*International Journal for Modern Trends in Science and Technology*, 6(12): 186-189, 2020
Copyright © 2020 International Journal for Modern Trends in Science and Technology
ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST061236
Available online at: http://www.ijmtst.com/vol6issue12.html

# Machine Learning House Price Prediction

Manu Shahi[1] | Abhay Singh[1] |Amita Goel[2]| Vasudha Bahl[3] | Nidhi Sengar[3]

[1]B-tech scholar, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India
[2]Professor, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India
[3]Assistant Professor, Department of Information Technology Maharaja Agrasen Institute of Technology, Delhi, India.

**To Cite this Article**

Manu Shahi, Abhay Singh, Amita Goel, Vasudha Bahl and Nidhi Sengar, "Machine Learning House Price Prediction", *International Journal for Modern Trends in Science and Technology*, 6(12): 186-189, 2020.

**Article Info**

Received on 10-November-2020, Revised on 30-November-2020, Accepted on 04-December-2020, Published on 07-December-2020.

## ABSTRACT

*This document present the implementation of Machine Learning algorithms for the prediction of the house and the real estate prices. As the house and real estate prices are subject to change with the market conditions, so it become very difficult to predict the real estate prices with the conventional methods as it may sometimes gives some exaggerated result that may incur losses. To predict the prices more accurately and precisely we predict the prices based on the statics of that particular area which has all the trends and factors on which the price is dependent. To analyse these data , several algorithms are used namely random forest, linear regression , lasso regression etc. Use of these algorithms decreases the margin of error and more precise result are achieved. So,we at this point recommend the real estate agents and house vendors as well as the people to look into the model for better valuation of the house. This model can also be integrated with the real estates websites to give better recommendation based on the prices using Machine Learning Algorithms.*

***KEYWORDS:*** *Linear Regression, Standard Deviation, Mean, Lasso Regression, Random Forest Regression, One Hot Encoding,Grid Search CV.*

## I. INTRODUCTION

In this modern era, when the organisations are receiving the millions of data every second, so it become a vital role to analysis all the data to enhance the user experience. To analyse these data efficiently several algorithms had been proposed and based on these algorithms, we came up with the concept of Machine Learning where we train the machines based on these data to take decisions. Machine Learning is not a different subject altogether but it is the subset of Artificial Intelligence where we train the models.Machine learning predictions works on the statistical methods and operations which requires the data. So , we can also say that data is the heart of the machine learning.

What is Learning? Truecaller application identifies the spam calls and pop up a notification based on the previous calls made from that number. This is done by the mean that if a receiver receives a call and mark it as a spam call, next time when humans receives a call from that particular number then based on the previous dataset and number of reports, it identifies the call as a spam and pop up a notification.

At some places this learning from memorization can be useful but what if a new number arrives, then this model will fail to identify the spam call.To make our model more accurate it must be something which can learn from the generalization. The generalization means classifying a specific number group as suspicious and more likely to be a spam.

These algorithms can be classified into two phases training phase and testing phase. Mainly these algorithms are of three types: supervised learning, unsupervised learning and reinforcement learning. In supervised learning the models are trained using well maintained dataset that means some target values are associated them already.Logistic Regression, Support vector machine and decision tree are some examples where the output is mapped with the input dataset.

In Unsupervised learning model has to act without any guidance as the dataset do not have any label and clusters and patterns are identified.Independent component analysis,Singular value decomposition, KNN,Principle component analysis are some examples.

Reinforcement learning is when the model is provided without any dataset and it has to act according to the given situation on its own from the self learning and by looking for the current best possible solution for the scenario.This model keeps on learning from it's previous experiences. Deep Deterministic policy gradient uses the reinforcement learning.

The price of a house depends on the several factors like number of bedrooms, bathrooms, locality, crime rate, pollution, amenities etc. So predicting the best price of a house depending on these factors is the purpose of the this model.

## II. METHODOLOGY

### A. DATASET

The dataset is taken for the banglore which have several parameter which are area type, number of bedrooms, bathrooms, total area of house, society , balcony, price. The area type is divided into three categories based on their built up.

### B. LINEAR REGRESSION

It is used to predict value of dependent variable on independent variable. As to predict the value of dependent variable it requires the set of data , so it comes under the category of supervised learning.

For a given regression line: Y=aX+b, where a is the slope of the line and b is the intercept made by the line on the axis.

