A machine learning method for real estate operation projects forecasting*

Sergey Bushuyev[1,*,†], Denis Bushuiev[1,†], Nikolay Poletaev[2,†], Mykola Malaksiano[2,†], and Dmitriy Kravtsov[2,†]

[1] *Kyiv National University of Construction and Architecture, 31, Povitroflotskyi avenue, Kyiv, Ukraine*
[2] *Odessa National Maritime University, 34, Mechnikov street, Odessa, Ukraine*

Abstract
The study examined the impact of heterogeneous urban night lighting on the possibilities of improving the forecast accuracy for real estate operations projects in a large city. By transforming high-resolution night satellite images into light clusters, new dataset features were obtained after calculating the centers of light clusters and the distances between them. The study used a dataset on real estate rentals in Houston, Texas, USA. The light clusters were linked to the terrain based on their geometric coordinates obtained using QGIS. These new features were integrated into a machine learning model based on the LightGBM regressor. Calculations showed that the reduction in forecast error (in the mean squared error metric, MSE) for our dataset was 11.8%, significantly exceeding the influence of other features investigated in the study. The results suggest that the "light" feature can be considered highly promising for real estate operations projects.

Keywords
machine learning, artificial intelligence, computer vision, neural networks, real estate, project forecasting, project modeling, satellite photos, geolocation, light clusters

1. Introduction

The real estate market is influenced by a variety of factors, the assessment of which is a complex task. On one hand, the state of the real estate market reflects the overall condition of economic, social, demographic, logistical, environmental, and other factors. On the other hand, effective prediction of real estate market trends enables timely resolution of various economic and social tasks. Investment companies and individuals, who consider real estate investment as one of the potential investment options, can significantly influence these market trends. Therefore, researching real estate market trends is of significant practical and theoretical interest. Promising research directions in this context include the

Proceedings of the 5th International Workshop IT Project Management (ITPM 2024), May 22, 2024, Bratislava, Slovak Republic
* Corresponding author.
† These authors contributed equally.
sbushuyev@ukr.net (S. Bushuyev); bushuyevd@gmail.com (D. Bushuiev); poletaev@ukr.net (N. Poletaev); malax@ukr.net (M. Malaksiano); dmkravtsov@gmail.com (D. Kravtsov)
0000-0002-7815-8129 (S. Bushuyev); 0009-0002-6477-0517 (D. Bushuiev); 0000-0002-1340-582X (N. Poletaev); 0000-0002-4075-5112 (M. Malaksiano); 0009-0005-0305-544X (D. Kravtsov)

development and use of artificial intelligence methods, big data and geospatial data analysis, as well as the improvement of project management methods capable of flexibly and efficiently handling large volumes of data and meeting the challenges and dynamics characteristic of the modern real estate market. In this regard, the development of modern project-oriented approaches to company management, as presented in various works, is of considerable interest [1, 2, 3]. The models for assessing the cognitive readiness of managerial teams in the implementation of infrastructure programs are studied in [1]. Entropy paradigm of project-oriented organizations management and portfolio structure dynamics of the organization development, taking into account information entropy, were studied in [2, 3]. Articles [4, 5] deal with SMART intelligence models in application to innovative projects management and natural-scientific methodologies, allowing for the expansion of the traditional view of the project. The Synergetic Effects and fuzzy method approach to project Management of Active Systems were discussed in [6, 7]. Papers [8, 9] propose some mathematical and simulation modeling approaches that can be effectively adopted to estimate projects which involve manipulations with assets that are prone to aging and depreciation. Also, in the context of real estate project management, certain attention should be paid to ecological and logistical aspects. Some promising approaches in this direction were proposed in [10, 11, 12].

Currently, there are many methods and their combinations for assessing real estate operations projects values. Classic approaches traditionally used for real estate operations project valuation included comparative assessment methods, income approach, cost approach, as well as combinations of these methods [13]. Machine learning models, such as decision trees, random forest [14, 15], more complex ensemble methods (XGBoost, LightGBM) [16], as well as neural networks [17, 18], are increasingly being used for analyzing and forecasting real estate operations project prices. These models, trained on property sales data, allow for considering a larger number of features and making more accurate forecasts compared to classic approaches. Textual description of a property can also contains valuable information for assessing its value [19]. Natural Language Processing (NLP) methods in combination with neural network models allow for extracting additional features from textual descriptions, considering factors such as property characteristics, its surroundings, and other features [20].

Geospatial data also play an important role in property operations projects valuation. The distance to various objects such as metro stations, schools or shopping centers can significantly affect the value of a property [21, 22]. Integrating this data into machine learning models can improve their performance and estimation accuracy [23].

The use of photographs of building facades and satellite images of urban infrastructure and the real estate objects themselves can be carried out as part of a modern method of real estate projects valuation. This approach improves the accuracy of valuations by vectorizing visual data, allowing for the extraction of features that reflect the visual characteristics of areas. It can be used as an independent tool or in combination with other analysis methods [24, 25].

