PROGRESS IN APPLIED SCIENCE AND TECHNOLOGY
Faculty of Science and Technology
Rajamangala University of Technology Thanyaburi
Prog Appl Sci Tech. 2024; 14(2):16-23
ISSN (Print): 2730-3012
ISSN (Online): 2730-3020
https://ph02.tci-thaijo.org/index.php/past

Research Article

Received: May 14, 2024
Revised: June 24, 2024
Accepted: June 27, 2024

DOI: 10.60101/past.2024.254102

Predicting Prices of Airbnb Accommodations in Thailand by SVM and XGBoost Methods

Sakuna Srianomai[1], Chayapat Natshivawong[2], Yuwadee Klomwises[1*] and Thanrada Chaikajonwat[1]
[1] *Department of Statistics, School of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand*
[2] *KMITL-Digital Analytics and Intelligence Center, School of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand*
*E-mail: yuwadee.kl@kmitl.ac.th

Abstract

In this study, our objective was to predict accommodation prices in Bangkok utilizing Airbnb data. The data went through necessary preparation procedures and was split into training and test sets. Both support vector machines and extreme gradient boosting methodologies were employed and optimized through hyperparameter tuning. However, the detection of overfitting necessitated a reassessment of feature selection. Several features were identified as having high importance values in both models, including the number of bedrooms, proximity to tourist destinations and landmarks in Bangkok, maximum property capacity, and the number of host listings. Additionally, support vector machines with the top 10 features outperformed other models, exhibiting the lowest mean absolute error (385.37) and root mean squared error (526.16) values. Crucially, features such as the number of bedrooms, proximity to tourist destinations, maximum property capacity, private room type, and provision of safety and facility information played significant roles. In conclusion, this study emphasizes the significance of machine learning in comprehending accommodation prices. The results highlight the importance of considering specific features, such as those identified, when setting accommodation prices.

Keywords: Airbnb Data, Accommodation Prices in Bangkok, Support Vector Machines, Extreme Gradient Boosting

1. Introduction

Tourism is considered as a major source of income for Thailand (1). In 2019, the Global Destination Cities Index from Matercard.com revealed that Bangkok, the capital of Thailand, is the No.1 destination with more than 22 million international overnight visitors (2). After the COVID-19 situation, Thai tourism has recovered once again. Nowadays the process of booking accommodations or hotels worldwide has become so simple because there are various online global platforms for booking and reservation.

Airbnb is an online marketplace that facilitates the rental of residential properties, connecting with those who are searching for accommodations. It allows hosts to list their properties, which can range from single rooms to entire homes, and helps travelers find places to stay in different locations worldwide. Airbnb provides a function called Smart Pricing, which helps hosts optimize booking rates by suggesting appropriate prices based on basic property data analysis. However, this function has been designed for hosts worldwide without specifying the unique detail of each location. According to research by (3) reveals that essential amenities in

accommodations, such as bathtub and breakfast, significantly influence pricing strategies. These comforts boost the potential for higher rates particularly in cities like Beijing (3). Additionally, the ambiance and locale of a property impact both its price and availability. Research in New York by Zhu, Li, and Xie indicate that areas near landmarks tend to command higher prices and greater demand compared to other locations (4). Moreover, many researchers have identified various factors that influence Airbnb pricing, such as accommodation type and room type, number of bedrooms, number of bathrooms, and facilities (5-7).

Therefore, this research aims to understand accommodation pricing trends and find the factors or features that impact accommodation prices in Bangkok by using data from the Airbnb website. Initially, the data was subjected to several preprocessing steps and was subsequently partitioned into training and test sets. In addition, two methods which are eXtreme Gradient Boosting (XGBoost) and Support Vector Machines (SVM) were used to predict accommodation prices and measure the effectiveness of these two techniques based on the root mean squared error and the mean absolute error. Furthermore, hyperparameter tuning was employed to optimize model performance. Moreover, the identification of the most significant feature, based on the optimal model, is also presented.

2. Methodology

In this study, we employed machine learning techniques for predicting accommodation prices. Specifically, we utilized Support Vector Machines (SVMs) and eXtreme Gradient Boosting (XGBoost) methods. Additionally, we performed hyperparameter tuning to optimize the performance of aforementioned models and evaluated their performance using various metrics. This section elaborates on the associated method as follows.

