Determining Optimal NYC Airbnb Listing Prices with Multi-Modal AI

Machine Learning Under a Modern Optimization Lens

Maya Nesen & Nidhish Nerur

Problem Statement & Significance

Help hosts optimize the price and features of a listing **in the *highly competitive* AirBnb market of New York City, in hopes of achieving high review score ratings and, ultimately, *greater customer satisfaction and higher revenue***

AirBnb Data & Preprocessing

	log_price	property_type	accommodates	review_scores_rating	bedrooms	amenities
0	5.010635	Apartment	3	100.0	1.0	{"Wireless Internet","Air conditioning",Kitche...
1	5.129899	Apartment	7	93.0	3.0	{"Wireless Internet","Air conditioning",Kitche...
8	4.605170	Apartment	2	93.0	1.0	{Internet,"Wireless Internet","Air conditionin...
13	3.688879	House	2	89.0	1.0	{Internet,"Air conditioning",Kitchen,"Smoking ...
16	5.003946	Apartment	6	100.0	3.0	{TV,"Cable TV",Internet,"Wireless Internet","A...
...	...	...	...	...	...	...
53051	3.688879	Apartment	2	86.0	1.0	{TV,Internet,"Wireless Internet","Air conditio...
53052	3.912023	Apartment	2	60.0	1.0	{TV,"Cable TV",Internet,"Wireless Internet","A...
53053	4.700480	Apartment	2	92.0	1.0	{Internet,"Wireless Internet","Air conditionin...
53060	4.605170	Apartment	1	NaN	1.0	{}
53061	5.220356	Apartment	5	94.0	2.0	{TV,Internet,"Wireless Internet","Air conditio...

24688 rows × 6 columns

Note: This is just a subset of our original columns

Feature Engineering

days_between_reviews

Number of days between first_review and last_review dates of a listing

price_per_capacity

To capture the price relative to the number of people the listing accommodates

We later dropped this feature due to multicollinearity with our price buckets as treatments for our counterfactual estimations

host_days_active

Number of days that a host has been active on AirBnb

distance_to_times_square

Using latitude and longitude coordinates of listings to calculate Euclidean distance from Times Square (tourist attractions)

room_capacity

How many people can the bedrooms of a listing accommodate

price_bucket

3 equal buckets for the price column, labeled as:

Low (0): $1 to $80

Medium (1): $80 to $150

High (2): $ 150 to $1999

rating_bucket

3 equal buckets for the ratings column, labeled as:

Low (0): rating < 90

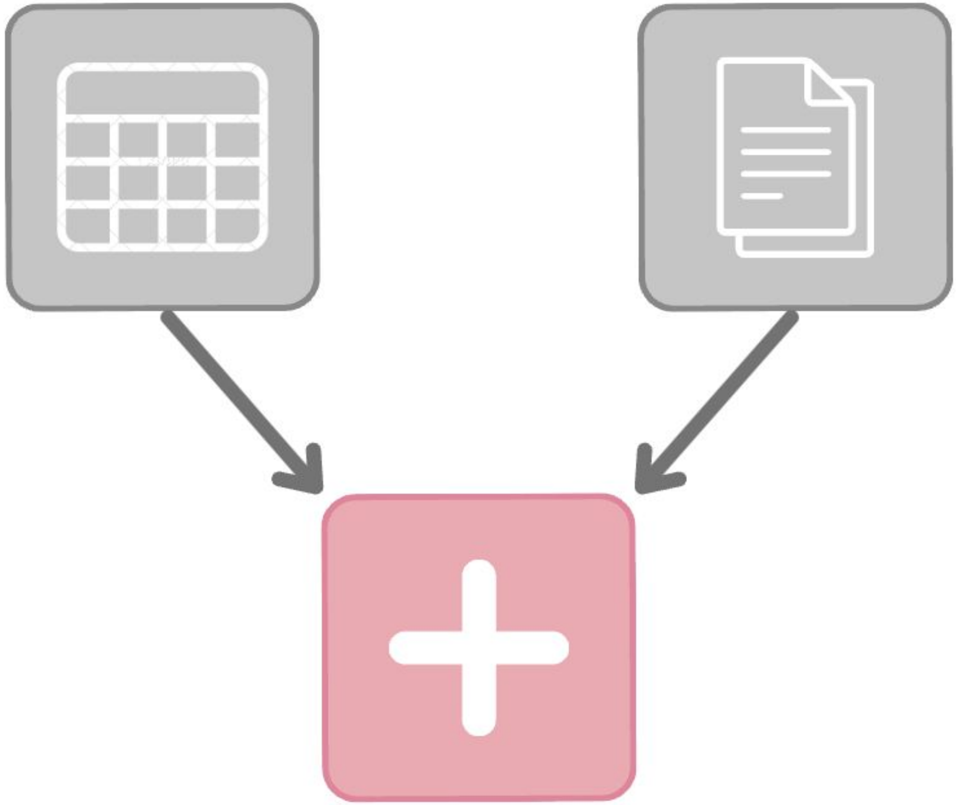
High (1): rating of 90 to 99

Perfect(2): rating = 100

Predictive Modeling: XGBoost

Tabular
Accuracy: 67.3%

Text (BERT)
Accuracy: 52.6%



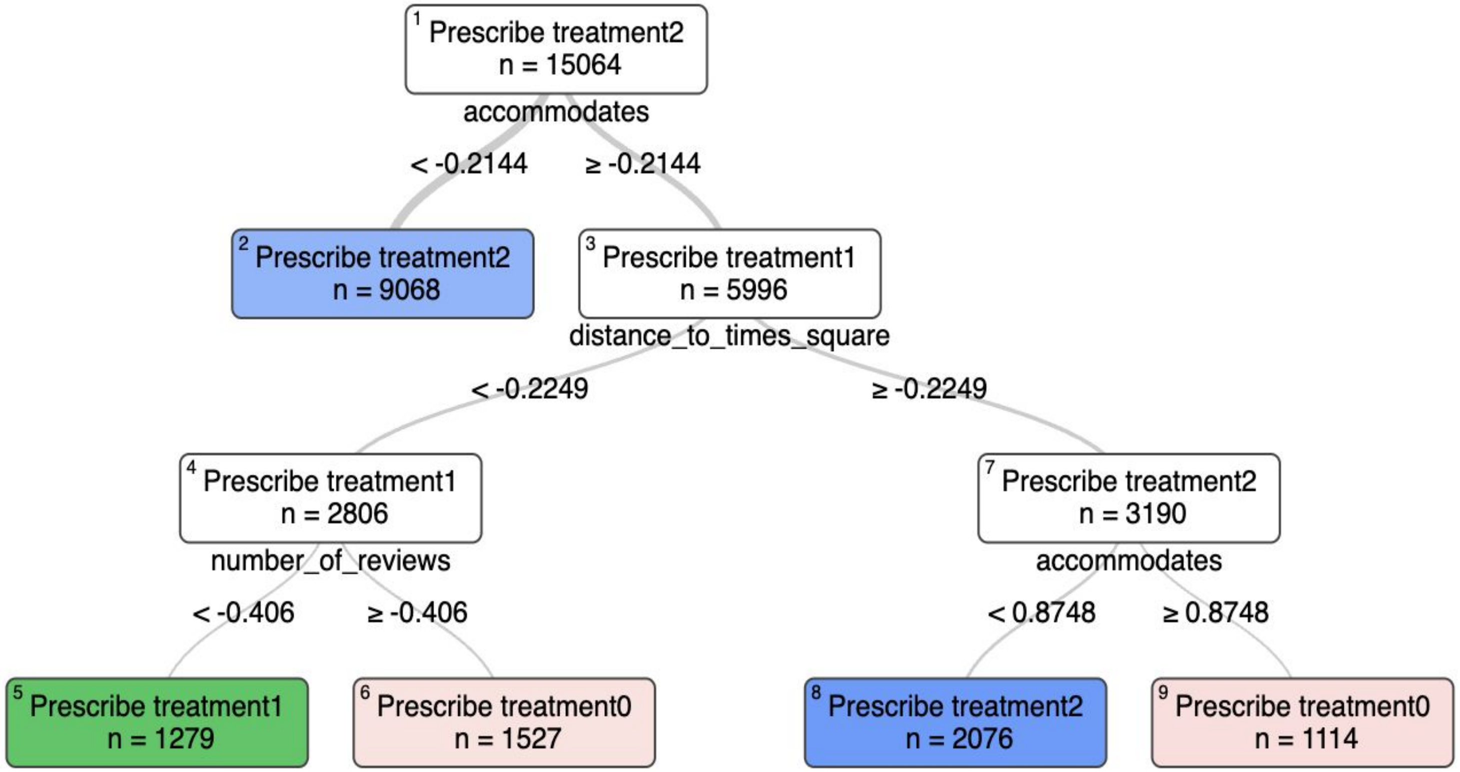
Tabular + Text
Accuracy: 67.1%

Counterfactual Estimates

Observed Treatments	P_0	P_1	P_2	
P_0	Observed P_0 Rating Outcomes	**Counterfactual P_0 if priced in P_1**	**Counterfactual P_0 if priced in P_2**	**6368 Observations**
P_1	**Counterfactual P_1 if priced in P_0**	Observed P_1 Rating Outcomes	**Counterfactual P_1 if priced in P_2**	**6910 Observations**
P_2	**Counterfactual P_2 if priced in P_0**	**Counterfactual P_2 if priced in P_1**	Observed P_2 Rating Outcomes	**5553 Observations**

= **18831 Total**

OPT with Direct Method



Treatment0: $1 to $80
