International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

Comparison of Machine Learning Algorithms for House Price Prediction using Real Time Data

Swarali M. Pathak, Prof. Archana K. Chaudhari
Department of Instrumentation and Control Engineering
Vishwakarma Institute of Technology, Pune

Abstract: **Housing prices are a crucial reflection of the economy, and property values are of great interest for consumers as well as sellers. Real Estate is the one of the least transparent industries in our ecosystem. Predicting house prices with real time factors is the main aim of this research project. This paper aims to make evaluations based on some basic parameters which are considered while determining the price of a house. In order to carry out the real time research, real time housing data of Pune City has been collected manually. The project tends to use Regression technique for Machine learning as we are dealing with continuous outcome variable. We have carried out a research by implementing different regression models to compare and determine the most effective model to resolve given problem statement. The goal of this research project is to create an effective machine learning model that is able to accurately estimate the price of the house based on given features and deploy the machine learning model in the form of a website to reach out individuals.**

Keywords: Linear Regression, Machine Learning, Random Forest, Real Estate, Real-time Data, Support Vector Regressor.

I.INTRODUCTION

In the past years, Machine learning has proven to be able to solve real world problems using various algorithms. It plays a major role in advances of medical imaging, spam and fraud detection, enhancements in automobile industry, security alerts and Business Analysis. In this paper, we have used machine learning algorithms to perform predictive analysis of house prices to provide an overview of real estate businesses and property demand. Data is the most important part for analysis of any problem. It provides the information in a detailed format which is able to be understood by machines. Real estate prices keep changing frequently based on certain parameters. In 2020, the average value of property prices in Pune costs around Rs 6,573 per sqft as per listings on Housing.com. For a Real estate Business, data is the most important source for analysis and predictions. It is always a perk to know about the predictions of variations of an entity which will be happening near future and business managers can act accordingly to avoid future loss. And for this we need a most accurate predicting Model for analysis. Similarly, we need a proper prediction on the real estate and the houses in the housing market to provide appropriate estimation of prices to help real estate managers know about prophecies.

Buying a house will be a life time goal for most of the individuals but there are a lot of people who make huge mistakes while buying the properties. One of the common mistakes is buying properties that are too expensive but it's not worth it. Various methods have been used in the price prediction. This project aims to predict the real estate price using the machine learning techniques with the help of the Real-Time Data of houses in Pune, India. The goal of this statistical analysis is to help us understand the relationship between house features and how these variables are used to predict house price. It uses comparison of Regression algorithms to find out best fitting model to predict the house price. So, it would be helpful for the people to avoid them from making mistakes. The results proven that this approach yields minimum error and most accuracy than individual algorithms applied. The goal of this project is to make a machine learning model that is able to accurately estimate the worth of the house given the options.

II.LITERATURE SURVEY

[1] Real Estate Price Prediction with Regression and Classification: In this paper house prices are predicted using explanatory variables that cover many aspects of residential houses. House prices are predicted with various regression techniques including Lasso, Ridge, SVM regression and Random Forest. According to this paper, for a regression problem, the best-performing model is SVR with Gaussian kernel, with RMSE of 0.5271, however, visualization for SVR was difficult due to its high-dimensionality. According to its analysis, living area square feet, material of the roof and neighborhood have the greatest statistical significance in predicting a house's sale price. **[CS 229 Autumn 2016 Project Final Report Hujia Yu, Jiafu Wu [hujiay, jiafuwu]@stanford.edu].**

[2] A SVR based forecasting approach for real estate price prediction: The support vector machine (SVM) has been successfully applied to classification, cluster, and forecast. This study proposes support vector regression (SVR) to forecast real estate prices in China. The aim of this paper was to examine the feasibility of SVR in real estate price prediction. The experimental results were calculated based on the mean absolute error (MAE), the mean absolute percentage error (MAPE) and the root mean squared error (RMSE) and the SVR based approach was an efficient tool to forecast real estate prices. **[Hong Zhao, Rong-Qiu Chen, Wei Xu, Da-Ying Li Published in: 2009 International Conference on Machine Learning and Cybernetics].**

[3] Using machine learning algorithms for housing price prediction: This study used machine learning to develop housing price prediction models. This study analyzes the housing data of 5359 townhouses in Fairfax County, VA. The 10-fold cross-validation was applied to C4.5, RIPPER, Bayesian, and AdaBoost. [**The case of Fairfax County,**

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

Virginia housing data, Byeonghwa Parka, Jae Kwon Baeb,⇑aDepartment of Business Statistics, Hannam University, 70 Hannam-ro, Daedeok-gu, Republic of Korea].

