Boston House Price Prediction Using Regression Models

Conference Paper · August 2022

DOI: 10.1109/CONIT55038.2022.9848309

Saptarsi Sanyal
National Institute of Technology, Silchar

Saroj Kr. Biswas
National Institute of Technology, Silchar

Dolly Das
National Institute of Technology, Silchar

Manomita Chakraborty
VIT-AP

2022 2nd International Conference on Intelligent Technologies (CONIT)
Karnataka, India. June 24-26, 2022

Boston House Price Prediction Using Regression Models

Saptarsi Sanyal, Saroj Kumar Biswas, Dolly Das, Manomita Chakraborty, Biswajit Purkayastha
Department of Computer Science and Engineering
National Institute of Technology Silchar
Silchar, Cachar, India
saptarsi_pg_21@cse.ac.in, bissarojkum@yahoo.com, dolly_rs@cse.nits.ac.in,
mou.look@gmail.com, biswajit_purakayastha@hotmail.com

***Abstract*— Everyone wishes to buy and live in a house which suits their lifestyle and which provides amenities according to their needs. There are many factors that are to be taken into consideration like area, location, view etc. for prediction of house price. It is very difficult to predict house price as it is constantly changing and quite often the prices are exaggerated for which people who want to buy houses, and various real estate agencies who want to invest in properties, find it difficult to buy or sell houses. For this reason, in this paper the author creates an advanced automated Machine Learning model using Simple Linear Regression, Polynomial Regression, Ridge Regression and Lasso Regression using the Boston house dataset to predict house price in future accurately, and to measure the accuracy of these models various measuring metrics like R-Squared, Root Mean Square Error (RMSE) and Cross-Validation are used. This paper also studies the correlation of various attributes of the Boston dataset using the heat map to see which attributes actually impact the prediction of the models. It removes the outliers which are present in the dataset to achieve good accuracy. In this paper it is observed that Lasso Regression performs better in all the measuring matrices whereas Simple Linear Regression performs poor in all of the measuring matrices.**

Keywords*— Machine Learning, Simple Linear Regression, Ridge Regression, Lasso Regression, house price

I. Introduction

Accurately predicting the value of a plot or house is an important task for many house owners, house buyers, plot owners, plot buyers or stake holders. Real estate agencies and people buy and sell houses all the time, people buy houses to live in or as an investment whereas real estate agencies buy it to run a business. But the problem arises in evaluation of the cost of the property. Over-validation / Under-validation have always been the issues faced in house markets due to lack of proper detection measures. It is also very difficult task. We know that features like size, area, location etc affect the price of the property but there are many other features also which affect the property such as inflation rates in market, age of the property etc. In order to overcome these problems a throw analysis is done using Machine Learning (ML) which is a branch of Artificial Intelligence (AI).

With the advancement of science and technology our daily life has become much easier. In today's world we use information and communication technology extensively. Every day a new technology emerges in our current digital age which improves the living standard of people. Sometimes these new technologies have negative effect but most of the times these technologies have positive effect [1]. AI or more widely known as AI is one of such technology which has improved the living standard of people [2] worldwide. AI is widely used now days in various fields like healthcare [3], real estate [4], stock market prediction [5], weather prediction [6], automobile [7] and also in many other fields [8][9]. AI has many subfields like Natural Language Processing, Machine Vision, Robotics, Expert System etc [10], but in this study ML is used.

ML is a branch of AI which deals with certain tasks using past data or recorded data and various algorithms. These tasks of ML involve classification, association, clustering and regression. ML can be used to make predictive models to make predictions for future or can be used to make descriptive models to make acquire some kind of knowledge from the given data. The main difference between ML programming and conventional programming is that, in conventional programming, programs are created manually by providing input data and based on the programming logic computer generates the output. However, in ML, the inputs and the outputs are fed into the algorithm creating the program. ML approaches are mainly divided into three categories these are Reinforcement Learning, Unsupervised Learning and Supervised Learning. In supervised learning the computers are given inputs and their desired outputs by a supervisor and the goal is to create a general rule using which a given input can be mapped into their desired output. In unsupervised learning there is no supervisor and the structures in the input are found on its own. The main goal of unsupervised learning is to find hidden structures or a mean towards an end. In reinforcement learning, in presence of a dynamic environment certain goals (such as driving a vehicle or playing against an opponent) are performed by the program. Using the feedback provided to the program as it navigates the problem space it is given reward and the goal is to maximize the rewards. Here machine is trained using various ML algorithms in Boston house dataset to create various models and using this trained machine models evaluation is done.

