Particle Swarm Optimization in Machine Learning Prediction of Airbnb Hospitality Price Prediction

Suraya Masrom¹, Norhayati Baharun², Nor Faezah Mohamad Razi³, Rahayu Abdul Rahman⁴,
Abdullah Sani Abd Rahman⁵

^{1,2,3,4}Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Tapah Campus, MALAYSIA

⁵Faculty of Science and Information Technology, Universiti Teknologi PETRONAS, Perak, MALAYSIA

Abstract—Particle Swarm Optimization is a meta-heuristics algorithm widely used for optimization problems. This paper presents the research design and implementation of using Particle Swarm Optimization to automate the features selections in the machine learning models for Airbnb price prediction. Today, Airbnb is changing the business models of the hospitality industry globally. While a bigger impact has been given by the Airbnb community to the local economic development of each country, there has been very little effort that investigates on Airbnb pricing issue with machine learning techniques. Focusing on Airbnb Singapore, the main problem on the dataset is the low correlation of the independent variables to the hospitality price. Choosing the best combination of the independent variables is essential, which can be achieved through features selection optimization. Particle Swarm Optimization is useful to optimize the best variables combination for automating the features selection in machine learning models. By comparing the magnitude of change of the R squared values before and after the use of PSO feature selection, the result showed that the automated features selection has improved the results of all the machine learning algorithms mainly in the linear-based machine learning (Linear Regression, Lasso, Ridge).

Keywords—Machine Learning, Automated Features Selection, Particle Swarm Optimization, Airbnb

I. INTRODUCTION

Particle Swarm Optimization (PSO) has been used widely in many kinds of real problems. A lot of advancement has been done to improve the PSO algorithms, which were found to be effective for many application domains. Currently, PSO has also begun to be used in machine learning techniques to automate some of the processes in the machine learning pipelines[1]. The PSO inclusion for optimizing the best features selections in machine learning has also become an appealing method.

However, more extensive research is needed on the PSO in machine learning by looking into the effectiveness on different cases of problem and domain.

Therefore, this paper presents the design and implementation of PSO in the prediction of Airbnb hospitality prices.

Airbnb[2] nowadays is a trusted community commercial center for individuals to find and book their hospitality around the world. Airbnb is one of the most excellent choices for a traveler to save their expenses on lodgings and hospitality. With more than 150 million, Airbnb has built up connections with a vast number of cities in the world in a bid to offer the foremost conducive space for travelers. As an online hospitality agent like booking.com and agoda.com, Airbnb has a different business model in terms of pricing, payment, and cancellation policies.

The pricing is one of the attractive aspects of Airbnb mainly for longer reservations. Therefore, it would be an interesting study to look precisely at the methods for predicting the pricing in Airbnb. While many recent studies have gained promising outcomes with the machine learning approach, very few attempts have been made on Airbnb pricing prediction with machine learning[3]. To the best of our knowledge, there is no single study that addressed the utilization of automated features selection for the Airbnb machine learning price prediction model.

The contribution of this paper is two-fold. First, it provides a fundamental design and implementation report on the predictor models of Airbnb pricing based on machine learning algorithms. Second, it extends the state-of-the-art of machine learning with automated features selection approach. Automated features selection has been proven in many research to improve machine learning in achieving high accuracy results.

This paper is organized as follows. Section two provides a literature of background related to research that has been conducted for Airbnb and the concept of PSO. Section three presents the design and implementation of the proposed PSO features selection on Airbnb price prediction models before summarizing the conclusions in section four.

II. RESEARCH BACKGROUND

A. Research on Airbnb

Based on the various research that is available in the literature for Airbnb. There exists a very extensive literature on Airbnb. The Airbnb business model as a policy distribution has been studied by researchers in [4] to improve some of the conventional processes of business operations. Furthermore, by looking at the aspects of interfaces and multimedia on the Airbnb website, researchers have produced a soft and social presence model to enhance customers' trust in the hospitality business[5]. In [6], the architecture of the Airbnb system using global navigation satellites has been highlighted. Additionally, how users' perceptions and acceptances of the Airbnb interfaces[7] and the website contents [8] have been extensively reported. In [9], the researchers tried to investigate travelers' repurchase intention based on the psychological factors and travelers' motivation. Additionally, there exists an enormous amount of work has been carried out in predicting different aspects of Airbnb, including user destination [10], hotel review[11], neighborhood, and price [3]. Researchers in [10] and [3] employed machine learning approaches for the Airbnb predictor models. XGBOOST machine learning algorithm has been used by researchers in [12] to predict the Airbnb users' destination. By using images and text features, researchers in [3] used an SVM classifier to predict the neighborhood of a listing is in San Francisco and the listing's price. Based on the existing works on machine learning for Airbnb, there is difficult to find any research report for automated features selection of the selected machine learning algorithms that tested on Airbnb pricing. Automated features selection has been proved beneficial to simplify the tasks of machine learning design mainly for the inexpert users[13]. In [14], the researchers used the Genetic Algorithm and Ant Colony Optimization for implementing automated features selection in the prediction of biogas production. Different in [15], hybridized Genetic Algorithm with Particle Swarm Optimization has been implemented in the Support Vector Machine to detect the category of computer network intrusion. The current progress on automated features selection research has shown that automated features selection has become important in machine learning.

