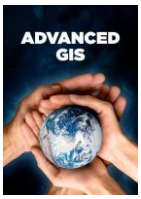
Advanced GIS

http://publish.mersin.edu.tr/index.php/agis/index

e-ISSN:2822-7026



Comparative evaluation of the performance of different regression models in land valuation

Sukran Yalpir*[1], Erol Yalpir [2]

[1]*Konya Technical University, Faculty of Engineering and Natural Sciences, Department of Geomatics Engineering, Konya, Türkiye*
[2]*Konya Provincial Directorate for National Education, Yazır Şehit Osman Küçükdillan Primary School, Selçuklu/Konya*

Keywords
Land Valuation,
Lasso,
Elastic-Net,
ML.Net,
OLS Regression

Research Article
Received: 18/01/2024
Revised: 02/02/2024
Accepted: 07/02/2024
Published: 02/04/2024

Abstract

Lands can play a dominant role in the real estate market, especially due to their legal zoning rights. These properties are preferred investment options compared to financial instruments due to factors such as high returns and long-term reliability. Today, Machine Learning (ML) algorithms are used to accurately determine the land value. Regression models, capable of handling complex relationships, integrating Geographic Information System (GIS), and providing a comparative approach, lead the way among these algorithms. In this study, Lasso, Elastic-Net, ML.Net, and Ordinary Least Squares (OLS) regression models were employed to predict land values in the central neighborhoods of Konya's Selçuklu, Meram, and Karatay districts. The datasets containing legal, physical, spatial, and local criteria of 440 lands were obtained, and GIS analyses were conducted to prepare the spatial data. Based on the modeling results, it can be observed that ML.Net exhibited successful performance with metric values of MAE=0.043, MSE=0.005, RMSE=0.060, and R2=0.82. Comparatively, ML.Net's 9% superior performance compared to the commonly encountered OLS in the literature is of significant importance. The results demonstrated the usability of various regression models for land valuation and highlighted that ML.Net can yield improved outcomes, particularly in modeling high-market-value lands.

1. Introduction

Quantitative determination of the value of the residential zoned lands, which are the basic building blocks of today's cities, with popular approaches and sustainable land management is of great importance. (Demetriou, 2016; Derdouri & Murayama, 2020). Lands can play a dominant role in the real estate market in developing cities, especially due to their legal zoning rights. They are among the most preferred investment items due to factors such as higher return amount and reliability compared to financial investment items such as foreign currency and stocks, and not losing value in the long run.

Lands, like other types of real estate, have many criteria that affect the value. Although it is very difficult to group these criteria, criteria can be defined in four main groups. These groups can be defined as legal, physical, spatial and local, respectively. The legal criteria group defines the criteria such as Base Area Coefficient (BAC), Floor Area Coefficient (FAC), and number of floors, which define the zoning status of the lands in terms of planning. The physical criteria group defines the criteria that express the situations such as the geometric shape of the land and the benefit from the infrastructure elements. The spatial criteria group defines the characteristics that define the effects of distance to urban points of interest such as education, health, and transportation on land value. The local criteria group, on the other hand, defines the characteristics that represent the social structure such as the education level of the people living in the region where the land is located, the population density, and the environmental criteria that have an effect on the land value such as air quality and noise pollution (Hu et al., 2016; Doan, 2023).

Therefore, the objective and criteria-based determining the land value with popular approaches such as Machine Learning (ML) algorithms has opened new horizons for many transactions from taxation to expropriation, from urban transformation to capital market activities (Krause & Bitter, 2012; Sisman & Aydinoglu, 2022). The usage of regressions in ML stands out as an effective approach for determining land values. Regression models can handle more complex relationships, while the integration of Geographic Information System (GIS) can enhance predictions by incorporating spatial data. Models based on linear or non-linear regression constitute one of the most commonly used approaches in practical implementation of statistical analysis of the market within a comparative approach for land valuation (Foryś & Gaca, 2018; Kokot & Gnat, 2019). In the literature, it is seen that the results obtained in the regression-based land value estimation give better results than the models based on other approaches (Zurada et al., 2011).

***Corresponding Author**

*(syalpir@ktun.edu.tr) ORCID 0000-0003-2998-3197
(eyalpir@gmail.coml) ORCID 0009-0002-9312-4354

Cite this article

Yalpir, Ş. & Yalpir, E. (2024). Comparative evaluation of the performance of different regression models in land valuation. *Advanced GIS,* 4(1), 10-14.

