Craig Bentzen

Airbnb Chicago Price Prediction

Problem Statement:

Getting started as a host on Airbnb is a series of daunting questions: What things are guests looking for in an Airbnb listing? How do I maximize the chance my guests will give me good reviews? How do I stand out from the other 100 or so Airbnb listings in the surrounding area? And the ultimate question - What should I charge per night for my listing? For the last question, Airbnb provides a tool that suggests a price per night. I want to examine the suggested price tool using data analysis and machine learning techniques on Airbnb listing data from Chicago.

After all the work and expense of setting up a listing, seeing what Airbnb recommended for my listing was fairly demoralizing and financially scary. Airbnb's price estimator tool recommended that I rent my listing for an average of $40 during the week and $60 on weekends. I couldn't figure out how Airbnb calculated these numbers so I immediately discounted them. Instead I looked at a local hotel's nightly price, discounted 10% and viola - nightly average price of $125. I didn't have immediate success but after one month I was making money and fully booked on weekends and partially booked during the week.

After six months of successful bookings and positive reviews, did the price estimator tool ever update its price? Not significantly. The highest suggested weekend price was $70. Still fairly far off from the $125 guests were willing to pay. I was always intrigued with why the price estimator tool priced my listing so poorly. To satiate my curiosity, I want to gain insight into the following questions:

- Which features of an Airbnb listing are most impactful to price?
- Can I create a machine learning model to accurately predict the nightly rate hosts are listing their properties for?

Data Wrangling:

The source data for the Chicago Airbnb data set used in this analysis can be found on InsideAirbnb.com's website. Inside Airbnb provides monthly snapshots of the publically available Airbnb data for free use. Inside Airbnb segments the data by city for many of the worlds largest Airbnb Market's. I choose Chicago because I live in Chicago.

The source data is scrapped from Airbnb's website and certain fields are pre-aggregated for analysis. The raw data consisted of 8,519 records, one record per Chicago listing, with 72 columns or features.

The below list of cleansing and cleanup tasks were performed on the data:

- 1,396 records were dropped because the listing has yet to receive a review. A listing without reviews will be considered unused, fake, or the host is being paid off the platform. In any case, the price hasn't been confirmed by guest reviews, so we will remove these listings.
- 27 features contained null entries. Categorical/string fields were filled with 'none' and numerical fields were filled with '0'.
- Price column had to be converted to numeric data type and special characters were removed.
- 32 features or parameters were dropped for various reasons: measured the same thing, not helpful towards analysis, didn't have enough data points.
- Outliers were reviewed to determine if they are true outliers or bad data that should be removed.
- Low outliers on Price were removed because it doesn't make sense that any Airbnb listing should be 1 or 2 dollars a night.
- Word count features were created to summarize number of amenities, length of listing description, length of house rules, length of the about host section
- Number of reviews per month were rounded to the nearest whole number for easier segmentation
- Average Neighborhood Price per Average Neighborhood accommodates was created because it was found to be useful in modeling.
- During the data analysis phase, the neighborhood parameter was seen to be very influential to price (which seems logical). To explore the impact neighborhood had in combination with other features the following parameters were created: Average price by neighborhood, Average price by neighborhood and property type, Average price by neighborhood and number of accommodations, Average price by neighborhood, beds, and bathrooms, Average price by neighborhood and rounded reviews per month.

The final shape of the cleaned data set was 5,631 records and 24 parameters.

Exploratory Data Analysis:

The target variable is price per night. The below figure displays the distribution of the nightly price. The distribution is highly skewed to the right and multi peaked. The peaks occur at somewhat regular intervals. For example: 200, 250, 300, 350 and 400 are all peaks. Hosts must round to these numbers at higher prices.

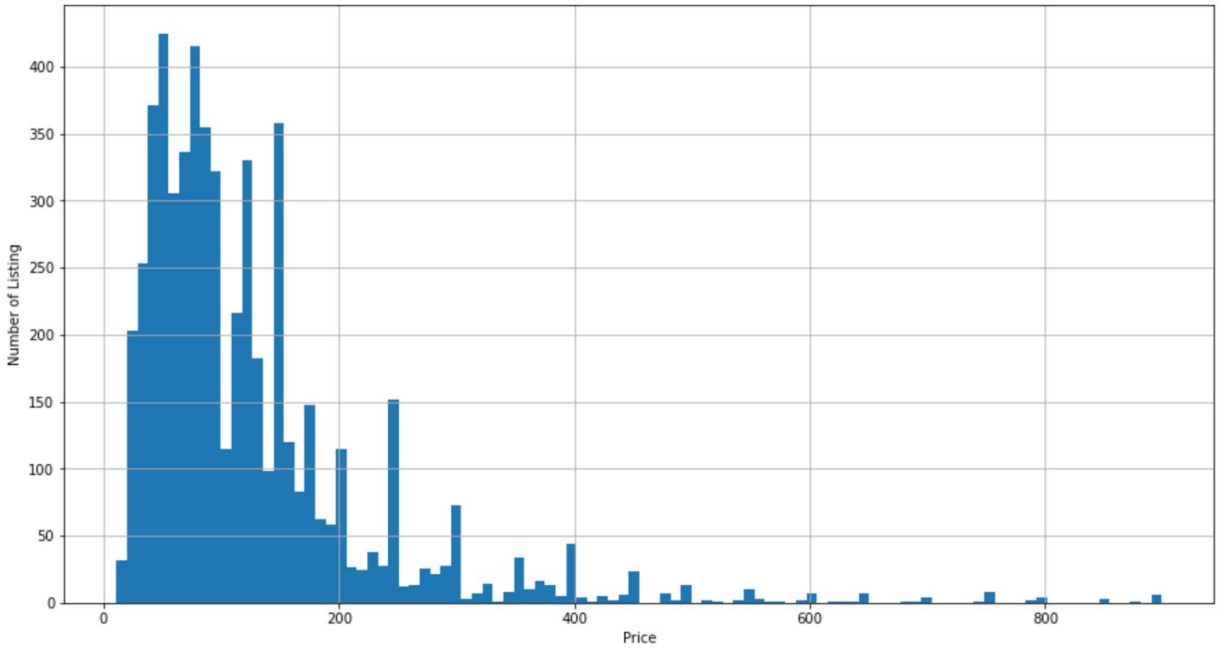


Figure 1: Histogram of Nightly Price

The first question I wanted to explore is how the neighborhood of an Airbnb listing affects price. The figure below displays the listing's neighborhood vs nightly price. The "Other" category are all neighborhoods that have less than 100 Airbnb listings.

