Property Rental Price Prediction Using the Extreme Gradient Boosting Algorithm

Marco Febriadi Kokasih [1,*], Adi Suryaputra Paramita [2]

[1,2] Program Studi Sistem Informasi, Universitas Ciputra, Indonesia
[1] marcofkokasih@gmail.com, [2] adi.suryaputra@ciputra.ac.id
[*]corresponding author

(Received August 28, 2020 Revised September 16, 2020 Accepted September 19, 2020, Available online September 30, 2020)

Abstract

Online marketplace in the field of property renting like Airbnb is growing. Many property owners have begun renting out their properties to fulfil this demand. Determining a fair price for both property owners and tourists is a challenge. Therefore, this study aims to create a software that can create a prediction model for property rent price. Variable that will be used for this study is listing feature, neighbourhood, review, date and host information. Prediction model is created based on the dataset given by the user and processed with Extreme Gradient Boosting algorithm which then will be stored in the system. The result of this study is expected to create prediction models for property rent price for property owners and tourists consideration when considering to rent a property. In conclusion, Extreme Gradient Boosting algorithm is able to create property rental price prediction with the average of RMSE of 10.86 or 13.30%.

Keywords: Rental Price; Prediction Model; Extreme Gradient Boosting; XGBoost.

1. Introduction

The online marketplace in the property rental sector is growing. One of the platforms from the online marketplace in the property rental sector is Airbnb which has more than 150 million users, 650 thousand property owners and more than 6 million properties that have been registered in 2019 [1]. With these figures, Airbnb has managed to attract attention not only to tourists as an alternative lodging place, but also to property owners as a source of additional income [12].

The more users on sites like Airbnb, the more things to consider when pricing the properties being offered [12]. Therefore, determining competitive rental rates is a challenging issue. [9]. This study aims to create a software that can be used to predict property rental prices based on the given dataset. In this study, the data used to build the model came from the Inside Airbnb project, a project that collects data from the Airbnb page. The variables that will be used in this research are house features, environment, reviews, date and property owner information, according to the variables in the data source. Then the prediction model will be made using the Extreme Gradient Boosting algorithm. The author chose this algorithm because it has been proven to have the ability to win various competitions [3].

2. Theoretical basis

2.1. Data Mining

Data mining is a step used to analyze a knowledge from a database or Knowledge Discovery in Databases [5]. Knowledge in databases can be found after going through the data cleaning process, data integration, data selection, data transformation, and data mining [4].

2.2. Data Collection

Data collection is the process of collecting data and measuring information about the variables in question in a systematic way that is well established and allows to answer the questions stated, test hypotheses and evaluate the results [11].

2.3. Data Preprocessing

Data preprocessing is needed to avoid problems that exist in data before processing such as missing data, data type errors, inconsistent data and others [10].

a. Correlation Analysis

Correlation analysis is a statistical analysis technique used to find the relationship between two variables [6]. Correlation analysis is used for feature selection process. Feature Selection is a technique used to reduce data dimensions by selecting relevant features for better learning performance [13].

There are several ways to analyze correlation, namely the theory of Pearson, Spearman and Kendall. Pearson correlation is used to measure the correlation between two continuous variables. Spearman correlation is used to assess the relationship between variables which are ordinal data [15]. While Kendall's correlation is used to measure the correlation between two variables which is ordinal data [6].

b. Data Cleaning

The definition of data cleaning is the process of preparing and selecting existing data through the analysis and processing of data which can affect the results [10]. The data cleaning process that will be carried out in this study is as follows.

1. Change data types on features.
2. Fill in the empty value with the appropriate value.
3. Deleting data that is too different from other data (outliers).
4. Remove features that are not relevant to the Machine Learning process.
5. Converts boolean values into binary numbers.

c. Data Aggregation

Data aggregation is a process in which raw data is collected and summarized in the form of statistical analysis [8].

d. Data Standardization

Data standardization is the process of giving standards to feature or attribute values so that data does not interfere with the computer learning process [7].

2.4. Extreme Gradient Boosting

Extreme Gradient Boosting, commonly known as XGBoost, is a development algorithm for Gradient Boosting that is more efficient and scalable [2]. Gradient boosting or gradient boosted tree is one of the algorithms used in solving supervised learning problems, where training data is used to predict the objective variables [3]. XGBoost provides linear model algorithms and tree learning which are efficient and capable of producing predictive models [2].

The basic model of XGBoost is a decision tree ensembles, an algorithm that contains a number of regression trees. The way this algorithm works is by adding the value of each tree for each variable so that it can generate a mathematical model like this where k is the number of trees, f is a function in functional space F and F is the set of all trees [3].

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

That way it can produce general objective functions that need to be optimized

$$\mathcal{L}(\phi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k)$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

In the Extreme Gradient Boosting algorithm, there are important feature terms that are obtained from the importance value of each feature. The more often a feature is used to make decisions in a decision tree, the higher its value. The main calculation of feature importance is the weight which indicates how important a feature is in creating a new branch. The importance of a feature changes based on how far the predictions change if a feature is replaced by another feature [14].

