Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Utilize the Machine Learning Models to Forecast Home Values.:Seattle U.S.

Yunqi Zhang *

Qian'an First Senior High School, No. 2019, Steel City Street, Qian'an City, Tangshan, China.

* Corresponding Author Email: 100564@yzpc.edu.cn

Abstract. Understanding housing price predictions assists the government in better adjusting and formulating relevant policies to promote economic stability and sustainable social development. In this study, researchers collected a large amount of data on the Seattle real estate market, including housing characteristics (such as area, number of bedrooms, number of bathrooms), geographical location (such as neighborhood, nearby facilities), and historical price data. This paper used this data to train and test linear regression, and KNN prediction models to forecast future housing price trends. The linear regression model, on the other hand, models the linear relationship between a single independent variable and the dependent variable, predicting housing prices by fitting an optimal line. The KNN prediction model, based on the nearest neighbor algorithm, predicts by searching for the K nearest neighbor samples closest to the target sample. Researchers will compare the accuracy and effectiveness of these three methods in predicting Seattle housing prices to determine which method is most suitable for housing price forecasting. Through this analysis, they aim to provide a reliable housing price prediction model for local residents to help them make wiser real estate decisions.

Keywords: Machine learning, ensemble learning, housing price forecast, neural network

1. Introduction

When expressing what they desire in a property, homebuyers are likely to give consideration to aspects other than floor height or accessibility to a metro station [1,2,3]. Their key priorities can be peaceful surroundings and comfort. Nevertheless, evidence points to the possibility that factors other than price discussions, such as the number of bedrooms and house finishing, may influence purchasers' selections [4]. Seattle has a relatively stable economy, mainly in agriculture, manufacturing and service industries, with relatively equal income, and compared with other large cities in the United States, Seattle's social environment is relatively peaceful and the cost of living is low, and there is a good educational environment. There are some positive effects for people living in Seattle, and these will also affect local housing prices.Thereby, a high-accuracy model in predicting the property values is entailed.According to the research of other researchers, the truth and accuracy of the database plays a very important role in the prediction of housing prices. The database used in this article is from kaggle, a housing data set for Seattle.This data set is true and accurate, but it's also very comprehensive, and these conditions allow me to experiment boldly, but because there are so many conditions, I only explored a few of them in relation to housing prices.

The following is how this paper is organized: In Section 2, we provide each category of housing price prediction in order to introduce the relevant work done by our colleagues. In Section 3, we go into further depth about our procedures, explaining the theories behind them as well as the rationale behind our selection of them. After that, in Section 4, we look over the experimental findings and carry out analysis. Finally, section 5 includes a list of the study's conclusions.

2. Related work

Many factors influence the cost of housing, including the style of property, location, and square footage. The more elements we evaluate, the easier it will be to do our investigation.First of all,

Monson Matt illustrate how to predict housing price using hedonic pricing models [1]. This model is characterized by having concerned both internal conditions of the housing and the external economic conditions. Chengke Zou uses multivariate linear regression, random forest and cat's eye algorithms to verify the factors affecting house prices in Jinan, China [2].The experiments in Petaling Jaya and Selangor in Malaysia proved that taking data from the data set that is unrelated to the research object would damage the correctness of the experimental results, thus we must first determine the correlation between these data and the research findings [3].Sayan Putatunda used many sophisticated machine learning techniques are used on actual datasets, and their respective performances are evaluated [4]. Examples of these algorithms include random forests, gradient boosting, and artificial neural networks [5, 6, and 7]. In terms of prediction accuracy, he discovered that the random forest approach fared the best.

3. Methodologies

We used linear regression and KNN in order to find a more accurate result. So we must compare the difference of the two methods. In both models, we use MAE, MSE and R^2 as evaluation indicators of prediction accuracy.

