WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Predicting the Future: Evaluating and Comparing Machine Learning Models for USA House Price Prediction

[1]Humaira Rashid Hiya, [2]Mahamoda Akter, [3]Abid Hasan, [4]Jahanara Suchi, [5]Ashadu Jaman Shawon

American International University-Bangladesh

ABSTRACT:

This paper compares the performance of k-Nearest Neighbors and Linear Regression in house price prediction, using a dataset drawn from Kaggle. It follows that the accuracy for the k-NN model was 98%, while the one for the Linear Regression model was 94%. The results evidently showed that when compared to the linear assumptions of the Linear Regression model, k-NN really captures complex and nonlinear relationships within the housing data. It can be used to compare and thus show the potential of advanced regression techniques in improving predictive accuracy in real estate analytics.

Keywords: House Price Prediction, k-Nearest Neighbors (k-NN), Linear Regression, Machine Learning, Data Normalization, Predictive Analytics

Introduction:

House price prediction is one of the intrinsic tasks of real estate analytics, providing insights into the view of buyers, sellers, investors, and policymakers alike [3]. From that alone, the impact of an accurate price prediction in decision-making processes cannot be overemphasized, since it affects investment strategies and market assessments [4]. Data science and machine learning in this domain further increase the relevance of predictive modeling.

The real estate market is intrinsically very complex, driven by several factors that include location, size, condition, and the prevailing economic conditions [2]. Traditional statistical methods such as linear regression have long been used in modeling these relationships. One reason this model could be considered the best is that, in comparison with other models, it is very easy and simple to understand how each feature contributes toward the target variable. Such methods, however, assume a linear relationship between features and target values in a tacit fashion, which may not be the case in real-world scenarios [3].

Modern machine learning techniques resolve this problem using a very different method [1]. K-nearest neighbors is an instance-based learning algorithm that is non-parametric and generates predictions based on proximity between data points in the feature space [5]. Unlike linear regression, k-NN does not assume the functional form between the features and target variable; hence, it can model quite complex nonlinear relationships.

The next paper contrasts k-NN with the linear regression model on house price prediction in the USA, based on a dataset obtained from Kaggle. It was in an attempt to bring out which approach is more appropriate for handling real estate data complexities by assessing the accuracy and effectiveness of these models [4]. The result might provide better predictive accuracy and insight into helpful activities by partners within the housing market [1].

It introduced the topic in detail with respect to the necessity of house price prediction and the relevance of different modeling approaches in achieving that, together with the objectives of the research.

