FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 2 | Pages 138-143 | e-ISSN: 2347-470X

Free Hold Price Predictor Using Machine Learning

Shivesh Singh[1], Shaurya Singh[2], Sudeept Singh Yadav[3] and Avneesh Kumar[4]

[1]School of Computing Science and Engineering, Galgotias University, Greater Noida, India, sshivesh@gmail.com
[2]School of Computing Science and Engineering, Galgotias University, Greater Noida, India, shaurya00700@gmail.com
[3]School of Computing Science and Engineering, Galgotias University, Greater Noida, India, sudeept999@gmail.com
[4]School of Computing Science and Engineering, Galgotias University, Greater Noida, India, avneesh.avn119@gmail.com

***Corresponding Author:** Sudeept Singh Yadav; E-mail: sudeept999@gmail.com

ABSTRACT- People who want to buy a new home tend to save more on their budgets and market strategies. The current system includes real estate calculations without the necessary forecasts for future market trends and inflation. The housing market is one of the most competitive in terms of pricing and the same has varied greatly in terms of many factors. Asset pricing is an important factor in decision- making for both buyers and investors in supporting budget allocation, acquisition strategies and deciding on the best plans as a result, it is one of the most important areas in which machine learning ideas can be used to maximize and accurately anticipate prices.

As a result, in this paper, we present the different significant factors that we employ to accurately anticipate property values. To reduce residual errors, we can utilize regression models with a range of characteristics. Some engineering aspects are required when employing features in the regression model for improved prediction.

To improve model fit, a set of multi-regression elements or a polynomial regression (with a set of varying strengths in the elements) is frequently utilized. In these models, it is expected to be significantly affected by the slope of the spine used to reduce it. Therefore, it directs the best use of regression models over other strategies to maximize the effect.

This paper's goal is to predict free hold prices for free hold consumers based on their budgets and goals. Prospective prices can be forecast by evaluating past market trends and price levels, as well as future developments.

General Terms: The experiments revealed that random for-set and gradient boosted tresses work better with higher accuracy percentages and lower error values when compared to other machine learning techniques. When the experiment's results are compared to the expected outcomes, these algorithms perform well.

Keywords: House price, Regression Analysis, Linear Regression, Supervised Learning, Healthcare, Machine Learning.

ARTICLE INFORMATION

Author(s): Shivesh Singh, Shaurya Singh, Sudeept Singh Yadav and Avneesh Kumar

Special Issue Editor: Dr. Vikash Yadav
Received: 30/03/2022; **Accepted**: 11/05/2022; **Published**: 22/05/2022;
e-ISSN: 2347-470X;
Paper Id: 0422SI-IJEER-2022-05;
Citation: 10.37391/IJEER.100215
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100215

This article belongs to the Special Issue on **Recent Developments in Communication Technology using Machine Learning Techniques**.

1. INTRODUCTION

The Study of land price patterns is considered crucial in assisting with urban planning decisions. The real estate market is a stochastic and intrinsically unpredictable process. Investors make choices based on market conditions in order to maximize their profits. For their decision-making, developers want to know about future patterns. To make accurate predictions and future trends, a large number of details for houses is required for predictions, modeling, and evaluating. When a past data weas examined it showed that the house prices have a non-linear pattern.

Because real estate is such a fast-growing market, all parties involved must research and estimate land prices using statistical modelling and other computer tools. [2] The recent increase in the knowledge area, particularly Information Technology (IT) and IT-enabled services, can be linked to a variety of factors, the most significant of which being the recent expansion in population and industrial activity. Land demand began to rise, and housing and real estate activities grew rapidly.

All barren regions and paddy fields were bulldozed to create space for multi-story and high-rise residences. The real estate industry has seen a large increase in investment over the years, and we've noticed a non-uniform trend in terms of land pricing. Everyone in the sector, including the government, regulatory authorities, banking institutions, developers, and investors, recognized the necessity to foresee land price patterns.

Over the last two decades, a large number of longitudinal reports on land prices have been published. To shape more practical relationships, economic criteria could be added.

