WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

House Price Prediction using Machine Learning

Komal Ghaytadkar[1], Smaruddhi Labhade[2], Leena Mahajan[3], Disha Patil[4], Prof. Swati Jadhav5

[1]Student, Department of Information Technology Sinhgad College of Engineering, Pune, Maharashtra, India
[2]Student, Department of Information Technology, Sinhgad College of Engineering, Pune Maharashtra, India
[3]Student, Department of Information Technology Sinhgad College of Engineering, Pune, Maharashtra, India
[4]Student, Department of Information Technology Sinhgad College of Engineering, Pune, Maharashtra, India
[5]Professor, Department of Information Technology Sinhgad College of Engineering, Pune, Maharashtra, India

ABSTRACT

On Bengaluru house price dataset, this paper demonstrates the use of machine learning algorithms in the prediction of real estate/house prices. This research will be really beneficial, to find the most important attributes to decide house values, especially for housing developers and academics and to recognise the most effective machine learning model for conducting research in this field. In the real estate sector, data mining is becoming widely used. The ability of data mining to retrieve useful information. It is highly useful to predict property values, essential housing features, and many other things utilising raw data information. Research has remarked that property price variations are frequently a source of anxiety for homeowners and the real estate sector. A review of the literature is conducted to determine the important criteria and the most effective models for forecasting house values. The results of this investigation confirmed the utilisation of linear regression. In comparison to other models, as the most efficient Furthermore, our data show that locational characteristics and House prices are heavily influenced by structural characteristics. The real estate market is one of the most price – sensitive and volatile. It is of the most important sector in which to apply machine learning concept. Learning how to improve and anticipate high cost accuracy. It will assist clients in putting resources into a bequest without resorting to a broker.

KEYWORDS: **House Price Prediction; Machine Learning; Deep Learning; Data Mining**

I. INTRODUCTION

As we know that house is one of the human life's most essential needs such as other primary needs like food, water and much more. Now a day's demands for houses grew rapidly over the years as people's living standards improve there are many people who make their house as an investment and property on the other hand some people around the world are buying a house as their shelter.

House price prediction can be done by using multiple various prediction models (Machine learning model) such as support vector regression, artificial neural network and many more. As an increase in house demand arise each year indirectly house price increases every year. The main problem come out when there are many variables such as property and location that may impact on house price that's why most stakeholders including house builders, buyers and developers and the real estate industry would like to know the exact features or the accurate factors manipulate the house price to help investors to make decision to help house builders set the house price. [1]

The primary aim of this paper is to use this machine learning techniques and curate them into ML model. Machine learning algorithms automatically build a mathematically model using sample data also refer to as training data which form decisions without being particularly programmed to make those decisions. Here's where machine learning comes in by training ML model with hundreds and thousands of data. A solution can be generated to predict prices accurately and provides to everyone's needs. [2]

Regression is a machine learning mechanism that motivates you to make expectations by taking in-from the current measurable assessable information-the connections between your goal parameter and many different independent parameters. According to this definition, a houses cost depends upon parameters, for example the number of rooms, living region, area, and so on. On the off chance that we apply forged figuring out how to this parameters, we calculate house valuations in a given land region. The target feature in this proposed model is the price of the real estate property and the independent features are: no. of bedrooms, no. of bathrooms, carpet area, built up area, the floor, zip code, age of the property, latitude and longitude of the property. Other than those of the mention features, which are generally required for predicting the house prise, we have covered two other features- air quality and prime rate. This feature provides a valuable contribution towards predicting property prices since the higher values of these features will lead to a reduction in the house prices. This regression model is built not only for predicting the price of the house which is ready for sale but also for houses that are under construction. [3]

II. LIMITATIONS OF EARLIER METHODOLOGIES

There is a notable amount of research done in the house price prediction department but very research has come up to any real-life solutions. There is very little evidence of a working house price predictor set up by a company. For now, very few digital solutions exist for such a huge market and most of the methods used by people and companies are as follows:

