Using Machine Learning to Predict Airbnb Rental Price in New York City

Andy Vu

University of Trento

November 6, 2020

Outline

A little bit about myself

- I'm a second year master student in Economics
- My background is in finance and banking. I spent 2 years working for a commercial bank in Vietnam.
- Research interest:My interest in applying statistical learning in forecasting started when I took a course on econometrics. I worked on a project on forecasting the electricity price using time series data. Since then, I have focused more on data mining, machine learning, and data science. My favorite book is "An Introduction to Statistical Learning" by Gareth James, Trevor Hastie. Besides reading, I took some courses on data analysis, applied machine learning, and programming.
- Technical Skills
  - Languages: R, Python, Java, SQL
  - Version Control: Git, Github
  - Markup Languages: Markdown, LaTeX

The Problem

Two Big Questions:

1. Which model has the best performance in predicting the Airbnb Rental Price in NYC?
2. Which features are essential in predicting the rental price?

The Solution- Machine Learning Approach

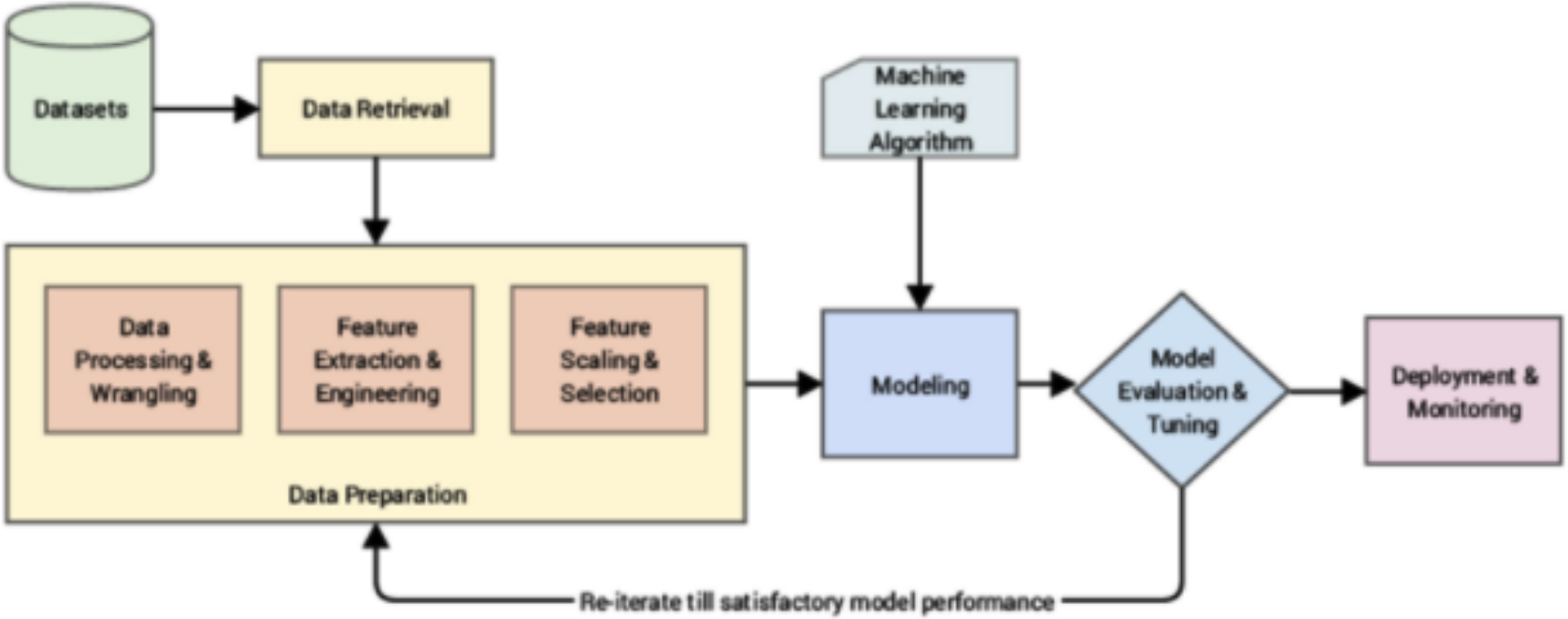
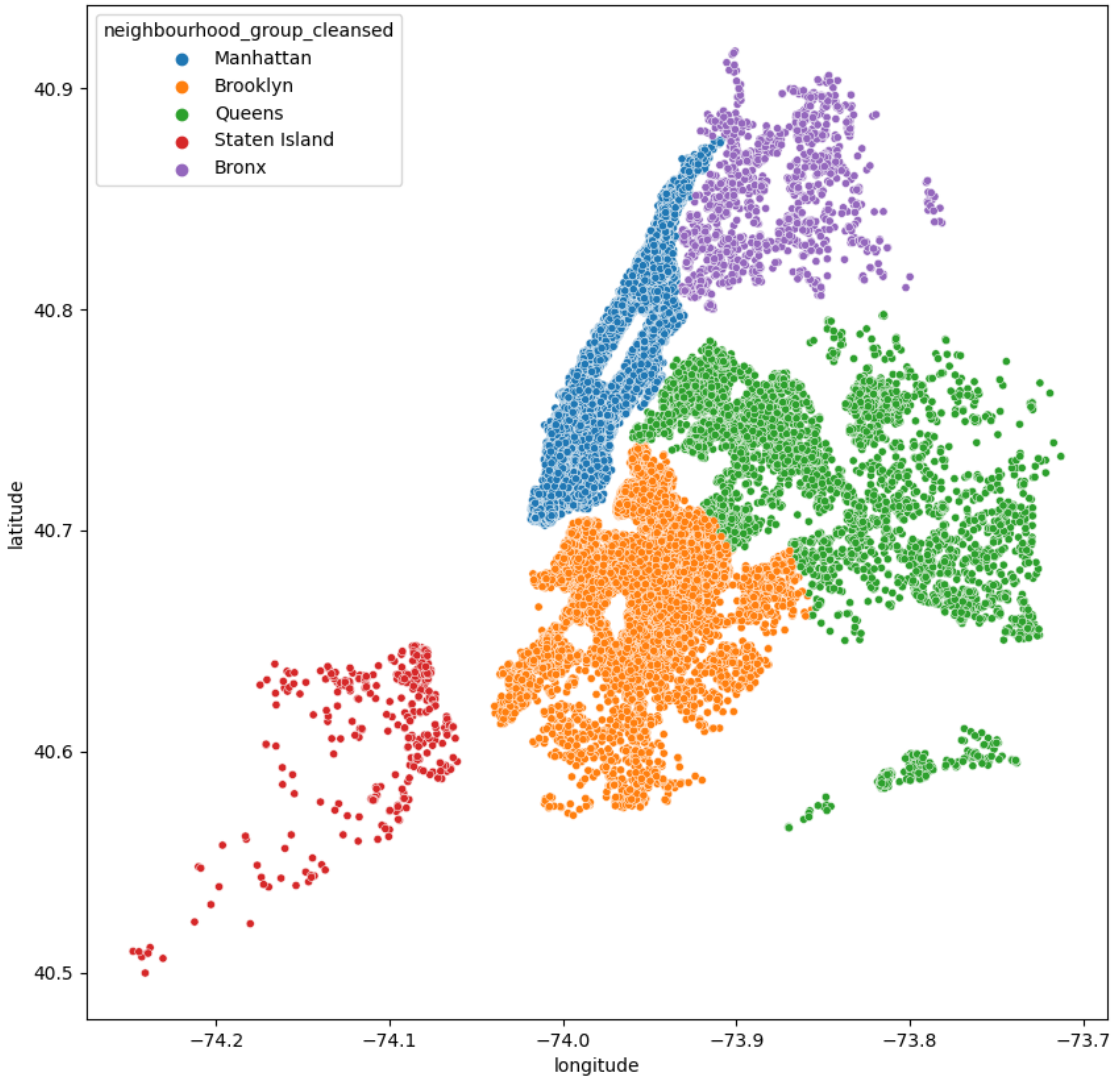


Figure: Machine Learning Pipeline

Airbnb New York City



The Dataset

- Source: The data comes from InsideAirbnb.com, a group that scraped the from the Airbnb website.
- Data Preprocessing: The data is untidy and not ready for model fitting, so we perform data filtering, feature selection, missing value handling, feature binning, data transformation.
- Features: The final dataset after the preprocessing step has 39907 observations with 279 features.

