AUG 2019

TEXAS
The University of Texas at Austin

AUSTIN AIRBNB DATA ANALYSIS

PRESENTER
Shuheng Ma
Yikang Wang
Yue Tian

Table of Contents

1. Descriptions of the Project goals

2. Exploratory Analysis

3. Solutions and insights

01

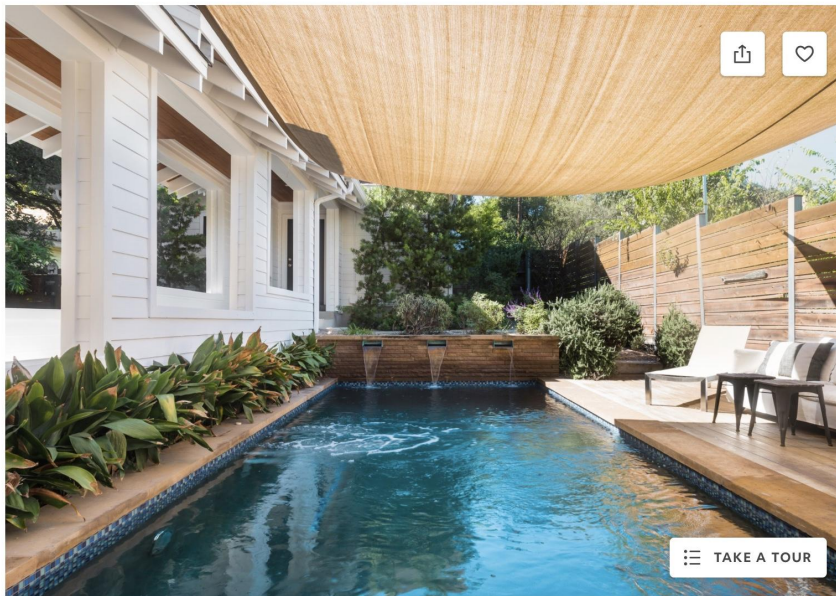## Description of the Project Goals
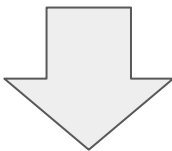
Pricing Problem

Description

Dataset:

- the Airbnb Dataset in Austin area.

Questions:

- Key features for influencing pricing of the listings

- Provide listings' recommendations for existing hosts
- Provide approximate pricing ranges for incoming hosts

Importance

- Number of listings: 11091

- Number of listings: 9963

- Economic Impact
  - Host: more profit
  - Customer: more pleasant experience
  - Airbnb: increase the quality of listings
  - Austin: boost the touring industry

