Data Article

HOUSE PRICE PREDICTION USING MACHINE LEARNING

Abstract: In our ecosystem, real estate is clearly a distinct industry. Predicting house prices, significant housing characteristics, and many other things is made a lot easier by the capacity to extract data from raw data and extract essential information. Daily fluctuations in housing costs are still present, and they occasionally rise without regard to calculations. According to research, changes in property prices frequently have an impact on both homeowners and the real estate market.

To analyze the key elements and the best predictive models for home prices, literature research is conducted. The analyses' findings supported the usage of artificial neural networks, support vector regression, and linear regression as the most effective modeling techniques. Our results also imply that real estate agents and geography play important roles in determining property prices. Finding the most crucial factors affecting housing prices and identifying the best machine learning model to utilize for this research would both be greatly aided by this study, especially for housing developers and researchers.

Keywords: House price prediction, linear regression, Machine learning

1. INTRODUCTION

In this report, we propose our system “House price prediction using Machine Learning”.Along with other fundamental requirements like food, water, and many other things, a place to call house is one of a person's most basic wants. In the real estate sector, predicting house prices is essential to work since it aids buyers and sellers in making wise choices. Numerous algorithms have been created to accurately anticipate property prices thanks to advances in machine learning. In this research, we use a dataset of real estate properties along with XGBoost, an advanced gradient boosting technique, to forecast house values. Powerful algorithm XGBoost effectively manages structured datasets. It has been found to perform well in forecasting complex datasets and has been utilized in a number of machine-learning competitions. In this experiment, we used XGBoost to solve the problem of predicting housing prices and assessed its effectiveness.

The aim of house price prediction is to create a model that can precisely estimate the price of a new house based on its attributes using previous data on house features (such as square footage, number of bedrooms and bathrooms, location, etc.) and their corresponding prices.In this project, we have applied these five algorithms namely linear regression, support vector machine, Lasso regression, Random Forest and XGBoost to predict house prices using a dataset of real estate properties. Because it can handle a large number of characteristics and capture intricate correlations between the features and the target variable (price), XGBoost is an effective algorithm for this purpose.

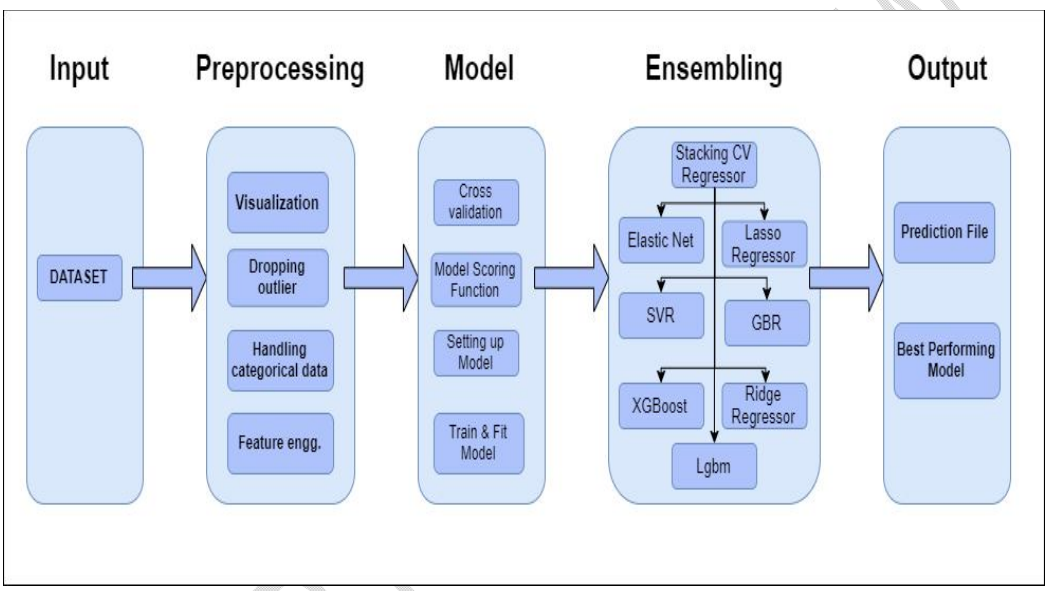

Fig 1. Flow of Execution

Explanation:

Input: The input section represents the initial stage of the house price prediction process. Here, one has to gather relevant data that could influence house prices, it can be the dataset. This data may include factors such as the size of the house, number of bedrooms, location, neighborhood amenities, historical sales data, and other relevant features.

