House Market Prediction Using Machine Learning

Nicușor-Andrei ANDREI
Bucharest University of Economic Studies
Faculty of Cybernetics, Statistics and Economic Informatics
Bucharest, Romania
andreinicusor20@stud.ase.ro

This study explores and compares the performance of tree-based machine learning algorithms' predictions for Bucharest real estate market prices. The dataset was obtained from a local platform in March 2025 and contains residential apartments for sale in Bucharest. The comprehensive data preprocessing step, including imputation of missing values, encoding of categorical variables, and the engineering of new key features such as distance to public transport, played a key role in the models' performance. The models were optimized using a grid search algorithm with 5-fold cross-validation and evaluated with key performance indicators including root mean squared error, mean absolute error, and coefficient of determination. The results indicate that XGBoost outperforms both Random Forest and a single Decision Tree, reducing all the key performance indicators used in analysis.
Keywords: *House market, Machine learning, Price prediction, Tree-based machine learning, XGBoost*

1 Introduction

The housing real estate market in Bucharest underwent a drastic transformation over time due to rapid urbanization, demographic changes, and shifting economic conditions. Homebuyers and investors require an effective price forecast based on newer techniques to provide a robust house market analysis. The performance of tree-based machine learning techniques, decision tree, random forest, and XGBoost, on this particular economic segment of Bucharest is presented in this paper. One Decision Tree provides simple interpretability but will probably overfit; Random Forests increase stability by averaging a large number of trees; and XGBoost continues to improve predictions using gradient boosting that adjusts residual errors iteratively. Despite the optimistic results presented by these approaches in different global markets, their comparative performance in the specific context of Bucharest is yet to be explored. The current research fills this gap by benchmarking the performance of Decision Tree, Random Forest, and XGBoost regressors against an elaborate dataset of real estate transactions in Bucharest, collected in March 2025 from a regional online marketplace. After an extensive preprocessing step including imputation of the missing values, outliers handling, and feature engineering, the algorithms are fine-tuned using the grid search method. The grid search is carried out using a 5-fold cross-validation mechanism with the best model determined as the minimizer of the mean squared error metric. The optimised versions of the three algorithms are next benchmarked against the key indicators relevant to regression analysis: root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and the coefficient of determination (R2). Apart from precision, the approach also considers the drivers of the prices of properties, with model interpretability highlighted using the feature importance score.

The structure of the paper is outlined as follows: Section 2 gives an overview of the literature relevant to the topic under investigation; Section 3 gives a comprehensive description of the

methodology used during this research; Section 4 presents the findings together with a related discussion; and Section 5 recapitulates the major conclusions of this research and makes recommendations for future research. This framework ensures an empirically supported comparison of tree-based models in the context of Bucharest's vibrant real estate market.

2 Literature Review

The rapid evolution of machine learning has opened new opportunities for modeling complex systems, including the real estate market, where price dynamics are difficult to capture. In the context of Bucharest's urban housing landscape, machine-learning-based price prediction promises to help buyers, investors, and policymakers navigate a market shaped by diverse offers and fluctuating demand. In the next paragraphs, studies related to this field are synthesized to highlight methodologies, key findings, and their implications for a Bucharest-focused investigation.

2.1 Hedonic Foundations

Before the beginning of modern machine learning, hedonic-price models provided a framework for decomposing sale prices into intrinsic and external attributes [1]. Early applications in Beijing and Paris confirmed the influence of building quality and neighborhood heterogeneity on prices [2, 3], and more recent work has extended hedonic indices to capture the effects of transit accessibility [4]. Although most hedonic studies focus on explanatory power rather than forecasting, they offer valuable guidance on feature selection and baseline model structure, insights that remain relevant when constructing machine-learning pipelines for Bucharest.

2.2 Ensemble and Tree-Based Methods

A recurring theme across multiple markets is the strong performance of tree-based ensemble models. For example, the authors of [5, 6, 7] show that Random Forest Regression achieves high accuracy when forecasting house prices in their case studies. Similarly, in Hong Kong, a Random Forest model outperformed both Support Vector Machines and Gradient Boosting Machines on a dataset of 40,000 transactions spanning 18 years [8]. Extreme Gradient Boosting algorithm (XGBoost) also stands out: a Vilnius study that scraped nearly 19,000 listings found XGBoost to be the most predictive among fifteen tested models [9]. In Bengaluru, XGBoost again outperformed Linear, Ridge, Lasso and SVM baselines [10], and hybrid approaches (for example stacking with CatBoost) further reduced error on 1.9 million transactions in Hong Kong [11]. Taken together, these ensemble successes suggest that machine-learning models for Bucharest will likely benefit from tree-based learners, provided that sufficient feature engineering and hyperparameter tuning are applied.

