Performance Comparison of Machine Learning Algorithms for Accurate Office Real Estate Price Estimation: Suggested Approaches

Jacek Maślankowski
University of Gdańsk
Gdańsk, Poland *jacek.maslankowski@ug.edu.pl*

Małgorzata Rymarzak
University of Gdańsk
Gdańsk, Poland *malgorzata.rymarzak@ug.edu.pl*

Abstract

This paper aims to test various machine learning algorithms on a real-world dataset of office real estate and identify the most accurate one. A comprehensive analysis was conducted using proprietary data on office real estate in Poland, obtained from a leading market intelligence provider specializing in commercial property analytics, covering a 20-year observation period. The research results indicate that among six tested algorithms including Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Decision Tree Classifier (DT), Gaussian Naive Bayes (GNB), Support Vector Machines (SVM), the Decision Tree Classifier (DT) appears to be the best-fit algorithm for selecting factors to estimate office real estate prices.

Keywords: big data, price estimates, artificial intelligence

1. Introduction

Office real estate price forecasting is a critical concern for both investors and policymakers. Investors rely on forecasts for effective portfolio allocations, risk management and strategic planning. Meanwhile, policymakers utilize price prediction for market assessments and the design, implementation, and adjustments of policies, particularly to prevent market overheating and stimulate businesses when necessary.

Although the literature presents various research on machine learning algorithms for estimating real estate prices, a notable research gap remains, particularly concerning the use of real-world databases of office real estate. Most existing studies primarily focus on algorithm comparisons, especially within the context of residential properties [15], which cannot be easily transferred to the commercial real estate sector due to its unique characteristics (i.e. greater heterogeneity). Furthermore, many of these studies are confined to specific geographical regions, particularly Western European nations, the USA and Asia [11, 19], owing to the limited availability of office real estate data in other areas. Therefore, this paper aims to test various machine learning algorithms on a real-world dataset of office real estate, specifically within the Polish market, and to identify the most accurate one. The research questions are as follows:

RQ1. Can a linear model be used to estimate the price of office real estate?
RQ2. Which classifiers provide the best estimates for the price of office real estate?
RQ3. Which machine learning model is optimal for estimating office real estate prices?

The paper is organized into five sections. Following the introduction is a literature review. The third section presents the data sources and methods used. The fourth section outlines the methodological framework, leading to the conclusions presented in the fifth section.

2. Literature review

The application of machine learning in real estate studies, compared to traditional models derived from comparative methods, offers a significant advantage in understanding

the complex relationships between different factors that influence real estate prices.

Various machine learning algorithms are described in the literature for accurately estimating real estate prices. Among these, Neural Networks are the most used (e.g., [2, 7, 9, 11, 18]). In the early 1990s, several authors highlighted issues associated with Neural Networks [18]. For instance, the average absolute error varied significantly depending on the algorithm employed in different software packages, leading to often unstable results [8]. Conversely, Nguyen and Cripps [13] demonstrated that Neural Networks are particularly effective when applied to large heterogeneous datasets.

Other algorithms reported to be effective include but are not limited to, k-Nearest Neighbors [1, 5], Decision Tree [14, 17, 20], and Random Forest [4, 6, 10, 12, 21].

3. Data source and methods

The research objective was accomplished through an exhaustive examination of a proprietary dataset of office real estate in Poland obtained from REDD AI, a premier market intelligence firm specializing in commercial property analytics. The longitudinal dataset comprises 1,223 transaction records collected over a 20-year observation window (2002–2022), encompassing 528 unique office real estate. Each entry in the dataset was characterized by 91 quantitative and qualitative variables, systematically categorized into two broad groups: market and property-related.

4. Methodological Framework for Estimations of Office Real Estate Prices Using Machine Learning

This study employed machine learning (ML) algorithms to develop a predictive model for office real estate prices and identify the most influential factors (market and property-related) to maximize the accuracy of price estimation. To achieve this overarching goal, the following subsidiary objectives were established:

O1. Implementation of Linear Regression for Real Estate Price Estimation

Linear regression was utilized to answer RQ1. Ordinary least squares (OLS) regression was applied to establish baseline predictive performance in property price estimation; however, the results were unsatisfactory. Consequently, a decision was made to reduce the dimensions of the data source and utilize classes of price ranges instead.