2.1 Support vector machine

Support Vector Machine (SVM) is a well-known machine learning algorithm, which is frequently employed for regression and classification problems. SVM regression is capable of capturing complex, non-linear relationships between the input features and the output variable. In regression, the goal is to find function $\boldsymbol{f(x)}$ that has at most $\boldsymbol{\varepsilon}$ deviation from the obtained target $\boldsymbol{y_i}$ across all training data points. In addition, the objective is to achieve the flattest possible function (8). The function $\boldsymbol{f(x)}$ can be expressed as:

$$f(x) = \langle \omega, x \rangle + b \text{ with } \omega \in X,\ b \in \mathbb{R}$$

where $\langle .\ ,\ . \rangle$ denotes the dot product in X. In other word, this algorithm focuses on finding the optimal hyperplane that best fits the training data, while minimizing the prediction error. This is achieved by minimizing the following loss function

$$\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-).$$

Subject to the constrains:

$$t_i \leq y_i + \varepsilon + \xi_i^+$$
$$t_i \geq y_i - \varepsilon - \xi_i^-$$
$$\xi_i^+, \xi_i^- > 0$$

where t_i is actual value,
ξ_i^+, ξ_i^- are slack variable penalties,
C is used to set the amount of regularization.

The region bound by $y_i \pm \varepsilon$ is called an ε-insensitive tube. Figure 1 illustrates the representation of the hyperplane generated by the SVM algorithm.

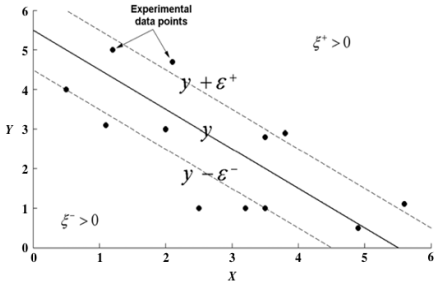

Figure 1 Hyperplane generated by the SVM algorithm (9).

2.2 Extreme gradient boosting

The XGBoost, also referred to as eXtreme Gradient Boosting, is considered to be an advanced algorithm in the field of Gradient Boosting. It is an ensemble learning technique that builds a strong predictive model by combining the predictions of multiple individual CART algorithms, typically decision trees. The prediction model of XGBoost can be expresses as follows:

$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$

where $\hat{y}_i$ is a prediction value, k is the number of decision tree, f_k denoted the k-th decision tree. The following outlines the procedure for initializing the model and sequentially incorporating decision trees.

Step 0: Initially, the model has no decision trees, and therefore, the prediction for any input is 0. This can be represented as:

$\hat{y}_i^{(0)} = 0$.

Step 1: Immediately, the first decision tree is added to the model

$\hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(x_i)$.

Step 2: The second decision tree is integrated to the model

$$\hat{y}_i^{(2)} = \hat{y}_i^{(0)} + f_1(x_i) + f_2(x_i)$$
$$= \hat{y}_i^{(1)} + f_2(x_i).$$

Step K: After updating the model t times, it becomes:

$\hat{y}_i^{(t)} = \sum_{k=1}^{K} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$,

the XGBoost can be mathematically represented as an optimization problem (10), where the objective function is formulated as follows:

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (3.1)$$

where $L(y_i, \hat{y}_i)$ is the loss function, and $\Omega(\cdot)$ represents the regularization term.

For *t*-th decision tree, the regularization term can be rewritten as

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \quad (3.2)$$

where γ is a minimum loss reduction required to make a further partition on a leaf node of the tree, T represents the number of leaves, λ is the L2 regularization term on the leaf weights, and ω denotes the score of the leaf mode. By adding *t*-th decision tree, the training error and regularization term from the preceding *t*-1 decision tree are aggregated and become constant (C). As a result, the Equation (3.1) becomes:

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \Omega(f_t) + C. \quad (3.3)$$

Therefore, from Equations (3.2) – (3.3), the objective function can be written as:

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 + C \ .$$

2.3 Hyperparameter tuning

Hyperparameter tuning is crucial for optimizing the performance of the machine learning. Effective hyperparameter tuning is essential for improving model performance and generalization capability (11). The grid search technique is one of the popular approaches for hyperparameter tuning. It involves exhaustively evaluating a predefined set of hyperparameter values and selecting the combination that gives the best performance (12). To achieve a balance between model complexity and generalization ability, we conducted grid search technique by setting a range of hyperparameter values for SVM and XGBoost methods, as outlined in Table 1 and Table 2 respectively.

