*buildings*

MDPI

*Article*

# Machine Learning Valuation in Dual Market Dynamics: A Case Study of the Formal and Informal Real Estate Market in Dar es Salaam

**Frank Nyanda [1], Henry Muyingo [2] and Mats Wilhelmsson [3,*]**

[1] Department of Business Studies, Ardhi University, Dar es Salaam P.O. Box 35176, Tanzania; nyanda@kth.se
[2] Division of Real Estate Business and Financial Systems, KTH Royal Institute of Technology, SE-10044 Stockholm, Sweden; henry.muyingo@abe.kth.se
[3] Division of Real Estate Economics and Finance, KTH Royal Institute of Technology, SE-10044 Stockholm, Sweden
* Correspondence: matswil@kth.se

**Abstract:** The housing market in Dar es Salaam, Tanzania, is expanding and with it a need for increased market transparency to guide investors and other stakeholders. The objective of this paper is to evaluate machine learning (ML) methods to appraise real estate in formal and informal housing markets in this nascent market sector. Various advanced ML models are applied with the aim of improving property value estimates in a market with limited access to information. The dataset used included detailed property characteristics and transaction data from both market types. Regression, decision trees, neural networks, and ensemble methods were employed to refine property appraisals across these settings. The findings indicate significant differences between formal and informal market valuations, demonstrating ML's effectiveness in handling limited data and complex market dynamics. These results emphasise the potential of ML techniques in emerging markets where traditional valuation methods often fail due to the scarcity of transaction data.

**Keywords:** machine learning; real estate valuation; thin market; Dar es Salaam; the formal and informal housing market

check for updates

**Citation:** Nyanda, F.; Muyingo, H.; Wilhelmsson, M. Machine Learning Valuation in Dual Market Dynamics: A Case Study of the Formal and Informal Real Estate Market in Dar es Salaam. *Buildings* **2024**, *14*, 3172. https://doi.org/10.3390/buildings14103172

Academic Editor: Rotimi Abidoye

Received: 23 August 2024
Revised: 29 September 2024
Accepted: 2 October 2024
Published: 5 October 2024

**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Tanzania has undergone rapid urbanisation since the early 1990s partially due to trade liberalisation policies that various administrations have implemented. New measures such as the Mortgage Act of 2008 and the Unit Titles Act of 2008 opened up the market to include institutional investors and a variety of property developers. Previously, Tanzania's primary source of housing provision was public housing schemes overseen by the National Housing Corporation (NHC) since the early 1990s to encourage private investment from domestic and foreign sources. According to projections, the percentage of Tanzanians residing in urban areas is expected to grow to 49% by 2040 and the population expansion in Dar es Salaam is anticipated to increase by 5–6% annually [1]. The surge in urbanisation has significantly increased the demand for affordable quality housing. Though Dar es Salaam has a housing deficit of approximately 432,000 units according to CAHF [2], the annual contributions of formally registered developers amount to less than 5000 homes [3]. Thus a substantial amount of new housing is provided outside of the formal sector.

Transaction costs are influenced by the time and effort necessary to locate an appropriate residence to purchase or rent. Property listings in Dar es Salaam are available through globally recognised agents such as Knight Frank Tanzania and RE/MAX Tanzania. However, a significant number of residences that are not included in these listings are available on the market. The housing market in Dar es Salaam is loosely divided into formal and informal sectors, with formal transactions primarily taking place in designated settlements.

Property transactions in the formal market necessitate a certificate of occupancy or title deed. Nevertheless, Alananga [4] and Panman and Gracia [5] have not identified any market premium for formal land or house ownership documentation. This can partly be due to the general scarcity of housing as well as availability through the informal market. As noted in, e.g., Andreasen et al. [6] and Kemwita et al. [7], urban households in sub-Saharan cities like Dar es Salaam frequently acquire land and housing through unregulated market transactions and use these acquisitions as income-generating assets for rental housing and home-based enterprises.

The growth of commercial housing projects has significantly increased investment opportunities in the sector, as well as the number of intermediaries involved in transaction processes. Real estate agents in Dar es Salaam fall into formal and informal categories. The Business Registration and Licensing Authority of Tanzania (BRELA) registers formal agents, often serving the well-established property consulting firms connected to global networks. However, Komu [8] found that unregistered practitioners, locally known as "Dalalis", and termed as informal agents in this paper, provide most real estate agency services in Dar es Salaam. These individuals often possess extensive knowledge of up-to-date transaction prices and rent levels in specific areas and share information through informal networks [3]. As both formal and informal agents are active in the market, there is a significant degree of information from both categories that could be utilised to create a housing index for Dar es Salaam. However, as described in Nyanda [9], gathering information in nascent markets is labour-intensive and costly.

Rapid urbanization in conjunction with advances in technology have led many governments such as Tanzania as well as various companies to increase their utilization of assembled data so as to deliver services in cost efficient ways. The e-government strategy in Tanzania aims at providing smart services to the citizens in areas such as taxation, passport renewal, and real estate-related subjects such as the residential address verification system. Crucial in the efforts to provide smart services is the demand for the services as well as the perception of the users in regards to their sense of gain of the smart services, an issue that is paramount in studies on smart government services SGSs (see for example [9,10]). The deployment of smart technologies that utilise personal data also raises the issue of ethics and legality in accessing the data. The Tanzanian government's right to access as well as secure the handling of the private data is analysed in [11] with the conclusion that several laws need to be improved in order to provide appropriate protection for privacy. Another major challenge in emerging economies is the scarcity of data. However, using machine learning techniques, limited quantities of data have successfully been utilised for purposes such as the prediction of droughts in Tanzania [12]. In a nascent market such as the one in Dar es Salaam in which a lot of information on transactions is costly to acquire, there is a need to augment the data available in order to increase efficiency in the market. Enhancing data availability through machine learning processes could create a more accurate index than one based solely on observations of formal transactions.

Property valuation has traditionally relied on the comparison of observed transactions in relation to the object of interest in order to predict a future price. Hedonic price valuations are built around a comparison of lots of variables that might have an effect on the willingness to pay for the property in question. Even the level of pollution in various geographic areas can be factored in where relevant especially in relation to the quest for sustainable development at the city or global level as discussed in, e.g., [13].

The study presented in this paper aims to apply machine learning techniques to estimate property values in a property market constrained to a small number of observations characterised by formal and informal transactions. The methodology used was a mixture of data collection through interviews in the formal and informal sectors followed by data processing and quantitative analysis based on applying the various ML techniques. The drawing of conclusions ended the study process. Due to the absence of a comprehensive centralised database for transaction data in Tanzania, the information was collected exclusively from individual agents in a process described in Nyanda [14]. The

dataset consists of 954 unique observations—430 from informal agents and 524 from formal agents. The proximity distances were determined using Google Maps, which allowed one to measure the distances between each dwelling and several proximity characteristics. The data analysis in this paper utilises machine learning techniques, which include regression, decision trees, neural networks, boosting, elastic net, nearest neighbour, random forests, and support vector machines [15]. The evaluations are conducted on both in-sample and out-of-sample data. Due to the "black-box" nature of the technologies the results have been presented without a lengthy description of the testing methods used other than the three that are given in the assumption that the reader will have some knowledge of machine learning technology.

The novelty of the research in the field of real estate valuation as presented in this paper is in the utilisation of sophisticated valuation techniques in a market characterised by a scarcity of properly documented transactions over time but with an abundance of data on different property attributes. The real estate market in Dar es Salaam, Tanzania, is analysed through various machine learning approaches that have not been previously utilised. The paper also recognises the Tanzanian economy's coexistence of official and substantial informal sectors.

The paper is organised as follows: Section 2 presents a brief review of relevant literature on machine learning and real estate valuation. Section 3 outlines the ML techniques applied, followed by Section 4 which contains the empirical analysis. The discussion of the results is in Section 5 while Section 6 presents the conclusions from the study.

## 2. Literature Review

Machine learning (ML) techniques have been applied to real estate mass appraisal in studies, such as Kontrimas and Verikas [16], McCluskey [17], and Hoxha [18], that find the techniques to be superior to the traditional econometric or hedonic valuation methods. Mullainathan and Spiess [19] analysed ML techniques such as LASSO and random forests and concluded that they outperform conventional methods in out-of-sample predictions. Valier [20] also finds that automated valuation models (AVMs) using ML outperform hedonic models in predictive accuracy. The results in Teoh et al. [21] point to the superiority of ML techniques in dealing with linear and non-linear relationships between housing prices and the attributes as well as providing the benefit of more flexibility. With a focus on the ANN, Kutasi and Badics [22] provide results that indicate that AVMs or ML models exceed the performance of traditional hedonic models in significant terms.

However, many of the studies limit themselves to just a few techniques. In their study based on data from Lithuania, Kontrimas and Verikas [16] limited their analysis to three ML techniques: regression and computational intelligence-based techniques (MLP and SVM). ML methods for valuation have previously been widely adopted for price prediction such as in Park and Bae [23] who use four ML techniques to predict house prices in Fairfax county, Virginia. Chen et al. [24] use the support vector machine (SVM) to predict the housing market dynamics for Taipei city in Taiwan with significant accuracy. Phan [25] also adopted the SVM algorithm for the case of Melbourne to compare prices in different locations. Zhang et al. [26] use three ML techniques, linear regression, random forests, and decision trees, to predict house price trends for Greater Toronto and Hamilton.

The geocoding effect within property valuation has also been investigated in studies such as Tchuente and Nyawa [27] who analysed the efficacy of integrating geocoding into machine learning models to forecast real estate prices in different French cities. The researchers evaluated seven machine learning algorithms using a publicly available dataset provided by the French government. The dataset covers five years of real estate transactions. The results indicate that the use of geocoding features significantly improves the accuracy of predictions. Deppner et al. [28] examined the problem of spatial autocorrelation in hedonic models employed for real estate pricing and suggested a spatial cross-validation technique to obtain error estimates that are more precise and applicable to a broader range of situations. The study examines flat rental prices in Frankfurt, Germany, using tree-based

algorithms and comparing geographical and non-spatial cross-validation techniques. The findings indicate that typical non-spatial resampling approaches lead to overly optimistic error estimates.

Sezer et al. [29] highlight the superior performance of deep learning models over traditional ML techniques. They suggest that future research should develop more interpretable models and integrate new data sources, such as social networks, to improve the accuracy of the forecast. Cerulli [30] notes that techniques like meta-learning and ensemble methods reduce error and variance, with ensemble learners balancing accuracy and variance. The author demonstrates that this approach is particularly effective for complex econometric data, where traditional models fall short.