Buyers/Customers:

- When people first think of buying a house/Real estate they tend to go online and try to study trends and other related stuff. People do this so they can look for a house which contains everything they need. While doing these people make a note of the price which goes with these houses. However, the average person doesn't have detailed knowledge and accurate information about what the actual price should be. This can lead to misinformation as they believe the prices mentioned on the internet to be authentic.
- The second thing that comes to mind while searching for a property is to contact various Estate agents. The problem with this is these agents need to be paid a fraction of the amount just for searching a house and setting a price tag for you. In most cases, this price tag is blindly believed by people because they have no other options. There might be cases that the agents and sellers may have a secret dealing and the customer might be sold an overpriced house without his/her knowledge.

Seller/Agencies:

- When an individual think of selling his/her property they compare their property with hundreds and thousands of other properties which are posted all around the world. Determining the price by comparing it with multiple estates is highly time-consuming and has a potential risk of incorrect pricing.
- Large Real estate companies have various products they need to sell and they have to assign people to handle each of these products. This again bases the prediction of a price tag on a human hence there is room for human error. Additionally, these assigned individuals need to be paid. However, having a computer do this work for you by crunching the heavy numbers can save a lot of time money and provide accuracy which a human cannot achieve.

M Thamarai, S P Malarvizhi experimented with the most fundamental machine learning algorithms like decision tree classifier, decision tree regression, and multiple linear regression. Work is implemented using the Scikit-Learn machine learning tool. This work helps the users to predict the availability of houses in the city and also to predict the prices of the houses. [4]

B.Balakumar, P.Raviraj, S.Essakkiammal used machine learning algorithms to predict house prices. We have mentioned the step-by-step procedure to analyse the dataset. These feature sets were then given as an input to four algorithms and a CSV file was generated consisting of predicted house prices. [5]

Akshay Babu, Dr. Anjana S Chandran expressed that There is a need to use a mix of these models a linear model gives a high bias (underfit) whereas a high model complexity-based model gives a high variance (overfit). The outcome of this study can be used in the annual revision of the guideline value of land which may add more revenue to the State Government while this transaction is made. [6]

III. PROPOSED WORK

The purpose of this system is to determine the price of a house by looking at the various features which are given as input by the user. These features are given to the ML model and based on how these features affect the label it gives out a prediction. This will be done by first searching for an appropriate dataset that suits the needs of the developer as well as the user. Furthermore, after finalizing the dataset, the dataset will go through the process known as data cleaning where all the data which is not needed will be eliminated and the raw data will be turned into a .csv file. Moreover, the data will go through data pre-processing where missing data will be handled and if needed label encoding will be done. Moreover, this will go through data transformation where it will be converted into a NumPy array so that it can finally be sent for training the model. While training various machine learning algorithms will be used to train the model their error rate will be extracted and consequently an algorithm and model will be finalized which can yield accurate predictions. Users and companies will be able to log in and then fill a form about various attributes about their property that they want to predict the price of. Additionally, after a thorough selection of attributes, the form will be submitted. This data entered by the user will then go to the model and within seconds the user will be able to view the predicted price of the property that they put in.

