Predicting Airbnb Listing Price with Different models

Haoqian Wang*

Physics, Nankai University, Tianjin, China

*Corresponding author: 2012340@mail.nankai.edu.cn

Abstract. Airbnb is a platform company that provides and directs connections between hosts and guests. People who have an open room or a vacant space can become a host on Airbnb and make it available to the world community. Airbnb offers hosts an easy way to turn otherwise wasted space into profitable space. Therefore, it is particularly necessary for hosts to forecast and analyze the price of the houses they own. Machine learning is the science of developing algorithms and statistical models. The regression model is a predictive modeling technique in machine learning. This technique is often used to discover causal relationships between variables, predictive analysis, and time series models. In this project, our goal is to predict Boston Airbnb listing prices through a variety of machine-learning methods. This paper chose four regression models, which are the random forest regression model, linear regression model, K-nearest neighbor regression model, and Gradient Boosting regression model. With one of the best regression models, this paper obtained R-squared values of 0.6593 in training and 0.7198 in testing on the Boston dataset.

Keywords: Airbnb, machine learning, regression model, R-squared.

1. Introduction

1.1. Background

A website called Airbnb offers short-term rentals of homes or individual rooms. Travelers can use the website or their mobile devices to find and reserve unique listings all over the world. As the sharing economy has grown quickly in recent years, Airbnb has established itself as one of its leading examples. Airbnb's pricing system differs from traditional hotels in that its prices are set by the host based on their own experience. Deciding on the perfect price without losing popularity is a big challenge for new landlords. While consumers may compare similar costs, it's always beneficial to understand whether the current price is reasonable and whether it's a good time to make a reservation.

With the rapid development of computer technology, machine learning has become a hot topic of concern for all walks of life, and with the development of the times, Machine learning is already being applied to every aspect of life. Due to the unique nature of Airbnb's pricing system, the purpose of this paper is to use a variety of machine learning methods to predict prices, to help consumers determine when the price is better, and to help hosts customize an optimal price.

1.2. Related work

Airbnb is a rental company that has seen rapid growth in recent years. It has surpassed its rival inns in terms of providing ephemeral facilities for visitors, making it important to meet the needs of visitors to re-visit the place.

Due to a large amount of information in the Airbnb dataset and a large amount of information that can be mined, analyzing the Airbnb dataset has become more and more popular among scholars in recent years. Yu and Wu [1] previously tried to predict real estate values using feature significance analysis, linear regression, SVR, and random forest regression. They attempted to categorize prices into 7 groups while using Random Forest, SVC, Logistic Regression, and Naive Bayes. They reported their PCA SVC model's best RMSE of 0,53 and classification accuracy of 69% for the SVR model. Li et al. [2] introduced a Multi-Scale Affinity Propagation technique in a different publication and demonstrated how it significantly increases the accuracy of rational price predictions. Nicolau and Wang [3] analyze Airbnb listings using Quantile Regression Analysis and Normal Least Squares to explore the elements influencing prices in the sharing economy. Masiero et al. [4] used quantile regression to examine the relationship between tourist attractions, vacation properties, and hotel rates. Recently, Lewis [5] made a prediction based on machine learning and deep learning on a London property market and found that XGBoost offers the best accuracy (R2 = 0.7274), which is superior to other Kaggle competitions.

1.3. Objection

To sum up, housing price prediction has always been a hot issue. This paper aims to compare the accuracy of price prediction with different machine learning models. Firstly, appropriate data is selected for data processing and missing value filling. Then the data visualization process is carried out to observe the relationship between different reasons and house rental prices. Finally, different machine learning models are used to predict prices, and RMSE and R-squared are used to explore which model is more accurate in predicting prices in this dataset.

2. Method

2.1. Source of data

The project uses the Boston Airbnb open data set from Kaggle, which includes 3, 585 listings between 2016 and 2017. All of these datasets include a detailed table of 95 original input columns, and one of the columns is ID, which is not relevant to this study. The dataset in this paper consists of 75% of the train set and 25% of the test set.