ML models are at times criticised for their obscure "black-box" nature, i.e., the models are not straightforward to interpret compared to the hedonic models. Rampini and Re Cecconi [31] note that the functional relationship between the ML model's inputs and outputs is rather intricate and that it just aims to provide a solution, rather than disclosing the workouts that would help to comprehend how the solution was arrived at. This limitation is also noted by other authors, e.g., Lorenz et al. [32], Molnar [33], Valier [20], and Glumac and Des Rosier [34], who recommended using interpretable machine learning techniques (IML) to improve transparency and understanding. This technique is also referred to by Lenaers et al. [35] as explainable artificial intelligence and they further note that IML allows for the comprehension of both explanations at the global as well as at the local level.

Osunsanmi [36] notes that modern valuation techniques such as ML are not used in African markets. The study presented in this paper provides a new insight on the evaluation of ML techniques to appraise real estate in formal and informal housing markets in the case of Dar es Salaam, a nascent real estate market in sub-Saharan Africa which has never been studied in this context.

## 3. Machine Learning Techniques

The literature review indicates gaps in the use of machine learning (ML) for real estate valuation. While many studies highlight the benefits of specific ML techniques in various settings, there is a lack of analyses that compare different ML methods, with Tchuente and Nyawa [27] and Abidoye et al. [37] as exceptions. This makes it harder to generalise the findings and determine the best approaches. To broaden the comparison between techniques, this paper tests and compares a total of eight ML techniques, including regression, decision trees, neural networks, Boost, elastic net, nearest neighbour, random forest, and support vector machines, which are briefly presented below. Machine learning can be classified as supervised and unsupervised. Supervised learning is the most common form of machine learning [15].

### 3.1. Learner: Regression

Regression *(Learner: Regression)* is supervised learning that aims to forecast a continuous dependent variable by utilising one or more independent variables [15,38]. Common methods include linear and polynomial regression. These methods presuppose a direct or polynomial correlation between the variables under investigation. Loss functions, such as the mean squared error, determine the optimal result. Linear regression is a straightforward and easily understandable technique that is very efficient when there is a linear or nearly linear correlation between variables, such as square footage, the number of bedrooms, location, and property value. However, the regression methodology may not accurately capture complex and non-linear connections between variables that could significantly influence real estate values [39,40].

### 3.2. Learner: Elastic Net

The elastic net (*Learner: Elastic net*) is a regularisation and variable selection technique that improves existing methods like the LASSO in situations with high-dimensional data

and multicollinearity among predictors [40]. It is a regression model that is especially effective when dealing with highly correlated data and predicts a continuous variable. This technique uses the advantages of both LASSO and ridge regression, yielding gains in situations where the number of predictors exceeds the number of observations. When there is a correlation among predictors, the elastic net tends to outperform the LASSO in terms of both prediction accuracy and variable selection. In their study, Zou and Hastie [40] proved that the elastic net method effectively minimises prediction errors in prostate cancer data, surpassing alternative approaches such as the LASSO and ridge regression. The elastic net method efficiently overcomes the limitations of LASSO regression by combining L1 and L2 penalties from LASSO and ridge regression, respectively. This approach helps to select groups among correlated predictors while ensuring stability. Therefore, this strategy effectively addresses the issues of multicollinearity and overfitting, resulting in a robust and reasonably easy-to-read model. However, the performance of this model may not be consistently comparable to that of more complex models. It is vital to choose proper hyperparameters carefully, such as the mixing parameter and the regularisation strength.

### *3.3. Learner: Tree*

A decision tree, or a regression tree (*Learner: Tree*), is a supervised machine learning algorithm that uses a tree structure to predict continuous output [41]. The methods partition the feature space into regions by making judgements at each node using feature values non-linearly. Regression trees offer a non-linear method that can capture more intricate patterns than linear regression without requiring extensive data. Trees partition the data into smaller subsets according to their features, enabling the collection of localised fluctuations in real estate values. The approaches effectively handle both category and numerical data and offer relevance rankings for various variables, helping to understand factors influencing property prices. Furthermore, regression trees can generate applicable models without extensive datasets and excel at capturing low variance in data, such as changes in real estate values. This attribute renders them well-suited for applications where the data displays intricate patterns that linear models may struggle to represent.

### *3.4. Learner: Forest*

The random forest (*Learner: Forest*) learner is a method that integrates many decision trees by utilising bootstrapped samples of the dataset and aggregating their predictions. Ensemble learning reduces prediction error by combining the predictions of multiple trees, either by taking the average in regression problems or voting in classification tasks [15]. The random forest method is commonly used to mitigate overfitting by aggregating predictions from many decision trees trained on different bootstrapped dataset samples [42].

### *3.5. Learner: Boost*

Boosting, also known as gradient boosting (*Learner: Boost*), is an ensemble technique that improves the performance of weak learners by iteratively training them on weighted versions of the dataset to create a robust prediction model. A weak learner, typically a decision tree, is trained using the data during the initial stage. During the subsequent phases, more trees are trained, utilising the residuals or errors of the combined model to improve performance-additive modelling [15].

### *3.6. Learner: SVM*

Support vector machines (*Learner: SVM*) are, according to Prosise [15], one of the newest algorithms at the forefront of machine learning research. Often, the algorithm is suitable for non-linear relationships. The SVM is commonly used for classification but can also be used for prediction [16]. Like other classifiers, the purpose of the SVM is to separate classes, but this can be done in infinite ways. The best boundary between the two classes is the one with the largest distance between the observations closest to the boundary (widest margin). SVM identifies these observations called support vectors. The algorithm

optimises a parameter C that specifies the distance in margin whereby low values of C represent wide margins, while high values indicate narrow margins.

### *3.7. Learner: Neural Network*

The neural network (*Learner: Neural Network*) is also known as deep learning within ML. In its simplest form, a neural network is an algorithm that takes a route from the inputs (the property attributes) called neurons, via several layers that also contain a number of neurons, to the output, which in the case of this study, are the property values. The relationships between all neurons are given a weight (w) and a bias (b) [15]. Given two inputs and a layer with two neurons, a neural network with five relationships can be created. With only one input, the neural network equals a simple regression model with intercept b (bias) and slope coefficient w (weight). We use two layers and optimise the number of neurons in each layer and the L2 penalty.

### *3.8. Learner: Nearest Neighbour*

K-nearest neighbours (*Learner: Nearest Neighbour*) is a straightforward prediction algorithm. The algorithm calculates the shortest distance to a number of neighbours [15,43]. The closest distance implies geographical distance and n-dimensional space. The factor that is optimised is the number of neighbours to include. Distance can be calculated in various ways, such as Euclidean or Minkowski.

## 4. Data Analysis and Evaluation of Models

### *4.1. The Research Design and Data Used*

The research design in this paper, which is based on a small data sample from a property market characterised by formal and informal transactions, is to apply several ML techniques that are then analysed on the basis of selected evaluation metrics. The evaluation metrics are presented first, followed by the descriptive statistics of the data in Table 1 and a brief explanation of the rationale of the models based on the characteristics of the data.

MAPE (mean absolute percentage error) has a superior interpretability advantage (having meaning on its own) over other metrics as it highlights the proportional magnitude of errors in percentage terms, making it easier to understand than other metrics [44,45]. MSE (mean squared error) provides the possibility for multiple perspectives in ML model evaluations. It has the unique ability to deal with large deviations by punishing significant errors [46], a situation not unique in real estate valuation. Like the MAPE, MSE is also computationally simple and easy to follow. Cross-validation, particularly k-fold, in conjunction with MAPE and MSE, provides the opportunity to identify the overfitting and underfitting of the ML models [26]. Whereas both MAPE and MSE can signal possible overfitting, cross-validation can easily confirm this. Cross-validation is also highly useful in small data samples as it ensures the optimal use of the available data [47].

Given the limited size of the data, sample cross-validation is useful for the aims of this paper. It evaluates a model's ability to generalise to an independent dataset, mainly focusing on how much variance in the target variable the model can explain. MAPE, which measures the accuracy of predictions, and MSE, which assesses the average squared difference between estimated and actual values, are two other evaluation metrics used in this paper in combination with cross-validation. These three evaluation metrics, MAPE, MSE, and k-fold cross-validation are also easy to use with STATA 17.0, a program that was utilised in processing the data. These reasons also provide an explanation as to why these three were picked out of several others given in the literature, that might not be the first choice for a small dataset managed with STATA.

**Table 1.** Descriptive statistics—in-sample and out-of-sample.

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| Price (1,000,000) | 954 | 193.124 | 184.342 | 5 | 1566 |
| Kimabu | 954 | 0.481 | 0.5 | 0 | 1 |
| Goba | 954 | 0.123 | 0.328 | 0 | 1 |
| Tabata | 954 | 0.129 | 0.335 | 0 | 1 |
| Kawe | 954 | 0.071 | 0.257 | 0 | 1 |
| No storeys | 954 | 1.06 | 0.246 | 1 | 3 |
| Roof cas | 954 | 0.158 | 0.365 | 0 | 1 |
| Roof asbestos | 954 | 0.021 | 0.143 | 0 | 1 |
| Roof clay tiles | 954 | 0.093 | 0.291 | 0 | 1 |
| Ceil gypchtng | 954 | 0.773 | 0.419 | 0 | 1 |
| Window wood | 954 | 0.637 | 0.481 | 0 | 1 |
| Floor cerrtiles | 954 | 0.405 | 0.491 | 0 | 1 |
| Floor terrazo | 954 | 0.006 | 0.079 | 0 | 1 |
| No. bedrooms | 954 | 3.351 | 0.944 | 1 | 8 |
| Plotsize | 954 | 303.32 | 268.455 | 24 | 2000 |
| Fence | 954 | 0.51 | 0.5 | 0 | 1 |
| 2010 | 954 | 0.072 | 0.259 | 0 | 1 |
| 2011 | 954 | 0.039 | 0.193 | 0 | 1 |
| 2012 | 954 | 0.048 | 0.214 | 0 | 1 |
| 2013 | 954 | 0.08 | 0.271 | 0 | 1 |
| 2014 | 954 | 0.078 | 0.268 | 0 | 1 |
| 2016 | 954 | 0.146 | 0.353 | 0 | 1 |
| 2017 | 954 | 0.151 | 0.358 | 0 | 1 |
| 2018 | 954 | 0.128 | 0.334 | 0 | 1 |
| 2019 | 954 | 0.101 | 0.301 | 0 | 1 |
| Distance road | 954 | 0.154 | 0.191 | 0.005 | 1.64 |
| Distance hospital | 954 | 1.612 | 1.479 | 0.045 | 10.544 |
| Distance airport | 954 | 14.864 | 7.49 | 3.161 | 33.707 |
| Distance food market | 954 | 2.439 | 2.716 | 0.131 | 12.53 |
| Y-coordinate | 954 | −6.746 | 0.069 | −7.004 | −6.579 |
| X-coordinate | 954 | 39.204 | 0.041 | 39.083 | 39.342 |

Note: The table shows descriptive statistics regarding the dependent variable price measured in TZS 1,000,000 and the independent variables (neighbourhoods, property attributes, transaction year, distance to amenities, and coordinates). The table presents mean values, standard deviation, kurtosis, skewness, and spatial dependency (Moran's I).

