DOI: 10.47899/ijss.1270433

The Effect of Outlier Detection Methods in Real Estate Valuation with Machine Learning*

Makine Öğrenimi İle Mülk Değerlemesinde Aykırı Değer Tespit Yöntemlerinin Etkisi

Cihan ÇILGIN[1], Yılmaz GÖKŞEN[2], Hadi GÖKÇEN[3]

Abstract

For those who invest in real estate as an investment tool, as well as those who buy and sell real estate, the price of real estate should be predicted realistically and with the highest accuracy. It should be noted that the predict model should be the most appropriate representation of the underlying fundamentals of the market. Otherwise, the mistake to be made in the real estate valuation will cause some undesirable results such as inconsistent and unhealthy increase or decrease of the property tax, excessive gains or losses in favor of some groups, and adverse effects on investors and potential real estate owners. At this point, data-driven real estate valuation approaches are preferred more frequently to create highly accurate and unbiased estimates. However, the consistency, precision and accuracy of the models realized with machine learning approaches are directly related to the data quality. At this point, the effects of outlier detection on prediction performance in real estate valuation are investigated with a large data set obtained in this study. For this purpose, a heterogeneous data set with 70.771 real estate data and 283 variables, 4 different outlier detection methods were tested with 3 different machine learning approaches. The empirical findings reveal that the use of different outlier detection approaches increases the prediction performance in different ranges. With the best outlier detection approach, this performance increase was at a high 21,6% for Random Forest, with a 6,97% increase in average model performance.

Keywords: Real Estate Valuation, Machine Learning, Outlier Detection, House Price Precition, Data Preprocessing

Öz

Konut alanlar ve satanlar kadar bir yatırım aracı olarak konut üzerinden yatırımda bulunanalar için de konut fiyatının gerçekçi ve en yüksek doğrulukta tahmin edilmesi gerekmektedir. Tahmin modelinin, piyasanın altında yatan temellerin en uygun temsili olması gerektiği unutulmamalıdır. Aksi takdirde konut değerlemesinde yapılacak hata emlak vergisinin tutarsız ve sağlıksız artırılması veya azaltılması, bazı gruplar lehine aşırı kazanç veya kayıp ve yatırımcılar ile potansiyel konut sahiplerini olumsuz etkilemesi gibi bazı istenmeyen sonuçlara neden olacaktır. Tam bu noktada günümüzde veri odaklı konut değerleme yaklaşımları yüksek doğrulukta ve önyargısız tahminler oluşturmada daha sık tercih edilmektedir. Fakat makine öğrenmesi yaklaşımları ile gerçekleştirilen modellerin tutarlılığı, kesinliği ve doğruluğu veri kalitesi ile doğrudan bağlantılıdır. Bu noktada bu çalışmada elde edilen geniş bir veri seti ile konut değerlemede özellikle aykırı değer tespitinin tahmin performansı üzerine etkileri araştırılmaktadır. Bu amaçla 70.771 konut verisi ve 283 adet değişkene sahip hetorejen bir veri seti ile 4 farklı aykırı değer tespiti yöntemi 3 farklı makine öğrenmesi yaklaşımı ile test edilmiştir. Elde edilen ampirik bul gular farklı aykırı değer tespiti yaklaşımlarının kullanılmasının tahmin performansını farklı aralıklarda artığını ortaya koymaktadır. En iyi aykırı değer tespiti yaklaşımı ile ortalama model performansında % 6,97'lik bir artışla birlikte Rastgele Orman için bu performans artışı % 21,6'lık yüksek bir oranda gerçekleşmiştir.

Anahtar Kelimeler: Mülk Değerleme, Makine Öğrenmesi, Aykırı Değer Tespiti, Konut Fiyat Tahmini, Veri Ön işleme

* In this article, the principles of scientific research and publication ethics were followed. This article was produced from the doctoral thesis carried out by Cihan Çılgın at Gazi University Informatics Institute, Department of Management Information Systems. / Bu makalede bilimsel araştırma ve yayın etiği ilkelerine uyulmuştur. Bu makale Cihan Çılgın tarafından Gazi Üniversitesi Bilişim Enstitüsü Yönetim Bilişim Sistemleri Anabilim Dalı'nda gerçekleştirilen doktora tezinden üretilmiştir.

[1] **Cihan Çılgın,** **ORCID ID:** 0000-0002-8983-118X
Arş. Gör., Bolu Abant İzzet Baysal Üniversitesi, Gerede Uygulamalı Bilimler Fakültesi, Yönetim Bilişim Sistemleri, Bolu, Türkiye. cihancilgin@ibu.edu.tr
Ress. Asst., Bolu Abant İzzet Baysal University, Faculty of Applied Sciences, Management Information Systems, Bolu, Türkiye. cihancilgin@ibu.edu.tr

[2] **Yılmaz Gökşen,** **ORCID ID:** 0000-0002-2291-2946
Prof. Dr., Dokuz Eylül Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Yönetim Bilişim Sistemleri, İzmir, Türkiye. yilmaz.goksen@deu.edu.tr
Prof. Dr., Dokuz Eylül University, Faculty of Economics and Administrative Sciences, Management Information Systems, İzmir, Türkiye. yilmaz.goksen@deu.edu.tr

[3] **Hadi Gökçen,** **ORCID ID:** 0000-0002-5163-0008
Prof. Dr., Gazi Üniversitesi, Mühendislik Fakülte, Endüstri Mühendisliği, Ankara, Türkiye. hgokcen@gazi.edu.tr
Prof. Dr., Gazi University, Engineering Faculty, Industrial Engineering, Ankara, Türkiye. hgokcen@gazi.edu.tr

Geliş Tarihi/*Received* : 24.03.2023
Kabul Tarihi/*Accepted* : 28.03.2023
Çevrimiçi Yayın/*Published* : 30.04.2023

