airbnb

Start your search

Get started

Our Team

Eunice Worifah
Data Scientist

Fandi Yi
Data Scientist

Pascal Nguyen Tang
Product Manager

Shivangi Soni
Data Analyst

Vivek Saahil
Business Analyst

euniceworifah | Bubbletea98 | pnguyentangmcgill | shivangi-soni | vsaahil

Business Context

Airbnb provides a platform for hosts to accommodate guests with short-term lodging and tourism-related activities

Use Case: Homeowners currently employ three types of revenue management strategies:

- 1 – Set one price for the entire year
- 2 – Airbnb's Smart Pricing tool
- 3 – A third-party intelligent pricing tool (e.g. PriceLabs or Wheelhouse)
  - ***Our Prediction Model**

Stakeholders:

- Current Hosts
- Prospective Hosts
- Prospective Guests

Objectives:

- Build a model which enables:
  - Existing Airbnb Hosts to update their pricing strategy
  - New Airbnb Hosts to find the best and most competitive price for their property
  - Airbnb guests to define their budget when looking for a place to book

Hypotheses

Starbucks effect

- Airbnb listings which are located in areas with a large number of Starbucks will be more expensive on average

Metro effect

- Airbnb listings which are located near a subway station will be more expensive on average

Rating of Listing

- An Airbnb listing's rating does not have a significant effect on the price of the listing

Approach

- **Data Acquisition**
  - Internal/External Sources
- **Data Transformation**
  - 6 data transformation tasks
- **Data Exploration**
  - Visualizations
- **Pre-Processing**
  - Flagging Missing Values
  - Iterative Imputation
  - Categorical Encoding
  - Correlation Analysis
  - Outlier Treatment
  - Feature Selection
- **Model Building**
  - **Unsupervised**
    - PCA
    - Autoencoders
  - **Supervised**
    - AutoML
    - Random Forest
    - SVR,
    - XGBoost
    - GBT
- **Hyperparameter Tuning**
  - MLFlow (Hypoeropt) / RandomSearchCV
- **Deployment**
  - Pandas UI
  - Docker

Data Acquisition

- Airbnb Data
  - Source: http://insideairbnb.com/get-the-data.html
- Starbucks Data
  - https://www.starbucks.com/store-locator?place=New%20York%2C%20NY%2010001%2C%20USA
- Metro Data
  - Source: https://catalog.data.gov/en/dataset/nyc-transit-subway-entrance-and-exit-data

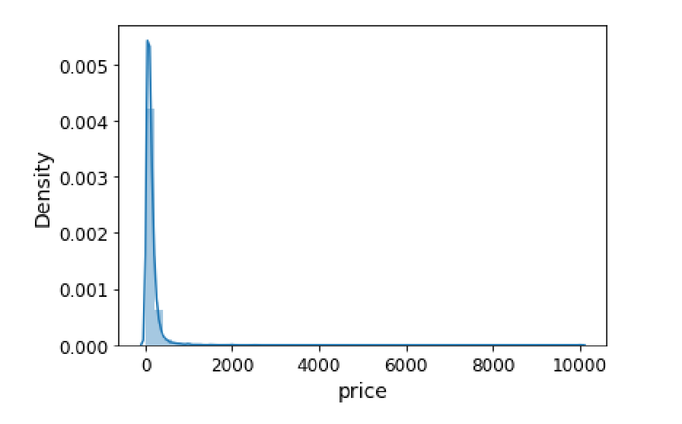
Overview Warnings 46 Reproduction

Dataset statistics

Number of variables	45
Number of observations	37012
Missing cells	186083
Missing cells (%)	11.2%
Duplicate rows	2
Duplicate rows (%)	< 0.1%
Total size in memory	12.7 MiB
Average record size in memory	360.0 B

