Proceedings of the International Conference on Applied Artificial Intelligence and Computing (ICAAIC 2022)
IEEE Xplore Part Number: CFP22BC3-ART; ISBN: 978-1-6654-9710-7

Real-Estate Price Prediction System using Machine Learning

[1]Veerraju Gampala, [2]Nalajala Yaznitha Sai, [3]Tadikonda Naga Sai Bhavya
[1, 2, 3]Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur, Andhra Pradesh 522502, India.

2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) | 978-1-6654-9710-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAAIC53929.2022.9793177

Abstract—
A common reason for home purchases to be done as investments is the desire to make a return on the property that is acquired. Usually, they are seeking for answers to the same questions: when and where should they purchase a property, and what may be the most favorable rent or selling price in the next few years, respectively, are all important considerations. In this project, the price of real estate can be anticipated via the use of machine learning approaches and algorithms. A range of algorithms, including logistic regression, Naive bias, decision trees, and Random Forest Classifier algorithms, are examined in more depth throughout this project. Following an initial analysis of the data using various charting methodologies, the value of the estate is estimated through a variety of algorithms, and then the most accurate algorithm among the cluster of algorithms utilized is evaluated to make the projections.

Keywords: Real-estate, housing, planning, price machine learning, Linear regression, Naive bayes, Decision Tree, Random Forest classification.

I. INTRODUCTION

Our project is based on the concept of using artificial intelligence and machine learning to real-estate investments, and our objective is to provide people with the benefit of making well-informed investment decisions. As a company, we have always focused on providing solutions to the following questions:

If you're looking to make the most money in real estate, what sort of property should you buy to optimize your profits?

Will they be able to purchase the property, and if so, when, and where will they be able to do this?

The most appropriate rental or selling price for a piece of real estate, whether it is being leased out or sold for a short or long period of time, is determined by many factors.

These kinds of questions not only help folks in making excellent financial choices, but they also assist them in gaining a better grasp of how things operate in general.

To overcome this challenge, machine learning makes use of the huge amount of data that is now available to the researcher. Data has been assessed to have a better understanding of its specialty before machine learning algorithms have been applied to our advantage in data processing.

The problem-solving process included employing a cluster of the best algorithms to determine which algorithm from a cluster of machine learning algorithms was most likely to produce the desired results in terms of problem-solving outcomes as part of the process.

Aim & Objectives-

The major purpose of this research is that we utilize a cluster of machine learning models to estimate the price of the home and perform a comparison analysis to determine which approach works the best.

II. LITERATURE SURVEY

[1] The researchers in this study article were able to predict the value of homes that were due to reach the market at the time the study was done by using machine learning algorithms that they had developed themselves. Providing a high-level overview of the approach that will be used to analyze the dataset and detect relationships between the parameters is the purpose of Section 3. Section 3 consists of the following items: Section 3: Because of this, people can choose components that are not tied to one another and are thus autonomous in their natural condition. In response to this information, a CSV file containing an estimated house price was created, and this file was then exposed to four distinct computational processes to arrive at the final findings.

After conducting an extensive investigation, it was discovered that the researchers had only implemented a small number of Machine Learning algorithms, which are essentially classifiers, and that they still needed to train many additional classifiers and better understand their predicting behavior for both continuous and discrete data. We anticipate that this research work will be beneficial in the development of applications for a variety of locations throughout the globe, as it will minimize the number of errors that are discovered during the development process.

The authors of this research build a prediction model for the expected selling prices of any real estate property in the future using the Decision Tree machine learning technique described in paper [2], which has been explored in detail elsewhere]. Several other variables, such as air quality and crime rate, were included into the dataset to improve the accuracy of the pricing estimates, as well. Because other prediction systems' datasets often do not contain these characteristics, our system stands out from the rest of the field. Because these characteristics have an impact on people's decisions when acquiring a home, it seems reasonable to include them into the process of projecting house prices. For users that supply input that is comparable to the system's predictions, researchers may conduct a comparison study between the expected price produced from the system and the predicted price derived from real estate websites such as

Housing.com. In addition, to make things easier for the client, they would propose real estate properties to the user based on the projected price of the property they were seeing. It is planned that in the future, the dataset would be expanded to include information on other Indian cities and states. Currently, the dataset solely contains information about the city of Mumbai. For everyone's benefit, Gmap will be integrated into the system to make it even more informative and user-friendly. If the given site is within a one-kilometer radius of the specified location, this will display the surrounding facilities, such as hospitals and schools, that are located within that radius of the stated location. Since the existence of such elements enhances the value of real estate property in the first place, it is possible to take this into account during the forecasting process.

