

DETERMINING THE INFLUENCE OF REAL ESTATE FEATURES ON PRICES WITH PARTIAL DEPENDENCE PLOTS: A CASE STUDY IN SZCZECIN, POLAND

Sebastian Gnat*¹

¹ Department Econometrics and Statistics, University of Szczecin, ul. Mickiewicza 64, 71-101 Szczecin, Poland, e-mail: sebastian.gnat@usz.edu.pl, ORCID: 0000-0003-0310-4254

* Corresponding author

ARTICLE INFO	ABSTRACT
Keywords: real estate market analysis, partial dependence plots	The study explores the application of Partial Dependence Plots (PDP) in the analysis of real estate features. The study centers on a selected real estate market in Szczecin, Poland, aiming to highlight the efficacy of PDP in understanding and interpreting the complex relationships between various features and property prices. The primary objective is to showcase the potential of PDP in capturing the nuanced interactions between real estate attributes and their impact on market prices. The CatBoost model, known for its robust handling of categorical features and strong predictive capabilities, is employed as the machine learning algorithm for this analysis. The performance of this model will be compared against a traditional multiple linear regression model, providing insights into the advantages of leveraging advanced machine learning techniques in real estate analysis. Results obtained from the analysis will be presented and discussed, shedding light on the interpretability and accuracy of the CatBoost model compared to the traditional linear regression approach. The presentation will conclude with implications for real estate practitioners and researchers, emphasizing the potential for PDP to enhance the transparency and understanding of complex models in the real estate domain. This research contributes to the growing body of knowledge on the application of advanced machine learning techniques in real estate analysis.
JEL Classification: C10, R30	
Citation:	
	Gnat, S. (2024). Determining real estate features influence on prices with partial dependence plots: A case study in Szczecin, Poland. <i>Real Estate Management and Valuation</i> , 32(4), 105-116. https://doi.org/10.2478/remav-2024-0039

1. Introduction

Market prices and valuation represent the culmination of multifaceted interactions between various economic, social, and environmental factors. Comprehension of these factors is paramount as it enables stakeholders to make informed decisions regarding property transactions, investments, and development strategies. In real estate, market prices are the result of a complex interplay of supply and demand dynamics. Factors such as population growth, employment opportunities, income levels, and migration patterns significantly influence demand for properties, thereby impacting prices. Similarly, the availability of land, construction costs, zoning regulations, and infrastructure development dictate the supply side of the market equation. Valuation, on the other hand, encompasses a rigorous assessment

of a property's worth based on numerous criteria. Traditional valuation methodologies encompass approaches such as the income approach, which considers the property's potential income generation capacity, the cost approach, which evaluates the cost of replacing the property, and the sales comparison approach, which benchmarks the property against comparable sales in the market. Furthermore, modern valuation techniques may incorporate advanced statistical models, machine learning algorithms, and geographic information systems (GIS) to provide more accurate and nuanced valuations. In the area of real estate mass appraisal, well established approaches to analyzing real estate data have relied on hedonic models (Metzner & Kindt, 2018; Sirmans et al., 2005; Yavuz Ozalp & Akinci, 2017; Yoo et al., 2012), which involve estimating the value of a property based on its

characteristics, such as location, size, age, amenities, and proximity to various amenities. However, with the advancements in technology and the availability of large-scale data, machine learning algorithms have emerged as a powerful tool for predicting property prices and valuations (Pai & Wang, 2020). These algorithms can analyze vast amounts of data and uncover complex patterns and relationships that may not be apparent through traditional methods (Bilgilioğlu & Yılmaz, 2023; Lee, 2022; Mora-Garcia et al., 2022; Trawiński et al., 2017). By utilizing machine learning models and actual transaction data, researchers are able to not only accurately predict real estate prices but also identify the key variables that significantly influence market dynamics such as supply and demand, economic conditions, demographic factors, and neighborhood amenities. This can provide valuable insights for real estate owners, agents, policymakers, and investors in making informed decisions regarding property investments, pricing strategies, and urban planning initiatives also to anticipate market trends, mitigate risks, identify investment opportunities, and optimize asset portfolios.

Although hedonic models and machine learning algorithms can be used, in some circumstances, interchangeably, there is an apparent trade-off between them. This trade-off concerns two elements – prediction accuracy and model interpretability (James et al., 2021). Some statistical models are inflexible (linear regression, for example) but highly interpretable; others, like random forests are more flexible, but their results are very hard to interpret. Greater flexibility in this context means the ability of the model or algorithm to better capture complex (non-linear, segmented) relationships between predictors and the explanatory variable. Greater flexibility results in a better fit to the real data. If the researcher is interested in making inferences about the phenomenon he or she is concerned with, it is better for him or her to use restrictive models, the predictive power of which is weaker. On the other hand, if the main objective is to make the most accurate prediction possible, it is worth turning to more flexible models. When considering the level of interpretability of models and their predictive power, it is worth pointing out that, while the former is due to the assumptions and design of the models and model comparisons due to interpretability are not controversial, comparing models based on the latter is no longer such a straightforward issue. For example,

there have been a number of studies in the area of the property market, which have used different statistical models and have led to inconclusive results.

The literature on utilizing machine learning models for property valuation is vast, and can be categorized into two main topics. The first one involves studies where researchers apply and enhance existing methods such as multiple regression (Zaddach & Alkhatib, 2014), regression trees (McCluskey et al., 2014), random forests (Antipov & Pokryshevskaya, 2012), support vector machines (Wang et al., 2014), or artificial neural networks (Demetriou, 2017; Zhou et al., 2018). The second one focuses on comparing various algorithms to ascertain which one produces better predictions. For instance, in one study, the effectiveness of forecasting property prices in Fairfax County, Virginia, was examined (Park & Bae, 2015). Another study evaluated the English housing rental market (Clark & Lomax, 2018) using generalized linear regression, machine learning, and an expert approach for mass appraisal. Apart from the conclusions regarding the fact that machine learning models proved to be superior to multiple regression, the authors argue that the use of machine learning is computationally demanding, which was also confirmed in this study. An interesting comparison of the random forest and multiple regression in the Cyprus market (Dimopoulos et al., 2018) demonstrated that the random forest outperformed the linear models. Morano et al. (2018) presented two mass appraisal procedures in the Italian residential property market. They tested a utility additive method, which views property price formation as a multi-criteria selection process based on property characteristics crucial in the real estate market. This method was compared to a hybrid data-driven technique called evolutionary polynomial regression, which employs multi-objective genetic algorithms to identify models that optimize both data accuracy and mathematical simplicity. In comparative studies, artificial neural networks are frequently utilized as representatives of machine learning. Their superiority over multiple regression models was showcased in the case of New York (Khamis & Kamarudin, 2014). Machine learning models have also been pitted against expert approaches, where the former often exhibited better performance (Trawiński et al., 2017). Zurada et al. (2011) conducted comparative research employing various regression methods and artificial intelligence for property appraisal, with results indicating a slight superiority of non-traditional