Variable types

Categorical	13
DateTime	3
Numeric	25
Boolean	4



Data Transformation & Preprocessing

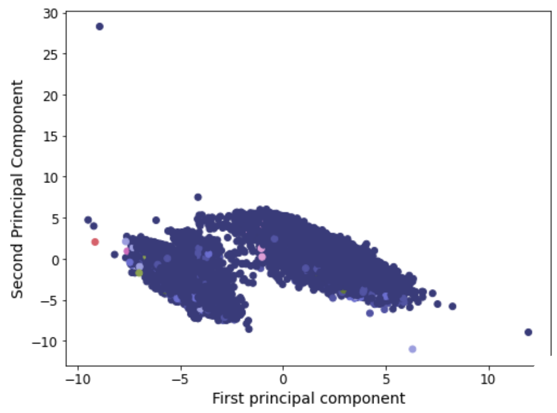
- **Conducted 6 data transformation tasks:**
  - **Bathroom**
    - Separated 'bathroom_text' (e.g. 3.5 baths) variable into 'num_bath' (3.5) and 'name_bath' (bath)
  - **Property Type**
    - Grouped categories into larger buckets, e.g. 'Townhouse', 'Apartment', 'other', etc...
  - **Amenities**
    - Found the top amenities and created dummy variables
  - **Dates**
    - Calculated duration of listing using 'host_since' and 'date_scraped' data
  - **Sentiment Analysis**
    - Obtained sentiment score for descriptive values such as 'description', 'host_about', 'neighbourhood_overview'
  - **Metro Distance**
    - Calculated the distance in KM from the nearest metro station to Airbnb location

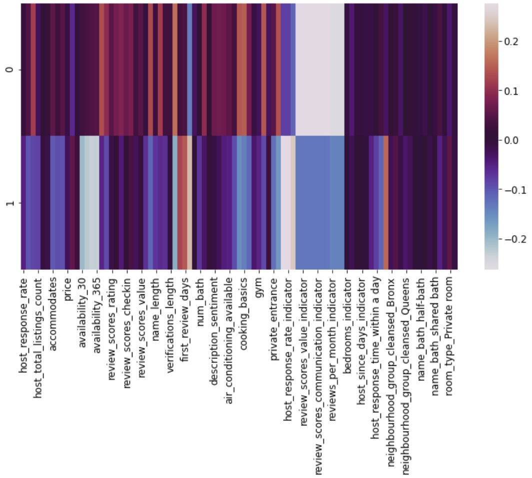
- **Imputations and Flagging:**
  - **Flagging before imputation:**
    - We flagged the nan value in the dataset into the new columns ('Feature_name+indicator')
    - Nan=1, valid =0
  - **Numerical columns Imputation:**
    - Applied iterative imputer in the sklearn package to impute the missing value in numerical cols
  - **Categorical columns Imputation:**
    - Set Nan value as the mode category
    - Set Nan value as the new category (ex: other, unknown)

Models and Results

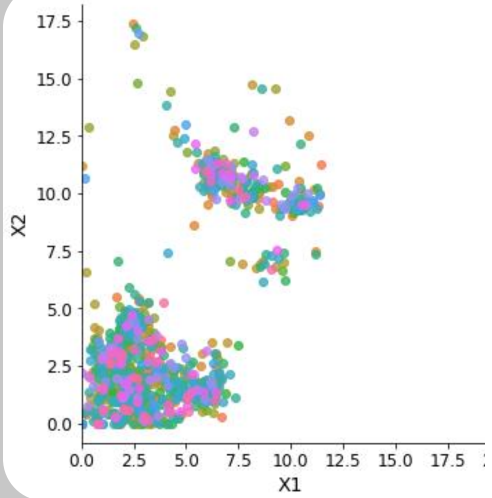
Results for Unsupervised Learning Models

