ESCUELA DE CONSTRUCCIÓN CIVIL
FACULTAD DE INGENIERÍA

Revista de la Construcción
Journal of Construction

Research Article

Machine learning methods for prediction real estate sales prices in Turkey

Cihan Çılgın[1, *], **Hadi Gökçen**[2]

[1] Bolu Abant Izzet Baysal University, Bolu (Turkiye); cihancilgin@ibu.edu.tr
[2] Gazi University, Ankara (Turkiye); hgokcen@gazi.edu.tr
*Correspondence: cihancilgin@ibu.edu.tr

Received: 26.06.2022; **Accepted:** 21.03.23; **Published:** 30.04.23

Citation: Çılgın, C., and Gökçen, H. (2023). Machine Learning Methods for Prediction Real Estate Sales Prices in Turkey. Revista de la Construcción. Journal of Construction, 22(1), 163-177. https://doi.org/10.7764/RDLC.22.1.163.

Abstract: Owning a house is one of the most important decisions that low and middle income people make in their lives. The real estate market is a significant factor of the national economy as much as it is important for individuals. Therefore, predicting real estate values or real estate valuation is beneficial and necessary not only for buyers, but also for real estate agents, economists and policy makers. This issue represents an active area of research, as individuals, companies and governments hold considerable assets in real estate. In this context, the aim of the study is to predict real estate prices with Machine Learning methods using the real estate sales data set in June and July 2021 belonging to the province of Ankara. In particular, it is to perform a comprehensive comparison on Machine Learning regression types methods that give successful prediction results in various but similar tasks, which are not included in the real estate literature. Real estate data obtained over the Internet was first included in a detailed data preprocessing process, and then Linear, Lasso and Ridge Regression, XGBoost and Artificial Neural Networks (ANN) methods were used on this dataset. According to empirical findings, XGBoost and ANNs appear as very important alternatives in predicting real estate sales prices.

Keywords: Machine learning, neural networks, real estate price, prediction, Ankara.

1. Introduction

Sheltering, which is at the beginning of the hierarchy of needs, has formed one of the most basic elements for the continuity of life throughout human history. In the ordinary course of life, individuals, after completing their education and personal development, need a job and then a house to live with their families in order to establish their own families (Aydemir et al., 2020). With the technological and social developments that have occurred over time, the acquisition of housing for the need for shelter has also gained a different dimension. Real estate is a long-term investment tool that is obtained by purchasing for a price after meeting the physiological needs of human beings (Işık, 2016). In addition to meeting the consumption drive and need, real estate also represents the largest asset class for many people, whose value increases over time and provides people with a safety net when they retire (Sing et al., 2021). Real estate indirectly affects the satisfaction and behavior of the residents in terms of the physical, psychological and socio-cultural environment in which it is located, and can shape the general health, happiness and well-being of the individual, family and society (Lawrence, 1987).

Real estate, which has changed shape with mass migrations and large population movements, is now more than just a shelter, it is also a durable consumer goods, a source of security for families and individuals, an investment tool, a place where

labor is reproduced, and a building block in the formation of the living environment (Yayar ve Deniz, 2014). As both a consumption and an investment good, the most important feature of the real estate is that it forms a large part of all assets of an average household during the household's life cycle. (Özsoy and Şahin, 2009). Owning a real estate is one of the most important, critical and in most cases rarely realized investments an individual can make in their lifetime. Financing a real estate represents a complex problem in terms of sheltering, health, safety and various sociocultural issues. In taking such an important decision, systems that make an accurate price prediction according to the real estate to be purchased are of great importance. (Aydemir et al., 2020). Therefore, it is essential for risk management for both individuals and firms to accurately know the real-time and relevant historical value of any real estate in order to sell and buy wisely. (Sangani et al., 2017). However, risk management is not possible without an appropriate and practical valuation tool (Doumpos et al., 2021). Therefore, the need to develop accurate real estate appraisal models is increasing in order to obtain a fair, accurate and appropriate real estate price prediction and to avoid human-specific subjectivity and bias in real estate appraisals. (Fan et al., 2018; Yazdani, 2021).

Appropriate and valid appraisal of a real estate is considered the key at any time and for any real estate related transaction, especially for sale or a loan application, so it is very significant that the price is a true reflection of its value (Alfaro-Navarro et al., 2020). However, real estate is perhaps as diverse as the people living in it, and the dynamics of the real estate market can directly reflect the situation of socio-economic diversity as well as affect the socio-economic conditions of countries (Wu and Sharma, 2012). A real estate appraisal is the process of developing a fair and acceptable market value for real estate for both buyer and seller. This process is a very comprehensive task that depends on many environmental, physical and macroeconomic factors and variables. The real estate market, unlike other markets, has an inelastic supply structure. This situation not only makes the real estate appraisal process more difficult, but also makes the real estate appraisal more significant. Real estate appraisal, in its simplest form, refers to a quantitative measurement of the benefits and liabilities arising from the ownership of the real estate. In this context, what is basically desired to be achieved is the market value of a real estate. Market value is predicted by applying valuation methods and procedures that reflect the nature of the property and the conditions under which the asset will likely be traded in the open market (Selim, 2009). Thus, in the real estate market, what is often referred to as a "valuation or appraisal" is the most accurate and consistent predict of the buying and selling price of real estate (Pagourtzi, et al., 2003). The objective and reliable determination of real estate values, the creation of models to be carried out simultaneously, and the production of value maps from the model are especially necessary for the future of developing economies (Yalpir et al., 2021). Although many methods and package programs are used in real estate appraisal today, there is a need to increase the success of these approaches in determining and predicting the value of real estate.

Especially in inflationary periods, as a traditional investment tool, real estate is an assurance tool against the loss of the value of money. It can earn rent as a result of sudden value increases in the market, and for this reason, many different reasons arise for its purchase. The course of real estate prices and whether the pricing is correct or not is an important indicator not only for the real estate sector but also for the whole economy. With the 2008 crisis, it became clear that the real estate market should be kept under constant scrutiny due to the potentially devastating impact on financial stability that the excessive asset price inflation may cause from this sector (Renigier-Biłozor and Wiśniewski, 2012). Therefore, real estate prices, price movements in the construction sector and economic realizations affect a large part of the society either directly or indirectly. Especially when today's Turkey is considered, the real estate sector represents a rapidly changing, competitive and non-transparent sector where information is difficult to access, as is the case in the rest of the world. Under the current conditions, the data mining process in such an industry can be a source of information for many stakeholders and can be used as an effective tool in rapidly responding to changing market conditions.

