The Predictive Capabilities of Airbnb Data: New York City as a Case Study

Christina cleveland
University of Colorado, Boulder
CSCI 5622: Machine Learning
Christina.Cleveland@Colorado.edu

Margarite Jacoby
University of Colorado, Boulder
CSCI 5622: Machine Learning
Margarite.Jacoby@Colorado.edu

1 Introduction

Long-term rental markets are complex systems with many unseen forces shaping them. Information asymmetry between owners and renters and the temporal structure of long-term leases means that property owners can be slow to respond to changes in the market. This makes rental trends difficult to predict or to witness in real-time.

The peer-to-peer, or sharing economy, has enabled the growth of a new, short-term accommodation market, through platforms such as Airbnb and VRBO. While many people have researched the effects that these platforms have had on the hotel industry [9], we are interested in the connection between these peer-to-peer rentals and the long-term rental market.

We hypothesize that the decisions informing long-term rental markets have sufficient overlap with those informing short-term rental markets to be able to predict one from the other. The ease of becoming a "host" on these short term rental platforms, coupled with the rapid rate of occupant turnover means the short-term market reacts more quickly than the traditional rental market to relevant contextual changes. Therefore, we will test whether changes to short-term rental markets can predict changes that will occur in long-term rental markets. We will also test whether trends in and information about the short-term market have predictive power within the short-term market.

Due to the complicated nature of the set of decisions influencing both the short and long-term rental markets, changes to these markets are difficult to predict using heuristic models. However, this complexity makes the problem a good target for machine learning. Additionally, any interpretable model we could develop would be of great interest to real-estate investors, city-planners, and consumers alike.

In the following analysis we are focusing just on the market in New York City. For simplicity we wanted to focus our work on one city, to eliminate complicating location-based variables (*e.g.*, whether a city is a vacation or business destination, total size of city, etc.). Of all the cities available, NYC has the largest amount of data: approximately 40,000 listings with 96 features each per time point, compiled approximately monthly over a time period of the past three years. NYC's geographic structure lends itself to this sort of analysis - the five separate boroughs, each with smaller, distinct neighborhoods, makes it easy to define our points. Finally, New York's real estate and rental markets have long been the subject of debate due to the high costs of living in the area. Recently, the debate over the Airbnb market in NYC has intensified, with the city comptroller blaming Airbnb for exacerbating the rental affordability crisis [8]. That claim was later refuted by Airbnb and Airdna, a for-profit company that provides Airbnb analytics to hosts [6].

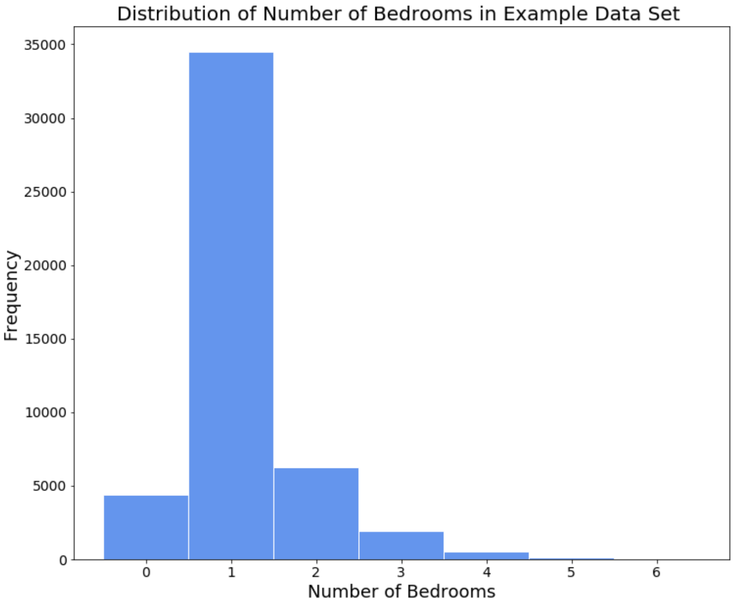


Figure 1: Distribution of number of bedrooms in July, 2018 data set

2 Data, Pre-processing, and Methods

2.1 Data Sources

Airbnb Data. Airbnb does not release its own data, however other groups have collected Airbnb data over time for major cities. One site, Inside Airbnb [1], has made their scraped data publicly available. This website scrapes and cleans Airbnb data from the organizations website, and compiles the data by time and city.

Inside Airbnb releases cleaned versions of data scraped from the Airbnb website in roughly monthly increments (beginning in 2015). Looking at a representative month of the data, July 2018, there are 47,854 individual listings. Of these, 50.3% are listed as 'Entire home/apt' (the remaining are private rooms in shared apartments or shared rooms). The distribution of number of bedrooms in entire homes for this representative month are shown in Figure 2.1. The fraction of entire home listins, as well as the distribution of bedroom types, is representative of the entirety of the data used.