O2. Factors Correlation Analysis and Dimensionality Reduction

A pairwise correlation analysis among real estate factors was conducted to eliminate non-predictive ones and mitigate multicollinearity, thereby enhancing model generalizability (Fig.1.).

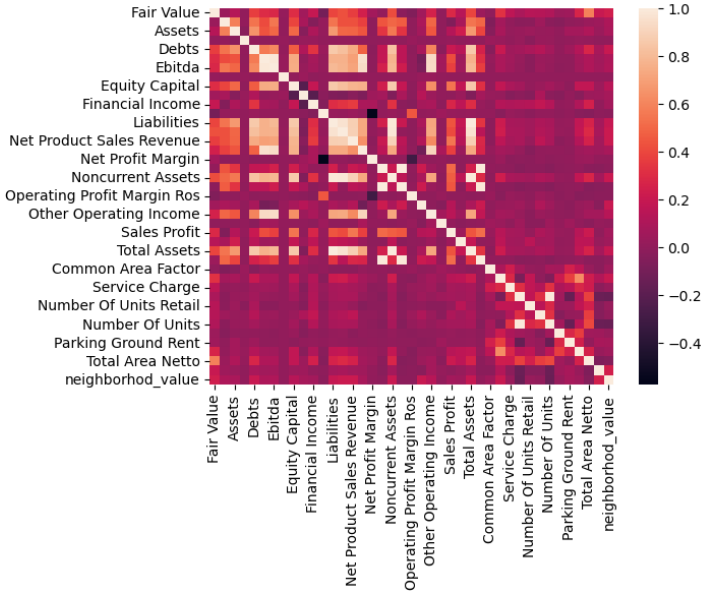

Fig. 1. Correlation heatmap for selected variables.

O3. Comparative Evaluation of Machine Learning Algorithms for Office Real Estate Price Classification

Six supervised machine learning algorithms were evaluated on stratified property price classes derived from a curated dataset of 790 complete records (from an initial 1,223 observations after handling missing data). The algorithms assessed included: Logistic Regression (LR); Linear Discriminant Analysis (LDA); K-Nearest Neighbors (KNN); Decision Tree Classifier (DT); Gaussian Naive Bayes (GNB), and Support Vector Machines (SVM).

Model performance was quantified using classification accuracy, reported as the mean (standard deviation) across validation runs: LR: 0.340 (±0.127); LDA: 0.602 (±0.041); KNN: 0.768 (±0.027); DT: 0.824 (±0.031); GNB: 0.543 (±0.052); SVM: 0.641 (±0.048). The validation size used in the test amounted to 10%.

Consistent with prior theoretical expectations, the decision tree algorithm demonstrated superior predictive performance, achieving the highest classification accuracy of 82.4% (± 3.1%). As a result, subsequent analytical efforts were directed toward optimizing this model through hyperparameter tuning and refining the training set to maximize generalizability (Fig.2).