3. Method

The research begins with the data collection stage that will be used for research or the data collection stage. The data used came from the Inside Airbnb project, namely the Singapore Airbnb data taken on September 25, 2019. Based on the variables used in this study, the initial data was taken from the listings, listings-details and reviews-details dataset. The data is combined into one dataset by doing feature selection first. The variables contained in the dataset are property features, neighborhood, reviews, date and property owner information.

The data will go through the data preprocessing stage where the data will be prepared to be used in the machine learning process. The first stage is to combine the dataset into one through the merge process. Data that has been merged has 107 features and 104725 rows. There are several features that have no effect on predictions such as id, listing id and reviewer_id that will be deleted. Once deleted, the data will have 53 features. In addition, there is one feature that can be divided into several features such as the amenities feature. New features are created based on random selection on the value of amenities features which have a total of 85,000, namely Laptop_friendly_workspace, TV, Microwave, Dishes and Silverware, Hot_water, Family_kid_friendly.

The feature selection stage is the selection of features from the data that will be used in the study. These features are selected based on correlation. Features that have a correlation above 0.9 will be removed because they have a similar effect to the other dependent variables. The final result of the dataset is 47 features.

Data cleaning is the process of removing or replacing values from data that have null values or are outliers. In the dataset, null data comes from features that describe reviews as well as prices. Then the null data is filled with a value of 0 and other data that has null data is filled with the average of these features. In addition to filling in null data, the comments feature is changed with polarity values, all data types are changed to Float, the date feature is separated into day, month and year.

Data Standardization is the process of assigning a certain value to each feature dataset. In the dataset, there are some data that need to be changed in value first. Categorical data is converted into numbers using the Label Encoding method and then the values will be separated into new features using the One Hot Encoding method. Label Encoding is a method used to convert categorical data into integers while One Hot Encoding is used to separate categorical data with nominal properties into separate features based on its value.

Data Aggregation is the process of converting raw data into a summary form that can be used for further data analysis. The data will be analyzed to understand the contents of the data in more detail. Then the feature_importance function of XGBoost will be carried out to find out which features have the highest value for rental price predictions based on the data provided.

The data that has been prepared is divided into two data frames, namely X and Y which contain independent variables and contain dependent variables. The two dataframes are then divided into train data and test data with a ratio of 80:20 using the train_test_split function.

The next step is to create an XGBoost class and prepare the fixed parameters and parameters that will go through the hyper-parameter tuning process. Fixed parameter is the early_stopping_rounds parameter which is 10, eval_metric is RMSE and the test dataset. Parameters that are searched for through the hyper-parameter tuning process are learning_rate, max_depth, gamma, colsample_bytree and n_estimators. Hyper-parameter tuning is performed using the RandomizedSearchCV function and the results will replace the original model that was created. The model that has been filled with parameters from the hyper-parameter tuning will go through the fitting and training process.

To generate a score from the model, 10 fold cross validation was performed using the KFold function. After that, the resulting model will be saved as a pickle file along with the resulting decision tree. Researchers will use RMSE to conduct model assessments. In this study, the average RMSE value is expected not to exceed 25.

Based on the results of implementing the XGBoost algorithm on the Airbnb Singapore dataset, it was obtained a value of 0.94. This value is obtained after the model goes through the cross-validation process between the train and test data. From this model can visualize 10 features that have Feature Importance value using the highest weight calculation which can be seen in Figure 1.

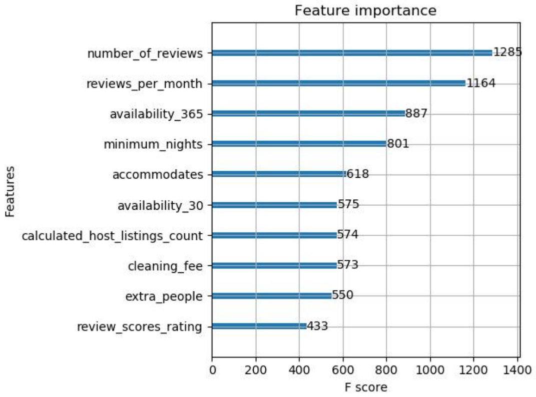

Fig. 1. Feature Importance (Top 10)

The model that has been created will be implemented into a website that has two features, namely Predict and Add Data. The Predict feature functions to predict property rental prices based on predetermined input. The Prediction Model used will be adjusted based on the location of the property. The Add Data feature serves to give users the ability to upload their own dataset to produce a prediction model according to the given dataset. The flow of the Predict and Add Data features can be seen in Figures 2 and 3.