As our country continues to expand and the construction industry struggles to keep up with demand, prices will continue to increase as interest rates rise. Secure an investment property now with rigorous due diligence and watch your money grow over time.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 2 | Pages 138-143 | e-ISSN: 2347-470X

This Real Estate Value Prediction can be used to quickly identify trends and patterns. As a result, the emphasis of this work is on improving the free hold price prediction model.

The paper is divided into further sections that are –
Section 2- Comparative Study
Section 3 - Materials and Methods
Section 4 - Conclusion.

2. RELATED WORK

For the time being, each framework could be shifted toward innovation due to the ease of claiming activities. E-taking will become a part of the training structure. Individuals are increasingly moving away from manual to automated processes. The main purpose of this will be to estimate the cost of housing in accordance with the clients' plans. Those show tactics could be a lengthy process that requires clients to contact the land operator. The land operators provide an acceptable A suggestion for estimating lodging prices. As a result of this strategy's high risk, the land operator may provide inaccurate customer information. They use those simple relapse calculations to calculate the expense. This analysis is also used to predict the optimal location for customers to buy residences. Since 2009, the information used in this article has come from the Mumbai lodging board.

[3]A late worth of effort was put in to increase the value of the house. Different fiscal issues may have an impact on the house's worth. China, as we all know, is one of the most populous countries on the planet. The information is gathered from the over-proliferation of Taipei lodging, after which the price is estimated using a machine learning technique called a neural network, and The Root Mean Square Error can be used to figure out how accurate a prediction is.

The literature review will provide a clear notion for each project and will serve as a starting point. By conducting this research, I was able to learn more about both the benefits and drawbacks of the project, and I was able to complete it effectively.

3. METHODOLOGY USED

3.1 Python

Python is an elevated level, deciphered, and object situated programming language. Python is intended to be a profoundly intelligible.

One of the significant or the strength is the standard library which can be utilized for the accompanying usage.

1. AI
2. GUI Applications
3. Picture Processing
4. Web Scrapping
5. Web Frameworks
6. Text Processing
7. Mixed media and some more…

3.2 Jupyter Notebook

Jupyter Notebook is an open source web application or we can say it is stage that permits one to make, make and offer the archives and record in its arrangement and outer as well and furthermore to compose codes, conditions andperception.

Chiefly jupyter note pad has two sections:

1. Web Application: An intuitive program based instrument which permits maths, text and all the computational work done at an awesome rate.
Note pad report: Collection of the relative multitude of records and envelopes in the web application, including it's I/P, O/P gadgets.

Generally, we use jupyter in any of the limited internet browser yet google chrome is most appropriate. Since it is an internet browser running application, it gets the URL in location bar with http://localhost: It essentially suggests that our framework is functioning as a worker. Since it is a worker-based application, it is most appropriate for the protection of the archives in the application. The record put away in the jupyter note pad is with the .ipynb expansion.

Libraries are Python frameworks that handle often needed activities. I strongly advise any aspiring data scientists to become acquainted with the following libraries:

- Pandas are a set of tools for working with structureddata.
- Scikit Learn is a machine learning library.
- NumPy is a Python library for linear algebra andmathematics.
- For data visualisation, use Seaborn.

The proposed model's architecture includes the following stages: data selection, data preprocessing, feature scaling, model creation, and model evaluation (*Figure 1- Architecture of price prediction*). This section provides adescription of each step in the model construction process.

