Analyzing Airbnb Listing Price and Host Type in New York City

Presented by Team 3B

Yulong Gong, Peter Mankiewich, Ruchika Venkateswaran, Phoenix Wang, Yangyang Zhou

About Us

Yulong Gong

Peter Mankiewich

Ruchika Venkateswaran

Phoenix Wang

Yangyang Zhou

airbnb

AGENDA

- Project Overview

- About the Dataset

- Data Cleaning

- Exploratory Data Analysis

- Machine Learning:

Classification & Regression

- Conclusion & Future Work

Project Overview

■ Objectives

- Analyzing neighborhood popularity for superhosts and non-superhosts

- Comparing the pre- and post-COVID listing prices

- Predicting the type of host and understanding what attributes contribute to the classification of a superhost

- Predicting the price for Airbnb listings in New York City and understanding what features contribute to profitable business opportunities for Airbnb

■ Data Source

The datasets have been taken from InsideAirbnb and were last updated in October 2020.

About the Dataset

Listings
There are 44666 unique observations (listings) and 74 features.

Neighborhoods
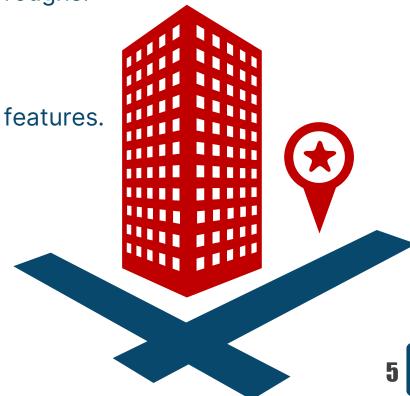There are 230 unique observations (neighborhoods) in 5 boroughs.

Reviews
There are 1003064 unique observations (comments) and 6 features.

Calendar
There are 18747307 observations and 7 features.

Data Cleaning

Listings

- **Step 1: Unifying Formats and Data Types**
  - Converted each column to the correct format and data type

- **Step 2: Dealing with Missing Values**
  - Dropped empty columns (E.g., "bathrooms")
  - Dropped 25 listings had price as $0 and 1595 listings had no beds

- **Step 3: Analyzing Outliers**
  - Some listing prices are greater than 5,000 USD
  - The maximum number of beds is 40 and it hosts 16 people

After cleaning, we now have **31880** unique observations.

◆ Data Cleaning - Outlier Examples

Data Cleaning

Reviews

- Dropped comments with missing values and those that have less than 3 characters, after which we have **1002383** comments for **34142** listings.

Calendar

- Converted "price" and "date" columns to float and datetime data types
- Brought in archived datasets to compare price fluctuations between 2019 and 2020

