Prediction of House Rental Status Based on Logistic Model of Airbnb on the Hotel Industry

Abstract

With the vigorous development of the sharing economy, the hotel industry has also spawned many emerging industries that belong to the sharing economy. We explored Airbnb in New York City's commercial housing rental information in April 2015. Through logistic models and quantitative analysis, we conducted statistical and visual analysis on the impact of different factors on the housing rental status.

Contents

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. 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. After cleaning and preprocessing 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, I also observed the impact of time and region on the state of housing rental through visual analysis.

2. Data processing and variable selection

For my study, I collect and combine data from various sources including the Airbnb website, I deleted the variables with too many missing values, and 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 information

variables				
Status	Min. :0.000	Median :0.000	Mean :0.297	Max. :1.000
Price	Min. :40	Median : 150	Mean : 179	Max. :2500
price_norm	Min. :-609.69	Median : -4.64	Mean : 2.90	Max. :2243.00
photo_room_ratio	Min. :0.20	Median : 6.00	Mean : 6.96	Max. :66.00
Number of Photos	Min. :1.0	Median : 12.0	Mean : 13.9	Max. :132.0
Count Reservation Days LTM	Min. :1.0	Median : 51.0	Mean : 91.7	Max. :350.0
Count Available Days LTM	Min. :0.0	Median : 74.0	Mean : 99.6	Max. :364.0
Count Blocked Days LTM	Min. :0	Median :115	Mean :138	Max. :365
Max Guests	Min. :1.00	Median : 2.00	Mean : 2.86	Max. :14.00
Cleaning Fee	Min. :5.0	Median : 60.0	Mean : 66.7	Max. :375.0
Annual Revenue LTM	Min. :85	Median : 8591	Mean : 16753	Max. :202412
Occupancy Rate LTM	Min. :0.033	Median :0.667	Mean :0.640	Max. :1.000
Number of Bookings LTM	Min. :0.0	Median : 9.0	Mean : 17.3	Max. :128.0
Number of Reviews	Min. :0.0	Median : 15.0	Mean : 29.7	Max. :248.0
Bedrooms	Min. :0.0	Median :1.0	Mean :1.1	Max. :7.0
Bathrooms	Min. :0.50	Median :1.00	Mean :1.09	Max. :4.00

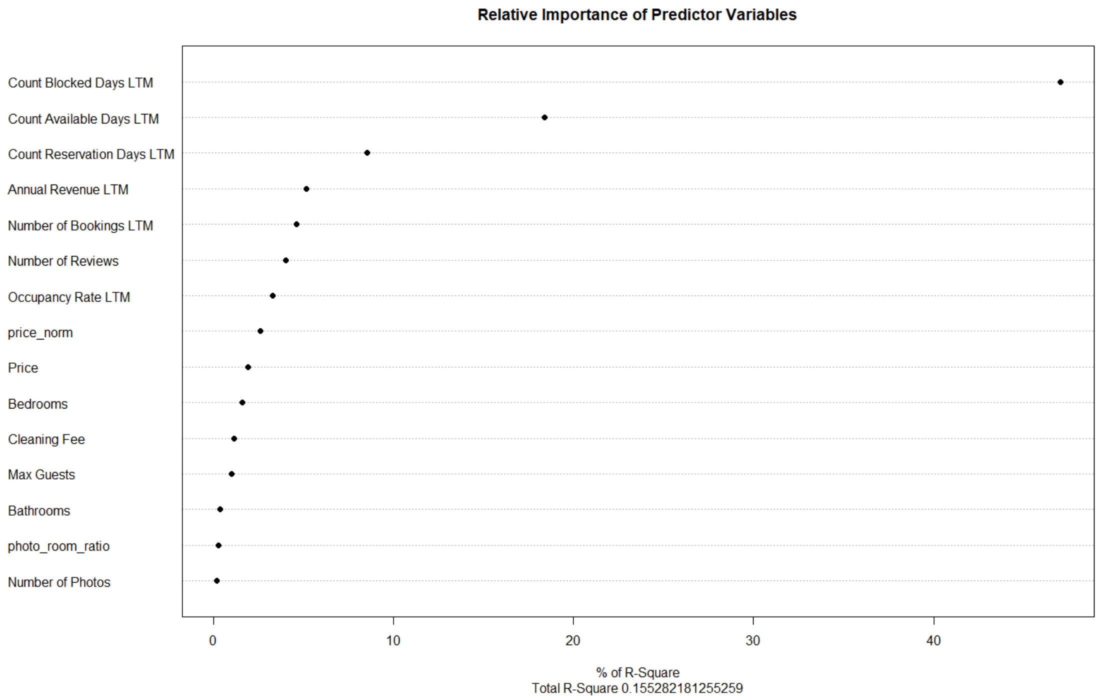
I first converted the three status variables of A, B, and R (A: available B: block R: reserved) into binary classification variables of 1 (Block) and 0 (Unblock) to facilitate the application of the logistic model. The basic statistics (minimum value, mean value, median value, and maximum value) of these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