4. EXPERIMENTAL ANALYSIS

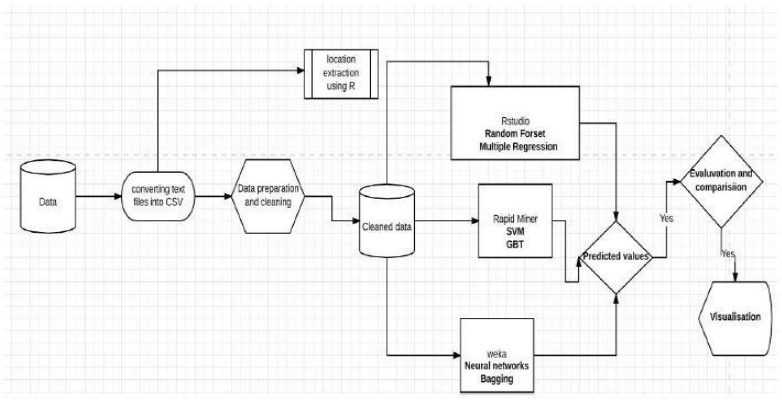

Figure 1: Architecture of price prediction

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 2 | Pages 138-143 | e-ISSN: 2347-470X

Machine learning is used to develop programs when provided some kind of data, understands it and then learns on its own. It is a part of artificial intelligence that has pre- defined libraries that can learn on its own without being externally programmed. Machine learning is focuses on areas where programs can be changed when exposed to new data.

The basic process starts with by observing the data, as to what the data is all about, such as instructions and/or direct experience, so that new trends can be searched in it ispossible to use the data to make informed decisions in the future based on examples [4]. However, text is treated as a series of keywords by using the data can be used to make informed decisions in the future based on the examples. The main goal is for computers to learn on their own, without human intervention, and modify their behavior according lying traditional machine learning algorithms; instead, a semantic analysis approach mimics the human ability to comprehend the context of a text.

It's commonly used in projects that include forecasting an outcome or identifying patterns. In such instances, a small amount of data is used to aid the machines in learning patterns that they can then apply to new input data to make right decisions. Supervised learning, unsupervised learning, and Reinforcement learning are the three major categories of machine learning.

4.1 Supervised Learning

Supervised learning is when we teach or train a computer using well-labeled data, which ensures that some of the data has already been marked with the correct response [5].

These algorithms basically labelled dataset to adapt what they've learned in the past and then make predictions on new data in order to predict future events.

The learning algorithm, in most of the cases, creates a function that uses a well-known training dataset to make predictions about target values. [6]The model will have targets for any new data after adequate planning.

4.2 Unsupervised Learning

Unsupervised machine learning algorithms, on the other hand, when training data that hasn't been categories or labelled, are used [7]. The system does not determine the correct performance, rather it explores the data and uses datasets to infer hidden structures from unlabeled data. Unsupervised learning is the process of teaching a computer to act on data that hasn't been categorized or labeled and allowing the algorithm to act on it without the need for human intervention without any prior data training, the machine's duty is to sort unsorted data into categories based on similarities, patterns, and differences [8]. Because there is no instructor present, unlike supervised learning, the machine will not be instructed. As a result, computers are limited in their ability to discover secret structure in unlabeled data on their own.

Reinforcement machine learning algorithms are a type of learning algorithm that interacts with its surroundings by generating actions and detecting errors or rewards. [9]Trial and error quest and delayed reward are two of the most significant features of reinforcement learning.

When there is a lot of ground truth data but no clear connection between the elements that generate the ground truth, machine learning shines [10]. Machine learning applications include virtual personal assistants, video surveillance, social networking services, email spam and virus filtering, and search engine result refinement, to name a few.

4.2.1 Data Collection

The first step in creating a machine learning model is to collect data. This is a critical phase with a cascading influence on the model's performance; the more data we collect, the better our model will work. The "Bengaluru house price prediction.csv" dataset is a raw dataset *(Figure 2– Importing CSV file and Data Set).* It implies that a significant amount of pre-processing is needed before any of the data can be used for evaluation. Our dataset is fairly large, with 7109 rows and 19 features that will aid us in predicting the sale of the property.

