Seattle Airbnb Open Data Analysis and Prediction

Airbnb is a global vacation rental online marketplace which offers arrangement for lodging, primarily homestays and tourism experiences. Many of us would like to use Airbnb when we travel around the world, since Airbnb is easy to order, usually less expensive than traditional hotels, and provide opportunities to connect travelers with good local hosts.

From the customer perspective, it is important to better understand the Airbnb price. Data scientists can help Airbnb and their customers to find the best rates! In this project, I used the Seattle Airbnb data to explore the Airbnb price in Seattle, and addressed three business problems.

1. Data

Since 2008, guests and hosts have used Airbnb to travel in a more unique, personalized way. As part of the Airbnb Inside initiative, this dataset describes the listing activity of homestays in Seattle, WA.

The following Airbnb activity is included in this Seattle dataset:

- *Listings, including full descriptions and average review score*

- *Reviews, including unique id for each reviewer and detailed comments*
- *Calendar, including listing id and the price and availability for that day*

2. Method

There are three main types of recommenders used in practice today and we User- based collaborative filtering system. Unfortunately, the data did not have very detailed information. The data is missing the transaction data, and the number of reviews. This would not have been enough data to provide an accurate content-based recommendation.

Why Price Prediction?

Price prediction is beneficial for Perfect Competition, a triple-win market.

- *Hosts have an intuitive opportunity to compare their services and amenities with competitors. As competition intensifies, the overall service quality and market - size of the rental housing market will be improved.*
- *Price prediction models are reliable references for data-driven decision-making. With price suggestions from Airbnb, hosts can make different business strategies.*
- *Airbnb users have more choices at lower price or higher quality.*
- *Airbnb can grow faster including attract more users and hosts, leading to operational and data center cost reduction and profit growth.*

Cons

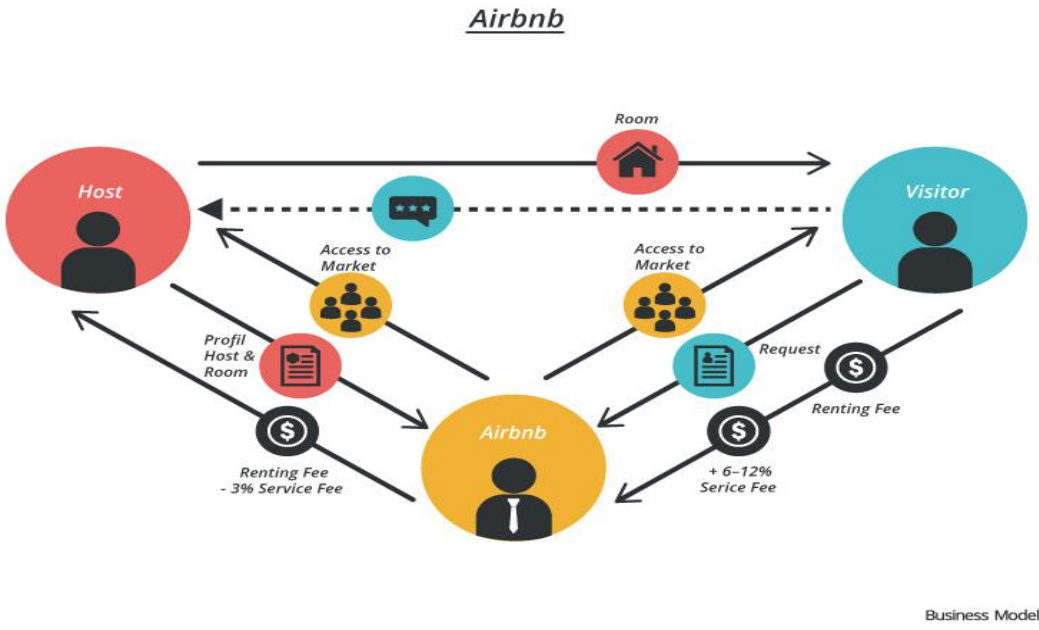
The model accuracy will be lower than expected due to missing some essential features.

