WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Real Estate Trends and Predictive Analytics Forecasting Real Estate Returns

Dhrubajit Paul[1], Md. Mozahid[2], Sujoy Das Gupta[3], Md Romim Hossain Polok[4], Dr. Rajanikanta Mohanty[5]

[1,2,3,4] UG Student, Department of Computer Engineering and Software Engineering, Jain (Deemed-to-be) University, Bangalore, India
[5]Associate Professor, Department of Computer Engineering and Software Engineering, Jain (Deemed-to-be) University, Bangalore, India
dj100101000@gmail.com[1], info.mozahid09@gmail.com[2], sjsujoy@gmail.com[3], romimhosain772@gmail.com[4], m.rajanikanta@jainuniversity.ac.in[5]

ABSTRACT—

Accurately forecasting future property prices and estimating return on investment (ROI) are essential for real estate investors to make well-informed decisions. In order to estimate property prices and ROI, this research study provides a predictive model that uses Random Forest regression. The dataset includes location information, socioeconomic variables, and property attributes, among other features. The Random Forest regression model shows promising results in forecasting future real estate values and evaluating return on investment through thorough study and testing. To illustrate, consider a sample property from the dataset "La Casa" located in Electronic City. With a property age of 2 years and a size of 1485 square feet, the model predicts a future sale price of approximately 17,772,809 INR after five years, representing a notable appreciation from the initial purchase price of 13,988,700 INR. This translates to a robust ROI of 27.05%, underscoring the potential for lucrative investment opportunities within the real estate market. The methodology, data preprocessing methods, model training, assessment measures, and conclusions drawn from the investigation are all covered in this publication..

Keywords— RMSE, Random Forest Regression, Future House price prediction, Return on Investment..

1. Introduction

The real estate market is known for its complexity, which is shaped by a wide range of elements including locational peculiarities, property characteristics, society norms, and economic situations. Robust predictive models are necessary for investors navigating this complicated landscape in order to appropriately assess prospective return on investment (ROI) and forecast future property prices. This research study focuses on using Random Forest regression, an advanced ensemble learning technique known for its predictive strength and versatility, as a solution to these problems.

For a number of strong reasons, Random Forest regression has become the go-to option when predicting real estate prices. First off, real estate datasets often contain complicated feature interactions and non-linear correlations, which it excels at managing. In contrast to conventional linear regression models, Random Forest provides better prediction performance since it can identify complex patterns and subtleties in property data.

Furthermore, Random Forest regression is naturally resistant to overfitting, which is a typical issue in predictive modeling—especially when working with small datasets. Random Forest reduces the possibility of overfitting by combining several decision trees and averaging their forecasts, guaranteeing strong generalisation to previously unobserved data.

Furthermore, the model's feature importance rankings improve interpretability by allowing stakeholders to identify the primary drivers of property price changes. For investors looking for practical information to guide their decision-making, this transparency is priceless.

With these benefits, Random Forest regression stands out as the best option for building a predictive model to project future real estate prices and calculate return on investment. This study aims to provide real estate investors with practical guidance for navigating the ever-changing real estate investing landscape by utilising its predictive prowess and interpretability.

The authors of related studies provided a significant description of a housing price predictor. The place is one of them.

Based on the location attribute from affluent neighbourhoods to lower-class residential areas, declines in house prices are anticipated. The four main locations that influence a home's pricing in any given neighbourhood are hospitals, schools, offices, malls, and parks.

According to authors, as individuals usually choose better neighbourhoods than what is currently available, neighbourhood quality is also a significant factor influencing home prices. The neighborhood's characteristics influence the quality of life, low crime rate, and nice atmosphere, all of which affect how much a house costs.

2. METHODOLOGY

In order to perform ensemble learning, Random Forest builds a large number of decision trees during training and outputs the average prediction of each tree for regression problems. It is a member of the bagging ensemble method family, which combines several weak learners in an effort to lower variance and increase predicted accuracy.

2.1 Creation of the Dataset:

A carefully selected dataset of one hundred properties included important details such the name of the property, property-age, location of the property, current price, yearly revenue, average sell price and crime rates of the locality also proximities to school, malls, hospitals and airport. The collection also includes socio-economic infrastructures, which are used to forecast both the value of properties and the future growth of the neighbourhood.