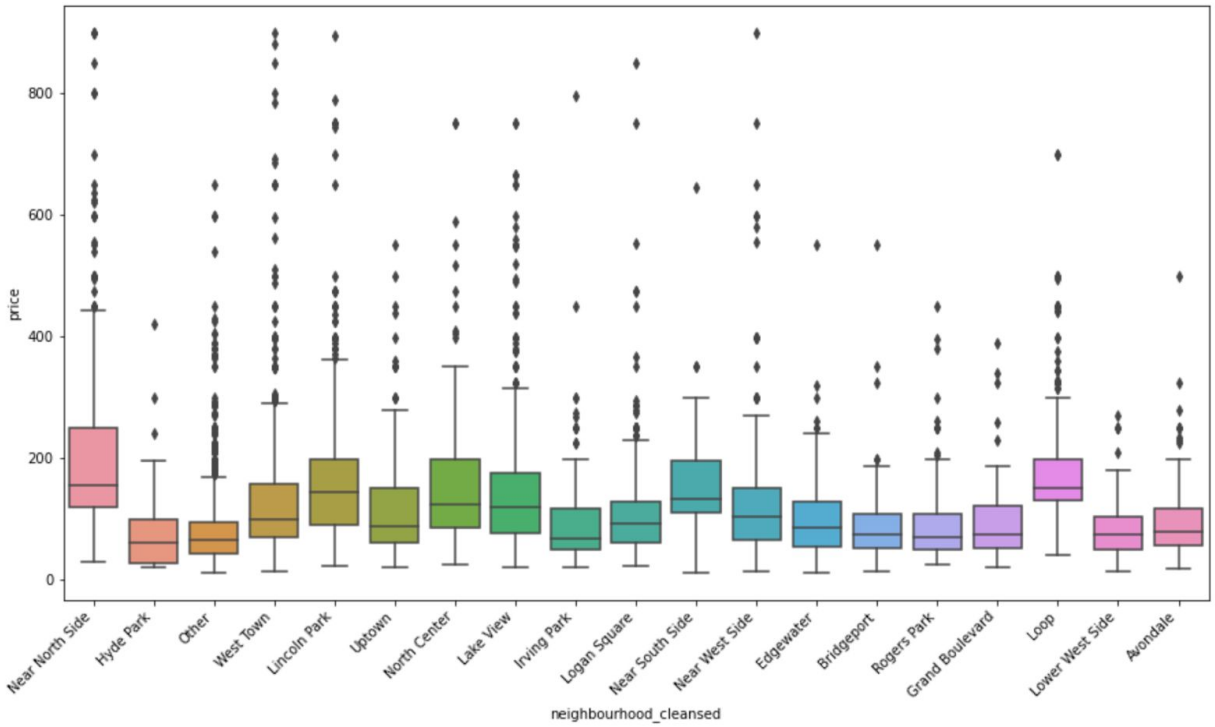


Figure 2: Box and whisker plot of Neighborhood vs. Nightly Price

There are many outliers in the above box and whisker plots. Is this bad data or the highly skewed nature of the distribution of price? In the next figure the price is segmented by the number of people the listing can accommodate.

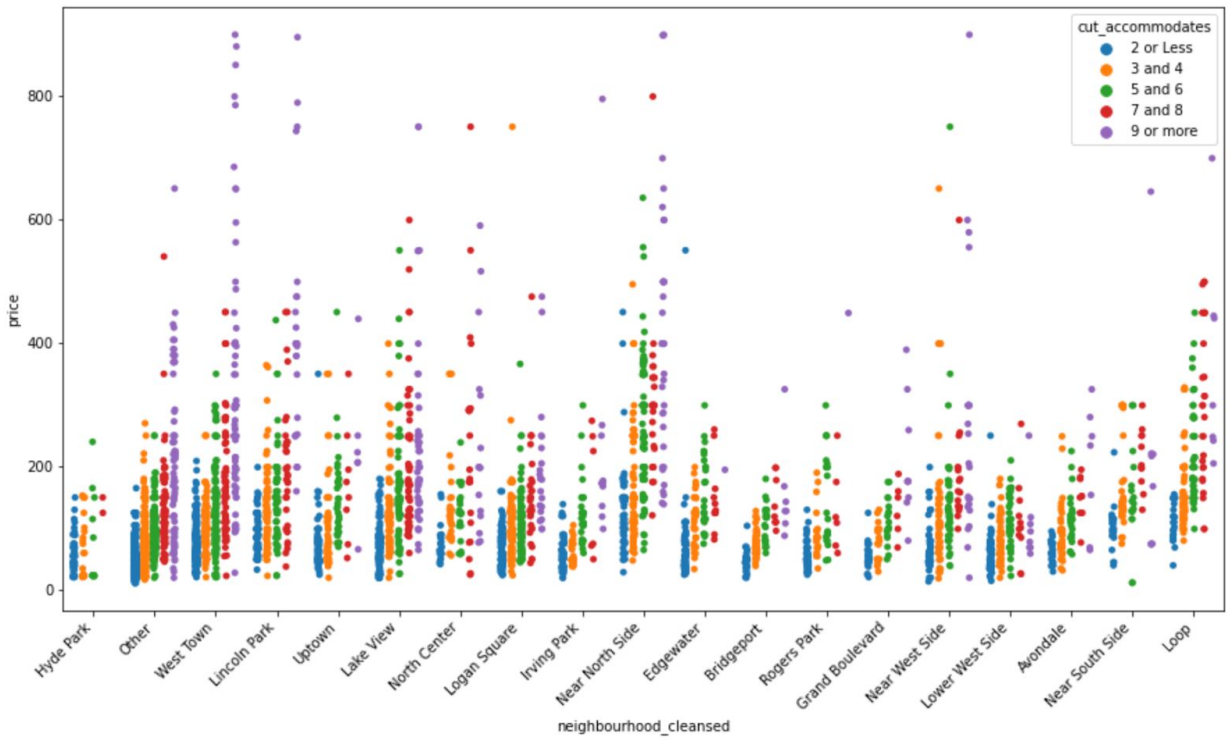


Figure 3: Scatter plot of Price vs Neighborhood and segmented by Accommodates

The vast majority of listings priced above $400 can accommodate 5 or more guests. For that reason we should leave the outliers because we want to be able to predict Airbnbs of any size, not just one bedroom listings.

To hosts I would use this information to mean that listings with five or more guests have a significantly larger range of values.

