Airbnb in NYC: Decipher the Truth About Pricing

Zhengfang Bao, Xiaoyang Xu,Yi (Vivian) Yu, Dekun Zhang, Yunlei Zhou

Introduction

Airbnb have entered a new market, which is more flexible, convenient, and diverse. Unlike traditional hotel business, house hosts in these platforms do not have efficient tools supporting their pricing strategy.

We are trying to utilize the programming and machine learning technique to build a pricing tool that can help hosts to estimate rental price with the attribute of the house and reviews.

Problem Statement

This project has two main goals:

- ❑ build a predictive price model
- ❑ integrate reviews information into our price model.

Dataset

- ❑ The dataset we are using comes from the community *Inside Airbnb*.
- ❑ We are using two different datasets: complete listings and complete reviews.

Methodology

I. Pricing Model

Fit initial price model. Steps:

- ❑ Exploratory Data Analysis
- ❑ Data cleaning, feature selection
- ❑ Fit a variety of price models using LASSO, XGBOOST & LIGHTGBM.

II. Review Text Analysis

We aim at counting for the details in reviews that are not shown in ratings.

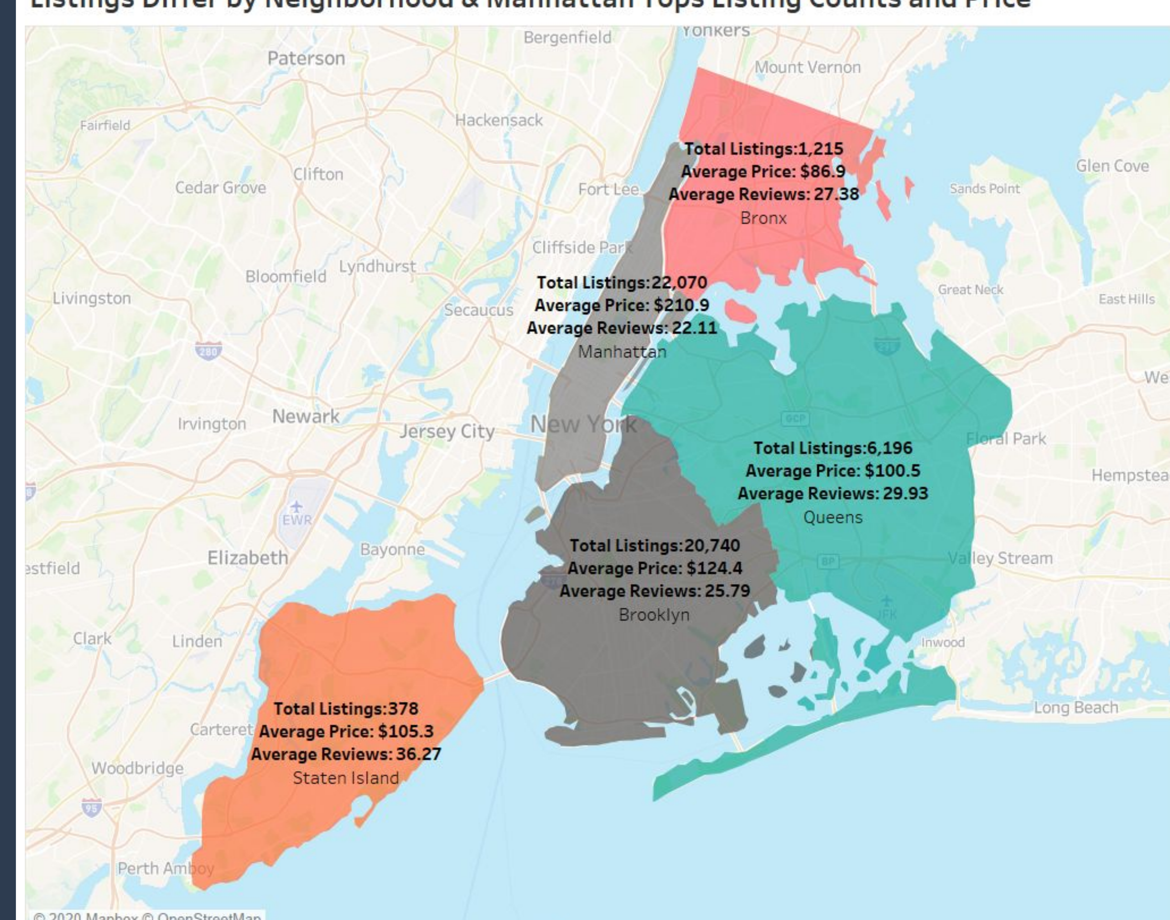
Steps:

- ❑ Applying the dictionary in SQL as the first attempt to filter for bad reviews
- ❑ Training DistilBERT+logistic regression model to sort the bad reviews from the filtered results of SQL.

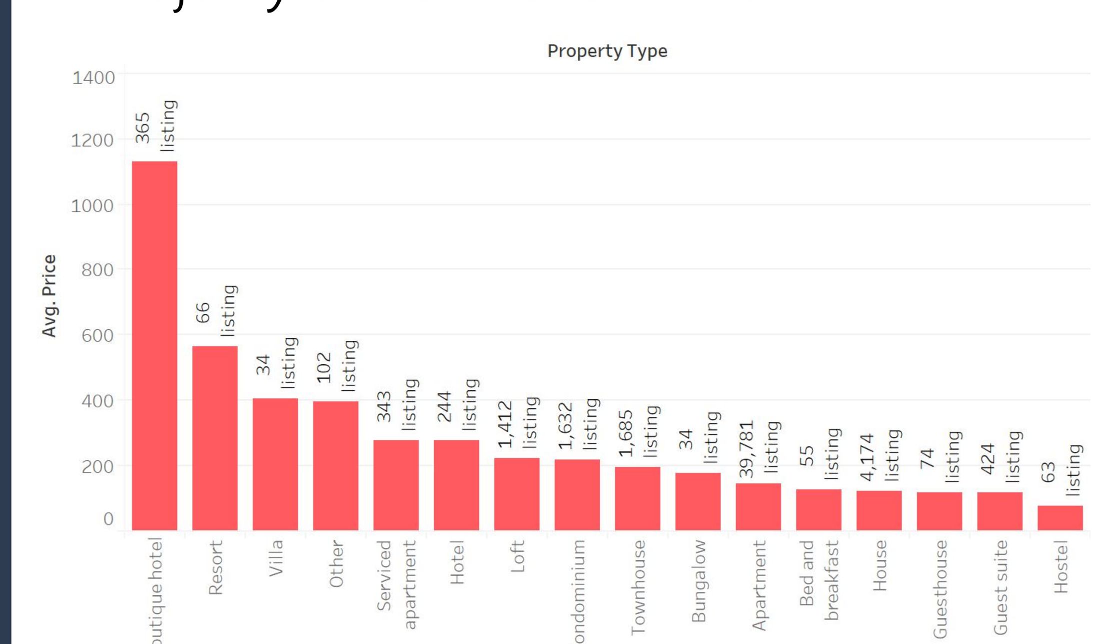
Exploratory Data Analysis

- ❑ Manhattan has most listings and highest average price

Listings Differ by Neighborhood & Manhattan Tops Listing Counts and Price

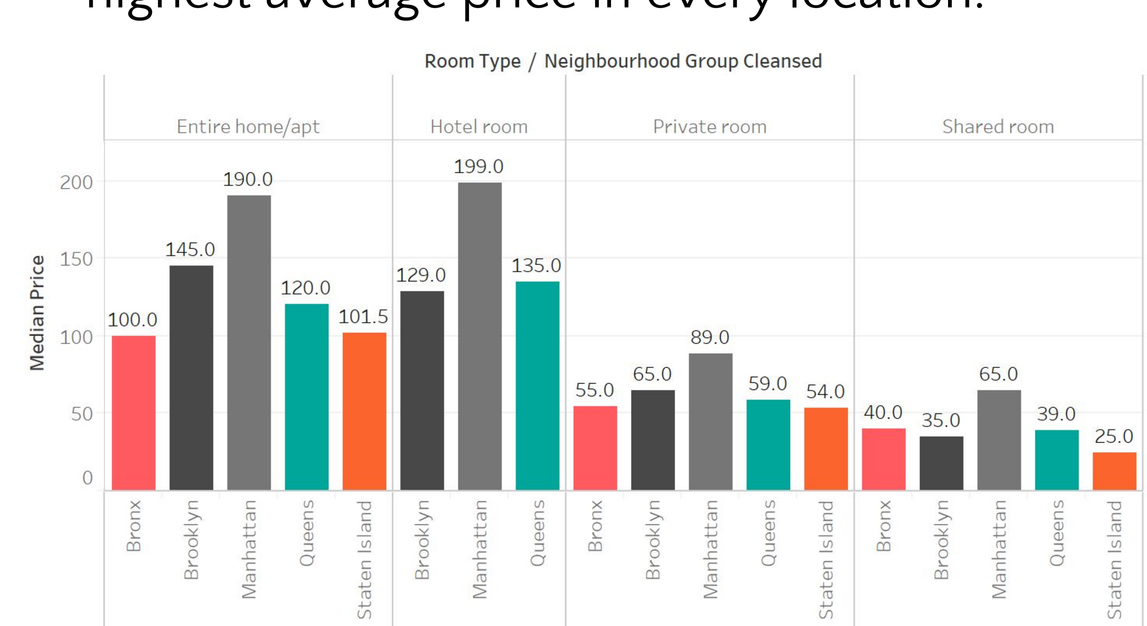


- ❑ Mid-price property types account for the majority of the Airbnb market.



Average of Price for each Property Type. The marks are labeled by sum of Number of Records. The view is filtered on Property Type, which keeps 16 of 35 members.

- ❑ House in Manhattan are generally most expensive, and entire apartment has the highest average price in every location.



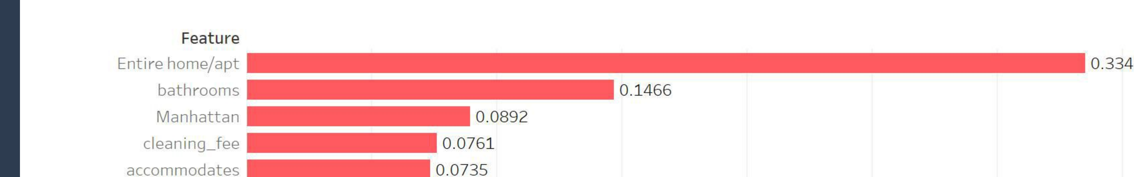
Median of Price for each Neighbourhood Group Cleansed broken down by Room Type. Color shows details about Neighbourhood Group Cleansed. The marks are labeled by median of Price.

Models & Influencoial Factors (w/o Bert)

LASSO Regression:

factor	coef	factor	coef
property_type_Resort	233.65	property_type_Condominium	6.16
property_type_Boutique hotel	50.84	reviews_per_month	0.86
neighbourhood_group_cleansed_Manhattan	43.47	review_scores_checkin	0.38
room_type_Entire home/apt	41.17	cleaning_fee	0.31
bathrooms	23.37	availability_30	0.03
Intercept	15.30	security_deposit	0.00
bedrooms	13.76	number_of_reviews	-0.01
accommodates	12.42	review_scores_value	-0.62
		room_type_Private	-7.36

XGBoost:



LightGBM:

DistilBERT Model & Review Analysis

We deployed DistilBERT model to help identifying negative reviews with hand-labelled review data.

