My Kaggle Project Report

Ning Yang

November 30, 2021

1 Initial exploration

First, I used dim()and str() functions to check data structure and data types. There are 41330 rows and 91 variables in the data set and some types were wrong so I would change them in next steps.

Explore data

Since the topic needs to predict the price of airbnb houses based on different factors, I first use ggplot2 to visualize the prices in the existing data, we can see the number of houses in each rental range. I will keep this in mind, and take measures before modeling.

Check Missing Data

First of all, I would like to see which variables contain missing values. Because some variables containing hidden missing values cannot be recognized, such as N/A, I used replace_with_na_all()function to replace them.Then I checked the total number of missing values and which variables they belong to for the next step to fill in missing values for different types of variables. I did the same thing to scoringData.

I noticed that there were more than 200,000 missing values, which is quite a large proportion. However, some variables, such as license, have all values missing, so they can not be added to the model. Some variables may not have relationships with prices. This would be studied in the next parts.

2.Data Preprocessing

Change data types

For character variables, I changed them to numeric or factor types. For the variables "host_response_rate" and "host_acceptance_rate" that contain " %", I remove the "%" and convert it to numeric. I did the same thing to scoringData.

Deal with data variables

Next, I changed date variables to numeric. First parsed date variables into "year- month-day". Since it is meaningless to reflect only specific dates, I converted them to the number of days from the day the event occurred to the present. We found that the "first_review" and "last_review" variables are the time away from the present, so I created a new variable to reflect the length of time between the first review and the last review to detect its impact on the price.Then I did the same thing to scoringData.

Create Dummy Variables

I found that the amenities variable has many levels and many words, so I created dummy variables. Also, I did the same thing to scoringData.

3 Feature Selection

Since “name”, “summary”, “description” and character variables are meaningless, we can't say that they are classified as numeric or factor variables. I took them out before feature selection. Variables such as “is_business_travel_ready” and “requirements_license” have only one value, which has no effect on the price, so they were also taken out. There are too many missing values for variables such as “weekly_price”, reaching one-third of the total number of rows. It is meaningless to impute these variables with the average value, and the average value is only meaningful when the data is missing 5%. So I also deleted them.

Then I tried to use forward selection to select variables for models, but I made mistakes so it didn’t succeed. So after I perform linear regression, I selected some variables that have strong relationships with price to create other models.

4 Model Comparison

After I performed Linear Regression, Decision tree and XGBoost, I listed the results here.

Model from previous section	RMSE on training data	RMSE on test data	RMSE on Kaggle	
Model 1: Linear Regression	94.31256	98.38247		
Model 2: Decision tree	92.12506	90.07974	96.34362	
Model 3: XGBoost	41.30136	66.69868	63.39710	

5 Mistakes

First, I spent less time on cleaning data, but paid more attention on performing analysis.Some character variables in initial data were ignored, like “zipcode", “city”, “neighbourhood_cleansed”.
Second, I didn’t use feature selection correctly, so the variables I choose may not have significant relationships with price.
I ignored some hidden missing values such as “N/A”, “N A”, and “Not Available” when I first created a model.

6 Lessons learned

1. The importance of Random Forest is useful to check the variables importance and select features. To improve the performance, I will try it for future explorations.
2. The model of linear regression and decision tree cost a lot of time to run when the datase. is large.XGBoost model saves almost half of time.
3. Missing values may have multiple specifications , such as “N/A”, “N A”, and “Not Available”.

7Future directions

If I have more time, I will focus more on cleaning data and feature encoding.Because some categorical variables such as neighborhood_cleased, zip code have many levels, the method of creating Dummy Variables is not feasible. I should create Dummy Variables for the most important values (such as the top 95% of Importance) based on Feature Importance or the frequency of these values in the data, and all other values fall into a category of "other" . Moreover, I will think about how to deal with the mutual influence between features, sometimes the interaction will cause the performance to decrease instead.

Appendix - Code for XGBoost

In my project, xgboost model has the smallest rmse, so I list the code of xgboost model on the appendix.