Harbin Institute of Technology

College of Economic and Management

Statistical Learning and Data Mining

Regression Project: Airbnb Pricing Analytics

June 20, 2022

Airbnb Pricing Analytics Report

Wang Tianye

Contents

1 Introduction

With the rapid development of the Internet, the sharing economy derived from it. In the rental space, Airbnb is the largest rental sharing platform. His business operating model is directly connected between landlords and customers, without intermediaries. Therefore, the rental transaction rate on the Airbnb platform is high, the competition is large, and the rental price is much lower than that of other traditional rental platforms. Price is always the most important piece of information for landlords and tenants to consider. But how the price is set is a tricky business for landlords. Because if the price is set too low, it will lead to a lower income for the landlord. If the price is set too high, the house cannot be rented out. So price setting is a very important thing for the sharing economy.

At present, many scholars have analyzed the price of the Airbnb platform.For the characteristic analysis aspect of the landlord,DanWang and Juan L. Nicolau analyzed 180533 rentals on Airbnb. They found that the price is mainly related to 5 variables host attributes, site and property attributes, amenities and services, rental rules, and online review ratings [13].Li, Y.Pan, Q Yang, T Guo, L. analyze the distance of the house from the nearest landmark and find that the rental price of the house has a lot to do with the distance from the landmark[10].Gutt, D. and Herrmann, P. analyzed 14,000 rental information from New York City and found that subjective evaluations have an impact on rental prices[6]. For the characteristic analysis of the house,in articles [3] and [14], the author analyzes the data and finds that the location and geographical environment of the house is closely related to the price of the house.Articles [4], [2] reveal that different types of houses and the facilities contained in the house have an impact on the rental price of the house.Article [12],[9] found that the customer's evaluation of the house can have an impact on the rental price of the house.

We used more than 10,000 pieces of rental information from Kagge's Airbnb London to analyze what factors rental prices are related to and build models to predict the rental price of the house, helping landlords better set rental prices.We also helped landlords analyze some of the favorable factors to improve the success rate of successful landlords renting out their homes.

Our analysis addresses two issues:

1.To develop a predictive model for the daily prices of Airbnb rentals based on state-of-the-art machine learning techniques. This model will and allow the company to advise hosts on pricing and to help owners and investors to predict the potential revenue of Airbnb rentals (which also depends on the occupancy rate).

2.To obtain some insights that can help hosts to make better decisions. What are the best hosts doing?

2 Dataset

2.1 The original dataset

The dataset we use is derived from information about rental homes on the Airbnb platform on the Kaggle Data Analytics website. The details is on the website [1]. The dataset is divided into two parts, the first of which is the training set, a dataset consisting of 12941 rows and 61 columns. The other is the test set, which consists of 5547 rows and 60 columns.We found that the test set has one

more id variable and less price variable than the training set.

2.2 Data Cleansing

1. Detect data types of variables

We use dabl packets for data cleansing. The data type is detected first, and the number of types of attribute variables is shown in the table 1.

Data types	numbers
continuous	19
dirty_float	0
low_card_int	14
categorical	10
date	3
free_string	12
useless	3

Table 1: Detected feature types

We can see from the table 1 that there are 19 continuous variables, 14 discrete integer variables, 10 categorical variables, 3 time type variables, 12 string type variables, and 3 unrecognizable variables.Therefore,We removed 3 variables that do not recognize the data type.

2. Data type format conversion

After a detailed look at the data we found that the data type of variables bathrooms_text, price is a string type, and we convert it to a numeric type. We convert data of time type (variable host_since) to the length of time (in days) to May 25th.We convert variables in percentages (host_acceptance_rate, host_response_rate) to numeric types.Table 2 shows the specific details of data type conversion.

Variables	Before	After
price	$100	100
bathrooms_test	2.5 baths	2.5
host_since	2017-08-07	1572
host_acceptance_rate	70%	70
host_response_rate	100%	100

Table 2: Data type conversion

3. Missing value detection and processing

i. Detect Missing Values

By detecting the missing values in the dataset, we found that there were a total of 91167 missing values in the dataset. There are a total of 22 variables with missing values and a total of 11897 tuples with missing values. Missing values as a percentage of the overall dataset were 12.1%.Table 3 shows the specific details of Missing values.

The distribution of missing values, as well as visualization of the median values in the dataset, is shown in the figure 1, 2.