Makale Atıf Önerisi /Citation (APA):
Çılgın, C., Gökşen, Y., Gökçen, H. (2023). The Effect of Outlier Detection Methods in Real Estate Valuation with Machine Learning. *İzmir Sosyal Bilimler Dergisi*, 5(1), 9-20. DOI: 10.47899/ijss.1270433

1. INTRODUCTION

Residential property is a long-term investment type where real estate investors generally hold regular or value increase and return for the future, investing to compensate for the invested capital (Shapiro et al., 2019:11). But a household or real estate owner needs to collect and process a lot of information in order to make the real estate and real estate market choices that maximize utility or profit (Hårsman and Quigley, 1991: 2). For this reason, the need for an impartial, objective and systematic evaluation of the house, which is a real estate, (Zurada et al., 2011: 349; Alexandridis et al., 2019: 1770; Peter et al., 2020: 2918) is an important phenomenon when the constant change in real estate prices is taken into account, and it is a phenomenon that closely concerns many stakeholders in this field (Janssen et al., 2001: 342). The real estate market is one of the markets that focuses on pricing the most and is needed among all the goods and services markets. Real estate valuation or housing price prediction can basically be defined as a regression task used to measure the value consumers attribute to real estates using objective data. The price prediction problem can be viewed as a regression problem where the dependent variable is the market value of a real estate and the independent variables are real estate characteristics such as size, age, number of bedrooms, etc. Considering the market value and characteristics of a large number of real estates, the aim is to obtain a function that relates the metadata of a real estate to its value (Poursaeed et al., 2018: 668).

For those who invest in real estate as an investment tool as well as those who buy and sell real estate, the price of real estate should be estimated realistically and with the highest accuracy (Bin, 2004: 69; Daşkıran, 2015). High-accuracy predictin of real estate prices has an important role in the decision of stakeholders to realize the potential of their investments (Kouwenberg and Zwinkels, 2014: 416). While investors in large pools of asset-backed securities cannot personally examine each asset, they want to have the comfort of knowing that these assets are valued reliably (Zurada et al., 2011: 350). In addition, as revealed by Gilbertson and Preston (2005), the type of real estate valuation methods adopted in a country can even affect the country's economy. For this reason, legal practices and concepts, environmental conditions (Küçükkaplan & Aldı, 2017), the model to be used should reflect the market culture and conditions during the valuation in the most accurate way. It should be noted that the model should be the most appropriate representation of the underlying fundamentals of the market (Pagourtzi, et al., 2003). Otherwise, the mistake to be made in the real estate valuation will cause some undesirable results such as inconsistent and unhealthy increase or decrease of the property tax, excessive gains or losses in favor of some groups, and adverse effects on investors and potential real estate owners.

At this point, data-driven real estate valuation approaches are preferred more frequently to create highly accurate and unbiased predictions. With these developments, the use of data-driven machine learning approaches that can learn the relationship or patterns between inputs and outputs and make inferences in order to minimize human involvement and prejudices in real estate valuations and improve the accuracy of residential real estate price prediction models is becoming more remarkable today. Especially in parallel with the developments in the volume and speed of information processing, the developments in the Machine Learning approach, which is a sub-field of Artificial Intelligence, triggered this situation more. In parallel with all these developments, a wide variety of machine learning approaches have been used by many researchers in the task of residential real estate price prediction. Random Forest(Aydemir et al., 2020; Yilmazer & Kocaman, 2020; Gupta et al., 2021; Tchuente & Nyawa, 2021; Bilgilioğlu & Yılmaz, 2021; Kim et al., 2021; Steurer et al., 2021; Yazdani, 2021; Imran et al. , 2021; Truong et al., 2020; Ho et al., 2021; Bergadano et al., 2021; Jui et al., 2020; Fu, 2018; Alkan et al., 2022), Support Vector Regression (Yacim and Boshoff, 2020; Manasa et al., 2020; García-Magariño et al., 2020; Pai and Wang, 2020; Tchuente and Nyawa, 2021; Bilgilioğlu and Yılmaz, 2021; Imran et al., 2021; Chou et al., 2022 ; Ho et al., 2021; Alkan et al., 2022), Decision Trees (Sawant et al., 2018; Aydemir et al., 2020; Pérez-Rave et al., 2020; Pai and Wang, 2020; Alfaro-Navarro et al., 2020; Mrsic et al., 2020; Bilgilioğlu and Yılmaz, 2021; Sing et al. , 2021; Sangha, 2021; Büyük and Ünel, 2021; Chou et al., 2022; Shi et al., 2022), Neural Networks (Štubňová et al., 2020; Yacim and Boshoff, 2020; Pai and Wang, 2020; Lee and Park, 2020; García-Magariño et al., 2020; Sevgen and Aliefendioğlu, 2020; Mankad, 2021; Rampini and Cecconi, 2021; Tchuente and Nyawa, 2021; Torres-Pruñonosa et al., 2021; Bilgilioğlu and Yılmaz, 2021; Kalliola et al., 2021; Steurer et al., 2021; Sa'at et al., 2021; Terregrossa and Ibadi, 2021; Tabar et al., 2021 ; Abhyankar and Singla, 2021; Yazdani, 2021; Chou et al., 2022; Seya and Shiroi, 2022), K-Nearest Neighbor (Zhao et al., 2019; Yıldırım, 2019; Mrsic et al., 2020; García-Magariño et al., 2020; Tchuente and Nyawa, 2021; Bergadano et al., 2021; Yazdani, 2021; Alkan et al., 2022), Gradient Boosting (Walthert and Sigrist, 2019; Truong et al., 2020; Mrsic et al., 2020; Manrique et al., 2020; Imran et al., 2021; Ho et al., 2021 ; Sangha, 2021; Bergadano et al., 2021) are just some of these approaches.

