www.degruyter.com/view/j/remav

sciendo

REAL ESTATE MARKET PRICE PREDICTION MODEL OF ISTANBUL

Mert Tekin
Business Analytics
University of Warwick, United Kingdom
e-mail: merttekinn98@gmail.com

Irem Ucal Sari
Department of Industrial Engineering
Istanbul Technical University, Turkey
e-mail: ucal@itu.edu.tr

Abstract

The Turkish Housing Market has experienced a steep increase in prices. Individual and corporate investors now possess tools to estimate the real estate evaluation while using smaller amounts of data with traditional techniques. Not having an analytical approach to evaluate the price of real estate could cause the investor to lose considerable amounts of money, especially in the case of individual investors. This study aims to determine how different machine learning algorithms with real market data can improve this process.

To be able to test this, over 30000 lines of housing market data with over 13 variables is scraped. Data is cleansed, manipulated and visualized, while predictive models such as linear regression, polynomial regression, decision trees, random forests, and XGboost are created and compared according to the CRISP-DM framework. The results show that using complex techniques to create machine learning models could improve the accuracy in predicting the listing prices of houses.

This paper aims to:

- analyze the effects of using a real and relatively large amount of data,
- determine the main variables that contribute to the evaluation of an estate,
- compare different machine learning models to find the optimal one for the real estate market,
- create an accurate model to predict the value of any house on the Istanbul market.

Key words: *price prediction, machine learning, real estate market*.

JEL Classification: *R30, R31*.

Citation: Tekin, M. & Sari, I.U. (2022). Real estate market price prediction model of Istanbul. *Real Estate Management and Valuation, 30*(4), 01-16.

DOI: https://doi.org/10.2478/remav-2022-0025.

1. Introduction

Turkey is an emerging country and the construction industry plays a vital role in its development. As it is one of the main investment tools for the Turkish people, there are many organizations and individuals interested in the correct valuation of housing prices.

The real estate market plays a crucial role in the Turkish economy. The country has one of the largest construction industries in the world. Forty-four Turkish companies are listed in the "Top 250 International Contractors of the World", and over 6% of the GDP of Turkey is directly related to the construction industry while employing over 1.5 million people. Thus, the housing prices are directly related to the macroeconomics of Turkey and correct valuation is crucial.

Many people are buying houses not to spend their life in a safe environment but by seeing the houses as an investment in their portfolio. In developed countries, real estate is one of the greatest parts of private a household's wealth. The valuation of the house might have a leading impact on the portfolio of the household (Case et al., 2004). Therefore, the correct valuation of the houses holds great importance to people working in banks, politicians, investors and real estate agents.

In this study, the indicators of the Turkish real estate market are discussed. The Istanbul housing market is chosen since it might be seen as a reflection of the Turkish real estate market in general and is one of the most stable markets in the country. The methods that can possibly create a correct valuation have been examined. These algorithms are examined on the Istanbul housing dataset collected by one of Turkey's most popular public real estate market websites (www.hepsiemlak.com) with over 30000 lines of housing data. The "beautiful soup library" in the python programming language is used to scrape this data from the web. This dataset answers the questions; ''Which district is the house located in?'', 'Which neighborhood is the house located in?'', ''How many rooms does the house have?'', ''How big is the living space in square meters?'', ''On which floor is the house located?'', ''How many floors exist in the building?'', ''How old is the building?'', ''Does the house have a connection to natural gas?'', ''How many bathrooms does the house have?", ''Is there an elevator in the building?'' and ''How much is the valuation of the house?''.

The creation of a valuation model can be defined as a regression problem. In machine learning, many methods are used to solve regression problems. In this study, five of these machine learning methods are discussed and applied. These methods are linear regression, polynomial regression, decision trees, random forests and gradient boosting. Their mathematical formulas are shown, explained, and will be used to create meaningful accurate models for predicting housing prices in Istanbul.

Over the last few decades, researchers mostly used hedonic approaches, support vector machines, fuzzy logic, artificial neural networks, and traditional machine learning algorithms to create house pricing prediction models. The purpose of this study is to provide an analytical approach to solving the housing valuation problem using the variables that are chosen and applied in the recent studies in this area. Prior studies carried out in this area have been reviewed. In the present study, several different methods are discussed, and their applications are explained in detail. Then, the results obtained from these methods are compared to find out which of these methods is performing better. In the application part of the study, data on houses being sold in the Istanbul market, containing over 30000 houses with their characteristics, is used. Lastly, the limitations of this study are mentioned in the conclusion with the discussion of future studies on the topic.

2. Literature review

Fan, Ong and Koh (2006) tried to understand the importance of the relationship between the prices of the housing market and the characteristics of houses in the Singapore resale public housing market by utilizing the decision tree approach. Park and Bae (2015) used C4.5, RIPPER, Naïve Bayesian, and AdaBoost machine learning algorithms to develop different housing price prediction models using over 5000 lines of housing data in Fairfax County, Virginia. As shown by the results of the study, the model in which they used the RIPPER algorithm outperformed the other developed models in terms of the accuracy of the test set. Yılmazel, Afşar and Yılmazel (2018) examined the housing market in Eskisehir by gathering data on over 5000 houses on sale along with their characteristics. During the study, they created 19 different neural network models. Neural network models in this study showed superior performance when the models are fed with huge amounts of data. Nghiep and Al (2001) compared the performance of artificial neural networks and multiple regression analysis on single-family residential properties in Rutherford County, Tennessee. As a result, artificial neural networks performed better than multiple regression analyses on their data.

Wang, Wen, Zhang and Wang (2014) used particle swarm optimization to optimize the parameters of a support vector machines algorithm to create a real estate price forecasting model with the data of Chongqing, China. The results have shown that the proposed approach has better results compared to genetic and grid algorithms. Yayar, and Gül (2014) suggested a hedonic pricing model for the Mersin real estate market. They used linear, semi and fully logarithmic regression models in their study as well. As a result of the study, the main factors that affect the price of a house are identified and explained. Liu and Wu (2020) suggested the combination of the whale optimization (WOA) algorithm and modified Holt's exponential smoothing (MHES) method to create house pricing models for

Kunming, Changchun, Xuzhou, and Handan cities. The examined results showed the models created by using WOA-MHES have lower prediction error and required less computational time than the traditional models.

Rui Liu & Lu Liu (2019) analyzed the long short-term memory (LSTM) method for creating a real estate pricing model for Shenzhen, China. The results of the study showed the LSTM approach to perform well and bring better results than support vector machines and back propagation neural networks. Kuşan, Aytekin and Özdemir (2010) have developed a fuzzy logic model to predict the housing prices of Eskisehir City in Turkey. The results of the study showed that the model performed very well on the Eskisehir real estate market data.

