Housing Price Prediction Using Machine Learning

Monika Sahu[1], Mrs. Awantika Singh[2], Mr. Rahul Chawda [3] [1]*M.techScholar,* [2]*Assistant Professor,* [3]*Assistant Professor*

Computer Science & Engineering Department, Kalinga University, Raipur (C.G.), India

Abstract

House costs increase reliably, so there is a requirement for a system to expect house costs later on. House cost assumption can help the architect with choosing the selling cost of a house and can help the customer with organizing the ideal opportunity to purchase a house. Three factors sway the expense of a house which incorporates state of being, thought and region. .House price forecasting is a crucial topic of land. Machine learning techniques are applied to research historical property transactions in World to get useful models for house buyers and sellers. In the current, paper we examine all the expectation of future lodging costs that are produced by AI calculation. For the choice of forecast strategies we look at and investigate different expectation techniques. For the decision of expectation strategies we analyze and investigate different forecast techniques. We use regression as our model because of its versatility and probabilistic. Our outcome display that our methodology of the trouble had the chance to make progress, and can handle expectations that may be similar to other house cost forecast models. This paper intends to help the seller to appraise the selling cost of a house impeccably and to help individuals to anticipate the exact time slap to collect a house. We utilize Linear Regression strategies during this pathway, and our outcome isn't a sole assurance of one procedure rather it's the weighted mean of shifted methods to offer the most exact outcomes. The outcomes demonstrated that this methodology yields the least blunder and greatest precision than singular calculation applied.

Keywords: House price prediction, Machine Learning, AI, Model, Linear Regression, Algorithm .

1. Introduction:

A huge number of homes are sold ordinary. Furthermore, houses cost expands each year. So there is a requirement for a framework to anticipate house costs later on. During this task, an AI model is proposed to anticipate a house cost upheld information related to the house (its size, the year it had been inbuilt, and so on) Expectation house costs are relied upon to help individuals that choose to purchase a house altogether that they can realize the value home in the more drawn out term, at that point they will design their money well. Furthermore, house value expectations likewise are advantageous for property financial backers to comprehend the pattern of lodging costs during a specific area. A House value forecast can assist the designer with deciding the requesting cost from a house and may assist the client with revamping the appropriate opportunity to get a house. There are three factors that impact the value of a house which incorporate states of being, idea and site. AI might be a subfield of AI that works with calculations and advancements to extricate helpful data from information. AI strategies are fitting in huge information since endeavouring to physically handle huge volumes of information would be inconceivable without the help of machines. AI in figuring endeavours to unwind issues algorithmically rather than simply numerically. Subsequently, it's upheld making calculations that let the machine discover. There are two sorts in AI which are managed and solo. Managed is the place where the program gets prepared on a pre-decided set to be prepared to foresee when a piece of substitution information is given. Solo is the place where the program attempts to search out the association and accordingly the secret example between the data Several Machine Learning calculations are wont to take care of issues inside the present reality. Subsequently, this paper endeavours to utilize Linear Regression calculations to coordinate with their exhibition when it includes foreseeing the upsides of a given dataset.