regression-based methods. It was emphasized that study outcomes are heavily influenced by the specifics of the property market, property types, or the size of the analyzed dataset. Such a statement could be considered debatable. It is therefore important to conduct research that tests the research tools and statistical methods used by other authors on other data sets, and other property markets, in order to draw firm conclusions. Despite existing studies showcasing the advantage of machine learning methods, other ones have reported no significant differences between, for instance, neural networks and multiple regression, or even some studies showed that neural networks were found to be inferior (Del Giudice et al., 2017). This ambiguity in research findings underscores the necessity for further exploration in comparing multiple regression with a broad set of machine learning models, especially considering the perspective that data science and big data are pivotal for the future of real property valuation (Doszyń, 2020).

In the area of machine learning, the algorithms that typically feature the most accurate predictions are algorithms that do not allow for an understanding of what, and to what extent, contributed to such and not other predictions - which are otherwise accurate. This situation has created a need for the development of methods that would enrich the precise, in terms of predictions, algorithms with the possibility of interpreting the causes of individual predictions. The research area that deals with increasing the transparency and interpretability of machine learning algorithms is called interpretable machine learning (IML) or explainable AI (XAI). The latest developments in this area are presented by researchers in various research fields (Molnar, 2022; Samek, 2023). A definition of interpretability and its extension can be found in the work of Miller T. (2019). He understands the interpretability of a model as: "The degree to which an observer can understand the cause of a decision". Adding a component to machine learning models to understand the results they generate is a very important issue for the development of science. In their book, Molnar and Freiesleben (2024) recognize, following in the steps of earlier researchers, that inference and understanding the causes of various phenomena is an indispensable part of science as such.

Applications of machine learning algorithms can be found in almost every sphere of interest for researchers around the world. It is therefore not

surprising that IML applications are also emerging in many areas, to mention medicine, finance, agriculture, meteorology, biology and many others. The real estate market area is no exception. Krämer et al. (2023) propose the use of the XGBoost algorithm, a non-parametric machine learning technique, in combination with the model-agnostic Accumulated Local Effects Plots (ALE), to provide a deeper representation of real estate market fundamentals, allowing for a better understanding of the determinants of residential real estate values. In the area of commercial real estate, the application of IML was presented by Deppner et al. (2023). They conclude that traditional appraisal methods in the commercial sector have remained unchanged for decades, while statistical learning methods have shown potential in capturing quickly changing market dynamics and complex pricing processes in the residential property sector. They also found that the level of interpretability varies for different types of commercial property - the understanding of the models is greatest for apartments and industrial properties, followed by office and retail buildings. The aforementioned trade-off between interpretability and accuracy have been taken under consideration by Kucklick & Müller (2023). The authors address the issue of applications of deep learning methods in property valuation. In this case they use satellite imagery. Their study raises the issue of the lack of interpretability of such models. Therefore, they suggest an innovative method of explanation, which they call Grad-Ram. Their proposed method of mass appraisal outperforms traditional hedonic regression models by 34% in terms of mean absolute error. They also conclude that satellite images are the second most important predictor after square feet in their model and that the neural network learns interpretable patterns about the neighborhood structure and density. To date, a number of different computational techniques have been proposed to interpret the results obtained by machine learning algorithms. One such technique is partial dependence plots (PDPs), which will be discussed in the following section of the paper. Lorenz et al. (2023) demonstrate the use of IML methods, including PDPs in real estate market analysis, allowing for a better understanding of the relationships between variables and algorithmic decision-making processes. An interesting issue raised here is the question of the impact of time on the relationship between property characteristics and their prices. The authors create interpretations for each of the years separately and conclude that the

relationships between variables and rental prices can change over time, emphasizing the need for periodic analysis. Interpretable machine learning can also be used in the field of property tax assessment (Lee, 2022). The author states that the low interpretability of black-box machine learning models, such as neural networks, hinders their adoption in industry. His study utilizes partial dependence plots to interpret the price estimations. According to Lee, the interpretable machine learning approach, using tools like partial dependence plots, is expected to facilitate the adoption of machine learning in the industry, particularly in property valuation. In his study, some real estate characteristics are qualitative, which is very interesting from the standpoint of the real estate market in Poland; they are not, however, extensively studied, which creates a research gap for this study.

The literature review conducted in the area of the application of statistical models in real estate market analysis, including mass real estate valuation, as well as in the area of computational techniques for interpreting advanced machine learning algorithms, resulted in the research objectives. The realization of these objectives will help to reduce the research gap in the area of advanced statistical model applications in the real estate market. The first research objective relates to the ongoing discussion on the greater predictive power of advanced models in real estate applications. The study will compare the accuracy of price predictions obtained using a multiple regression model and a state-of-the-art CatBoost model. Given that real estate markets are often local in nature, meaning that the regularities within them may vary from market to market, the results of such a

comparison will contribute to the growing body of research on the application of quantitative methods in real estate price predictions.

The second, and indeed main, research objective is to verify the feasibility of using partial dependence plots to discover the impact of property characteristics on their prices. If the CatBoost model turns out to give better predictions, then enriching it with the possibility of interpreting what characteristics and to what extent influenced property prices will be, as to speak, a combination of both worlds, i.e. the inference capability typical of multiple regression models with the greater predictive power of machine learning models. In this respect, the article will also contribute to reducing the research gap. To the author's current knowledge, research to date has focused little on the qualitative characteristics of real estate, which in some countries (e.g. Poland) are the main means of describing it. The second element that has not been found much in research is the inclusion of the time variable in the statistical modeling. The property market and the prices quoted on it are subject to constant change. In the publications to date on the adoption of partial dependence plots, the authors divide the data sets into individual years and analyze each of the years separately. In this study, time will be incorporated as one of the explanatory variables.