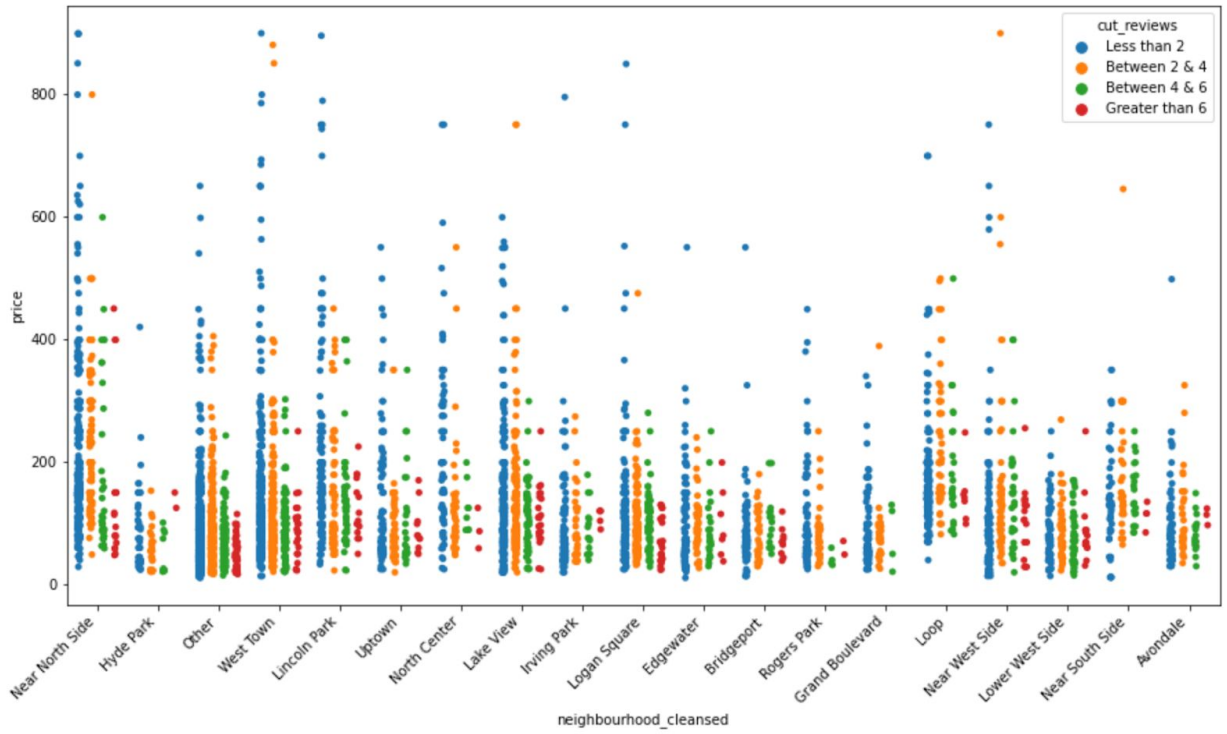


Figure 4: Scatter plot of Price vs Neighborhood segmented by Number of Reviews per Month

The above figure segments price by the number of reviews per month. Almost all of the listings with six or more reviews a month have a price lower than $200 a night. Listings with a price of over $400 per night get rented at a decreased rate because the fast majority receive two reviews or less a month.

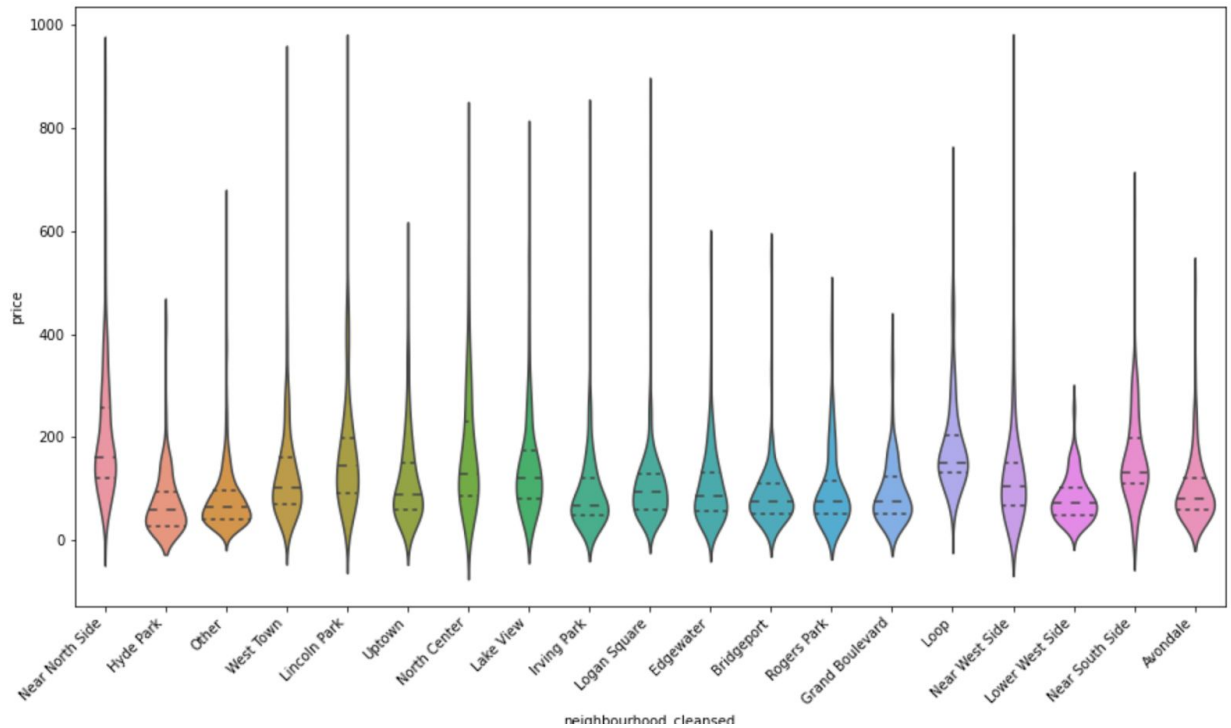


Figure 5: Violin Plot of Price vs Neighborhood

The above violin plots visualize the distribution of neighborhood nightly price. The violin plots with the widest distribution spread are generally the most expensive neighborhoods. Near North Side, Lincoln Park, North Center, Lake View, and Near West Side are all expensive neighborhoods with a lot of tourism. As a host, the take away is that the price for these neighborhoods will have a large range of values that guests are willing to pay.