Table 2: Logistic model results

```
Deviance Residuals:
    Min       1Q   Median      3Q      Max
-1.640  -0.755  -0.532   0.992   2.482

Coefficients:
                                 Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)                    -6.52e-01   1.35e-01   -4.84   1.3e-06 ***
Price                            -5.91e-03   4.11e-04  -14.39   < 2e-16 ***
price_norm                        6.20e-03   4.06e-04   15.28   < 2e-16 ***
photo_room_ratio                 1.58e-02   8.69e-03    1.81     0.070 .
Number of Photos                 -8.52e-04   4.52e-03   -0.19     0.850
Count Reservation Days LTM   -3.30e-03   6.22e-04   -5.31   1.1e-07 ***
Count Available Days LTM     -2.76e-03   3.30e-04   -8.37   < 2e-16 ***
Count Blocked Days LTM         5.96e-03   2.45e-04   24.34   < 2e-16 ***
Max Guests                       -1.12e-01   1.55e-02   -7.25   4.2e-13 ***
Cleaning Fee                    -1.28e-03   5.56e-04   -2.30     0.022 *
Annual Revenue LTM              -1.49e-06   1.99e-06   -0.75     0.452
Occupancy Rate LTM              -8.48e-01   8.80e-02   -9.64   < 2e-16 ***
Number of Bookings LTM           1.15e-02   2.28e-03    5.05   4.4e-07 ***
Number of Reviews               9.28e-04   7.67e-04    1.21     0.226
Bedrooms                          7.11e-01   5.61e-02   12.68   < 2e-16 ***
Bathrooms                         2.75e-01   6.00e-02    4.59   4.4e-06 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29957  on 24899  degrees of freedom
Residual deviance: 26009  on 24884  degrees of freedom
  (5190 observations deleted due to missingness)
AIC: 26041

Number of Fisher Scoring iterations: 4
```

It can be found from the results of logistic model that a total of 12 variables including the intercept term have passed the significance test at a significance level of 0.05, indicating that these 12 variables are of significant importance to the model and should be retained. This also shows that the variables to be selected through subjective judgment in advance do have some reasonableness. However, in the case that the 12 variables all pass the significance test, we cannot understand intuitively which of the 12 variables has a relatively greater impact on status, so we use the contribution of each variable to the R-squared to Evaluate the contribution of each variable to the model, that is, the weight corresponding to the change in status.

Figure 1: Relative importance of Predictor Variables



It can be found from the results that the rental status of the merchant in the past 12 months has greatly affected the future situation, which is very similar to the prediction of future data based on the past data of the sample in the time series model, but here I use It is a logistic model. I prefer to select variables that are not related to the past situation of the house rental status, and to ensure that the accuracy of the prediction results is as high as possible, the fewer predictive variables are required. Therefore, based on my personal subjective judgments, I only selected the following three variables for logistic modeling.