2. Material and methods

The survey is based on a dataset of 1170 dwelling transactions in Szczecin, Poland between 2017 and 2020. The dwellings in the analyzed set were described by 9 features. A brief overview of the dataset is presented in Tables 1 and 2.

Table 1

Descriptive statistics in unit prices (in PLN - Polish zlotys) of properties and their attributes defined for a set of 1170 units

Statistics	Price of 1m ²	Usable area [sqm.]	Right to the premises	Technical condition	Electronic surveillance	Separate kitchen	Elevator	Number of rooms	Year of constr.	Year of trans.
Minimum	1 763.64	16	1	1	1	1	1	1	1	2017
First Quartile	4 303.86	40.51	1	3	1	1	1	2	2	2017
Median	5 074.47	51	2	4	1	2	1	2	2	2018
Third Quartile	6 027.73	63.64	2	4	1	2	2	3	3	2019
Maximum	11 621.47	166.8	2	5	2	2	2	4	3	2020
Interquartile range	861.93	11.57	0.5	0.5	0.0	0.5	0.5	0.5	0.50	1
Positional Coefficient of Variation V_Q (%)	16.99	22.69	25.00	12.50	0.0	25.00	50.00	25.00	25.00	0.05

Source: own elaboration.

Table 2

Qualitative features of properties and their variants

No.	Feature	Feature variant (state)
1	Right to the premises	Cooperative (1) Ownership (2)
2	Technical condition	Major overhaul needed (1) Minor overhaul needed (2) Good (3) Very good (4) Perfect (5)
3	Electronic surveillance	No (1) Yes (2)
4	Separate kitchen	No (1) Yes (2)
5	Elevator	No (1) Yes (2)
6	Number of rooms	From 1 to 4
7	Year of construction	Before 1940 (1) 1940-1989 (2) After 1989 (3)
8	Year of transaction	2017 (1) 2018 (2) 2019 (3) 2020 (4)

Source: own elaboration.

As mentioned, two kinds of regression models are used in the study. The first is a classical multiple regression model – in a log-linear form:

$$\ln(p_i) = \alpha_0 + \alpha_1 x_{1i} + \sum_{k=2}^K \sum_{p=2}^{k_p} \alpha_{kp} x_{kpi} + u_i \quad (1)$$

where:

- p_i – unit price of i -th property,
- N – number of properties ($i = 1, 2, \dots, N$),
- α_0 – constant term (intercept),
- α_1 – usable area coefficient,
- x_{1i} – usable area of i -th property,
- K – number of properties qualitative attributes ($k = 2, 3, \dots, K$),
- k_p – number of states of k -th attribute,
- α_{kp} – impact of p -th state of attribute k ,
- x_{kpi} – zero-one (dummy) variable for p -th state of attribute k of i -th property,
- u_i – random component (residuals).

The target variable is a natural logarithm of real estate unit prices. Real estate attributes (apart from usable area) are qualitative characteristics measured on an ordinal scale, so they are introduced into the model (1) through dummy variables for each state of an attribute. In model (1), there is a constant term. To avoid strict collinearity of the explanatory variables, each dummy variable for the worst attribute states is skipped. In the interpretation, the ignored state of an attribute serves as a point of reference for the

remaining states.

The second regression model is a representative of the group of gradient boosting algorithms – a regression variant of the CatBoost algorithm (Dorogush et al., 2018). Gradient boosting is an effective machine learning method renowned for delivering top-class performance across various real-world tasks. It has long been the go-to tool for tackling challenges involving diverse features, noisy datasets, and intricate relationships, such as those encountered in web search, recommendation systems, weather prediction, and numerous other domains. The main advantage of the CatBoost algorithm in the context of dataset used in current study is that it is designed to effectively manage categorical features, incorporating them into the training process rather than preprocessing them separately, as other boosting algorithms do. This approach makes the algorithm better able to deal with the problem of over-fitting, which lowers the accuracy of the prediction. The mathematical notations that lay behind this algorithm are beyond the scope of this article. This is explained in detail by its authors (Prokhorenkova et al., 2018). Although this algorithm has high predictive power, as will be verified using the real estate dataset utilized in this study, its mathematical design allows it to be referred to as a black-box model, i.e. one in which it is difficult to understand how the explanatory variables contributed to the prediction. Therefore, this algorithm was chosen for comparison with a multiple regression model - it has high predictive power but is difficult to interpret. The partial dependence plots technique will be used to interpret the model.

Partial dependence plots (PDP) were introduced by Hastie et al. (2009). PDPs show the dependence between the target response and a set of input features of interest, marginalizing over the values of all other input features (the “complement” features). Intuitively, we can interpret the partial dependence as the expected target response as a function of the input features of interest (Pedregosa et al., 2011). PDP indicates whether the relationship between an input variable and the prediction is linear, monotonic, or more complex (Lee, 2022). Good explanation of how PDPs are created is presented by Molnar (2022): “The partial dependence function for regression is defined as:

$$\hat{f}_s(x_s) = E_{X_C}[\hat{f}(x_s, X_C)] = \int \hat{f}(x_s, X_C) d\mathbb{P}(X_C) \quad (2)$$

The x_s are the features for which the partial dependence function should be plotted, and X_C are

the other features used in the machine learning model $\hat{f}$, which are treated here as random variables. Usually, there are only one or two features in the set S . The feature(s) in S are those which we want to know the effect on the prediction for. The feature vectors x_S and x_c combined make up the total feature space X . Partial dependence works by marginalizing the machine learning model output over the distribution of the features in set C , so that the function shows the relationship between the features in set S we are interested in and the predicted outcome. By marginalizing over the other features, we get a function that depends only on features in S , interactions with other features included. The partial function tells us what the average marginal effect on the prediction is for given value(s) of features S . Partial dependence plots can be produced for both continuous and categorical variables. The PDPs represents a global approach, analyzing all instances to provide an assessment of the overall relationship between a feature and the predicted outcome. The PDP is a curve (in the case of continuous explanatory variables) and a bar plot (in the case of categorical explanatory variables) that shows the input variable and the average predictions.

