Data Mining (15.062), Final Project

"Explaining and predicting Airbnb prices across the US"

Michele Marinucci, MIT Master of Finance 2021

Table of Contents

Introduction

Airbnb is a platform that allows homeowners to seamlessly rent apartments and rooms for both short- and long-term periods. Ever since its inception in 2008, the housing market underwent a dramatic revolution. Both new entrants and incumbents on the platform certainly need a systematic way to optimally price their apartments; simultaneously, estimates of Airbnb rental rates are an ever more relevant datapoint to consider when purchasing real estate. Given this vivid interest, the purpose of this project is to explain and predict the prices of Airbnb rentals in the US as accurately as possible.

The first part of this paper goes over data pre-processing, including feature engineering and selection, handling of missing data and outliers, handling of categorical and numerical variables, data exploration, variable correlation, and principal component analysis. The second part iterates through several data-mining models, and for each it attempts to both predict and explain Airbnb prices across the US. In particular, every model is fine-tuned to maximize out-of-sample (or validation set) prediction and, whenever possible, the model's output is analyzed to explain which determinants are the strongest predictors of US Airbnb prices. I ultimately show that the best model to predict Airbnb prices across the US is a simple decision tree, and that the most important variables to explain such prices are the host's number of listings, latitude, availability during the year and number of the reviews. Hence, Airbnb owners cold leverage these characteristics to boost prices of their listings, whereas Airbnb user should beware and lookout for them to save money.

Data pre-processing

Raw data description

This project analyzes a data set provided by Kaggle and whose URL can be found in the references section. The data contains nearly 226029 Airbnb listings in US, and it lets us consider a wide range of potentially significant predictors, including latitude, longitude, city, neighborhood, neighborhood group, room type, number of reviews, reviews per month, date of last review, availability throughout the year, minimum number of nights required to book, host's total number of listings, host's name, the host's id, the listing's id, and the listing's name. Please refer to Exhibit 1 and Exhibit 2 in the appendix for a snapshot of the data and for further details about the individual variables. Notice that the data is tidy, meaning that every raw is a different Airbnb listing and every column is a different variable.

Feature engineering and selection

The first step is to properly engineer the features at hand. While the other independent variables seemed ready-to-use in their raw versions, the *last_review* variable, i.e. the date on which last review was received, may be more useful if transformed into the number of days since last reviews. Hence, I took December 9th, 2020 as the reference day and calculated the difference in days from this date for all the datapoints. This simple transformation implies that now the *last_review* indicates the days passed since the last review.

The next logical step is feature selection. Variables such as *name, host name, host id, neighborhood group,* and *neighborhood* were either redundant or particularly hard to extract meaningful information from. One of the reasons was that such variables were not numerical and at the same time they could not be meaningfully converted to categorical because it would have yielded dozens or hundreds of categories for each variable, as most of them represented unique values. Hence, I ultimately decided to drop them to focus on the remaining variables. However, future analysis could potentially use Natural Language Processing techniques to extract explanatory and predicting power from this information.

Missing data and outliers

Fortunately, the data did not present many missing values. Of the remaining variables, only *last_review, reviews_per_month* had NaNs. After considering several approaches, I decided to fill them in with the variable's median for simplicity. Clearly, I used median rather than mean to avoid outliers and to deal with the high skewness of the data at hand.

Taking note of the data's skewness, which is mainly attributable to the inherent nature of the variables under observation, I still checked that the data did not present extreme observations that could rather be attributed to flawed data entries et similia. To this aim, I sorted all observations for each variable and looked at the top and bottom results. It quickly became noticeable that the largest value in *minimum_nights* was an erroneous data entry – it is value was 100 million nights, which is clearly non-sensical. No other significant outlier emerged as clearly.

Exploration and handling of categorical data

There are two categorical variables, these being *city* and *room_type*. As far as handling of these categorical variables is concerned, I first explicitly transformed both in categorical variables and then got dummies out of them, dropping the first dummy for each.

Moving onto exploration of these categorical variables, a quick look at the graphs in Exhibit 3 shows that these variables may be important, as the variability in price for each of their different categories is quite visible. Moreover, to have a more comprehensive view of the data at hand, I used the latitude and longitude data in the dataset to plot the several AirBnbs on the US map, as

shown at the bottom of Exhibit 3. A quick glance of this map reveals that the dataset mainly contains observations for densely populated areas spread across all the US.

Exploration and handling of numerical data

There are 8 numerical variables, these being *latitude, longitude, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count,* and *availability_365.* As these numerical variables has markedly different scales, I standardized all of them by subtracting means and dividing by variance.

Moving onto exploration of these numerical variables, Exhibit 4.1 shows boxplots for all variables excluding latitude and longitude, whereas Exhibit 4.2 shows a matrix of histogram and scatterplots combinations across all variables excluding latitude and longitude. As clearly shown by the boxplots, all variables under analysis show right-skewness, though each to a different degree, with *minimum_nights* showing the highest skewness and *availability_365* showing the lowest one. Furthermore, the scatterplot combinations for *review_per_month* and *number_of_reviews* seem to have some degree of correlation. This claim will be further investigated in the next section with the correlation matrix.

