Revenue Prediction and Price Optimization - Airbnb

Jiayang Nie[1]

[1]Department of Statistics, University of California, Berkeley, CA

[*]jnie@berkeley.edu

May 11, 2022

1 Introduction

1.1 Problem Description

This project is from the point of view of Airbnb hosts, both potential and existing. For the hosts who haven't entered the market, our model can help them gain a concrete understanding of the potential return-on-investment based on their real-estates characteristics. This includes an analysis of the impact of house size and house type on return-on-investment. For the existing hosts, our project will give them a tool to set an optimal renting price.

To be more specific, this project analyzes return-on-investment of Airbnb hosts in Los Angeles in 2021 through Airbnb dataset. In this project, we are researching on all the different covariates that may affect revenue and build a model to predict revenue. We first observe that revenue can be decomposed of price and demand. So, we build a model to predict return-on-investment. We will define demand and return-on-investment in the following subsection. The models we considered are: ridge regression and XGBoost models. We performed cross-validation to tune hyper parameters and examine the error to decide which model to use. By the model, we are able to answer the following questions:

1. What to improve if you want to increase return-on-investment?
2. How does location affect return-on-investment?
3. How does demand changes with booking price, assuming everything else is fixed? (Price Elasticity)
4. Where and what type of real estate to purchase for new Airbnb hosts?
5. For current Airbnb hosts, how do they set an optimal price?

Questions 1, 2, 4, and 5 are crucial to investors while question 3 has economical values.

1.2 Definitions

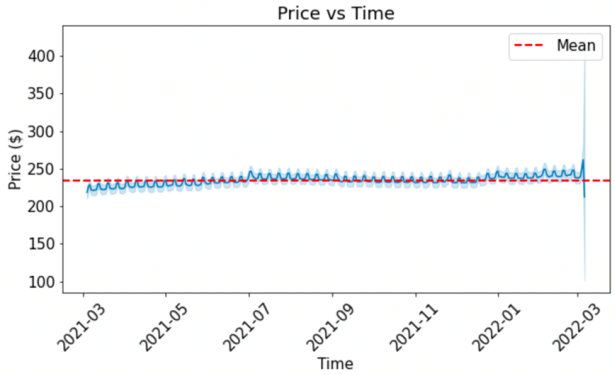
Definition 1.1. *Let $D_i(t) = 1$ if Airbnb i is occupied by a party of guests at date t and $D_i(t) = 0$ if Airbnb i is not occupied at date t*

In the raw dataset, we do not observe $D_i(t)$. Instead, we observe $\hat{D}_{i,t'}(t)$ where t' is the time the data was collected and t is sometimes in the future in compare to t'. Thus, $\hat{D}_{i,t'}(t)$ is subject to change and the two quantities are not necessarily equal unless t' = t because booked guests might cancel booking and new guests may book after t'.

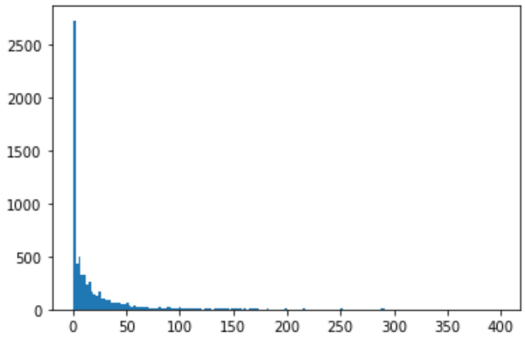
Definition 1.2. *$OR_i = \mathbf{E}(\mathbf{D_i})$ denote the occupancy rate for Airbnb i. Empirically, we estimate $\hat{OR}_i(n, t') = \frac{1}{n}\sum_{t=t'}^{t'+n} \hat{D_{i,t'}}(t)$ with fixed n, t. n is the number of days, and t is a date. Take n = 30 and t' = 2021.12.21 as an example, $\hat{OR}_i(n, t')$ means, at 2021.12.21 and we look into the future 30 days, record how many days are booked and then divide it by 30.*

At each time point where the data was scraped, we compute the occupancy rate, or the demand, for a thirty-day interval. As InsideAirbnb only gives four time points in 2021, the general occupancy rate for each Airbnb listing is the average of the four occupancy rates we have. We will explain the reason for pooling in the next section. The four time points are: 2021.3, 2021.6, 2021.9, 2021.12.