- *The transaction data: The datasets did not contain transaction data.*
- *We don't know if the listings were booked or just unavailable.*
- *We will explore how to find the confirmed orders under the Exploratory Data Analysis part.*
- *The search engine data: The datasets do not contain searching histories.*

3. Business Understanding

- *Airbnb is a platform of accommodation which match the needs of staying and of lending.*

- *Their main source of income is the **fee for host**. Basically, as the number of transactions between the host and the guest increases, their profit also increases. So, it is important to their business and I expect it to be one of their KPIs.*



What can we do to increase the transactions? We considered three below questions to explore its way.

- ***How long is the period available for lending by rooms?***
  Is there rooms which is available all for years? or almost rooms are available on spot like one day or one week?
  Here, I want to know the trend in the outline of the data.
- ***Is there a busy season?***
  If the demand for accommodation is more than the number of rooms available for lending, it leads to the loss of business opportunities.
  So, I want to know whether is there the busy season. If this is true, we must create a mechanism to increase the number of rooms available for lending during the busy season.
- ***Are there any trends of popular rooms?***
  If this question's answer is true, we can suggest host to ways make the room easier to rent.
  In this part, I'll use machine learning technique.

4. Business Problem

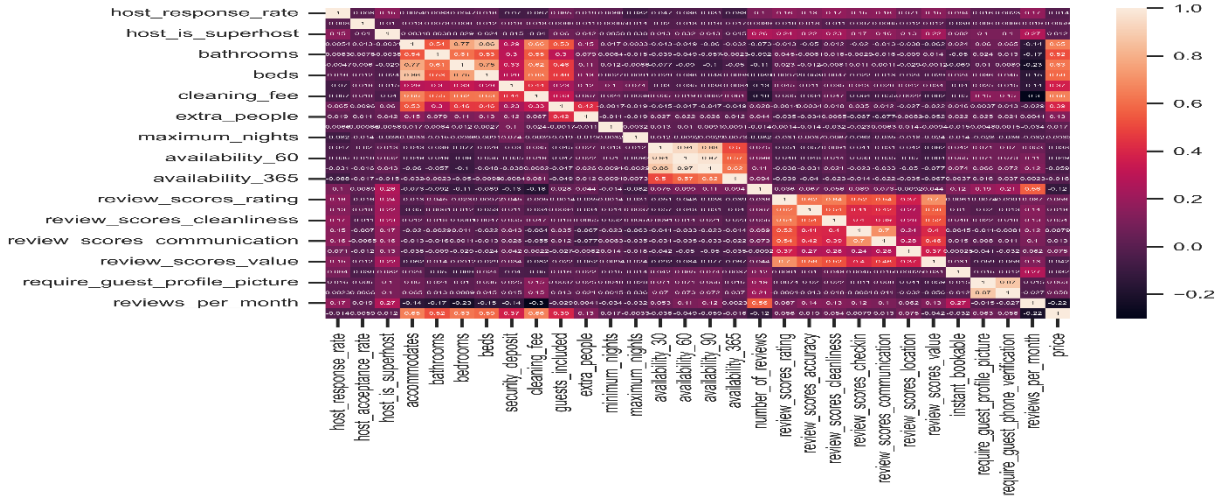
- *The key stakeholders in Airbnb are the guests and the hosts. Airbnb goes by the tag – "What you charge is always up to you". So, the hosts have freedom to set prices for*

their listings. Therefore, while listing a new property, hosts tend to compare similar properties and put a price close to similar listings. This may not be very accurate as the price depends on a lot of factors. So, the challenge for the host is to put an optimal price for the listing.