Positive Review

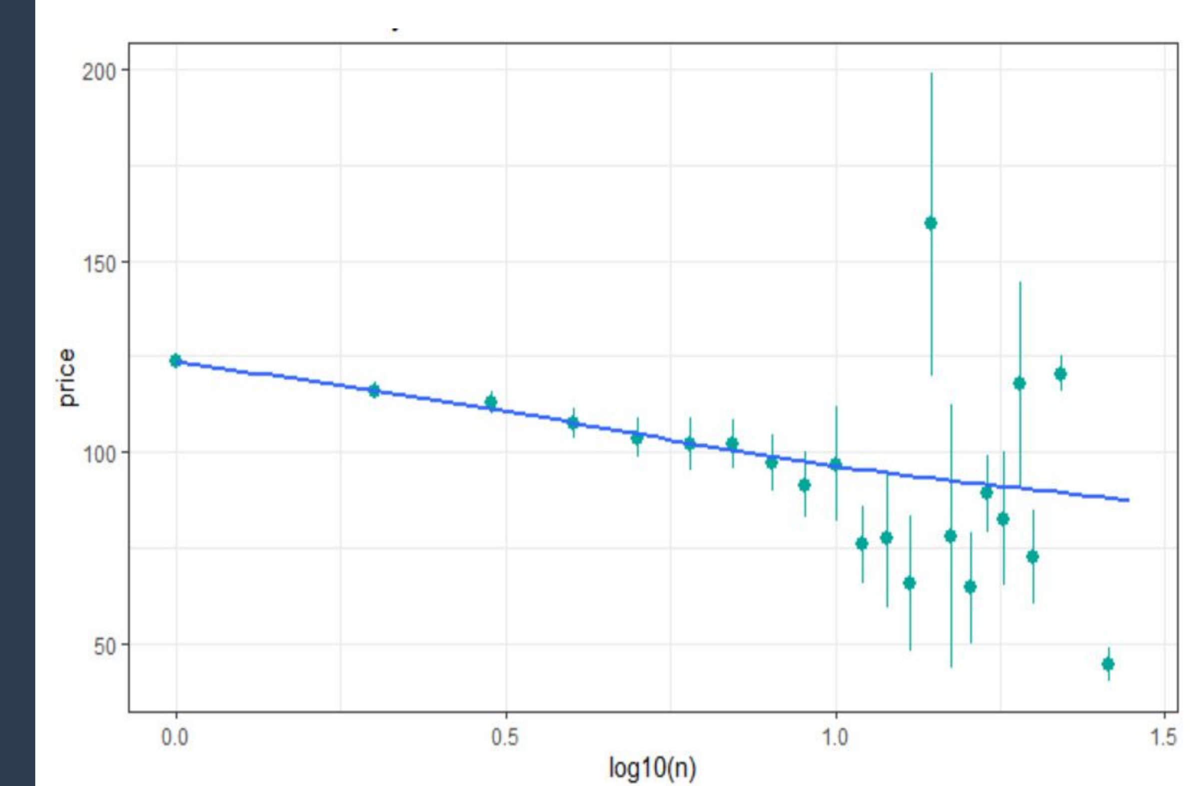
Negative Review

Model Results

Model	Accuracy(RMSE) Training	Accuracy(RMSE) Testing
Linear Regression	69.47	69.75
LASSO	69.38	85.37
XGBoost	57.55	83.19
LightGBM	61.95	60.72

Conclusion

House type, room type, location, and services affects price, and negative reviews also influence price negatively.



Suggestions for Host:

- ❑ Provide better amenities
- ❑ Update calendar frequently (for better availability)
- ❑ Respond to message promptly (for smooth booking)
- ❑ Verify your account (for smooth booking)
- ❑ Encourage your guests to write reviews – you may need to impress!
- ❑ Review scores matter, but read reviews carefully for improvements

Acknowledgements

We want to appreciate Prof. Andrey Fradkin, Prof. Brock Tibert, and Prof. Mohammad Soltanieh-ha for the kindly help and patience during the project.

Reference:

Zervas, Georgios, et al. "The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry." *Journal of Marketing Research*, vol. 54, no. 5, 2017, pp. 687–705.