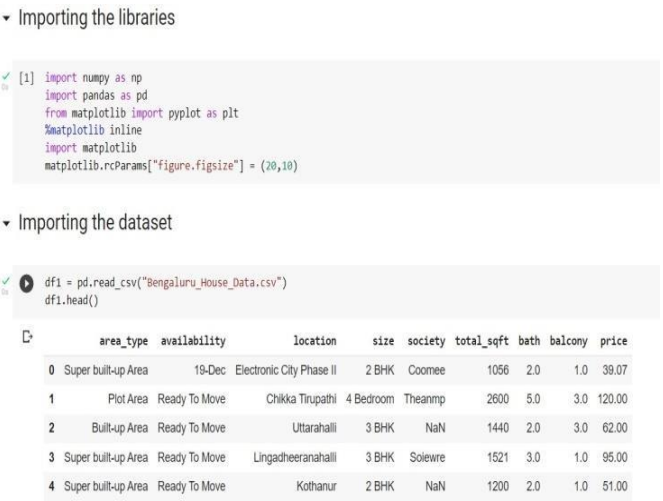

Figure 2: Importing CSV file and Data Set

4.2.2 Data Preprocessing

It entails translating the data set into the best format possible so that we can extract all of the features needed to predict the house price. This is where all of our data is cleaned, whether it's missing values, redundant values, or the inclusion of various features based on our requirements *(Figure 3– Data Preprocessing).* The "NaN" or "Null" indicators are often used to describe missing values. There are many options for dealing with them once they've been detected –

4.2.2.1 Samples or features with missing values should be removed to avoid deleting any valuable information or too many samples.

4.2.2.2 Use pre-built estimators, such as Scikit-Imputer Learn's class, to fill in the missing variables. We'll fit the data, then alter it to figure out where the missing numbers are. The mean value of the remaining samples is often used to fill in the missing values.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 2 | Pages 138-143 | e-ISSN: 2347-470X

	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2	3699.810606
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4	4615.384615
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3	4305.555556
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3	6245.890861
4	Kothanur	2 BHK	1200.0	2.0	51.00	2	4250.000000
5	Whitefield	2 BHK	1170.0	2.0	38.00	2	3247.863248
6	Old Airport Road	4 BHK	2732.0	4.0	204.00	4	7467.057101
7	Rajaji Nagar	4 BHK	3300.0	4.0	600.00	4	18181.818182
8	Marathahalli	3 BHK	1310.0	3.0	63.25	3	4828.244275
9	other	6 Bedroom	1020.0	6.0	370.00	6	36274.509804

Figure 3: Data Preprocessing

4.2.3 Data Analysis

Then we analyze the data and select the features. We get to know the number of features, their mean values, standard deviation, min and max values, etc.

4.2.3.1 Univariate Analysis

An analysis of data using a univariate approach is the simplest analysis of data. A univariate analysis analyses data with only one variable, while a multivariate approach analyzes data with many variables [11]. The major purpose of a descriptive analysis is not to find relationships between causes or effects, but to summarize and look for patterns. Using different methods, we analyze randomly selected features, such as bar graphs, frequency distributions, plots, etc.

4.2.3.2 Bivariate Analysis

In a bivariate analysis, two features are taken together and analyzed to determine if there is a relationship between them *(Figure 4 – Bivariate Analysis Graph).* This is one of the simplest types of analysis techniques. We use the same methodology of randomly selecting any two features, a pair at a time, and analyzing them via histograms, bar graphs, plots, etc.

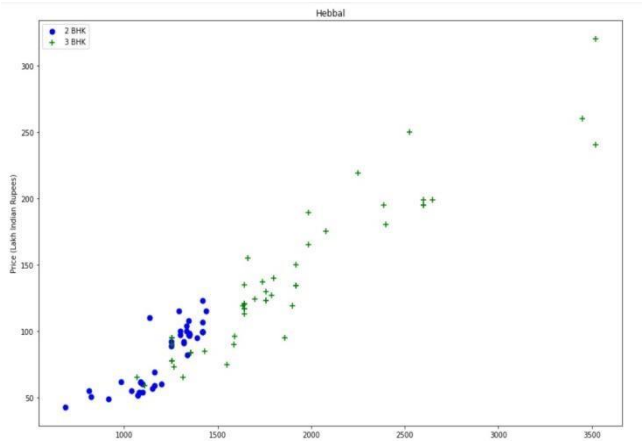

Figure 4: Bivariate Analysis Graph

4.2.4 Feature Scaling

During the preprocessing phase, this process is crucial because the majority of machine learning algorithms work much better when they deal with characteristics with the same scale[12]. Techniques most commonly used include: Normative scaling involves rescaling the features to an interval of [1,0], which constitutes a special case of min- max scaling. We will only need to scale each feature columnusing the min-max method to normalize the data.

