Estimating the Value of Commercial Real Estate Using Machine Learning Approaches

Ireneusz Czarnowski *[0000-0003-0867-3114]*
Faculty of Computer Science
Gdynia Maritime University
Gdynia, Poland *i.czarnowski@umg.edu.pl*

Abstract

The paper addresses the problem of predicting the attractiveness of the commercial real estate market based on market valuation. The aim is to estimate market value, and models of the commercial real estate market need to be established. The complexity of the problem lies in accounting for market changes resulting from the COVID-19 pandemic and, more broadly, its variability in response to geopolitical shifts, which are reflected in the available data. To address this, a machine learning-based framework is proposed. The proposed models for solving the problem have been validated using data from recent years. Finally, a summary is provided.

Keywords: commercial real estate market, machine learning, clustering, classification, regression.

1. Introduction

The market consists of a large range of different economic and financial phenomena. These phenomena are very much dependent on different internal and external factors. These factors have an influence on the market, the price of goods, business position and business attractiveness [6]. The real estate market is an example of a financial market and different economic fluctuations and global events can have an impact on it. The real estate market is very sensitive to shifts in global economic conditions. This is not helped by variable rates and inflation, and all of these factors can impact demand, investor confidence and the availability of affordable finance. A special example of this real estate market is the commercial segment.

Intelligence technologies can be used for identifying trends and investment opportunities in the real estate sector. Although there are broad applications, this work has been narrowed down to estimating the value of a given real estate, in terms of a long-term and secure investment. The intelligence technologies, thinking especially on artificial intelligence tools, including machine learning methods, need to be used when the historical data and efficient hardware are available [5].

A range of machine learning algorithms—such as Extra Trees, kNN, Random Forest, Lasso, Ridge, XGBoost, and ANN—have been employed to predict property prices, often compared to traditional models like the hedonic price model (see for example [1], [10]. Studies show that while linear models like ARIMA may perform well for short-term forecasts, machine learning and deep learning methods (including LSTM and ensemble models) offer more promising results for long-term estimations [2], [7, 8].

Machine learning models for commercial real estate have been discussed especially in [4], [11]. Additionally, these examples of research emphasise the importance of big data in commercial real estate valuation, discussing its benefits, challenges, and application processes.

Machine learning algorithms have been validated for predicting house values, particularly in the context of the COVID-19 pandemic. The study in [8] covers model selection, feature engineering, and optimisation using methods like Random Forest, Gradient Boosting, and LightGBM. Similarly, [4] examined the Lithuanian real estate market using linear regression and decision trees to assess pandemic-related impacts.

In this paper, an appropriate modelling approach based on machine learning algorithms

is proposed. Specifically, it involves clustering to identify market behaviour patterns, classification to assign current real estate to the identified behaviour models, and ultimately, estimation of market value. The paper demonstrates that machine learning algorithms can serve as effective tools for price prediction in the commercial real estate sector, particularly when market behaviour changes dynamically over time, as observed during and after the COVID-19 pandemic [3]. Furthermore, the paper addresses the critical role of data pre-processing, which can significantly affect the accuracy of the estimation process.

This paper is organised as follows. Section 2 is dedicated to a detailed presentation of the proposed framework. The results of computational experiments are presented in Section 3. Finally, concluding remarks are provided in Section 4.

2. Methodology and Proposed Framework

This chapter presents a methodology and framework for estimating the value of commercial real estate.

2.1. Data Collection and Properties

The models for estimating commercial real estate value were developed using data from the Polish market, encompassing 90 diverse features—both quantitative and qualitative—reflecting financial, tax-related, and economic aspects. These features aimed to capture a comprehensive picture of each property's attractiveness, including location, size, infrastructure proximity, architectural characteristics, and financial indicators such as local taxes, inflation, loan interest rates, and housing price indexes. Some factors could not be fully described due to data limitations, resulting in missing values in the dataset. Additionally, the data contained noise in the form of outliers. To address this, statistical techniques and hierarchical clustering (using Ward's method) were applied to detect outliers at the level of individual attributes and entire instances. It should be added that this evaluation was performed on instances formed from qualitative and quantitative data. This process ensured a more reliable dataset for training the prediction models, when finally the outliers have been eliminated. To ensure the appropriate characteristics of the data for the machine learning model training process, the quantitative values were standardised.