Source:http://insideairbnb.com/austin/
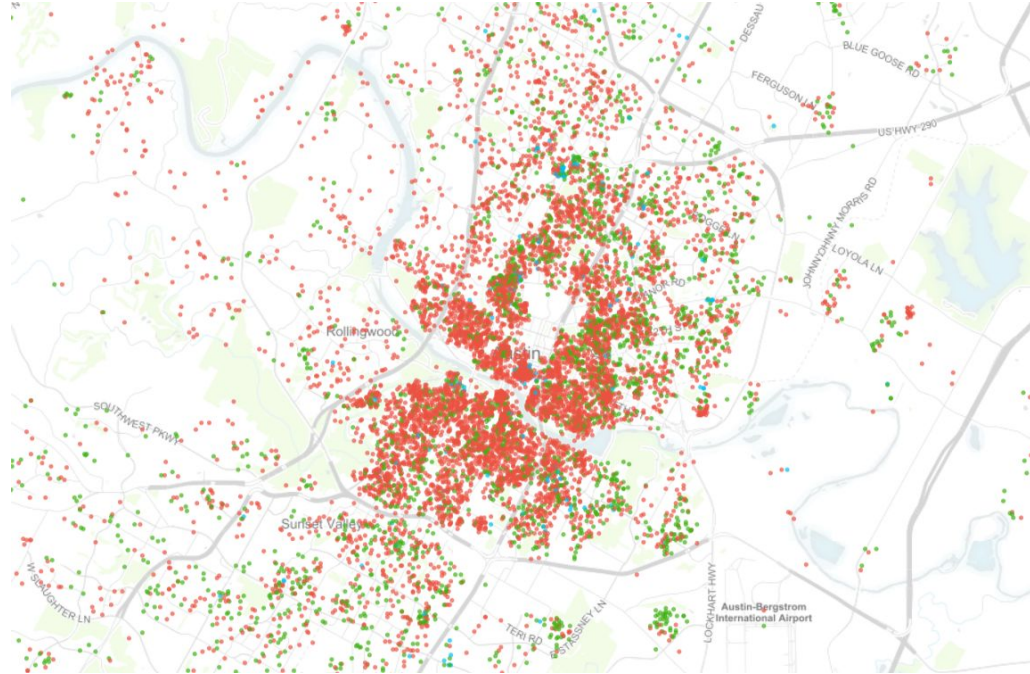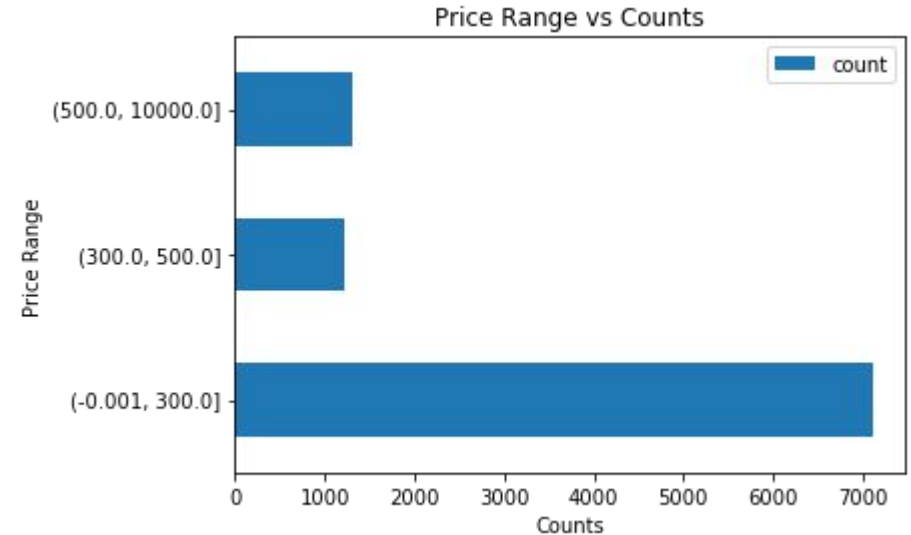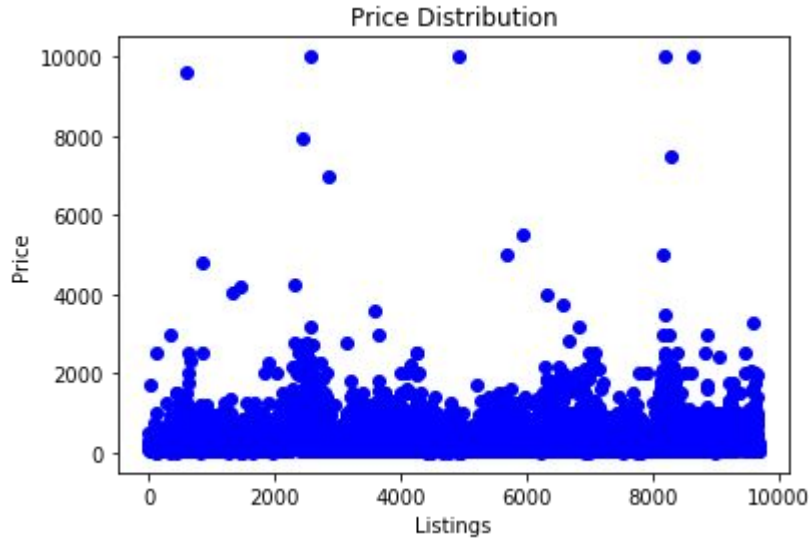
02

Exploratory Analysis

First Look: Features

- Superhost Status
- Host Verifications
- Host's Profile Picture
- Host's Identity
- Accurate Location
- Room Type
- Instant Bookable

