Predicting Properties of Airbnb Listings

Project by:
Grant Gittes (gbg2)
Shubhrika Sehgal (shs53)
Tejasvi Medi (tem73)

Introduction

- Airbnb is a very popular, new, and disruptive technology company by which individuals can rent living space as an alternative to a hotel
  - Accommodations range from an extra pull-out couch, to million-dollar mansions, listed by individual hosts
  - Lengths of stays range from a single night to months, depending on the listing
  - Typically Airbnb has been used for tourism purposes, but it has been recently attracting business travelers looking for superior accommodations at attractive rates, or a more "home like" alternative to a hotel
- The company does not own any of the real estate listings; it acts as a broker, receiving commissions from each booking

Dataset & Source

- We have sourced the Data from Inside Airbnb, an independent, non-commercial set of tools and data that allows exploration of how Airbnb is being used in cities around the world
- The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb site
  - It is updated about monthly; our dataset is from March 14, 2020
- We chose Los Angeles data as we wanted to analyze a diverse community, with many "hotspots". We wanted to see how attractions like Hollywood, and the beach nearby can affect listings, price, amenities, and other factors

Description

- The original dataset has 106 parameters and 38,481 observations
    - Includes listings information about the property and also host information
- The dataset contains many missing values, biased variables, outliers or entry errors, and has both categorical and continuous variables
- There are no duplicate records for any listing
- After the data cleaning, we were left with 37 parameters and ~30,000 observations
    - After text extraction of amenities, we added 12 additional dummy variables for included amenities

airbnb

Data Cleaning

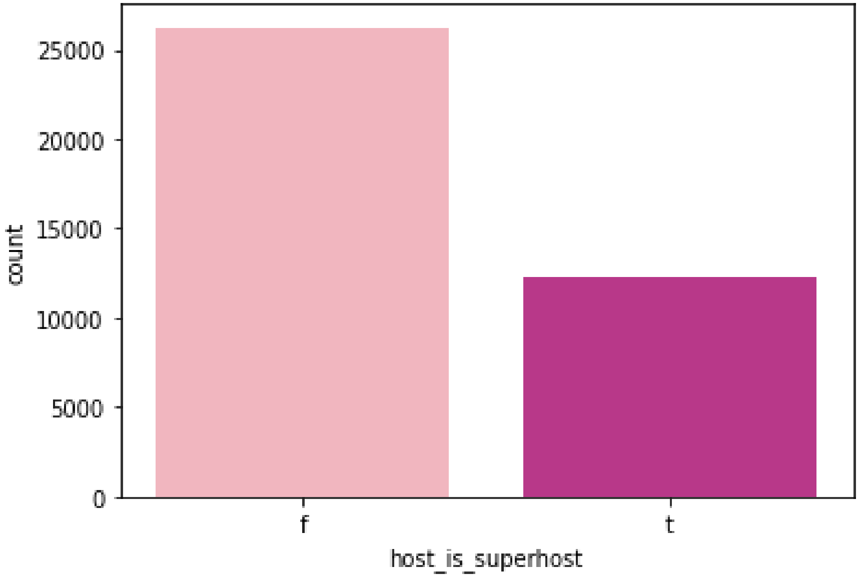
- Dropped columns with more than 25% null values
- The ‘State’ column had values like ‘NY’, ‘NV’, ‘Ny’. Since we are using dataset for “Los Angeles” listings, we dropped the rows with such values
- Some columns like Price, Cleaning fee, security deposit had a ‘$’ sign on every field. We removed that and changed the type to a float value
- Converted Boolean text features to binary 1 and 0 for analysis
- Removed identify variables like listing id, host id, url, etc.