Definition 1.3. *Booking price $Price_i$ for Airbnb i is the average of the four reported booking prices for Airbnb i.*



(A) Seasonal trend of booking price



(B) Distribution of standard deviation of booking price for each host

Figure 1: Proof of Assumption 1.1

Definition 1.4. *Return-on-investment:* $ROI_i = \frac{OR_i * Price_i}{Sales_i}$ *where* $Sales_i$ *is the Sale price for Airbnb* i *in real-estate market in year 2021.*

The interpretation for ROI is the yearly return rate. The inverse of ROI is the number of years expected for the Airbnb investors to recover its capital from rent.

Assumption 1.1. $Price_i$ *remains fixed across year 2021.*

As can be seen from Figure 1 that the seasonal variation of $Price_i$ is generally small. Also, from the distribution of standard deviation of booking price for each host, we observe that many hosts did not change price at all, with most hosts change price within 50 dollars. Thus, it is justifiable to make Assumption 1.1.

1.3 Return-on-investment Decomposition

It is clear that ROI can be decomposed into demand, price and housing price. Furthermore, once demand is predicted, ROI can also be determined. Therefore, it seems that building a model to predict demand is sufficient. However, through building the model that predicts ROI, we will be able to look into the components that directly affect ROI. Thus, there are actually two models: 1. the model that predicts ROI, 2. the model that predicts demand. To know what factors are important to ROI, we will use model 1. But a downside of model 1 is that it sometimes overestimate demand. For instance, if we fix everything else and increase the price, then model 1 will always predict ROI to be bigger. This is not ideal and thus we need model 2 to set an optimal price.

2 Data Description

The main data was collected from *InsideAirbnb*. The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb site. The data has been analyzed, cleansed and aggregated where appropriate to faciliate public discussion (Claim in Inside Airbnb Website). The free data from *InsideAirbnb* only contains four dataframes for March, June, September, December 2021 as mentioned in section 1. To gain access to achieved data for data prior to 2021, we would have to pay 500 dollars and we decide to do data analysis without the achieved data for this stage.

The dataset we have can be treated as a (11375 x 72 x 4) 3-dimensional data. 11375 is the number of observations, 72 is the number of covariates and 4 is the time-wise dimension. That is, we have four (11375 x 72) tabular tables for March, June, September and December in 2021. The main variables include the house type, the number of bedrooms, the number of accommodates, the booking price, the occupancy rate, the neighborhood, the review status, longitude and latitude. A reasonable assumption is that most of those features will remain fixed across the four time points and only the price and demand may change significantly. In the previous section, we have proved that price can be assumed to be fixed. Thus, the only variable that will change is demand, and of course, ROI. This is the variable of interest to be predicted. One variable that is missing is the housing price. We retrieve the housing price from Zillow based on zipcode and number of bedrooms. For example, for an Airbnb listing with zipcode 93111 and 3 bedrooms, we look for the average sales price for a real estate with the same zipcode and number of bedrooms on Zillow in 2021. That will be assumed as the housing price for this Airbnb. This is not exactly accurate but is the best method we can find with zero monetary cost.

The features we are interested in are: property type (e.g. apartment or house), room type (e.g. entire home or shared room), accommodates (maximum number of guests), number of bathrooms, number of bedrooms, review score, location, and price per night. The variable to be predicted is the occupancy rate and ROI (they are equivalent).

The hypothesis is that the higher the price, the lower the demand. However, the relationship between price and ROI should be a curve, with a maximum point somewhere. We also think entire home has more advantages. Better review status should also have a positive effect on ROI and demand.

2.1 Data Cleaning

We observe that many Airbnbs are receiving great reviews, has low price, and has been doing Airbnb for many years, but have 0 booking rate for 2021, and thus have 0 ROI. We make the assumption that they did not enter the market at all. Thus, we filter out all the Airbnbs that have 0 demand.

As we mentioned in the previous section, to make the estimator more robust and has less variance, we use an average of $\hat{OR}_i(30, 2021.3)$, $\hat{OR}_i(30, 2021.6)$, $\hat{OR}_i(30, 2021.9)$, $\hat{OR}_i(30, 2021.12)$ as the final definition of an estimator of OR_i.

Besides, minimum nights is an interesting feature. Over 50 percent of airbnb hosts set their booking requirement to be at least 30 nights. We make the assumption that those Airbnb hosts are doing long-term rental, which we decide to not include in our mdel.

Another thing to note here is that the reasons for a day not being available for booking could be: 1. the day is already booked by another party, 2. the host refuses to open the Airbnb for guests that day. To avoid situation 2, we filtered out Airbnbs with $\hat{OR}_i(90, t') = 1$ for a given t'. Here, we assume that if Airbnb is fully booked for the next consecutive 30 days, then this host probably did not open his/her Airbnb for booking. In summary, we filter out the Airbnbs that has 0 or 1 occupancy rate, and have minimum nights over 30 nights.