Table 1 provides an overview of the properties in the dataset, measured by the average and the variability in price, the types of buildings, and their geographical distribution.

The dependent variable price (measured in TZS 1,000,000 (TZS 1,000,000 = EUR 373.150000 (25 April 2024)) has a high standard deviation relative to the mean, indicating a wide variability in property prices: the mean price is 193.124 million with a standard deviation of 184.342 (range: 5 to 1566). Figure 1 illustrates the distribution of prices. There is a certain skewness in the material with more transactions that have lower prices than in a normal distribution and fewer properties that are more expensive. There is a higher kurtosis than a normal distribution and a number of outliers in the material.

The numbers in Table 1 indicate a geographic distribution of properties with about 48.1% in the Kimabu neighbourhood, 12.3% in Goba, 12.9% in Tabata, and 7.1% in Kawe. The data further indicate that most properties are low-rise, with the majority having one storey, and that the predominant roofing material at 15.8% is 'roof cas', with 2.1% using 'roof asbestos', and 9.3% using 'roof clay tiles'. Ceiling type (ceil gypchtng) indicates a predominant use of gypsum ceiling technology (proportion: 77.3%). The majority of properties (63.7%) have wooden windows. Floor types (floor cerrtiles and floor terrazzo) proportions are 40.5% for 'floor cerrtiles' and 0.6% for 'floor terrazzo', indicating a higher usage of ceramic tiles over terrazzo flooring. The number of bedrooms (No. bedrooms) is, on average, 3.351 with a standard deviation of 0.944, which indicates a moderate to high number of bedrooms per property suitable for family living. Plot size is, on average,

303.32 $m^2$ with a standard deviation of 268.455 $m^2$, indicating a wide variation in plot sizes, reflecting a diverse property market. About half of the properties have fencing.

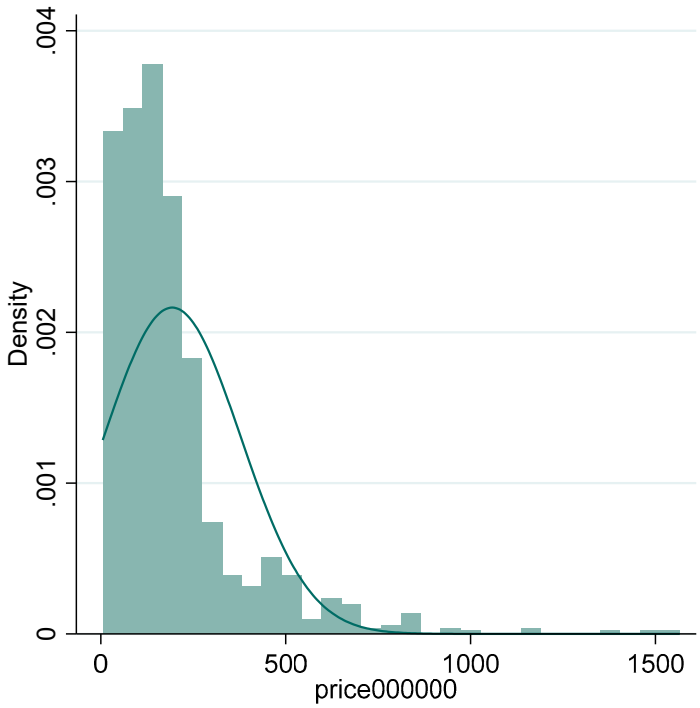



**Figure 1.** Histogram dependent variable.

Properties were sold from 2010 to 2019, with fewer transactions each year, indicating a steady but low volume of new transactions throughout the decade. Distances to key amenities like arterial roads, hospitals, airports, and food markets vary, with properties being relatively closer to arterial roads and food markets on average compared to hospitals and airports. Nyanda [14] presents a map with the locations of all of the transactions used in the dataset.

We use this dataset for machine learning applications to predict property prices. It comprises a wide range of property characteristics used in the predictive modelling, including numerical variables such as price, the number of floors, the number of bedrooms, plot size, distances to amenities, and geographical coordinates. It also incorporates categorical and binary variables such as neighbourhood location, the type of roofs, ceilings, windows, floors, fencing, and the year of construction.

The wide range of characteristics provides the possibility to develop models capable of capturing various aspects of real estate valuation. The substantial variability indicated by the high standard deviation in price and plot size variables is important for understanding property valuation's underlying distribution and dynamics. Models must be robust to handle this variability effectively. Several binary variables, such as roof type or neighbourhood, and features like specific roofing materials or construction years, suggest a potential under-representation of some categories. This could pose challenges in training machine learning models. Including geographical coordinates enhances the potential for spatial dependency, which could reveal trends and patterns based on location, such as clusters of high-value properties. The dataset also captures temporal characteristics, such as the year of sales.

Given the mixed nature of data, numbers, categorical data, and binary data, various machine learning models such as linear and polynomial regression, decision trees and random forests, gradient boost machines, and neural networks can be employed. The choice of model depends significantly on the model's ability to handle specific statistical properties of the data, such as outliers, multicollinearity, and heteroskedasticity. Given the

evident differences in data subsets (in-sample and out-of-sample), it is important to utilise cross-validation techniques to ensure the model's generalisability and to prevent overfitting.

### *4.2. Machine Learning Results*

Table 2 compares eight machine learning approaches: regression, elastic net, regression tree, boost, forest, neural network, SVM, and nearest neighbour. The models are evaluated by presenting metrics for both the training data (in-sample) and the testing data (out-of-sample). The models were refined by adjusting the optimal parameters, such as the tree depth for the regression tree and boost and the number of neurons or layers for the neural network. Using a dataset covering 419 transactions in the training dataset and 104 transactions in the testing dataset, the following results were obtained.

**Table 2.** Real estate valuation performance for the formal market.

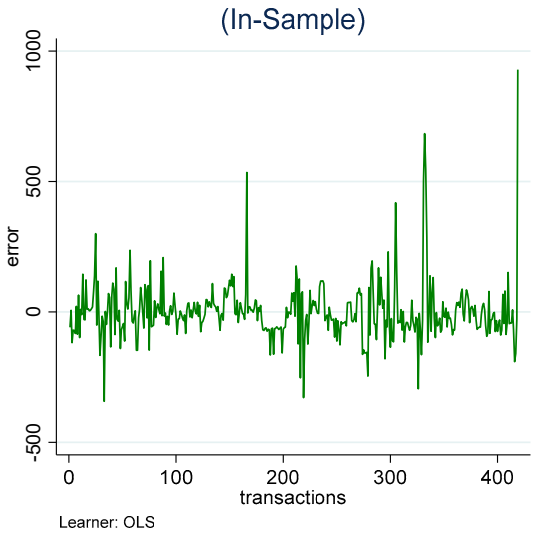
| | **MAPE** | | **MSE** | | **Cross-Validation** | |
|---|---|---|---|---|---|---|
| | **Training** | **Testing** | **Training** | **Testing** | **Training** | **Testing** |
| Regression | 68.142 | 92.410 | 11,503.529 | 13,731.797 | 0.745 | 0.229 |
| Elastic net | 80.727 | 89.890 | 22,128.828 | 20,498.686 | 0.495 | 0.336 |
| Regression tree | 74.329 | 94.020 | 17,022.891 | 26,605.547 | 0.599 | 0.045 |
| Boost | 83.225 | 101.460 | 16,299.616 | 13,646.772 | 0.604 | 0.188 |
| Forest | 35.586 | 56.420 | 7844.813 | 15,963.183 | 0.838 | 0.477 |
| Neural network | 88.532 | 108.594 | 23,820.044 | 21,733.466 | 0.562 | 0.392 |
| SVM | 9.306 | 52.399 | 9031.117 | 17,819.553 | 0.250 | 0.154 |
| Nearest neighbour | 0.000 | 37.611 | 0.000 | 14,385.118 | 1.000 | 0.335 |

Note: The table shows the results of our analysis of eight machine learning algorithms. The valuations are based on the different methods used and evaluated with MAPE (mean absolute percentage error) and MSE (mean squared error). The result is shown for both in-sample and out-of-sample transactions from the formal market. The target variable is the price measured in Tanzanian shillings (TZS 1,000,000). The number of features equals 24, the number of training transactions is 419, and the number of testing transactions is 104. The cross-validation results are based on five folds, and the accuracy measures how much variance is explained. Optimal parameters for each learner: elastic net (optimal penalising parameter = 0.75 and optimal elastic parameter = 1), regression tree (optimal tree depth = 3), boost (optimal learning rate = 1, optimal tree depth = 4, and the optimal number of trees = 3), forest (optimal number of splitting features = 10, optimal tree depth = 6, and the optimal number of trees = 5), neural network (optimal number of neurons in layer 1 = 4, optimal number of neurons in layer 2 = 5, and optimal L2 penalisation =3), SVM (optimal C parameter = 300 and optimal Gamma parameter = 0.1) and nearest neighbour (optimal number of nearest neighbours = 6). We used Stata 17.0 (command: r_ml_stata_cv ) and Python 3.12 (pandas, numpy, and scikit-learn). See [30] for further details.

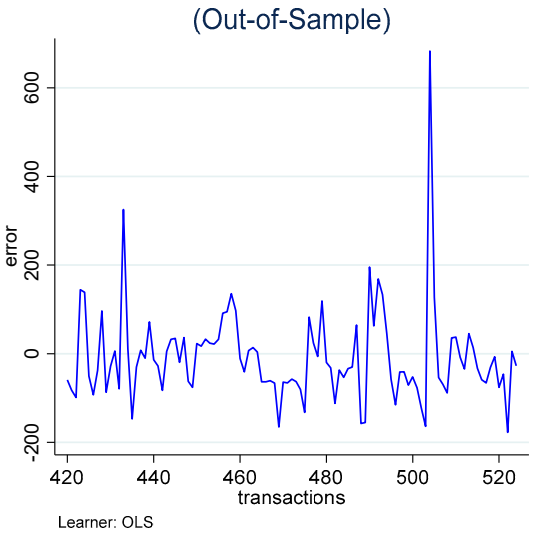
The SVM (support vector machine) shows the best MAPE and the fourth-best testing MSE, indicating strong generalisation despite its more moderate performance on training data. The forest model demonstrates the second lowest MAPE for training and a relatively low testing MAPE, suggesting good predictive performance. Nearest neighbour achieves perfect training MAPE, implying that it fits the training data without error but decreases performance on testing data. The neural network and elastic net models exhibit relatively high error rates and MSE across training and testing datasets compared to other models.