II. Literature Survey

A lot of past works have been done for predicting house prices. Different levels of accuracies and results have been achieved using different methodologies, techniques and datasets. A study of independent real estate market forecasting on house price using data mining techniques was done by Bahia [11]. Here the main idea was to construct the neural network model using two types of neural network. The first one is Feed Forward Neural Network (FFNN) and the second one is Cascade Forward Neural Network

978-1-6654-8407-7/22/$31.00 ©2022 IEEE

(CFNN). It was observed that CFNN gives a better result compared to FFNN using MSE performance metric.

Mu et al. [12] did an analysis of dataset containing Boston suburb house values using several ML methods which are Support Vector Machine (SVM), Least Square Support Vector Machine (LSSVM) and Partial Least Square (PLS) methods. SVM and LSSVM gives superior performance compared to PLS. Beracha et al. [13] proved that high amenity areas experience greater price volatility by investigating the correlation between house prices volatility, returns and local amenities. Law [14] finds that there is a strong link between house price and street based local area compare to the house price and region based local area. Binbin et al. [15] to study London house price build a Geographically Weighted Regression (GWR) model considering Euclidean distance, travel time metrics and Road network distance. Marco et al. [16] to reduce the prediction errors, a mixed Geographically weighted regression(GWR) model is used that emphasize the importance and complex of the spatial Heterogeneity in Australia. Using State level data in USA, Sean et al. [17] have examined the correlation among common shocks, real per capita disposable income, house prices, net borrowing cost and macroeconomic, spatial factors and local disturbances and state level population growth. Joep et al. [18] using the administrative data from the Netherlands have found that wealthy buyers and high income leads to higher purchase price and wealthy sealer and higher income leads to lower selling price.

III. Methodology

ML has various analyses like classification, regression, clustering and association and this analysis use various algorithms like logistic regression, Simple linear regression, support vector machine, Naïve Bayes etc. The analysis used here is regression, as regression is used for predicting continuous variable and house price prediction is prediction of continuous value. Various statistical methods are used for regression analysis to model the relationship between a dependent/ target variable and the independent or the predictor variable with one or more independent variables. The change of a dependent variable corresponding to an independent variable when other variables are kept fixed is understood using regression analysis. Certain features such as temperature, age, salary, house price etc. continuous/real values are predicted using regression.

A. *Model Selection*

Model selection is one of the most important tasks in ML for doing accurate prediction. Correct models must be selected to get good accuracy. There are various models available under regression analysis but for this paper four regression models are used which are Simple Linear Regression, Polynomial Regression, Lasso Regression and Ridge Regression on the Boston house dataset. After implementation of these models we measure the accuracy by splitting the dataset into two parts which are training dataset and test dataset. We use 80% of the dataset as training data and 20% is used as test data. The fitting of our models is done using the training dataset and evaluation of the model is done in test dataset.

B. *Techniques Used*

The techniques that are implemented on the Boston house dataset in this paper are Simple Linear Regression, Polynomial Regression, Lasso Regression and Ridge Regression.

1) *Simple Linear Regression*

In this type of regression model a linear relationship is established among the target variable which is the dependent variable (Y) and a single independent variable (X). Linear Relationship between dependent and independent variable is established by fitting a regressor line between them. The equation of the line is given by:

$$Y=a+bX \tag{1}$$

where "a" and "b" are the model parameter called as regression coefficients. When we take the value of X as 0, we get the value of "a" which is the Y intercept of the line and "b" is the slope that signifies the change of Y with the change of X. If the value of "b" is large then it means with a little change in X there will be a huge change in Y and vice versa. To compute the values of "a" and "b" we use the Ordinary Least Square Method. The values predicted by the model Linear Regression may not always be accurate. There may be some difference hence we add an error term to the original equation (1), it helps for better prediction of the model.

$$Y=a+bX+\mathcal{E} \tag{2}$$

There are some assumptions that are to be made in case of simple linear regression and those are as follows:

1. The number of observations must be greater than the number of parameters present.
2. The validity of the regression data is over a restricted period.
3. The mean of the error term has expected value of 0, which means that the error term is normally distributed.