2. Literature Survey

All through the latest twenty years, there have been a gigantic number of careful assessments analyzing land costs. Kilpatrick showed the accommodation of time-course of action backslide model which used monetary data to give a figure of Central Business District (CBD) land cost in moving business area. Wilson et al thought about the private property market address a huge degree of UK money related development. Valuers check property assessments subject to current bid costs. In this paper, the public housing trade data was arranged using machine learning, which gauges future example of the housing market. Engraving and John cultivated a backslide model with void land bargains. The model uncovered up to 93% of the market regards. Wang and Tian used the wavelet Neural Network (NN) to guess the land esteem document. This kind of wavelet NN fused the worth of the wavelet assessment and the custom NN. It furthermore differentiated the deciding outcome and smoothing system and the NN figure. Zhangming assessed the land esteem list by using the Back Propagation (BP) NN. The BPN used the sigmoid limit. Tinghao used the Auto-Regressive Integrated Moving Average (ARIMA) model and passed on the illustrative examination on year data from 1998 to 2006. He used the set-up model to make the check to the land esteem record of 2007. A liberal backslides on the expense of land suggested that genuine course of action contrasts between political domains have fundamentally influenced land costs between1970 and 1980. Steven and Albert used 46,467 private properties crossing 1999 - 2005 and displayed that using composed with sets that near with straight profligate assessing models, ANN produces lower dollar assessing botches, had more essential esteeming precision out-of-test, and extrapolate better from more flighty assessing conditions. ANN is more able to liberal models that utilization gigantic amounts of components. Sampath Kumar and Santhi inspected the land esteem example of Sowcarpet which is the central part. They made a quantifiable model using monetary factors and expected that the yearly rising inland cost would be 17%. Urmila uncovered that the past designs were analyzed to decide the speed of improvement or decline and the examples are used in deciding. Financial limits might be familiar with plan more reasonable relationship. A segment of various systems they Mansural Bhuiyan and Mohammad Al Hasan 2016 use is backslid, significant sorting out some way to take in models from the previous results (the property/land which were sells in advance which are used as getting ready data). There are different models used, for instance, straight model data using only one component, multivariate model, using a couple of features as its data and polynomial model using the data as cubed or squared and consequently decided the root mean squared slip-up (RMS a motivating force) for the model.

3. Methodology:

Dataset: The dataset utilized during this work is obtained from land Agents within the US to guage the prices of homes. This dataset was intended to help them find the foremost appropriate cost of the house within the in respective locations within the US. For predicting the prices of the homes, the following attributes are identified and included, in conjunction with the acronym used for his or her representation within the dataset snippet as shown. Base Price (price), Size of Plot, House age, Number of Bedrooms, Number of rooms, Area population, Address, of the house. These attributes, or predictor variables, are the factors which are majorly considered during house purchase and thus influence the pricing of homes .The dataset utilized in this work is obtained from land Agents within the US to judge the costs of homes.

The data-set we have used is a group of houses of USA .The size of the dataset is of 5,000 houses which are divided into training data and testing data.The dataset contains 7 columns and 5000 rows with CSV extension. The data contains the following columns :

‘Avg. Area Income’ – Avg. The income of the householder of the city house is located.
‘Avg. Area House Age’ – Avg. Age of Houses in the same city.
‘Avg. Area Number of Rooms’ – Avg. Number of Rooms for Houses in the same city.
‘Avg. Area Number of Bedrooms’ – Avg. Number of Bedrooms for Houses in the same city.
‘Area Population’ – Population of the city.
‘Price’ – Price that the house sold at.
‘Address’ – Address of the houses.

Fetch the data set by the help of the pandas in python platform and analyze the data set.

Data preprocessing is a process to convert raw data into meaningful data using different techniques. Data in real word are incomplete, noisy, duplicate or inconsistent, so we can convert this data into accuracy, completeness, consistency by various data preprocessing techniques.

Major steps in data preprocessing:

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation
- Data Discretization

Train Test Split Evaluation It is a technique for evaluating the performance of a machine learning algorithm. We have segregated dependent and independent variable in two different vectors respectively x and y, we handled our by data preprocessing. We have 500 records inside of our Data set and we split the data into training set and test set. Vector y is dependent or target variables. Training part of data is there that will pass to our machine learning model for the training purpose and one machine learning model is trained, the rest remaining data that is test data will use to validate machine learning model. We split a data in percentage between 0 and 1 for either the train or test datasets. Here we take test set with the size of 0.40 (40 %) means that the remainder percentage 0.60 (60 %) is assigned to the training set.

Linear Regression:

Linear regression is a statistical model that attempts to point out the connection between two variables with the equation. It is the one of the easiest algorithm in machine learning. It is calculated by using formula,

$$\mathbf{Y = MX + B}$$

Y = Dependent variable

M = Coefficient rate and slope of line

X = Independent variable

B = Where line crosses the y-axis

From the above formula we try to find the value of x and y that every value of x has a corresponding value of y in it if it is continuous. The reason for this is linear regression is always continuous. The output of the linear regression is the value of the variable. The accuracy or the great fit is calculated by using the r squared method.

Why we select linear Regression

Regression model predicts a continuous variables and linear Regression predicted value is continuous and it also remove the outliers there for our model will perform in a better way. Linear regression are easily comprehensive and transparent. They can be understood very easily because it is represented by simple mathematical notation.

