JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PROPERTY PRICE ESTIMATOR USING MACHINE LEARNING

A Phase 1 Report on Developing a Predictive Model for Bangalore Real Estate

[1]DHANVISHYAM U B, [2]Dr. Shashidhar Kini K
[1]Student, [2]Professor & Head
[1]Department of Master of Computer Applications,
[1]Srinivas Institute of Technology, Valchil Mangaluru, Karnataka, India

Abstract : This study undertakes the development of a Property Price Estimator for the Bangalore residential real estate market using machine learning regression models. To predict property prices, key features derived from property listings are used. These features include location, size, number of bedrooms, bathrooms, and property type. For this purpose, contemporary cross-sectional data was scraped from Indian real estate portals (Makaan.com, Commonfloor.com). The analytical framework involves data collection, preprocessing, feature analysis, and the training and evaluation of predictive models like Linear Regression and XGBoost.

IndexTerms **- Property Price Prediction, Machine Learning, Regression Analysis, XGBoost, Linear Regression, Data Scraping, Real Estate Analytics, Bangalore Housing Market**

I. INTRODUCTION

Real estate investment is a significant financial decision and often reflects wealth and status. Unlike more volatile assets, property values tend not to decline abruptly, making real estate an attractive investment alternative [9]. However, accurately determining the value of a property is crucial for all stakeholders involved, including investors, banks, policymakers, buyers, and sellers.

The advent of online real estate portals has provided vast amounts of information, yet significant discrepancies often exist in listed prices, even for similar properties, leading to confusion for potential buyers [7]. Furthermore, relying solely on real estate agents can incur fees and may not always align with the buyer's best interests [9]. For sellers, accurately pricing a property requires extensive market comparison, a time-consuming process prone to errors, even within large real estate firms [9]. These challenges highlight a need for automated, data-driven approaches to property valuation.

This project addresses the need for precise forecasting of residential property sale prices within the specific market of Bangalore, India. The primary goal is to develop a system that estimates accurate and justified prices, thereby promoting transparency for both buyers and sellers [7]. The system aims to leverage machine learning techniques trained on relevant, localized data.

Accurate house price predictions offer substantial benefits:

- **Buyers:** Can set realistic budgets and negotiate fair prices [7].
- **Sellers:** Can set competitive asking prices and make informed decisions about renovations [7].
- **Real Estate Agents:** Can better advise clients and negotiate deals effectively [8].
- **Lenders:** Can make more informed decisions when evaluating loan applications [8].
- **Government & Policymakers:** Can utilize predictions for developing housing policies and market stabilization programs [8].

The scope of this project is specifically focused on residential properties (houses, apartments, villas) within the Bangalore region, utilizing data scraped from Indian real estate websites like Makaan.com and Commonfloor.com [8, 13]. This localized approach allows the model to capture the unique characteristics and trends of the Bangalore housing market. This paper details the methodology employed in Phase 1 of this project, covering data acquisition, preprocessing, analysis, and the planned model development strategy, along with the expected outcomes.

II. EASE OF USE

The proposed property price estimator is designed with end-user accessibility and practicality in mind. Once trained, the machine learning models can be embedded within a user-friendly interface that allows users to input key property features—such as location, square footage, number of bedrooms and bathrooms, and property type—and instantly receive a predicted price estimate. This model eliminates the need for domain expertise or advanced technical knowledge, making it suitable for a broad range of users including individual buyers, sellers, and real estate consultants.

The ease of integration with existing real estate platforms is also considered. By structuring the prediction system as an API, it can be embedded into web-based portals, allowing real-time evaluation of property listings. Users benefit from an intuitive experience where minimal input yields actionable insights. Additionally, the model's adaptability ensures that as new data becomes available, retraining can occur seamlessly, ensuring continued relevance and accuracy in a dynamically evolving market.

To ensure that the estimator meets practical needs, emphasis is also placed on performance optimization and inference speed. Lightweight models like Linear Regression offer fast predictions suitable for mobile devices, while more complex models like XGBoost provide enhanced accuracy for professional analysis tools. Overall, the system is built to deliver accurate, scalable, and interpretable results with minimal user effort.