Selim (2009) utilized two different methods to create a price modeling for houses in Turkey by including many features in the dataset. Hedonic-based regression and artificial neural networks are compared in this study and, as a result, it is found that the artificial neural networks outperformed when given sufficient data features.

Abidoye and Chan (2017) suggested the use of artificial neural networks to create a housing valuation prediction model in Nigeria. Since the records of the households being sold were not recorded properly, Abidoye and Chan could only use 370 data in the dataset. As the result, the model created by artificial neural networks was observed to perform sufficiently and could be used to reduce misvaluations of housing prices. Abraham (2016) used artificial neural networks to create a house pricing model for New Zealand. In his studies, the New Zealand House Price Index is used to track the changes in prices for over ten years. The correlation between prices and immigration is also examined. As the result of the study, it was proven that the use of artificial neural networks performs sufficiently to predict the housing prices in New Zealand.

Eriki and Udegbunam (2007) examined the performances of the multiple regression model (MRM) and brain maker neural network (BMNN) in evaluating the housing unit prices in Nigeria between the period 1980-2001. The study concluded that the use of BMNN performs better than MRM to determine the housing unit prices in Nigeria. Ecer (2014) suggested the use of the hedonic regression approach and artificial neural networks approach to create a model to predict the housing prices in Turkey by using 610 lines of housing data in Izmir city, Karsiyaka. The study supported the use of artificial neural networks outperforming the hedonic approach to valuating housing prices.

Górak (2017) used linear artificial neural networks to set a model for real estate valuation and appraisal. The study revealed that the model can be deployed for property valuation. Mora-Esperenza (2004) used artificial neural networks (ANN)to create a real estate valuation model for Madrid City, Spain. The study revealed ANN is capable of capturing the collective behavior of housing market factors and is applicable in prediction models. Khalafallah (2008) utilized artificial neural networks to develop a model to forecast the behavior of the housing market in the United States. The study showed that the use of the forecast model is applicable for short time periods.

Zurada, Levitan and Guan (2006) utilized fuzzy logic and artificial neural networks to create a valuation model for the real estate market. Their parameters were: the size of the garage, the number of bathrooms, and the size of the household. The study showed that both methods provide satisfactory results as long as there are enough data samples. Mimis, Rovoli, and Stamou (2013) studied the Athens real estate market to create a model for valuation prediction using artificial neural networks. Their data included over 3000 data and a six-year period. The study showed that artificial neural networks could explain 87% of the variation in the dataset. Cebula (2009) used a hedonic approach to create a house pricing model for the City of Savannah, Georgia. The data of 2888 households and a period of 5 years was used and included 25 variables in the model.

Thanasi (Boçe) (2016) suggested the construction of a hedonic pricing model for apartments in the capital city of Albania, Tirana. According to the study, location was found to be the most important variable that affects pricing. Other variables include the square meters of living space, the number of rooms, access to parking, furniture, the view, and surface area.

3. Data and Methods

The real estate of the Istanbul market was selected in this study after reviewing the reports of the Residential Property Price Index of the Central Bank of the Republic of Turkey (2020).The Residential Property Price Index covers the indicators that are constructed for monitoring the price movements in the Turkish housing market, and it is used as a general inflation indicator for house prices in Turkey.

The aim of the study is to create a valuation model for a metropolitan city in Turkey that can be used for a relatively long time period. Figure 1 from the report has shown that the pricing fluctuations of the real estate market in Istanbul are relatively lower than in other metropolitan cities of Turkey.

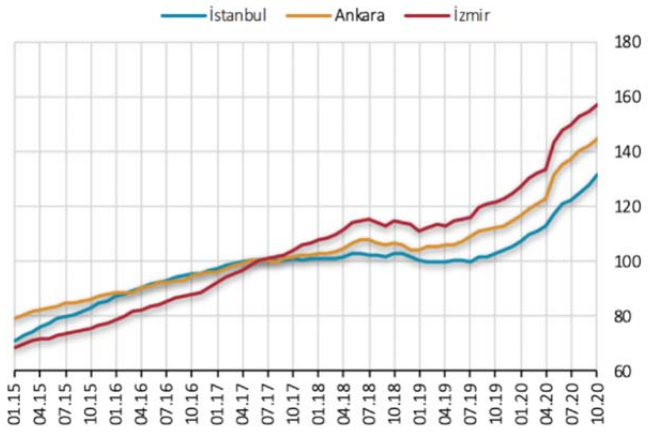


Fig.1. Graph for the major cities of Turkey, (Residential Property Price Index Report, 2020).*Source*: own study.

In this study, we apply different ML methods. Different ML methods might require different parameters, however, all of the ML methods benefit from the abundance of useful data. Therefore, more data carries over into better results obtained from ML models. Since the housing prices are not publicly shared by the notaries, the best way to gather this data was by web scraping using the Beautiful Soap with Python programming language. This method was used to gather 36759 lines of housing data from a well-known real estate listing website (www.hepsiemlak.com) in Turkey.

3.1. Data Characteristics

The initially collected data consisted of 36759 lines of information on housing. However, the data was cleaned to get rid of heteroskedasticity and some meaningless publications on the website and the number of data used in the models amounted to somewhere under 36759 but over 20000.

The collected data includes 10 variables that are used in listing and also the listing price. Table 1 shows the initial variables that were scraped.

Table 1

Collected data characteristics

Feature Name	Description
District	Indicates the district where the house is located.
Neighborhood	Indicates in which neighborhood the house is located.
Room_Number	Indicates how many rooms the house has.
M_square	Indicates the square meters of living space the house has.
Which_flooritis	Indicates the floor the house is located on in the building.
Floor_number	Indicates total floor number of the apartment.
Building_age	Indicates the age of the house.
Heating_type	Indicates whether or not the house has a connection to natural gas.
Shower_numbers	Indicates the number of bathrooms in the house.
Elevator_exists	Indicates whether there is an elevator in the building.
Listing_price	Indicates the listing price of the house.

Source: own study.

The study assessed the houses in Istanbul sold publicly in December 2020. The total number of the collected housing data is 36759 before data manipulation. The variables in the publicly listed houses are used to predict the listing price of the houses. For policymakers, investors, banks and many other organizations, as well as individuals, it is crucial to perform the correct valuation. This study aims to explore the relationship between the listing variables and the listing price of the houses with some ML methods. The main aim is to create a house pricing prediction model using these listing variables. The listing prices of the houses are chosen as the dependent variable in this study and the rest of the variables are chosen as possible independent variables/features of the ML models that are created. In statistics, the regression problem is to predict the value of a new observation based on the similarities between new observations and existing ones. Therefore, the prediction of a housing price can be defined as a regression problem. During this study, polynomial regression, random forests, and gradient boosting methods are applied to solve this problem.