Preprocessing: In the preprocessing stage, the collected data goes through various cleaning and transformation steps to ensure its quality and suitability for analysis. This involves tasks like handling missing values, removing outliers, normalizing or scaling the data, and encoding categorical variables. Preprocessing helps to prepare the data for effective modeling.

Model: The model section represents the core of the house price prediction process. Here, one has to select an appropriate machine learning algorithm or ensemble of algorithms to build a predictive model. Commonly used algorithms for house price prediction include linear regression, decision trees, random forests, support vector machines, or neural networks. The model takes the preprocessed data as input and learns patterns and relationships within the data to make predictions on house prices.

Ensembling: Ensembling refers to the practice of combining multiple predictive models to improve the accuracy and robustness of the predictions. In this stage, one has to employ techniques such as averaging, bagging, boosting, or stacking to create an ensemble model. By leveraging the strengths of different models, ensembling aims to achieve more accurate and reliable predictions by reducing bias and variance.

Output: The output section represents the final stage of the house price prediction process. Here, the trained model or ensemble provides predictions on house prices based on the given input data. The predictions can be in the form of specific price values or in percentage.

2. PROBLEM STATEMENT

The asking price and general description are frequently presented independently from the generic and standardized real estate attributes. These qualities may be easily compared across the entire spectrum of potential houses because they are given separately and in a systematic manner. House sellers might list a summary of all the key aspects of the house in the description because every house also has distinctive elements, such as a particular view or style of washbasin. Potential purchasers can take into account all provided real estate features, but owing to the great diversity, it is almost not possible to provide an automatic comparison of all variables.This also applies in the opposite direction: house sellers must evaluate the worth based on the attributes of the house in relation to the current market price of comparable houses. It is difficult to determine a fair market price due to the variety of features. In addition to outlining the property's essential features and capturing the reader's curiosity, the house description functions as a persuasive tool.

Housing prices are a significant indicator of the health of the economy, and both buyers and sellers are keenly interested in price points. In this study, explanatory variables that encompass a wide range of residential dwelling characteristics will be used to forecast

house values. The objective of this project is to develop a regression model that can precisely calculate the house's price given its attributes.

3. LITERATURE REVIEW

1. Sushant Kulkarni. (2021)Testing the dataset using four distinct retrogression algorithms—Velicet Lasso Regression, Logistic Retrogression, Decision Tree, and Support Vector Regression—is one of the approaches suggested in the study by Neelam Shinde and Kiran Gawande. When comparing error criteria such as R-Square Value, Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error, Decision Tree emerged as the fashionable algorithm with the highest delicacy score of 86.4 and the lowest error values, while Lasso Regression performed the worst with a delicacy score of 60.32.
2. To predict the cost of resale homes, P. Durganjali suggested using classification algorithms. The selling price of a property is predicted in this study using a variety of classification methods, including Linear regression, Decision Tree, K-Means, and Random Forest. A home's price is influenced by its physical attributes, its geographic location, and even the state of the economy. Here, they apply these techniques, use RMSE as the performance matrix for different datasets, and find the best accurate model that predicts better results.
3. Bengaluru has been chosen by Manasa and Gupta as the case study city. The square footage of the property, its location, and its amenities are all significant determinants of price. There are 9 different qualities employed. For experimental work, Multiple Linear Regression (Least Squares), Lasso/Ridge Regression, SVM, and XGBoost are employed.
4. According to Panjali and Vani, especially for those who plan to live there for a long time before selling it again. It also applies to people who want no risks taken when building their houses. To determine the house's resale value, authors use a variety of classification techniques, including Logistic Regression, Decision trees, Naive Bayes, and Random Forest. Additionally, it uses the AdaBoost method to help weak students become strong ones. The resale price of a home is determined by its physical attributes, location, as well as numerous economic factors that are persuasive at the moment. In order to release the best-selling strategy for each dataset, accuracy is employed to gauge performance.