Exploratory Data Analysis

Hosts

Neighborhood

Price

Reviews

airbnb

9

Exploratory Data Analysis for Neighborhoods

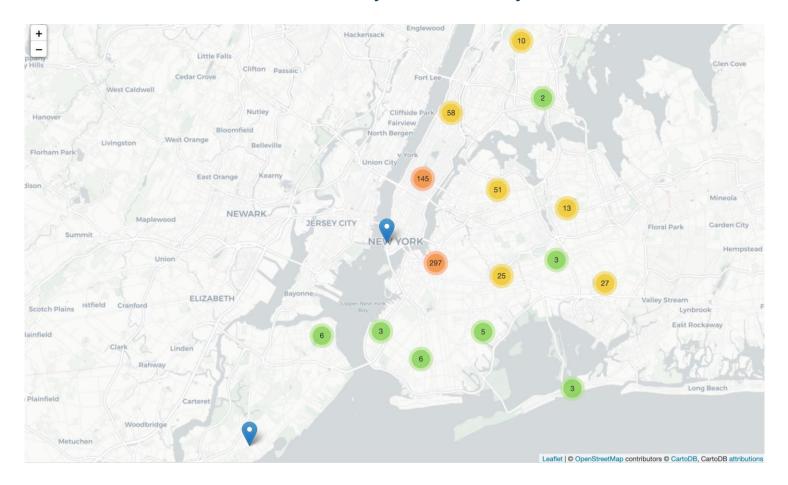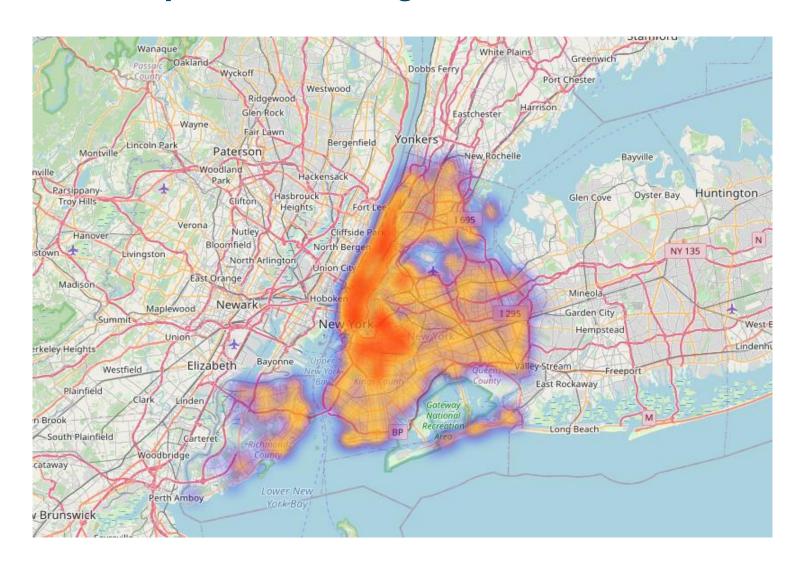## Most Reviewed Listings

- There are 672 apartments with more than 200 reviews, out of which 256 apartments are from Manhattan.
- 278 reviews are from Brooklyn followed by 115 in Queens.

Exploratory Data Analysis for Neighborhoods

Heatmap of NYC Listings

Exploratory Data Analysis for Neighborhoods
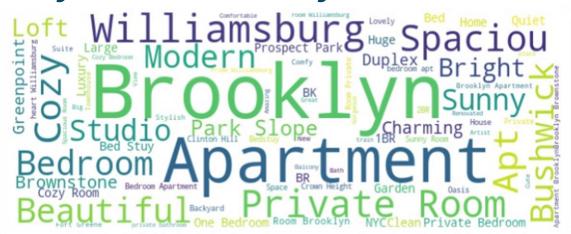
Keywords in Manhattan