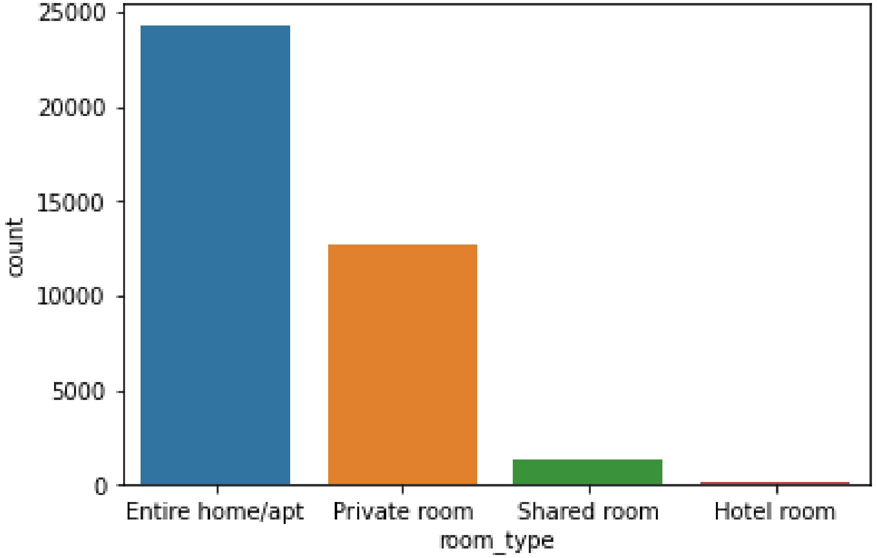
Data Cleaning

- Cleaned the data in the column 'zipcode' which had additional strings like CA etc
  - Those values were used to join with the household income by zipcode dataset
- Replaced the null with an appropriate value or forward fill for parameters such as 'host_acceptance_rate'
- We decided to ordinally map some features that had limited number of distinct values, into integer values with ascending "value"
  - Ex. Response time was mapped ('a few days or more':1,'within a day':2, 'within a few hours':3, 'within an hour':4), to allow the models to assign importance to more "desirable" features

