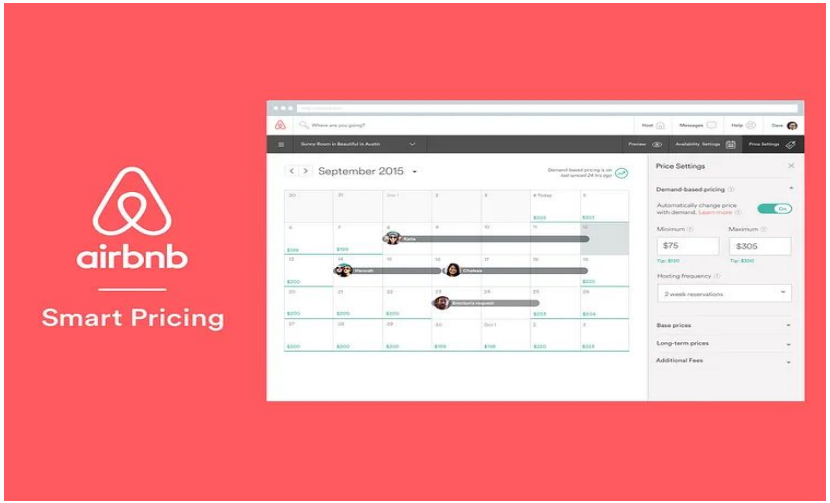
Airbnb Listing Prediction

APRIL 4, 2025

Tam Trinh

PROBLEM

Pricing can be tricky. Hosts may not know the best price for their property, they can under-price and lose out on revenue, or over-price and lose out on bookings. There are dynamic pricing tools, but they tend to take a black-box approach, and hosts usually do not know how the pricing is calculated.



OPPORTUNITY

Airbnb is a major homestay booking service. In 2024*, there were an estimated 490 million bookings of nights and experiences, giving Airbnb an estimated 83 billion market capitalization, and generating about 11 billion dollar revenue in 2024. Airbnb has listings worldwide, with over 5 million hosts, listing an estimated 7.7 million listings. There is an opportunity to help with better pricing, more transparency, and better customization.

* Source: Statistica

IMPACT

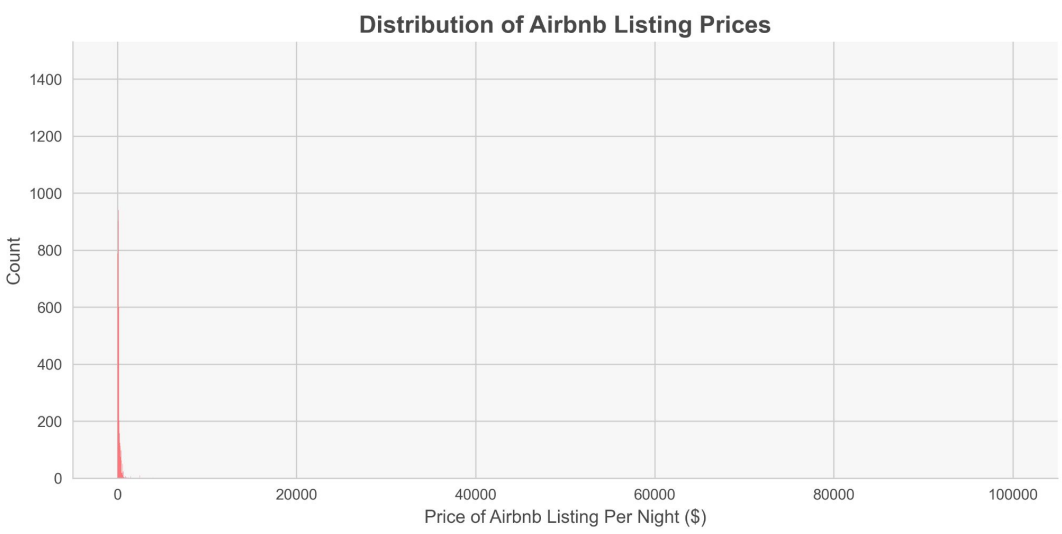
Fairer prices for renters, allowing them more freedom of travel.

