frontiers | Frontiers in Applied Mathematics and Statistics

TYPE Original Research
PUBLISHED 13 February 2024
DOI 10.3389/fams.2024.1327376

Comparative analysis of machine learning algorithms for predicting Dubai property prices

Abdulsalam Elnaeem Balila* and Ani Bin Shabri

Department of Mathematics, Faculty of Science, University of Technology Malaysia, Skudai, Johor, Malaysia

Introduction: Predicting property prices is a crucial task in the real estate market, and machine learning algorithms offer valuable tools for accurate predictions. In this study, we introduce a comprehensive comparison of eight well-known machine learning algorithms, namely, ensemble empirical mode decomposition (EEMD)–stochastic (S) + deterministic (D)–support vector machine (EEMD-SD-SVM), support vector machine (SVM), gradient boosting, random forest, K-nearest neighbors (KNN), linear regression, artificial neural networks (ANN), and decision trees. The focus is on predicting property prices in Dubai, with the primary objective of assessing the predictive performance of these algorithms within this specific market context.

Methods: The evaluation is based on four key performance metrics: R-squared (R^2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). These metrics provide insights into prediction errors, accuracy in percentage terms, and the proportion of variance in property prices explained by independent variables. The study compares the strengths and limitations of each algorithm for predicting property prices in Dubai, highlighting scenarios where certain algorithms excel based on the nature of decision boundaries, handling complex data, capturing localized patterns, and offering interpretability.

Results: Findings from the comparative analysis shed light on the performance of each algorithm in predicting property prices in Dubai. EEMD-SD-SVM and SVM excel in scenarios requiring precise decision boundaries, while gradient boosting and random forests demonstrate robust performance with complex and noisy property price data. KNN captures localized patterns effectively, linear regression is suitable for straightforward regression tasks, ANN excels with extensive datasets, and decision trees offer interpretability in understanding factors influencing property prices.

Discussion: The study emphasizes the significance of model tuning, feature selection, and data pre-processing to enhance predictive power. Additionally, practical aspects such as computational efficiency, model interpretability, and scalability in real-world applications are discussed. The comparative analysis provides valuable guidance for stakeholders, including real estate professionals, data scientists, and stakeholders interested in selecting the most suitable machine learning algorithm for predicting property prices in Dubai, with a focus on the essential evaluation metrics of MSE, RMSE, MAPE, and R^2. This study offers insights into the applicability and performance of different machine learning algorithms for predicting property prices in Dubai. Stakeholders such as real estate agents, buyers, sellers, or investors can leverage these insights to make informed decisions in the Dubai real estate market.

KEYWORDS

machine learning algorithms, EEMD-SD-SVM, SVM, KNN, ANN, gradient boosting, random forest, linear regression

EDITED BY
Xueying Zeng,
Ocean University of China, China

REVIEWED BY
Deepak Gupta,
National Institute of Technology Arunachal Pradesh, India
Mhamed Mesfioui,
Université du Québec à Trois-Rivières, Canada

*CORRESPONDENCE
Abdulsalam Elnaeem Balila
✉ abdalsalam153@hotmail.com

RECEIVED 24 October 2023
ACCEPTED 15 January 2024
PUBLISHED 13 February 2024

CITATION
Elnaeem Balila A and Shabri AB (2024) Comparative analysis of machine learning algorithms for predicting Dubai property prices.
Front. Appl. Math. Stat. 10:1327376.
doi: 10.3389/fams.2024.1327376

1 Introduction

The Dubai real estate market has long been a focal point of international interest, characterized by its rapid growth and dynamic nature. Dubai is a dynamic and cosmopolitan city in the United Arab Emirates, characterized by its constant growth and evolving landscape. Accurate predictions of property prices in Dubai are crucial for a variety of stakeholders, including buyers, sellers, investors, and policymakers. In this context, machine learning algorithms have emerged as powerful tools for making precise predictions and informed decisions.

Accurate prediction of property prices in such a dynamic market is essential for buyers, sellers, investors, and policymakers. This study aims to provide a comprehensive comparative analysis of eight popular machine learning algorithms used for predicting Dubai property prices for 1BHK. The hospitality market recovery strengthened during 2022 as key performance indicators tracked higher than the previous year despite total visitor numbers remaining below pre-pandemic levels, and Dubai welcomed 10.1 million overnight visitors in the first 9 months of 2022. Average sales prices for residential property in Dubai increased by approximately 12% between Q3 2021 and Q3 2022 to reach AED 1,203 per sq ft. Average rents also increased by approximately 19% over the same period, rising to AED 73 per sq ft at the end of 2022. Gross yields reflect 6.1% compared to 5.7% in 2021 (1). Residential values across Dubai rose by 4.8% during Q2 2023, marking the 10th consecutive quarter of price rises. The latest increase leaves values 24% higher than Q1 2020; however, average prices remain 11% below the 2014 peak. On an annualized basis, prices are up by 17% (2).

Property insurance companies also rely on proper valuation to determine the appropriate coverage amount for insuring the property (3). Real estate development and investment in Dubai have also been made easier by creating various free zones and business-friendly regulations (4).

The algorithms under investigation include EEMD-SD-SVM, support vector machine (SVM) (5), gradient boosting (6), random forest (7), K-nearest neighbors (KNN) (8), linear regression, artificial neural networks (ANN) (9), and decision trees (10). Each of these algorithms offers unique advantages and trade-offs, making them suitable for various prediction tasks (5). A comparative study has been conducted to evaluate the effectiveness of five different machines. Machine learning algorithms perform better than traditional linear models because they are better adapted to the non-linearities of complex data such as real estate market data (11). Machine learning models predict the final output with respect to correlated attributes in the dataset (12). Machine learning algorithms outperform traditional linear models due to their superior adaptation to the nonlinear complexities inherent in datasets like real estate market data (13). Support vector regression with wavelet conjunction has been applied for modeling suspended sediment load in a river using extreme learning machine and twin support vector regression with wavelet conjunction, and the outcomes reveal that the hybrid models based on the coiflet wavelet offer good performance (14).