Table 1 Hyperparameter ranges for grid search in SVM optimization

Hyperparameter	Values
Regularization parameter (C)	[0.1, 1, 10, 100]
Kernel coefficient (gamma)	[scale, auto]
Kernel type (kernel)	[rbf, linear]

In SVM regression, the regularization parameter, denoted as C, controls the trade-off between minimizing the error on the training data and minimizing the complexity of the decision function. The kernel coefficient (Gamma) is a parameter specific to kernelized SVM regression, particularly when using the radial basis function (RBF) kernel. Moreover, the kernel type determines the type of function used to map the original input space into a higher-dimensional space in SVM regression. Common kernel types include linear, polynomial, and radial basis functions.

Table 2 Hyperparameter ranges for grid search in XGBoost optimization

Hyperparameter	Values
weights of regularization term (alpha)	[0,1, 5, 10, 15]
subsample parameter (colsample_bytree)	[0.1, 0.3, 0.5, 0.7, 0.8, 0.9]
step size (learning_rate)	[0.001, 0.01, 0.05, 0.1, 0.2]
Maximum depth (max_depth)	[3, 5, 7, 10, 15]
Number of trees (n_estimators)	[50, 100, 200, 300, 500]

For hyperparameter setting of the XGBoost method in Table 2, alpha represents L1 regularization term on weights, colsample_bytree parameter specifies the subsample ratio of columns when constructing each tree, learning_rate is step size shrinkage used in update to prevents overfitting, n_estimators denotes maximum depth of a tree, n_estimators refer to the number of trees.

2.4 Performance matric

The model performance is assessed using two widely used evaluation metrics: the root mean squared error (RMSE) and the mean absolute error (MAE). The RMSE is a measure of the average magnitude of the errors between predicted values and actual values. Mathematically, it is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y - \hat{y})^2}$$

In addition, MAE considers the absolute differences between predicted and observed values.

$$\text{MAE} = \frac{1}{n}\sum|y - \hat{y}|$$

where n is the number of observations,
y is the actual value, and
$\hat{y}$ is the predicted value.

The lower values of RMSE and MAE indicate better model performance. This implies that the model's predictions are closer to the actual values on average.

3. Data preprocessing

The accommodation price data utilized in this study was sourced from www.insideairbnb.com, and the was updated on September 22, 2023. Prior to analysis, the data was preprocessed in a number of steps in order to make it suitable for further modelling development. The main steps can be outlined as follows.

1) Cases with null entries were eliminated in order to cope with missing values.

2) Features exhibiting high intercorrelation, exceeding an 80% threshold, were removed from the dataset in order to reduce multicollinearity.

3) The categorical column transformations for the amenities column involved a process of grouping and creating new labels based on both frequency and similarity criteria. Then, we obtained 4 groups for amenities column as shown in Table 3.

4) Values within the property type column were categorized and labeled based on their similarity, as shown in Table 4.

5) A new column named "Located near the tourist destinations and landmarks in Bangkok" was created by cross-referencing the mentioned locations in Bangkok from travel-related websites.

6) Qualitative features were standardized.

7) One-hot dummy encoding was applied to all qualitative features.

Table 3 New label for amenities column based on transformations process.

Groups	Values
Comfort & Basics	Air conditioning, dedicated workspace, long term stays allowed, essentials, bed linens, shampoo, hangers, and hot water
Appliances & Technology	Microwave, refrigerator, washer, hair dryer, iron, wifi, and TV
Safety & Facilities	Smoke alarm, fire extinguisher, free parking on premises, elevator, and kitchen
Convenience	Self check-in, cooking basics, dishes and silverware

Table 4 New label for property type column based on transformations process.

Groups	Some values
Entire units	Entire rental unit, entire townhouse, entire condo, entire loft, entire home/apt, entire guesthouse
Shared rooms	Room in aparthotel, room in hotel, room in boutique hotel, room in hostel
Private rooms	Private room in rental unit, private room in hostel, private room in bed and breakfast, private room in guesthouse
Special accommodation	Casa particular, treehouse, farm stay, barn, pension, shipping container

Moreover, the dependent variable has the right skewed as shown in Figure 2. It was then preprocessed by removing outliers and transformed using a logarithmic function prior to analysis.

