Beijing City Feature-Based Airbnb Price Prediction

Team GeekHub

Brian Feng
Mike Liu
Tony Li
Logan Luo

Outline

Motivation

Problem Definition

State-of-the-Art

Methodology

Evaluation

Motivation

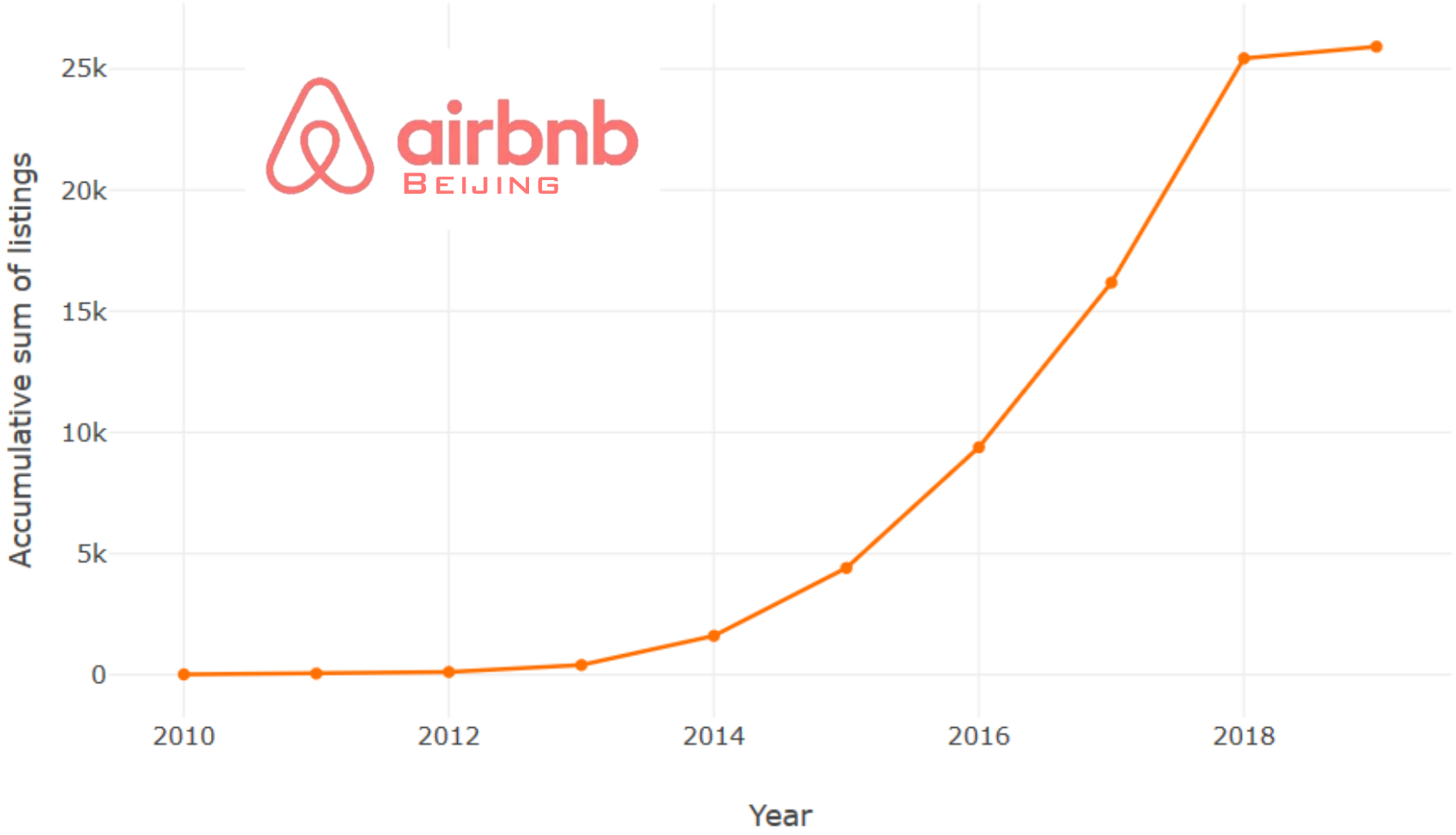


Figure 1 Accumulative number of listings year by year since 2010 in Airbnb Beijing

