Airbnb Short-term Housing Rental Status Prediction Model Under the Impact of the COVID-19 Pandemic

Zhixiang Lu*

School of Mathematics and Science, Leshan Normal University, 614000 Leshan, China

Abstract. With the vigorous development of the sharing economy, the short-term rental industry has also spawned many emerging industries that belong to the sharing economy. However, due to the impact of the COVID-19 pandemic in 2020, many sharing economy industries, including the short-term housing leasing industry, have been affected. This study takes the rental information of 1,004 short-term rental houses in New York in April 2020 as an example, through machine learning and quantitative analysis, we conducted statistical and visual analysis on the impact of different factors on the housing rental status. This project is based on the machine learning model to predict the changes in the rental status of the house on the time series. The results show that the prediction accuracy of the random forest model has reached more than 94%, and the prediction accuracy of the logistic model has reached more than 74%. At the same time, we have further explored the impact of time span differences and regional differences on the housing rental status.

1 Introduction

The development of the Internet-based industry has always been in a hot area where emerging industries are constantly flowing, especially those related to the sharing economy, are growing more rapidly[1]. However, too fast development often brings many unavoidable potential problems. Utilizing the powerful computing power of modern computers and combining statistical methods such as machine learning to predict the future state of the studied variables, we can take measures in advance to prevent or even avoid problems. In this paper, I obtained more than 50 kinds of information on 1003 rental houses on Airbnb in New York in April 2020. After processing the data, the top 15 variables that have the most influential variables on the state of the house (Block / unblock) were retained and the state of the house was predicted based on the logistic model. At the same time, We also observed the impact of time and region on the state of housing rental through visual analysis.

2 Data processing and analysis

This project collects more than 50 types of information (variables) about 1003 rental houses on Airbnb in New York, USA in 2020 through the Internet, excavates the corresponding important characteristics and establishes the corresponding rental status prediction model. After cleaning and preprocessing the data, the feature engineering is used to extract the top 15 variables (rental price, price level, proportion of room photos, total number of photos, number of days of different status last year, house habitable population, cleaning cost of house, average annual rental income last year, rented status ratio in last year, number of bookings last year, total number of comments, the number of bedrooms, the number of bathrooms).

2.1 Variable description

We collect and combine data from various sources including the Airbnb website, and deleted the variables with high missing values ratio, then judged by subjective experience to retain 15 variables. Finally, there were 10 significant variables through multiple linear regression. The contribution of each variable to the R-squared of the model was Evaluate the impact of different variables on the model.

Table 1. Variable description.

Variables	Description
Status	*Housing status (available/ block)*
Price	*Rent price (per month)*
Price norm	*Regional price average*
Photo room ratio	*Ratio of room photos to total photos*
Number of Photos	*Number of total photos*
Count Reservation Days LTM	*Number of days reserved last 12 months*
Count Available Days LTM	*Number of days available last 12 months*
Count Blocked Days LTM	*Number of days blocked last 12 months*
Max Guests	*Habitable population*
Cleaning Fee	*Cleaning cost (per month)*
Annual Revenue LTM	*Average annual rental income*
Occupancy Rate LTM	*Rented status ratio in last 12 months*

* Corresponding author: luzhixiang1998@gmail.com

Number of Bookings LTM	*Number of Bookings last 12 months*
Number of Reviews	*Number of Reviews*
Bedrooms	*Number of bedrooms*
Bathrooms	*Number of bathrooms*

We first converted the three status variables of A, B, and R (A: available B: block R: reserved) into binary classification variables of 1 (Block) and 0 (Unblock) to facilitate the application of the logistic model. The basic information of these 15 variables do not give us a intuitive understanding of the corresponding relationship with rental status, so we use a logistic model to try to observe the relationship between each variable and the statistical significance of rental status.

2.2 Highly competitive features in rental market

For renters, one of the most concerned issues is that their rented houses can be rented out as soon as possible and can be rented to a price that they are satisfied with. However, because many lessors do not understand the leasing market and cannot quantify or obtain the competitiveness of their rental houses in the market, this is likely to cause the lessors to make mistakes in the evaluation of their rental houses, and to pay too high or too low prices. Renting a house may not only unintentionally infringe on the personal interests of the lessor or lessee, but may disrupt the market order to a certain extent. Therefore, a reasonable price evaluation based on the properties of the house is a very important part of the house leasing market[2].