IV. FLOW CHART

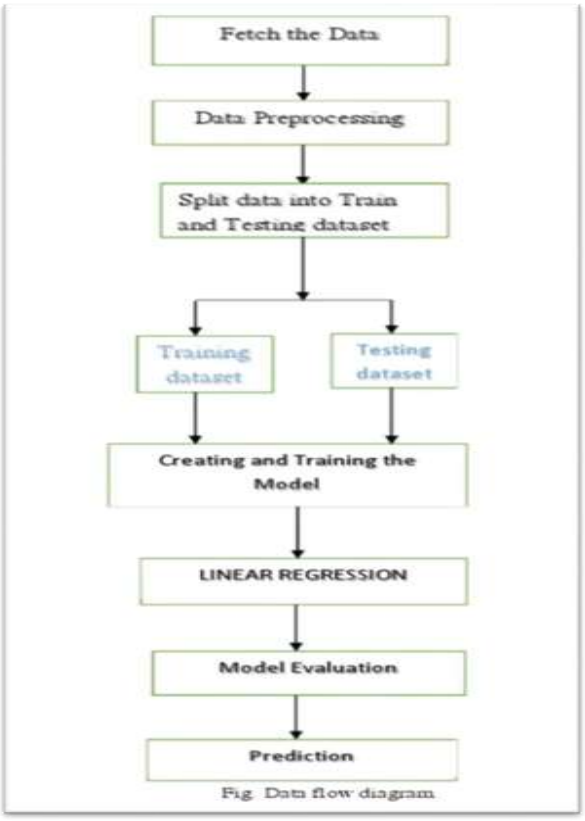
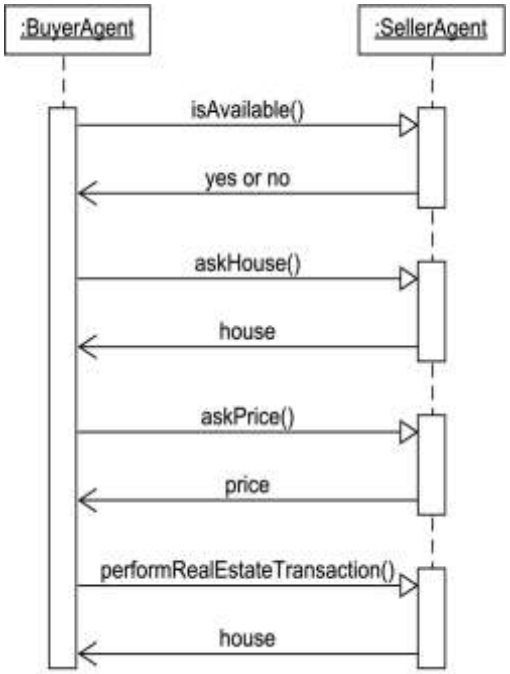


Fig. Data flow diagram

V. SEQUENCE DIAGRAM



VI. METHODOLOGY

We implemented a work for predicting house price for the buyers who are interested to buy house and for those peoples who are watching home in their affordable budget. Our main focus was providing comfort home in low price. So, we worked on that. For that, we had done analysis on the particular dataset which is Bengaluru house dataset. So, in that dataset we focus on those features on which buyers are interested to find out their fit one. So, the features are mainly **location, areas (in sqft), BHK, bathroom.**

To take response from users we create UI (user interface). In which we used some important tools like **HTML, CSS, Java Script** etc. Also we used local server by using **Nginix.** And using **Postman** application to check the **APIs**.

While working on back-end we use Anaconda software, in which we used notebook to apply our machine learning algorithm. Also we had used some advanced libraries of python, which works in machine learning that's are **NumPy, Pandas, Matplotlib, Seaborn** etc. By using this libraries, we **load** the data, **clean** the data, **analyse** the data to **remove outliers** and **visualize** the data.

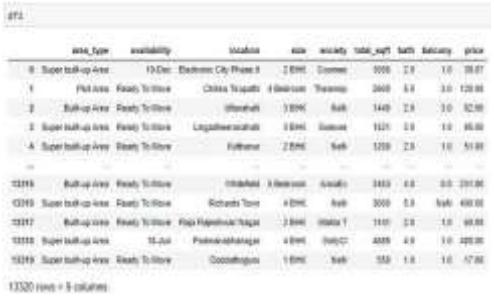


Fig. Bengaluru house price dataset.

```
df1.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13320 entries, 0 to 13319
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   area_type     13320 non-null  object
 1   availability  13320 non-null  object
 2   location      13319 non-null  object
 3   size          13304 non-null  object
 4   society       7818 non-null   object
 5   total_sqft    13320 non-null  object
 6   bath          13247 non-null  float64
 7   balcony       12711 non-null  float64
 8   price         13320 non-null  float64
dtypes: float64(3), object(6)
memory usage: 936.7+ KB
```