Table 3: Coefficients Table

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-7.708e-01	3.338e-02	-23.093	< 2e-16	***
Price	1.704e-03	1.742e-04	9.783	< 2e-16	***
Annual.Revenue.LTM	-2.590e-05	8.827e-07	-29.343	< 2e-16	***
price_norm	-5.948e-04	1.984e-04	-2.999	0.00271	**

The results are also that these three variables are very significant for the model, then I will use the logistic model based on all 15 variables and 3 variables to perform status pre-judgment respectively. First, all the pre-judgments based on the original sample data. That is, modeling from all the sample data, which may cause the model to overfit, which may make the model's status prediction accuracy for the original sample very high, but the prediction accuracy for other samples is often not satisfactory.

3. Model prediction

Table 4: Model prediction results

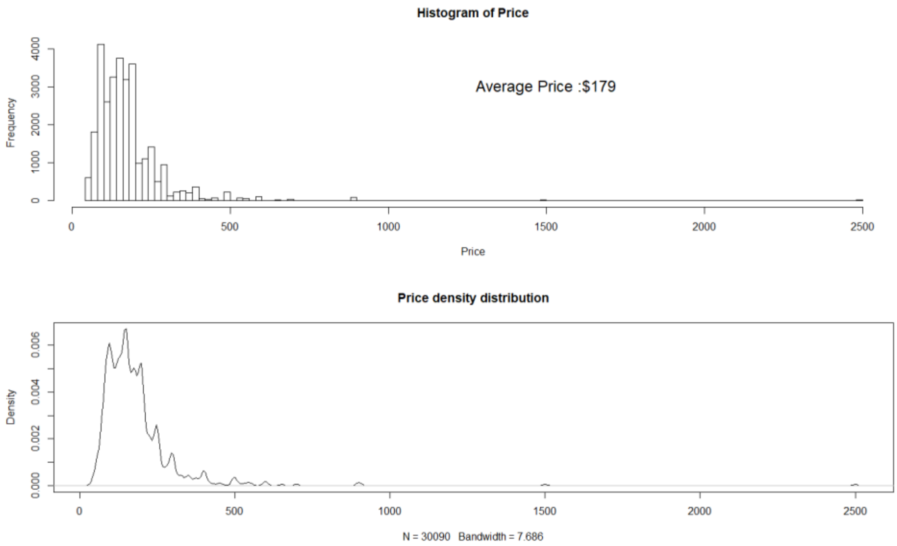
Model	Status	Correct	Wrong	Accuracy	MAPE
Based on 15 variables (in sample)	Unblock	15890	4774	73.5743%	0.442421
	Block	2430	1806		
Based on 3 variables (in sample)	Unblock	20142	6	70.31696%	0.2973222
	Block	46	8516		
Based on 15 variables (out sample)	Unblock	4736	1463	73.19946%	0.4414082
	Block	732	539		
Based on 3 variables (out sample)	Unblock	6015	2587	69.96401%	0.3003599
	Block	11	0		

By calculating the corresponding accuracy and MAPE based on the prediction results matrix of 3 variables and 15 variables of in sample and out sample respectively, it can be found that the prediction accuracy of outsampe has slightly decreased, but the overall accuracy can still be maintained to 70% , MAPE values of 3 variables have obvious advantages.

So next I would like to first analyze and explore the price and the past housing rental status of merchants.