Consistency, precision and accuracy of models realized with machine learning approaches are directly related to data quality. At this point, in order to accurately prediction the real value of the residential real estate, it is imperative to support the obtained data with accurate and appropriate models and methods (Küçükkaplan & Aldı, 2017; Almond et al., 1997: 2). The fundamentals of real estate valuation or price prediction are also directly related to the collection, analysis and interpretation of comparable data (McGreal et al., 1998: 58). At this point, data quality and data preprocessing steps are of critical importance. Appropriate and careful data preparation is one of the most time-consuming and direct factors on performance in the use of machine learning methods. Model performance is affected not only by the performance of the methods, but also by the quality of the data set and its ability to represent the final goal to be achieved. The applicability of machine learning applications, which offer an alternative approach for smart system design in the real estate valuation process, can only be guaranteed when a suitable large pool of transaction data is available to work with and this data set is prepared under appropriate conditions. Although dataset size is an important factor, completeness and representativeness are even more important. The more appropriate representative samples the data set contains, the more robust (Kalliola et al., 2021: 2) and high predictive performance models can be created. However, data preprocessing steps are often ignored or not given special attention, except for a small number of studies in the field of real estate valuation (Sandbhor and Chaphalkar, 2019; Jha et al., 2020; Sing et al., 2021; Sisman and Aydinoglu, 2022).

At this point, the effects of outlier detection in residential real estate valuation are investigated with a large data set obtained in this study. For this purpose, 4 different outlier detection methods were tested with 3 different machine learning approaches with a heterogeneous data set with more than 70 thousand residential real estate data and 283 variables. The empirical findings reveal that the use of different outlier detection approaches increases the prediction performance in different ranges. In addition, this situation is a proof that increasing the time allocated to the stages that increase data quality, such as outlier detection, which is a data preprocessing process, will increase the success in the real estate valuation task.

In the following sections of the study, the data and preprocessing steps used in the second part of the study are presented. Information on machine learning approaches and outlier detection methods used in the price prediction task is given in Chapter 3. In Chapter 4, the implementation steps and findings with these methods are presented. The last part of the study is the conclusion part, which includes the discussion of the findings.

2. DATA

Within the scope of this study, Ankara, the capital of Turkey, which has a residential real estate stock of nearly 3 million, has been discussed. Between 02.12.2021 and 10.07.2022, 159.244 advertisement data were obtained from real estate sites with a web scraper developed. As a result of removing the missing and inconsistent data and repetitive data from the obtained data, 72.873 flats, 2.157 villas, 40 detached houses, 585 residence flats were obtained. In this study, only flats and residences are considered. As a result, after a general data preprocessing process, there is a total of 73.458 residential real estate data. 156 variables were obtained, including the total price information of these real estates. Variables such as *"Announcement Number"* and *"Real Estate Type"*, which do not represent any information in the real estate price prediction task, have been removed from these variables. In addition, a total of 82 variables were added to the data set, with 4 air quality features on a district basis, 12 demographic characteristics on a district basis, and 66 features related to proximity to various points (bus stops, schools, hospitals, banks, etc.). In addition to all these, in order to perform healthier analyzes within the data obtained from 25 different districts, data with less than 100 observations on a district basis were removed from the data set. For this reason, the data of Bala, Beypazarı, Çamlıdere, Elmadağ, Evren, Haymana, Kahramankazan, Kalecik, Kızılcahamam, Nallıhan, Şereflikoçhisar districts, which have less than 100 housing data, have been removed. In the new data set obtained as a result of this process, the distribution of the data on the residences by district is shown in Figure 1.

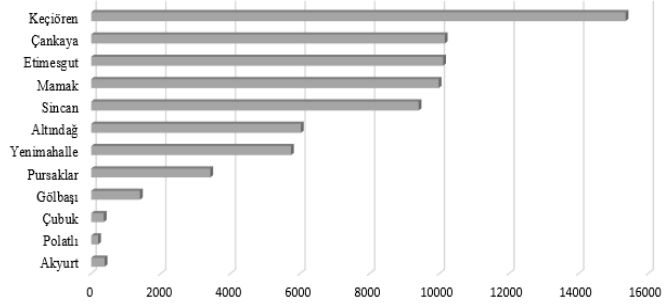

Figure 1. Number of *real estates* by districts

In addition to the monthly and district-based cumulative distribution of real estate prices, the statistical distribution is an exploratory data analysis that contains very important information about the obtained data set. For this reason, the price distributions of the flat data set are given in Figure 2. As can be seen in Figure 2(a), there is a high level of right

skew in the price variable. The main reason for this situation is that the deviation of some flat prices from the average, albeit a small number, is very high. Figure 2(b) shows a clearer conclusion about the effects of this situation. Figure 2 (b) shows the distribution of houses with a price of less than 5 million TL, excluding the data for only 479 houses. As it can be understood from here, the skewness of the flat price distribution has decreased significantly. Although Figure 2(b) is less right-skewed than Figure 2(a), the deviation from the mean is very significant. The most obvious conclusions that can be drawn from this figure are that the dataset clearly has outliers and the residential real estate price data needs various transformations under the assumption of linearity when linear models are considered.