In this study, the authors hypothesized that the intensity of urban lighting in an area can indicate the business and other activities of its inhabitants. Indeed, it can be assumed that the distribution of light flows of urban infrastructure, captured in satellite images of the area at night, is connected with its vital pulse - the level of foot traffic, visitation, and activity in these places. Indirectly or directly, such activity is reflected in the level of rental rates for real estate.

The goal of this study was to investigate the impact of the "light" feature on the accuracy of real estate operations projects price forecasting in large cities, where light clusters have a complex structure and can be highly informative.

2. Machine Learning Model Considering the Light Feature

2.1. Method for Obtaining the Light Feature from Satellite Photography

In the context of researching the impact of illumination on rental property prices, data on housing rentals were enriched with new features based on the degree of area lighting. During the preliminary processing of images, geographically tagged using QGIS, the places of greatest illumination were grouped into light clusters. The centers of these clusters were then identified, and their coordinates were used to create new features in the mathematical model.

We used a dataset on real estate rentals in the city of Houston, Texas, USA. The data was collected from Redfin [26], a real estate website that handles residential real estate brokerage operations. This dataset consists of 9,260 rental properties and includes 9 features including the target "Price", such as geographic coordinates, year of construction, number of bedrooms and bathrooms, total area, size of the adjacent plot, and the rental price of the property in U.S. dollars per month (Table 1). In our calculations, we used the logarithm of the target variable "Price", which helped to smooth the distribution of the target variable, making it more normal and reducing the impact of extreme values.

Table 1
Descriptive statistics of dataset features

Index	count	mean	min	25%	50%	75%	max	std
Latitude	9260	29.7702	29.5326	29.7229	29.7505	29.8034	30.1351	0.09
Longitude	9260	-95.444	-95.801	-95.527	- 95.428	-95.3804	-95.0742	0.11
Year Built	9260	1952.00	1893.0	1913.0	1940.0	1993.0	2023.0	989.78
Beds	9260	2.0	0.0	1.0	2.0	3.0	8.0	01.08
Baths	9260	1.8	0.0	1.0	2.0	2.0	8.0	0.90
BuildingSize	9260	1268.7	101.0	748.0	1088.0	1608.0	5000.0	683.25
lotSize	9260	18726.1	0.0	0.0	0.0	5502.0	1175401.0	74470.5
PostalCode	9260	77050.3	77002.0	77018.0	77044.0	77077.0	77598.0	54.16
Price	9260	1848.2	619.0	1269.0	1698.0	2200.0	12500.0	876.50

This also helped to decrease heteroscedasticity (variance inhomogeneity) in the data, leading to more stable and interpretable model estimates (Figure 1a, 1b). Such an approach usually leads to improved performance of regression models, especially in the presence of outliers or a non-normal distribution of the target variable [27].

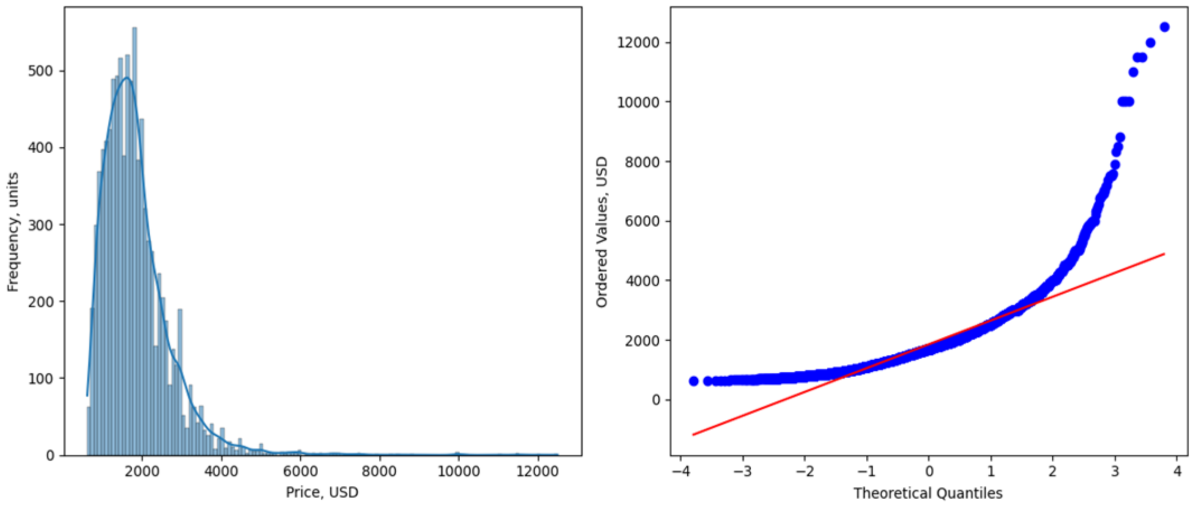

Figure 1a: In the original coordinates.