Correlation matrix and principal component analysis

As a final step in data pre-processing, I computed the correlation matrix across all variables first and then across all numerical variables. Exhibit 5 shows heatmaps of these two correlation matrices, with blue representing more positive correlations and red representing more negative correlations. As it emerges clearly, high correlations (i.e., with absolute value greater than 0.5) are present between New York and longitude, between Hawaii and longitude/latitude, and between number of reviews and reviews per month. Hence, I decided to drop reviews per month and longitude to balance between loss of information and reduction of correlation.

For the sake of completeness, I also ran a principal component analysis to understand whether the high-dimensionality of the data could be reduced. However, as shown by the first six principal components (out of 6 numerical variables) presented in Exhibit 6, while the first two components account for roughly 45% of variability, the remaining components are almost evenly distributed. Furthermore, utilizing these principal components would have hindered the explanatory objectives of this paper. Because of these two reasons, I decided to move forward without utilizing these principal components.

Model tuning and selection

After all the data pre-processing detailed above, the objective now is to develop an efficient model to predict and possibly explain Airbnb rental prices in the US. In the following sections, I will use several supervised learning algorithms, including simple linear regression, regression trees, random forests, K-NN for regression and an ensemble of all these models. A summary table for all results is available in Exhibit 11, which compares RMSE across all the models under analysis.

Multiple linear regression

First and foremost, I ran a linear regression to treat it as a baseline model. Additionally, as shown in Exhibit 7, the linear regression output along with the corresponding p-values provide useful information regarding the explanatory power of the predictors under consideration. Overall, these p-values show that the coefficients for several of these predictors is statistically significant, suggesting some degree of linear relationship. Nevertheless, the following sections will show that other models may better capture the relationships between our independent and dependent variables.

Decision Tree

The next model I used is decision trees. Of course, regression trees have a wide array of hyperparameters as well as strong potential to overfit. Hence, I initially ran a random grid search across a large range of values for maximum depth, minimum impurity decrease, and minimum samples split to obtain an initial ballpark for the hyperparameters that would maximize out of sample predictability. After that, I focused a more specific grid search around such values. The final tree has a maximum depth of 6, a minimum impurity decrease of 0.001, and a minimum sample split of 22; this tree is depicted in Exhibit 8.

Random Forest

A similar combination of random search and grid search to the one used in the decision tree was also utilized for the random forest. Nevertheless, the search focused on more parameters this time, including number of estimators, maximum number of features, maximum depth, minimum samples split, minimum samples leaf and presence of bootstrap. Besides the predictive power that will be discussed below, this random forest also reports variable importance, as shown in Exhibit 9. As it emerges at a quick glance, the host's number of listings, latitude, availability during the year and number of the reviews are the most important variables in explaining AirBnb rental prices in the US.

K-Nearest Neighbors

The next model I used is K-NN. Following a similar procedure to the one outlined in the previous methods, I initially trained and evaluated the model across a wide range of k-values; subsequently, I reduced the range around the point that seemed to be closer to the minimum RMSE. As show by the graphs and table in Exhibit 10, the best value for parameter k seems to be 7, as the out of sample RMSE is minimized for that value.

Ensemble

Finally, I attempted to create an ensemble that used all the models previously discussed. The intuition is that combining all the models would yield better results if the individual models are not entirely stable.

Results and conclusion

In this paper, I used several data mining models to attempt to predict and explain AirBnb rental prices in the US.

As far as the prediction aspect is concerned, Exhibit 11 shows a table that summarizes all these models along with their corresponding RMSE sorted from best (i.e., lowest RMSE) to worst (i.e., highest RMSE). As it clearly emerges, the best model is the decision tree with a RMSE of roughly 534. This result was unexpected, as random forest or other ensembles usually perform better than individual models. There might be two potential explanation for this. First, I could have tuned certain parameters in different ways across decision tree and random forest; however, this was not the case. A second and more likely explanation is that the decision tree was already a stable model with the best possible parameters, and the random forest only added noise. Hence, this explanation is a quick reminder of the fact that, while likely, random forests do not universally and unequivocally always outperform simple tree. Overall, the predictive power of this model is quite satisfactory.

As far as the explanation aspect is instead concerned, the linear regression, the decision tree, and the random forests all provide potential answers. Of these, the random forests probably provide the most reliable results. As mentioned above and as it emerges at a quick glance from Exhibit 9, the host's number of listings, latitude, availability during the year and number of the reviews are the most important variables in explaining AirBnb rental prices in the US. Hence, AirBnb owners could potentially focus on these parameters to increase the price of their listings, whereas Airbnb user should beware and lookout for them to save money.