A Glimpse of some features:

Table: The variable list

Variables	Definition
experience_offerd	recommended category of travel type, e.g. business
host_since	date that the host first joined Airbnb
host_response_time	average amount of time the host takes to reply to messages
host_response_rate	proportion of messages that the host replies to
host_is_superhost	whether or not the host is a superhost, which is a mark of mark of quality for the top-rated and most experienced hosts, and can increase your search ranking on Airbnb

Data Visualization

Listing capacity

A listing's price seems to be positive associated with the number of guests it can accommodate.

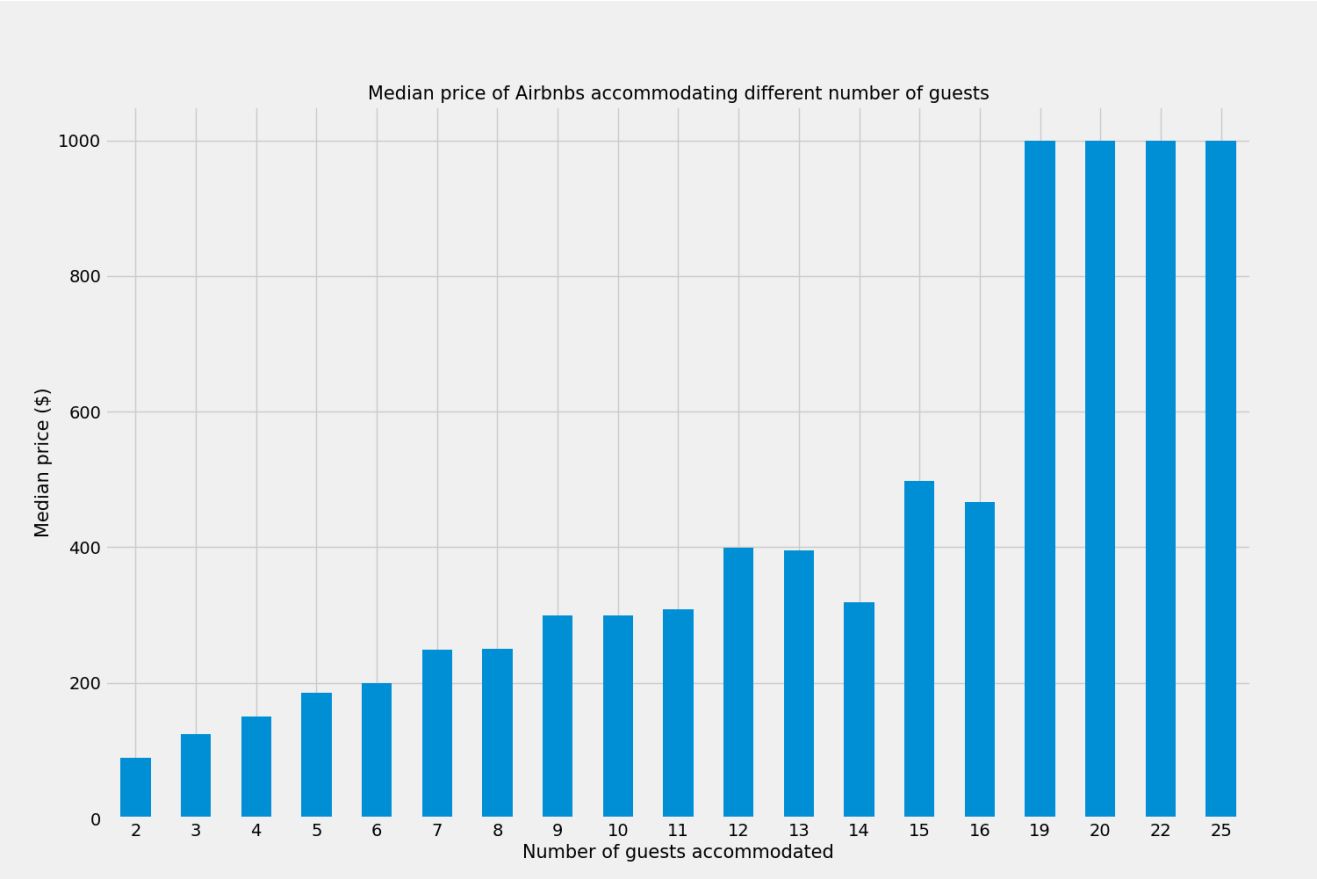


Figure: Median Price according to accomodates

Data Visualization

Location Features

The location might have an important role in determining the price. Unsurprisingly, listing in the city center (Manhattan and Brooklyn) has a higher median rental price.

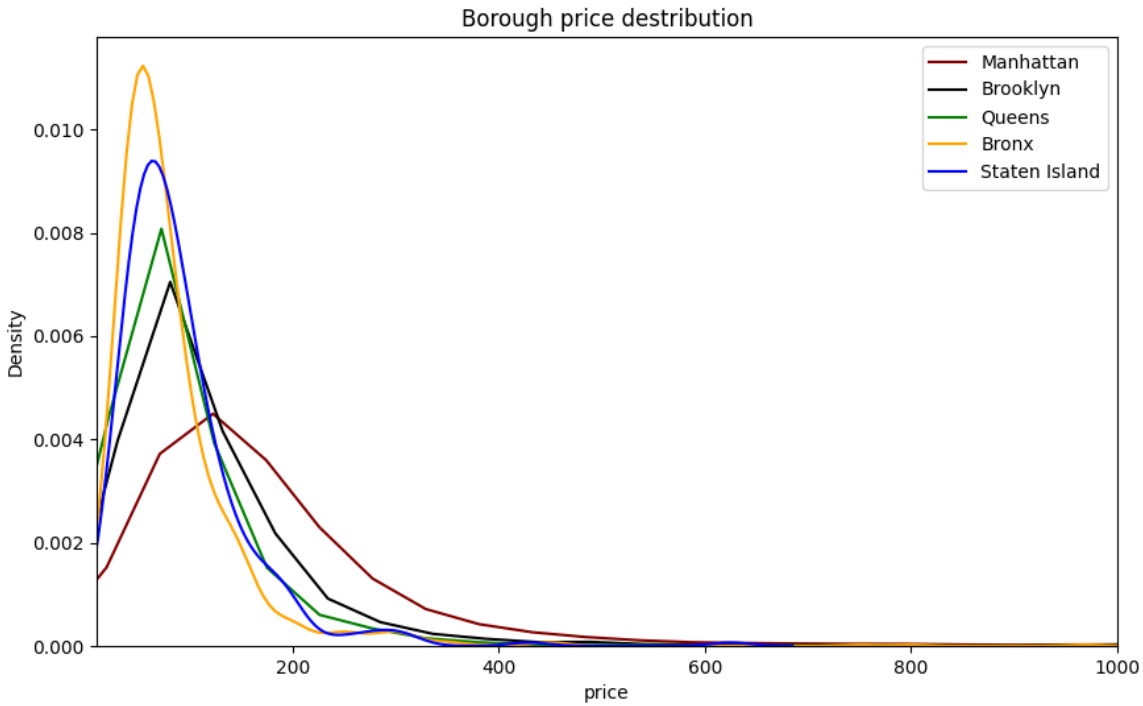
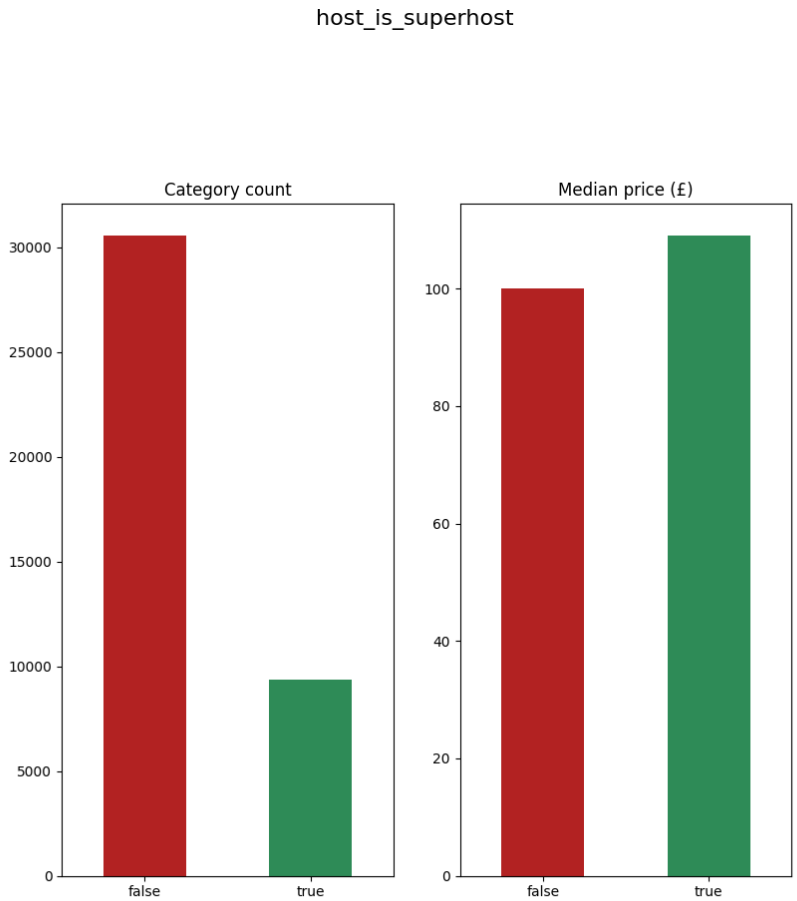