2.2. The Problem of Price Group Identification

The development of an effective method for estimating commercial real estate value, particularly in response to disruptions caused by the COVID-19 pandemic, is a priority of the research. Before the pandemic, standard regression models were used, but these became unreliable as the relationship between dependent and independent variables was no longer stable [9]. As a solution, the proposed approach introduces a two-stage framework. In the first stage, the real estate data is divided into price groups using a quartile-based method.

To increase the quality of consideration in the price groups, clustering was also applied, with the aim of better capturing the similarities between instances in the dataset and to highlight potential price classes, based on these similarities. The clustering is performed using the DBSCAN algorithm. DBSCAN was chosen due to its ability to automatically determine the number of clusters. Additionally, the algorithm supports clustering based on different subsets of attributes, allowing for flexible modelling of price patterns and market behaviour.

The resulting clusters define consistent and distinguishable price groups, which are then used in the second stage of the process: value estimation within those groups using regression models. This structured approach aims to restore the accuracy of price prediction despite dynamic changes in the market. In other words, the clustering results in independent characteristics for the price groups, which can be marked independently. In conclusion, this stage ends the data modelling, which is ready to be used at the classifier induction stage.

2.3. The Problem of Estimating the Commercial Real Estate

The dataset prepared as it has been described in previous section, can be used for the

training of a machine learning model. It should be added that the goal is to obtain a tool that will assign new data to an appropriate price group. As a result, the value of the property can be determined but this is already the task of machine leaning-based regression methods.

So, the proposed framework consists of two main processes: classifier learning to assign real estate to a price group, and regression modelling to estimate its value. XGBoost was chosen as the classification algorithm due to its popularity and performance. Once a new instance is assigned to a price group, its value can be estimated using a regression model specific to that group. This can involve linear or nonlinear models, or machine learning methods such as CART. The approach decomposes the estimation task into k regression sub-problems, where k is the number of identified price groups, with each model trained only on instances from its respective group.

The framework for estimating the value of commercial real estate is presented using Algorithms 1 and 2.

Algorithm 1 Classifier induction and regression model training
Input: D –dataset;
Begin
Standardise the quantitative values in D;
Let $D_1,..,D_4$ be the subset of D produced using the quartiles approach;
For i:=1 **to** 4 **do**
Run clustering on D_i;
Let C_i be the set of clusters resulting from clustering D_i;
For j:=1 **to** $|C_i|$ **do**
Let L_{ij} be a classifier induced on cluster $c_j \in C_i$;
Let R_{ij} be a regression model on cluster $c_j \in C_i$;
End For;
End For
Return L and R as sets of classifiers and regression models;
End

Algorithm 2 Estimation of the value of the commercial real estate
Input: d – new data instance;
L – set of classifiers;
R – set of regression models;
Begin
Let k denote the number of models, where |R| and |L| represent the number of regression models and classifiers, respectively;
Let a denote the price group predicted by $L_i(d)$, where $i=1,\ldots,k$;
Let v denote the value of the commercial real estate predicted by $R_a(d)$;
Return v;
End

3. Computational Experiments

Computational experiments were conducted to evaluate the effectiveness of the proposed approach in estimating commercial real estate value and its potential as an alternative to traditional econometric methods in which regression models are built. The study also examined whether the choice of classifier and regression models influences the quality of the results.

The experiments used real market data from 2010–2022, consisting of 3,800 instances described by 90 qualitative and quantitative features. The dataset was split into 80% for training and 20% for testing.

