7th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Comparison of machine learning regression methods for mass real estate valuation

Batuhan Kamil Sağlam [*1], **Muhammed Oğuzhan Mete** [1], **Ufuk Özerman** [1], **Reha Metin Alkan** [1]

[1] *Istanbul Technical University, Department of Geomatics Engineering, Istanbul, TÜRKİYE*

Keywords
Real Estate Valuation
Mass Valuation
Machine Learning
Prediction Model
Inverse Distance Weighting

Abstract

Efficient management of real estate requires an objective assessment of their values by using scientific approaches. Valuation is key for value-related applications such as purchase and sale, taxation, expropriation, and urban regeneration. Mass valuation reduces time and costs by evaluating multiple properties simultaneously. Leveraging statistical analysis and predictive capabilities of machine learning enhances accuracy and speed in real estate valuation. This study focuses on applying many regression models for mass valuation of residential properties in Melbourne, Australia, aiming to improve accuracy and efficiency for stakeholders. Evaluating various algorithms, including Linear Regression, Decision Trees, Random Forest, Bagging, AdaBoost, Gradient Boosting, and XGBoost, on Kaggle's open data, performance metrics are calculated. Notably, ensemble methods like Random Forest and XGBoost consistently outperformed others by capturing nonlinear relationships of determinants and predicting the value accurately. Finally, applying the Inverse Distance Weighting (IDW) interpolation method, a real estate value map is generated for the study area. This study aims to uncover machine learning's role and limitations in real estate valuation by comparing the performance of different ensemble learning methods. The findings highlight the significance of advanced regression models in improving valuation practices, supporting decision-making, and enhancing market efficiency.

1. Introduction

Real estate valuation is a comprehensive process crucial for determining the value of a property, serving as a cornerstone for various transactions. Conventionally, methods like sales comparison, income, and cost approaches have been employed. Nevertheless, recent advances in methodologies such as Multiple Regression Analysis (Benjamin, Guttery, & Sirmans, 2020; Yilmazer & Kocaman, 2020), Hedonic Pricing (Lisi, 2019; El Yamani, Ettarid, & Hajji, 2019), Nominal Valuation (Mete & Yomralioglu, 2019; Yomralioglu, 1993), Geographically Weighted Regression (Dimopoulos & Moulas, 2016; Sisman & Aydinoglu, 2022; Wang, Li, & Yu, 2020), Ensemble Learning (Alfaro-Navarro et al., 2020; Aydinoglu, Bovkir, & Colkesen, 2021; Gnat, 2021), and Artificial Neural Networks (Demetriou, 2017; Lee, 2023; Yalpır, 2018) have led more widespread usage of those approaches in mass valuation. In the context of real estate valuation, these modern techniques have been instrumental in providing more accurate estimations and expediting decision-making processes (Doldur & Alkan, 2021). Furthermore, the integration of machine learning has catalyzed a transformative shift in the real estate valuation landscape, promising even greater precision and efficiency for assessments (Ngiam & Khor, 2019). Mass valuation methods can be broadly classified into two groups: hedonic-based regression and machine-learning regression approaches. Hedonic pricing, a prevalent technique for forecasting housing prices, considers both internal and external property characteristics to ascertain value. While known for producing robust predictive results, implementing this approach often requires specialized knowledge in statistical analysis and model specification (Mete & Yomralioglu, 2023). The city of Melbourne, renowned for its dynamic urban landscape and vibrant real estate market, serves as a focal point for this study. Melbourne's diverse property types and market fluctuations provide an ideal ground for testing the robustness and effectiveness of different machine learning regression models. The integration of advanced methodologies in mass valuation processes has facilitated more precise

* **Corresponding Author**

*(saglamba17@itu.edu.tr) ORCID ID 0000-0003-0287-787X
(metemu@itu.edu.tr) ORCID ID 0000-0002-9312-1965
(ozerman@itu.edu.tr) ORCID ID 0000-0001-9812-2185
(alkanr@itu.edu.tr) ORCID ID 0000-0002-1981-9783

Cite this study

Saglam BK, Mete MO, Ozerman MU & Alkan RM (2023). Comparison of Machine Learning Regression Methods for Mass Real Estate Valuation. Intercontinental Geoinformation Days (IGD), 7, 106-110, Peshawar, Pakistan

estimations and informed decision-making processes. Machine learning, in particular, has played a pivotal role in revolutionizing the industry, enabling the analysis of complex data structures and relationships with high accuracy and efficiency (Ngiam & Khor, 2019).

