OYO hotels rental price prediction in China

Introduction

OYO hotels, as known as OYO hotels and homes. In recent years, OYO hotels have developed rapidly in China, mainly in franchise and chain of lease, occupying many second and third-tier cities' market.
This report tries to predict the rental price of the OYO hotels in China. So, the target variable is "rental price”, and it's a regression problem. Because OYO's market layout is concentrated in second and third-tier cities, it is also the main city in China, occupying a large part of economic development. Therefore, the research on it has a great reference value.
This report downloads the dataset from Kaggle [1], but all the information can be gathered form Chinese OTA platforms. This dataset has 5834 data points and 25 features.
Through searching, it is found that some people use this data set to predict rent. The author encodes target variable with OrdinalEncoder and tunes parameters with XGBRegressor. The author chooses two models. The prediction of y (y_train, y_pred) in model 1 has mse value of 4.47 and smse value of 2.11; (y_val, y_pred) has a mse value of 45.13 and a smse value of 6.71. Model 2 has a mse value of 46.12. The author finally chose Model 1.

Exploratory Data Analysis

1. The target variable "rental price" is a regression problem, so I use "describe ()" function to exploratory the data. And the max of the price is 10000, the min of it is 0, and the average of it is 286. Because

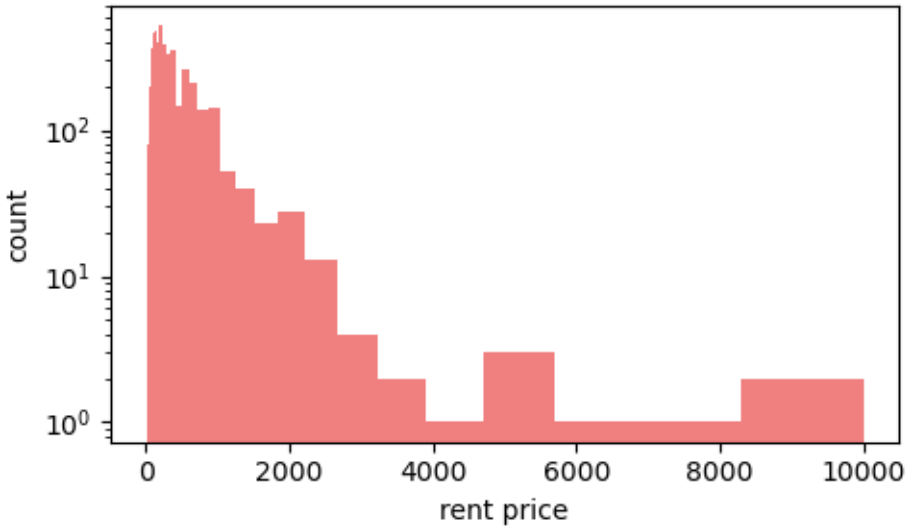

Figure 1

the value difference is too large, so I enlarge the axis by log function like the figure 1 shows. It is clearly to see that the distribution is right-skewed.

2. Relationship between Accommodate and Rent price. (Accommodate which means somewhere provide lodging or sufficient space for.) V This is a numerical classification feature, which should have been

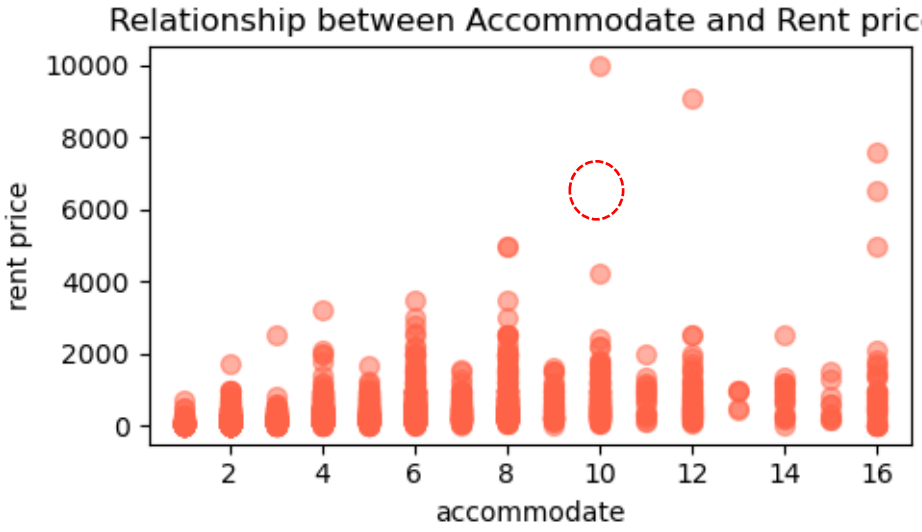

Figure 2

plotted with bar plot. But there are some special points that I hope to express through scatter.