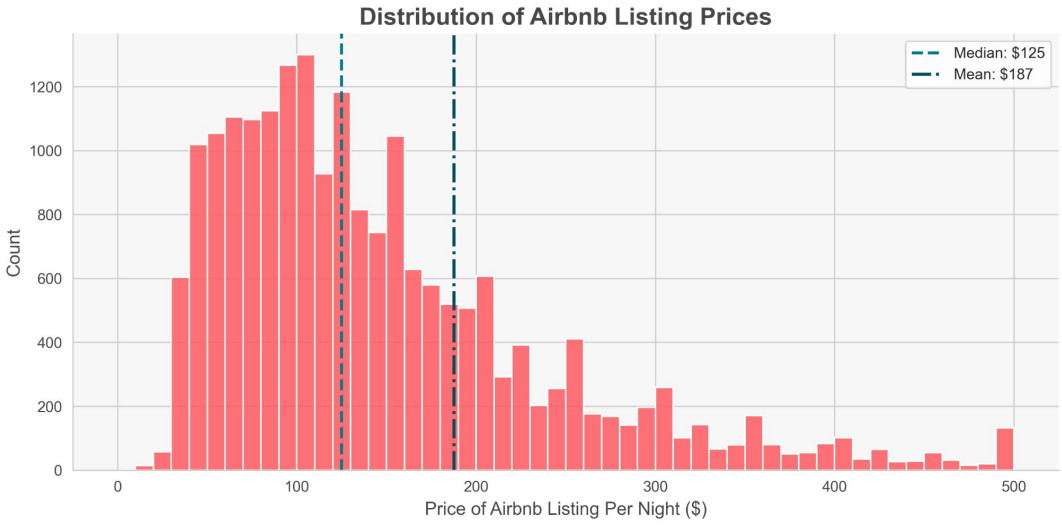
More transparency and control for hosts, allowing them more competitive rates and increased revenue.

Better pricing efficiencies and satisfaction for Airbnb.

DATASET

- ➔ 20,758 observations
- ➔ 22 columns (listing name, neighborhood, rental type, bedrooms, rating)
- ➔ Target variable: Price
- ➔ 3.5% of outlier prices above $500