In this analysis, the focus is on the evaluation of these algorithms based on four critical performance metrics: R-squared (R^2) (15, 16), mean squared error (MSE) (17), root mean squared error (RMSE) (17), and mean absolute percentage error (MAPE) (18, 19). These metrics are widely accepted standards for assessing the predictive accuracy of machine learning models in regression tasks, and their application in the context of the Dubai property market will enable a robust comparison of these algorithms. Shuzlina Abdul-Rahman and others used the metrics MAE and RMSE to evaluate the performance of the ML models and found useful results (20). Nor Hamizah Zulkifley and others evaluated three machine learning algorithms using the RMSE in predicting property prices using survey data and concluded that the SVM is better than others (21). A novel selection method to evaluate the accuracy of specific predictions using the correlations between the eight accuracy metrics found R^2 and RMSE to be better than other metrics (22).

The objective of this study is to assist real estate professionals, data scientists, and stakeholders in Dubai in selecting the most appropriate machine learning algorithm for property price prediction.

The findings of this analysis will serve as a valuable resource for real estate professionals, data scientists, and other stakeholders seeking to make data-driven decisions in the Dubai property market. By examining the trade-offs between accuracy, interpretability, and computational efficiency, this study offers essential insights for selecting the most appropriate machine learning algorithm for predicting property prices in Dubai, thus facilitating more informed and strategic decision-making in this dynamic real estate landscape.

In this study, several contributions should be highlighted:

- Algorithm Selection and Comparison: This study explores and compares various machine learning algorithms commonly used for regression tasks.
- Feature Engineering: The research delves into feature selection and engineering, identifying and analyzing the most influential features or creating new features that better capture the dynamics of Dubai's real estate market. This could involve factors such as location, property size, amenities, market trends, and economic indicators.
- Data Preprocessing: Discussing the preprocessing steps undertaken to clean and prepare the data for analysis. This includes handling missing values, normalization or standardization of features, encoding categorical variables, and dealing with outliers.
- Model Evaluation Metrics: Evaluating the performance of the algorithms using appropriate metrics for regression tasks. Metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), R-squared, or others are employed to compare the predictive accuracy of different models.
- Friedman test: Used to assess if there are statistically significant differences in the ranks of the models across datasets.
- Results and Recommendations: Presenting the findings and discussing which machine learning algorithms perform better for predicting Dubai property prices based on the dataset and analysis. This study suggest swhich features or algorithms are more influential in determining property prices in the context of Dubai's real estate market.
- Limitations and Future Work: There are limitations to this study, such as data availability, quality, or constraints of the models used. This might suggest areas for future research or

improvements to the methodology employed, which can be a reconstruction of the series.

This study provides valuable insights into the applicability and performance of different machine learning algorithms for predicting property prices in the specific context of Dubai, aiding stakeholders such as real estate agents, buyers, sellers, or investors in making informed decisions.

2 Problem statement

The Dubai real estate market is renowned for its rapid growth and dynamic nature. With properties ranging from luxury apartments to upscale villas, the market is highly diversified, making accurate predictions of property prices a complex and critical task. Various stakeholders, including property buyers, sellers, investors, and policymakers, rely on precise price predictions to make informed decisions in this ever-evolving market. However, the volatility and unique characteristics of the Dubai real estate market present challenges for accurate price predictions. To address this challenge, this study aims to conduct a comprehensive comparative analysis of eight machine learning algorithms for predicting Dubai property prices.

The objective is to evaluate and compare the predictive performance of these algorithms in the specific context of the Dubai property market. By doing so, the study seeks to answer the following key questions:

1- Which machine learning algorithm provides the most accurate predictions of property prices in Dubai?
2- What insights can be gained from the comparative analysis of these algorithms to support data-driven decision-making in the Dubai real estate market?
3- What is the important feature of property prices in Dubai?

3 Methodology

3.1 Data source and variables

The dataset was extracted from an open data source (23).[1] The dataset of historical unit sales transactions contains details about all projects registered in the Dubai Land Department (DLD). The data frame contains 10 years of Dubai property prices for one bedroom (1BHK) properties, and the sample size is 247,517 observations from 25/5/2004 to 9/10/2023. The data provide information on property prices on eight variables (features) namely, the transaction group (sales, mortgages, gifts), procedure (sell, grant, lease,…, etc.), property type (unit, villa), registration type (existing properties, off-plan properties), area name (Burj Khalifa, Al Jadaf, Nadd Hessa,…, etc.), availability of parking (yes, no), area size (area size in squared meter), and unit price (target). The dataset is divided into 80% training and 20% testing subsets to test and evaluate the models (24).

3.2 Data analysis

In this comparative analysis of machine learning algorithms for predicting Dubai property prices using R-squared (R^2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) as evaluation metrics, the Friedman test is used to test the significant differences between the models, and each algorithm is applied within a similar framework for consistency. Python and R software are used for data cleaning and modeling; the Python and R codes are provided in Appendix (5).

To measure the performance of our models and make the comparison, we use the above metrics to evaluate the errors of the forecasting results, which are calculated by Eqs. (1, 2, 3, 4).

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} \left(A_t - F_t\right)^2}{n-1}} \tag{1}$$

$$MAPE = \frac{\sum_{t=1}^{n} \left|\left(A_t - F_t\right)\right|}{F_t} \tag{2}$$

$$MSE = \frac{\sum_{t=1}^{n} \left(A_t - F_t\right)^2}{n} \tag{3}$$

where: A_t is the actual values and F_t is the forecasted values, and Forecast error (F_e) = A_t–F_t

$$R^2 = 1 - \frac{\sum_{t=1}^{n} \left(yi - \hat{y}i\right)^2}{\sum_{i=1}^{n} \left(yi - \bar{y}\right)} \tag{4}$$

where

- yi is the observed value of the dependent variable.
- y_i is the predicted value of the dependent variable.
- $\bar{y}$ is the mean of the observed values of the dependent variable.
- n is the number of observations.