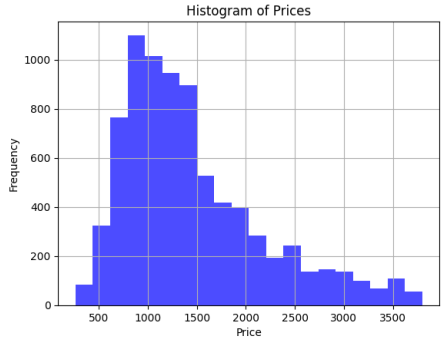

Figure 2 Histogram of accommodation prices

4. Results and Discussion

The dataset was adjusted to contain 7,874 records after preprocessing. It was split 80:20 into training and test subsets to ensure reliable model evaluation, which is a common practice in machine learning providing a good trade-off between training and validation.

In this study, we evaluate the performance of models under three distinct scenarios:

Selected features: All features with an importance value greater than zero are incorporated into the model.

Top 5 features: The model uses the top five features with the highest importance values.

Top 10 features: The model includes the top ten features with the highest importance values.

The first scenario is conducted to ensure that any feature contributing to the predictions is considered. The second and third scenarios focus on a limited number of highly influential features. Additionally, the latter two scenarios aim to simplify the model and reduce the risk of overfitting while retaining the most critical predictors. Subsequently, the RMSE and MAE metrics on the test dataset are employed to identify the optimal model.

4.1 Results of SVM

Based on optimal hyperparameter value, the SVM methodology was implemented to training dataset. Then, we can identify the main factors that influence accommodation price prediction by considering important values. As an illustration in Table 5, the features such as located near the tourist destinations and landmarks in Bangkok, the maximum capacity of the property, the number of bedrooms, and the availability of the listing 30 days in the future are concerned as significant impact on the prediction of accommodation prices.

Table 5 Top 10 feature importances derived from the SVM method

No.	Features	Value
1	Located near the tourist destinations and landmarks in Bangkok.	0.1819
2	The maximum capacity of the property	0.1748
3	The number of bedrooms	0.1724
4	The availability of the listing 30 days in the future	0.0459
5	The number of listings the host has	0.0442
6	Review Scores	0.0425
7	Minimum number of night stay	0.0349
8	Contact email	0.0337
9	The number of reviews	0.0316
10	Superhost status	0.0299

Table 6 Performance metrics of SVM method in training and test datasets

SVM	Training Data		Test Data	
	MAE	RMSE	MAE	RMSE
Selected features	296.20	437.25	361.73	506.49
Top 5 features	401.78	551.40	408.24	545.93
Top 10 features	367.97	516.23	385.37	526.16

In addition, the model evaluation derived from applying the SVM methodology to predict accommodation prices is detailed in Table 6. In the scenario of selected features, a notable indication of overfitting is also apparent. Subsequently, examination of the test data highlights the efficacy of the SVM method when employed in the top 10 features case, evidenced by the achievement of the lowest MAE and RMSE values.

4.2 Results of XGBoost

XGBoost method was executed using the best hyperparameters obtained through hyperparameter tuning. Table 7 presents the top 10 features that play crucial roles in predicting accommodation prices such as the number of bedrooms, located near the tourist destinations and landmarks in Bangkok, the maximum capacity of the property, private room type.

Table 7 Top 10 feature importances derived from the XGBoost method

No.	Features	Value
1	The number of bedrooms	0.1912
2	Located near the tourist destinations and landmarks in Bangkok.	0.1513
3	The maximum capacity of the property	0.0652
4	Private room type	0.0555
5	Providing safety and facility information	0.0403
6	Shared room type	0.0065
7	Contact phone numbers	0.0329
8	The number of listings the host has	0.0324
9	Providing a response within a day	0.0294
10	Verified identity of host	0.0278

In the process of applying the features to XGBoost, MAE and RMSE values are computed for both the training and test datasets. Moreover, this fitting summary can be found in Table 8.

Table 8 Performance metrics of XGBoost method in training and test datasets

XGBoost	Training Data		Test Data	
	MAE	RMSE	MAE	RMSE
Selected features	52.09	97.18	308.54	442.88
Top 5 features	409.08	557.78	413.62	548.52
Top 10 features	298.91	440.10	350.59	497.28

In selected features case and top 10 features case, they show a significant sign of overfitting, as evidenced by the discrepancy between the MAE and RMSE values for the training and test datasets. Consequently, a reconsideration of the number of features was performed. This led to the discovery that employing XGBoost with the top 7 features eliminated the overfitting issue. The updated results are presented in Table 9.

Table 9 Performance metrics of XGBoost method in training and test datasets

XGBoost	Training Data		Test Data	
	MAE	RMSE	MAE	RMSE
Selected features	52.09	97.18	308.54	442.88
Top 5 features	409.08	557.78	413.62	548.52
Top 7 features	403.31	551.04	410.56	544.02

Based on the result of Table 9, it can be concluded that the XGBoost method, when applied with the top 7 features exhibiting the highest values of feature importances, yields the lowest values for both MAE and RMSE.