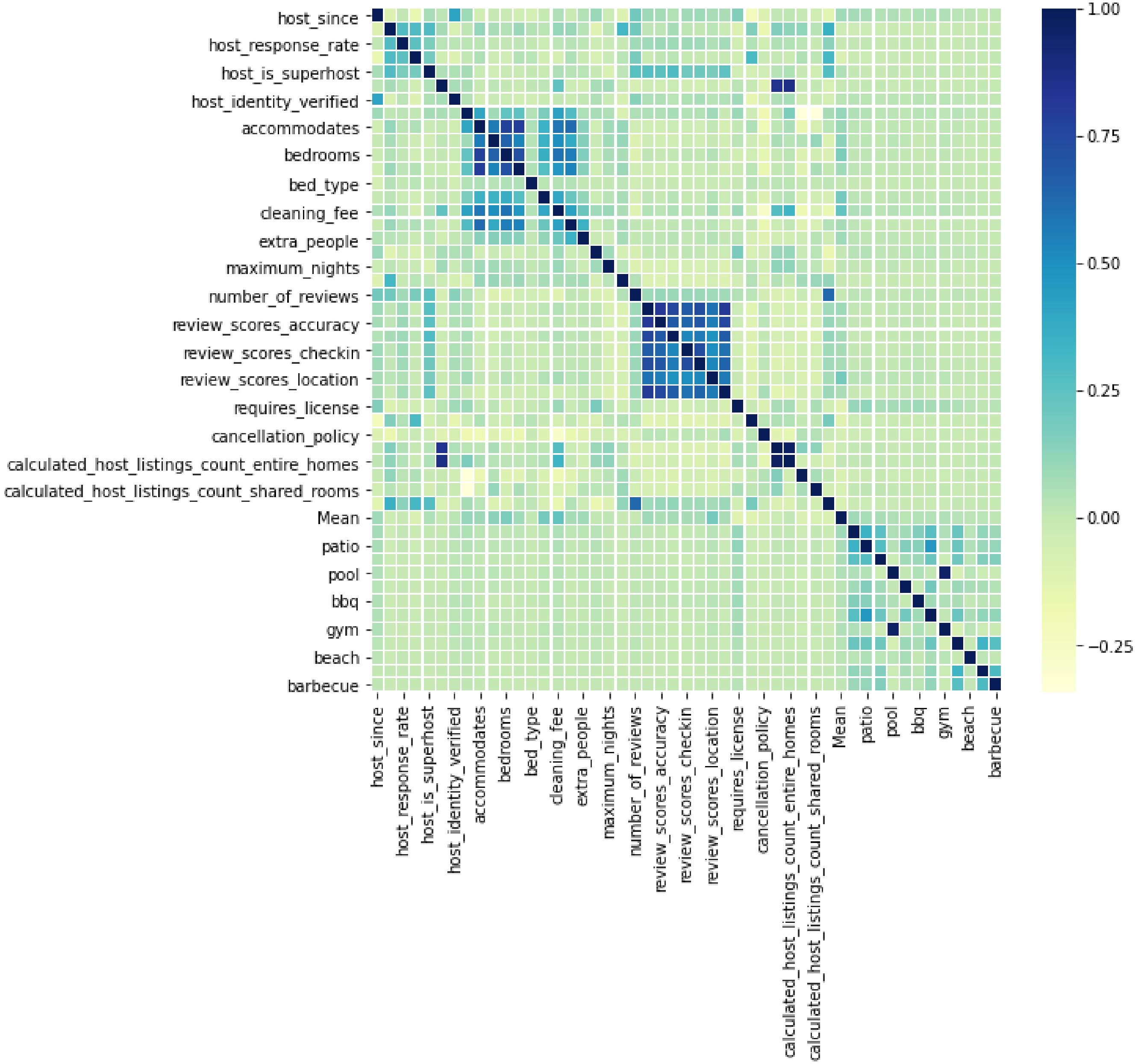
Summary Statistics

- There are only 4 room types, with more than 60% of them being an entire home/apartment
- Majority of hosts in Los Angeles area have response time of within an hour, with only about 2% with response time of more than 1 day
- 75% of the listings are priced under $189/night with cleaning fees not exceeding $125. Mean price of a listing in Los Angeles area was ~$229/night
- Majority of listings are for 1 or 2 guests only