Within the scope of this study, as a part of land management and valuation, ML models were implemented in the central neighborhoods of Konya's Selçuklu, Meram, and Karatay districts. In the modeling process; Lasso, Elastic-Net, ML.Net and Ordinary Least Squares (OLS) regression models were used together with legal, physical, spatial and local criteria affecting the land value. The datasets were organized for the market values, properties and spatial features of 440 land offered for purchase/sale. The valuation of land-type real estate with GIS-integrated use of different methods has been realized. Results obtained from the models were compared according to the performance metrics such as determination coefficients (R^2), Mean Absolute Error (MAE), Mean-Square Error (MSE), and Root-Mean-Square Error (RMSE).

2. Materials and Methods

2.1. Study area and dataset

Neighborhoods in Selçuklu, Meram, and Karatay districts, which are among the central districts of Konya city, were determined as the study area. In the determination of the study area, the location of the land-type real estate in high market activity were taken into consideration. As seen in Figure 1, the study area also includes many urban service functions such as education, industry, agriculture and transportation facilities. All these factors were effective in determining the study area.

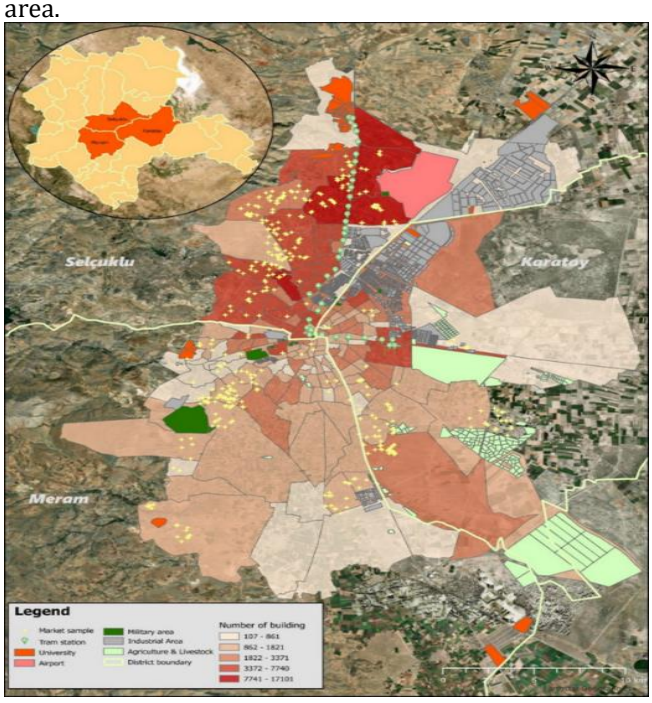


Figure 1. Determining the study area and distribution of market samples

A total of 440 market samples within the study area and geographical data representing legal (5), physical (6), spatial (11) and local (9) criteria affecting the land value were obtained. All data were organized in a GIS environment. Relationships between market samples and geographic datasets were defined. In this context, the geographical distribution of the market samples is presented in Figure 1, and summary statistics on some numerical legal and physical variables in the dataset are given in Table 1.

Table 1. Summary statistics about some numerical variables in the dataset

Variable	Min.	Max.	Mean	Standard Deviation
Market value (₺)	550000	23500000	3339727	3458173
BAC	0.08	0.90	0.24	0.07
FAC	0.15	3.60	0.59	0.34
Number of floors	1.00	8.00	2.44	0.90
Area (m^2)	117.00	5462.00	875.20	648.11
Facade length (m)	4.00	100.00	25.41	13.43
Number of facades	1.00	4.00	1.47	0.62
Road with (m)	4.00	40.00	11.87	4.70

Then, the local influence distances for each of the spatial criteria were defined by literature research as given in Table 2. Each criterion was analyzed by using its related distance values. The maps of the analysis results for some criteria such as distance to transportation, city center, shopping center and green areas are given in Figure 2. In this way, the analyzes of all criteria were completed and the results were spatially brought together with market samples. Thus, an enriched dataset was prepared for modeling studies.