- Bathrooms
- Bedrooms
- Amenities
- Security Deposit
- Number of Reviews
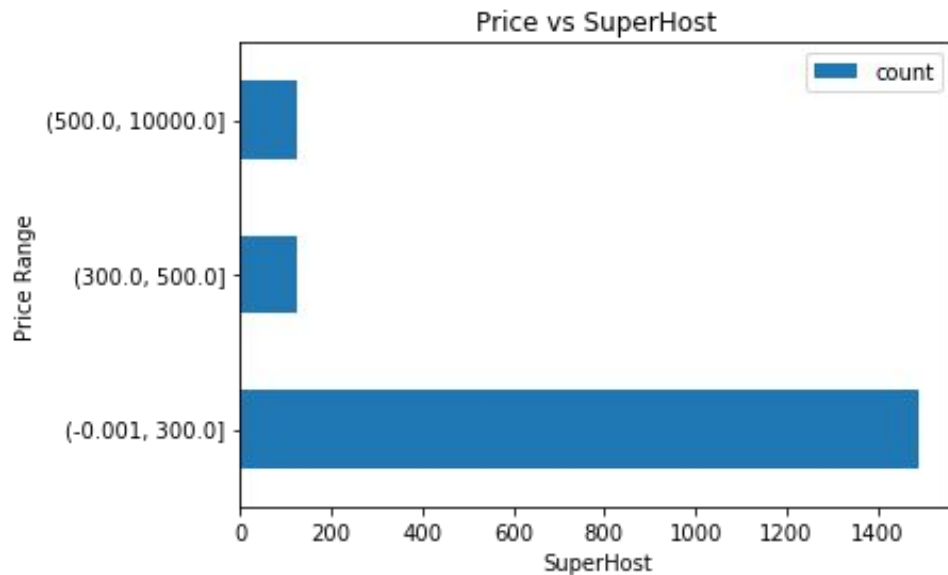- Review Scores
- Cancellation Policy

9663 Observations

Listings' Price Range

Nearly **70%** of the listings have prices lower than $300 Dollar Per night

Price vs Superhost

Price vs SuperHost

Price Range	%	count
0~300	85%	1492
300~500	7%	122
500~	8%	126

Total: 1740 listings are posted by Superhosts

Price vs Review Score

Price Listing	Review Score	count
Low	Low	56.8%
	High	**43.2%**
Medium	Low	45.3%
	High	**54.7%**
High	Low	40.5%
	High	**59.5%**

Higher listing price => **higher** percentage of high review score

Amenities

Amenities Frequency Chart
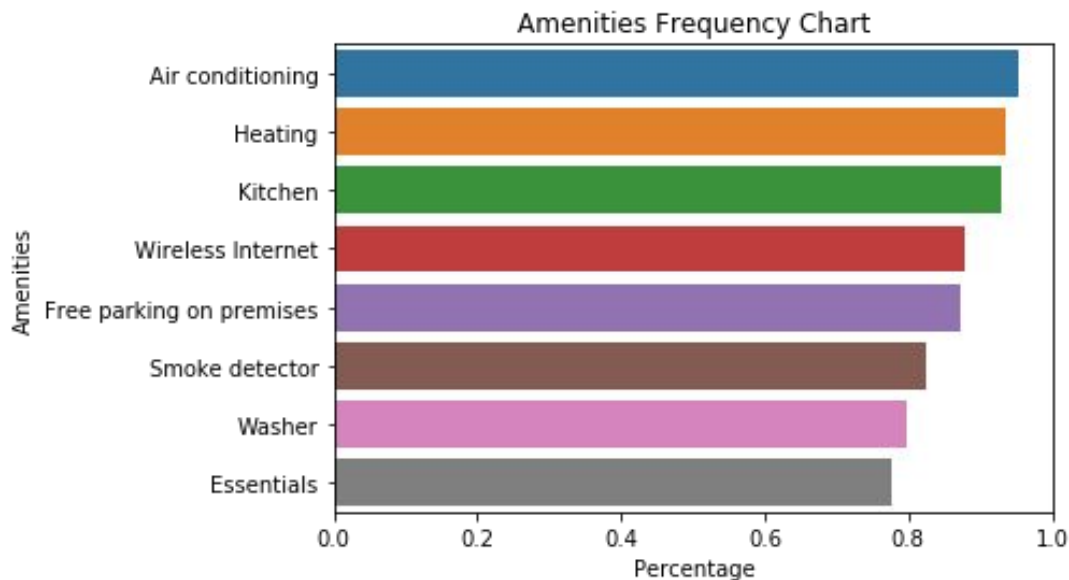
Rank	Amenities	Ratio
1	Air conditioning	0.95
2	Heating	0.93
3	Kitchen	0.93
4	Internet	0.88
5	Parking	0.87
6	Smoke Detector	0.82
7	Washer	0.80
8	Essentials	0.77

