doi.org/10.1109/ICMNWC56175.202

https://doi.org/10.1109/MLBDBI54094.202
1.00059