[4] House Price Prediction Using Machine Learning and Neural Networks: This paper aims to make evaluations based on every basic parameter that is considered while determining the price. This model used various regression techniques in its pathway, and the results are not solely determined by one technique rather it is the weighted mean of various techniques to give the most accurate results. The results proved that this approach yields minimum error and maximum accuracy than individual algorithms applied. **[Ayush Varma, Abhijit Sarma, Sagar Doshi, Rohini Nair, Computer Engineering Department, KJ Somaiya College Of Engineering, Mumbai, Published in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)].**

[5] Valuation Of House Prices Using Predictive Techniques: This paper uses machine learning algorithms to predict the house prices. In this paper, algorithms such as logistic regression and support vector regression, Lasso Regression technique and Decision Tree are employed to build a predictive model. It had considered housing data of 3000 properties. Logistic Regression, SVM, Lasso Regression and Decision Tree show the R-squared value of 0.98, 0.96,0.81 and 0.99 respectively. Further comparisons of these algorithms are based on parameters such as MAE, MSE, RMSE and Accuracy. **[International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-5, Issue-6, Jun.-2018].**

III.METHODOLOGY

A. Dataset Description:

Dataset is the Real-time data of Houses for sale in Pune city. This is a manually collected data or often called as manual web scraping and it contains more than 1635 entries and 8 features. It is collected from housing websites like 99acres and magic bricks and housing.com. Initially, the data contains seven columns or attributes namely area type, location and name of the Society, number of BHK (bedroom, hall and kitchen), number of bathrooms, total square feet area, price and availability of the property. Amongst these price is our dependent variable. In the data cleaning process the removal of unnecessary data columns is done for the sake of accuracy and over fitting of the model. This Dimensionality reduction avoids the model from the curse of dimensionality and eventually the model predicts more accurately. In the cleaning process we reduce the area type, society and availability. After the data cleaning process now the model has five features amongst which four are independent and one output feature i.e. Price is a dependent variable.

area_type	location	Society	size	Baths	total_sqft	price	availability
Super Built-up Area	Bavdhan	Satyam Shrey	2 bhk	2	640	55.96	Under construction
Built-up Area	Sus	Kiran Sanskriti	2 bhk	2	690-720	45	Under construction
Built-up Area	Tathawade	Vivanta	2 bhk	2	688-693	60	Under construction
Super Built-up Area	Koregaon Park	Sunit Anant	3 bhk	3	1659	210	Ready to move
Built-up Area	Sinhgad Road	Nanded city	3 bhk	3	1450	78	Ready to move
Super Built-up Area	Koregaon Park	New Akshaya	2 bhk	2	1300	100	Ready to move
Super Built-up Area	Bibwewadi	Sukhsagar Nagar	2 bhk	2	850-950	50	Ready to move
Super Built-up Area	Koregaon Park	Tulip	2 bhk	2	1080	95	Ready to move
Built-up Area	Dhankavadi	Ganga Altus	3 bhk	3	850-950	87.5	Under construction
Built-up Area	Dhankawadi	Ganga Altus	2 bhk	2	650-750	67.5	Under construction
Built-up Area	Hinjewadi	Rohan Ipsita	2 bhk	2	645-716	59.43	Under construction
Built-up Area	Hinjewadi	Rohan Ipsita	1 bhk	1	346-384	37	Under construction
Plot Area	Moshi	Kasturi	2 bhk	2	627-628	72	Under construction
Super Built-up Area	Baner	Rachana	2 bhk	2	573-606	69	Under construction
Built-up Area	Katraj	Laxmi Golden Palm	2 bhk	2	850	55	Ready to move
Built-up Area	Katraj	DNK Business Bay	1 bhk	1	660	41.5	Ready to move
Super Built-up Area	Bibwewadi		1 bhk	1	500	45.6	Ready to move
Built-up Area	Kanchan Nagari	Mnik Moti Complex	3 bhk	2	1140	70	Ready to move
Built-up Area	Katraj		2 bhk	2	1020	70	Ready to move
Super Built-up Area	Katraj	KNN Yashashree	2 bhk	2	905	65	Ready to move
Plot Area	Mangdewadi		3 bhk	3	1800	99	Ready to move
Super Built-up Area	Bibwewadi	Laketown	2 bhk	2	1050	102	Ready to move
Super Built-up Area	Katraj	Wondercity	4 bhk	4	2100	164	Ready to move
Plot Area	Vighnaharta Nagar	Kulshree	1 bhk	1	607	35	Ready to move
Plot Area	Katraj	Hemant kamala city	1 bhk	1	605	45	Ready to Move
Built-up Area	Katraj		4 bhk	4	1620	150	Ready to move
Super Built-up Area	Katraj		5 bhk	6	6000	305	On Sale
Super Built-up Area	Katraj	Belvalkar Chaitranjan	1 bhk	1	680	37	Ready to move
Built-up Area	Bibwewadi	Pride Purple	2 bhk	2	865-1168	105.5	Under construction
	Bibwewadi	Mahesh Society	1 RK	1	350	25	Ready to move