The R^2 value ranges from 0 to 1, where 0 indicates that the model does not explain any variance in the dependent, while 1 indicates that the model perfectly explains the variance in the dependent variable.

However, it is important to note that R^2 alone might not fully describe the goodness of fit of a model, and it is advisable to consider other evaluation metrics and context-specific information when assessing model performance. Friedman test will be used to assess if there are statistically significant differences in the ranks of the models across datasets, which is determined by Eq. (5).

$$\text{Friedman test} = \frac{12}{NK\left(K+1\right)} \sum R_i^2 - 3\text{N}\left(\text{K}+1\right) \tag{5}$$

[1] dubaipulse.gov.ae

where k = number of columns (Models), n = number of rows (blocks), and R_i = sum of the ranks.

3.3 The methodologies for each algorithm are outlined as follows

3.3.1 Support vector machine (SVM)

SVM is applied using both linear and non-linear kernels to assess its performance. Hyperparameters, such as the kernel type and regularization parameter (C), are optimized through grid search or cross-validation (5). In linear SVM, the goal is to find the hyperplane that best separates the classes. The formula for the hyperplane is given by Eq. (6):

$$w^T x + b = 0 \tag{6}$$

where w is the weight vector, x is the input vector, and b is the bias term.

3.3.2 Gradient boosting

Gradient boosting is implemented with an ensemble of decision trees, with boosting iterations and learning rates as hyperparameters. Cross-validation is used for hyperparameter tuning (6). The formula for gradient boosting involves a series of steps where multiple weak learners (often decision trees) are combined to create a strong learner. For regression, the objective function is to minimize the loss function (often mean squared error for regression) as in Eq. (7):

$$\text{Loss} = \sum_{i=1}^{N} L\left(y_i, F\left(x_i\right)\right) \tag{7}$$

where N is the number of training examples, y_i is the true target value, $F\left(x_i\right)$ is the predicted value from the model, and L is the loss function that measures the difference between the predicted and true values.

Gradient calculation:

- Calculate the negative gradient of the loss function with respect to the predicted values.
- This gradient is calculated for each data point and represents the direction and magnitude of the error.

3.3.3 Random forest

Random forest is configured with an ensemble of decision trees. The number of trees, maximum depth, and other hyperparameters are tuned (7). For classification, the output of a random forest can be represented in Eq. (8) as follow:

$$\hat{y} = mode\left(T_1\left(x\right), T_2\left(x\right), T_2\left(x\right), \ldots, T_n\left(x\right)\right) \tag{8}$$

where $\hat{y}$ is the predicted class and $T_i\left(x\right)$ represents the prediction of the ith tree for input x.

For regression, the output can be represented as:

$$\hat{y} = \frac{1}{n} \sum_{i=1}^{n} T_i\left(x\right) \tag{9}$$

where $\hat{y}$ is the predicted class and $T_i\left(x\right)$ represents the prediction of the ith tree for input x.

This ensemble approach in random forest reduces overfitting and improves predictive accuracy by combining multiple individual decision trees, each trained on different subsets of data and features.

3.3.4 K-nearest neighbors (KNN)

KNN is applied by varying the number of neighbors (k) and distance metrics. Cross-validation is used to determine the optimal k (5).

- The algorithm starts by calculating the distance between the query point (the data point you want to classify or predict) and all other points in the dataset.
- The most commonly used distance metric is the Euclidean distance, though other metrics such as Manhattan distance or cosine similarity can also be used.
- For two points p and q in an n-dimensional space, the Euclidean distance is calculated by Eq. (10):

$$\text{Distance} = \sqrt{\sum_{i=1}^{n} \left(q_i - p_i\right)^2} \tag{10}$$

where

- n is the number of dimensions (features) in the dataset.

q_i and p_i are the ith feature values of points q and p, respectively.

3.3.5 Linear regression

Linear regression is employed to create a baseline model. Feature engineering may be applied to improve its predictive performance (25). In simple linear regression, the formula to predict the target variable y from a feature x is presented in Eq. (11):

$$y = mx + b \tag{11}$$

where y is the predicted value, x is the input feature, m is the slope of the line (weight or coefficient), and b is the y*-intercept*.

3.3.6 Artificial neural networks (ANN)

ANN is implemented as a deep learning model with multiple layers and neurons. Hyperparameters include the number of hidden layers, neurons per layer, activation functions, and learning rate. Validation sets are used to select the best architecture (26). Neural networks have layers of interconnected nodes/neurons. The forward pass for a simple neural network involves matrix multiplication and activation functions. For example, the output of a single neuron with an activation function σ is calculated by Eq. (12).

$$\text{Output} = \sigma\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{12}$$

where w_i are the weights of connections, x_i are the input values, and b is the bias.

3.3.7 Decision tree

Decision trees are employed to create interpretable models. Hyperparameters such as maximum depth and splitting criteria are adjusted (27). Decision trees split the data based on feature thresholds. The splitting criterion often involves measures such as Gini impurity or entropy. For instance, the Gini impurity for a node (t) is calculated by Eq. (13).

$$Cini(t) = 1 - \sum_{i=1}^{c} p_i^2 \tag{13}$$

where c is the number of classes and p_i is the probability of class i at node t.