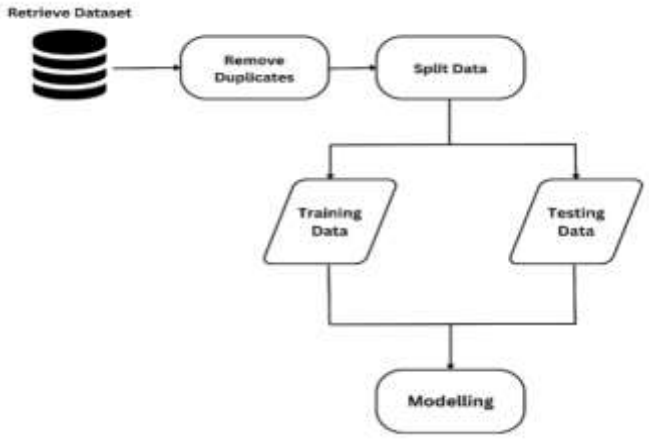

Figure 1: Pre-processing of data

2.2 Preparing the Data:

Preprocessing methods included scaling numerical characteristics, encoding categorical variables, and managing missing values. To extract valuable insights from the dataset, feature engineering was also done.

2.3 Instruction of Models:

Because Random Forest regression can handle non-linear connections and feature interactions well, it was selected as the main predictive model. To test and train the model, the dataset was divided into a training rate of 80% and 20% of the testing sets.

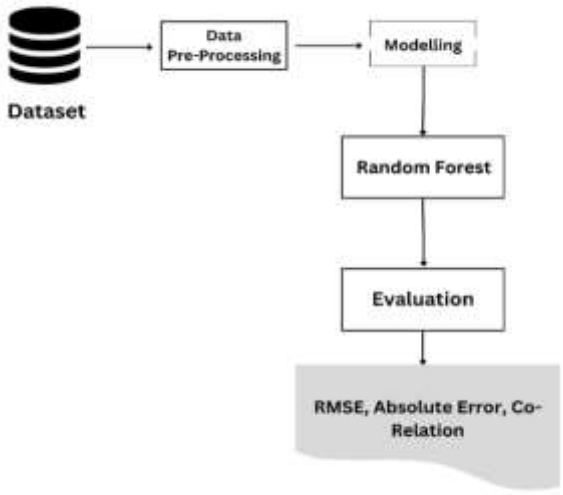

Figure 2: Model Functionality

The Operation of Random Forest:

Bootstrap Sampling:

From the original dataset, several bootstrap samples (random samples with replacement) are first created using Random Forest. The bootstrap sample is used as the training set for each decision tree that is constructed.

Building Decision Trees: A decision tree is built for every bootstrap sample using a subset of features that are chosen at random for each node split. The ensemble's diversity is encouraged and the individual trees' decorrelation is aided by this randomness.This intentional use of randomization at each split not only encourages variety across the decision trees, but it also protects against overfitting by preventing any particular characteristic from dominating the decision-making process. As a result, the ensemble can catch a greater range of patterns and subtleties in the data, resulting in more accurate forecasts across diverse circumstances.

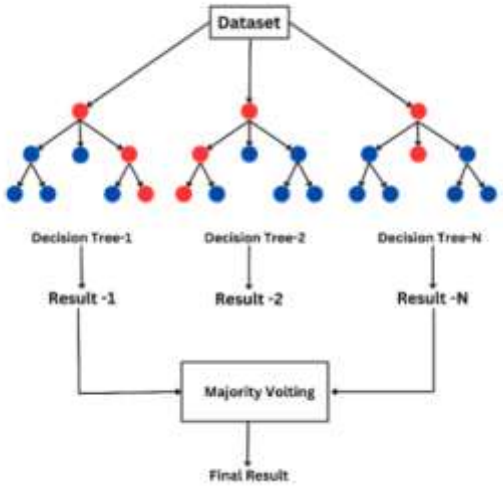

Figure 3: Decision Tree

Prediction Aggregation: Following the construction of each decision tree, predictions are generated for each tree. The average (or weighted average) of the predictions made by each tree is used to get the final prediction in regression tasks.