Fig 1. Dataset

The data is split into training and testing sets. The 80-20 split used is a typical ratio for this purpose; 80% of the data has been considered as a training set and 20% as a test set. According to the analysis, Location plays the most important role in predicting the price. Then in the second place, the size of the property followed by the number of BHK and baths. We will analyze this is the implementation process.

B. Regression Techniques:

Machine Learning comes with two approaches of learning namely Supervised and Unsupervised Learning. Supervised Learning used needs that the data which is being used to the train the algorithm, has already some samples of correct outcomes. This is the most commonly used method as it increases the chances of getting much accurate results. We have performed supervised learning approach for this project. Supervised Learning is further divided into two groups Regression and Classification. The difference between the two is that the dependent attribute or outcome (predicted) variable is Numerical for regression which has numeric continuous value and Categorical for classification which has results in the form of categories.

Regression analysis is a type of predictive modeling technique that analyses the relation between the target or dependent variable and independent variables in a dataset. It involves determining the best fitting line that passes through all the data points in such a way that distance of the line from each data point is minimal. For most accurate Predictions we are trying different Regression techniques on given problem statement to find out best fitting model. This includes linear regression, Support Vector Regressor and Decision Trees.

1. Linear Regression:

The main aim of Linear Regression model is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized. Error is defined as the difference between the actual value and Predicted value. The goal is to reduce this error or difference. Linear Regression is of two types based on number of independent variables: Simple and Multiple. Simple Linear Regression contains only one independent variable and the model has to find the linear relationship between this and the dependent variable. Whereas, Multiple Linear Regression contains more than one

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

independent variables for the model to find the relationship with the dependent variable.

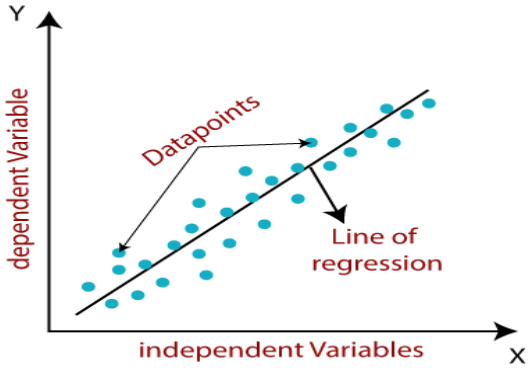

Fig 2. Linear Regression plot

Equation of Simple Linear Regression is,

$$y = b_o + b_1 x$$

Where b_0 is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.
Equation of Multiple Linear Regression is,

$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

Where b0 is the intercept, $b_1, b_2, b_3, b_4 \ldots, b_n$ are the coefficients or slopes of the independent variables $x_1, x_2, x_3, x_4 \ldots, x_n$ and y is the dependent variable.