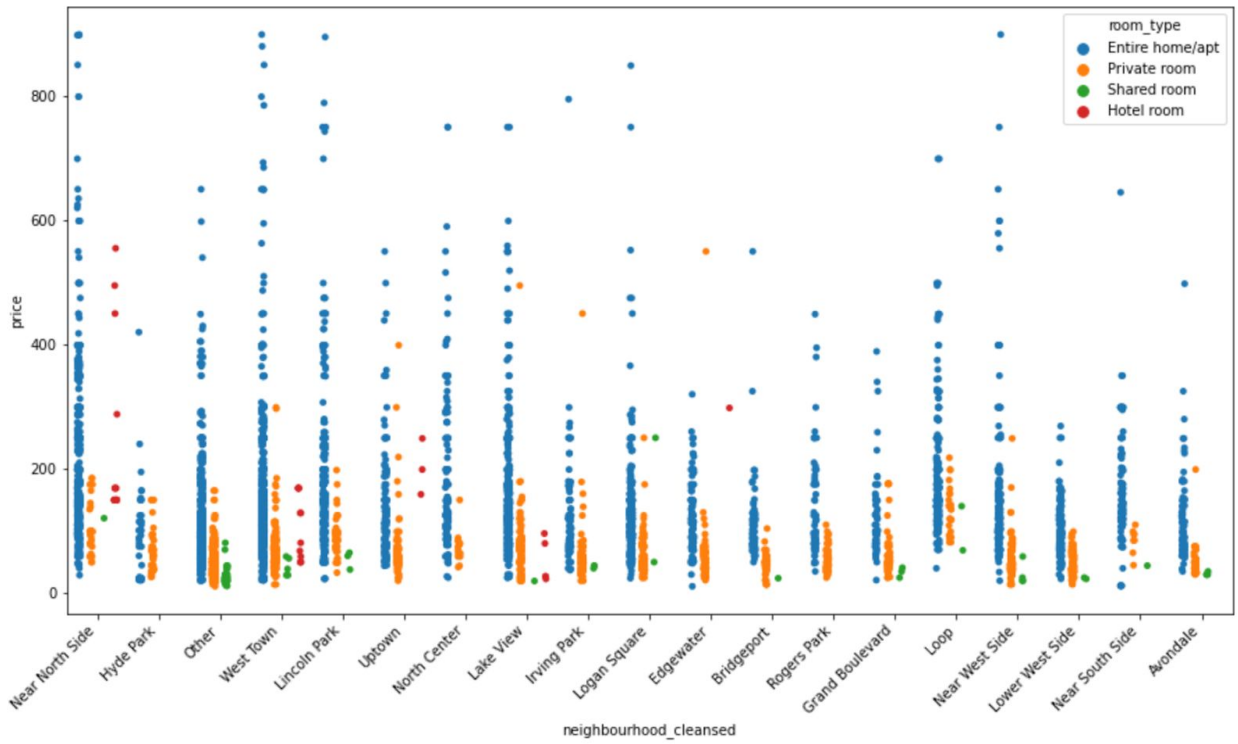


Figure 6: Scatter plot of Price vs Neighborhood segmented by Listing Type

The vast majority of listings on Airbnb are for Entire Homes and Apartments. Very few private rooms are rented for above $200.

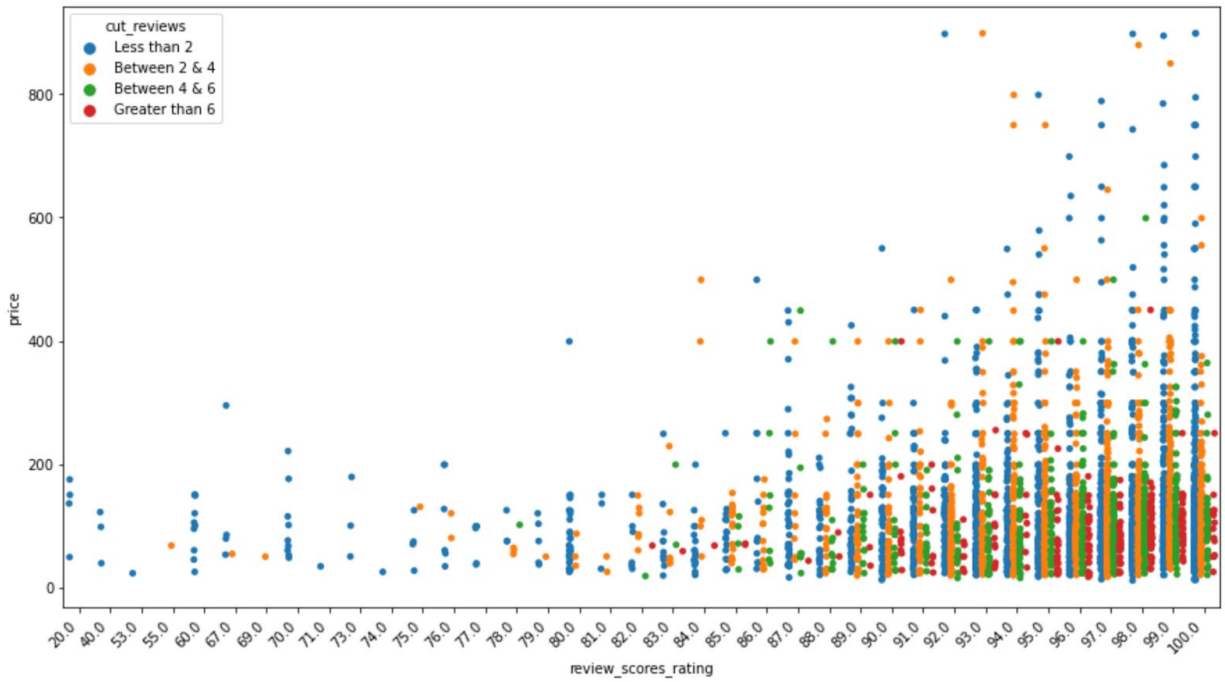


Figure 7: Price by Review Rating

The above figure shows the review rating and price segmented by the number of reviews. It is interesting to note that the majorting of listings with 6 or more ratings a month have a review rating of at least 90%. Very few listing that want to charge more than $400 a night have a review rating less than 90%.

