ISSN: 1304-4796
Manisa Celal Bayar Üniversitesi Sosyal Bilimler Dergisi
Manisa Celal Bayar University Journal of Social Sciences
2020; 18 (4); 251-260

Regresyon Analizleri mi Karar Ağaçları mı?[a]

Burcu Kocarık Gacar [b, c], İpek Deveci Kocakoç [d]

Özet

Karar ağaçları algoritması, veri madenciliği teknikleri içinde önemli bir sınıflandırma yöntemidir. Karar ağacı, kök düğümü, dalları ve yaprak düğümleri olan ağaç yapısında sınıflandırma ve regresyon modelleri oluşturur. Bağımlı değişken iki kategorili olduğunda regresyon analizine alternatif bir yöntem olarak tercih edilen lojistik regresyon analizi, sınıflandırma amacıyla kullanılan bir diğer tekniktir. Bu araştırma kapsamında aynı veri seti üzerinde lojistik regresyon, doğrusal regresyon, sınıflandırma ağacı ve regresyon ağacı yöntemleri uygulanmıştır. Bu dört yöntem kullanılarak konut fiyatını belirleyen en önemli değişkenler belirlenmiştir. Modellerin performansları ve tahmin güçleri karşılaştırılmış; en iyi sınıflandırma yapan model belirlenmeye çalışılmıştır. Bu karşılaştırma, 5 bağımsız değişken ve bağımlı değişken ev fiyatı olmak üzere, 414 gayrimenkul verisi kullanılarak yapılmıştır. Analiz sonucunda elde edilen bulgular, gayrimenkul değerleme verisi için sınıflandırma ağacı modelinin standart yaklaşımlardan daha iyi performans sergilediğini göstermiştir.

Anahtar Kelimeler

Regresyon
Lojistik Regresyon
Sınıflandırma ve Regresyon Ağaçları
Karar Ağaçları

Makale Hakkında

Geliş Tarihi: 16.09.2020
Kabul Tarihi: 25.12.2020
Doi: 10.18026/cbayarsos.796172

Regression Analyses or Decision Trees?

Abstract

Decision tree algorithm is an important classification method in data mining techniques. A decision tree creates classification and regression models like a tree that has a root node, branches, and leaf nodes. Logistic regression which is an alternative method to regression analysis when the dependent variable is a dichotomy, is another technique used for classification purposes. Within the scope of this research, logistic regression, linear regression, classification tree, and regression tree were applied on the same data set. This study explores the most important variables determining the house price by using these four methods. Models' performances and predictive powers were compared and the best model is determined. This comparison was performed using 414 real estate data on 5 independent variables and the dependent variable is house price. The findings showed that the classification tree model for real estate valuation data performs better than standard approaches.

Keywords

Regression
Logistic Regression
Classification and Regression Trees
Decision Trees

About Article

Received: 16.09.2020
Accepted: 25.12.2020
Doi: 10.18026/cbayarsos. 796172

[a] Bu çalışma, 5th International Researchers, Statisticians and Young Statisticians Congress (IRSYSC2019)'de özet bildiri olarak sunulmuştur.

[b] İletişim Yazarı: burcu.kocarikgacar@deu.edu.tr.
[c] Arş. Gör., Dokuz Eylül Üniversitesi, İİBF, Ekonometri Bölümü, Dokuz Çeşmeler Kampüsü, İzmir, Türkiye. ORCID: 0000-0001-5944-4456
[d] Prof. Dr., Dokuz Eylül Üniversitesi, İİBF, Ekonometri Bölümü, Dokuz Çeşmeler Kampüsü, İzmir, Türkiye. ORCID: 0000-0001-9155-8269

Introduction

In statistical applications, it is desirable to determine and analyze the relationship structure between independent variables and dependent variables. At this point, the classification and regression methods included in machine learning methods are important data analysis techniques used to define the relationship between dependent and independent variables. In data analysis, numerous statistical techniques are used, including classification and regression. Of these, multiple regression and logistic regression analysis are traditional/classic methods.

In recent years, there has been an increase in the use of non-linear, hierarchical, rule-based methods since the relationships are not linear in many places and many of the variables examined are not normally distributed. One of the most preferred of these methods is the classification and regression tree technique. Classification and regression tree analysis is a type of decision tree methodology. The purpose of classification and decision trees involved in data mining is to develop a model based on the data and to estimate the result values of the datasets (Güner, 2014). Classification and regression trees are easy to explain and provide an easy to interpret visualization of model outcomes. In statistical analysis methods, it is very important to determine the most appropriate method to be applied according to the structure of the data.