**a**= $\{(\sum_0^n y)(\sum_0^n x^2) - (\sum_0^n x)(\sum_0^n xy)\} \div \{n(\sum_0^n x^2) - (\sum_0^n x)^2\}$

$$b = \{n\left(\sum_0^n xy\right) - \left(\sum_0^n x\right)\left(\sum_0^n y\right)\} \div \{n\left(\sum_0^n x^2\right) - \left(\sum_0^n x\right)^2\}$$

### C. MULTIVARIABLE REGRESSION

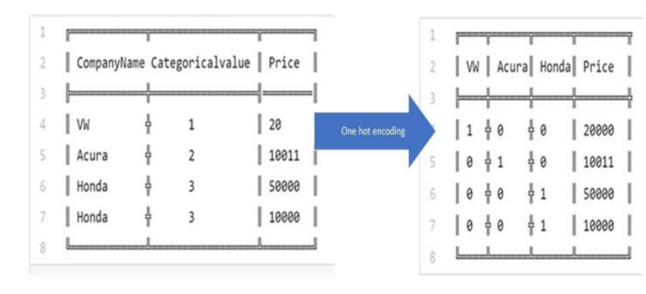
In this, the value of a dependent variable depends on more than one independent variable.

k(xi)=b0+b1xi1+b2xi2+···+bpxip

Here, b0, b1,b2,b3... these are the regression coefficient and k(xi) is the predicted value.

### D. ONE HOT ENCODING

It is the encoding method in which categorical variable are converted into the numerical values so that it can apply the machine learning algorithms properly and provide a better prediction with higher accuracy. It assign 1 to the current column and 0 to the rest of the variables in other columns. So in this way it generates a separate binary code for each variable.If a datasheet has n separate categories then it will generate the n column for each category.



| CompanyName | Categoricalvalue | Price |
|---|---|---|
| VW | 1 | 20 |
| Acura | 2 | 10011 |
| Honda | 3 | 50000 |
| Honda | 3 | 10000 |

| VW | Acura | Honda | Price |
|---|---|---|---|
| 1 | 0 | 0 | 20000 |
| 0 | 1 | 0 | 10011 |
| 0 | 0 | 1 | 50000 |
| 0 | 0 | 1 | 10000 |

Figure 1

### E. *LASSO REGRESSION*

LASSO stands for Least Absolute Shrinkage and Selection Operator.

It is shrinkage and variable selection method for linear regression model and is used to minimise the prediction error in the model by shrinking the regression coefficient towards zero. Variables which have regression coefficient zero are exempted from this model.

$$\sum_{i=1}^{n}\left(y_i - \sum_j x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

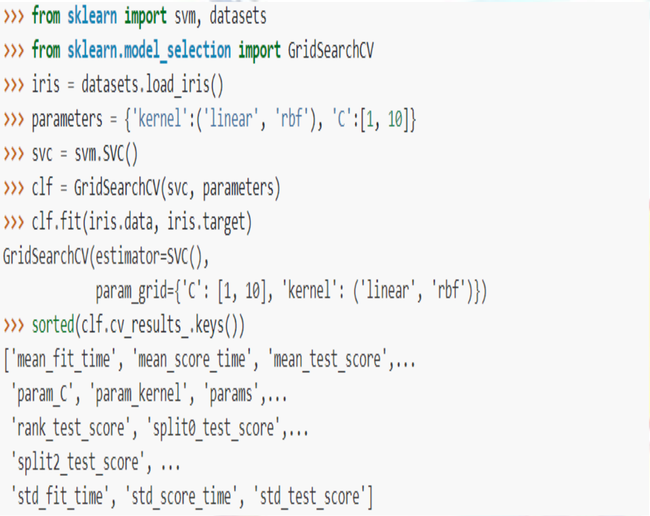
Here λ denotes the amount of shrinkage .The value of λ vary from 0 to ∞.

λ =0 means all factors are considered and is equals to linear regression .

λ = ∞ means no factor is considered.