4. SYSTEM DESIGN AND ARCHITECTURE

Phase I: collection of data

In this phase, relevant data pertaining to house prices is gathered from reliable sources such as real estate websites and public datasets. The data may include features such as location, size, number of rooms, area_type, availability, and sale prices. Care should be taken to ensure the data is diverse and representative of the target market.

Phase II: Data pre-processing

This phase involves cleaning and preparing the collected data for model training. Tasks such as handling missing values, removing outliers, normalizing numerical features, and encoding categorical variables are performed. Feature selection techniques can be applied to identify the most relevant attributes for predicting house prices. Additionally, data splitting techniques such as stratified sampling can be used to create training and testing datasets.

Phase III: Training the model

In this phase, various machine learning algorithms are applied to train a predictive model using the pre-processed data. Common approaches include linear regression, decision trees, random forests, or more advanced techniques like gradient boosting or neural networks. The training process involves fitting the model to the training data, optimizing hyperparameters, and evaluating the model's performance using appropriate metrics such as mean squared error or R-squared.

The training set is used to train the machine learning model. It comprises a majority portion of the dataset and is used to teach the model the patterns and relationships between the input features (e.g., number of rooms, location, square footage) and the target variable (i.e., house prices). The model learns from the training set to make predictions.

Phase IV: Testing the model

Once the model is trained, it is evaluated using the testing dataset to assess its predictive capabilities. The model's performance is measured by comparing its predictions with the actual house prices in the testing set. Evaluation metrics such as mean absolute error or root mean squared error can be used to quantify the accuracy of the predictions.

The testing set, on the other hand, is a separate subset of the dataset that is used to evaluate the performance and generalization ability of the trained model. It is unseen by the model during the training phase and is used to assess how well the model can predict house prices on new, unseen data.

The division of the dataset into training and testing sets is typically done randomly, ensuring that the two subsets have similar distributions and characteristics. A common practice is to allocate around 80% of the data to the training set and the remaining 20% to the testing set and we followed the same.

Regarding the number of rounds of training, it depends on various factors such as the complexity of the dataset, the chosen machine learning algorithm, and the performance requirements. Generally, multiple rounds of training are effective to improve the model's accuracy and fine-tune its performance.

5. METHODOLOGY

To estimate housing values in this study, we used a number of well-known machine learning methods. Support vector machines (SVM), random forest, XGBoost, Lasso regression, and linear regression were some of the methods used in our investigation.

1. Algorithms: In the process of developing this model, various machine learning algorithms were studied. The model is trained on Support vector machines (SVM), random forest, XGBoost, Lasso regression, and linear regression . Out of this Random Forest gives highest accuracy in prediction of housing prices and the next highest accuracy achieved is by XGBoost algorithm and this algorithm is preferred due to its ability to handle complex, structured datasets and its ability to automatically handle missing values and outliers. Therefore, we recommend the use of XGBoost for house price prediction tasks in the real estate industry. The decision to choose the algorithm depends on the dimensions and type of data used. XGBoost is the best fit for our model.

2. XGBoost: The XGBoost observes features of an attribute and train the model by analyzing given features. XGBoost from the graph, attribute combination, labels including features and according to the system analyzes the data.

6. IMPLEMENTATION

Here are the steps that we followed in implementation

1. Data Collection:

Gather a dataset either from github or it will be also available on Kaggle, that includes relevant features of houses such as location, number of rooms, square feets, and sale prices. Ensure the dataset that has the features that you are about to consider.