Keywords in Brooklyn

Exploratory Data Analysis for Neighborhoods
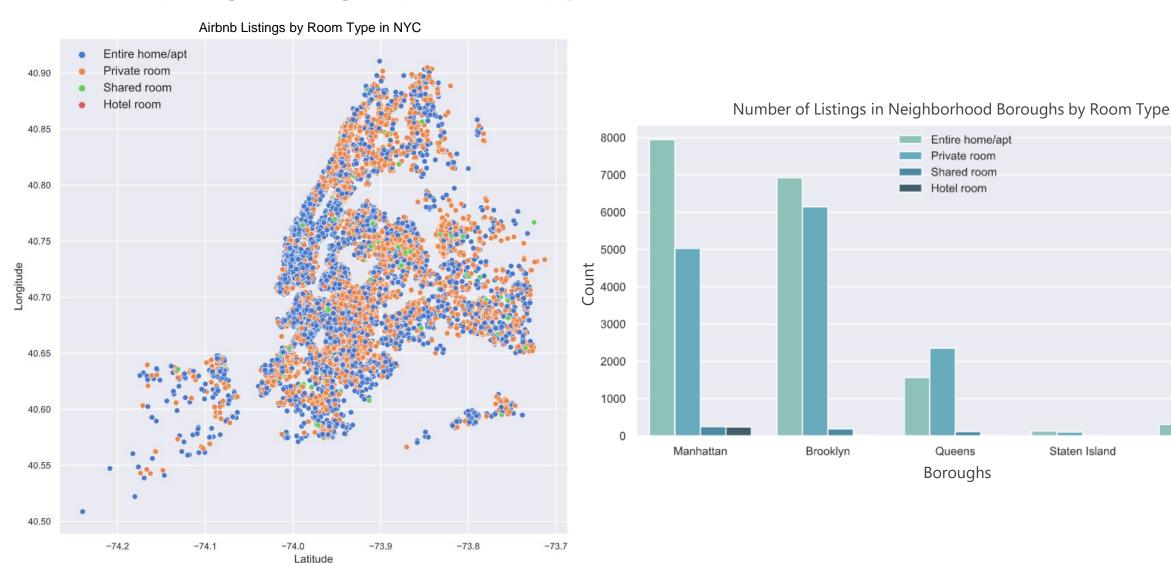
Analyzing Listings by Room Type
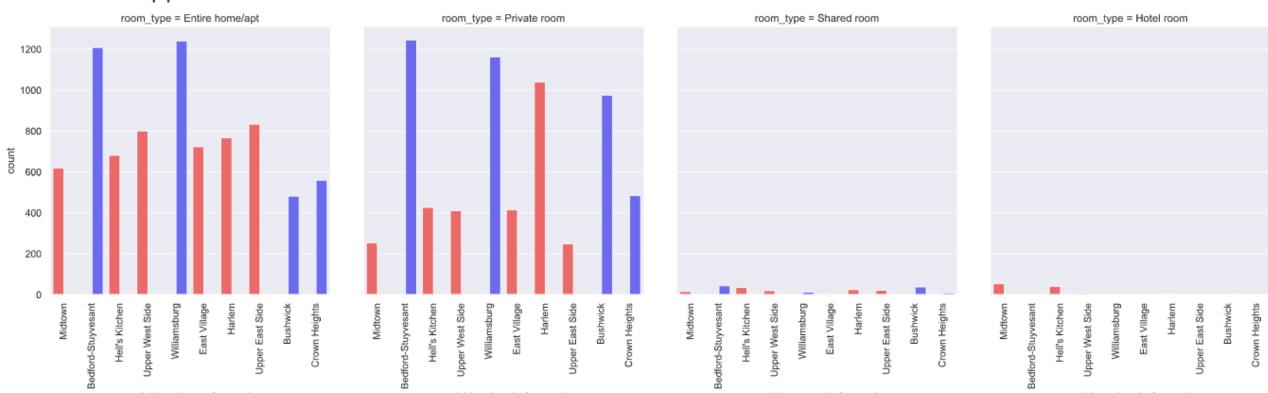
Airbnb Listings by Room Type in NYC

Number of Listings in Neighborhood Boroughs by Room Type
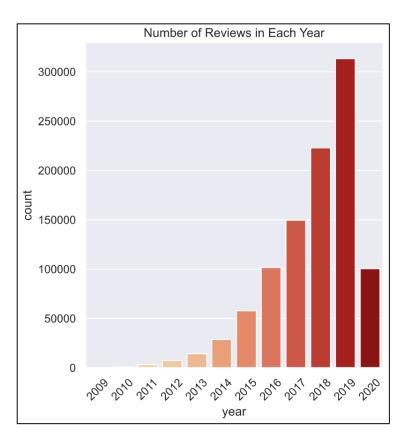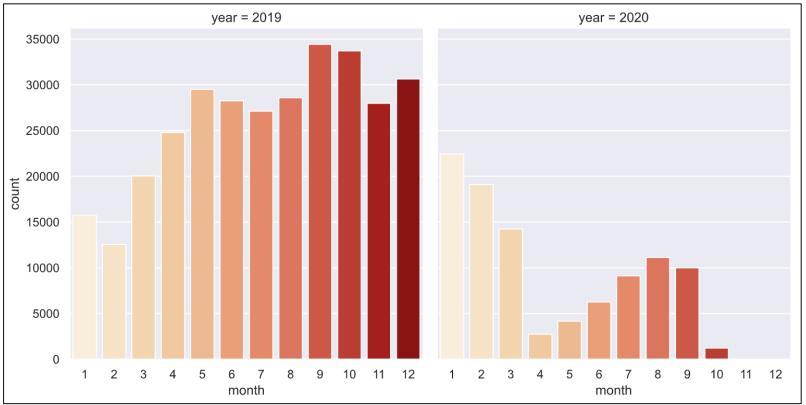
Exploratory Data Analysis for Reviews

Impact of COVID-19 on the Number of Reviews

- The number of reviews drastically reduces approximately 66% in 2020.

Exploratory Data Analysis for Reviews

Comments for the Top-Rated Listing

Comments for the Lowest-Rated Listing

Exploratory Data Analysis for Hosts

Superhost is a host who consistently receives good reviews and provides guests with extraordinary experiences over at least a year.