3.2. Data Manipulation

The initially collected dataset had over 36 thousand data including housing information of the houses that were currently being sold in Istanbul in November 2020. However, the data collection was made by the data scraping technique using the python programming language. This resulted in the collection of some data with missing information. Some rows did not have the district name, neighborhood name, square meter value, floor number, building age, number of rooms, number of living rooms, number of showers, the existence of an elevator, heating type, number of floors in the apartment, and/or price information. Healthy use of machine learning cannot proceed without a complete dataset. Therefore, to be able to transform the dataset into a usable form for machine learning algorithms, rows with missing values were deleted from the dataset.

After getting rid of rows with missing values, the column with the information on the presence of an elevator of the house being sold is observed. In the collected dataset, the listing was made as string type as “yes” or “no”. To be able to use it in the machine learning algorithms, this column was transformed into dummy values and added to the dataset, while the original elevator column was deleted. To make the data computationally friendly, only elevator existence is taken as the binary in the column.

The same procedure is followed for data regarding the heating type of the house being sold. The information on heating type fell into one of four techniques, i.e.: houses heated with oil, natural gas, electricity or coal. All this information was listed in a column as string type. This column was transformed into dummy variables and added to the dataset as four different columns.

Also, all 39 districts of Istanbul were transformed into dummy variables and added to the dataset as binary data. Similarly, 632 neighborhoods of Istanbul were transformed into dummy variables and added to the dataset as binary data, excluding one less variable for not having the dummy variable trap during the study. This procedure seemed to increase the performance of the machine-learning algorithms.

A similar procedure has been followed for the floor number column. But instead of creating dummy variables, floor numbers have been assigned as their counterpart values as integers. Based on the data analysis, it is assumed that the floor number data is ordinal for the price information.

After the technical correction of the data is complete, the logical correction of the dataset is commenced, with the dataset being checked after the technical correction. From over 36 thousand lines of data, 10 thousand had missing information. The distribution of the 26 thousand lines of data showed that most of the houses (over 23 thousand) were priced between 200 thousand Turkish Liras and 3 million Turkish Liras. To get better results from the machine learning algorithms, the density of the data should be higher. For this reason, price data which was over 3 million Turkish liras was erased from the dataset.

A similar procedure is done for the number of floors existing in the apartment, which the flat is located on; the number of living-room type rooms and the number of regular-type rooms.
This dataset also contained some NaN values, and these values were also erased.

After the manipulation of the dataset, 36759 unusable numbers of housing data decreased to 22176 meaningful, usable rows for the machine learning algorithms. A sample part of the manipulated data can be seen in Figure 2 below.

m_square	which_flooritis	floor_number	building_age	shower_numbers	yes_elevator	Oil	Natural_gas	Electric	Coal	room_number	saloon_number	Price
90	2	4	5	1	1	0	1	0	0	2	1	200000
65	1	5	30	1	1	0	1	0	0	2	1	200000
65	1	4	12	1	1	0	1	0	0	1	1	200000
68	4	16	3	1	1	0	1	0	0	1	1	200000
75	1	5	30	1	1	0	1	0	0	2	1	200000
130	1	6	20	1	1	0	0	0	1	3	1	200000
42	0	3	25	1	1	0	1	0	0	1	1	200000
95	9	5	7	1	0	0	1	0	0	2	1	200000
70	4	4	25	1	1	0	1	0	0	1	1	200000
72	6	30	8	1	0	0	1	0	0	1	1	200000
100	1	5	12	2	1	0	1	0	0	3	1	200000

Fig.2 . Sample of the manipulated data. *Source*: own study.

3.2. Data Analysis

The dataset has been analyzed to understand its nature. This procedure allows us to see the correlation between the variables and the price column. Unnecessary variables can therefore be removed from the dataset for computational ease, as well as to gain an understanding of the relationships and have an idea of how effective some variables in determining the price of the house.

A correlation matrix is a table form of the variables that displays the correlation coefficients between each of them. Every cell displays the correlation between two variables. The coefficient numbers are values between -1 to 1.A value closer to -1 means that the variables are negatively correlated, whereas closer to 1 means the variables are positively correlated with each other. However, not all the columns after the manipulation can be listed in the correlation matrix. Figure 3 shows the dataset without information regarding the district and neighborhood.

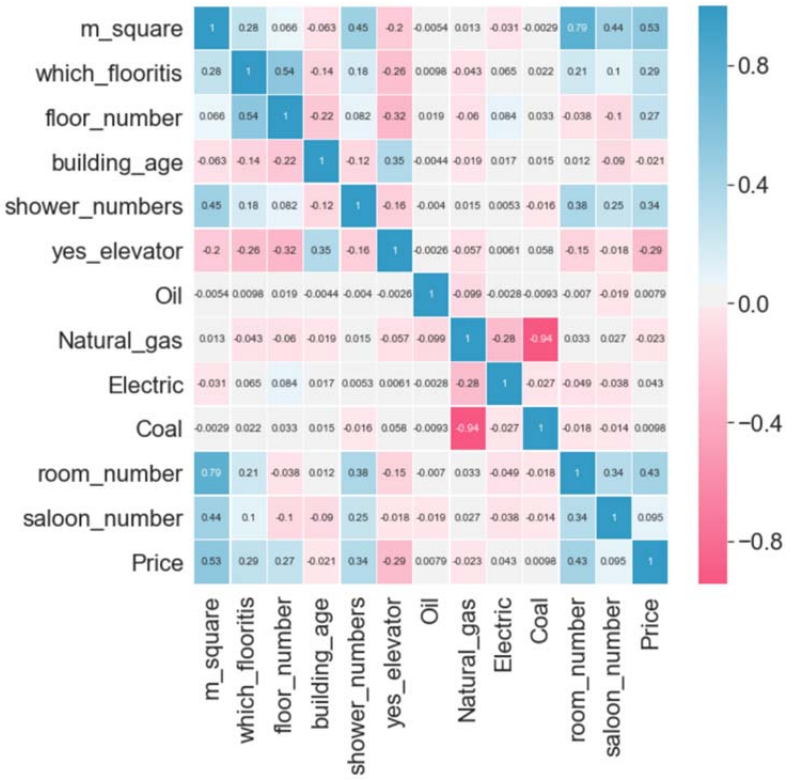


Fig.3. Correlation matrix of the features. *Source*: own study.