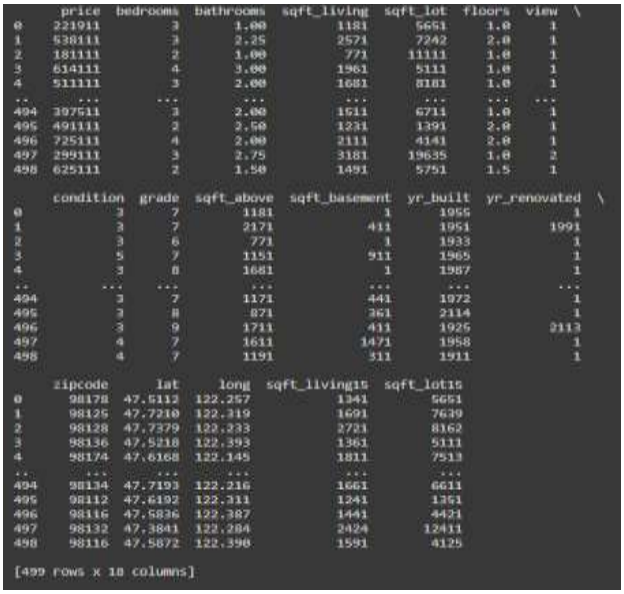
Literature Review

Studies on house prediction using machine learning (ML) has witnessed substantial growth in recent times. The raising complexity of house has urged experimenters to claw into the realm of deep learning (DL) algorithms, with the combined CNN- LSTM model, a group of researchers achieved the highest delicacy to date which is 95.1 [6]. Despite the wide research with multiple ML algorithms, certain classifiers remain underutilized or entirely neglected, challenging a focused disquisition of their performance in house Prediction [6]. Addressing the challenge of overfitting in house prognostic, paper [7] introduces a new prediction system using a deep literacy neural network. By incorporating powerhouse layers to alleviate overfitting, the proposed neural network surpasses other state- of- the- art styles, arising as the optimal pantomime for the Pima Indians House. Data Set [7], Papers [8] and [9] emphasize the significance of early house discovery, emphasizing the eventuality for ML models to not only prognosticate the circumstance of house but also discern the type of the complaint. The proposed DLPD (Deep Learning for Predicting House) model, constructed using retired layers of a

deep neural network and incorporating powerhouse regularization, achieves high delicacy in training datasets [8]. The pressing global impact of house, as stressed by the International House Federation (IDF) [9], underscores the need for effective prognostic tools to grease early discovery and life interventions. Likewise, paper [10] emphasizes the adding frequency of house and the critical need for accurate opinion. The paper explores data analytics and machine learning algorithms for enhancing delicacy in house prediction, admitting the significance of examination of retired patterns in business data [10]. In synthesizing the literature, it becomes apparent that the integration of Deep Learning and Machine Learning ways, coupled with a focus on underutilized classifiers, powerhouse styles for overfitting forestallment, and the bracket of house types, contributes significantly to the field of house vaticinator. The proposed models parade estimable delicacy and outperform being styles, italicizing the eventuality for ML in addressing the global challenge of house.

Methodology

Dataset The process of collecting the precise coffers involved the operation of a many technologies and data sources. Coffers from Google Scholar, IEEE Xplore, Research Gate, ACM, and IGI Global have been gathered with delicacy. The theme that was chosen meant that there were not as numerous coffers available. Every source has been precisely named. Using the coffers, it was doable to detect some comprehensive instructions for enforcing the automated fashion for relating house. The design of this model was done in phases. In the morning, the model helps to collect the needed datasets. In this dataset than 500 customer data are present, it's a business data and useful for machine literacy. The data includes features similar as price, gender, body mass indicator (BMI), hypertension, heart complaint, smoking history, HbA1c position, and blood glucose position. The descriptions of the attributes are shown in figure 1.



```
      price  bedrooms  bathrooms  sqft_living  sqft_lot  floors  view  \
0    221911         3       1.00         1181      5651     1.0     1
1    538111         3       2.25         2571      7242     2.0     1
2    181111         2       1.00          771     11111     1.0     1
3    614111         4       3.00         1961      5111     1.0     1
4    511111         3       2.00         1681      8181     1.0     1
..      ...       ...        ...          ...       ...     ...   ...
494  307511         3       2.00         1511      6711     1.0     1
495  491111         2       2.50         1231      1391     2.0     1
496  725111         4       2.00         2111      4141     2.0     1
497  299111         3       2.75         3181     19635     1.0     2
498  625111         2       1.50         1491      5751     1.5     1

     condition  grade  sqft_above  sqft_basement  yr_built  yr_renovated  \
0            3      7        1181              1      1955             1
1            3      7        2171            411      1951          1991
2            3      6         771              1      1933             1
3            5      7        1151            911      1965             1
4            3      8        1681              1      1987             1
..         ...    ...         ...            ...       ...           ...
494          3      7        1171            441      1972             1
495          3      8         871            361      2114             1
496          3      9        1711            411      1926          2113
497          4      7        1611           1471      1958             1
498          4      7        1191            311      1911             1

     zipcode      lat     long  sqft_living15  sqft_lot15
0      98178  47.5112  122.257           1341        5651
1      98125  47.7210  122.319           1691        7639
2      98128  47.7379  122.233           2721        8162
3      98136  47.5218  122.393           1361        5111
4      98174  47.6168  122.145           1811        7513
..       ...      ...      ...            ...         ...
494    98134  47.7193  122.216           1661        6611
495    98112  47.6192  122.311           1241        1361
496    98116  47.5836  122.387           1441        4421
497    98132  47.3841  122.284           2424       12411
498    98116  47.5872  122.398           1591        4125

[499 rows x 18 columns]
```

Figure 1: Description of the attributes.