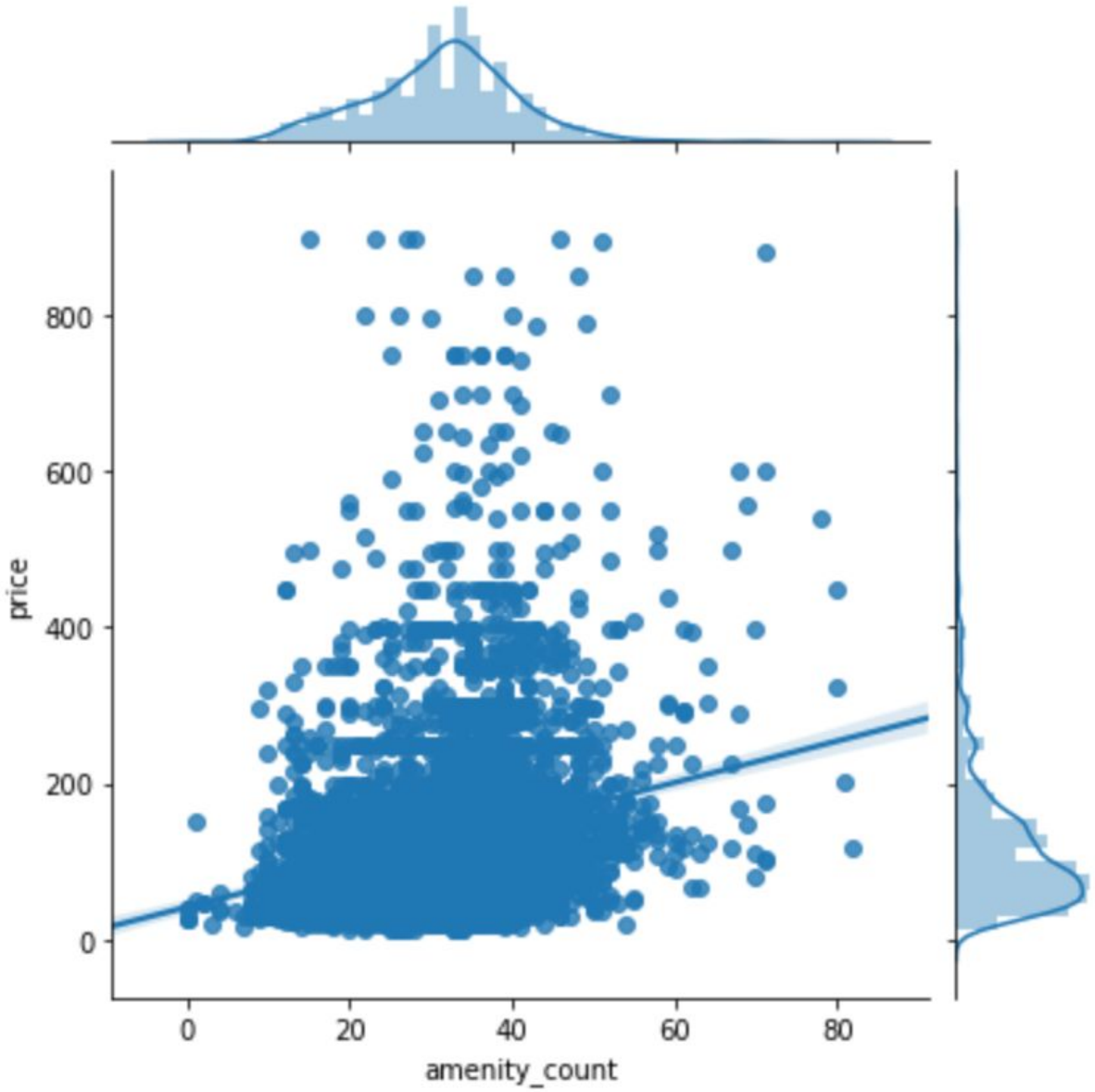


Figure 7: Joinplot of Amenities Count vs Price

Amenities are a free text field on the Airbnb platform. Hosts can list as many amenities as they feel warranted. But do these amenities affect price at all. The above join plot does find a slight positive correlation to price.

Is it possible to glean any understanding from high performing Airbnb listings? I'm going to define a high performing Airbnb as any listing that has a greater than 98% rating and more than 3.2 ratings per month. The top 25% in each category.

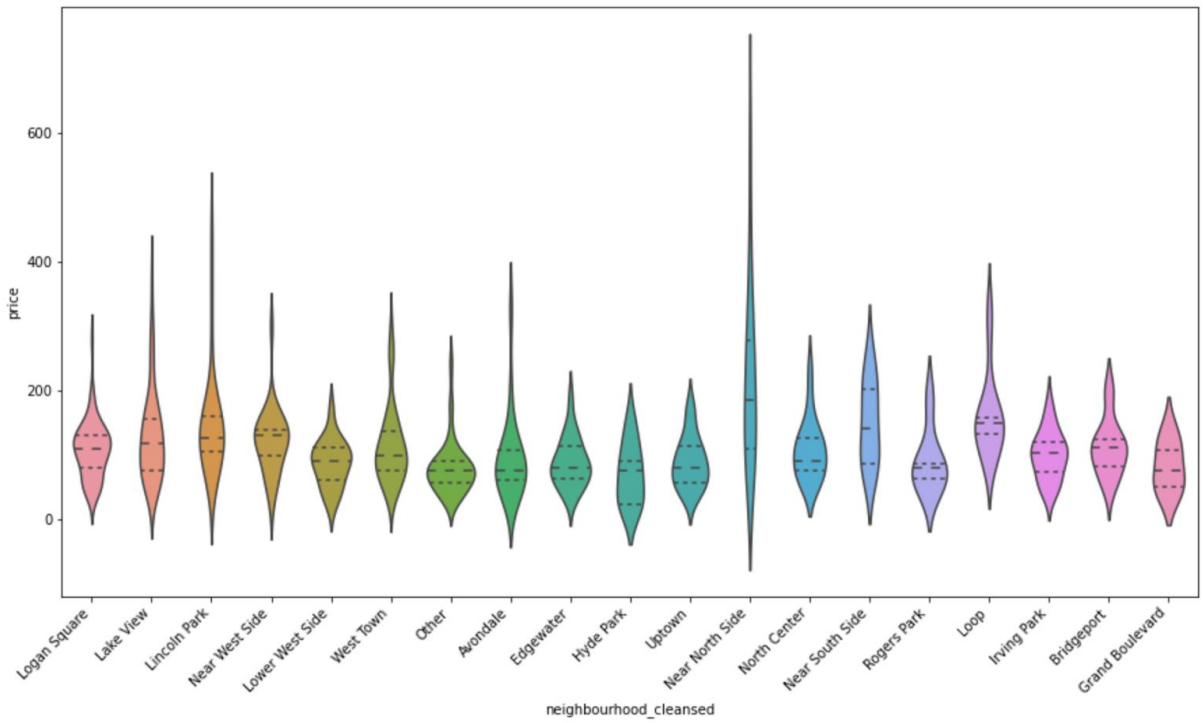


Figure 8: Violin Plot of Neighborhood vs Price for High Performing Airbnb's

The above violin plot of neighborhood and price for high performing Airbnb's shows that there are fewer outliers and the distributions generally are more concentrated. This makes sense as high performing Airbnb's know their worth. Also the neighborhoods with the most tourists again have the distributions with the greatest spread and number of outliers.

