Optimizing Real Estate Prediction - A Comparative Analysis of Ensemble and Regression Models

Runkana Durga Prasad[1[0009-0001-0866-0316]], Vemulamanda Jaswanth Varma[1[0009-0007-8757-6414]], Uppalapati Padma Jyothi[1[0000-0001-8393-8157]], Sarakanam Sai Shankar[1[0009-0001-7874-6932]], Mamatha Deenakonda[2[0000-0002-2723-2026]], and Kandula Narasimharao[1[0000-0001-7120-0593]]

[1]Department of Computer Science and Engineering, Vishnu Institute of Technology
[1]Department of Electric Engineering, Vishnu Institute of Technology
Bhimavaram, Andhra Pradesh
padmajyothi64@gmail.com

Abstract. Valuation is a fundamental aspect of real estate for businesses. Land and property serve as factors of production, and their value is derived from the use to which they are put. This value is influenced by the demand and supply for the product or service produced on the property. Valuation involves determining the specific amount for which a property would transact on a given date. Accurate prediction of real estate prices is crucial for investors, house owners and industry professionals. In this article, analysis of USA real estate prediction using regression and ensemble models was presented, also evaluating the best model out of all the models that have been applied. The objective of this article is to provide accurate predictions for the real estate market, by making use of Multi-Variate Regression, Random Forest Regressor, Decision Tree Regressor, XGB Regressor and CatBoost Regressor. This analysis offers valuable insights for making wise and right choices in the real estate market.

Keywords: Real Estate, United States, Machine Learning, Economy, Regression, Ensembling techniques.

1 Introduction

Real estate is any property, including the rights and interests attached to it, that consists of land, structures, and natural resources. It consists of both residential and commercial properties. A significant asset class that has the potential to increase in value over time is real estate [1]. It is an important part of the investment portfolios of many people. Real estate transactions entail legal procedures, discussions, and financial concerns. The importance of real estate to the economy is demonstrated by its contributions to economic expansion, job creation, wealth generation, housing market stability, commercial activity, and tax receipts [2].

Predictions of real estate prices in the US are influenced by several things. Even though it is difficult to predict real estate prices with complete accuracy, there are a few key variables that can have an impact on price trends, including supply and demand, economic growth, interest rates, location, governmental policies and regulations, demographics, housing market inventory, development, and other outside factors [3].

The real estate market in the USA holds immense significance as a key sector of the economy, attracting investors, house owners and industry professionals. With a vast range of property types and a significant contribution to the country's GDP, the USA real estate market attracts domestic and international investors in the same way [3].

It is important to note that the covid-19 pandemic had an impact on the real estate market of US. People were conscious about buying properties in largely populated areas that had the high chances of spread of covid-19 [4]. Hence it was found that there was a higher demand for the places with low population density [5].

The objective of this work is to predict the price of the real estate market using different machine learning techniques based on regression [6] like Linear Regression, Random Forest, Decision Tree Regressor and ensembling techniques were applied like XGBoost Regressor and CatBoost Regressor. Among all the techniques, this work illustrate which technique is more adaptable for real estate data in United States [7].

The flow of this paper as follows, Section 2 describes about the related work done by other researchers in this area, Section 3 describes the *Exploratory data analysis* to identify the patterns and also understanding the data in a clear manner, Section 4 describes the methodology that was followed during the experiment while Section 5 describes the brief about various algorithms that were used, Section 6 discusses about the metrics used to evaluate the performance of models, Section 7 and 8 describes the results of the experiment conducted and conclusions and future work of this experiment.

2 Related Work

Truong et. al worked for the best results in prediction of real estate, three different machine learning approaches—Random Forest, XGBoost, and LightGBM—as well as two machine learning methods—Hybrid Regression and Stacked Generalisation Regression—are contrasted and examined. Even if all of those techniques produced pleasing outcomes, various models each have advantages and disadvantages [8].

A study looked at how to forecast the asking and sales prices of Nissan Pow land properties using a variety of factors, including location, living space, and the number of rooms. Direct regression, Support Vector Regression (SVR), k-Nearest Neighbours (kNN), and Regression Tree/Random Forest Regression were some of the techniques they used. According to their research, the asking price may be predicted using a kNN and Random Forest algorithm combination with an error rate of 0.0985. The details of the prediction models, examination of the real estate listings, and testing and validation outcomes from the numerous algorithms employed in the study all played a role in the researchers' conclusions [9].