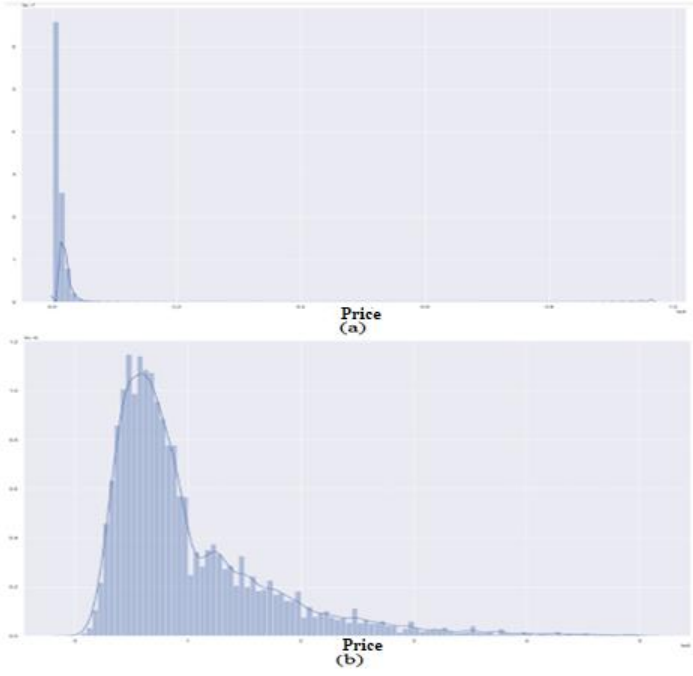

Figure 2. Distribution of flat prices

In addition, it is possible to comment on some variables and variable values by looking at the explanatory statistics of the variables in general. In particular, although some continuous variables have healthy mean and standard deviation values, this is not the case for a few continuous variables. The most obvious example of this situation is seen in the *"Net Square Meter"* and *"Gross Square Meter"* variables, which show the size of the house. The maximum values of both variables are observed as 995 and 869, respectively. In addition, the minimum value of both variables was found to be 1 square meter. These findings suggest that outliers need to be excluded according to a clear bound on the values of these variables. A similar situation is also valid for the dependent variable "Real Estate Price" as mentioned before. This variable, which has a very high standard deviation value, has very unrealistic values in terms of minimum values of 1,330 TL and maximum values of 97,000,000 TL. As in the *"Net Square Meter"* and *"Gross Square Meter"* variables, an outlier analysis should be made on this variable as well. In addition, the significant differences between the value ranges of the continuous variables examined in this context reveal the necessity of scaling the features especially within the scope of these variables. In addition to all these, all categorical variables in the variable set were converted into binary variables using the One Hot Encoding method. This means that any N-valued categorical variable will be converted into a total of N binary variables. Thus, a total of N new binary variables are created, one for each possible category in a variable, and these new variables take the values of one and zero. For this purpose, the categorical variables *"Real Estate Type"*, *"Building / Residence Age"*, *"Heating Type"*, *"Usage Condition"*, *"Sold By Who"*, *"District"*, *"Year"* and *"Month"* are converted into a new variable that takes each category value into binary value by using One Hot Encoding. Although the variables *"Real Estate Loan Opportunity"* and *"Furniture Status"* are categorical variables, no conversion was needed as they are currently only binary variables. A total of 283 variables have been reached in the new real estate dataset obtained as a result of these transformations.

3. METHOD

The methods used in this study should be evaluated under two different headings. The first of these is the methods used in the detection of outliers, while the other is the machine learning approaches used in the real estate price prediction task.

3.1. Outlier Detection Methods

Identification of outliers in a data set is of critical importance in terms of both improving the quality of the data and reducing the effect of outliers in the process of knowledge discovery from data. Such outliers can complicate the process of discovering useful patterns during data analysis, but they must also be detected for more consistent and reliable information. These outliers can be isolated with a wide variety of methods and these data can be analyzed externally (Rahman et al., 1998: 23). Outliers are observations that clearly differ from other observations in the same dataset and raise doubts about the source and originality of the data (Barnett and Lewis, 1984: 4). Outliers may occur due to various malfunctions, changes in the behavior of the system under study, fraudulent behavior, human error, data recording errors or simply natural deviations in sampling (Hodge and Austin, 2004: 85). Outliers are a very important step in the data preprocessing process, which directly affects the data

quality and indirectly affects the performance of the prediction models. The main purpose of outlier detection and removal from the dataset is to narrow the range of the dataset to make it suitable for producing better predictive results. Especially for residential real estate price prediction models where heterogeneity is a big problem and directly affects performance, the detection of outliers is very important to obtain a homogeneous data set.

There are many methods available on the detection of outliers, with both statistical and supervised or unsupervised learning. For this reason, in this study, various methods that are frequently preferred especially in the residential real estate valuation literature are discussed.

3.1.1. Interquartile range method

This method, also called the box chart or Tukey method, basically aims to determine the lower outlier gate and upper outlier gate using the 25th Quarter and 75th Quarter. For this purpose, Equation 1, Equation 2 and Equation 3 are given below. In this method, where the Inter Quantile Range (IQR) value is calculated primarily, the IQR proximity rule limits are calculated by multiplying the IQR by 1,5. However, extreme values can be determined by multiplying IQR by 3 (Galli, 2020: 38). The values outside the lower bound and upper bound thus obtained represent outliers.

Upper outlier gate = 75th Quarter + (IQR * 1.5) (1)

Lower outlier gate = 25th Quarter + (IQR * 1.5) (2)

IQR = 75th Quarter – 25th Quarter (3)

3.1.2. Standard deviation method

Similar to the IQR procedure, the standard deviation method can detect outliers, depending on the limit that can be detected as 2 standard deviations or 3 standard deviations. Equation 4 and Equation 5 given below are generally used to determine the lower and upper limits for two standard deviation ranges. Thus, the values outside the lower limit and upper limit obtained as in the quarter span method represent outliers.

$$\text{Lower limit } = \bar{x} - 2 * \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}} \quad (4)$$

$$\text{Upper limit} = \bar{x} + 2 * \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}} \quad (5)$$

3.1.3. Modified Z Score method

The Z-score method, which is very similar to the standard deviation method, reveals how many standard deviations a value is from the mean. However, Z-scores are very sensitive to data values that are too large or smaller than the mean; therefore, a more robust way to detect outliers is to use a modified Z-score based on the median. According to Iglewicz and Hoaglin (1993), observations with modified Z-scores less than -3,5 or greater than 3,5 are identified as potential outliers. In the Z-score calculation modified for this purpose, the Equation 6 and Equation 7 are used.

$$Z_{\text{Modified}} = 0{,}6745 * \frac{x_i - \bar{x}}{MAD} \quad (6)$$

$$MAD = median(|x_i - median(x_i)|) \quad (7)$$

3.1.4. Isolation Forest method

Unlike the statistical outlier detections mentioned above, Isolation Forest is a different model-based method that explicitly isolates anomalies rather than profiling normal samples. While other methods focus only on the dependent variable, model-based methods such as Isolation Forest take both dependent and independent variables into account. Using two quantitative features for outlier detection, Isolation Forest searches for minorities with fewer samples and for samples with very different quality 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: 414). Unlike other anomaly detection algorithms, Isolation Forest focuses on detecting exception data and different characteristics, rather than distance or density, to detect anomalies. Thus, it can quickly separate outliers from normal data with low linear time complexity (Xu et al., 2017: 288).

