SSURJET

Sir Syed University Research Journal of Engineering & Technology
2022, Vol. 12, No. 2,
https://doi.org/10.33317/ssurj.504

House Price Prediction of Real-Time Data (DHA Defence) Karachi using Machine Learning

Lata Bai Gokalani[1], Bhagwan Das[1], Dilip Kumar Ramnani[2], and Mazhar Ali Shah[1]

[1]*Department of Electronic Engineering, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan*
[2]*Department of Electronic Engineering, Dawood University of Engineering and Technology Karachi, Pakistan*

Correspondence Author: Bhagwan Das (engr.bhagwandas@hotmail.com)

Received June 05, 2022; Revised August 23, 2022; Accepted October 05, 2022

Abstract

Pakistan's real estate market has a large impact on GDP growth. Investment in the real estate sector in Pakistan is encumbered with lucrative opportunities. The market demand for housing is ever-increasing year by year. House sales prices, keep on changing and increasing frequently, so there is a need for a system to forecast house sales prices in the future. Several factors that influence house sales price includes; location, physical attributes, number of bedrooms as well as several other economic factors. This research paper mainly focuses on real-time DHA Defence Karachi data, applying different regression algorithms like Decision Tree, Random Forest, and Linear regression to find the sales price prediction of the house and compare the performance of these models. The random Forest algorithm gives 98 % of accuracy.

Index Terms: *Housing Sale Price Prediction Models, Machine Learning, Real Estate Market, Real-Time Data, Regression Algorithms.*

I. INTRODUCTION

Real estate is one of the most famous and usually practiced investment ideas in Pakistan. The best thing about real estate investment in Pakistan is that it is a safe investment option that yields higher returns, making it a good option for small investments. Residential and commercial real estate are most commonly used. Residential real estate provides houses; flats and plots for families while commercial real estate is usually used for business activities such as offices, shopping malls, and departmental stores. In Pakistan, Karachi is the major city; investors' main focus is on investing in houses/property, resulting in the prices of properties touching millions of rupees. But, the major problem of ordinary Pakistanis is that they do not compute property prices using factors such as location, square footage, number of bedrooms, and number of bathrooms. Take external factors, to refine prediction accuracy [1].

Machine Learning (ML) is used for image recognition, text generation, and many other uses, case applications in a real-world scenario. Linear Regression (LR) and Artificial Neural Networks (ANN) algorithms were successfully applied to housing price prediction by [2] and [3] respectively. Deep Learning has also made significant progress in low-level tasks [4] and various Spatio-temporal remote sensing tasks [5]. ML is an emerging research field in real estate analysis that helps us to predict the best type of property for ordinary persons thereby helping them to predict property prices.

There are many ML models available but in this research paper, we have used three machine learning techniques; Decision Tree, Random Forest, and Linear Regression to perform predictive analysis of house price predictions. Data is the heart of ML. We used real-time DHA, Defence Karachi data set to sell a house. The dataset comprises of eleven (11) input features and one target output feature. The results confirmed that this method provides good accuracy and minimum error than single algorithms used.

II. PROBLEM STATEMENT

There are certain problems regarding house price prediction such as; it is difficult to choose the algorithm for predicting accurately the price of a house among the multiple regression algorithms available. We are only concerned with the specific area but the problem is; an accurate estimate of house prices can prove challenging. We also face some problems with available tools mostly because they may provide an inaccurate estimation and also the problem of misguidance by property agents about valuation. House buyers and sellers want help predicting the house price for DHA Karachi, Pakistan. We take the dataset to work on and decided to use regression algorithms/models.

III. LITERATURE REVIEW

Machine Learning or ML is a fast-growing field of self-evaluating algorithms that predict future activity based on past data. House price prediction is based on the same principle. This section presents various ideas and prevailing studies in this specific field.

The previous study completed and reported has its own set of benefits and drawbacks in terms of developing a system based on a particular problem of estimating value when

purchasing, selling, covering, lending, or taxing residence property.