An automated method has been developed with the purpose of producing an accurate estimate of future property values, according to the information provided in article [3]. To get the best potential results, the approach takes considerable use of other statistical methods, such as Linear Regression, Forest Regression, and boosted regression. It has been feasible to further increase the efficiency of the algorithm via the use of neural networks, allowing it to become even more efficient in the process. Because the technique would provide exact findings while also decreasing the probability of purchasing the erroneous home, customers will be overjoyed with the outcomes. The ability to make system enhancements that are beneficial to the client without interfering with the system's basic function is also available.

Many ways may be employed to improve the accuracy of the system. The following are some examples: It will be possible to allow the incorporation of a higher number of cities if the system's overall size and processing capabilities are both enhanced. Furthermore, they may use Augmented Reality to include a range of user interface and user experience approaches to provide a more realistic representation of the results in a more engaging manner.

[4] Multiple Linear Regression, Ridge Regression, LASSO Regression, Elastic Net Regression, Ada Boosting Regression, and gradient boosting were among the machine learning algorithms tested in the authors' work, with the comparison of these algorithms in the context of housing price prediction analysis being the primary focus. If the results of the previous experiment are followed, the gradient boosting algorithm exceeds all other algorithms in terms of accuracy when it comes to assessing the value of real estate assets, according to the findings of the study. To compute the accuracy value of the technique on the King County Source data set, it is important to utilize the [MSE] Mean Square Error as well as the [RMSE] Root Mean Square Error. Our team was able to determine the King County Source, which was obtained from a publicly available data source and used in this analysis. Using the approaches described above to forecast the resale value of a property in the future might be used to further enhance the study's conclusions in the future. It is still vital for researchers to train many classification algorithms and understand their predicting behavior for both continuous and discrete values, since only a few Machine Learning approaches, which are truly classifiers, have been applied. The results of this study have the potential to lead to improved error values being used in the development of apps for various cities all over the globe because of the findings of this research.

These three machine learning algorithms [5], which include the decision tree classifier, the decision tree regression, and the multiple linear regression, are among the most fundamental machine learning algorithms now accessible to academics, and they are all freely available on the internet. This study utilizes all three approaches at the same time to analyze data. Thanks to the assistance of the machine learning application Scikit-Learn, the work was completed with success. The tactics used to aid customers in anticipating the availability of properties in the city, as well as the price of the residences they are interested in acquiring, are customized to each customer's demands, and used in this position. It was necessary to utilize two distinct forecasting approaches, both of which were employed in combination with one another, to accurately predict the prices of the dwellings. The techniques employed in this study were decision tree regression and multiple linear regression, which were both used in tandem with one another. It was discovered via a review of the literature that when it comes to estimating the value of real estate properties, the performance of multiple linear regression consistently beats the performance of decision tree regression in the great majority of cases. A model for estimating the value of a home in the future may be developed using sophisticated machine learning methods, and this is a doable goal. Aside from this, it is likely that the dataset will be expanded to include more elements in the not-too-distant future if the need arises. Researchers may opt to conduct a comparison study between the projected price from the system and the price from real estate websites such as Housing.com, utilizing the same user input that was used in the prior study in order to get their results in the subsequent study.

III. Data analysis

An in-depth data analysis of the information we have obtained regarding the pricing of dwellings has been carried out by our team of analysts.

The pair plots were designed to discover the most efficient collection of attributes for describing a relationship between two variables or for producing the most separated clusters that could be formed. We may also develop some simple categorization models in our data set by drawing some basic lines or generating linear separations in it, which will be important later.

Figure 1: Pair plots of the dataset.

The plot below depicts the evolution of the price of a house over time.