Determination of real estate prices is important not only for households, but also for businesses and various institutions. Especially in Turkey, in the real estate sector, which is growing rapidly and gaining more importance with the urban transformation and various other projects, the determination of real estate prices is very necessary for many parties. In this context, the aim of this study is to predict real estate prices using Machine Learning and Deep Learning methods, using the real estate training data set in June and July of Ankara, a metropolis with a population of more than five million and the capital city of Turkey. In this context, a comprehensive examination is made on Machine Learning regression types and deep learning models, which are not widely used in the real estate appraisal literature, but give popular and successful prediction results in

similar tasks. XGBoost model showed superior performance compared to other modems used in the study. Although there is not much difference between XGBoost and the Neural network model proposed in the study, there is a substantial difference between Linear, Lasso and Ridge regressions. In addition, similar studies conducted in Turkey usually have a very small data set, while a much larger data set was used in this study, unlike other studies. In addition to the large number of observations, the number of variables used in the study is also quite high, and these variables were determined using feature selection methods. Another important point of the study that differs from the literature is that it does not focus on a single region within the scope of Ankara province, but makes and prediction on a data set with high heterogeneity covering all districts in Ankara. Thus, in the study, it is aimed to realize a universal modeling process without putting any extra effort into the discovery of sub-markets.

In the light of this information, this study reveals that machine learning approaches, especially for the Turkish real estate market, are quite successful in the task of house price estimation. In addition, unlike the literature, while performing this task, it takes into consideration not only a region but also a whole city. Of course, evaluating more homogeneous sub-housing markets with machine learning approaches will yield better results. However, this study shows that machine learning approaches can yield successful results in real estate price prediction, especially in a large real estate market, which is heterogeneous in contrast to existing studies. In addition, the empirical findings of this study show that alternative current machine learning approaches can produce successful results in real estate price prediction. It also provides evidence that machine learning approaches such as XGBoost, which are used relatively little in this field, can be a better alternative to methods such as ANN and traditional multiple regression analysis, which are frequently preferred especially in the task of real estate price prediction.

In the later stages of the study, in the second part, studies on the prediction of real estate prices in the literature are presented. In the 3rd Chapter, information about Machine Learning and Deep Learning approaches within the scope of the methods used in the study, in the 4th Chapter, the dataset to be used in the application and the preparation processes of the dataset before the implementation, in the 5th Chapter, the information about the implementation process and the discussions on the results of the implementation and in Chapter 6, the conclusion of the study are presented.

2. Literature

Real estate appraisal is a phenomenon that closely concerns many stakeholders in this field around the world. A wide variety of real estate evaluate for multiple purposes (Janssen et al., 2001). Inflationary pressures experienced after the COVID pandemic in Turkey, as in the whole world, are attracting more individuals and companies to the real estate sector, both to protect the current investment and due to many profit opportunities. Although this type of analysis is a very difficult task that involves many areas, requires multivariate and fair market value determination (Afonso et al., 2019), studies on real estate appraisal are increasing day by day as a result of this increasing demand. The development of new models in this area and the improvement of data pre-processing processes positively affect real estate appraisal performances.

Today, especially extensive studies on machine learning and deep learning lead to the acceleration of developments in this field and its application in a wide variety of fields. In this context, many studies on the determination of real estate prices are available in this study area. Especially the presence of too many parameters in determining real estate prices makes machine learning and deep learning models more attractive in this field. Varma et al. (2018) developed a system to accurately predict real estate prices in their work with Linear Regression, Random Forest and Boosted Regression machine learning algorithms and ANN on real estate in Mumbai. The main purpose of the developed system is to prevent the risk of investing in the wrong real estate for system users by providing correct output.

Baldominos et al. (2018) in their study on the Spanish real estate sector, tested various machine learning algorithms that determined the advantages and handicaps of each method using various methods. The results of the study revealed that the better performing models are always ensemble models consisting of regression trees. Peterson and Flanagan (2009), using a large sample of 46,467 residential properties covering the years 1999-2005, found that ANNs produced significantly lower dollar pricing errors than linear hedonic pricing models. Selim (2009), in his study to determine house prices in Turkey using

ANNs and hedonic regression model, carried out an analysis with 46 variables on 5741 houses belonging to both rural and urban areas throughout Turkey. When the prediction performance results of the two methods compared in the study are examined, it is emphasized that ANNs show a higher prediction success in determining real estate prices in Turkey and therefore can be a good alternative. With a similar approach, Ecer (2014) used Hedonic models and ANNs in his study on İzmir province. Among the models developed on 610 different real estates and 87 variables, it was observed that ANNs gave the best prediction results, as in Selim's (2014) study. Kuru et al. (2021) considered real estate price prediction as a classification problem. Like Ecer (2014), Kuru et al. (2021) determined the functions that determine the real estate sales price ranges for the province of Izmir.

Alexandridis et al. (2019) compared linear and nonlinear models based on hedonic regression and ANNs in their study on the real estate market in Greece, and examined the strengths and performance of each method and applied a combined prediction rule to increase the prediction accuracy. Aydemir et al. (2020) developed an intelligent system that makes real estate price prediction with the data of real estates for sale. For 14 different methods applied in the study, a price prediction study was carried out with 176 features. The most successful price prediction for 852 houses in Ataşehir district of Istanbul was obtained by Random Forest algorithm. In their study, Yılmazel et al. (2018) used ANNs to predict the prices of real estate for sale in Eskişehir. As a result, ANNs have been shown to be an effective tool in predicting real estate price with 19 different ANN models.

Do and Grudnitski (1992) found in their study based on single-family real estate market transactions that neural networks provide nearly twice as accurate predictions as traditional regression models. Rossini (1997) concluded that neural networks show better results for smaller datasets, but multiple linear regression is clearly better for larger datasets. In addition, the study findings show that price prediction can be calculated very quickly regardless of the size of the problem with multiple linear regression, while the time required to use neural networks can increase exponentially with the size of the data set. However, contrary to Rossini (1997), the findings of a more recent and much larger sample size study conducted by Seya and Shiroi (2022) reveal that neural networks give better results with larger sample sizes. Liu et al. (2006) reveal that the success of fuzzy neural networks in predicting real estate prices is directly dependent on the quality of the data used. However, considering the implementation dates of these studies and the developments in the field of technology and machine learning at that time, it can be easily said that many of the criticisms made on ANN in these studies are in the past. In support of this situation, Rossini (1997) emphasizes that although ANN does not exhibit high estimation accuracy in his study, it may become more suitable as an analytical tool over time with the developments in computer technology, just like in regression models.