2.3 Linear and Kernel Methods

Although less dominant than ensemble models, linear and kernel-based regressors remain valuable for their interpretability and computational efficiency. In the Bengaluru comparison, Ridge and Lasso regressions, as well as a support-vector machine, delivered reasonable accuracy but were ultimately outperformed by XGBoost [10]. In Melbourne, an SVM achieved the lowest mean squared error at the expense of the longest training time, underscoring a trade-off between predictive power and computational cost [12]. A Multiple linear regression with the right data split ratio is validated as being more efficient than simple linear regression for house market predictions [13]. In Beijing, after a rigorous data preprocessing step, which included outlier elimination, feature engineering, and one-hot encoding algorithm, Hybrid and Stacked Generalization Regression delivered promising results on the training set and test set [14].

For Bucharest, where data volumes may vary significantly across districts, it may be

worthwhile to benchmark a well-tuned SVM alongside faster linear models as baseline approaches.

2.4 Classification Formulations

Some researchers have reframed house-price forecasting as a binary classification task, predicting whether a property's selling price will exceed its listing price. In Fairfax County, Virginia, the RIPPER rule-learning algorithm yielded fewer classification errors than C4.5, Bayesian classifiers, and AdaBoost [15]. A related study applied SVM, Random Forest, and a neural network, finding that Random Forest delivered the highest accuracy, precision, sensitivity, and specificity [16]. While classification methods do not provide information on the magnitude of price changes, they may prove useful in Bucharest for stakeholders interested primarily in upside versus downside risk.

2.5 Deep and Hybrid Architectures

Deep-learning models, especially when enriched with a good data engineering step, have great potential but also require large datasets and careful regularization to avoid overfitting. One study combined street-view imagery and socio-economic indicators with a gradient boosting machine to forecast neighborhood-scale appreciation rates with high accuracy [17]. Comparative work found that both deep networks and classical regression models tended to overestimate control-sample prices, indicating a need for larger datasets and ensemble hybrids [18]. The most advanced approaches leverage Transformer architectures optimized via Bayesian hyperparameter search, achieving substantial RMSE reductions on large Hong Kong datasets [11]. For Bucharest, integrating aerial or street-view data, if such data are available, could yield similar improvements, although attention must be paid to model complexity and generalization.

2.6 Research Gaps and Outlook

Despite the extensive literature, few studies have applied these advanced machine-learning techniques explicitly to the Bucharest housing market. Key knowledge gaps include the integration of geographic data with house attributes and the adaptation of machine learning models and deep-learning networks to conform to the data structure of Romania. By overcoming these limitations and integrating best practices across the different stages of preprocessing of the data, model development, and model evaluation, future studies can provide accurate and interpretable price forecasts customized to the specific urban context of Bucharest.

3 Methodology

The following chapter, summarized in Figure 1, presents the methodology steps in depth, starting with data acquisition and preprocessing, continuing with the architectures of the proposed models, and closing with the metrics used for comparison.