3.1. Linear Regression

Linear regression is a simple yet powerful statistical method used to establish the relationship between a continuous target variable (such as house prices) and one or more independent variables. The basic assumption is that there is a linear relationship between the dependent variable and the independent variable. In house price prediction, we need to process a large amount of data, and linear regression has strong computational efficiency, so it is very suitable to be applied in this experiment [8]. For a dataset $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$, in which $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{id})^T, y_i\in R$, the objective of Linear Regression is to fit a linear model with coefficients $w=(w_1,w_2,\cdots,w_m)$ to minimize the residual sum of squares (RSS) between the true values and the predicted values. Mathematically this problem can be formalized as:

$$\min_w \|Xw-y\|_2^2 \tag{1}$$

Where

$$X=\begin{pmatrix}\boldsymbol{x}_1^T & 1\\ \boldsymbol{x}_2^T & 1\\ \vdots & \vdots\\ \boldsymbol{x}_m^T & 1\end{pmatrix},\quad y=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_m\end{pmatrix} \tag{2}$$

3.2. KNN

The KNN algorithm is a fundamental supervised learning method used for classification and regression problems. Its principle is based on the idea that similar samples in the feature space are more likely to belong to the same category, akin to the concept of "birds of a feather flock together". The basic idea of the KNN algorithm is simple: when given an unlabeled sample, it calculates the distance between the sample and all labeled samples in the training set, selects the K nearest samples based on the calculated distances, and then uses a voting mechanism to assign the unlabeled sample to the class with the most votes (for classification problems) or calculates the average value (for regression problems).

The main steps of the KNN algorithm include:

(1). Selecting a suitable distance metric method, such as Euclidean distance, Manhattan distance, Minkowski distance, etc.

(2). Determining the value of K, which represents the number of nearest neighbor samples to consider.

(3). for a given unlabeled sample, calculating its distances to all labeled samples in the training set.

(4). Selecting the K nearest samples based on distance.

(5). for classification problems, determining the class of the unlabeled sample through a voting mechanism; for regression problems, calculating the average value of the K nearest neighbor samples as the predicted value.

The KNN algorithm is simpler and more intuitive than linear regression. It's easy to understand and implement. There is no complex mathematical model, it is a lazy learning algorithm that does not require building a model during the training phase. Instead, it directly uses existing training data when making predictions. KNN and linear regression, have their own advantages and disadvantages. In terms of complexity, linear regression only needs to build a linear model to easily calculate all data, while KNN needs to calculate the distance between all samples. Therefore, KNN will have higher computational cost when processing large amounts of data.

4. Result

4.1. Experimental Settings

In this research, all models were implemented in Python 3.7.11 environment, with Pandas, Scikit-Learn, Tensorflow and XGBoost packages. The hardware configurations comprise a 2.40GHz i5-9300H CPU, a GTX 1660Ti GPU and 16GB RAM.

4.2. Evaluation Metrics

Root Mean Squared Error (RMSE)

$$\text{RMSE}=\sqrt{\text{MSE}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{3}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value and N is the number of predictions.

RMSE represents the standard deviation of the prediction errors (residuals), which evaluates the degree of scatter of these residuals [9,10]. In other words, it indicates how centralized the data is around the line of best fit. Notice that RMSE value is scale-dependent, which increases significantly if the scale of error increases.

Adjusted R^2

The coefficient of determination, or R^2 is defined by

$$R^2=\frac{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{4}$$

Where $\sum_{i=1}^{N}(y_i-\bar{y})^2, \sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2, \sum_{i=1}^{N}(y_i-\hat{y}_i)^2$ denote the total sum of squares (TSS), the explained sum of squares (ESS) and residual sum of squares (RSS) respectively.

The adjusted R^2 is calculated by

$$R_{adj}^2 = 1 - \frac{(1-R^2)(N-1)}{N-k-1} \quad (5)$$

Where k is the number of variables in the regression model.

Both R^2 and R_{adj}^2 represent the proportion of the variance for a dependent variable that can be explained by independent variables in a regression model. However, R_{adj}^2 is used as an evaluation of how successful a regression model predicts responses for new observations. It will increase when more useful variables are added to the model, and decrease reversely.

4.3. Model Evaluation

A linear connection between a response and an explanatory variable may be modeled using simple linear regression. Price is the response variable, and my goal is to forecast housing prices. Nevertheless, in order to build a basic model, we must select a feature. Based on my examination of the dataset's columns, Living Area (square feet) appears to be the feature. According to the correlation matrix, pricing has the largest coefficient when it comes to living area (square feet).

It is straightforward to depict the simple regression because it just has two dimensions. The outcome of the simple regression is shown in the figure 1. Although it doesn't appear to be a perfect fit, finding a perfect fit might be challenging when working with real-world statistics.