The filtering and pooling method, indeed makes sense and is supported by empirical evidence. It is natural to think that reviews-per-month should be positively correlated with occupancy rate. Before the filtering, corr(reviews-per-month, $\hat{OR}(30, 2021.12)) = 0.0024$. After filtering, corr(reviews-per-month, $\hat{OR}(30, 2021.12)) = 0.2188$. After we averaging out to get the pooled occupancy rate, corr(reviews-per-month, Pooled Occupancy Rate) = 0.4291. This proves that the filtering method indeed works.

After all the filtering and merging, the dataset now has a dimension of 5487 observations, meaning over a half of the data is filtered out.

2.2 Data Visualizations

In this section, we examine the relationship between of various variables with demand and ROI. Firstly, we examine the location's effect on ROI.

From the map of ROI [Figure 2], we can see that areas around UCLA, Beverly Hills, Santa Monica, and Malibu tend to have high ROI. However, this is only observational results, which we cannot make causal claims. There could some other factors that affect both location and ROI.

House size and type are also important features: Airbnb with too few bedrooms might be popular because guests usually appear in party and need two or three bedrooms; Airbnb with too many bedrooms may also experience lack of demand due to the high price.

Besides house type, the type of amenities and property is also useful. To see what amenities information looks like in our dataset, takes the amenity information for a random airbnb listing as an example: ["Stove", "Cooking basics", "Bathtub", "Kitchen", "Hangers", "Essentials", "Iron", "Oven", "Heating", "Hair dryer", "Long term stays allowed", "Hot water", "Lock on bedroom door", "Dedicated workspace", "Smoke alarm", "Refrigerator", "Shampoo", "Host greets you", "Dishes and silverware", "Carbon monoxide alarm", "Wifi", "Free street parking"]. We take the top 100 frequent words, generate 100 dummy variables accordingly and apply dimension reduction tools to reduce the dimension to 20. We will discuss the methodology in the Method Section. We also apply the same process to property type. Property type includes 'Private room in apartment', 'Room in hotel', 'Entire guesthouse', 'Entire bungalow', 'Private room in island', 'Entire house', 'Private room in house', etc. After we transforming all categorical variables, the total number of variables in our model is: 24.

Lastly, in all economic models, unit price is an important component. Thus, we divide the price per night by the number of accommodates allowed to generate the new feature: price per person per night. This is the unit price and is independent of most of the covariates.

To have a general visualization of all the variables, we construct a correlation pair plot, as [Figure 3]. From the Pearson correlation, we can see that bedrooms, beds, price per person have strong correlation with ROI and demand. We expect to see them in our modelling evaluation process.