4.8+ overall rating

10+ stays in the past year

90% response rate

<1% cancellation rate

Exploratory Data Analysis for Hosts
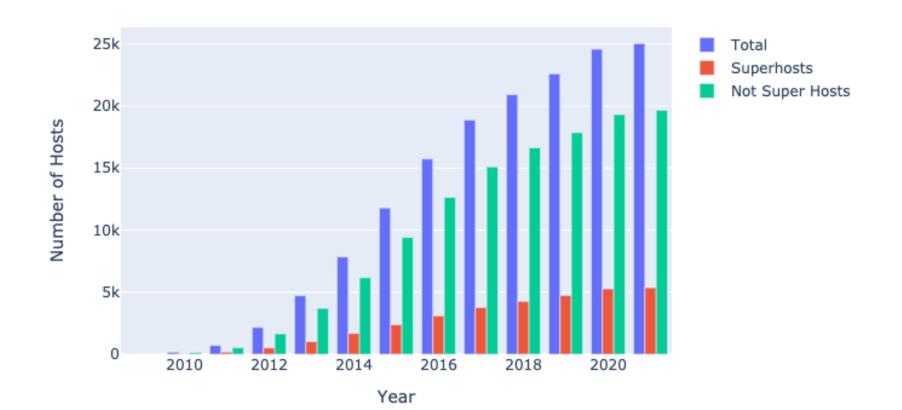
Breakdown on Number of Hosts by Type (Superhosts/Non-Superhosts)
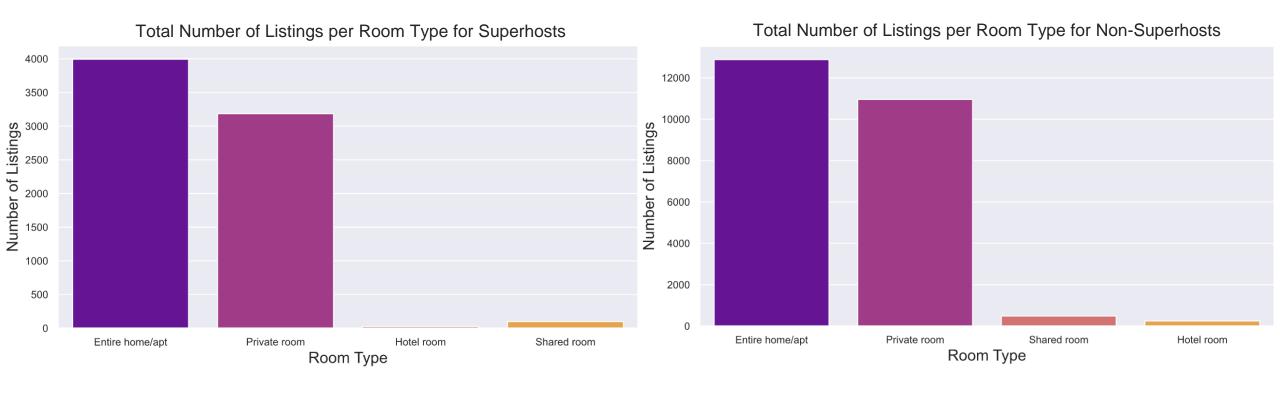
- There is an upward trend in the number of superhosts and non-superhosts.
- The number of non-superhosts is higher for each year.