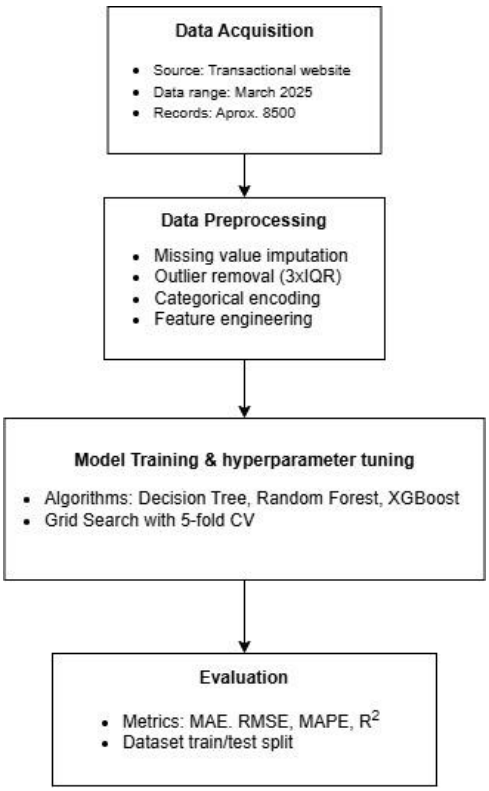

Fig. 1 Methodology applied

3.1 Data Acquisition and Preprocessing

Transactional data for the Romanian residential real estate market was obtained by web scraping a local listings website in March 2025. Initial preprocessing involved the removal of variables presenting over 5% missing values. Discrete features with missing entries were imputed according to domain-specific rules: absent bathroom counts were assigned a value of 1, and missing balcony counts were set to 0. The year built field, with a low percentage of missing values, was median-imputed to maintain discreteness, while missing "built area" values were replaced by the median built area stratified by room count.

Feature engineering produced several binary and continuous variables to enhance model performance. The categorical attributes housing_type, comfort, and compartmentalization_type were encoded as the binary indicators is_penthouse, is_luxury, and is_detached. Vertical position within the building was captured by the flags is_first_floor and is_last_floor, together with floor_ratio, defined as the apartment's floor number divided by the total building floors.

Spatial features were derived from latitude and longitude coordinates by computing distances (in meters) to the city center and the nearest metro station using the Haversine formula [19]:

$$\varphi_i^* = \varphi_i \cdot \frac{\pi}{180},\ \lambda_i^* = \lambda_i \cdot \frac{\pi}{180},\ i = 1,2 \quad (1)$$

$$\Delta\varphi = \varphi_2^* - \varphi_1^*, \qquad \Delta\lambda = \lambda_2^* - \lambda_1^* \quad (2)$$

$$a = \sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos(\varphi_1^*)\cos(\varphi_2^*)\sin^2\left(\frac{\Delta\lambda}{2}\right) \quad (3)$$

$$d = 2R\ \arcsin\left(\sqrt{a}\right), \quad (4)$$

where φ represents latitude, λ represent longitude and R represents the radius of the Earth, which is 6371 km. [20]

Interquartile filtering was applied to mitigate the influence of extreme outliers. For each variable of interest—sale price, usable area, built area, bathroom count, and floor count—the first (Q1) and third (Q3) quartiles were calculated, and the interquartile range (IQR = Q3 − Q1) was determined. Observations falling outside the interval [Q1 − 3·IQR, Q3 + 3·IQR] were removed, resulting in a homogeneous and robust dataset suitable for subsequent inferential and predictive modeling.

3.2 Model Architectures

This study presents a comparative analysis between the Decision Tree, Random Forest, and XGBoost regression algorithms. To minimize the mean squared validation error, the exhaustive grid search approach was employed to select the optimal hyperparameters for each model, with the aid of 5-fold cross-validation. The exhaustive grid search provided the best parameter settings that achieved the lowest average validation MSE for each of the respective algorithms, thus attaining a good trade-off between bias and variance [21].

The Decision Tree algorithm partitions the feature space by using splits that increase variance reduction according to the squared-error criterion. To avoid the over-splitting and overfitting, three complexity parameters were tuned as explained below:

- **max_depth**: (3, 5, 7, 10, 12, 15, 20, 25) - Limits the depth of the tree to balance representation ability with overfitting risk.
- **min_samples_split**: (2, 5, 10, 20) - Specifies the minimum number of samples to be used when making a split for a node.
- min_samples_leaf: (1, 2, 4, 10) - Guarantees terminal nodes to have a reasonable quantity of observations, smoothening predictions in regions of sparsity.

The Random Forest algorithm builds a series of decision trees by bootstrap aggregation (bagging) and random feature selection, thus reducing variance by averaging [22]. Four hyperparameters were explored:

- **n_estimators**: (10, 50, 100, 200, 400, 600, 800) - Controls the size of the forest of trees; larger forests

reduce variance more effectively at increased computational cost.

- **max_depth**: (3, 5, 7, 10, 12) - Max out the depth of each tree to control individual complexity.
- **max_features**: ("sqrt", "log2") - Controls the number of predictors considered at each split, introducing randomness which decorrelates individual trees further.
- **bootstrap**: (True, False) - Switches sampling with or without replacement, evaluating the effect of bootstrap aggregation on ensemble stability.