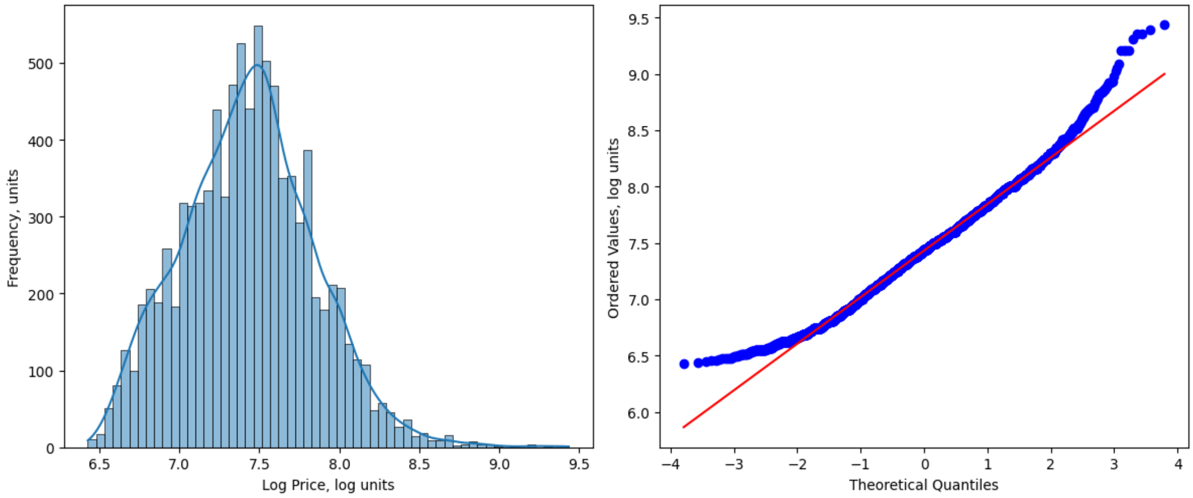

Figure 1b: In logarithmic coordinates.

Outliers and anomalous values were removed from the dataset to ensure the credibility of the data and avoid distortions in the model. We also paid attention to the range of values for each parameter to ensure that the data remained realistic and interpretable.

A high-resolution satellite image taken at night was obtained from open sources [28]. As a result of preprocessing, the resulting black and white image was displayed for visual assessment (Figure 2).

To determine the centers of light clusters in the image of Houston city, a two-stage method was used. In the first stage, we optimally adjusted the brightness and contrast of the image. This step allowed us to highlight the local centers of light areas in the image. For this purpose, we loaded the image using the tools of the OpenCV library [29]. The image was then converted to grayscale to simplify the analysis. After that, we conducted brightness and contrast correction, which helped us to better distinguish the bright areas against the night cityscape. For further analysis, threshold binarization was applied to highlight the bright spots in the image. We then found the contours of these bright areas and determined their centers.

Figure 2: Satellite image of the city of Houston (Black and White). [28] (https://earthobservatory.nasa.gov).

Visualization allowed us to clearly represent the detected bright spots and their centers in the image (Figure 3).

At the same stage, a transformation of the pixel coordinates of the cluster centers, obtained in the previous step, into geographical coordinates was carried out. For this, we used information about the geographical coordinates of the corners of the image (top left and bottom right corners), as well as the size of the image. The latitude and longitude of each cluster center were calculated based on its pixel coordinates and the proportionally distributed geographical coordinates of the image corners. The obtained geographical coordinates of the cluster centers provide additional information about the location of the highlighted areas in the image and can be used for further analysis of spatial data.