References

Data has been taken from the Kaggle website and it is available for free to the public:
https://www.kaggle.com/kritikseth/us-airbnb-open-data/tasks?taskId=2542

The Python code utilized can be found in the following GitHub repository:
https://github.com/michele-marinucci/Data_Mining_15.062

Appendix

Exhibit 1: Data Snapshot

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type
0	38585	Charming Victorian home - twin beds + breakfast	165529	Evelyne	NaN	28804	35.65146	-82.62792	Private room
1	80905	French Chic Loft	427027	Celeste	NaN	28801	35.59779	-82.55540	Entire home/apt
2	108061	Walk to stores/parks/downtown. Fenced yard/Pet...	320564	Lisa	NaN	28801	35.60670	-82.55563	Entire home/apt
3	155305	Cottage! BonPaul + Sharky's Hostel	746673	BonPaul	NaN	28806	35.57864	-82.59578	Entire home/apt

	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365	city
0	60	1	138	16/02/20	1.14	1	0	Asheville
1	470	1	114	07/09/20	1.03	11	288	Asheville
2	75	30	89	30/11/19	0.81	2	298	Asheville
3	90	1	267	22/09/20	2.39	5	0	Asheville

Exhibit 2: Variable Description

#	Variable Name	Description
1	id	unique listing id
2	name	name of listing
3	host_id	unique host Id
4	host_name	name of host
5	neighbourhood_group	group in which the neighbourhood lies
6	neighbourhood	name of the neighbourhood
7	latitude	latitude of listing
8	longitude	longitude of listing
9	room_type	room type (categorical)
10	price	price of listing per night
11	minimum_nights	minimum number of nights required to book
12	number_of_reviews	total number of reviews on listing
13	last_review	date on which listing received its last review
14	reviews_per_month	average reviews per month on listing
15	calculated_host_listings_count	total number of listings by host
16	availability_365	number of days in year the listing is available for rent
17	city	region of the listing (categorical)

Exhibit 3: Exploration of categorical variables (i.e., City and Room type)

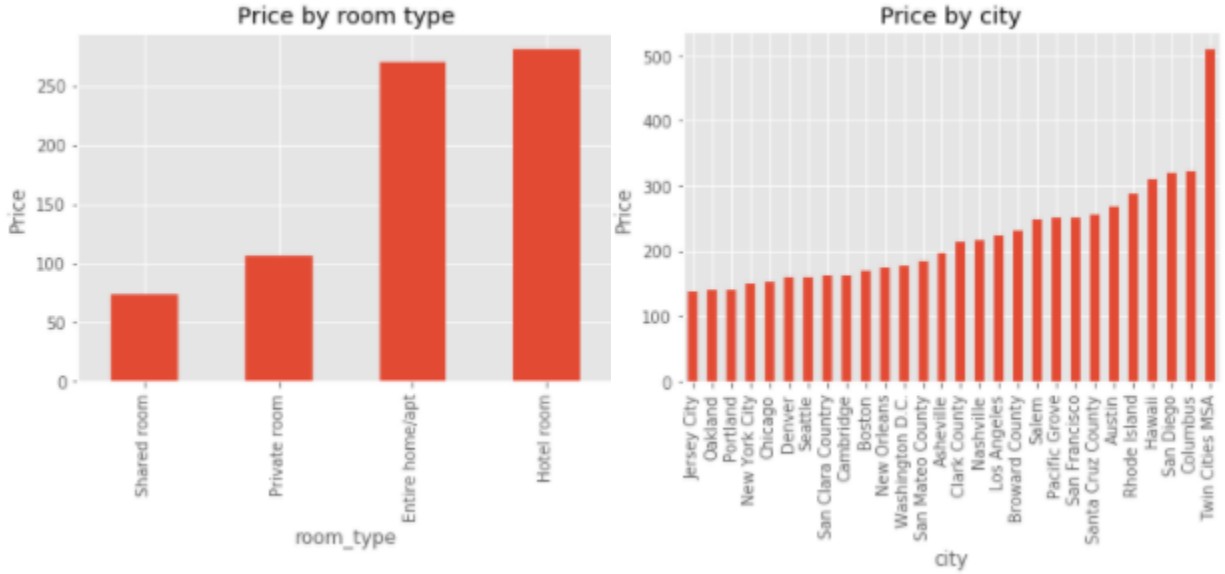
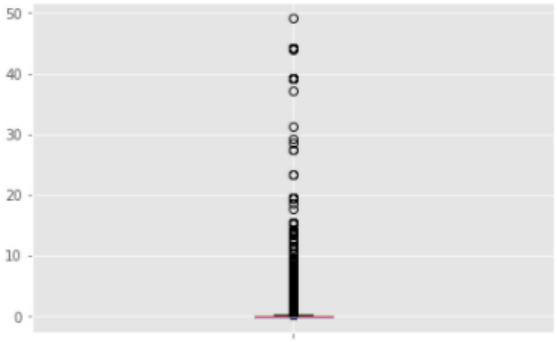
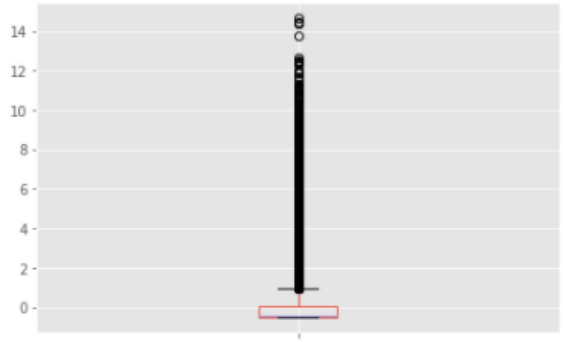
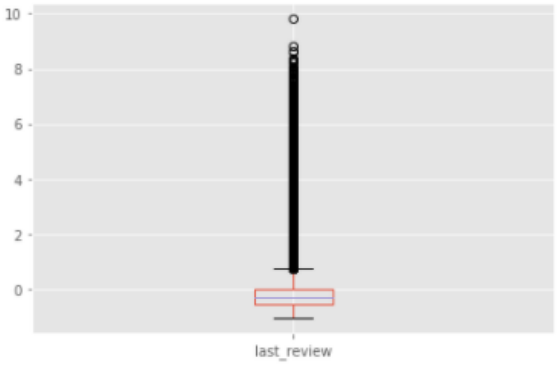
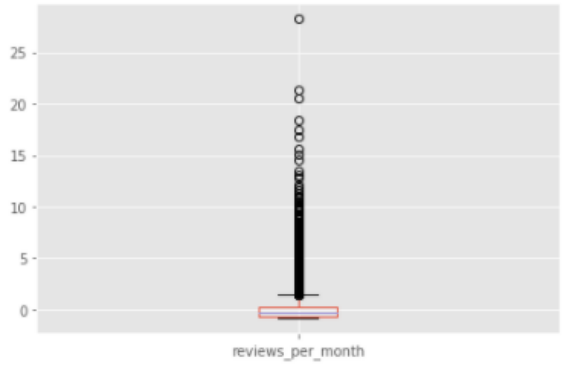