4.3 Comparing results between SVM and XGBoost

By considering the values of MAE and RMSE, it can be concluded that SVM outperformed XGBoost. In addition, the SVM with top 10 features have the lowest value of MAE at 385.37 and RMSE at 526.16. We can

perceive that the feature that have impact on predicting accommodation prices include located near the tourist destinations and landmarks in Bangkok, the maximum capacity of the property, the number of bedrooms, the availability of the listing 30 days in the future, the number of listings the host has, review scores, minimum number of night stay, contact email, the number of reviews, and superhost status, respectively.

5. Conclusions

The objective of this paper is to identify the factor influencing accommodation prices on Airbnb in Bangkok. As the data was not initially in a format suitable for modeling, many preprocessing steps were conducted. Subsequently, the cleaned data was divided into training and test sets. For model development, the SVM regression and XGboost were applied to training data, including hyperparameter tuning step.

Feature importance values were used to select the appropriate features. The result shows only selected features with importance values greater than zero led to overfitting. We also considered models with the top 5 and top 10 features based on their importance values. For SVM regression, the overfitting issue was resolved. However, XGBoost still exhibited overfitting when using the top 10 features. Upon further consideration, we found that the overfitting problem could be eliminated by carefully selecting the top 10 features.

During hyperparameter tuning and model fitting, it's important to note that XGBoost is more prone to overfitting compared to SVM. In contrast, SVM tends to be more time-consuming during hyperparameter tuning. Additionally, the RMSE and MAE values of models without overfitting were compared. The SVM model with the top 10 features demonstrated the lowest RMSE and MAE values. Consequently, the most impactful features on price were identified based on the importance values from the SVM model with the top 10 features. These essential features include located near the tourist destinations and landmarks in Bangkok, the maximum capacity of the property, the number of bedrooms, the availability of the listing 30 days in the future, the number of listings the host has, review scores, minimum number of night stay, contact email, the number of reviews, and superhost status, respectively.

This finding will provide insight into the key factors that influence accommodation pricing strategies. Moreover, it could serve as a guideline for rental hosts in Bangkok area to optimize accommodation pricing, improve their service delivery to tourists and enhance provision for tourism business in the future.

Declaration of Conflicting Interests

The authors declared that they have no conflicts of interest in the research, authorship, and this article's publication.

References

1. Bank of Thailand. Bank of Thailand Annual Report 2022 [Internet]. Bangkok: Bank of Thailand; 2022 [cited 2024 Feb 20]. Available from: https://www.bot.or.th/content/dam/bot/documents/en/research-and-publications/reports/annual-report/AnnualReport2022.pdf
2. Mastercard. Global Destination Cities Index 2019 [Internet]. Purchase, New York: Mastercard; 2019 [cited 2024 Feb 24]. Available from: https://www.mastercard.com/news/media/wexffu4b/gdci-global-report-final-1.pdf
3. Yang S. Learning-based Airbnb Price Prediction Model. 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT). 2021. p. 283-288.
4. Zhu A, Li R, Xie Z. Machine Learning Prediction of New York Airbnb. 2020 Third International Conference on Artificial Intelligence for Industries (AI4I). 2020. p. 1-5.
5. Chen Y, Xie K. Consumer valuation of Airbnb listings: A hedonic pricing approach. Int J Contemp Hosp M. 2017;29(9):2405-24.
6. Wang D, Nicolau JL. Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. Int J Hosp Manag. 2017;62: 120-31.
7. Gibbs C, Guttentag D, Gretzel U, Morton J, Goodwill A. Pricing in the sharing economy: A hedonic pricing model applied to Airbnb listings. J Travel Tour Mark. 2018;35(1):46-56.
8. Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput. 2004;14:199-222.

9. Sánchez AS, Nieto PG, Fernández PR, del Coz Díaz J, Iglesias-Rodríguez FJ. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). Math Comput Model. 2011;54(5-6):1453-66.
10. Sharma H, Harsora H, Ogunleye B. An Optimal House Price Prediction Algorithm: XGBoost. Analytics. 2024;3(1):30-45.
11. Probst P, Boulesteix AL, Bischl B. Tunability: Importance of hyperparameters of machine learning algorithms. J Mach Learn Res. 2019;20(53):1-32.
12. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13(2):281-305.

