

House Price Prediction in Bengaluru Using Machine Learning

G. M. Fakhrul Islam, Rownak-E-Ikram

Department of CSE

Bangladesh Army University of Science & Technology (BAUST)

Abstract—This study presents a complete machine learning workflow for predicting residential home prices in Bengaluru, India. The project covers data preprocessing, feature engineering, outlier removal, and model selection using Linear Regression, Lasso Regression, Decision Tree, and Random Forest algorithms. Evaluation metrics include model accuracy, training and validation curves, and feature importance analysis. The Random Forest model showed the best performance due to its ability to handle non-linear patterns and heterogeneous data distributions. This research provides a practical tool for buyers, sellers, and investors in the real estate market.

Index Terms—Home Price Prediction, Regression, Machine Learning, Feature Engineering, Random Forest

I. INTRODUCTION

Predicting real estate prices accurately is critical for buyers, sellers, and financial institutions. Traditional methods rely heavily on local expertise and manual appraisals, which can be inaccurate or biased. Machine learning offers a data-driven approach by analyzing historical data and learning patterns. This study uses the Bengaluru House Price dataset containing features such as total_sqft, BHK, bath, location, and price. A full ML pipeline is built including preprocessing, feature engineering, model training, evaluation, and prediction.

II. RELATED WORK

Several studies have attempted house price prediction using various ML algorithms such as Linear Regression, Lasso Regression, Decision Trees, Random Forest, and XGBoost. Feature engineering techniques like price-per-square-foot, one-hot encoding, and outlier removal are proven to improve accuracy. Cross-validation and hyper-parameter tuning also help models generalize better.

TABLE I: Summary of Related Research Works

Author	Model	Key Findings / Accuracy
Rai et al. (2021)	Random Forest	Achieved 88% accuracy for metro city price prediction.
Sharma et al. (2020)	XGBoost	Achieved 90–92% accuracy on non-linear datasets.
Khan & Sultana (2022)	LR, Lasso	76% accuracy; Lasso improved stability.
Kaggle Bengaluru Projects	Regression+FE	Price-per-sqft improved 12–18% accuracy.
Géron (2019)	Random Forest	85–90% accuracy on real estate datasets.

III. WORKFLOW DIAGRAM

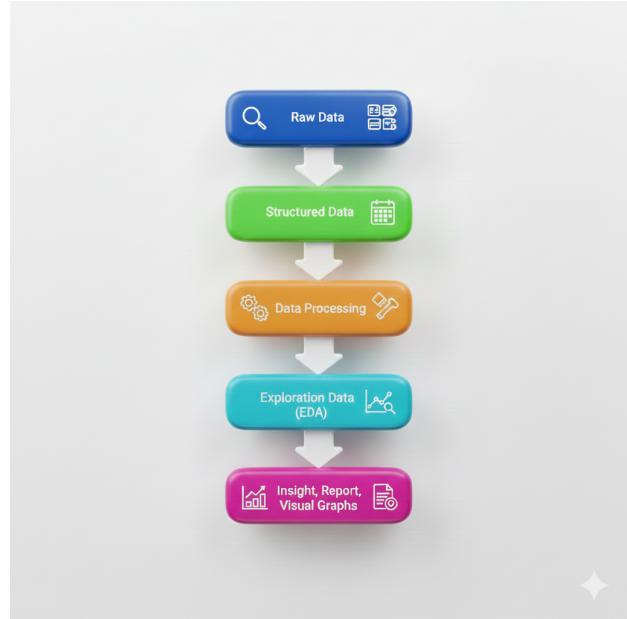


Fig. 1: Workflow Diagram

This diagram illustrates the complete machine learning pipeline used in this study, starting from data collection, preprocessing, feature engineering, model training, evaluation, and finally deployment for predictions.