2. Support Vector Regression:

Support Vector Regression (SVR) uses the same method as Support Vector Machine (SVM) but for regression problems. In SVR, the straight line that is required to fit the data is referred to as hyperplane. The objective of a SVR algorithm is to find a hyperplane in an n-dimensional space that classifies the data points. The data points on either side of the hyperplane that are closest to the hyperplane which are called Support Vectors.

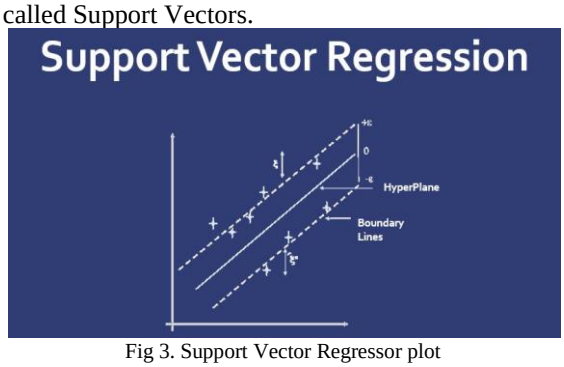

Fig 3. Support Vector Regressor plot

The best fit line is the hyperplane that has the maximum number of points. Unlike other Regression models that try to minimize the error between the real and predicted value, the SVR tries to fit the best line within a threshold value. The threshold value is the distance between the hyperplane and boundary line. The problem with SVR is that they are not suitable for large datasets.

3. Decision Tree Regression:

Decision Tree is a tree-structured algorithm with three types of nodes; Root Node, Interior node and Leaf node. The Interior Nodes represent the features of a data set and the branches represent the decision rules.

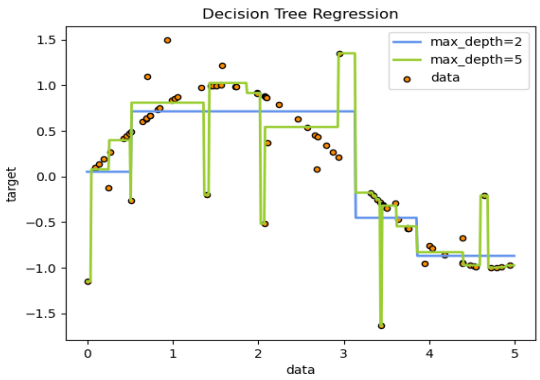

Fig 4. Decision Tree Regression

The Decision Tree Regressor observes features of an attribute and trains a model in the form of a tree to predict data in the future to produce meaningful output. Decision tree Regressor learns from the max depth, min depth of a graph and according to system analyzes the data.

C. Flask framework:

To deploy our Machine Learning model we need flask which is framework for deploying functional webpage for our created model. There are more options for that but flask is one of the effective and instant ways for creating UI for proposed Machine Learning model. It is also easier to integrate Flask with the model. Flask allows us to create a UI for our model. Flask provides us with tools, libraries and technologies that allow us to build a web application.

IV. IMPLEMENTATION:

The Flow of Implementation goes as follows:

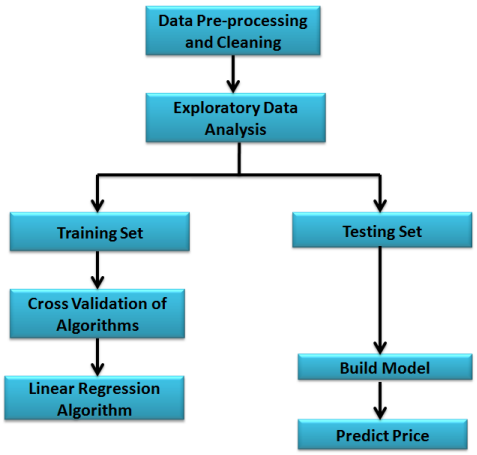

Fig 5. Flow of the Project

1. Data Preprocessing

After the manual collection of data through web scraping, there could be some mistakes in the collected entries, some null or blank values, human errors or some impractical values which we call as outliers. So to overcome these inaccuracies, we need to Pre-process and clean the data from these clutter values. There is a high need of Data Pre-processing because if the Data that we are providing to our model is accurate and faultless, then only the model will be able to give precise estimations which are very close the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