There are 221 neighborhoods listed in this set, spread out over the five boroughs in NYC - Manhattan, Brooklyn, Queens, Bronx, and Staten Island. The Zillow data that we used (see below) contained the same five counties, minus Staten Island, and so we excluded Staten Island listings from the Airbnb data that we looked at.

Zillow Data. To get information about the rental market, we downloaded data from zillow [2]. This data set gave the median rental price for New York City, broken down by neighborhood and number of bedrooms, reported monthly since 2010.

2.2 Pre-processing.

Even though Inside Airbnb cleans the scraped data, the data sets still required much pre-processing and analysis before they could be used for any prediction tasks since the cleaned data is still information input by humans and humans are largely imperfect. Feature columns listed below were hand selected from the set of all columns since many features were unusable/could not be converted to a float or int to be used in prediction. Remaining columns were converted to floats, ints or bools as needed.

Although the neighborhoods in the Airbnb data and Zillow data generally matched up, there were quite a few listed neighborhoods in the Airbnb data that were not represented in the Zillow data. This is likely due to the informality of neighborhoods names and boundaries, and the fact that neighborhoods are reported by the hosts, not assigned by the Airbnb organization. In order to not throw away these data points, we looked at the largest "orphan neighborhoods" (those not present in Zillow), and either matched them to their larger umbrella neighborhood, or decided they were outside our boundaries and remove them from the data set.

Once we had a mapping of listed neighborhood to umbrella neighborhood, we replaced each neighborhood with it's umbrella neighborhood. We also modified the neighborhood group to represent the county the neighborhood is in: New York, Queens, Kings, and Bronx.

To ensure the Airbnb listings we are looking at are consistent and generally full-time listings, *i.e.*, listings that are the most similar to rental listings, we only kept listings where the room type was the Entire home/apt.

We also dropped listings that had a nan value in the following columns: 'id','zipcode','neighbourhood_cleansed','bedrooms','accommodates','price','guests_included', 'extra_people', 'minimum_nights'. Some listings did actually have prices of $0.

Methods. Data pre-processing and cleaning was performed using Pandas [5].

Data analysis was performed using Numpy[4].

Data prediction and clusterting were performed using Scikit-Learn[7] (clustering, regression, and random forests) and XGBoost[3] (gradient tree boosting).

2.3 Rent Change Predictor by Neighborhood and Number of Bedrooms

The goal of this section was to analyze trends in the Airbnb listings over the past N months for a given neighborhood and number of bedrooms and predict the relative change in the median rent price for that neighborhood and number of bedrooms over the next M months.

Since the Zillow median rents were listed for a given number of bedrooms and neighborhood, we first found the sets of Airbnb listings that matched Zillow neighborhoods and number of bedrooms (neigh/bd). For each time point the number of listings per neigh/bd set was calculated and the neigh/bd set was only considered if there were sufficient listings in the set (n > 30).

The features were calculated generally by looking at the relative change for that feature in the Airbnb data over the past N months. The set of features we calculated are shown in Table 1

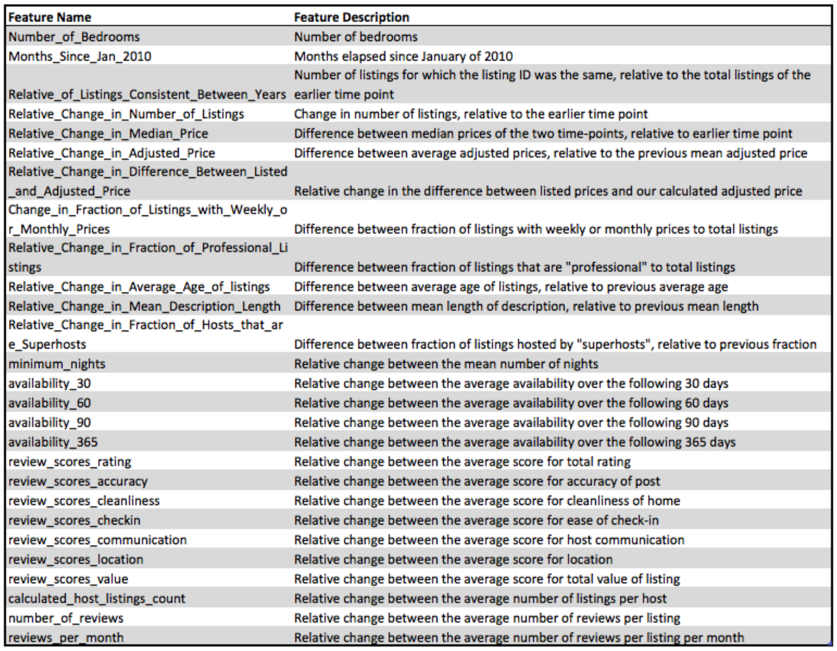
The features were calculated as described. We suspect that hosts sometimes increase their revenue and decrease the appearance of the cost of the listing by adding disproportional cleaning fees, or charging high prices for additional people, even if the space is designed for many. For instance, and unit may say $85 per night, but have a 2 night minimum and charge a $50 cleaning fee. Similarly, a 4 bedroom unit may advertise for $150 per night, but charge $30 per person after 2 people, and allow up to 8 people to stay there. To accurately reflect these differences in advertised versus actual cost, we calculated what we are referring to as the "adjusted price", the calculation of which is shown below.