Although many studies have been carried out with many different methods and data sets in the field of real estate appraisal, this is not the case in Turkey. Especially the studies carried out within the scope of Turkey generally have very small observation and feature sets, but the variety of the models developed is quite limited. Most of the existing studies are linear regression models based on the hedonic model, and the number of studies that perform real estate price prediction with Machine Learning methods is very few. In addition, most of the studies have been conducted on a homogeneous data set as a sample for a specific housing sub-market, and the models developed are not universal. For this reason, in this study, alternative approaches are examined by comparing various Machine Learning regression types that give popular and successful prediction results for the Turkish real estate market. In addition, one of the main contributions of this study is to present a general model for Turkey with a large non-homogeneous data set collected for Ankara. In addition, this study has a very wide observation number and feature set compared to previous real estate appraisal studies conducted in Turkey. One of the most fundamental factors affecting the success of machine learning approaches is the data preprocessing process. For this reason, this study pays special attention to the data preprocessing process, which is neglected in the literature. Thus, it is aimed to increase the prediction performance.

3. Methods

Predicting real estate prices is basically a regression problem. For this reason, regression-based models from Machine Learning approaches were applied in the study.

3.1. *Liner regression*

According to the most widely used definition of Şahinler (2000), *"Regression analysis is a statistical analysis technique that characterizes the relationship between two or more variables with a cause-effect relationship, with a mathematical model called a regression model in order to make predictions about that subject".* Linear regression tries to model the relationship between two variables by fitting a linear equation to the observed data. One variable is considered an explanatory variable and the other is considered a dependent variable.

$$Y = a + bx + \varepsilon \tag{1}$$

where, *Y* is the dependent variable, *x* is the independent variable, *a* is the constant term, *b* is the slope of the line, and ε is the error term.

3.2. *Ridge regression*

Ridge regression is used as a very useful tool to overcome the multicollinearity problem (Walker ve Birch, 1988). Ridge regression, which is a linear regression method, is a regression technique using the L2 regularization norm and is expressed by the equation given below.

$$\beta^{Ridge} = min \sum (y - (\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k))^2 + \lambda \sum \beta^2 \tag{2}$$

where, y is the dependent variable, x is the independent variable, β is the coefficients, λ is the penalty term, and ε is the error term. Basically, Ridge regression tries to determine the coefficients with the smallest variance by compromising the concept of unbiasedness of the model.

3.3. *Lasso regression*

Lasso Regression is similar in nature to Linear Regression as Ridge regression, but uses a number of improvements. It has a form of regulation that aims to impose various restrictions to prevent over-fitting. For this purpose, L1 uses the regularization norm. In Lasso Regression, weights are constrained by penalizing their absolute values. This forces less important features to have weights of 0 and implicitly weeds out unnecessary variables in the process (Bin et al., 2017).

$$\beta^{Lasso} = min \sum (y - (\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k))^2 + \lambda \sum \|\beta\| \tag{3}$$

where, y is the dependent variable, x is the independent variable, β is the coefficients, λ is the penalty term, and ε is the error term. As can be seen in Equation 3, Ridge regression and Lasso regression are quite similar, but the main difference between the two is the regularization norms they use.

3.4. *Extreme gradient Boosting-XGBoost*

Developed by Chen ve Guestrin (2016), XGBoost is a high-impact, scalable machine learning algorithm for tree growing. XGBoost is one of the most effective machine learning methods for tree growing and is widely used by data scientists to achieve cutting-edge results in many research fields.

Unlike a traditional tree learning, XGBoost not only extracts information from its predecessors, but also aggregates the scores on the corresponding leaves to reduce the errors of the previous tree and get more accurate results at the end. XGBoost has many other features such as parallel and distributed computing that accelerate learning and can predict high accuracy (Zhao et al., 2019). This algorithm consists of multiple decision trees and gradient descent method is used to construct each

tree. Based on all single decision trees, optimization is performed by minimizing the loss function as the goal (Liang et al., 2019) and uses the following equation to minimize the loss function:

$$L(f_t) = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{t-1}) + \Omega(f_t) + C \quad (4)$$

where, $\Omega(f_t)$is the loss function used to explain the difference between the estimated value $\hat{y_i}^{t-1}$and the actual value y_i.

3.5. *Neural networks*

Neural networks are making great advances in solving problems that have resisted attempts by the artificial intelligence community for years (LeCun et al., 2015). Neural network approaches that have evolved over the years have proven to be very good at discovering complex structures in high-dimensional data and are therefore applied in many fields by science, business, and governments. Neural networks are computer systems developed with the aim of automatically realizing abilities such as deriving new information, creating new information and discovering new information through learning, which consists of the characteristics of the human brain, without any assistance. In short, artificial neural networks are inspired by the working mechanism of the human brain (Çılgın et al., 2020). Just as biological neural networks have nerve cells, artificial neural networks also have artificial nerve cells. The smallest units that allow artificial nerve cells to work are called the processing element. Each processing element has five basic elements: inputs, weights, summation function, activation function, and output (Öztemel, 2003).

Inputs include information from the outside world to an artificial neuron. The weights show the effect and importance of the information coming to the neuron on that neuron. As seen in Equation 5, the addition function is added by multiplying the input values and weights in order to calculate the net input.

$$net = \sum_{i=1}^{n} w_i \, x_i + \mathrm{b} \quad (5)$$

In neural networks, neurons are in layers. These layers are expressed as input layer, hidden layers and output layer. In addition, Python programming language was preferred in the realization of all the models mentioned. In particular, the applications of the relevant models were carried out by using the Sklearn and Keras libraries of this programming language.

4. Data

The data used within the scope of the study consists of the real estate information of the province of Ankara for the months of June and July 2021. These real estate data were obtained by a software developed from various sources over the internet. A total of 70.360 records were obtained from 25 districts of Ankara within the specified date range. The data obtained consists of 86 different variables, as summarized in Table 1 below, with some of the variables considered important. While 85 of them are independent variables, the real estate price variable is the dependent variable.