Missing Cells	Numbers
Missing Cells	91167
Missing Cells(%)	12.1%
Missing Columns	22
Missing Rows	11897
Avg Missing Cells per Column	1571.84
Avg Missing Cells per Rows	7.04

Table 3: Missing value details

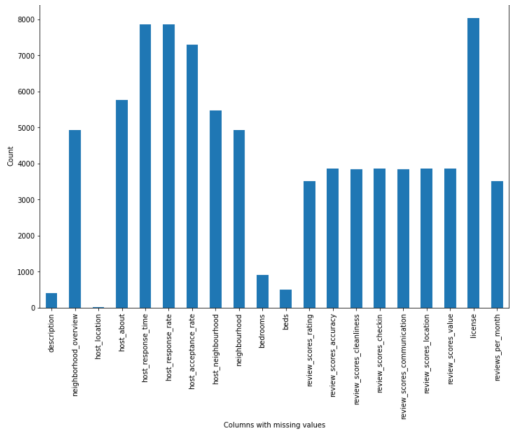

Figure 1: The histogram of missing values variables

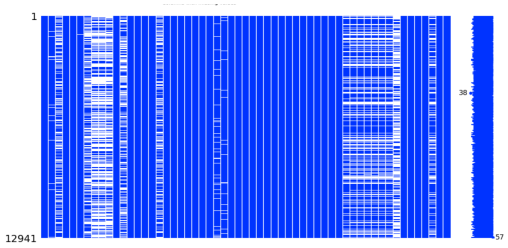

Figure 2: The distribution of missing values

ii. Process Missing values

We populate variables with more missing values and remove tuples of variables with fewer missing values.We populate the missing values in the numeric variable as the average of the variable, the missing values in the literal descriptive variable as none, and the missing values in the proportional variable as 0.

The specific missing values are populated as shown in the table 4.

Variables	processing method	Variables	processing method
bedroom	mean	neighborhood_overview	none
beds	mean	host_response_time	median
review_scores_value	median	host_neighbourhood	none
review_scores_rating	median	neighbourhood	none
review_scores_accuracy	median	description	none
review_scores_cleanliness	median	license	none
review_scores_checkin	median	host_location	none
review_scores_communication	median	host_about	none
review_scores_location	median	reviews_per_month	0
host_response_rate	0	host_acceptance_rate	0

Table 4: Missing value Processing

iii. Processing Other variables

We created a new variable review_gap and removed last_review,first_review variables.The relationship between them is shown in the formula.

$$review\_gap = last\_review - first\_review$$

3 Exploratory data analysis

3.1 response

For the first question, we select the target variable price. First make a histogram of the price variable to observe the distribution of the variables. We found that the target variable was heavily skewed to the right. So we transform the variables logarithmically. The histogram distribution of the target variables and Q-Q plot is shown in the figure 3.