actual value. In Data Pre-processing and Cleaning, we remove the null values, take an overview of the dataset and also removal of unnecessary data columns (independent attributes) is done for the sake of accuracy and over fitting of the model. After cleaning the data looks like this:

	location	size	Baths	total_sqft	price	bhk
0	Bavdhan	2 bhk	2.0	640.0	55.96	2
1	Sus	2 bhk	2.0	705.0	45.00	2
2	Other	2 bhk	2.0	690.5	60.00	2
3	Koregaon Park	3 bhk	3.0	1659.0	210.00	3
4	Sinhgad Road	3 bhk	3.0	1450.0	78.00	3
5	Koregaon Park	2 bhk	2.0	1300.0	100.00	2
6	Bibwewadi	2 bhk	2.0	900.0	50.00	2
7	Koregaon Park	2 bhk	2.0	1080.0	95.00	2
8	Dhankawadi	3 bhk	3.0	900.0	87.50	3
9	Dhankawadi	2 bhk	2.0	700.0	67.50	2
10	Hinjewadi	2 bhk	2.0	680.5	59.43	2

Fig 6. Data after Pre-processing

2. Fitting the Model

After the data has been cleaned and free from the outliers, Feature Engineering and Exploratory Data Analysis have to be done.

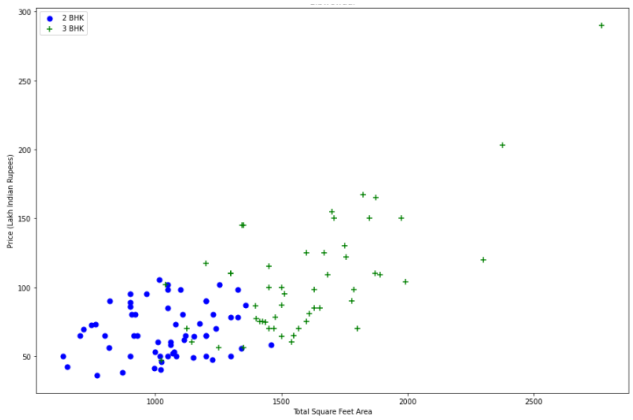

Fig 7. Plot: Total Square feet area v/s Price of the house

Then the data was split into training and testing sets for the classification of the best fitting machine learning model. The standard 80-20 split ratio is used, a typical ratio for this purpose; 80% of the data is considered as a training set and 20% as a testing set. To allow the implementation of the model, Scikit-Learn have to be imported. It is a Python Library which provides machine learning algorithms for implementation and many more features for modeling. We are performing Supervised Learning and to find out best model we will be implementing some regression algorithms which are likely to do a precise estimation of prices. The model which gives least error and most nearer value prediction will be our final model.

To test the results of Different models and compare them, we will provide same input values for all the models. Let's take an example of Area Bibwewadi in Pune and check price for 900 sqft. House having 2 bedrooms and 2 baths and Compare the Price given by different Algorithms.

First Algorithm we are using is Linear Regression. The Output is as follows:

```
In [164]: #Linear Regression
          predict_price('Bibwewadi', 900, 2, 2)

Out[164]: 70.42921553067562
```

Fig 8. Linear Regression Output Prediction

The Confusion Matrix of Accurate and Predicted value is shown in the figure below:

Actual	Predicted
156.00	179.586682
18.50	21.055223
190.00	112.623569
160.00	207.937531
59.95	58.745553

Fig 9. Linear Regression Confusion Matrix

Here it gives us the Price of approximately Rs.70.42 lakhs which is nearly similar to original value.
Second Algorithm we are using is Support Vector Regression. The Output is as follows:

```
In [165]: #svm
          pred_price('Bibwewadi', 900, 2, 2)

Out[165]: 61.49530758268064
```

Fig 10. Support Vector Regression Output Prediction

Actual	Predicted
156.00	55.054865
18.50	40.313997
190.00	87.660781
160.00	155.643540
59.95	59.978320