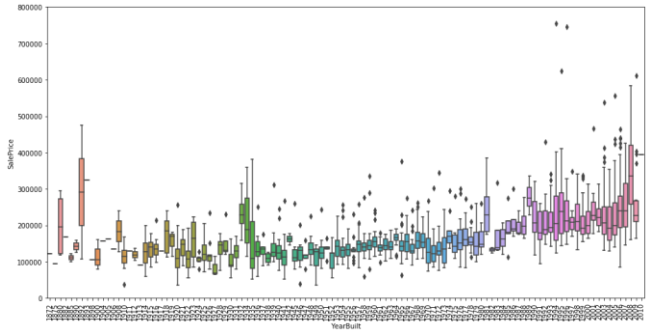

Figure 2: Box Plot

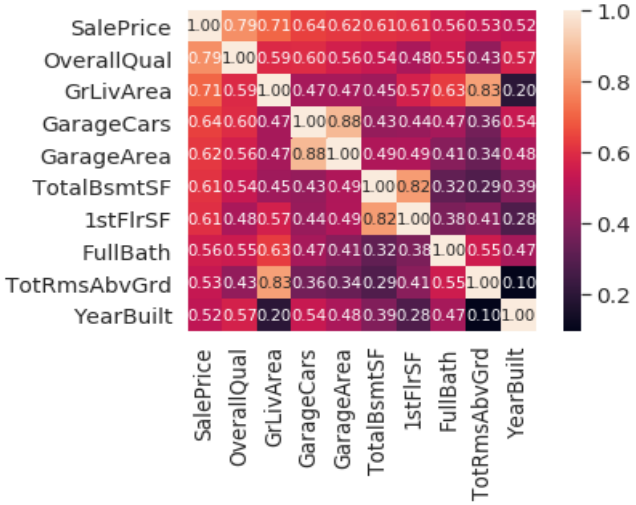

Figure 3: Heat Map

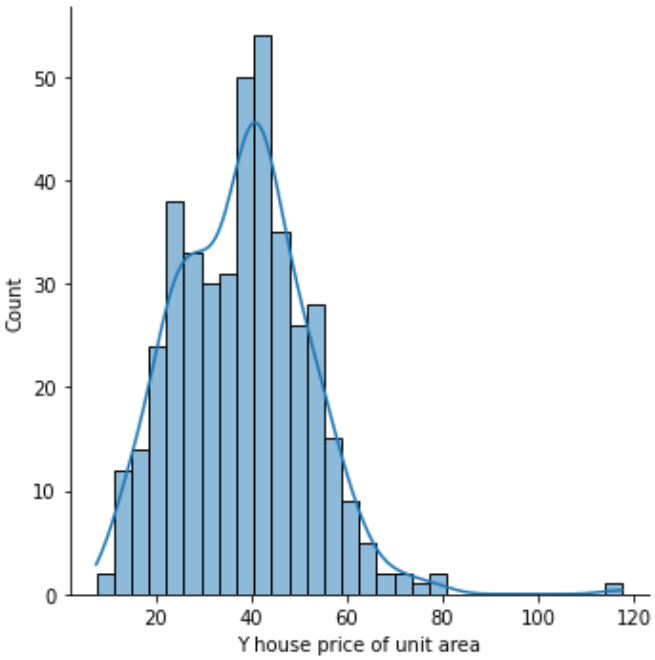

Figure 4: Evolution of the price of houses over time.

Outliers are seen in the following graphic, which builds a boxplot for the full dataset.

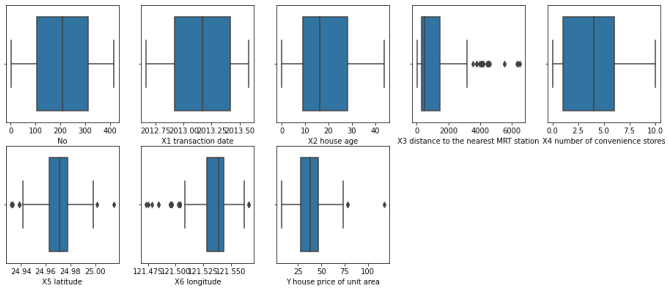

Figure 5: Box plot

For better understanding and interpretation of the behavior of the full dataset, we have created a heat map.

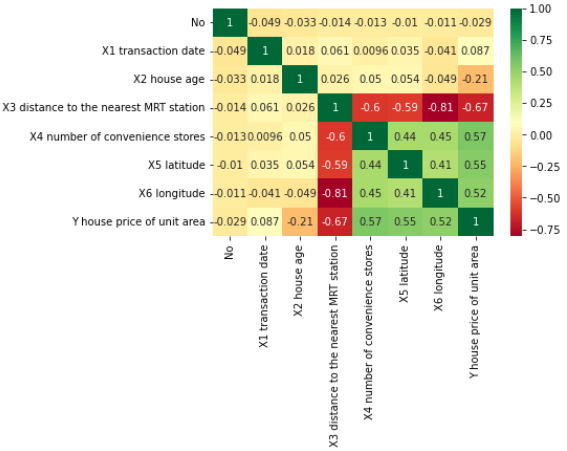

Figure 6: Heat Map

IV. Methodology

We have made use of the clustering machine learning methods that are listed below.

1) Linear Regression
2) Naive Bayes
3) Decision Tree
4) Random Forest Classification.