Exploratory Analysis: Review Score vs Bathroom
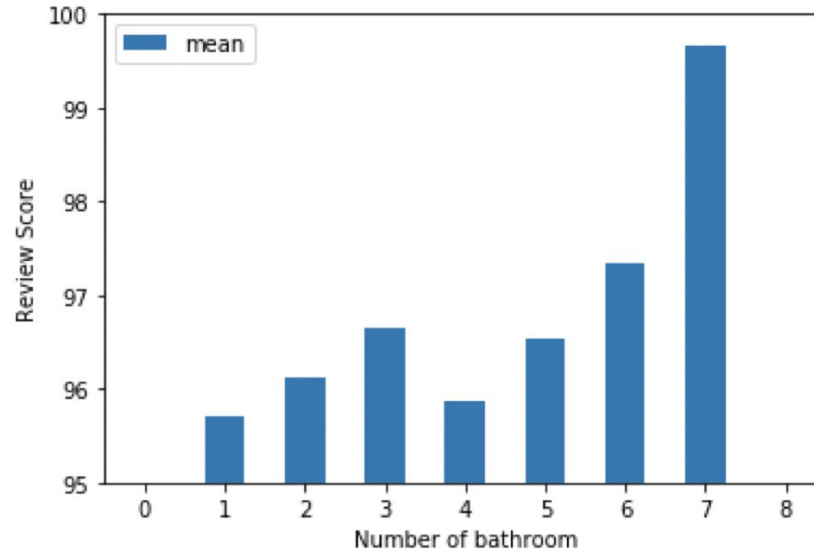
More bathrooms suggest **higher** review score

03
Solutions and Insights
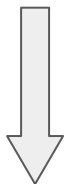
Classifiers

- Logistic Regression Model

- Tree Models

    - Decision Tree

    - Bagging

    - Random Forests

    - Gradient Boosting
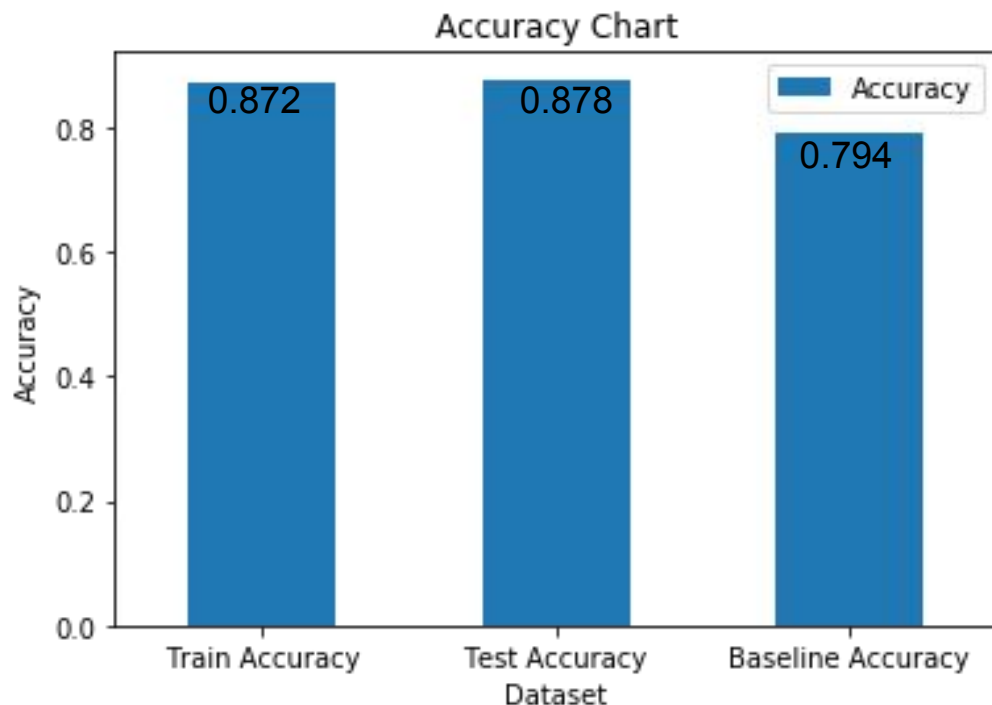
Summary - Logistic Regression

