IRANIAN ECONOMIC REVIEW

Print ISSN: 1026-6542
Online ISSN: 2588-6096

https://ier.ut.ac.ir/

University of Tehran

House Price Prediction Model Selection Based on Lorenz and Concentration Curves: Empirical Evidence from Tehran House Market

Mohammad Mirbagherijam[a,*]

a. Department of Economics, Shahrood University of Technology, Shahrood, Iran

* Corresponding Author, E-mail: m.mirbagherijam@shahroodut.ac.ir

Article Info

Article Type: Research Article

Article History:
Received: 27 January 2021
Received in revised form:
16 December 2021
Accepted: 24 January 2022
Published online:
03 December 2023

Keywords:
ABC Criteria,
City of Tehran,
Concentration Curves,
House Features,
Housing Price,
Model Selection Metrics,
Price Prediction Models.

JEL Classification:
C51, C52, R30.

ABSTRACT

This study provides a selection of the house price prediction model for the Tehran city based on the area between the Lorenz curve (LC) and the concentration curve (CC) of the forecast price using 206,556 observed transaction data from March 21, 2018, to February 19, 2021. Several different methods such as generalized linear models (GLM) and recursive partitioning and regression trees (RPART), random forests (RF) regression models, and neural network (NN) models for predicting housing prices. We used 90% of all randomly selected data samples to estimate the parameters of pricing models and 10% of the remaining data sets to test the accuracy of the prediction. The results showed that the area between the LC and CC curves (known as the ABC criterion) of reals and forecast prices in the test data sample of the random forest regression model was less than that of other models examined. The comparison of the calculated ABC criteria leads us to conclude that the nonlinear regression, like the RF regression model, provides an accurate prediction of housing prices in Tehran City.

Cite this article: Mirbagherijam, M. (2023). House Price Prediction Model Selection Based on Lorenz and Concentration Curves: Empirical Evidence from Tehran House Market. *Iranian Economic Review, 27*(4), 1285-1306. DOI: 10.22059/IER.2022.85775

Publisher: University of Tehran Press.

DOI: 10.22059/IER.2022.85775

1. Introduction

Modeling and forecasting house prices is an interesting topic for researchers. That could be due to the advantage of an accurate house price forecast for those involved in the housing market. Accordingly, prediction accuracy measured by an association of pairs of actual and predicted prices is the primary concern of models/methodologies for predicting the house price.

Price prediction accuracy has a long history in modeling property prices, so there have been several solutions to improve it. 1- Developing models for forecasting house price and using a modern model/method in forecasting; Developed from simple forms such as the hedonic pricing model by Malpezzi (2002) to novel forms such as machine learning models; For more details on this method, see Truong et al. (2020). 2- a combination of different models and approaches to price prediction; See recent cases by Glennon et al. (2018) and Wei et al. (2020) support forecast combination methods. 3- Incorporate more relevant variables into a model using large data sets. An example of this is the study by Atrianfar et al. (2013), who improved the accuracy of the price prediction of the Tehran housing market by including 81 effect variables in the model. 4- Using different forecasting models and measuring and then comparing the accuracy of price forecasting models and selecting the more accurate model.

The latter solution can be found in studies by Zietz and Traian (2014) as well as Ghorbani and Afgheh (2017) and Mukhlishin et al. (2017) and recently by Dinarzehi and Shahiki Tash (2020). Although the examined models and the criteria used for the measurement accuracy of the models in the cited studies are different and varied.

Measuring the predictive accuracy of each competitive model by an appropriate metric to compare the accuracy of predictors is the real challenge in using the fourth approach to improving predictive accuracy because of the variety of accuracy metrics. Indeed, choosing a powerful criterion/metric to measure prediction accuracy or model performance in order to get feedback to improve the prediction appears to be the main problem in practice. MSE, RMSE, MAE, and MAPE are the most common metrics used in many studies. Therefore, these metrics have been exemplified as the top three performance metrics in a report (Botchkarev, 2019). Surveys such as Denuit et al. (2019) have shown that Concentration Curves (CC) and Lorenz Curves (LC) are powerful tools for evaluating or comparing the performance of different price-prediction models. They showed that the area between two CC and LC curves, represented below by ABC, is a better indicator of the performance of a particular predictor. The prediction error size and the error distribution across predicted samples are both included in the measure.

Although the models for forecasting house prices have expanded and there have been numerous studies in the field, there is still a need for discussion about how to measure and compare the predictive accuracy of models. Some previous studies provided statistical tools to compare the predictive accuracy of two predictions. A well-known example of this is the Diebold-Mariano test (1995), used to compare the performance of two models. This test next had expanded by Mariano and Preve (2012) to a multivariate version of the DM test for multiple models. Mariano and

Preve's experiment might be worthwhile testing the null hypothesis of equal predictive accuracy; however, it does not show which model provides a more accurate prediction.

So far, the predictive accuracy of competitive models not been conclusively compared. None of the previous studies used the ABC criterion to measure and compare the accuracy of competing models for forecasting house prices. This criterion has only been used in the insurance industry to compare the performance of insurance pricing models.

This article attempts to measure the accuracy of predictors of house prices using ABC criteria and then compare them. The aim is to present an accurate model for predicting the housing price city of Tehran. Using Tehran City as a research sample has significant advantages. That is the capital and most populous city of Iran and about 11% of the total share of the Iranian housing market based on the resident population, and more than 43% of the resident households are renters. However, more than 20% of the residential units are currently vacant. Therefore, the results of an accurate price prediction model for tax collection proportional to the property price will be helpful to both policy-makers and other market participants.

Measuring and comparing the accuracy of any residential property predictive model with a great fit metric, such as the ABC measure, will help us correctly select a more accurate predictive model. Several families of house price prediction models had examined. These include generalized linear models (GLM), recursive partitioning and regression trees (RPART), random forest (RF) models, and neural network (NN) models. The parameters of the research model had estimated by the statistical software R. Real estate transaction prices had collected from the Iranian Ministry of Roads and Urban Development website. We used 90% of all randomly selected data samples to estimate the parameters of predicted pricing models and 10% of the remaining data set to test forecast accuracy. According to Denuit et al. (2019), both concentration and Lorenz curves had used simultaneously to measure the performance of the estimated models. The more precise model for house price prediction had selected according to the ABC criterion (area between CC and LC curve) based on the amount of actual and forecast house prices.

The study provides insights into the Tehran housing market as it applied the application of CC and LC carving as model selection metrics. The empirical results of the research contribute to the literature on housing price prediction and model selection metrics.