PREPROCESSING

BEDROOMS & BATHS COLUMNS

String values cleaned and unspecified values filled in

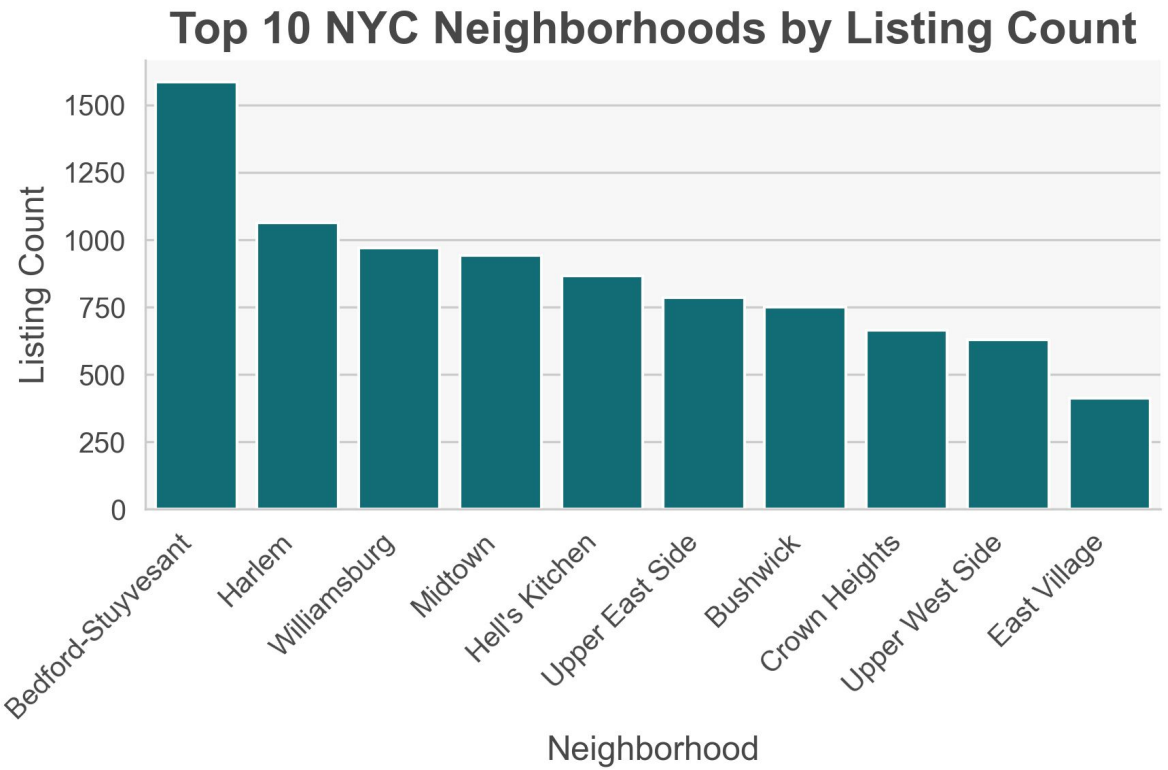
NEIGHBOURHOOD COLUMN

Neighbourhood mean price imputed

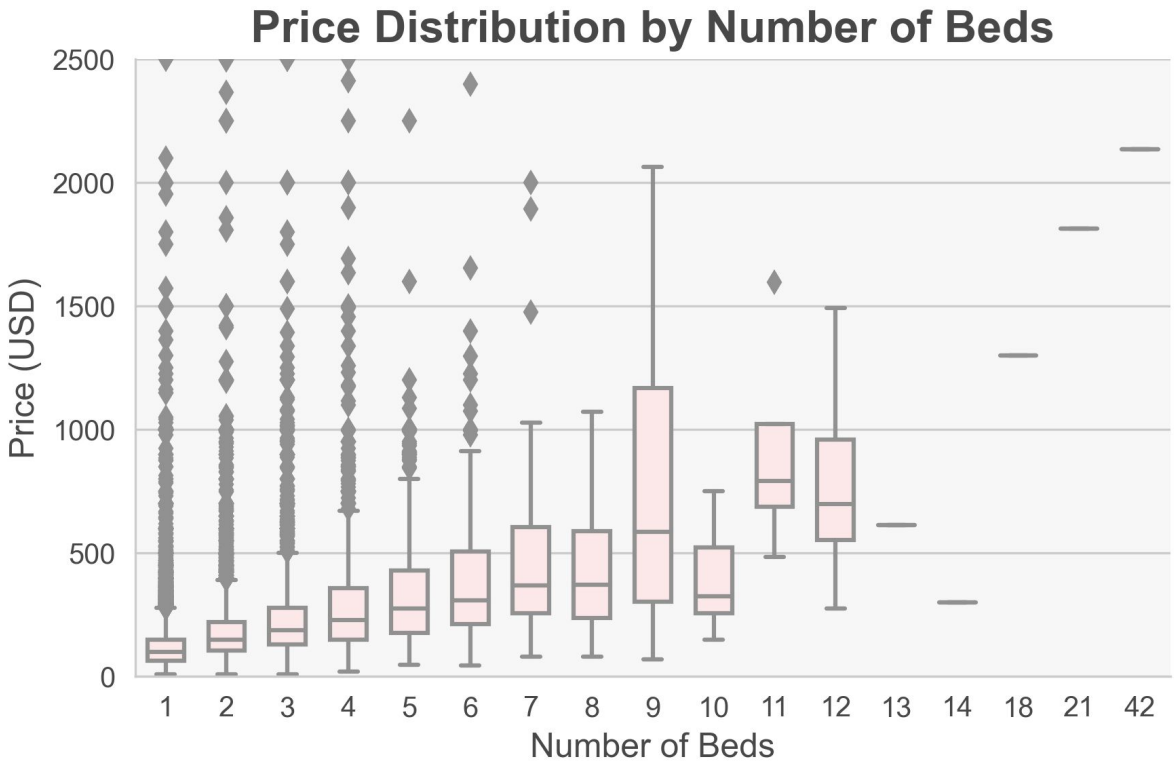
RATING COLUMN

Rating binned and OneHotEncoded

Top 10 Neighborhoods



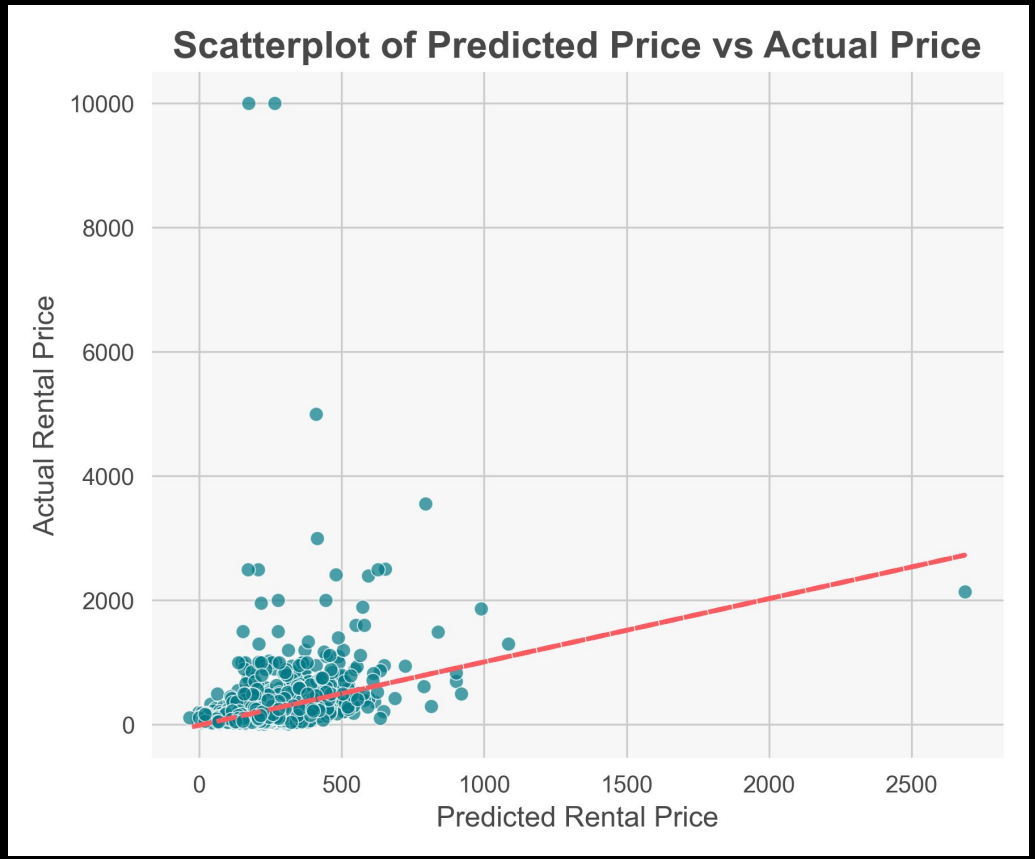
Prices are more predictable for listings with less beds



MODELING

Predicted prices miss outliers

- ➔ Actual rental prices has extreme outliers
- ➔ Outliers heavily punish linear regression model



SEGMENTING CAN BE GOOD FOR PREDICTION

Scores before price limit

Score	Linear Regression	XGBoost	Random Forest
Train score	0.008	0.99	-0.15
Test score	0.13	-5.87	-0.064
RMSE	284.44	800.15	315.01
MAE	98.53	93.94	91.52