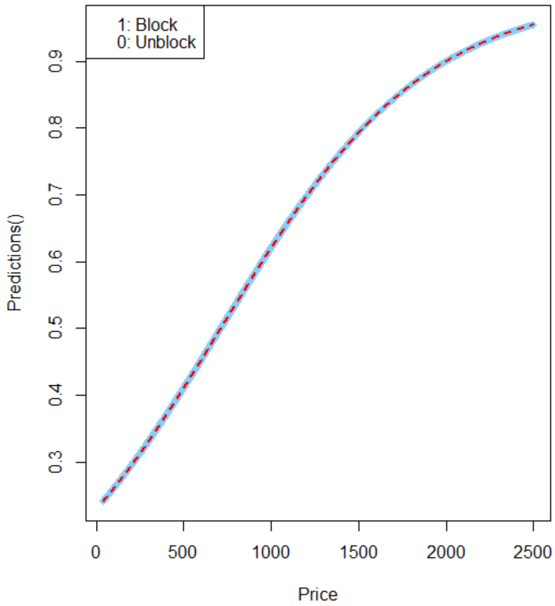
4. Price VS Status

Figure 2: Price distribution



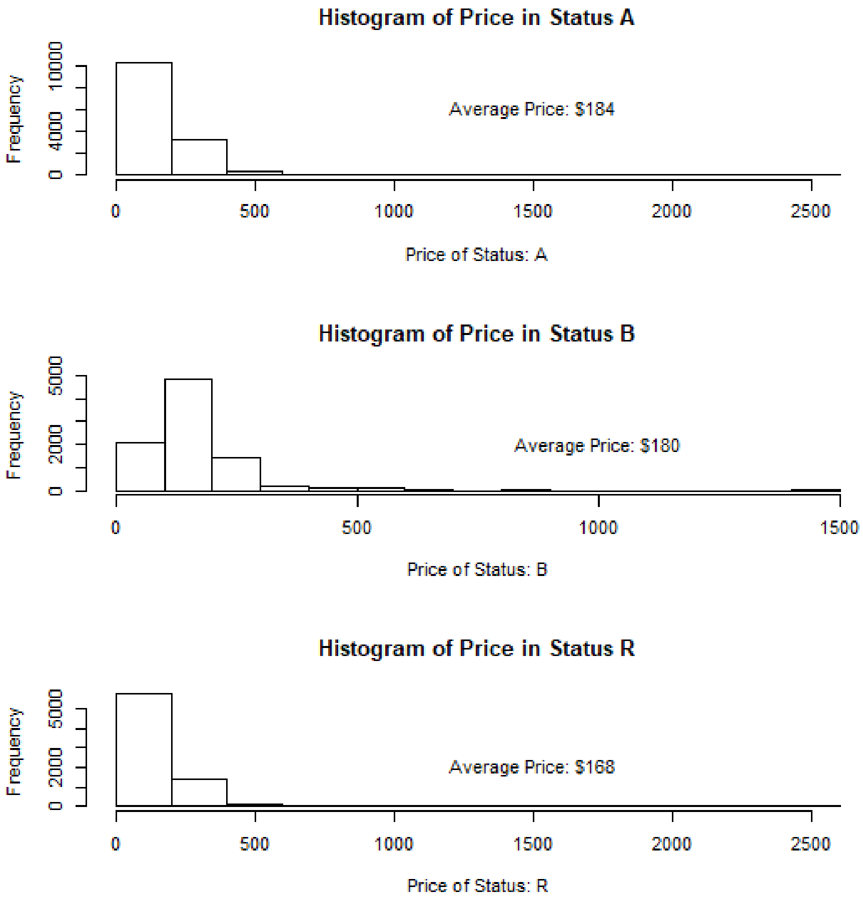
As shown in the figure, the frequency histogram of the house price in New York in April 2015 and its density curve show that most of the house prices in this month are concentrated between $110 and $200 (About 70% of the price).

Figure 3: Status Forecast of Price Changes Based on Logistic Model



When the state of other variables is set to the sample mean, the probability of Stauts being Block is shown in the figure, which is very close to the linear function. Obviously, the probability of the occurrence of Block state is significantly positively correlated with Price.