2) *Polynomial Regression*

It is a special case of Simple Linear Regression. Unlike in linear regression where the model tries to fit a straight regression line between the dependent and independent variable, here a line cannot be fit as there doesn't exists any linear relationship between the target variable and the predictor variable. Here instead of straight line a curve is being fitted against the two variables. This is accomplished by fitting a polynomial equation of degree n on the non-linear data which forms a curvilinear relationship between the dependent and independent variables. In polynomial regression the independent variable may not be independent of each other unlike that in case of simple linear regression. The equation of polynomial regression is as follows:

$$Y = a+b_1X^1+b_2X^2+b_3X^3+........+b_nX^n \tag{3}$$

The advantages of polynomial regression are as follows:

1. Polynomial Regression offers the best estimate of the relationship between the dependent and independent variable.
2. The higher the degree of the polynomial the better it fits the dataset.
3. A wide range of curves can be fit into polynomial regression by varying the degree of the model.

The disadvantage of polynomial regression is as follows:

1. These are too sensitive towards the presence of outliers in the dataset, as the presence of outliers will increase the variance of the model. And when the model encounters any unseen data point it under performs.

3) Ridge Regression

In data which are suffering from multi-collinearity ridge regression is used. It is a tuning process which is used to analyze data having multi-collinearity. Here approximation of coefficient of regression model is done in case where the independent variables are highly associated. Bias is introduced to get better prediction. The complexity of the model is reduced using this regularization technique which is also known as L2 Regularization. In ridge regression, by adding penalty term, cost function is altered. Bias which is added is known as Ridge Regression penalty. It is calculated as λ* (Squared weight of individual features), where λ is the parameter tuning. In Ridge Regression, Regularization of the co-efficient of the model is done by the penalty term and hence Ridge regression reduces the amplitude of the co-efficient that decreases the complexity of the model. It is observed that the equation (4) becomes cost function of Linear Regression model if the value of λ tends to 0. The model resembles that of linear Regression if the value of λ is linear. A general Polynomial Regression or Linear Regression will fail if there is highly co-linearity between the independent variables. For this reason Ridge Regression is used. If the parameters are more than samples then it can be solved by Ridge Regression. The least Square determines the values of the parameters for the equation (4), which diminishes the sum of squared residuals. But in contrast the Ridge Regression regulates the value for parameters that results in minimization of the sum of squared residuals along with an additional term λ*b^2. Ridge Regression performs L2 regularization.

$$Loss_{Ridge} = \sum(y_i - y^*)^2 + \lambda b^2 \quad (4)$$

where y* = a + bX is the predicted value.

4) Lasso Regression

Lasso or Least Absolute Shrinkage and Selection Operator are very much similar to Ridge regression. In ML for selection of significant subset of variables Lasso regression is used. The prediction accuracy of Lasso regression is usually higher when compared to interpretations of other model. Similar to ridge, lasso also adds a little amount of bias to its result which thereby decreases the variance of the model. Lasso Regression is evaluated by the following:

Residual Sum of Squares + λ * (b=Sum of the absolute value of the magnitude of coefficients)

Here λ denotes the amount of shrinkage. The main difference between ridge and lasso is that, ridge reduces the slope asymptotically close to zero, whereas Lasso reduce the slope all the way down to zero which results in the elimination of useless parameters from the equation that don't have any significance role for predicting the value of the target variable. When the predictors have huge coefficients, Lasso shows better performance.

$$Loss_{Lasso} = \sum(y_i - y^*)^2 + \lambda|b| \quad (5)$$

where y* = a + bX is the predicted value.