As can be seen from the table, the square meter variable, floor variable, number of showers variable, and the number of regular rooms variable are positively correlated with the listing price. On the other hand, the presence of an elevator is negatively correlated with the listing price. The building

age and heating type variables, as well as the number of living rooms seem to have no correlation with the listing price. This information gathered from the correlation matrix might be useful for the machine learning algorithms, as some of these variables can be removed from the dataset. To be able to determine which variables are not improving the performance of the algorithm significantly, a T-test can be applied. The T-test shows how significant the differences between the groups are.

Plot visualization of the data allows us to have some idea about how the variables are shaped and how to list them in our data frame. For that reason, some graphs have been created and discussed to understand the data better. Data regarding the floor the flat is located on is added as an ordinal variable to the dataset; however, if there was no ordinal relationship between the price and the floor location variable, this would affect the performance of the algorithms badly. Hence, the visual in Figure 4 allowed us to see that this variable is in fact ordinal.

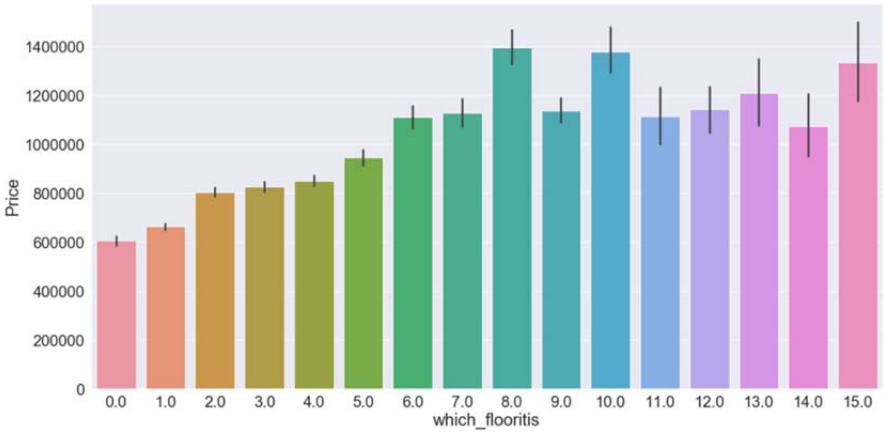


Fig. 4. Price graph according to floor information of the flats. *Source*: own study**.**

However, it can be seen in Figure 5, that this relationship cannot be seen in the total number of floor data in the dataset. Therefore, it might be harmful to algorithm performances to take this data as ordinal instead of binary.

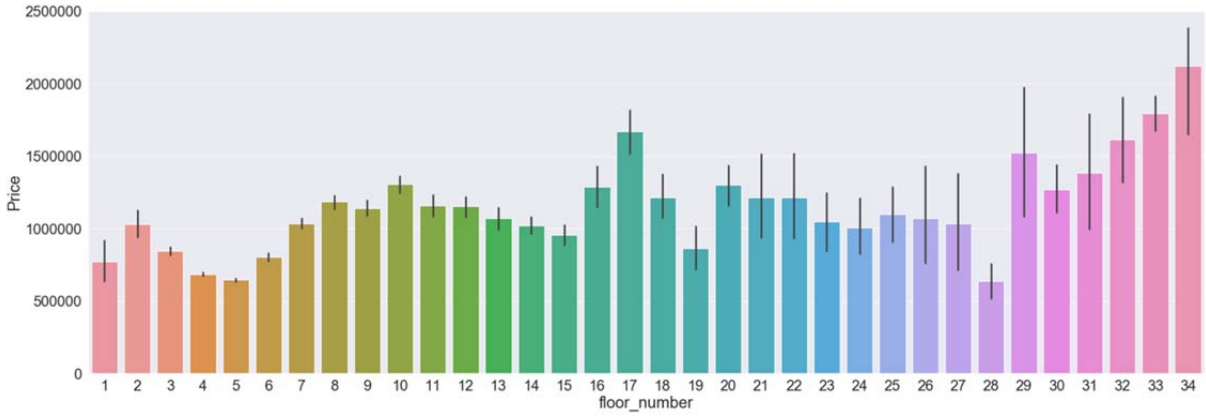


Fig. 5. Price graph according to floor information of the apartments. *Source*: own study.

From Figure 6, it is possible to see that most of the houses being sold generally contain 2 or 3 regular rooms. This seems to fit the normal distribution, with over 2134 dwellings with 1 regular room, 9474 dwellings with 2 regular rooms, 8722 dwellings with 3 regular rooms, 1697 dwellings with 4 regular rooms, and only 481 dwellings with 5 regular rooms.

According to the correlation matrix, it is possible to see that the number of rooms variable and the price variable have a positive correlation, as illustrated in Figure 7.

Similarly, living room-type rooms of flats seem to have a positive relationship with the price in Figure 8. Most of the data of the houses have 1 living room-type room, with over 21 thousand lines of

data; 1182 of them have 2 living room-type rooms and only 98 of them have no living room-type rooms.

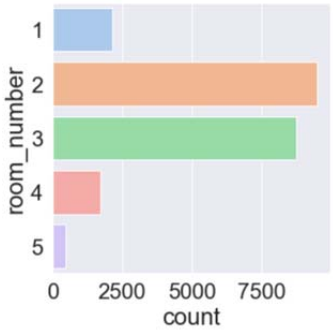


Fig. 6. Count graph of dwellings according to the number of rooms. *Source*: own study.

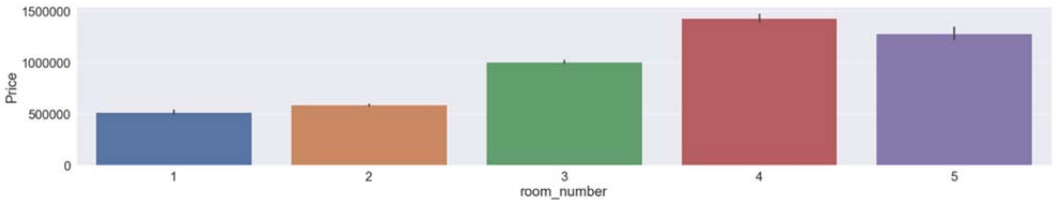


Fig.7. Price graph of the dwellings according to the number of rooms. *Source*: own study.