3.3.8 EEMD-SD-SVM

First, decompose the time series into IMFs using EEMD:

To better understand the decomposition and reconstruction approaches, consider a time series y(t) = {y (1), y (2),. .., y(T)} having T observations. To obtain the random and deterministic components, y(t) is initially decomposed using EEMD, £(·), which returns a set of IMF monotonic components, for example, $h_n(t)$, ∀n ∈ {1, 2,. .., N}, where N represents the number of IMFs extracted from y(t), plus a residue e(t). It is important to emphasize that each IMF is a time series in itself, the size of which is equal to the original. Moreover, the total number of IMFs returned by EEMD depends on the time series is displayed in formula (14).

$$£(y(t)) = \{h_1(t), h_2(t), \ldots, h_N(t), e(t)\} \tag{14}$$

Reconstruct IMFs into stochastic (S) and deterministic (D) components using PACF:

Second, calculate the partial autocorrelation function (PACF) for each IMF by the Eq. (15).

$$\rho(y_i, k) = \frac{Covariance\begin{pmatrix}[y_i | y_{(i-1)}, y_{(i-2)} \cdots y_{(i-k+1)}], \\ \left[y_{i-k} | ,y_{(i-1)} | ,y_{(i-2)} \cdots y_{(i-k+1)}\right]\end{pmatrix}}{\sigma[y_i | y_{(i-1)}, y_{(i-2)} \cdots y_{(i-k+1)}] * \\ \sigma\left[y_{i-k} | ,y_{(i-1)} | ,y_{(i-2)} \cdots y_{(i-k+1)}\right]} \tag{15}$$

Third, we look for the value of the ρ immediately below 0.95, which is used as the cutoff point. The conclusion is that hi(t), ∀i ∈ {1, 2,…, k} are considered stochastic (S_t), while all hj(t), ∀j ∈ {k + 1, k + 2,…, N} plus the residues are taken as deterministic (D_t). These two components are determined formally by Eqs. (16, 17).

$$S_t = \sum_{n=1}^{k} h_n(t) \tag{16}$$

$$D_t = \sum_{n=k+1}^{N} h_n(t) + e(t) \tag{17}$$

It is important to highlight the sum of the IMFs and the original time series residuals. Hence, by summing up the stochastic and deterministic components, the original time series are obtained, i.e., y (t) = s(t) + d(t). The EEMD-SD algorithm is summarized in the following.

ALGORITHM 1 EEMD-SD.

```
Input: Noisy time series Z(t) = {$z_1$,$z_2$, …, $z_t$};
Output: Stochastic component R(t) = {$r_1$,$r_2$,…, $r_t$};
Deterministic component D(t) = {$d_{1,}d_2$,…,$d_t$};
Start:
* Decompose Z(t) into a several of N different components
{$h_1$(t), $h_2$(t),…, $h_N$(t), r(t)} = E(Z(t));
* Define an array of length N-1, V = array{1…N-1};
* For each pair of components, perform the following analysis.
For n ϵ {1,2,3,…,N-1} do
* Define two arrays of length t
$C_n$ = array{1,2,3,…,T};
$C_{n+1}$ = array{1,2,3,…,T};
For k ϵ {1,2,3,…,N-1} do
* Calculate the PACF between the phase components
* Estimate the cutoff point
$Z = \arg {}_{i=1}^{N-1} \max\{p | v[p] < \rho\}$;
* Compute the deterministic and stochastic components
stochastic: $R_t = \sum_{n=1}^{p} h_n(t)$ and deterministic: $D_t = \sum_{n=p+1}^{N} h_n(t) + r(t)$
End.
```

Finally, after reconstruction into stochastic and deterministic components, we apply SVM on each component and sum up the prediction results to get EEMD-SD-SVM.

Common to all methodologies: Data preprocessing, including data cleaning, transformation, standardization, feature selection, and scaling, is performed to prepare the dataset for modeling. All predictions of the model are evaluated using R^2, MSE, RMSE, and MAPE.

4 Results and discussions

Log transformation is valuable for improving the performance and interpretability of machine learning models (28). Figure 1 displays the log transformation compressing the range of values. To apply the log transformation, we calculate the natural logarithm (ln) of each value (29), and the data after transformation tend to have a more symmetric, bell-shaped distribution. The original chart was negatively skewed and became closer to a normal distribution after transformation. The variance of the data is more consistent across the range, which helps stabilize the variance and make the distribution more symmetric, making the data more amenable to machine learning techniques.

Figure 2 displays the trend of average property prices over time, which can help us know how prices have changed over the years. Internationally, economic growth is expected to slow down in 2023 by

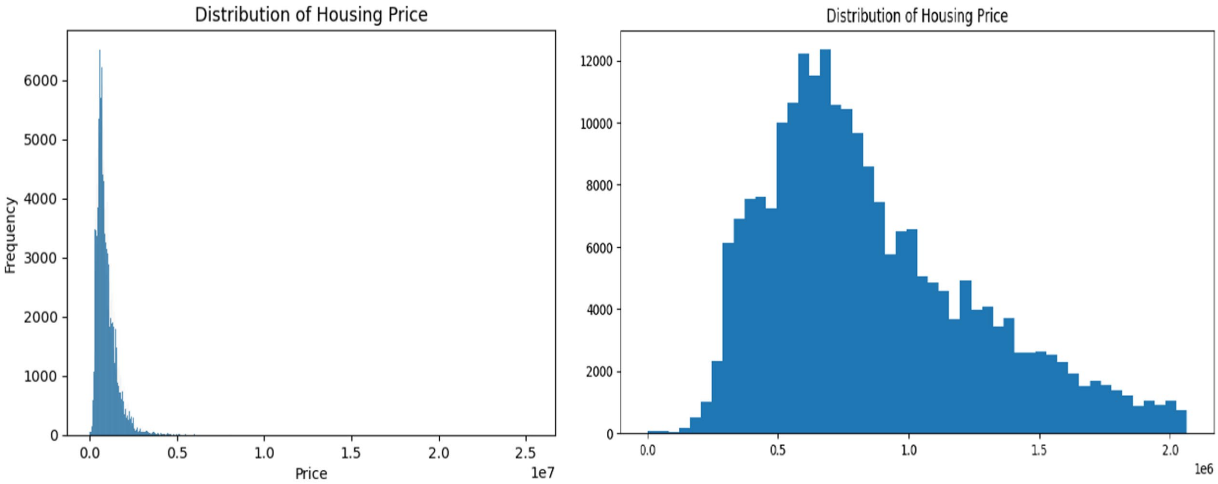


FIGURE 1
Distribution of property price in AED before and after the log transformation.