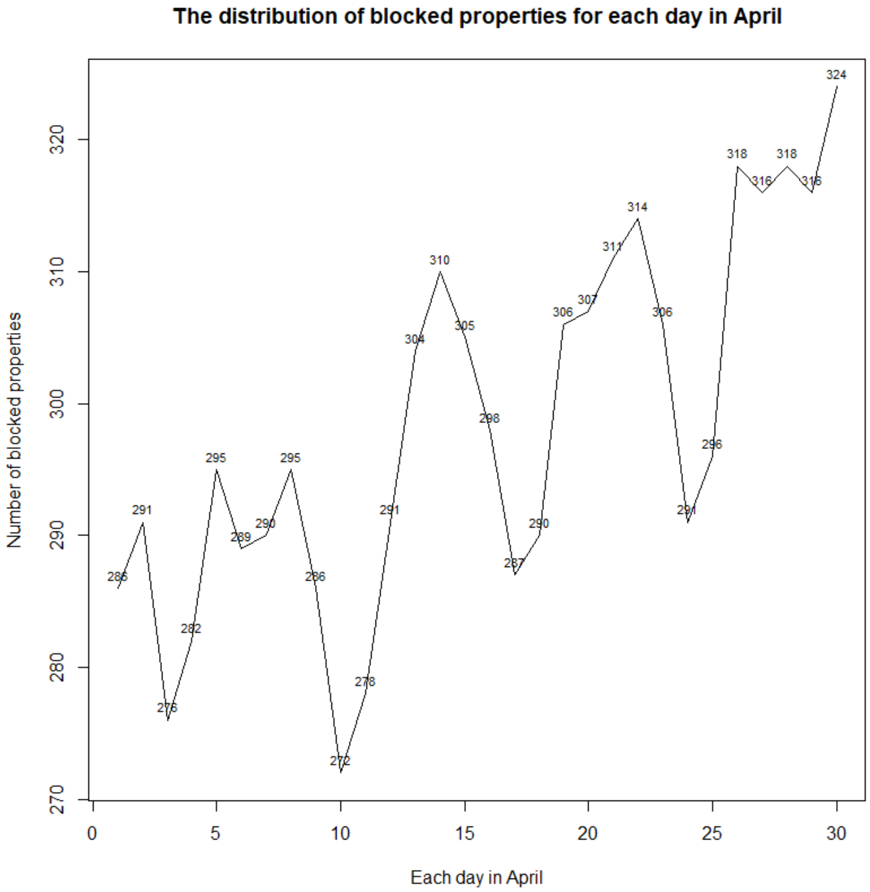
Figure 4: Price distribution of Different Status



With a histogram plot and a violin plot we can definitely observe a couple of things about distribution of property prices of different property status in NYC. First, we can state that property in status A has the highest range of prices for the listings with $184 price as average observation, followed by status B with $180 per night, while the property status A has the enormous amount compared with other status. All status appeared to have very similar distributions. Status R is the cheapest of them all. This distribution and density of prices were indicate that is property in Status A has higher cost to live in, where property in Status R on other hand appears to have lower standards of living.