Fig 11. Support Vector Regression Confusion Matrix

Here it gives us the Price of approximately Rs.61.49 lakhs which is near to the original value.
Next Algorithm we are using is Decision Tree Regression. The Output is as follows:

```
In [167]: #Decsion tree
          tree_price('Bibwewadi', 900, 2, 2)

Out[167]: 26.0
```

Fig 12. Decision Tree Regression Output Prediction

Actual	Predicted
156.00	72.800000
18.50	35.100000
190.00	64.000000
160.00	240.000000
59.95	56.400000

Fig 13. Decision Tree Regression Confusion Matrix

Here it gives us the Price of approximately Rs.26 lakhs which is nowhere around the original value.
So from the above observations it is much clear that Linear Regression is giving most precise results and selected as predictive model for the House Price Prediction. The model

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

is ready to be used as an analytic tool for Real Estate Business Managers as well as buyers.

3. Deployment of model

Once the Implementation is done the model is predicting us the price of the property (house) in that particular location. We will deploy the model using Flask framework and create UI where the user will enter the desired values and our Model will predict the output. This is made Possible by using the python package for creating an API called Flask. For building the web application and linking the Model with the web application, first we need to extract our model into pickle and json files and design webpage using HTML, CSS and JavaScript. With this the Model is ready to be displayed and make predictions on the web application.

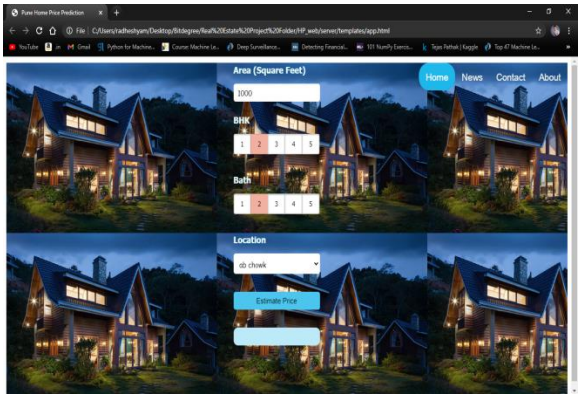

Fig 14. Deployment of Model using Flask

V. RESULTS AND DISCUSSIONS

Cross-validation of different Algorithms has proven to be a suitable method to find an acceptable best fitting algorithm for the Model. Linear Regression Algorithm is giving very precise Estimation of the house prices. For different Locations it is giving much accurate estimations. Also, according to confusion matrix linear regression is giving nearly accurate predictions. Linear Regression fits our dataset and gives the highest accuracy of 85.64%. Decision Tree gives the least accuracy of 56.02%. Support Vector Regression gives an accuracy of 62.81%.

```
In [303]: predict_price('Koregaon Park', 1500, 2, 3)
Out[303]: 124.99803191532536

In [304]: predict_price('Katraj', 850, 1, 1)
Out[304]: 70.81005225210848

In [305]: predict_price('Sinhgad Road', 700,2,2)
Out[305]: 58.199792341718556

In [307]: import pickle
          with open('home_prices_model.pickle','wb') as f:
              pickle.dump(lr_clf,f)

In [308]: import json
          columns = {
              'data_columns' : [col.lower() for col in X.columns]
          }
          with open("columns.json","w") as f:
              f.write(json.dumps(columns))
```

Fig 15. Linear Regression Output Predictions for Different locations

The Model has also proved that Location and square feet area plays an important role in deciding the price of a property. This is helpful information for Sellers and buyers to act accordingly. The GUI has provided Ease of access to the model, hence improving quality of accessibility.