$$\text{additional} = min[2 + \text{number of listed beds} - \text{included occupants}, 0] \tag{1}$$

$$\text{adjusted} = \text{listed price} + \text{additional} * \text{price per additional person} + \text{cleaning fee} \tag{2}$$

We created 36 labeled sets with different times used for each of the following: number of months in past to look at Airbnb data, number of months in past to look at Zillow data, and number of months in the future to try to predict the Zillow relative rent change (airbnbLag, zillowLag, monsPred). All combinations of the three values were tried for the following sets of values, respectively: [1,6,12], [1,6,12], and [1,3,6,12]. So, for example, if the three dates were 12 months, 12 months, and 3 months, then looking at the time point of 01/2017, the Airbnb relative change is between 01/2016 and 01/2017, and the Zillow relative change is between 04/2016 and 04/2017. We created so many different time

Table 1: Features used in prediction of median rent change



Feature Name	Feature Description
Number_of_Bedrooms	Number of bedrooms
Months_Since_Jan_2010	Months elapsed since January of 2010
Relative_of_Listings_Consistent_Between_Years	Number of listings for which the listing ID was the same, relative to the total listings of the earlier time point
Relative_Change_in_Number_of_Listings	Change in number of listings, relative to the earlier time point
Relative_Change_in_Median_Price	Difference between median prices of the two time-points, relative to earlier time point
Relative_Change_in_Adjusted_Price	Difference between average adjusted prices, relative to the previous mean adjusted price
Relative_Change_in_Difference_Between_Listed_and_Adjusted_Price	Relative change in the difference between listed prices and our calculated adjusted price
Change_in_Fraction_of_Listings_with_Weekly_or_Monthly_Prices	Difference between fraction of listings with weekly or monthly prices to total listings
Relative_Change_in_Fraction_of_Professional_Listings	Difference between fraction of listings that are "professional" to total listings
Relative_Change_in_Average_Age_of_listings	Difference between average age of listings, relative to previous average age
Relative_Change_in_Mean_Description_Length	Difference between mean length of description, relative to previous mean length
Relative_Change_in_Fraction_of_Hosts_that_are_Superhosts	Difference between fraction of listings hosted by "superhosts", relative to previous fraction
minimum_nights	Relative change between the mean number of nights
availability_30	Relative change between the average availability over the following 30 days
availability_60	Relative change between the average availability over the following 60 days
availability_90	Relative change between the average availability over the following 90 days
availability_365	Relative change between the average availability over the following 365 days
review_scores_rating	Relative change between the average score for total rating
review_scores_accuracy	Relative change between the average score for accuracy of post
review_scores_cleanliness	Relative change between the average score for cleanliness of home
review_scores_checkin	Relative change between the average score for ease of check-in
review_scores_communication	Relative change between the average score for host communication
review_scores_location	Relative change between the average score for location
review_scores_value	Relative change between the average score for total value of listing
calculated_host_listings_count	Relative change between the average number of listings per host
number_of_reviews	Relative change between the average number of reviews per listing
reviews_per_month	Relative change between the average number of reviews per listing per month

sets because we were not certain what the temporal relationship between the Airbnb changes and Rent changes might be.

Our initial, small batch analysis of two time points implied that there was some correlation between the features and the rent change, however, once we had collected the full data sets for each set of time lags, the correlation coefficients between our features and our labels had dropped to effectively zero 2.3.

We attempted to simplify the problem by turning it into a binary classification task. The y labels were converted to 0 and 1. The cutoff was chosen based on the fraction of a year over which the Zillow rent was observed and an approximate inflation rate of 0.03 (3%). So, for a zillowLag of 6 months, the cutoff$= 0.03 * 6/12 = 0.015$. We tested the binary classification on all of the data sets as well, for example the original and binary labels are shown for the (airbnbLag, zillowLag, monsPred) = (12,6,6,) data set 2.3.

When we took the binary labeled data and put it through random forest classifiers (n_estimators=1000, max_depth=12), all of the training accuracies were 95-98%, however the validation set accuracies were 60-70%. Additionally, when we calculated the fraction the the validation sets that were 0 or 1, the accuracies tended to nearly match the larger fraction. After an analysis of the labels we found that the tree would overpredict or only predict the label that matched the greater fraction of the training data. Basically, the predictor was doing nothing. We attempted other predictors, such as kernel perceptron (including polynomial and Gaussian kernels), but continued to see nothing.