In the literature, a comparison of the classification and regression trees with logistic regression analysis has been made in different studies. Firstly, Long, Griffth, Selker and Agostino (1993) compared the performance of logistic regression to decision-tree in classifying patients as having acute cardiac ischemia. As a result, the logistic regression performed better than the decision tree. In the study of Lemon, Roy, Clark, Friedmann, and Rakowski (2003), it was concluded that classification and regression trees can better identify populations at risk in public health research than logistic regression. In another paper, Irimia-Dieguez, Blanco-Oliver and Vazquez-Cueto (2015) compared the predictive performance of classification and regression trees against logistic regression by employing the smallest enterprises, which included financial, non-financial, and macro-economic factors. Findings show that classification and regression trees outperform logistic regression. The study of Rudd and Priestley's (2017) aim was to predict the worst non-financial payment status among businesses and evaluate decision tree model performance against the logistic regression model. Their findings show that both methods performed the same.

So it can be said that logistic regression and decision tree classification are two of the basic classification algorithms being used today. Some studies show that classification and regression trees outperform the standard approach in the literature, others show that decision trees have been found better than logistic regression. None of them is better than the other and the performance of each generally changes depending on the nature of the data.

This study compares the performance of a non-parametric methodology (namely classification and regression trees), against traditional methods (like linear regression and logistic regression). Classification techniques predict categorical responses unlike regression techniques predict continuous responses. An important aspect of the study was the choice of the scale type of the dependent variable. In the scope of this paper, logistic regression, linear regression, classification and regression trees were compared on the same data set, in addition to the literatüre. Our aim in this study is to reveal the performance differences between the four models which predicted by these four methods.

The main purpose of real estate valuation is to assist in the decisions to be taken regarding real estate because real estate is preferred both as a necessity and as an investment tool. Real estate has been an important portion of wealth for thousands of years (Garay, 2016; Krulický & Horák, 2019). The present text focuses on whether the price of the house is correctly classified based on the characteristics of the house.

Methods and Materials

Methodology

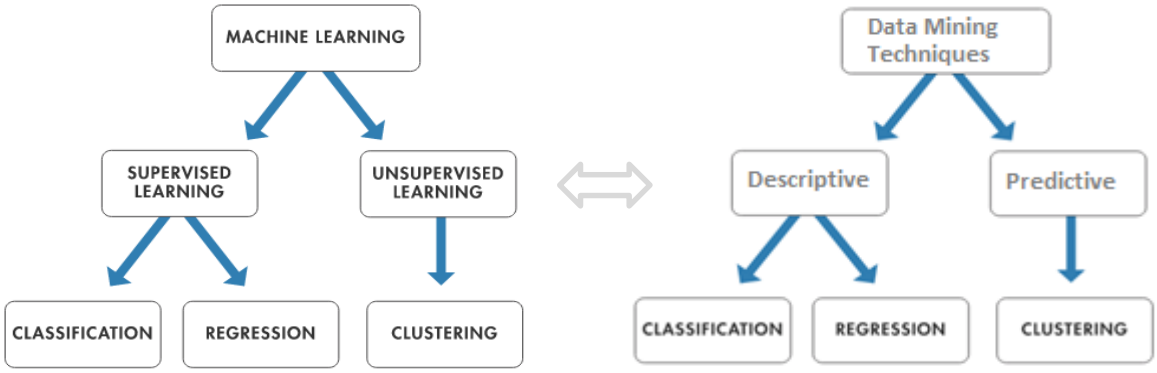

Figure 1. Classification of machine learning and data mining techniques.
(*What is Machine Learning?, Matlab, Mathworks)*

Both data mining and machine learning are drawing from the same foundation but in different ways. Machine learning embodies the principles of data mining, but can also learn from them to apply to new algorithms. Machine learning uses two types of learning methods. Supervised learning trains a model on known input and output data so that it can predict future outputs. Unsupervised learning finds hidden patterns structures in input data. Machine learning has changes the working principles in many areas (Aery & Ram, 2017). Decision tree algorithm is an important classification method in data mining. A decision tree is a flow-chart like a tree structure, where each internal node denotes a decision on a variable, each branch depicts an outcome of the decision, and leaf nodes represent classes (Ru-ping, 2010).