In the study using data from Melbourne, Australia, the power of machine learning was utilized and a comprehensive evaluation of the performance criteria was made, and the contribution of using machine learning methods in the real estate valuation process to the efficiency and accuracy of the valuation processes was tried to be revealed. Within this paper, it is not only prioritized highlighting the accurate machine learning models but also conducting a detailed comparison of performance metrics. By meticulously comparing the results generated from traditional valuation methodologies with those derived from the machine learning algorithms, predictive capability and efficiency of the modern approaches are highlighted.

2. Machine Learning-based Real Estate Valuation

A field of computing algorithms called machine learning is constantly developing and aims to replicate human intelligence by learning from the environment (El Naqa & Murphy, 2015). With the advent of machine learning, data driven approaches have gained prominence in the field, revolutionizing the way real estate professionals evaluate properties. In this article, we explore a range of machine learning algorithms used in real estate valuation, including linear regression, decision trees, random forests, bagging regressor, AdaBoost regressor, gradient boosting regressor, and extreme gradient boosting regressor.

Linear Regression is a statistical modeling technique used in real estate valuation, estimating the relationship between independent variables and the dependent variable as given Eq. (1).

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n \qquad (1)$$

where, $\hat{y}$ is dependent variable; X_n is independent variable; b_0 is y-constant value and b_n is coefficients.

It helps predict property prices based on factors like size, location, and amenities. It captures linear trends, aiding in informed decisions and laying the groundwork for advanced modeling techniques in dynamic markets.

Real estate valuation is crucial in property investment and market analysis. Decision trees, a machine-learning model, have revolutionized property assessment and pricing. They provide structured, intuitive predictions based on input variables, allowing for accurate valuation. This paper explores the application and benefits of decision trees in real estate valuation, highlighting their predictive capabilities. The structure of the algorithm is given in Figure 1.

The Bagging Regressor model is a machine learning technique that enhances prediction performance by combining multiple decision trees. It's useful in real estate valuation due to its ability to capture complex interactions and nonlinear relationships, allowing accurate property value predictions. It can handle large datasets and deliver reliable results. The Random Forest Regressor is a machine learning method that combines multiple decision trees to create a decision forest, improving prediction performance. It's particularly useful in real estate valuation, as it can handle complex datasets, non-linear relationships, and noisy data, enhancing accuracy and decision-making.

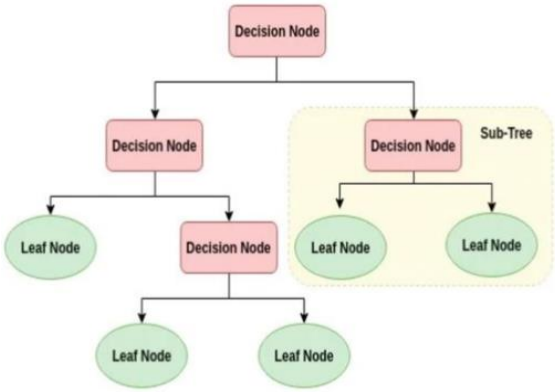

Figure 1. Basic structure of decision tree algorithm

The Gradient Boosting Regressor model is a robust real estate valuation tool that utilizes multiple decision trees to identify significant predictors and make precise property price estimations. XGBoost is a powerful regression tool that uses gradient boosting and regularization techniques to analyze complex data, identify influential variables, and make informed investment decisions. On the other hand, Ada-Boost Regressor is a machine learning strategy that uses boosting to produce highly accurate predictions. Iteratively focusing on misclassified instances and assigning higher weights improves accuracy. This versatile approach helps real estate professionals understand market dynamics, identify influential variables, and estimate property values.