$$X_{changed} = (X - X_{min}) / (X_{max} - X_{min}) \quad .....(1)$$

"Standardization" simply involves standardizing each feature column at mean zero with a standard deviation of 1, so that the columns have the same parameters as a standard normal distribution"[13]. By doing this, it is much easier for the algorithms to determine what parameters to learn. This ensures the algorithms are less sensitive to outliers, while also retaining useful information on them.

$$Z = (X - mean) / std. \quad .....(2)$$

4.2.5 Model Building

During this step, the actual machine-learning algorithms will be implemented. According to the Chennai house price prediction model, we are using linear regression machine learning algorithms to predict house prices.

4.2.5.1 Separating dependent and independent variables

The independent variables are the inputs to the process under study *(Figure 5- Graph Analysis).* Variables that depend on each other are the results. For example:

$$y = f(x) \quad .....(3)$$

Where, "X" equals an independent variable "Y" stands for the dependent variable.

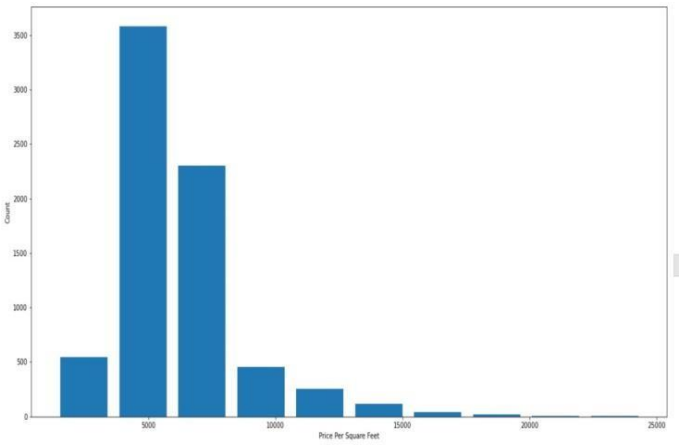

Figure 5: Graph Analysis

There change can be either positive or negative. All other features are considered independent variables in our model, with the target/dependent variable being "SALES_PRICE".

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 2 | Pages 138-143 | e-ISSN: 2347-470X

4.2.5.2 Splitting the Data Set into Train and Test Dataset

In our data analysis, we will split it into the training sets, testing sets, and validating sets. Our model is first trained with training data, evaluated with validation data then tested one last time on test data. In the end, the modelshould be able to generalize unseen data well, i.e. predict accurate results based on the parameters that were adjusted during training and validation.

In our Model, we have divided our dataset into a 70:30 ratio, which is to say, 70 percent of the dataset is training data, while the remaining 30% is testing data.

4.2.5.3 Linear Regression

A linear regression analysis involves the identification of a relationship between predictor variables and a continuous response variable, which can be used to predict a continual outcome.

A straight line, for instance, can be fitted for X and Y given methods to determine the coefficients between the fitted line and the sample points.

We will use the intercept and slope found to forecast the outcomes of new data using the fitted line.
Straight lines are defined by the following formula:

$$y = B0 + B1*x + u \quad .....(4)$$

Considering only the variables x and y, the only variables affecting the result are B0 and B1. These two values (B0 and B1) are the "weights" of the predicting function.

Taking these weights and biases and arranging them into a matrix produces the results. Repeating the process one step ata time iterates the process. As the line is iterated, it becomes more accurate and closer to the ideal one.

Random Forest: The regression forests are a type of random forest technique that can be used to predict both classification and regression. The basic technique entails the creation of a huge number of decision trees from a random set of data and variables, and then assigns a class of dependent variable to each tree.