3.2. Machine Learning Models Used in Real Estate Price Prediction

Regression analysis creates a mathematical model of the connection between dependent and independent variables with mutual cause-effect relationships under various assumptions and allows predictions to be made through this model (İlhan and Semih, 2020: 176). In this context, regression analysis methods, which express the collection of methods used to analyze the relationship between a dependent variable and one or more

independent variables (Zhang and O'Donnell: 2020: 123) and thus to estimate the dependent variable, are frequently used in real estate valuation. In this context, K-Nearest Neighbor, Lasso and Random Forest machine learning models, which are frequently used in regression tasks, were used in this study to test the success of outlier detection methods.

3.2.1. Random Forest

Random Forest, a supervised learning algorithm that uses the ensemble learning approach for classification and regression, is basically a decision tree machine learning method, consisting of combining many tree structures with ensemble learning and combining the prediction of these decision trees to create a more accurate final prediction (Truong et al., 2020: 436). In addition to prediction and classification tasks, the fact that they can be used with a small number of observations and a large number of variables, and also that they can provide information about the importance levels of the variables constitute the power behind the popularity of the Random Forest approach. Random Forest (Breiman, 2001) is a machine learning algorithm based on a method of bagging trees, which is mainly used for classification problems and can also be applied to regression tasks.

3.2.2. Lasso

Lasso (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996) is a linear regression analysis method with L1-norm arrangement, used for both variable selection and regularization. The accuracy of the Lasso algorithm largely depends on the selection of the regularization parameter (alpha), and small values selected lead to over-learning, while relatively large alpha values up to a certain level lead to more pure accuracy (Viktorovich et al., 2018: 3). By increasing the regularization parameter, solutions using fewer features are implemented by the model in order to improve the prediction accuracy and interpretability of the model (Gao et al., 2022: 15). In Lasso Regression, penalty and limitation are applied by considering the absolute values of the sum of the regression coefficients. This forces less important features to have zero weight values and enables implicit removal of unnecessary variables from the model (Bin et al., 2017: 210).

3.2.3. K-Nearest Neighbor

The K-nearest neighbor algorithm, one of the most widely used in the family of distance-based machine learning algorithms, is a non-parametric (Cover and Hart 1967: 22) machine learning algorithm used for classification and regression problems. The k-nearest neighbor algorithm is to calculate the distance to be predicted to all samples in the training dataset based on some distance function (such as Euclidean, Manhattan or cosine distance). After all these distance calculations, depending on the distance of each calculated observation, the *k* samples closest to the prediction subject in the training set are determined, and the averages of the known output values of these observations represent the output value of the observation to be predicted.

4. APPLICATION AND FINDINGS

All outlier detection methods mentioned in the method section are handled separately with different parameters, variable sets or transformations. During the first outlier detection process performed with the Quarter Range method, it is clearly noticed that the lower limit for the detection of outliers takes a negative value due to the right skew of the data set. As a result of this situation, very low-priced real estate records, which are very obvious to have an outlier in the current data set, cannot be detected as outliers. Therefore, both Standard Deviation and Quarter Range methods are used for both the normal dataset and the dataset with the dependent variable whose natural logarithm is taken. Thus, it is possible to detect a more appropriate outlier with the dependent variable *"Price"*, which exhibits a normal distribution. In addition, separate outliers were detected for the 2 sigma and 3 sigma intervals for the Standard Deviation method within the scope of both the normal data set and the logarithmic transform data set. Contrary to all these variables, which focus only on the dependent variable, Isolation Forest performs outlier detection by considering the values of both dependent and independent variables. In addition, Isolation Forest has a *"contamination"* parameter that takes a value between 0 and 0,5 to determine the outlier detection intensity. For a more successful outlier detection, outlier detection was performed with both 0,1 and 0,05 *"contamination"* values and all combinations of different variable sets separately. For this purpose, in addition to the *"Price"* dependent variable for the 1st Isolation Forest, all variables that take continuous values; 2. Variables of *"Price", "Number of Rooms", "Net Square Meter", "Number of Bathrooms", "Floor"* for Isolation Forest; 3. *"Price", "Number of Rooms", "Net Square Meters"* variables for Isolation Forest; 4. For Isolation Forest, outlier detection was performed using only the *"Price"* and *"Net Square Meter"* variables. In addition, automatic outlier detection methods such as Local Outlier Value Factor and One-Class SVM, apart from Isolation Forest, were also tested within the scope of the study, but they were not reported separately because they did not have any obvious superiority over other methods. In order to select the most

suitable dataset among the 15 new datasets obtained by all these outlier detection methods, as mentioned before, more than 10.000 models were carried out with parameter adjustments with 3 different machine learning approaches. These machine learning approaches used for this purpose were applied separately with each data set and the Mean Absolute Percentage Error (MAPE) was calculated. The MAPE values obtained by each model over each data set and the new observation numbers obtained after outlier detection are presented in Table 1. In addition, the grid search process was used to determine the most appropriate parameters during the development process of all these models. As part of the Grid Search approach, model performance is tested with all possible combinations of all parameter values. As a result, the model performances presented in Table 1 are obtained from the models that show the best performance as a result of this parameter optimization. In addition, within the scope of this whole process, the datasets were divided into training and test datasets at a rate of 80:20 percent, and the MAPE values presented in Table 1 were obtained from the test dataset.