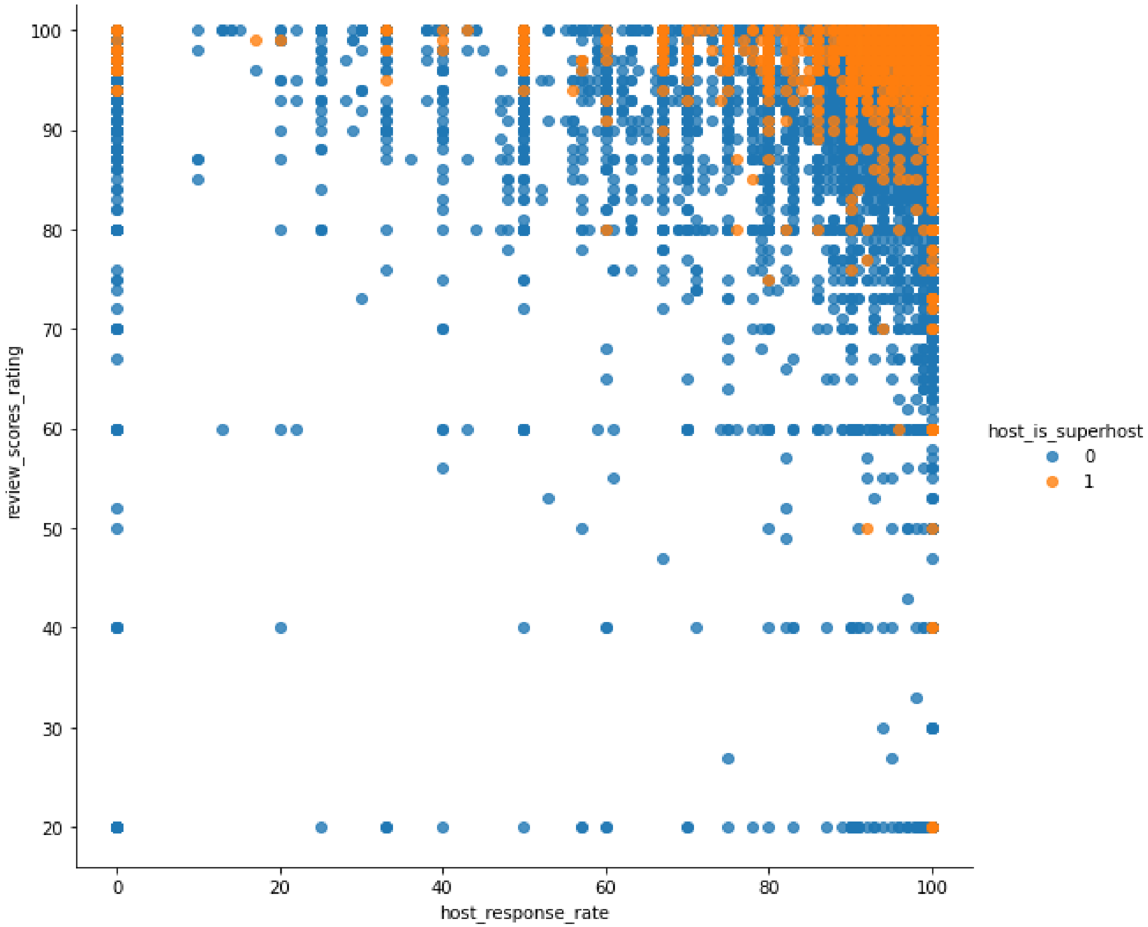




Summary Statistics - Correlation Heatmap



Cluster Analysis



The cluster analysis between the superhost, review_scores_rating and host_response_rating shows that as superhost a higher rating than the average host and a faster response rate

Problem Statement

- What price should a new market entrant price a listing based on its attributes?

- What distinguishes hosts that have superhost status? Do superhosts command a higher price?

- What will be rating of the listing based on various listing features or host attributes?

- What are the most important attributes for a listing?

Hypothesis and Expected Results

- Null Hypothesis
  - There is no relationship between price, superhost and rating with the features like response time, number of beds, duration of listing etc.
- Research Hypothesis
  - The price, superhost and rating is dependent on directionally, proportionally or inversely combination of features
- Expected Results are to predict price for the listing

Why This Problem is Interesting

- This problem is interesting because hosts on the Airbnb platform are independently setting prices in reaction to the market, and have many more competitors than a standard hotel, while able to offer different amenities and experiences over a hotel
    - It is interesting to analyze how efficient this marketplace is and what consumers seem to value most
- The solution can help new as well as existing hosts to adjust to the market trend and increase their profits
- As the popularity of Airbnb is increasing due to the comfort of staying in an entire house for families together instead of separate rooms
    - Also, with the advent of ridesharing, consumers are more likely to opt for a private/shared room in someone's home for favorable rates compared to a hotel room

airbnb

Techniques Used

As the dataset is diverse, we wanted to work on 3 different problem statements and use different techniques

- Classification for whether the host is superhost or not
- Regression to determine the price of a listing
- Classification to predict the ratings of the host (whether more than 90 or not, based on non-rating factors)