2. Data Pre-processing:

Clean and prepare the collected data for model training. Handle missing values, perform feature scaling to bring features to a similar range, encode categorical variables, and address outliers. Additionally, one can explore feature engineering techniques to create new meaningful features.

```
[ ] df3.isnull().sum()
    df3.head()
```

	area_type	availability	location	size	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	1200	2.0	1.0	51.00

Fig 2. Data Pre-processing

3. Model Selection:

Choose a suitable machine learning algorithm for house price prediction, considering factors such as the dataset size, feature complexity, and interpretability requirements. Common algorithms include linear regression, decision trees, random forests, or more advanced techniques. We trained and tested five algorithms and finally XGBoost is performing better.

4. Exploratory Data Analysis:

In the exploratory data analysis (EDA) conducted for the house price prediction project, an image was generated to visualize the relationship between the variables "balcony," "bath," and "price."

Fig 3. Exploratory Data Analysis

The findings from this EDA analysis could be valuable for potential homebuyers, real estate agents, and property developers, as it sheds light on the factors that influence house prices

5. correlation heatmap:

In our exploratory data analysis (EDA) for house price prediction, we created a correlation heatmap to examine the relationships between the variables bath, balcony, and price. The correlation heatmap visually represents the strength and direction of correlations between these variables.

```
[ ]  num_vars = ["bath", "balcony", "price"]
     sns.heatmap(df[num_vars].corr(),cmap="coolwarm", annot=True)
```

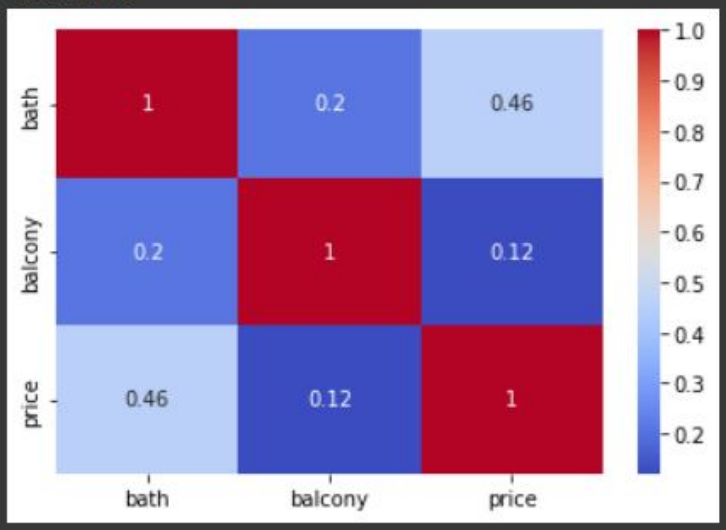

Fig 4. Correlation heatmap

The correlation heatmap reveals valuable insights regarding the influence of bath and balcony on house prices. We observed a positive correlation between the number of bathrooms and the price of the house, indicating that properties with more bathrooms tend to have higher prices. Additionally, we observed a positive correlation between the number of balconies and the house price, suggesting that houses with more balconies may command higher prices as well.

6. Training and Testing the model:

In the training and testing phase of our house price prediction model, we applied a comprehensive approach by training and testing the data using five different algorithms. This approach allowed us to evaluate the performance and effectiveness of each algorithm in predicting house prices.

The five algorithms employed in our study include linear regression, Lasso regression, XGBoost, random forest, and support vector machines (SVM). Each algorithm was trained on a portion of the preprocessed dataset and then tested on a dataset to assess its predictive accuracy.

By utilizing multiple algorithms, we aimed to capture a wide range of modeling techniques and identify the best-performing approach for our specific house price prediction task.

The training and testing phase involved tuning hyperparameters for each algorithm, using techniques such as cross-validation and grid search, to optimize their performance.

Evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared were utilized to compare and assess the accuracy and predictive power of the trained model.

7. RESULTS AND ANALYSIS

To use various machine learning algorithms for solving this problem.

Random Forest achieves a high accuracy score of 0.903 and a low root mean squared error (RMSE) value of 44.032. This suggests that the Random Forest model captures the underlying patterns and relationships in the data effectively, resulting in accurate predictions of house prices.

Similarly, XGBoost achieves a commendable accuracy score of 0.887 and a reasonably low RMSE value of 47.733. XGBoost is a boosting algorithm that builds an ensemble of weak learners iteratively. It has the ability to handle complex feature interactions and can effectively capture non-linear relationships, resulting in accurate predictions. The regularization techniques employed in XGBoost help prevent overfitting and improve generalization performance.

