Analyzing Airbnb Features for a Smoother Travel Experience

Karim Hazimeh

Mohammed AlAlem

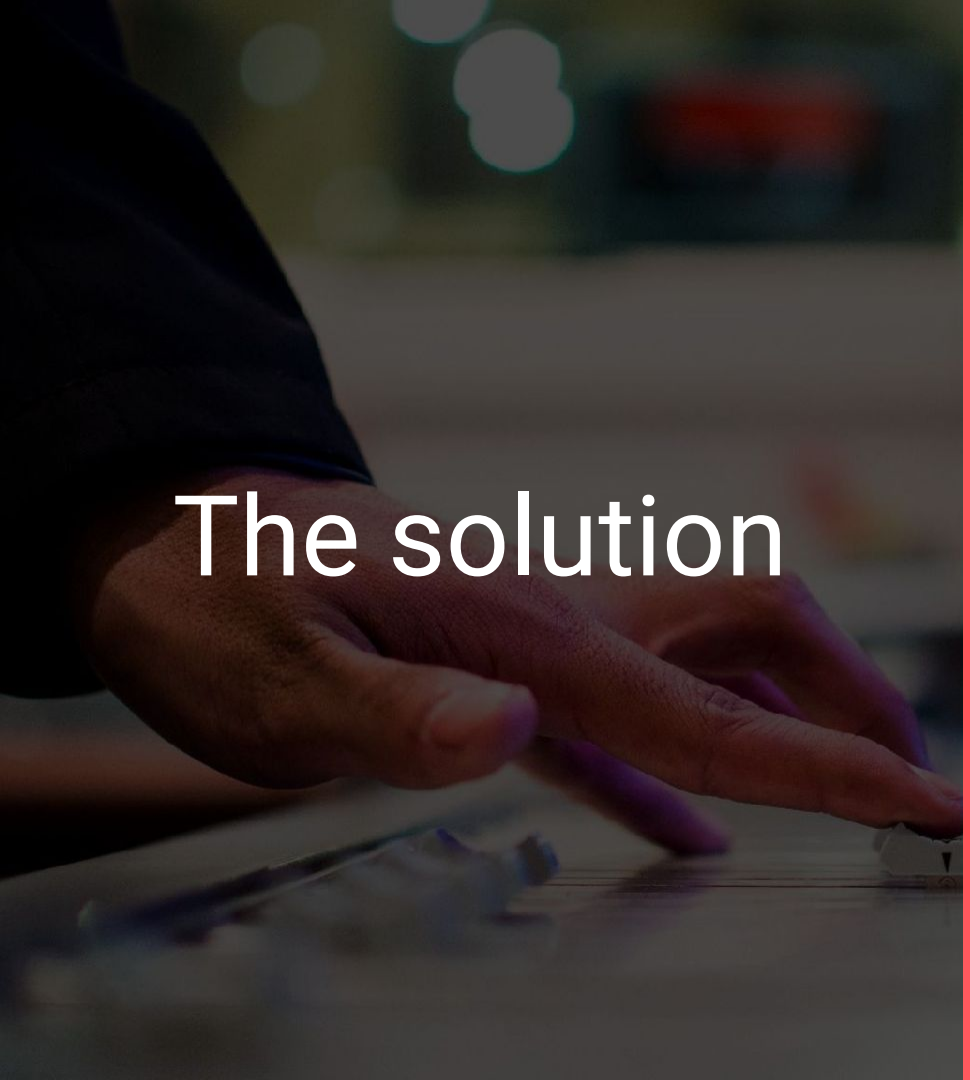
The Problem

As an Airbnb traveler:

- The issue of booking within a short time frame and picking the best staying based on location, facilities and price is always a big challenge to face.

As an Investor:

- Investors who are looking to buy or invest in a property, and list their own to get an idea about its price in the market.



The solution

- ✓ Analysis of Airbnb data
- ✓ Price prediction with machine learning models
- ✓ Recommender engine to help users choose their stayings