PCA

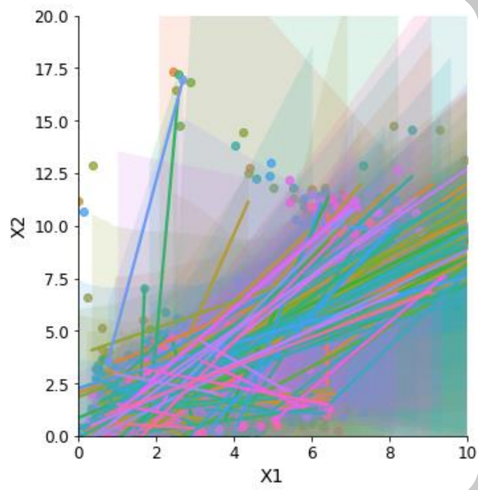




Autoencoder



fit_reg = False

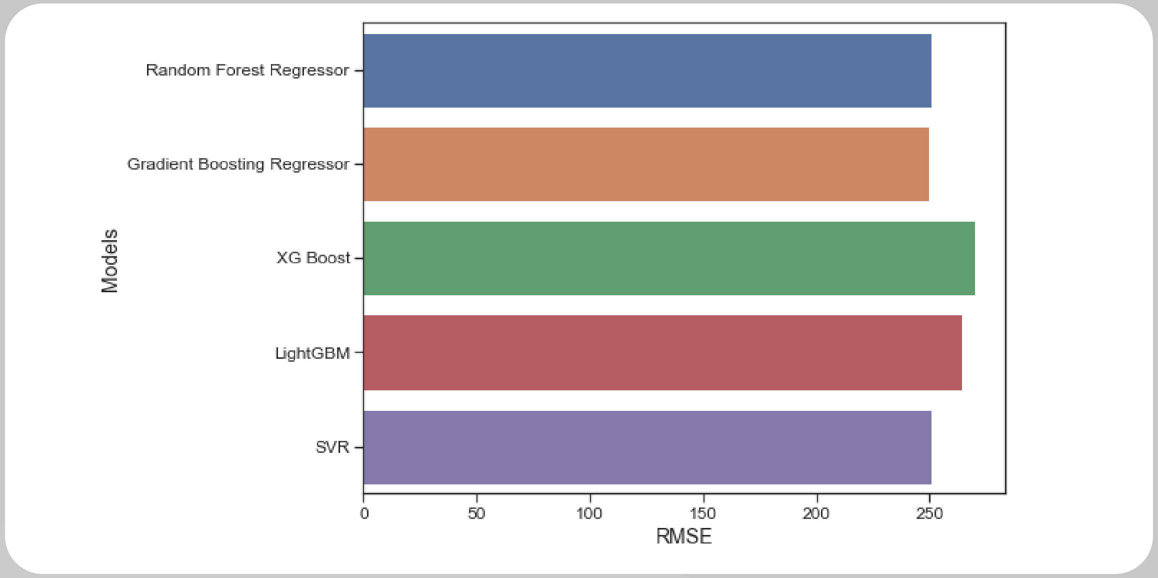


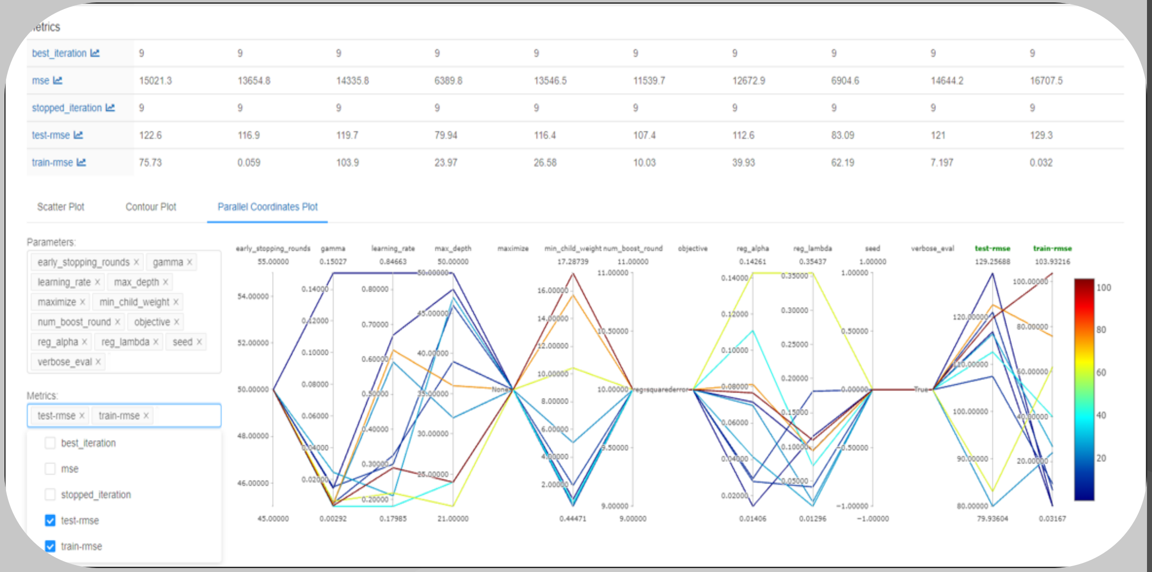
fit_reg = True

Models and Results

Results for the models to predict price of the existing listings