A ranking based on the results in Table 2 is as follows: (1) SVM (support vector machine), as it offers a robust generalisation capability, with relatively low errors on unseen data; (2) forest, as it exhibits solid predictive accuracy and generalises well across new data; (3) nearest neighbour, as it shows the best performance metrics in testing but the perfect training MAPE suggests possible overfitting; (4) elastic net, as it has moderate performance, with a balance between error metrics and generalisation reflected in its cross-validation score; (5) regression comes in fifth place, as it offers moderate to good generalisation but with higher prediction error; (6) regression tree, as it has lower performance with the highest MSE, indicating significant prediction errors; (7) boost is in place number seven, as it is comparable to the regression model but with a slightly worse error in MAPE; and (8) neural network, which ranks last, having the poorest performance on testing data both in terms of MAPE and MSE, indicating less predictive accuracy and reliability.

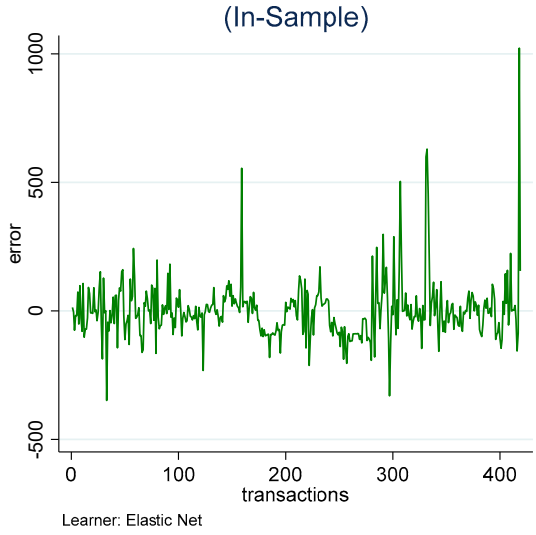
The prediction errors for the machine learning models applied to the formal market can be presented in graphic form as shown in Figure 2 below.



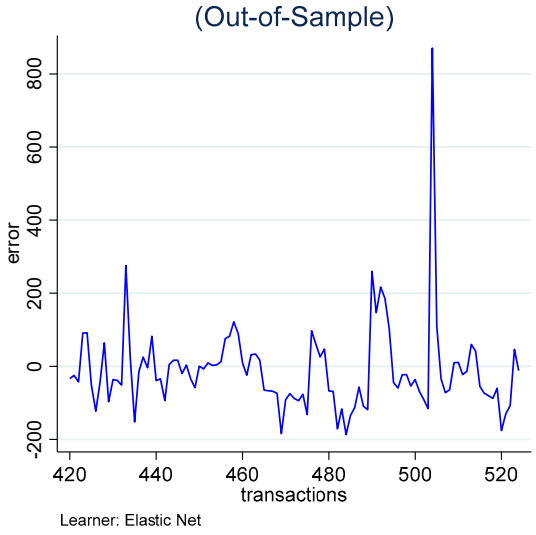

(**a**)



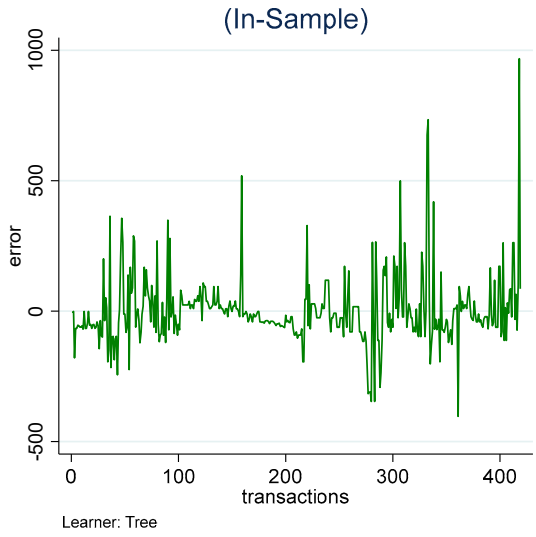

(**b**)



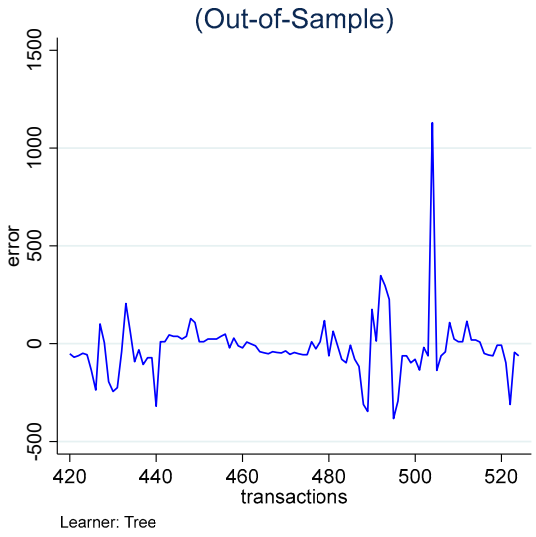

(**c**)




(**d**)

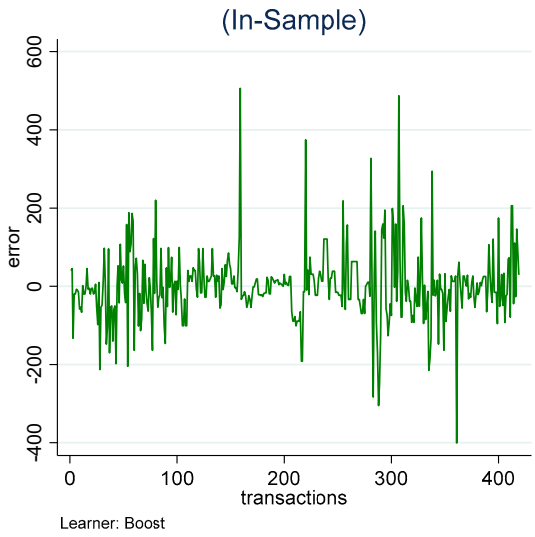



(**e**)

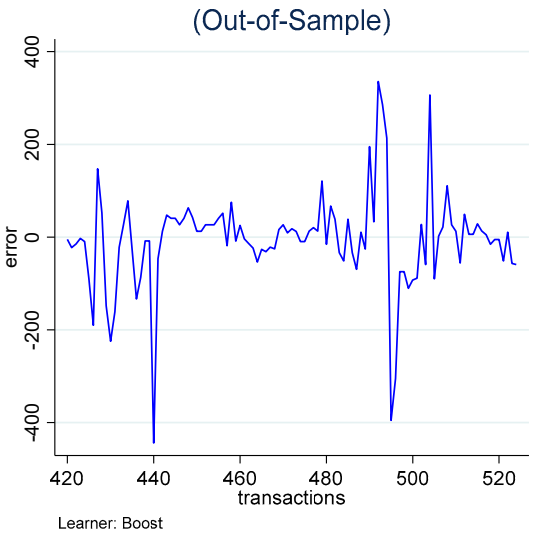



(**f**)

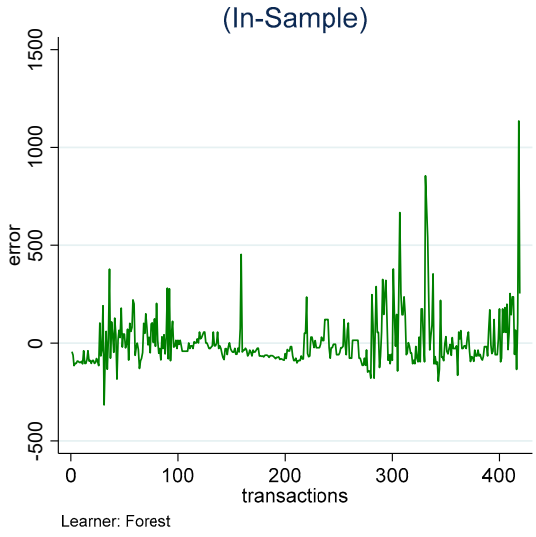
**Figure 2.** *Cont*.



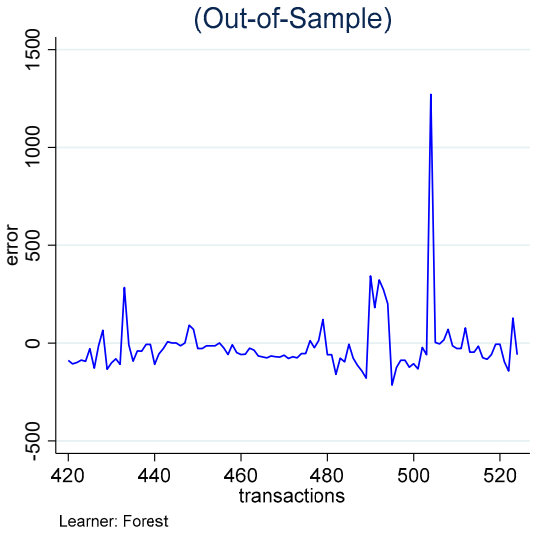

(**g**)



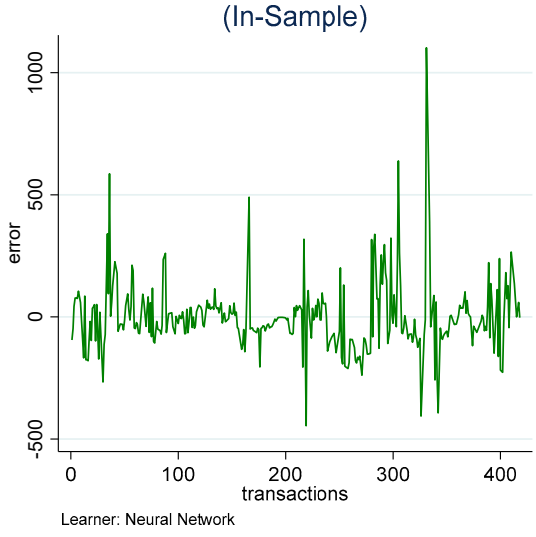

(**h**)