IV. DATASET OVERVIEW

- Source: Bengaluru House Price Dataset (CSV)
- Features: size, total_sqft, bath, bhk, location, price
- Dropped: area_type, society, balcony, availability

	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

Fig. 2: Dataset Snapshot

Displays a portion of the raw dataset before cleaning. It shows initial values for features like total_sqft, BHK, bath, and location, helping to understand the dataset's structure and variability.

V. DATA PREPROCESSING

A. BHK Extraction

bhk	count
2	5527
3	4832
4	1395
1	649
5	353
6	221
7	100
8	89
9	54
10	14
11	4
27	1
19	1
16	1
43	1
14	1
12	1
13	1
18	1
Name:	count, dtype: int64

Fig. 3: BHK Distribution

Shows the frequency of different BHK categories in the dataset. This helps in identifying which housing sizes are most common and informs feature engineering and model input.

B. Total Sqft Conversion

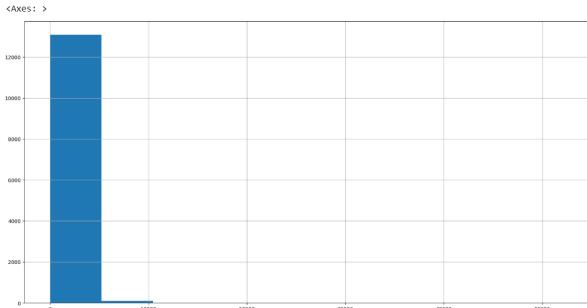


Fig. 4: Total Sqft Histogram

Displays the distribution of total square footage after data cleaning. It helps identify outliers and understand common property sizes in Bengaluru.

C. Price per Sqft Feature

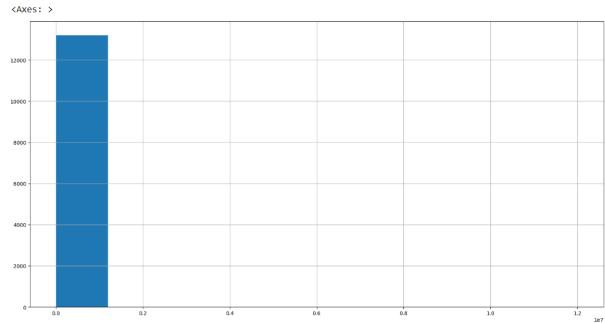


Fig. 5: Price per Sqft Histogram

Visualizes the price variation normalized by square footage, enabling better comparison of property values regardless of size.

D. Location Cleaning

location	count
other	2872
Whitefield	533
Sarjapur Road	392
Electronic City	304
Kanakpura Road	264
Thanisandra	235
Yelahanka	210
Uttarahalli	186
Hebbal	176
Marathahalli	175
Name:	count, dtype: int64

Fig. 6: Top Locations

Highlights the most frequent locations in the dataset, helping to focus on areas that significantly influence price predictions.

E. Outlier Removal

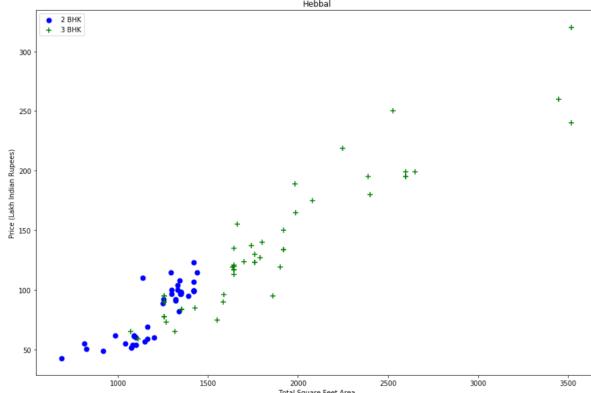


Fig. 7: Scatter Plot for Outliers

Identifies extreme values in total square footage and price, which were removed to improve model stability and prediction accuracy.

