AIRBNB PRICING

DECEMBER 5, 2016

HAN DONG HD324
ZHENGYU LONG ZL478

I. Introduction

Airbnb is an online marketplace that enables people to list, find, and then rent vacation homes for a processing fee. It connects hosts (vendors of rooms/ accommodations) and travels via its website, enables transactions between two entities by charging a ‘service fee’ without directly owning any rooms. In this process of business model, the pricing of the accommodations is critical as it is important for stakeholders of both hosts and customers.

The aim of our project is to study the price of Airbnb property, answering questions such as “what affects the price of an Airbnb property”, “what would be the most important feature of a property in terms of price”. The potential audience of this project are new Airbnb property owners. By solving this problem, we aim to help them set appropriate prices for their properties. Additionally, while owners apparently benefit from this project, lessees may also benefit from this project since they are renting at an appropriate price. The appropriate solution by this project would provide informative solution for both property owners and the customers, providing a win-win situation for both stakeholders. In this project, we assume the prices on the listing are appropriate in a sense that they have been rented out several times.

Our project scope will mainly focus on studying the city of San Francisco at United States. San Francisco is the birthplace of the Airbnb business, and it is one of most popular cities in its business. It includes more than 8000 records of valid listings every year. The insights from this problem could also be useful for other cities.

II. Data

The data from “Inside Airbnb” provides detailed information on the listing of accommodations, host information and historical reviews. The original data consists of 95 columns and 8620 observations. In our preliminary analysis, some features as the demographical information of owners are excluded, as in rare circumstances would these features be a significant effect on price. The website also provides a simpler version of the data which only consists of 16 features. According to the analysis conducted during the midterm, neighborhood and room type are more important than others. Thus these two features should be selected. For observation with missing categorical values, we deleted the whole observation. For observation missing numerical values, we assign average to it. However, while some features might seem important, such as amenities, many observations contain empty information on it. Thus, such feature should be dropped since the remaining data is insufficient. Other features, such as summary, contain pure text information which are substantially more difficult than others and we believe are beyond the scope of this class. The only feature that contains text information yet we still selected is house rules, and we extracted two pieces of information from it: whether pets allowed and whether smoking allowed. Although it may contain more house rules, it would be too difficult to extract those pieces of information. The response of the model is not price but appropriate price, so we want to eliminate some inappropriate values, that is, price that is too high or non-positive. Note that while some prices are high, however, are reasonable. Intuitively, longitude and altitude are redundant with neighborhood, so in order to prevent collinearity, we only kept neighborhood for simplicity. Thus, out of 95 features and 8620 observations, we selected 30 features and 8646 observations.

Owner info	**Property info**
length of host`s account	house rules: smoking and pets allowed or not
is host local	Neighborhood
host response time	location is exact or not
host response rate	property type (house, apartment, boat, etc.)
host listings	room type (entire, private, shared room)
host verification methods (email, phone, facebook, etc.)	number of people can accommodate
host has profile picture or not	number of bathrooms
host total listings	number of bedrooms
host identity verified	number of beds
	bed type (real bed, futon, airbed, couch, pullout bed)
	security deposit
	cleaning fee
	extra people fee
	minimum nights
	maximum nights
	number of reviews
	review scores

	whether requires id
	instant bookable
	cancellation policy
	whether requires phone verification
	requires guest profile picture

In order to test the performance of our model, the data was randomly split into a training set which consists of 2/3 of the data and a test set which contains the rest of the data. We will use 5 methods to fit a model and find out which features are more important than the others.

III. Methods

1. Loss Function

We fit linear regression models using different loss functions and regularizers. The mean square errors are reported for comparing purposes. Cross validation methods were used to determine the best parameter for both the loss functions and regularizers.

The first models are linear regression with quadratic loss with l1 and l2 regularizers. Lasso and ridge regression, while easy to compute, provide powerful tools to do feature selection as well as prevent overfitting.

As discussed before, although some prices are much higher than others and some prices are much lower than others, we still consider them appropriate if they are within the tolerable range. However, tolerable range is hard to define and mostly by intuition. Thus, it`s possible that we put some outliers in the training data. Huber loss is less sensitive to outliers, so we tried Huber loss with l2 regularizer. The detailed mean square errors are given in Table 3.1.1.