Techniques Used - Feature Selection

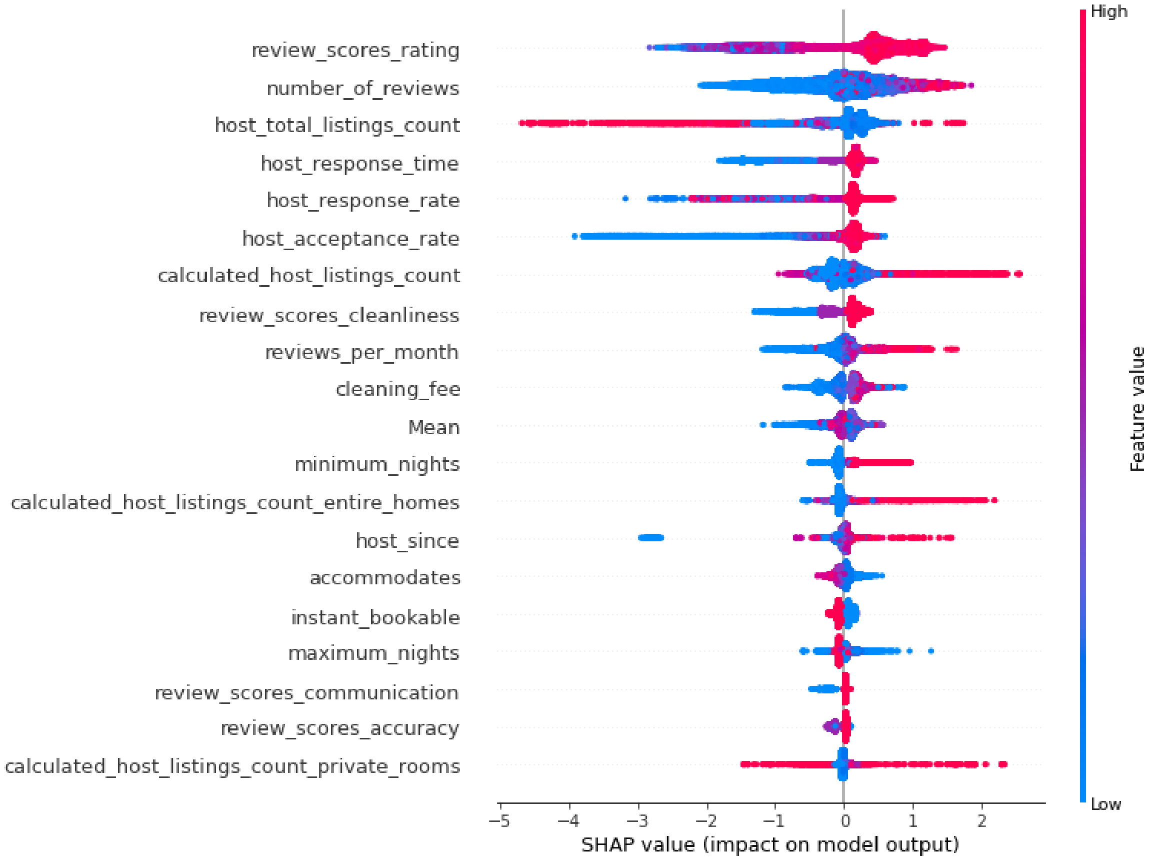
- As there are many factors which impact the price, listing rating and the title of superhost
  - The numerical data derived from the dataset was a good start, but we felt it needed to be augmented a bit
- Mean income by zip code was imported externally as a proxy for how "nice" a neighborhood was (as it would command a higher price)
- We worked with the textual amenities columnize to tokenize the words in each row and assessed which were most important "amenities" that would drive up the value of a listing. Certain amenities like a TV were excluded, so we chose to identify tokens that would drive more interest like balcony, patio, pets, pool, gym, etc.
- The dataset was split into train and test sets (with a val set for the boosted trees) based on the specific problem and target (ex. for price regression of the listing problem, the target was price)