This work cannot provide a comprehensive overview of the model performance metrics due to practical limitations. In addition, not all house price forecasting models and approaches explored in research, including STSM, PCA, PLS, SPLS approaches, fuzzy logic models, and machine learning models.

The remaining part of the work proceeds as follows: Section 2 gives an overview of the related literature. Section 3 explains the model selection metric. Section 4 has four subsections; the first subsection describes the research variables; the relevance of the selected variables for the house price compared in the second subsection; the third subsection presents the estimation results of the models; the fourth subsection evaluates the performance of the predictive models examined. Section 5 discusses the results. Finally, section 6 concludes the work.

2. Literature Review

In practice, modeling and accurately forecasting housing prices have accompanied several problems. Determining the factors that determine house prices and quantifying the influence of each or group of price determinants variable/feature are two factors in modeling and accurately forecasting house prices. Additionally, measuring the accuracy of forecasts to get feedback to improve forecasting seemed like another topic to consider here.

In general, the nature of housing assets made it very difficult to model and predict housing prices. House is an immovable multi-purpose property traded for residential and investment purposes, so various characteristics and variables affect its price. The determinants of house price vary from changing factors such as macroeconomic variables to fixed attributes such as the area and skeleton type of a building. Hong et al. (2020) classified the housing price factors into four categories: housing structural attributes (Construction year, area, floor level, and type of heating system), neighborhood attributes (Apartment brand, available units in the building, number of buildings in the apartment complex, parking lot, floor area ratio, building coverage ratio, and the top/lowest floor of the building), location attributes[1] (such as accessibility to nearby facilities), and macroeconomic variables factors. Inherent characteristics of a transacted house, such as construction year, area, floor level, and type of heating system, show the structural attributes. Apartment brand, available units in the building, number of buildings in the apartment complex, parking lot, floor area ratio, building coverage ratio, and the top/lowest floor of the building are known as neighborhood attributes. Latitude, longitude, and accessibility to nearby facilities are related to the geographical position of the house and categorized as the locational attributes. Factors that determine property prices can separate into demand-side factors and supply-side factors. Domestic disposable income, household net financial wealth, interest rates, demographic trends (such as population size and structural change) are fundamental demand-side factors identified by Geng (2018). Housing prices are

[1]. Latitude, longitude, and accessibility to nearby facilities are related to the geographical position of the house and categorized as the locational attributes.

positively dependent on the disposable income and wealth of households and demographic needs, and it negatively depends on housing construction and usage costs. The cost of using the apartment itself will be affected by interest rates, taxes, and expected capital gains Kulikauskas (2017). Land availability and land costs, construction costs, and investments in housing to increase the housing stock are the main factors influencing housing prices on the supply side. Structural, political, and institutional factors such as the regulation of land, buildings, mortgages, and rental and home loans as invisible components such as structural changes in the markets and changing market participant preferences can also influence the dynamics of housing prices.

The relative importance of the relevant housing price factors are not the same and may have changed over time as this could have changed from place to place. This is due to various causes, including different economic conditions, changes in government policy, changes in the building technology of the building, and changes in the spatial determinants of the price of houses due to the lack of uniform development of cities and the increasing density of cities. It implies why the modeling and accurate prediction of house prices has remained an open object in the literature. In addition to the change in the relative importance of house price determinants in different spatial and temporal dimensions, the diversity of house price determinants is another cause. Someone can conveniently categorize this problem into two sub-problems. The first sub-problem relates to variable/characteristic (s) selection relevant to the house price response variable.

One example of this is the current cross-sectional and longitudinal study by Ajija et al. (2021), which shows the size of the influencing coefficient of the demand-side factors, unlike in industrialized countries, has a more pronounced influence than in developing countries.

Regarding the determination of housing price determinants, there is extensive literature on methods for quantifying housing price determinants and for modeling and forecasting housing prices. The models for forecasting house prices have expanded. Therefore, different versions of it can see in the literature. Typical examples are: Hedonic price models (Malpezzi, 2002; Hu et al., 2013; Oladunni et al., 2017), Principal Component Analysis (PCA), Partial Least Squares (PLS), and Sparse PLS (SPLS) approaches (Bork and Mller, 2018), structural time series models (STSM) (Mousavi and Doroodian, 2016), random forest (RF) method (Antipov and Pokryshevskaya, 2012; Čeh et al., 2018; Hong et al., 2020), artificial neural network methods (ANN) (Selim, 2009; Chiarazzo et al., 2014; Fachrurrazi et al., 2017), fuzzy logic (FL) models (Kuan et al., 2010; Sarip et al., 2016) and algorithms of machine learning (ML) (Park and Kwon Bae, 2015; Trawinski et al., 2017; Banerjee and Dutta, 2017; Prez-Rave et al., 2019; Jarosz et al., 2020; Truong et al., 2020).

However, with this growth in the modeling and forecasting of housing prices, concerns about forecasting accuracy grow. As pointed out in the introduction, some studies attempted to solve this issue by comparing the accuracy of the models under consideration. Despite this, due to the variety of predictive accuracy measures, the criteria used to measure the performance of the models are not the same in the references cited. Zietz and Traian utilized the root mean square error (RMSE) criteria to measure the performance of three classes of univariate time series techniques such as ARIMA models, switching regression models, and state-space/structure time series models (STSM). They found that the STSM method gives the most accurate predictions. While Zietz and Traian compared predictive models with only one performance metric, Ghorbani and Afgheh used multiple metrics to measure forecast accuracy. The metrics considered in their research were the coefficient of determination (R2), the mean square error (MSE), the RMSE, the mean absolute percentage error (MAPE), and the mean absolute deviation (MAE), and the Theils inequality coefficient (TIC). They found that the predictive accuracy of the artificial neural network model used to predict property price in Ahvaz city is higher than that of the hedonic model. Mukhlishin et al. use fuzzy logic, an artificial neural network, and the K-Nearest Neighbor to predict the selling price of a house. By applying the MAPE metric to compare the performance of the models under consideration. They show that the fuzzy method is superior to neural networks as the k-nearest neighbor for house price prediction in restricted data training. Dinarzehi and Shahiki Tash examined the Black-Scholes model, the Merton model and the geometric Brownian motion model for price jump diffusion in the real estate market. By comparing the maximum likelihood statistics, they show that the geometric Brownian model with stochastic NGARCH-based fluctuations has more explanatory power than the Merton model and than the geometric Brownian model with constant-based fluctuations. Overall, these studies underscore the importance of accuracy metrics in choosing predictor models.