## F. GRID SEARCH CV

It is used to determine the optimal value of the model by performing hyperparameter tuning.
As there is no way to determine the best value for hyperparameters in advance.So, GridSearchCV checks all the permutations and combinations to know optimal values.GridSearchCV is the function of SciKit Learn package .
How to use GridSearchCV:
*class*
sklearn.model_selection.**GridSearchCV**(estimator, param_grid, *, scoring=None, n_jobs=None, iid='deprecated', refit=True, cv=None, verbose=0, pre_dispatch='2*n_jobs', error_score=nan, return_train_score=False)



```
>>> from sklearn import svm, datasets
>>> from sklearn.model_selection import GridSearchCV
>>> iris = datasets.load_iris()
>>> parameters = {'kernel':('linear', 'rbf'), 'C':[1, 10]}
>>> svc = svm.SVC()
>>> clf = GridSearchCV(svc, parameters)
>>> clf.fit(iris.data, iris.target)
GridSearchCV(estimator=SVC(),
             param_grid={'C': [1, 10], 'kernel': ('linear', 'rbf')})
>>> sorted(clf.cv_results_.keys())
['mean_fit_time', 'mean_score_time', 'mean_test_score',...
 'param_C', 'param_kernel', 'params',...
 'rank_test_score', 'split0_test_score',...
 'split2_test_score', ...
 'std_fit_time', 'std_score_time', 'std_test_score']
```

Figure 2

## G. K-FOLD CROSS VALIDATOR

In this all the samples are divided into k groups called folds of equal size. All the predictions then are done using k-1 fold and the last fold is used to test.

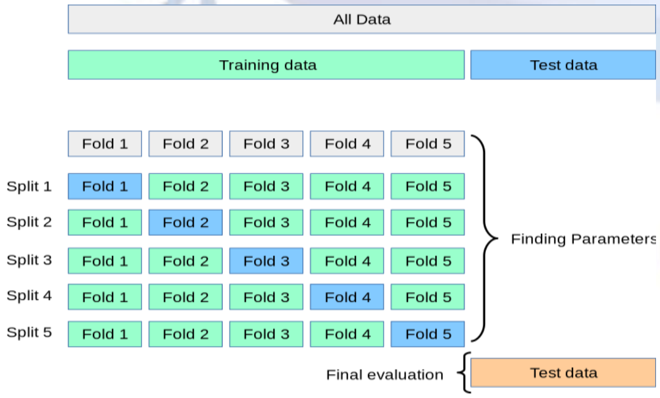


Figure 3

the subjects was tested using a 12-minute Cooper test and the corresponding equation was calculated.

## III. WORKING OF PROPOSED MODEL

The proposed model first loads the dataset which is available in .csv format. Once the dataset is extracted , hot encoding is used to remove the categorical values and missing values is filled using the proper algorithms. After cleaning the dataset is appended to original dataset and we start using machine learning algorithms to predict the prices of the area. We have used decision tree , Linear regression and Lasso regression to predict the prices. After this , the GridSearchCv which is a function of SK learn is used to find the optimal prediction.

## IV. PREVIOUS WORKS

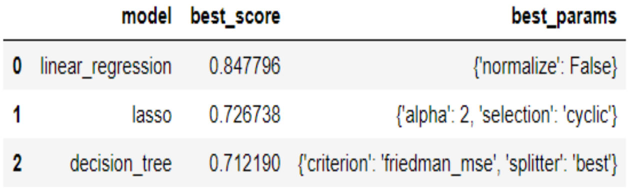
The real estate prices of an area depends on its socio-economic conditions and interest of buyers and sellers. The GDP and per capita income of that country also plays a valuable role in deciding the prices. As the Covid-19 has hit globally, the real estate market is under serious pressure and crumbling. As per many report[1] and articles published in newspapers it will take few years before the real estate market will come to its place. The price of a house depends on the number of bed rooms , size of area, bathrooms and many more other factors and several studies has been done in this area for the prediction of the house prices based on these factors.To predict house price manually with all this factors will produce unrealistic and unreliable results. So to enhance the accuracy and Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh [2] suggested a model “advanced house prediction system using linear regression”which is base on the regression relations to predict the prices. This uses the supervised learning where the datasets of the entire location is given with all the factors and then to accurately predict the price using the statistical modelling on those data. For this purpose he uses several algorithms like Linear Regression, Random Forest etc. to view the best price of the given location entered by the user.
Li and Chu; 2017 [3]proposed the model to check the accuracy of the prediction that has been made by the model. For this purpose Li and Chu used the Root Mean Square Error(RMSE) and Mean Absolute Percentage Error (MAE). This model also used the supervised learning in which the dataset is given to the model to predict the values.
Sebil selim (2008)[4] proposed a hedonic model to examine the effect of characteristic of houses on their price. He used the least square method to

estimate the hedonic model and the result tells that the houses near the pools,type of room, water system , number of rooms, area type, location characteristic have the higher prices as compare to other aspects.