Number of Hosts on Airbnb By The Last Day of Each Year

Exploratory Data Analysis for Hosts
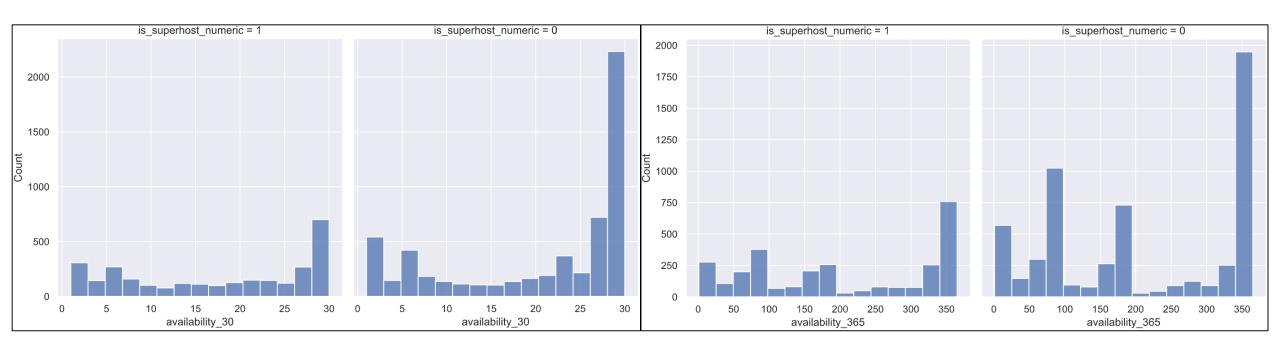
Comparison of Superhost and Non-Superhost by Room Type

- The most prolific room type is an entire home/apartment, and the least prolific appears to be the hotel room.
- A critical observation is that non-superhosts have 68% more listings that are entire homes/apartments than superhosts, and 75% more listings that are private rooms.

Exploratory Data Analysis for Hosts

Superhosts and Non-Superhosts Availability (Over 30 days and 365 days)

- Generally, non-superhosts have higher number of listings available over a period of 30 and 365 days.

Exploratory Data Analysis for Price

Average Daily Room Price for Superhosts and Non-Superhosts in 2019
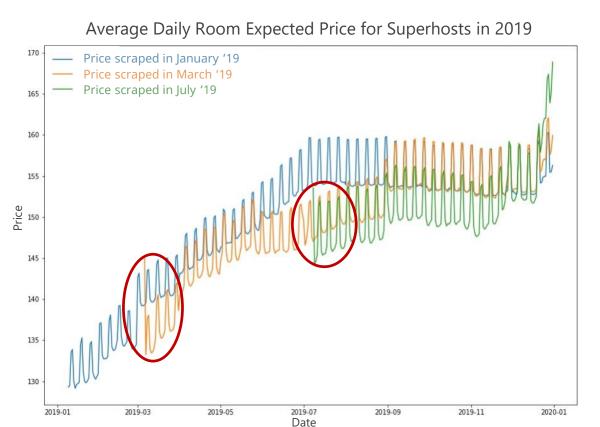
- During a year, the price rises between Jan. and Jul., and fluctuates between Jul. and Dec.
- For each week, the price on weekends is higher than that of weekdays.
- The average daily room price for super hosts tends to be higher than that of non-super hosts.
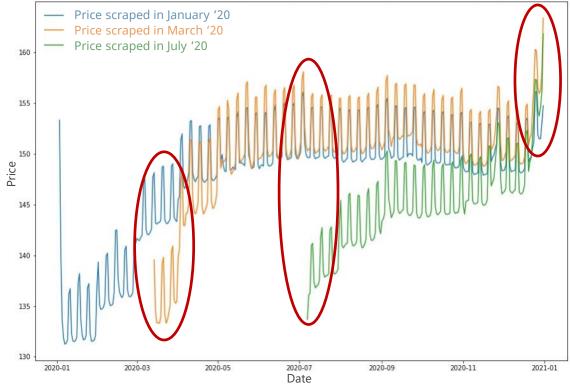
Exploratory Data Analysis for Price