As discussed above, performance measures of the model are varied and numerous. In 1995, Makridakis and Hibon stated that fourteen accuracy measures existed in the forecasting literature. Based on the structures of performance metrics, Botchkarev (2019) divided them into four categories: primary metrics, advanced metrics, composite metrics, and hybrid metrics. By studying the properties and typology of performance metrics, he found that the method of determining point distance, the method of normalization, and the method of aggregating point distances over a data set are the three (3) key components (dimensions) that are structure and properties determine the primary metrics. Research into predictive accuracy metrics and the selection of appropriate metrics can continue. When asked which measure is better, Denuit et al. (2019) showed in a study that the ABC criterion is better. This performance metric includes both the amount of error and its frequency in measuring the accuracy of the prediction.

3. Model Selection Metrics

In order to select the more accurate prediction model, the accuracy of several competing house price prediction models was measured and compared according to ABC criteria. ABC is used by Denuit et al. (2019) and refers to the area between CC and LC.

Assume that y_i^a and y_i^p are the real and predicted price of house price transacted in the ith house transaction. Predicted price y_i^p, obtained by the house price prediction model π(x), i.e $y_i^p = \pi(x_i)$, and all house details and explanatory variables of the house price were in the vector x=x(x$_1$, x$_2$,…,x$_k$). The predictive model of π(.) is unknown, and we have assumed that there are alternative predictive models of $\pi^1, \pi^2, \dots, \pi^m$ that someone used to predict the house price. To decide which model is better than another, the ABC index of each model is calculated in the following steps: First, house prices are predicted using several different models. It is worth noting that before this step, the coefficients of the assumed models are estimated with the available data using the R software. Then the Lorenz curve and the concentration curve of the prediction results of each model are plotted simultaneously in a graph and the area between these two is calculated. Finally, the calculated ABC indices are compared and the model with the lowest value of the ABC index is selected as a suitable model. The package of IC2 in R was used to estimate and display the CC and LC curves. For a better explanation of the calculation method, the definition of concentration and Lorenz curves is based on Denuit et al. (2019).

3.1 The Concentration and Lorenz Curve

For each probability level α, the concentration curve of the real price Y with respect to the predicted price $\pi(X)$ based on the information in the vector x is defined as follows[1].

$$CC[Y, \pi(X); \alpha] = \frac{E[Y|[\pi(X) \leq F_\pi^{-1}(\alpha)]]}{E[Y]} \tag{1}$$

where $F_\pi(t)$ is the distribution function of the predicted price ($\pi(X)$), and F_π^{-1} is that with the quantile function defined as the generalized inverse of F_π, i.e.

[1]. Assuming the samples (y_i^a, y_i^p), i=1,…,n, to be independent and identically distributed, the empirical concentration curve and Lorenze curve of the real price can be estimated as follows:

$$\widehat{CC}[\boldsymbol{Y}, \pi(\boldsymbol{X}); \alpha] = \frac{1}{n\bar{Y}} \sum_{\boldsymbol{i}|[\hat{\pi}(\boldsymbol{X_i}) \leq \hat{F}_\pi^{-1}(\alpha)]} Y_i = \frac{\sum_{\boldsymbol{i}|[\hat{\pi}(\boldsymbol{X_i}) \leq \hat{F}_\pi^{-1}(\alpha)]} Y_i}{\sum_{i=1}^n Y_i}$$

$$\widehat{LC}[\pi(\boldsymbol{X}); \alpha] = \frac{\sum_{\boldsymbol{i}|[\hat{\pi}(\boldsymbol{X_i}) \leq \hat{F}_\pi^{-1}(\alpha)]} \hat{\pi}(\boldsymbol{X_i})}{\sum_{i=1}^n \hat{\pi}(\boldsymbol{X_i})}$$

$F_{\pi}^{-1}(\alpha) = \inf\{t|F_{\pi}(t) \geq \alpha\}$ for a probability level α. Equation 1 can be interpreted as the proportion of real price observations to which Y is attributable to a subset of predicted price observations as a percentage of the lowest forecast transactions price.

The Lorenze curve LC associated with the predicted price $\pi(X)$ is as Equation 2:

$$LC[\pi(X);\alpha] = CC[\pi(X),\pi(X);\alpha] = \frac{E[\pi(X)|[\pi(X) \leq F_{\pi}^{-1}(\alpha)]]}{E[Y]} \qquad (2)$$

If the predicted price is the same as the actual price, there is no need to distinguish CC from LC. This is because if $Y = \pi(X)$, then $LC[\pi(X);\alpha] = CC[Y,\pi(X);\alpha]$.

3.2. Properties of CC and LC

According to Denuit et al. (2019), CC and LC curves have several specific properties. The CC and LC curves are non-decreasing (or monotone) and convex functions. The monotony of CC curve satisfies $\lim_{\alpha\to 0} CC[\pi(X);\alpha] = 0$ and $\lim_{\alpha\to 1} CC[\pi(X);\alpha] = 1$.

The concentration curve is the copula of pairs $(Y,\pi(X))$. In addition, the area between 45-degree line and CC measure the dependence on variables. When two variables are independent of each other, the concentration curve is the 45-degree line. This line is referred to as an independent line in the literature. Because If two variables Y and $\pi(X)$ are independent of each other, then $CC[Y,\pi(X);\alpha] = \frac{E[Y]P[\pi(X) \leq F_{\pi}^{-1}(\alpha)]}{E[Y]} = \alpha$.

The positive dependenc on variables results in a convex concentration curve. Conversely, if the CC is convex, then the two variables are positively dependent. Nevertheless, the predictor $\pi_1(X_1)$ is more discriminatory than $\pi_2(X_2)$ for response Y if and only if the following inequality exist for all levels of α.

$$CC[Y,\pi_1(X_1);\alpha] \leq CC[Y,\pi_2(X_2);\alpha] \qquad (3)$$

In other words, the CC curve of predictor π_2 is below the CC curve of predictor π_1. If the respective concentration or Lorenz curves of two predictors intersect, the ICC index is used instead of the CC index to compare the distinctive power of two predictors (for further details, see Denuit et al., 2019).

Since the Lorenz curve is a special case of the concentration curve, it has its own special properties in addition to the properties of the concentration curve. LC is derived by dividing the cumulative value of the variable by its expected value. That is related to the Gini's mean difference (GMD)[1] and the Gini coefficient. The area ratio is between 45-degree line (line of equality or identity) and the LC over the

[1]. As explained by Yitzhaki and Schechtman (2012), GMD has more than 14 alternative representations. The most convenient presentation of the GMD to be used is the covariance presentation, i.e. $E[|X_1 - X_2|] = 4Cov[X,F(X)]$.

entire region below the line of equality, known as the Gini coefficient. It can be shown that this region is equal to $2Cov[\pi(X), F_\pi(\pi(X))]$.