We've undertaken a thorough investigation to determine which of them will provide the best outcomes in the long run.

We first select the goal and feature variables to do linear regression in this manner. Then we split the dataset into two parts: the training dataset and the testing dataset, which we will discuss later.

It is then used to train a machine learning model, which is subsequently tested against the testing data set to see if the trained machine learning model properly predicts the outcome of the experiment.

It is a kind of regression that is linear in nature, which is the linear regression. It is also found that a coefficient matrix, which holds the values of the coefficients of the variables in a set of linear equations, may be constructed. When working with systems of linear equations, the matrix may be quite useful in many situations. A system of linear equations is a collection of one or more linear equations that all include the same set of variables and that are all solved at the same time in mathematics.

The same dataset that was used for linear regression is then fitted into a Naive Bayes machine learning model, Decision Tree, and Random Forest machine after which the model is then evaluated using the testing dataset that was previously used for linear regression.

Since in statistics and optimization, errors and residuals are two closely related and frequently misunderstood measurements of the difference between an observed value of an element in a statistical sample and its "theoretical value," errors and residuals are often used interchangeably. In statistics, the error (or disturbance) of an observed value is the deviation of a quantity of interest from its (unobservable) true value (for example, the population mean), and the residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest (for example, the mean of the population) (for example, a sample mean). There is a particularly noticeable difference in regression analysis, where the concepts of regression errors and regression residuals are commonly referred to as regression errors and regression residuals, respectively, and where they lead to the concept of studentized residuals, which is derived from the concepts of regression errors and regression residuals.

A line of best fit is produced whenever a simple linear regression (or any other kind of regression analysis) is performed. The data points do not always fall exactly on the line formed by the regression equation; rather, they are distributed across the data set because of the distribution of the data points. A residual is a vertical gap between a data point and the regression line that is used to estimate the relationship between the two variables. Each data point has a single residual, which is represented by the symbol R. The names of them are as follows: When the points are located above the regression line, they are deemed positive; when they are located below, they are considered negative; and when they are in the middle, they are designated as zero. Because residuals are defined as the difference between any data point and a regression line, they are usually referred to as "errors" in statistical analysis. This kind of inaccuracy does not necessarily mean that the study was flawed; rather, it just implies that there is some unexplainable difference between the two findings. So, the residual is the amount of error that cannot be explained by the regression line because of this process:

It is also feasible to express the residual(e) using an equation in addition to the method described above. The difference between the predicted value () and the actual value () is represented by the e symbol (). The scatter plot depicts a set of data points that have been seen, but the regression line depicts a prediction.

As soon as we have finished training and testing all the machine learning models, we can move on to the assessment step, where we will view the Model as a Case in Point. The evaluation phase of the model generating process is a vital step in the entire process of model building and should not be overlooked. It supports the selection of the most suitable model to describe our data, and it also assists in the forecast of how well the selected model will work soon. It is not acceptable in data science to evaluate model performance using data that was used for training since it is likely to result in overoptimistic and overfitted models, which is not desired given the nature of the data. When it comes to testing models in data science, the hold-out and cross-validation procedures are two of the most often used techniques. Both methodologies evaluate model performance to a test set that is not accessible to the model to avoid overfitting. This approach divides a huge dataset into three portions, each of which includes a distinct random number generated by a random number generator. An example of a training set is a subset of a dataset that is used to construct prediction models while creating prediction models. Following the generation of a model during the training phase, a validation set is constructed from a subset of the dataset that is used to evaluate the model's performance. Model parameters are fine-tuned in a testing environment, and the best overall performance is determined by choosing the model with the greatest overall performance from many candidates. When it comes to modeling approaches, it is not always required to employ a validation

set. Test sets are subsets of a dataset that are used to predict the projected future performance of a modeling system. It is often referred to as "unexpected occurrences." When a model performs much better on the training set than it does on the test set, overfitting is most likely to be the cause of the poor performance.

V. RESULTS AND DISCUSSION:

When it comes to determining the overall quality of the machine learning model, it is probable that residuals will play a major role.

If the residual of a specific machine learning model is zero, this suggests that the machine learning model predicts perfectly in that situation. The results of the machine learning model are shown in the following images.