Prepare Your Paper Before Styling

Before finalizing the structure and format of the paper, significant emphasis was placed on the completeness and clarity of its content. The initial phase involved gathering and storing raw property data scraped from credible sources such as Makaan.com and Commonfloor.com. This raw data included features such as locality, property type, number of bedrooms and bathrooms, built-up area, and price.

The preprocessing phase focused on cleaning inconsistent entries, handling missing data, and standardizing feature formats—for instance, converting different area units to a common scale (square feet), filtering out outliers, and encoding categorical variables such as location and property type. These steps were essential for ensuring the quality and integrity of the input data before modeling.

The core model development involved experimenting with multiple regression algorithms. Linear Regression served as a baseline due to its simplicity and interpretability, while XGBoost was employed for its robustness, handling of non-linearities, and superior performance on structured data. A variety of evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score were used to assess model performance during cross-validation.

Only after completing all data-centric and model-driven tasks was the paper structured using the IEEE format. The organization of content adheres to academic writing standards, with logical flow across the sections: Introduction, Methodology, Model Evaluation, Results, and Conclusion. All figures, tables, and references were integrated at this stage, ensuring technical accuracy and readability. Care was taken to avoid typographical and grammatical errors through rigorous proofreading.

2. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even if they have already been defined in the abstract. Common abbreviations such as IEEE, SI, and units of measurement (e.g., kg, km, and sqft) do not need to be defined. Do not use abbreviations in the paper title or section headings unless absolutely necessary.

In this paper, the following abbreviations and acronyms are used:

- **ML** – Machine Learning
- **API** – Application Programming Interface
- **MAE** – Mean Absolute Error
- **RMSE** – Root Mean Squared Error
- **R^2** – Coefficient of Determination
- **XGBoost** – Extreme Gradient Boosting
- **BHK** – Bedroom, Hall, Kitchen
- **CSV** – Comma-Separated Values
- **GUI** – Graphical User Interface
- **HTML** – HyperText Markup Language

RESEARCH METHODOLOGY

This section outlines the methodology adopted to develop a machine learning model for estimating residential property prices in Bangalore. It includes the universe and sample of the study, data sources, theoretical framework, and statistical tools employed to analyze and model the dataset.

3.1 Population and Sample

The population of this study comprises residential properties listed in the Bangalore real estate market. These properties include apartments, independent houses, and villas listed on prominent Indian real estate portals such as Makaan.com and Commonfloor.com. The population is defined by all available residential property listings across various localities in Bangalore at the time of data collection.

From this population, a representative sample of approximately 6,000 property listings was selected through web scraping. The sample includes listings from diverse neighbourhoods, covering both premium and affordable housing segments to ensure comprehensive market representation. Properties were filtered to exclude commercial listings, duplicate entries, and those with missing or inconsistent pricing or area data.

The sampling strategy focused on actively listed properties with complete information regarding location, area (in square feet), number of bedrooms and bathrooms, and property type. The data used is cross-sectional, primarily collected during the year 2024, ensuring the relevance of market trends and pricing dynamics for the current housing landscape in Bangalore.

3.2 Data and Sources of Data

The study uses **secondary data** gathered via web scraping from real estate platforms. Data was collected during the first quarter of 2024 to capture current trends in Bangalore's housing market.

Key features extracted from the listings include:

- **Location** (e.g., Whitefield, Electronic City)
- **Area (in sqft)**
- **BHK configuration** (e.g., 2BHK, 3BHK)
- **Number of bathrooms**
- **Property type** (Apartment, Independent House, Villa)

- **Price (target variable)**

Additional location-based data such as proximity to metro stations, amenities, and local pricing trends were optionally integrated through geospatial APIs. The data was preprocessed to address outliers, handle missing values, and normalize units for consistency.

3.3 Theoretical framework

Variables The objective of this study is to build predictive models that estimate the **market value of residential properties** using supervised machine learning regression techniques. The dependent variable is the **price** of the property, while the independent variables include:

- **Location** (categorical)
- **Size/Area in square feet** (numerical)
- **Number of Bedrooms (BHK)** (numerical)
- **Number of Bathrooms** (numerical)
- **Property Type** (categorical)

The relationship between the price and these features is assumed to be **non-linear**, justifying the need for advanced regression models beyond traditional linear methods.