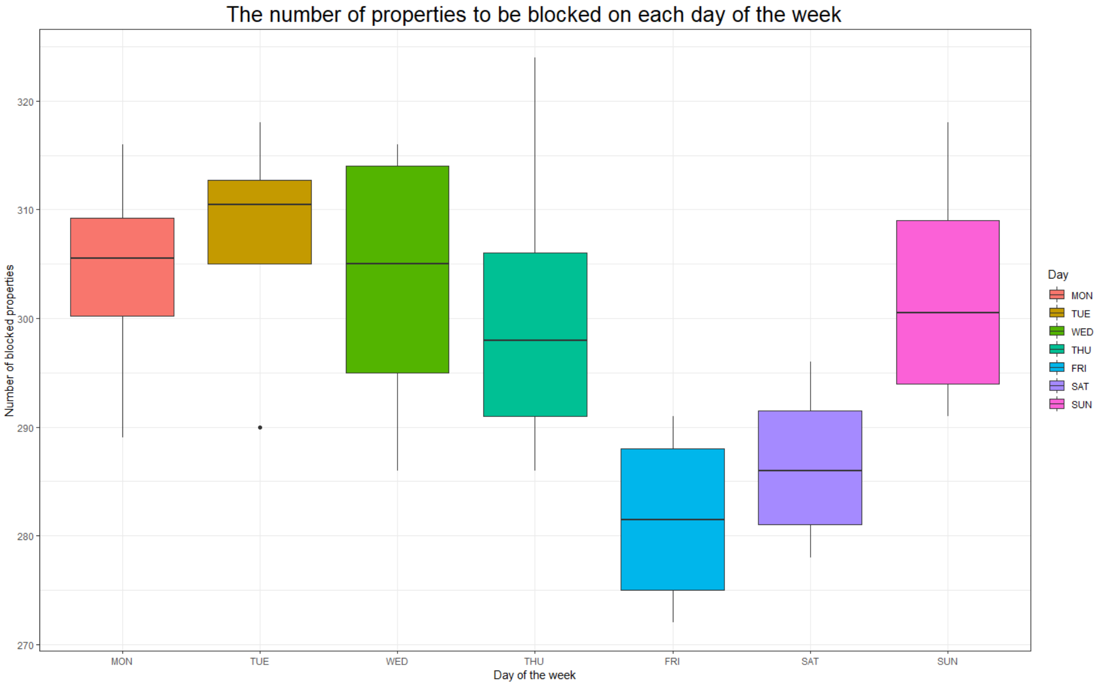
5. Time VS Status

Figure 5: The time series plot of blocked properties for each day in April



According to the time series distribution graph of the number of blocked properties per day in April, it can be seen that 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.

Figure 6: The number of properties to be blocked on each day of the week

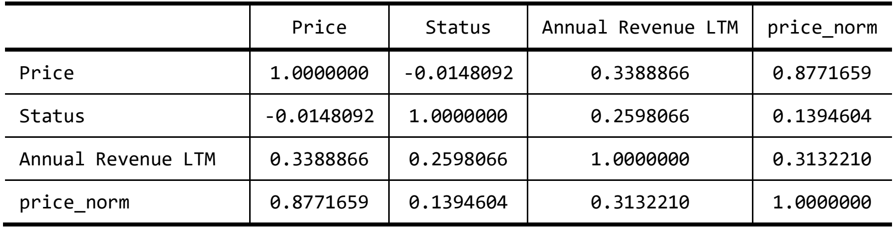


According to a boxplot of the number of blocked properties for all properties in April corresponding to each day of the week, we can see that Tuesday and Thursday are relatively have more blocked properties, while Friday and Saturday have the least blocked properties.

6. Regional Difference

Rental prices in different regions are not the same, so this may also affect the rental status of the house, so I will explore the rental status of the two places separately.

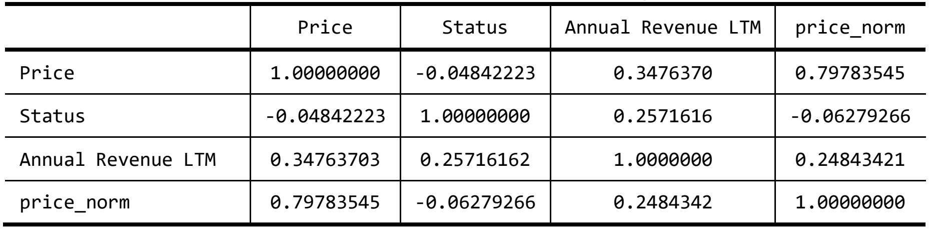
Table 4: Correlation coefficient between Stuyvesant Town variables



	Price	Status	Annual Revenue LTM	price_norm
Price	1.0000000	-0.0148092	0.3388866	0.8771659
Status	-0.0148092	1.0000000	0.2598066	0.1394604
Annual Revenue LTM	0.3388866	0.2598066	1.0000000	0.3132210
price_norm	0.8771659	0.1394604	0.3132210	1.0000000

From the strong correlation between the rent and the real estate price, the correlation coefficient (0.338866), we can find that there is a certain correlation between the real estate price and the Status, the correlation coefficient (0.2598066).

