Berkeley Datathon

Factors Affecting AirBnB Booking Ratio

Xinhui Gu, Guyu Xue, Jiying Zhou, Shiyu Du

Not all AirBnB listings are created equal; some are booked more frequently than others. What drives this discrepancy in frequency? In this report, we define a metric “booking rate” as the proportion of days that a listing is booked; in other words, an estimated likelihood that the property is booked based on historic data.

$$Booking\ Ratio\ =\ \#\ Times\ Booked\ /\ \#\ Total\ Times\ Listing\ Available$$

Being able to explain a booking ratio trend plays an important role in business planning for both Airbnb and the hosts, since understanding this trend could both help Airbnb better negotiate contracts and provide hosts with better expectations about the popularity of their listings. Examining the booking ratio trend can also lead to a deeper understanding about listing prices, assisting Airbnb in optimizing profits and guiding hosts in pricing their listings.

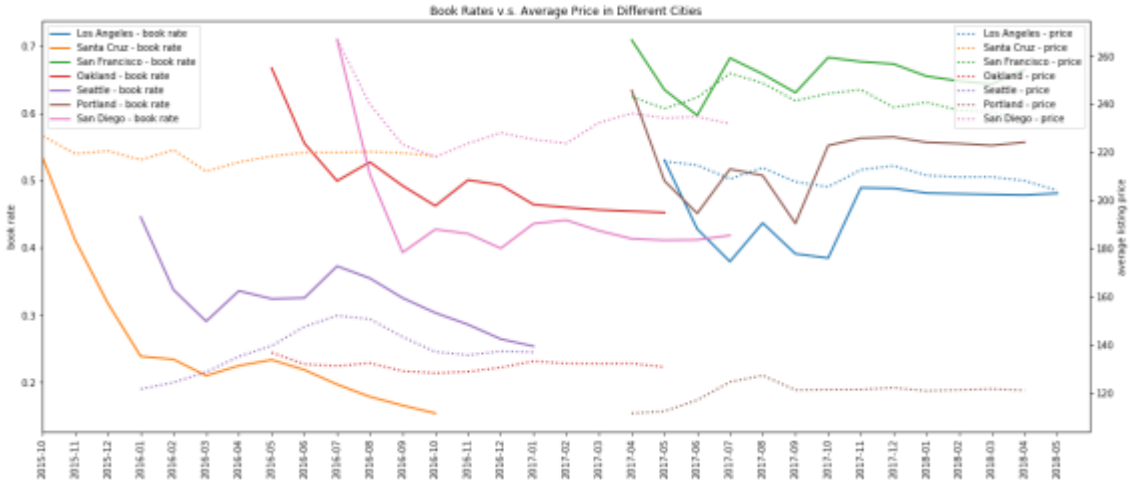
Our analysis focuses on following points:

- The time series relationship between booking ratio and local housing prices
- Explain booking rates with in-residence and external features provided in the data
    - In-house features are features that are specific to each listing (e.g. wifi, AC, kitchen, free parking, etc). Out-house features are the vicinity features (e.g. restaurant, food, gas station, ATM, etc.)

The data under analysis come from three datasets. One dataset provides information on details of individual Airbnb listings, another provides calendar information about the times listings were available, and a third details information about venues near listings.

Non-Technical Executive Summary:

- Observation: booking rates have seasonal trends, and tend to behave similarly in different cities:



For example, in 2017-2018 Los Angeles, Portland, and San Francisco shared very similar booking rate patterns, lowest in June/July and peaking in October/November. From 2016-2017, Oakland and San Diego had the opposite trend, possibly because San Diego is more of a tourist location, while Oakland leans more residential. Thus, the opposing trends are reasonable.

- Booking ratio trends are heavily correlated with listing prices:

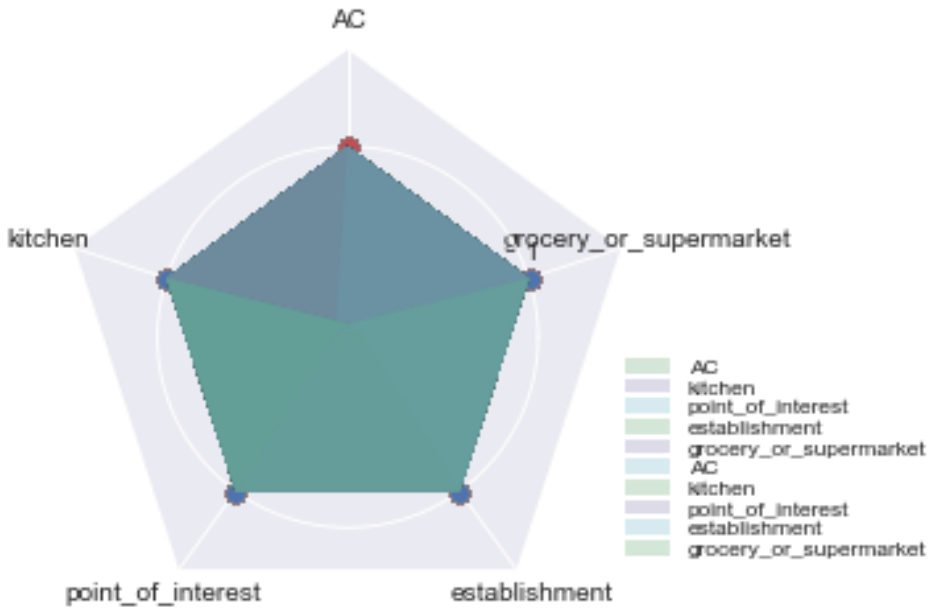
	Los Angeles	Santa Cruz	San Francisco	Oakland	Seattle	Portland	San Diego
r-squared	0.166086	0.353449	0.084545	0.374140	0.010693	0.032374	0.751304
pvalue	0.000000	0.000000	0.000110	0.000000	0.000008	0.000001	0.000000
const	196.066977	214.491201	206.414282	120.171052	143.903363	128.297118	175.787437
coef	31.568726	18.395249	53.889595	22.652756	-19.249578	-15.042029	125.979375

Thus, studying booking ratios have the double advantage of understanding consumer behavior and providing insight into listing prices.

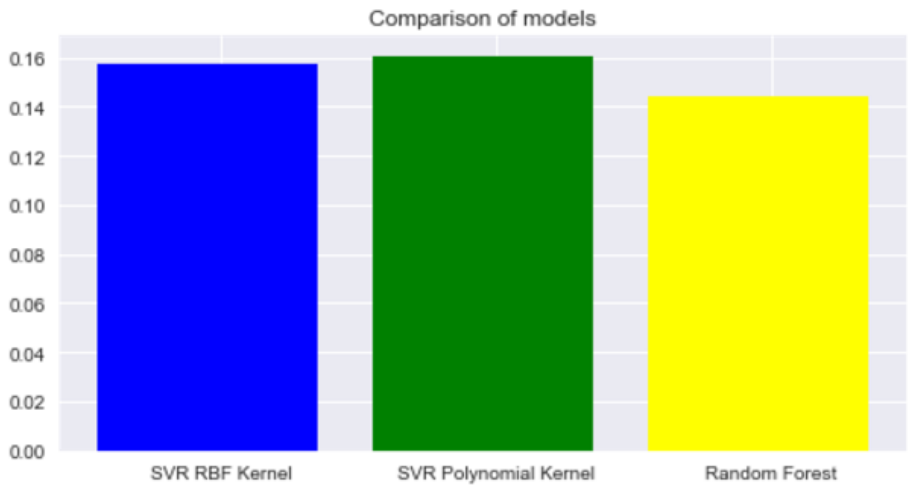
- Trend-driving factors analysis: Availability of certain in-house and neighborhood facilities have significant impact on booking ratio. Based on our analysis we found the main factors affecting booking rate are:
    - In-house facilities: availability of air conditioning, a kitchen, a workspace, breakfast, living essentials, and a family friendly residence
    - Neighborhood facilities: liquor store, point of interest, gas station, home goods store, museum, transit station, subway station, grocery store or supermarket

We analyze the most revalent features from the features pool and draw the star pie charts. The features selected are 'AC', 'kitchen', 'point_of_interest', 'establishment', 'grocery_or_supermarket'. Lightness of the color is the explanatory power, darker color

shows more explanatory power. From the polygon chart, the three top explanatory variables are AC, kitchen and grocery_or_supermarket.



We built 3 predictive models to predict booking ratio based on availability of facilities and yield relatively good output (low predictive error). This means, if we are given information like whether the listed house have a kitchen, or whether it is near a night bar, we can predict the its booking ratio. The comparison of three models are as follows:

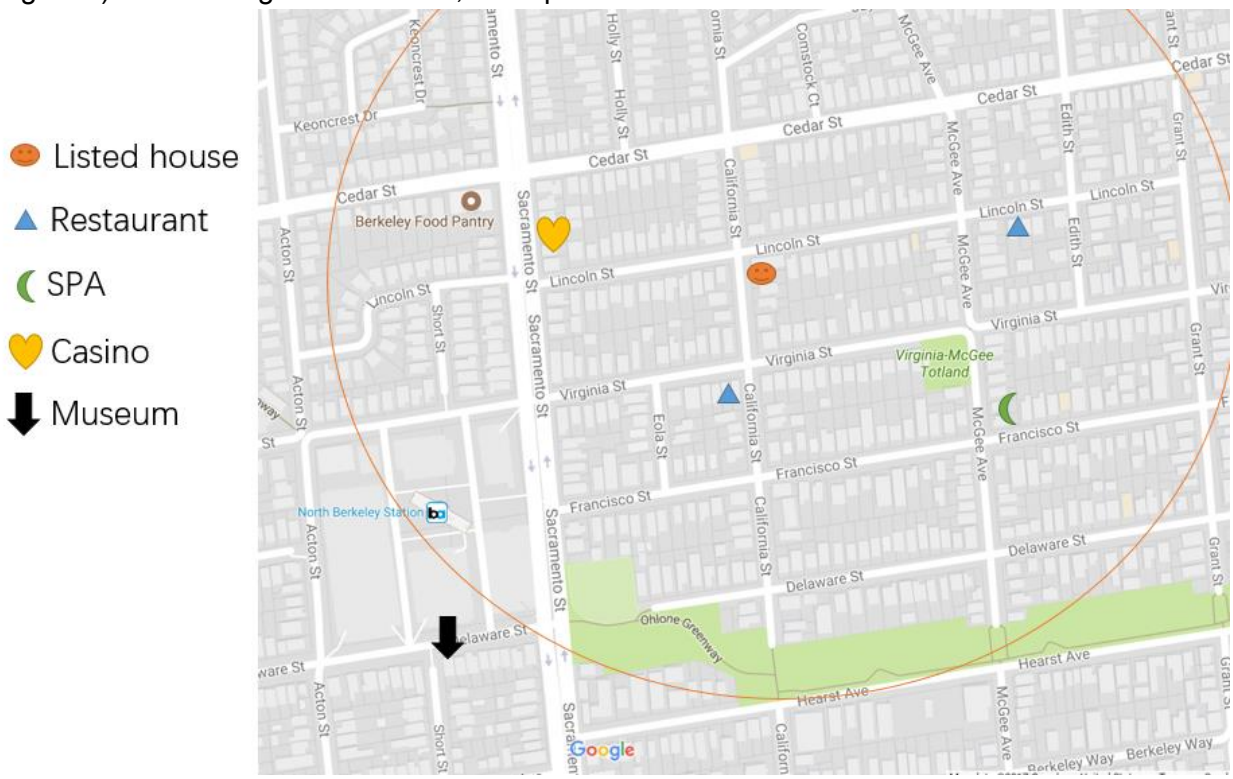


Technical Exec Summary

Data Cleaning & Variable Selection

The availability of features inside and outside the rental property is likely to affect a property's booking rate. For example, a property providing wifi or with restaurants nearby is more likely to be booked than a property without. To assess this hypothesis, we feature engineer dummy variables indicating whether a property has certain internal and external features.

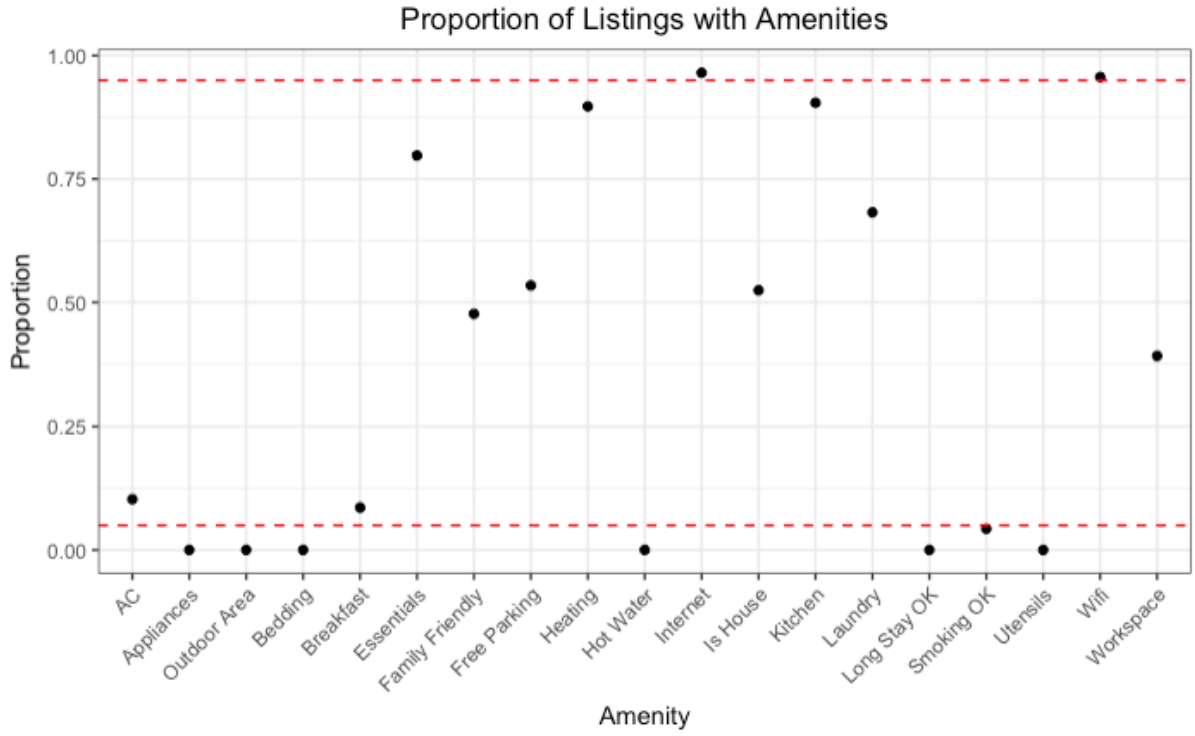
Indicators for internal features (e.g. internet, kitchen, backyard, family friendly) are extracted from 'listings.csv' using regular expressions on the 'amenities' column, and take the value 1 if the property provides it, and 0 otherwise. Indicators for external features (e.g. police station, restaurant, shopping mall) are extracted by determining which type of venues (from 'venues.csv) are located within the "neighborhood" of a listing location, with "neighborhood" defined as within a four-block radius (using Google Maps, roughly a radius of 1.35e-5 latitude or longitude) of the listing's coordinates, exemplified below:

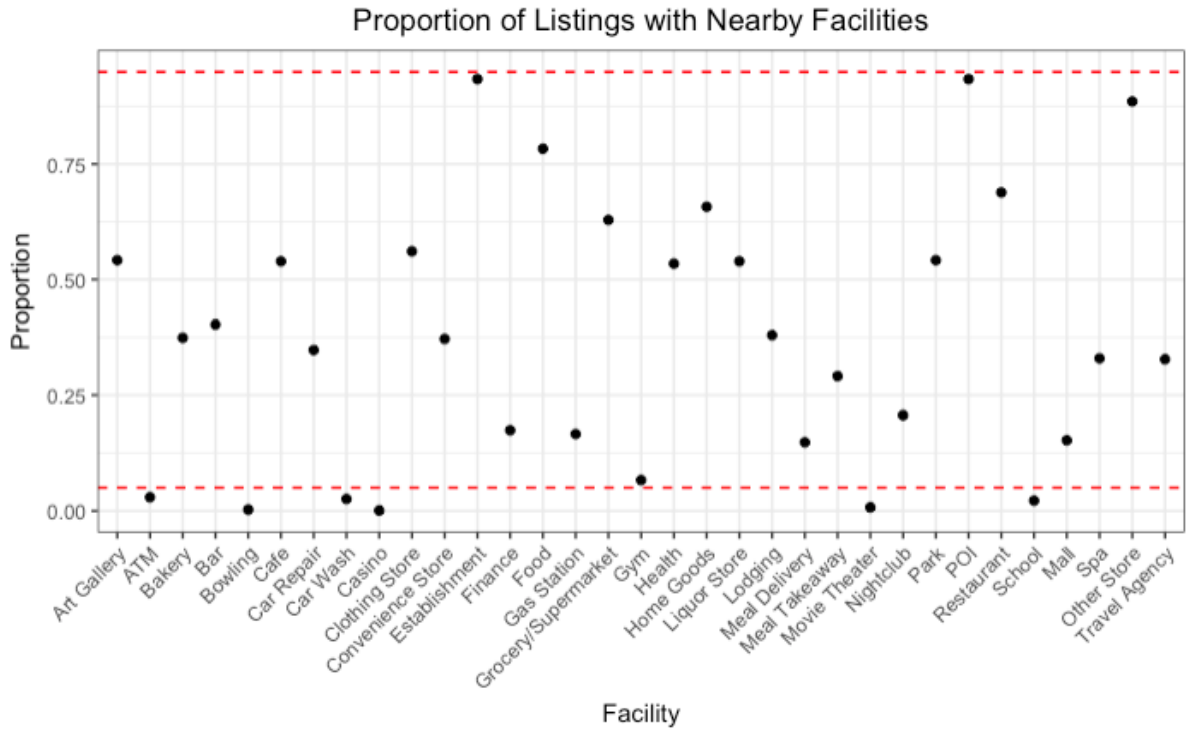


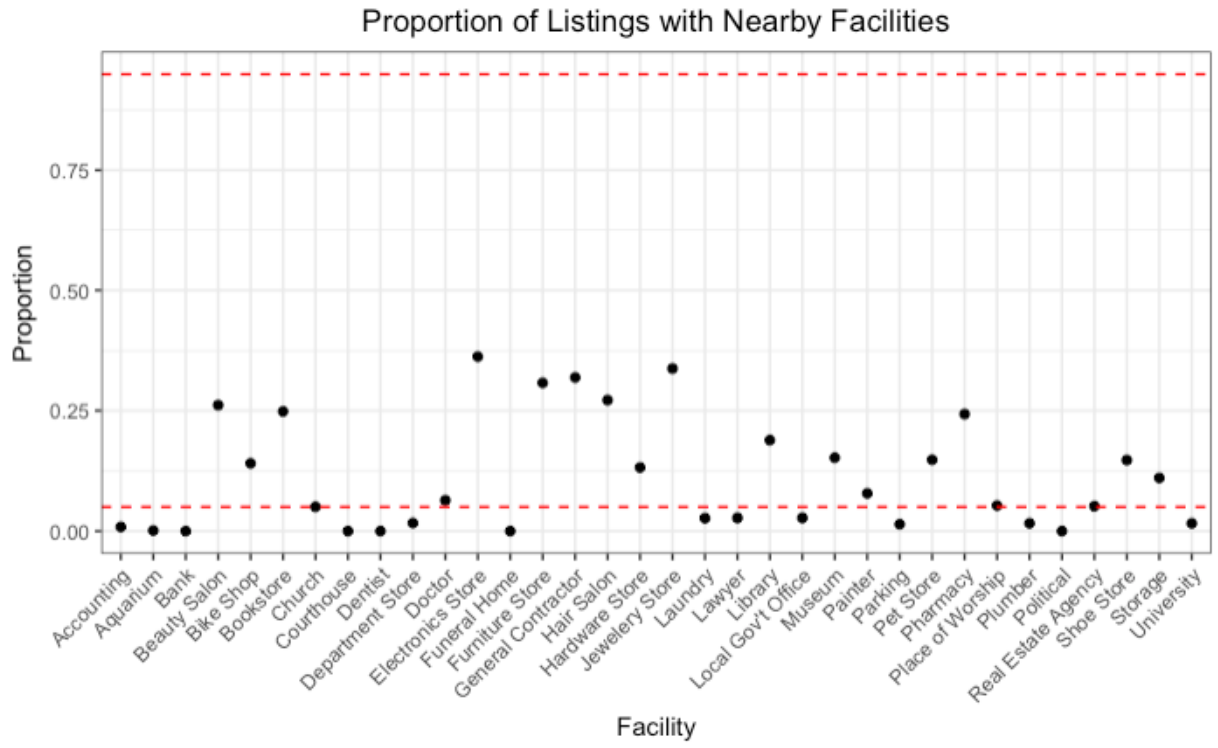
Distance between a listing and a venue is calculated as:

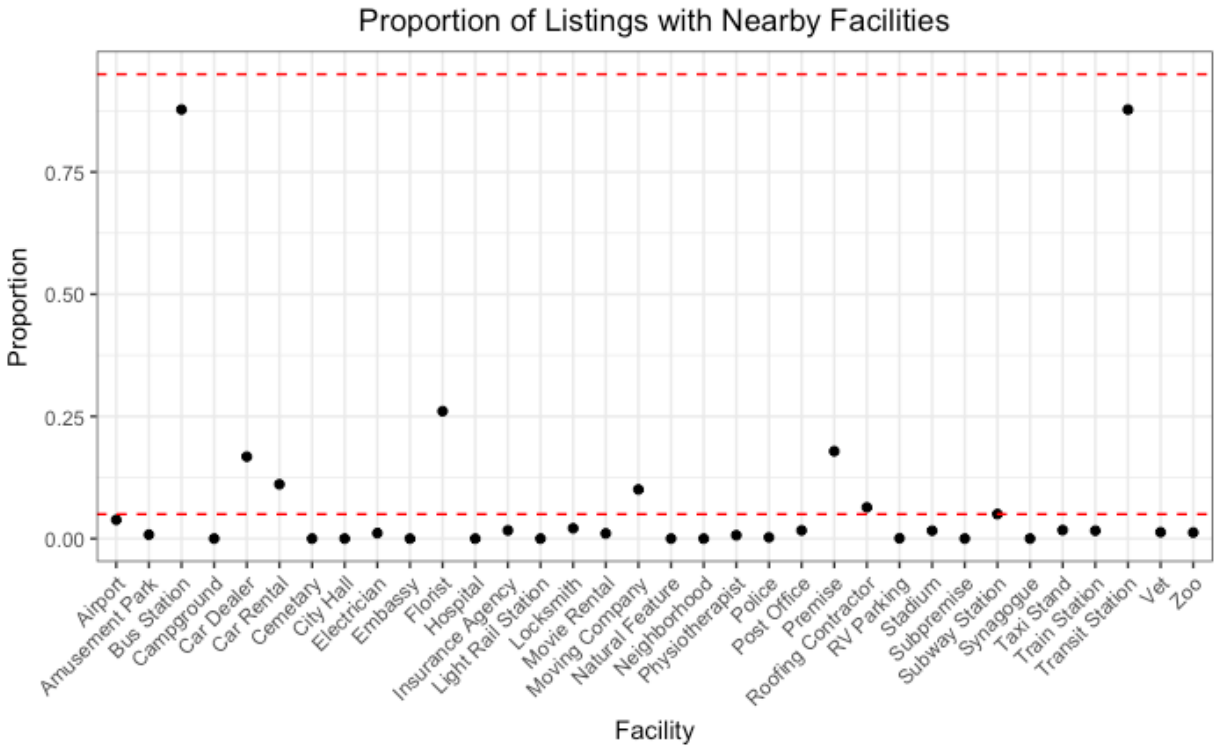
$$dist = \left(latitude_{house} - latitude_{facility}\right)^2 + \left(longitude_{house} - longitude_{facility}\right)^2$$