Regression and Logistic Regression Analysis

Regression is an analysis method used to model the linear relationship between two or more variables. When the dependent variable is continuous, a linear regression model is often used. Linear models have an important assumption that the error terms have a normal distribution.

In the case that the dependent variable is categorical, the use of the logistic regression model is the alternative in cases where the linear model cannot be applied due to the distortion of the normality assumption (Hosmer & Lemeshow, 1989). Least square estimation method is used for regression estimation.

Logistic regression assumes a logit relationship between dependent and independent variables. Logistic Regression assumes that the data is (curvy) linearly separable in space. Therefore, logistic regression can produce nonlinear (exponential or polynomial) models. In

linear regression, the dependent variable has an estimated value, while in logistic regression the possibility of realization of one of the values that the dependent variable can take is estimated. The great advantage of this method is that it produces a probabilistic formula of classification (Tatlıdil, 2002; Khemphila & Boonjing, 2010). It is illustrated by the logistic regression model (1).

$$L = \ln(p_i/(1-p_i)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (1)$$

Where b_0 is the fixed value and b_i (i = 1,2,… k) indicates the regression coefficients of each independent variable. $\ln(p_i/(1-p_i))$ represents the logistic transformation, and p_i denotes the probability of the desired situation being realized. Thus, $p_i/(1-p_i)$ is called the odds ratio and can be defined as the ratio of preference to non-preference in any event. The probability value p_i is indicated by (2).

$$p_i = \exp(\textstyle\sum_{k=0}^{n} \beta_k x_{ik}) \,/(1 + exp(\sum_{k=0}^{n} \beta_k x_{ik})) \quad (2)$$

The dependent variable can be either categorical or continuous. If the dependent variable is continuous, analysis is called regression, and if the dependent variable is categorical, analysis is called logistic regression (Hosmer & Lemeshow, 1989; Alpar, 2017). If there are two groups, binary logistic regression is used.

Classification and Regression Trees Analysis

Decision trees form classification and regression models like a tree structure by asking questions and creating decision rules according to the structure of the datasets that constitute a problem. For this process, questions are started to be asked at the root node, which is the basic element of the tree structure, and the tree grows by branching until the leaves are reached, which is the last element of the tree structure (Kurt, Türe, & Kurum, 2008; Deveci Kocakoç & Keser, 2019). The tree structure indicates that the root node is the most important independent variable, while the sub-branches show that it is the other independent variables. So the leaves show the values of the dependent variable.

Classification and regression trees within the decision trees are frequently preferred because of their advantages such as being easy to understand, not requiring assumptions and visual presentation of the results in the form of diagrams (Lewis, 2000). It is a powerful alternative to parametric techniques because it requires no assumptions on the data set examined. The classification and regression tree, which is constructed by splitting subsets of the data set using all independent variables, is a recursive partition method to be used both for regression and classification. The best predictor is chosen using measures, such as gini, least-squared deviation, entropy, twoing, ordered twoing, information gain and information gain ratio, chi-square probability value (Khemphila & Boonjing, 2010). The aim here is to produce the most homogeneous data subgroups possible for variables. Classification and regression trees are able to model the relationship between dependent variables and independent variables for homogeneous subclasses, taking into account the heterogeneity in the data set and visualizing this relationship in the form of a tree structure.

With classification and regression trees, both categorical and continuous variables can be modeled. The dependent variable can be either categorical or continuous. If the dependent

variable is categorical, analysis is called a classification tree and if the dependent variable is continuous, analysis is called the regression tree (Deconinck, Hancock, Coomans, & Massart, 2005). There are no classes in the regression trees and the data is continuous. For this reason, classification, separation rules, gini index or entropy measurement can not be applied in the regression tree technique. The branching process in the regression tree is performed according to the algorithm of reducing the squares of residues, which means that the total variance estimated for the node resulting in two results must be minimized (Lewis, 2000).

Data Set

The real estate valuation dataset provided by a UCI Machine Learning Repository was used in this study. The dataset consists of 414 real estate valuation data were collected from New Taipei City in Taiwan between January 2012 and December 2013 (Yeh & Hsu, 2018).

This study used the following 5 variables as independent (explanatory) variables; the house age (unit: year), the distance to the nearest MRT (Mass Rapid Transit) station (unit: meter), the number of convenience stores in the living circle on foot (integer), the geographic coordinate - latitude (unit: degree), the geographic coordinate - longitude (unit: degree). The dependent variable included house price of unit area (10000 Taiwan Dollar/Ping).