Figure: Borough Price Distribution

Data Visualization

Superhost status

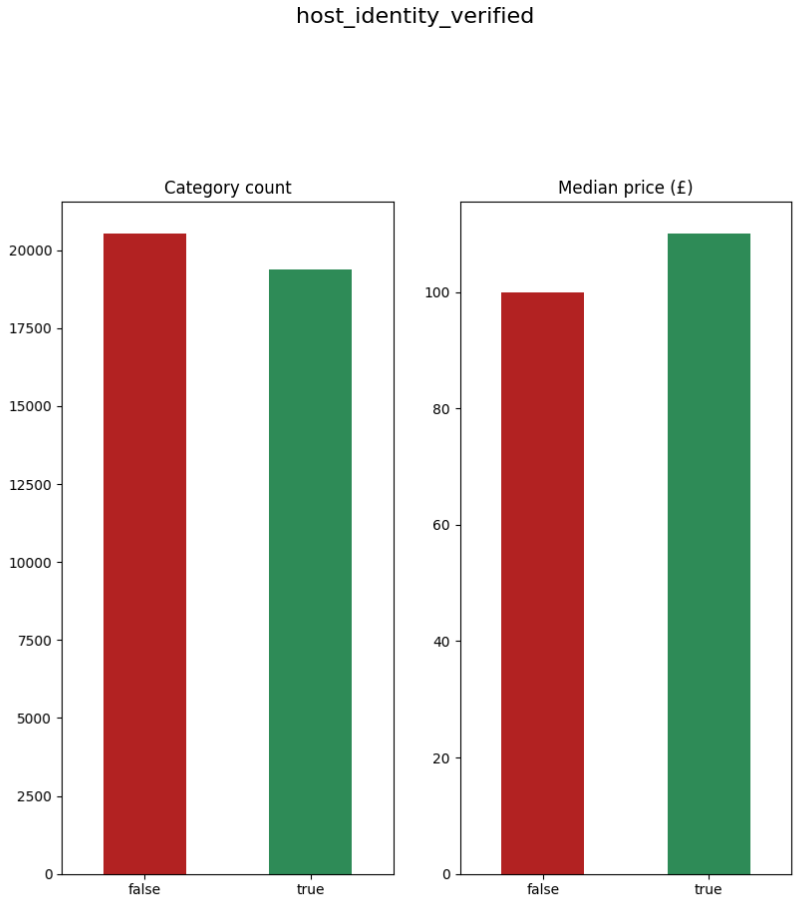
Hosts with superhost status usually charge higher prices. A possible explanation for this might be that people are willing to pay a premium price because they consider superhost status a mark of quality.