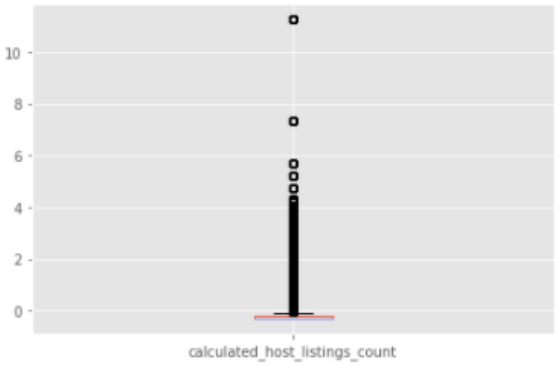
Exhibit 4.1: Exploration of numerical variables











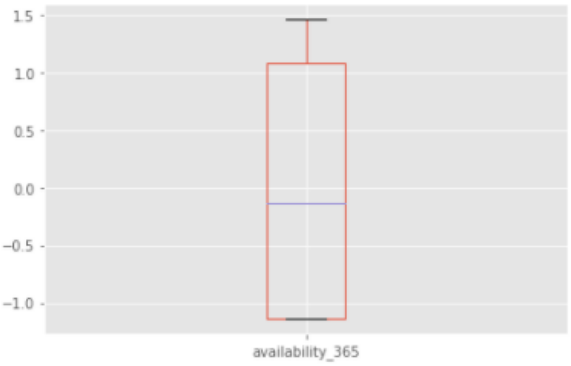


Exhibit 4.2: Exploration of numerical variables (continued)