3.4 Statistical tools and econometric models

This section elaborates the proper statistical/econometric/financial models which are being used to forward the study from data towards inferences. The detail of methodology is given as follows.

3.4.1 Descriptive Statistics

Descriptive Initial descriptive analysis was performed to understand the distribution of the data. Metrics such as **mean, median, standard deviation, minimum, and maximum** were computed for each numeric variable. Categorical features were analyzed through frequency distributions. This step was crucial to understand price variability across locations and property types.

3.4.2 Regression Models Used

The study employs and compares multiple regression models to estimate property prices.

a) LinearRegression (Baseline Model)

A basic model that establishes a linear relationship between the features and property prices. This serves as a baseline for performance comparison.

b) XGBoost Regression

An advanced ensemble technique based on gradient boosting that is capable of modeling non-linear interactions. It handles outliers and missing values effectively, making it suitable for real-world property price prediction.

c) Random Forest Regression (Optional)

Used for robustness comparison, this model provides high accuracy and resistance to overfitting.

All models were evaluated using **cross-validation** techniques, and performance was assessed using the following metrics:

- **Mean Absolute Error (MAE)**
- **Root Mean Squared Error (RMSE)**
- **R^2 Score**

3.4.3 Comparison of the Models

The models were compared using statistical performance metrics on a held-out validation dataset. The best-performing model was selected based on its **lowest MAE and RMSE** and **highest R^2 score**. Feature importance plots were generated (especially in XGBoost and Random Forest) to understand which features contribute most to property price predictions.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

Table 4.1: Descriptive Statics

Variable	Minimum	Maximum	Mean	Std. Deviation
Area	400	8000	1470	710
Price	25	800	98	64
Bedrooms	1	5	2.6	0.9
Bathrooms	1	5	2.4	0.8

Table 4.1 The results from descriptive statistics show that the data is **reasonably distributed**, with some variability in price due to area and location. Further analysis revealed that location had a strong influence on property prices.

Visual inspection confirmed the need for **non-linear models** due to heteroscedastic patterns between predictors and the target variable.

III. ACKNOWLEDGMENT

The author wishes to express sincere gratitude to the Project Guide and Head of the Department of MCA, Dr. Shashidhar Kini K, for his invaluable guidance, constant encouragement, and kind support throughout this research work. Appreciation is also extended to the Principal, Dr. Shrinivasa Mayya D, for fostering an environment conducive to completing this project within the institution. The author thanks the management of Srinivas Institute of Technology for their direct and indirect support. Gratitude is also due to all the faculty members and non-teaching staff of the MCA department for their constant help and support. Finally, the author is indebted to parents and friends for their unwavering support and belief throughout this endeavor.

REFERENCES

[1] Zhang, Z. (2021). Decision trees for objective house price prediction. *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, [Preprint]. https://doi.org/10.1109/mlbdbi54094.2021.00059.

[2] Sivasankar, B., Ashok, A. P., Madhu, G., & Fousiya, S. (2020). House Price Prediction. *International Journal of Computer Science and Engineering (IJCSE)*, *8*(7).

[3] Singh, A. P., Rastogi, K., & Rajpoot, S. (2021). House Price Prediction Using Machine Learning. *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 203-206. doi: 10.1109/ICAC3N53548.2021.9725552.

[4] Kuvalekar, A., Manchewar, S., Mahadik, S., & Jawale, S. (2020, April 8). House Price Forecasting Using Machine Learning. *Proceedings of the 3rd International Conference on Advances in Science Technology (ICAST) 2020*.

[5] Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering and Electronic Business (IJIEEB)*, *12*(2), 15-20. DOI: 10.5815/ijieeb.2020.02.03.

[6] Dabreo, S., Rodrigues, S., Rodrigues, V., & Shah, P. (2021). Real Estate Price Prediction. *(Internal Report/Conference Proceeding)* Fr.Conceicao Rodrigues College of Engineering, Mumbai.