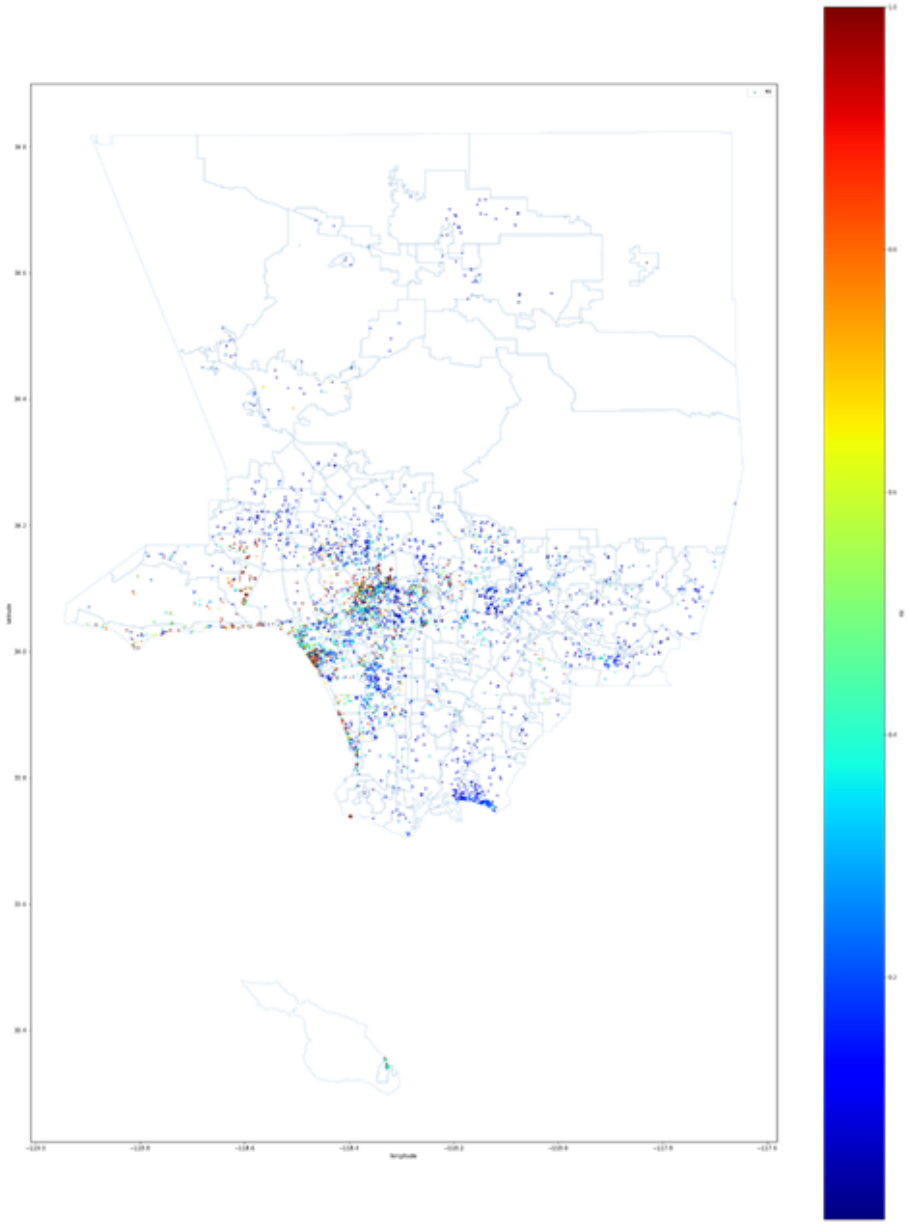


Figure 2: Observational ROI Map

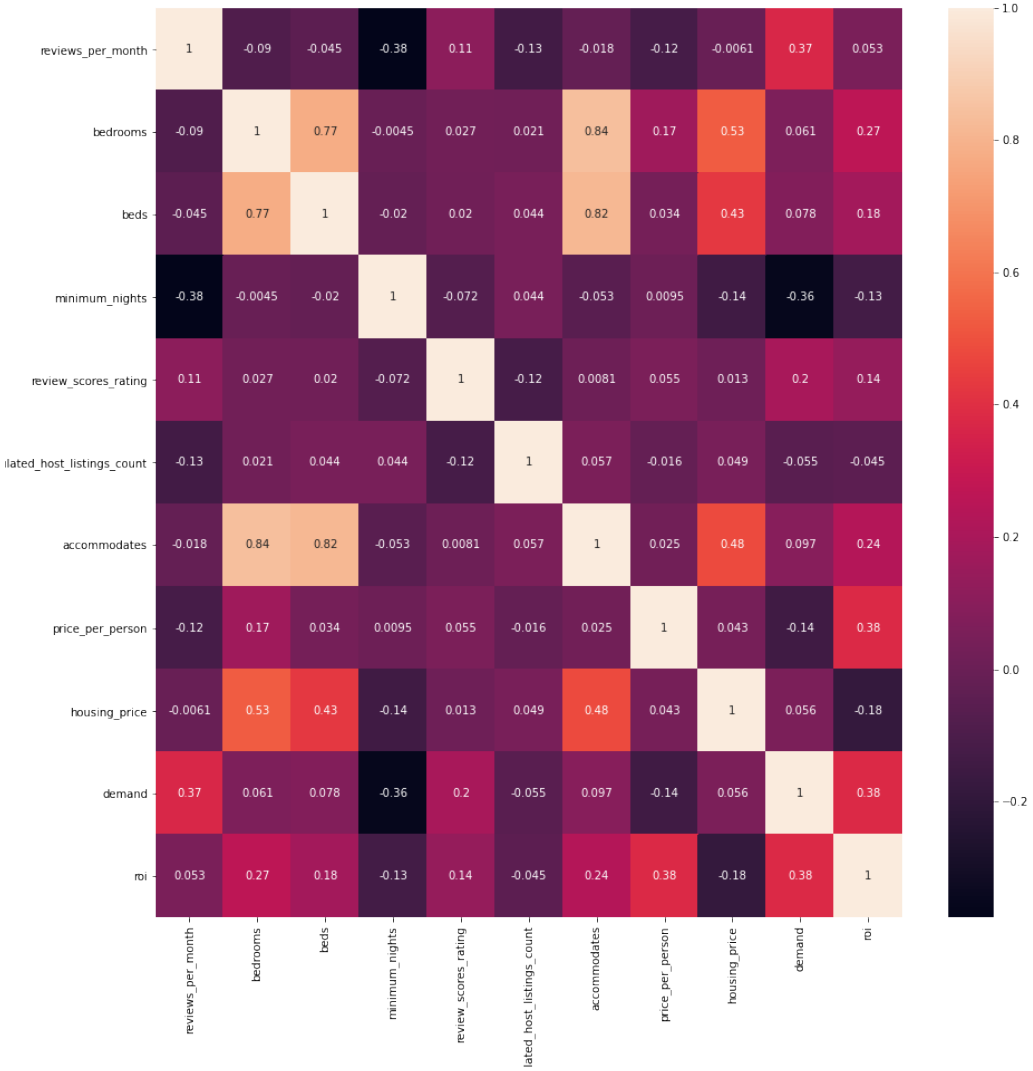


Figure 3: Correlation Pairs

We are also aware that quantitative modeling does not always work and beat the market. For instance, the downfall of Zillow's attempt to quantitatively predict housing price is a lesson to learn. However, rental market seems to be more robust than the housing market, as renting price will not likely to change significantly. The only risky factor in our model is the occupancy rate, which ranges from 0 to 1. Therefore, we believe our model is safe to use.

3 Methods

Let us define $X \in R^{24}$ as a vector containing all features except for ROI and occupancy rate. Let us define $y_{roi} >= 0$ as the return-on-investment value, and $y_{demand} \in [0, 1]$ as the occupancy rate. Also, $y_{roi} = y_{demand} * X_{price}/X_{housingprice}$. We assume that there exists a function f such that $f(X) = y_{roi} + \epsilon$ and $\epsilon \sim Normal(0, \sigma)$ for some σ. There also exists a function $\hat{f}$ which is a subset of f such that $\hat{f}(X) = y_{demand} + \bar{\epsilon}$. This is because that once $\hat{f}$ is known, f will also be determined as roi is composed of demand.

For ridge regression, the model is

$$y_{roi} = \beta^t X + \epsilon \tag{1}$$

where

$$\hat{\beta} = argmin[(y_{roi} - \beta^t X)^2 + \lambda \beta^t \beta] \tag{2}$$