This extensive search guaranteed the choice of a forest configuration that minimized validation error while ensuring computational efficiency. XGBoost uses gradient tree boosting to train trees sequentially on the residuals of earlier iterations, with a squared-error loss combined with regularization to avoid overfitting. [23] Three of the most important parameters were optimized:

- **learning_rate**: (0.01, 0.1, 0.3) - Scales the contribution of each new tree, with lower values promoting slow learning and improved generalization.
- **max_depth**: (3, 5, 7, 10, 12) - Cuts off the depth of each boosted tree, balancing the ability to model complex patterns against overfitting.
- **n_estimators**: (10, 50, 100, 200, 400, 600, 800) – The total number of trees to be trained, more iterations allowing for more accurate residual correction, but also using more training time.

By applying the grid search technique over these grids, the optimal balance among learning rate, tree complexity, and ensemble size was determined.

3.3 Evaluation Measures

Models' performance was compared on an unseen test set of data against four complementary metrics chosen for their individual interpretive strengths in housing-market forecasting contexts: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2).

Root Mean Squared Error was selected since it punishes huge deviations more severely than small ones, aligning with the fact that huge pricing errors (e.g., underpricing a home by tens of thousands euros) are particularly unperformant for real-estate applications. By squaring residuals before averaging, RMSE rewards models that avoid occasional but big mispredictions, so it is a sensitive metric of worst-case performance.

Mean Absolute Error provides a straightforward, linear measure of average prediction error in the same currency units as the target, which is euros. Unlike RMSE, MAE treats all errors equally and is thus more robust to outliers and more interpretable for stakeholders who require an intuitive sense of actual deviation from the actual sale price.

To measure errors concerning scale, the Mean Absolute Percentage Error (MAPE) was applied. MAPE is a measure of the average deviation from the true price, thus allowing effective comparisons between segments with different levels of prices (e.g., comparing central business district penthouses with suburban apartments). The ratio measure is useful in providing decision-makers with the ability to measure the relative accuracy of the models.

Finally, the coefficient of determination quantifies the degree to which observed price variation is accounted for by the model. A value of R^2 close to 1 signifies strong explanatory power for market movements, while values close to zero imply that the predictor set fails to adequately capture systematic pricing patterns. Together, these four measures offer a complete and balanced evaluation of accuracy, robustness, relative error, and

goodness-of-fit, and thus allow for a strict comparison of the Decision Tree, Random Forest, and XGBoost models.

4 Results and Discussions

This chapter presents an in-depth view of both dataset features and exploratory analysis, along with each individual model's performance, and concludes with a comparative view of the trained models.

4.1 Dataset Presentation

The final dataset used in the analysis has 8500 entities, each with the following independent variables based on type:

- Continuous variables:
  - Built_area
  - Usable_surface
  - Lat
  - Lon
  - Metro_distance_in_meters
  - Distance_to_center
  - Floor_ratio
- Integer Variables:
  - Year_built
  - Floor_number
  - Number_of_floors
  - Bedroom_count
  - Bathroom_count
- Dummy variables:
  - Is_penthouse
  - Is_luxury
  - Is_first_floor
  - Is_last_floor
  - Is_detached

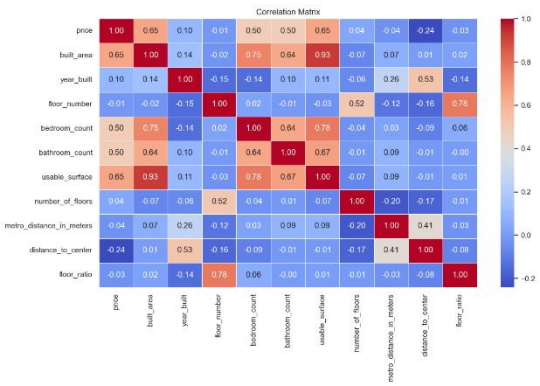

Fig. 2. Correlation Matrix

The correlation table in Figure 2 gives a general overview of the association between predictors and between predictors and the target variable. There is a moderate positive relationship between price and bathroom_count, bedroom_count, and usable_surface, which reflects the major role of building size and layout in influencing the price. In contrast, there is a weak negative relationship between distance_to_center and price, which shows that an increase in distance from the city center has the general tendency to lower the value of the properties. Furthermore, distance_to_center and metro_distance_in_meters have moderate positive relationships with year_built, which shows that newer constructions are most likely to be located far from the city center and metro stations. Finally, the strong relationship between usable_surface, built_area, bedroom_count, and bathroom_count supports the fact that larger living spaces usually have more bedrooms and bathroom counts. These findings reinforce the significant part played by structural and location variables in establishing the value of Bucharest residential properties.