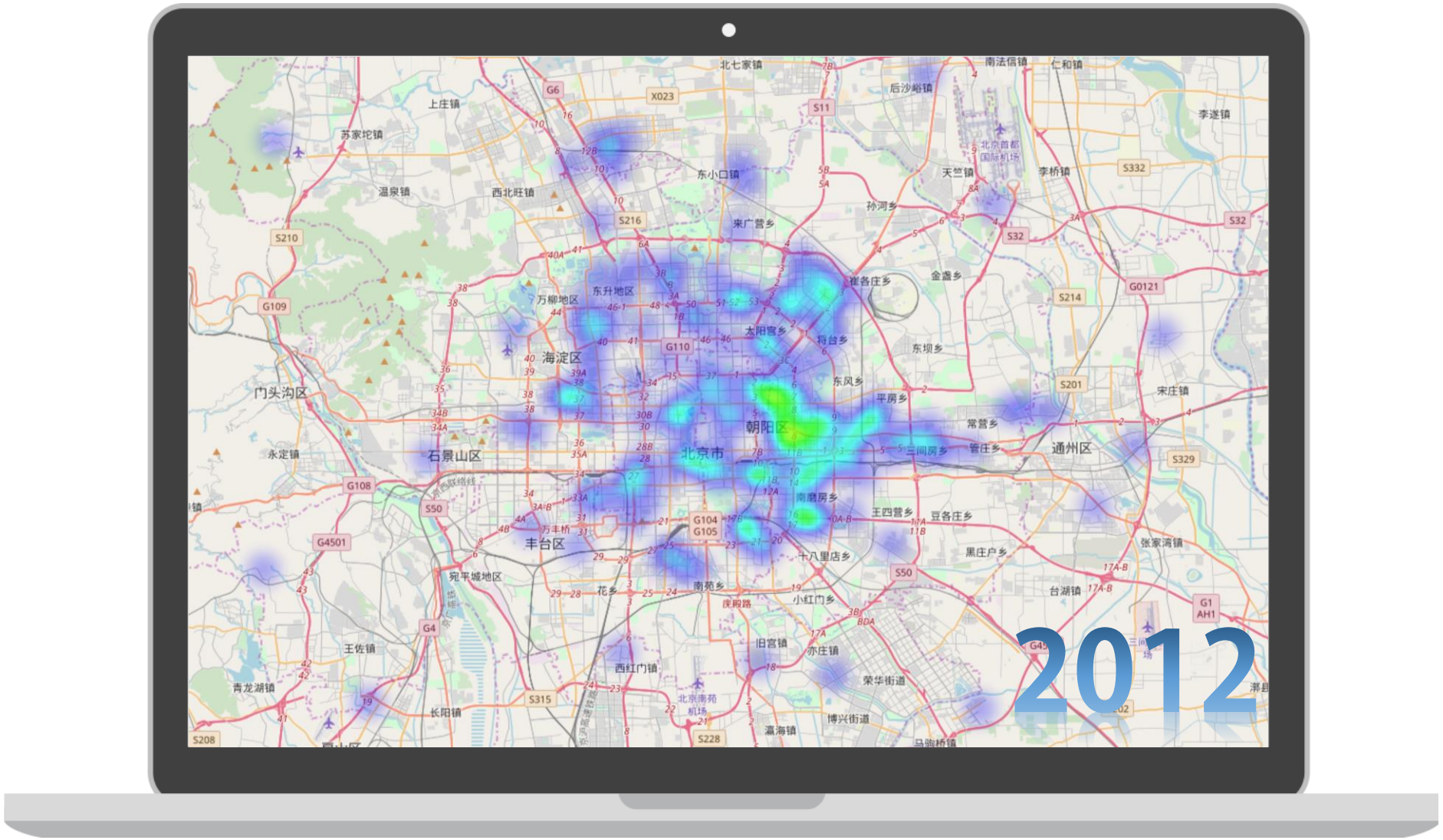