Table 1. Variables used in the study

Variable	Variable Type	Range	Variable	Variable Type	Range
Property Type	Categorical	1,2,3,4,5,6,7	Glass Insulation	Binary	0,1
Number of rooms	Continuous	1-10	Jacuzzi	Binary	0,1
Number of Halls	Continuous	1-6	Covered Balcony	Binary	0,1
Gross Size (m^2)	Continuous	1-175000	Air Condition	Binary	0,1
Net Size (m^2)	Continuous	0-150000	Laminate Kitchen	Binary	0,1
Number of Floors	Categorical	1,2,3,4,5,….,37	Marble Floor	Binary	0,1
Building Age	Continuous	0-52	Is it furnished?	Binary	0,1
Heating system	Categorical	1-10	Is there natural gas in the kitchen ?	Binary	0,1
Floor level	Continuous	1-52	Blinds	Binary	0,1
Credit Eligibility	Categorical	1,2,3	Parquet	Binary	0,1
Furniture Condition	Categorical	1,2,3	Laminate Flooring	Binary	0,1
Number of bathrooms	Continuous	1-6	Lamina Flooring	Binary	0,1
Fuel Type	Categorical	1,2,3,4,5,6,7,8,9	Satin Plaster	Binary	0,1
Using Status	Categorical	1,2,3,4	Satin Wall Paint	Binary	0,1
Is it on the Street?	Binary	0,1	Sauna	Binary	0,1
Is the Street Near?	Binary	0,1	Ceramic Floor	Binary	0,1
Is the Airport Near?	Binary	0,1	Spot Light	Binary	0,1
Any Landscape	Binary	0,1	Fireplace	Binary	0,1
Landscape	Binary	0,1	Terrace	Binary	0,1
Lake View	Binary	0,1	Cloakroom	Binary	0,1
City View	Binary	0,1	Underfloor Heating	Binary	0,1
Is it in the city center?	Binary	0,1	Elevator	Binary	0,1
Close to Metro	Binary	0,1	Garden	Binary	0,1
Close to Minibus	Binary	0,1	Mosaic Coating	Binary	0,1
Close to the Highway	Binary	0,1	Fitness Room	Binary	0,1
Close to Public Transport	Binary	0,1	Security	Binary	0,1
ADSL infrastructure	Binary	0,1	Water Booster	Binary	0,1
Wood Material	Binary	0,1	Thermal İnsulation	Binary	0,1
Alarm	Binary	0,1	Generator	Binary	0,1
Built-in Kitchen	Binary	0,1	Doorman	Binary	0,1
Balcony	Binary	0,1	Outdoor Parking	Binary	0,1
Household Appliances	Binary	0,1	Parking Garage	Binary	0,1
Laundry Room	Binary	0,1	Playground	Binary	0,1
Steel Door	Binary	0,1	PVC window	Binary	0,1
Shower Cabin in Bathroom	Binary	0,1	Siding	Binary	0,1
En-suite Bathroom	Binary	0,1	Is it on the site?	Binary	0,1
Dressing room	Binary	0,1	Water tank	Binary	0,1
Fixed Cabinet	Binary	0,1	Tennis Court	Binary	0,1
Carpeting	Binary	0,1	Fire Escape	Binary	0,1
Kitchen Utensil	Binary	0,1	Swimming Pool	Binary	0,1
Hilton Bathroom	Binary	0,1			

4.1. *Data preprocessing*

Before applying the models for house price prediction, the dataset needs to be preprocessed and cleaned. As the first step, the records with the same record number in the data set were deleted. Especially due to the announcements made on different dates, the records of the same houses are quite high in the obtained data set. Investigation of missing data was carried out as the second step. Many missing patterns have been rigorously evaluated as they play an important role in deciding on appropriate methods to handle missing data. Although most of the general information about the records obtained is available, it has been observed that there are many deficiencies in the data of interior and location characteristics. For this reason, since most of the missing variables are binary discrete variables such as 0-1, it was decided to completely remove the relevant records from the dataset instead of filling in these gaps. After these two processes, 18.420 records were obtained from the process that started with a total of 70.360 records.

4.2. *Data Transformation*

The first variable that needs attention in the data transformation process is Price as the dependent variable. As can be seen in Figure 1, the distribution obtained from the Price variable is right skewed. Considering that the models we will run on are linear regression models, this skewness in the price variable will affect the predictive power.

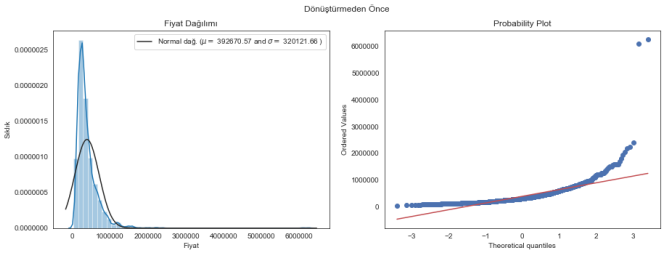

Figure 1. Distribution of the pre-transformation price variable.

In order to eliminate this effect and eliminate the skewness, logarithmic transformation was performed on the Price variable. The logarithmic transformation is used to convert an exponential and skewed distribution into a normal distribution, with the simplest expression. Indirectly, it is possible to obtain a more accurate prediction. This transformation process is performed using the mathematical expression given in Equation 6 below.

$$x_i = log\ (1 + x_i) \tag{6}$$

The distribution of the Price variable obtained after this transformation is shown in Figure 2. As can be seen in the figure, the price distribution now has a more normal distribution, which reduces the possibility of errors in the predictions.

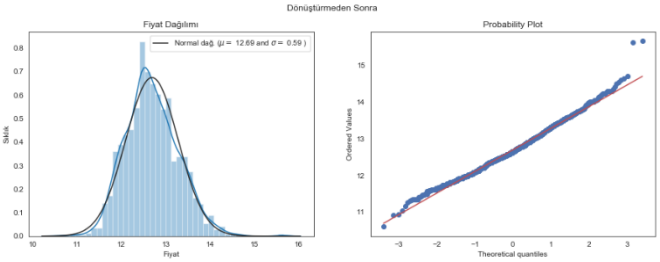

Figure 2. Distribution of the post- transformation price variable.