These attributes are summed up as shown in table I.

Table I: Methods to Predict Future House Prices with the help of Machine Learning Historically

Author-Year	Problem	Method	Results	Disadvantages
Satish et al. in (2019) [6]	Method to Predict Future House Prices with the help of Machine Learning.	XGBoost, Neural System, and Lasso Regression	Lasso Regression Algorithm, with high accuracy, reliably outperforms alternate models in the execution of housing cost prediction.	LASSO selects at most n variables before it saturates. LASSO cannot do group selection.
Alfiyatin et al. in (2017) [7]	Discussed the Prediction Model is based on Regression Analysis and Particle Swarm Optimization (PSO).	PSO is a Stochastic Optimization Technique used for the selection of affect variables and regression is used to determine the Optimal Coefficient in-house prediction.	The results obtained Minimum Prediction Error.	In a PSO system, it can be difficult to define initial design parameters.
Segnon et al. in (2020) [8]	Logistic Smooth Transition Autoregressive Fractionally Integrated Process to predict housing price volatility in the US.	It analyzed complicated Statistical Models based on assumptions of the Variance Process.	Measurement results provide satisfied Forecasting Accuracy.	Its works on high-frequency data.
Kuvalekar et al. in (2020) [9]	Prediction of the market values of a Real Estate property.	A Decision Tree Regressor helps find a starting price for a property based on Geographical Variables. By breaking down past market patterns and value ranges and coming advancements future costs will be predicted.	It provides 89 % accuracy.	Instability, higher time to train the model, and complex calculation.
Singh et al. in (2020) [10]	The concept of Big Data to predict Housing Sale Data in Iowa.	Linear Regression, Random Forest, Gradient Boosting.	The Gradient Boosting Model outperforms the other Forecasting Models in terms of Forecasting Accuracy.	The gradient Boosting Algorithm is sensitive to outliers.
ZHANG in (2021) [11]	Housing Price Prediction based on Multiple Linear Regressions.	Multiple Linear Regression Model.	It effectively predicts and analyzes the housing price to some extent.	In multiple Linear Regression Models, the Prediction Accuracy is still limited to some extent.
LIU in (2022) [12]	Prediction and Analysis of the Real Estate Market based on the Multiple Linear Regression Model.	Multiple Linear Regression Model uses the Least Square Method.	The Maximum Error of the Real Estate Price Prediction Model is no more than 8 %.	The Maximum Prediction Error of the Real Estate Price Prediction Model is 7.6 %; the Minimum Prediction Error is −0.2 %.

IV. RESEARCH METHODOLOGY

As shown in figure I, the proposed design method involves the collection of data, pre-processing of data, inventive feature engineering, the different regression models such as Linear, Decision Tree, and Random Forest regression, and the result.

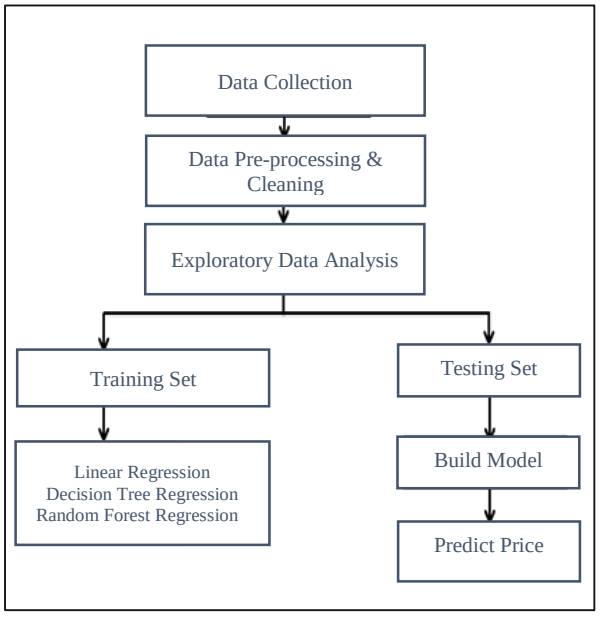


Figure I: Design Methodology

A. Phase I: Collection of Data

The collection of Data is one of the most important steps in conducting research. We have gathered data on DHA Defence Karachi from different agents of real estate. Dataset consists of thousand (1000) Rows and has twelve (12) Columns as shown in figure II.