In order to uncover potential underlying structures to the data we performed hierarchical clustering using Ward's Method as a distance metric, with two bins. In order to do this we removed obviously correlated items such as number of bedrooms, and normalized all the variables left. We then plotted a histogram showing the distribution of the true y-values for each bin (Figure 2.3). Unfortunately the clustering did not seem to group the values in a way that represented our own binary classifications.

We believe the initial correlations we saw were simply a product of statistical fluctuations and small batch size. Each of the 36 data sets we ended up with was a slightly different size and we saw that the sum of the correlation coefficients was correlated with the size of the data set2.3.

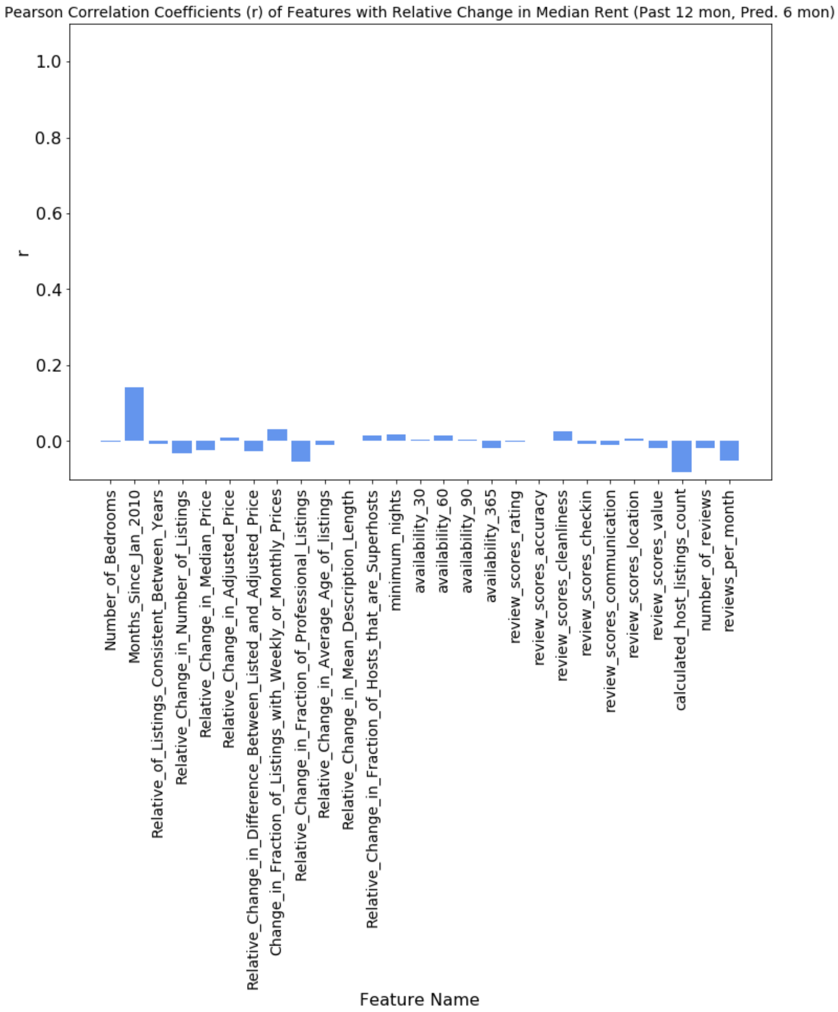


Figure 2: Correlation Coefficients of Features for Rent Predict: airbnbLag = 12 months, zillowLag = 6 months, monsPred = 6 months

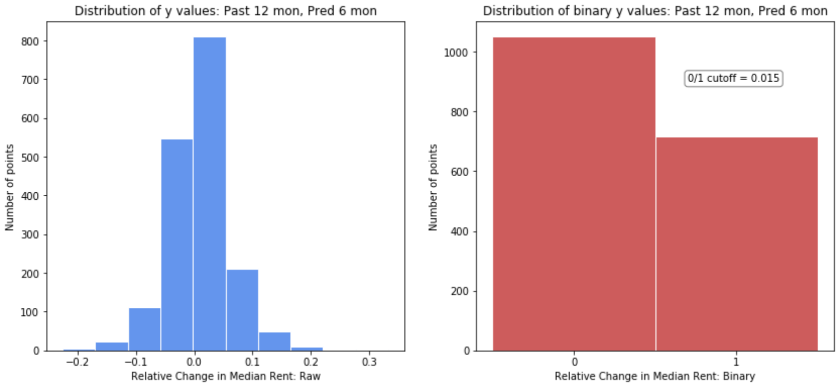


Figure 3: Original y labels and Binary y Labels for Rent Predict: airbnbLag = 12 months, zillowLag = 6 months, monsPred = 6 months