Predictions produced with the multiple linear model and CatBoost algorithm will be compared in terms of relational (relative) root mean square error ($rRMSE$):

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\hat{p}_i - p_i)^2}{N}} \quad (3)$$

$$rRMSE = \frac{RMSE}{\bar{p}_i}, \quad (4)$$

where:

p_i – unit price of i -th property,

$\hat{p}_i$ – predicted price of i -th property,

$\bar{p}_i$ – mean unit price,

$RMSE$ – root mean square error,

$rRMSE$ – relational root mean square error.

3. Results

3.1. Statistical modeling of property prices

The results of the conducted study are divided into two parts. The first will compare the accuracy of the modeling of the prices of the analyzed apartments obtained with both types of models, i.e. the multiple regression model and the regression CatBoost algorithm. In the second part, the interpretation possibilities of the models used will be presented, which will make it possible to assess how individual property characteristics affected transaction prices.

Studies conducted around the world indicate that, as a rule, models of the same type as CatBoost often have better accuracy, though not in every situation. This study will verify whether this is also the case in the analyzed market.

Estimated using the data described in Tables 1 and 2, the multiple regression model (1) had an adjusted R^2 coefficient of determination of 73.2%. In contrast, that of the CatBoost algorithm was 89.9%. This means a clearly higher fit of the model to the real data. The relative mean squared error (RMSE) for the multiple regression model was 13.24%, while for the CatBoost algorithm, it was 8.21%. The accuracy of the second approach used was therefore about 61% higher.

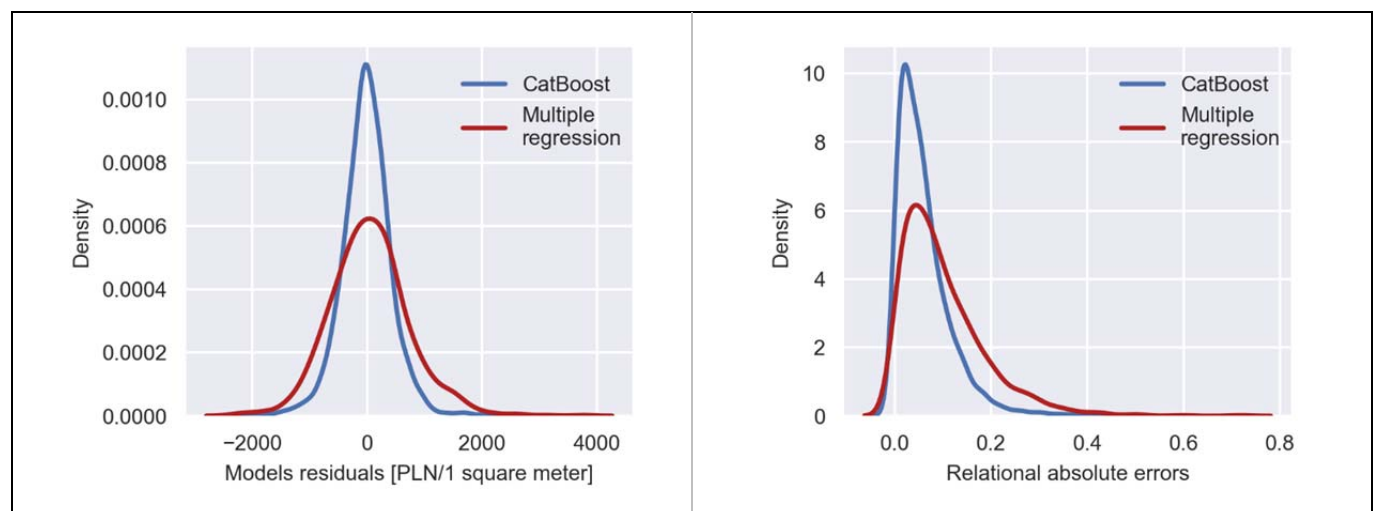


Fig. 1. Kernel density estimations of models' residuals and relational absolute errors distributions. Source: own elaboration.

A more in-depth analysis of the fit of the algorithms used to the real data is illustrated in Figure 1. The left part of the figure shows the distributions of the residuals of both algorithms. It can be seen that, for the CatBoost algorithm, the distribution of residuals is much slimmer. The range of residuals for the multiple regression model was from -2291 PLN/1 sqm. to 3756 PLN/1 sqm. For the CatBoost algorithm, it was from -2040 PLN/1 sqm. to 2101 PLN/1 sqm. The right side of Figure 1 shows the distributions of relative absolute errors. Here, it is also confirmed that the CatBoost algorithm enabled smaller errors. The least accurate price predictions for the CatBoost algorithm were considerably less (11.88% of the predictions of the multiple regression model had a prediction error of more than 20%, as compared to 2.65% for the CatBoost algorithm). Table 3 shows the descriptive statistics of the absolute percentage errors obtained using both types of regression algorithms.

Table 3

Descriptive statistics of absolute percentage errors

Statistics	Multiple regression model	Catboost algorithm
Mean	10.05%	6.03%
Standard deviation	8.62%	5.57%
Minimum	0.02%	0.00%
First Quartile	3.92%	1.99%
Median	7.74%	4.71%
Third Quartile	13.81%	8.13%
Maximum	71.79%	40.16%

Source: own elaboration.

3.2. Interpretation of models

In the first part of the study, results were obtained indicating that the CatBoost algorithm better captures regularities in the relationship between real estate features and their price. Lacking an analytical form, the algorithm does not offer the possibility of interpreting the results it produces, with interpretation being

understood as determining the impact of individual property characteristics on prices. This part of the study will discuss the interpretive capabilities of model (1) and the use of PDPs to understand the functioning of the CatBoost algorithm.