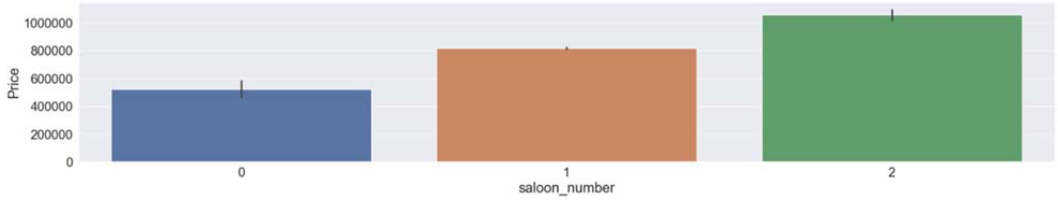


Fig. 8. Price graph of the dwellings according to the number of saloons. *Source*: own study.

The importance of the districts is significant for the listing price. Thus, including the districts as variables for predicting the listing price is necessary for the algorithms. The visualization of these districts in Figure 9according to the listing price helps us understand which districts most affect the price. According to the analysis, Sariyer, Kadikoy, Bakirkoy, and Adalar districts have the most expensive housing prices on average, whereas Silivri, Sultangazi, Esenyurt, and Sile districts are the least expensive locations.

Visualization of neighborhood variables was also carried,but including it is not possible since there are over 600 neighborhoods of Istanbul. When analyzed, it was observed that the Ayazaga, Burgazada, Tesvikiye, Bebek, Poligon, Emirgan, Senlikkoy, Acarlar, and Ulus neighborhoods are some of the most expensive locations according to the dataset. Malkocoglu, Sultaniye, Rami Cuma, and Kazim Karabekir neighborhoods are the least expensive locations when the data is analyzed according to the dataset.

The count plot visualization shows the number of houses being sold in each district. Kadikoy is the hotspot for the housing market of Istanbul according to our database with 2930 houses. Catalca is the district with the least houses being sold, i.e. only 6 houses. Districts with less than 100 houses sold were deleted from the secondly tried model to have better performance and will be discussed in the results of the algorithm section.

4. Results of The Algorithms

4.1. Linear and Polynomial Regression

Linear and polynomial regression are potentially two of the most well-known algorithms in machine learning. The linear regression model assumes the relationship between the dependent variable and independent variables is linear, and that the dependent variable can be calculated from the linear

combination of independent variables. Similarly, the polynomial regression model assumes the relationship between the dependent variable and independent variables is polynomial. Getting the most influential features for linear and polynomial regression is vital for computational reasons. To be able to do so forward feature selection algorithms are used with 5 cross-validation sets. The most influential 27 feature subset is created and used in the regression algorithms. These features include 8 of the original collected value features, such as meter square information of the house etc., and the rest of the 19 features regard district information.

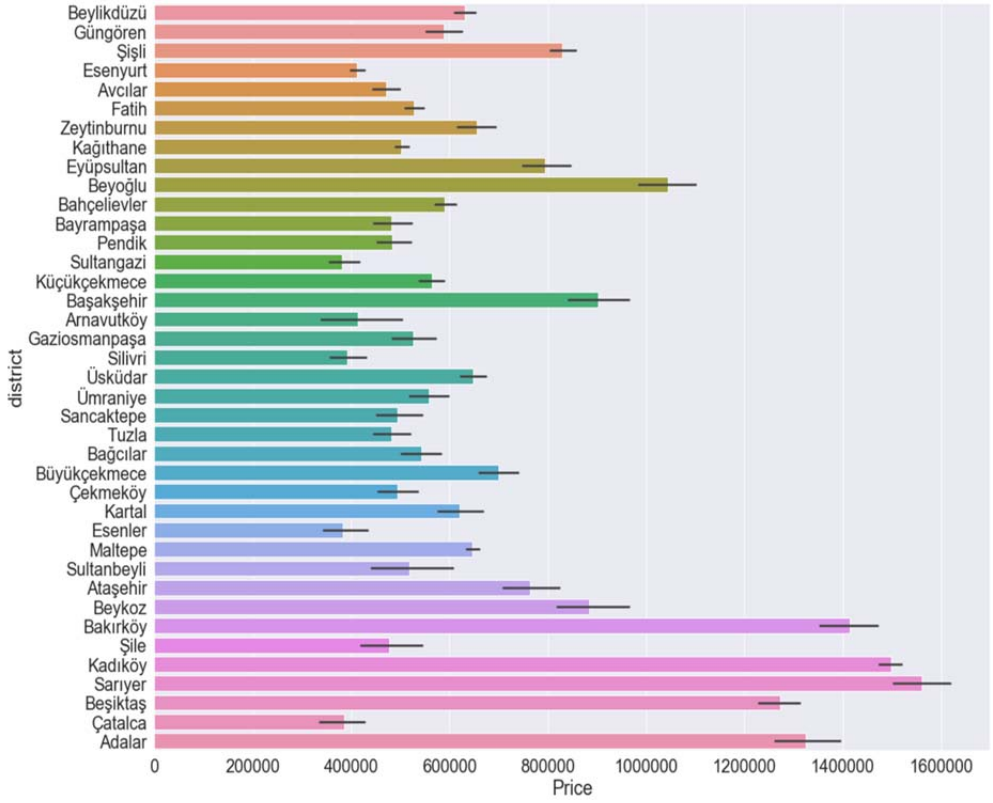


Fig. 9. Average price graph of dwellings according to district-based information. *Source*: own study.

After this procedure, the dataset is scaled with standardization and split into train and the test sets with the test size of 1/7 of the original dataset.

R square is selected as the performance indicator for the algorithm's performance. As it is used in the linear regression, a 62% r square score has been gathered whereas it goes up to 70% r square score in the polynomial regression with the degree of 2. Also, mean absolute percentage error (MAPE) is used as another performance indicator. Linear regression's MAPE score is 34.42%, whereas polynomial regression showed a superior performance with a 30.15% MAPE score.

	Actual	Predicted	Difference
count	3168.00	3168.00	3168.00
mean	806283.71	811981.40	216066.65
std	596691.85	514191.38	243326.60
min	200000.00	-1013738.00	38.00
25%	390000.00	440150.00	62090.00
50%	595000.00	650504.00	137397.50
75%	980000.00	1050946.50	282566.00
max	2995000.00	3551888.00	2151156.00

Fig.10. Outcome of the predicted houses using the polynomial regression model. *Source*: own study.

Even though the polynomial regression seemed to acquire a 70% prediction score because of the weights of the variables in the polynomial regression function, it is seen that the algorithm predicted negative prices. Overall polynomial regression seems like an okay running model for the prediction of houses but the performance is not sufficient.