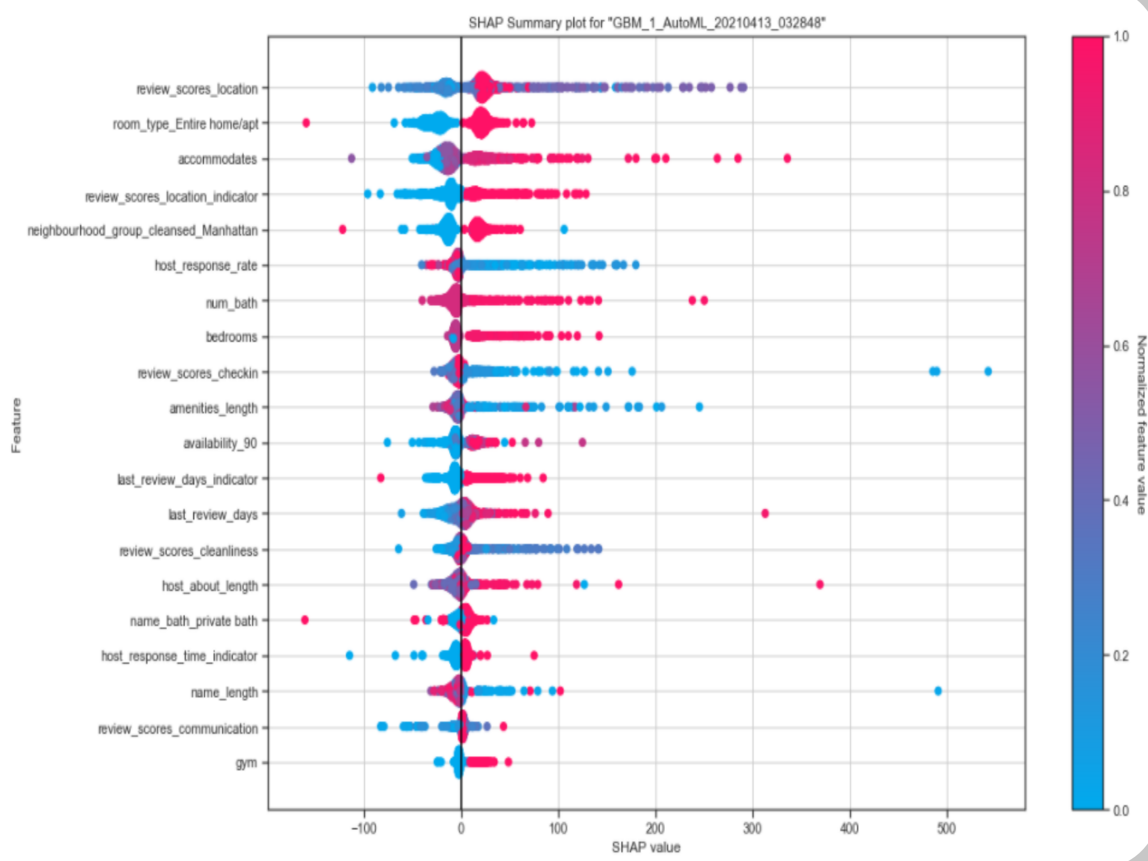
- Recursive Feature Analysis was used to conduct feature engineering
- Several models were tested
    - RMSE did not vary a lot between different models
    - ML Flow used to hypertune XG Boost model