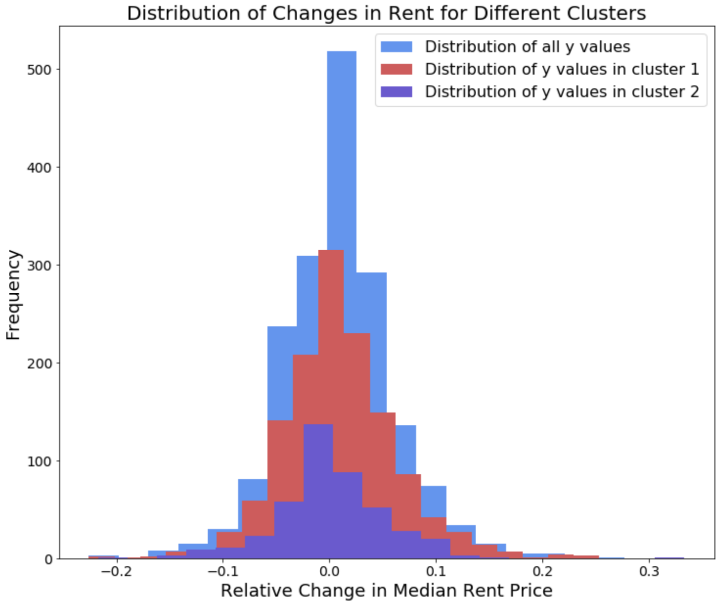


Figure 4: Distribution of all y values among the two clusters

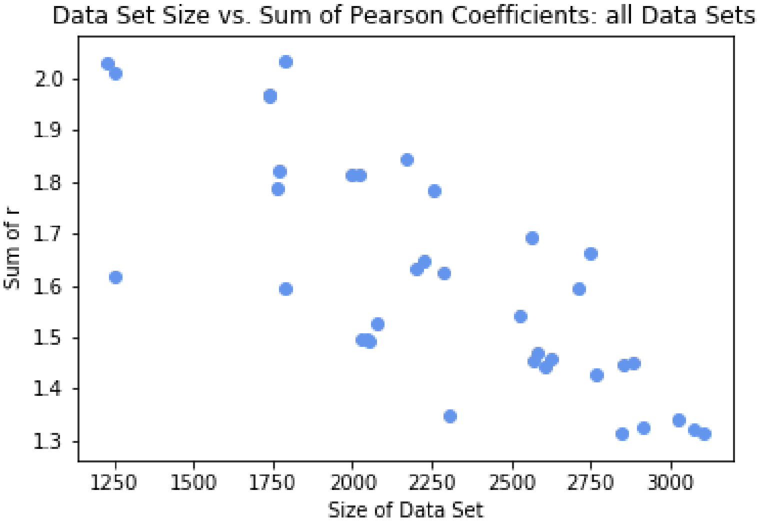
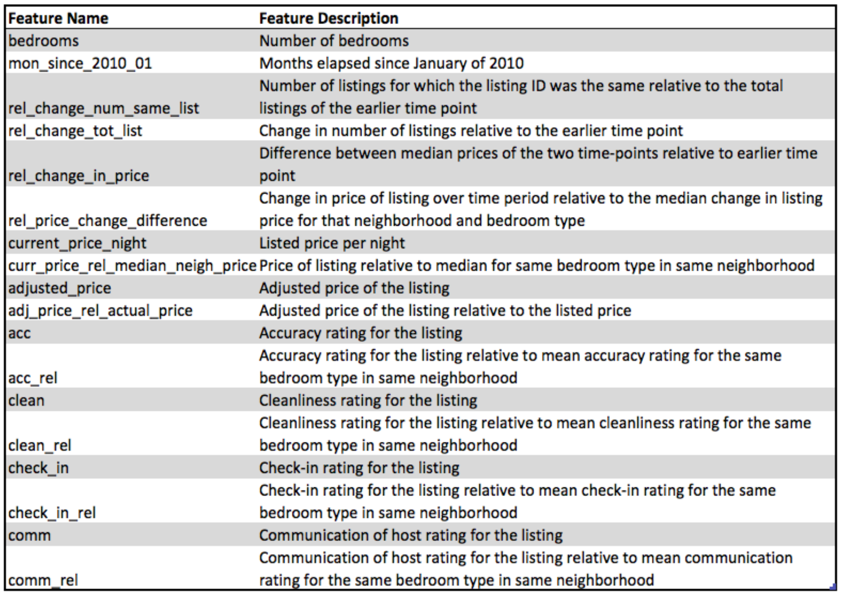