Part 1:
Airbnb Data Analysis

Overview analysis

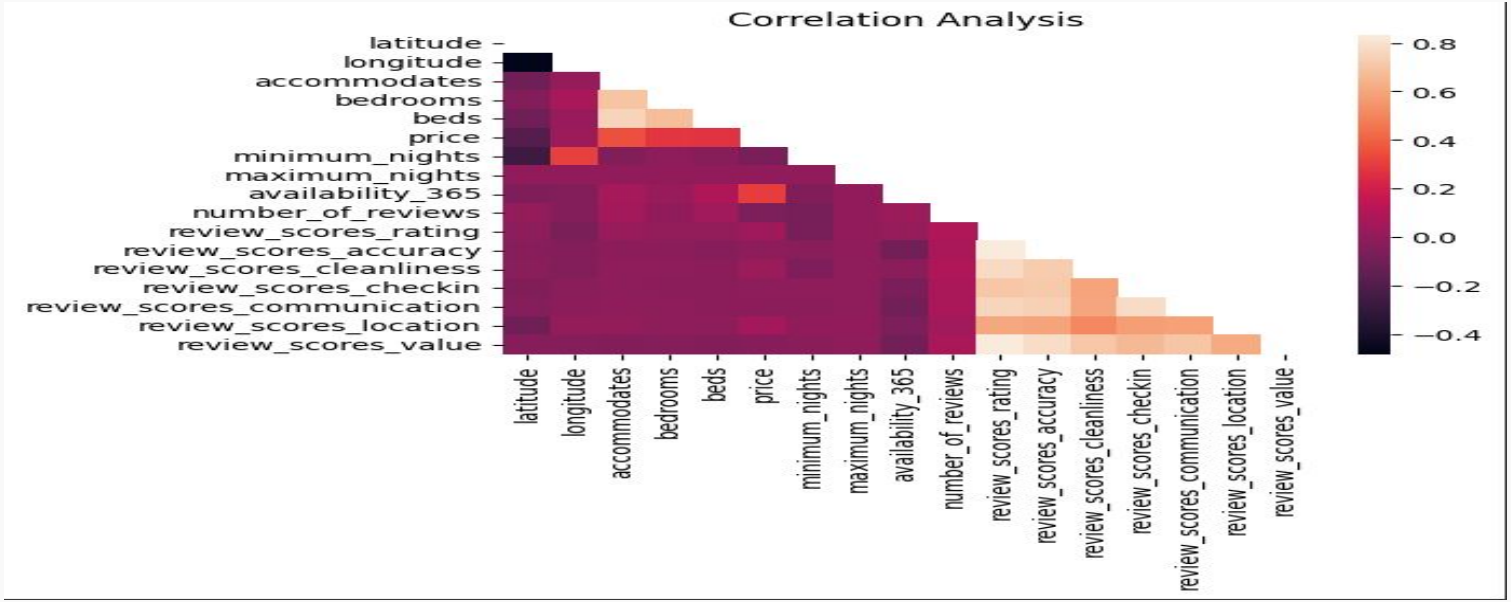
- We extracted airbnb listings from Inside Airbnb open source dataset
  - We picked 7 cities from London, Barcelona, Singapore, Sydney, Istanbul, New York and Rio de Janeiro
  - It has 211030 rows and 76 columns

- Removed features that will not contribute to our analysis and will not be with any added value (e.g: Scrape_id, Last_Scraped_Date, Source, Host_Url, redundant features, etc…). The total number is reduced to **42 rows.**

- Data cleaning stage included converting all data types to the right format (Price with dollar sign from string to a float), removing outliers, dropping all null values which exceeded 30% of the data and doesn't contribute to the study we're conducting

Data exploration

- We realize that the price has a positive correlation with bedrooms and accommodates higher than other features.
- Most of features that are positively correlated are the review scores since the higher the rating the higher the other scores like cleanliness, check-in etc



Data exploration

- The amenities at each listing also influence the price especially as we were investigating the added value of listing with higher prices.
- Listings with long-term stays, parking premises, wifi, hot water, kitchen, tv, air conditioning, and with smoke alarms will have greater prices.

Data exploration

- The following map illustrates sydney listings based on location, price, and neighborhood.
- The closer the listings are to the bays like Bandi Bay the higher the prices are (For example 3 BR Home in Bondi Beach property


Part 2:
Machine Learning

Data preparation and pre-processing

- We selected 6 features which are room type, property type, bedrooms, amenities number_of_reviews for machine learning models.
- We transformed all categorical columns to integers by using a technique called one hot encoding (converting each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns.)
- Cleaning the “amenities” column by grouping similar meanings under one umbrella, for example, we had “Step-free access”, “Wheelchair”, and “Accessible”. If any “amenities” column entry has any of these they will be turned to “accessible”
- Models used are multiple linear regression as a baseline model and the other 2 ones are xgboost regressor and random forest regressor.

Experiments and results

The measure used to evaluate is **root mean squared error**.