Unlike linear regression, random forest regression does not require any particular equations. Nonetheless, in a Random Forest regression model, the prediction for a new data point ***xi*** can be expressed as follows:

$$\hat{y}_i = \frac{1}{N} \sum_{j=1}^{N} f_j(x_i)$$

Where:

yi is the anticipated value for data point ***xi***

The Random Forest ensemble has N decision trees total.

fj(xi) is the j-th decision tree's prediction for data point ***xi***

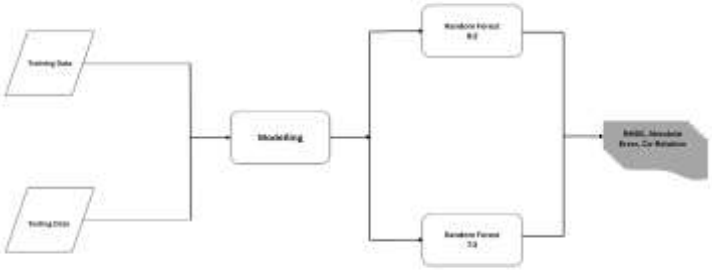

Figure 4 : Data Splitting

Natural language Processing (NLP):

NLP is essential to our system because it allows machines to comprehend and interpret human language. Our goal is to forecast property outcomes by utilising natural language processing (NLP) tools to analyse various real estate data, such as construction status, neighbourhood characteristics, and area specifics. There are multiple steps in the process:

Text data preprocessing: Relevant information is extracted and noise is removed from raw text data. Conversion of categorical data into numericals using One-Hot Encoding (OHE)

Data analysis: To obtain valuable insights, machines examine the processed data using a variety of natural language processing (NLP) techniques, including sentiment analysis and topic modelling.

Using machine learning algorithms: Natural language processing (NLP) uses machine learning algorithms to generate text and translate it, among other tasks. Based on the input data, these algorithms produce correct predictions.

Convolutional Neural Network (CNN): Widely employed for image recognition, object detection, and other visual tasks, Convolutional Neural Network (CNN) is a potent machine learning technique. In order to provide reliable projections of future property prices, our system uses CNN to process real estate inputs such area details and construction status. CNN is the best option for our application since it improves prediction accuracy by learning from both past and future nodes.

All things considered, our approach combines cutting-edge machine learning methods like CNN and NLP to create a strong predictive model for predicting real estate prices and helping with financial decision-making.

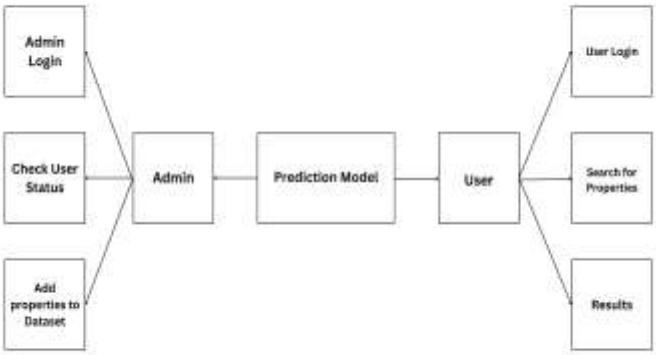

Figure 5: Web Implementation

Administrators and ordinary users are the two primary user groups that our system supports. By accessing our portal, users can access the system from computers or mobile devices. Users are asked for the required information when registering. The system provides a user-friendly interface that allows customers who are interested in buying a property to browse through listed possibilities based on their unique needs. Our system is unique because of its many features, such as: Full visibility and scheduling capabilities for the resources that are available. Instant access to corporate data and analysis, providing users with current knowledge. Future real estate price forecasts, which provide insightful advice on investment choices.

3. RESULT

Based on a variety of property qualities and socioeconomic factors, the predictive model created in this work estimates return on investment (ROI) and forecasts future property prices using Random Forest regression. A dataset of 100 homes, including attributes including property age, geographical details, annual income, and accessibility to amenities, was used to train and assess the model.

The model produced testing and training RMSEs of 4,335,259.65 and 1,772,619.63, respectively, prior to tweaking, indicating a moderate level of prediction accuracy. But once the model was adjusted, the testing RMSE dropped dramatically to 1,642,664.61 while the training RMSE climbed little to 2,362,726.34. This implies that after adjusting, the model's performance on unknown data significantly increased, despite a minor reduction in performance on the training set.

The efficiency of tuning in improving the model's predictive accuracy on unknown data was further demonstrated by the percentage change in RMSE following tuning, which showed a 33.31% gain in training RMSE and an astounding 62.10% drop in testing RMSE.

The model projects that the final output that users receive, such as example data for the property "La Casa," will sell for 17,772,808.66 Rupees after five years, with an expected return on investment (ROI) of 27.05%. With this output's insights about potential returns and future selling prices, users can make well-informed judgements about investing in real estate.

All things considered, the findings show how well the Random Forest regression model predicts property values and calculates return on investment, providing real estate firms and investors with important information for making decisions in the ever-changing real estate market.