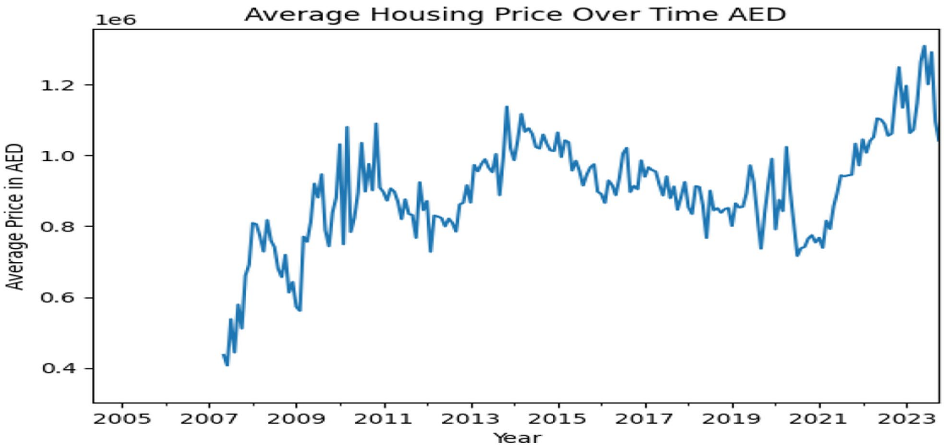


FIGURE 2
Property price average in AED over time.

the central banks of some countries to keep the inflation under control. However, Dubai's economy is expected to be strong in 2023, with growth majorly driven by steady property prices.

Figure 3 displays the data plots before and after removing outliers of property prices (sale price) and area size, which helped us address issues of outliers and extreme values. The plot, after removing transformation, becomes more normally distributed.

Figure 4 displays the pairwise correlations between the variables. In the context of predicting Dubai property prices using machine learning algorithms, a correlation matrix provides valuable insights into how the independent variables (features) are related to each other and the target variable (property prices). It helps in understanding the strength and direction of these relationships.

Figure 5 displays time series plots, which provide a clear picture of the trend, variation, and help in the interpretation process.

Table 1 displays the metrics values (R2, MSE, RMSE, and MAPE) for each model in the context of predicting Dubai property prices. R^2 is an important evaluation metric, which represents the proportion of the variance in the dependent variable (property prices) that can be explained by the independent variables (features) in the predictive model. The R^2 value ranges from 0 to 1, where $R^2 = 0$ indicates that the model explains none of the variance in the data, suggesting that it is not useful for predicting property prices; $R^2 = 1$ indicates that the model perfectly explains all the variance in the data, which is rarely achievable in real-world applications. A higher R^2 value suggests that the model's predictions are better at explaining the variance, while a lower R^2 value indicates that the model is less effective at capturing the variance (30).

In the comparative analysis of machine learning algorithms for predicting Dubai property prices, R^2 serves as an essential metric to assess how well each algorithm explains the variability in property prices. A high R^2 value indicates that the algorithm is successful in capturing the underlying patterns in the data, while a lower R^2 value suggests that there may be room for improvement in the predictive capabilities of the model. On the other hand, a lower value of MSE, RMSE, and MAPE indicates better model performance (18, 25, 31). Together, MSE, RMSE, MAPE, and R^2 provide a comprehensive view of how well each algorithm performs in the specific context of predicting Dubai property

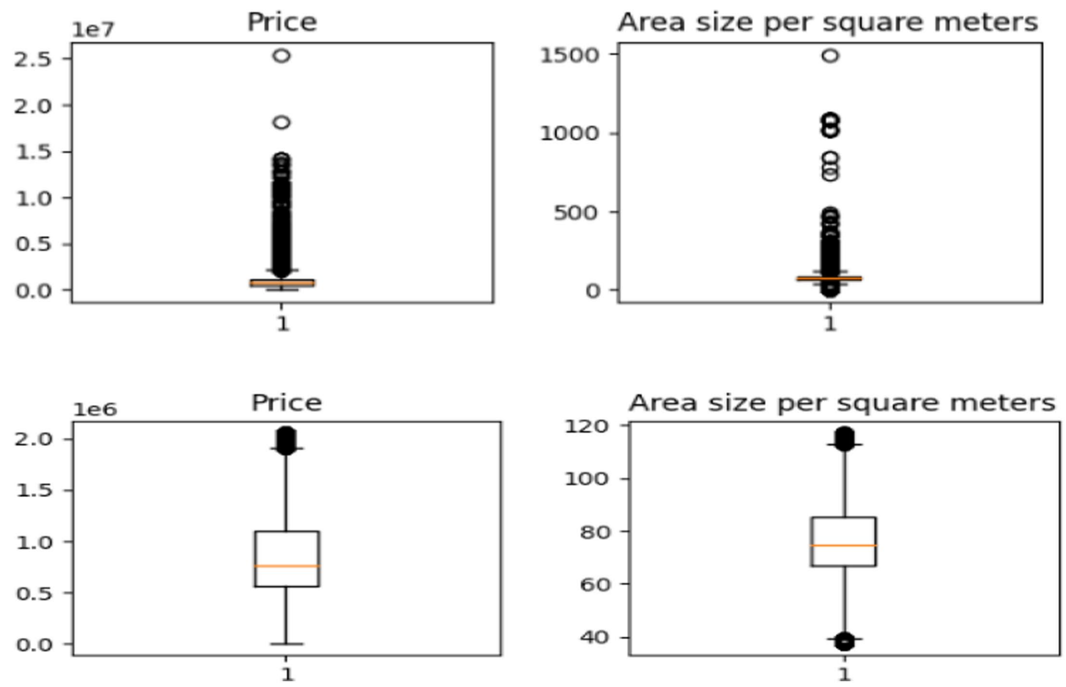
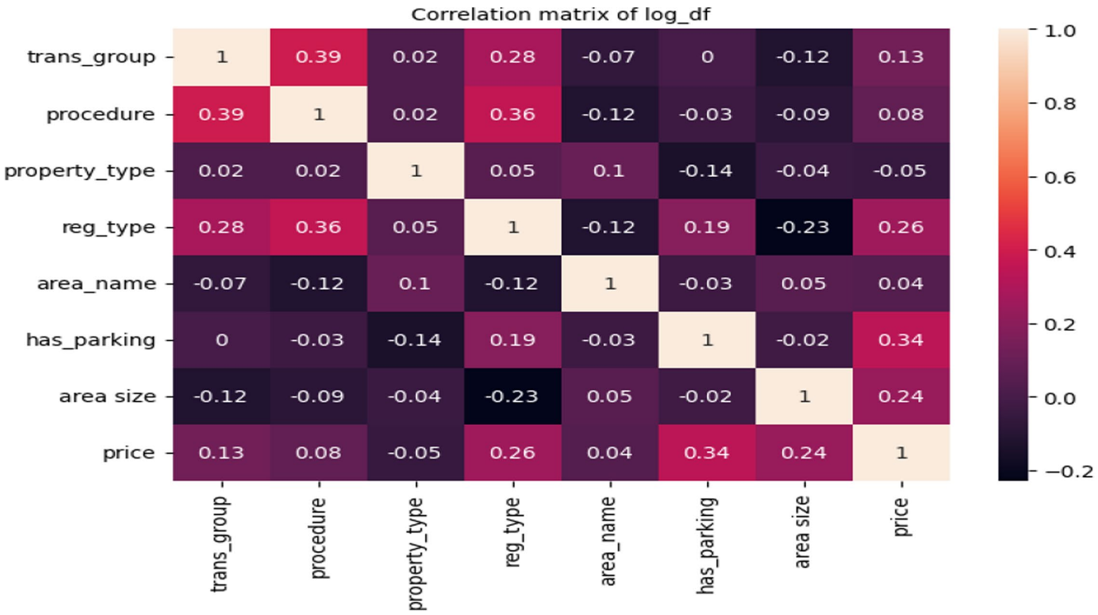