Notice in the visuals below that some internal/external features are present in almost all or no properties (proportion of listings with feature is < 5% or > 95%). Since these variables are highly monotone, they are unlikely to contribute greatly to our model, thus features with proportions above or below the red dashed lines are removed from consideration.


Proportion of Listings with Amenities
Proportion
1.00
0.75
0.50
0.25
0.00
AC
Appliances
Outdoor Area
Bedding
Breakfast
Essentials
Family Friendly
Free Parking
Heating
Hot Water
Internet
Is House
Kitchen
Laundry
Long Stay OK
Smoking OK
Utensils
Wifi
Workspace
Amenity


Proportion of Listings with Nearby Facilities
Proportion
0.75
0.50
0.25
0.00
Art Gallery
ATM
Bakery
Bar
Bowling
Cafe
Car Repair
Car Wash
Casino
Clothing Store
Convenience Store
Establishment
Finance
Food
Gas Station
Grocery/Supermarket
Gym
Health
Home Goods
Liquor Store
Lodging
Meal Delivery
Meal Takeaway
Movie Theater
Nightclub
Park
POI
Restaurant
School
Mall
Spa
Other Store
Travel Agency
Facility





The remaining variables are the variables of interest utilized in model building.

Linear Regression

Regression of Booking rates:

The booking ratio data is calculated from the dataset('calendar.csv'). We calculated the booking rates and volatilities of the prices for each of the listed id. We joined dataset('calendar.csv', 'listings.csv' and 'venues.csv') by matching the listing_id. We run three separate regressions on in-house variables, venue variables, and all variables. For each regression, we remove the insignificant variables after running on the full set of variables.

Regression the booking rates on the in-house variable (after removing insignificant variables):

OLS Regression Results

Dep. Variable:	hit_rate	R-squared:	0.033
Model:	OLS	Adj. R-squared:	0.033
Method:	Least Squares	F-statistic:	176.8
Date:	Sat, 09 Sep 2017	Prob (F-statistic):	0.00
Time:	11:33:09	Log-Likelihood:	-25223.
No. Observations:	56417	AIC:	5.047e+04
Df Residuals:	56405	BIC:	5.058e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5481	0.006	84.623	0.000	0.535	0.561
AC	0.0457	0.003	13.537	0.000	0.039	0.052
kitchen	-0.0705	0.005	-12.892	0.000	-0.081	-0.060
free_parking	0.0619	0.004	17.670	0.000	0.055	0.069
smoking_ok	0.0264	0.007	3.892	0.000	0.013	0.040
breakfast	0.0072	0.005	1.419	0.156	-0.003	0.017
heating	0.0370	0.005	7.203	0.000	0.027	0.047
fam_friendly	0.0251	0.003	7.542	0.000	0.019	0.032
laundry	-0.0122	0.004	-3.116	0.002	-0.020	-0.005
essentials	-0.0802	0.005	-16.767	0.000	-0.090	-0.071
workspace	-0.0391	0.003	-11.513	0.000	-0.046	-0.032
is_house	0.0604	0.003	18.182	0.000	0.054	0.067

Omnibus:	52.358	Durbin-Watson:	1.890
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5509.145
Skew:	-0.075	Prob(JB):	0.00
Kurtosis:	1.476	Cond. No.	11.2

Regression the booking rates on the venue variable (after removing insignificant variables):