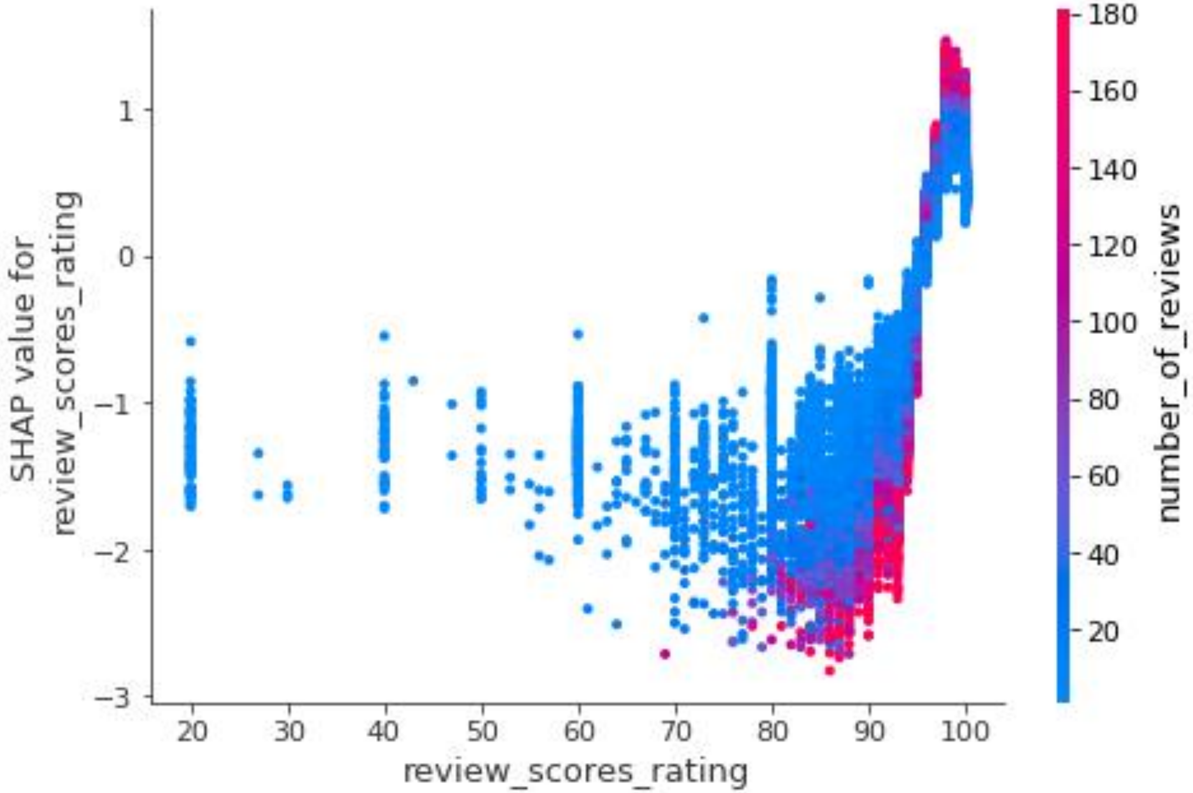
Techniques Used – Models

1. Classifying superhost (yes/no)
   - Logistic Regression
   - XGBoost
   - XGBoost with Hyperopt
   - LightGBM
   - LightGBM with Hyperopt and Hyperband
2. Regression for Price
   - Linear Regression
   - Linear Regression (with ElasticNet)
   - XGBoost
3. Binomial Classification for Rating
   - XGBoost

Feature Importance & Evaluation

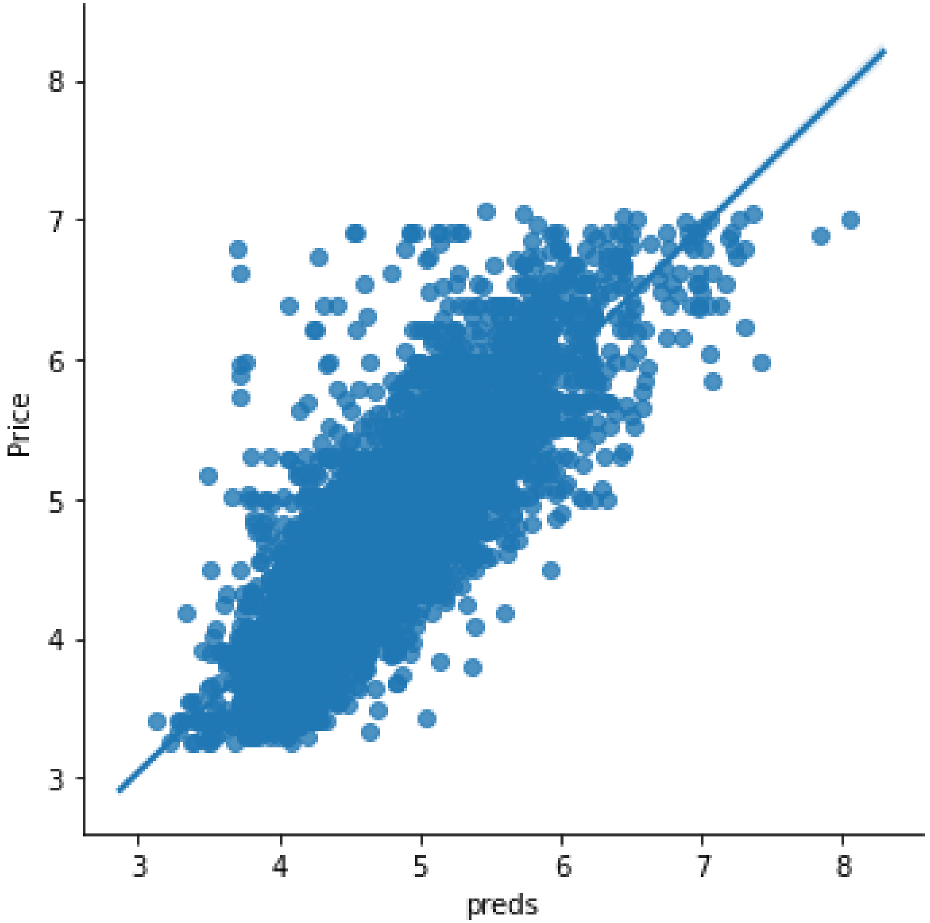
- For the classification problem, we evaluated the model for AUC and then shap plots to determine the important features.