Impact of COVID-19 on Average Daily Room Price for Superhosts

Average Daily Room Expected Price for Superhosts in 2019
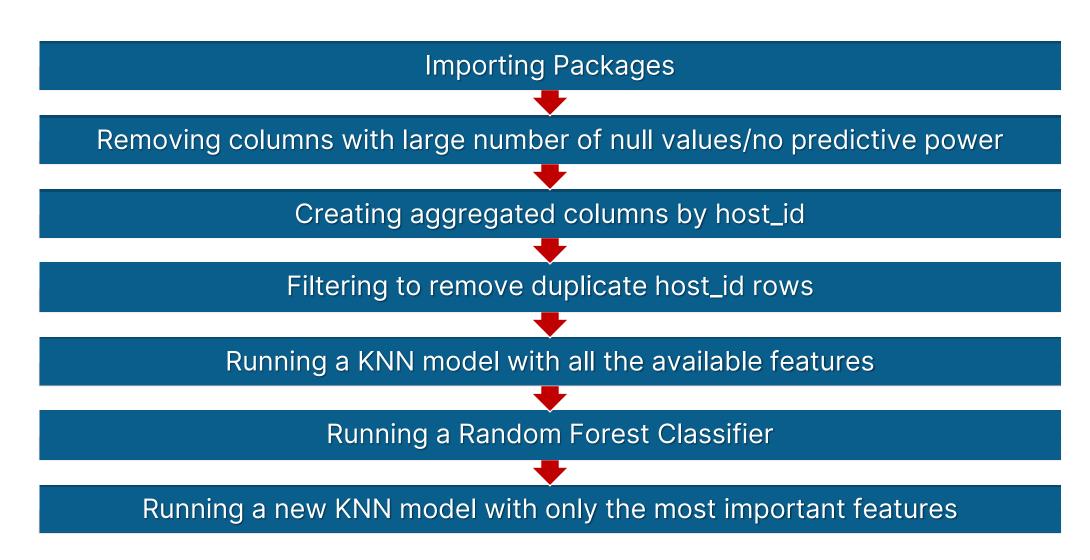
Average Daily Room Expected Price for Superhosts in 2020

Machine Learning – Classification Model
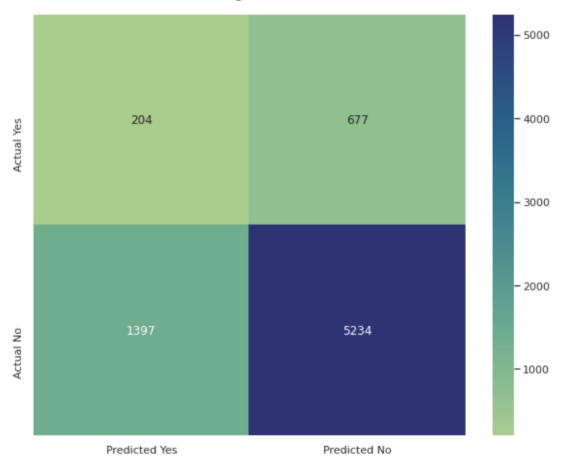
Steps implemented to build the model to predict if the host is a superhost/non-superhost

Importing Packages

Removing columns with large number of null values/no predictive power

Creating aggregated columns by host_id

Filtering to remove duplicate host_id rows

Running a KNN model with all the available features

Running a Random Forest Classifier

Running a new KNN model with only the most important features
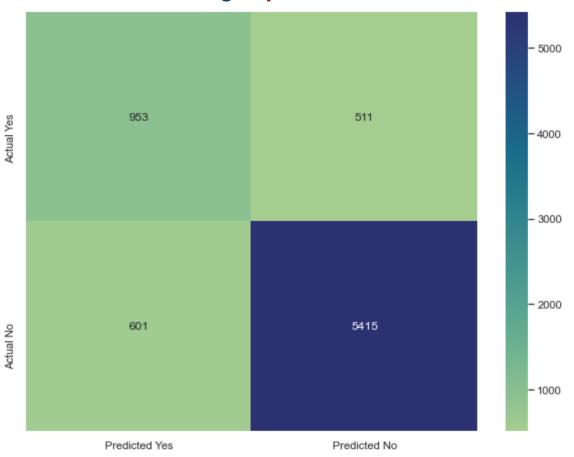
Machine Learning – Classification Model