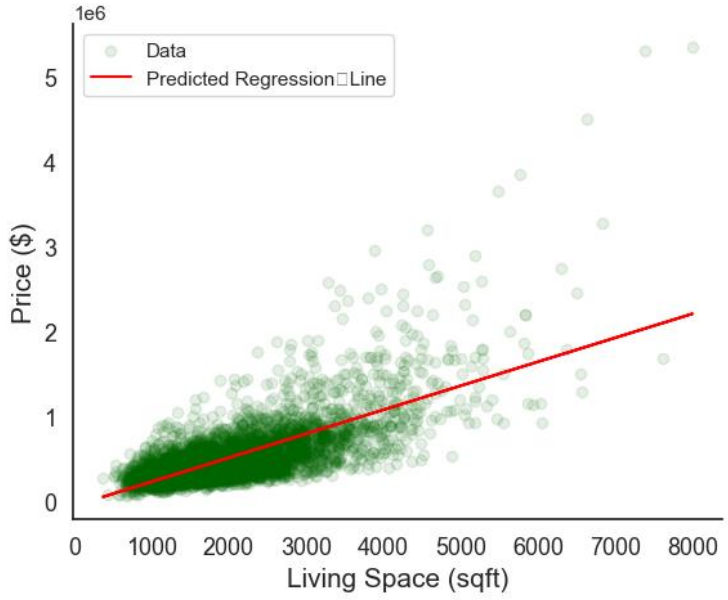

Figure 1. Living Space vs. Predicted Price (Picture credit :Original)

We don't have a lot of features, and this is not a very huge dataset. As a consequence, we can map the majority of them and derive some insightful analytical conclusions. It is best practice to create charts and analyze the data before using a model, as we could find any potential outliers or choose to normalize the data.

We plotted the price against a few variables, and it appears that the price and these features do not have a perfectly linear connection (Figure 2). Nevertheless, what is the nature of their interpersonal relationships? To depict this, we employed three-dimensional visualizations. Moreover, the color scheme was adjusted to a vibrant green hue. Regions of dark green denote high density, whereas the overlapping of numerous light green points contributes to a darkened appearance.

The visual representations presented below illustrate the concurrent increase of sqrt_lot, bedrooms, and bathrooms/bedrooms with sqrt_living. The interconnection between floors, bedrooms, and bathrooms/bedrooms or sqrt_living exhibits distinct characteristics (Figure 3).