B. Phase II: Data Preprocessing

Data processing is used to improve the quality and efficiency of data. Data cleaning is a subset of data pre-processing. Before applying different ML algorithms to house price prediction, data preprocessing is applied. The data preprocessing smooths out the noisy data, remove outliers, and fill out the missing values. We used it 80 % of the time to clean the data. After preprocessing, the dataset consists of thousand (1000) Rows and four (4) Columns.

C. Phase III: Training and Testing Model

In this step, data is split into two parts: 'Training' and 'Testing'. Here 30 % of the data is used for testing purposes and 70 % is used for training purposes. The training set included three (3) target variables and in testing set price was to be predicted. We trained models for different algorithms i.e., Decision Tree, Linear, and Random Forest Regression, to test the dataset for house price prediction and get the result. The Decision Tree Algorithm gives the highest accuracy.

```
df1 = pd.read_csv("lsk.csv")
df1.head()
```

	Price	Sq. Yd	Bedrooms	Bathrooms	Locality	Latitude	Longitude	Area_marla	City	Province	Location	Purpose
0	35600000	115.6	4	4.0	DHA Phase 7 Extension	24.8196	67.0777	4.62 marla	Karachi	Sindh	DHA Defence	Sale
1	30400000	112.7	3	3.0	DHA Phase 7 Extension	24.8196	67.0777	4.51 marla	Karachi	Sindh	DHA Defence	Sale
2	29300000	108.0	4	5.0	DHA Phase 7 Extension	24.8196	67.0777	4.32 marla	Karachi	Sindh	DHA Defence	Sale
3	36300000	118.5	4	4.0	DHA Phase 7 Extension	24.8196	67.0777	4.74 marla	Karachi	Sindh	DHA Defence	Sale
4	42500000	100.0	5	5.0	DHA Phase 8	24.7732	67.0762	4 marla	Karachi	Sindh	DHA Defence	Sale

Figure II: Dataset Description

D. Phase IV: Programming Language

In our proposed work Python programming language is used. Python is a high-level interpreted language that has easy-to-use/understand syntax and versatile functionality. Python has many standard libraries that make coding much easier. Python comes with a variety of common libraries that make coding considerably easier. Python has a huge number of built-in libraries that runs ML algorithm. To build Machine Learning Model, Python language is used. Python is undoubtedly the best choice for ML.

V. IMPLEMENTATION

A. Read Dataset

Using the 'read_csv ()' function is the first step for reading the dataset from the 'Pandas Python' package.

B. Dealing with Missing Values

While dealing with missing values, ML Models would not accept data with missing values. Missing values have been removed with NAN values.

C. Feature Selection

Feature selection is an important step that highly impacts the performance of the model. Identifying those features which are less or more important for house price prediction. Those features which are less important were removed. We select features with the Correlation Matrix Method. Correlation Matrixes are important because they show the kind of relationship that exists between parameters. Figure III shows the Correlation Matrix. The highest correlation of 86 % exists between bedrooms and bathrooms. It is important to find out how the dataset variables relate to each other and how the predictor variables relate to the target variable. For example, we check how much 'Price' and 'Area' are correlated: is it decreasing and increasing simultaneously (positive correlation)? One of them decreases then the other increase (negative correlation). Does neither have a correlation?