C. Data Description

The dataset used in this project comes from the UCI Machine Learning Repository which concerns housing values in the suburbs of Boston. This data was collected in 1978 and contains 506 entries which give information about 14 attributes of homes from various suburbs located in Boston and one "target" attribute. The attribute description of this dataset is given below:

- CRIM: This is the per capita crime rate by town
- ZN: the proportion of residential land zoned for lots over 25,000 sq.ft
- INDUS: the proportion of non-retail business acres per town.|
- CHAS: Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10 million)
- RM: The average number of rooms per dwelling
- AGE: the proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centers
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per $10,000
- PTRATIO: the pupil-teacher ratio by the town
- B: - 1000(Bk - 0.63)^2, Where Bk is the proportion of blacks by the town
- LSTAT: "% lower status of the population"
- MEDV: "The median value of owner-occupied"

The "target" variable in this dataset is the MEDV variable on which will be predicted by the ML models. The rest of the variables are used for training the models. Statistics of the features are described in the table below:

TABLE I. DATASET STATISTICS

S. No.	Attribute	Mean	Minimum	Maximum
1.	CRIM	3.61	0.006	88.97
2.	ZN	11.36	0.00	100.00
3.	INDUS	11.13	0.46	27.74
4.	CHAS	0.069	0.00	1.00
5.	NOX	0.55	0.385	0.871
6.	RM	6.284	3.56	8.78
7.	AGE	68.574	2.90	100.00
8.	DIS	3.795	1.129	12.126
9.	RAD	9.549	1.00	24.00
10.	TAX	408.23	187.00	711.00
12.	PTRATIO	18.455	12.60	22.00
13.	B	356.67	0.32	396.90
14.	LSTAT	12.65	1.73	37.97
15	MEDV	22.53	5.00	50.00

The dataset doesn't contain any null value nor does it contain any duplicate row. The dataset contain 13 numerical values and only one categorical value which is "CHAS". For the attribute "ZN" it is observed that the 25^{th} and 75^{th} percentile are 0, implying that the data is highly skewed. This is because the attribute "ZN" is a conditional variable.

For the attribute “CHAS” it is observed that the 25th, 50th and the 75th percentile is 0 meaning that it is also highly skewed. This is because “CHAS” is a categorical variable and it contains values either 0 or 1. Another important observation is made and that is that the maximum value of “MEDV” is 50.00, so it seems from the description of data that “MEDV” is censored at 50.00 (corresponding to a median price of $50,000). Based on this observation about “MEDV” it seems that values above 50.00 will not be helpful so we remove them. It is also observed that the attributes “CRIM”, “ZN”, “RM” and “B” are having outliers so we remove them. From the histogram it is observed that the attributes “CRIM”, “ZN” and “B” are having highly skewed distributions. The attribute “MEDV” has normal distribution whereas other attributes have either normal or binomial distribution except “CHAS” as it is a discrete variable. After that heat map is implemented to see the correlation of the attributes.

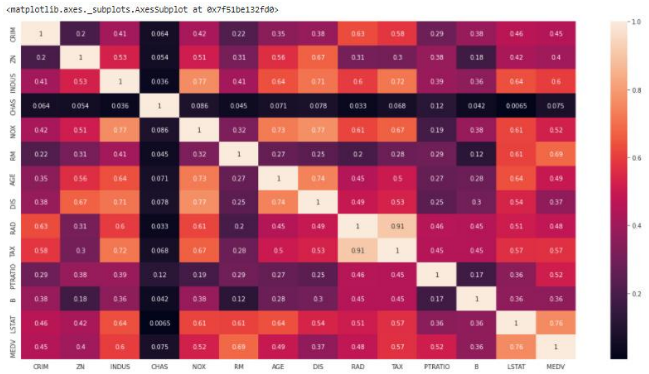

Fig. 1. Heatmap

From heat map it is observed that the attributes “TAX” and “RAD” are highly correlated. As both of them are highly correlated they are having similar behavior and will also have similar impact while doing prediction calculation. So rather than keeping redundant attributes it is always better to remove them as it will save space and computation time for complex algorithms. From heat map it is also observed that the attributes “LSTAT”, “INDUS”, “RM”, “TAX”, “NOX”, and “PTRAIQ” are having correlation score of above 0.5 with the “MEDV” which is a good indication of using them as predictors, so keeping only these eight attributes we discard other attributes. Then skewness of the data is removed using log transformation. These are the steps taken to refine the data. Refining of data is important to get accurate and good evolution of the models. If data pre-processing is not done then we will not get good result.