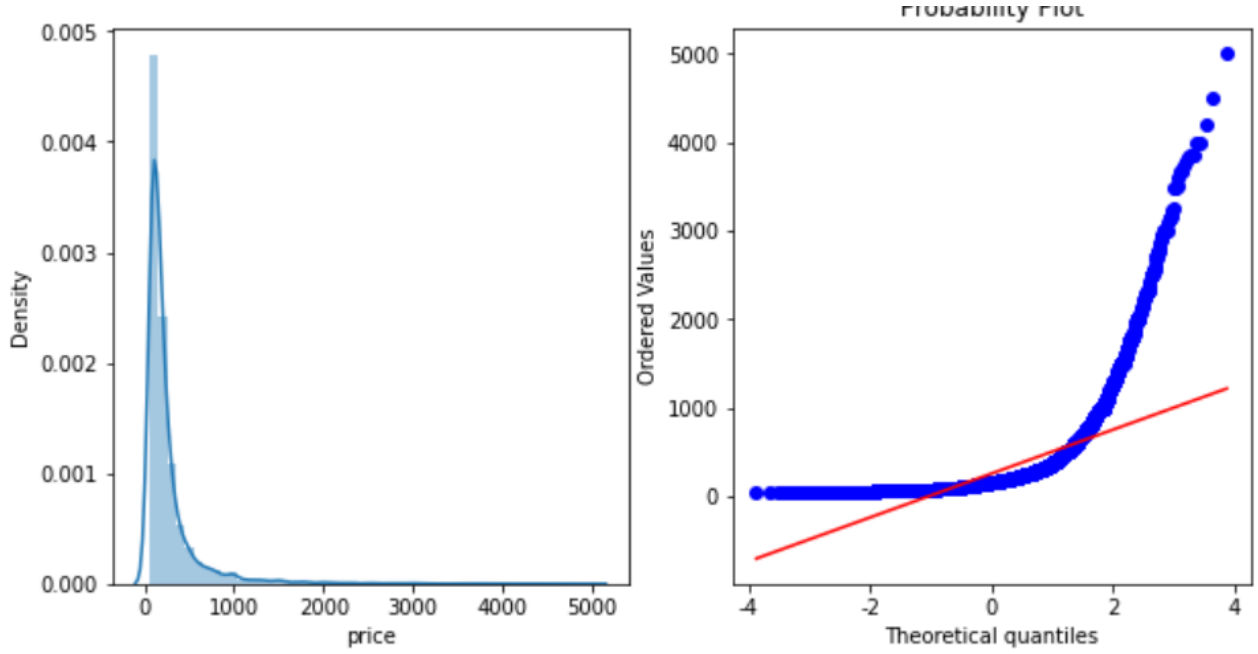

Figure 3: Target variables

3.2 Predicted Variables

For numeric variables, we make a histogram of the variables of the training and testing sets and observe the distribution of the predictors and the comparison of the training set and the data set.The histogram of the comparison is shown in the figure 4,5,6.As we can see from the figure, the variables are almost identical in the training set test set. The distribution of individual variables is heavily skewed.

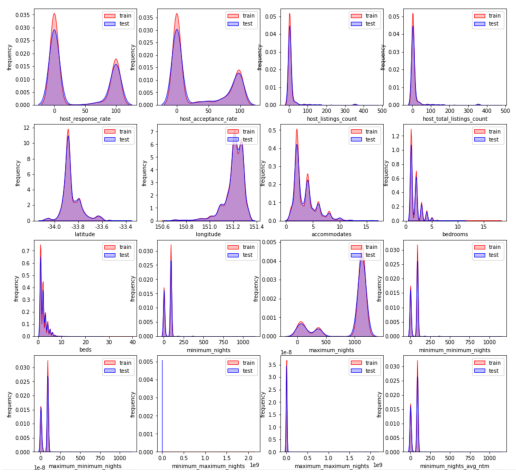

Figure 4: The histogram of predicted variables

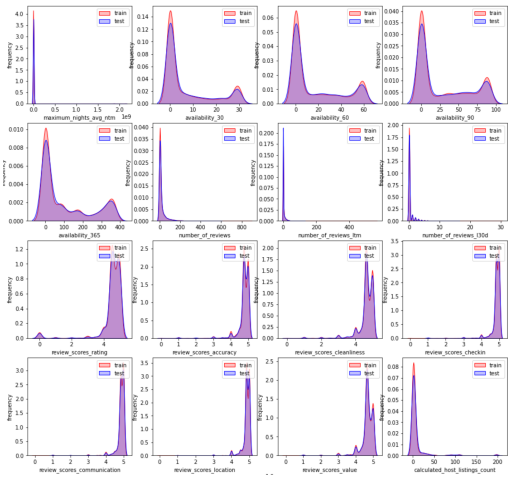

Figure 5: The histogram of predicted variables

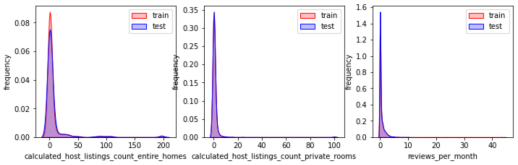

Figure 6: The histogram of predicted variables

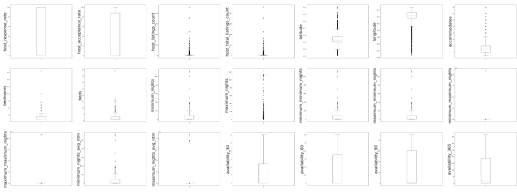

Figure 7: The Boxplot of predictors

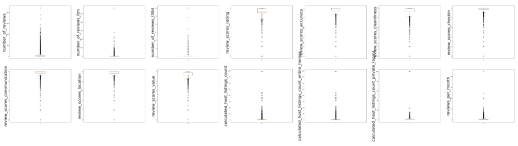

Figure 8: The Boxplot of predictors

Figures 9 and 10 are box plots of numeric variables, and we can see that many variables have many outliers.