Only features that are instructive and likely to be related to pricing are chosen as features. Therefore, functions like id and host_id that appear to be noise will be excluded.

2.1.1 Dependent variable

The dependent variable is the listing price. In order to avoid the influence of a few overpriced data on the subsequent prediction results, this paper selects the data priced between 0 and 800, which contains more than 99% of the data in the dataset.

2.1.2 Independent variables

There are many possible factors that influence the listing price. For example, the type of property, the type of rooms, the type of beds, accommodation quantity, amenities, etc. In this paper, this paper chooses 18 independent variables. These 18 independent variables are highly correlated with price and independent of each other.

2.2. Data processing

In order to make it easier to program, all variables are transformed into numeric variables during the data processing. Supplement null values for bathrooms, bedrooms, and review_scores_rating to the median value. In view of the high cost of the dataset, this paper proposed a data threshold approach to reduce this problem. This paper took out data that cost over $800 a night, and that eliminated about 1% of the total.

2.3. Machine learning models

Linear regression with a full set of features was used as a standard against which another performance was measured. After using Lasso to select a set of features, multiple machine-learning models are considered in order to find the best model. The Scikit-learn library is used to implement all models, and the results are shown in the following section.

2.3.1 Random forest regression

Random forest [6] is the act of constructing a forest at random. There are a large number of decision trees in the random forest, and there is no relationship between them. After getting the forest, when a fresh input sample is introduced, each decision tree in the forest should make a distinct determination as to which category the sample should belong to. If one class is selected the most, it is expected that the sample belongs to that class. Random forests may handle attributes with discrete and continuous values. Moreover, random forests can be utilized for unsupervised learning clustering and outlier detection.

2.3.2 Linear regression

The purpose of linear regression [7], a method of statistical analysis, is to establish the quantitative relationship between two or more variables by the application of regression analysis, which is a branch of mathematical statistics. It's approximately a linear function of characteristic x $h_\theta(x) = \theta^T x$. The goal is to minimize the loss function$J(\theta) = (y - h_\theta(x))^2$.

2.3.3 K-nearest neighbor regression

If the majority of K-samples (that is, the nearest one in the characteristic space) in the characteristic space belongs to the category, then the sample [8] also belongs to the category. This method shall specify only the category of samples to be broken down by category of one or more of the closest samples.

2.3.4 Gradient Boosting regression

Multiple weak learners are produced in sequence [9]. The objective of every weak learner is to match the negative gradient in the previous accumulation model, and then decrease the amount of the accumulated model loss in the negative gradient.

2.4. Evaluation metric

The main evaluation metrics are R-squared [10] and RMSE (root mean squared error) [11]. R-squared is used to evaluate the goodness of fit of the regression model coefficient after regression.$R^2 = \frac{SSR}{SST} = \frac{\sum(pre_y - mean_y)^2}{\sum(y - mean_y)^2} = 1 - \frac{SSE}{SST}$. R-squared ranges from zero to one. When the value of R squared is closer to 1, it indicates that the regression line provides a better fit to the observations and that the model provides more accurate results. The more closely the value of R squared approaches 0, the less well the regression line fits the data that was collected, and the less accurate the model is RMSE measures the mean of the squared errors in statistics. $RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Autual_i)^2}{N}}$. The lower the amount of the root mean square error (RMSE), the lower the average absolute error between the model's prediction and the actual value, and the more accurate the model is. The greater the value of the root mean square error (RMSE), the greater the average value of the absolute error between the predicted value and the actual value, and the less accurate the model is.

3. results and discussion

3.1. Data visualization

The Airbnb Boston dataset contains the following three datasets.

Listings: including a detailed description and overall rating of the room type

Reviews: including a unique ID for each tenant's review and detailed comments.

Calendar: including the price and listing id as well as availability for that day.