FIGURE 3
Box plot for price and area size before and after removing outliers.



	trans_group	procedure	property_type	reg_type	area_name	has_parking	area size	price
trans_group	1	0.39	0.02	0.28	-0.07	0	-0.12	0.13
procedure	0.39	1	0.02	0.36	-0.12	-0.03	-0.09	0.08
property_type	0.02	0.02	1	0.05	0.1	-0.14	-0.04	-0.05
reg_type	0.28	0.36	0.05	1	-0.12	0.19	-0.23	0.26
area_name	-0.07	-0.12	0.1	-0.12	1	-0.03	0.05	0.04
has_parking	0	-0.03	-0.14	0.19	-0.03	1	-0.02	0.34
area size	-0.12	-0.09	-0.04	-0.23	0.05	-0.02	1	0.24
price	0.13	0.08	-0.05	0.26	0.04	0.34	0.24	1

FIGURE 4
Correlation matrix.

prices, enabling informed decision-making when selecting the most suitable algorithm for this task.

4.1 Friedman test

The Friedman test is used when comparing more than two models. It assesses if there are statistically significant differences in the ranks of the models across datasets. Assumption: The data are ranked.

Table 2 displays the Friedman test, and if the value of p is less than the chosen significance level (α =0.05), we will reject the null hypothesis. This indicates that there is a significant difference in performance among the models across the data.

When comparing multiple models using RMSE, focusing on the model with the lowest RMSE, from Figure 6, we conclude that the ensemble empirical mode decomposition-support vector machine (EEMD-SD-SVM) model is the best to predict Dubai property price data and minimizing the average prediction error, and it is the most

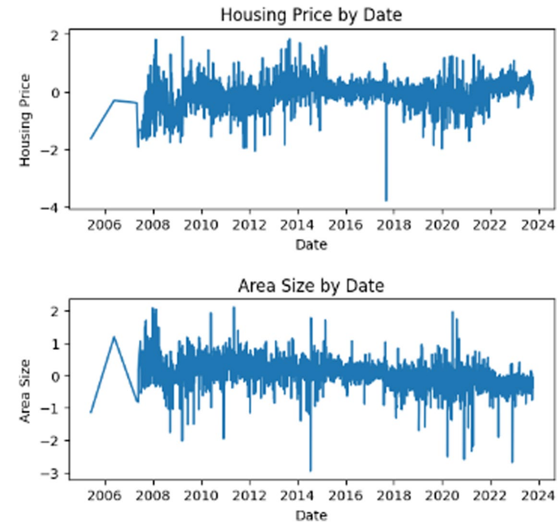


FIGURE 5
Data trend grouped by month for price and area size.

TABLE 1 Metric values of each model.

Model	R^2	MSE	RMSE	MAPE
EEMD-SD-SVM	0.541	0.090	0.3010	3.3310
Support vector machine	0.3440	0.1232	0.3510	25.1032
Gradient boost	0.3402	0.1239	0.3520	25.3897
Linear regression	0.3093	0.1297	0.3602	26.3541
Random forest	0.2952	0.1324	0.3638	26.0751
KNN	0.2569	0.1396	0.3736	27.4998
ANN	0.2559	0.1397	0.3738	27.6510
Decision tree	−0.2841	0.2412	0.4911	35.9393

SD, Stochastic and deterministic.

TABLE 2 Friedman test.

Test name	df	Test statistic	*p*-value
Friedman test	7	9.833	0.0446

accurate among the models being evaluated according to the RMSE scores; next is the support vector machine (SVM), gradient boost, linear regression, random forest, KNN, ANN, and decision tree.

Feature importance can guide feature selection or dimensionality reduction. Removing low-importance features can simplify the model while preserving predictive power. Feature importance involves understanding the relative importance of features within the context of the predictive model (32, 33). Feature selection is the process of selecting the most significant and relevant features from a vast set of features in the given dataset. So, feature selection helps in finding the smallest set of features, which results in training a machine learning algorithm faster, reducing the complexity of a model and making it easier to interpret, building a sensible model with better prediction power, and reducing over-fitting by selecting the right set of features. This can inform decision-making, model improvement, and understanding the driving factors behind the predictions and helps us uncover relationships between features and the target variable. From Figure 7 in our analysis, we can see that the area size is the most important feature, which means that the buyer will first check the area size, which is very important, and next is whether the unit has parking, area name, registration type, property type, procedure, and transaction group.