Fig. Attributes in the dataset.

```
df2=df1.drop(["area_type","society","balcony","availability"],axis="columns")
df2
```

	location	size	total_sqft	bath	price
0	Electronic City Phase II	2 BHK	1056	2.0	39.07
1	Chikka Tirupathi	4 Bedroom	2600	5.0	120.00
2	Uttarahalli	3 BHK	1440	2.0	62.00
3	Lingadheeranahalli	3 BHK	1521	3.0	95.00
4	Kothanur	2 BHK	1200	2.0	51.00
...	...	...	...	...	...
13315	Whitefield	5 Bedroom	3453	4.0	231.00
13316	Richards Town	4 BHK	3600	5.0	400.00
13317	Raja Rajeshwari Nagar	2 BHK	1141	2.0	60.00
13318	Padmanabhanagar	4 BHK	4689	4.0	488.00
13319	Doddathoguru	1 BHK	550	1.0	17.00

13320 rows × 5 columns

Fig. Drop unused features.

```
df2.isnull().sum()

location       1
size          16
total_sqft     0
bath          73
price          0
dtype: int64
```

Fig. Drop null values.

In our project, we used **linear regression algorithm**. Because, in this project we have to predict house price and house price is continuous data. So, for continuous data we used regression algorithm. While doing grid search cv for selecting the best fit algorithm, we got the highest accuracy score in linear regression. So, we used linear regression algorithm.

In linear regression algorithm, we used the concept of **coefficient** and **bias**. Because, linear means in line. And equation of line comes with two concepts, i.e. **slope** and **intercept**.

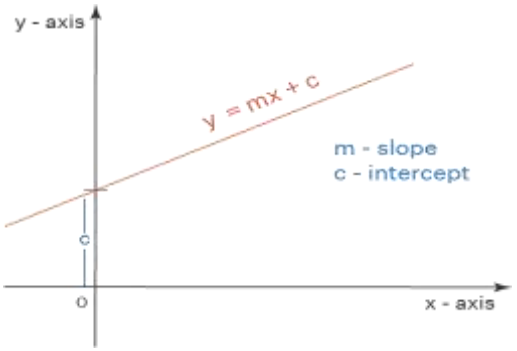


Fig. Equation of line.

By using this equation, we can draw the line which going through all the data points. But this line having the slope which based on the concept of **residual=0**. So for that, our linear regression algorithm works to find such slope in which we get residual=0. Residual means summation of error and our algorithm try to make residual=0. The difference between an observed value of the response variable and the value of the response variable predicted from the regression line called the residual.

The line which we find will get when all data points make residual=0. that means a line going through all data point is such type of line which based on the actual data points. Because, some predicted data points are above of the actual data points and some of the predicted data points are below of the actual data points and the actual data points and it's predicted points are always **parallel to Y-axis**. Which predicted data points are above of the actual data points, those distance are considered as a **(+) positive** distance and which predicted points are below of the actual data points those distances are considered as a **(-) negative** distance. When the total sum of these distances is become **zero (0).** At that time, we get **residual=0** and the **line** which have **coefficient** is very useful. Because, by using this **coefficient** we can **predict house price** according to the client's choice.

The **linear regression algorithm** always tries to make **residual=0**. This is the **methodology** which we study from this algorithm. And by using this we can easily predict the remaining values which user wants to find. This is our study according to linear regression algorithm.