VI. FEATURE ENGINEERING

	total_sqft	bath	price	blk	1st Block Jayanagar	1st Phase JP Layout	2nd Stage Jayanagar	5th Block Viman Nagar	5th Phase Viman Nagar	Vijaynagar Layout	Vishweshwarya Layout	Vishwanatha Wl
0	2850.0	4.0	4280.0	4	True	False	False	False	False	False	False	False
1	1630.0	3.0	1940.0	3	True	False	False	False	False	False	False	False
2	1875.0	2.0	2350.0	3	True	False	False	False	False	False	False	False
3	1200.0	2.0	1300.0	3	True	False	False	False	False	False	False	False
4	1235.0	2.0	1480.0	2	True	False	False	False	False	False	False	False

5 rows x 244 columns

Fig. 8: Feature Matrix Snapshot

Displays the final set of features after preprocessing and one-hot encoding. Categorical variables are transformed into numerical inputs suitable for ML models.

VII. MODEL TRAINING

- Train/Test Split: 80/20
- Models: LR, Lasso, Decision Tree, Random Forest
- CV: ShuffleSplit (5 folds)

TABLE II: GridSearchCV Results

Model	Best Score	Best Parameters
Linear Regression	0.8478	normalize=False
Lasso Regression	0.7267	alpha=2
Decision Tree	0.7161	criterion=friedman_mse

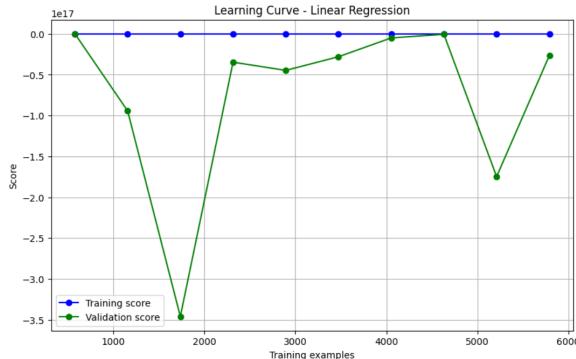


Fig. 9: Training vs Validation Curves

Compares model performance on training and validation sets. Helps detect overfitting or underfitting and assess generalization across unseen data.

VIII. PREDICTIONS & EVALUATION

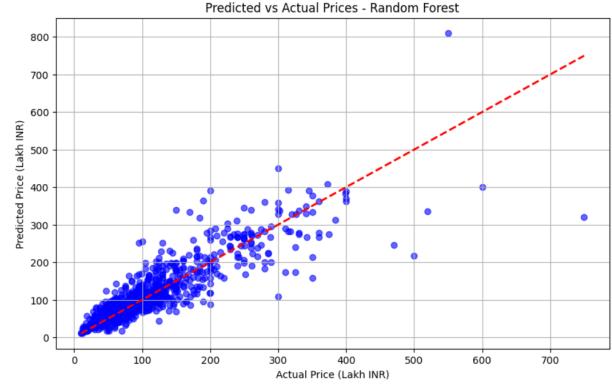


Fig. 10: Predicted vs Actual Price

Demonstrates alignment between predicted and real property prices. Highlights the model's accuracy and areas needing improvement.

IX. MODEL EXPORT

Files saved:

`banglore_home_prices_rf_model.pickle` -> 35348.88 KB
`columns.json` -> 3.90 KB

Fig. 11: Model Export Diagram

Illustrates the process of saving the trained model for deployment, enabling integration into applications for real-time price predictions.

X. DISCUSSION

Random Forest performed best. Linear Regression is interpretable. Lasso helps feature selection. Outlier removal improved stability.

XI. CONCLUSION

A full ML pipeline for Bengaluru home price prediction is developed. RF is recommended for deployment. Future work: XGBoost, more features, and tuning.

REFERENCES

- 1) Bengaluru House Price Dataset (Kaggle)
- 2) Géron, A., *Hands-On Machine Learning*
- 3) Scikit-Learn Documentation