S.No	Model	Score	RMSE
1	Linear Regression	0.790384	64.898435
2	Lasso Regression	0.803637	62.813243
3	Support Vector Machine(SVM)	0.206380	126.278064
4	Random Forest	0.903507	44.032172
5	XGBoost	0.886607	47.732530

Fig 5. Model outputs

The superior performance of Random Forest and XGBoost can be attributed to their ability to handle high-dimensional datasets, capture complex relationships, and effectively manage feature interactions. These algorithms are known for their robustness, scalability, and versatility in handling a wide range of machine learning tasks.

8. CONCLUSION

The goal of the project "House Price Prediction Using Machine Learning" is to forecast house prices based on various features in the provided data. Our best accuracy was around 90% after we trained and tested the model. To make this model distinct from other prediction systems, we must include more parameters like tax and air quality. People can purchase houses on a budget and minimize financial loss. Numerous algorithms are used to determine house values. The selling price was determined with greater precision and accuracy. People will benefit greatly from this. Numerous elements that influence housing prices must be taken into account and handled.

References

1.https://www.researchgate.net/publication/347584803_House_Price_Prediction_using_a_Machine_Learning_Model_A_Survey_of_Literature

2. House price prediction using a hedonic price model vs an artificial neural network. American Journal of Applied Sciences. 3:193–201. Limsombunchai, Christopher Gan, and Minsoo Lee.

3. https://www.ijraset.com/research-paper/house-price-prediction-using-ml

4. https://ieeexplore.ieee.org/document/8473231

5. Fabian Pedregosa et al., "Python's Scikit-learn library for machine learning," Journal of Machine Learning Research, 12:2825–830.

6. Joep Steegmans and Wolter Hassink. an empirical investigation of how wealth and income affect one's financial status and ability to purchase a home. Journal of Housing Economics, 2017; 36:8–24.

7. Ankit Mohokar, Nihar Baghat, and Shreyash Mane. "House Price Forecasting Using Data Mining," International Journal of Computer Applications, 152:23–26.

8. Joao Gama, Torgo, and Luis. Logic regression using Classification Algorithms. Intelligent Data Analysis, 4, 275-292.

9. Real Estate Economics, 46:582–611, Heidelberg, Bork M. and Moller V. S., "House Price Forecast Ability: A Factor Analysis."

10. Hy Dang, Minh Nguyen, Bo Mei, and Quang Troung. Improvements to home price prediction methods using machine learning. Precedia Engineering, 174:433-442.

11. Atharva Chogle, Priyankakhaire, Akshata Gaud, and Jinal Jain. A article titled House Price Forecasting Using Data Mining Techniques was published in the International Journal of Advanced Research in Computer and Communication Engineering, 6: 24-28.

12. Kai-Hsuan Chu, Li, and Li. "Prediction of real estate price variation based on economic parameters," 2017 International Conference on.IEEE, Applied System Innovation (ICASI).

13. Jae Kwon Bae and Byeonghwa Park. Housing price forecast using machine learning algorithms, volume 42, pages 2928–2934.

14. Subhani Shaik and Dr. Uppu Ravibabu “Classification of EMG Signal Analysis based on Curvelet Transform and Random Forest tree Method” Paper selected for Journal of Theoretical and Applied Information Technology (JATIT), Vol. 95, December.

15. Shiva Keertan J and Subhani Shaik,” Machine Learning Algorithms for Oil Price Prediction”, International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-8.

16. KP Surya Teja, Vigneswar Reddy and Subhani Shaik,” Flight Delay Prediction Using Machine Learning Algorithm XGBoost”, Jour of Adv Research in Dynamical & Control Systems, Vol. 11, No. 5.

17. Dr. Subhani Shaik, Dr. K. Vijayalakshmi and Dr. Ramakanth Reddy “Location based house prediction using data science techniques”, Asian Journal of Advanced Research and Reports, Vol. 17, Issue 4.