Positive Class: >$300

Negative Class: <=$300

Train Accuracy: 0.872
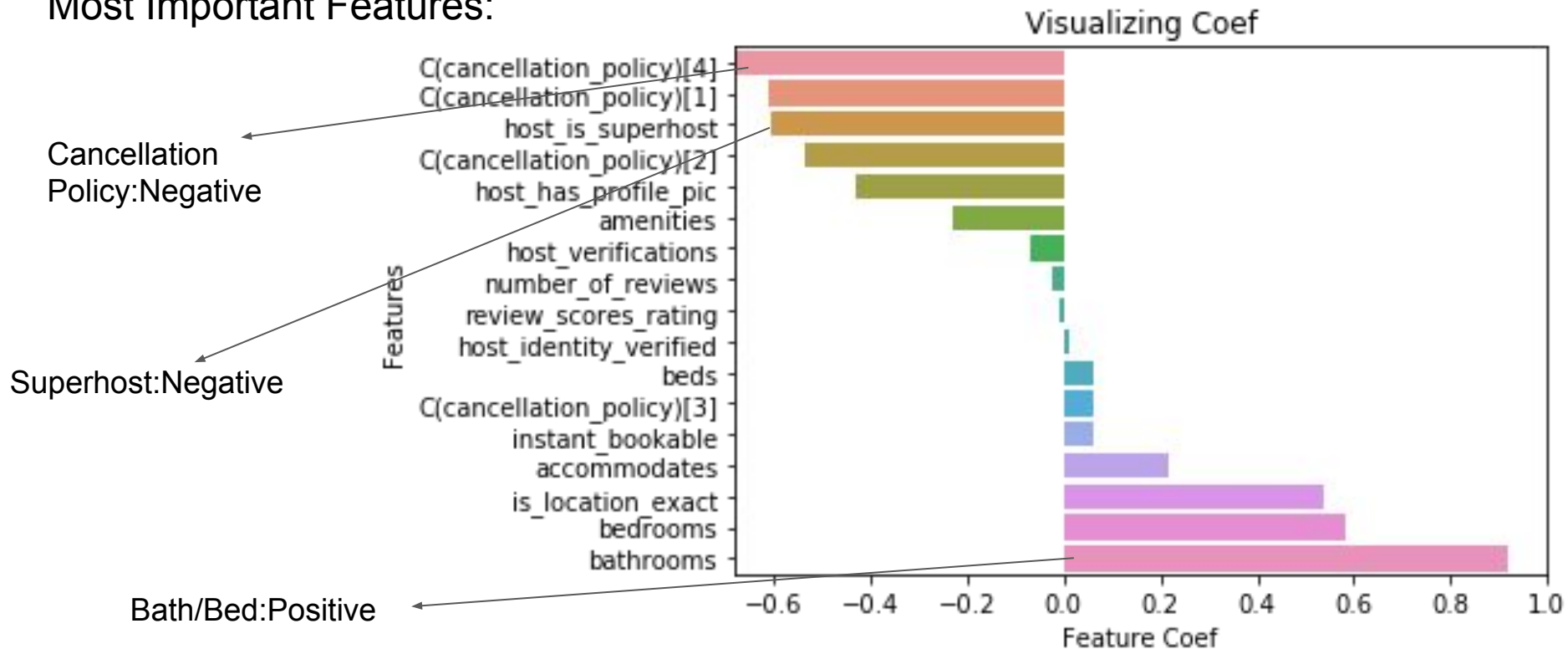
Test Accuracy:0.878