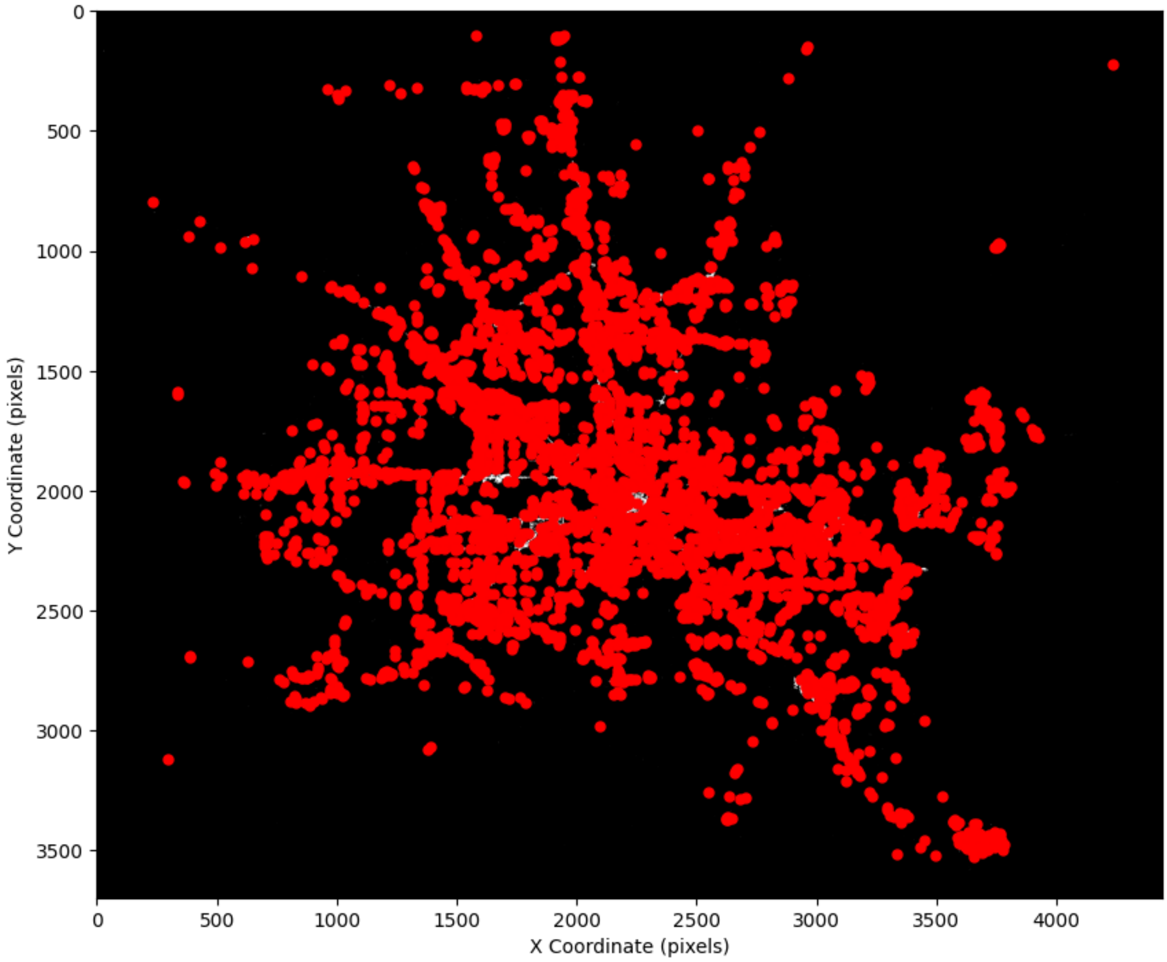

Figure 3: The image with the centers of bright spots (4815 centers).

The second stage involved the use of the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm to identify local centers of light spots. With its help, we combined close centers into clusters and identified the centers of these clusters as global centers of light areas.

These cluster centers are displayed on a scatter plot, where each center is marked in red, and the original points are in blue (Figure 4). This visual representation allows for a better understanding of the data structure and the identification of spatial patterns or groupings in geographical data.