3.3 Calculation of ABC Indicator

The ABC indicator is given by Equation 4:

$$\begin{aligned} ABC[\pi(X)] &= \int_0^1 (CC[Y, \pi(X); \alpha] - LC[\pi(X), \pi\alpha])\, d\alpha \\ &= \frac{1}{E[\pi(X)]} \int_0^1 (E[Y|[\Pi \leq \alpha]] - E[\pi(X)|[\Pi \leq \alpha]])\, d\alpha \\ &= \frac{1}{E[\pi(X)]} \int_0^1 \int_0^\infty (P[\pi(X) \leq y, \Pi \leq \alpha] \\ &- P[Y \leq y, \Pi \leq \alpha])\, dyd\alpha \\ &= \frac{1}{E[\pi(X)]} (cov[\pi(X), \Pi] - cov[Y, \Pi]) \end{aligned} \tag{4}$$

We use Equation 4 as a powerful model selection metric to make decisions which model is better than another.

4. Data and Estimation Results

4.1 The Descriptive Statistics of Research Variables

The raw data used in research includes both registered transaction information and macroeconomic variables. Table 1 shows the source and description of research variables. The data sample comprised 206,556 observed transaction data in the period from March 21, 2018 to February 19, 2020.

Table 1. Research Variables and The Relevant Data Sources

Variables (Unit)	Description	Data Type	Source
Price (thousand IRR)	House price per square meter		http://www.mrud.ir
Regional	Regional municipality		
Area	Area (square meter)		
Age	Building age (years)		
Skeleton	Skeleton type: concrete, metal, brick or cement block, concrete and metal, skeletonless, clay, wooden		
Dollar	Closed price of 1$ per IRR	Daily	https://www.tgju.org
Euro	Closed price of 1€ per IRR	Daily	
Emami coin	Closed price of 1Gold Emami coin per IRR	Daily	
TSE	Total price index of Tehran stock exchange	Daily	https://tse.ir
Land price (thousand IRR)	The average sale price of one square meter of land or residential building land	Quarterly	http://www.mrud.ir
Rent (IRR)	Average monthly rent plus 3% of the deposit payment on rent of 1 sq.m.	Quarterly	
CPI	Urban consumer price index (2016=100)	Monthly	https://www.amar.org.ir
Materials price	Building materials price index (2011=100)	Quarterly	http://www.mrud.ir
Age level			
Total price (thousand IRR)	Price * Area		
En Date	Contract registration date		

Source: Research finding.

The number of observed house transactions per municipality region of Tehran is shown in Figure 1. In addition, the average price of transacted buildings, in each region is compared in Figure 2. Figure 2 show how the municipality’s brand affects the house price in Tehran. Figures 3 and 4 show that house details such as age and area vary depending on the municipality number. A possible explanation for this could be that some attributes of a house such as the house area, are influenced by the brand of municipality.

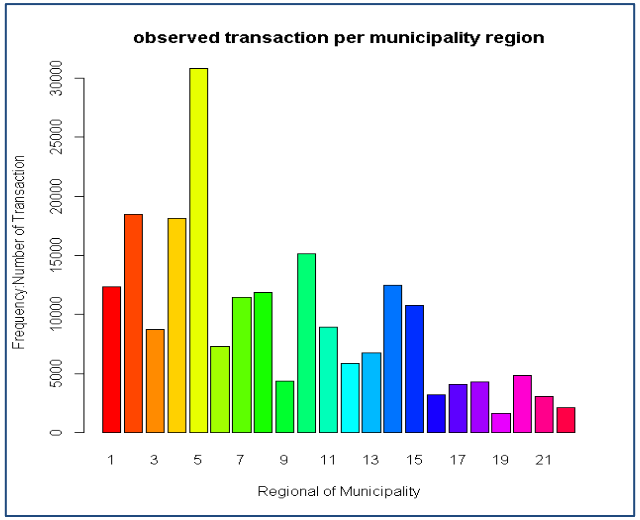

Figure 1. The Number of Transactions Observed per Region
Source: Research finding.

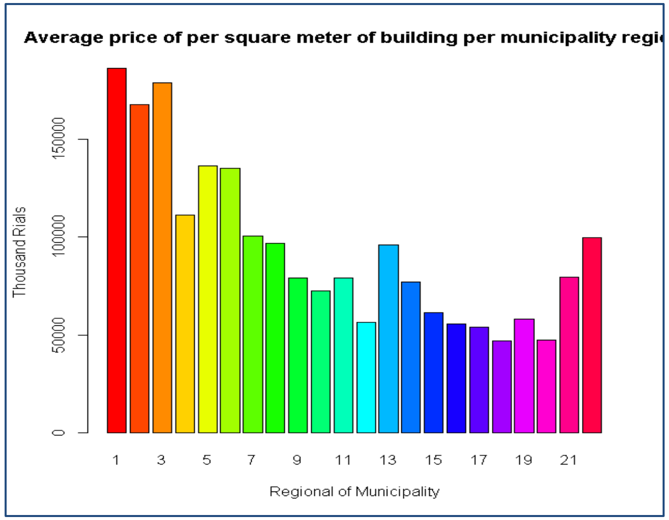

Figure 2. The Comparison of the Average Prices of the Buildings Carried Out per Region
Source: Research finding.

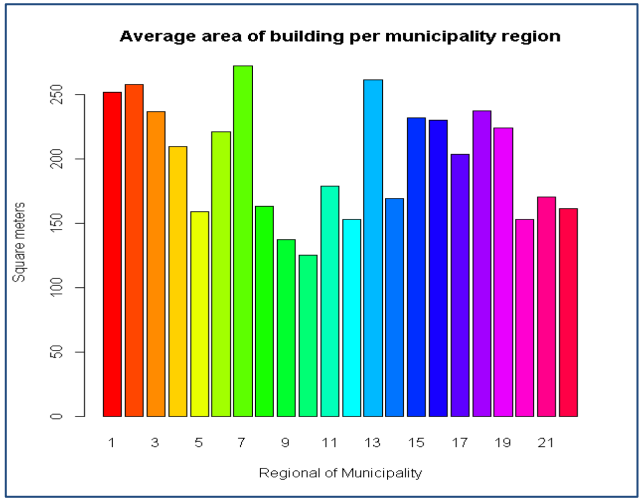

Figure 3. The Comparison of the Average Area of Transacted Buildings per Region
Source: Research finding.