Data Visualization

Host Verification

Hosts with verified profiles gain a price premium. The relationship may be explained by the fact that verified profiles (e.g., by providing ID and verifying your phone number and email address) can increase their trustworthiness and, therefore, can charge a higher rental price.



How to Assess One Model Performance?

We can assess how well a model performs by using mean square error criteria

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{1}$$

The best model is the one that has the lowest *test* MSE (not *training* MSE)

Variance-Bias Tradeoff

The expected test MSE, for a given value x_0, can be broken down into three parts as followed:

$$E(y_0 - \hat{f}(x_0))^2 = \mathrm{Var}(\epsilon) + [\mathrm{Bias}(\hat{f}(x_0)]^2 + \mathrm{Var}(\hat{f}(x_0)) \tag{2}$$

Comments:

- Minimizing the test MSE means reducing the combination of bias and variance.
- However, it is impossible to do both simultaneously. Too simple model (low variance) tends to have high bias and too complex model (low bias) is likely to vary significantly.

Modelling Strategy

We employ some commonly-used algorithms to find the one with the lowest `Test MSE` :

1. Linear Regression with the full number of features - This model is expected to be overfitting because it tries to fit a huge number of features.
2. Ridge Regression: An improvement to Linear Regression as it tries to overcome overfitting by constraining the coefficient parameters.
3. Lasso Regression: Similar to Ridge Regression but have an advantage of being able to perform *feature selection*.
4. XGboost: decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.

Which model performs the best?

ML Algorithm	Training MSE	Test MSE	Training R^2	Test R^2
Linear Regresion	0.1291	8.5E21	0.7019	-1.9E22
Ridge Regression	0.1291	0.138	0.7019	0.6857
Lasso Regression	0.1351	0.1441	0.688	0.6718
XGboost	0.0798	0.1173	0.8157	0.7328

Comments:

- As expected, Linear Regression has the worst performance. The `Test MSE` and R^2 is very large and much higher than other models.
- Ridge and Lasso do much better both in terms of Test MSE and R^2. While Lasso performance is not as good as Ridge, it has the advantage of producing of *sparser* model with just 153 features compared to that of 278 features of Ridge.
- XGboost has the best performance among all models.

Which features are essential to predict the rental price?

Table: XGBoost Top 10 Feature Weights

	weight
room_type_Entire home/apt	0.336396
bathrooms	0.032001
neighbourhood_Midtown	0.025008
neighbourhood_Hell's Kitchen	0.018545
neighbourhood_East Village	0.015763
property_type_Other	0.015168
neighbourhood_Bedford-Stuyvesant	0.014314
neighbourhood_West Village	0.014031
neighbourhood_Chelsea	0.013612
neighbourhood_Lower East Side	0.011874

Which features are importance in predict the rental price?

Comments:

- The most important positive features are whether the type of listing is the entire home.
- Features related to the location are in the top 10. Being in Hell's Kitchen, Midtown, East Village, Chelsea, West Village, Upper West Side, Williamsburg, Upper East Side, SoHo, Lower East Side neighborhood is associated with an increase in the listing price.

Future Works

- Experiment the data with deep learning algorithms.
- Find a way to incorporate customer reviews feature through sentiment

Software Used

Python : a general-purpose programming language used in many data science applications

Pandas : a Python library for data wrangling and analysis

Scikit-learn : the most promient Python library for machine learning.

Matplotlib and Seaborn : the primary plotting libraries in Python

Other libraries and packages : geopandas (for plotting geographical data), xgboost (for XGboost algorithms)