As a value between -1 and +1, correlation is shown as -1 denoting the highest negative correlation, +1 denoting the highest positive correlation, and 0 reveals that there is no correlation at all. Using a 'Heatmap' graph, we will demonstrate the correlation between variables in our dataset. Our dataset contains correlated variables. It is reasonable to notice that bathrooms and bedrooms have high positive correlations. As far as negative correlations are concerned, we can observe that latitude and longitude are negatively correlated. Most importantly, we want to examine the target variable (Price) that is correlated with the predictor variables. There is a positive correlation between the area and target variable in the first row of the Heatmap. The bedroom and bathroom variables are also positively correlated to the target variable [13].

In the above discussion, we see that area, the number of bedrooms and the bathrooms are highly correlated to the target variable.

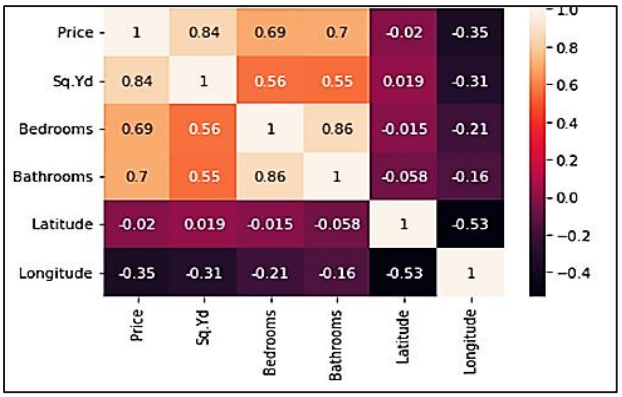


Figure III: Feature Correlation Matrix for the House Price Dataset

VI. TRAINED MODEL

The designed model is achieved by subsequent applying the dataset to the ML algorithms. Our findings provide a house-searching framework, depending on the regression algorithm which is dependent on the features of the house given as an output to the model. The algorithm will require a house i.e., the most alike in value to the provided values. The result will be the factors of price, house with high results suitable for the input values. In the whole dataset the algorithm search for similarities. The result is determined dependent on the nearest values, as illustrated in figure IV.

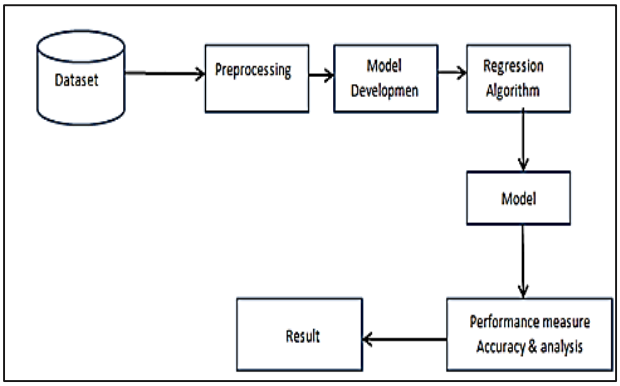


Figure IV: Diagram of Model Training

VII. REGRESSION ANALYSIS

Using regression analysis to examine the relationship between dependent and independent variables is a useful statistical technique. Regression analysis in ML is widely used for prediction and forecasting. Linear, multiple, polynomial, and some other regression analysis techniques are most often utilized/used for prediction.

A. Linear Regression

A Linear regression algorithm can be used for supervised ML. This algorithm provides less complexity and overfitting problem. An algorithm is used to find the relationship between two independent variables. In order to predict the unseen house price from the test set, a Linear regression model will find the best line that fits the training set.

B. Decision Tree Regression

Decision Tree is a practical approach for supervised ML. Both classification and regression problems can be dealt with using this technique. The tree structure is used to build the model. A Decision Tree is developed by breaking down the dataset into smaller and smaller portions. As a result, a tree contains decision nodes and leaf nodes.

To develop a model that can predict a target variable's value, basic decision rules must be learned using data attributes. This algorithm is easy to understand, interpret and visualize.

C. Random Forest Regression

Random Forest is capable of performing both classification and regression tasks. It uses Ensemble learning and combines several different classifiers to get a better result for more complex problems. Random Forest algorithms are created using many decision trees. Decision trees are used to predict the outcome of the Random Forest algorithm.