2.2.1 Relationship between rental price and status

The analysis results show that the house price in New York in April 2020 and its density curve show that most of the house prices in this month are concentrated between $110 and $200 (About 70% of the price). Compared to New York in 2019 The average rent has been reduced by about 55%. This may be due to the impact of the new crown epidemic to some extent.

Taking house rent feature as an example, we further explored the influence of house rent feature on house occupancy rate based on the previously established logistic regression model. We took other variables as the median of the overall sample, and restricted the scope of house rent to the minimum and maximum values in the overall sample.

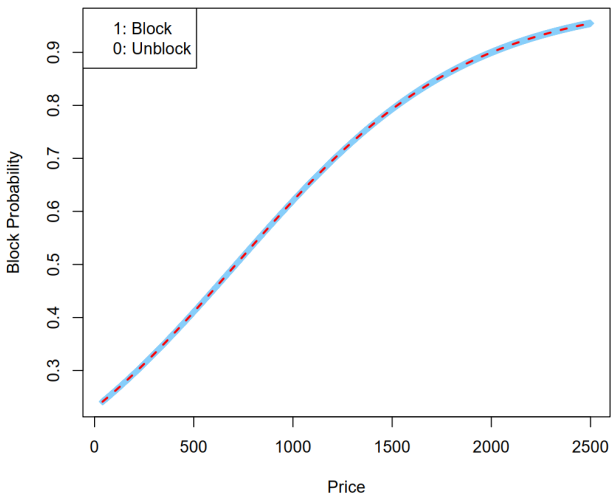

Fig. 1. Changes in the rental rate based on the rental feature of 1003 short-term rental houses in New York on Airbnb

As shown in Figure 1, the occupancy rate obviously decreases as the price of house rent increases (the probability of Block gets closer to 1). The specific calculation formula of the logistic model is as follows:

$$ln\frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (1)$$

$$\pi = \frac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)} \quad (2)$$

$$P = \begin{cases} 0 & \pi \le A \\ 1 & \pi > 1-A \end{cases} \quad (3)$$

Where x_k is the independent variable, π is the monotonic continuous probability function, and P is the dependent variable (rental status, 1 corresponds to Block; 0 corresponds to Unblock).

2.2.2 Housing location and occupancy rate

Rental prices in different regions are not the same, so this may also affect the rental status of the house, so we explored the rental status of different location.

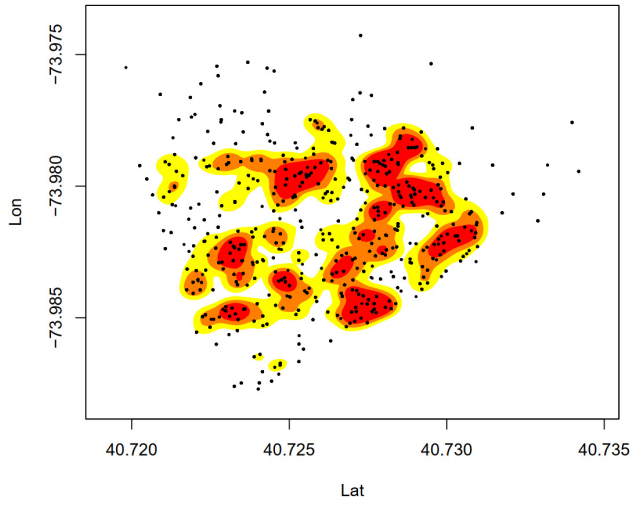

Fig. 2. The density heat map of blocked properties locations

By examining the impact of regional differences on the housing rental situation (Figure 2), it can be found that the rental rate in a specific area is significantly higher than that in other areas. The higher the rental rate, the more tenants in the corresponding area, and the greater its density (the red area in the figure is a high-density area).

The density heat map composed of blocked properties projected in latitude and longitude shown that the red area in the figure is the high-density area of properties in the block state. This area has the most blocked state properties, and the number of blocked properties in yellow area are second only to the red area. It can be seen that the properties of the block state are relatively concentrated, and it seems to have a cluster effect[3].