To understand the distribution of data sets before model prediction analysis, this section first analyzes the distribution of dependent variables and the characteristics of independent variables. Then, the distribution of variables is analyzed through the graph, the specific description is as follows. Table 1 shows the distribution of house rental prices. There are a total of 3585 sets of data, among which the median house rental price is 173.92. It can see from the table that 75% of the prices are within 220, while the highest price is 4000, indicating that the following data are very discrete. In order to avoid the influence of some discrete data, In this paper, data in the price range of 0-800 is selected for research, but the number of data above 800 only accounts for less than 1%, which can be ignored.

Table 1. statistics of the dependent variable

	Price
count	3585.000000
mean	173.925802
std	148.331321
min	10.000000
25%	85.000000
50%	150.000000
75%	220.000000
max	4000.000000

Table 2 shows the independent variables used in this article. In the process of machine learning, 18 independent variables are used in this paper to predict the price. These 18 variables are easy to understand and easy to process data, and the results are clear, and highly correlated with prices which are very suitable for the subsequent work of machine learning model price prediction.

Table 2. descriptions of variables

Variable	Category	Description
Host_is_superhost	int	Whether the host is an experienced host with high scores
host_identity_verified	int	Whether the homestay owner's identity is verified
Host_has_profile_pic	int	whether the host has a profile picture
Is_location_exact	int	Whether the website has the exact location of the host
Requires_license	int	Whether the host has a permit license
Instant_bookable	int	Is the room available for immediate reservation
Require_guest_profile_picture	int	Whether the guest profile picture is required
Require_guest_phone_verification	int	Whether the guest phone verification is required
Security_deposit	float	Do guests need to pay a security deposit
Cleaning_fee	float	Do guests need to pay cleaning fee
Host_listings_count	int	Guests expect a certain level of cleanliness and are willing to pay more for a perfect stay of a high standard
Minimum_nights	int	minimum stay policies
Bathrooms	float	Bathroom quantity
bedrooms	int	Bedroom quantity
Guests_included	int	*Discrete values used to estimate cost per person*
Number_of_reviews	int	The number of the host reviews
Review_scores_rating	float	The rating of the host by guests
price	float	The price of the B&B for one night

Monthly price changes of Airbnb properties in Boston, USA, from September 2016 to September 2017. In Fig. 1, the horizontal axis represents prices and the vertical axis represents months. From the chart above, Airbnb's prices are highest in September and October. However, prices are lowest in January and February.