R (programming language) and IBM SPSS 25.0 were used for analysis. For each trial, approximately 30% of the cases are randomly selected for validation. The rest is selected as the training set which was 290 houses.

Analyses Results

Regression Analysis Results

According to the analysis of variance, the regression model is significant ($p\_value < 0.05$). As shown in Table 1, The coefficients of all variables are also significant ($p\_value < 0.05$). Slopes and signs of variables follow expectations.

Table 1. Coefficients of the Regression Model Fitted to the Dataset

Model	Unstandardized Coefficients		t	Sig.
	Coefficients	Std. Error		
constant	-34343,895	5253,746	-6,161	0,000
houseage	-0,249	0,051	-4,875	0,000
stores	1,463	0,184	6,334	0,000
latitude	357,628	52,098	6,897	0,000
longitude	209,406	44,305	4,745	0,000

dependent variable is house price

The distance to the nearest MRT station argument is highly correlated with the other independent variables and was not included in the analysis because of the occurrence of multicollinearity. Therefore, four of the independent variables were used in the analysis.

Logistic Regression Analysis Results

The dependent variable was converted to a binary category according to its arithmetic mean to implement logistic regression. For all cases, the dependent variable is taken the value 1 when the house price is higher than the arithmetic mean (count = 214) and 0 otherwise (count = 200).

Table 2. Variables in the Logistic Regression Model Fitted to the Dataset

	Coefficients	Std. Error	Wald	df	Sig.
houseage	-0,056	0,015	14,296	1	0,00
stores	0,374	0,069	29,761	1	0,00
latitude	90,039	16,963	28,155	1	0,00
longitude	53,920	15,886	11,512	1	0,00
constant	-8802,002	2066,109	18,209	1	0,00

dependent variable is house price

The coefficients of all variables are also significant ($p\_value < 0.05$). Slopes and signs of variables follow expectations (Table 2.). According to the analysis, logistic model is significant ($p\_value < 0.05$). When the Hosmer-Lemeshow test is examined to evaluate the model's goodness of fit, it is concluded that there is a model fit at the 5% significance level ($p\_value = 0.132$). The performance of logistic regression is achieving about 82.3% correct classification (Table 3.).

Table 3. Logistic Regression Classification Table

			Predicted		
			House Price		Percentage
	Observed		Low	High	Correct
Train set	House Price	Low	115	26	81,6
		High	25	118	82,5
	Overall Percentage				**82,3**
Validation set	House Price	Low	45	14	76,3
		High	9	62	87,3
	Overall Percentage				**81**

Table 4. Model Summaries

	Regression		Logistic Regression	
	R Square	Adjusted R Square	-2 Log likelihood	Nagelkerke R Square
Train set	0,51	0,498	237,379	0,56
Validation set	0,60	0,58	108,448	0,57

As it is seen that in Table 4., The adjusted R Square for regression is about 49,8%, almost 50% for the training set and about 58% for the validation set. Nagelkerke R Square for logistic regression is 56% for the training set and almost equal to it for the validation set. For the real estate valuation date, this is the expected situation.

Regression Tree Analysis Results

As reading the regression tree in Figure 2, it can be seen that the houses with a distance < 249 and houseage < 11.7 have the highest price (70.65). Houses with distance > 761 and located under 121.5 longitudes have the lowest price (16.57). Continuing to the left of the root node of the tree, the decision rule is distance > 761, otherwise, the decision rule is distance < 761.