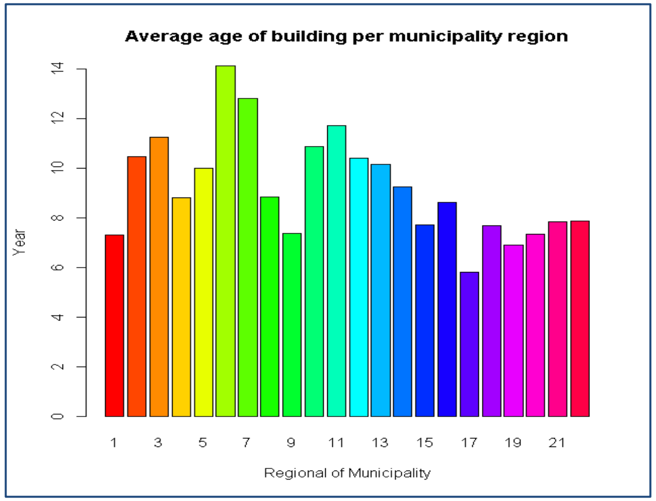

Figure 4. The Comparison of the Average Age of Transacted Buildings per Region
Source: Research finding.

4.2 The Relevance of the Selected Variables to the House Price

To identify the relevant house price factors, and to determine the intensity and direction of each factor on house prices, the strength of the association between the house price and the selected variables is measured (Table 2) by three different types of metrics: a- The Pearson correlation statistic used to study the linear relationship. b- The statistic based on Spearman’s Rho rank used to assess the monotonic relationship (linear or not)[1]. c- The CC index for measuring any type of dependency (whether linear or monotonic or non-monotonic)[2].

According to Denuit et al. (2019), in order to determine the most relevant variables for the house price, the CC index of the house price and the selected variable is calculated first. then the variables are sorted and ordered according to the absolute value of the CC index.

Based on the statistical value of all three types of association metrics in Table 2, it is clear that house prices in Tehran are inversely proportional to regional municipality and age of construction. The relative importance of the variables for house prices is shown in the third column of Table 2. According to the size of the CC index, the influence of the land price variables on the building price is stronger than for other variables, therefore the land price is identified as the most relevant variable for house prices in Tehran.

Table 2. The Values and the Comparison of The Relevance of Selected Variables for The House Price

Selected variable	Concentration curve index	Relevancy rank	Pearson R2	Spearman’s Rho
Land price	0.25840	1	0.1249	0.6862
Rent	0.25192	2	0.0786	0.4686
Regional municipality	-0.18887	3	-0.0922	-0.5272
CPI (urban consumer price index)	0.16289	4	0.0769	0.4618
Materials price index	0.16289	5	-0.0366	-0.1699
Stock price index (TSE)	0.15952	6	0.1265	0.7186
Gold price (Emami coin)	0.13916	7	0.0726	0.4572
Exchange rate $	0.12519	8	0.0580	0.3268
Exchange rate €	0.10962	9	0.0684	0.4038
Building age (age level)	-0.06048	10	-0.0277	-0.1690
Area	0.05456	11	0.0657	0.3665
Skeleton type	0.01496	12	-0.0104	0.3827

Source: Research finding.

[1]. That is equal to the Pearson correlation between the rank values of the two variables.

[2]. As mentioned in Section 2, the CC index is the copula of one variable (here house price) and the rank of another variable (such a house feature).

The concentration curves of the house price in relation to each variable are shown in Figure5. From Figure 1 it can be seen that there is a significant difference in the concentration curve of the determining variables for the housing price. The CC curve of some variables such as the regional municipality and the age of the building is completely above the 45-degree line due to their negative influence on the house price. Some variables such as the price of land and rent are lower due to their positive effects. However, some variables can have a threshold effect on the house price. For example, the area of the house has a threshold effect, so its CC curve intersects the 45 degree line and is above and below the line.

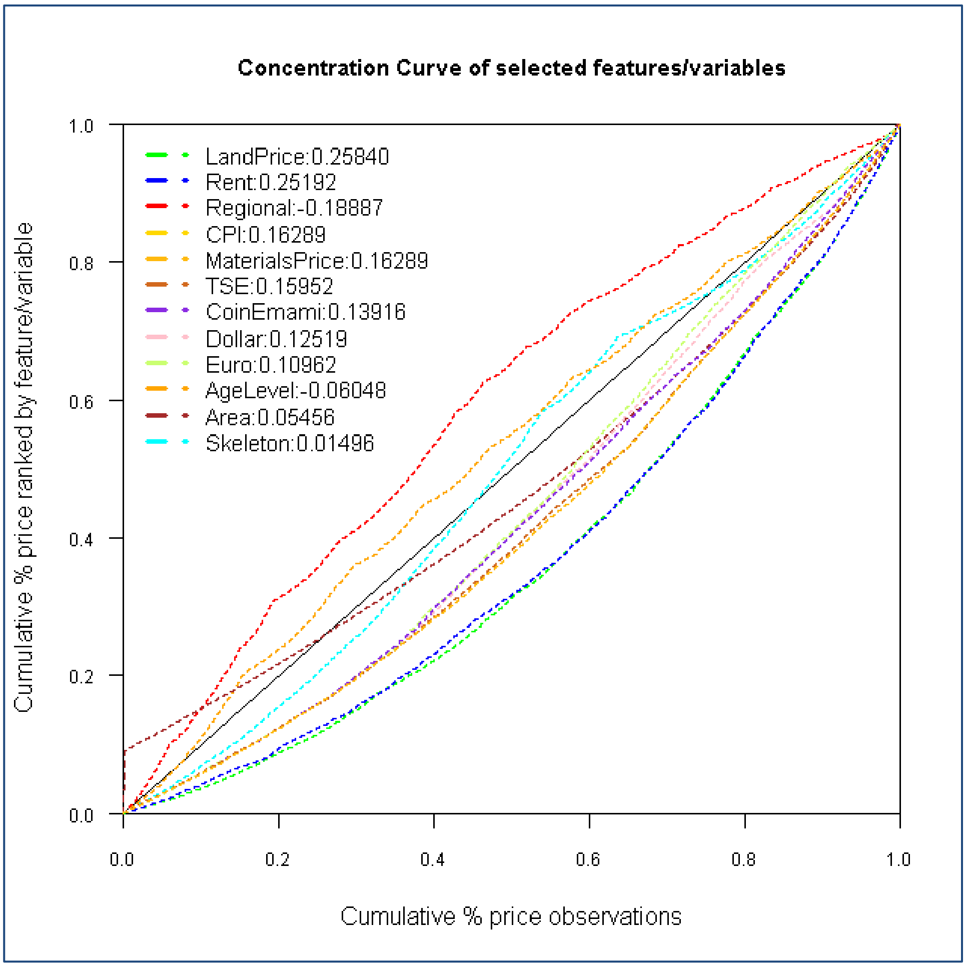

Figure 5. The Comparison of Concentration Curve Of House Price To The Selected Variables

Source: Research finding.