2.2.3 Periodicity of rental status

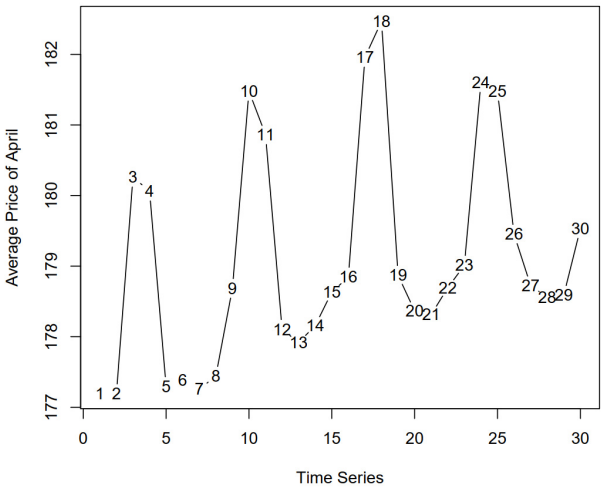

Fig. 3. Time series of blocked properties for each day in April

According to the time series distribution of the number of blocked properties per day in April, we found the change of blocked properties has a clear periodicity, and the interval between the periods is approximately one week. Looking at the whole of April, blocked properties generally showed a continuous increase trend, and reached the peak of the number of blocked properties on April 30th. Based on our life experience, it is generally believed that more tenants will come to rent on weekends or holidays. We further dig out the average number of blocked days in each week in the research data.

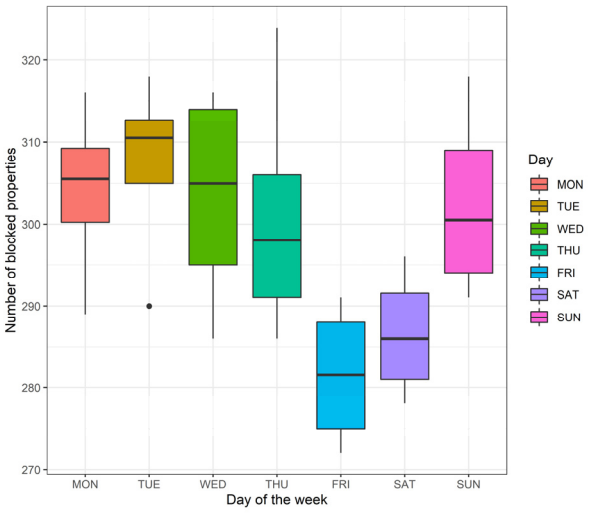

Fig. 4. Number of properties blocked on each day of week

According to a boxplot of the number of blocked properties for all properties in April corresponding to each day of the week, we can see that Tuesday and Thursday are relatively have more blocked properties, while Friday and Saturday have the least blocked properties. Housing features also include, but are not limited to, the rental method (whole rental/shared rental/unknown), the floor of the house, the total number of floors of the house, the decoration of the house, the area of the house, the number of views of the house's webpage in the current month, etc. These features, individually or comprehensively, may have a significant impact on housing rents and housing occupancy rates, and they are also the objects we need to focus on[4,5].

2.3 Model prediction

Table 2. Prediction results of logistic regression model on the rental status of 1003 short-term rental houses

Model	Status	Correct	Wrong	Accuracy
Based on 15 variables (in sample)	Unblock	15890	4774	73.574%
	Block	2430	1806	
Based on 3 variables (in sample)	Unblock	20142	6	70.316%
	Block	46	8516	
Based on 15 variables (out sample)	Unblock	4736	1463	73.199%
	Block	732	539	
Based on 3 variables (out sample)	Unblock	6015	2587	69.964%
	Block	11	0	

Based on the results of predicting the rental status of 1003 short-term rental houses in New York on the Airbnb platform in April 2020 using logistic regression model[6], it is found that only the three variables of house rental price, average annual rental income, and regional price average can well predict the rental status of the house.