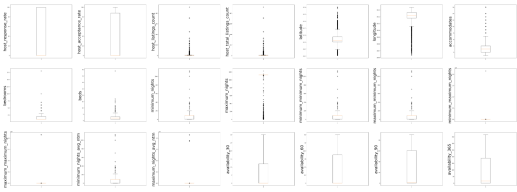

Figure 9: The Boxplot of predictors

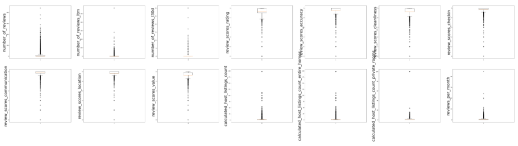

Figure 10: The Boxplot of predictors

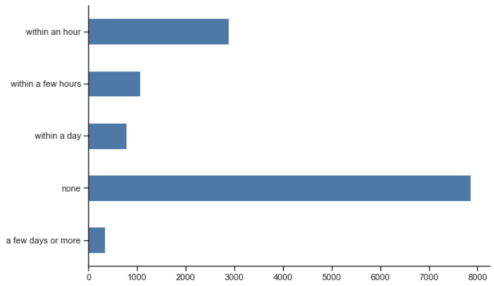

Figure 11: The Barplot of Categorical variables host_response_time

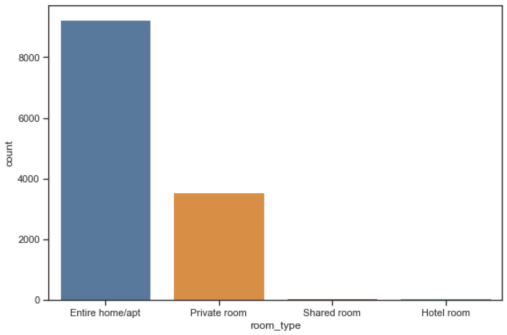

Figure 12: The Barplot of Categorical variables room_type

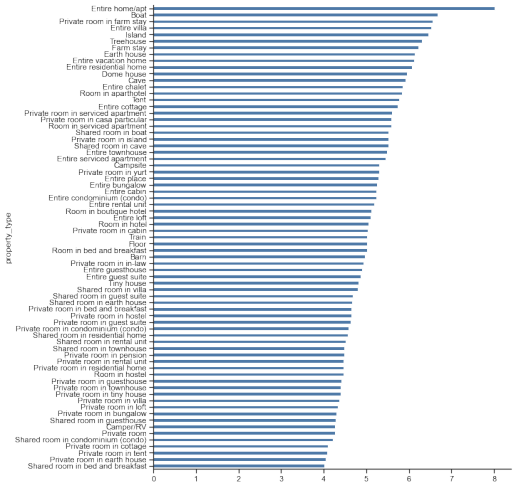

Figure 13: The Barplot of Categorical variables property_type

Figures 11, 12, 13 show a bar chart of categorical variables, and we find that there are some categories that are particularly rare. The hosts mostly did not react to the time and the second is within one hour. Property_type Variable has a particularly large number of categories.Type Entire home is the most common type and the second is private room.

Figure 14: The Wordcloud of neighborhood_overview

Figure 15: The Wordcloud of description

Figures 14 and 15 show the word cloud distribution of text-descriptive variables.We found that in the text of the description of the neighborhood, more attention is paid to the distance between the house and the surrounding shops.In the elaborated text, more emphasis is placed on the house's proximity to the beach, and there are many words of praise for the house.

4 Feature Engineering

4.1 Data Transform

1. Response Variable

We observed that the response variable was heavily skewed, so the response variable was transformed logarithmically, and the transformed histogram and Q-Q plot are shown in the figure 16

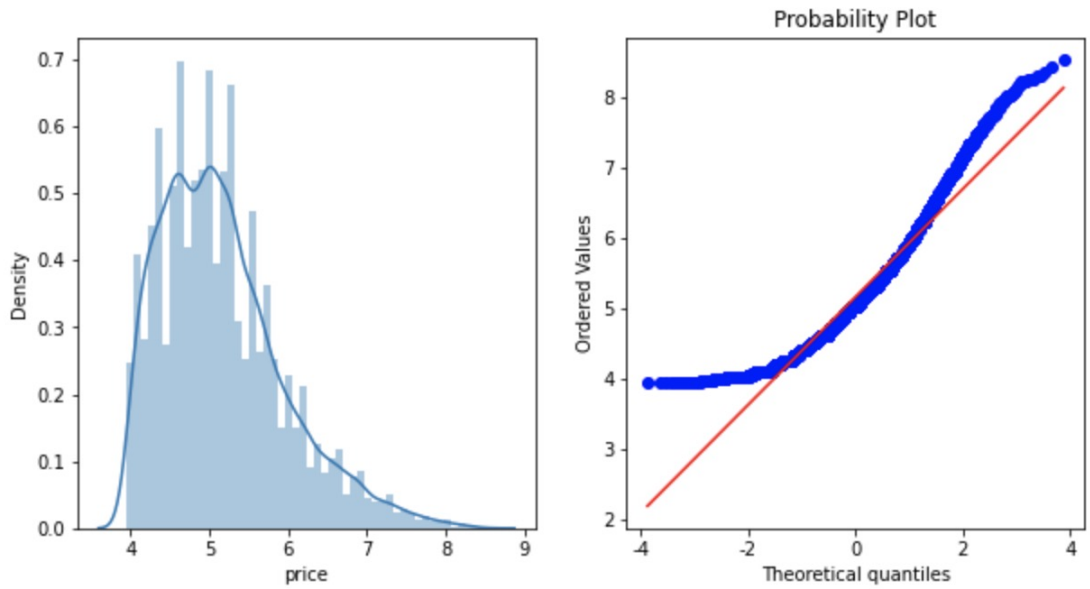

Figure 16: Target variables After Transformation

We found that after the variable conversion, the converted variables are approximately normally distributed.