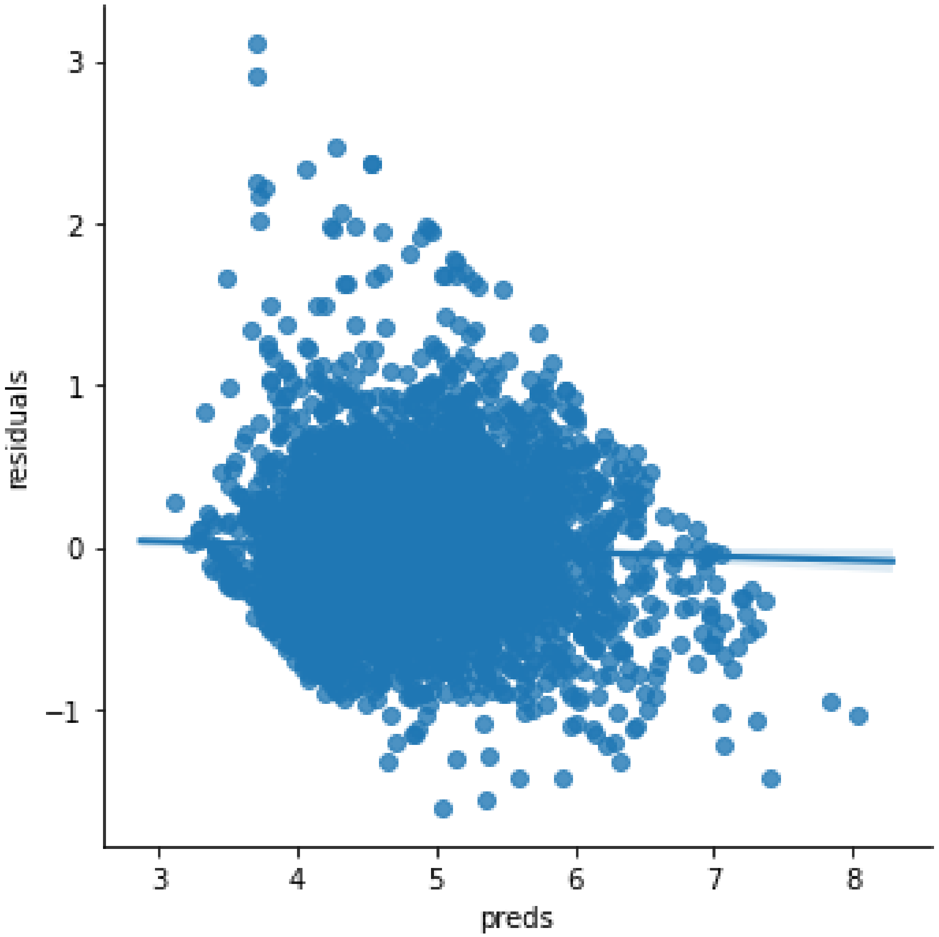




Evaluation

- For the regression, we used Mean Absolute Error and R2 score to evaluate our model and used scatter and residue plot to determine the result