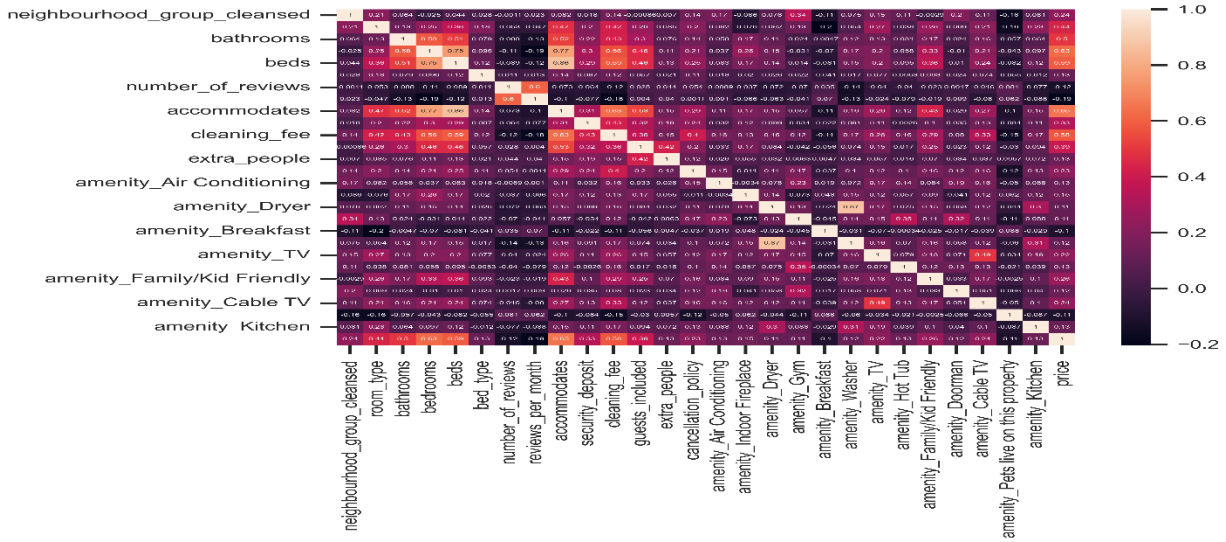
- *Airbnb generates revenue from the host by charging a service fee = 3% of (Nightly Rate + Cleaning Fee + Add. Guest fee) + VAT.*
- *So, when a host is guided with pricing tips, the possibility of the host putting a very low or high price is less. This helps in the host increasing guests which in turn generates revenue for Airbnb.*
- **Problem questions:**
- ● *Does Seattle Airbnb price show seasonal or daily variation?*
  ● *What are the most important ingredients/features for the Airbnb price?*
  ● *How to understand the most important features for the price?*
- **Solutions suggestions:**
- ● *Providing pricing tips for a new listing to assist hosts in accurately pricing their properties.*
- ● *What are the most important ingredients/features for the Airbnb price?*
  ● *Pricing tips combines the information provided by the host along with the neighborhood scores to incorporate the importance of location.*

5. EDA and Feature Engineering

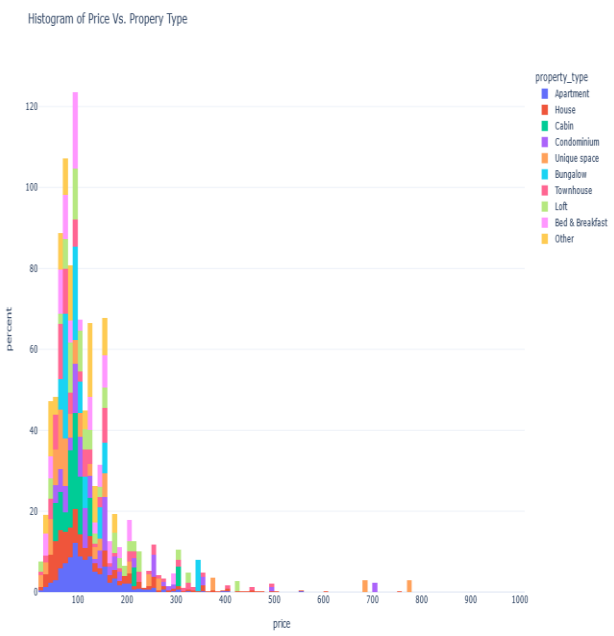
Pandas .corr() method calculates the Pearson correlation coefficient between every two features. And heatmap shows the relationship more clearly. Here, accommodates, bathrooms, bedrooms, beds, security_deposit, cleaning_fee, guests_included, extra_people have relatively higher correlation coefficient than other features with price.