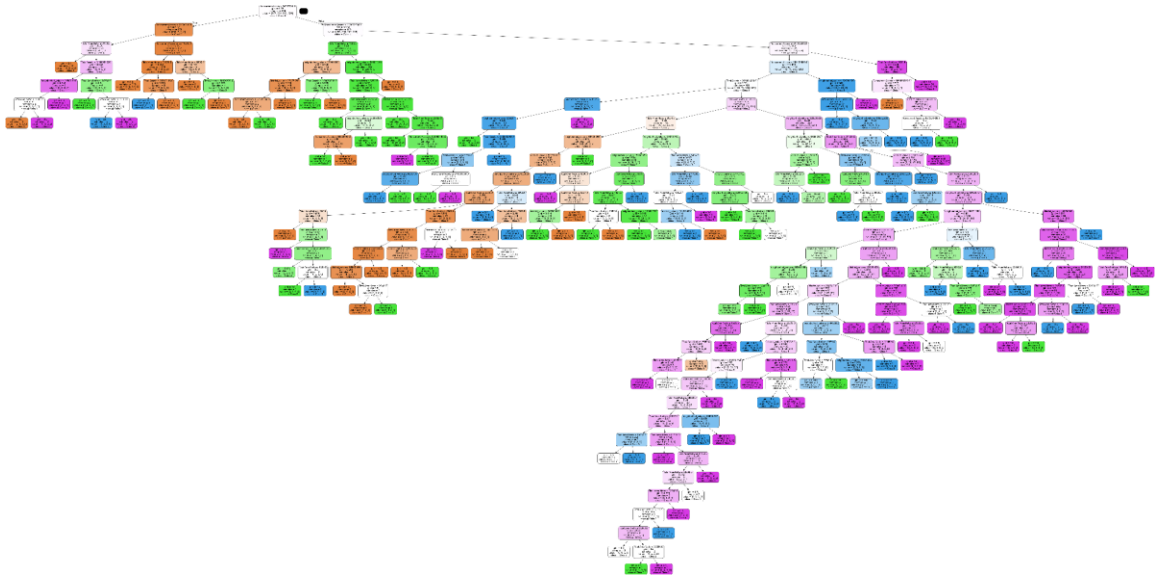

Fig. 2. Decision Tree for selecting factors to estimate office real estate prices.

O4. Algorithmic Optimization and Interpretability Analysis: Data Imputation and Factors Selection Methodology

A comprehensive evaluation of the top-performing model was conducted, including hyperparameter tuning, validation metrics, and factor importance analysis to ensure both predictive accuracy and interpretability.

The subsequent phase of the analysis involved addressing missing data through systematic imputation. This critical preprocessing step served two primary purposes: (1) preserving dataset completeness to maintain statistical power, and (2) enhancing the representativeness of property classifications. Without such imputation, the effective sample size would have been reduced by approximately 35.4% (from n=1,223 to n=790), potentially introducing selection bias.

Consistent with established protocols for handling missing data [16], mean substitution was employed for continuous variables. This conservative approach demonstrated satisfactory performance in preliminary analyses, yielding imputed values that preserved the predictive accuracy of subsequent machine learning applications (see Methods 3.2 for validation details).

The final factor space for model training comprised:

- Predictor variables (X): Neighborhood economic index ('Neighborhood_value'),

Administrative district valuation ('District_value'), Net leasable area ('Total Area Netto'), Fixed asset holdings ('Noncurrent Assets'), Balance sheet total ('Total Assets')

- Target variable (y): Property classification ('Class')

Notably, multicollinearity was observed between 'Noncurrent Assets' and 'Total Assets' (Variance Inflation Factor VIF = 4.2, $r = 0.81$, $p < 0.001$). However, both variables were retained due to their established theoretical importance in property price models [3] and their complementary information content—where 'Noncurrent Assets' captures capital intensity, while 'Total Assets' reflects overall financial scale.

5. Conclusions

Various machine learning algorithms have been presented in the literature for estimating real estate prices. In this study, six algorithms were tested. Among them, the Decision Tree Classifier (DT) appears to be the best-fitted algorithm for selecting factors to estimate office real estate prices.

In response to RQ1, as presented in objective 1 of the study, linear regression yielded less accurate results than the algorithms that utilized labelled training datasets. The labels represent the ranges of office real estate prices. Therefore, in the case study the decision was to use ranges of values instead of numeric variable as the output.

RQ2 was answered through objectives 2 and 4 of the study, where selected factors were presented as part of the final factors space in the fourth section of the paper. In particular, the inclusion of all features may provide noise in the data. The suggested approach shows that it is important to find the most accurate features and not to include feature which may have accidental impact on the result values.

Finally, RQ3 was answered by objective 3, which showed the general accuracy of machine learning algorithms tested on the office real estate database. Of them, the most accurate was the decision trees, trained with the data currently available in the dataset provided by the REDD AI company.