2.Predictors i.property_type

We found that there are more items in the property collection of property_type variable, so we only keep more of the first 6 items by sorting, and the others are unified to others.

ii.host_response_time

We found through preliminary exploratory analysis that the discrete variable host_response_time contains within an hour, within a few hours, within a day, none, a few days or more terms, so we set 1, 2, 3, 4, 5 in order.

iii.room type

We found through preliminary exploratory analysis that the discrete variable room type contains Shared room, Private room, Hotel room, Entire home apt terms. However Shared room and Hotel room are rare in the dataset, so we set 1 to Entire home apt, 2 to Private room, 3 to Both Shared room and Hotel room.

iv.neighborhood_overview

We extract the number of numbers mentioned in variable neighborhood_overview and set as neighborhood_overview. Because we believe that the description of the surrounding environment with numbers, the description will be more specific, more conducive to the choice of rental customers. So the more numbers appear in the description, the more specific the description.

v.name

We extract the number of all capitalized words in variable 1 as the value of the variable. We believe that a name full of capital letters would be detrimental to renting a home. So the greater the number of capitalized words, the lower the probability of renting out the house.

vi.host_verification,amenties

We changed variable host_verification to the type of service provided by the landlord and variable amenties to the type of room facility.We think the more the quantity, the higher the price of the house relatively high.

vii. host_together_t

We check if the landlord and tenant are in the same district, and if they are in the same district, we set it to 1, otherwise we set it to 0.

viii. description

We detect the number of words that appear positive in the description. We believe that the greater the number of positive words that appear, the higher the house price. Positive words: new, comfort, good, nice, enjoy, close.

viiii. license

We will set the licensed one to 1 and the unlicensed one to 0.We think the price of a licensed house will be higher.

3. Dummy variables

We encode the variables host_is_superhost,neighbourhood_cleansed,license,instant_bookable solitally. We have removed the variables host_name, host_location, host_about. We believe that there is no connection between these variables and the price.