3. Mass valuation with Machine Learning: A case study of Melbourne City, Australia

In this study, different machine learning regression methods such as Linear Regression, Random Forest, Ada-Boost, XGBoost are used for mass valuation of residential properties, and model performances are compared. Mass valuation processes contain data preprocessing, model development, model accuracy assessment, and valuation map production (Figure 2).

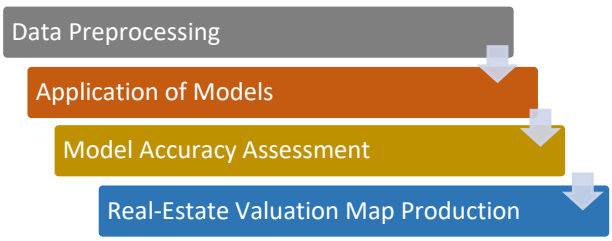

Figure 2. Workflow diagram of the study

While developing mass valuation regression models, The Melbourne Housing Snapshot dataset, which is derived from the Kaggle platform, is used. This dataset can help us understand the general state of the market and predict future trends since it provides valuable

insights about the real estate market in the City of Melbourne, Australia. Melbourne is the capital of the Victoria state of Australia. After Sydney, it is the second most crowded city with a 9,993 km² metropolitan area.

The key features of the Melbourne Housing Snapshot dataset include the number of rooms, type, price, distance, number of bedrooms, land size, year built, municipality, latitude, longitude, region name, and number of properties. The dataset contains a total of 13,580 samples and the highest price value is determined as 9,000,000 AUD, while the lowest price value is recorded as 85,000 AUD (Figure 3). This range indicates that the real estate market in Melbourne has a fairly wide price range. The standard deviation of price is calculated as 1,075,684 AUD. This shows how volatile housing prices in Melbourne are overall. Visualizing the data on a map helps determine possible price trends, as well as the geographical distribution of housing prices.

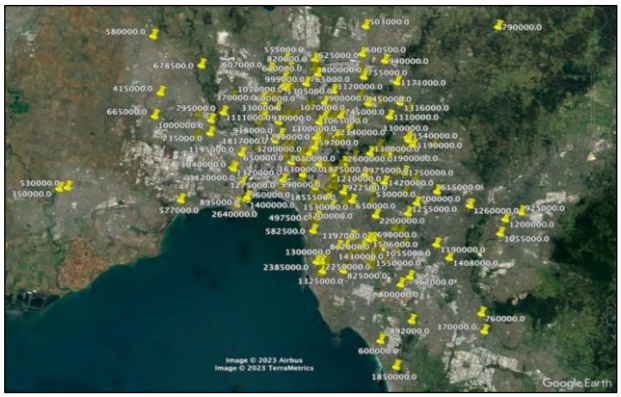

Figure 3. Spatially distributed samples of the Melbourne Housing Snapshot dataset

Within the scope of regression analysis, Exploratory Data Analysis (EDA) is first performed to obtain detailed information about the data. EDA is an indispensable step for data science projects, providing a better understanding of the data using statistics and various visualization techniques. At this stage, descriptive statistics (minimum, maximum, mean, standard deviation, etc.) are calculated, null values in the data are checked, correlation matrix, histogram, box plot, scatter plot, and pairwise comparison graphs are created.

Box plot analysis is used to detect and remove outliers. In this method, points outside the limits determined by the quartile range and median of the data distribution are considered outliers, and they are removed from the data set. Then, the 2D graph created by removing these outliers from the data set is shown in Figure 4.

In Figure 5, 3D visualization of the remaining data was performed. This process enabled a more homogeneous and representative visualization of the data set and increased the reliability of the analysis results.

The correlation matrix of determinants is used to establish the relationships between variables in the dataset (Figure 6). The correlation matrix contains the correlation coefficients that show the relationship between each pair of variables. When the correlation matrix is examined, it is seen that the number of rooms, bedrooms, bathrooms, and building area have a high correlation with property prices.