The Random Forest algorithm uses either bagging or bootstrap aggregation to train its 'forest'. The bagging approach is based on an ensemble meta-algorithm that maximizes accuracy by combining multiple approaches. The forecasting can be done by aggregating or averaging the output of multiple trees. A higher number of trees mean more precise results. The Random Forest algorithm handle missing values provide higher accuracy and overcomes the overfitting problem.

VIII. RESULT AND DISCUSSION

In order to address the issue of the ever-increasing value of the real estate market in Pakistan, the above ML system is suggested/proposed, designed, and implemented.

This ML system predicts 'House Prices', and suggests investors/house buyers identify the maximum/most accurate returns on a small amount of investment. In house price forecasting, three features such as; the number of bedrooms, number of bathrooms, and area square yard will highly impact the price of a house.

In below figure i.e., figure V, illustrate the relationship between bedrooms and total house unit sale. It also shows the preference of house buyers with respect to the number of bedrooms. The maximum number of buyers prefer to buy four-bedroom houses.

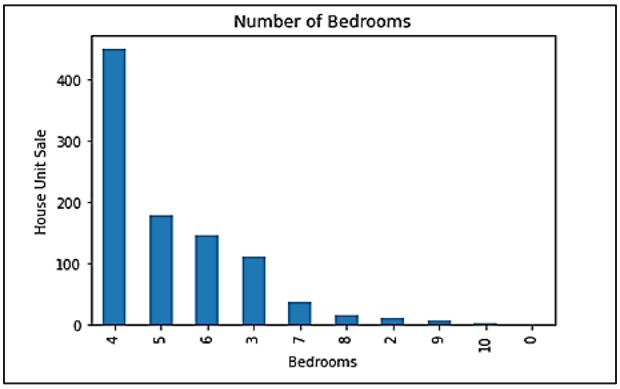


Figure V: Relation between the Number of Bedrooms and Total House Unit Sale

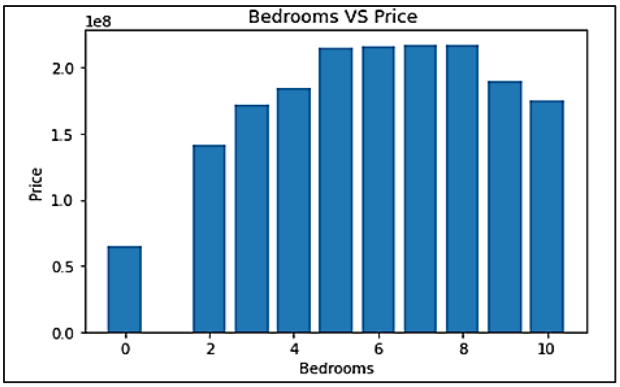


Figure VI: Relation between the Number of Bedrooms and the Price

Figure VI, shows the relationship between 'Price' (in million) and 'Number of Bedrooms' preferred. It was observed that the six-bedroom and eight-bedroom houses were available at the highest prices as compared to the ten-bedroom houses.

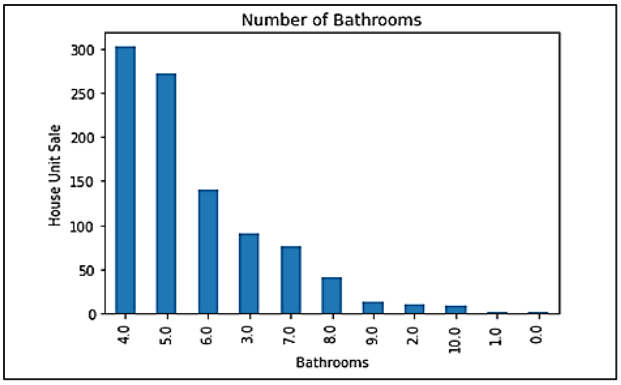


Figure VII: Relation between the Number of Bathrooms and Total House Unit Sale

Figure VIII: Relation between the Number of Bathrooms and the Price

Similarly, the analysis is performed in figure VII, between 'Number of Bathrooms' and 'Total House Unit Sale'. It shows that buyers' preference is for four numbers of bathrooms which they like to buy more.