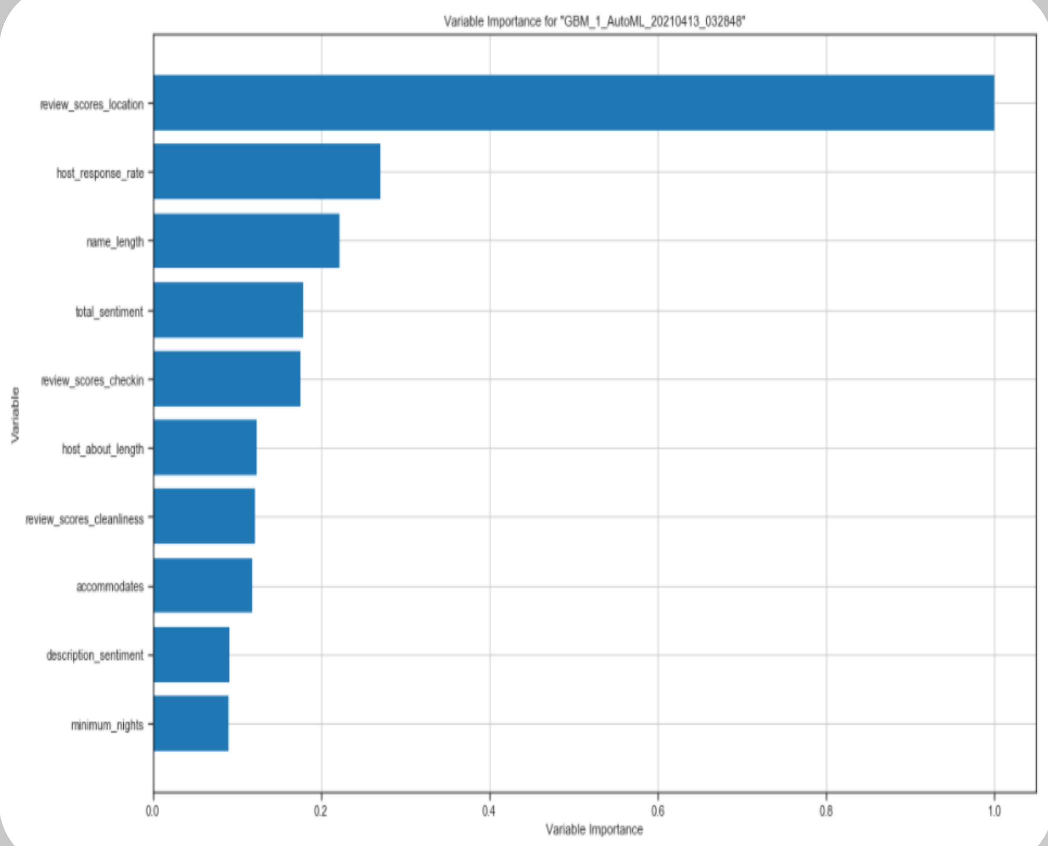




Explainability & Feature Importance – Best Model

Results for AutoML in the best performance model:

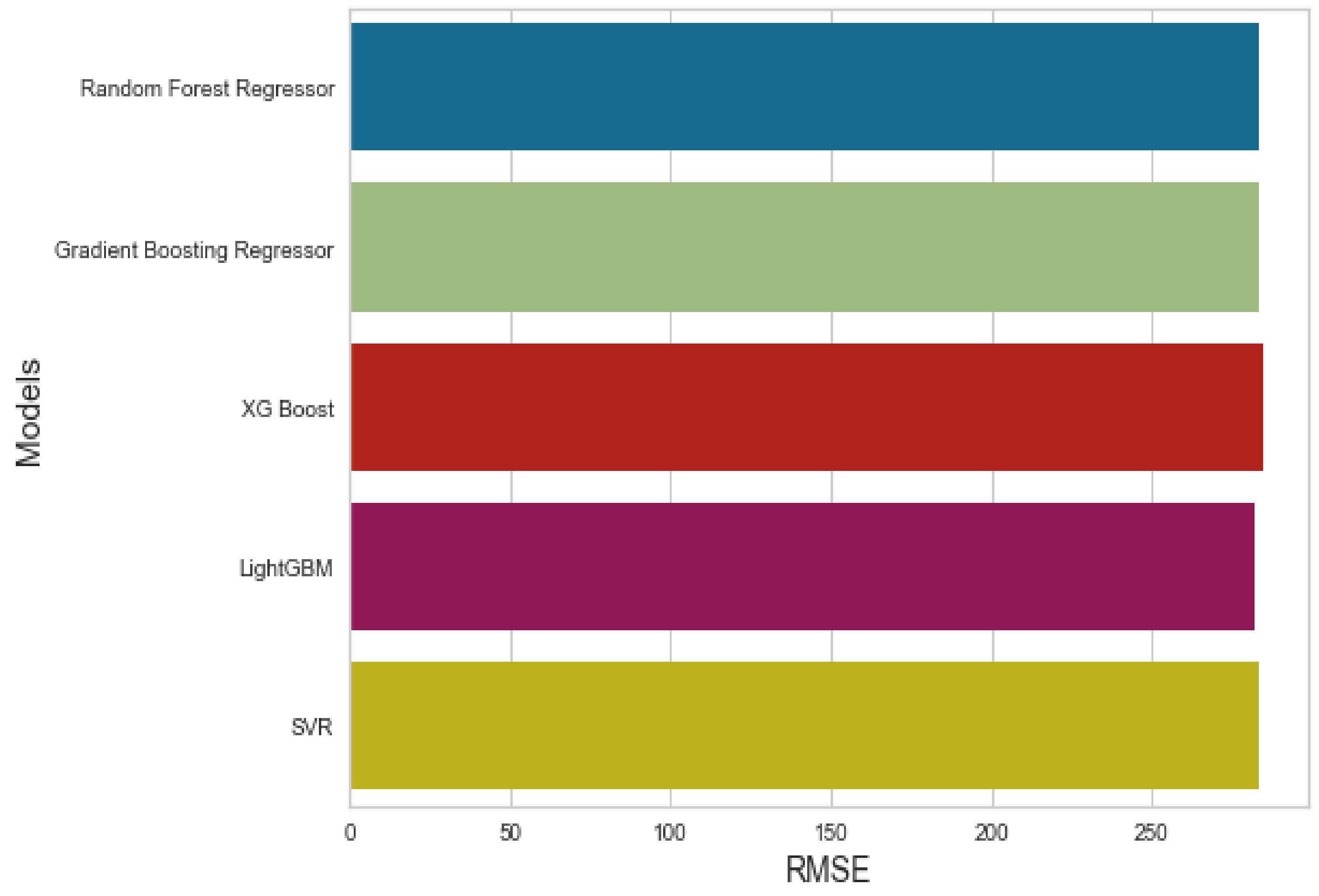




Models and Results

Results for model to predict price of new listings

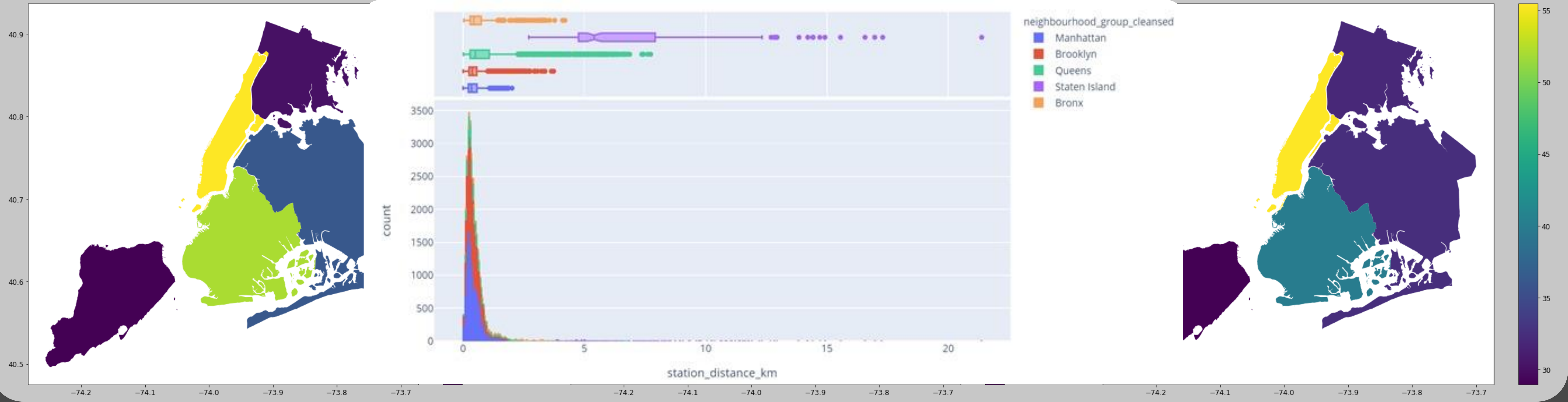
- All features leading to data leakage were dropped



Hypothesis Results – Starbucks Effect & Station Distance

Data Exploration and Causal ML using DoWhy

- Airbnb Data: http://data.insideairbnb.com/united-states/ny/new-york-city/2021-02-04/data/listings.csv.gz
- Starbucks Data: https://www.starbucks.com/store-locator?place=New%20York%2C%20NY%2010001%2C%20USA



Number of Airbnbs in NY

The Distribution of the Metro Distance Across Boroughs

Price per accommod. in NY

Model 1

Number of Starbucks & Review Scores Rating

Mean value: 0.003231172176185737

p-value: 0.04047677

Inference: Significant (*)

Model 2

Closest Train Station Dist. & Review Scores Rating

Mean value: -0.002887094771580223

p-value: 0.19418372

Inference: Not Significant (.)