Figure 5: Sum of r values vs. size of data set

Table 2: Features used in prediction of reviews per month



Feature Name	Feature Description
bedrooms	Number of bedrooms
mon_since_2010_01	Months elapsed since January of 2010
rel_change_num_same_list	Number of listings for which the listing ID was the same relative to the total listings of the earlier time point
rel_change_tot_list	Change in number of listings relative to the earlier time point
rel_change_in_price	Difference between median prices of the two time-points relative to earlier time point
rel_price_change_difference	Change in price of listing over time period relative to the median change in listing price for that neighborhood and bedroom type
current_price_night	Listed price per night
curr_price_rel_median_neigh_price	Price of listing relative to median for same bedroom type in same neighborhood
adjusted_price	Adjusted price of the listing
adj_price_rel_actual_price	Adjusted price of the listing relative to the listed price
acc	Accuracy rating for the listing
acc_rel	Accuracy rating for the listing relative to mean accuracy rating for the same bedroom type in same neighborhood
clean	Cleanliness rating for the listing
clean_rel	Cleanliness rating for the listing relative to mean cleanliness rating for the same bedroom type in same neighborhood
check_in	Check-in rating for the listing
check_in_rel	Check-in rating for the listing relative to mean check-in rating for the same bedroom type in same neighborhood
comm	Communication of host rating for the listing
comm_rel	Communication of host rating for the listing relative to mean communication rating for the same bedroom type in same neighborhood

2.4 Reviews per Month Predictor for Listings on Airbnb

Reviews per month for a given listing can be viewed as a proxy for the occupancy rate for that listing. Occupancy rates are of interest to both the host of the listing, for obvious reasons, and to regulatory boards, however Airbnb does not release almost any of its data, including occupancy rates. The goal of this portion was to predict the reviews per month over the next N months for a given Airbnb listing.

To perform this prediction task we created feature sets that included information per listing. We looked at changes in the listing over the past M months and current information about the listing (both raw and relative to similar listings) and labeled the listing with the (number of reviews of the next L months)/L. For a listing to be used as a point it had to persist throughout that time frame. For example, if the past lag was 12 and the months predicted in the future was 12, that exact listing had to persist for that whole 24 month time period (using the unique listing ID to identify it).

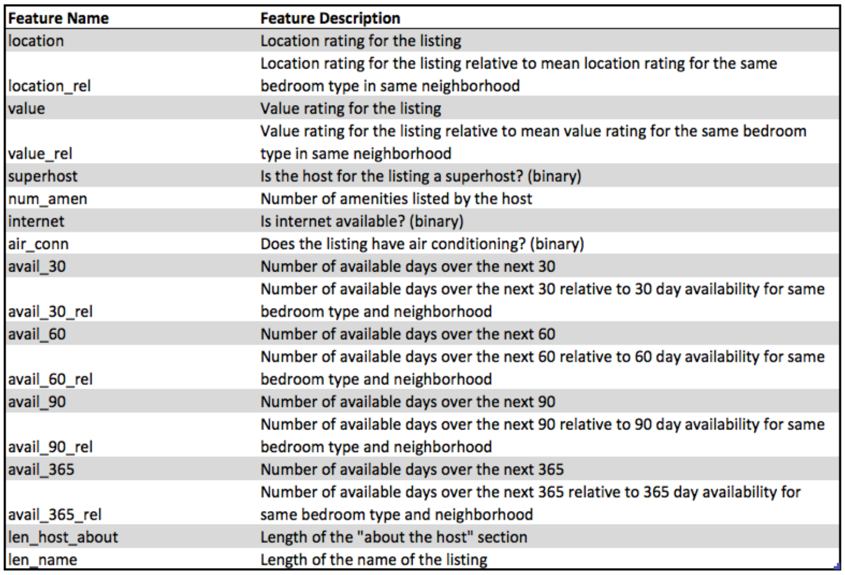
Since we wanted to look at trends for a single listing raw and relative to trends for listings of that same type (same neighborhood and number of bedrooms) and current features of the listing raw and relative to listings of the same type, the data sets were constructed only from neighborhood and bedroom pairs that had a sufficient number of listing at each time point examined (n > 40).

For each data point, we calculated 47 features, as described in the Tables2 and 3. The adjusted price was calculated as described for the rent predictor.

We created 5 different data sets with different time pairs (num.months in past, num. months in future to predict): (12,12), (12,6), (6,6), (6,3) and (3,3). We examined the correlation coefficients for each data set and found that there was siginificant correlation, with the best results for the (12,12) data set. We also decided to examine the (6,6) set since it performed relatively well and we wanted a predictor that would work for newer listings as well. The (3,3) set performed to poorest, probably because at these shorter time scales the seasonality is no longer smoothed out of the data. The correlation coefficients for the (12,12) set are shown, but all data sets had roughly the same pattern just scaled down a bit.2.4 We converted to data to a multiclass classification problem by round the reviews per month to the nearest whole number.2.4 The test and validation sets were all at least the number of predict months (either 6 or 12) months in the future relative to the last time point in the training set to ensure our train, validation and test sets would best represent how the data would be spread out in a real world application.