4.2 Correlation Analysis

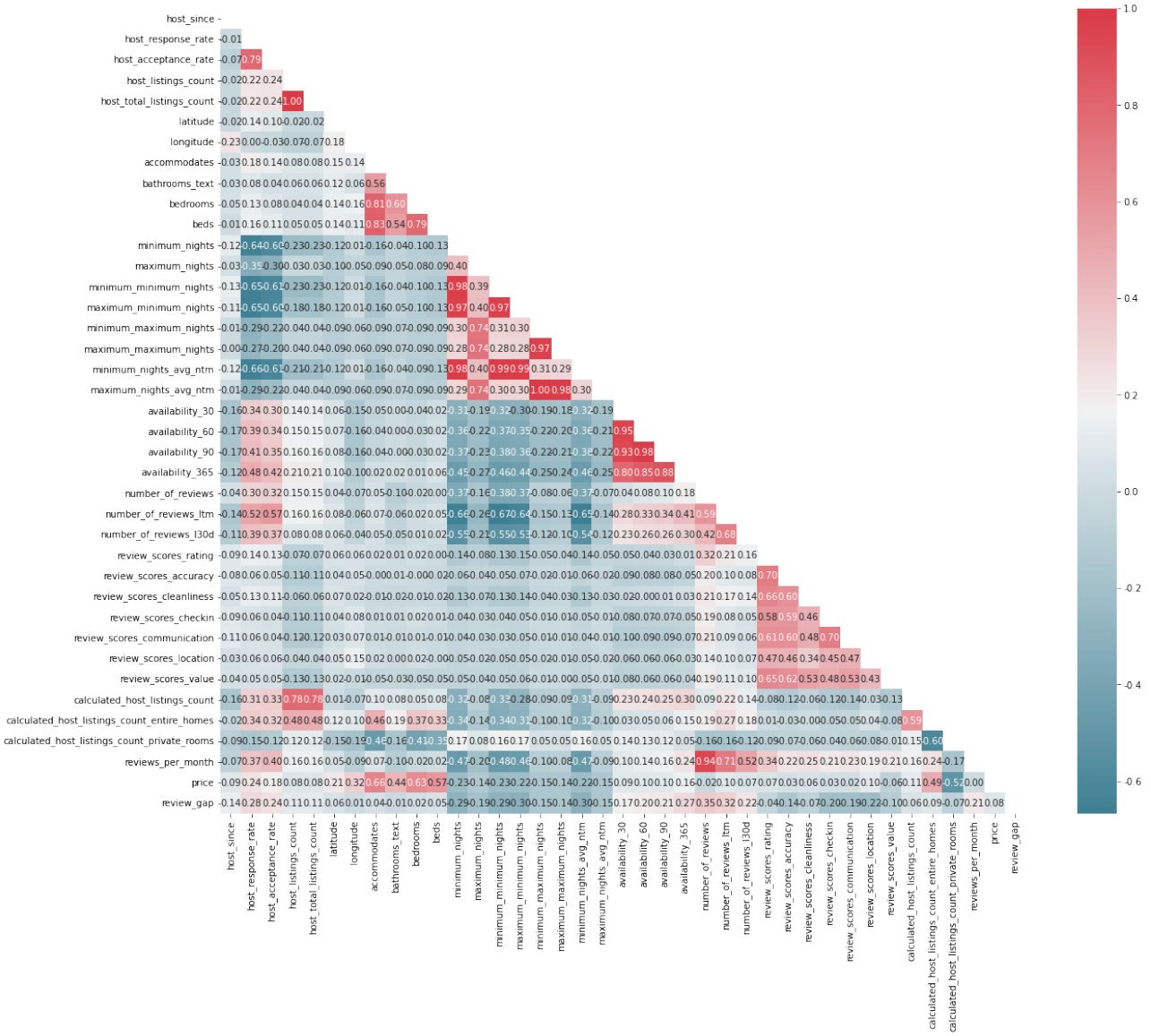

Figure 17: Target variables After Transformation

The stronger the positive correlation, the darker the red, and the stronger the negative correlation, the darker the blue.

5 Model building

We build Ordinary least squares Regression,Xgboost,Random Forest,Light Gradient Boost,Stacking model using variables other than the target variable to make predictions for the target variable. The validation criteria used for the test model are r squared, rmse, mse. The larger the Rsquare, the better the prediction of the model. Rmse, the smaller the mse, the better the model prediction.

5.1 Ordinary least squares Regression

The least squares method (also known as the least flat method) is a mathematical optimization technique. It finds the best function match for the data by minimizing the sum of squares of the error. By minimizing the error's sum of squares, it finds the best function match for the data. The least squares method makes it easy to find unknown data and minimize the sum of squares of the error between these obtained data and the actual data.

The least squares method is the most commonly used method to solve the curve fitting problem. The basic idea is:

$$f(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + ... + \beta_n x_n$$

where $beta_k ispending coefficient(k = 1, 2, ..., n)$. The fitting criterion is to minimize the sum of squares of the distance between y_i and f(x).

Let (x,y) be a pair of observational measurements, and x = $[x_1, x_2, ..., x_n]^T \in R^n, y = R$ satisfies the following theoretical functions.

$$y = f(x, \omega)$$

where$\omega = [\omega_1, \omega_2, ..., \omega_n]^T$ is pending coefficient.

To find the optimal estimate of the function's argument ω, for a given set of$m$ (usually $m > n$) observational data $(x_i, y_i)(i = 1, 2, ..., m)$, the solution objective function is

$$L(y, f(x, \omega)) = \sum_{i=1}^{m} [y_i - f(x_i, w_i)]^2$$

[7]

5.2 Xgboost

XGBoost is an improvement on the gradient boosting algorithm, using Newton's method to solve the extreme value of the loss function, expanding the loss function Taylor to the second order, and adding regularization terms to the loss function. The objective function at training consists of two parts, the first part is the gradient boosting algorithm loss, and the second part is the regularization term. The loss function is defined as

$$L(\phi) = \sum_{i=1}^{n} l(y_i', y_i) + \sum_{k} \Omega(f_k)$$

where n is the number of training function samples, l is the loss of a single sample, assuming it is a convex function, y_i' is the predicted value of the model for the training sample, and y_i is the true label value of the training sample. Regularization terms define the complexity of the model:

$$\Omega(f) = \gamma T + 1/2\lambda ||\omega||^2$$

where y is a manually set parameter, w is the vector formed by the values of all leaf nodes in the decision tree, and T is the number of leaf nodes. Regularize the term[5].

5.3 Random Forest

A random forest is a classifier that consists of multiple decision trees in machine learning, with the output category determined by the mode of the category output by each individual tree. Leo Breiman and Adele Cutler created a random forest deducing algorithm. Their trademark is "Random Forests." Tin Kam Ho of Bell Labs coined the term in 1995 to propose random decision forests. To create a set of decision trees, this method combines Breimans' "Bootstrap aggregating" concept with Ho's "random subspace method."

The algorithm

The following algorithm [1] is used to construct each tree: The number of training cases (samples) is denoted by N, and the number of features is denoted by M. To determine the decision outcome of a node on the decision tree, enter the number of features, m, where m should be much smaller than M. Samples are taken N times from N training cases to form a training set (bootstrap sampling), and the unsampled use cases are used to make predictions and assess their errors. The m features are chosen at random for each node, and the decision of each node on the decision tree is based on these characteristics. Calculate the best splitting method based on this m characteristic. Each tree will grow to its full potential without being pruned, which can be used after a normal tree-like classifier has been built.