Gradient Boosting: Both regression and classification can benefit from gradient boosting. Gradient boosting works by repeating the process of learning to calculating the error residual using a basic regression predictor for the data. We learn a novel model to estimate the error residual based on the amount of error per data point.

Root Mean Square Error: The root mean square error (RMSE) is a popular formula for calculating a regression model's error rate; however, it can only be used to compare models whose mistakes are quantified in the same units.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \quad ..... (5)$$

4.2.5.4 Evaluation of the model

Using appropriate evaluation matrices is the last step of the modeling process. A score function [score ()] and R2- squared metrics [14] were used to assess our model as they were perfectly suited to our model.

5. CASE STUDY

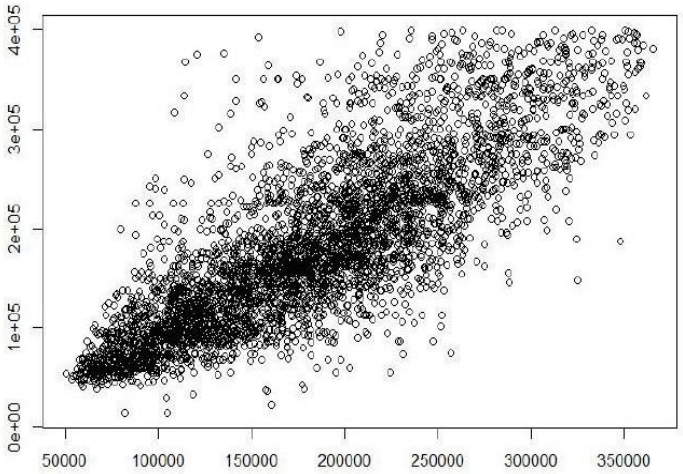

Figure 6: Random Forest Predicted vs Actual

The actual and anticipated values are distributed throughout linearly in this scatter plot *(Figure 6 – Random Forest Predicted vs Actual),* resulting in a price prediction accuracy of roughly 90.1% with a Root Mean Square Error of 0.0120.

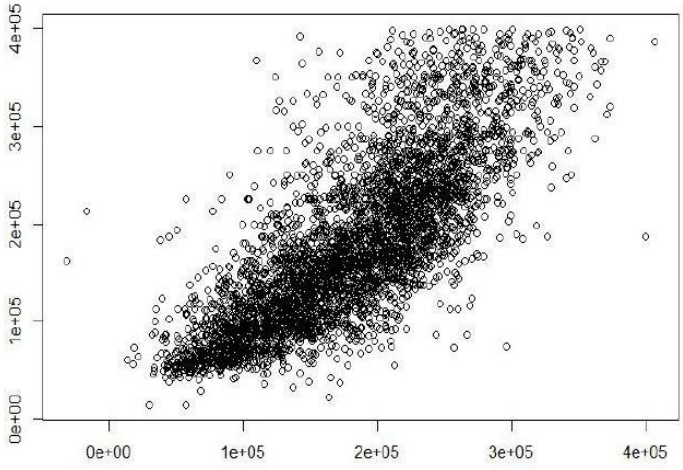

Figure 7: Multiple regression vs Actual

The preceding scatter plot for multiple regression reveals that values are not spread linearly *(Figure 7 - Multiple regression vs Actual),* therefore the actual and predicted price values are not identical, but the predicted percentage is around 54.9% with a Root Mean Square Error of 0.701.

When the results of this experiment are compared to the predicted results, these algorithms perform admirably.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 2 | Pages 138-143 | e-ISSN: 2347-470X

6. CONCLUSION

The primary goal of this study is to generate a price forecast, which we have accomplished effectively using a Random forest, gradient boosted trees, etc., among other machine learning techniques.