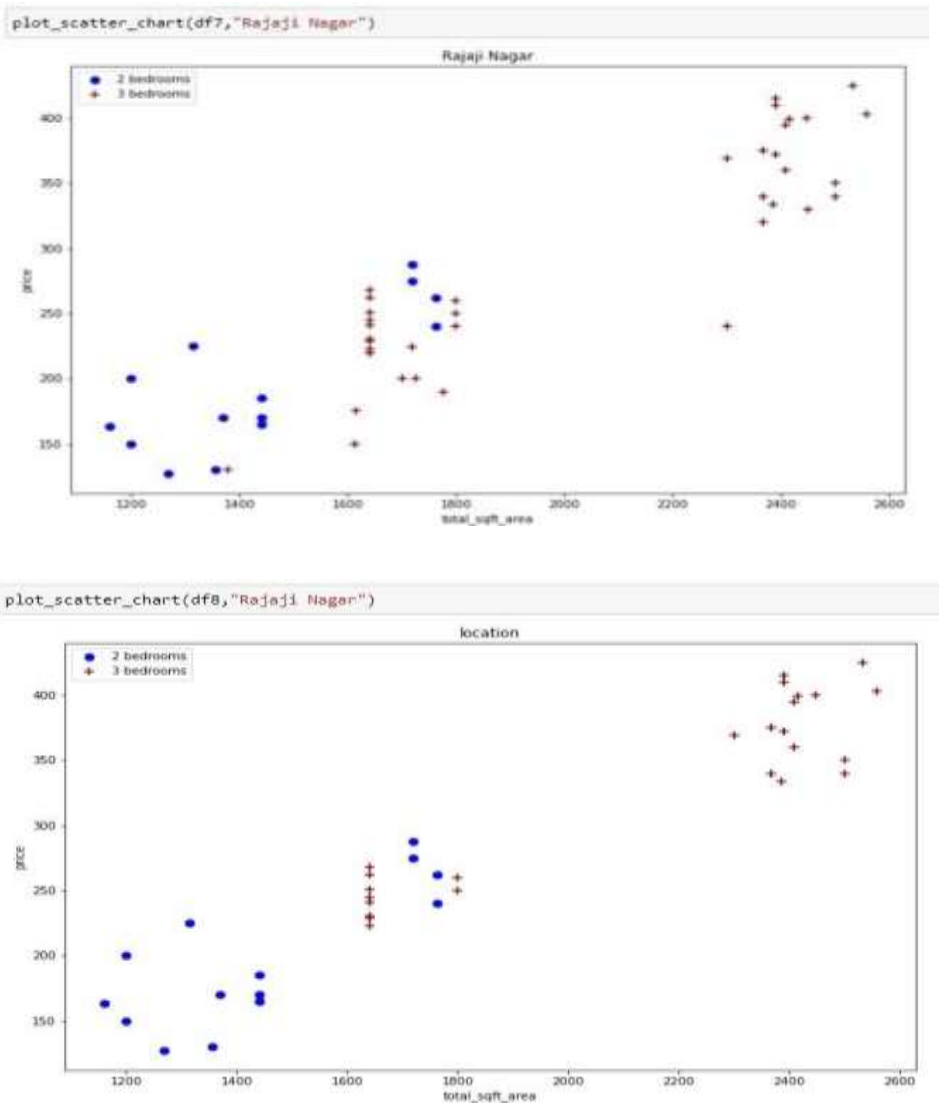


Fig. Remove outliers which having higher price of 2bhk than 3bhk.