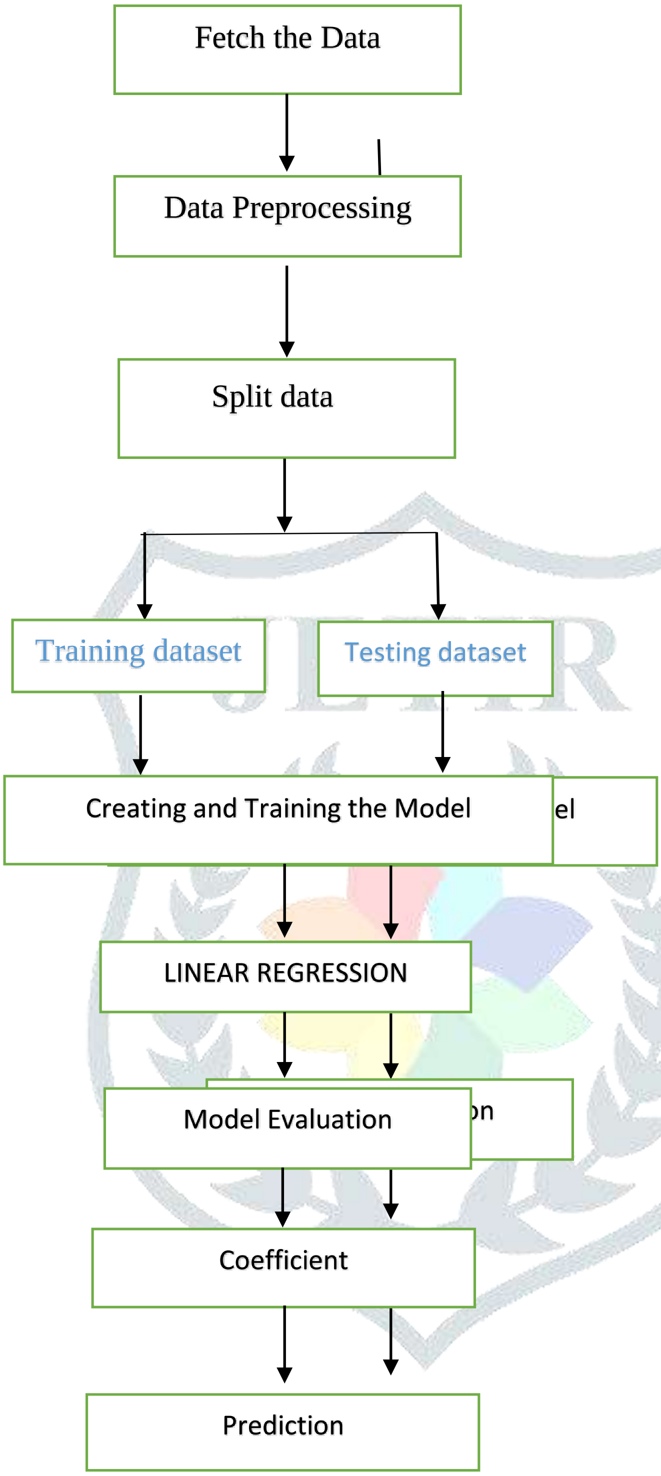
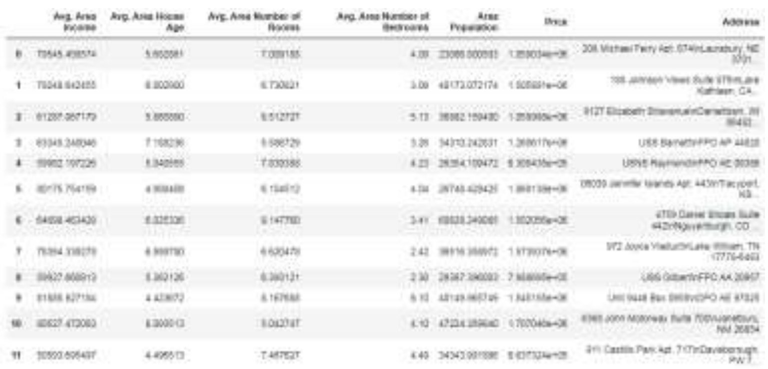