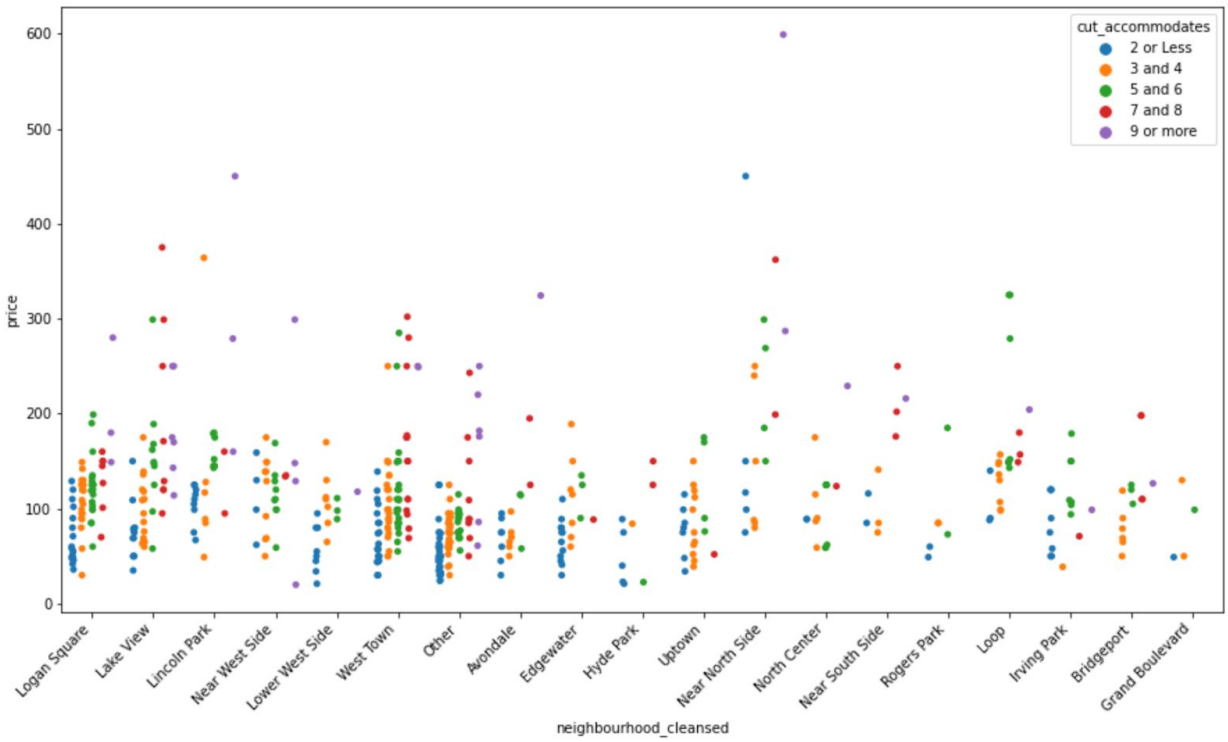


Figure 9: Price vs Neighborhood segmented by Accommodates of High {erforming Airbnbs

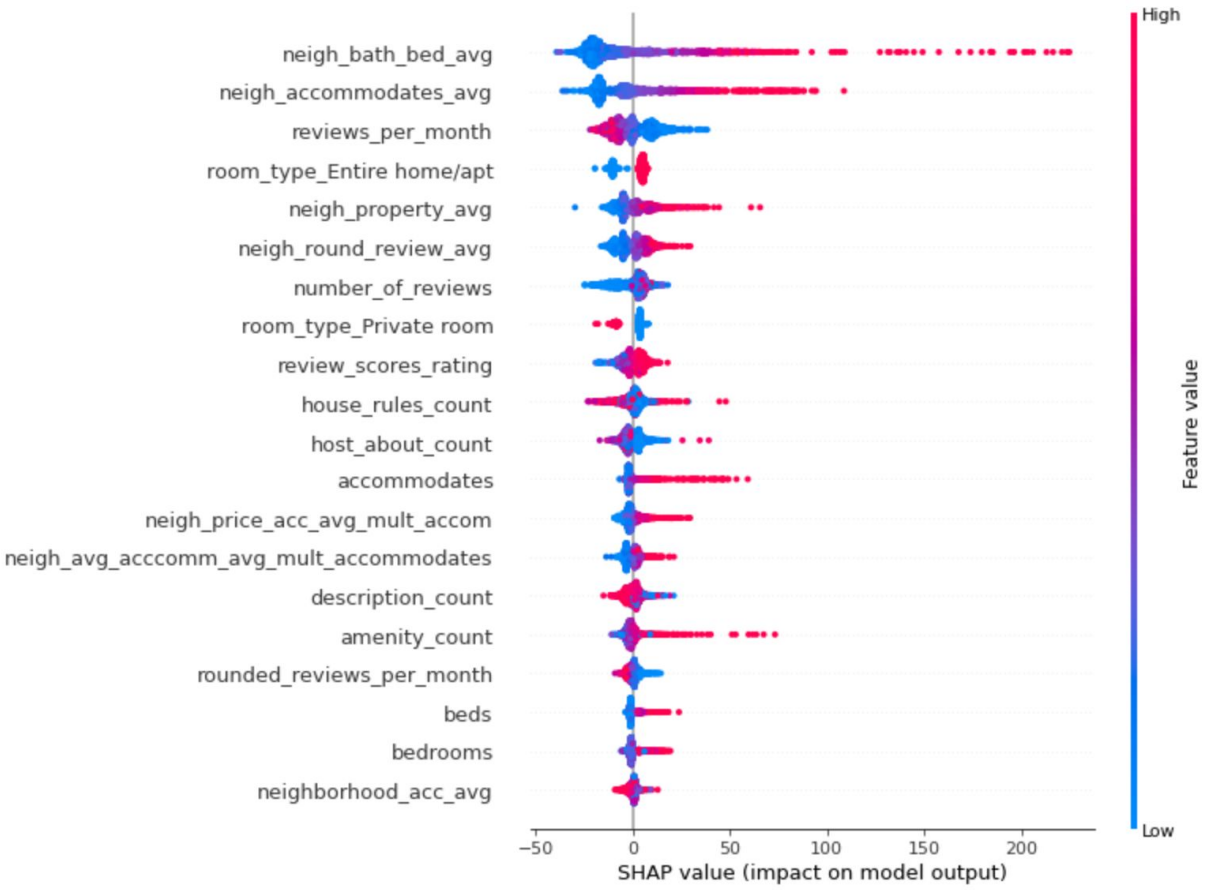
The above shows that Airbnb's that have more accommodations know to charge more on average.