Business practitioners may find this helpful in identifying customers who may be at risk for house and in creating specialized treatment regimens. Researchers can also use the dataset to examine associations between different demographic and business characteristics and the risk of getting house.

Pre-processing: In this work here, pre-processing steps includes data loading, data type finding, missing value detect and remove, normalize the data, data type conversion all are done.

At first all the data has been loaded and the dataset has been checked to be run. All the row and column viewed. By finding missing value the data become more precise and error free. The missing value has been deleted from the dataset. By normalize the data all the data min, max and median has been shown and the values are normalizing (Figure 2). It increases the accuracy of the data.

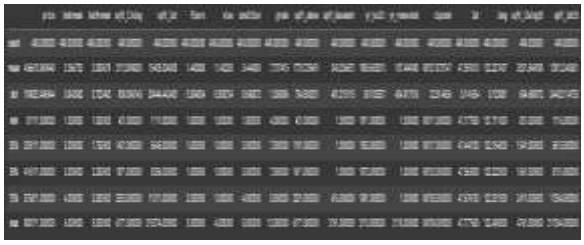


Figure 2: Normalized Data.

The target value of this research is to find out the diabetic. Plot makes easy decision to visualize all the data. SNS plot help to explore and understand data (Figure 3).

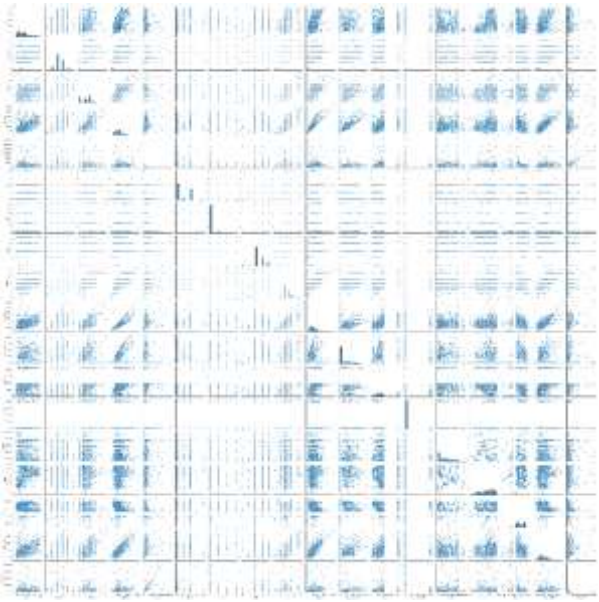


Figure 3: SNS Plot.

Another visualization plot is the heat map. Heat map help to visualize in a numerical and graphical way (Figure 4).

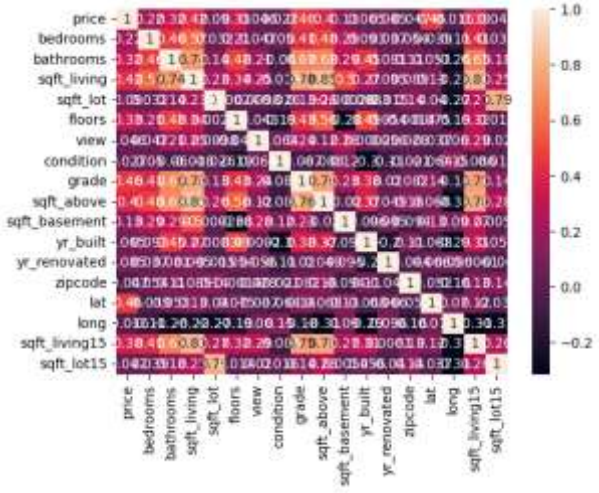


Figure 4: Heat Map.