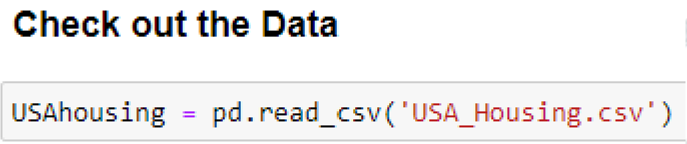
Fig. Data Flow Diagram

4. Implementation

Data set:

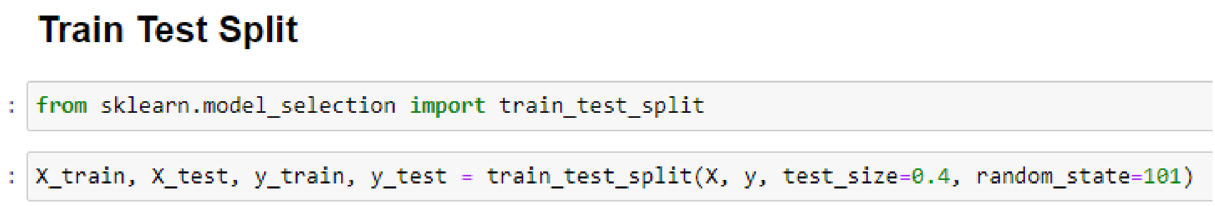


	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386
5	80175.754159	4.988408	6.104512	4.04	26748.428425	1.068138e+06	06039 Jennifer Islands Apt. 443\nTracyport, KS...
6	64698.463428	6.025336	8.147760	3.41	60828.249085	1.502056e+06	4759 Daniel Shoals Suite 442\nNguyenburgh, CO...
7	78394.339278	6.989780	6.620478	2.42	36516.358972	1.573937e+06	972 Joyce Viaduct\nLake William, TN 17778-6483
8	59927.660813	5.362126	6.393121	2.30	29387.396003	7.988695e+05	USS Gilbert\nFPO AA 20957
9	81885.927184	4.423672	8.167688	6.10	40149.965749	1.545155e+06	Unit 9446 Box 0958\nDPO AE 97025
10	80527.472083	8.093513	5.042747	4.10	47224.359840	1.707046e+06	6368 John Motorway Suite 700\nJanetbury, NM 26854
11	50593.695497	4.496513	7.467627	4.49	34343.991885	6.637324e+05	911 Castillo Park Apt. 717\nDavisborough, PW 7...



Check out the Data

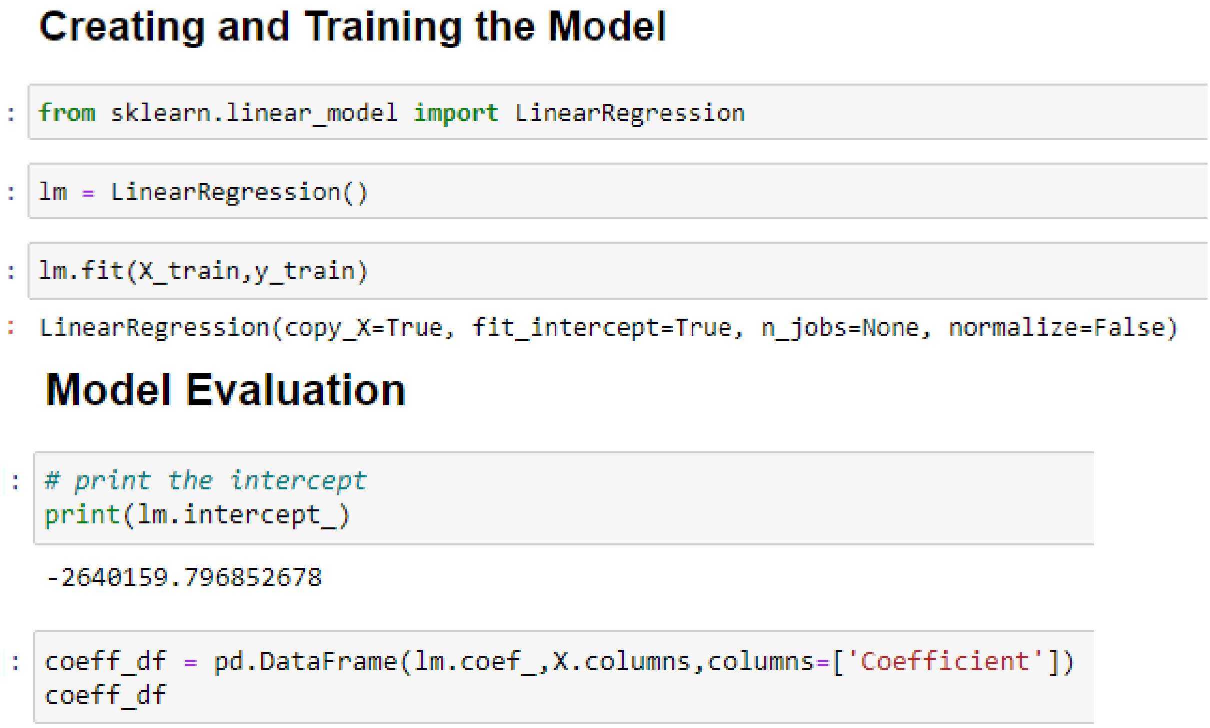
```
USAhousing = pd.read_csv('USA_Housing.csv')
```