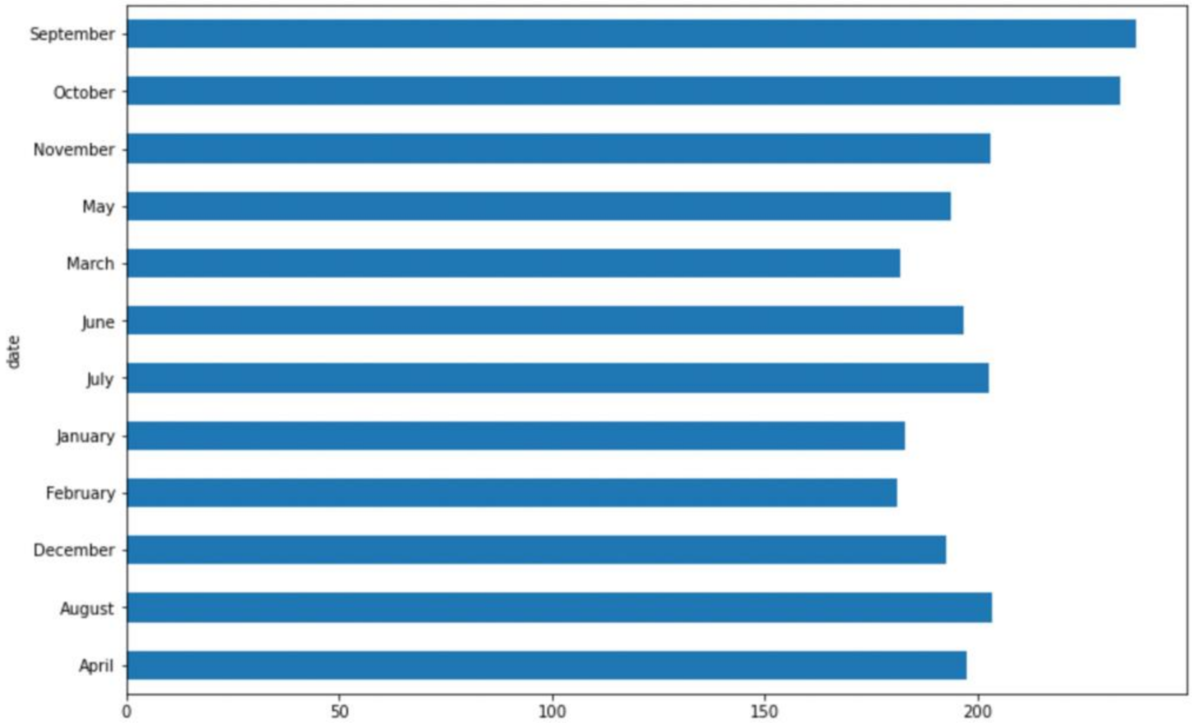

Fig. 1 Bar chart of price movement by month (Photo credit: Original)

Fig. 2 shows the relationship between the day of the week and the price. As can be seen from the chart, the price of the room is very low before Wednesday, and after Wednesday, the price rises sharply and peaks on Saturday. This paper analyzes the reason may be that people have no time to take a vacation on weekdays, so few people rent rooms and the room price is very low. On weekends, the number of people who take a vacation suddenly increases, so the number of people who rent rooms also increases, resulting in a sharp rise in the room price.

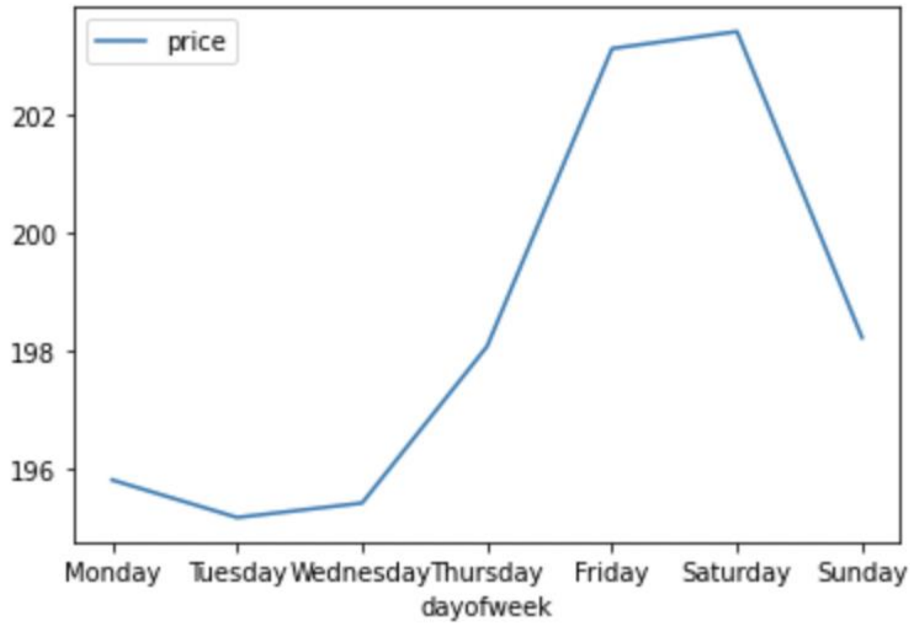

Fig. 2 Line chart of price movement by day of the week (Photo credit: Original)

Fig. 3 provides us with the relationship between the number of accommodations and room prices in the Airbnb listings in Boston, USA. As can be seen from the chart that follows, the price of the house goes up proportionally with the number of rooms available, and when there are 12 rooms available, the price of the room is at its highest.