Fig. 3. Price per m2 based on the location of the apartment

The hexbin map in Figure 3 illustrates price per square meter variability across the geography of the city. The higher values cluster together in the north-central suburbs, where there is high access speed to central amenities and a well-connected transport network, increasing demand. Prices fall off gradually from the center,

then fall lowest in the southern, eastern, and western peripheries—suburbs dominated by newly developed areas that haven't yet had time to experience the benefits of a fully settled urban infrastructure.

4.2 Models Overview

This chapter presents the final resulted models by the grid search algorithm along with their parameters and feature importance.

For the Decision Tree Regressor, the optimal model based on the MSE value resulting from the grid search algorithm has the following parameters:

- Max_depth: 10
- Min_samples_leaf: 10
- Min_samples_split: 2

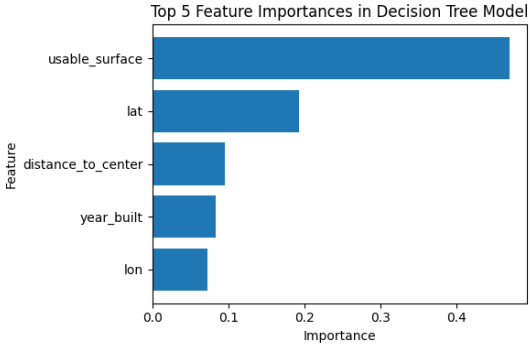

Fig. 4. Top 5 feature importances in Decision Tree Regressor

The barplot presented in Figure 4 presents the top 5 characteristics by importance, which are used by the decision tree model to predict the final price of the residential building. The most important variable used in forecasting is the usable surface, but the geospatial data represented by latitude and longitude, along with the distance from the city center, also play a key role in predictions.

Based on the same search algorithm, the optimised Random Forest algorithm has the following parameters:

- Max_depth: 12
- Max_features: ‘sqrt’
- N_estimators: 200
- Bootstrap: False

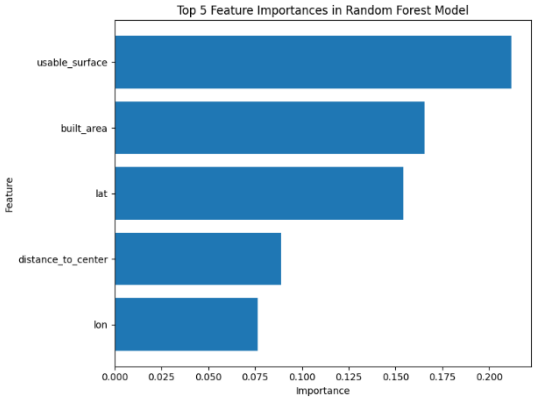

Fig. 5. Top 5 feature importances in Random Forest Regressor

Figure 5 illustrates the top 5 features by importance used by the Random Forest model. The barplot also suggests that distance to city center, geospatial location, and the surface of the apartment are among the most important features used for price prediction. Regarding the Decision tree model, the Random forest model uses the built area of the apartment instead of the year built as one of the top 5 features.

For the XGBoost ensemble model, the optimal parameters after training using the grid search algorithm are:

- Learning_rate: 0.1
- Max_depth: 7
- N_estimators: 800
-

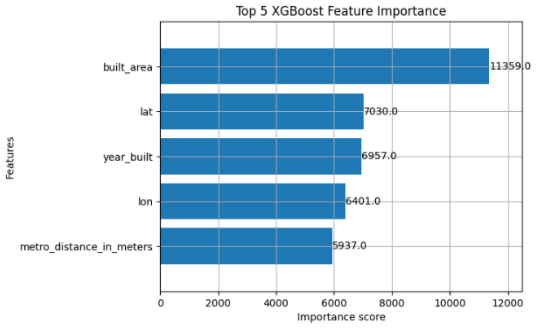

Fig. 6. Top 5 XGBoost feature importance

Figure 6 presents the top 5 features used by the XGBoost model in predictions. In the same manner as the last models, the surface of the residential unit, the location, and the distance to the center are the main factors used in predictions.