Loss Function	Regularizer	MSE
Quadratic	l1	129420
Quadratic	l2	129567
Huber	12	168860

Table 3.1 1

Based on mean square error, Lasso has the best performance. Analyzing the coefficients of Lasso confirmed our conclusion from the midterm report that room type and neighborhood are among the important features that affect prices. Other important features are given in Table 3.1.2.

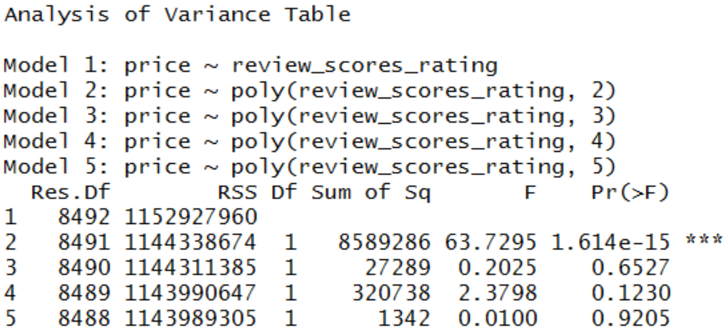
Owner info	**Property info**
host response time	Smoking allowed or not
host identity verified	Neighborhood
host verification methods	property type (house, apartment, boat, etc.)
	room type (entire, private, shared room)
	number of people can accommodate
	bed type (real bed, futon, airbed, couch, pullout bed)
	number of bathrooms
	number of bedrooms
	number of beds
	cleaning fee
	minimum nights
	review scores
	whether requires id
	instant bookable
	cancellation policy

Table 3.1.2

Out of 9 categories of owner information, only 3 are important while out of 22 categories of property information, 15 are important. These features are indeed important from our point of view. For example, number of people can accommodate, it makes sense that the more people that an Airbnb property can accommodate, the higher the price in general, so this feature can affect prices. On the other hand, maximum nights is not important, because in reality, most Airbnb lessees seek short term rental so maximum nights should not affect anything at all but rather minimum nights should matter. Indeed, minimum nights is among those important features. Therefore, we conclude that people are more concerned with the property itself when judging prices rather than its owner. Again, this is the truth in reality.

2. Polynomial Regression and General Additive Model

Although linear model is simple, it suffers from inflexibility. We first determine the appropriate order of features one by one. Since polynomial regression is much more flexible, there is a potential danger of overfitting. In order to build a model with enough flexibility to capture certain trends of the data and not to be to flexible to capture the noises, we decided only to transform two features: number of reviews and review score. This is because these two features are significant in our analysis of the simple version of the data. The output from R is shown in Figure 3.2.1 and Figure 3.2.2.

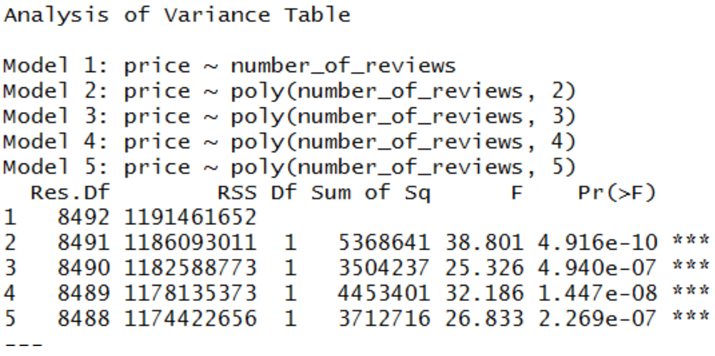


```
Analysis of Variance Table

Model 1: price ~ review_scores_rating
Model 2: price ~ poly(review_scores_rating, 2)
Model 3: price ~ poly(review_scores_rating, 3)
Model 4: price ~ poly(review_scores_rating, 4)
Model 5: price ~ poly(review_scores_rating, 5)
  Res.Df        RSS Df Sum of Sq       F    Pr(>F)    
1   8492 1152927960                                   
2   8491 1144338674  1   8589286 63.7295 1.614e-15 ***
3   8490 1144311385  1     27289  0.2025    0.6527    
4   8489 1143990647  1    320738  2.3798    0.1230    
5   8488 1143989305  1      1342  0.0100    0.9205    
```