The interpretation of the multiple regression model coefficient (1) is straightforward. The intercept term $e^{8.3491} = 4226.58$ PLN/1 sqm represents the theoretical unit price for a property characterized by the lowest values of all features used to describe the given properties. Other coefficients inform us about the percentage change of theoretical price with the unit change of each variable (property feature). In the case of usable area, there is, on average, a $(e^{-0.0058} - 1) \cdot 100\% = 0.58\%$ decrease of unit price with an increase of this feature by its unit (1 square meter). For the remaining variables, which are qualitative in nature, the interpretation tells of the percentage increase in the unit price compared to the lowest state of each characteristic. For example, for perfect technical condition, a coefficient of 0.2967 informs us that housing units in this condition were, on average, $(e^{0.2967} - 1) \cdot 100\% = 34.54\%$ more expensive than units in need of major overhaul. It is also worth noting the time variable, the values (the year of the transaction) of which inform us about the percentage increase in prices compared to the first year of the survey, i.e. 2017. Prices in 2020, for example, were, on average, over 43% higher than in 2017. The interpretations of the other coefficients are analogous. Thus, it can be said that thanks to the analytical form of the model applied (1) very informative feedback can be obtained and the model gives a clear understanding of the analyzed data. Some model coefficients proved statistically insignificant. Such coefficients are: the right to the premises, electronic surveillance in the building, and the number of rooms being equal to 3 and 4.

Table 3

Results of model (1) estimation

	coefficient	standard error	t statistic	p-value
intercept	8.3491	0.025	336.096	0.000
Usable area (x1)	-0.0058	0.000	-15.736	0.000
Right to the premises_ownership (x2)	0.0163	0.010	1.696	0.090
Technical condition_minor overhaul (x3)	0.0775	0.018	4.334	0.000
Technical condition_good (x4)	0.1365	0.015	9.066	0.000
Technical condition_very good (x5)	0.2132	0.015	14.508	0.000
Technical condition_perfect (x6)	0.2967	0.017	17.617	0.000
Electronic surveillance_yes (x7)	0.0298	0.015	1.942	0.052
Separate kitchen_yes (x8)	0.0636	0.010	6.389	0.000
Elevator_yes (x9)	0.0449	0.009	4.799	0.000

	coefficient	standard error	t statistic	p-value
Number of rooms_2 (x10)	-0.0349	0.014	-2.538	0.011
Number of rooms_3 (x11)	0.0093	0.018	0.531	0.596
Number of rooms_4 (x12)	0.0405	0.026	1.551	0.121
Year of construction_1940-1989 (x13)	0.0962	0.013	7.452	0.000
Year of construction_after 1989 (x14)	0.1852	0.012	15.396	0.000
Year of transaction_2018 (x15)	0.1177	0.011	10.689	0.000
Year of transaction_2019 (x16)	0.2560	0.011	23.520	0.000
Year of transaction_2020 (x17)	0.3618	0.012	29.329	0.000

Source: own elaboration.

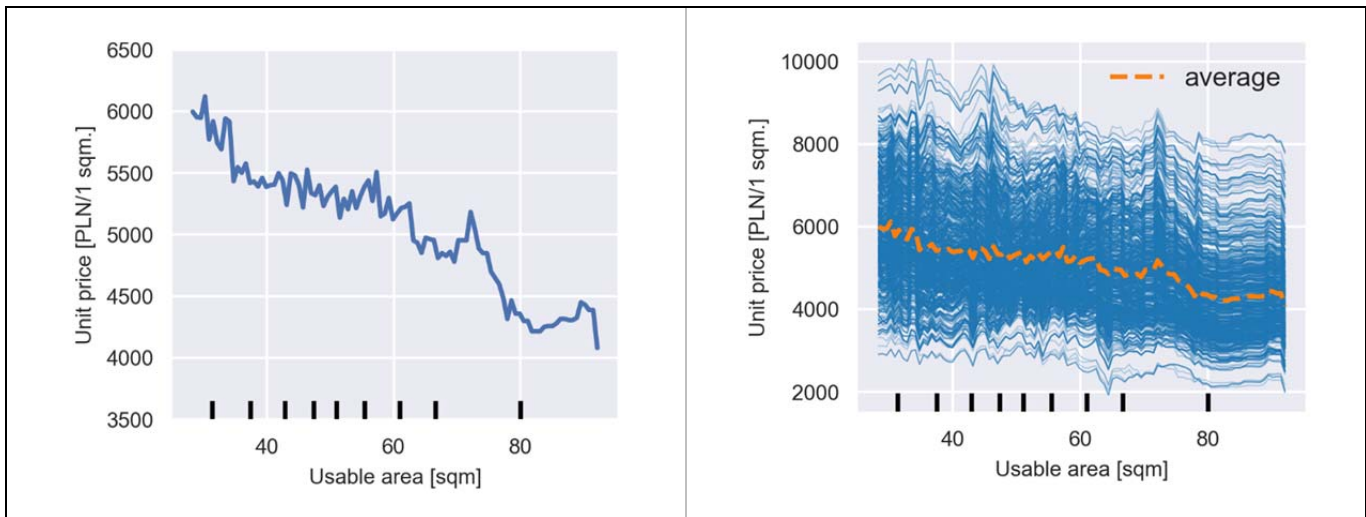


Fig. 2. Partial dependence plots for the usable area of the analyzed properties (on the left - average effect on prices, on the right - complete data). Source: Own elaboration.

Partial dependence plots were used to determine how individual property characteristics in the CatBoost algorithm affected the results of modeling the transaction prices of the residential units analyzed. Since there were two types of property characteristics (continuous - usable area and other categorical), two types of PDPs were used, i.e. line charts for usable area and bar charts for the other characteristics. The relationship between the usable floor area and prices per square meter captured by the CatBoost algorithm is shown in Figure 2. The average theoretical unit prices as a function of usable floor area are shown on the left. It can be seen that, as with the multiple regression model, the relationship is negative. The larger the usable area, the lower the price per square meter. This interpretation (found in both approaches) is consistent with real estate theory and practice. The right side of Figure 2 shows the total data for all 1170 properties. Short vertical lines shown in Figure 2 represent deciles of usable area. An analysis of Figure 2 shows that the relationship between the variables may be more complicated than a single coefficient in a multiple regression model might