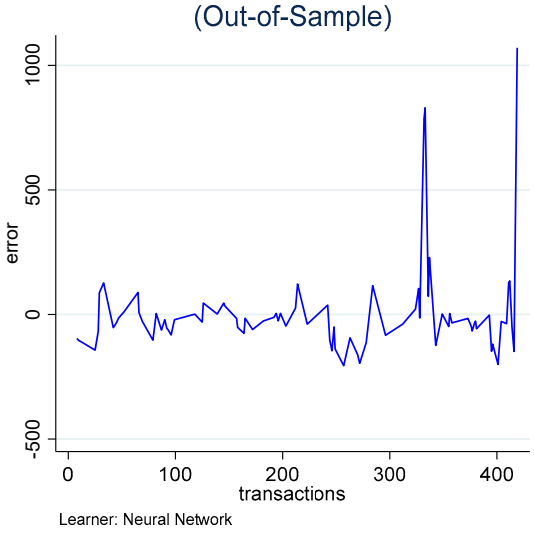

(**i**)




(**j**)




(**k**)




(**l**)

**Figure 2.** *Cont*.

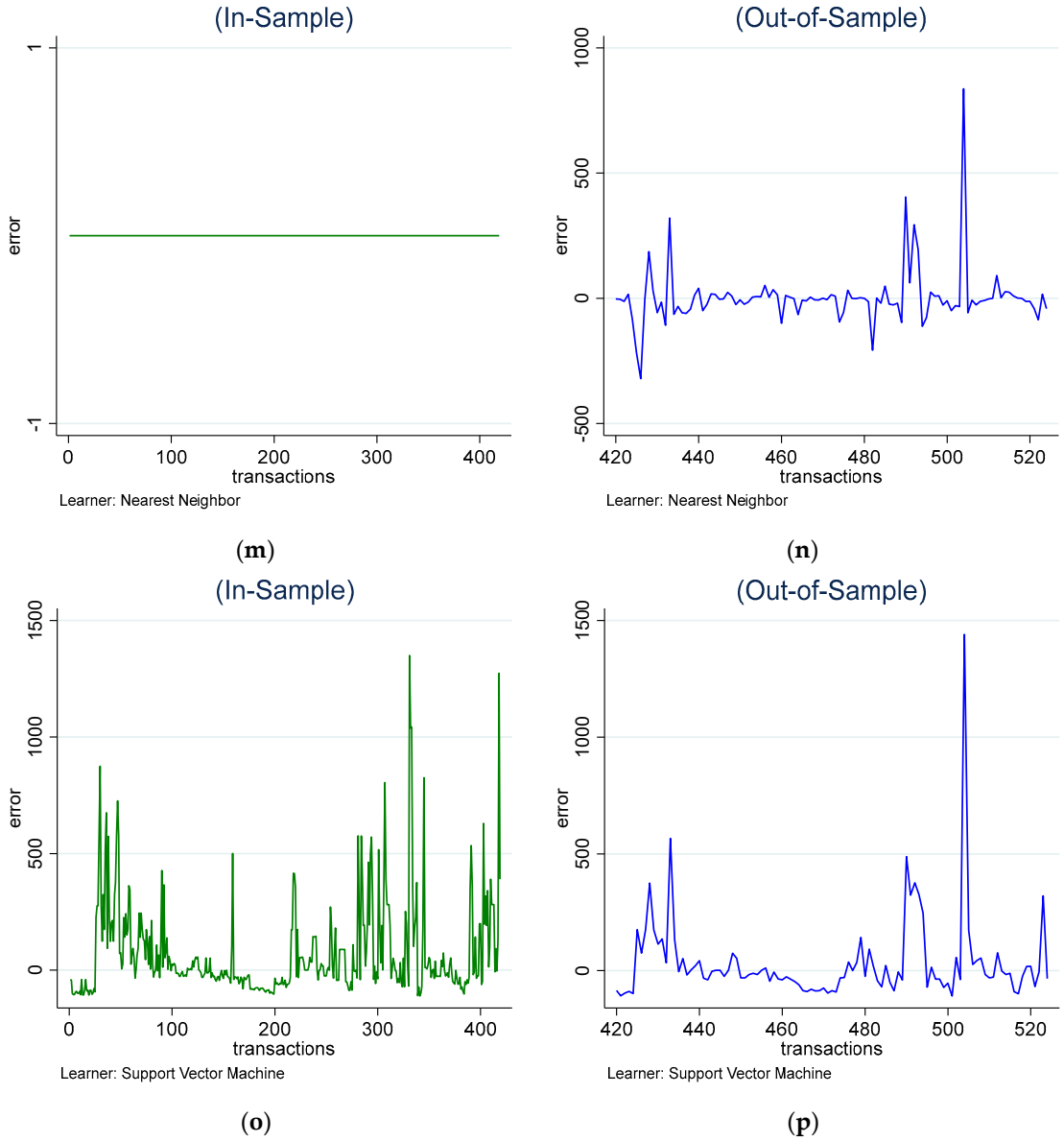



**Figure 2.** (**a**–**p**) The eight ML models' in-sample and out-of-sample performance (prediction error) on the formal market. Note: The figure shows the prediction error defined as actual minus predicted values regarding in-sample (left in green) and out-of-sample (right in blue) data. Each pair of diagrams relates to a specific learner. The vertical axis measures the prediction error in Tanzanian shillings (TZS 1,000,000), and the horizontal axis measures each transaction's identification number. The figure refers only to valuations made on the formal market. The results are based on the estimates in Table 2.

In Figure 2, the left side (green) illustrates in-sample errors, showing how well each model fits the training data. In contrast, the right side (blue) displays out-of-sample errors, highlighting the models' generalisation ability on unseen data. The vertical axis quantifies the prediction error in Tanzanian shillings (TZS 1,000,000), and the horizontal axis corresponds to individual transaction IDs. The figure emphasises the varying performance of each model, indicating a trade-off between fitting the training data and predicting new, unseen transactions accurately.

Table 3 compares ML learners for the combined formal and informal housing market transactions. Using a dataset covering 763 transactions in the training dataset and 191 transactions in the testing dataset, the following findings were obtained.

**Table 3.** Real estate valuation performance for the formal and informal market.

| | MAPE | | MSE | | Cross-Validation | |
|---|---|---|---|---|---|---|
| | Training | Testing | Training | Testing | Training | Testing |
| Regression | 80.256 | 73.098 | 11,796.093 | 12,532.698 | 0.672 | 0.370 |
| Elastic net | 74.886 | 69.075 | 12,586.828 | 13,439.529 | 0.647 | 0.424 |
| Regression tree | 152.035 | 137.899 | 24,895.351 | 28,213.442 | 0.425 | 0.040 |
| Boost | 37.652 | 47.985 | 4426.025 | 13,670.108 | 0.842 | 0.203 |
| Forest | 30.897 | 52.652 | 4709.234 | 13,728.414 | 0.880 | 0.503 |
| Neural network | 59.752 | 63.486 | 14,859.287 | 15,296.036 | 0.344 | 0.206 |
| SVM | 3.830 | 92.332 | 2937.825 | 15,161.066 | 0.921 | 0.239 |
| Nearest neighbour | 0.000 | 75.58 | 0.000 | 13,702.390 | 1.000 | 0.067 |

Note: The table shows the results of our analysis of eight machine learning algorithms. The valuations are based on the different methods used and evaluated with MAPE (mean absolute percentage error) and MSE (mean squared error). The result is shown for both in-sample and out-of-sample transactions from formal and informal markets. The target variable is the price measured in Tanzanian shillings (TZS 1,000,000). The number of features equals 24, the number of training transactions is 763, and the number of testing transactions is 191. The cross-validation results are based on five folds, and the accuracy measures how much variance is explained. The results in the table are the results of the cross-validation grid search. Optimal parameters for each learner: elastic net (optimal penalising parameter = 1 and optimal elastic parameter = 1), regression tree (optimal tree depth = 1), boost (optimal learning rate = 1, optimal tree depth = 3, and the optimal number of trees = 4), forest (optimal number of splitting features = 10, optimal tree depth = 8, and the optimal number of trees = 5), neural network (optimal number of neurons in layer 1 = 5, optimal number of neurons in layer 2 = 2, and optimal L2 penalisation = 3), SVM (optimal C parameter = 700 and optimal Gamma parameter = 0.1) and nearest neighbour (optimal number of nearest neighbours = 5). We used Stata 17.0 (command: r_ml_stata_cv ) and Python 3.12 (pandas, numpy, and scikit-learn). See [30] for further details.

The results suggest that the regression model exhibits moderate prediction errors, performing slightly better on testing data. For example, the elastic net model shows lower prediction errors than the basic regression model. However, an increase in MSE in out-of-sample data suggests high, though stable, errors. On the other hand, cross-validation scores reflect good generalisation capabilities.

The regression tree model reveals significant errors in valuation predictions, but shows some improvement on testing data. This model shows the highest MSE, indicating substantial average errors. The low cross-validation score in testing (0.040) underscores poor generalisation. The results indicate that the boost learner model outperforms other models in terms of prediction accuracy. The MSE increase using test data indicates sensitivity to new data. This ML method has a training cross-validation score of 0.842 that in testing drops significantly to 0.203, emphasising the risk of overfitting.

The forest model has the second lowest MAPE (52.652), following the boost model, and demonstrates relatively good predictive accuracy, with a low MSE in both training and testing. Robust cross-validation scores, especially in testing (0.503), suggest strong generalisation capabilities.

The neural network learner, with moderate MAPE values that are better than the basic regression but not as good as more complex models like boost or forest, shows moderate MSE values that remain stable across training and testing. However, lower cross-validation scores suggest a limited ability to generalise.

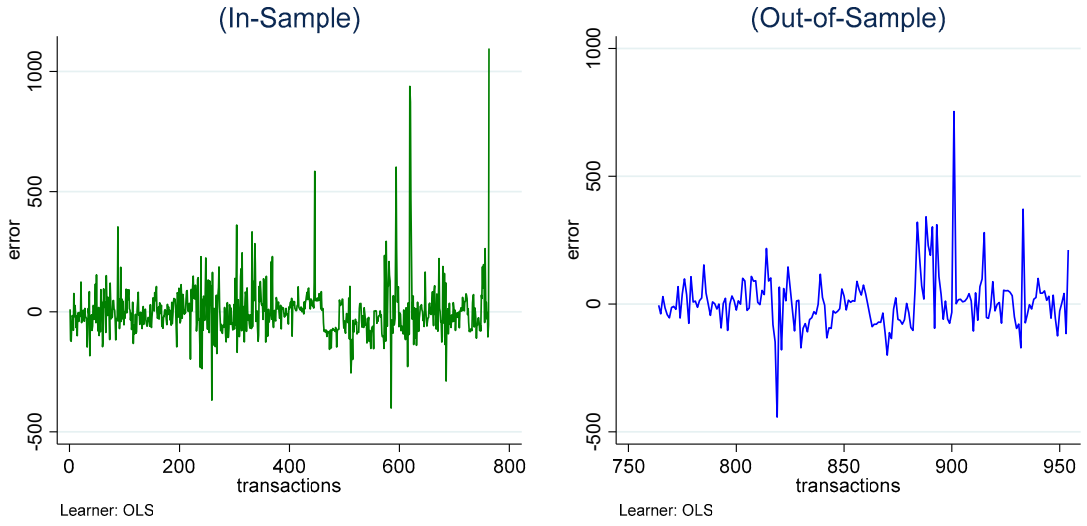
The SVM (support vector machine) model exhibits a very low training MAPE (3.830) but a high testing MAPE (92.332), indicating an excellent fit to training data but poor generalisation. Additionally, a low training MSE that increases during testing further underscores overfitting. A reasonable training cross-validation score (0.921) that drops substantially in testing (0.239) confirms concerns about overfitting.