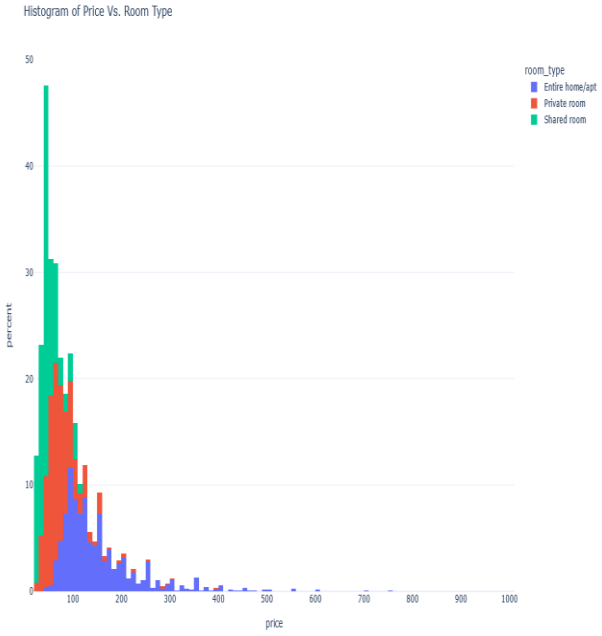


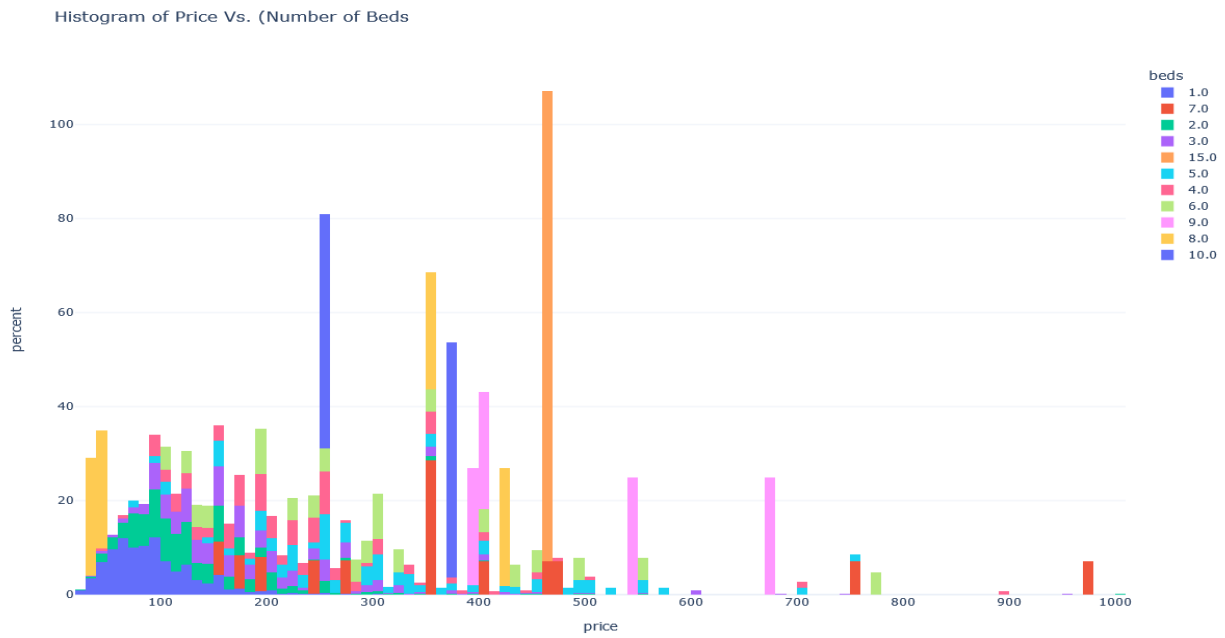
Since Pandas .corr() only calculates the numeric data. I performed target encoding then drew the heatmap again. room_type, neighbourhood_group_cleansed, bed_type, cancellation_policy were categorical data that cannot be calculated the correlation coefficient. But after target encoding, they do have a very high impact on price. For the amenities, TV, Hot Tub, Kitchen, Indoor Fireplace, Dryer, Family/Kid Friendly, Doorman, Gym, Cable TV, Washer, Air Conditioning are the more critical than other amenities.



Next, we took a glance at the following high correlation coefficient features: accommodates, bathrooms, bedrooms, beds, security_deposit, cleaning_fee, guests_included, extra_people, room_type, neighbourhood_group_cleansed, bed_type, cancellation_policy







6. Algorithms & Machine Learning

The models in this project are XGBoost and LightGBM.

The baseline can provide an insight into the performance of different data preprocessing strategies, such as encoding methods. Here, I chose target encoding. First, it had a better performance than ordinal encoding. Second, we already knew the categorical data have potential levels for a different price. e.g. different roomt_type has different price histogram.