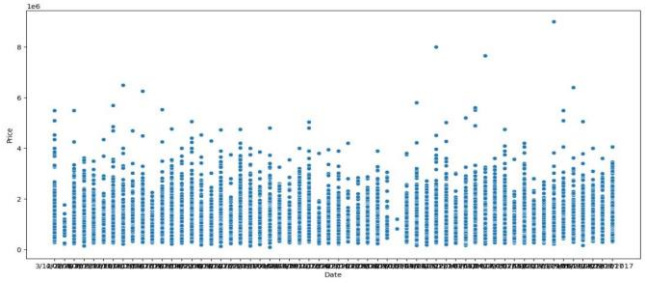

Figure 4. Display of data on graph before outliers are removed

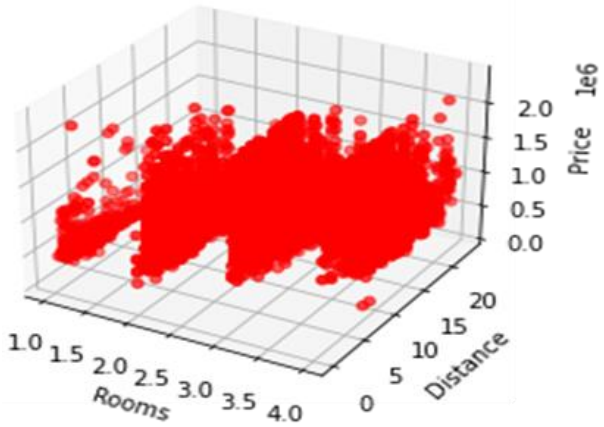

Figure 5. 3D visualization of data after outlier removal

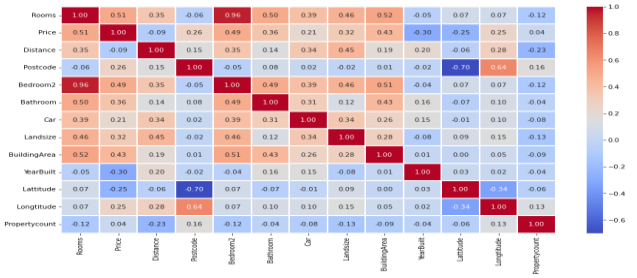

Figure 6. Correlation matrix of the determinants

In the EDA process, a positive correlation is observed between independent variables and property price. However, outliers can significantly affect this correlation. After outliers are removed, a higher correlation is observed between property price and the other determinants.

4. Results

Interpreting the results of the study, Table 1 contains significant metrics used to evaluate the performance of different regression algorithms. Among these metrics, Prediction Score (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) stand out. MAE represents the mean value of the absolute differences between actual and predicted values, while RMSE is the square root of the mean value of the squares of these differences. The results were taken into account to determine which model performed best and which model was less suitable for certain datasets.

Table 1. Regression models and performance metrics

Linear Regression			
Train-Test Split (%)	R^2	MAE	RMSE
75 - 25	0.659310	196989.181160	254281.379577
80 - 20	0.657725	197020.394116	254333.875054
85 - 15	0.662422	197041.308263	256050.971488
90 - 10	0.663747	199722.787849	259143.666717
Decision Trees Regressor			
Train-Test Split (%)	R^2	MAE	RMSE
75 - 25	0.627048	188136.539305	266049.036833
80 - 20	0.624262	185960.863014	266476.603100
85 - 15	0.636636	187174.043451	265650.051759
90 - 10	0.628760	187521.209632	272291.957281
Random Forest Regressor			
Train-Test Split (%)	R^2	MAE	RMSE
75 - 25	0.810018	133228.888338	189885.386483
80 - 20	0.811294	132797.522498	188846.789752
85 - 15	0.816629	133367.614061	188714.254893
90 - 10	0.823304	131936.028623	187854.172165
Bagging Regressor			
Train-Test Split (%)	R^2	MAE	RMSE
75 - 25	0.782739	143070.818058	203060.631339
80 - 20	0.794855	134483.965722	196900.253935
85 - 15	0.791527	142722.343037	201216.525166
90 - 10	0.807556	138637.171001	196046.780131
AdaBoost Regressor			
Train-Test Split (%)	R^2	MAE	RMSE
75 - 25	0.554980	244615.349281	290619.562314
80 - 20	0.505043	255277.041956	305843.963484
85 - 15	0.542380	247881.764403	298120.210653
90 - 10	0.541683	253421.545375	302545.302392
Gradient Boosting Regressor			
Train-Test Split (%)	R^2	MAE	R RMSE
75 - 25	0.783971	147336.171724	202484.172220
80 - 20	0.785251	147152.509979	201456.943290
85 - 15	0.794758	146142.888033	199651.220882
90 - 10	0.794952	148050.077514	202365.239109
Extreme Gradient Boosting Regressor			
Train-Test Split (%)	R^2	MAE	RMSE
75 - 25	0.816906	132114.238361	186411.051174
80 - 20	0.817800	133547.754318	185562.469265
85 - 15	0.825908	131702.227114	183877.622347
90 - 10	0.830326	132555.493272	184083.676625