Tablo 1. Comparison of outlier detection methods

Outlier Detection Methods	Number of Observations	K-Nearest Neighbor	Lasso	Random Forest	Mean
1. Isolation Forest (contamination: 0.1)	63.694	0,2245	0,2639	0,2292	0,2392
1. Isolation Forest (contamination: 0.05)	67.233	0,2333	0,2659	0,2415	0,2469
2. Isolation Forest (contamination: 0.1)	63.694	0,2243	0,2469	0,2338	0,235
2. Isolation Forest (contamination: 0.05)	67.232	0,2332	0,2646	0,2497	0,2492
3. Isolation Forest (contamination: 0.1)	63.700	0,2211	0,2134	0,2025	0,2123
3. Isolation Forest (contamination: 0.05)	67.232	0,2291	0,2332	0,2033	0,2219
4. Isolation Forest (contamination: 0.1)	63.698	0,2222	0,2157	0,2105	0,2161
4. Isolation Forest (contamination: 0.05)	67.236	0,2294	0.2303	0,2132	0,2213
Interquartile range method (Log)	70.164	0,2378	0,2575	0,1620	0,2191
Interquartile range method (Normal)	66.327	0,2264	0,2380	0,2596	0,2413
Standard deviation method (Log - 2 sigma)	67.834	0,2313	0,2523	0,2723	0,2519
Standard deviation method (Log - 3 sigma)	70.397	0,2392	0,2605	0,1688	0,2228
Standard deviation method (2 sigma)	**68.677**	**0,2334**	**0,2355**	**0,1597**	**0,2095**
Standard deviation method (3 sigma)	70.179	0,2367	0,2536	0,1676	0,2193
Modified Z Score method	69.456	0,2339	0,2303	0,1658	0,2133
Data Set for which Outlier Detection was not Performed	70.771	0,2468	0.2801	0,2037	0,2252

As can be seen in Table 1, each method used reveals different results despite each outlier detection method. Although the K-Nearest Neighbor and Lasso approach offered similar error rates to all outlier detection methods, the Random Forest method showed significantly better results with various outlier detection methods. For this purpose, it is not possible to select a data set by evaluating only the results of a machine learning method. Both for this purpose and because this study uses an ensemble learning approach as the general model architecture, the average MAPE value of the results obtained from each method was used to select the dataset. When the average MAPE values were examined, the outlier detection method with the lowest MAPE value was the Standard Deviation method with 2 sigma intervals without any logarithmic transformation. Thus, with a MAPE value of 20,95%, a 1% error rate improvement in overall performance was achieved compared to its closest competitor. In addition, the Standard Deviation model with a 2-sigma interval in terms of the mean MAPE value provided a 6,97% improvement in the error rate compared to the data set without any outlier detection. Considering the situation in terms of the Random Forest approach, which exhibits the best estimation result among all estimation models, the Standard Deviation outlier detection approach with 2 sigma intervals performs much better than the average values. In addition, all outlier approaches, except for only a few, showed better prediction performance in all prediction models compared to the data set in which no outliers were detected. The new dataset created after the removal of the outliers detected by the Standard Deviation outlier detection approach with a 2-sigma range has a total of 68.677 residential records.

5. CONCLUSION

Increasing perdiction performance, which is of critical importance in the field of real estate price prediction or real estate valuation, directly depends on data quality when data-driven approaches such as machine learning are used. The applicability of machine learning applications, which offer an alternative approach for smart system design in the real estate valuation process, can only be guaranteed when a suitable large pool of transaction data is available to work with and this data set is prepared under appropriate conditions. Although dataset size is an important factor, completeness and representativeness are even more important. At this point, the effects of outlier detection in residential real estate valuation are investigated with a large data set obtained in this study. For this purpose, a heterogeneous data set with 70.771 real estate data and 283 variables, 4 different outlier detection methods were tested with 3 different machine learning approaches. The empirical findings reveal that the use of different outlier detection approaches increases the prediction performance in different ranges. With a 6,97% increase in average model performance, this performance increase for Random Forest was a high of 21,6%. In this context, the results obtained in this study show that with an appropriate outlier detection approach and process, the data quality can be increased and therefore the prediction performance will also increase. In addition, this study also shows that, contrary to the literature, using more than one method rather than a single outlier detection method may yield better results. As the empirical findings show, the prediction model and data set to be used can also change the outlier detection method that should be used. In other words, the performance of outlier detection methods may vary according to the data set and prediction model used. For this reason, the necessity of increasing the processes to be allocated to data pre-processing processes and the use of alternative models at this stage, as in the same predicting models, is a necessity for this area where the limits of predicting performance are pushed.

REFERENCES

Abhyankar, A. A., & Singla, H. K. (2021). Comparing predictive performance of general regression neural network (GRNN) and hedonic regression model for factors affecting housing prices in "Pune-India". *International Journal of Housing Markets and Analysis*.

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.

Alfaro-Navarro, J. L., Cano, E. L., Alfaro-Cortés, E., García, N., Gámez, M. and Larraz, B. (2020). A fully automated adjustment of ensemble methods in machine learning for modeling complex real estate systems. *Complexity*, *2020*.

Alkan, T., Dokuz, Y., Ecemiş, A., Bozdağ, A., & Durduran, S. S. (2022). Using Machine Learning algorithms for predicting real estate values in tourism centers.

Almond, N., Lewis, O., Jenkins, D., Gronow, S., & Ware, A.

(1997, September). Intelligent systems for the valuation of residential property. In *RICS Cutting Edge, Conference, Dublin* (pp. 1-19).

Aydemir, E., Aktürk, C., & Yalçınkaya, M. A. (2020). Yapay zekâ ile konut fiyatlarının tahmin edilmesi. *Turkish Studies*, *15*(2), 183-194.

Aydemir, E., Aktürk, C., & Yalçınkaya, M. A. (2020). Yapay zekâ ile konut fiyatlarının tahmin edilmesi. *Turkish Studies*, *15*(2), 183-194.