4.3 Modeling the House Price Prediction

Any house price prediction model should take into account the relevant determining variables of the house price in the model. Therefore, here we set up the basic equation of house price prediction models based on the hedonic price equation as follows:

$$\begin{aligned} House\ price = f(&Structural\ attributes, \\ &Neighborhood\ attributes, Locational\ attributes, \\ &Macroeconomic\ variables\ and\ Other\ factors) \end{aligned} \tag{5}$$

It only allows us to include the residential area, the age and the type of skeleton as structural information and the number of the municipality as house features. However, in order to capture the influence of building construction costs on the purchase price of a home, variables such as building material price index and property price are included in the forecast models. In addition, the apartment rent and the prices of foreign currencies (dollars and euros) as well as the stock market index and gold prices are included in order to record the effects of the optimal portfolio of investors in the competitive housing market on the house price. Because of the research objectives, the above basic house price equation is estimated by the four different approaches below.

1. The GLM with two subfamilies, i.e. Gaussian GLM regression model and Poisson-GLM regression model;
2. RPART with two rules for dividing the ANOVA and Poisson models, i.e. rpart.anova and rpart.poission;
3. RF regression models;
4. NN model with three kinds of hidden configurations (2,1), (3,1) and (3,2) i.e. NN.21, NN.31 and NN.32 respectively.

As mentioned in Section 1, the research data sample is randomly split into two sub-samples, namely learning samples and testing samples. These subsamples cover 90% and 10%, respectively, of all data samples that are used for estimation or prediction purposes[1]. In order to predict the house price, the following two-step process is carried out in each method approach:

1. Model building and its estimation using the learning sample data;
2. Price forecast based on the estimated model using test pattern data.

A summary of the estimated outputs of models is presented in the supplementary Excel file. To compare the models used in modeling and forecasting property prices, Tables 3 and 4 summarize the descriptive statistical results of real prices with the adjusted (or predicted) prices of each model based on both learning sample and test sample data sets based. As indicated in Tables 3 and 4, the mean value of the real price is smaller than its median value, and therefore the distribution of the

[1]. It is noteworthy that in the estimation process of all models except neural network models, apart from building age, skeleton type, and regional municipality, other research variables are used in the natural logarithms form. However, for the neural network models, research variables are used in the normalization form which is normalized by max-min normalization technique.

observed house prices is asymmetrical with negative (left) skewness. However, according to the mean and median values, it appears that some models such as Gaussian GLM and NN.32 provide a positive (right) skew prediction. The minimum value and the 1st quartile value of the adjusted / predicted price are higher than the minimum value and the 1st quartile of the real price, while the maximum value of the adjusted / predicted prices is smaller than the maximum value of the real price.

The results in Tables 3 and 4 show that there is a significant difference between the statistical parameters of adjusted / predicted prices and the statistical parameters of real prices in both learning and testing samples.

Table 3. The Comparison of the Descriptive Statistics of Actual and Fitted Price Value of by Learning Sample

Statistical parameters		Min	1st Quantile	Median	Mean	3rd Quantile	Max
Real price		0.501	10.866	11.347	11.267	11.773	17.814
Fitted value by:	**Poisson GLM**	7.804	10.910	11.267	11.267	11.651	14.476
	Gaussian GLM	7.261	10.919	11.279	11.267	11.654	14.031
	tree.anova	10.440	10.980	11.270	11.270	11.620	12.050
	tree.poisson	10.440	10.980	11.270	11.270	11.620	12.050
	RF	4.138	10.909	11.308	11.266	11.662	13.818
	NN.21	11.616	11.616	11.616	11.616	11.616	11.616
	NN.31	11.605	11.605	11.605	11.605	11.605	11.605
	NN.32	11.172	11.238	11.395	11.549	11.838	12.446

Source: Research finding.

Table 4. The Comparison of the Descriptive Statistics of Actual And Predicted Value of ln Price by "Test-Sample"

Statistical parameters		Min	1st Quantile	Median	Mean	3rd Quantile	Max
Real price		0.621	10.866	11.350	11.269	11.775	17.577
Predicted value by:	**Poisson GLM**	7.972	10.911	11.267	11.268	11.651	14.143
	Gaussian GLM	7.498	10.919	11.280	11.268	11.654	13.766
	tree.anova	10.440	10.980	11.270	11.270	11.620	12.050
	tree.poisson	10.440	10.980	11.270	11.270	11.620	12.050
	RF	4.494	10.910	11.309	11.267	11.659	13.843
	NN.21	11.616	11.616	11.616	11.616	11.616	11.616
	NN.31	11.605	11.605	11.605	11.605	11.605	11.605
	NN.32	11.172	11.237	11.391	11.548	11.836	12.445

Source: Research finding.

4.4 Evaluation of the Models

To evaluate the performance of models, the accuracy of each price prediction model is measured using ABC criteria[1]. Table 5 illustrates the estimated LC, CC, and ABC indices of the forecast price[2]. It also indicates the rank of the models studied based on the accuracy of the prediction. In addition, Figure 6 shows the concentration curves of the forecast price of all models in the test sample. In Figure 6, the Lorenz curve of the real price is drawn in black, but the concentration curves are drawn in color.

Table 5. The Comparison of the Models' Performance

Models	LC	CC	ABC	Accuracy rank
Poisson GLM	0.391548	0.291207	0.100341	3
Gaussian GLM	0.391548	0.292435	0.099113	2
rpart.anova	0.391548	0.276273	0.115275	5
rpart.poisson	0.391548	0.276273	0.115275	6
random.forest	0.391548	0.325611	0.065937	1
NN.21	0.391548	-0.168691	0.560239	8
NN.31	0.391548	-0.098172	0.489720	7
NN.32	0.391548	0.279327	0.112221	4

Source: Research finding.

As Table 5 shows, the forecast of some NN models such as NN.21 and NN32 is inversely proportional to real price data. Therefore, the estimated CC index for these models is a negative number. As a consequence, the corresponding concentration curve of these models in Figure 6 lies above the 45-degree line. Logically, the positive relationship between forecast prices and real prices is a necessity for choosing the appropriate price forecasting model. Therefore, two models of the artificial neural network family with hidden configuration (2,1) and (3,2) are omitted, and the more accurate predictive model is selected from among other models based on ABC criteria. Both Table 5 and Figure 6 show that the accuracy of the models in predicting house price, as measured by ABC criteria, is not the same and there is a significant difference between them. Interestingly, the lowest value of the calculated ABC index is related to the Random Forest model because its associated concentration curve is closer to the Lorenz curve. These results confirm that the choice of the random forest model as an eminently suitable model is a good choice for house price prediction.