In general, expressing one variable in smaller units and another in larger values will result in a wider range for these variables and therefore a large impact on the resulting predictive structure. To help avoid dependence on the choice of units of measurement, data should be standardized. By standardizing the measurements, it is tried to give equal weight to all variables (Han et al., 2011). In this context, considering the data used in our study, the Z-Score standardization method was used for standardization because the data consisted of both intermittent data and different continuous variables. When measuring for a variable f, standardization is done as follows (Liu et al., 2008):

$$s_f = \frac{1}{n}(|x_{1f} + m_f| + |x_{2f} + m_f| + \cdots + |x_{nf} + m_f|) \tag{7}$$

$$Z_{if} = \frac{x_{if} - m_f}{S_f} \tag{8}$$

where, $x_{1f}, \ldots, x_{nf}$ is the nth measurements of f'in and $m_f$is the mean value of f. m_f is calculated as $m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf})$. Since the mean absolute deviation, S_f, is stronger than the standard deviation in outliers (limit values), σ_f, the mean

absolute value is used. Within the scope of the study, the Z-Score standardization method, the details of which were explained above, was applied to all variables in the data set.

4.3. *Feature selection*

Generally, a dataset can contain many variables. Within the scope of this study, the data set consists of 85 very different variables. However, all these variables in the dataset may not be useful in creating a machine learning algorithm to perform the necessary prediction. In some cases, its use may even reduce its predictive power. Huang (2019) supported by empirical findings that heuristic deletion of features in a real estate valuation task can be very dangerous and that feature selection is a very important process. Therefore, feature selection plays a very important role in building a machine learning model. The first step that can be done for this purpose is to create the correlation matrix in order to control the mutual correlation of the variables.

Correlation is a statistical term that expresses how close two variables are to a linear relationship with each other in common usage. Variables with a high correlation are more linearly dependent and therefore have almost the same effect on the dependent variables. That is, when two features are highly correlated, one of the two variables needs to be removed from the dataset. For this purpose, the correlation matrix giving the mutual correlation coefficients is given in Figure 3.

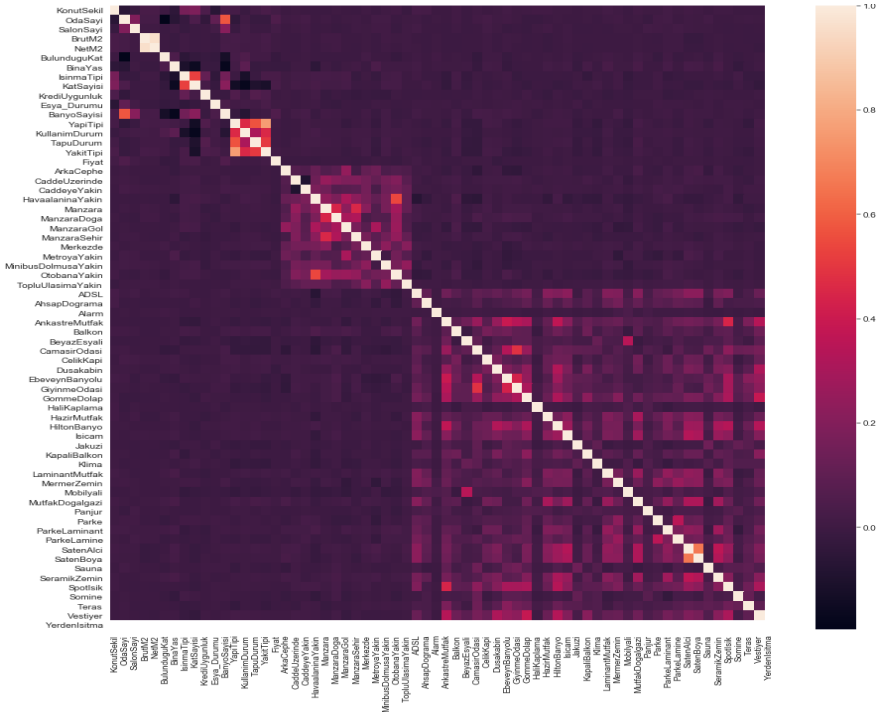

Figure 3. Correlation matrix.

According to the correlation matrix given in Figure 3, there is a high correlation between the GrossM2 and NetM2 variables. Therefore, removing one of these variables from the data set will not only shorten the processing time, but also positively affect the forecast performance. In this framework, the data with a correlation coefficient higher than 0.90 were excluded from the study. Since only BurtM2 and NetM2 variables provided this threshold value among all variables, the GrossM2 variable was excluded from the dataset. In addition, some features were removed from the dataset by using the Back Elimination method in feature selection. Each independent variable has an effect on the dependent variable. While some variables have a high effect on the system, some have less. The elimination of independent variables that have little effect on the system allows a better model to be established. For this purpose, the Back Elimination method is used.

In this method, which is basically based on the logic of performing a hypothesis test, the following steps are followed in order:

1) The level of significance at which the test will be performed is determined, usually a value of 0.05.
2) A simple linear regression model with all variables is created.
3) The model including all the variables is tested over this model.
4) The variable is extracted by comparing the calculated P value to the significance level.
5) If it does not provide a level of significance, this process is continued by removing each variable step by step.

As presented in Table 2, the total number of independent variables was reduced from 83 to 47 with the Back Elimination method.

Table 2. Variables used in the study after feature selection

Variable	Variable Type	Range	Variable	Variable Type	Range
Property Type	Categorical	1,2,3,4,5,6,7	Glass İnsulation	Binary	0,1
Number of rooms	Continuous	1-10	Marble Floor	Binary	0,1
Number of Halls	Continuous	1-6	Parquet	Binary	0,1
Gross Size (m^2)	Continuous	1-175000	Laminate Flooring	Binary	0,1
Building Age	Continuous	0-52	Ceramic Floor	Binary	0,1
Heating system	Categorical	1-10	Spot Light	Binary	0,1
Floor level	Continuous	1-52	Cloakroom	Binary	0,1
Credit Eligibility	Categorical	1,2,3	Elevator	Binary	0,1
Number of bathrooms	Continuous	1-6	Garden	Binary	0,1
Using Status	Categorical	1,2,3,4	Mosaic Coating	Binary	0,1
Is the Airport Near?	Binary	0,1	Fitness Room	Binary	0,1
Any Landscape	Binary	0,1	Security	Binary	0,1
Landscape	Binary	0,1	Water Booster	Binary	0,1
City View	Binary	0,1	Thermal İnsulation	Binary	0,1
Is it in the city center?	Binary	0,1	Generator	Binary	0,1
Close to Metro	Binary	0,1	Doorman	Binary	0,1
ADSL infrastructure	Binary	0,1	Parking Garage	Binary	0,1
Wood Material	Binary	0,1	Playground	Binary	0,1
Balcony	Binary	0,1	Siding	Binary	0,1
Laundry Room	Binary	0,1	Is it on the site?	Binary	0,1
En-suite Bathroom	Binary	0,1	Water tank	Binary	0,1
Dressing room	Binary	0,1	Fire Escape	Binary	0,1
Fixed Cabinet	Binary	0,1	Swimming Pool	Binary	0,1
Carpeting	Binary	0,1			

4.4. *Cleaning outliers*

Anomalies or outliers are data patterns that have data characteristics that differ from normal samples. Detection of outliers is of significant relevance and often provides critical actionable information in a variety of application areas, but in a prediction model these outliers directly affect prediction performance.