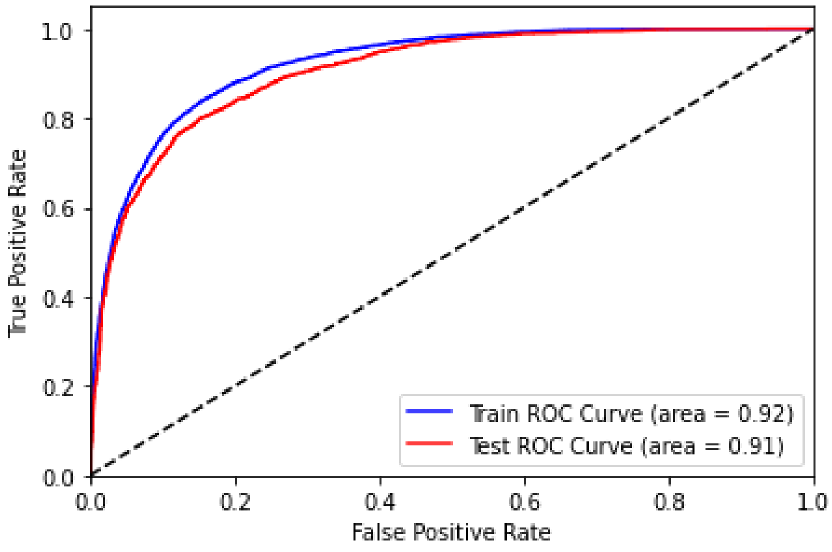


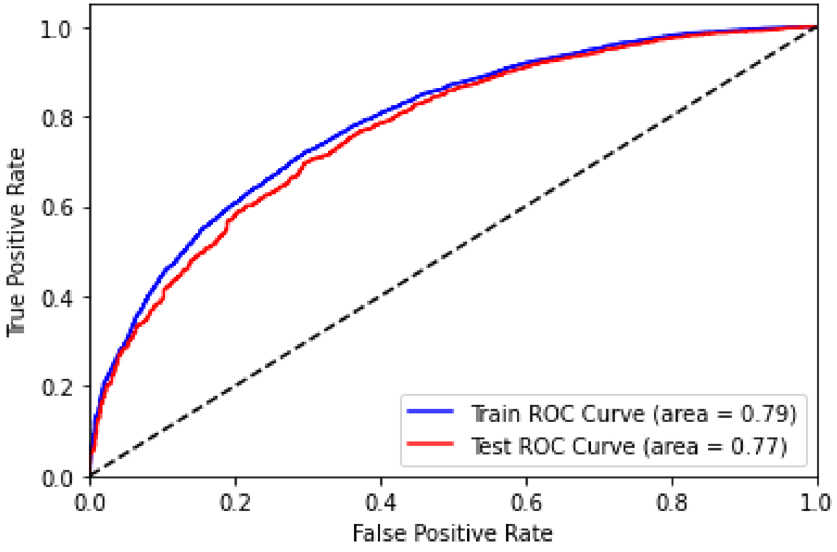


Findings & Conclusions

Best model for

1. **Classifying Super host**
   Light gbm with Hyperopt
   AUC Train :0.92
   AUC test 0.91
   Overfit/Underfit N/A
2. **Classifying Rating**
   Xgboost
   AUC Train :0.79
   AUC test 0.77
   Overfit/Underfit N/A
3. **Regression Price**
   Xgboost
   r2 score train :0.68
   r2 score test:0.67
   Overfit/Underfit N/A





Potential Deployment Strategies

- Classification of superhost
  - Performs well with the ROC curve. This classification can help hosts know what to focus on to achieve superhost status
- Pricing
  - The model predicts pricing reasonably well for the parameters analyzed, but it is clear that new and existing hosts must analyze additional factors not found in the dataset such as exact location (relative to desired locations), pictures, etc. that are not found in this dataset
- Rating Classification
  - Though the model is neither overfit nor underfit, most of the ratings are higher than 90, we would need to make it a multinomial classification, or to better understand if nuances in rating between 90-100 make a large difference to make it usable for the real world

Confidence

- The models for classifying super_host have good AUC (best one with auc ~0.91 and accuracy of ~81%). No other model had significantly lower AUC (except for logistic regression)
- The regression for price gave a similar R2 score for all models at a level of ~.67. Thus we are confident there are other factors besides type of model that must go into the pricing of different units
- For predicting classification as 90+, we are reasonably confident with an AUC of ~.77, but this analysis would probably have to be refined before implementation since so many hosts are 90+, and the model would need to determine which factors separate 90 from 100

Thank You