House Price Prediction Using Machine Learning Techniques: A Data-Driven Approach

Ravi Joshi
Assistant Professor, Department of CSE, Global Institute of Technology, Jaipur
ravi.joshi@gitjaipur.com

Dr. Reema Ajmera
Professor, Department of CSE, Global Institute of Technology, Jaipur
reema.ajmera@gitjaipur.com

Pradeep kumar
Assistant Professor, Department of CSE, Global Institute of Technology, Jaipur
pradeep.kumar@gitjaipur.com

ABSTRACT: Accurate house price prediction is essential for effective decision-making in the real estate market. This study proposes a machine learning–based approach to estimate property prices using various features such as location, property characteristics, and market-related factors. Historical housing data is preprocessed through techniques such as handling missing values, encoding categorical variables, and scaling numerical attributes. Several machine learning algorithms, including Linear Regression, Decision Trees, Random Forest, Support Vector Machines, and Gradient Boosting, are implemented and evaluated using performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. The experimental results demonstrate that machine learning models can effectively predict house prices and identify key factors influencing property valuation. The findings highlight the potential of data-driven techniques in supporting accurate real estate market analysis and informed decision-making.

KEYWORDS: House Price Prediction, Machine Learning, SVM, ML Algorithms.

1. INTRODUCTION

The real estate market is one of the most important sectors of the global economy and represents a significant portion of individual wealth and investment portfolios. Property transactions, housing investments, and real estate development play a crucial role in economic growth and financial stability. In such a dynamic market environment, accurately forecasting house prices is essential for various stakeholders including homeowners, buyers, sellers, investors, financial institutions, and policymakers. In recent years, the increasing availability of large datasets and advancements in data analytics has transformed the way property valuation is performed. Traditional property valuation methods often rely on manual appraisal techniques and expert judgment, which may be time-consuming and prone to human error. With the development of machine learning and statistical modeling techniques, it has become possible to analyze large volumes of housing data and identify patterns that influence property prices.

House price prediction involves analyzing multiple factors such as property size, location, number of rooms, neighborhood characteristics, economic indicators, and market demand. Machine learning algorithms can process these complex datasets and build predictive models capable of estimating property values with improved accuracy. These models help

stakeholders make informed decisions regarding property investment, purchasing, selling, and financial planning.

This project focuses on developing a predictive system for house price forecasting using advanced data analytics and machine learning techniques. By analyzing historical housing data, the project aims to identify important factors affecting house prices and build models that can accurately estimate property values. The study integrates data preprocessing, feature engineering, model training, and evaluation techniques to create an effective prediction system. The insights obtained from this project can support better decision-making in the real estate market by providing reliable forecasts and highlighting key factors that influence housing prices. Such predictive systems contribute to improving market transparency and efficiency while assisting investors and policymakers in understanding housing market trends.

2. DATASET DESCRIPTION

The dataset used in this project contains historical information about real estate properties and is used to develop machine learning models for house price prediction. It includes various attributes such as location, property size, number of bedrooms and bathrooms, amenities, year built, and previous sale prices. The main objective is to predict the current market value of a property, which serves as the target variable. The dataset was obtained from Kaggle and consists of 1332 records with 8 features describing different characteristics of residential properties. Important features include location details (city, neighborhood, or zip code), property size, number of rooms, available amenities such as garages or swimming pools, construction year, and historical sale prices. These features provide valuable insights into the factors influencing property valuation.

Before training the machine learning models, several data preprocessing steps are performed. These include handling missing values, detecting and removing outliers, and performing feature engineering to create meaningful attributes such as property age or price per square foot. The dataset is then divided into training and testing sets, typically using an 80:20 ratio, to train the model and evaluate its performance. Additionally, exploratory data analysis (EDA) is conducted to understand the distribution of data and identify relationships between different features. Feature selection techniques are applied to determine the most relevant variables influencing house prices. Data visualization methods such as histograms, scatter plots, and heatmaps are used to identify trends and correlations within the dataset.

Further steps such as data quality checks, transformation of categorical variables into numerical form, and normalization of numerical features are applied to ensure data consistency and improve model performance. These processes prepare the dataset for building accurate and reliable machine learning models capable of predicting house prices effectively.