It is a linear model that has L2 regularization. The purpose of including L2 regularization is that many covariates have high correlation, which leads to heterogeneity. Having a penalization term could help reduce the variance of estimate. One problem with linear regression is that the predicted value may go pass 0, which does not make sense. Therefore, we truncate the model such that any value predicted to be below 0 will be set to 0.

For XGBoost models, it is essentially an ensemble model.

$$y_{roi} = \sum_{i=1}^{N} a_i * g_i(X) + \epsilon \tag{3}$$

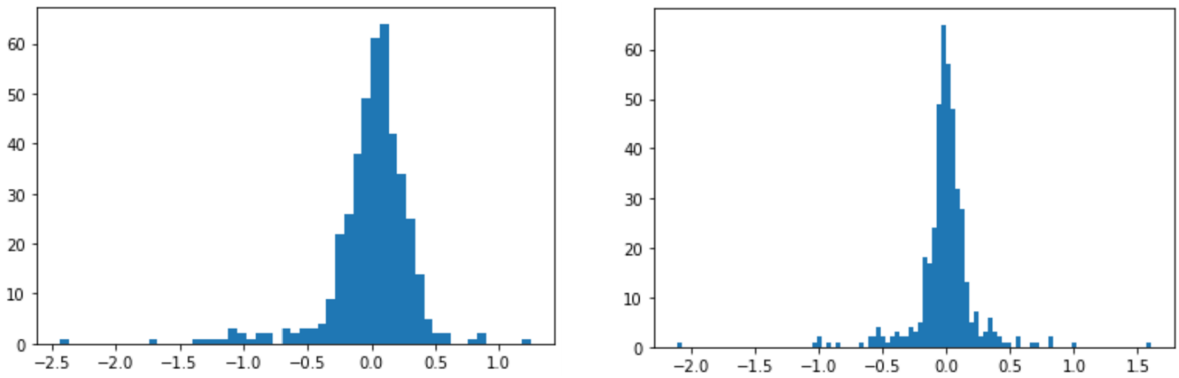
where g_i is a weak learner, and a_i is the weight of the weak learner, and N is the number of weak learners. Each weak learner is constructed based on the previous weak learners to boost the model performance.

We first split 90 percent of the data into training set and the other 10 percent into testset. We ran 10-fold cross-validation on training set to select λ for ridge regression, max depth, number of estimators, and learning rate for XGBoost. Then we test the best models' performance on the testset. The number 10 is an arbitrary number we choose such that each subset of the dataset is not too small. [Figure 4] shows the error comparison. Clearly, XGBoost is the better model and it makes more sense as tree-based models capture non-linear relationship. Another good feature of tree-based models is that, in theory, every prediction is a weighted average of a basket of values in tree nodes. Therefore, its range will always be the same as the range in training dataset. This means that the predicted ROI will not fall below 0. This is a property that linear regression does not have.