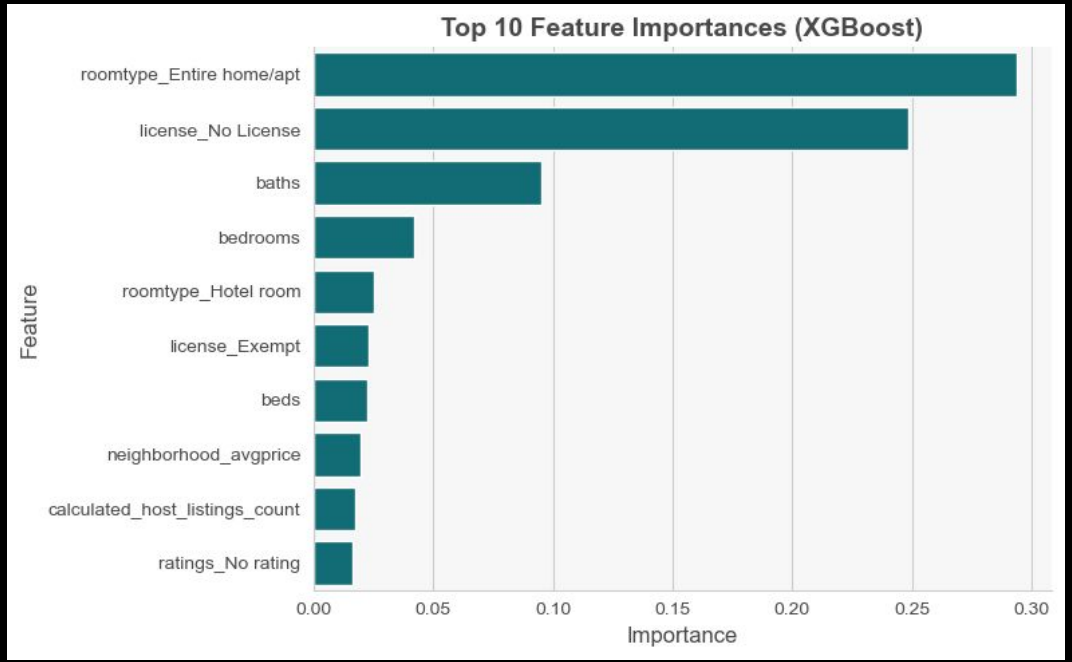
Scores after price limit

Score	Linear Regression	XGBoost	Random Forest
Train score	0.42	0.80	0.94
Test score	0.39	0.52	0.52
RMSE	71.5	62.91	63.61
MAE	51.32	43.58	43.72

MODELING

XGBoost: Feature Importance

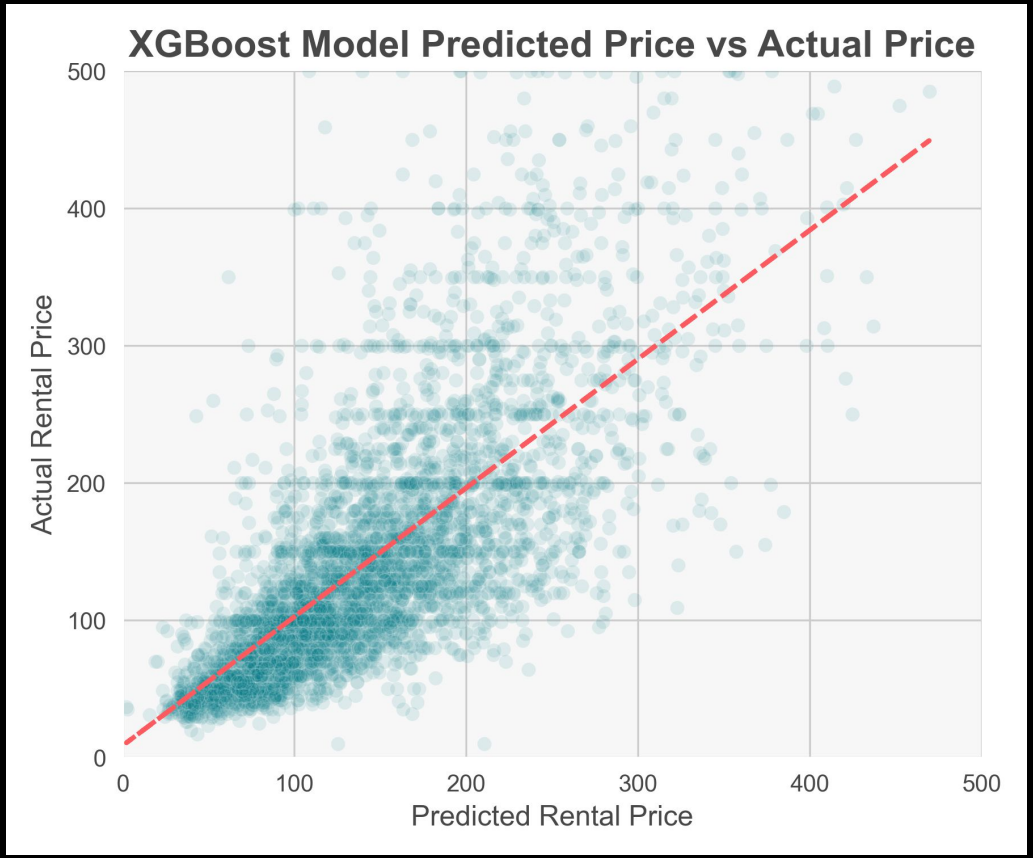
- ➔ Entire homes have largest impact
- ➔ Licensing might show an issue
- ➔ Baths and bedrooms, intuitively have a large impact



MODELING

XGBoost

- ➔ Improved predictions, closer to the line
- ➔ Dense area under $200
- ➔ Uniform prediction



Next steps

- ➔ Tune features to improve model performance
- ➔ Build application tool to take in user input

Thank you

https://github.com/tam3ourine/Capstone_Airbnb/tree/main