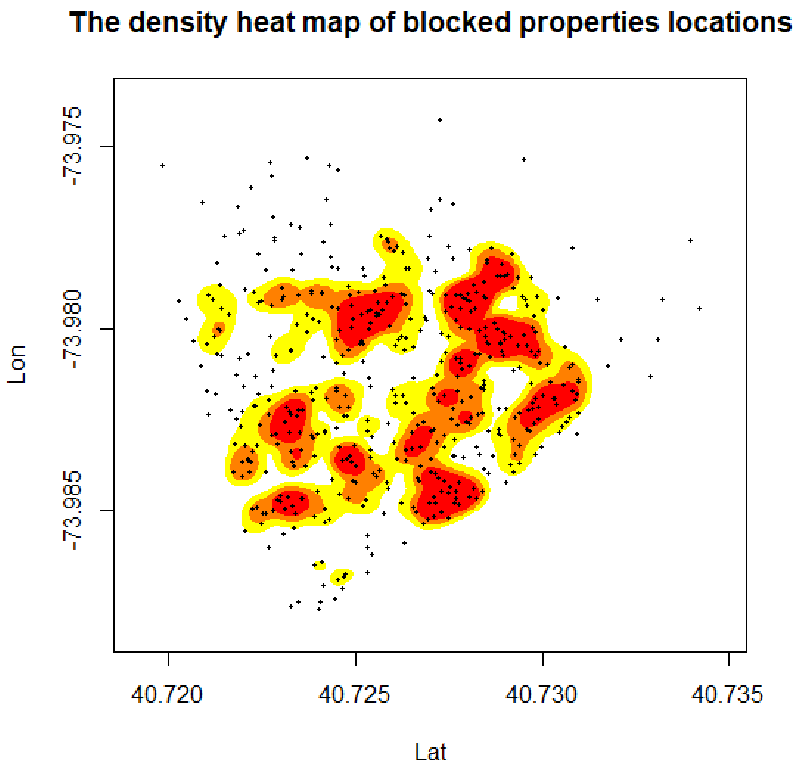
Table 5: Correlation coefficient between East Village variables



	Price	Status	Annual Revenue LTM	price_norm
Price	1.00000000	-0.04842223	0.3476370	0.79783545
Status	-0.04842223	1.00000000	0.2571616	-0.06279266
Annual Revenue LTM	0.34763703	0.25716162	1.0000000	0.24843421
price_norm	0.79783545	-0.06279266	0.2484342	1.00000000

From the strong correlation between the rent and the real estate price, the correlation coefficient (0.34763703), we can find that there is a certain correlation between the real estate price and the Status, the correlation coefficient (0.25716162).

Figure 7： The density heat map of blocked properties locations



The density heat map composed of blocked properties projected in latitude and longitude can be found 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.

7. Summary

Through this research, I found that based on Airbnb's merchant rental information, I found through logistic models based on 15 variables. The most influential state is often the corresponding rental status information of the merchant in the past. Time period is relevant. The prediction results of in sample and out sample have reached an accuracy rate of more than 70%, which seems acceptable. I then continued to conduct 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. At the same time, 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 this paper 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. Through the model in this study, tenants can better understand the future rental status of the house, so that they can better choose a suitable house.

REFERENCES

Bolton, Gary, Ben Greiner, and Axel Ockenfels. 2013. “Engineering Trust: Reciprocity in the Production of Reputation Information,” Management Science, Vol. 59, No. 2, 265-285.

Card, David, Alexandre Mas. and Jesse Rothstein, 2008. “Tipping and the Dynamics of Segregation,” *Quarterly Journal of Economics*, Vol. 123, No. 1, 177-218.

Benjamin Edelman, Michael Luca. 2014. “Digital Discrimination:The Case of Airbnb.com,” Working Paper, 14-054.

Anderson, Simon P., and Victor A. Ginsburg. 1994. “Price Discrimination via Secondhand Markets.” *European Economic Review*, 38(1): 23–44.

Attanasio, Orazio. 2000. “Consumer Durables and Inertial Behavior: Estimation and Aggregation of (S,s) rules for Automobile Purchases.” *Review of Economic Studies*, 67: 667–696.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. “Automobile Prices in Market Equilibrium.” *Econometrica*, 63(4): 841–890