B. Particle Swarm Optimization Page Style

Particle Swarm Optimization (PSO) is a population-based of meta-heuristic algorithm[16] widely used in bio-inspired optimization[17]. PSO has been developed based on social behaviors of swarm flocking that moved towards different directions. In the PSO algorithm, each movement of the particle is calculated based on the current position and velocity of the particle, which is directed towards the optimization objective (maximum or minimum value). The best position that each particle can achieve at a particular iteration is called as the particle's personal fitness (*pbest*), while the best position of the whole particles in the swarm population is called as the global fitness (*gbest*). The latest global best at the end of an iteration is the optimal solution that the PSO can reach for the optimization problem. In the case of machine learning focused in this study, the optimal solution is the optimized combination of features or independent variables that are found to be suitable for the machine learning model. Figure 1 presents the flowchart of the PSO algorithm.

Figure 1 The PSO algorithm

To use PSO, one important step is to create a dataset that complies with the particle representation, which will be described in the methodology part.

III. METHODOLOGY

A. Dataset

The dataset used for testing the models was taken from Kaggle websites on Singapore Airbnb. It consists of 10 independent variables (IVs) or features that were expected to influence the hospitality price (Singapore dollar). These IVs are *host_id*, *calculated_host_listings_count*, *availability*, *reviews_per_month*, *number_of_reviews*, *neighbourhood*, *minimum_nights*, *longitude*, *latitude* and *room_type*. The hospitality *price* is the dependent variable for the predictor model. The output of Python codes to display the hospitality *price* data distribution is depicted in Figure 2.

Figure 2: Price data distribution, skewness, and kurtosis values from the Python code

It appears that the hospitality *price* is highly skewed (Skewness > 1) and a transformation like a logarithm would make it more normally distributed. Machine Learning models tend to work much better with normally distributed targets, rather than greatly skewed targets. By transforming the normality of *price*, the relationship to the IVs can be strengthened. Figure 3 shows the data distribution of hospitality *prices* labeled as *price2* after the transformation.

Skewness: 0.121051
Kurtosis: 0.648937

Figure 3: Price data distribution (label as price2), skewness and kurtosis values after normalization

The target hospitality *price2* has been moderately skewed between -0.5 and -0.5 means that the data is approximately symmetric. Figure 4 shows the Pearson correlation between the IVs to the dependent variable (*price*).

```
correlations = training.corr()
correlations = correlations["price"].sort_values(ascending=False)
features = correlations.index[1:6]
correlations
```

price	1.000000
price2	0.546197
host_id	0.045319
calculated_host_listings_count	0.035464
availability	0.016743
reviews_per_month	0.000920
number_of_reviews	-0.033523
neighbourhood	-0.033917
minimum_nights	-0.039486
longitude	-0.046606
latitude	-0.090695
room_type	-0.237742

Name: price, dtype: float64

Figure 4: Pearson correlation of IVs to price before the normalization step

International Journal of Emerging Technology and Advanced Engineering

Website: www.ijetae.com (E-ISSN 2250-2459, Scopus Indexed, ISO 9001:2008 Certified Journal, Volume 12, Issue 01, January 2022)

It can be seen that all the 10 IVs have a very weak correlation to the *price*. However, by transforming the *price* distribution (named as *price2*), a great improvement of the correlation coefficient can be seen from the *room_type* variable (from 0.23 to 0.69). Additionally, a slight improvement correlation has been presented by the rest of the nine IVs. Figure 5 shows the correlation coefficient between all the ten IVs to the *price2* (the transformed price).

```
correlations = training.corr()
correlations = correlations["price2"].sort_values(ascending=False)
features = correlations.index[1:6]
correlations
```

```
price2      1.000000
price       0.546197
calculated_host_listings_count  0.168679
host_id      0.086999
reviews_per_month  0.047137
availability  0.038934
longitude    0.021345
number_of_reviews -0.033780
minimum_nights -0.088042
neighbourhood -0.208440
latitude     -0.236000
room_type    -0.695830
Name: price2, dtype: float64
```

Figure 5: Pearson correlation of IVs to price after the normalization step

Although the strength can be improved with the *price* data transformation, most of them remain very weak with 6 out of 10 below 0.1 value of correlation coefficient. Therefore, using PSO to automate the selection of IVs is expected to be useful for the prediction model.