The advantages of random forests are:

1) It can produce a highly accurate classifier for a variety of data types;

2) It is capable of dealing with a large number of input variables;

3) When deciding on categories, it can assess the importance of variables.

4) After internal generalization, when building a forest, it can produce an unbiased estimate of the error;

5) It has a good way to estimate missing data, and the accuracy can be maintained even if a large portion of the data is lost;

6) It provides a method for detecting variable interactions through experimentation.

7) It can balance the error in an unbalanced classified dataset.

8) It calculates the affinity for each case, which is highly useful for data mining, detecting outliers, and visualizing data.

9) Make use of the preceding. It can be applied to unlabeled data, which is typically clustered using unsupervised methods. It can also detect deviations and display data; and

10) the learning process is really quick. [11].

5.4 Gradient Boosting Decison Tree

GBDT uses regression trees, GBDT is used to make regression prediction, after adjustment can also be used for classification, set thresholds, greater than the threshold for positive examples, vice for negative examples, you can find a variety of distinguishing features and feature combinations. The goal of GBDT is to add up all of the tree conclusions to arrive at a final conclusion, the core of GBDT is that each tree is the residual of all the previous tree conclusions and residuals, this residual is the accumulation of a prediction value to get the true value.

For example, if A is 18 years old, but the predicted age of the first tree is 12 years old, the residual difference is 6 years old. Then, in the second tree, we set A's age to 6 years old to learn; if the second tree can truly divide A into 6-year-old leaf nodes, the conclusion of the two trees is the true age of A; if the conclusion of the second tree is 5 years old, A still has a 1-year-old residual, and A's age in the third tree becomes 1 year old and continues to learn. The most significant benefit of Boosting is that the residual calculation of each step inadvertently increases the weight of the incorrect instance, while the divided instance tends to 0. This allows the tree behind to focus more and more on those in front of the wrongly divided instance. The gigabit partitioning standard for the gigabit by default is friedmanmse you can see the difference between Gradient Boost and traditional Boost Each calculation is made to reduce the residual of the previous one, and to eliminate the residual, we can build a new model in the gradient direction of the residual reduction.

So, in Gradient Boost, each new model is built so that the residuals of the previous model are reduced in the gradient direction. Shrinkage's idea suggests that the effect of gradually approximating the result with each small step is easier to avoid overfitting than approaching the result quickly with each big step.

That is, it does not fully believe in each remnant tree, and he believes that each tree learns only a part of the truth, and only accumulates a small part when accumulating, and makes up for the deficiency by learning more trees at a time.

Essentially, Shrinkage sets a weight for each tree, multiplied by that weight when added up, but has nothing to do with Gradient. The advantages of GBRT are: 1)Natural handling of data of mixed type (heterogeneous features)

2)Can handle properties of different properties, numerical features category features, Numeric

3)features require preprocessing of the data

4)Predictive power

The disadvantages of GBRT are:

Scalability is limited due to the sequential nature of boosting, which cannot be parallelized.[8]

5.5 Stack

Stacking specific processes

First, divide the data into 5 parts, In the first layer of stacking defines 5 base models [model_1, model_2, model_3, model_4, model_5], where each model chooses to make a 50The output prediction vector [1,1,1,1,1,0] of the first layer of the five base models is trained as the characteristics model_6 the second layer model. When doing test, the test data is directly fed to the 5 base models trained by the first layer before, and the 5 models predicted to average are used as inputs to the second layer model.

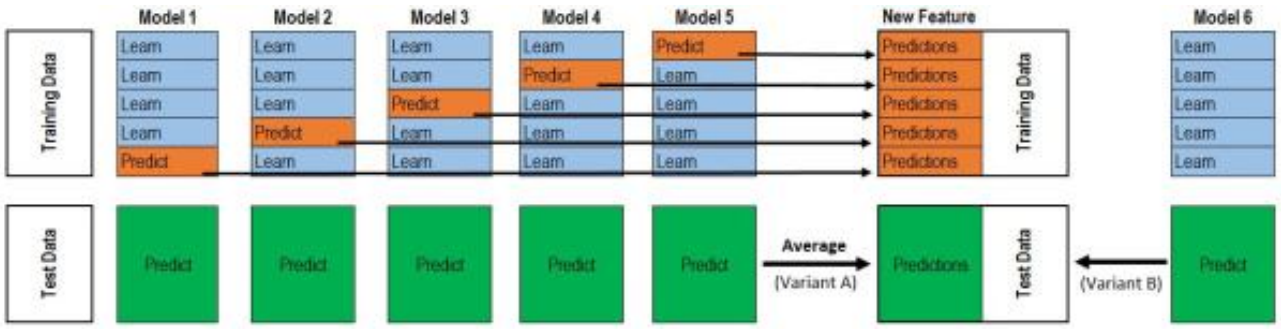

Figure 18: Model Stacking

We fused the first four models to form a new model to train the dataset.