4.3 Model Comparison

Table 1. Performance of the models based on key indicators

	MAE	RMSE	MAPE	R2
Decision Tree	20487.65	32589.67	15.86	0.77
XGBoost	15089.36	24296.89	11.88	0.87
Random Forest	17134.86	26573.65	13.7	0.85

This chapter will analyze and compare the performance of the studied models presented in Table 1. The single Decision Tree model, although very interpretable, performs the worst in terms of prediction among the three methods. It has a Mean Absolute Error (MAE) of approximately 20,488 and a Root Mean Squared Error (RMSE) of 32,590 on the test set, meaning its point-predictions tend to be short of actual values by a significant margin. Its MAPE of 15.9 % shows that, on average, predictions are different from actual prices by nearly one-sixth, and its R^2 of 0.77 shows that only 77% of the variability in sale prices is explained. This result illustrates the model's tendency to overfit patterns in the training data and not to smooth out noise between different regions of the feature space.

By comparison, the Random Forest ensemble substantially reduces the bias and variance through the application of bootstrap aggregation. Its MAE is reduced to approximately 17,135 (a decrease of 16 % relative to the Decision Tree), and its RMSE is reduced to 26,574, while the MAPE is reduced to 13.7 %. Such reductions yield an R^2 equal to 0.85, indicating that 85 % of residential prices' variance is accounted for by the model. The improved performance comes from averaging the predictions of 200 decorrelated trees, each with a maximum depth of 12, which stabilizes estimates and lowers over-fitting at the expense of less interpretability.

Finally, the XGBoost model achieves the highest overall accuracy by sequentially correcting the residuals of its predecessor. It gives an MAE of 15,089 (26 % lower than the Decision Tree), an RMSE of 24,297, and a MAPE of just 11.9 %, which corresponds to an R^2 of 0.87. Practically, this means the gradient-boosted ensemble reduces average percentage error by nearly one-quarter compared to the tree baseline and explains 87 % of price variation. Though slightly more complex to train and to tune, XGBoost's ability to learn subtle feature interactions and to penalize overly complex trees makes it the go-to when minimizing prediction error is top priority.

5 Conclusion

In this research, a thorough comparison of tree-based regression algorithms Random Forest, Decision Tree, and XGBoost for predicting apartment prices in residential apartments in Bucharest has been performed. Using a true transactional dataset collected in March 2025, which was subject to detailed preprocessing, missing-value imputation, one-hot encoded, and creating new spatial indicators, each model was optimized using grid search combined with five-fold cross-validation to reduce the mean squared error (MSE) as the goal function. Among the independent variables involved, geospatial variables like proximity to the city center, longitude, latitude, distance to metro station, and unit size measurements (usable area and built area) were the most critical indicators of the price.

The results reveal a clear and noteworthy trend in the predictive accuracy. The single Decision Tree, while illustrating higher interpretability, incurred the largest testing errors, with the mean absolute percentage error being 15.9%. Application of the bootstrap averaging in the Random Forest approach reduced the mean error by 16%,

thereby attesting to the importance of variance reduction via ensemble methods. Finally, the gradient-boosted XGBoost model reported the lowest errors, demonstrating its ability to accurately make incremental corrections and discern complex, non-linear relationships.

Practically, these findings suggest that ensemble methods, particularly XGBoost, are significantly better than single trees when reducing pricing error is the goal. Random Forest is a good compromise if interpretability and computational cost need to be balanced against performance. In the meantime, the Decision Tree remains a good benchmark for rapid prototyping and stakeholder communication because it has interpretable decision rules.

Despite these advancements, some of these limitations remain to be explored in future research. First, our spatial attributes depend on fixed coordinates and do not account for dynamic urban factors such as traffic congestion or planned infrastructure development. Second, temporal dynamics (for instance, seasonality, macroeconomic data) were not explicitly modeled; employing time-series methods or adding lagged market indicators might enhance predictions further. Investigating these directions will contribute to real-time appraisal systems and support data-informed decision-making in Bucharest's transitioning real estate market.

References

[1] S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy,* vol. 82, no. 1, 1974.