Table 3. Confusion matrix of random forest model prediction result base on the rental status of 1003 short-term rental houses

A^* / P^*	Rental Status			Class Error
	A	B	R	
A	728	11	10	0.028037383
B	2	384	0	0.005181347
R	17	31	197	0.195918367
OOB estimate of error rate: 5.14% A^*:*Actual value;* P^*:*Predictive value*				

After modeling the rental status of 1003 New York short-term rental houses on the Airbnb platform in April 2020 using random forest based on full variables, the error rate of the prediction results is only 5.14%. Random Forest is a versatile machine learning algorithm that can perform regression and classification tasks. In a random forest, a lot of decision trees will be generated. When a new object is classified and discriminated based on certain fetures, each tree in the random forest will give its own classification choice, and then weight it. The forest overall output will be the classification option with the highest weight. In essence, the random forest algorithm is an improvement to the decision tree algorithm. It combines many decision trees into a forest. The growth of each decision tree is only related to the corresponding sample. Every tree in the forest is subject to the same distribution, the effect of random forest classification and the size of the error are determined by the ability of each decision tree and the correlation between trees.

It can be seen from Table 3 that the OOB (Out-of-bag) error of the model is 5.14%, and the model has the highest error rate when it is classified in the R category (Reserved), and the error rate is the lowest for classification in A category (Available).

$$OOB\ error = \sum_{1}^{N} \frac{Erri_i}{N}_i$$

Where N is the number of samples. For a sample x_i, a small random forest composed of numbers trained by x_i is not used to predict x_i. The prediction error is the OOB Error in the entire random forest model, that is $Erri_i$. Generally speaking, the OOB Error will be larger than the cross-validation error, because only some of the trees in the random forest are selected and the model is not used completely, which limits the performance of the model. However, it saves the huge amount of calculation that requires multiple training in cross-validation and is more efficient[7].

3 Conclusion

Through this research, based on the short-term rental information of merchants on Airbnb in New York in April 2020, we found that the most influential factor on the status of rental housing is usually the corresponding rental status information of the merchants in the past. At the same time, house rental price, average annual rental income, and regional price have the greatest impact on housing rental status. We used logistic model to predict the rental status of rental houses with in-sample and out-of-sample model tests, the model prediction accuracy rate exceeds 70%. And further combined with the random forest model, we increased the model prediction accuracy to 94.86%. After conducting further research on the impact of time, price, and region on the state of housing rental. It was found that the state of house rental is significantly related to the price of rent. The higher the price of the house, the more likely it is that the house is in a block state, and the price of rent is largely affected by the region. Meanwhile, businesses are more inclined to block the house on Tuesdays. Further research on the block status of houses in the region found that it seems that some specific areas are more prone to block status, and these areas are often some areas with higher rent prices, which is in line with the previous findings of Price and Status research. The final conclusion from our research is that the rental status of houses is largely affected by time, area, and rent, and there is a great correlation between rent and area.

This research results can allow renters to better evaluate the prices of houses based on their own house advantages, which can improve the effectiveness of the landlord's decision-making on rent prices, and at the same time increase the rents according to the priority of the tenant's attention to the house advantages among the market competitiveness; it can also allow residents to refer to the forecast of the future rental status of the house to avoid missing the best time to rent a house due to personal hesitation and other issues. It can help tenants to better understand the future rental status of the house, so that they can better choose a suitable house.

Acknowledgment

This research was financially supported by 2020 Leshan Normal University National College Student Innovation and Entrepreneurship Training Program Fund Project (Project No. 202010649034).

References

1. Miller, S. R. . "Transferable Sharing Rights: A Theoretical Model for Regulating Airbnb and the Short-Term Rental Market." Social Science Electronic Publishing (2014).
2. Leon, Adlpd , et al. "Impact Of Trust On The Platform And On The Host In The Airbnb Accommodation Rental Market El Impacto De La Confianza En La Plataforma Y En El Anfitrion En La Renta De Alojamiento En Airbnb." Revista Internacional Administracion & Finanzas 13(2020).
3. J Lladós-Masllorens, and A. Meseguer-Artola . "Price Determinants of Tourist Accommodation Rental: Airbnb in Barcelona and Madrid." Research & Innovation Forum (2019).
4. Williams, and Rudi. "Clarification of ROA's Non-Partisan Status and Building Rental Use. " Officer (2004).
5. Cooper, M. . "A-Plant forecasts UK rental market growth." Contract Journal (2005).

6. Guerriero, F. , and F. Olivito . "Revenue Models and Policies for the Car Rental Industry." Journal of Mathematical Modelling & Algorithms **13.3**(2014):247-282.
7. Dong, Y. , B. Du , and L. Zhang . "Target Detection Based on Random Forest Metric Learning." IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing **8.4**(2017):1830-1838.