As shown in above figure VIII, the relationship between 'Price' (in million) and 'Number of Bathrooms' is preferred. It was observed that the six-bathroom and eight-bathroom houses were available at the highest prices as compared to ten-bathroom houses.

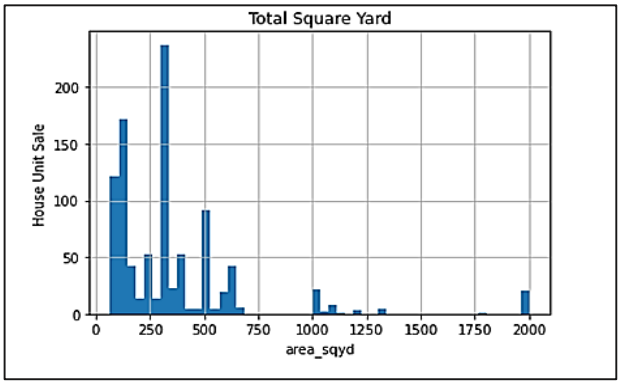


Figure IX: Relation between Area and Total House Unit Sale

Figure IX, shows the relationship between the 'Area' in the square yard and 'Total House Unit Sale'. It is shown that the maximum number of houses is available in the range of 100 and 300 square yards.

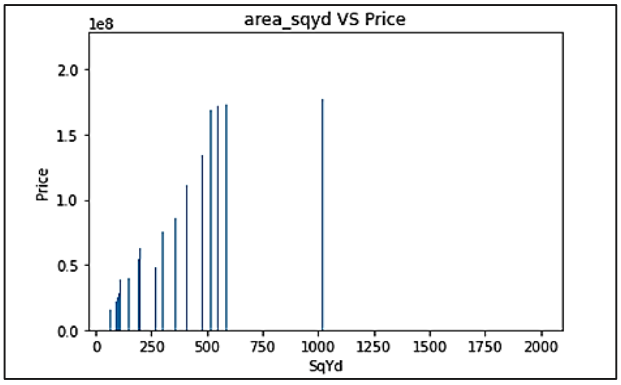
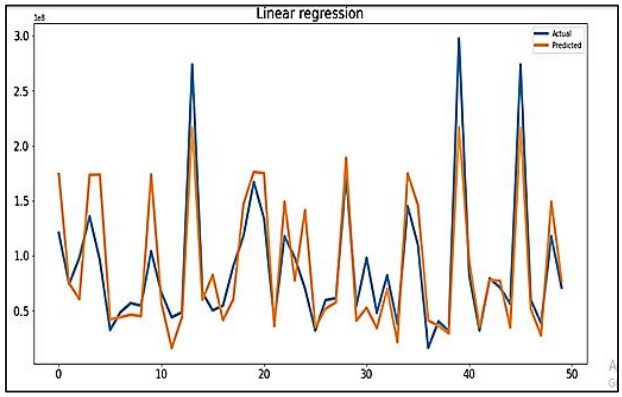


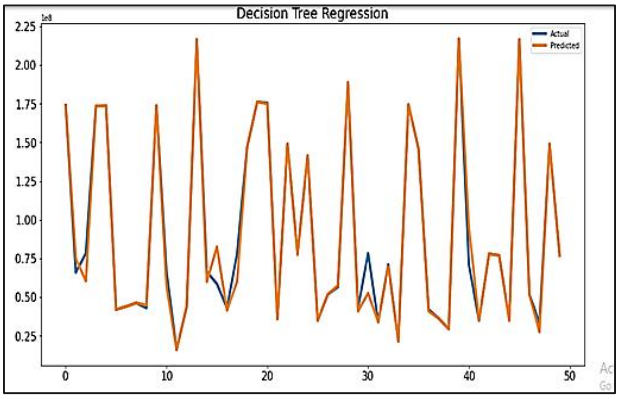
Figure X: Relation between Area and Price

Figure X, shows the relationship between square yards, i.e., 'Area' and 'Price'. Area squared yard is one of the most important features which affect the prices of the house.