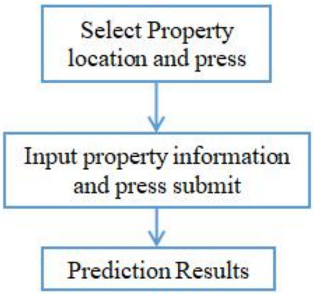
Select Property location and press
Input property information and press submit
Prediction Results

Fig. 2. Predict feature usage flow

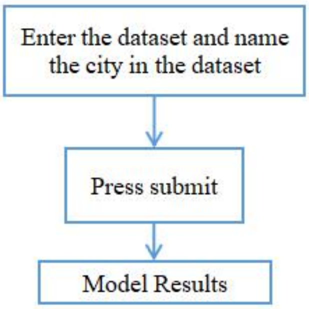

Fig. 3. Add Data feature usage flow

4. Results

Based on the model that has been created, tests are carried out in predicting rental prices using data taken from the testing dataset. The test results can be seen in Table 1.

Table. 1. Test Result

Test Number	Actual Prices	Prediction Prices	RMSE
Test 1	105	101	**3.62 (3.45%)**
Test 2	66	73	**6.99 (10.59%)**
Test 3	45	46	**1.37 (3.04%)**
Test 4	61	66	**4.77 (7.82%)**
Test 5	70	99	**28.84 (41.2%)**
Test 6	200	205	**5.08 (2.54%)**
Test 7	217	215	**2.15 (0.99%)**
Test 8	85	118	**32.72 (38.49%)**
Test 9	184	189	**5.59 (3.04%)**
Test 10	80	63	**17.43 (21.79%)**

5. Conclusion

In this paper, we have presented Property rental price prediction model Extreme Gradient Boosting algorithm is able to create property rental price prediction with the average of RMSE of 10.86 or 13.30%. The highest RMSE is 38.49% on test 8 and the lowest RMSE is 0.99% on test 7.

References

[1] G. Zervas, D. Proserpio, and J. Byers, “A First Look at Online Reputation on Airbnb, Where Every Stay is Above Average,” SSRN Electron. J., pp. 1–22, 2018, doi: 10.2139/ssrn.2554500.

[2] T. Chen and T. He, “xgboost: Extreme Gradient Boosting,” R Lect., no. 2016, pp. 1–84, 2014, doi: 10.1145/2939672.2939785>.This.

[3] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-August-2016, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[4] L. Markusheski, I. Zdravkoski, and M. Andonovski, “Data Mining Process Ljupce,” Ibaness Congr. Ser. Econ. Bus. Manag., pp. 71–79, 2019, [Online]. Available: https://www.researchgate.net/publication/332876172.

[5] T. Hendrickx, B. Cule, P. Meysman, S. Naulaerts, K. Laukens, and B. Goethals, “Mining association rules in graphs based on frequent cohesive itemsets,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9078, no. 3, pp. 637–648, 2015, doi: 10.1007/978-3-319-18032-8_50.

[6] D. R. Hardoon, S. Szedmak, and J. Shawe-taylor, “Canonical correlation analysis ; An methods,” Science (80-.)., vol. 16, no. 12, pp. 2639–64, 2003, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15516276.

[7] M. S. Gal and D. L. Rubinfeld, “Data standardization,” New York Univ. Law Rev., vol. 94, no. 4, pp. 737–770, 2019, doi: 10.2139/ssrn.3326377.

[8] P. Jesus, C. Baquero, and P. S. Almeida, “A Survey of Distributed Data Aggregation Algorithms,” IEEE Commun. Surv. Tutorials, vol. 17, no. 1, pp. 381–404, 2015, doi: 10.1109/COMST.2014.2354398.

[9] P. R. Kalehbasti, L. Nikolenko, and H. Rezaei, “Airbnb Price Prediction Using Machine Learning and Sentiment Analysis,” 2019, [Online]. Available: http://arxiv.org/abs/1907.12665.

[10] S. B. Kotsiantis and D. Kanellopoulos, “Data preprocessing for supervised leaning,” Int. J. …, vol. 1, no. 2, pp. 1–7, 2006, doi: 10.1080/02331931003692557.

[11] R. B. Davis, S. Ounpuu, D. Tyburski, and J. R. Gage, “Davis_1991.pdf,” Human Movement Science, vol. 10. pp. 575–597, 1991.

[12] E. Tang and K. Sangani, “Neighborhood and Price Prediction for San Francisco Airbnb Listings,” 2015.

[13] C. C. Aggarwal, X. Kong, Q. Gu, J. Han, and P. S. Yu, “Active learning: A survey,” Data Classif. Algorithms Appl., pp. 571–605, 2014, doi: 10.1201/b17320.

[14] H. Zheng, J. Yuan, and L. Chen, “Short-Term Load Forecasting Using EMD-LSTM neural networks with a xgboost algorithm for feature importance evaluation,” Energies, vol. 10, no. 8, 2017, doi: 10.3390/en10081168.

[15] N. H. Trang, “Limitations of Big Data Partitions Technology,” J. Appl. Data Sci., vol. 1, no. 1, pp. 11–19, 2020.