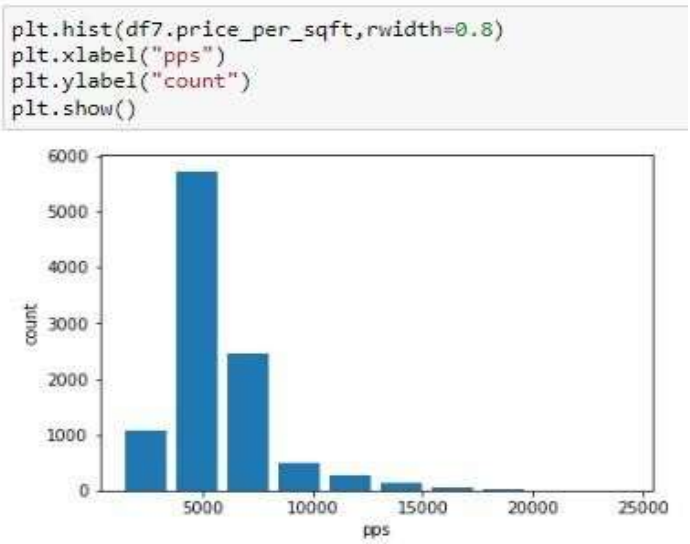


Fig. Check the price per sq.ft. to remove outliers.

```
X=df13[["total_sqft","bhk","bath"]]
y=df13["price"]
```

```
lr.fit(X,y)
```

```
LinearRegression()
```

```
lr.score(X,y)
```

```
0.7007730682300131
```

Fig. Model fitting without location.

```
y_predict=lr.predict(X)
y_predict
```

```
array([230.64210877, 112.13373285, 131.32345102, ...,  86.4015307 ,
        34.69593526, 309.34453578])
```

```
y.values
```

```
array([428., 194., 235., ..., 110.,  26., 400.])
```

```
plt.scatter("total_sqft","price",data=df13)
plt.scatter(df13.total_sqft,y_predict,c="red")
plt.show()
```

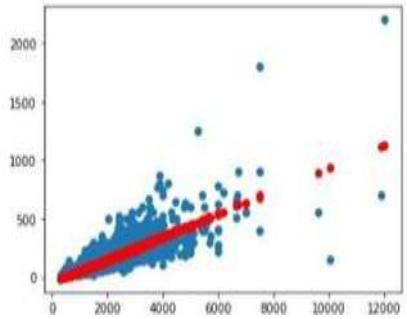


Fig. Predicted values vs actual values.

```
X=df15.drop("price",axis="columns")
y=df15["price"]
```

```
lr.fit(X,y)
```

```
LinearRegression()
```

```
lr.score(X,y)
```

```
0.8544273054686995
```

Fig. Model fitting with location.

```
y_predict=lr.predict(X)
y_predict
```

```
array([230.64210877, 112.13373285, 131.32345102, ...,  86.4015307 ,
        34.69593526, 309.34453578])
```

```
y.values
```

```
array([428., 194., 235., ..., 110.,  26., 400.])
```

```
plt.scatter("total_sqft","price",data=df13)
plt.scatter(df13.total_sqft,y_predict,c="red")
plt.show()
```

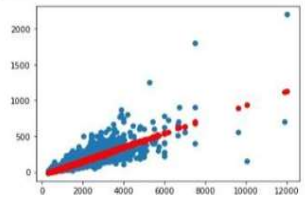


Fig. Predicted values VS actual values.

```
from sklearn.model_selection import ShuffleSplit

from sklearn.model_selection import cross_val_score

cv=ShuffleSplit(n_splits=7,test_size=0.30,random_state=0)

cross_val_score(LinearRegression(),X,y,cv=cv)
```

```
array([0.79970448, 0.82497628, 0.84492394, 0.86266323, 0.87080754,
       0.86873092, 0.82739684])
```

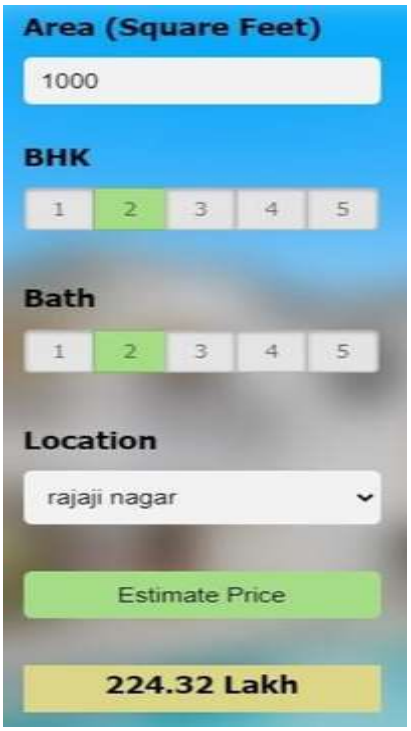
Fig. Shuffle split & CV.