Confusion Matrix for Prediction of Superhosts Using All Features

Confusion Matrix for Prediction of Superhosts Using Top Features

True Positive Rate: 12.74%
False Positive Rate: 21%
Accuracy Score: 72.39%
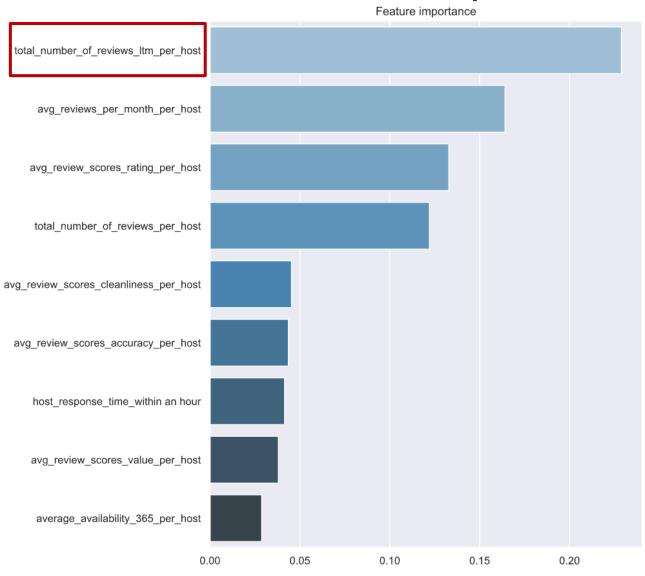
True Positive Rate: **62.96%**
False Positive Rate: **8.94%**
Accuracy Score: 85%
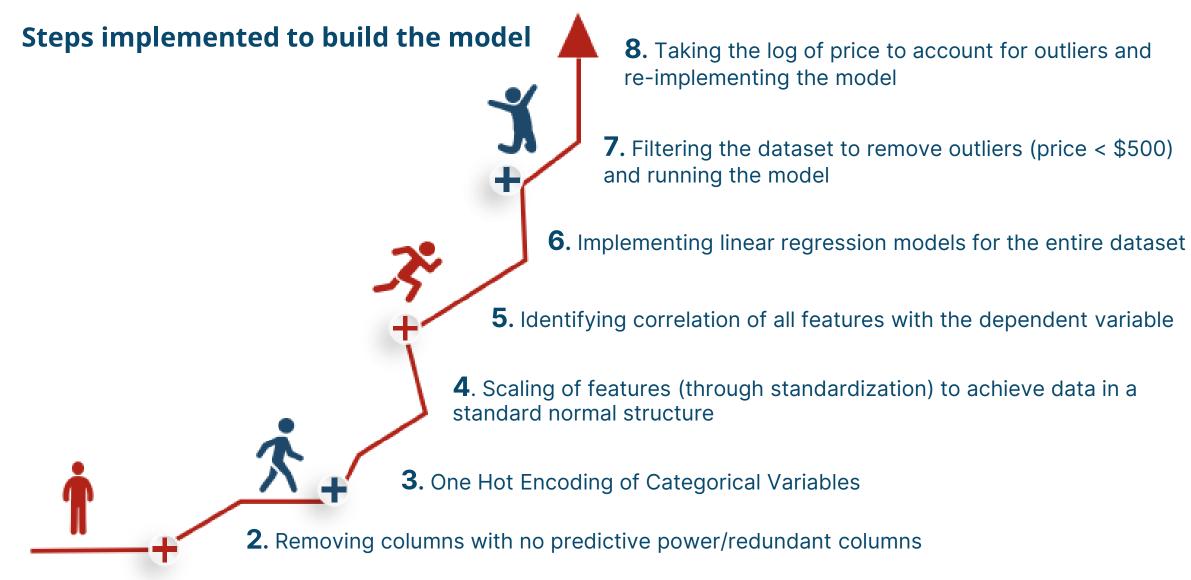
Machine Learning – Classification Model

Random Forest Classifier with Top Features

Machine Learning – Linear Regression Model

Steps implemented to build the model

8. Taking the log of price to account for outliers and re-implementing the model

7. Filtering the dataset to remove outliers (price < $500) and running the model

6. Implementing linear regression models for the entire dataset

5. Identifying correlation of all features with the dependent variable

4. Scaling of features (through standardization) to achieve data in a standard normal structure

3. One Hot Encoding of Categorical Variables

2. Removing columns with no predictive power/redundant columns

1. Import Packages

Machine Learning – Linear Regression Model

Linear Regression Model with Total Price

- MAE: 61.83
- MSE: 49282.50
- RMSE: 221.99
- R-Squared: 0.1228

Linear Regression Model with Price < $500
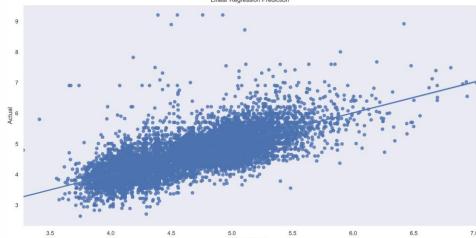
- MAE: 39.40
- MSE: 3265.60
- RMSE: 57.14
- R-Squared: 0.47