Data Modeling:

The below table depicts various regression models and techniques. Using the below as a starting point I want to examine and tune some of the more effective models and find the most important features.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
0	CatBoost Regressor	32.6401	3099.3449	55.3468	0.6994	0.3730	0.3284	4.6758
1	Light Gradient Boosting Machine	33.4460	3196.5163	56.1212	0.6910	0.3769	0.3325	0.2075
2	Random Forest	33.2856	3244.8917	56.5886	0.6869	0.3842	0.3389	1.8154
3	Extra Trees Regressor	32.6423	3344.4685	57.4112	0.6775	0.3806	0.3264	1.2979
4	Gradient Boosting Regressor	35.3971	3484.5399	58.6816	0.6611	0.4032	0.3691	1.1972
5	Extreme Gradient Boosting	35.5073	3545.7716	59.1790	0.6555	0.4038	0.3698	0.3640
6	TheilSen Regressor	37.6975	3846.5750	61.6648	0.6272	0.4373	0.3905	4.8615
7	Ridge Regression	38.0998	3858.1641	61.7590	0.6247	0.4591	0.3974	0.0126
8	Linear Regression	38.1821	3870.9229	61.8620	0.6232	0.4543	0.3984	0.0247
9	Bayesian Ridge	37.9036	3909.3343	62.1405	0.6213	0.4376	0.3951	0.0416
10	Lasso Regression	38.2298	3955.6069	62.5065	0.6170	0.4405	0.4013	0.1115
11	Elastic Net	38.4015	3973.5930	62.6526	0.6151	0.4365	0.4033	0.1229
12	Orthogonal Matching Pursuit	38.4343	4004.3323	62.9077	0.6122	0.4479	0.4045	0.0109
13	Huber Regressor	38.1736	4011.8711	62.9461	0.6114	0.4384	0.3863	0.3566
14	K Neighbors Regressor	39.7071	4145.3565	64.0042	0.5998	0.4438	0.4046	0.0627
15	Passive Aggressive Regressor	48.0867	5060.4552	70.7406	0.4988	0.5922	0.5540	0.0309
16	Support Vector Machine	44.1171	6442.6081	79.3800	0.3988	0.4693	0.4194	1.5928
17	Decision Tree	45.6164	6513.9001	80.2936	0.3709	0.5243	0.4214	0.0720
18	Lasso Least Angle Regression	60.8185	8182.8585	89.7399	0.2290	0.6650	0.7815	0.0137
19	AdaBoost Regressor	78.5526	8853.0355	93.9313	0.1296	0.8055	1.2040	0.6357

Initial impression is that the model is okay. The Decision Tree based models are better at maximizing the R squared value and minimizing the Root Mean Squared Error. The RMSE of $55.34 for the CatBoost Regressor is okay considering the average listing is priced at $125 per night. To evaluate the model we also need to look at the residual plots.



The above figure displays the SHAP values for the CatBoost Regressor model. The SHAP values for the various features show which features continue the most to the model positively or negatively. To read a SHAP value chart. Values to the right of 0 on the X-axis contribute positively to the target variable and values to the left contribute negatively to the target. Values that are pink have a higher impact on the target variable than values that are blue. So as Price per Accommodation increases it positively affects price. Alternatively, when a listing only accommodates one guest (accommodates_1), it negatively impacts price.

The means that the most important features relate to the neighborhood average of prices for the number of bedrooms, bathrooms, and accommodates. Reviews per month, property type, review rating, and amenity count were also contributing features in the model.

Figure 10: Random Forrest Regressor Residual Plot

There are a couple things to note in the above residual plot. First, at high predicted values the residuals are not uniform. Second, the R squared values are significantly different. Both of these indicators mean that the model is not ideal and overfitting is an issue.

Various techniques to address overfitting were employed on the data without success. Normalizing and features did not affect overfitting.

The modeling process is iterative. To better model the data the high performing Airbnb's were selected from the data. A high performing Airbnb has 3.2 reviews per month and a rating review score of 98% - top 25% in each category. These Airbnb's are confirmed by guest bookings and reviews to have a certain value. Consequently, modeling these Airbnb's might produce better results.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
0	Lasso Regression	23.4939	1199.3977	34.3761	0.6812	0.2820	0.2392	0.0078
1	Bayesian Ridge	24.2946	1252.2015	34.9647	0.6747	0.2906	0.2458	0.0095
2	Huber Regressor	24.4018	1259.8590	35.0793	0.6695	0.2904	0.2418	0.0664
3	Orthogonal Matching Pursuit	23.9952	1229.1324	34.8963	0.6665	0.2920	0.2422	0.0077
4	TheilSen Regressor	25.0490	1303.6608	35.6863	0.6572	0.3084	0.2508	2.1970
5	Elastic Net	25.3708	1393.6484	36.6726	0.6571	0.3064	0.2698	0.0063
6	Ridge Regression	25.0900	1317.7917	35.8903	0.6558	0.3018	0.2568	0.0054
7	Linear Regression	25.3495	1335.0791	36.1301	0.6515	0.3045	0.2599	0.0069
8	CatBoost Regressor	25.9363	1566.3945	38.0969	0.6372	0.3170	0.2758	1.9864
9	Extra Trees Regressor	26.3986	1633.6921	39.1266	0.6084	0.3216	0.2785	0.2883
10	Passive Aggressive Regressor	27.3124	1452.2862	37.7389	0.6031	0.3766	0.2855	0.0057
11	Extreme Gradient Boosting	26.5153	1679.8188	39.5945	0.6027	0.3120	0.2727	0.1625
12	Random Forest	26.8254	1650.3381	39.6076	0.5974	0.3200	0.2804	0.2977
13	Gradient Boosting Regressor	26.6456	1670.7803	39.8191	0.5861	0.3174	0.2766	0.1423
14	Lasso Least Angle Regression	28.9257	1707.6341	40.7468	0.5794	0.3623	0.3362	0.0065
15	Light Gradient Boosting Machine	29.1047	1790.9360	41.1227	0.5678	0.3330	0.2971	0.1019
16	AdaBoost Regressor	29.3696	1789.6038	41.3646	0.5613	0.3663	0.3480	0.1337
17	K Neighbors Regressor	30.6068	2010.8860	43.8908	0.5119	0.3706	0.3323	0.0067
18	Decision Tree	37.7678	3178.5736	55.0438	0.2259	0.4480	0.3776	0.0100
19	Support Vector Machine	38.9906	3854.8120	59.7697	0.1457	0.4647	0.4059	0.0202

The above regression results for high performing Airbnbs are significantly better than the original models on all the data. A RMSE, or standard deviation, of $34.3 is a solid improvement. It is also interesting that linear regression models do better than the decision tree based models. Next, let's take a look at the residual plot.