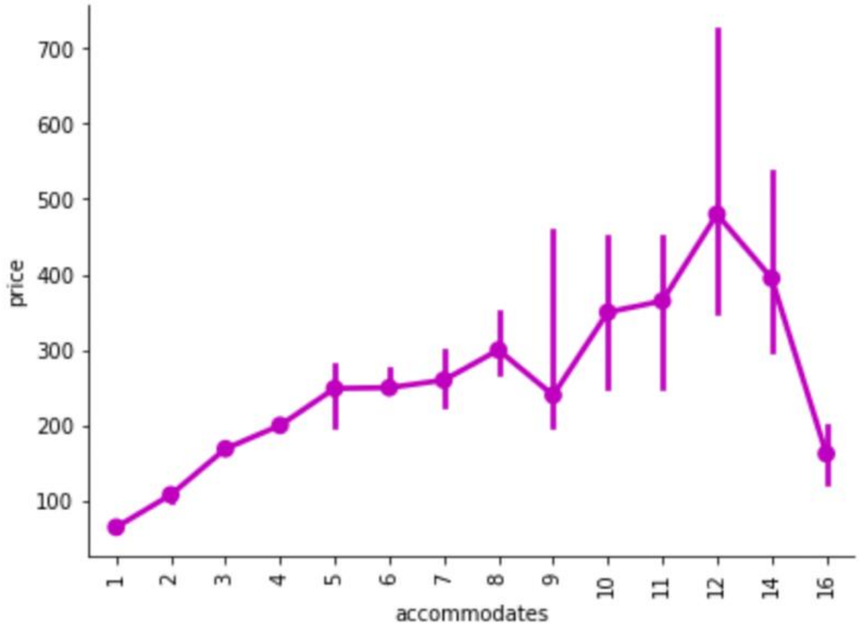

Fig. 3 The relationship between accommodates and price(Photo credit: Original)

3.2. Results

In each model, the data set was divided into test set and training set data. The training data and test data were divided into 75% and 25% respectively. Table 3 shows the accuracy of different regression models in predicting prices. In this paper, the accuracy of the model is judged by comparing the value of R-squared. It is clear from the table that the R-squared value of the Gradient Boosting Regression model's training set is 0.7164, whereas the R-squared value of the model's test set is 0.6593. Among all models, the Gradient Boosting Regression model is the most accurate prediction tool since its R-squared value is the closest to 1. The results indicate that the Gradient Boosting Regression model is the one that has the maximum accuracy when it comes to predicting prices using the dataset that was chosen. However, the R-squared value of the training set and the test set of KNN are 0.5420 and 0.3360, respectively. The K-nearest Neighbor Regression model has the lowest R-square value among all of the regression models, which indicates that the K-nearest Neighbor Regression model is not appropriate for regression prediction using this dataset.

The accuracy of the first two models differs greatly from one another to a significant degree, although the accuracy of the random forest regression model and the linear regression model is, for the most part, comparable. The R-squared values of the two models are identical in the training set, while the difference between the R-squared values of the two models is only 0.0001 in the test set. The R-squared values of the two models are identical in the test set.

Table 3. the RMSE and R-squared of each regression model

Different models	Test		Train	
	RMSE	R-squared	RMSE	R-squared
Random forest	66.523	0.6256	67.856	0.6151
Linear regression	66.438	0.6257	67.856	0.6151
K-nearest neighbor regression	84.887	0.3360	79.871	0.5420
Gradient Boosting regression	62.430	0.6593	57.888	0.7198
Random forest	66.523	0.6256	67.856	0.6151