Co-relation Matrix: Correlation functions characterize the link between microscopic variables at distinct locations, such as spin and density. All the attribute relation has been shown by viewing the co-relation. The co-relation matrix range between -1 to 1.

Figure 5: Co-relation Matrix.

Model Train:

Separately model has been trained for this prediction. Same data has been used for those models. For the first model linear regression model has been used. Where the test size is 40% and the random state is 42. All the attributes are in the same side and the target value is the diabetics. For the second model the Knn model has been used. Where the test size is 20% and the random state is 42.

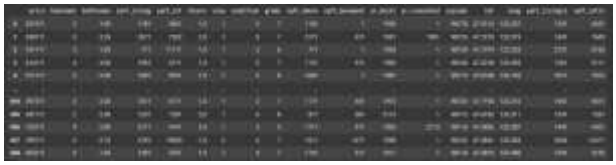


Figure 6: Train model using linear regression and Knn model.

Result and Analysis:

In Knn model the accuracy is 98.33% on the other hand in the linear regression the accuracy is 94.66%. Knn model is better than the linear regression for the diabetic prediction.

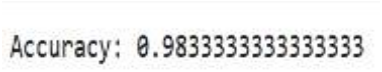


Figure 7: Knn Model

Figure 8: Linear Regression Model

References:

[1] Khosravi, M., Arif, S.B., Ghaseminejad, A., Tohidi, H. and Shabanian, H., 2022. Performance evaluation of machine learning regressors for estimating real estate house prices.

[2] Satish, G. Naga, Ch V. Raghavendran, MD Sugnana Rao, and Ch Srinivasulu. "House price prediction using machine learning." *Journal of Innovative Technology and Exploring Engineering* 8, no. 9 (2019): 717-722.

[3] Truong, Q., Nguyen, M., Dang, H. and Mei, B., 2020. Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, *174*, pp.433-442.

[4] Park, B. and Bae, J.K., 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert systems with applications*, *42*(6), pp.2928-2934.

[5] Zulkifley, N.H., Rahman, S.A., Ubaidullah, N.H. and Ibrahim, I., 2020. House price prediction using a machine learning model: a survey of literature. *International Journal of Modern Education and Computer Science*, *12*(6), pp.46-54.

[6] Thamarai, M. and Malarvizhi, S.P., 2020. House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering & Electronic Business*, *12*(2).

[7] Adetunji, A.B., Akande, O.N., Ajala, F.A., Oyewo, O., Akande, Y.F. and Oluwadara, G., 2022. House price prediction using random forest machine learning technique. *Procedia Computer Science*, *199*, pp.806-813.

[8] Rawool, A.G., Rogye, D.V., Rane, S.G. and Bharadi, V.A., 2021. House price prediction using machine learning. *Int. J. Res. Appl. Sci. Eng. Technol*, *9*, pp.686-692.

[9] Vineeth, N., Ayyappa, M. and Bharathi, B., 2018. House price prediction using machine learning algorithms. In *Soft Computing Systems: Second International Conference, ICSCS 2018, Kollam, India, April 19–20, 2018, Revised Selected Papers 2* (pp. 425-433). Springer Singapore.

[10] Singh, A.P., Rastogi, K. and Rajpoot, S., 2021, December. House price prediction using machine learning. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 203-206). IEEE.

[11] Banerjee, D. and Dutta, S., 2017, September. Predicting the housing price direction using machine learning techniques. In *2017 IEEE international conference on power, control, signals and instrumentation engineering (ICPCSI)* (pp. 2998-3000). IEEE.

[12] Madhuri, C.R., Anuradha, G. and Pujitha, M.V., 2019, March. House price prediction using regression techniques: A comparative study. In *2019 International conference on smart structures and systems (ICSSS)* (pp. 1-5). IEEE.