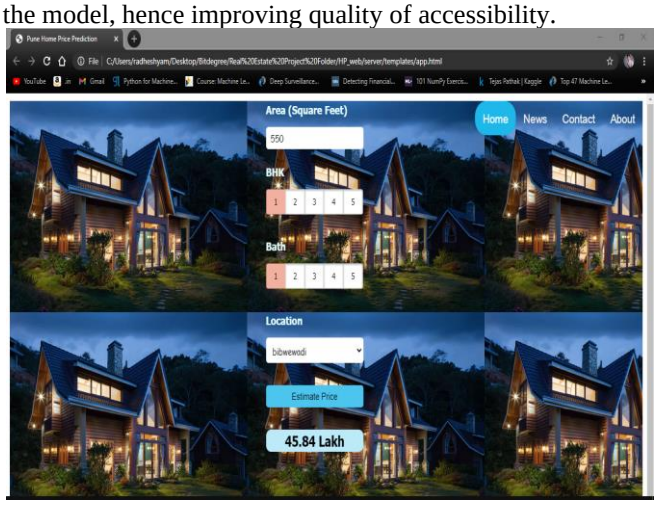

Fig 16. Final Output Prediction with UI

VI. FUTURE SCOPE

- In the future, the GUI can be made more attractive and interactive. It can also be turned into any real estate sale website where sellers can give the details and house for sale and buyers can contact according to the details given on the website.
- To simplify it for the user, there can also be a recommending system to recommend real estate properties to the user based on the predicted price. The current dataset only includes a few locations of Pune city, expanding it to other cities and states of India is the future goal.
- To make the system even more informative and user-friendly, Google maps can also be included. This will show the neighborhood amenities such as hospitals, schools surrounding a region of 1 km from the given location. This can also be included in making predictions since the presence of such factors increases the price of real estate property.

VII. CONCLUSION

In this research paper, we have used machine learning algorithms to predict the house prices. We have performed step by step procedure to analyze the dataset and found the correlation between the parameters. The manually collected Real-time Dataset has been collected which contains 1635 entries and independent variables. We analyze and pre-process this dataset before performing Exploratory Data Analysis. This analyzed feature set was given as an input to machine learning algorithms and calculated the performance of each model to compare based on Accuracy score. We found that Linear Regression fits our dataset and gives the highest accuracy of 85.64%. Decision Tree gives the least accuracy of 56.02%. Support Vector Regression gives an accuracy of 62.81%. Thus we conclude that we implemented regression techniques to check how well an algorithm fits to given problem statement of House price prediction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

VIII.REFERENCES

[1] Maharshi Modi, Ayush Sharma, Dr. P. Madhavan "Applied Research On House Price Prediction Using Diverse Machine Learning Techniques", International Journal of Scientific & Technology Research Volume 9, Issue 04, April 2020.

[2] G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu "House Price Prediction Using Machine Learning", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-9, July 2019.

[3] Dr. M. Thamarai, Dr. S P. Malarvizhi "House Price Prediction Modeling Using Machine Learning", I.J. Information Engineering and Electronic Business, 2020, 2, 15-20.

[4] Neelam Shinde, Kiran Gawande "Valuation Of House Prices Using Predictive Technique", International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-5, Issue-6, Jun.-2018.

[5] Ayush Varma, Abhijit Sarma, Sagar Doshi, Rohini Nair, Computer Engineering Department, KJ Somaiya College Of Engineering, Mumbai "House Price Prediction Using Machine Learning and Neural Networks", 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT).

[6] Hong Zhao, Rong-Qiu Chen, Wei Xu, Da-Ying Li "A SVR based forecasting approach for real estate price prediction", 2009 International Conference on Machine Learning and Cybernetics.

[7] Uysal, İ., Güvenir, H. A. "An overview of regression techniques for knowledge discovery", The Knowledge Engineering Review, Vol. 14:4, 1999, 319±340 (KER 14404) Printed in the United Kingdom.

[8] Bhuriya, Dinesh, et al. "Stock market predication using a linear regression.", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of. Vol. 2. IEEE 2017.

[9] Limsombunchai "House price prediction: hedonic price model vs. artificial neural network.", New Zealand Agricultural and Resource Economics Society Conference 2004.

[10] S. C. Bourassa, E. Cantoni, and M. Hoesli, "Predicting house prices with spatial dependence: a comparison of alternative methods," Journal of Real Estate Research, vol. 32, no. 2, pp. 139–160, 2010.

[11] Li, Li, and Kai-Hsuan Chu "Prediction of real estate price variation based on economic parameters", Applied System Innovation (ICASI), 2017 International Conference in IEEE, 2017.

[12] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python", Journal of machine learning research 12.Oct (2011): 2825-2830.