Afterwards, the rows which belonging to Catalca, Sile, Sultanbeyli, and Arnavutkoy were deleted from the dataset because there was not enough data for the algorithms to make accurate predictions. Then, when the model was run once again, linear regression had an r square score of 63.5%, which is an over 1.5 percent improvement.

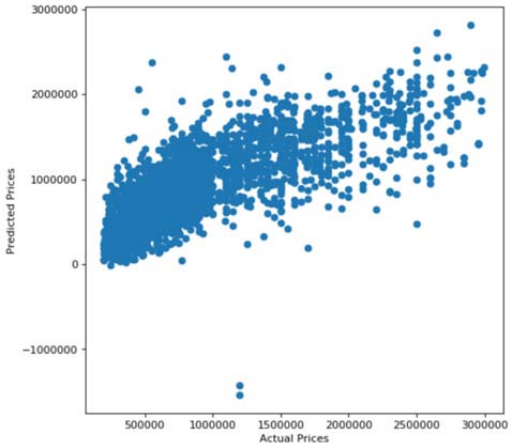


Fig. 11. Actual price and predicted price comparison graph of the linear regression. *Source*: own study.

However, the polynomial regression performing badly when these districts are removed from the dataset, with almost 10% down from the earlier model. 60% r square score is acquired from the polynomial regression model with these districts removed.

4.2. Tree Models

4.2.1. Decision Tree Regressor

The decision tree solves the machine learning regression problem by transforming the data into a tree structure. It requires less effort in the data preparation stage and is very intuitive. However, it is known that the decision tree algorithm is under performs in regression problems.

Firstly, the heating type features and the number of living-rooms information feature is deleted from the dataset as it is seen to be unnecessary according to the correlation matrix. These variables are deleted after being put into T-test and observed not to be improving the model accuracy significantly. Then the dataset is split into train and test subsets with 1/7 of it as the test set size. The MAPE score is used as a performance indicator in this model as well. Without any parameter set, the decision tree regressor runs a relatively good model with around 25.42% MAPE score.

	Actual	Predicted	Difference
count	3168.00	3168.00	3168.00
mean	820309.72	821467.96	193673.79
std	608826.69	618247.33	279546.04
min	200000.00	200000.00	0.00
25%	399000.00	385000.00	30000.00
50%	600000.00	600000.00	100000.00
75%	1000000.00	985000.00	240000.00
max	2995000.00	2995000.00	2300000.00

Fig. 12. Outcome of the predicted houses using the decision tree model. *Source*: own study.

Unlike polynomial regression, the decision trees did not predict negative housing prices in these models since its algorithm only allows for the categorization of existing prices and predicting the exact prices of some houses in the test set.

	Actual	Predicted	Difference
count	346.00	346.00	346.00
mean	1172517.34	1172517.34	0.00
std	742275.12	742275.12	0.00
min	210000.00	210000.00	0.00
25%	520000.00	520000.00	0.00
50%	1090000.00	1090000.00	0.00
75%	1695000.00	1695000.00	0.00
max	2995000.00	2995000.00	0.00

Fig. 13. Outcome of the perfectly predicted houses using decision tree model. *Source*: own study.

Also, the algorithm performed perfectly on 346 predictions of3168 predictions with 0 deviations from the original price of the house, which is over the 10% of the tested dataset.

Afterwards, the rows which belong to Catalca, Sile, Sultanbeyli, and Arnavutkoy were deleted from the dataset because there was not enough data for the algorithms to predict accurately enough. When the model is run once again, r square score has changed to 67%. However, this time the model perfectly predicted the prices of 379 houses in the dataset and 892 houses within the 30000 liras price range out of a total of3168 predictions.

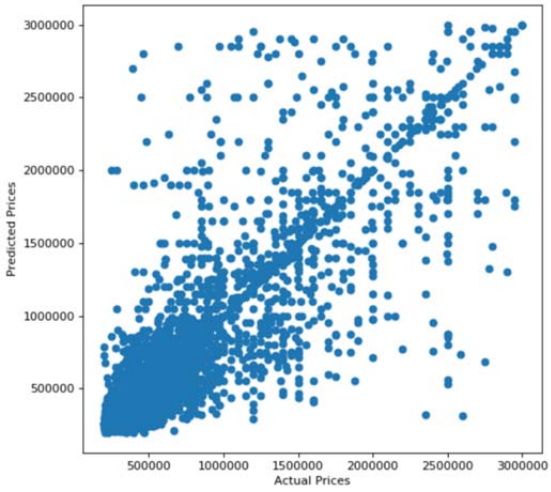


Fig. 14. Actual and predicted price comparison graph of the decision tree model. *Source*: own study.

The tree structure has not been included since the depth of the trees is too deep.

4.2.2. Random Forest Regressor

The Random Forest algorithm is an advanced version of the decision trees algorithm. It is less prone to overfit than other algorithms since it uses bagging. The algorithm selects random variables and creates hundreds of trees. Eventually, it combines all the trees resulting in increased accuracy. Random forest regressors' performance can be improved using its parameters, such as the number of estimators and the maximum features which will be used in the model. To be able to find the best parameters for our dataset, a subset of 36 models with a cross-validation value of 10 with possible parameters were investigated and the best performing model has been chosen among the360 possible models. The best-performing models have 1300 estimators and 36 features. This procedure took over 230 minutes.

	Actual	Predicted	Difference
count	3168.00	3168.00	3168.00
mean	820309.72	812139.81	155090.72
std	608826.69	515377.41	199034.22
min	200000.00	239562.22	0.00
25%	399000.00	439448.17	38532.44
50%	600000.00	631652.50	90206.85
75%	1000000.00	1019806.20	189133.57
max	2995000.00	2992027.78	2093009.56

Fig. 15. Outcome of the predicted houses using the first random forest model. *Source*: own study.

After the model is fit to the dataset and the model is run, we acquired an outperforming of20.58% of the MAPE score on this model which surpasses that of the models that had been discussed earlier.

The model predicted the exact prices of the 7 models in the dataset and predicted 651 house prices within the 30000 liras price range. Overall, it outperformed all of the models discussed above.

	Actual	Predicted	Difference
count	651.00	651.00	651.00
mean	721419.20	723022.33	14197.24
std	518744.82	517051.30	9013.17
min	220000.00	240320.00	0.00
25%	375000.00	374177.05	5998.46
50%	520000.00	522601.92	14175.38
75%	832500.00	848560.43	22057.31
max	2995000.00	2992715.38	29964.23

Fig. 16. Outcome of the predicted houses within the 30000 lira range using first random forest model. *Source*: own study.