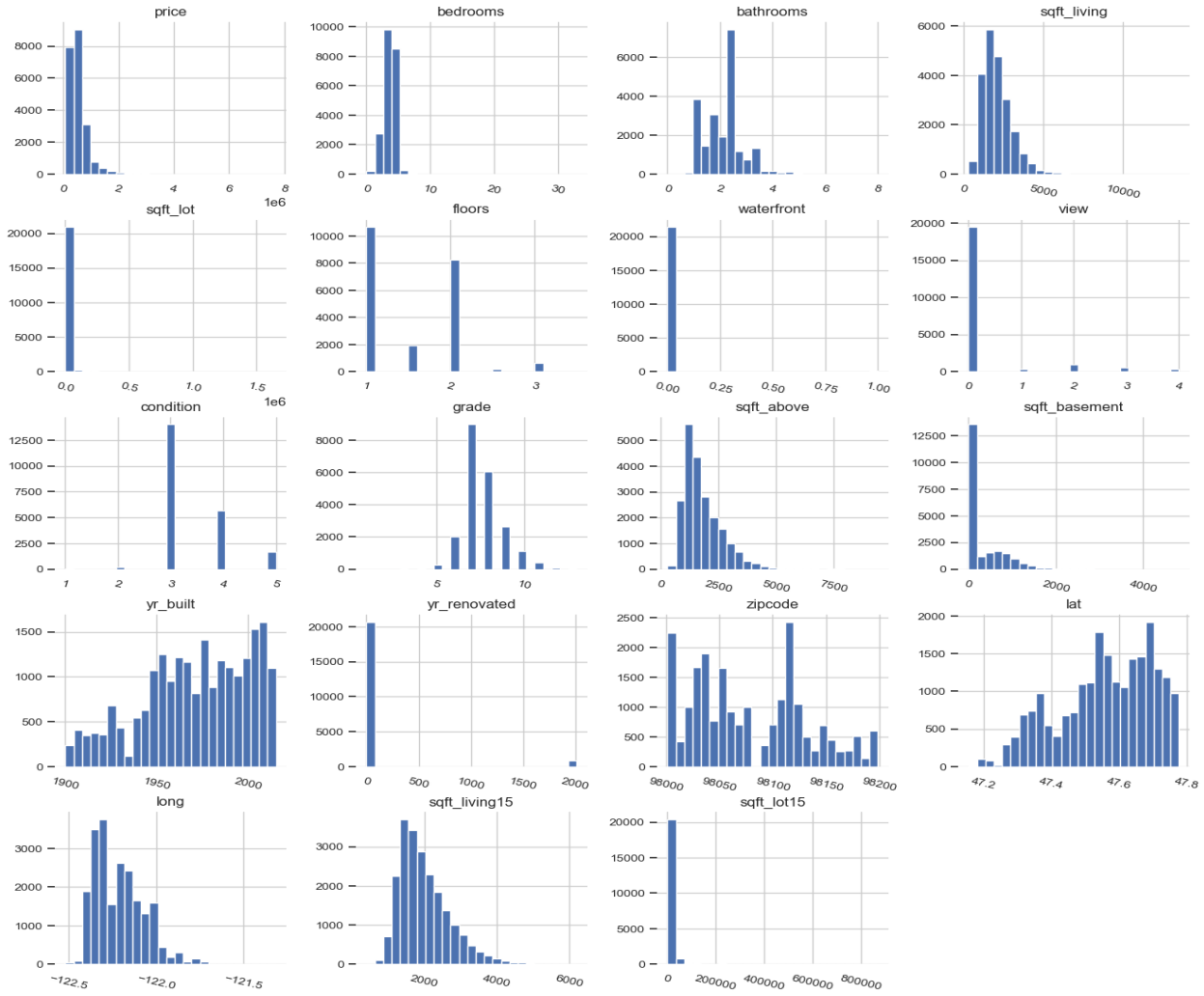

Figure 2. Price against a few variables (Picture credit :Original)

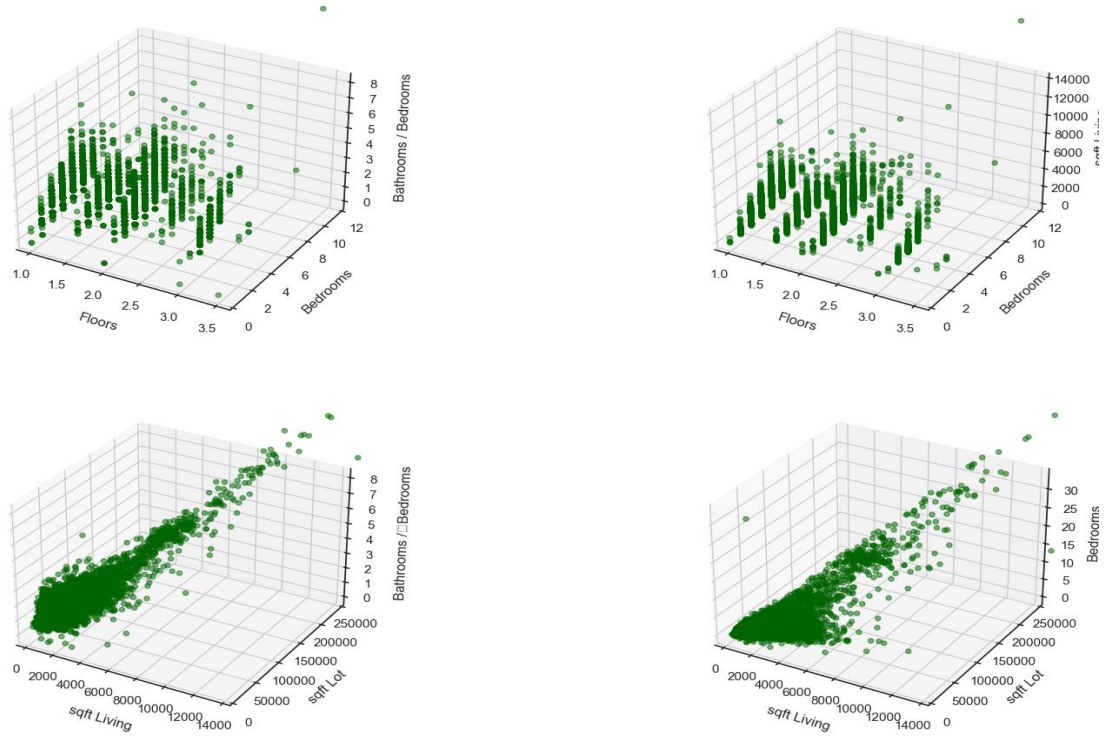

Figure 3. Concurrent increase of sqrt_lot, bedrooms, and bathrooms/bedrooms with sqrt_living (Picture credit: Original)