Table 2. Buffer distances for spatial criteria (Sisman et al., 2023)

Spatial criteria	Analysis Distance (m)
Distance to healthcare facilities	250-500-750-1000-1250
Distance to education facilities	750-1000-1500-2000-2500
Distance to public agencies	1000-2000-3000-4000-5000
Distance to security units	1000-2000-3000-4000-5000
Distance to shopping malls	500-1000-1500-2000-2500
Distance to cultural facilities	250-500-750-1000-1250
Distance to entertainment facilities	500-1000-1500-2000-2500
Distance to green areas	750-1500-2250-3000-3750
Distance to transportation facilities	600-1200-1800-2400-3000
Distance to insanitary areas	250-500-750-1000
Distance to city center	1000-2000-3000-4000-5000

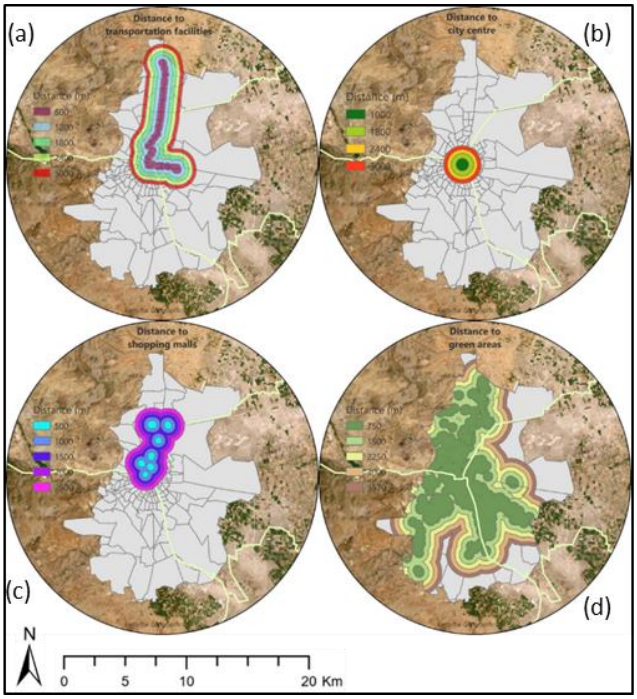


Figure 2. Examples of spatial buffer analysis for (a) distance to transportation, (b) city center, (c) shopping center and (d) green area criteria

2.2. Lasso regression

Lasso is a method used in regression analysis, particularly aimed at reducing model complexity, eliminating unnecessary variables, and preventing overfitting. Essentially, it aims to simplify the model by driving the coefficients of irrelevant variables closer to zero. This method is especially employed in high-dimensional data to diminish the impact of irrelevant features and enhance the model's generalization ability (Tibshirani, 1996). The mathematical equation for Lasso regression is as follows (Equality 1):

$$\text{Lasso} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (1)$$

where:

y_i and $\hat{y}_i$ are the observed and predicted values,

λ: Regularization parameter. This parameter controls the complexity of the model and encourages coefficients to approach zero.

$\sum_{j=1}^{p} |\beta_j|$: Using the L1 norm, it represents the sum of the absolute values of all coefficients. This part constitutes the fundamental property of Lasso as it performs variable selection by driving coefficients towards zero.

2.3. Elastic-Net regression

Elastic-Net regression, developed by Zou & Hastie (2005), builds upon the Ridge and Lasso regression methods. Similar to Ridge regression, the correction process is carried out by following the same procedural step. The λ2 parameter applies correction to the β coefficients based on the role of each coefficient in the sum of squared errors. Variable selection is performed similar to the Lasso regression. The coefficients of insignificant variables are set to zero, thus achieving automated variable selection. The equation for Elastic-Net regression is as follows (Equality 2):

$$ElasticNet = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha \left[\frac{1}{2}(1-\lambda_2)\sum_{j=1}^{p}\beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|\right] \quad (2)$$

where:

n is the number of data points,

p is the number of features,

y_i and $\hat{y}_i$ are the observed and predicted values,

β_j are the regression coefficients,

α is a parameter controlling the regularization,

λ is a parameter controlling the balance between L1 and L2 regularization.

Elastic-Net regression is often effective when dealing with feature selection or datasets with multicollinearity. However, proper tuning of regularization parameters like λ and α is crucial. These parameters can be determined using hyperparameter tuning techniques.

2.4. ML.Net model

ML.Net is an open-source machine learning framework developed by Microsoft. ML.Net allows .Net developers to add machine learning capabilities to their applications while continuing to utilize the .NET platform. It supports various machine learning tasks such as image recognition, natural language processing, classification, regression, clustering, and many more (Ramel, 2018).

Resources and examples related to ML.Net can be found in Microsoft's official documentation as well as in resources provided by the open-source community. ML.Net can assist both novice and experienced developers in easily developing machine learning projects within the .Net ecosystem (Microsoft, 2018).

2.5. Ordinary least squares (OLS) regression

OLS is a statistical method and one of the fundamental techniques in linear regression analysis (Dismuke & Lindrooth, 2006). Its objective is to model the relationship between one or more independent variables and a dependent variable. This relationship is expressed using a linear equation (Equality 3):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_{ij} \quad (3)$$

Y_i is the dependent variable (the value to be predicted). Xik, are the independent variables. βi are the regression coefficients and εij is the error term. OLS aims to predict these coefficients based on a given dataset. The predicted coefficients are calculated to minimize the sum of squared errors. Hence, it's often referred to as the "Least Squares Method".

