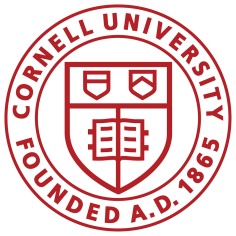

ORIE 4741

Fall 2017

Project Midterm Report
Airbnb Price and Occupancy Rate Prediction in Seattle

Kerou Gao (kg486)
Kartikay Gupta (kg477)

Cornell University, Ithaca, NY
October 27, 2017

Acknowledgement: We would like to thank **Mr. Jialin Liu**, a PhD student in the Electrical and Computer Engineering Department, for providing valuable inputs in this project.

Introduction

After considering our primary objective that using review score prediction to reveal important features of Airbnb's listing and improve customer satisfaction, we find that the best way for owners to improve performances is reading reviews, which makes a complicated machine learning model unnecessary. Thus, we change our objective to price and occupancy rate prediction, which can be used for:

- Guiding new owners to set prices for their properties.
- Optimizing revenues from both Owners' and Airbnb's perspective. Our models can be used to reveal relationship between price, occupancy rate and other features as well as the trade-off between price and occupancy rate. Since revenue = price * occupancy rate, it is possible to optimize revenue by adjusting price and some characteristics of houses.

Data Set Description and Feature Selection

The data set that we use for training our models has been obtained from "Inside Airbnb" and contains information about 3818 Airbnb listings in Seattle, Washington. There are in total 93 features that are provided in the data set. These features give us a plethora of information ranging from:

- Listing Location - Coordinates, Neighbourhood, Street
- Listing Characteristics - Room type, Amenities, Number of Bedrooms, Annual Availability, Number of People that can be accommodated, Transit proximity information etc.
- Listing Reviews - Statements by guests, number of reviews, review scores etc.
- In addition to all this, some columns contain details such as host id, scrape id etc. These are obviously useless and will not be used for the development of the predictive model.

Thus the data set that we use has a a mix of discrete, continuous and nominal values. There are some ordinal values as well such as the cancellation policy and response fequency of the Airbnb host.
For our preliminary analysis, we extracted all the continuous and discrete values from our data set and created a new data frame from these values. The new data set consisted of 30 features in total excluding the price. Out of these features, we used data visualization tools such as bubble plots (Figure 3) and strip plots (Figure 2) to select 7 features that displayed correlation with the listing price. In addition, we also added an offset to get a better fit to our data.
The feature space that we used consisted of:

- accommodates: number of persons that can be accommodated in a listing
- bathrooms: number of bathrooms
- bedrooms: number of bedrooms
- beds: number of beds
- guests included: number of guests included in a listing price
- latitude & longitude: coordinates of the listing