6 Result

The scoring criteria we use are the scores on Kaggle.

Model	Scores
OLS	0.96793
GBDT	0.63443
Random Forest	0.67928
Xgboost	0.64775
Model Stacking	0.44589

Table 5: The Result of outcome

We found that model stacking results were the best.

After statistical analysis, as well as an analysis of the correlation between price and other variables, we found that tenants paid more attention to the facilities in some rooms and were very concerned about how many bathing places in the house.

Houses with more bathrooms tend to be more expensive. The price is inversely proportional to the number of days of rental. The more days you rent, the lower the price of the house. We believe that there may be an increase in rents, and the landlord will give some preferential treatment to the rent.

At the same time, we also found that the price of the whole room type was higher than the price of only the dead man's room.

We also found a very interesting thing, often reply to more timely landlords, the maximum rental time is relatively short. Landlords who are not very efficient in replying will take longer to lease. We think landlords who respond more promptly prefer to rent short-term homes to tourists.

6.1 Suggestion

We recommend that the platform can first classify the characteristics of tenants, and assign tourists and short-term tenants to landlords with higher response efficiency. This avoids the loss of the customer.

At the same time, it is recommended that the landlord can install several more bathrooms in the house when decorating the house, which will help increase the price of the house.

7 Discussion

We used Airbnb's public dataset to make predictions on rental prices and got a Kaggle score of 0.40 (the smaller the score, the better the model). We found that using model fusion worked better than using a single model to make predictions about the target variable. But the model only applies to studying the situation of rentals in Australia, not to projections of rental prices in other countries. The rental environment is different in every country. The report also examines only the characteristics of housing that Australian tenants are concerned about.

And this data analysis only uses a part of the data features in the dataset, and only mines the data features from the aspects of housing conditions and the quality of service of landlords. This data analysis report uses only the panel dataset. Other researchers can also re-examine the problem from other perspectives, or add time series data information.

References

[1] QBUS6810 2021 Semester 2. *QBUS6810_2022_Sem1_Regression*. https://www.kaggle.com/competitions/qbus6810-2022-sem1-regression. Accessed 2022/5/26.

[2] Manuel Becerra, Juan Santaló, and Rosario Silva. "Being better vs. being different: Differentiation, competition, and pricing strategies in the Spanish hotel industry". In: *Tourism management* 34 (2013), pp. 71–79.

[3] Adrian O Bull. "Pricing a motel's location". In: *International Journal of Contemporary Hospitality Management* (1994).

[4] Ching-Fu Chen and Rochelle Rothschild. "An application of hedonic pricing analysis to the case of hotel rooms in Taipei". In: *Tourism Economics* 16.3 (2010), pp. 685–694.

[5] Tianqi Chen et al. "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4 (2015), pp. 1–4.

[6] Dominik Gutt and Philipp Herrmann. "Sharing Means Caring? Hosts' Price Reactions to Rating Visibility". In: *Proceedings of the 23rd European Conference on Information Systems (ECIS), Münster*. 2015.

[7] Graeme D Hutcheson. "Ordinary least-squares regression". In: *L. Moutinho and GD Hutcheson, The SAGE dictionary of quantitative management research* (2011), pp. 224–228.

[8] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017).

[9] Seul Ki Lee and SooCheong Jang. "Premium or discount in hotel room rates? The dual effects of a central downtown location". In: *Cornell Hospitality Quarterly* 53.2 (2012), pp. 165–173.

[10] Yang Li et al. "Reasonable price recommendation on Airbnb using Multi-Scale clustering". In: *2016 35th Chinese Control Conference (CCC)*. IEEE. 2016, pp. 7038–7041.

[11] Mahesh Pal. "Random forest classifier for remote sensing classification". In: *International journal of remote sensing* 26.1 (2005), pp. 217–222.

[12] Christer Thrane. "Examining the determinants of room rates for hotels in capital cities: The Oslo experience". In: *Journal of revenue and Pricing Management* 5.4 (2007), pp. 315–323.

[13] Dan Wang and Juan L Nicolau. "Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb. com". In: *International Journal of Hospitality Management* 62 (2017), pp. 120–131.

[14] Honglei Zhang et al. "Modeling hotel room price with geographically weighted regression". In: *International Journal of Hospitality Management* 30.4 (2011), pp. 1036–1043.