Models	**Number of features**	**Score**
Multiple Linear Regression	23	164931338
Xgboost regressor	6	150.87
Random Forest	6	149.9

We notice how number of features improved our model drastically , additionally the first one was a baseline model only.

Experiments and results

We came up with **random forest model** as the best one and we evaluate the importance of the features accordingly

Features	**Importance**
number_of_reviews	21.35%
bedrooms	8.75%
Entire home/apartment	4.5%
Private room	3.94%
air_conditioning	3.36%


Part 3:
Image similarity
models

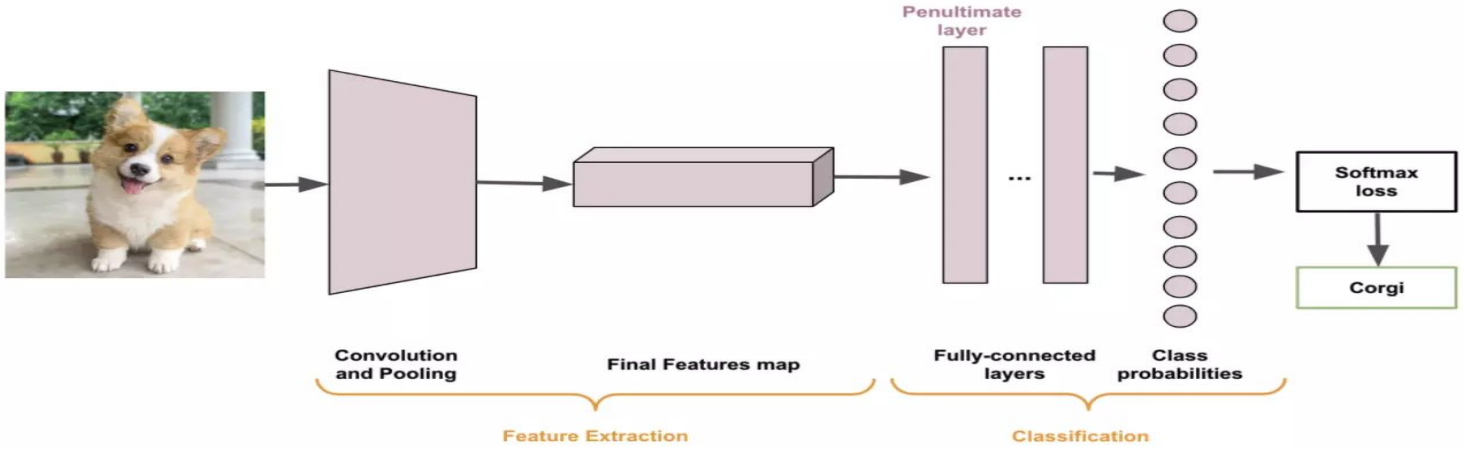
Loading Images

- Images in our data were in form of URLs and hence we created a function in python looping over each link, reading the image and add it to a file accordingly.
- Load our images and add them to the data with the URLs linked to them.
- Apply the regular expression to find and capture strictly the images with extensions, load them into separate data then implemented a left merge to get the paths of our images associated with each listing.

Applying Image Embedding

- An image embedding is a lower-dimensional representation of the image. In other words, it is a dense vector representation of the image which can be used for many tasks such as classification

- A **convolutional neural network** (CNN) is used to create the following:
  - We began with pre-processing the images by loading them using Keras and OpenCV packages in python and then resizing them and turning them into an array so that the model will be able to be fed by it.



Our Approach

- We used 2 models for image embedding and which are well-known in the research field. **ImageNet** is a project intended to label and categorize images manually.

- The pre-trained models used were first **Xception** which is a convolutional neural network that can classify images into 1000 object categories and is 71 layers deep after passing the images into the following model and obtaining vectors, we used these vectors as cosine similarity metrics throughout KNN model.