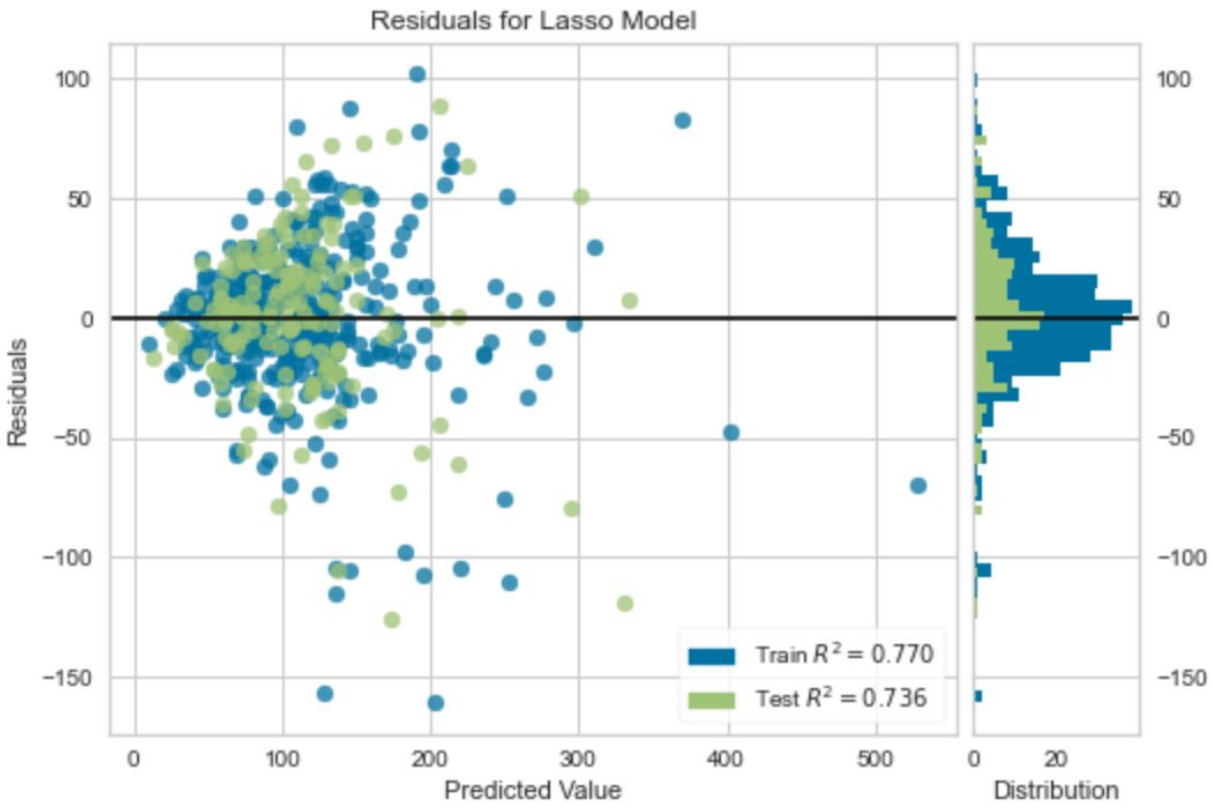


Figure 10: LASSO Regression Residual plot for High Performing Airbnb's

The residual plot of the LASSO Regression model has a uniform distribution and the training and test R squared values do not suggest overfitting.

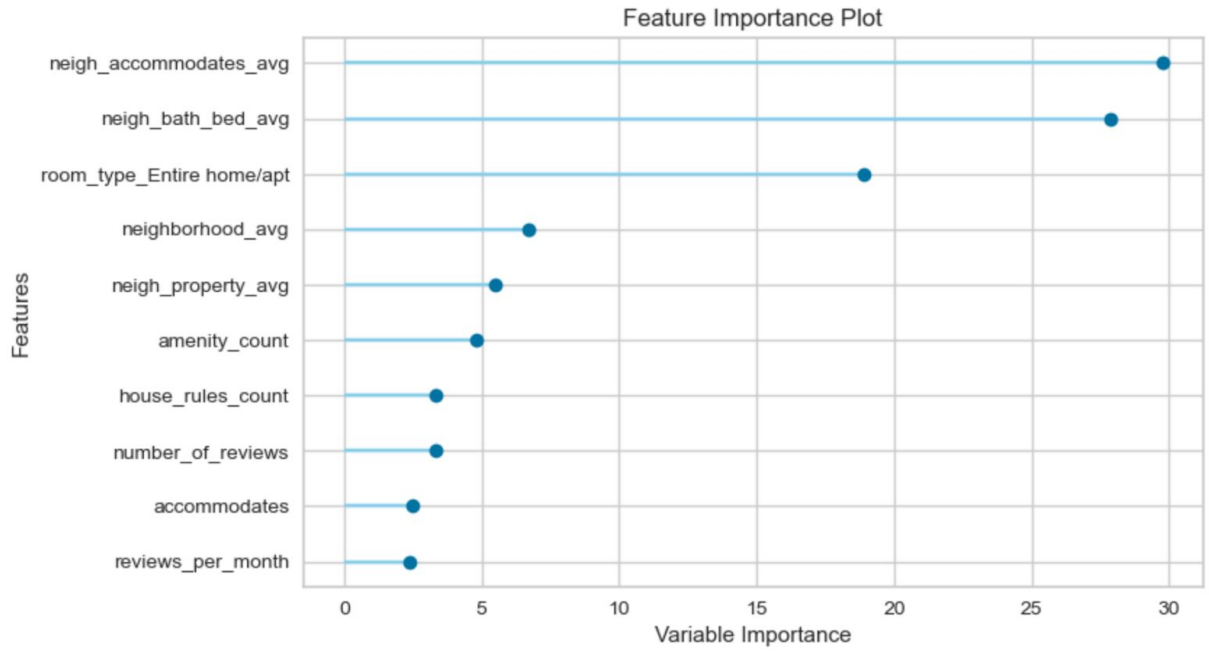


Figure 11: LASSO Regression Feature importance plot

Regardless of high performing Airbnbs or not, the same features impact the model.

Conclusions:

With and RMSE of $34.4 and Mean Absolute error of around $23.00 it is possible to build an okay model to predict Airbnb nightly price for high performing listings. The most important features relate to the number of accommodations and neighborhood. Creating features around the most important features allowed the model to become far more accurate.

Future Research:

I think it would be interesting to inspect the records that cause the most error. There were a lot of very low nightly prices that don't make a lot of sense to me, example a listing in a very nice neighborhood for $15 a night. This requires a lot of manual effort and review.

I also think gathering the photo data on the listings would be useful. Having just the count of photos and relative brightness might be useful enough. I have to imagine that guests won't rent rent an Airbnb with only two dark photos and would negatively affect price. Going down the rabbit hole you could use computer vision to confirm all the rooms, beds, and bathroom exist.

Doing some natural language processing on the title, description, about, and reviews would also be interesting. Do successful listing use similar phrases? Do some phrases contribute to price?