This model is to predict ROI directly and it is easier to interpret the features that affect ROI. It estimates f, and thus is equivalent to a model that estimate the demand, $\hat{f}$. However, in this model, its predictive power of implicitly predicting demand is not good enough. Therefore, we also create a model to predict demand (estimate $\hat{f}$). The reason for still including the model to predict ROI is that it is easier to interpret the features that affect ROI, which is the target of this project.

Another part of this project focuses on optimizing price. Let $\bar{p}$ denote the optimal price.

$$\bar{p} = argmax_{Xprice}(\hat{f}(X_{-price}, X_{price}) * X_{price}) \tag{4}$$

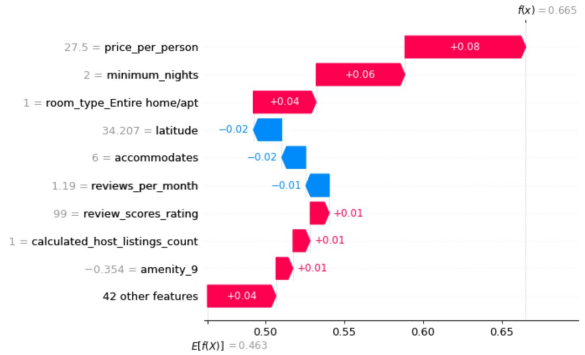


(A) Error Distribution of Ridge Regression

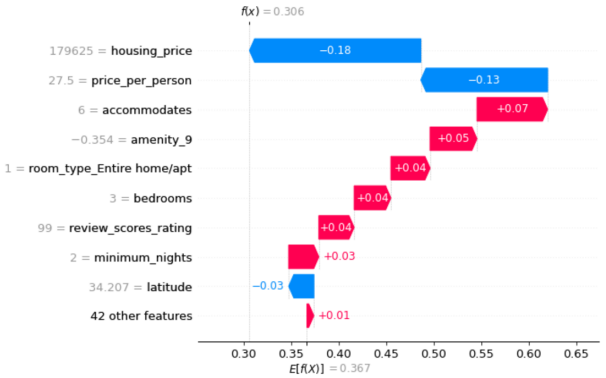
(B) Error Distribution of XGBoost

Figure 4: Error Comparison

Essentially, by predicting demand, and estimating a demand function of price, we can derive the optimal price, assuming a global maximum exists.