3. PROPOSED METHODOLOGY

Methodology for Home Valuation Forecasting Using Machine Learning:

Problem Definition: Clearly define the problem: Develop a machine learning model to predict the market value of residential properties based on various features. Identify the target variable: The target variable is the home valuation, representing the estimated current market price of the property.

Data Collection and Preprocessing: Gather relevant data: Collect historical data on real estate properties from trusted sources such as real estate databases or APIs. Clean the data: Handle missing values, remove duplicates, and address any inconsistencies in the dataset.

Feature Engineering: Create new features if necessary, such as calculating price per square foot or deriving the age of the property from the year built.

Exploratory Data Analysis (EDA): Understand the dataset: Explore the distribution of features, identify outliers, and detect any patterns or correlations. Visualize the data: Utilize histograms, scatter plots, and heatmaps to visualize relationships between variables and gain insights into the data.

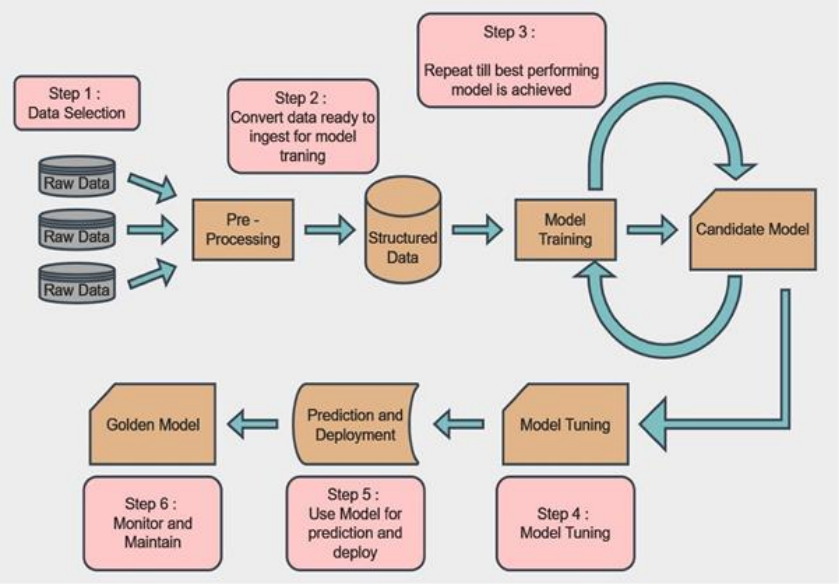

Figure 1: Proposed Model

Feature Selection: Identify relevant features: Determine which features have the most significant impact on home valuations. Techniques may include correlation analysis, feature importance from tree-based models, or domain knowledge.

Data Splitting: Divide the dataset into training and testing sets: Allocate around 80% of the data for training and 20% for testing. Ensure that the split maintains the distribution of the target variable to prevent bias in model evaluation.

Model Selection: Choose appropriate machine learning algorithms: Consider regression techniques such as linear regression, decision trees, random forests, or gradient boosting. Select models based on their ability to handle the complexity of the data and their Interpretability.

Model Training and Evaluation: Train the selected models using the training dataset. Evaluate model performance using appropriate metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score. Utilize cross-validation techniques to assess model generalization and robustness.

Hyperparameter Tuning: Optimize model hyperparameters using techniques like grid search or random search to improve model performance.

Model Interpretation: Interpret model results: Understand the coefficients or feature importance values to identify the factors driving home valuations. Analyze any patterns or trends observed in the predictions versus actual values.

Model Deployment: Deploy the trained model for real-world applications, such as providing home valuation estimates on a website or mobile app. continuously monitor and update the model to maintain its accuracy and relevance over time.

Documentation and Reporting: Document the entire process, including data sources, preprocessing steps, model selection criteria, and evaluation results. Prepare a comprehensive report summarizing the methodology, findings, and recommendations for stakeholders and future reference.

This methodology aims to leverage machine learning algorithms to develop a reliable and accurate predictive model for forecasting home valuations, thereby providing valuable insights for real estate stakeholders and industry professionals.