IV. Results

After necessary data pre-processing is done various models are implemented and evaluated. For evaluation of models Train-Test split method is used in this paper. Data is split into 80% and 20%, 80% of the data is used as training data and the rest of the data is used as test data. The first partition which is the training data is used for fitting the model and the second partition which is the test data is used for testing the accuracy of the model. There are various ways of splitting the data like 70-30, 70% for training and 30% for testing or 75-25, 75% for training and 25% for testing etc there is no hard and fast rule, but in this paper the dataset is split into the ratio mentioned above which is 80% for training and 20% for testing. After data set is split the models are implemented which are Simple Linear Regression, Polynomial Regression, Ridge Regression and Lasso Regression and the accuracy of the models are measured. To measure the accuracy of the models the metrics used are RMSE, R-Squared and cross validation. These are described as follows:

A. RMSE

For evaluation of the quality of predictions Root Mean Square Error (RMSE) is used. It measures the standard deviation of the prediction errors (Residuals). Residuals are basically the measure of the Euclidean distance of the data points from the regression line. To compute RMSE the following formula is used: The absolute fit of the model is shown by RMSE. The lower the value of RMSE is the better the fit of the model is and better the accuracy is achieved.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - y^*)^2}{n}} \quad (6)$$

B. R-Squared

A good measure for evaluation of the fitness of the model is R-Squared. The value of R-Squared ranges from 0 to 1, 0 being 0% accurate while 1 being 100% accurate. The larger the value of R-Squared is the better the fit is achieved. For evaluation of R-Squared the following formula is used: Here SSE is the squared sum of error terms i.e. the sum of the squared residuals. Residuals are the difference between the observed and predicted values. SST is the sum of the squared total and it is the difference of each observation from the overall mean.

$$R^2 = 1 - \frac{SSE}{SST} \quad (7)$$

where SSE (Squared sum of error): sum of the squared residuals, which is squared differences of each observation from the predicted value. $\sum(y_i - y^*)^2$and,

SST (Sum of Squared Total): squared differences of each observation from the overall mean. $\sum(y_i - \hat{y})^2$.

where y_i is the observed value, y^* is the predicted value and $\hat{y}$ is the mean of the observed values.

C. Cross-Validation

It is also known as rotation estimation or out of sample testing. It is a resampling technique in which different portions of the data is used for training and testing on different iterations. It is mainly used for predictions and cases where evaluation of the accuracy of the accuracy of a predictive model is needed. After implementation of the models it is observed that the best accuracy in R-Squared metric is achieved by Lasso Regression which is 88.72% and the second best accuracy is achieved by Ridge Regression which is 88.28% in the same metric. Then the third best accuracy was achieved by Polynomial Regression which is 74.27% when the degree of the polynomial is set at 2. The least accuracy was shown by Simple Linear Regression which achieved an accuracy of 73.66%. In case of cross validation the best accuracy is given by the Lasso Regression which is 85.57% and the least accuracy is observed in both Simple Linear Regression and Polynomial Regression both achieving an accuracy of 73.17%. Ridge Regression also performed well and got an accuracy of 85.53% in cross validation metric. In RMSE metric Ridge Regression and Lasso Regression got 2.88 and 2.83 score respectively whereas Linear Regression got score of 4.32 and Polynomial Regression got 4.27 score.