Table 1: Represents Error and Accuracy

ROOT MEAN SQUARE ERROR			
S. No.	**Algorithm**	**RMSE Error**	**Accuracy**
1	Random forest	0.012	90.1%
2	Neural Network	0.590	60%
3	Gradient Boosted	0.573	65%
4	Bagging	0.563	70%
5	Support Vector machine	0.636	58%
6	Multiple Regression	0.70	54.9%

Many factors, such as surrounds, marketplaces, and other associated variables, can be added to the properties to improve price prediction. The projected data may be saved in databases, and an app can be developed for users so that they can get a quick idea and invest their money safely.

In order to estimate prices, a variety of strategies have been utilized, such as hedonic regression, and in this study, I am integrating machine learning techniques and past research to estimate future real estate prices. As a result, it would be beneficial for people to be aware of both current and future conditions in order to prevent making mistakes.

The experiments revealed that random for-set and gradient boosted tresses work better with higher accuracy percentages and lower error values (*Table 1 – Represents Error and Accuracy*) when compared to other machine learning techniques. These algorithms perform brilliantly when compared to the projected results of this experiment.

7. FUTURE WORKS

The scope of the project is achieved to what it was implemented the accuracy is good and the working stability is also high. In the upcoming phase of our project we will be able to connect an even larger dataset to this model so that the training can be even better.

Also we will try out other dimensionality reduction techniques [15] like Uni-variate Feature Selection and Recursive feature elimination in the initial stages.

REFERENCES

[1] Fisher, R.A. "The goodness of fit of regression formulae, and the distribution of regression coefficients". Journal of the Royal Statistical Society, 1992.

[2] Sampathkumar, V. M. Helen Santhi, and J. Vanjinathan. “Forcasting the Land Price UsingStatistical and Neural Network Software”, 2015.

[3] Koskela, T., Lehtokangas, M., Saarinen, J. and Kaski, K. Time series prediction with multilayer perceptron, fir and elman neural networks, Proceedings of the World Congress on Neural Networks, INNS Press San Diego, USA, 1996.

[4] Ying He, Fei Richard Yu, Nam Zhao, Hongxi Yin, Haipeng Yao, Robert C. Qiu. “Big Data Analyticsin Mobile Cellular Network”, IEEE, 2016.

[5] Yan, Xin, Linear Regression Analysis: Theory and Computing, World Scientific, 2009.

[6] “Proceedings of Data Analytics and Management”, Springer Science and Business Media LLC, 2022.

[7] Vapnik, V. N. The Nature of Statistical Learning Theory (2nd Ed.), Springer Verlag, 2000.

[8] Abhirami K, Alok Kumar Johri and Udayan Chanda. “Software Testing Resource Allocation and Release Time Problem: A Review”, InternationalJournal of Modern Education and Computer Science, 2014.

[9] Roman, Victor (2019-04-21). "Unsupervised Machine Learning: Clustering Analysis". Medium. Retrieved 2019.

[10] Quoc-Viet Pham, Fang Fang, Vu Nguyen Ha, Md. Jalil Piran, Mai Le, Long Bao Le, Won-Joo Hwang, Zhiguo Ding. “A Survey of Multi-Access Edge Computing in 5G and Beyond: Fundamentals, Technology Integration and State-of-theArt”, IEEE Access, 2016.

[11] Kaelbling, Leslie P., Littman, Michael L., Moore, Andrew W. "Reinforcement Learning: A Survey". Journal of Artificial Intelligence Research, 1996.

[12] Grus, Joel. Data Science from Scratch. Sebastopol, CA: O'Reilly, 2015.

[13] Ganjisaffar, Y., Caruana, R. and Lopes, C. V. Bagging gradient-boosted trees for high precision, low variance ranking models, Proceedings of the 34th international ACM SIGIR conference onResearch and development in Information Retrieval,2011.

[14] Li, L. and Chu, K.-H. Prediction of real estate price variation based on economic parameters, Applied System Innovation (ICASI), International Conference on, IEEE, 2017.

[15] Van der Maaten, Laurens; Postma, Eric; van den Herik, Jaap. "Dimensionality Reduction: A Comparative Review", 2019.