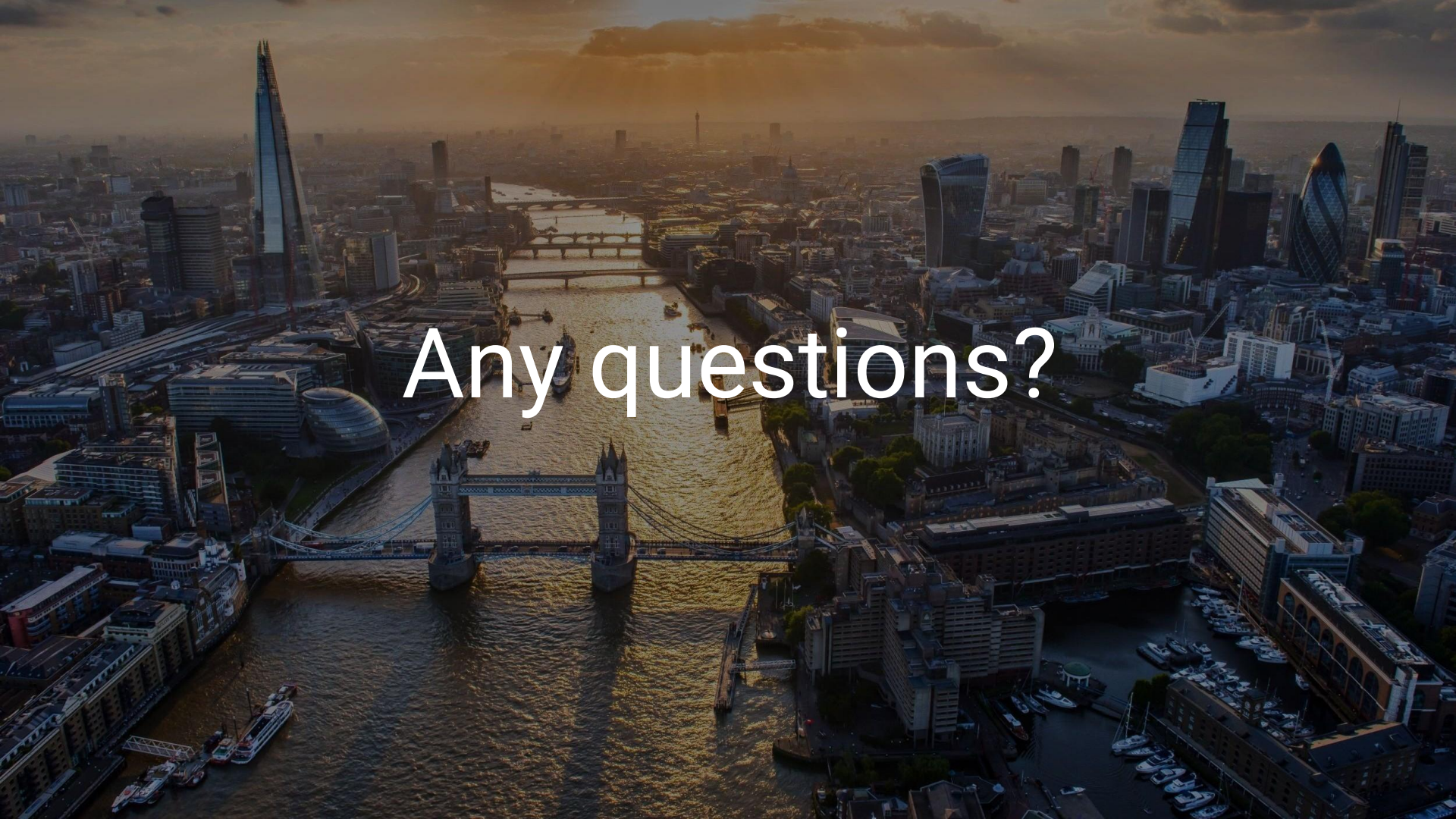
- The second model **Resnet50** is a 50-layer convolutional neural network (48 convolutional layers, one MaxPool layer, and one average pool layer).
    - The weights used are also imagenet and the similarity measure used is cosine similarity but we computed using pairwise distances.
    - As an output, the embeddings received by CNN models are fed into similarity measures and the listings will be recommended based on the most similar ones

Final Output & Deployment

- Deployed locally an app on streamlit.
  - Decomposed into two parts:
    - First part, **data exploration** where:
      - The user may explore different aspects of data:
        - The content of it through rows and columns values.
        - Data visualization through correlation plots, bar plots, interactive maps and customizable plots.
    - Second part, deployment of the **image similarity** model:
        - Once users click on get listing:
          - An image of the listings with all the details will be uploaded
          - When users scroll-down they will check all the similar listings with their URLs to visit the listing and get all the details for their desired stayings.

Future Work

- Tourism is a vast world where great ideas could turn into actions and we are glad that we proposed a new approach to booking tourism where the user may check similar listings based on the images of hosts.
- We were limited with the data strategies if we had data about customer transactions on these listings. The model would have been more robust, especially with a history of purchases of such listings.
- An approach to even get the recommendation more accurate is that hosts may categorize the most appealing images of their listings at first so that customers will click right away on the first image and get recommendations accordingly.


Any questions?