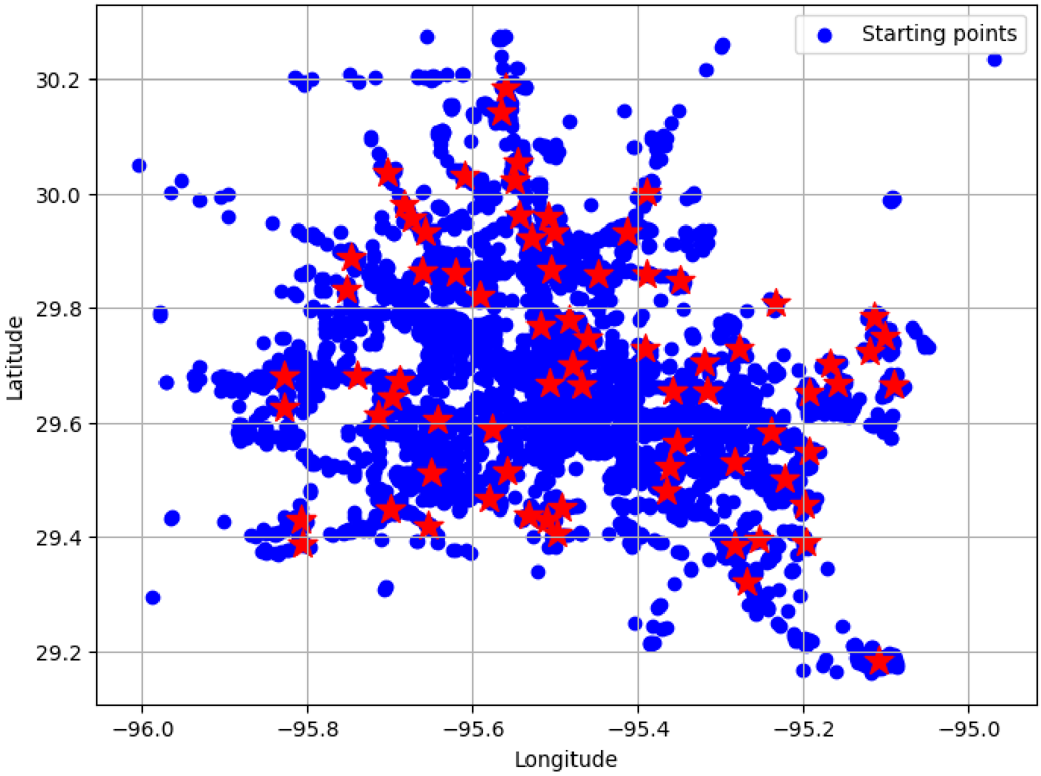

Figure 4: Clustering with DBSCAN.

This two-stage approach allowed us to more efficiently and accurately identify the centres of light spots in the image, ensuring high throughput and accuracy of analysis.

2.2. Algorithm for predicting real estate prices

After processing the satellite images, 16 new features based on light centers in the city of Houston were added to the original dataset (Table 1). For this, we used functions that determine the number of light centers within a certain radius (0.1, 0.3, 0.5, 1, 2, 4, 6, 8, 10 km) from each real estate object, the distance to the nearest light center, and the average distance from the real estate object to several (2, 4, 6, 8) nearest centers.

The resulting dataset was processed using the Recursive Feature Elimination with Cross-Validation (RFECV) algorithm, which automatically selects the most important features for the model based on their impact on the target variable. This method allows iteratively removing redundant features with the least impact, assessing the quality of the model at each iteration using cross-validation.

Thus, the proposed model for assessing the values of real estate operations projects, based on light streams from night cities captured in high-resolution satellite photographs, is depicted in the following scheme (Fig. 5).