For the nearest neighbour model, perfect training performance (MAPE and MSE are zero) indicates that the model completely memorises the training data. However, testing performance significantly weakens (MAPE: 75.580, MSE: 13,702.390). Near-perfect training cross-validation (1.000) with a dramatic drop in testing (0.067) indicates severe overfitting.

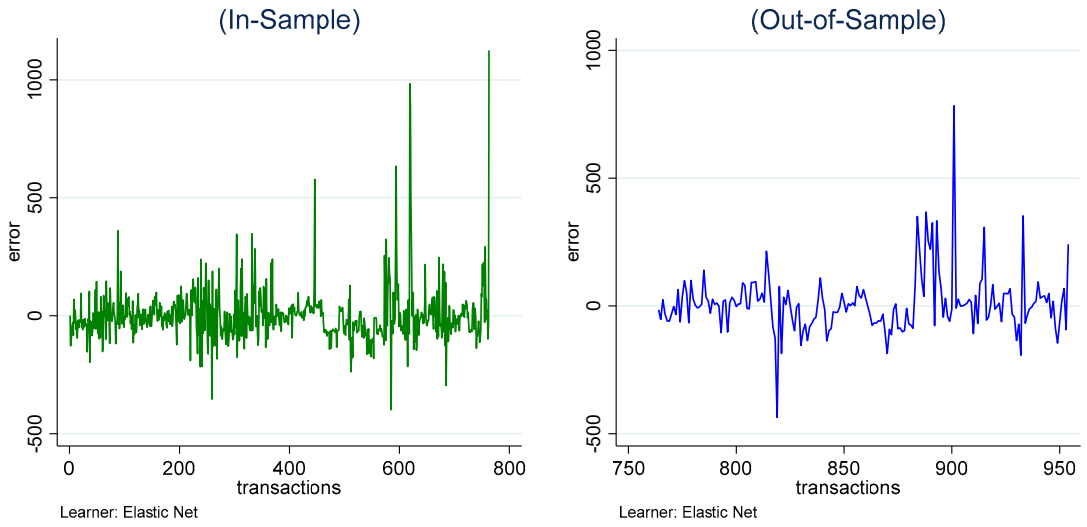
Based on their performance in both training and testing phases, considering MAPE, MSE, and cross-validation scores, the models can be ranked as follows: (1) the forest model, with low MAPE values and consistent MSE, indicating good predictive accuracy and strong

generalisation capabilities; (2) the boost model, offering the lowest MAPE in training while maintaining reasonable accuracy in testing despite a relative increase in MSE; (3) the elastic net, which consistently shows lower MAPE in training and testing than simpler models like regression, albeit with a slightly higher MSE, suggesting reasonable generalisation capability despite a drop in cross-validation scores; (4) the regression model, which is ranked in fourth place due to slightly better performance on testing data and relatively decent generalisation; (5) the SVM comes in fifth place due to poor generalisation in testing despite reasonable generalisation in training; (6) the neural network with moderate errors and lower generalisation follows; (7) the regression tree, with the poorest generalisation capabilities ranks second last; and (8) nearest neighbour, with severe overfitting, is ranked last.

Figure 3 shows the prediction errors for both the in-sample and out-of-sample phases in various machine learning models applied to the formal and informal real estate markets. The errors, calculated as the difference between actual and predicted values, are represented on the vertical axis in millions of Tanzanian shillings. In contrast, the horizontal axis shows each transaction's identification number. The left section (green) depicts in-sample errors, and the right section (blue) shows out-of-sample errors. This visualisation focuses on valuations within the formal market, comparing metrics in Table 3.




(**a**) (**b**)




(**c**) (**d**)

**Figure 3.** *Cont.*

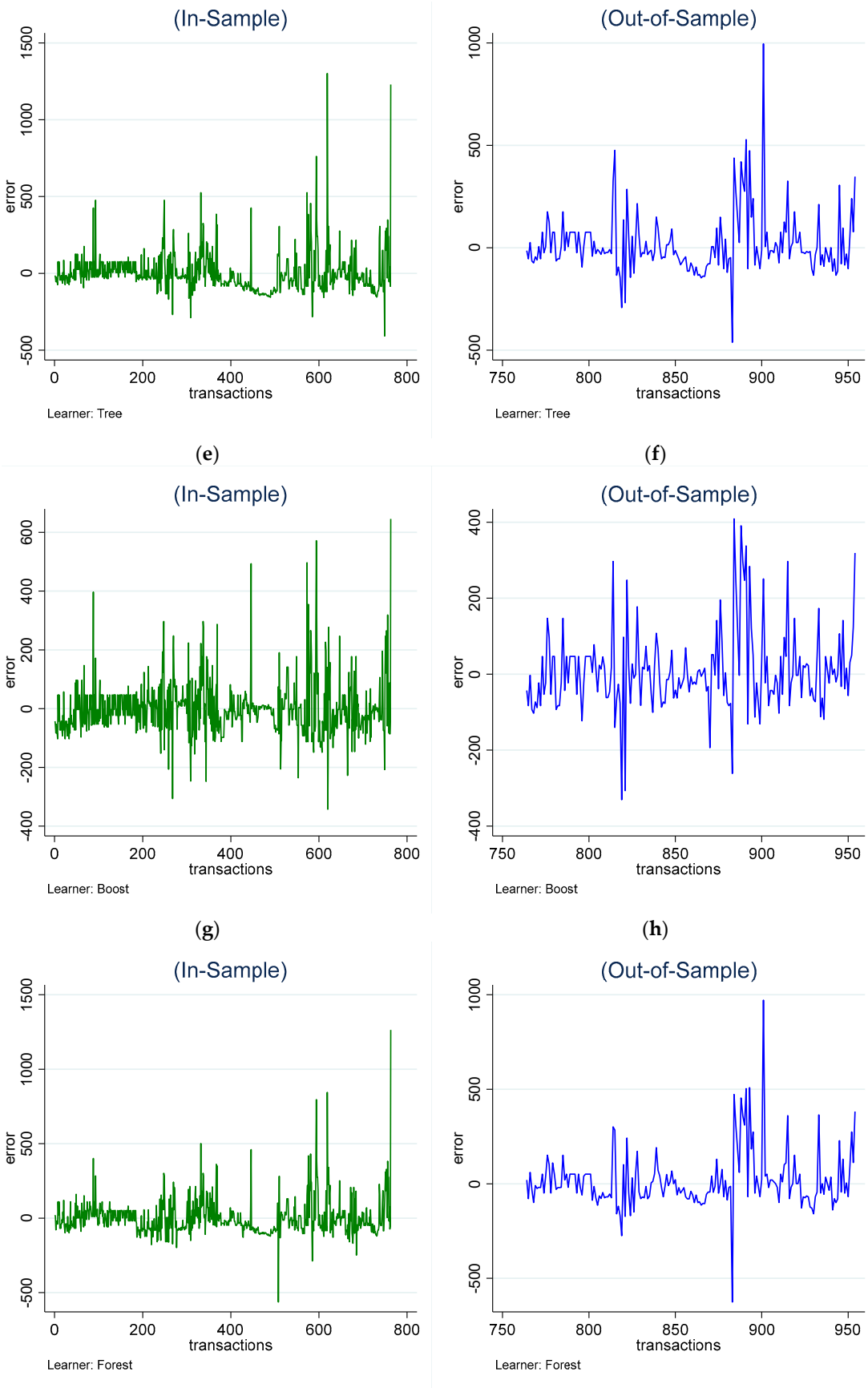
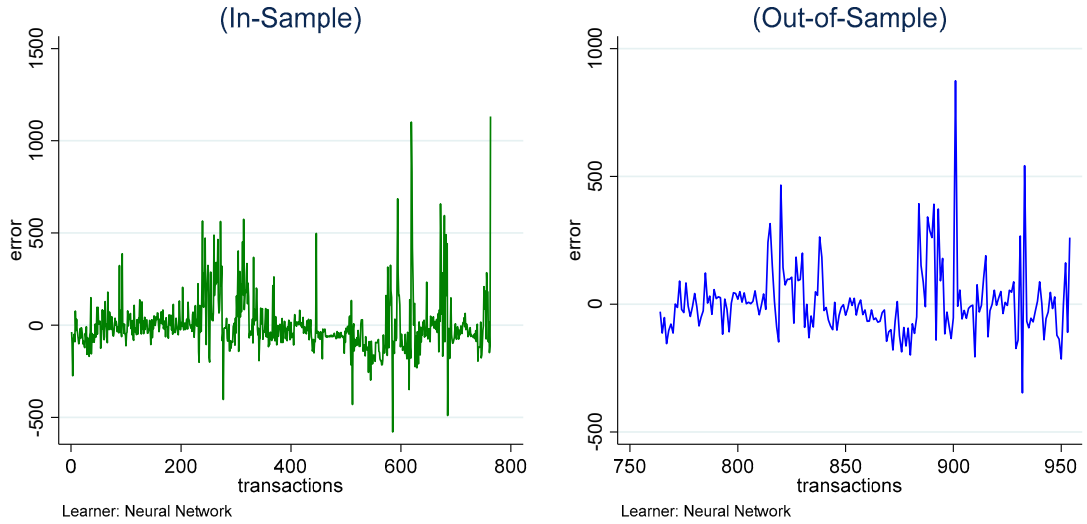
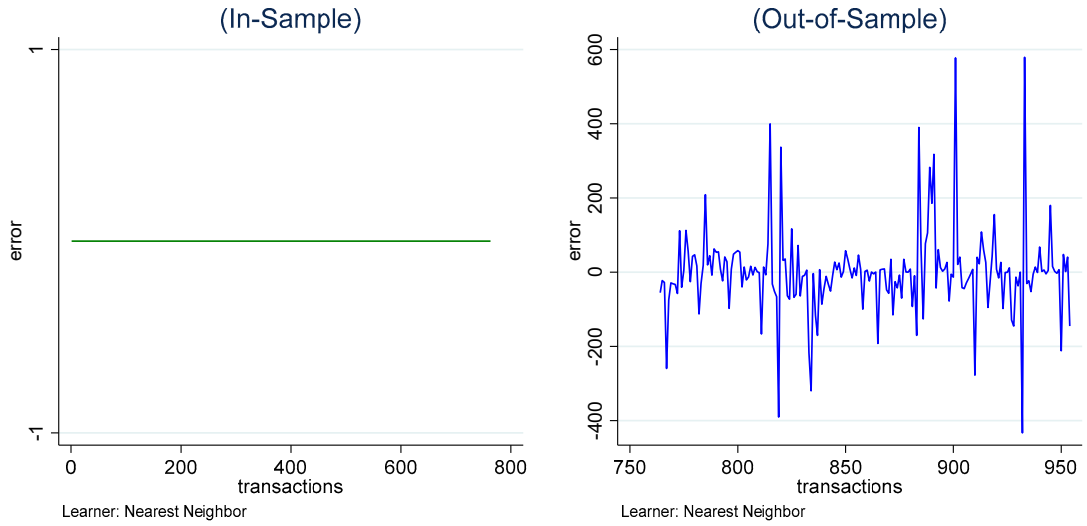