OLS Regression Results

Dep. Variable:	volatility	R-squared:	0.020
Model:	OLS	Adj. R-squared:	0.019
Method:	Least Squares	F-statistic:	28.77
Date:	Sat, 09 Sep 2017	Prob (F-statistic):	3.98e-71
Time:	14:03:43	Log-Likelihood:	-97864.
No. Observations:	18439	AIC:	1.958e+05
Df Residuals:	18425	BIC:	1.959e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
food	4.1939	1.348	3.112	0.002	1.553	6.835
point_of_interest	6.6136	0.725	9.124	0.000	5.193	8.034
establishment	6.6136	0.725	9.124	0.000	5.193	8.034
gas_station	-3.4294	1.203	-2.850	0.004	-5.788	-1.070
grocery_or_supermarket	-3.2839	0.944	-3.479	0.001	-5.134	-1.434
lodging	4.5944	0.885	5.194	0.000	2.861	6.328
spa	1.9505	0.894	2.181	0.029	0.197	3.704
home_goods_store	-3.4311	0.991	-3.461	0.001	-5.374	-1.488
real_estate_agency	9.2754	0.903	10.271	0.000	7.505	11.045
museum	1.9307	0.944	2.046	0.041	0.081	3.781
moving_company	-4.6844	1.512	-3.098	0.002	-7.648	-1.721
transit_station	-2.9014	1.414	-2.051	0.040	-5.674	-0.129
subway_station	-5.1570	2.412	-2.138	0.033	-9.885	-0.429
car_dealer	2.2663	0.897	2.528	0.011	0.509	4.024
roofing_contractor	7.8134	2.076	3.763	0.000	3.744	11.883

Omnibus:	52436.932	Durbin-Watson:	1.929
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5669818334.133
Skew:	37.550	Prob(JB):	0.00
Kurtosis:	2718.536	Cond. No.	1.09e+17

Regression the booking rates on the all variables: (after removing the insignificant variables)

OLS Regression Results

Dep. Variable:	hit_rate	R-squared:	0.043
Model:	OLS	Adj. R-squared:	0.042
Method:	Least Squares	F-statistic:	55.28
Date:	Sat, 09 Sep 2017	Prob (F-statistic):	3.17e-163
Time:	14:06:07	Log-Likelihood:	-6536.9
No. Observations:	18439	AIC:	1.311e+04
Df Residuals:	18423	BIC:	1.323e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
AC	0.0260	0.006	4.447	0.000	0.015	0.037
kitchen	-0.0295	0.008	-3.680	0.000	-0.045	-0.014
breakfast	-0.0248	0.008	-2.982	0.003	-0.041	-0.008
fam_friendly	-0.0070	0.005	-1.346	0.178	-0.017	0.003
essentials	-0.0800	0.008	-10.374	0.000	-0.095	-0.065
workspace	-0.0818	0.005	-15.319	0.000	-0.092	-0.071
point_of_interest	0.3938	0.007	59.274	0.000	0.381	0.407
establishment	0.3938	0.007	59.274	0.000	0.381	0.407
gas_station	0.0149	0.008	1.760	0.078	-0.002	0.031
grocery_or_supermarket	-0.0407	0.006	-6.603	0.000	-0.053	-0.029
home_goods_store	-0.0319	0.007	-4.702	0.000	-0.045	-0.019
real_estate_agency	0.0148	0.006	2.495	0.013	0.003	0.026
museum	-0.0192	0.006	-3.048	0.002	-0.031	-0.007
moving_company	0.0275	0.011	2.603	0.009	0.007	0.048
transit_station	-0.0287	0.010	-2.944	0.003	-0.048	-0.010
subway_station	-0.1055	0.017	-6.196	0.000	-0.139	-0.072
roofing_contractor	-0.0456	0.015	-3.114	0.002	-0.074	-0.017

Non-linear Predictive Models

We aim to build predictive model that can predict booking rate of a listed house, given whether the house has certain in-house facilities (e.g. free parking, breakfast) and neighborhood facilities (e.g. restaurant, night club, SPA). Due to the nature of data, we explored some non-linear models

Data format

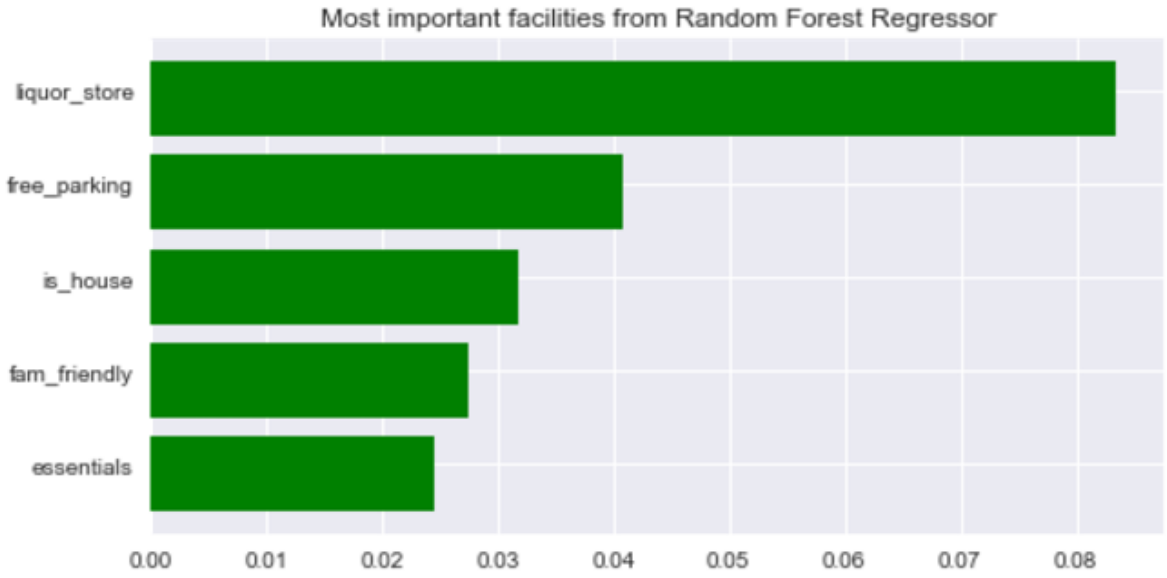
Each listed house is a record of data. It has hit rate (y), and a list of boolean values that indicate whether it has certain facilities. The data record looks like this