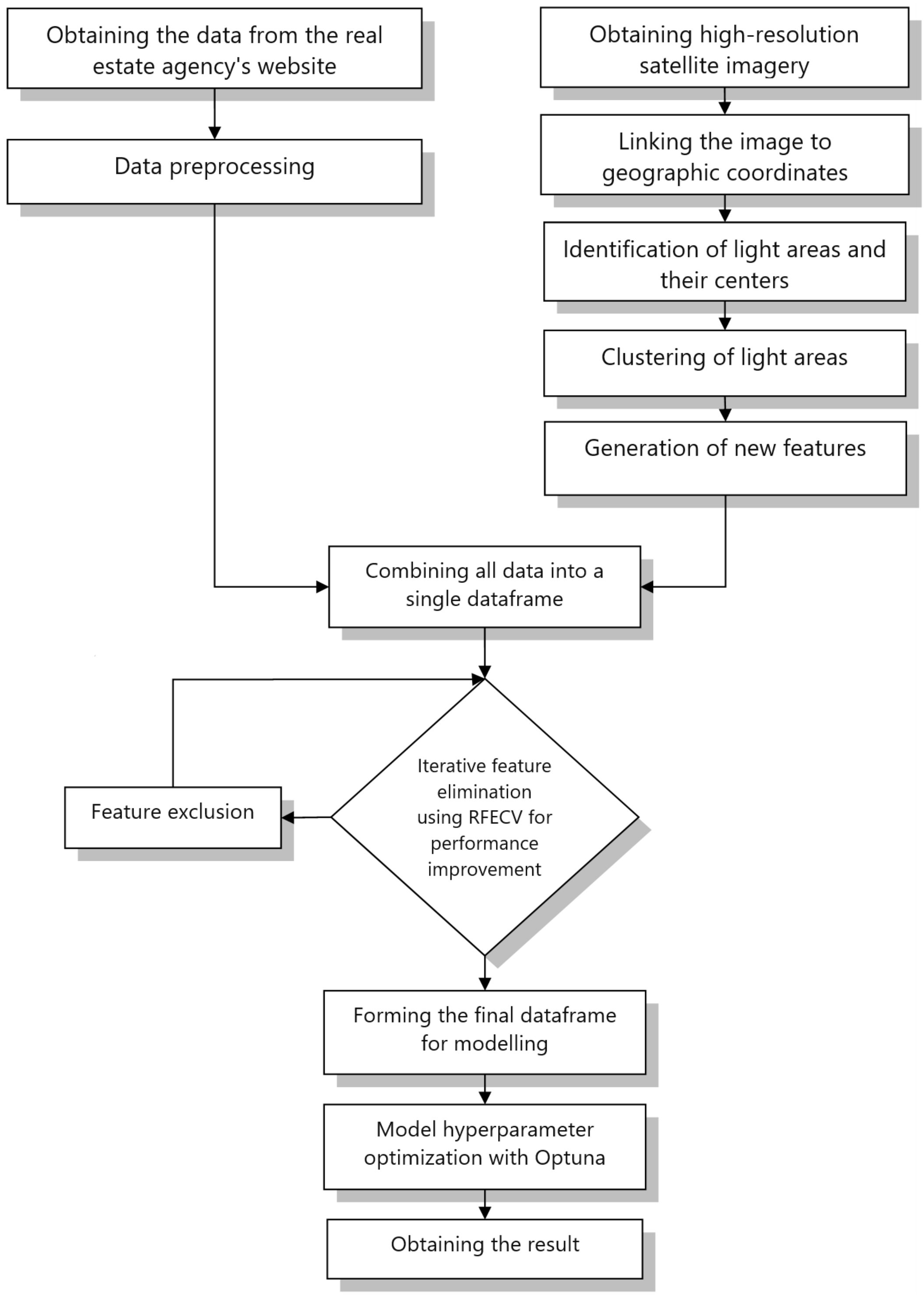

Figure 5: Scheme of the model for assessing real estate operations projects values based on light streams from night cities on high-resolution satellite photographs.

For training the model on each iteration, RFECV uses the DecisionTreeRegressor model. This decision was made due to its ability to efficiently work with large-scale data and high dimensionality, which allows for fast and accurate feature selection. As a result of the selection of optimal features using the RFECV algorithm, we reduced the feature space from 24 to 16 (Table 2).

Table 2
Optimal features for a predictive model

Feature	Description
Latitude	Latitude of the geographical location of the object
Longitude	Longitude of the geographical location of the object
Year Built	Year of construction
Beds	Number of bedrooms in the property
Baths	Number of bathrooms in the property
buildingSize	Object area
lotSize	Area of the plot on which the building is located
PostalCode	Postcode
clusters_within_4km	Number of clusters within a 4 km radius from the object
clusters_within_8km	Number of clusters within a radius of 8 km from the object
clusters_within_10km	Number of clusters within a 10 km radius from the object
distance_to_nearest_clusters	Distance to nearest cluster
avg_dist_nearest_clusters_2	Average distance to two nearest clusters
avg_dist_nearest_clusters_4	Average distance to the four closest clusters
avg_dist_nearest_clusters_6	Average distance to the six closest clusters
avg_dist_nearest_clusters_8	Average distance to eight nearest clusters

As the base model for predicting real estate rental prices, we used the LightGBM gradient boosting algorithm. This choice of algorithm is explained by its high performance, efficiency, and ability to quickly process large volumes of data [30].

The optimization of model parameters was conducted using the Optuna library. The optimized model parameters included tree depth (max_depth), learning rate (learning_rate), number of tree leaves (num_leaves), maximum number of bins (max_bin), number of trees (n_estimators), and the fraction of random features for each split (colsample_bytree).

Optuna is a library for optimizing model hyperparameters, which automates the process of their selection. It uses optimization algorithms such as the Tree-structured Parzen Estimator (TPE) and genetic optimization algorithms for efficient searching of optimal parameter values. These are based on a Bayesian approach and effectively find optimal hyperparameters, even in large search spaces. This is a significant advantage over optimizers like RandomizedSearchCV and GridSearchCV, which explore the hyperparameter space by enumeration or random selection. In contrast, Optuna aims for a more intelligent approach, selecting the next points of optimization based on the results of previous steps. The advantages of Optuna include ease of use, flexibility in configuration, and the possibility of parallel computation, which speeds up the optimization process. It is a powerful tool for optimizing model parameters and enhancing their accuracy and efficiency [31].

3. Results and discussion

For each parameter combination, we used 5-fold cross-validation to evaluate model performance. We calculated the mean squared error (MSE), root mean squared error (RMSE), coefficient of determination (R^2), and median absolute percentage error (MDAPE) for each iteration.

After conducting 200 optimization iterations, we selected the best combination of parameters for both variations of the datasets, which minimized the MSE. The results of the optimization allowed us to determine the optimal values of the model parameters for our data, ensuring the best performance and accuracy of the forecasts.

After selecting the optimal hyperparameters for the LightGBM model, we obtained the following results on cross-validation (Table 3).

Table 3
Model efficiency by the chosen metric

Parameter	Until it is enriched with new functions	After enrichment with new features	Improvement percentage
MSE	139998	123431	11.85
RMSE	374	351	6.15
R^2	0.817752	0.839	2.61
MDAPE	7.830909	7.386	5.67

From the table, it is evident that the new data enrichment method led to improved model performance across all evaluation metrics. The differences in model efficiency depending on the chosen metric are explained by their specific characteristics. MSE penalizes larger errors more than smaller ones since errors are squared. A significant improvement in MSE may indicate a reduction in these larger errors.

RMSE is similar to MSE but is scaled back to the original units of measurement, making it more interpretable. The smaller improvement compared to MSE might be due to RMSE being less sensitive to very large errors. R^2 measures what proportion of variability in the dependent variable is explained by the model. An increase in R^2 by 2.61% suggests that the model has become better at explaining the data but not necessarily at reducing every error.