27

Machine Learning – Linear Regression Model

```
                        OLS Regression Results
========================================================================
Dep. Variable:              price   R-squared:                   0.534
Model:                        OLS   Adj. R-squared:              0.534
Method:             Least Squares   F-statistic:                 1828.
Date:            Sat, 28 Nov 2020   Prob (F-statistic):           0.00
Time:                    16:54:11   Log-Likelihood:             -20708.
No. Observations:           31880   AIC:                       4.146e+04
Df Residuals:               31859   BIC:                       4.163e+04
Df Model:                      20
Covariance Type:        nonrobust
========================================================================
                                     coef    std err         t      P>|t|      [0.025     0.975]
------------------------------------------------------------------------
const                              4.6056     0.017    277.020     0.000      4.573      4.638
accommodates                       0.1743     0.005     36.997     0.000      0.165      0.184
bedrooms                           0.1088     0.004     26.263     0.000      0.101      0.117
beds                              -0.0260     0.004     -5.831     0.000     -0.035     -0.017
availability_30                    0.1705     0.011     15.725     0.000      0.149      0.192
availability_60                   -0.1020     0.025     -4.142     0.000     -0.150     -0.054
availability_90                   -0.0015     0.019     -0.080     0.937     -0.038      0.035
availability_365                  -0.0070     0.004     -1.663     0.096     -0.015      0.001
number_of_reviews                  0.0035     0.004      0.855     0.392     -0.004      0.011
number_of_reviews_ltm             -0.0232     0.005     -4.502     0.000     -0.033     -0.013
number_of_reviews_130d            -0.0184     0.003     -5.442     0.000     -0.025     -0.012
reviews_per_month                 -0.0221     0.006     -3.547     0.000     -0.034     -0.010
is_superhost_numeric               0.0674     0.007     10.113     0.000      0.054      0.080
neighbourhood_group_cleansed_Brooklyn       0.1886     0.017     11.414     0.000      0.156      0.221
neighbourhood_group_cleansed_Manhattan      0.4666     0.017     28.194     0.000      0.434      0.499
neighbourhood_group_cleansed_Queens         0.0745     0.018      4.259     0.000      0.040      0.109
neighbourhood_group_cleansed_Staten Island -0.0662     0.034     -1.955     0.051     -0.133      0.000
room_type_Hotel room               0.4463     0.029     15.377     0.000      0.389      0.503
room_type_Private room            -0.5633     0.006    -90.531     0.000     -0.576     -0.551
room_type_Shared room             -0.7556     0.020    -37.598     0.000     -0.795     -0.716
instant_bookable_t                -0.0122     0.006     -2.159     0.031     -0.023     -0.001
========================================================================
Omnibus:                     9770.630   Durbin-Watson:                1.853
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         81763.289
Skew:                           1.238   Prob(JB):                      0.00
Kurtosis:                      10.445   Cond. No.                      31.1
========================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Number of reviews as well as features related to availability for 90 and 365 days and Staten Island are not statistically significant in predicting price (p-value > 0.05).

Conclusion

Higher competition for Airbnb hosts in Manhattan and the most expensive neighborhoods (Flatiron District and Tribeca) are in downtown Manhattan

Hosts in Manhattan might look at Airbnb listings as a business opportunity, viz-av-z hosts in Brooklyn, who would be looking to save rent/money by renting out private rooms

COVID-19 impacted the price and number of reviews of the listings which decreased

The number of reviews plays a significant role in determining the prediction of host types, but is not significant in predicting price of listings

Features related to accommodation, room type and neighborhoods in Manhattan, Queens and Brooklyn play an important role in determining future price of the listings

Recommendations for Prospective Hosts

- Regular updates of listing images and descriptions

- Adjustment of listing prices based on the availability of competing listings within the neighborhood

- Airbnb hosts can place more emphasis in finding guests who are opinion leaders and bloggers who can leave positive reviews and increase exposure

- Prospective hosts can penetrate markets/boroughs with popular neighborhood locations where listings are in higher demand

airbnb

Future Work

- Exploring other models to predict price

- Finding new alternatives to deal with outliers and missing values

- Gaining more insight on the lack of number of listings for shared room types in NYC

THANK YOU!

Any questions?