TABLE II. RESULT TABLE

Model Name	R-Squared	RMSE	Cross-Validation
Simple Linear Regression	73.66%	4.329	73.17
Polynomial Regression (degree=2)	74.27%	4.279	73.17
Ridge Regression	88.28%	2.888	85.83
Lasso Regression	88.79%	2.833	85.57

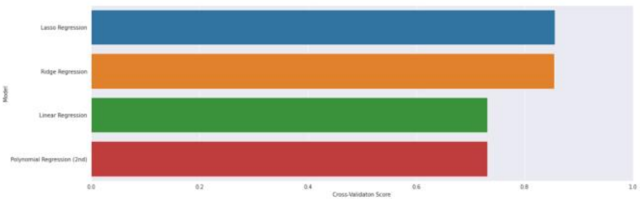

Fig 2: Cross Validation Score

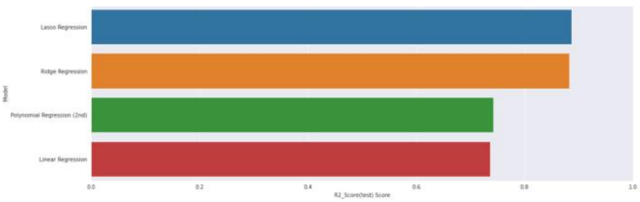

Fig 3: R-Squared

V. OBSERVATION

The Boston dataset used here contains only 506 instances which are very small however some important observations are made using this dataset. Firstly after splitting the data in 80-20, 80% for training and 20% for testing using Train-Test split method and using metrics like R-Squared, RMSE, Cross-Validation, it is seen that in all the cases Lasso Regression performs the best and the second best performance is given by Ridge regression. So the models which use regularization techniques perform the best in this dataset. This is achieved by reducing the over fitting of the models. It is also observed that Simple Linear Regression gives the least evaluation in all the cases. By keeping the degree of the polynomial Regression fixed at 2 it is observed that it achieves accuracy of over 70% when R-Squared and Cross-Validation metric are used. These are the observations made using this dataset however it doesn't imply that every time Simple Linear Regression will give the least performance out of the rest. For different datasets we will have different observations.

VI. FUTURE WORK

This model can be considered as the baseline for predicting house price. Further evaluation can be done here by increasing the data. More data can be collected and more attributes can be increased for getting a much better evaluation of the model. The data collected in Boston house dataset is from 1978 which is almost 50 years old and since then a lot of changes have occurred in house price due to inflation rate. Thus, new data can be collected and further evaluation can be made on the new collected data. In this paper four models are implemented which are Simple Linear Regression, Polynomial Regression, Lasso Regression and Ridge Regression on the Boston House dataset. More advanced models like Support Vector Machine, Decision Tree, Random Forest, Multiple Linear Regression etc can be implemented and the results can be compared. Other ensemble learning techniques can be used like Adaboost, Xgboost etc and the results can be compared to the previous models. Feature selection techniques like Linear Discriminant Analysis, Principle Component Analysis, Independent component Analysis etc can be used before implementing the models and a study can be made on the performance of the models before applying feature selection methods and after implementing feature selection methods. The observation can be made on how each of the feature selection methods impacts the performance of the model. Neural network and deep learning methods can also be applied and the performance can be studied. To increase the performance of the models and reduce the time complexity of models we can use optimization techniques like Particle Swarm optimization, Genetic Algorithm, and Ant Colony optimization etc. By implementing the optimization techniques an observation can be made on how each of these techniques impacts the models. There are various scopes of work that can be done on this field which will be very helpful to people who want to buy plot or house and also to real estate agencies for investing on houses.

VII. CONCLUSION

It is very important to predict house price accurately. To accurately predict house price various variables must be taken into consideration like location of the house, the views that are visible from the house, crime rate around that area etc. A lot of time people pay overprice from the actual market price for a real estate property, similarly a lot of time sellers get very low price compare to the actual market price of the property. Not only people, various estate agencies also face the same problem where they are not sure whether to invest toward a certain property or not. They are confused as they are not able to predict what the price of the house can be in future. The main purpose of this paper is to help people who are facing these issues to predict the house price in future years. In this paper an intelligent system is made using the Regressor models which are Simple Linear Regression, Polynomial Regression, Ridge Regression and Lasso Regression on the Boston House Dataset to predict the house price. In this paper it is observed that using the Boston house dataset, and implementing various data preprocessing techniques which are needed on the dataset and then splitting the dataset into 80-20, 80% for training and 20% for testing, Lasso Regression performs the best. Then the next best performance is given by Ridge Regression and the least performance is observed in Simple Linear Regression model. The best performance is given by all the methods which use regularization techniques. This may change if we use different dataset or different pre-processing approaches but in this case this is the results that we get.