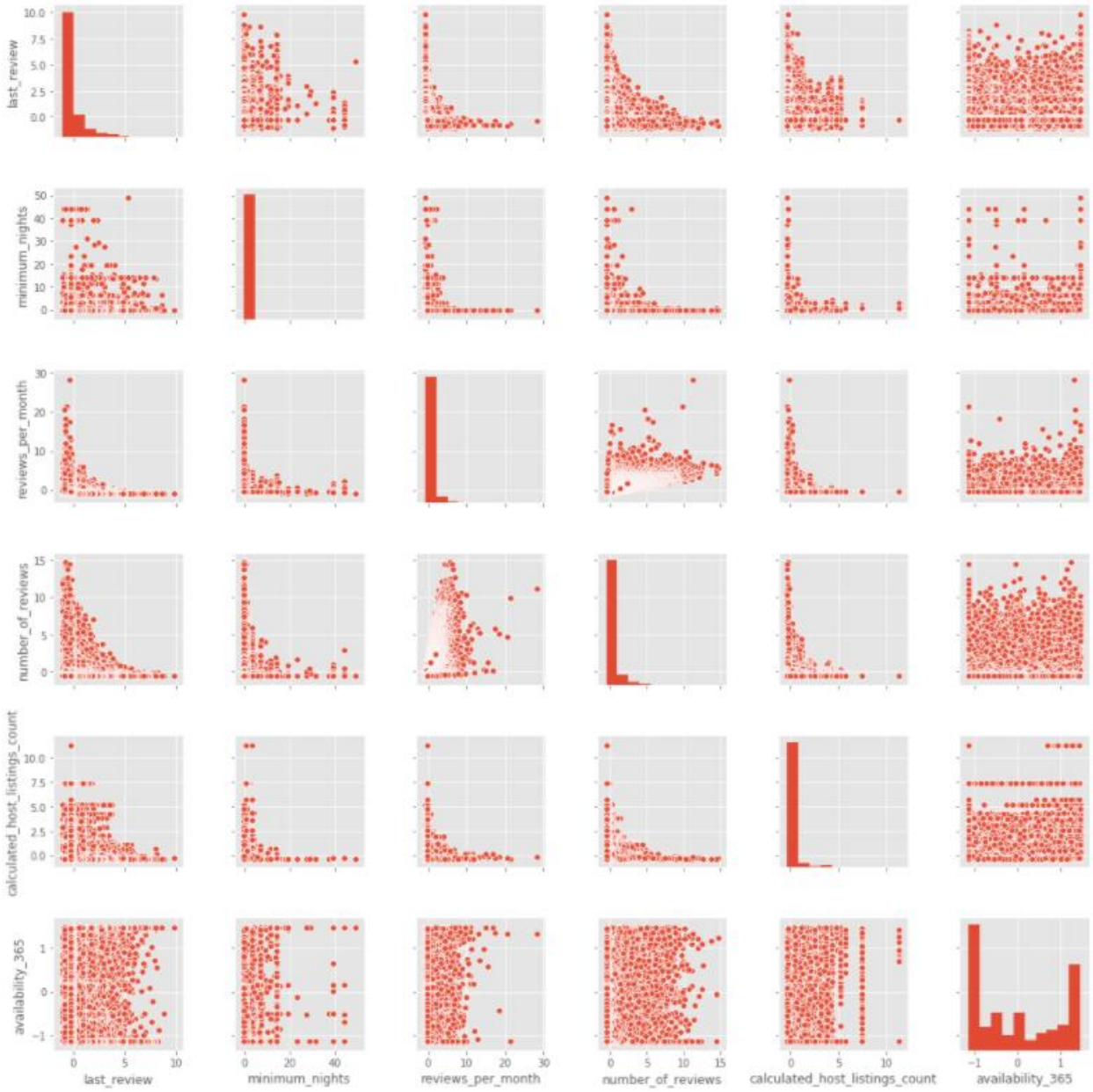


Exhibit 5: Heatmap of correlation matrix

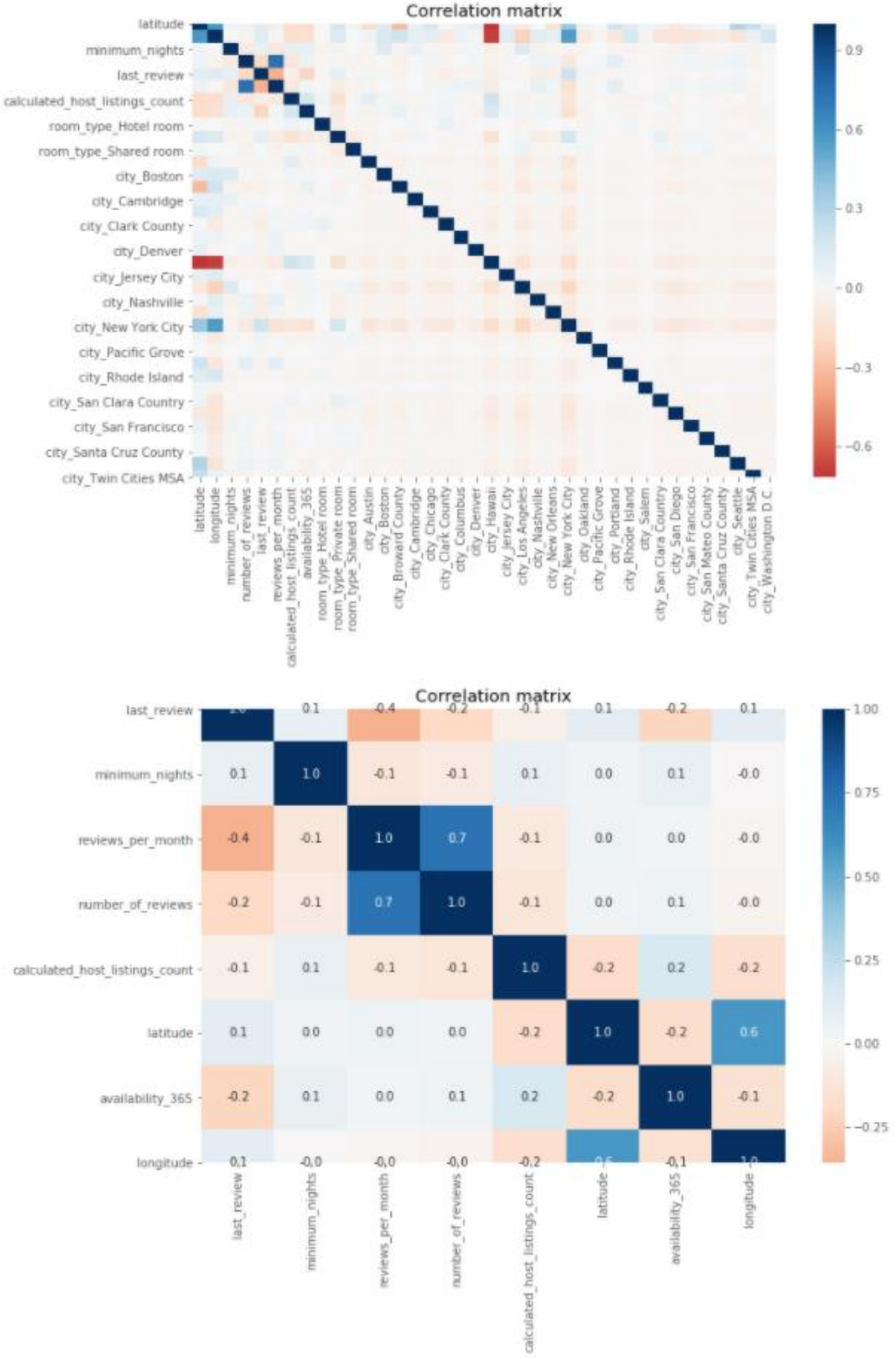


Exhibit 6: Principal component analysis

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.208459	1.126625	1.005937	0.888867	0.867984	0.845548
Proportion of variance	0.243396	0.211547	0.168652	0.131681	0.125566	0.119158
Cumulative proportion	0.243396	0.454943	0.623595	0.755275	0.880842	1.000000

Exhibit 7: Multiple linear regression output

	coef	std err	t	P>|t|	[0.025	0.975]
const	207.1626	31.221	6.635	0.000	145.968	268.357
latitude	230.1360	77.974	2.951	0.003	77.304	382.968
minimum_nights	-7.7366	2.949	-2.623	0.009	-13.518	-1.955
number_of_reviews	-42.8911	3.121	-13.742	0.000	-49.009	-36.774
last_review	-5.6782	3.138	-1.809	0.070	-11.829	0.473
calculated_host_listings_count	8.3750	3.201	2.616	0.009	2.100	14.650
availability_365	11.4348	3.130	3.654	0.000	5.300	17.569
room_type_Hotel room	-19.7854	33.695	-0.587	0.557	-85.828	46.258
room_type_Private room	-153.4850	6.842	-22.433	0.000	-166.896	-140.074
room_type_Shared room	-210.5658	21.953	-9.592	0.000	-253.594	-167.537
city_Austin	268.3017	69.316	3.871	0.000	132.440	404.163
city_Boston	-210.6965	86.722	-2.430	0.015	-380.675	-40.718
city_Broward County	369.9849	113.131	3.270	0.001	148.244	591.726
city_Cambridge	-190.8942	94.839	-2.013	0.044	-376.783	-5.006
city_Chicago	-216.2420	80.318	-2.692	0.007	-373.668	-58.816