We used the (6,6) data to tune the parameters of a random forest, specifically the n_estimators, max_depth, and max_feat. We found that the max_depth was the most important feature 2.4 and

Table 3: Features used in prediction of reviews per month (continued)



Feature Name	Feature Description
location	Location rating for the listing
location_rel	Location rating for the listing relative to mean location rating for the same bedroom type in same neighborhood
value	Value rating for the listing
value_rel	Value rating for the listing relative to mean value rating for the same bedroom type in same neighborhood
superhost	Is the host for the listing a superhost? (binary)
num_amen	Number of amenities listed by the host
internet	Is internet available? (binary)
air_conn	Does the listing have air conditioning? (binary)
avail_30	Number of available days over the next 30
avail_30_rel	Number of available days over the next 30 relative to 30 day availability for same bedroom type and neighborhood
avail_60	Number of available days over the next 60
avail_60_rel	Number of available days over the next 60 relative to 60 day availability for same bedroom type and neighborhood
avail_90	Number of available days over the next 90
avail_90_rel	Number of available days over the next 90 relative to 90 day availability for same bedroom type and neighborhood
avail_365	Number of available days over the next 365
avail_365_rel	Number of available days over the next 365 relative to 365 day availability for same bedroom type and neighborhood
len_host_about	Length of the "about the host" section
len_name	Length of the name of the listing

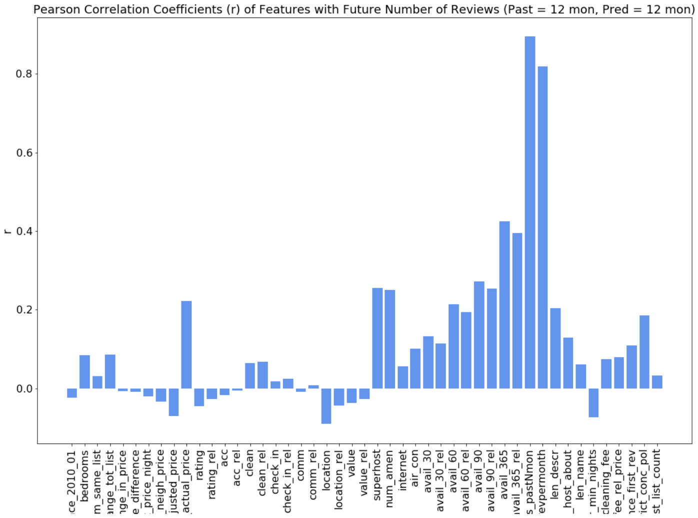


Figure 6: Correlation Coefficients of Features for Review Predict: past = 12 months, predict = 12 months

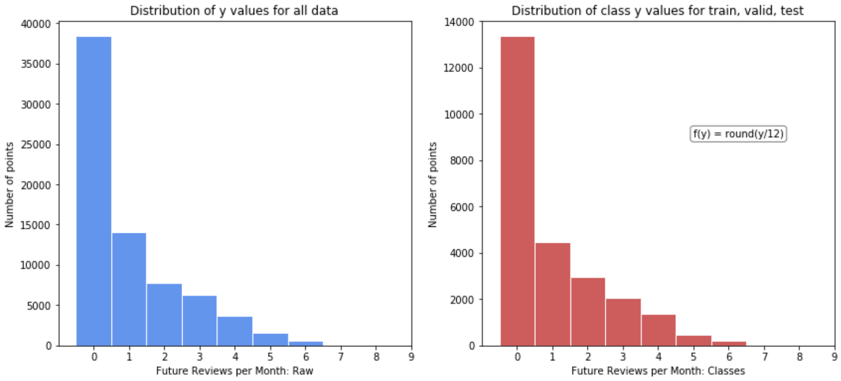
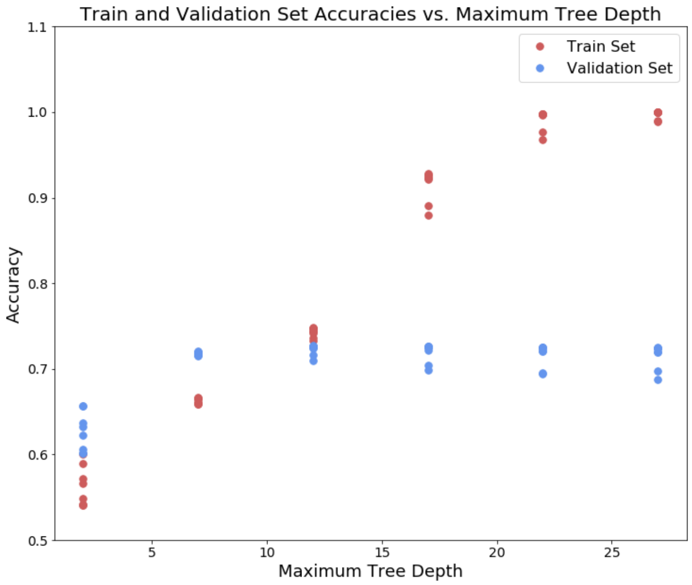


Figure 7: y distributions for the continuous and multiclass labels: past = 12 months, predict = 12 months



(1).png

Figure 8: Accuracy vs. Max Depth for Random Forest: past = 6 months, predict = 6 months

chose the following parameters for our final random forest: num_estimators = 1000 , depth = 12, max_feat = sqrt. The results for both the (12,12) set and (6,6) set are shown.2.42.4

We did attempt to use linear regressors (lasso and ridge) and the gradient boosting tree (xgboost, varied eta, num_round, gamma, and max depth) classifier as well, however the random forest outperformed both of these.