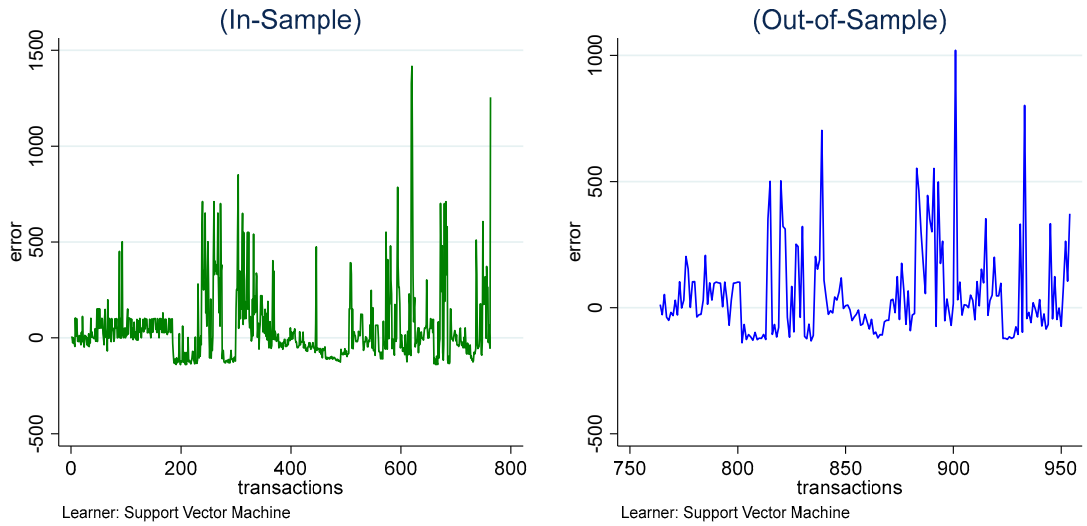

**Figure 3.** *Cont*.




(**k**) (**l**)




(**m**) (**n**)




(**o**) (**p**)

**Figure 3.** (**a**–**p**): The eight ML models' in-sample and out-of-sample performance on the formal and informal market. Note: the figure shows the prediction error defined as actual minus predicted values

regarding in-sample (left in green) and out-of-sample (right in blue) data. Each pair of diagrams relates to a specific learner. The vertical axis measures the prediction error in Tanzanian shillings (TZS 1,000,000), and the horizontal axis measures each transaction's identification number. The figure refers only to valuations made on the formal market. The results are based on the estimates in Table 3.

Most models achieve low in-sample errors in the green sections, indicating a good fit. However, the nearest neighbour model, with exceptionally low in-sample errors, suggests potential overfitting, as shown by perfect MAPE and MSE values in Table 3. The out-of-sample predictions (blue) highlight how well each model generalises to new data. The boost and forest models exhibit the lowest prediction errors. This suggests a balance between fitting the training data and adapting to unseen data.

## 5. Discussion

The differences between Table 2, which exclusively evaluates machine learning models in the formal real estate market, and Table 3, which assesses these models on informal and formal real estate markets in Dar es Salaam, show the impact of including informal market data on the machine learning model. Generally, including the informal market results in higher errors and reduced generalisation capability, highlighting the challenges associated with modelling more heterogeneous and less structured datasets.

Models on the formal market (Table 2) exhibit a lower mean absolute percentage error (MAPE) and mean squared error (MSE) when applied solely to the formal market. This suggests that the data characteristics of the formal market could be more uniform and better structured, facilitating more effective learning and prediction. Models show higher cross-validation scores, especially in the testing phase, indicating superior generalisation when trained and tested on data from the same market type (formal only).

When models are applied in both markets (as shown in Table 3), there is a tendency for the mean absolute percentage error (MAPE) and mean squared error (MSE) to go up. This indicates that including the informal market introduces complexity and variation, resulting in problems for the models. The informal market is expected to have less structured or more diverse data, leading to more prediction inaccuracies. For example, the cross-validation scores in the testing phase decrease compared to Table 2, indicating a reduction in the model's capacity to generalise when trained on a dataset using both market types. The lower overall performance can be attributable to the increased heterogeneity of the merged datasets.

Models in the formal market generally perform better in terms of predictive accuracy when limited to the formal market. This can be attributed to the more consistent and possibly better-documented transaction data in the formal market, which supports the assumptions and capabilities of machine learning models. Predictive accuracy declines when models incorporate data from both markets, as indicated by higher MAPE and MSE values. This reflects the challenges of dealing with less structured data from the informal market, which may include inconsistencies, irregular transactions, and reduced pricing transparency. Including informal market data requires that models handle a broader range of data types and relationships, which might not be as necessary when models are applied strictly to formal market data.

The random forest model seems to perform consistently well with different data sources (formal market vs. both formal and informal markets). The model is ranked second with formal transaction data and first with data from both formal and informal markets. The elastic net and regression models are also not far behind, based on how they consistently perform well with various data sources (Tables 2 and 3). This denotes that machine learning techniques could be adopted for real estate valuation even in data-limited nascent markets such as Dar es Salaam. The results in this study that show that the random forest is of significant importance are consistent with earlier studies that indicated that it is superior compared to other ML models (see, e.g., [48–50]). The superiority of the model stems from its ability to deal with high dimensional data, class imbalances, complexities

arising from non-linearity, and outliers [51]. More importantly, the model is robust against overfitting, particularly when dealing with limited data [52], as is the case with the Dar es Salaam housing market.

The results from this study have shown that data collected from the formal as well as informal markets can be utilised within the same model for price predictions from limited data samples. The study has demonstrated that the ML models have a higher predictive accuracy in the formal market. However, the results have also highlighted the need for increased caution when basing an investment decision on a market infused with information from informal agents as these do not always have well-kept data corresponding to their vast knowledge of the market. The reduced predictive accuracy of the ML models when run with data from both market segments highlights the need for the adjustment of risk assessment and pricing strategies particularly for investors with a focus on both formal and informal market segments.

A large degree of investors makes transactions in both the formal and the informal markets in Dar es Salaam. Thus, there is a clear need to bridge the gap between formal and informal markets in terms of data quality as addressing this gap will lead to an increased availability of reliable market predictions and provide a solid foundation for informed decision-making across the real estate sector. Policymakers may implement targeted policies to improve the stability and transparency of the housing market by improving informal data collection and streamlining procedures. There need to be policies to influence the awareness among informal real estate agents of the need for quality data or improved data keeping and sharing and the key benefits of house price indices.

## 6. Conclusions

The focus in this study was not so much on discussing traditional valuation methods as on the issue of how advances in technology can be harnessed within the valuation sector in an environment of limited market information. The study demonstrates that machine learning (ML) can enhance property appraisal in Dar es Salaam, Tanzania, and the results show that machine learning can be used to handle data that is both sparse and diverse. It also provides indications on the issues to address in consideration of the dual market dynamics of the formal and informal real estate sectors during the valuation process. Unlike in previous studies, eight ML models were used here and the research presented shows that machine learning techniques can provide well-needed easily managed valuation tools in emerging economies, where traditional methods often fail due to market segmentation and limited data accessibility even though the inclusion of information from the informal sector decreases performance.

The use of only formal market data would significantly reduce the data sample, whereas incorporating the informal market data enhances data volume, although with the introduction of noise due to the increased data heterogeneity. To benefit from data richness, the structure of the informal market data could be improved in various ways such as through the enforced standardisation of data collection. The informal agents lack structured record-keeping systems leading to the possession of inconsistent/noisy data. One of the solutions for standardising data collection and minimising inconsistencies is to use identical well-tailored data templates. The government of Tanzania has introduced e-government and committed itself to advancing local government service provision through structures related to smart governance [53]. However, it still has a long way to go in many sectors. A step within the housing investment sector could be to implement a template in its e-government structure that would require informal agents to provide some basic information on all transactions carried out as part of the registration of property ownership. This would also increase the volume of data that could be utilised in creating a reliable nation-wide property index, as results from this study provide proof that it is possible to increase predictive accuracy given well-documented data even from small samples.

Further research based on the results of this study could be the use of augmentation techniques to increase the sample size and to focus on evaluating the four most highly

ranked ML techniques in this study: random forest, SVM, boost and elastic net with a larger dataset. This way, the ML models are likely to have less memorisation of data features, hedging against the problem of overfitting, and thereby improving generalisability. For policy implications in regards to technological advances in this nascent market there is a need for further research on how the information from the informal sector can be captured reliably in a cost-effective way. The "Dalalis" possess a fountain of information that should not go to waste. Capturing this data can be combined with other efforts related to studies of AI technology within the real estate sector such as the development of superior real-time data for better ML-based house price indices [44]. Such indices could improve urban planning through a better allocation of resources, thereby improving the supply of Smart Governance Services (SGSs) such as affordable housing.

Advances in technology invariably raise the issue how AI-generated data are used. There are calls for regulatory measures and ethical accountability. However, the ML techniques used in this study rely on data that has already been gathered to produce results that largely are not related to a single identifiable property owner. Nevertheless, of importance to the investors, policy makers, and other stakeholders such as urban planners there is a need for further research that also looks into the ethical behaviour of the agents in this housing market to increase the reliability of the information provided as a basis for decision-making.