Train Test Split

```
: from sklearn.model_selection import train_test_split
```

```
: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=101)
```



Creating and Training the Model

```
: from sklearn.linear_model import LinearRegression
```

```
: lm = LinearRegression()
```

```
: lm.fit(X_train,y_train)
```

```
: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

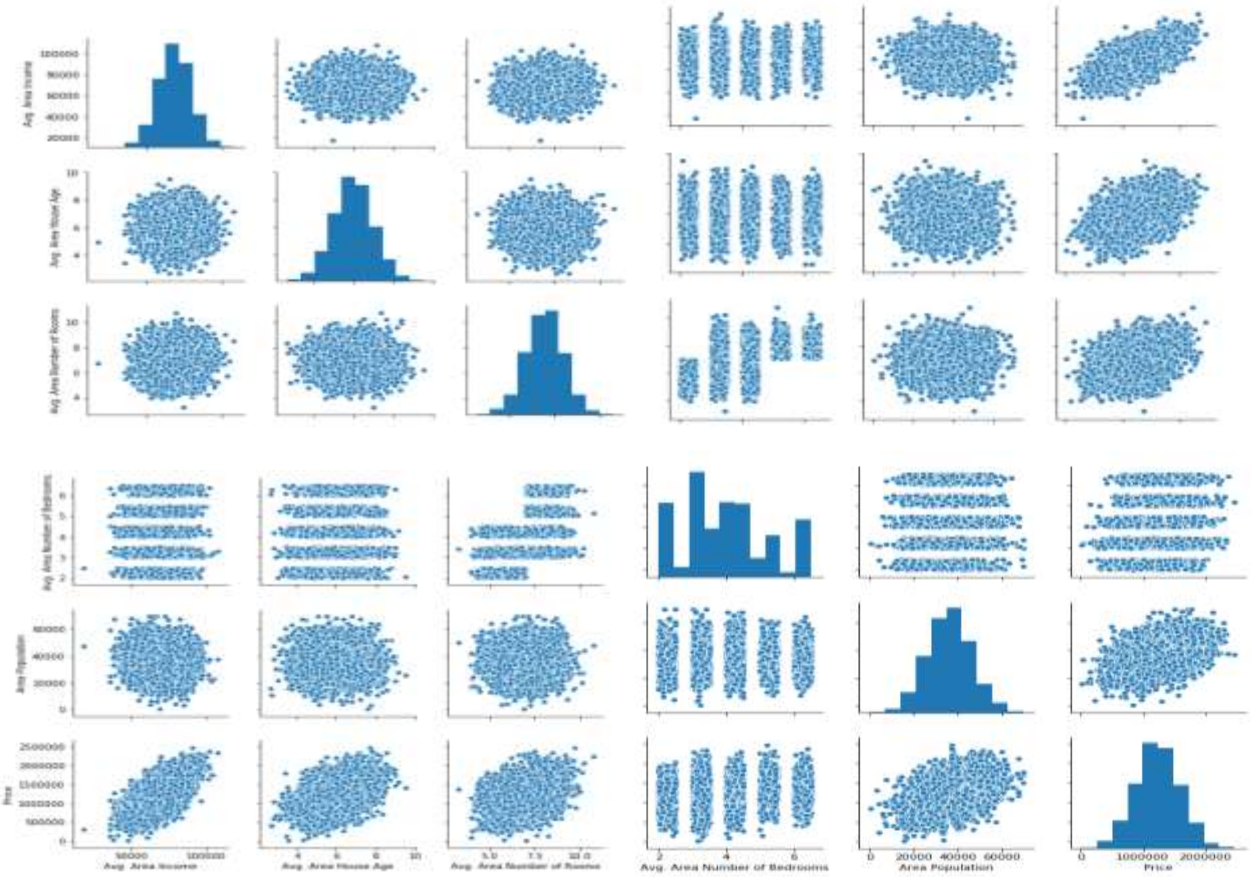
Model Evaluation