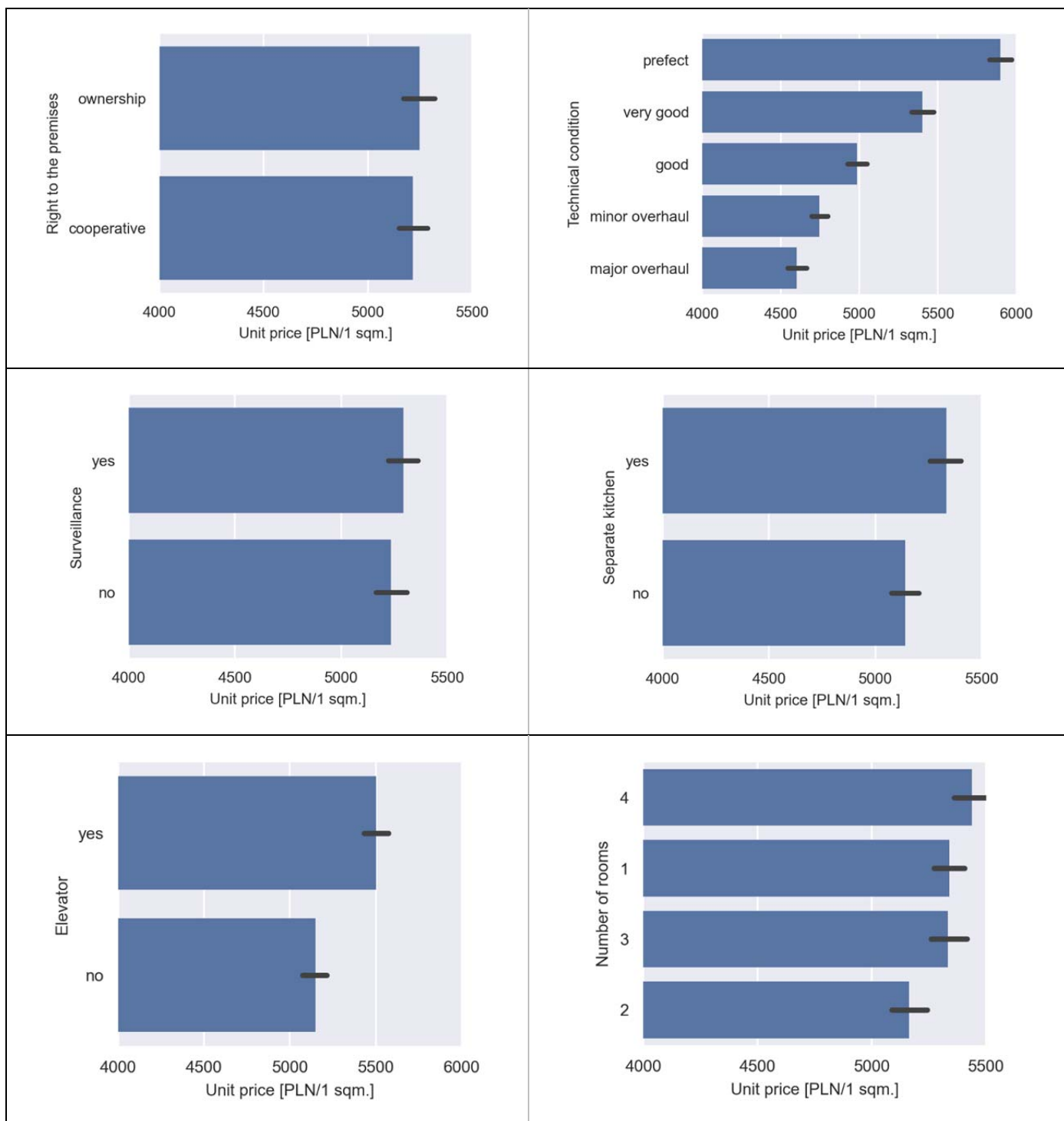
indicate. Partial dependence plots are model agnostic type of IML so it can be used for MLR models as well.

Figure 3 illustrates partial dependence plots for categorical characteristics of the analyzed properties. The bars on the charts represent average theoretical prices per square meter. The black lines indicate the limits of the price ranges for each state of each feature. For example, the theoretical average unit price for apartments in need of major overhaul was PLN 4600.90, while for apartments in perfect condition it was PLN 5899.00, a difference of 28%. This is slightly less than the value of the coefficient for the variable x6 in model (1). The time variable can serve as another example of categorical variable. Apartments sold in 2020 cost, according to CatBoost algorithm estimates, an average of PLN 6350/1 sqm. while four years earlier, their prices had been PLN 4389/1 sqm., or 44.7% less. This is very similar to the coefficient of the x17 variable in Model (1). The effects of individual characteristics on prices are slightly different, which is expected since the fit to real data was different in the two approaches (higher for the CatBoost algorithm).

When analyzing the similarities and differences in

the impact of individual property characteristics on their prices, it is also worth noting the partial dependence plots of those characteristics that were found to be statistically insignificant in Model (1). In the case of the right to the premises and electronic surveillance, the differences of average prices are very small, whereas in the case of the number of rooms, the average theoretical prices take values which are rather inconsistent with the knowledge and practice of

the real estate market. In conclusion, partial dependence plots proved to be an effective way to evaluate the impact of property characteristics on their prices in an algorithm that does not directly allow such analyses. PDPs enriched the CatBoost algorithm, whose accuracy and fit to real data was higher. It can be said that this algorithm, together with PDPs, is a valuable tool in real estate market analysis.