**Author Contributions:** Conceptualization, F.N. and M.W.; Methodology, F.N. and M.W.; Formal analysis, M.W.; Resources, F.N. and H.M.; Data curation, F.N.; Writing—original draft, F.N., H.M. and M.W.; Writing—review & editing, F.N., H.M. and M.W.; Supervision, H.M. and M.W.; Funding acquisition, H.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** SIDA (Swedish International Development Agency), Project name: Real Estate Market Dynamics and Housing Finance, Project number: 51170073/2193.

**Data Availability Statement:** The datasets presented in this article are not readily available because of copyright issues. Requests to access the datasets should be directed to Frank Nyanda.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. United Nations. *2018 Revision of World Urbanization Prospects*; United Nations Department of Economic and Social Affairs: New York City, NY, USA, 2018.
2. Centre for Affordable Housing Finance Africa. *Housing Finance in Africa Yearbook*, 13th ed.; Centre for Affordable Housing Finance Africa: Johannesburg, South Africa, 2022.
3. Centre for Affordable Housing Finance Africa. *Housing Finance Yearbook: Tanzania*; Centre for Affordable Housing Finance Africa: Johannesburg, South Africa, 2023.
4. Sanga, A. The value of formal titles to land in residential property transactions: Evidence from Kinondoni municipality Tanzania. *Int. J. Hous. Mark. Anal.* **2018**, *11*, 117–148. [CrossRef]
5. Panman, A.; Lozano Gracia, N. Titling and beyond: Evidence from Dar es Salaam, Tanzania. *Land Use Policy* **2022**, *117*, 105905. [CrossRef]
6. Andreasen, M.H.; McGranahan, G.; Steel, G.; Khan, S. Self-builder landlordism: Exploring the supply and production of private rental housing in Dar es Salaam and Mwanza. *J. Hous. Built Environ.* **2021**, *36*, 1011–1031. [CrossRef]
7. Kemwita, E.F.; Kombe, W.J.; Nguluma, H.M. Acquisition of land in flood risk informal setlements in Dar es Salaam: Choices and Compromises. *Afr. J. Land Policy Geospat. Sci.* **2023**, *6*, 188–208.
8. Komu, F. Analysis of real estate value determinants—The case of valuation practice in Tanzania. In Proceedings of the 19th Annual AfRES Conference, Arusha, Tanzania, 10–13 September 2019.
9. Huang, G.; Li, D.; Ng, S.T.; Wang, L.; Wang, T. A methodology for assessing supply-demand matching of smart government services from citizens' perspective. *Habitat Int.* **2023**, *138*, 102880. [CrossRef]
10. Huang, G.; Li, D.; Yu, L.; Yang, D.; Wang, Y. Factors affecting sustaintbility of smart city services in China: From the perspective of citizens' sense of gain. *Habitat Int.* **2022**, *128*, 102645. [CrossRef]
11. Makulilo, A.B. Analysis of the regime of systematic government access to private sector data in Tanzania. *Inf. Commun. Technol. Law* **2020**, *29*, 250–278. [CrossRef]
12. Lalika, C.; Mujahid, A.U.H.; James, M. Machine learning algorithms for the prediction of drought conditions in the Wami River sub-catchment, Tanzania. *J. Hydrol. Reg. Stud.* **2024**, *53*, 101794. [CrossRef]

13. Das, R.C.; Chatterjee, T.; Ivaldi, E. Nexus between housing price and magnitude of pollution: Evidence from the panel of some high-and-low polluting cities of the world. *Sustainability* **2022**, *14*, 9283. [CrossRef]
14. Nyanda, F. The effect of proximity and spatial dependence on the house price index for Dar es Salaam. *Int. J. Hous. Mark. Anal.* **2024**, *17*, 945–963. [CrossRef]
15. Prosise, J. *Applied Machine Learning and AI for Engineers*; O'Reilly Publishing: Newton, MA, USA, 2022.
16. Kontrimas, V.; Verikas, A. The mass appraisal of the real estate by computational intelligence. *Appl. Soft Comput.* **2011**, *11*, 443–448. [CrossRef]
17. McCluskey, W.J.; Zulkarnain Daud, D.; Kamarudin, N. Boosted regression trees: An application for the mass appraisal of residential property in Malaysia. *Financ. Manag. Prop. Constr.* **2014**, *19*, 152–167. [CrossRef]
18. Hoxha, V. *Comparative Analysis of Machine Learning Models in Predicting Housing Prices: A Case Study of Prishtina's Real Estate Market*; Emerald Publishing Limited: Bradford, UK, 2024. [CrossRef]
19. Mullainathan, S.; Spiess, J. Machine learning: An applied econometric approach. *J. Econ. Perspect.* **2017**, *31*, 87–106. [CrossRef]
20. Valier, A. Who performs better? AVMs vs. hedonic models. *J. Prop. Invest. Financ.* **2020**, *38*, 213–225. [CrossRef]
21. Teoh, E.Z.; Yau, W.C.; Ong, T.S.; Connie, T. Explainable housing price prediction with determinant analysis. *Int. J. Hous. Mark. Anal.* **2023**, *16*, 1021–1045. [CrossRef]
22. Kutasi, D.; Badics, M.C. Valuation methods for the housing market: Evidence from Budapest. *Acta Oecon.* **2016**, *66*, 527–546. [CrossRef]
23. Park, B.; Bae, J.K. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Syst. Appl.* **2015**, *42*, 2928–2934. [CrossRef]
24. Chen, J.H.; Ong, C.F.; Zheng, L.; Hsu, S.C. Forecasting spatial dynamics of the housing market using Support Vector Machine. *Int. J. Strateg. Prop. Manag.* **2017**, *21*, 273–283. [CrossRef]
25. Phan, T.D. *Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia*; IEEE Publication: New York City, NY, USA, 2019. [CrossRef]
26. Zhang, Y.; Rahman, A.; Miller, E. Longitudinal modelling of housing prices with machine learning and temporal regression. *Int. J. Hous. Mark. Anal.* **2023**, *16*, 693–715. [CrossRef]
27. Tchuente, D.; Nyawa, S. Real estate price estimation in French cities using geocoding and machine learning. *Ann. Oper. Res.* **2022**, *308*, 571–608. [CrossRef]
28. Deppner, J.; von Ahlefeldt-Dehn, B.; Beracha, E.; Schaefers, W. *Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Approach*; Springer: Berlin/Heidelberg, Germany, 2023. [CrossRef]
29. Sezer, O.B.; Gudelek, M.U.; Ozbayoglu, A.M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Appl. Soft Comput.* **2020**, *90*, 106181. [CrossRef]
30. Cerulli, G. Improving econometric prediction by machine learning. *Appl. Econ. Lett.* **2021**, *28*, 1425. [CrossRef]
31. Rampini, L.; Re Cecconi, F. Artificial intelligence algorithms to predict Italian real estate market prices. *J. Prop. Invest. Financ.* **2022**, *40*, 588–611. [CrossRef]
32. Lorenz, F.; Willwersch, J.; Cajias, M.; Fuerst, F. Interpretable machine learning for real estate market analysis. *Real Estate Econ.* **2023**, *51*, 1178–1208. [CrossRef]
33. Molnar, C. *A Guide for Making Black Box Models Explainable*; Leanpub Publishing: Victoria, BC, Canada, 2020.
34. Glumac, B.; Des Rosiers, F. Towards a taxonomy for real estate and land automated valuation systems. *J. Prop. Invest. Financ.* **2021**, *39*, 450–463. [CrossRef]
35. Lenaers, I.; Boudt, K.; De Moor, L. Predictability of Belgian residential real estate rents using tree-based ML models and IML techniques. *Int. J. Hous. Mark. Anal.* **2024**, *17*, 96–113. [CrossRef]
36. Osunsanmi, T.O.; Olawumi, T.O.; Smith, A.; Jaradat, S.; Aigbavboa, C.; Aliu, J.; Oke, A.; Ajayi, O.; Oyeyipo, O. Modelling the drivers of data science techniques for real estate professionals in the fourth industrial revolution era. *Prop. Manag.* **2024**, *42*, 310–331. [CrossRef]
37. Abidoye, R.B.; Chan, A.P.C.; Abidoye, F.A.; Oshodi, O.S. Predicting property price index using artificial intelligence techniques: Evidence from Hong Kong. *Int. J. Hous. Mark. Anal.* **2019**, *12*, 1072–1092. [CrossRef]
38. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*; Springer: Berlin/Heidelberg, Germany, 2021.
39. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: Berlin/Heidelberg, Germany, 2013. [CrossRef]
40. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
41. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Taylor and Francis: Abingdon, UK, 2017. [CrossRef]
42. Scornet, E. Trees, forests, and impurity-based variable importance in regression. *Ann. Inst. H. Poincaré Probab. Stat.* **2023**, *59*, 21–52. [CrossRef]
43. Cerulli, G. Machine learning using Stata/Python. *Stata J.* **2022**, *22*, 772–810. [CrossRef]
44. Neves, F.T.; Aparicio, M.; Neto, M.C. The impacts of open data and eXplainable AI on real estate price predictions in smart cities. *Appl. Sci.* **2024**, *14*, 2209. [CrossRef]

45. Yağmur, A.; Kayakuş, M.; Terzioğlu, M. House price prediction modeling using machine learning techniques: A comparative study. *Aestimum* **2023**, *81*, 39–51. [CrossRef]
46. Meharie, M.G.; Mengesha, W.J.; Gariy, Z.A.; Mutuku, R.N.N. Application of stacking ensemble machine learning algorithm in predicting the cost of highway construction projects. *Eng. Constr. Archit. Manag.* **2022**, *29*, 2836–2853. [CrossRef]
47. Nguyen, N.; Cripps, A. Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. *J. Real Estate Res.* **2001**, *22*, 313–336. [CrossRef]
48. Ho, W.K.; Tang, B.S.; Wong, S.W. Predicting property prices with machine learning algorithms. *J. Prop. Res.* **2021**, *38*, 48–70. [CrossRef]
49. Lohith, O.; Jha, A.; Tamboli, S.C. Comparative Analysis of Random Forest Regression for House Price Prediction. *Int. J. Creat. Res. Thoughts* **2023**, *11*, h336–h343.
50. Rico-Juan, J.R.; Taltavull de La Paz, P. Machine learning with explainability or spatial hedonic tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Syst. Appl.* **2021**, *171*, 114590. [CrossRef]
51. Han, S.; Williamson, B.D.; Fong, Y. Improving random forest predictions in small datasets from two-phase sampling designs. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 322. [CrossRef]
52. Qi, Y. Random forest for bioinformatics. In *Ensemble Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2012.
53. Pellatt, J.; Palfreman, J. Smart technology for cleaner city: A case study of Dar es Salaam, Tanzania. *GeoJournal* **2023**, *88*, 5221–5245. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.