city_Clark County	23.5681	35.355	0.667	0.505	-45.728	92.864
city_Columbus	-158.1478	69.233	-2.284	0.022	-293.846	-22.449
city_Denver	-142.7514	60.721	-2.351	0.019	-261.767	-23.736
city_Hawaii	621.8791	170.608	3.645	0.000	287.481	956.277
city_Jersey City	-192.9756	72.569	-2.659	0.008	-335.213	-50.738
city_Los Angeles	133.9378	36.650	3.654	0.000	62.102	205.773
city_Nashville	14.7933	36.358	0.407	0.684	-56.469	86.056
city_New Orleans	182.6375	73.238	2.494	0.013	39.087	326.188
city_New York City	-153.8329	66.796	-2.303	0.021	-284.756	-22.910
city_Oakland	-64.2531	47.280	-1.359	0.174	-156.925	28.418
city_Pacific Grove	38.3904	119.766	0.321	0.749	-196.355	273.136
city_Portland	-342.8192	119.388	-2.871	0.004	-576.824	-108.815
city_Rhode Island	-79.9264	78.317	-1.021	0.307	-233.430	73.577
city_Salem	-122.9782	143.883	-0.855	0.393	-404.995	159.038
city_San Clara Country	-18.3025	40.748	-0.449	0.653	-98.170	61.565
city_San Diego	217.8312	46.458	4.689	0.000	126.772	308.890
city_San Francisco	18.4662	43.577	0.424	0.672	-66.947	103.879
city_San Mateo County	-17.8624	46.603	-0.383	0.702	-109.206	73.481
city_Santa Cruz County	55.8784	51.557	1.084	0.278	-45.176	156.933
city_Seattle	-393.4943	141.682	-2.777	0.005	-671.197	-115.791
city_Twin Cities MSA	38.7905	112.843	0.344	0.731	-182.385	259.966
city_Washington D.C.	-100.8102	51.719	-1.949	0.051	-202.181	0.560