In this study, the performance of different regression algorithms for predicting property prices based on different rates of train-test split is evaluated. Evaluating the results, Linear regression performed good performance, but high Mean Absolute Error and Mean Squared Error values indicated deviation between predicted and actual prices. Decision Trees Regressor showed comparable prediction scores but slightly higher error metrics, suggesting model instability. XGBoost had the highest prediction scores and lower mean absolute and squared errors, indicating better data complexities and improved accuracy. On the other hand, Bagging Regressor yielded satisfactory results, while AdaBoost Regressor had the lowest prediction scores and highest errors, suggesting limitations in capturing underlying patterns. Further analysis and experimentation may be necessary to fine-tune these algorithms.

In order to produce mass valuation map, the Inverse Distance Weighting (IDW) method is used to interpolate estimated real estate prices across the study area using the ArcGIS platform. The IDW calculates the estimated value of a point based on the values of other known points around that point and their distances to those points. The map on the ArcGIS platform shown in Figure 7 visualizes the points predicted by the Extreme Gradient Boosting Regressor model. This visualization shows how close the model's predictions are to actual data. Considering how close the points on the map are to the actual data and how accurate the predictions are overall, it can be concluded that this model is a robust prediction tool applied in the Melbourne property market with high accuracy.

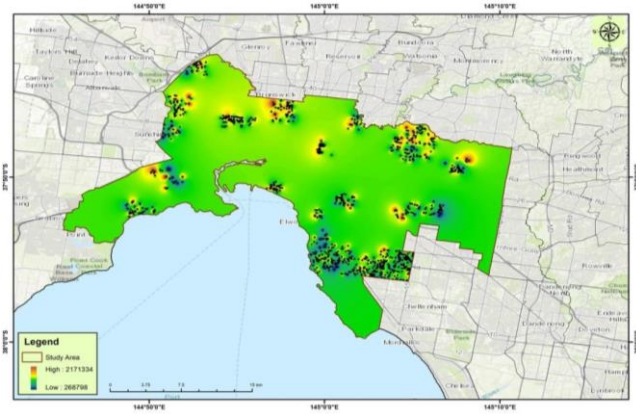

Figure 7. Display of predicted points on the map on the ArcGIS platform

5. Conclusion

Mass valuation is a method used in real estate that involves the assessment of multiple properties at once, typically utilizing statistical models to determine the value based on various factors and characteristics. Regression models for real estate valuation have been analyzed, showing their strengths and limitations. In contrast, machine learning methods, specifically XGBoost, have offered a data-driven and objective approach that leverages the power of algorithms to analyze vast amounts of real estate data, including property characteristics, market trends and location factors, as we see in our study. This allows for the creation of more accurate and reliable property valuations, reducing the risk of overvaluation or undervaluation. Aside from their superior accuracy, machine learning methods also offer several advantages over nominal methods. Machine learning models can be constantly updated and improved as new data becomes available. This ensures that valuations remain updated and reflect current market conditions.

Moreover, machine learning algorithms can identify complex patterns and relationships within real estate data that appraisers cannot see, leading to a more comprehensive understanding of property values. Considering these advantages, it is clear that machine learning methods represent a significant advance in mass valuation applications, as seen in the study. By adopting machine learning techniques, Türkiye can increase the efficiency, accuracy, and transparency of its property

valuation system, ultimately benefiting both homeowners and government agencies.