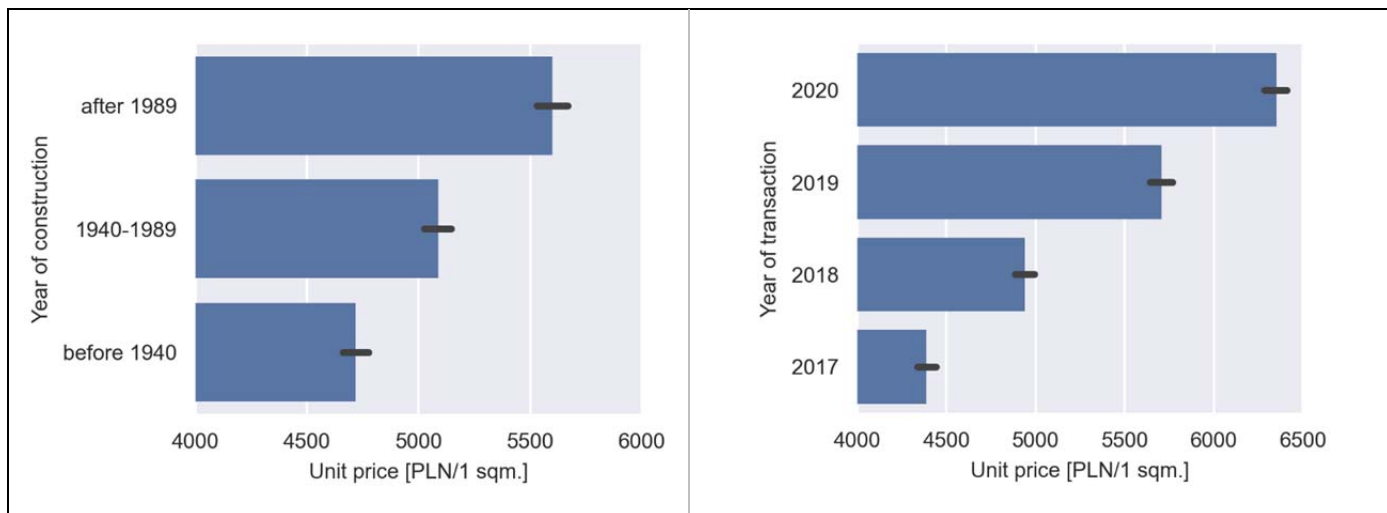


Fig. 3. Partial dependence plots for categorical features. *Source:* own elaboration.

4. Discussion

The available literature on the application of machine learning algorithms in the real estate market is quite extensive. As shown in the introduction, the results presented by different researchers are inconclusive and indicate better prediction of prices or property values for classical multiple regression models as well as situations where machine learning algorithms more accurately predict prices based on the characteristics of properties and the markets in which they are located (Baldominos et al., 2018; Pérez-Rave et al., 2019). Mayer et al. (2019) conclude that the gradient boosting method yields the greatest accuracy (the CatBoost model used in the current study is this kind of model), while the robust methods (such as multiple regression models) provide the least volatile predictions. There is a clear trade-off across methods depending on whether the goal is to improve accuracy or avoid accidentalism. As in the above study, the algorithm used here - the CatBoost model, a representative of the field of machine learning, showed greater predictive power. Prediction accuracy of this algorithm was 61% higher in terms of *RMSE*. The results obtained therefore contribute to the ongoing scientific discussion on whether modern algorithms predict property prices and values more effectively. However, the specifics of these types of algorithms do not provide a direct way of establishing why the prediction result is as it is and not otherwise (Lipton, 2018). Machine learning focuses on obtaining the most accurate predictions possible (Molnar & Freiesleben, 2024) and, as the results of this study also show, it performs this task to a satisfactory effect. In the era of Big Data, the use of machine learning is becoming increasingly common, and it seems that it

will be difficult to stop such a trend (Kok et al., 2017). In the property market, though not exclusively, the recipient of a prediction would often like to know what lies at its foundation. It is indisputable to say that knowing how individual property characteristics affect their price and value is invaluable knowledge. Hence, it is important to follow the development of methods for interpreting the results generated by a machine learning algorithm and testing their usefulness on as many datasets as possible. The development of computational techniques within the area of machine learning and their application to the real estate market has, like any human creation, its advantages and disadvantages (Seagraves, 2024). These techniques need to be recognized and tested to benefit from their use and mitigate weaknesses. The partial dependence plots used in the study provided an understanding of how the different levels of each of the variables describing the properties affected the price prediction. The visualizations obtained were very useful, as they enabled a representation of how the CatBoost algorithm would "understand" the property market being analyzed. The study therefore confirmed the usefulness of PDP in studying the impact of property characteristics on their prices.

5. Conclusions

This study contributes to the research area related to the applications of machine learning algorithms in the real estate market and the evaluation of the usefulness of techniques for interpreting real estate price forecasts that result from such algorithms. The results obtained as a result of the study made it possible to draw several conclusions. First, the accuracy of forecasts obtained using the CatBoost

algorithm was higher than those obtained using the multiple regression model. Such a situation is more common in studies conducted to date. Therefore, the results presented here provide further confirmation of the superiority of advanced algorithms in terms of predictive power. Second, it was confirmed in another real estate market that partial dependence plots are a useful tool in analyzing the extent to which the levels of individual property characteristics affect their prices. Combining an algorithm with higher predictive power with the ability to interpret model results creates an effective combination of computational techniques and confirms their applicability. Thirdly, the treatment of time as a qualitative variable also made it possible to assess the impact of time on the prices in the analyzed real estate market. As for the difficulties and obstacles connected with using this type of model, it is necessary to point out the necessity of the existence of good quality data, which should be available in sufficient quantity so that the model can learn market regularities. Such availability of data, despite the ongoing so-called Big Data era, is not an obvious fact. In the case of atypical properties, or in markets where data is not recorded in sufficient detail or the number of transactions is small, expert knowledge is still indispensable, and it is reasonable to assume that human participation in such situations will still be necessary for years to come. Further research in the area presented in this article should focus on the impact of data quality and availability on the accuracy of real estate price predictions and the interpretability of those predictions.