As the figure2 show, if the mark point is equal to 10, the price of this point is particularly high. Because this sample point is not an ordinary hotel of OYO, but a private villa in cooperation with OYO. That' s the reason why this point' s price is so high.

3. Relationship between Property Type and AVG Price.

Figure 3

About the property type feature, it' s a categorical data. Therefore, I calculate the average by using the "groupby()" function to classify them.

4. The OYO hotels locations in China.

Because the dataset has two features, longitude and latitude, I choose to use basemap to draw the country map and scatter to represent the hotel distribution location.

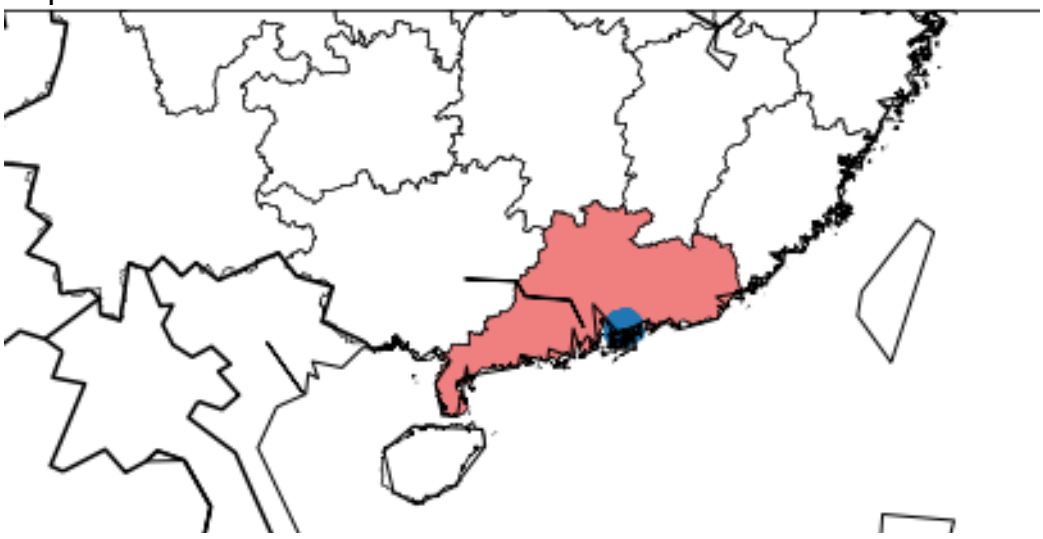

Figure 4

In figure4, the blue dots are gathered by more than 5000 data points, but due to the close location of hotels, the dots seem to be dense. It can be seen that this data set collects the information of OYO hotels in Guangdong Province.

Data preprocessing

Since this data set has more than 5,000 sample points, it is not a large data set, and it’ s an i.i.d. dataset, so I choose 60% of the data as X_train and y_train, and the remaining data sets are 50% as validation and test data sets.
There are 46 features be used by OneHot Encoder, they are classifications and can’ t be ordered. The remaining categories features “cancellation_policy” and “property_type” apply the OrdinalEncoder.
These 14 regression features are encoded by MinMaxScaler. And the “maximum_nights” feature, because this feature has a tail distribution feature, I chose StandardScater for coding.

References

[1] https://www.kaggle.com/code/budibudi/oyo-rental-price-prediction-in-china-003/data
[2] https://www.kaggle.com/code/budibudi/oyo-rental-price-prediction-in-china-003/notebook#Model

GitHub repository

My GitHub: https://github.com/AstrosiosaurQ7/data1030_mid_proj