Real estate price volatility has been found by Li et.al to complicate non-linear behaviors and introduce some uncertainty. The author employed a cost-free mathematical model neural network algorithm characteristic. The nonlinear model for real residences value variety expectation is set up using back propagation neural system (BPN) and outspread premise work neural system (RBF), two plans that take into account driving and concurrent financial lists.The two lists of the value variety that are picked as the execution list are the mean absolute value and root mean square error. As

a result, the author has come to the conclusion that the fluctuation in house price trends is not particularly true [10].

The predictive performance of the random forest machine learning technique in comparison to commonly used hedonic models based on multiple regression for the prediction of apartment prices is analyzed by Čeh et al . A dataset that consists of 7407 records of apartment transactions referring to real estate sales from 2008–2013 in the city of Ljubljana, the capital of Slovenia, was used in order to test and compare the predictive performances of both models. All performance measures of both the models such as R2 values, sales ratios, mean average percentage error (MAPE), coefficient of dispersion (COD) revealed significantly for better results for predictions obtained by the random forest method [11].

3 Exploratory Data Analysis

To accomplish the objective of prediction, a comprehensive real estate dataset named ***USA Real Estate Dataset***[1] is brought into use, which was downloaded from Kaggle. The dataset consists of approximately 10,000 records and 10 columns. This dataset encompasses major features that contribute to the pricing of the houses, including the number of bedrooms, number of bathrooms, land area, and location among others. The target variable of the interest is pricing of the house. One challenge with the features is that they may contain some null values. Handling these null values requires specific strategies.

Handling Null Values: The dataset consists of null values in the following columns: bed, bath, acre-lot, city, and house-size. These features contain a significant number of null values. These null values can be handled either by removing the entire row that contains a null value or replacing the null value by mean, mode and median of the respective column. There are different methods such as fillna and replace to accomplish this task.

In this dataset, null values are handled by replacing them with the mean of the corresponding column using the fillna method. Null values must be addressed since they can cause inconsistencies and degrade the data's quality. The selection of the handling strategy is influenced by several variables, including the type of data, the quantity of missing values, and the particular issue that needs to be resolved. The consequences of each approach and any potential effects on the functionality and interpretability of our model must be carefully considered [12][13][14].

Dimensionality Reduction: Dimensionality reduction refers to the process of removing unwanted features and identifying the necessary features in a dataset. One way to identify influential features is by using a correlation matrix. Features that exhibit positive correlation with the target variable are considered important, while features with negative correlation can be removed. In this dataset, the important features that show positive correlation with the target variable are bed, bath, acre-lot, city, and house-size. Therefore, these features are retained as they have a significant influence

[1] USA Real Estate Dataset

on the target variable. However, the features "status", "zip-pincode", and "prev-sold-date" are removed since they do not demonstrate a strong positive correlation with the target variable [15].

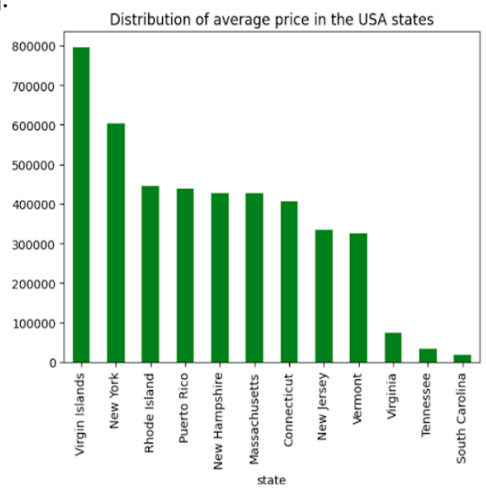

Fig 1: Distribution of average price in the USA states

The figure 1 above illustrates the bar plot plotted for visualizing the average prices of different states in the USA. When examining the plot, it becomes evident that the state "Virgin Islands" has the highest average price compared to the other states. On the other hand, the state "South Carolina" has a relatively lower average price. Additionally, the states "Puerto Rico", "New Hampshire", and "Massachusetts" display similar average prices.