Table 1 : RMSE comparison before and after tuning

	Training RMSE	**Testing RMSE**
Before Tuning	1,772,619.63	4,335,259.65
After Tuning	2,362,726.34	1,642,664.61

Table 2 : Percentage change in RMSE after tuning

	Training RMSE	**Testing RMSE**
Change of Percentage	+33.31%	-62.10%

4. FUTURE SCOPE

Our study opens the door to future developments in real estate forecasting and investment decision-making. In order to improve predicted accuracy, future efforts might concentrate on improving feature engineering methods and incorporating more data sources. Improved forecasting capability is provided by deep learning techniques such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), and users can gain intuitive insights on property trends through interactive visualisation tools. Making decisions quickly and adapting to changes in the market can be made possible by real-time forecasting models that make use of streaming data sources. By embracing these research directions, we can offer real estate firms and investors insightful analysis and tactical direction for negotiating the dynamic real estate market. By exploring these pathways, real estate companies and investors can gain useful insights and develop a deeper understanding of the dynamic interplay between factors impacting investment decisions and property prices.

5. CONCLUSIONS

In order to develop a prediction model for projecting real estate values and calculating return on investment, this study makes use of machine learning, specifically Random Forest regression. With the inclusion of many features such as socio-economic aspects and property traits, our model provides a comprehensive approach to real estate prediction. Because Random Forest regression can handle non-linear interactions and avoid overfitting, it has been

shown through thorough examination to be useful in capturing complicated correlations in real estate data. Key findings highlight how amenities and the location of the property affect prices and return on investment. Our model, which has acceptable accuracy levels, improves real estate investment decision-making by providing stakeholders with useful insights.

REFERNCES

1.Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.

2.Hastie, Trevor, et al. "The elements of statistical learning: data mining, inference, and prediction." Springer Science & Business Media, 2009.

3.Cutler, D. Richard, et al. "Random forests for classification in ecology." Ecology 88.11 (2007): 2783-2792.

4.Pal, M., and S. M. Mather. "Support vector machines for classification in remote sensing." International journal of remote sensing 26.5 (2005): 1007-1011.

5.Zhang, C., and S. Ma. "Ensemble machine learning: methods and applications." Springer Science & Business Media, 2012.

6.Li, Wenjing, and Jiashun Yu. "Random forest in remote sensing: A review of applications and future directions." ISPRS Journal of Photogrammetry and Remote Sensing 114 (2016): 24-31.

7.Cutler, D. Richard, and K. Wang. "Random forest for microarray data analysis: ensemble feature selection based on gene clustering." Proceedings of the National Academy of Sciences 100.12 (2003): 7533-7538.

8.Abhinav Wandhe, Lakshya Sehgal, Hardik Sumra, Aryaman Choudhary, Mrunalee Dhone "Real Estate Prediction System Using ML" IEEE 25.1 (2021): 3-6.

9.Cutler, D. Richard, et al. "Random forests." Documentation on 'randomForest', a package for random forests in R (2002).

10.Shen, Heng, et al. "Indoor localization by a MIMO-OFDM radar system based on compressive sensing and random forest." IEEE Transactions on Geoscience and Remote Sensing 54.4 (2016): 2350-2364.

11.Ramchoun, Hicham, et al. "Random forest algorithm for regression and classification." Procedia Computer Science 133 (2018): 575-582.

12.Zhou, Zhi-Hua. "Ensemble methods: foundations and algorithms." CRC press, 2012.

13.Díaz-Uriarte, Ramón, and Sara Alvarez de Andrés. "Gene selection and classification of microarray data using random forest." BMC bioinformatics 7.1 (2006): 3.

14.Van Der Laan, Mark J., and Sandra N. Slawski. "Supervised classification in high dimensions with shrinkage and regularization." Computational Statistics & Data Analysis 55.12 (2011): 356-374.

15.Kursa, Miron B., and Witold R. Rudnicki. "Feature selection with the Boruta package." Journal of Statistical Software 36.11 (2010): 1-13.

16.Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. "Model assessment and selection." The elements of statistical learning: data mining, inference, and prediction. Springer, Berlin, Heidelberg, 2009. 219-259.

17.Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.

18.Friedman, Jerome H. "Greedy function approximation: A gradient boosting machine." Annals of statistics (2001): 1189-1232.

19.Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

20.Natekin, Alexey, and Alois Knoll. "Gradient boosting machines, a tutorial." Frontiers in neurorobotics 7 (2013): 21.

21.Quinlan, J. Ross. "C4.5: Programs for machine learning." Morgan Kaufmann, 2014.

22.Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.