Model 3

Number of Starbucks & Price per accommod.

Mean value: -0.06782232293168278

p-value: 0.22349717

Inference: Not Significant (.)

Model 4

Closest Train Station Dist. & Price per accommod.

Mean value: -4.216259842350333

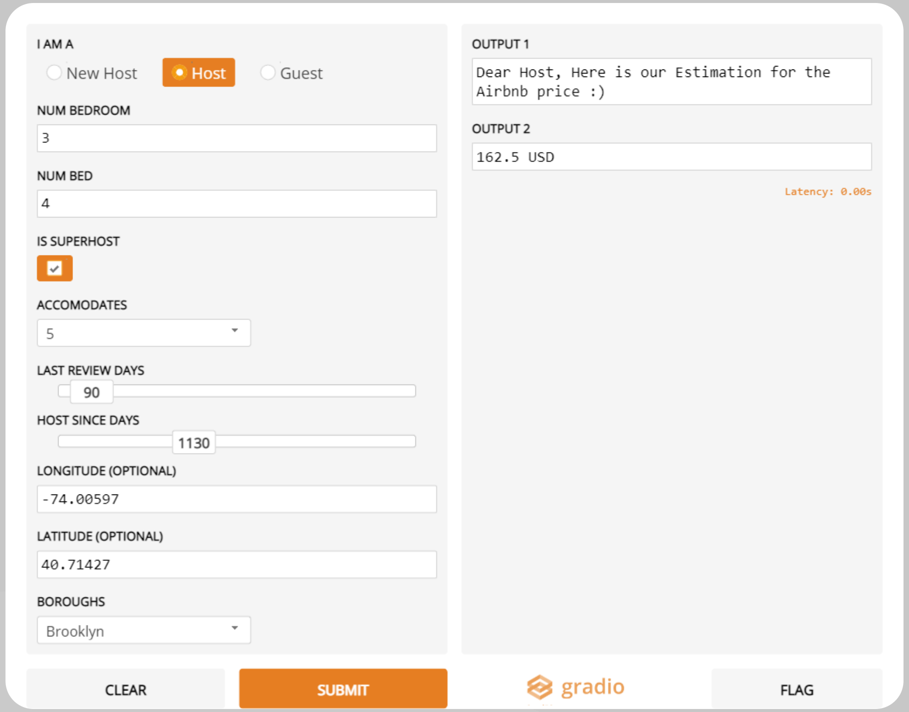
p-value: 0.03357141

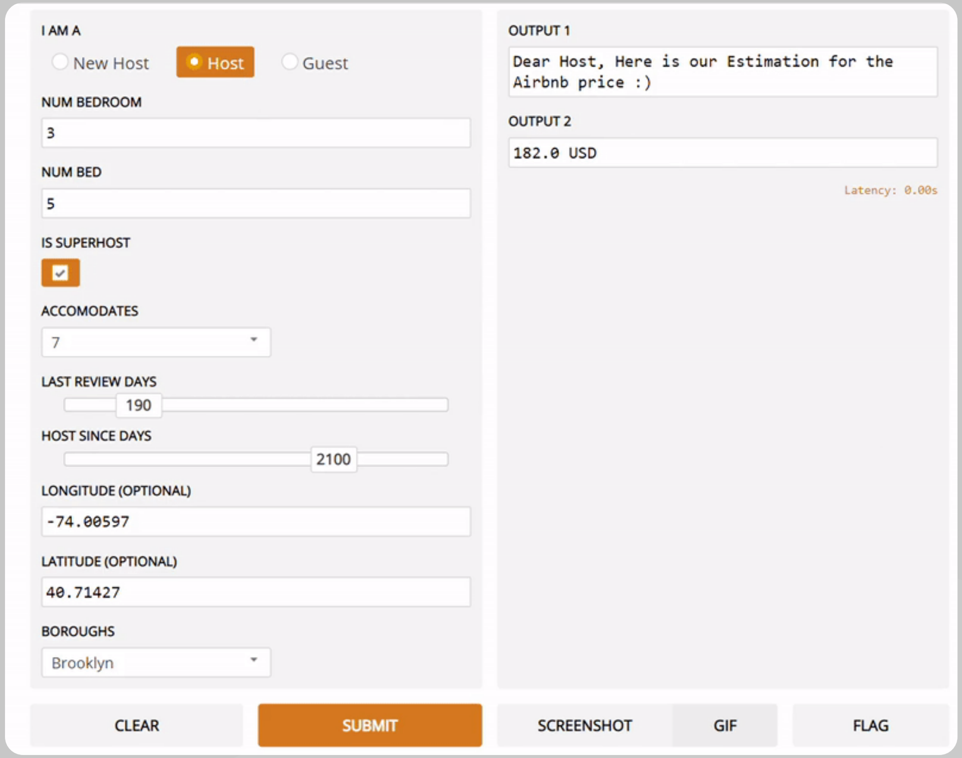
Inference: Significant (*)

Insights and Business Implications

Homeowner

- Pricing Strategies – balanced between the host, the customers and Airbnb
- Gain an understanding of what features to add to a listing to increase value of property
- Gain an understanding of which locations to rent/purchase a home as an investment
- In our UI interface, we targeted our users to new host, host and guest.



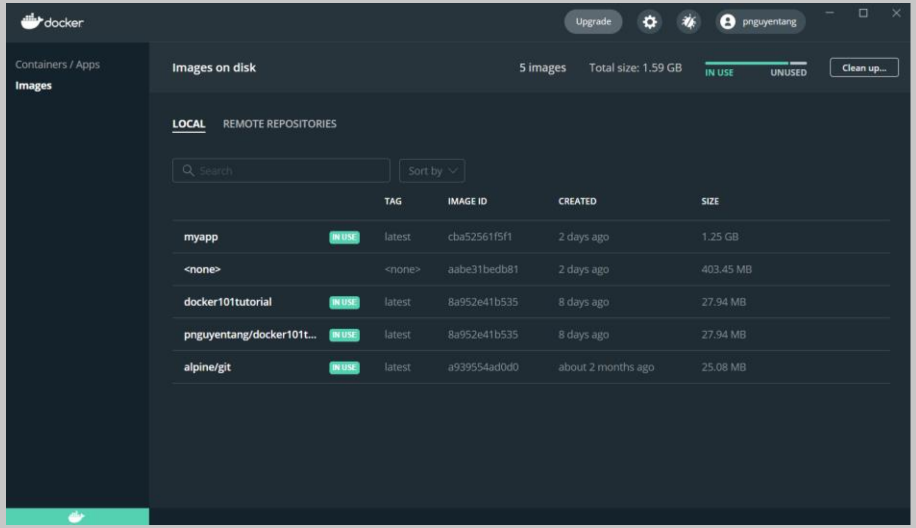


Threats to Validity

- Exogenous Shocks
  - **Effect of COVID-19 on target variables:** Historical data used does not reflect post-COVID shifts in demands, rating and prices
  - **Effect of COVID-19 on Starbucks overlay:** Due to COVID-19 restrictions on businesses, the Starbucks overlay may no longer be significant
- Extraneous Factors
  - **Omission of other potential restaurants/shops:** The analysis did not include other restaurants or shops that may also have a similar effect than Starbucks
  - **Analysis only performed on NYC:** The analysis did not include other cities, for which the selected overlays may not apply

Lessons learnt & Next steps

- Importance of external data
  - Worldwide
- Utilization of standard industry tools
  - AutoML
  - MLFlow
  - Docker
  - AutoEncoder
  - SHAP plot



Next Steps & Future Improvements:

- Expand our model to different geographic locations and use post- pandemic data to enlarge our user base.
- Docker
  - The infrastructure for Docker has been set up
  - Image containing the necessary requirements is created
  - Must test further to ensure the entire project can be run through the container

Thank You

Hypothesis Results – Rating effect on Price