Figure 3.2.1

The output showed that the second order is significantly better than the linear one while all the others are not significantly better than the second order. Thus, we should choose the second order to transform review score.



```
Analysis of Variance Table

Model 1: price ~ number_of_reviews
Model 2: price ~ poly(number_of_reviews, 2)
Model 3: price ~ poly(number_of_reviews, 3)
Model 4: price ~ poly(number_of_reviews, 4)
Model 5: price ~ poly(number_of_reviews, 5)
  Res.Df        RSS Df Sum of Sq      F    Pr(>F)    
1   8492 1191461652                                  
2   8491 1186093011  1   5368641 38.801 4.916e-10 ***
3   8490 1182588773  1   3504237 25.326 4.940e-07 ***
4   8489 1178135373  1   4453401 32.186 1.447e-08 ***
5   8488 1174422656  1   3712716 26.833 2.269e-07 ***
---
```

Figure 3.2.2

The output showed that the fifth order is significantly better than all the others. Thus, we should choose the fifth order to transform number of reviews.

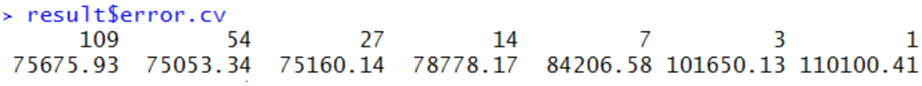
Note that we only consider orders from linear up to 5, because we want the model to be simple as well as to prevent overfitting.

For the purpose of incorporating what we found in polynomial regression, we fit a general additive model with number of reviews and review score transformed. General additive models(GAMs) provide a general framework for extending a standard linear model by allowing non-linear functions of each of the variables, while maintaining additivity. Again, general additive model is flexible so we cannot guarantee our model is not overfitting. Additionally, general additive model is not a feature selection technique, so this model is more useful if it is applied to predict the appropriate price for new Airbnb property owners. The mean square error is

129988, which is not an improvement of linear models. The significant features of general additive model are almost the same as Lasso.

3. Random Forest

We finally decided to implement random forest because it has only one tuning parameter and is a non-linear technique. Random forest is a decision tree based technique. Decision tree is a rule of stratifying and segmenting the predictor space into a number of simple regions. While single decision trees suffer from high variance, random forest averaged over several trees and thus provides a more accurate result. Random forest has many nice properties: 1. Increasing number of trees will not lead to overfitting 2. It prevents overfitting by randomly selecting a subset of features at each step 3. It gives importance graph that is easy for visualization. Given these merits, we fitted a model using random forest. The cross-validation result for determining the best number of features to select is given in Figure 3.3.1.



```
> result$error.cv
      109        54        27        14         7         3         1
 75675.93  75053.34  75160.14  78778.17  84206.58 101650.13 110100.41
```

Figure 3.3.1

Thus, we fitted a random forest model with 3000 trees and at each split, 54 features are randomly selected. Note that while the more trees the better, it will take much longer time for random forest model to give out a result. The mean square error is 111079, which is an improvement over previous techniques. The variable importance graph is shown in Figure 3.3.2.