```
: # print the intercept
  print(lm.intercept_)
```

```
-2640159.796852678
```

```
: coeff_df = pd.DataFrame(lm.coef_,X.columns,columns=['Coefficient'])
  coeff_df
```

5. Execution and Outputs:



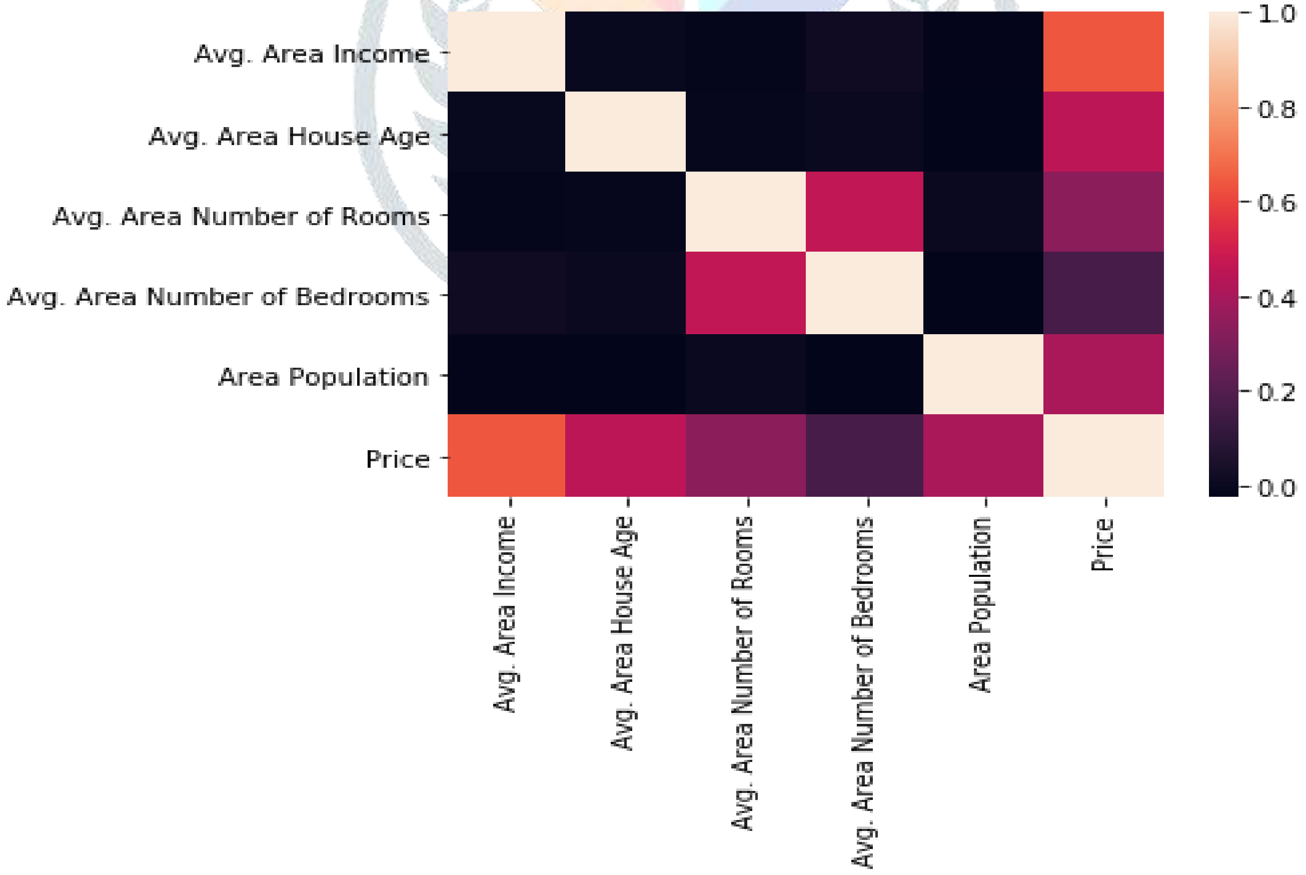


Fig. Heatmap (Represent correlation between them)

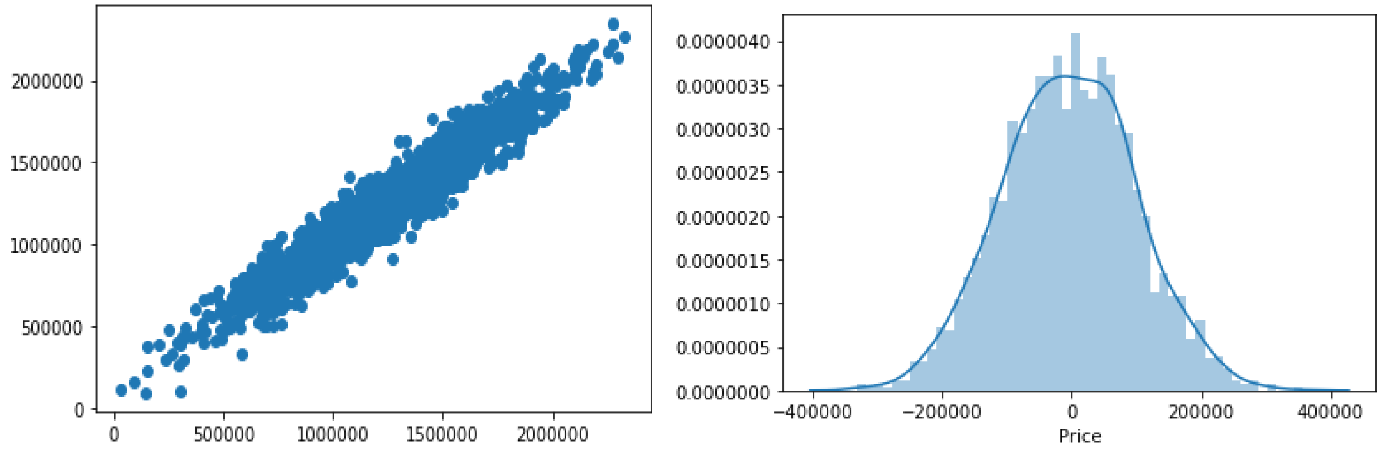


Fig. Scatter plot

6. Conclusion:

Linear Regression implies we will foresee a variable from an autonomous one, so at whatever point we'd prefer to comprehend from the beginning at whatever point we add data. The relapse bend is indispensable on the grounds that it makes the assessment of a variable more precise and it permits the assessment of a reaction variable for individuals with upsides of the transporter variable excluded inside the information. We additionally gathered there are two strategies for foreseeing a variable either from inside the scope of upsides of an exploratory variable of the example given (interjection) or outside this reach (extrapolation). The house cost and the linear regression is the best model for our dataset.

REFERENCES

1. Akash Dagar and Shreya Kapoor, “A Comparative Study on House Price Prediction”, *International Journal for Modern Trends in Science and Technology*, 6(12): 103-107, 2020.

2. Puneet Tiwari[1], Varun Singh Thakur ‘Review on house price prediction through Regression technique’ International Journal of Scientific Progress and Research (IJSPR), Issue 173, volume 73.

3. R Manjula, Shubham Jain, Sharad Srivastava and Pranav Rajiv Kher ‘Real estate value prediction using multivariate regression models’ *et al* 2017 *IOP Conf. Ser.: Mater. Sci. Eng.* **263** 042098