Linear Regression

Iterations	Y_Test	Y_Pred	Residuals
176	19.2	12.0802665	6.397335
347	11.2	9.549151	1.650849
307	24.7	22.516894	2.183106
299	46.1	48.213227	-2.113227
391	31.3	31.972364	-0.672364

Table 1: Test results for Linear Regression

Naive Bayes

Iterations	Y_Test	Y_Pred	Residuals
176	19.2	12.0802665	6.397335
347	11.2	9.549151	1.650849
307	24.7	22.516894	2.183106
299	46.1	48.213227	-2.113227
391	31.3	31.972364	-0.672364

Table 2: Test results for Naive Bayes

Decision Tree

Iterations	Y_Test	Y_Pred	Residuals
176	19.2	12.0802665	6.397335
347	11.2	9.549151	1.650849
307	24.7	22.516894	2.183106
299	46.1	48.213227	-2.113227
391	31.3	31.972364	-0.672364

Table 3: Test results for Decision Tree

Random Forest

Iterations	Y_Test	Y_Pred	Residuals
176	19.2	12.0802665	6.397335
347	11.2	9.549151	1.650849
307	24.7	22.516894	2.183106
299	46.1	48.213227	-2.113227
391	31.3	31.972364	-0.672364

Table 4: Test results for **Random Forest**

VI. Conclusion and Future Scope:

Considering our findings and extensive study, we have concluded that Linear regression has done the best, achieved an accuracy of 95 percent while had a mean squared error of just five percent, and that it is the best method for predicting house values. When used in conjunction with a machine learning model that performs the best for house price prediction, the provision of a fully featured user interface is an optional feature that may now be included in our proposed system now that we have identified the machine learning model that performs the best for house price prediction. With the machine learning model in place, customers will be able to access a few capabilities from a variety of different geographical areas. This interface will assist clients in making more informed decisions about which home to purchase and in receiving better customer service.

VII. References

1. Shinde, Neelam & Gawande, Kiran. (2018). Survey on predicting property price. 1-7. 10.1109/ICACE.2018.8687080.
2. Kuvalekar, Alisha and Manchewar, Shivani and Mahadik, Sidhika and Jawale, Shila, House Price Forecasting Using Machine Learning (April 8, 2020). Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST) 2020, Available at SSRN: https://ssrn.com/abstract=3565512 or http://dx.doi.org/10.2139/ssrn.3565512
3. A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.
4. C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," 2019 International Conference on Smart Structures and Systems (ICSSS), 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.
5. M. Thamarai, S P. Malarvizhi, " House Price Prediction Modeling Using Machine Learning", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.12, No.2, pp. 15-20, 2020. DOI: 10.5815/ijieeb.2020.02.03
6. Heidari, Maryam, and Setareh Rafatirad. "Semantic convolutional neural network model for safe business investment by using bert." 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE,
7. Abutahoun, Bushra, Maisa Alasasfeh, and Salam Fraihat. "A framework of business intelligence solution for real estates analysis." Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems. 2019.

8. Huy, Tran Duong, and Anh Ngoc Le. "Clustering helps to improve price prediction in online booking systems." International Journal of Web Information Systems (2021).

9. Koncilja, Aleš. Napovedovanje vrednosti nepremičnin iz podatkov Evidence trga nepremičnin. Diss. Univerza v Ljubljani, 2018.

10. S.C. Dharmadhikari, Veerraju Gampala, Ch. Mallikarjuna Rao, et al., A smart grid incorporated with ML and IoT for a secure management system, Microprocessors and Microsystems, Volume 83, 2021, https://doi.org/10.1016/j.micpro.2021.103954.

11. V. Gampala, J. Vallapuneni, P. Kumar Ande, R. Kumar Indurthi and N. Rajesh, "Comparative Study on Telugu text Classification using Machine Learning and Deep Learning models," *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2021, pp. 1393-1398, doi: 10.1109/ICOEI51242.2021.9453040.

12. Rajendran, R., Piali, B., Chandrakala, P., Gampala, V. and Majji, S. (2022), "Role of digital technologies to combat COVID-19 pandemic", World Journal of Engineering, Vol. 19 No. 1, pp. 72-79. https://doi.org/10.1108/WJE-01-2021-0043

13. Gampala, V., Nandankar, P.V., Kathiravan, M., Karunakaran, S., Nalla, A.R. and Gaddam, R.R. (2021), "Early prediction and analysis of corona pandemic outbreak using deep learning technique", World Journal of Engineering, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/WJE-03-2021-0145

14. Gampala, V., Maram, B., Vigneshwari, S., and Cristin, R. (2022), "Glaucoma detection using hybrid architecture based on optimal deep neurofuzzy network", International Journal of Intelligent Systems, Vol. ahead-of-print No. ahead-of-print. http://doi.org/10.1002/int.22845