Exhibit 8: Regression Tree

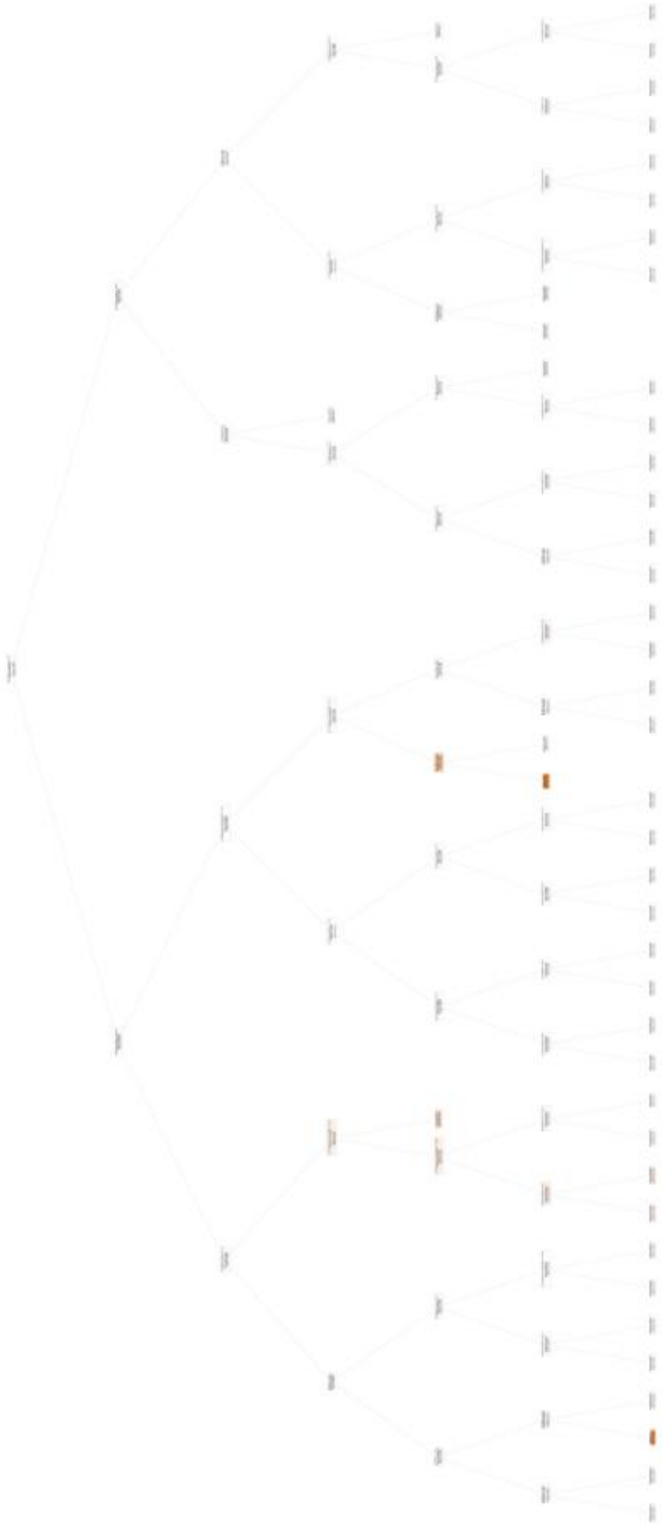


Exhibit 9: Random Forest, variable importance

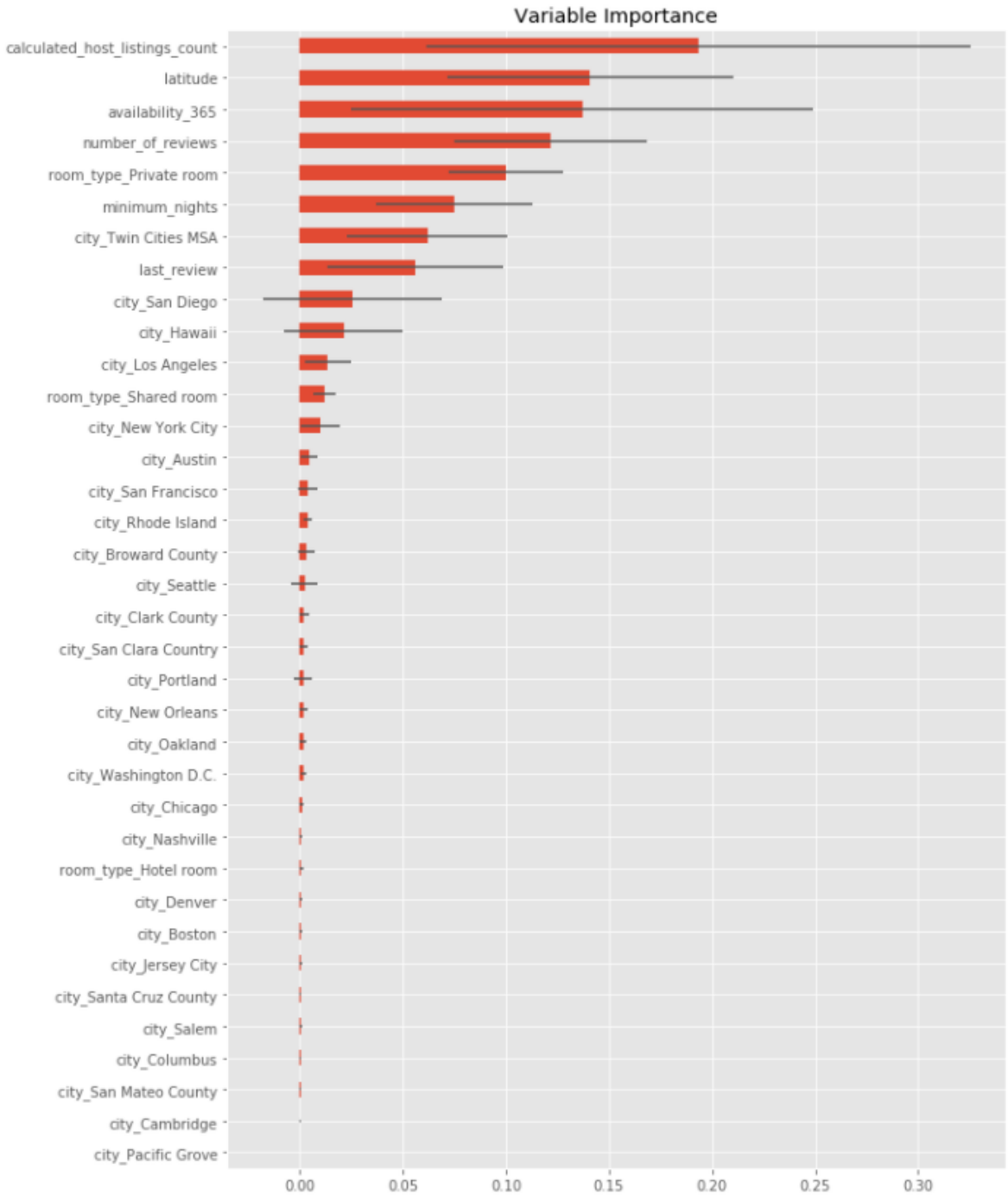
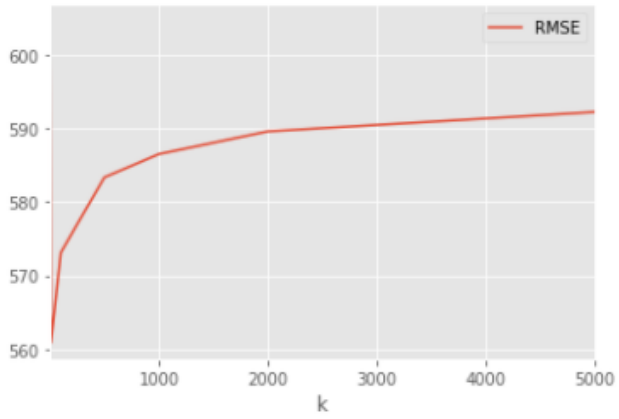
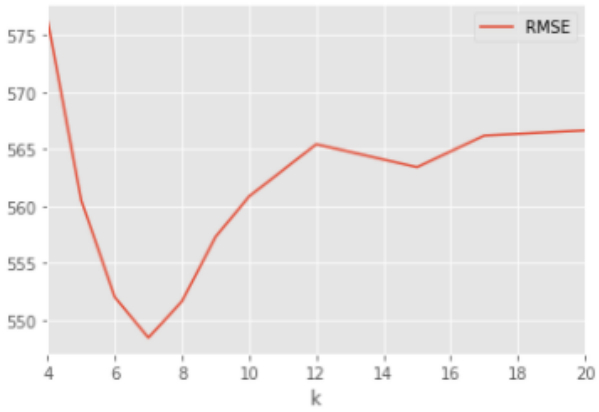


Exhibit 10: K-Nearest-Neighbors hyperparameter tuning

K-tuning (range 3-5000)



K-tuning (range 4-20)



k	RMSE
4	576.263504
5	560.493532
6	551.992074
7	548.431394
8	551.635106
9	557.287595
10	560.839229
12	565.418531
15	563.398688
17	566.178252
20	566.628117

Exhibit 11: summary of results

Model	RMSE
Decision Tree	534.3
K-Nearest Neighbors	548.4
Random Forest	548.8
Ensemble	555.0
Linear Regression	586.2