[1]. In addition, the results of the multivariate Diebold-Mariano test show that the null hypothesis was rejected at a confidence level of 99%, so by accepting the alternative hypothesis, we conclude that equal predictive accuracy does not hold.

[2]. The estimated LC and CC indices of predicted price are used to calculate the ABC index.

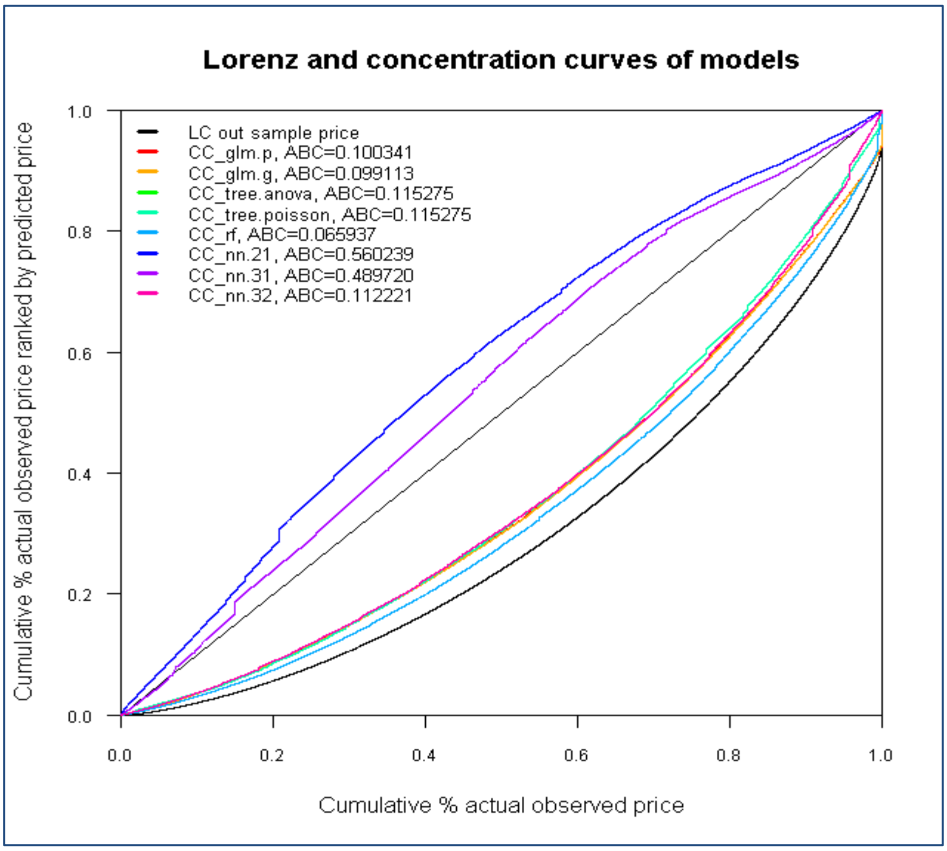

Figure 6. The Comparison of the Concentration Curve Of Predicted Price of All Models

Source: Research finding.

5. Discussing Finding

Various models for forecasting residential property prices have been found in the literature. In addition, various metrics have been used to assess the performance of predictive models in the literature. The present study was designed to use a powerful model selection metric developed by Denuit et al. (2019) to choose the more accurate predictive model among several competing house price prediction models. We found that the accuracy of the random forest method in house price prediction was higher than that of other models examined. In addition, the GLM offers a better forecast. It is somewhat surprising that the random forest technique allows for a more accurate prediction of the house price. This finding is consistent with that of Antipov and Pokryshevskaya (2012) who suggested using random forest models for tasks with missing values and multi-level categorical variables[1].

[1]. They used both coefficients of dispersion and MAPE indicators to compare the accuracy of different methods.

In addition, this agrees with the latest findings from Čeh et al. (2018) and Hong et al. (2020). This result can be explained by the fact that the house price is influenced by several categorical variables.

Another important result of the present work was the relevance of the variables that determine the house price for the house price. Based on the CC index measure, the most positive relevance is firstly for the price of the property and secondly for the house rent. As expected, the house price is negatively related to the municipality and the age of the building. These results agree with those of previous studies (Sabbagh Kermani et al., 2010; Mohammadian Mosammam, 2015).

In addition, the results obtained from Figure 5 showed that some variables such as the residential region had threshold effects on the house price. This result can be explained by the fact that the purchasing power of households when buying a house was clearly unequal. Due to the low purchasing power of most of Tehran's residents, the number of small house buyers is higher than that of large houses.

These results can help policymakers to adopt appropriate tax policies in the housing market. More work is needed to determine the threshold effect of house price determinants.

6. Conclusion

This article has evaluated several competitive house price prediction models and highlighted the importance of proper model selection for accurate house price prediction. In addition, we assessed the relevance of the house price determinant variables for the house price in Tehran. The research results were provided in order to select an eminently suitable model for the housing price forecast and to determine the main determinant of the housing price. In the current study, the ABC criterion was used to measure the accuracy of the examined predictors. It is not specifically designed to measure and compare model accuracy prediction with other performance metrics. Since other price prediction models such as fuzzy logic and STSM models were not addressed.

Research Highlights

✓ Relevance of the housing price determinant measured and ranked by concentration curve index.

✓ The more accurate price prediction model among several competitive considered models determined for Tehran housing market.

✓ The research results lead us to select an excellently suitable method/model for the housing price forecast and determine the determinant of the housing price.

References

Ajija, S., Pratiwi, I., & Wasiaturrahma, W. (2021). How to Control the House Prices Through the Demand Sides? *Iranian Economic Review*, *27*(1), 1-15.

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a CART-Based Approach for Model Diagnostics. *Expert Systems with Applications, 39*(2), 1772-1778.

Atrianfar, H., Barakchian, S., & Fatemi, S. (2013). Evaluation of Forecast Combination Methods. *Journal of Applied Economics Studies in Iran, 2*(6), 123-138.

Banerjee, D., & Dutta, S. (2017). Predicting the Housing Price Direction Using Machine Learning Techniques. *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)* (2998-3000). Retrieved from https://ieeexplore.ieee.org/abstract/document/8392275

Bork, L., & Møller, S. V. (2018). Housing Price Forecastability: A Factor Analysis. *Real Estate Economics, 46*(3), 582-611.