The incorporation of an excessive number of features in the model can lead to overfitting and the generation of unreliable predictions when applied to new data. Therefore, it is imperative to eliminate features that do not significantly enhance the model's predictive capacity.

Relevance serves as an additional critical consideration. In certain scenarios, retaining two closely related features may not be the optimal choice. For instance, in cases of overfitting, it may be necessary to exclude features such as sqrt_living or sqrt_above due to their strong correlation. While the relationship between these features may be discernible from the dataset definitions, a correlation matrix assessment is essential to confirm its significance. Nonetheless, the existence of a strong correlation does not mandate the elimination of one of the closely associated traits. As exemplified by the relationship between bathrooms and sqrt_living, despite their strong correlation, it may differ from that observed between sqrt_living and sqrt_above (Figure 4).

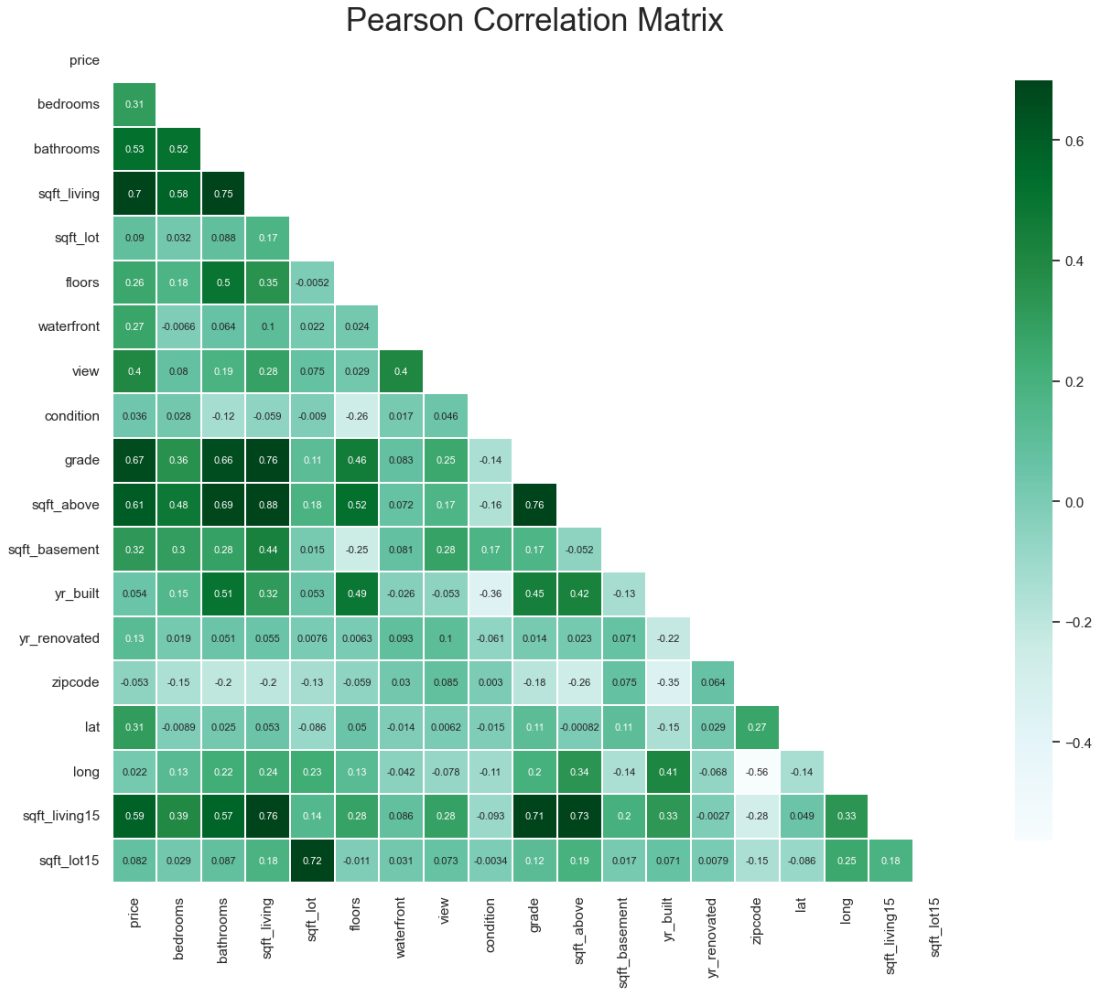

Figure 4. Attribute Correlation Matrix (Picture credit :Original)

With a simple linear regression, we were able to find a poor match. In order to improve this model, we want to include more features. Multiple regression is the term used to describe a linear regression that includes more than one feature. Next, we develop a few intricate models.

This paper expanded the features list to include extra features in addition to the preceding subsection. In addition, I printed the model's coefficients as they were in the preceding part. Based on the evaluation measures, we can see that they have greatly improved. I created a model without any preprocessing that included every attribute so that differences could be readily observed. Assessment metrics proved to be amazing once more. This time, this paper made advantage of the preprocessed data that was collected.

In this project, we used k-NN regression, although k-NN did not provide much information. The algorithm's basic notion is the same as that of the k-NN classification, making it a fairly

straightforward technique. In a nutshell, it employs the k-nearest instances' weighted average, median, or any other desired statistic.

The test set, training set, and different values of k are listed together with the assessment metrics in the accompanying table 1.

Table 1 Model Evaluation Results

The number of K	RMSE	R-squared
15	242834.42	0.562
25	247032.235	0.529
27	247414.263	0.523

5. Conclusion