The rows which are belonging to Catalca, Sile, Sultanbeyli, and Arnavutkoy have been deleted from the dataset because there was not enough data to allow for sufficient predictions of the algorithms. After this event, the prediction score has improved from 20.58%to 20.56%. The model predicted 651 houses within the 30000 liras price range same as the model without deleted features out of a total of 3168 predictions.

4.2.3.Xgboost

The XGBoosting algorithm is another advanced tree algorithm. Similar to the random forest, it creates trees from randomly selected variables, but instead of combining all the trees at the end of the process, it sequentially adds trees to increase the accuracy. XGBoosting algorithm is first run before deleting the districts with insufficient data in the dataset. To be able to find the best parameters for the XGboost algorithm, a grid search is run with 72 different models with a cross-validation value of 3. The searched parameters were subsample ratio of all the features of the dataset when constructing each tree, regularization terms on weights, maximum depth of the trees, and the number of estimators for each model. After a computationally expensive 6-hour search, best-performing parameters are selected and then the model is applied for prediction. The MAPE score of the model came out as 21.81%, outperforming linear regression, polynomial regression, and the decision tree models and performing similarly with the random forest algorithm.

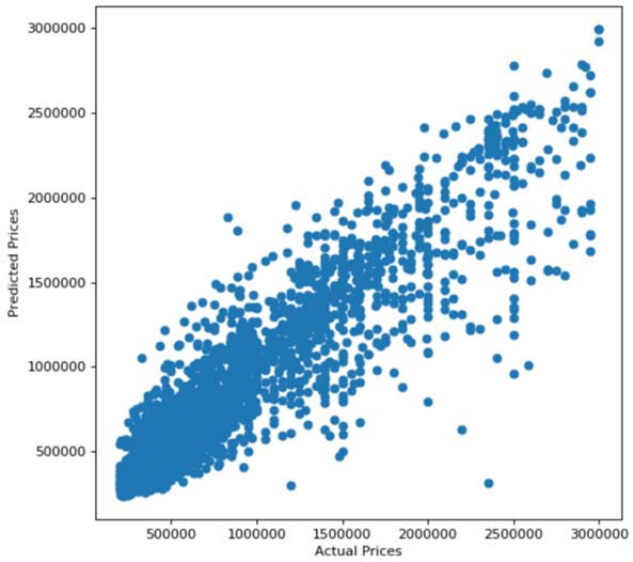


Fig. 17. Actual and predicted price comparison graph of the second random forest model. *Source*: own study.

	Actual	Predicted	Difference
count	3150.00	3150.00	3150.00
mean	828434.63	821370.94	148497.59
std	604108.06	521379.42	187577.48
min	200000.00	238726.74	0.00
25%	400000.00	436585.38	36082.56
50%	610000.00	639175.90	87545.19
75%	1050000.00	1064592.60	182441.71
max	2995000.00	2994837.18	2033079.62

Fig. 18. Outcome of the predicted using second random forest model. *Source*: own study.

	Actual	Predicted	Difference
count	3168.00	3168.00	3168.00
mean	806283.71	811003.06	152709.77
std	596691.85	550582.50	198914.73
min	200000.00	122959.21	36.98
25%	390000.00	420944.09	37871.42
50%	595000.00	608389.06	90011.88
75%	980000.00	1026695.19	187800.53
max	2995000.00	3062752.75	2091142.31

Fig. 19. Outcome of the predictions using first XGboost model. *Source*: own study.

The firstly run model successfully predicted 641 houses with 30000 lira range over 3168 predictions.

Afterward, the districts with insufficient data are deleted from the dataset and the model is run once again with a better r square score of 83.1%. This is proving that, in ensemble models, clearing the insufficient data from the dataset improves the performance of the model.

	Actual	Predicted	Difference
count	3150.00	3150.00	3150.00
mean	808746.49	820950.81	154625.10
std	591813.27	562171.31	187584.14
min	200000.00	187064.08	52.25
25%	389250.00	418561.29	39939.28
50%	598000.00	607561.59	94015.67
75%	998000.00	1040964.17	191224.50
max	2995000.00	3061608.75	1971437.00

Fig. 20. Outcome of the predictions using second XGboost model. *Source*: own study.

This time the predictions within the 30000 liras price range got down to 598 out of 3150 predictions. The comparison graph can be seen below.

5. Conclusion

Istanbul has one of the most active housing markets, with over 260 thousand dwellings sold in 2020 according to the Turkish Statistical Institute. This makes the decision-making process of many people and corporations extremely hard when it comes to buying a house or property. Before buying a dwelling, several features must be considered.

This study aims to build machine learning models using different algorithms with improved parameters for predicting listing prices of the real estate market in Istanbul, Turkey.

To be able to build a highly accurate model, there needs to be a sufficient amount of data for the researcher. Many studies acquired some data from real estate agents or other institutions. In this study, the studied data is taken from a public website and manipulated. After the manipulation, the dataset is tried on different machine learning algorithms and models.

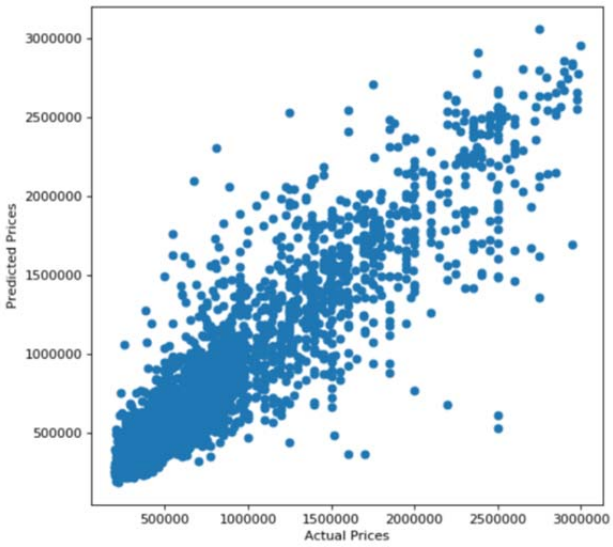


Fig. 21. Actual and predicted price comparison graph of the second Xgboost model. *Source*: own study.

In this study, 5 different algorithms are used and over 100 models are created. Linear regression, polynomial regression, decision trees, random forests, and XGboost algorithms are used for creating

these models. Linear regression, polynomial regression, and decision trees performed poorly compared to ensemble algorithms. The study showed that the random forests model and the XGboost model outperform the other models with a similar mean absolute percentage error score of around 20%. This study also showed the parameters of the models should be set accordingly to the dataset. The improvements from the parameters are average around 4% to 6%. It is possible to see the comparison of MAPE and r square scores of the models in the table below.