For this purpose, Isolation Forest method was used to determine outliers in the study. Isolation Forest is a different model-based method that explicitly isolates anomalies rather than profiling normal samples. To achieve this, the method used makes use of two quantitative characteristics of the anomaly: i) they are a minority of fewer samples, and ii) they have very different attribute values from normal samples. In other words, anomalies are "few and different," making them more susceptible to isolation than normal spots. Due to their susceptibility to isolation, anomalies are isolated close to the root of the tree; normal points are isolated at the deep end of the tree. This isolation feature of the tree forms the basis of this method of detecting anomalies (Liu et al., 2008). As a result of the Isolation Forest method applied on the data set, 1.842 records were detected as anomaly. As a result of removing these records from the data set, 16.578 records were obtained. With the removal of outliers, the operations performed on the data that will increase the performance of the tahini models have been concluded, and the last 16.578 records obtained in all models have been studied.

5. Application and findings

Within the scope of the study, five different methods described above were used to create real estate price predictions. Especially Linear, Lasso and Ridge regression methods are traditional and linear regression methods that are frequently used in real estate pricing from past to present, while XGBoost and Neural Networks are more up-to-date and more effective methods in these matters. Especially in various subject areas, the XGBoost method provides an effective perspective in the comparison of application results due to its remarkable performance recently and the fact that deep neural networks show a universal approach by giving effective results in all kinds of linear or non-linear problems. Before moving on to the application findings, three different metrics used in comparing the methods used in the study and measuring the errors in the training dataset are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{9}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(e_i)^2 \tag{10}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(e_i)^2}{n}} \tag{11}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - y_i}{x_i}\right| \tag{12}$$

If these values, which are used for measuring error values, approach zero, it means that the predictive power of the model is higher. Another important point to be noted is how the process of comparing and selecting the model will be carried out. For this purpose, k-fold cross validation was used to prevent bias and random over-optimism that may arise during model testing. Within the scope of the study, 5-fold cross-validation was used for model evaluation.

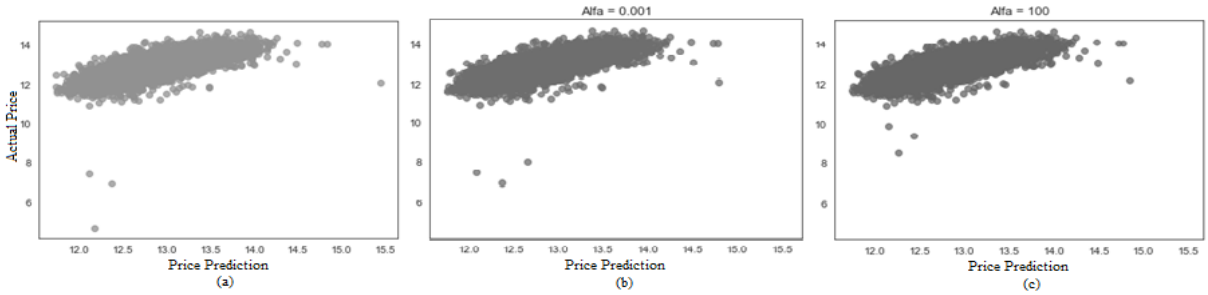

Figure 4. Test dataset predictions of linear, lasso, ridge regression.

The first model to focus on is the linear regression model. This model, which is very easy to use, can be applied to all linearly related data sets. It does not need to use any extra parameters during use. Figure 4(a) visualizes the comparison of estimated and actual values performed on 4.974 test data. Lasso regression inherently requires an alpha parameter. After trying various alpha values, it was observed that the value of 0.001 gave the best result. Figure 4(b) visualizes the comparison of the predicted and actual values performed on the test data, and it can be observed that Lasso regression shows a better improvement than linear regression.

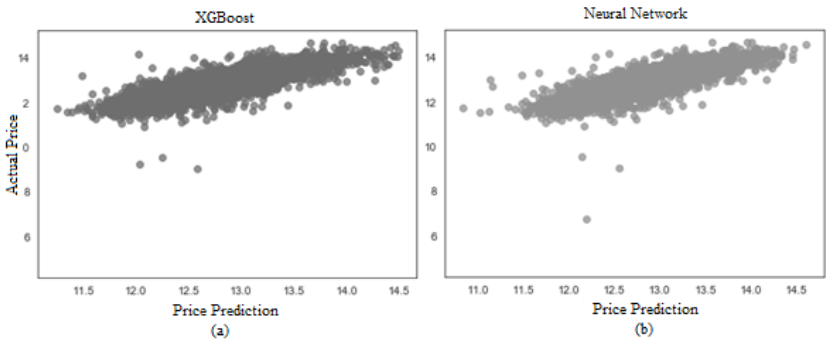

Figure 5. Test dataset predictions of XGBoost and neural network.

As we mentioned before, Ridge regression and Lasso regression are very similar to each other and differ only in terms of regularization norms. Ridge regression also needs an alpha parameter like Lasso. Likewise, as a result of various trials, the alpha value giving the best results was found to be 100. In Figure 4(c), the comparison of the predicted and actual values performed on the test data is visualized and it can be observed on this graph that Ridge regression shows a better improvement than Lasso and linear regression.