(A) Average Shap values of XGBoost that predicts demand



(B) Average Shap values of XGBoost that predicts ROI

Figure 5: Average Shap Values

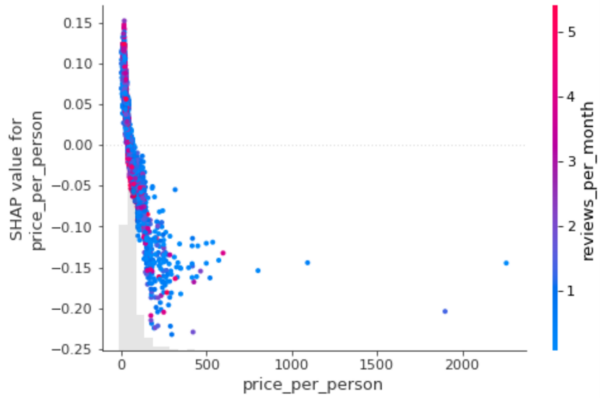


Figure 6: Shap values for price per person in predicting demand

4 Results

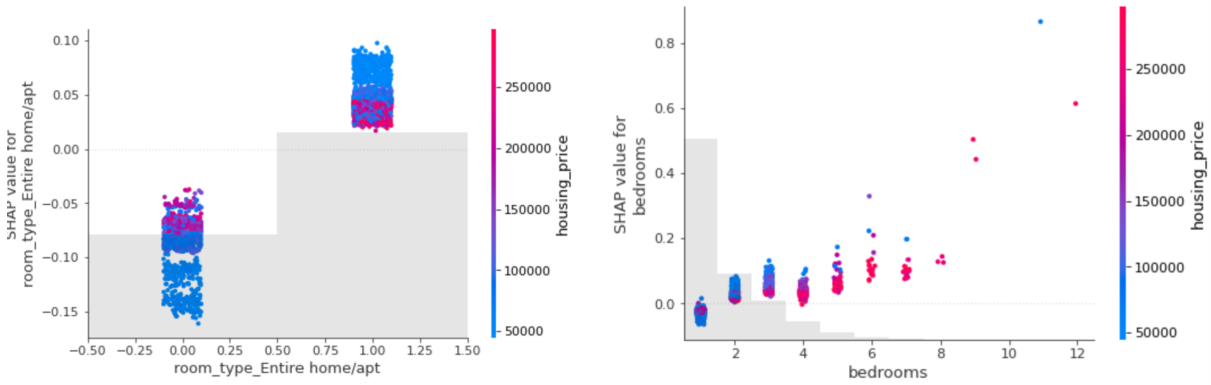
4.1 Improve Return-on-investment

To analyze the model performance of XGBoost, we refer back to Figure 4. In this case, the error is approximately symmetric can left skewed

We report the average Shapley(Shap) values of the XGBoost model in Figure 5. Shap values is essentially a tool for measuring feature importance. The difference between Shap values and the ordinary permutation feature importance is that permutation feature importance is based on the decrease in split criteria while SHAP is based on the magnitude of feature attributions to final prediction.

Average Shap value can be interpreted as: in compare to a model without the information of feature k, what is the average change in prediction if the model is given the new information about feature k across the training dataset. For instance, for Figure 5 (A), price per person has an average of 0.08 percent increase in predicting demand. However, this does not mean that the higher the unit price, the more the demand. This just means that including a unit price feature into our tree-based model tends to increase the prediction for 0.08. Here, the x-axis for (A) is percent in demand, and x-axis in (B) is percent in ROI. The average Shap values follow a vertical descending order. So, the top features are the most important features.

Figure 6 shows the shap values of price per person for each single observation in training



(A) Shap values of Room Type

(B) Shap values of number of bedrooms

Figure 7: Shap Values for Room Type and Number of Bedrooms in Determining ROI

data set. We can interpret it as: in compare to a model without the information about price per person, the impact of the final predicted demand when telling the model the price per person for that observation. Here, the Shap value fixes everything else and looks at the impact of price per person. Therefore, it is possible to make causal claims. Note the trend shows that the higher the unit price, the lower the demand. However, the average of all the values is 0.08.

We also find that the factors that are important for demand are minimum nights (the minimum number of nights required by the host for the guests to book), room type of entire home, and reviews per month. Clearly, entire homes are more popular than any other room types.

For ROI, the important features are housing price, price per person, number of bedrooms, room type of entire home and reviews scores rating. This is unsurprisingly similar to the demand analysis because ROI is composed of demand.

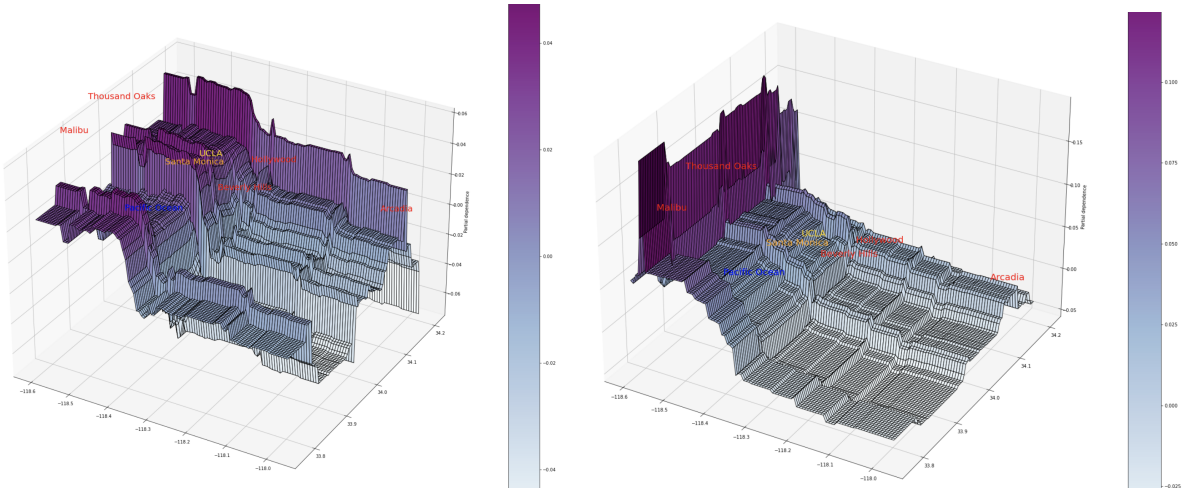
We first interpret the housing price effect. Clearly, keep everything else fixed, and specifically price per person fixed, the higher the housing price, the lower the expected ROI. This is reasonable because if price per person is the same, then the higher the investment cost, the lower the return on investment.