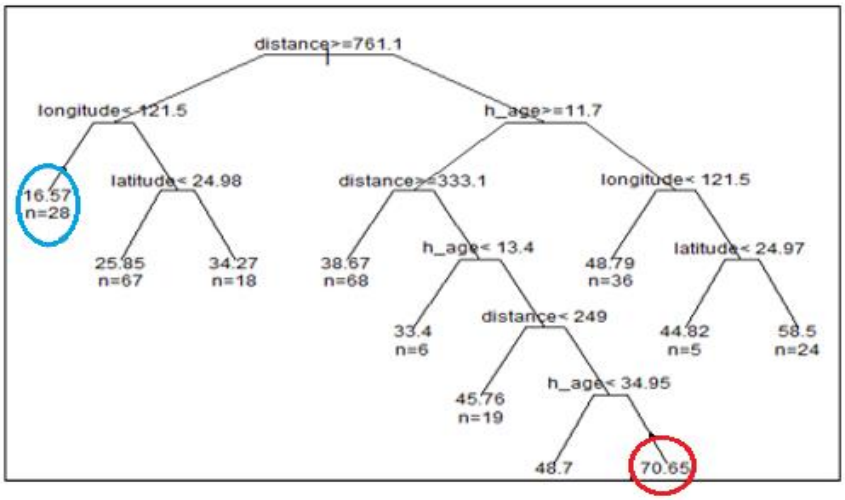

Figure 2. Regression tree results

Classification Tree Analysis Results

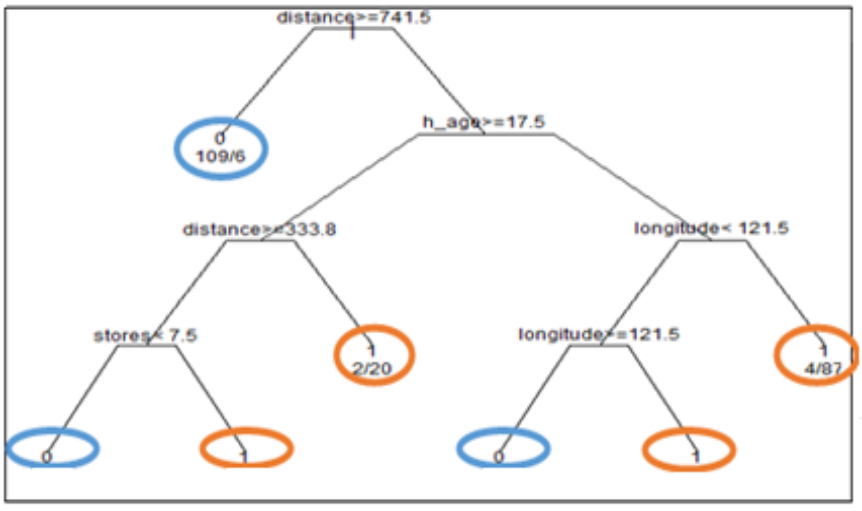

Figure 3. Classification tree results

As reading the classification tree in Figure 3, it can be seen that the houses with houseage < 17.5 or distance < 333.8 or longitude < 121.5 have high prices (value 1). Houses with distance >

741 or located far from stores have low prices (value 0). Continuing to the left of the root node of the tree, distance > 741, otherwise distance < 741.