4. Housing Price Prediction using Machine Learning Yashraj Garud , Hemanshu Vispute , Nayan Bisai and Prof. Madhu Nashipudimath4 Volume: 07 Issue: 05 | May 2020 (IRJET)

5.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.

6. Darshil Shah, Harshad Rajput, Jay Chheda “house price prediction using machine learning and RPA”International Research Journal of Engineering and Technology (IRJET)
Volume: 07 Issue: 03 | Mar 2020

7. “Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization “(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017

8. Eli Beracha, Ben T Gilbert, Tyler Kjorstad, Kiplan womack, "On the Relation between Local Amenities and House Price Dynamics", Journal of Real estate Economics,Aug. 2016.

9. T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, “How many trees in a random forest?” In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in

Bioinformatics), vol. 7376 LNAI, 2012, pp. 154–168,ISBN: 9783642315367. DOI: 10 . 1007 / 978 - 3 - 642 -31537-4\ 13.

10. S. Ray, “CatBoost: A machine learning library to handle categorical (CAT) data automatically,” CatBoost: Analytics Vidhya, 14-Aug-2017.

11. R. J. Shiller, “Understanding recent trends in house prices and home ownership,” National Bureau of Economic Research, Working Paper 13553, Oct. 2007. DOI: 10.3386/w13553.

12. 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.

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

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

15. Byeonghwa Park , Jae Kwon Bae (2015). Using machine learning algorithms for housing price prediction , Volume 42, Pages 2928-2934.

16. Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, 2015. Introduction to Linear Regression Analysis.