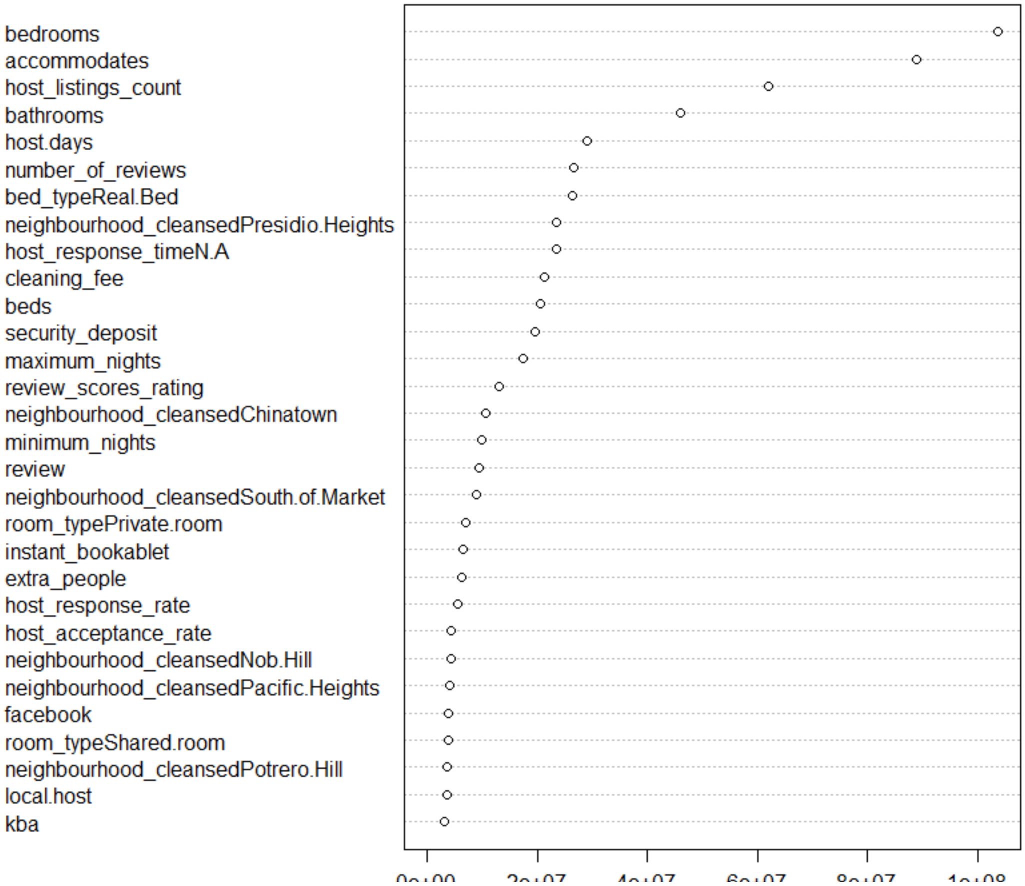


Figure 3.3.2

The method for determining the importance of variables is increase in node purity. As we can see from the graph, number of bedrooms, number of people can accommodate, host listing counts and number of bathrooms are the four most important features that can affect the price of an Airbnb property.

IV. Conclusion

We established 5 models and the mean square error of each model is reported in Table 4.1.1.

Model	MSE
Lasso	129420
Ridge Regression	129567
Huber loss with l2 regularizer	168860
General Additive Model	129988

Random Forest	111079

Table 4.1.1

Of all the 5 models, random forest outperforms other models regarding mean square error. Models other than Huber loss with l2 regularizer all have similar mean square errors. Since random forest takes substantially longer to run, if accuracy is not the main concern, these 3 models should be considered especially Lasso and ridge regression because we always want the simplest model to explain our data.

Based on the result of random forest and Lasso, we concluded that the following features are important to the price of an Airbnb property.

Owner info	**Property info**
host response time	Neighborhood
	property type (house, apartment, boat, etc.)
	room type (entire, private, shared room)
	number of people can accommodate
	bed type (real bed, futon, airbed, couch, pullout bed)
	number of bathrooms
	number of bedrooms
	number of beds
	cleaning fee
	minimum nights
	review scores
	instant bookable

Given which features may affect the prices of an Airbnb property, we hope provide new Airbnb owners a way of evaluating their properties and thus set up appropriate prices for their properties. However, due to the high mean square error, we don`t think our models are appropriate to predict an appropriate price for them. Since our models were based on the assumption that the prices in the data are appropriate, it is very unlikely to be true in reality since we lack the criteria to evaluate whether a particular price is reasonable or not. It is very likely that some Airbnb owners are charging arbitrarily. Nevertheless, our models will still provide new Airbnb owners some references to consider when determining prices for their properties. This will also help lessees to determine if the property owners are charging them at an appropriate price.

V. Bibliography

1. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. "An Introduction to Statistical Learning - with Applications in R | Gareth James | Springer." *With Applications in R | Gareth James | Springer*. Springer-Verlag New York, n.d. Web. 05 Dec. 2016.
2. http://insideairbnb.com/get-the-data.html