MDAPE focuses on the median of the errors, making it robust to outliers. A decrease in MDAPE indicates an overall improvement in the model's accuracy but, again, is less sensitive to extreme values compared to MSE and RMSE.

Different metrics assess different aspects of model performance, so improvements may vary depending on how exactly the new data enrichment method impacts these aspects.

Thus, the method we proposed for incorporating the light feature significantly improved the accuracy of the real estate project price prediction model, confirming its effectiveness and practical applicability.

The technique for assessing the level of urban lighting is a complement to existing approaches in real estate operations project value analysis described in the literature. Differing from methods focusing on the physical characteristics of properties or urban infrastructure, our approach adds an analysis of area lighting. This allows for a deeper understanding of the factors influencing the attractiveness of areas and, consequently, real estate operations project value, enriching traditional valuation with new dimensions.

4. Conclusion

This study presents an innovative approach to assessing the real estate operations project value in large cities. The essence of the approach lies in integrating new features into the regression model for forecasting, obtained from high-resolution satellite photographs by analyzing light streams from night cities.

The study demonstrates that applying this approach to a dataset consisting of 9260 rental real estate objects in Houston (Texas, USA) has reduced the mean squared error (MSE) of the baseline regression model by 11.8%. This gives the authors reason to believe that the proposed approach can become an effective tool for analyzing and assessing the real estate market.

The comparative simplicity and accessibility of using satellite images suggest that their analysis can be adapted to assess other environmental factors, such as proximity to water resources, greenery, parks and forests, as well as industrial zones, to obtain effective features for regression models. It should be noted that one of the limitations of using satellite images to create new features for regression models is the limited availability of quality satellite night images for all cities. For the successful implementation of the method, high-resolution images and their linkage to geographical coordinates are necessary.

References

[1] S. Bushuyev, N. Bushuyeva, D. Bushuiev, V. Bushuieva, Cognitive readiness of managing infrastructure projects driven by SMARTification, 2022 IEEE European Technology and Engineering Management Summit, E-TEMS 2022 - Conference Proceedings, (2022), pp. 196–201. doi:10.1109/E-TEMS53558.2022.9944458.

[2] A. Bondar, S. Bushuyev, S. Onyshchenko, H. Tanaka, Entropy paradigm of project-oriented organizations management. CEUR Workshop Proceedings. 2565, (2020), pp. 233–243.

[3] S. Bushuyev, S. Onyshchenko, N. Bushuyeva, A. Bondar, Modelling projects portfolio structure dynamics of the organization development with a resistance of information entropy, International Scientific and Technical Conference on Computer Sciences and Information Technologies, IEEE, (2021), pp. 293–298. doi:10.1109/CSIT52700.2021.9648713.

[4] S. Rudenko, T. Kovtun, T. Smokova, I. Finohenova, The genetic approach application and creation of the project genetic model. International Scientific and Technical Conference on Computer Sciences and Information Technologies, (2022), pp. 434–437. doi:10.1109/CSIT56902.2022.10000822.

[5] S. Bushuyev, N. Bushuyeva, V. Bushuieva, D. Bushuiev. SMART intelligence models for managing innovation projects. CEUR Workshop Proceedings, (2023) pp. 1463–1474.

[6] S. Chernov, L. Chernova, N. Kunanets, V. Piterska, The Synergetic Effect in the Management of Active System with Distributed Control. 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, 2023, pp. 1–4. doi:10.1109/CSIT61576.2023.10324123.

[7] O. Melnyk, O. Onishchenko, S. Onyshchenko, O. Shumylo, S. Volyanskyy, A. Bondar, N. Cheredarchuk, Application of fuzzy controllers in automatic ship motion control systems. International Journal of Electrical and Computer Engineering, 13 (4), (2023), pp. 3948–3957. doi:10.11591/ijece.v13i4.

[8] I. Lapkina, M. Malaksiano, Elaboration of the equipment replacement terms taking into account wear and tear and obsolescence. Eastern-European Journal of Enterprise Technologies, 3(3-93), (2018), pp. 30–39. doi:10.15587/1729-4061.2018.133690.

[9] O. Melnyk, S. Onyshchenko, O. Onishchenko, O. Shcherbina, N. Vasalatii, Simulation-Based Method for Predicting Changes in the Ship's Seaworthy Condition Under Impact of Various Factors. Studies in Systems, Decision and Control, 481, (2023), pp. 653–664. doi:10.1007/978-3-031-35088-7_37.

[10] V.V. Romanuke, A.Y. Romanov, M.O. Malaksiano, Crossover Operators in a Genetic Algorithm for Maritime Cargo Delivery Optimization. Journal of Eta Maritime Science, 10(4), (2022), pp. 223–236. doi:10.4274/jems.2022.80958.