Table 1 presents values of the prediction quality assessment metric, i.e. the Mean Absolute Error (MAE) - as normalized values, relative to the actual values. Based on the results presented in Table 1, it can be observed that the proposed approach (using CART algorithm) improved the predictive performance across all other selected regression models (i.e. LASSO and RIDGE). The proposed framework, combined with regression models, improves prediction quality over standalone regression. Notably, pairing with CART reduced MAE from 0.22 to 0.18, confirming the framework's effectiveness, especially for nonlinear models in commercial real estate price prediction.

Table 2 presents the validation results for various classifiers (XGBoost, C4.5, kNN, Random Forest, SVM, Naive Bayes) based on price group prediction accuracy and F1-score. XGBoost achieved the best performance overall, followed closely by Random

Forest. SVM also performed well, while C4.5, Naive Bayes, and kNN showed moderate to lower effectiveness. XGBoost is recommended, with Random Forest as a strong alternative.

Table 1. Mean Absolute Error of Compared Regression Models

	Proposed approach + CART	Proposed approach + LASSO	Proposed approach + RIDGE	CART	LASSO	RIDGE
MAE	0.18	0.24	0.26	0.22	0.33	0.31

Table 2. Performance Comparison of Classification Models

	XGBoost	C4.5	kNN	Random Forest	SVM	Naive Bayes
Accuracy (%)	93.0	89.2	84.1	91.5	90.4	86.5
F-score	92.5	88.1	83.0	90.7	89.4	84.8

4. Conclusion

The paper presents a framework for estimating commercial real estate value, based on identifying price groups and applying suitable prediction models. The approach is validated and shows promising results. Future work will focus on statistical and multivariate analysis, using broader and more diverse datasets. The method proves competitive and offers stable predictions in dynamic market conditions, including those influenced by events like COVID-19.

References

1. Choy, L.H.T., Ho, W.K.O.: The Use of Machine Learning in Real Estate Research. Land 12, 740 (2023), doi:10.3390/land12040740
2. Çılgın, C., Gökçen, H.: Machine Learning Methods For Prediction Real Estate Sales Prices In Turkey, Revista de la Construcción. Journal of Construction 22, 163–177 (2023) doi:10.7764/rdlc.22.1.163
3. Di Liddo, F., Anelli, D., Morano, P., Tajani, F.: The Impacts of COVID-19 on Real Estate Market Dynamics: A Systematic Literature Review of Emerging Trends. Buildings 13, 2334 (2023), doi:10.3390/buildings13092334
4. Grybauskas, A., Pilinkienė, V., Stundžienė, A.: Predictive Analytics Using Big Data For The Real Estate Market During The COVID-19 Pandemic, J. Big Data 8(1) (2021), doi:10.1186/s40537-021-00476-0.
5. Isada, F.: The Impact Of Inter-Organisational Network Structures On Research Outcomes For Artificial Intelligence Technologies. Int. J. Econ. Sci. 11, 1–18 (2022)
6. Kutner, R., Ausloos, M., Grech, D., Di Matteo, T., Schinckus, C., Stanley, H.E.: Econophysics and Sociophysics: Their Milestones & Challenges. Physica A: Statistical Mechanics and its Applications 516, 240-253 (2019), doi:10.1016/j.physa.2018.10.019
7. Milunovich, G.: Forecasting Australia's Real House Price Index: A comparison of Time Series and Machine Learning Methods. Journal of Forecasting 39(7), 1098 1118 (2020)
8. 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, 2100 (2022)
9. Ouchen, A.: Econometric Modeling of the Impact of the COVID-19 Pandemic on the Volatility of the Financial Markets. Eng. Proc. *39*, 14 (2023), doi:10.3390/engproc2023039014
10. Wu, X, Yang, B.: Ensemble Learning Based Models for House Price Prediction, Case Study: Miami, U.S, In: Proceedings of the *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, pp. 449-458, Wuhan, China (2022), doi:10.1109/AEMCSE55572.2022.00095
11. YongLin, X.: Big Data For Comprehensive Analysis Of Real Estate Market. Electronic Theses, Projects, and Dissertations 1596, (2022)