4. RESULTS AND DISCUSSION

A. Data Set

First five rows of the dataset by df.head() Function;

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

Figure 2: First five rows of the dataset by df.head() Function

Information by df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13320 entries, 0 to 13319
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   area_type     13320 non-null  object
 1   availability  13320 non-null  object
 2   location      13319 non-null  object
 3   size          13304 non-null  object
 4   society       7818 non-null   object
 5   total_sqft    13320 non-null  object
 6   bath          13247 non-null  float64
 7   balcony       12711 non-null  float64
 8   price         13320 non-null  float64
dtypes: float64(3), object(6)
memory usage: 936.7+ KB
```

Figure 3: First five rows of the dataset by df.head() Function

B. Data processing

In this phase, the missing attribute is handled by using the mean value. The target feature is dropout. By using Pandas library the operation is performed. Forvisualization of dataset graph use Matplotlib python function. After that try to catch some attribute combinations and set the missing values. We split the data in the proportion of 80% for Training and the remaining 20% use for Testing. Once data processingis done, create a suitable pipeline for the execution of the model

Loading a dataset of Bengaluru house data using Pandas and then plotting histograms for each feature using Matplotlib. However, there's a small syntax issue in your code. The function should be outside the loop. Here's the corrected version:

This code will load your dataset, print the first few rows, and then display histograms for each feature in the dataset. Make sure to replace path to your CSV file containing the Bengaluru house data.

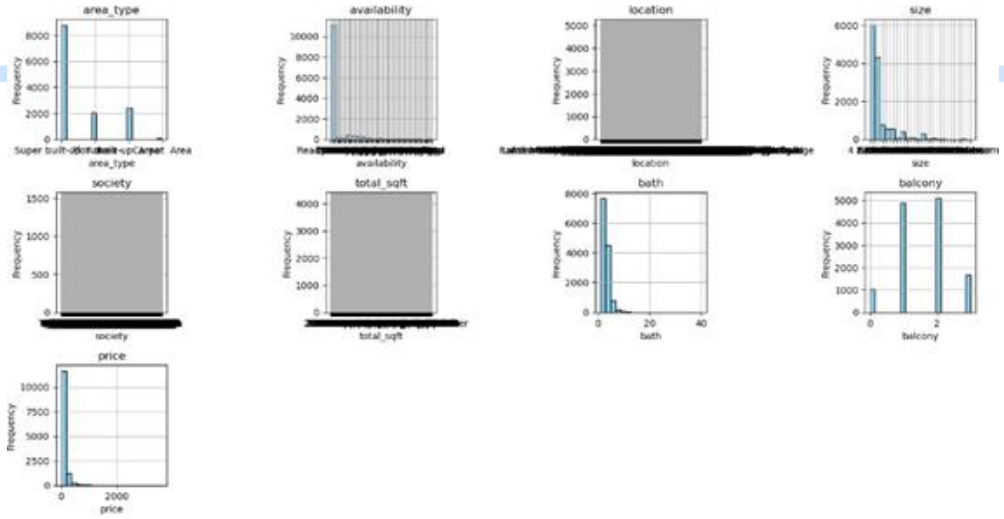

Figure 4: Visualization

C. Looking for correlations

We are trying to find out some new correlations between various attributes. The code loads a Bengaluru housing dataset from a CSV file, calculates the correlation matrix for all numeric features, and then creates a heatmap to visualize the correlations between features, including the target variable ('Price'). The heatmap helps identify relationships between features, indicating which ones are strongly correlated.

Here's a brief description of correlation:

- Strength of Correlation:
- Correlation values range from -1 to 1.
- A correlation of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
- A correlation of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- A correlation of 0 indicates no linear relationship between the variables.
- Direction of Correlation:
- Positive correlation: When one variable increases, the other variable tends to also increase. The correlation coefficient is positive.

- Negative correlation: When one variable increases, the other variable tends to decrease. The correlation coefficient is negative.
- Zero correlation: There is no systematic relationship between the variables.
- Interpretation:
- Correlation values closer to 1 or -1 indicate stronger relationships between variables.
- Correlation values closer to 0 indicate weaker relationships or no relationship between variables.
- A correlation of 0 does not necessarily mean there is no relationship between variables; it just means there is no linear relationship.

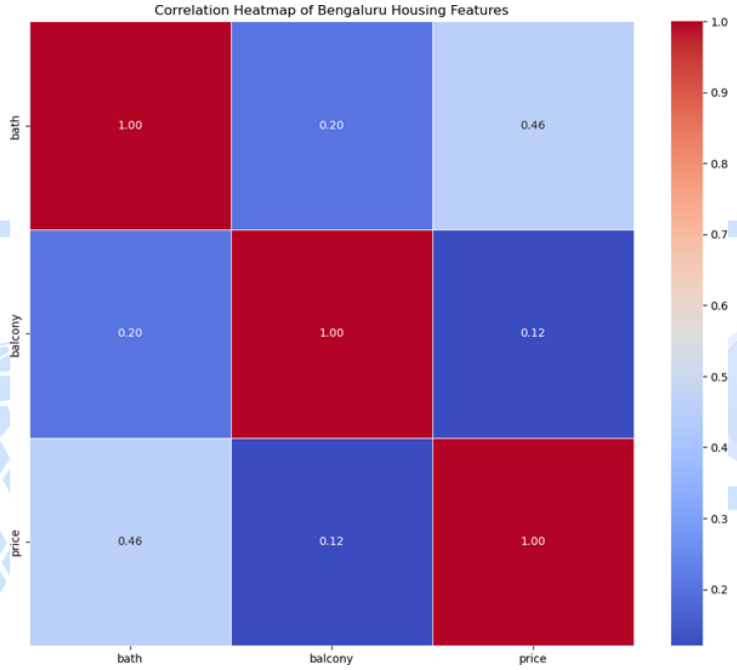

Figure 5: Heatmap

D. Finding Missing values

In the "After Handling Missing Values" section, I've used forward fill (method='ffill') as an example method for handling missing values. You can replace it with any other method that suits your data and analysis needs, such as mean imputation, median imputation, or dropping rows/columns with missing values.

Before handling the missing values:

```
Dataset Before Handling Missing Values:
             area_type   availability                  location       size  \
0  Super built-up  Area         19-Dec  Electronic City Phase II      2 BHK
1            Plot  Area  Ready To Move          Chikka Tirupathi  4 Bedroom
2        Built-up  Area  Ready To Move               Uttarahalli      3 BHK
3  Super built-up  Area  Ready To Move        Lingadheeranahalli      3 BHK
4  Super built-up  Area  Ready To Move                  Kothanur      2 BHK

   society total_sqft  bath  balcony   price
0  Coomee        1056   2.0      1.0   39.07
1  Theanmp       2600   5.0      3.0  120.00
2      NaN       1440   2.0      3.0   62.00
3  Soiewre       1521   3.0      1.0   95.00
4      NaN       1200   2.0      1.0   51.00
```

Figure 6: Before handling the missing values

After Handling the missing values:

Dataset After Handling Missing Values:

	area_type	availability	location	size	\
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	

	society	total_sqft	bath	balcony	price
0	Coomee	1056	2.0	1.0	39.07
1	Theanmp	2600	5.0	3.0	120.00
2	Theanmp	1440	2.0	3.0	62.00
3	Soiewre	1521	3.0	1.0	95.00
4	Soiewre	1200	2.0	1.0	51.00

Figure 7: Before handling the missing values

E. Fitting the Model

Description: Linear regression is a simple and widely used statistical technique for modeling the relationship between a dependent variable (often denoted as *y*y) and one or more independent variables (often denoted as *X*X). It assumes a linear relationship between the independent variables and the dependent variable.

Working Principle: The model tries to find the best-fitting straight line through the data points. It estimates the coefficients (weights) for each independent variable to minimize the difference between the actual and predicted values of the dependent variable.

```
import sklearn.linear_model as linear_model
LR = linear_model.LinearRegression()
test_model(LR)
```

```
[-5.005870284604971e+22]
```

```
# Cross validation
cross_validation = cross_val_score(estimator = LR, X = X_train, y = y_train, cv = 10)
print("Cross validation accuracy of LR model = ", cross_validation)
print("\nCross validation mean accuracy of LR model = ", cross_validation.mean())
```

```
Cross validation accuracy of LR model =  [-5.66875617e+18 -1.88227478e+17 -1.42898101e+16 -1.26026619e+18
 -9.26426701e+19 -4.59746420e+19 -8.52571745e+20 -1.48349165e+20
 -8.12256274e+19 -1.90227254e+18]

Cross validation mean accuracy of LR model =  -1.2297976618742694e+20
```

Random Forest:

Description: Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

Working Principle: It builds multiple decision trees by randomly selecting subsets of features and data samples. Each tree in the forest is trained independently using a technique called bagging (bootstrap aggregation). The final prediction is then made by averaging the predictions of all the trees (regression) or taking a majority vote (classification).