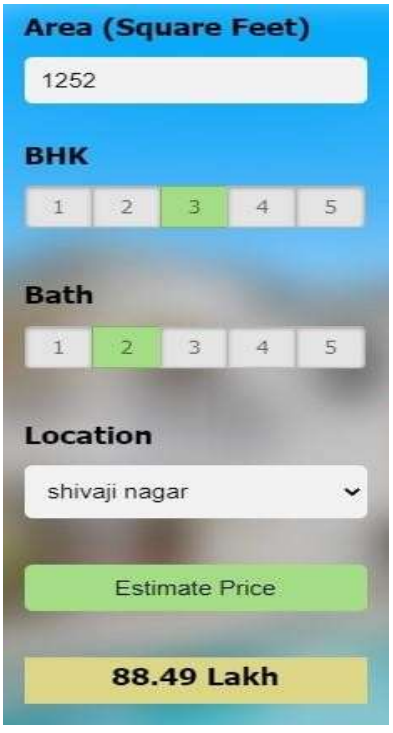
```
LinearRegression()
Lasso()
DecisionTreeRegressor()
```

	model	best_score	best_params
0	linear_regression	0.818354	{'normalize': False}
1	lasso	0.687478	{'alpha': 2, 'selection': 'random'}
2	decision_tree	0.733502	{'criterion': 'mse', 'splitter': 'random'}

Fig. Various algorithms score using grid search CV.

VII. RESULT





VIII. CONCLUSION

The paper entitled “House Price Prediction Using Machine Learning” has presented to predict house price based on various features on given data. We'll predict a variable from an independent one using linear regression, thus we like to understand from the start anytime we add information. The regression curve is important because it improves the accuracy of variable estimate and allows the estimation of a response variable for people whose carrier variable values are not included in the data. We also deduced that there are two ways to forecast a variable: from within the range of values of the experimental variable in the sample or from beyond the range. The house price and linear regression is most important effective model for our dataset. In conclusion, the impact of this research was intended to help and assist other researchers in developing a real model which can easily and accurately predict house prices. Further work on a real model needs to be done with the utilization of our findings to confirm them. It helps people to buy house in budget and reduce loss of money.

REFERENCES

[1]. Hamizah Zulkifley, Shuzlina Abdul Rahman, Hasbiah Ubaidullah, Ismail Ibrahim, House Price Prediction using a Machine Learning Model: A Survey of Literature I.J. Modern Education and Computer Science, 2020, 6, 46- 54, Published Online December 2020 in MECS, DOI: 10.5815/ijmecs.2020.06.04

[2]. Smith Dabreo, Shaleel Rodrigues, Valiant Rodrigues, Parshvi Shah, Real Estate Price Prediction, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278- 0181, Vol. 10 Issue 04, April-2021

[3]. ALISHA KUVALEKAR, SHIVANI MANCHEWAR, SIDHIK MAHADIK, SHILA JAWALE, House Price Forecasting Using Machine Learning, Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST) 2020

[4]. Byeonghwa Park, Jae Kwon Bae “Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data”. 2017

[5]. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh - “A hybrid regression technique for house prices prediction” 2017, IEEE.

[6]. 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

[7]. Sayan Putatunda, "PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market".

[8]. Thuraiya Mohd, Suraya Masrom, Noraini Johari, "Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia ", International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-2S11, 2019

[9]. M Thamarai, S P Malarvizhi, " House Price Prediction Modeling Using Machine Learning", International Journal of Information Engineering and Electronic Business(DJIEEB), VoL12, No.2, pp. 15- 20, 2020. DOI: 10.5815/ijieeb.2020.02.03

[10]. Akshay Babu, Dr. Anjana S Chandran, "Literature Review on Real Estate Value Prediction Using Machine Learning”, International Journal of Computer Science and Mobile Applications, Vol: 7 Issue: 3, 2019