For any project the literature review acts as a foundation of the idea and as most of the authors have concluded that price prediction depend on many number of factors so we take all the available algorithms into consideration to predict the prices and to understand the pros and cons of every algorithm and find the best algorithm for prediction.

## V. RESULTS

In the model three algorithms linear regression , decision tree and lasso is tested under different permutations and combinations. The table 1 below shows the accuracy of each algorithms under these parameters and linear regression comes out to be the best algorithm to predict the house prices based on the given params with the accuracy of 84.7. The decision tree has the least accuracy with just the accuracy of 71.2.



|   | model | best_score | best_params |
|---|---|---|---|
| 0 | linear_regression | 0.847796 | {'normalize': False} |
| 1 | lasso | 0.726738 | {'alpha': 2, 'selection': 'cyclic'} |
| 2 | decision_tree | 0.712190 | {'criterion': 'friedman_mse', 'splitter': 'best'} |

Table 1

After evaluating all the algorithms on different parameters we have managed to propose a model that can predict the prices more accurately using the linear regression. Further it will be integrated into the web development project so that the user can come to check the predictions more easily.The performance of the model will vary as per the features changes. We have proposed a model that can give consistent and accurate result to customers which satisfies their need by showing the correct output and preventing the risk of investing in the wrong house.

## VI. DISCUSSION

As the every human being is different and so as their needs and their economic conditions and requirement. Therefore, it become very difficult to predict those requirement of every human beings. This model works on the dataset of what people like to buy and sell and their basic parameters while looking for a house. So to get a 100% true result is not possible because of uncertainty of human psychology. This model works more on generalization of the humans' need and the available dataset as per requirements.From generalization we can say that after reading the various research paper published on this topic that people generally prefer to live in a place where schools, hospitals , employment opportunities and standard of life is better. The price of the houses at such places are higher than other places. Data is the fuel of this model and if data is not available for any part or place , this model will fail to predict the prices.

## REFERENCES

[1] https://www.thehindu.com/news/cities/mumbai/real-estate-industry-may-take-up-to-a-year-to-recover-says-study/article31936403.ece

[2] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, "A Hybrid Regression Technique for House Price Prediction", December 2017.

[3] Li, L. and Chu, K.-H. (2017). Prediction of real estate price variation based on economic parameters, Applied System Innovation (ICASI), 2017 International Conference on, IEEE, pp. 87–90.

[4] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.891.7453&rep=rep1&type=pdf

[5] W. T. Lim, L. Wang, and Y. Wang, ―Singapore Housing Price Prediction Using Neural Networks,‖ Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov., vol. 12, pp. 518–522, 2016.

[6] Azme Bin Khamis, Nur Khalidah Khalilah Binti Kamarudin -Comparative Study On Estimate House Price Using Statistical And Neural Network Model IJSTR VOLUME 3, ISSUE 12, December 2014 ISSN 2277-8616 .

[7] https://www.nber.org/system/files/working_papers/w13553/w13553.pdf

[8] K. W. Chau and K. W. Chau, "A critical review of literature on the hedonic price model," Int. J. Hous. Sci. Its Appl., 74(852), 2003, pp. 3–18.

[9] S. Sirmans, D. Macpherson, and E. Zietz, "The composition of Hedonic pricing models," J. Real Estate Lit., 13(1), 2005, pp. 1–44.

[10] M. Sasaki and K. Yamamoto, "Hedonic price function for residential area focusing on the reasons for residential preferences in Japanese metropolitan areas," J. Risk Financ. Manag., 11(3), 2018, pp. 2–18.

[11] C. K. Wing, S. K. Wong, and L. W. C. Lai, "Hedonic price modelling of environmental attributes: A review of the literature and a Hong Kong case study," Underst. Implement. Sustain. Dev., 2002, pp. 87–110.

[12] Quick Fact: Resident Demographics. National Council on Multi-Housing. Accessed:11/11/2017

[13] Sherwin Rosen. "Hedonic Prices and the Underground Market: Product Differences Pure competition. ' Submitted by: Journal of Political Economy 82.1 (1974)

[14] David E. Rapach , Jack K. Strauss " Forecasting real housing price growth in the Eighth District states"

[15] KAUKO T., (2003), "On current neural network applications involving spatial modelling of property prices", Journal of Housing and the Built Environment 18, pp. 159-181.

[16] Ayush Varma, Abhijit Sharma, Sagar Doshi, Rohini Niar, "House Price Prediction Using Machine Learning and Neural Networks", 2018.