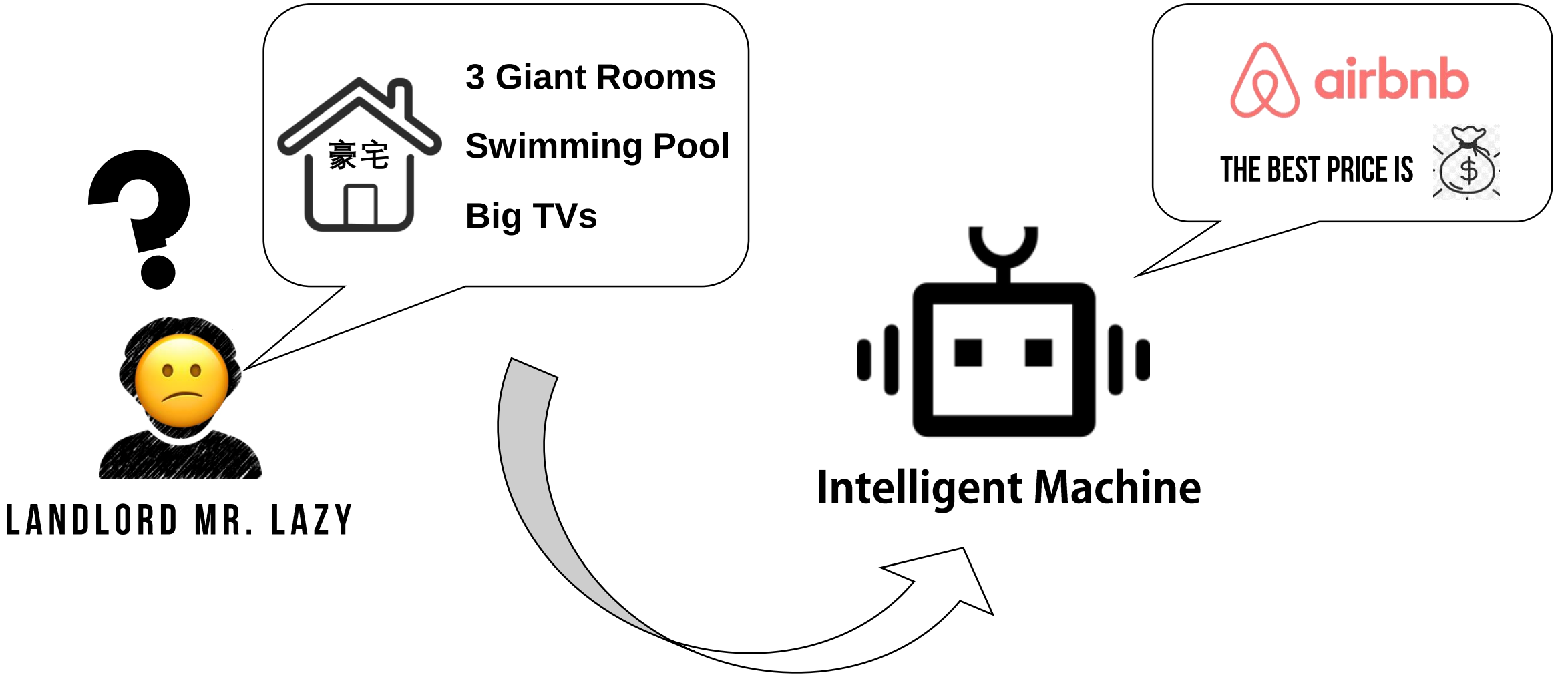