References

- Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 39(2), 1772–1778. <https://doi.org/10.1016/j.eswa.2011.08.077>
- Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying real estate opportunities using machine learning. *Applied Sciences (Basel, Switzerland)*, 8(11), 2321. Advance online publication. <https://doi.org/10.3390/app8112321>
- Bilgilioglu, S. S., & Yilmaz, H. M. (2023). Comparison of different machine learning models for mass appraisal of real estate. *Survey Review*, 55(388), 32–43. <https://doi.org/10.1080/00396265.2021.1996799>
- Clark, S. D., & Lomax, N. (2018). A mass-market appraisal of the English housing rental market using a diverse range of modelling techniques. *Journal of Big Data*, 5(1), 43. Advance online publication. <https://doi.org/10.1186/s40537-018-0154-3> PMID:30931238
- Del Giudice, V., De Paola, P., Forte, F., & Manganelli, B. (2017). Real estate appraisals with bayesian approach and Markov Chain Hybrid Monte Carlo Method: An Application to a Central Urban Area of Naples. *Sustainability (Basel)*, 9(11), 2138. Advance online publication. <https://doi.org/10.3390/su9112138>
- 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>
- Deppner, J., von Ahlefeldt-Dehn, B., Beracha, E., & Schaefer, W. (2023). Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Approach. *The Journal of Real Estate Finance and Economics*, 1–38. Advance online publication. <https://doi.org/10.1007/s11146-023-09944-1> PMID:38625136
- Dimopoulos, T., Tyralis, H., Bakas, N. P., & Hadjimitsis, D. (2018). Accuracy measurement of Random Forests and Linear Regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus. *Advances in Geosciences*, 45, 377–382. <https://doi.org/10.5194/adgeo-45-377-2018>
- Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. *ArXiv, abs/1810.11363*. <https://api.semanticscholar.org/CorpusID:26037613>
- Doszyń, M. (2020). Econometric support of a mass valuation process. *Folia Oeconomica Stetinensia*, 20(1), 81–94. <https://doi.org/10.2478/fole-2020-0005>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd Edition). Springer. <https://doi.org/doi.org/10.1007/978-1-0716-1418-1>
- Khamis, A. Bin, & Kamarudin, N. K. (2014). Comparative study on estimate house price using statistical and neural network model. *International Journal of Scientific & Technology Research*, 3(12), 126–131.
- Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *Journal of Portfolio Management*, 43(6), 202–211. <https://doi.org/10.3905/jpm.2017.43.6.202>
- Krämer, B., Nagl, C., Stang, M., & Schäfers, W. (2023). Explainable AI in a real estate context – Exploring the determinants of residential real estate values. *Journal of Housing Research*, 32(2), 204–245. <https://doi.org/10.1080/10527001.2023.2170769>
- Kucklick, J.-P., & Müller, O. (2023). Tackling the accuracy-interpretability trade-off: Interpretable deep learning models for satellite image-based real estate appraisal. *ACM Transactions on Management Information Systems*, 14(1), 1–24. Advance online publication. <https://doi.org/10.1145/3567430>
- Lee, C. (2022). Training and interpreting machine learning models: Application in property tax assessment. *Real Estate Management and Valuation*, 30(1), 13–22. <https://doi.org/10.2478/remav-2022-0002>
- Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43. <https://doi.org/10.1145/3233231>
- Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2023). Interpretable machine learning for real estate market analysis. *Real Estate Economics*, 51(5), 1178–1208. <https://doi.org/10.1111/1540-6229.12397>
- Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2019). Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research*, 12(1), 134–150. <https://doi.org/10.1108/JERER-08-2018-0035>

- McCluskey, W. J., Zulkarnain Daud, D., & Kamarudin, N. (2014). Boosted regression trees: An application for the mass appraisal of residential property in Malaysia. *Journal of Financial Management of Property and Construction*, 19(2), 152–167. <https://doi.org/10.1108/JFMPC-06-2013-0022>
- Metzner, S., & Kindt, A. (2018). Determination of the parameters of automated valuation models for the hedonic property valuation of residential properties: A literature-based approach. *International Journal of Housing Markets and Analysis*, 11(1), 73–100. <https://doi.org/10.1108/IJHMA-02-2017-0018>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.).
- Molnar, C., & Freiesleben, T. (2024). *Supervised machine learning for science* (1st ed.).
- Mora-Garcia, R. T., Cespedes-Lopez, M. F., & Perez-Sanchez, V. R. (2022). Housing price prediction using machine learning algorithms in COVID-19 times. *Land (Basel)*, 11(11), 2100. Advance online publication. <https://doi.org/10.3390/land11112100>
- Morano, P., Tajani, F., & Locurcio, M. (2018). Multicriteria analysis and genetic algorithms for mass appraisals in the Italian property market. *International Journal of Housing Markets and Analysis*, 11(2), 229–262. <https://doi.org/10.1108/IJHMA-04-2017-0034>
- Pai, P. F., & Wang, W. C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences (Basel, Switzerland)*, 10(17), 5832. Advance online publication. <https://doi.org/10.3390/app10175832>
- Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42(6), 2928–2934. <https://doi.org/10.1016/j.eswa.2014.11.040>
<https://doi.org/10.1016/j.eswa.2015.03.005>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pérez-Rave, J. I., Correa-Morales, J. C., & González-Echavarría, F. (2019). A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *Journal of Property Research*, 36(1), 59–96. <https://doi.org/10.1080/09599916.2019.1587489>
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 6639–6649.
- Samek, W. (2023). Explainable deep learning: concepts, methods, and new developments. *Explainable Deep Learning AI: Methods and Challenges*, 7–33. <https://doi.org/10.1016/B978-0-32-396098-4.00008-9>
- Seagraves, P. (2024). Real Estate Insights: Is the AI revolution a real estate boon or bane? *Journal of Property Investment & Finance, ahead-of-print*(ahead-of-print). <https://doi.org/10.1108/JPIF-05-2023-0045>
- Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, 13(1), 1–43. <http://www.jstor.org/stable/44103506>
<https://doi.org/10.1080/10835547.2005.12090154>
- Trawiński, B., Telec, Z., Krasnoborski, J., Piwowarczyk, M., Talaga, M., Lasota, T., & Sawiłow, E. (2017). Comparison of expert algorithms with machine learning models for real estate appraisal. *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 51–54. <https://doi.org/10.1109/INISTA.2017.8001131>
- Wang, X., Wen, J., Zhang, Y., & Wang, Y. (2014). Real estate price forecasting based on SVM optimized by PSO. *Optik (Stuttgart)*, 125(3), 1439–1443. <https://doi.org/10.1016/j.ijleo.2013.09.017>
- Yavuz Ozalp, A., & Akinci, H. (2017). The use of hedonic pricing method to determine the parameters affecting residential real estate prices. *Arabian Journal of Geosciences*, 10(24), 535. <https://doi.org/10.1007/s12517-017-3331-3>
<https://doi.org/10.1007/s12517-017-3351-z>
- Yoo, S., Im, J., & Wagner, J. E. (2012). Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning*, 107(3), 293–306. <https://doi.org/10.1016/j.landurbplan.2012.06.009>
- Zaddach, S., & Alkhatib, H. (2014). Least squares collocation as an enhancement to multiple regression analysis in mass appraisal applications. *Journal of Property Tax Assessment & Administration*, 11(1), 47–66.
- Zhou, G., Ji, Y., Chen, X., & Zhang, F. (2018). Artificial neural networks and the mass appraisal of real estate. *International Journal of Online Engineering*, 14(3), 180–187. <https://doi.org/10.3991/ijoe.v14i03.8420>
- 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>