Barnett, V., & Lewis, T. (1984). Outliers in statistical data. *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*.

Bergadano, F., Bertilone, R., Paolotti, D., & Ruffo, G. (2021). Developing real estate automated valuation models by learning from heterogeneous data sources. *International Journal of Real Estate Studies*, *15*(1), 72-85.

Bilgilioğlu, S. S., & Yılmaz, H. M. (2021). Comparison of different machine learning models for mass appraisal of real estate. *Survey Review*, 1-12.

Bin, J., Tang, S., Liu, Y., Wang, G., Gardiner, B., Liu, Z., & Li, E. (2017, September). 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)* (pp. 209-213). IEEE.

Bin, O. (2004). A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics*, *13*(1), 68-84.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Büyük, G., & Ünel, F. B. (2021). Comparison of modern methods using the python programming language in mass housing valuation. *Advanced Land Management*, *1*(1), 21-26.

Chou, S. M., Lee, T. S., Shao, Y. E., & Chen, I. F. (2004). Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. *Expert systems with applications*, *27*(1), 133-142.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

Daşkıran, F. (2015). Denizli kentinde konut talebine etki eden faktörlerin hedonik fiyatlandırma modeli ile tahmin edilmesi. *Journal Of International Social Research*, *8*(37).

Fu, T. (2018, June). Forecasting second-hand housing price using artificial intelligence and machine learning techniques. In *2018 8th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2018)* (pp. 269-273). Atlantis Press.

Galli, S. (2020). *Python feature engineering cookbook: over 70 recipes for creating, engineering, and transforming features to build machine learning models*. Packt Publishing Ltd, 42-25.

Gao, G., Bao, Z., Cao, J., Qin, A. K., & Sellis, T. (2022). Location-centered house price prediction: A multi-task learning approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *13*(2), 1-25.

García-Magariño, I., Medrano, C., & Delgado, J. (2020). Estimation of missing prices in real-estate market agent-based simulations with machine learning and dimensionality reduction methods. *Neural Computing and Applications*, *32*(7), 2665-2682.

Gilbertson, B., & Preston, D. (2005). A vision for valuation. *Journal of Property Investment and Finance*, *23*(2), 123-140.

Gupta, R., Marfatia, H. A., Pierdzioch, C., & Salisu, A. A. (2021). Machine Learning predictions of housing market synchronization across us states: the role of uncertainty. *The Journal of Real Estate Finance and Economics*, 1-23.

Hårsman, B., & Quigley, J. M. (Eds.). (1991). *Housing markets and housing institutions: an international comparison*. Massachusetts: Kluwer Academic Publishers, 2-3.

Ho, W. K., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48-70.

Hodge, V., & Austin, J. (2004). A survey of outlier detection

methodologies. *Artificial intelligence review*, *22*(2), 85-126.

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers* (Vol. 16). Asq Press.

Imran, I., Zaman, U., Waqar, M., & Zaman, A. (2021). Using machine learning algorithms for housing price prediction: the case of Islamabad housing data. *Soft Computing and Machine Intelligence*, *1*(1), 11-23.

İlhan, A. T., & Semih, Ö. Z. (2020). Yapay sinir ağlarının gayrimenkullerin toplu değerlemesinde uygulanabilirliği: Gölbaşı ilçesi örneği. *Hacettepe Üniversitesi Sosyal Bilimler Dergisi*, *2*(2), 160-188.

Jha, S. B., Babiceanu, R. F., Pandey, V., & Jha, R. K. (2020). Housing market prediction problem using different machine learning algorithms: A case study. arXiv preprint arXiv:2006.10092.

Jui, J. J., Molla, M. I., Bari, B. S., Rashid, M., & Hasan, M. J. (2020). flat price prediction using linear and random forest regression based on machine learning techniques. In *Embracing Industry 4.0* (pp. 205-217). Springer, Singapore.

Kalliola, J., Kapočiūtė-Dzikienė, J., & Damaševičius, R. (2021). Neural network hyperparameter optimization for prediction of real estate prices in Helsinki. *PeerJ Computer Science*, *7*, e444.

Kim, J., Won, J., Kim, H., & Heo, J. (2021). Machine-Learning-Based prediction of land prices in Seoul, South Korea. *Sustainability*, *13*(23), 13088.

Kouwenberg, R., & Zwinkels, R. (2014). Forecasting the US housing market. *International Journal of Forecasting*, *30*(3), 415-425.

Küçükkaplan, İ,, & Aldı, F. A. (2017). Denizli ilinde konut fiyatlarına etki eden faktörlerin panel verilerle analizi. *Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *20*(37), 219-236.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413-422). IEEE.

Manasa, J., Gupta, R., & Narahari, N. S. (2020, March). Machine learning based predicting house prices using regression techniques. In *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)* (pp. 624-630). IEEE.

Mankad, M. D. (2021). Comparing OLS based hedonic model and ANN in house price estimation using relative location. *Spatial Information Research*, 1-10.

Manrique, M. A. C., Otero Gomez, D., Sierra, O. B., Laniado, H., Mateus C, R., & Millan, D. A. R. (2020). Housing-Price Prediction in Colombia using Machine Learning. *OSF Preprints*, (w85z2).

McGreal, S., Adair, A., McBurney, D., & Patterson, D. (1998). Neural networks: the prediction of residential values. *Journal of Property Valuation and Investment*, *16*(1), 57-70.

Mrsic , L., Jerkovic, H., & Balkovic, M. (2020). Real estate market price prediction framework based on public data sources with case study from croatia. In: Sitek, P., Pietranik, M., Krótkiewicz, M., Srinilta, C. (eds*) Intelligent Information and Database Systems*. ACIIDS 2020. Communications in Computer and Information Science, vol 1178. Springer, Singapore. https://doi.org/10.1007/978-981-15-3380-8_2.

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.