Findings

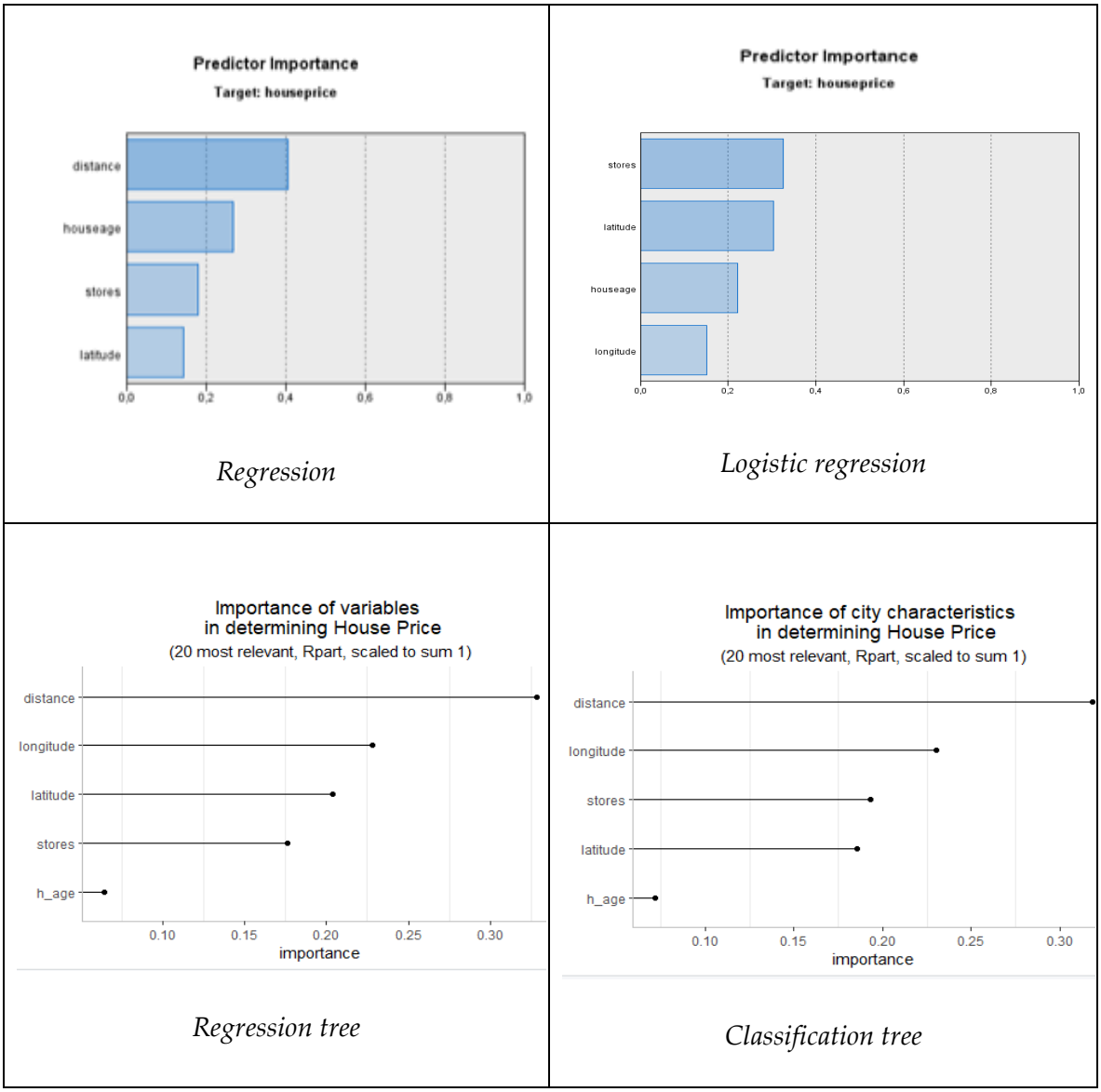

Figure 4. Importance of predictor on house price

According to Figure 4 on the top of the left, the most important variable in determining the house price is “Distance”. Houseage and stores follow it in regression. According to the figure on the top of the right, important variables are stores and latitude of the house in logistic regression. According to the figure at the bottom of the left, the most important variable in determining the house price is also "Distance".

The longitude and latitude of the house follow it in the regression tree. According to the figure on the bottom of the right, the most important variable in determining the house price is again "Distance". Longitude and stores around the house follow it in the classification tree.

Table 5. Models' Performances

	Regression	Logistic Regression	Regression Tree	Classification Tree
Train Set	71%	82%	82%	91%
Validation Set	77%	81%	50%	80%

The cross-tabulation in Table 5. can be examined to evaluate the models' performances. In the analyzes, 70% of all data was used as training data for the purpose of creating a model, and the remaining 30% was used as validation data to test the accuracy of classification rules. The validation sets yielded results close to the training sets for regression and logistic regression, on the other hand, the validation sets have a lower classification rate than training sets for regression tree and classification tree formed with the real estate data.

The regression tree formed with the same data has an accuracy of 82% for the train set but 50% for the validation set. The classification tree model has an accuracy of 91% which is the highest classification score for the train sets. The fact that the classification accuracy is lower than other methods in the regression and regression tree may be due to the dependent variable scale type is continuous.

For this data, results show that best performances with logistic regression model which has 81% classification accuracy and as same as the classification tree model which has the 80% classification accuracy. It has been seen that the models created do not differ in terms of classification success and give very close results. It can be noted that both models have a classification success of over 80%, while the classification tree analysis is found to have a higher classification accuracy. From this point on, it has been found more appropriate to use the classification tree analysis technique in order to minimize the risk of error in the studies conducted with the classification tree and logistic regression analysis.

Conclusions and Discussion

According to the findings, it can be said that confirm the distinction of the non-parametric techniques on the classic regression analyses. It will not be wrong to say that the classification tree model is more reliable because of the high classification success on the training data repeats on the validation data. Also, classification tree models have the advantage of their visualization like a flowchart diagram and there are no assumptions about distribution because of the nonparametric nature of the algorithm.

In New Taipei City-Taiwan between January 2012 and December 2013, it was observed that the ''Distance'' factor had a significant effect on the selection of house. Since complete data for 2014 and beyond real estate were not published, this study analyzes 2012-2013 data. Having a single data source related to this study constitutes the limitation of the study.