REFERENCE

[1] Mansi Bosamia,'Positive and Negative Impacts of Information and Communication Technology in our Everyday Life', International Conference on "Disiplinary and Interdisciplinary Approaches to Knowledge Creation in Higher Education: CANADA and INDIA (GENESIS 2013).

[2] Rahul Reddy Nadikattu, 'THE EMERGING ROLE OF ARTIFICIAL INTELLIGENCE IN THE MODERN SOCIETY', International Journal Of Creative Research Thoughts, 2016.

[3] Ravi Manne and Sneha C Kantheti, 'Application of Artificial Intelligence in Healthcare: Changes and Challenges', Current Journal

of Applied Science and Technology, 2021, DOI:10.9734/CJAST/2021/v40i631320

[4] Woubishet Zewdu Taffese, ‘A Survey on Application of Artificial Intelligence in Real Estate Industry’, 3rd International Conference on Artificial Intelligence in Engineering andTechnology, 2006.

[5] Ferreira, F. G. D. C., Gandomi, A. H., & Cardoso, R. T. N. ‘Artificial Intelligence Applied to Stock Market Trading: A Review’, IEEE, 2021, doi:10.1109/access.2021.3058133.

[6] Anandharajan, T. R. V., Hariharan, G. A., Vignajeth, K. K., Jijendiran, R., & Kushmita. (2016).’ Weather Monitoring Using Artificial Intelligence’. 2016 2nd International Conference on Computational Intelligence and Networks (CINE). doi:10.1109/cine.2016.26.

[7] Tong, W., Hussain, A., Bo, W. X., & Maharjan, S., ‘ Artificial Intelligence for Vehicle-to-Everything: a Survey’, IEEE, 2019 ,doi:10.1109/access.2019.2891073.

[8] Sumit Das, Aritra dey, Akash Paul and NAbamita Roy, ‘Applications of Artificial Intelligence in Machine Learning: Review and Prospects’, International Journal of Computer Applications, 2015, DOI:10.5120/20182-2402.

[9] Avneet Pannu, ‘Artificial Intelligence and its Application in Different Areas’, International Journal of Engineering and Innovative Technology (IJEIT), Volume 4, Issue 10, April 2015.

[10] John A. Bullinaria, ‘IAI : The Roots, Goals and Sub-fields of AI’, 2005.

https://www.cs.bham.ac.uk/~jxb/IAI/w2.pdf

[11] Bahia, I. S. (2013). A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study. International Journal of Intelligence Science,03(04), 162- 169. doi:10.4236/ijis.2013.34017.

[12] Mu, J., Wu, F., & Zhang, A. (2014). Housing Value Forecasting Based on Machine Learning Methods. Abstract and Applied Analysis, Volume 2014 (2014), Article ID 648047, 7 pages. Retrieved April 2017, from https://www.hindawi.com/journals/aaa/2014/648047/.

[13] 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.

[14] Stephen Law, "Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of Metropolitan London", Cities, vol. 60, Part A, pp. 166–179, Feb. 2017.

[15] Binbin Lu, Martin Charlton, Paul Harris & A. Stewart Fotheringham, "Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data", International Journal of Geographical Information Science, pp. 660-681, Jan 2014.

[16] Marco Helbich, Wolfgang Brunauer, Eric Vaz, Peter Nijkamp, "Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria", Urban Studies, vol. 51, Issue 2, Feb. 2014.

[17] Sean Holly, M. Hashem Pesarana, Takashi Yamagata, "A spatio-temporal model of house prices in the USA", Journal of Econometrics, vol. 158, Issue 1, pp. 160–173, Sep. 2010.

[18] Joep Steegmans, Wolter Hassink, "Financial position and house price determination: An empirical study of income and wealth effects", Journal of Housing Economics, vol. 36, pp. 8-24, June 2017.