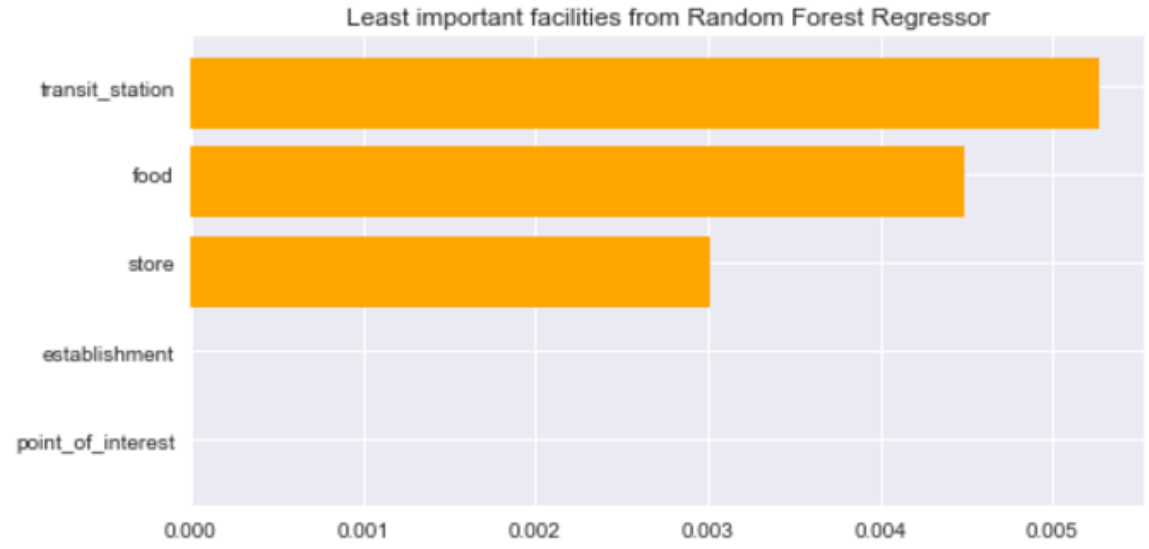
hit_rate	kitchen	free_parking	breakfast	...	restaurant	night_club	spa
0.5924	0	1	1	...	1	0	1
0.3251	1	0	0	...	0	1	0

0 indicates that the house does not have the facility, while 1 indicates that it has.
The features in blue are in-house facilities, while features in orange are neighborhood facilities.

Random Forest Regressor

We run the regressor with grid search, and find the best model with **mean squared error = 0.1446**. That means, the true hit rate is within +- 0.1446 of predicted hit rate. We also analyzed the most and least important facilities based on regression result.





Some interesting conclusion:
Liquor store and free parking are most important features from Random Forest regression; transit_station and food are least most import features. Hence those features are less considered when a customer are about to book his airbnb.

SVM Regressor - RBF model

We run the SVM regressor and get **mean squared error = 0.1574**. That means, the true hit rate is within +- 0.1574 of predicted hit rate.

SVM Regressor - Polynomial model

We run the SVM regressor and get **mean squared error = 0.1607**. That means, the true hit rate is within +- 0.1607 of predicted hit rate.

The comparison of models are as follows:

Appendix

Data Limitations:

The listings analyzed are from 2007-2017 and are located in or near six major West Coast cities: San Diego, Santa Cruz, San Francisco, Oakland, Portland, and Seattle. Our results are thus generalizable only to this general West Coast area.

Future Research Areas:

The future research is needed to better explain the relationship of the likelihood of booking and the relevant features. Our methods are linear regressions and random forest. We are incentivized to use more non-linear model to dig into the relationships. An important question remained unsolved is the analysis of causes and the results.