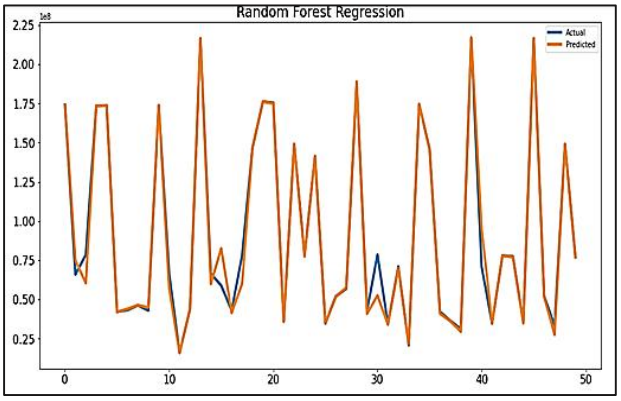
In figure XI, we compare and examine the result of different ML models on similar data input values. To train and evaluate supervised ML algorithms, we need a training dataset for the model and a test dataset for evaluating the model. Our dataset will be split into two parts, one for testing and the other for training. In order to make a prediction on the test data to fit the model on train data, 'The Python Scikit Learn library' called 'train_test_split ()' is used to train and test the model. For predicting the actual and predicted prices of the house, we used Random Forest, Linear, and Decision Tree algorithms. We used 70 % data for the training set and 30 % for testing. We have four number of data subsets: 'y_train', 'X_train', 'y_test' and 'X_test'. 'y_train' and' X_train' is used to train our model, for evaluate the test model 'y_test' and 'X_test' is used. 'y_train' and 'y_test' represent the target; 'X_test' and 'X_train' are features (predictors). The 'X_train' and 'y_train' datasets will be referred to as the training datasets, and the test dataset will be referred to as 'X_test' and 'y_test'.



(a)



(b)



(c)

Figure XI: Comparison between Predict and Test Values of Houses based on (a) Linear Regression; (b) Decision Tree Regression and (c) Random Forest Regression

IX. EVALUATION METRICS

In order to determine the efficiency and performance of a Regression model, several factors must be considered. Based on the data, we will evaluate each Regression algorithm using the following measures/metrics.

A. Mean Absolute Error (MAE)

It is one of the most commonly used metrics to summarize and assess the quality of a Machine Learning or (ML) model. Absolute errors are measured using the Mean Absolute Error (MAE). An absolute error is the number of errors a model makes when trying to predict future values.

The difference between actual and predicted values is called an absolute error. One of the parameters/metrics that we used in our experiment to evaluate the efficiency of Regression models was Mean Absolute Error.
MAE can be expressed mathematically as follows:

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

Table II: Analysis of Models

S. No.	MAE		
	Method	**Model Accuracy**	**Mean Absolute Error**
1	Linear Regression	76.93%	18406282.997433938
2	Decision Tree	98.01%	3815628.66737928
3	Random Forest	98.08%	3799622.3401656877

Table II shows the Mean Absolute Error and the Model Accuracy. The results are based on a variety of ML algorithms.