[2] Z. Yang, "An application of the hedonic price model with uncertain attribute - The case of the People's Republic of China," *Property Management,* vol. 19, no. 1, pp. 50-63, 2001.

[3] R. P. M. &. S. S. Maurer, "Hedonic price indices for the Paris housing market," *Allgemeines Statistisches Archiv,* vol. 88, pp. 303-326, 2004.

[4] Y. C. N. X. R. Z. K. C. S. H. Linchuan Yang, "Place-varying impacts of urban rail transit on property prices in Shenzhen, China: Insights for value capture," *Sustainable Cities and Society,* vol. 58, p. 102140, 2020.

[5] M. a. J. a. Y. S. a. P. R. a. N. P. a. Indervati, "House Price Prediction Using Machine Learning," in *1st International Conference on Advances in Computing, Communication and Networking (ICAC2N),* Greater Noida, India, 2024.

[6] Y. L. M.-H. L. a. S.-Y. H. Jeonghyeon Kim, "A Comparative Study of Machine Learning and Spatial," *Sustainability,* vol. 14, no. 15, p. 9056, 2022.

[7] O. N. A. F. A. A. O. O. Y. F. A. G. O. Abigail Bola Adetunji, "House Price Prediction using Random Forest Machine Learning Technique," *Procedia Computer Science,* vol. 199, pp. 806-813, 2022.

[8] B.-S. T. S. W. W. Winky K.O. Ho, "Predicting property prices with machine learning algorithms," *Journal of Property Research,* vol. 38, no. 1, pp. 48-70, 2021.

[9] P. V. &. S. A. Grybauskas, "Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic," *Journal of Big Data,* vol. 8, p. 105, 2021.

[10] R. G. a. N. S. N. J. Manasa, "Machine Learning based Predicting House Prices using Regression Techniques," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA),* Bangalore, India, 2020.

[11] Y. L. Z. W. M. Z. T. W. C. Choujun Zhan, "A hybrid machine learning framework for forecasting house price," *Expert Systems with Applications,* vol. 233, p. 120981, 2023.

[12] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," in *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, Sydney, NSW, Australia, 2018.

[13] A. a. A. H. a. S. R. Rai, "Predicting Housing Sale Prices Using Machine Learning with Various Data Split Ratios," *Data and Metadata*, vol. 3, 2024.

[14] M. N. H. D. B. M. Quang Truong, "Housing Price Prediction via Improved Machine Learning Techniques," *Procedia Computer Science*, vol. 174, pp. 433-442, 2020.

[15] J. K. B. Byeonghwa Park, "Using machine learning algorithms for housing price prediction:," *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928-2934, 2015.

[16] B. a. S. Dutta, "Predicting the housing price direction using machine learning techniques," in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Chennai, India, 2017.

[17] F. Z. W. P. S. G. J. R. F. D. C. R. Yuhao Kang, "Understanding house price appreciation using multi-source big geo-data and machine learning," *Land Use Policy*, vol. 111, 2021.

[18] Foryś, "Machine learning in house price analysis: regression models versus neural networks," *Procedia Computer Science*, vol. 207, pp. 435-445, 2022.

[19] R. A. A. a. F. Darnis, "Use of Haversine Formula in Finding Distance Between Temporary Shelter and Waste End Processing Sites," *Journal of Physics: Conference Series*, vol. 1500, 2020.

[20] J. &. H. S. Carroll, "Using a video camera to measure the radius of the Earth," *Physics Education*, vol. 48, no. 6, p. 731., 2013.

[21] P.-B. G. W. Iwan Syarif, "SVM Parameter Optimization Using Grid Search and Genetic Algorithm to Improve Classification Performance," *TELKOMNIKA*, vol. 14, no. 4, pp. 1502-1509, 2016.

[22] Y. W. Y. Z. J. Liu, "New Machine Learning Algorithm: Random Forest," in *Information Computing and Applications. ICICA 2012.*, 2012.

[23] G. Tianqi Chen, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKD International Conference on Knowledge Discovery and Data Mining*, 2016.

Nicușor Andrei earned his bachelor’s degree in Economic Informatics in 2023 at the Academy of Economic Studies from Bucharest. He is a master's student at the same university and is currently working as a Data Analyst with expertise in ETL processes, dashboard designs, and database optimizations. His fields of interest are: data analysis, machine learning, deep learning, and LLM models.