[11] S. Rudenko, T. Kovtun, V. Smrkovska, Devising a method for managing the configuration of products within an eco-logistics system project. Eastern-European Journal of Enterprise Technologies, 4(3-118), (2022), pp. 34–42. doi:10.15587/1729-4061.2022.261956.

[12] V.V. Romanuke, A.Y. Romanov, M.O. Malaksiano, Pseudorandom number generator influence on the genetic algorithm performance to minimize maritime cargo delivery route length. Pomorstvo, 36(2), (2022), pp. 249–262. doi:10.31217/p.36.2.9.

[13] M. Stang, B. Krämer, C. Nagl, W. Schäfers, From human business to machine learning - methods for automating real estate appraisals and their practical implications. Z Immobilienökonomie, 9, (2023), pp. 81–108. doi:10.1365/s41056-022-00063-1.

[14] A. Mahale, V. Bhistannavar, N. Chauhan, V. Matey, A. Shitole, Housing Price Prediction Using Supervised Learning, in: Proceedings of the 3rd International Conference on Advances in Engineering, Technology & Business Management (ICAETBM-2022), Pune, 2022. doi:10.1109/PuneCon55413.2022.10014818.

[15] M. Heidari, S. Zad, S. Rafatirad, Ensemble of Supervised and Unsupervised Learning Models to Predict a Profitable Business Decision, in: Proceedings of the 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), (2021). doi:10.1109/IEMTRONICS52119.2021.9422649.

[16] S. A. Rahman, N. H. Zulkifley, I. Ibrahim, S. Mutalib, Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur, International

Journal of Advanced Computer Science and Applications 12(12), (2021). doi:10.14569/IJACSA.2021.0121291.
[17] S. Peterson, A. Flanagan, Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal, Journal of Real Estate Research, 31(2), (2009), pp. 147–164. doi:10.1080/10835547.2009.12091245.
[18] F. Mostofi, V. Toğan, H. B. Başağa, Real-estate price prediction with deep neural network and principal component analysis. OTMCJ, 14(1), (2022), pp. 2741–2759. doi:10.2478/otmcj-2022-0016.
[19] H. Ahmed, M. Moustafa, House Price Estimation from Visual and Textual Features. In Proceedings of the 8th International Joint Conference on Computational Intelligence (IJCCI 2016) - NCTA; ISBN 978-989-758-201-1, SciTePress, (2016), pp. 62–68. doi: 10.5220/0006040700620068.
[20] M. Heidari, S. Rafatirad, Semantic Convolutional Neural Network model for Safe Business Investment by Using BERT, in: Proceedings of the 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), (2020). doi:10.1109/SNAMS52053.2020.9336575.
[21] P.-Y. Wang, C.-T. Chen, J.-W. Su, T.-Y. Wang, S.-H. Huang, Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism, IEEE, (2021), pp. 55244–55259. doi:10.1109/ACCESS.2021.3071306.
[22] S. Orford, Valuing Locational Externalities: A GIS and Multilevel Modelling Approach, Environment and Planning B Planning and Design, 29(1), 2002, pp. 105–127. doi:10.1068/b2780.
[23] S. S. S. Das, M. E. Ali, Y.-F. Li, Y. B. Kang, T. Sellis, Boosting house price predictions using geo-spatial network embedding, Data Mining and Knowledge Discovery, 35, (2021), pp. 2221-2250. doi:10.1007/s10618-021-00789-x.
[24] S. Law, B. Paige, C. Russell. Take a Look Around: Using Street View and Satellite Images to Estimate House Prices, ACM Transactions on Intelligent Systems and Technology, 10(5), (2019), pp. 1–19. doi:10.1145/3342240.
[25] S. Law, Y. Shen, C. I. Seresinhe, An application of convolutional neural network in street image classification: The case study of London, in: Proceedings of the 1st Workshop, (2017). doi:10.1145/3149808.3149810.
[26] Redfin. URL: https://www.redfin.com.
[27] M. Yin, Model For Predicting London House Prices, Highlights in Business Economics and Management, 5, (2023), pp. 124–133. doi:10.54097/hbem.v5i.5039.
[28] NASA Earth Observatory. URL: https://earthobservatory.nasa.gov.
[29] OpenCV. URL: https://opencv.org.
[30] L. M. John, R. Shinde, S. Shaikh, D. Ashar, Predicting House Prices using Machine Learning and LightGBM, in: Proceedings of the 7th International Conference on Innovations and Research in Technology and Engineering (ICIRTE-2022), organized by VPPCOE & VA, Mumbai-22, INDIA, (2022). doi:10.2139/ssrn.4108744.
[31] M. Mao, A Comparative Study of Random Forest Regression for Predicting House Prices Using, in: Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023), Advances in Intelligent Systems Research, Atlantis Press, 2024. doi:10.2991/978-94-6463-370-2_63.