X. CONCLUSION

In this paper, we found effective methods for predicting accurately property prices. It also offers information on the DHA, Defence Karachi house price market. To predict the price of a house, Machine Learning or ML techniques are most effective. In this research paper, three modeling techniques have been used, such as the Random Forest, Decision Tree, and Linear Regression to detect the feature selection to estimate the sale price in real estate by taking into account the DHA Defence dataset that contains 1000 records and 11 features including price, longitude, latitude, city, property type (House), number of bedrooms, number of bathrooms, and area.
Initially, the data is prepared and transformed into a cleaned/ transparent test process and found out the features that best match the predicted value. The purpose of this system is to create a Predictive Model based on the factors that affect the price for evaluating the price.
Results will be utilized by the developer to determine the selling price of a house and help the buyer or seller to purchase a house according to their budget. The House prediction system is capable of filling the information gap and improving Real Estate efficiency. Mean Absolute Error and Model accuracy are being calculated using Random Forest, Decision Tree, and Linear Regression algorithm. Comparing the results of three models; Random Forest provides/outperformed the best results, both in terms of a high rate of accuracy/predictions and Mean Absolute Error (MAE).

Acknowledgment

The authors would like to thank Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan, for all the support provided to accomplish this research work.

Authors Contributions

Lata Bai Gokalani's contributions to the study were conceptualization, formal analysis, validation, and original draft preparation. Bhagwan Das's contributions to the study were, methodology, supervision, administration, and correspondence. Dilip Kumar Ramnani's contributions to the study were investigation, visualization, review, and editing. Mazhar Ali Shah's contributions to the study were, software selection, final review, and final editing.

Conflict of Interest

The authors declare no conflict of interest and confirm that this work is original and not plagiarized from any other source, i.e., electronic or print media. The information obtained from all of the sources is properly recognized and cited below.

Data Availability Statement

The testing data is available in this paper.

Funding

This research received no external funding.

References

[1] Satish, G. N., Raghavendran, C. V., Rao, M. S., & Srinivasulu, C. (2019). House price prediction using machine learning. *Journal of Innovative Technology and Exploring Engineering*, *8*(9), 717-722.

[2] Alfiyatin, A. N., Febrita, R. E., Taufiq, H., & Mahmudy, W. F. (2017). Modeling house price prediction using regression analysis and particle swarm optimization case study: Malang, East Java, Indonesia. *International Journal of Advanced Computer Science and Applications*, *8*(10).

[3] Segnon, M., Gupta, R., Lesame, K., & Wohar, M. E. (2021). High-frequency volatility forecasting of US housing markets. *The Journal of Real Estate Finance and Economics*, *62*(2), 283-317.

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

[5] Singh, A., Sharma, A., & Dubey, G. (2020). Big data analytics predicting real estate prices. *International Journal of System Assurance Engineering and Management*, *11*(2), 208-219.

[6] Zhang, Q. (2021). Housing Price Prediction Based on Multiple Linear Regression. *Scientific Programming, vol. 2021, pp. 1-9.*

[7] Liu, G. (2022). Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model. *Scientific Programming*, *vol. 2022, pp. 1-8.*

[8] Dharejo, F. A., Deeba, F., Zhou, Y., Das, B., Jatoi, M. A., Zawish, M., ... & Wang, X. (2021). TWIST-GAN: Towards wavelet transform and transferred GAN for spatio-temporal single image super-resolution. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *12*(6), 1-20.

[9] Du, Y., Wang, H., Cui, W., Zhu, H., Guo, Y., Dharejo, F. A., & Zhou, Y. (2021). Foodborne disease risk prediction using multigraph structural long short-term memory networks: algorithm design and validation study. *JMIR Medical Informatics*, *9*(8), e29433.

[10] Deeba, F., Dharejo, F. A., Zawish, M., Memon, F. H., Dev, K., Naqvi, R. A., ... & Du, Y. (2021). A novel image dehazing framework for robust vision-based intelligent systems. *International Journal of Intelligent Systems*.

[11] Tang, Z., Qiao, Z., Hong, X., Wang, Y., Dharejo, F. A., Zhou, Y., & Du, Y. (2021, August). Data augmentation for graph convolutional network on semi-supervised classification. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data* (pp. 33-48). Springer, Cham.

[12] Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert systems with Applications*, *36*(2), 2843-2852.

[13] Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy* (pp. 89-118). University of Chicago Press.