Analysis results can be compared for real estate and other sectors with more different and up-to-date datasets. Artificial neural networks and support vector machine methods, which are other classification algorithms in machine learning, can also be included in the analysis and compared.

References

Aery, M., & Ram, C. (2017). A Review on Machine Learning: Trends and Future Prospects. https://www.researchgate.net/publication/323377718

Alpar, R. (2017). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. Detay Yayıncılık. Dördüncü Baskı, Ankara.

Deconinck, E., Hancock, T., Coomans, D., & Massart. (2005). "Classification of Drugs in Absorption Classes Using the Classification and Regression Trees (CART) Methodology", Journal of Pharmaceutical and Biomedical Analysis, 39: 91–103.

Deveci Kocakoç, İ., & Keser, İ. (2019). Exploring Decision Rules for Election Results by Classification Trees. In Economies of the Balkan and Eastern European Countries, Kne Social Sciences, Pages 107--115. DOI 10.18502/Kss.V4i1.5982economies Of The Balkan and Eastern European Countries (EBEEC 2019), Conference Paper.

Garay, U. (2016). Real Estate as an Investment (Chapter 14). N Book: Alternative Investments: Caıa Level Iı (Pp.343-358.)Edition: 3rd Chapter: Real Estate as an Investment publisher: Wıley https://www.researchgate.net/publication/309415671

Güner, Z. B. (2014). ''CART and Logistic Regression Analysis in Data Mining: An Application on Pharmacy Provision System Data'' . Sosyal Güvenlik Uzmanları Derneği, Sosyal Güvence Dergisi, Sayı 6.

Hosmer, D. W., & Lemeshow, S. (1989). "Applied Logistic Regression", John Wiley & Sons, New York, 5-50.

Irimia-Dieguez, A., Blanco-Oliver, A., & Vazquez-Cueto, M. (2015). ''A Comparison of Classification/Regression Trees and Logistic Regression in Failure Models''. Procedia Economics and Finance 26, 23 – 28.

Khemphila, A., & Boonjing, V. (2010). ''Comparing Performances of Logistic Regression, Decision Trees, and Neural Networks for Classifying Heart Disease Patients'' 978-1-4244-7818-7/10/$26.00_C 2010 Ieee.

Krulický, T., & Horák, J. (2019). Real Estate as an Investment Asset. Web of Conferences 61, 01011 https://Doi.Org/10.1051/Shsconf/20196101011.

Kurt, İ., Türe, M., & Kurum, A. T. (2008). Comparing Performances of Logistic Regression, Classification and Regression Tree, and Neural Networks for Predicting Coronary Artery Disease. . Expert Systems Xith Applications, 34, 366-374.

Lemon, S., Roy, J., Clark , M., Friedmann, P., & Rakowski , W. (2003). Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison with Logistic Regression. Ann Behav Med. 2003; 26 (3) : 172 - 181. Doi:10.1207/S15324796abm2603_02. © 2003 By The Society of Behavioral Medicine, 26.

Lewis, R. J. (2000). ''An Introduction to Classification and Regression Tree (CART) Analysis'' Ucla Medical Center Torrance, California Presented at the 2000 Annual Meeting Of The Society For Academic Emergency Medicine İn San Francisco, California.

Long, W., Griffth, J., Selker, H., & Agostino, R. (1993). A Comparison of Logistic Regression to Decision-Tree Induction in a Medical Domain. Reprinted from Computers in Biomedical Research,26: 74-97, 1993.

What is Machine Learning?, Matlab, Mathworks, Statistics and Machine Learning Toolbox, Adress: https://www.mathworks.com/discovery/machine-learning.html

Rudd, J., & Priestley, J. (2017). "A Comparison of Decision Tree with Logistic Regression Model for Prediction of Worst Non-Financial Payment Status in Commercial Credit. Grey Literature from PhD Candidates. 5. http://digitalcommons.kennesaw.edu/dataphdgreylit/5

Ru-Ping, L. (2010). Research of Decision Tree Classification Algorithm in Data Mining. Journal of East China Institute of Technology, Natural Science 2010-02.

Tatlıdil, H. (2002). Uygulamalı Çok Degişkenli İstatistiksel Analiz, Ankara: 1. Basım, Cem WebOfset.

Yeh, I., & Hsu, T. (2018). Building Real Estate Valuation Models with Comparative Approach Through Case-Based Reasoning. Applied Soft Computing, 65, 260-271.