Treatment1: $80 to $150
Treatment2: $ 150 to $1999

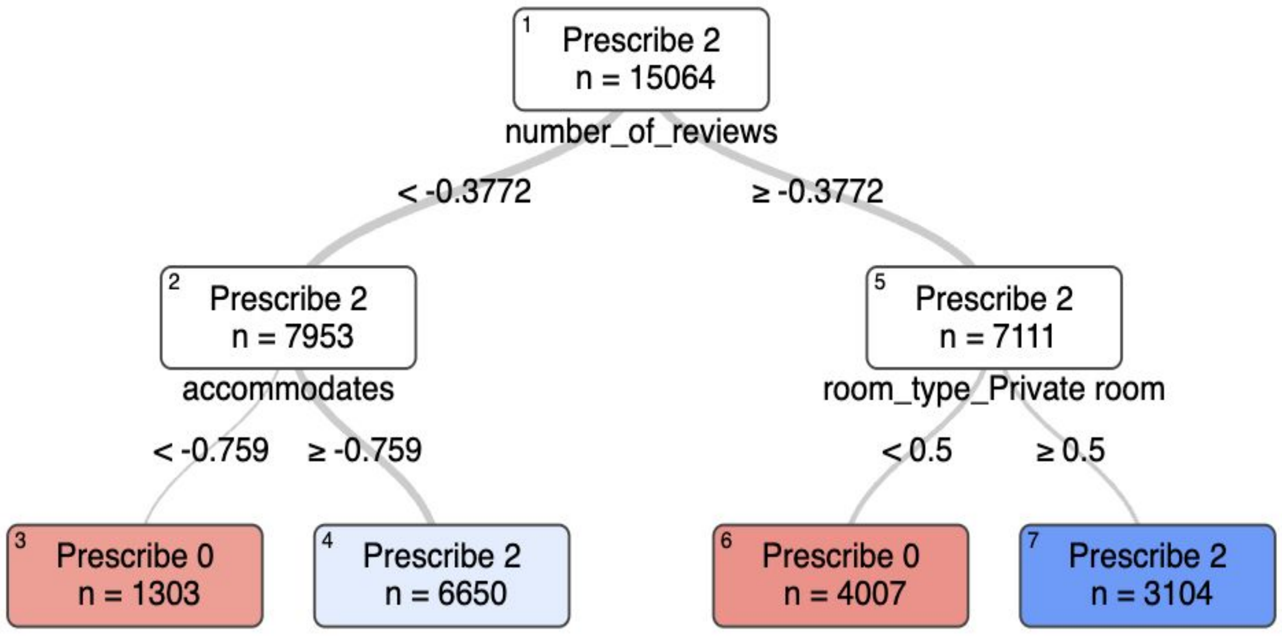
Averaged Rating for Observed Treatments (Baseline): 1.09636

Average Rating for Prescribed Treatments: 1.32598

Percentage Increase in Average Rating with our Model:

↗ 20.94%

OPT with Doubly Robust



Low: rating < 90

High: rating of 90 to 99

Perfect: rating = 100

Low/High: rating of 0 to 99

Perfect: rating = 100

Averaged Rating for Observed Treatments (Baseline):
0.29344

Average Rating for Prescribed Treatments:
0.36039

Percentage Increase in Average Rating with our Model:

↗ 22.81%

Future Direction & Conclusion

1. Incorporate Image Data (Computational Limitation)
2. Adjust Price and Rating Buckets (Data Limitation)
3. Include Customer Reviews (Text Limitation)

Thank you!

Questions?