```
from sklearn.svm import SVR
svr_reg = SVR(kernel='rbf')
test_model(svr_reg)
```

```
[0.8466007618141084]
```

Support Vector Machine (SVM)

Description: Support Vector Machine is a powerful supervised learning algorithm used for classification and regression tasks. In the context of regression, it's often referred to as Support Vector Regression (SVR).

Working Principle: SVR tries to find a hyperplane in the feature space that has the maximum margin, i.e., the maximum distance between the hyperplane and the nearest data points (support vectors). The goal is to minimize the error while keeping the magnitude of the weights small.

```
from sklearn.ensemble import RandomForestRegressor
rf_reg = RandomForestRegressor(n_estimators = 1000, random_state=51)
test_model(rf_reg)
```

```
[0.8564798523284844]
```

5. CONCLUSION

In conclusion, house price prediction is an important application of machine learning in the real estate sector. This study explored different machine learning algorithms such as Linear Regression, Random Forest, and Support Vector Machine (SVM) to predict house prices based on factors like location, property size, and number of bedrooms. The results indicate that machine learning models can effectively analyze housing data and provide reasonably accurate predictions. The study also highlights the importance of proper data preprocessing, feature selection, and model evaluation to improve prediction accuracy. Among the models used, ensemble methods like Random Forest demonstrate better performance in handling complex relationships within the dataset. Overall, the proposed system shows that machine learning techniques can support buyers, sellers, investors, and real estate professionals in making informed decisions. Future work may focus on incorporating additional features such as economic indicators and real-time market data, as well as applying advanced techniques like deep learning to further enhance prediction accuracy.

REFERENCES

[1] G. Hu, J. Wang, and W. Feng, “Multivariate regression modelling for home value estimates with evaluation using maximum information coefficient,” International Journal of Data Analysis Techniques and Strategies, 2016.

[2] B. Park and J. K. Bae, “Using machine learning algorithms for housing price prediction,” Expert Systems with Applications, vol. 42, pp. 2928–2934, 2015.

[3] R. Joshi, A. Maritammanavar, "Deep Learning Architectures and Applications: A Comprehensive Survey", International Conference on Recent Trends in Engineering & Technology (ICRTET 2023), pp. 1-5, 2023.

[4] P. Jain, R. Joshi, "Bridging the Divide Between Human Language and Machine Comprehension", International Conference on Recent Trends in Engineering & Technology (ICRTET 2023), 2023.

[5] R. Ajmera et al., "Prediction analysis for diabetic patients using clustered based classification," Journal of Emerging and Innovative Research, vol. 5, no. 7, pp. 770–775, Jul. 2018.

[6] H. Sharma and R. Ajmera, "Comprehensive review and analysis on machine learning based Twitter opinion mining framework," Tuijin Jishu/Journal of Propulsion Technology, vol. 44, no. 5, 2023.

[7] H. Sharma and R. Ajmera, "Comprehensive review and analysis of elderly fall detection system using machine learning," Tuijin Jishu/Journal of Propulsion Technology, vol. 44, no. 5, 2023.

[8] X. Wang, J. Wen, Y. Zhang, and Y. Wang, "Real estate price forecasting based on SVM optimized by PSO," Optik – International Journal for Light and Electron Optics, vol. 125, no. 3, pp. 1439–1443, 2014.

[9] P. Pushpa, S. Khan IR, Q. Aziz, T. Anwar, and M. Arfath, "House pricing prediction using ML algorithm – A comparative analysis," International Journal of Scientific Research in Science and Technology, vol. 9, no. 2, pp. 262–266, 2022.

[10] S. Naik and A. Puthran, "Metro cities house price prediction using ML," in 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, pp. 588–594, 2023.