17. A. Azadeh, B. Ziaei, and M. Moghaddam, ―A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations,‖ Expert Syst. Appl., vol. 39, no. 1, pp. 298–315, 2012.

18. F. S. Gharehchopogh, T. H. Bonab, and S. R. Khaze, ―A Linear Regression Approach to Prediction of Stock Market Trading Volume: A Case Study,‖ Int. J. Manag. Value Supply Chain., vol. 4, no. 3, pp. 25–31, 2013.

19. H.-I. Hsieh, T.-P. Lee, and T.-S. Lee, ―A Hybrid Particle Swarm Optimization and Support Vector Regression Model for Financial Time Series Forecasting,‖ Int. J. Bus. Adm., vol. 2, no. 2, pp. 48–56, 2011.

20. F. Marini and B. Walczak, ―Particle swarm optimization (PSO). A tutorial,‖ Chemom. Intell. Lab. Syst., vol. 149, pp. 153–165, 2015.

21. A. Hayder M. Albehadili Abdurrahman and N. . Islam, ―An Algorith for Time Series Prediction Using,‖ Int. J. Sci. Knowl. Comput. Inf. Technol., vol. 4, no. 6, pp. 26–33, 2014.

22. Y. P. Anggodo and W. F. Mahmudy, ―Automatic Clustering and Optimized Fuzzy Logical Relationship for Minimum Living Needs Forecasting,‖ J. Environ. Eng. Sustain. Technol., vol. 4, no. 1, pp. 1–7, 2017.

23.Y. P. Anggodo, W. Cahyaningrum, A. N. Fauziyah, I. L. Khoiriyah, K. Oktavianis, and I. Cholissodin, ―Hybrid K-means Dan Particle Swarm Optimization Untuk Clustering Nasabah Kredit,‖ J. Teknol. Inf. dan Ilmu Komput., vol. 4, no. 2, pp. 1–6, 2017.

24. Y. P. Anggodo, A. K. Ariyani, M. K. Ardi, and W. F. Mahmudy, ―Optimation of Multi-Trip Vehicle Routing Problem with Time Windows using Genetic Algorithm,‖ J. Environ. Eng. Sustain. Technol., vol. 3, no. 2, pp. 92–97, 2017.

25. R. A. Rahadi, S. K. Wiryono, D. P. Koesrindartotoor, and I. B. Syamwil, ―Factors influencing the price of housing in Indonesia,‖ Int. J. Hous. Mark. Anal., vol. 8, no. 2, pp. 169–188, 2015.