3. Findings

In the study, the data of 31 variables that affect the land value were organized, and the performance of Lasso, Elastic-Net, ML.Net, and OLS regression models were evaluated. The total number of market samples used for the modeling stage was 440. To achieve high-performance results from the models, the model needed to be trained.

This means that approximately 80% of the dataset was used for training the models, while the remaining portion was used for testing. Optimization was performed to determine the most suitable parameters for each model, and hyperparameter tuning was conducted using 10-fold cross-validation. Table 3 presents the optimal parameters of the regression-based approaches used for land value prediction.

Table 3. Optimal hyperparameters of the various regression models

Model type	Hyperparameters	Optimum value
Lasso Regression	λ	0.004
Elastic-Net Regression	λ_1	0.355
	λ_2	0.331
ML.Net Regression	*Learning_rate*	0.009
	Max_depth	3
	Min_samples_leaf	2
	Other parameters	default
OLS Regression	*Coefficients (Intercept)*	-0.2438
	Other parameters	(31 variables coefficients)

In Table 3, it can be observed that the parameter to be determined for Lasso regression is the λ parameter. According to the analysis results, when the optimum value of λ is 0.004, it implies that the model is most successful under these conditions. In other words, it has been concluded that Lasso regression should be trained with the parameter "λ=0.004".

As for the Elastic-Net regression model, the parameters that need to be determined are λ1 and λ2. Here, as λ1 and λ2 approach 0, it signifies that the parameter has no effect and the equation transforms into the least squares method, while moving towards infinity indicates that the parameter is increasing and regression coefficients will be almost equal to zero. For the training of the Elastic-Net regression, the optimal hyperparameters have been found as "λ1=0.355" and "λ2=0.331".

In the ML.Net application conducted through Microsoft Visual Studio 2022, the optimal model selected is "FastTreeRegressionTrainer" and the optimal hyperparameter values are shown in Table 3. Finally, for the OLS regression, the constant and variables coefficients of the mathematical function were found.

After determining the optimal parameters, regression models were created using the training dataset. Model validation was performed using the test dataset. Table 4 shows the R2, MAE, MSE, and RMSE over the Lasso, Elastic-Net, ML.Net and OLS regression models for the land value prediction.

Table 4. Optimal hyperparameters of the regression models

Performance metrics		R^2	MAE	MSE	RMSE
Lasso	Training	0.88	0.034	0.003	0.053
	Test	0.74	0.050	0.006	0.076
Elastic-Net	Training	0.89	0.031	0.003	0.051
	Test	0.79	0.043	0.004	0.065
ML.Net	Training	0.94	0.028	0.002	0.047
	Test	0.82	0.043	0.005	0.060
OLS	Training	0.84	0.039	0.004	0.060
	Test	0.73	0.052	0.007	0.081

Among various regression methods, it has been concluded that ML.Net (FastTreeRegressionTrainer) model yields superior results. The results of the Lasso and Elastic-Net models have shown similarity between training and test data. The success of these models is slightly higher compared to OLS regression (Table 4). In comparison to the most commonly used method in land valuation, which is OLS, the fact that ML.Net performs 9% better in terms of test data is highly significant within the scope of the study. This situation implies that ML.Net improves the prediction results by 9% compared to OLS in predicting land market values. The accordance between predicted values from regression models and the market value of the lands subject to sale has been assessed using test data (Figure 3).

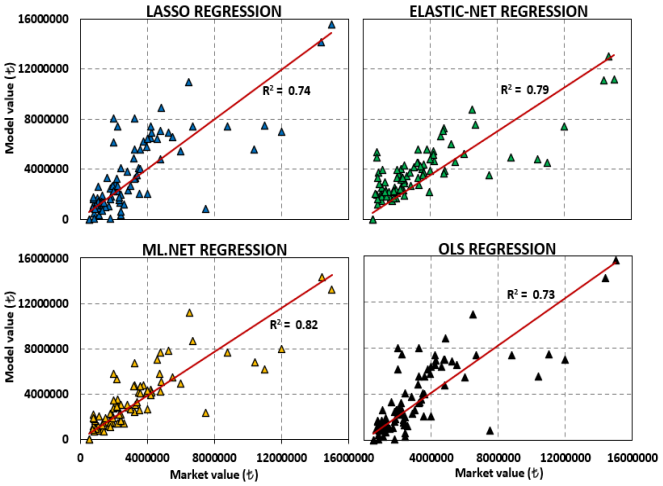


Figure 3. The accordance between model and market values (for test data)