Baseline Accuracy = 0.794

Accuracy Chart

Analysis - Logistic Regression

Most Important Features:

Visualizing Coef

Cancellation Policy:Negative
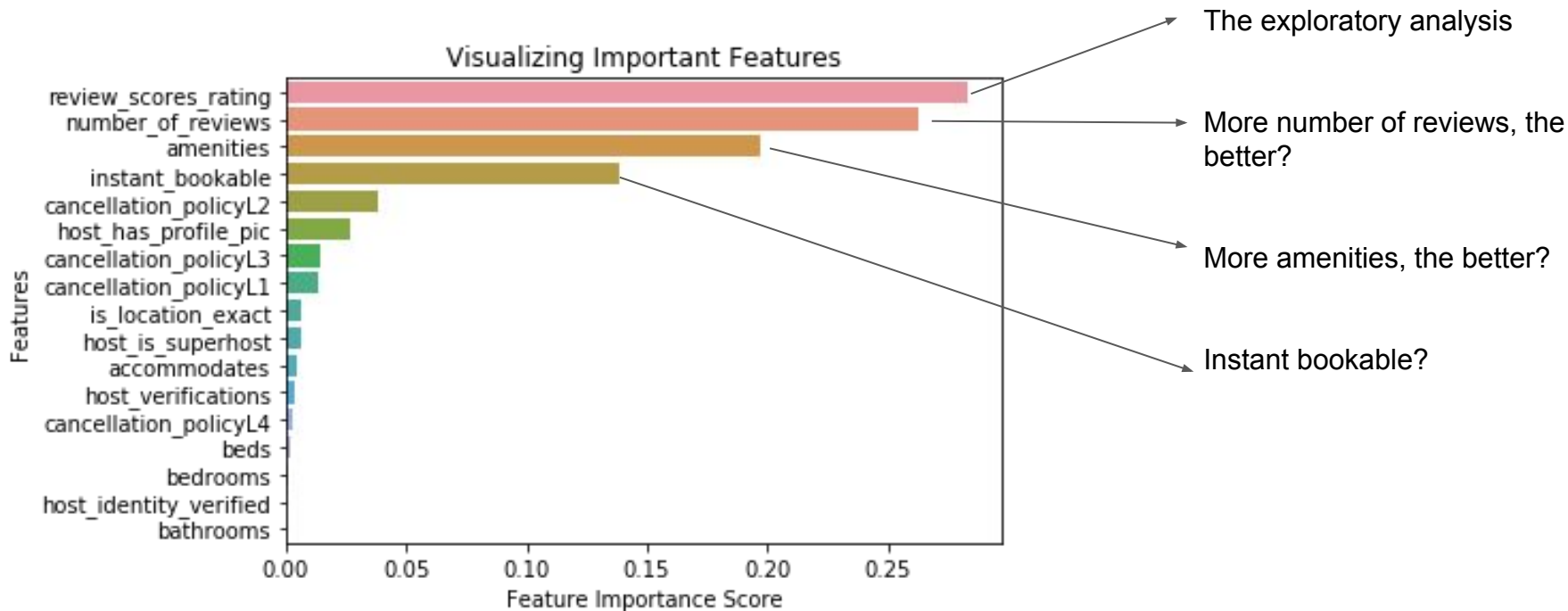
Superhost:Negative

Bath/Bed:Positive

Summary - Tree Models

- Split Listings' Price into three bins:

    - <= $200

    - $200 ~ $400

    - > $400

Random Forest **outperforms** all three other models with an accuracy of nearly 80%

Analysis - Random Forest

Visualizing Important Features

The exploratory analysis

More number of reviews, the better?

More amenities, the better?

Instant bookable?

Recommandation

- Improve Review score

- Increase the number and the quality of amenities

- Try to make the listing 'Instant bookable'

- Try to make the customer to write reviews after staying

- Try not to make the cancellation policy too strict

- Become the Superhost: Try to make same listings at a lower price point

04
Q&A

THAT IS AN EXCELLENT QUESTION

Thank You