Features marked as ML1 were defined by Airbnb for Machine Learning models. Then, as you can see, ML1 + ML2 is better than ML1, which means we found more features that were useful for Machine Learning.

Model Tuning

- *The Hyperparameter tuning platform I used to be Optuna. We implemented a logger to write the tuning results in the local log file. After all tunings are finished, the program will send an email to my mailbox with the best hyperparameters.*
- *To enable this feature, go to configure your Gmail first.*
- *P.S: If your computer does not support GPU acceleration, uncomment code For CPU and comment code for GPU.*
- *If you want to train your model, DO NOT RUN the following code in this notebook. Instead, make another notebook for model tuning.*

Model Blending

After hyperparameter tuning, we have a set of better hyperparameters for XGBoost and LightGBM. Then I performed a model blending for a better ML performance.

Model Stacking

Model Stacking is a way to improve model predictions by combining the outputs of multiple models and running them through another machine learning model called a meta-learner.

After hyperparameter tuning, we have a set of beter hyperparameters for XGBoost and LightGBM. Then I performed a model blending for a better ML performance.

7. Predictions

- *The model stacking result is not much better than any single model since I only used two models with target encoding. You can combine different models with different encoding strategies even different features to improve the overall performance.*
- *For instance, you can combine XGBoost, LightGBM, and CatBoost with one-hot encoding, target encoding, ordinal encoding, and polynomial encoding. Then you have 3×4 models for model stacking*

8. Summary for Key Concerns

- *How long is the period available for lending by rooms?*
  The histogram of maximum nights shows that there are two groups.
  One is a listing that can be used at spots such as the maximum number of nights within a week.
  The other is a listing that supports a wide range of stay from the super long-term stay of the maximum number of stays for three years or more and the minimum number of nights for around two days to the spot use.
- *Is there a busy season?*
  The answer is Yes.
  Apart from the increase in the number of Airbnb users, there was definitely a timely increase in the number of reviews at the same time each year.
  It is thought that it is about one month around early September and overlaps with the summer vacation time. It is important that the number of properties that can be provided at this time exceeds demand.

- *Are there any trends of popular rooms? I could not derive it from my analysis. However, I learned that the score improves by logarithmic transformation.*

9. Conclusions

To help customers better understand the root causes of the price and get better rates. We found that the Airbnb price shows a seasonal variation. We built a price predictor based on features from the listings table. Among many features; we explored the most important features for the price prediction, and studied the correlation between these features and the price (such as weekly price*,* number of bedrooms*,* bathrooms*,* reviews per month*,* cleaning fee, room types, *and* neighborhood*). These features can be considered as the most important indictors for evaluating the Airbnb price.*

The model stacking result is not much better than any single model since I only used two models with target encoding. You can combine different models with different encoding strategies even different features to improve the overall performance.

For instance, you can combine XGBoost, LightGBM, and CatBoost with one-hot encoding, target encoding, ordinal encoding, and polynomial encoding. Then you have 3×4 models for model stacking.

Hosts could adjust the following features for improving their competitiveness.

- ***Decrease the following fees***
    - *Cleaning_fee*
    - *Security_deposit*
- ***Policies Change Suggestions***
    - *Cancellation_policy*
    - *Bed_type*
    - *Beds*
    - *Guests_included*
    - *Extra_people*
    - *Accommodates*
- ***Take Extra Care of the Following Amenities:***
    - *Indoor Fireplace*
    - *Cable TV*

    - *TV*
    - *Doorman*
    - *Dryer*
    - *Air Conditioning*
    - *Gym*
    - *Family/Kid Friendly*
    - *Kitchen*
    - *Washer*
- ***Hard to Change but Good to Update (Remodel) (Cost-Benefit-Analysis is still required):***
    - *Neighbourhood_group_cleansed*
    - *Room_type*
    - *Bathrooms*
    - *Bedrooms*

10. Suggestions and Future Improvements

Although we don't have the transaction data, and the number of reviews did not play an essential role in price prediction, I'm sure such features will influence the rank on the search engine.

Airbnb uses machine learning models to intensify the competition. They can provide Price Tips, Smart Pricing, and improvement suggestions to the host. As a host, you can combine the models and good strategies to maximize your competitiveness and profit. Airbnb users are also benefited from the perfect competition.

11. Credits

Thanks to Tony Paek, my Springboard mentor.