Table 2

Comparison of algorithms

Algorithm	Mean Absolute Percentage Error	R square
Linear Regression	34.42%	62%
Polynomial Regression	30.15%	70%
Decision Trees	25.42	
Random Forest	20.56%	
XGBoost	21.81%	

Source: own study.

Because the data is collected using data scraping, although nearly half of the firstly collected 36000 lines of data is cleansed during the data manipulation part, outliers without explanations still remain. This might be caused by randomly placing a listing in the public website, or there might still be many features not listed in the dataset, such as the safety score of the neighborhood, view score of the house, etc.

The authors admit that there are limitations to this study. The models are only trained with the data collected in December 2020 for the houses being sold in Istanbul. The housing market prices keep surging along with inflation and many other factors. Future work can be made by considering building a dynamic model that keeps updating the dataset and the models with an updated dataset from each month, with more features and more data.

6. References

Abidoye, R., & Chan, A. (2017). Modeling property values in Nigeria using artificial neural network. *Journal of Property Research, 34*, 36–53. https://doi.org/10.1080/09599916.2017.1286366

Abraham, M. (2016). *Determinants of residential property value in New Zealand: a neural network approach*. Department of Applied Business, New Zealand Government Institute of Technology. Whitireia.

Breiman, L. (2001). Machine Learning. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Case, B., Clapp, J., Dubin, R., & Rodriguez, M. (2004). Modeling spatial and temporal house price patterns: A comparison of four models. *The Journal of Real Estate Finance and Economics, 29*, 167–191. Advance online publication. https://doi.org/10.1023/B:REAL.0000035309.60607.53

Cebula, R. (2009). The hedonic pricing model applied to the housing market of the city of Savannah and its Savannah Historic Landmark District. *The Review of Regional Studies, 39*(1), 9–22.

Central Bank of the Republic of TurkeyResidential Property Price Index Report (October 2020). https://www.tcmb.gov.tr/wps/wcm/connect/21c8c007-4006-45ee-bbc2-852f396a23f0/RPPI.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-21c8c007-4006-45ee-bbc2-852f396a23f0-npHk6EW

Ecer, F. (2014). Comparision of hedonic regression method and artificial neural networks to predict housing prices in Turkey. International Conference On Eurasian Economies, 1-10.

Eriki, P., & Udegbunam, R. (2007). Application of neural network in evaluating prices of housing units in Nigeria: a preliminary investigation. *Journal of Artificial Intelligence, 1*(1), 21-27. https://doi.org/10.3923/jai.2008.21.27.

Fan, G.-Z., Ong, S. E., & Koh, H. C. (2006). Determinants of House Price: A DecisionTreeApproach. *Urban Studies (Edinburgh, Scotland), 43*(12), 2301–2315. https://doi.org/10.1080/00420980600990928

Fix, J. F., Frezza-Buet, H. F. B., Geist, M. G., & Pennerath, F. P. (2019). *Machine learning* (Revised ed.). CentraleSupélec., http://sirien.metz.supelec.fr/depot/SIR/CoursML/Poly-ML-SIR.pdf

Górak, M. (2017). Employing linear artificial neural networks in property appraisal and valuation - possible applications. *Geomatics, Land Management and Landscape, 1*(1), 17–24. https://doi.org/10.15576/GLL/2017.1.17

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2016). *An introduction to statistical learning: with applications in R* (Springer Texts in Statistics). 1st ed. 2013, Corr. 7th printing 2017 ed.. Springer.

Khalafallah, A. (2008). Neural network based model for predicting housing market performance. *Tsinghua Science and Technology, 13*, 325–328. https://doi.org/10.1016/S1007-0214(08)70169-X

Kuşan, H., Aytekin, O., & Özdemir, İ. (2010). The use of fuzzy logic in predicting house selling price. *Expert Systems with Applications, 37*, 1808–1813. https://doi.org/10.1016/j.eswa.2009.07.031

Liu, L., & Wu, L. (2020). Predicting housing prices in China based on modified Holt's exponential smoothing incorporating whale optimization algorithm. *Socio-Economic Planning Sciences, 72*, 100916. https://doi.org/10.1016/j.seps.2020.100916

Liu, R., & Liu, L. (2019). Predicting housing price in China based on long short-term memory incorporating modified genetic algorithm. *Soft Computing, 23*, 11829–11838. Advance online publication. https://doi.org/10.1007/s00500-018-03739-w

Mimis, A., Rovolis, A., & Stamou, M. (2013). Property valuation with artificial neural network: The case of Athens. *Journal of Property Research, 30*, 128–143. Advance online publication. https://doi.org/10.1080/09599916.2012.755558

Mora-Esperanza, J. G. (2004). Artificial intelligence applied to real estate valuation: An example for the appraisal of Madrid. *Catastro, 1*, 255–265.

Nghiep, N., & Al, C. (2001). Predicting housing value: a comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research, 22*, 313–336. https://doi.org/10.1080/10835547.2001.12091068

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications, 42*, 2928–2934. Advance online publication. https://doi.org/10.1016/j.eswa.2014.11.040. Erratum: https://doi.org/10.1016/j.eswa.2015.03.005.

Que, A. Q. (2015). *Mathematics of polynomial regression*. Polynomial Regression. http://polynomialregression.drque.net/math.html

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

Thanasi (Boçe), M. (2016). Hedonicappraisal of apartments in Tirana. *International Journal of HousingMarkets and Analysis., 9*, 239–255. https://doi.org/10.1108/IJHMA-03-2015-0016

Wang, X., Wen, J., Zhang, Y., & Wang, Y. (2014). Real estate price forecasting based on SVM optimized by PSO. *Optik (Stuttgart), 125*, 1439–1443. https://doi.org/10.1016/j.ijleo.2013.09.017

Yayar, R., & Gül, D. (2014). Mersin kent merkezinde konut piyasası fiyatlarının hedonik tahmini [Hedonic estimation of housing market prices in Mersin city center]. *Anadolu University Journal of Social Sciences Anadolu Üniversitesi Sosyal Bilimler Dergisi., 14*, 87–100.

Yılmazel, Ö., Afşar, A., & Yılmazel, S. (2018). Konut Fïyat Tahmïnïnde Yapay Sïnïr Ağlari Yöntemïnïn Kullanilmasi. *International Journal of Economic &Administrative Studies, 20*, 285–299. https://doi.org/10.18092/ulikidince.341584

Zurada, J., Levitan, A., & Guan, J. (2006). Non-conventional approaches to property value assessment. *Journal of Applied Business Research, 22*, 1–14. https://doi.org/10.19030/jabr.v22i3.1421