B. Automated features selection with PSO

Figure 6 is the particle representation for the features to be used in the PSO.

Figure 6: PSO particle representation for the Airbnb features

The dimension size of each particle is the total number of features (*n*), which in this study is ten IVs. Each particle holds the weight values of each feature a_1 to a_{10} for the dataset. The a value is bounded between 0 to 1, which only value 0.5 and above is selected. The PSO setting is given in Table I.

TABLE 1
PSO SETTING

Attribute	Value
Particles number, n	30
Particle dimension, dim	Number of features, 10
Personal learning rate, c1	0.9
Social learning rate, c2	0.1
Iterations number, i	100

Regardless of each algorithm, the number of function evaluations is the number of particles multiply the number of iterations. Therefore, for the 30 particles with 100 iterations, the total number of evaluations is 3000. It is important to identify the suitable number of particles and iterations based on the input dataset. Figure 7 depicts the features selection optimization with PSO in machine learning.

Figure 7: PSO features selection in the machine learning models

International Journal of Emerging Technology and Advanced Engineering

Website: www.ijetae.com (E-ISSN 2250-2459, Scopus Indexed, ISO 9001:2008 Certified Journal, Volume 12, Issue 01, January 2022)

The machine learning algorithms used in this research were Linear Regression, Lasso, Ridge, Decision Tree, and Random Forest. Table II lists the machine learning algorithms with the parameters setting implemented in Python codes with the Scikit-Learn library.

TABLE II
PARAMETERS SETTING OF THE MACHINE LEARNING ALGORITHMS

	Algorithm call and the parameters passing
Linear regression	<i>LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)</i>
Lasso	<i>Lasso(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=1000, normalize=False, positive=False, precompute=True, random_state=None, selection='cyclic', tol=0.0001, warm_start=False)</i>
Ridge	<i>Ridge(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=None, normalize=True, random_state=None, solver='auto', tol=0.001)</i>
Decision tree	<i>DecisionTreeRegressor(criterion='friedman_mse', max_depth=None, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=5, min_weight_fraction_leaf=0.0, presort=False, random_state=None, splitter='best')</i>
Random forest	<i>RandomForestRegressor(bootstrap=True, criterion='mae', max_depth=None, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=3, min_weight_fraction_leaf=0.0, n_estimators=15, n_jobs=1, oob_score=False, random_state=None, verbose=0, warm_start=False)</i>

Pyswarm library was used to implement the PSO algorithm. The hardware for running the machine learning models is a Lenovo notebook Intel i7 7th generation processor with 16GB RAM. Table 2 is the parameters of the machine learning algorithms written in Table 2.

C. Training Approach

Based on the 5150 number of records from the dataset, Split training and validation approach with 70:30 percentages ratio was set to all algorithms

IV. RESULTS

Firstly, Table III presents the results of each tested algorithm in predicting the price of the Singapore Airbnb dataset. In these machine learning models, PSO was not used hence all the 10 independent variables were included for the features selection.

TABLE III
RESULTS OF MACHINE LEARNING WITH ALL FEATURES SELECTION

Algorithm	R ²	RMSE
Linear Regression	0.528	0.498
Lasso	0.034	0.714
Ridge	0.563	0.480
Decision Tree	0.997	0.037
Random Forest	0.998	0.032

Without PSO features selection, Table III shows that Decision tree and Random Forest are the two outstanding algorithms compared to the other three algorithms with RMSE values less than 0.1 at R squared almost closed to 1. Lasso and Ridge that were originally introduced to handle multicollinearity problems in the dataset (as also exists in the tested Airbnb) have produced less accurate results. Lasso in this case is a little worse than Ridge. Furthermore, Table IV is the results for Airbnb prediction with PSO features selection.

International Journal of Emerging Technology and Advanced Engineering

Website: www.ijetae.com (E-ISSN 2250-2459, Scopus Indexed, ISO 9001:2008 Certified Journal, Volume 12, Issue 01, January 2022)

TABLE IV
RESULTS OF MACHINE LEARNING WITH PSO FEATURES SELECTION

Algorithm	R ²	RMSE
Linear Regression	0.752	0.190
Lasso	0.092	0.310
Ridge	0.770	0.256
Decision Tree	0.997	0.017
Random Forest	0.998	0.022

The results in Table 4 show performance benefit to the price prediction model with the PSO features selection mainly on the prediction error of all algorithms. By comparing the magnitude of change in R squared values recorded in Table 3 and Table 4, it appears in the Lasso algorithm, there was a slight improvement on the R² result as well as a great reduction of the RMSE. Similarly, the changes also appear in Linear Regression and Ridge algorithms. Nevertheless, Decision Tree and Random Forest have recorded similar accuracy with the PSO features selection which recorded the same correlation results (R²).