In conclusion, the study on housing price predictions in the Seattle real estate market has significant implications for informing government policies aimed at promoting economic stability and sustainable social development. By leveraging a wealth of data encompassing housing attributes, geographical factors, and historical price trends, researchers have employed linear regression and KNN prediction models to forecast future housing prices accurately.

The linear regression model utilized in this study establishes a linear relationship between independent variables and housing prices, providing valuable insights through optimal line fitting. Meanwhile, the KNN prediction model, leveraging the nearest neighbor algorithm, offers predictive accuracy by identifying the K closest neighbor samples to the target sample. The comparative analysis among these models aims to ascertain the most suitable approach for housing price forecasting in Seattle. The primary goal is to develop a dependable housing price prediction model that empowers local residents to make informed real estate decisions. Notably, the experimental results highlight the superior performance of KNN regression with K=27 and all features considered. The RMSE values were 242834.420, 247032.235, and 247414.263, respectively, while the R^2 values stood at 0.541, 0.525, and 0.523. Additionally, the adjusted R^2 values were found to be 0.537, 0.521, and 0.520, demonstrating the effectiveness of the KNN regression approach in predicting Seattle housing prices.

Reference

[1] M. Monson, "Valuation using hedonic pricing models." Journal of Property Research, 26(1), 75-88 (2009).

[2] C. Zou, "The House Price Prediction Using Machine Learning Algorithm: The Case of Jinan, China." Journal of Real Estate Data Science, 8(2), 123-136 (2023).

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

[4] O.I. Abiodun, A. Jantan, A., Omolara, "State-of-the-art in artificial neural network applications: A survey." Heliyon 4.11, e00938 (2018).

[5] E. Ahmed, M. Moustafa, "House price estimation from visual and textual features." arXiv preprint arXiv: 1609.08399, (2016).

[6] Q. Truong, M. Nguyen, H. Dang, H., B. Mei, "Housing price prediction via improved machine learning techniques." Procedia Computer Science 174, 433-442 (2020).

[7] S Levantesi, G. Piscopo, "The importance of economic variables on London real estate market: A random forest approach." Risks 8.4, 112 (2020).

[8] T.D. Phan, "Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia." 2018 International Conference on Machine Learning and Data Engineering (iCMLDE). 35-42, (2018).

[9] "Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia." Journal of Real Estate Analytics (2020).

[10] S. Putatunda, "PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market." International Journal of Real Estate Technology and Innovation (2021).