5 Conclusion

Predicting property prices is a critical challenge in the real estate market, and the choice of a suitable machine learning algorithm can significantly impact the accuracy of predictions. In this study, we conducted an in-depth comparison of eight popular machine learning algorithms with a focus on their performance in predicting property prices in Dubai using mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE), and R^2 as evaluation metrics. Our analysis has provided valuable insights into the strengths and weaknesses of each algorithm in this specific context:

EEMD-SD-SVM and SVM showed competitive performance, particularly when clear decision boundaries were crucial. It is a robust choice for scenarios where property price prediction requires a well-defined separation between data points.

Gradient boosting and random forest: These ensemble methods demonstrated exceptional predictive power and adaptability. They effectively handle complex, noisy data and tend to produce accurate property price predictions.

Linear regression: Linear regression performed well for straightforward regression tasks but may not capture the complexity of property price prediction in a dynamic market such as Dubai.

K-nearest neighbors (KNN): KNN excelled at capturing localized patterns in the property price data. It is a valuable choice when the spatial proximity of similar properties significantly influences the price.

Artificial neural networks (ANN): ANN showed promise, especially when dealing with large, diverse datasets. Its deep learning capabilities make it a suitable choice for property price prediction.

Decision trees: Decision trees provide interpretable models and insights into the factors affecting property prices, making them an ideal choice when transparency and understanding of the prediction process are vital.

The sorting of the models from the first to the last is as follows: EEMD-SD-SVM, support vector machine (SVM), gradient boosting, linear regression, random forest, K-nearest neighbors (KNN), artificial neural networks (ANN), and decision trees.

In our analysis, we emphasized the importance of model tuning, feature engineering, and data preprocessing in optimizing the predictive performance of these algorithms. Furthermore, the computational efficiency and scalability of these models were also considered as these factors are critical for practical real-world applications.

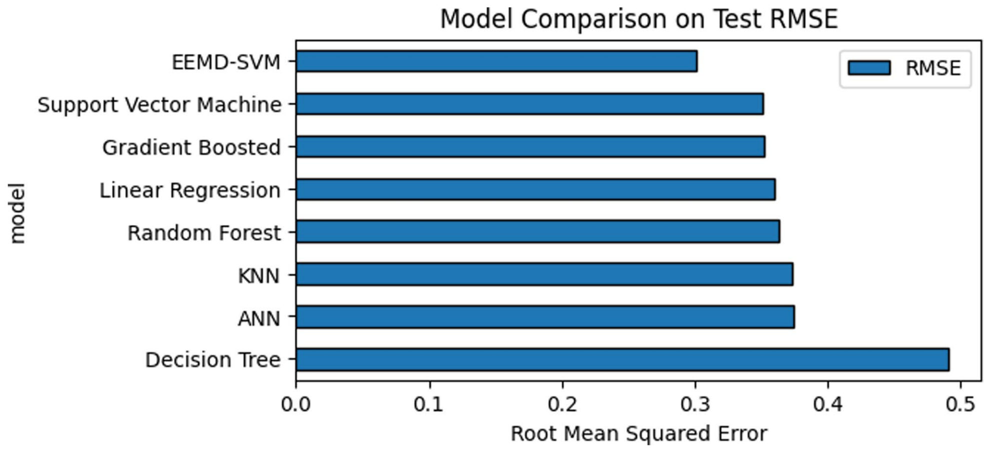


FIGURE 6
Models' comparison using the RMSE.

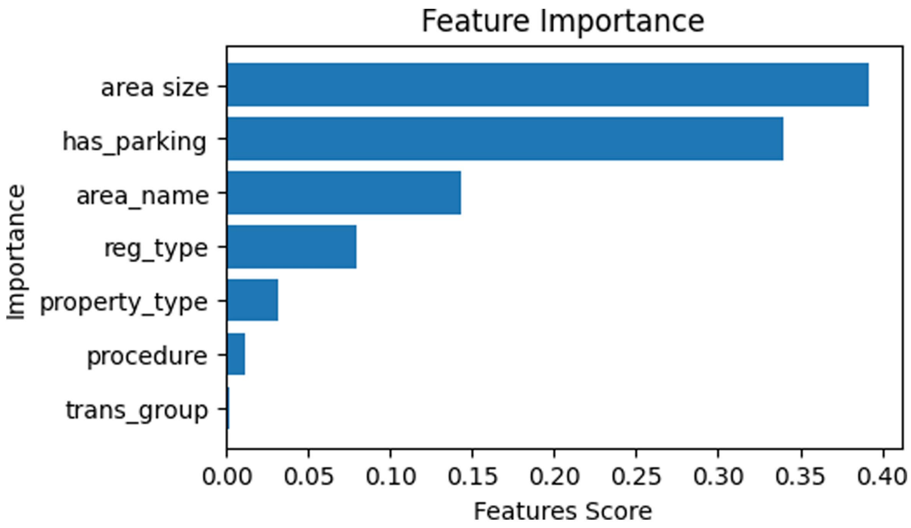


FIGURE 7
Features importance.

Selecting the most appropriate algorithm for predicting property prices in Dubai should consider not only the predictive accuracy, as indicated by R^2, MSE, RMSE, and MAPE, but also the interpretability of the model, computational resources available, and the specific characteristics of the dataset. No single algorithm is universally superior; the choice should align with the goals and requirements of the prediction task at hand.