Acknowledgement

This study was carried out as part of the Design Project thesis prepared by B.K. Sağlam at ITU Geomatics Engineering.

References

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

Aydinoglu, A.C., Bovkir, R., & Colkesen, I. (2021). Implementing a mass valuation application on interoperable land valuation data model designed as an extension of the national GDI. Survey Review, 53(379), 349–365. https://doi.org/10.1080/00396265.2020.1771967

Benjamin, J.D., Guttery, R. S., & Sirmans, C. F. (2020). Mass appraisal: an introduction to multiple regression analysis for real estate valuation. Journal of Real Estate Practice and Education, 7(1), 65–77. https://doi.org/10.1080/10835547.2004.12091602

Demetriou, D. (2017). A spatially based artificial neural network mass valuation model for land consolidation. Environment and Planning B: Urban Analytics and City Science, 44(5), 864–883. https://doi.org/10.1177/0265813516652115

Dimopoulos, T., & Moulas, A. (2016). A proposal of a mass appraisal system in Greece with CAMA system: evaluating GWR and MRA techniques in Thessaloniki municipality. Open Geosciences, 8(1), 675–693. https://doi.org/10.1515/geo-2016-0064

Doldur, M., & Alkan, R. M. (2021). Producing GIS-based land value maps by using nominal valuation method: case study in Avanos/Nevşehir. Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi, 21(4), 846–863 (in Turkish). https://doi.org/10.35414/akufemubid.888502

El Naqa, I., & Murphy, M.J. (2015). What is machine learning? In: El Naqa, I., Li, R., Murphy, M. (eds) Machine Learning in Radiation Oncology. Springer, Cham. https://doi.org/10.1007/978-3-319-18305-3_1

Gnat, S. (2021). Property mass valuation on small markets. Land, 10(4):388. https://doi.org/10.3390/land10040388

Lee, C. (2023). Designing an optimal neural network architecture: an application to property valuation. Property Management. 41(1), 84-96. https://doi.org/10.1108/PM-12-2021-0106

Lisi, G. (2019). Property valuation: the hedonic pricing model-location and housing submarkets. Journal of Property Investment & Finance, 37(6), 589–596. https://doi.org/10.1108/JPIF-07-2019-0093

Mete, M.O., & Yomralioglu, T. (2019). Creation of nominal asset value-based maps using GIS: a case study of Istanbul Beyoglu and Gaziosmanpasa districts. GI_Forum Journal for Geographic Information Science, 7(2), 98–112. https://doi.org/10.1553/giscience2019_02_s98

Mete, M.O., & Yomralioglu, T. (2023). A hybrid approach for mass valuation of residential properties through geographic information systems and machine learning integration. Geographical Analysis, 55(4), 535-559. https://doi.org/https://doi.org/10.1111/gean.12350

Ngiam, K.Y., & Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. The Lancet Oncology, 20(5), e262–e273. https://doi.org/10.1016/S1470-2045(19)30149-4

Sisman, S., & Aydinoglu, A.C. (2022). A modelling approach with geographically weighted regression methods for determining geographic variation and influencing factors in housing price: a case in Istanbul. Land Use Policy, 119(106183). https://doi.org/10.1016/j.landusepol.2022.106183

Wang, D., Li, V.J., & Yu, H. (2020). Mass appraisal modeling of real estate in urban centers by geographically and temporally weighted regression: a case study of Beijing's core area. Land, 9(143). https://doi.org/10.3390/LAND9050143

Yalpır, Ş. (2018). Enhancement of parcel valuation with adaptive artificial neural network modeling. Artificial Intelligence Review, 49(3), 393–405. https://doi.org/10.1007/s10462-016-9531-5

El Yamani, S., Ettarid, M., & Hajji, R. (2019). Building information modeling potential for an enhanced real estate valuation approach based on the hedonic method. WIT Transactions on the Built Environment, 305–316. https://doi.org/10.2495/bim190261

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. https://doi.org/10.1016/j.landusepol.2020.104889

Yomralioglu, T. (1993). A nominal asset value-based approach for land readjustment and its implementation using geographical information systems. University of Newcastle upon Tyne.