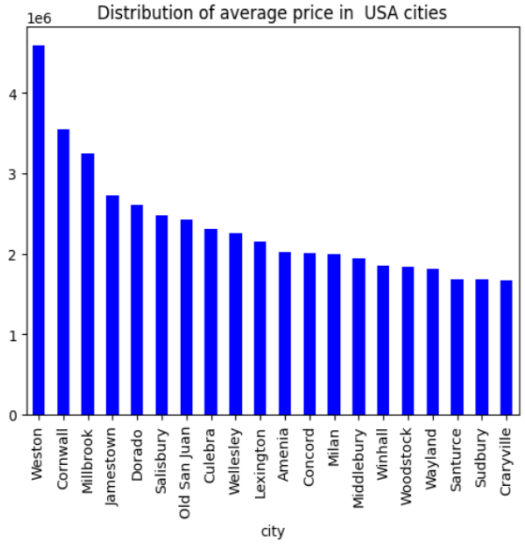

Fig 2: Distribution of average price in the USA states

The figure 2, illustrates the average prices of different cities in the USA. It is true to fact that the city named Weston has the highest average price compared to all other cities. Conversely, the cities of Craryville, Sudbury, and Santurce have relatively lower average prices.

Feature Engineering: It is important to convert string features into numeric format because machine learning models typically cannot understand string data. In this dataset, there are two string columns: State and City. To convert these string features into numeric format, the LabelEncoder method was used. The LabelEncoder method is a commonly used technique for encoding categorical variables into numeric labels. It assigns a unique integer value to each distinct category in the string column. By doing so, it transforms the string values into numerical representations that can be processed by machine learning algorithms [16].

4 Methodology

The datasets were retrieved from Kaggle's machine learning repository, which offer information on American real estate. Then unnecessary features are removed, missing data are handled, and feature engineering is applied during preprocessing. Data is separated into training and testing sets. We investigated many algorithms to develop a real estate price forecast model. They consist of ensembling techniques like XGBoost, CatBoost, and Supervised techniques like Decision Trees, Random Forests Regressor, and Multi- Variate Regression. Finally, with test data, performance of the models is evaluated using several statistical metrics, including MSE, RMSE, MAE, and R2-Score. Fig. 3 depicts the methodology that was chosen.

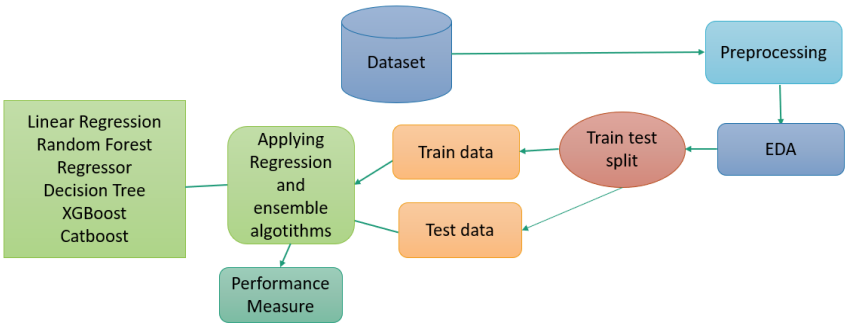

Fig 3: Methodology of the work flow

5 Modeling

Multi-Variate Regression: Regression including more than one independent variable is referred to as multivariate regression or multiple regression. It enables the simultaneous examination of the relationships between a dependent variable and several predictors. In this regression model the predictions are formed from multiple features of the data [17]. In this dataset we provide features are bed, bath, acre_lot, city, state, house_size by using these features we predict the price of the house.

The equation of multivariate regression is

$$y=w_0+x_1w_1+x_2w_2+....$$

Where $x_1,x_2\ldots$ is set of input features and y is output

RandomForest Regressor: It is the best algorithm that belongs to supervised machine learning. This one of the best algorithms which is applied on complex problems and improves the performance of the model. This algorithm will avoid the overfitting. In the random forest classifier, which consists a set of decision trees. The random forest will take predictions from all decision trees and select the one of best predictions among the all-decision trees predictions [18][19].

Decision Tree Regressor: Decision tree regressor is flowchart-like tree structure. This Decision Tree consists of three types of nodes. First one is Root node that has no any incoming edges and 0 or more outgoing edges. Second one is Internal Node Which consists of features or attributes. last one is leaf node consists of prediction value. Edges representing the decision rule. Decision tree is traversed from root node to leaf node [20].