3.3. Discussion

The findings of this study have to be seen in the light of the following limitations. Firstly, The selected data set has limitations. This paper choose is Airbnb data from 2016 to 2017 in Boston. The data set is too far away from today. Due to the changes in the global economic situation in recent years and the outbreak of COVID-19 in 2020, it is of little reference value to use the data from 2017 for forecasting at present. The newer data have not yet been collected and published. Besides, this paper only uses the data of Boston, which is not of great u generalisability. Then, In the process of data visualization, there are also some deficiencies. Due to a large amount of data, in this paper, not all the data with a high price correlation coefficient are visualized, but only a few representative data are selected for visualization. Finally, when the regression model is used to predict the price, only four machine learning models are selected to predict the price, and it is not certain that the Gradient Boosting Regression model chosen in this paper is the regression model with the highest accuracy. During model training, the model parameters used in this paper are default parameters, which cannot represent the optimal training results of the model. In future work, how to improve the accuracy of the regression model and finding a more accurate machine model will be the focus of work.

4. Conclusion

The analysis in this paper shows visualizes the data from the Airbnb Boston dataset and uses machine learning models to predict prices and compares which model is more accurate.

In the data visualization work, this paper explores the relationship between the month and the number of rental guests, and the room price. By looking at the chart, we find that the prices of rooms in September and October are higher than those in other months, while the prices in January and February are much lower. This paper speculates that the reason for this phenomenon is that in September and October, Boston has a pleasant climate and tourists have enough holidays, so more people rent rooms, so the price goes up; while in January and February, the cold weather and the off-season tourism, fewer people rent rooms for short periods, so the price goes down. In the line chart observing the relationship between the number of accommodations and the room price, we find that when the number of tenants is 12, the room price is the highest.

When using regression models for price prediction, this paper chooses four regression models (random forest regression model, linear regression model, K-nearest neighbor regression model, and Gradient Boosting regression model) and uses RMSE and R-squared to judge the accuracy of the models. The results showed that the accuracy of the Gradient Boosting Regression model was the highest, with the R-squared value of the training set reaching more than 0.7 and the R-squared value of the test set reaching more than 0.65.

In future work, researchers will do more work on price forecasting to help consumers choose a better room for them and help hosts choose a better and appropriate room price.

The research will make the R-square value of the model prediction result closer to 1 through parameter adjustment and other methods. Moreover, based on the four models selected in this paper, more suitable machine-learning models will be selected to improve the accuracy of the model. The research will also learn relevant knowledge of neural networks, and use neural networks to analyze users' preferences, demands, and evaluation of various room types, facilities, and other aspects, so as to more accurately predict the price and push their favorite rooms to users.

References

[1] H. Yu and J. Wu, “Real estate price prediction with regression and classification,” CS229 (Machine Learning) Final Project Reports, 2016.

[2] Y. Ma, Z. Zhang, A. Ihler, and B. Pan, “Estimating warehouse rental price using machine learning techniques.,” International Journal of Computers, Communications & Control, vol. 13, no. 2, 2018.

[3] D. Wang and J. L. Nicolau, “Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb. com,” International Journal of Hospitality Management, vol. 62, pp. 120–131, 2017.

[4] Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, and Hoormazd Rezaei. Airbnb price prediction using machine learning and sentiment analysis. arXiv preprint arXiv:1907.12665, 2019.

[5] Laura Lewis. Predicting airbnb prices with machine learning and deep learning, 2019.

[6] Segal M R. Machine learning benchmarks and random forest regression[J]. 2004.

[7] Seber G A F, Lee A J. Linear regression analysis[M]. John Wiley & Sons, 2003.

[8] Burba F, Ferraty F, Vieu P. k-Nearest Neighbour method in functional nonparametric regression[J]. Journal of Nonparametric Statistics, 2009, 21(4): 453-469.

[9] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001:

[10] Kasuya E. On the use of r and r squared in correlation and regression[R]. Hoboken, USA: John Wiley & Sons, Inc., 2019.

[11] Chai T, Draxler R R. Root mean square error (RMSE) or mean absolute error (MAE)[J]. Geoscientific Model Development Discussions, 2014, 7(1): 1525-1534. Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.