Botchkarev, A. (2019). A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management, 14,* 45-76.

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS International Journal of Geo-Information, 7*(5), 168-184.

Chiarazzo, V., Caggiani, L., Marinelli, M., & Ottomanelli, M. (2014). A Neural Network Based Model for Real Estate Price Estimation Considering Environmental Quality of Property Location. *Transportation Research Procedia, 3,* 810-817.

Denuit, M., Sznajder, D., & Trufin, J. (2019). Model Selection Based on Lorenz and Concentration Curves, Gini Indices and Convex Order. *Insurance: Mathematics and Economics, 89,* 128-139.

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics, 13*(3), 253-263.

Dinarzehi, K., & Shahiki Tash, M. (2020). Price Jump Diffusion in Iranian Housing Market (Merton Model and NGARCH Approach). *Iranian Economic Review, 26*(2), 369-388.

Fachrurrazi, Husin, S., Tripoli, & Mubarak (2017). Neural Network for the Standard Unit Price of the Building Area. *Procedia Engineering, 171*, 282-293.

Geng, M. N. (2018). Fundamental Drivers of House Prices in Advanced Economies. *International Monetary Fund, WP/18/164*, 1-24.

Ghorbani, S., & Afgheh, S. M. (2017). Forecasting the House Price for Ahvaz City: the Comparison of the Hedonic and Artificial Neural Network Models. *Journal of Urban Economics and Management, 5*(19), 29-44.

Glennon, D., Kiefer, H., & Mayock, T. (2018). Measurement Error in Residential Property Valuation: An Application of Forecast Combination. *Journal of Housing Economics, 41*, 1-29.

Hong, J., Choi, H., & Kim, W. S. (2020). A House Price Valuation Based on the Random Forest Approach: the Mass Appraisal of Residential Property in South Korea. *International Journal of Strategic Property Management, 24*(3), 140-152.

Hu, G., Wang, J., & Feng, W. (2013). Multivariate Regression Modeling for Home Value Estimates with Evaluation Using Maximum Information Coefficient. *Studies in Computational Intelligence, 443*, 69-81.

Jarosz, M., Kutrzyński, M., Lasota, T., Piwowarczyk, M., Telec, Z., & Trawiński, B. (2020). Machine Learning Models for Real Estate Appraisal Constructed Using Spline Trend Functions. In *Asian Conference on Intelligent Information and Database Systems* (636-648). Cham: Springer.

Kulikauskas, D. (2017). The User Cost of Housing in the Baltic States. *Journal of European Real Estate Research, 10*(1), 17-34.

Kuşan, H., Aytekin, O., & Özdemir, I. (2010). The Use of Fuzzy Logic in Predicting House Selling Price. *Expert Systems with Applications, 37*(3), 1808-1813.

Makridakis, S., & Hibon, M. (1995). Evaluating Accuracy (or Error) Measures. *ECONIS Working Paper*, Retrieved from http://www.opengrey.eu/item/display/10068/56650

Malpezzi, S. (2002). Hedonic Pricing Models: A Selective and Applied Review (67-89). In *Housing Economics and Public Policy*. New York: John Wiley & Sons, Ltd.

Mariano, R. S., & Preve, D. (2012). Statistical Tests for Multiple Forecast Comparison. *Journal of econometrics, 169*(1), 123-130.

Mohammadian Mosammam, A., & Abbasi, M. (2015). House Price Analysis of Tehran City Using Generalized Additive Models. *Ijoss Iranian Journal of Official Statistics Studies*, *25*(2), 161-174.

Mousavi, M., & Doroodian, H. (2016). Analyzing the Determinants of Housing Prices in Tehran City. *Quarterly Journal of Economic Modeling*, *9*(3), 103-127.

Mukhlishin, M. F., Saputra, R., & Wibowo, A. (2017, November). Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor. 2017 *1st International Conference on Informatics and Computational Sciences (ICICoS)*, Retrieved from https://ieeexplore.ieee.org/abstract/document/8276357

Oladunni, T., Sharma, S., & Tiwang, R. (2017). A spatio-temporal hedonic house regression model. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Retrieved from https://ieeexplore.ieee.org/abstract/document/8260698

Park, B., & Kwon Bae, J. (2015). Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data. *Expert Systems with Applications*, *42*(6), 2928-2934.

Pérez-Rave, J. I., Correa-Morales, J. C., & González-Echavarría, F. (2019). A Machine Learning Approach to Big Data Regression Analysis of Real Estate Prices for Inferential and Predictive Purposes. *Journal of Property Research*, *36*(1), 59-96.

Sabbagh Kermani, M., Ahmadzadeh, K., & Musavi Nik, S. (2010). Determinants of House Price Using Causality Relations Approach in Vector Error Correction Model: Case Study Tehran. *Economic Research*, *10*(37), 267-293.

Sarip, A. G., Hafez, M. B., & Nasir Daud, M. (2016). Application of Fuzzy Regression Model for Real Estate Price Prediction. *Malaysian Journal of Computer Science*, *29*(1), 15-27.

Selim, H. (2009). Deter minants of House Prices in Turkey: Hedonic Regression Versus Artificial Neural Network. *Expert Systems with Applications*, *36*(2 Part 2), 2843–2852.

Trawiński, B., Telec, Z., Krasnoborski, J., Piwowarczyk, M., Talaga, M., Lasota, T., & Sawiłow, E. (2017). Comparison of Expert Algorithms with Machine Learning Models for Real Estate Appraisal. *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, Retrieved from https://www.semanticscholar.org/paper/Comparison-of-expert-algorithms-with-machine-models-Trawinski-Telec/a29a1f38f5ce13fb0f36d6de7375aaecd1d8b266

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science, 174*, 433-442.

Wei, C., Du, C., & Zheng, N. (2020). A Changing Weights Spatial Forecast Combination Approach with an Application to Housing Price Prediction. *International Journal of Economics and Finance, 12*(4), 1-11.

Yitzhaki, S., & Schechtman, E. (2013). *The Gini Methodology: A Primer on A Statistical Methodology*. New York: Springer.

Zhang, S., Dang, X., Nguyen, D., Wilkins, D., & Chen, Y. (2019). Estimating Feature-Label Dependence Using Gini Distance Statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Retrieved from https://arxiv.org/pdf/1906.02171.pdf

Zietz, J., & Traian, A. (2014). When Was the U.S. Housing Downturn Predictable? A Comparison of Univariate Forecasting Methods. *Quarterly Review of Economics and Finance*, *54*(2), 271-281.