XGB Regressor: XGB regressor best algorithm to handle the large data sets. It is an ensemble learning method that combines weak prediction and makes it as Strong Prediction. It is very useful to handle Missing values in a data set and avoid causing of overfitting. This Algorithm uses the gradient boosting library. This library will improve the performance of the model [21][22].

CatBoost Regressor: The mathematical formula for CatBoost Regressor is more complex than that of linear regression. Instead, it iteratively incorporates more decision trees into the ensemble to optimize a loss function. The loss function varies according to the issue being resolved, but it often quantifies the difference between the projected values and the actual target values. This also uses the gradient boosting library that will improve the performance of a model[23][24].

6 Evaluation Metrics

The effectiveness of machine learning algorithms for predicting the real estate price can be assessed using a variety of evaluation approaches. All these techniques were

evaluated using metrics such as Mean absolute error, Mean Square error, Root Mean square error and R2-Score [25].

If y_i represents the actual real estate price and $\hat{y}_i$ represents the predicted real estate price then,

Mean absolute Error (MAE): It is the mean of the difference between ground real estate price and predicted real estate price as shown in the equation (1).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i}) \quad (1)$$

Mean Squared Error (MSE): It is the mean of the squares of difference between ground real estate price and predicted real estate price as shown in the equation (2).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2 \quad (2)$$

Root Mean Square Error (RMSE): It is the square root of Root Mean Squared Error i.e., the square root of mean of the squares of difference between ground Data estate price and predicted real estate price as shown in the equation (3).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2} \quad (3)$$

R2 or R2-Score: It is the proportion of the variation in the real estate price that is predictable from the features available in the real-estate dataset [26].

If y is the mean of the actual real estate prices as shown in the equation (4)

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad (4)$$

Variability of the dataset can be measured with two sum of squares formulas as shown in the equation – (7)

$$\text{Residual sum of squares } (SS_{res}) = \sum_{i=1}^{N}(y_i - \hat{y_i})^2 \quad (5)$$

$$\text{Total sum of squares } (SS_{tot}) = \sum_{i=1}^{N}(y_i - \bar{y})^2 \quad (6)$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (7)$$

7 Results

In the conducted analysis, various regression algorithms were assessed for their performance in predicting the target variable. The results revealed intriguing insights into their capabilities. Multi-variate regression, while showing potential, displayed relatively higher errors with a Mean Absolute Error (MAE) of 1.926 and a R2-Score of 0.14. On the other hand, both the random forest regression and CatBoost regressor stood out with impressive performances. The random forest regression demonstrated exceptional accuracy, yielding a negative MAE of -0.651 and an impressive R2-Score of 0.94. Similarly, the CatBoost regressor showcased strong predictive abilities, as evidenced by its negative MAE of -0.643 and a remarkable R2-Score of 0.94. These findings suggest that the random forest regression and CatBoost regressor are well-suited for the given task, displaying superior performance in accurately predicting the

target variable as seen in Table 1. Even the real estate prediction can be done with the advanced neural network models[27].

Table 1. Performance of the models.

Algorithm	MAE	MSE	RMSE	R2-Score
Multi-variate Regression	1.926	1.999	1.997	0.14
Random Forest Regression	-0.651	-0.523	-0.556	0.94
DecisionTree Regressor	-0.669	-0.477	-0.445	0.93
XGBoost Regressor	0.037	-0.478	-0.448	0.93
CatBoostRegressor	-0.643	-0.520	-0.548	0.94

8 Conclusion and Future Work

This research paper describes how machine learning algorithms effectively predict the value of real estate based on various factors like location, number of bedrooms, square feet, etc. Random Forest and CatBoost Regressor have high accuracy and a low error rate compared to the other algorithms. The dataset that is available has a limited number of features. Future extension work is identified in working with high-dimensional data, and instead of applying the core machine learning techniques, deep learning techniques like long-term shortest memory and GRUs can be applied to work with huge data.