Then, we interpret the effect of price per person in ROI. As we see above that price per person is also important for determining demand. It also makes sense that it affects ROI significantly.

We now focus on the effect of room type and number of bedrooms. In Figure 7, we can see that while keeping everything else the same, Airbnbs being claimed to be an entire home tend to yield higher return-on-investment. If we look at the y-axis on the right, which shows housing price in different colors, we may find that the lower the housing price, the higher the ROI for entire-room Airbnbs. Furthermore, three-bedroom Airbnbs tend to yield higher ROI as well.

4.2 Location effect on return-on-investment?

Then, we look at the geographical effects on demand and ROI. We use a partial dependence plot to see this. A partial dependence plot is another model interpretation tool like Shapley value. It also keeps everything else fixed, and check for the average change in one feature. The difference between Shapley value is that partial dependence does not compare with a model



(A) Partial Dependence Plot of Geographical Effect on Demand

(B) Partial Dependence Plot of Geographical Effect on ROI

Figure 8: Partial Dependence Plot of Geographical Effects

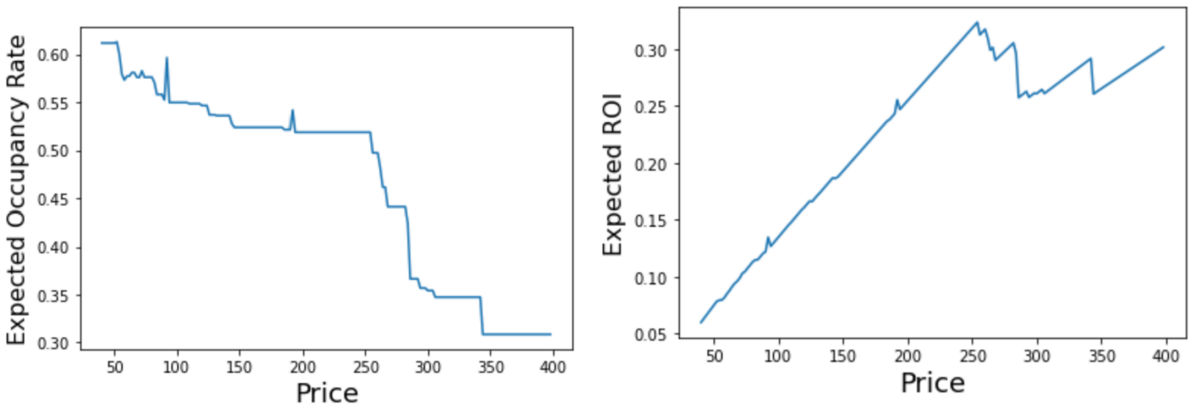
without information about that feature. Instead, it directly computes the average effect of that feature being at a specific value. In Figure 8, we see that UCLA, Malibu, and Santa Monica all have a strong positive effect on determining both demand and ROI. Note that a 3-D partial dependence plot of longitude and latitude is like a square on the map. This square inevitably includes some weird locations like the Pacific Ocean, which we shall ignore.

4.3 Pricing Model

We also briefly develop a model to set optimal price, as described in Section 3 equation (4). Figure 9 gives an showcase example. It is a real-world Airbnb host with 36 percent occupancy rate, 13.2 percent return on investment, and 160 dollars per night. We feed different price to the demand model and see how it changes with price. We then can see that an optimal price is between 200 and 250. Certainly, there are some noise in the graph like the sudden jumps or drops. We will need more data and finer considerations of modelling technique to give a better pricing model. But we show that it is possible.

4.4 Summary

In summary, from the Shap values, we recommend new Airbnb hosts or current Airbnb hosts to rent out Entire room, with three bedrooms. From the partial dependence plot, we have advice for real estate investors targeting Airbnb: try purchasing houses/apts at UCLA, Malibu, and Santa Monica and avoid South Eastern of LA in general. The other unsurprising suggestions from our model is to suggest hosts to earn higher review rating and more reviews. These answer questions: 1. What to improve if you want to increase return-on-investment? 2. How does location affect return-on-investment? 3. How does demand changes with booking price, assuming everything else is fixed? (Price Elasticity) 4. Where and what type of real estate to purchase for new Airbnb hosts? The last part about pricing model answer question 5. For current Airbnb hosts, how do they set an optimal price? We also discuss some of the criticism and downsides along the way. Next steps include a finer tuning of the model and more data



(A) Expected Occupancy Rate v.s. Price

(B) Expected Return-on-investment v.s. Price

Figure 9: Pricing Model

collected for a better pricing model.