Pai, P. F., & Wang, W. C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences*, *10*(17), 5832.

Pérez-Rave, J. I., González-Echavarría, F.,, & Correa-Morales, J. C. (2020). Modeling of apartment prices in a Colombian context from a machine learning approach with stable-important attributes. *Dyna*, *87*(212), 63-72.

Peter, N. J., Okagbue, H. I., Obasi, E. C., & Akinola, A. O. (2020). Review on the application of artificial neural networks in real estate valuation. *International Journal*, *9*(3), 2918-2925.

Poursaeed, O., Matera, T., & Belongie, S. (2018). Vision-based real estate price estimation. *Machine Vision and Applications*, *29*(4), 667-676.

Rahman, S. K., Sathik, M. M., & Kannan, K. S. (2012). Multiple linear regression models in outlier detection. International Journal of Research in Computer Science, 2(2), 23-28.

Rampini, L., & Cecconi, F. R. (2021). Artificial intelligence algorithms to predict Italian real estate market prices. *Journal of Property Investment & Finance*.

Sa'at, N. F., Maimun, N. H. A., & Idris, N. H. (2021). Enhancing the accuracy of malaysian house price forecasting: a comparative analysis on the forecasting performance between the hedonic price model and artificial neural network model. *Planning Malaysia*, *19, 249- 259*.

Sandbhor, S., & Chaphalkar, N. B. (2019). Impact of outlier detection on neural networks based property value prediction. In *Information systems design and intelligent applications* (pp. 481-495). Springer, Singapore.

Sangha, A. (2021). Property valuation by machine learning for the Norwegian real estate market. *ScienceOpen Preprints.* DOI: 10.14293/S2199-1006.1.SOR-.PP0TP9I.v1

Sawant, R., Jangid, Y., Tiwari, T., Jain, S., & Gupta, A. (2018, August). Comprehensive analysis of housing price prediction in pune using multi-featured random forest approach. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-5). IEEE.

Sevgen, S. C., ve Aliefendioğlu, Y. (2020). Mass apprasial with a machine learning algorithm: random forest regression. *Bilişim Teknolojileri Dergisi*, *13*(3), 301-311.

Seya, H., & 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.

Shapiro, E., Mackmin, D., & Sams, G. (2019). *Modern methods of valuation*. Estates Gazette

Shi, D., Guan, J., Zurada, J., and Levitan, A. S. (2022). Predicting home sale prices: A review of existing methods and illustration of data stream methods for improved performance. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(2), e1435.

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.

Sisman, S., & Aydinoglu, A. C. (2022). Improving performance of mass real estate valuation through application of the dataset optimization and Spatially Constrained Multivariate Clustering Analysis. *Land Use Policy*, *119*, 106167.

Steurer, M., Hill, R. J., & Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research*, *38*(2), 99-129.

Štubňová, M., Urbaníková, M., Hudáková, J., & Papcunová, V. (2020). Estimation of residential property market price: comparison of artificial neural networks and hedonic pricing model. *Emerging Science Journal*, *4*(6), 530-538.

Tabar, M. E., Başara, A. C. ve Şişman, Y. (2021). Çoklu Regresyon ve Yapay Sinir Ağları ile Tokat ilinde konut değerleme çalışması. *Türkiye Arazi Yönetimi Dergisi*, *3*(1), 1-7.

Tchuente, D., & Nyawa, S. (2021). Real estate price estimation in French cities using geocoding and machine learning. *Annals of Operations Research*, 571-608.

Terregrossa, S. J., & Ibadi, M. H. (2021). Combining housing price forecasts generated separately by hedonic and artificial neural network models. *Asian Journal of Economics, Business and Accounting*, *1*, 130-148.

Tibshirani. R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

Torres-Pruñonosa, J., García-Estévez, P., & Prado-Román, C. (2021). Artificial neural network, quantile and semi-log regression modelling of mass appraisal in housing. *Mathematics*, *9*(7), 783.

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

Viktorovich, P. A., Aleksandrovich, P. V., Leopoldovich, K. I., & Vasilevna, P. I. (2018, August). Predicting sales prices of the houses using regression methods of machine learning. In *2018 3rd Russian-Pacific Conference on Computer Technology and*

Applications (RPC) (pp. 1-5). IEEE.

Walthert, L., & Sigrist, F. (2019). Deep learning for real estate price prediction. *Available at SSRN 3393434*.

Xu, D., Wang, Y., Meng, Y., & Zhang, Z. (2017, December). An improved data anomaly detection method based on isolation forest. In *2017 10th international symposium on computational intelligence and design (ISCID)* (Vol. 2, pp. 287-291). IEEE.

Yacim, J. A., & Boshoff, D. G. B. (2020). Neural networks support vector machine for mass appraisal of properties. *Property Management*, *38*(2), 241-272.

Yazdani, M. (2021). Machine Learning, Deep Learning, and Hedonic Methods for real estate price prediction. *arXiv preprint arXiv:2110.07151*.

Yıldırım, H. (2019). Property value assessment using artificial neural networks, hedonic regression and nearest neighbors regression methods. *Selçuk Üniversitesi Mühendislik, Bilim ve Teknoloji Dergisi*, *7*(2), 387-404.

Yilmazer, S., & Kocaman, S. (2020). A mass appraisal assessment study using machine learning based on multiple regression and random forest. *Land use policy*, *99*, 104889.

Zhang, F., & O'Donnell, L. J. (2020). Support vector regression. In *Machine Learning* (pp. 123-140). Academic Press.

Zhao, Y., Chetty, G., & Tran, D. (2019, December). Deep learning with XGBoost for real estate appraisal. In *2019 IEEE symposium series on computational intelligence (SSCI)* (pp. 1396-1401). IEEE.

Zurada, J., Levitan, A., & Guan, J. (2011). A comparison of regression and artificial intelligence methods in a mass appraisal context. *Journal of real estate research*, *33*(3), 349-388.