The research was funded by the project "Development of an Intelligent Tool for Estimating the Value of REDD AI Office Real Estate", under the RPLD.01.02.02-IP.02-10-077/21 competition within the Regional Operational Programme of the Łódź Voivodeship for 2014-2020.

References

1. Borde, S., Rane, A., Shende, G., Shetty, S.: Real estate investment advising using machine learning. International Research Journal of Engineering and Technology (IRJET). 4(3), 1821-1825 (2017)
2. Chou, J. S., Fleshman, D. B., Truong, D. N.: Comparison of machine learning models to provide preliminary forecasts of real estate prices. Journal of Housing and the Built Environment. 37(4), 2079-2114 (2022)
3. Geltner, David and Kumar, Anil and Van de Minne, Alex, Riskiness of Real Estate Development: A Perspective from Urban Economics & Option Value Theory (2018). http://dx.doi.org/10.2139/ssrn.2907036
4. Guo, J. Q., Chiang, S. H., Liu, M., Yang, C. C., Gou, K. Y.: Can machine learning algorithms associated with text mining from internet data improve housing price prediction performance? International Journal of Strategic Property Management. 24(5), 300-312 (2020)
5. Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., Cai, Z.: Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modelling to inform equitable housing policies. Land Use Policy. 82, 657-673 (2019)
6. Huang, Z., Lai, G.: A House Price Prediction Model Based on K-means Clustering and

Random Forest in Guangzhou. Frontiers in Business, Economics and Management. 10(2), 377-381 (2023)

7. Kauko, T.: On current neural network applications involving spatial modelling of property prices. Journal of Housing and the Built Environment. 18(2), 159-181 (2003)
8. Kontrimas, V., Verikas, A.: The mass appraisal of the real estate by computational intelligence. Applied Soft Computing. 11(1), 443-448 (2011)
9. Liu, J., Zhang, G., Wu, W.: Application of fuzzy neural network for real estate prediction. LNCS. 3973, 1187-1191 (2006)
10. Mohd, T., Harussani, M., Masrom, S.: Rapid modelling of machine learning in predicting office rental price. International Journal of Advanced Computer Science and Applications. 13(12), 543-549 (2022)
11. Mohd, T., Jamil, S., Masrom, S.: Machine learning building price prediction with green building determinant. IAES International Journal of Artificial Intelligence (IJ-AI). 9(3), 379-386 (2020)
12. Mora-Garcia, R. T., Cespedes-Lopez, M. F., Perez-Sanchez, V. R.: Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. Land. 11(11), 2100 (2022)
13. Nguyen, N., Cripps, A.: Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. The Journal of Real Estate Research. 22(3), 313-336 (2001)
14. Park, B., Bae, J. K.: Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications. 42(6), 2928-2934 (2015)
15. Rico-Juan, J.R., de La Paz, P.T.: Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain, Expert Systems with Applications, 171, 114590 (2021), DOI 10.1016/j.eswa.2021.114590
16. Rubin, D.B., Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc., New York. (1987) http://dx.doi.org/10.1002/9780470316696
17. Sarı, P., Bedirhan, İ. D., Sel, Ç.: Comparing the Practical Differences Between Decision Tree and Random Forest Algorithms in Estimating Housing Prices. Current Trends in Computing. 1(2), 149-157 (2024)
18. Worzala, E., Lenk, M., Silva, A.: An exploration of neural networks and its application to real estate valuation. Journal of Real Estate Research. 10(2), 185-201 (1995)
19. Xu, X., Zhang, Y.: Office property price index forecasting using neural networks. Journal of Financial Management of Property and Construction. 29(1), 52-82 (2024)
20. Yücebaş, S., Doğan, M., Genç, L.: A C4. 5–Cart Decision Tree Model for Real Estate Price Prediction and the Analysis of the Underlying Features. Konya Journal of Engineering Sciences. 10(1), 147-161 (2022)
21. Zhang, Y., Rahman, A., Miller, E.: Longitudinal modelling of housing prices with machine learning and temporal regression. International Journal of Housing Markets and Analysis. 16(4), 693-715 (2023)