References

1. Brueggeman, W. B., & Fisher, J. D.: Real estate finance and investments. New York: McGraw-Hill Irwin (2011).
2. Ghysels, E., et al.: Forecasting real estate prices. In: Handbook of economic forecasting, 2, 509-580 (2013).
3. Saiz, A., & Salazar Miranda, A.: Real trends: The future of real estate in the United States. MIT Center for Real Estate Research Paper 5 (2017).
4. Kumari, K. R., Gayathri, T., & Madhavi, T.: Machine Learning Technique with Spider Monkey Optimization for COVID-19 Sentiment Analysis. In: 2022 International Conference on Computing, Communication and Power Technology (IC3P). IEEE (2022).
5. Del Giudice, V., De Paola, P., & Del Giudice, F. P.: COVID-19 infects real estate markets: Short and mid-run effects on housing prices in Campania region (Italy). Social Sciences, 9(7), 114 (2020).
6. Kovvuri, A. R., Uppalapati, P. J., Bonthu, S., & Kandula, N. R.: Water Level Forecasting in Reservoirs Using Time Series Analysis – Auto ARIMA Model. In: Gupta, N., Pareek, P., Reis, M. (eds.) Cognitive Computing and Cyber Physical Systems. IC4S 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 472. Springer, Cham. https://doi.org/10.1007/978-3-031-28975-0_16 (2023).
7. Wang, D., & Li, V. J.: Mass appraisal models of real estate in the 21st century: A systematic literature review. Sustainability, 11(24), 7006 (2019).
8. Truong, Q., et al.: Housing price prediction via improved machine learning techniques. Procedia Computer Science, 174, 433-442 (2020).

9. Pow, N., Janulewicz, E., & Liu, D.: Applied Machine Learning Project 4 Prediction of real estate property prices in Montreal (2016).
10. Li, L., & Chu, K. H.: Prediction of Real Estate Price Variation Based on Economic Parameters. Department of Financial Management, Business School, Nankai University (2017).
11. Čeh, M., et al.: Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS International Journal of Geo-Information, 7(5), 168 (2018).
12. Jyothi, U. P., et al.: Comparative Analysis of Classification Methods to Predict Diabetes Mellitus on Noisy Data. In: Machine Learning, Image Processing, Network Security and Data Sciences: Select Proceedings of 3rd International Conference on MIND 2021. Singapore: Springer Nature Singapore (2023).
13. ang, C., et al.: Subtle bugs everywhere: Generating documentation for data wrangling code. In: 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE (2021).
14. Emmanuel, T., et al.: Handling Null: A survey on missing data in machine learning. Journal of Big Data, 8(1), 1-37 (2021).
15. Van Der Maaten, L., Postma, E., & Van den Herik, J.: Dimensionality reduction: A comparative. Journal of Machine Learning Research, 10, 66-71 (2009).
16. Milo, T., & Somech, A.: EDA: Automating exploratory data analysis via machine learning - An overview. Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (2020).
17. Heidari, M., Zad, S., & Rafatirad, S.: Ensemble of supervised and unsupervised learning models to predict a profitable business decision. In: 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). IEEE (2021).
18. Levantesi, S., & Piscopo, G.: The importance of economic variables on London real estate market: A random forest approach. Risks, 8(4), 112 (2020).
19. Liaw, A., & Wiener, M.: Classification and regression by random Forest. R News, 2(3), 18-22 (2002).
20. Fan, G. Z., Ong, S. E., & Koh, H. C.: Determinants of house price: A decision tree approach. Urban Studies, 43(12), 2301-2315 (2006).
21. Satish, G. N., et al.: House price prediction using machine learning. Journal of Innovative Technology and Exploring Engineering, 8(9), 717-722 (2019).
22. Avanijaa, J.: Prediction of house price using xgboost regression algorithm. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(2), 2151-2155 (2021).
23. Fedorov, N., & Petrichenko, Y.: Gradient boosting–based machine learning methods in real estate market forecasting. In: 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020). Atlantis Press (2020).
24. Kumar, G. K., et al.: Prediction of house price using machine learning algorithms. In: 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI). IEEE (2021).
25. Botchkarev, A.: A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdisciplinary Journal of Information, Knowledge, and Management, 14, 045-076 (2019).
26. Chicco, D., Warrens, M. J., & Jurman, G.: The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623 (2021).
27. Khalafallah, A.: Neural network-based model for predicting housing market performance. Tsinghua Science and Technology, 13(S1), 325-328 (2008).