XGBoost has a wide variety of parameters compared to the methods mentioned above, and therefore, using each parameter together with other parameters produces different results. In order to find the best result, the cross-validation method was used. For this purpose, the values of 0.001, 0.1, 0.12, 0.13, 0.14 for the *"learning rate"* parameter; 100, 300, 400, 500, 700 values for the *"n_estimators"* parameter; 1, 2, 3, 4, 5, values for the *"max depth"* parameter, 125 different combinations were evaluated, and the parameter vector that gave the best result was selected. In this direction, the "learning rate" parameter that gave the best results was determined as 0.1, the *"n_estimators"* parameter as 4, and the *"max depth"* parameter as 400. The comparison of the estimated and actual values performed on the test data of the model obtained with these parameters is visualized in Figure 5(a).

The neural network topology used in the study is a 5-layer, fully connected feedforward network consisting of an input layer, 3 hidden layers, and an output layer. While determining these layers, alternative network structures have been tried just like in the XGBoost method. As a result of the experiments on the alternatives of this network structure used, it was observed that the success of the network did not increase significantly, and it was observed that an increase in the number of layers significantly increased both the training time and the evaluation process of the network, considering the amount of data to be evaluated. In the input layer, as many as the number of variables in the data set, that is, 52 neurons are used and *"tanh"* is chosen as the activation function of this layer. For the hidden layers, on the other hand, 720 samples and the *"relu"* activation function were used as the activation function. In addition, there is a dropout layer (p=0.2) between each hidden layer. The output layer consists of a single neuron and the activation function is used as *"elu"*. The *"adam"* function was used as the optimization function. Adam (Adaptive Moment Estimation) is a more efficient, adaptive optimization algorithm that can be preferred over stochastic gradient reduction method. The Adam algorithm dynamically updates the learning rate for each parameter. Here, the value of 0.0015 was determined as the learning rate, and the values of 0.9 and 0.999 were assigned to the values of β_1 and β_2, respectively.

As the update time of the weights, the network representation of the whole training dataset, that is, epoch-based training, was used. In this approach, all the examples in the training dataset are shown to the network one by one, and then the errors that occur in the training set are collected after the entire training dataset is displayed to the network. After the examinations, no change was observed in the error rate after 500 epochs and the learning of the network was completed. Figure 5(b) visualizes the comparison of the deep neural network model and the prediction and actual values performed on the test data.

Table 3. Comparison of model performances

	MAE	MAPE	MSE	RMSE
Linear Regression	0.2989	0.0220	0.1655	0.4068
Lasso Regression	0.2988	0.0224	0.1654	0.4066
Ridge Regression	0.2978	0.0219	0.1652	0.4065
XGBoost	0.2420	0.0198	0.1214	0.3484
Neural Network	0.2421	0.0199	0.1216	0.3488

When Table 3 is examined, it is seen that XGBoost shows the best performance according to the mean square error, mean absolute error and root mean square error performance measures. Although there is not much difference between the neural network model and the Linear, Lasso and Ridge regressions, there is a significant difference. In this context, XGBoost and deep learning models appear as very important alternatives in predicting real estate prices.

6. Conclusions

Today, especially extensive studies on machine learning and deep learning lead to the acceleration of developments in this field and its application in a wide variety of fields. In this context, many studies on the determination of real estate prices are available in this study area. Especially the presence of too many parameters in determining real estate prices makes machine learning and deep learning models more attractive in this field. In this study, research was carried out on the determination of real estate prices with various regression methods and XGBoost and neural network approach, which are frequently used in current applications, on a large dataset consisting of real estate records belonging to the province of Ankara. A total of 16.578 real estate records were obtained with data cleaning and feature extraction performed on the dataset obtained from various sources on the Internet. When the performances of the models are compared with the remaining data after this preprocessing, XGBoost gives the best metric results and there is no significant difference between the proposed neural network model; however, it was observed that there was a significant difference between linear, Lasso and Ridge regressions. Especially in the prediction of real estate prices, the fact that every country and even every city has its own different texture makes this process very difficult. Here, the following point is also important that it is possible that the methods used in a study that can be done for different cities, that is, a study that can be carried out for a different dataset, may give different results, and it is possible that other methods will give more successful results. At this point, it is revealed with this study that XGBoost and deep learning methods can be used as a good alternative, especially in predicting real estate prices in Ankara. Based on the data obtained from the experimental study, the following conclusions could be drawn:

1. XGBoost model showed superior performance compared to other modems used in the study.
2. Although there is not much difference between XGBoost and the Neural network model proposed in the study, there is a substantial difference between Linear, Lasso and Ridge regressions.
3. In addition, similar studies conducted in Turkey usually have a very small data set, while a much larger data set was used in this study, unlike other studies.
4. In addition to the large number of observations, the number of variables used in the study is also quite high, and these variables were determined using feature selection methods.
5. This article reveals that machine learning approaches, especially for the Turkish real estate market, are quite successful in the task of house price estimation. In addition, unlike the literature, while performing this task, it takes into consideration not only a region but also a whole city.
6. In addition, the empirical findings of this study show that alternative current machine learning approaches can produce successful results in real estate price prediction.

In future studies, it may be possible to compare different and up-to-date approaches with the methods used in this study, as well as testing hyper-parameters with wider intervals in the methods used within the scope of the study. In addition, a collective method that can be created with more than one model that can be used is likely to affect the prediction performance positively.

Author contributions: Cihan Çılgın, writing, experimental investigation, data collection and application; Hadi Gökçen, writing, planned the experiments, critically revised the article. In addition to all these, this article is derived from the doctoral thesis carried out by Cihan Çılgın in Gazi University Institute of Informatics, Department of Management Information Systems.

Funding: The authors received no financial support for this article.

Conflicts of interest: The authors declare that there is no conflict of interest for the article.

References

Afonso, B., Melo, L., Oliveira, W., Sousa, S., & Berton, L. (2019). Housing Prices Prediction with a Deep Learning and Random Forest Ensemble. In Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional,389-400.

Alexandridis, A. K., Karlis, D., Papastamos, D., & Andritsos, D. (2019). Real Estate valuation and forecasting in non- homogeneous markets: A case study in Greece during the financial crisis, Journal of the Operational Research Society, 70(10), 1769-1783. https://doi.org/10.1080/01605682.2018.1468864

Alfaro-Navarro, J. L., Cano, E. L., Alfaro-Cortés, E., García, N., Gámez, M., & Larraz, B. (2020). A fully automated adjustment of ensemble methods in machine learning for modeling complex real estate systems. Complexity, Article ID: 5287263. https://doi.org/10.1155/2020/5287263.