V. CONCLUSION

In today's competitive environment, Internet technology allows hospitality and tourism industries to become lean making their internal processes more seamless and effective at low cost. It is beneficial for hospitality websites to engage visitors longer on the site and ended with confirmed reservations. One attractive factor for the majority of the customer to make the reservation is pricing. This paper reports the design and implementation of Airbnb price reservation with machine learning tools and introduced PSO automated features selection. The results were reported based on the Singapore Airbnb dataset and can be further tested on different countries as potential future research studies.

Acknowledgment

This research was funded by the Ministry of Higher Education (FRGS/1/2019/SS01/UITM/02/36). Thank you to the Ministry of Higher Education and the Universiti Teknologi MARA.

REFERENCES

- [1] G. N. Kouziokas, "A new W-SVM kernel combining PSO-neural network transformed vector and Bayesian optimized SVM in GDP forecasting," *Eng. Appl. Artif. Intell.*, vol. 92, p. 103650, 2020.
- [2] Z. Zhang, R. J. C. Chen, L. D. Han, and L. Yang, "Key factors affecting the price of Airbnb listings: A geographically weighted approach," *Sustainability*, vol. 9, no. 9, p. 1635, 2017.
- [3] E. Tang, "Neighborhood and Price Prediction for San Francisco Airbnb Listings," 2015.
- [4] E. Biber, S. E. Light, J. B. Ruhl, and J. Salzman, "Regulating business innovation as policy disruption: From the model T to Airbnb," *Vand. L. Rev.*, vol. 70, p. 1561, 2017.
- [5] S. Ye, T. Ying, L. Zhou, and T. Wang, "Enhancing customer trust in peer-to-peer accommodation: A 'soft' strategy via social presence," *Int. J. Hosp. Manag.*, vol. 79, pp. 1–10, 2019.
- [6] P. D. Matasaru, "Smart lock system for Airbnb Platform using global navigation satellite system services," in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 572, no. 1, p. 12085.
- [7] C. R. Wang and M. Jeong, "What makes you choose Airbnb again? An examination of users' perceptions toward the website and their stay," *Int. J. Hosp. Manag.*, vol. 74, pp. 162–170, 2018.
- [8] H.-Y. Lin, M.-H. Wang, and M.-J. Wu, "A STUDY OF AIRBNB USE BEHAVIOR IN THE SHARING ECONOMY," *Int. J. Organ. Innov.*, vol. 10, no. 1, 2017.
- [9] Z. Mao and J. Lu, "Why travelers use Airbnb again?: An integrative approach to understanding travelers' repurchase intention *International Journal of Contemporary Hospitality Management* Article information:" no. June 2017, doi: 10.1108/IJCHM-08-2016-0439.
- [10] H. Ulfsson, "Predicting Airbnb user's desired travel destinations Förutsäga nya Airbnbs användares första resemål," pp. 1–27, 2017.
- [11] Y. Hu and K. Chen, "International Journal of Information Management Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings," vol. 36, pp. 929–931, 2016.
- [12] S. Avireddy, S. N. Ramamirtham, and S. S. Subramanian, "Predicting Airbnb user destination using user demographic and session information," pp. 1–5.
- [13] O. Almomani, "A feature selection model for network intrusion detection system based on PSO, GWO, FFA and GA algorithms," *Symmetry (Basel)*, vol. 12, no. 6, p. 1046, 2020.
- [14] T. Beltramo, M. Klocke, and B. Hitzmann, "Prediction of the biogas production using GA and ACO input features selection method for ANN model," *Inf. Process. Agric.*, vol. 6, no. 3, pp. 349–356, 2019.
- [15] M. Moukhafi, K. El Yassini, and S. Bri, "A novel hybrid GA and SVM with PSO feature selection for intrusion detection system," *Int. J. Adv. Sci. Res. Eng.*, vol. 4, no. 5, pp. 129–134, 2018.
- [16] S. K. Joshi and J. C. Bansal, "Parameter tuning for meta-heuristics," *Knowledge-Based Syst.*, vol. 189, p. 105094, 2020.
- [17] E. Osaba, J. Del Ser, D. Camacho, M. N. Bilbao, and X.-S. Yang, "Community detection in networks using bio-inspired optimization: Latest developments, new results and perspectives with a selection of recent meta-heuristics," *Appl. Soft Comput.*, vol. 87, p. 106010, 2020.