In conclusion, this comparative analysis equips real estate professionals, data scientists, and stakeholders with the knowledge needed to make informed decisions when selecting a machine learning algorithm for predicting property prices in Dubai, using R^2, MSE, RMSE, and MAPE as critical guides for evaluating algorithm performance. By leveraging the strengths of these algorithms and acknowledging their limitations, we can develop more accurate and reliable models to navigate the dynamic Dubai property market.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.dubaipulse.gov.ae/main_search?searchword=real+estate&as_sfid=AAAAAAUqIS2timdlB54bi6Ligo9kVyKuVPunVy3mmcIDapeqri-VZLi89ERdxr87FZc2lhkVpoQk-9LqpPLG-QdIB1nHCxQXJ2mal3H-aJK7gNnCl_d8awy1vSVczmLhXC3Dfz9FJywkb3XtmHT5N399-3F4Fnq27wt9yvdNJ-VBuhRN6Q%3D%.

Author contributions

AE: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft. AS: Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fams.2024.1327376/full#supplementary-material

References

1. Deloitte. *Middle East real estate 2023 Dubai market review*. Dubai: Deloitte (2023).

2. Frank K. *Dubai residential market review*. Dubai: Knight Frank (2023).

3. Folger J. What you should know about real estate valuation. (2021). Available at: www.investopedia.com/ and https://www.investopedia.com/articles/realestate/12/real-estate-valuation.asp.

4. Dubai FDI. Dubai FDI. (2020). Available at: https://www.dubaifdi.gov.ae/why-dubai.

5. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. (1995) 20:273–97. doi: 10.1007/BF00994018

6. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. (2001) 29:1189–232. doi: 10.1214/aos/1013203451

7. Breiman L. Random forests. *Mach Learn*. (2001) 45:5–32. doi: 10.1023/A:1010933404324

8. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inform Theory*. (1967) 13:21–7. doi: 10.1109/TIT.1967.1053964

9. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*. (1943) 5:115–33. doi: 10.1007/BF02478259

10. Quinlan JR. Induction of decision trees. *Mach Learn*. (1986) 1:81–106. doi: 10.1007/BF00116251

11. Mora-Garcia R-T, Cespedes-Lopez M-F, Perez-Sanchez VR. Housing price prediction using machine learning algorithms in COVID-19 times. *Land*. (2022) 11:–2100. doi: 10.3390/land11112100

12. Kiran Kumar G, Malathi Rani D, Koppula Neeraja, Ashraf Syed. Prediction of house price using machine learning algorithms. Tirunelveli, India: IEEE. (2021). Proceedings of the 5th international conference on trends in electronics and informatics (ICOEI).

13. Truong Q, Nguyen M, Dang H, Mei B. Housing price prediction via improved machine learning techniques procedia computer science. *Proc Comput Sci*. (2020) 174:433–42. doi: 10.1016/j.procs.2020.06.111

14. Hazarika BB, Gupta D, Berlin M. Modeling suspended sediment load in a river using extreme learning machine and twin support vector regression with wavelet conjunction. *Environ Earth Sci*. (2020) 79:234. doi: 10.1007/s12665-020-08949-w

15. Devore Jay L. (2011) *Probability and statistics for engineering and the sciences*. Boston, MA: Cengage Learning. 508–510.

16. Lee M, Yoseb Y, Cheon Y, Baek S, Kim Y, Kim K, et al. Machine learning-based prediction of controlled variables of APC systems using time-series data in the petrochemical industry. *Processes*. (2023) 11:2091. doi: 10.3390/pr11072091

17. Montgomery DC, Peck EA, Vining GG. *Introduction to linear regression analysis*. Hoboken: Wiley (2012).

18. Hyndman RJ, Athanasopoulos G. *Forecasting: principles and practice*. Melbourne: OTexts (2018).

19. Willmott CJ. Some comments on the evaluation of model performance. *Bull Am Meteorol Soc*. (1983) 63:1309–13. doi: 10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2

20. Abdul-Rahman S, Zulkifley NH, Ibrahim I, Mutalib S. Advanced machine learning algorithms for house price prediction: case study in Kuala Lumpur. *Int J Adv Comput Sci Appl*. (2021) 12:736–45.

21. Zulkifley Nor Hamizah, Rahman Shuzlina Abdul, Ubaidullah Nor Hasbiah, Ibrahim Ismail (2020). House price prediction using a machine learning model: a survey of literature. Int J Modern Educ Comput Sci 12, 46–54. doi: 10.5815/ijmecs.2020.06.04

22. Jierula A, Wang S, Oh T-M, Wang P. Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. *Appl Sci*. (2021) 11:2314. doi: 10.3390/app11052314

23. dubaipulse. dubaipulse. (2023). Available at: https://www.dubaipulse.gov.ae/main_search?searchword=real+estate&as_sfid=AAAAAAUqIS2timdlB54bi6Ligo9kVyKuVPunVy3mmcIDapeqri-VZLi89ERdxr87FZc2lhkVpoQk-9LqpPLG-QdIB1nHCxQXJ2mal3H-aJK7gNnCl_d8awy1vSVczmLhXC3Dfz9FJywkb3XtmHT5N399-3F4Fnq27wt9yvdNJ-VBuhRN6Q%3D%.

24. Wang L., Wang Z., Xu W., Li G. S.L.. *The prediction of real estate land price in Dubai based on machine learning algorithms*. IEEE. (2018). 12:91–96.

25. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning* Springer (2013) 103.

26. Goodfellow I., Bengio Y., Courville A., Bengio Y. Deep learning. (2016). MIT Press: Cambridge

27. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. New York: CRC Press (1984).

28. Fuchs Caro, Spolaor Simone, Kaymak Uzay, Nobile Marco S.. (2022). *The impact of variable selection and transformation on the interpretability and accuracy of fuzzy models*. Proceedings of the IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, Ottawa, ON, Canada.

29. Stuart A, Ord KJ, Arnold S. *Kendall's advanced theory of statistics*, vol. *1*. London: Edward Arnold (1999).

30. Moore DS, Mccabe GP, Craig BA. *Introduction to the introduction to the practice of statistics practice of statistics*. New York: W. H. Freeman and Company (2009).

31. Chatfield C. *Time-series forecasting*. Boca Raton: CRC Press (2000).

32. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. (2003) 3:1157–82. doi: 10.1162/153244303322753616

33. Kuhn M., Johnson K. Applied predictive modeling. (2013). New York, NY: Springer