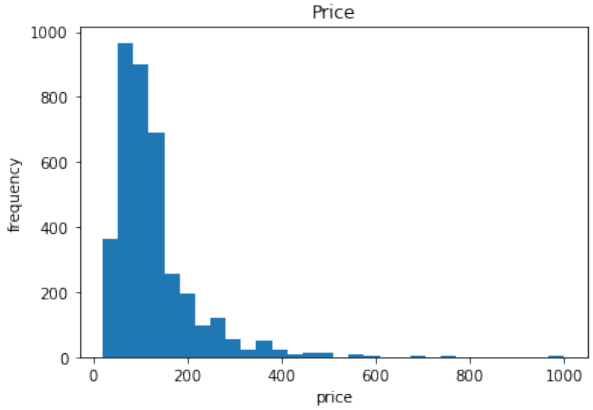

Figure 1: Price frequency distribution diagram

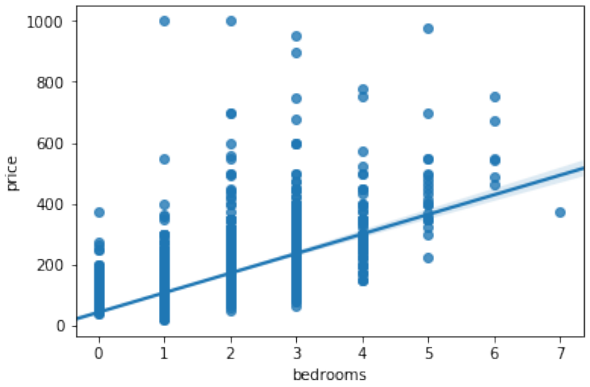

Figure 2:Correlogram Price-bedrooms

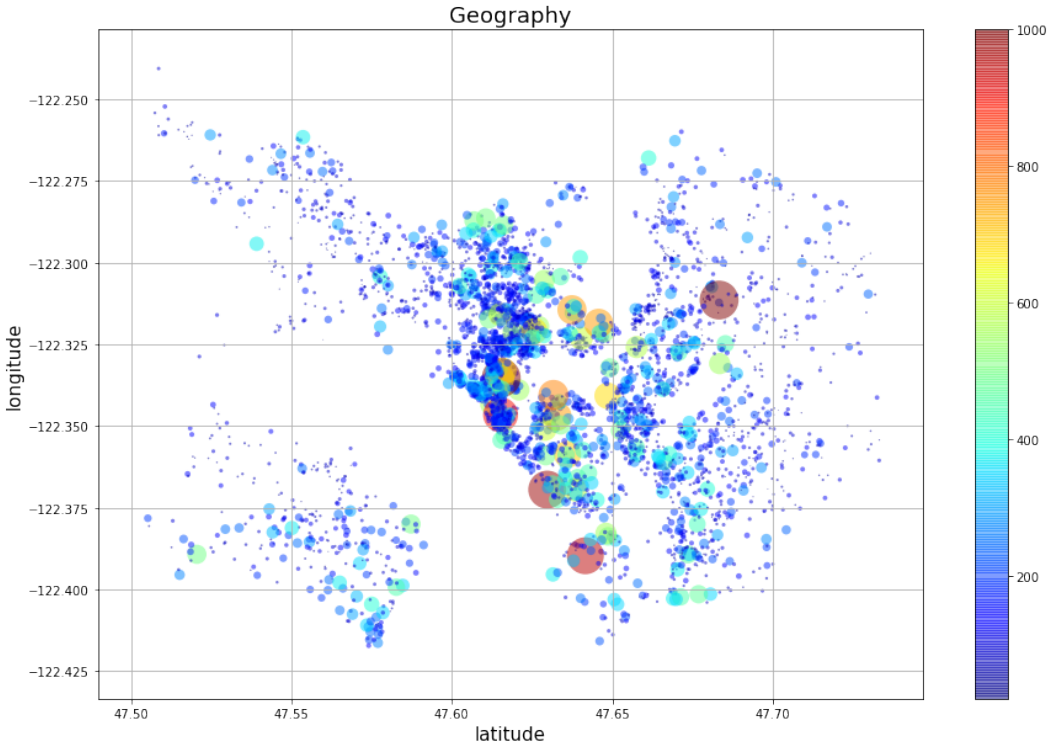

Figure 3: Bubble plot depicting listing price as a function of location

We find that most features such as number of bedrooms, accommodates etc have positive correlations with price. In terms of geographical location, prices are likely to be highers in certain areas.
We also observed a key anomaly in our data set namely that the area of the listing was missing. This certainly would have played a major role in predicting the price of the listing as one would expect a higher price for a larger property.

Preliminary Analysis and Results

After we dropped all the samples with missing values in one or more columns, there were 3796 samples left. This data was then split with 80% data in the training set and the remaining 20% in the test set.Our preliminary analysis consisted of fitting some regression models using Scikit Learn in Python to a reduced data set.
We started by fitting some models such as linear, lasso and ridge regression models and random forest ensemble learning method. We used the root mean squared error (RMSE) as a metric to judge the performance of these learning methods and the table below gives the results obtained for this metric:

Table 1: Performance of Different Supervised Learning Methods

Supervised Learning Method		
	Training RMSE ($)	**Test RMSE ($)**
Linear Regression	64.26	66.60
Ridge Regression	64.28	66.60
Lasso Regression	64.59	66.90
Random Forest	52.16	72.04

Further Work

After running preliminary analyses, we can see that both training set errors and test set errors are highwhich means that our model is under-fitting. Additionally, we also observe that the errors that we obtain from fitting different regression models are similar. This seems to indicate that there are heavy correlations among our features. Thus the solution to this is to increase the size and diversity of our feature space or use a more complex model. Our plan is listed below:
(1) Go though non-numerical features and select more into the data set in the sequence of binary value, ordinal values, nominal values and text(description).
(2) Try different feature engineering methods to increase complexity of our models and also use cross validation for hyper parameter tuning. Our ultimate goal is to develop a predictive model that generalizes well on unseen data.
(3) Once we have an accurate predictive model for predicting the listing price, we will develop a model for predicting the number of reviews per month for a listing which will act as a proxy for measuring the occupancy rate.