Based on the graphical distributions in Figure 3, an R^2 value approaching 1 and data points nearing the trend line indicate that the model results are closer to market values. The R^2 between the predicted land values and market values are found to be 0.74, 0.79, 0.82, and 0.73 for Lasso, Elastic-Net, ML.Net, and OLS methods, respectively. From the results, it can be observed that Lasso and OLS regression produce similar outcomes, while Elastic-Net regression yields slightly different results compared to these two methods. Due to the similarity in the mathematical equations underlying Lasso and OLS models, there are no significant prediction differences.

Particularly, in modeling lands with high-market values within the dataset, the ML.Net algorithm has shown better performance and achieved good results compared to other methods. Therefore, for datasets with

a small number of lands with high market values, ML.Net modeling can be recommended.

4. Conclusion

Valuation methods play a crucial role in accurately predicting and comprehending values in the real estate market. This study has examined the significance and impact of Lasso, Elastic-Net, ML.Net and OLS regression models in the field of land valuation. As a result of the ML.Net modeling, successful performance (test data) results were obtained with MAE=0.043, MSE=0.005, RMSE=0.060, and R2=0.82 metric values. Compared to OLS, which is frequently encountered in the literature when determining market value, the better performance of ML.Net (about 9%) is quite notable for this study. ML.Net provides users with a wide range of features, assisting in tasks ranging from data analysis for land valuation to predictions. Features such as fast model training, data preprocessing tools, multi-platform support, and integration of pre-trained models make ML.Net a strong choice for users predicting real estate values.

Author Contributions

Author1*:* Conceptualization, methodology, visualization, investigation. **Author2**: Data curation, writing-original draft preparation, writing-reviewing and editing

Statement of Conflicts of Interest

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

Research and publication ethics were complied with in the study.

References

Demetriou, D. (2016). The assessment of land valuation in land consolidation schemes: The need for a new land valuation framework. *Land use policy*, 54, 487-498. https://doi.org/10.1016/j.landusepol.2016.03.008

Derdouri, A., & Murayama, Y. (2020). A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across Fukushima prefecture, Japan. *Journal of Geographical Sciences,* 30, 794-822. https://doi.org/10.1007/s11442-020-1756-1

Dismuke, C., & Lindrooth, R. (2006). Ordinary least squares. *Methods and designs for outcomes research*, 93(1), 93-104.

Doan, Q. C. (2023). Determining the optimal land valuation model: A case study of Hanoi, Vietnam. *Land use policy*, 127, 106578. https://doi.org/10.1016/j.landusepol.2023.106578

Foryś, I., & Gaca, R. (2018). Intuitive methods versus analytical methods in real estate valuation: preferences of Polish real estate appraisers. *In Problems, Methods and Tools in Experimental and Behavioral Economics: Computational Methods in Experimental Economics (CMEE) 2017 Conference*. Łódź, Poland 79-87.

Hu, S., Yang, S., Li, W., Zhang, C., & Xu, F. (2016). Spatially non-stationary relationships between urban residential land price and impact factors in Wuhan city, China. *Applied Geography*, 68, 48-56. https://doi.org/10.1016/j.apgeog.2016.01.006

Kokot, S., & Gnat, S. (2019). Simulative verification of the possibility of using multiple regression models for real estate appraisal. *Real Estate Management and Valuation*, 27(3), 109-123. https://doi.org/10.2478/remav-2019-0029

Krause, A. L., & Bitter, C. (2012). Spatial econometrics, land values and sustainability: Trends in real estate valuation research. *Cities*, 29, 19S25. https://doi.org/10.1016/j.cities.2012.06.006

Microsoft. (2018). ML.Net: Machine Learning made for .Net. Microsoft.

Ramel, D. (2018). Open Source, Cross-Platform ML.Net Simplifies Machine Learning -- Visual Studio Magazine. *Visual Studio Magazine*.

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. https://doi.org/10.1016/j.landusepol.2022.106167

Sisman, S., Akar, A. U., & Yalpir, S. (2023). The novelty hybrid model development proposal for mass appraisal of real estates in sustainable land management. *Survey Review*, 55(388), 1-20. https://doi.org/10.1080/00396265.2021.1996797

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso*. Journal of the Royal Stat. Society Series B: Statistical Methodology,* 58(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

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. https://doi.org/10.1080/10835547.2011.12091311