Aydemir, E., Aktürk, C., & Yalçınkaya, M. A. (2020). Estimation of Housing Prices with Artificial Intelligence, Turkish Studies, 15(2), 183-194.

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying Real Estate Opportunities Using Machine Learning, Applied sciences, 8(11), 2321. https://doi.org/10.3390/app8112321.

Bin, J., Tang, S., Liu, Y., Wang, G., Gardiner, B., Liu, Z., & Li, E. (2017). Regression model for appraisal of real estate using recurrent neural network and boosting tree. In 2017 2nd IEEE international conference on computational intelligence and applications (ICCIA), 209-213. IEEE.

Chen, T. & Guestrin, C. (2016). XGBoost: "A Scalable Tree Boosting System", *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794.

Çılgın, C., Ünal, C., Alici, S., Akkol, E., & Gökşen, Y. (2020). In Text Classification, Bitcoin Prices and Analysis of Expectations in Social Media with Artificial Neural Networks. Mehmet Akif Ersoy University Journal of Applied Sciences, 4(1), 106-126.

Do, A. Q., & Grudnitski, G. (1992). A neural network approach to residential property appraisal. The Real Estate Appraiser, 58(3), 38-45.

Doumpos, M., Papastamos, D., Andritsos, D., & Zopounidis, C. (2021). Developing automated valuation models for estimating property values: a comparison of global and locally weighted approaches. Annals of Operations Research, 306(1), 415-433.

Ecer, F. (2014). Comparision of Hedonic Regression Method and Artificial Neural Networks to Predict Housing Prices in Turkey. In International Conference On Eurasian Economies,1-10.

Fan, C., Cui, Z., & Zhong, X. (2018, February). House prices prediction with machine learning algorithms. In Proceedings of the 2018 10th International Conference on Machine Learning and Computing (pp. 6-10).

Han, J., M. Kamber & J. Pei. (2011). Data Mining: Concepts and Techniques, 2011, USA: Elsevier.

Işık, C. (2016). Environment, Structure And Social Differentiation Of Housing Prices In Erzurum: Hedonic Pricing Method. Erzincan University Journal of Social Sciences Institute, 8(2), 23-32.

Janssen, C., Söderberg, B., & Zhou, J. (2001). Robust estimation of hedonic models of price and income for investment property. Journal of Property Investment & Finance, 19(4), 342-360. https://doi.org/10.1108/EUM0000000005789.

Kuru, M., Erdem, O. and Calis, G. (2021). Sale price classification models for real estate appraisal. Revista de la Construcción. Journal of Construction, 20(3), 440-451. https://doi.org/10.7764/RDLC.20.3.440.

Lawrence, R. J. (1987). Housing, dwellings and homes: Design theory, research and practice.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Liang, Y., Wu, J., Wang, W., Cao, Y., Zhong, B., Chen, Z., & Li, Z. (2019). Product marketing prediction based on XGboost and LightGBM algorithm. In Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition, 150-153.

Liu, J. G., Zhang, X. L. and Wu, W. P. (2006, May). Application of fuzzy neural network for real estate prediction. In International Symposium on Neural Networks (pp. 1187-1191). Springer, Berlin, Heidelberg.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining, 413-422. IEEE.

Özsoy, O., & Şahin, H. (2009). Housing price determinants in Istanbul, Turkey: An application of the classification and regression tree model. International Journal of Housing Markets and Analysis, 2(2), 167-178. https://doi.org/10.1108/17538270910963090

Öztemel, E. (2003). "Artificial neural networks". PapatyaYayincilik, Istanbul, 21-22.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: a review of valuation methods. Journal of Property Investment & Finance, 21(4), 383-401. https://doi.org/10.1108/14635780310483656.

Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal". Journal of Real Estate Research, 31(2), 147-164.

Renigier-Biłozor, M., & Wiśniewski, R. (2012). The impact of macroeconomic factors on residential property price indices in Europe. Folia Oeconomica Stetinensia, 12(2), 103-125. https://doi.org/10.2478/v10031-012-0036-3.

Rossini, P. (1999, January). Accuracy issues for automated and artificial intelligent residential valuation systems. In International real estate society conference, 1-10.

Şahinler, S. (2000). The basic principles of constructing a linear regression model using the least squares method., Mustafa Kemal Üniversitesi Ziraat Fakültesi Dergisi, 5(1-2), 57-73.

Sangani, D., Erickson, K., & Al Hasan, M. (2017, October). Predicting zillow estimation error using linear regression and gradient boosting. In 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS) (pp. 530-534). IEEE.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. Expert systems with Applications, 36(2), 2843-2852. https://doi.org/10.1016/j.eswa.2008.01.044.

Seya, H. and Shiroi, D. (2022). A comparison of residential apartment rent price predictions using a large data set: Kriging versus deep neural network. Geographical Analysis, 54(2), 239-260.

Sing, T. F., Yang, J. J., & Yu, S. M. (2021). Boosted Tree Ensembles for Artificial Intelligence Based Automated Valuation Models (AI-AVM). The Journal of Real Estate Finance and Economics, 1-26.

Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018). "House Price Prediction Using Machine Learning and Neural Networks". In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, 1936-1939. IEEE.

Walker, E., & Birch, J. B. (1988). Influence measures in ridge regression, Technometrics, 30(2), 1988, 221-227.

Wu, C., & Sharma, R. (2012). Housing submarket classification: The role of spatial contiguity. Applied Geography, 32(2), 746-756. https://doi.org/10.1016/j.apgeog.2011.08.011.

Yalpir, S., Sisman, S., Akar, A. U., & Unel, F. B. (2021). Feature selection applications and model validation for mass real estate valuation systems. Land use policy, 108, 105539.

Yayar, R., & Deniz, G. Ü. L. (2014). Hedonic Estimation of Housing Market Prices in Mersin City Center. Anadolu UniversityJournal of Social Sciences, 14(3), 87-100.

Yazdani, M. (2021). Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction. arXiv preprint arXiv:2110.07151.

Yılmazel, Ö., Afşar, A., & Yılmazel, S. (2018). Using Artificial Neural Network Method to Predict Housing Prices, International Journal of Economic & Administrative Studies, (20), 285 – 300.

Zhao, Y., Chetty, G., & Tran, D. (2019). Deep Learning with XGBoost for Real Estate Appraisal. In 2019 IEEE Symposium Series on Computational Intelligence (SSCI), 1396-1401. IEEE.