Similarly to the way we performed clustering on the rent prediction data set, we also performed hierarchical clustering on the review prediction data set. We removed the binary classifiers, number of bedrooms, and price, and normalized the remaining features. We performed clustering using a number of different cluster options, and then plotted histograms of the true distributions of calculated reviews per month. As in the earlier clustering exercise, we saw that the distributions within the clusters seemed to roughly match the distributions of the results in the full set, and so were not able to draw conclusions regarding the features clustered on.

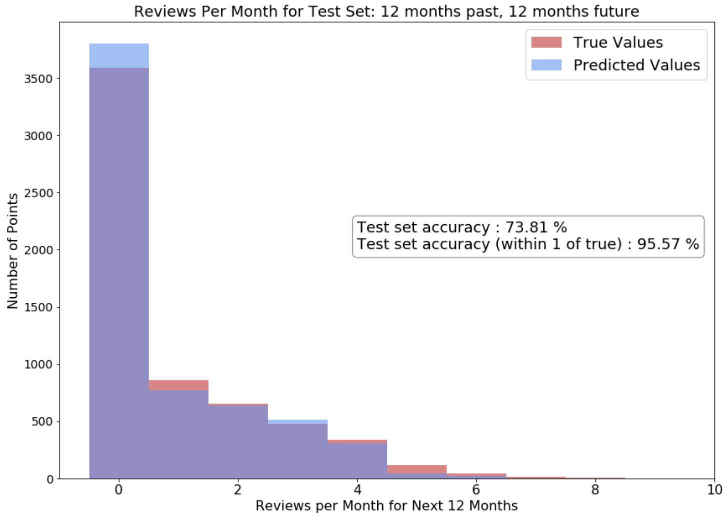


Figure 9: Random Forest Results for the (12,12) Review Predictor

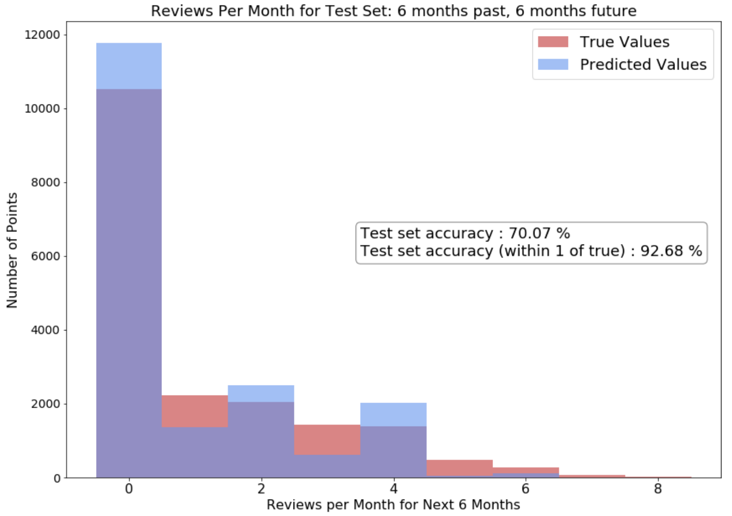


Figure 10: Random Forest Results for the (6,6) Review Predictor

3 Conclusion

Our analysis showed that the Airbnb data was unable to predict rental changes in New York City over the time periods measured. This supports the hypothesis that the two markets are sufficiently decoupled and that they neither predict each other, nor significantly influence each other.

We were able to predict, with an accuracy of 74% and 70% the number of reviews that a given listing would get over the next twelve and six months respectively. Allowing for a prediction within one we achieved accuracies of 96% and 93%, respectively.

References

[1] Inside airbnb. `http://insideairbnb.com/`, 2018.

[2] Zillow research data. `https://www.zillow.com/research/data/`, 2018.

[3] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[4] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed <today>].

[5] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[6] Elliot Mest. Data provider says airbnb report founded on 'flawed conclusions'. `https://www.hotelmanagement.net/operate/data-provider-says-airbnb-report-founded-flawed-conclusions`, May 2018.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[8] Scott Stringer. The impact of airbnb on nyc rents. `https://comptroller.nyc.gov/wp-content/uploads/documents/AirBnB_050318.pdf`, 2018.

[9] Georgios Zervas, Davide Proserpio, and John W. Byers. The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. *Journal of Marketing Research*, 54(5):687–705, 2017.