Figure 2 Accumulative distribution of listings in Year 2012

Figure 3 Accumulative distribution of listings in Year 2019


豪宅
3 Giant Rooms
Swimming Pool
Big TVs
LANDLORD MR. LAZY
airbnb
THE BEST PRICE IS
Intelligent Machine

Problem Definition

Problem Statement

Given a series of data describes the property's features

Output the reasonable/best price point for the host.

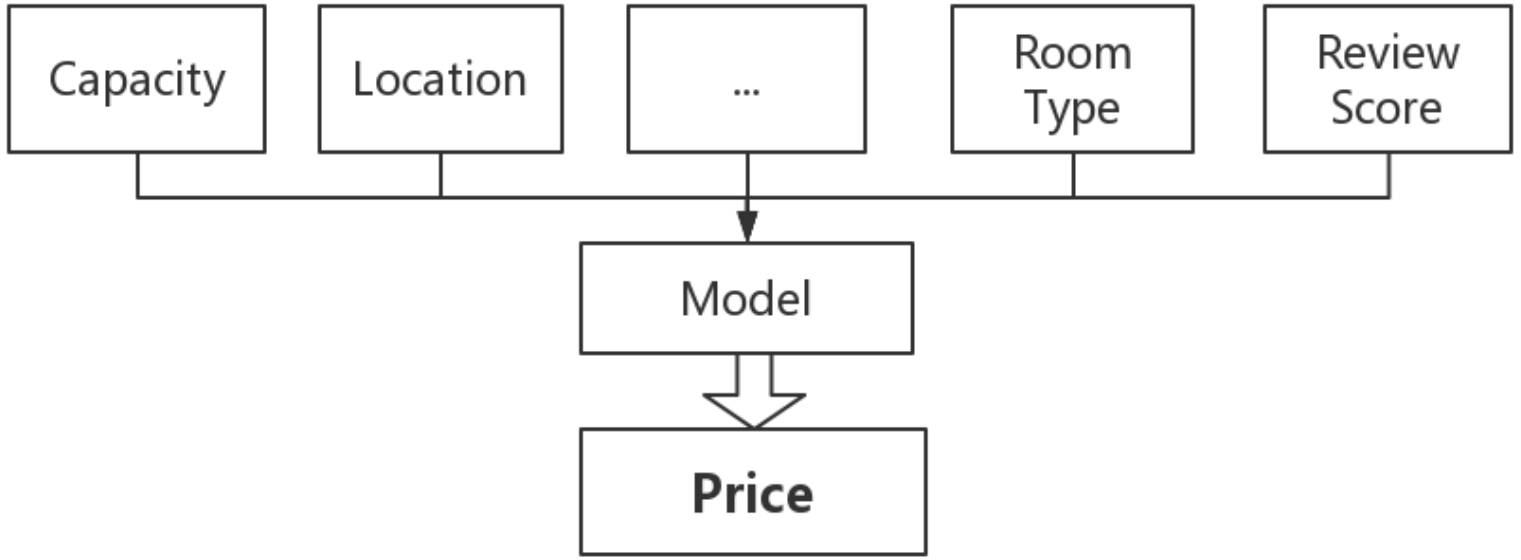


Figure 4 Problem Basic Solving Method

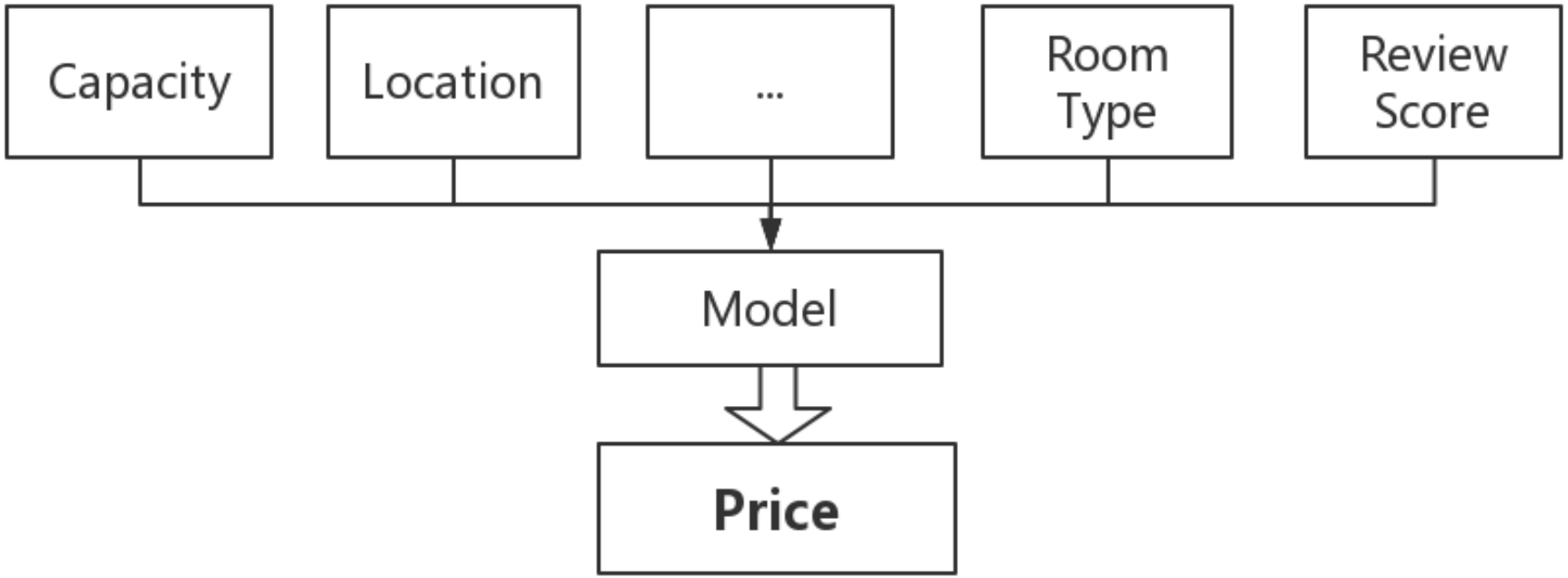


Figure 4 Problem Basic Solving Method

$$f(Capacity, Location, \dots, RoomType, ReviewScore) + \beta_0 \Rightarrow Price$$

State-of-the-Art

Table 1 A brief review of the literature on price influencing factors

Dimension	Factors	Effects	Author
External	Price in cities	N/A	Stephen and Andrew(2016)
	Number of owners Distance between owners	Pos/Neg	Balaguer and Pernias(2013), Becerra(2013)
Location	Distance	Negative	Bull(1994), Wang & Nicolau(2017), Li (2016), Huang(2010)
House	Room type	Negative	Wang & Nicolau(2017), Zhang(2017)
	Room type	Positive	
Room	Bedroom, bathroom, bed	Positive	
Rule	Instant bookable	Negative	
	Cancellation policy	Positive	
Sociality	Number of reviews per year	Negative	Wang & Nicolau(2017), Zhang(2017)
		Positive	Zhu Lv(2017), Gutt & Herrmann(2015)
	Host identity verified	Pos/Neg	Wang & Nicolau(2017)
		Positive	Zhu Lv(2017)

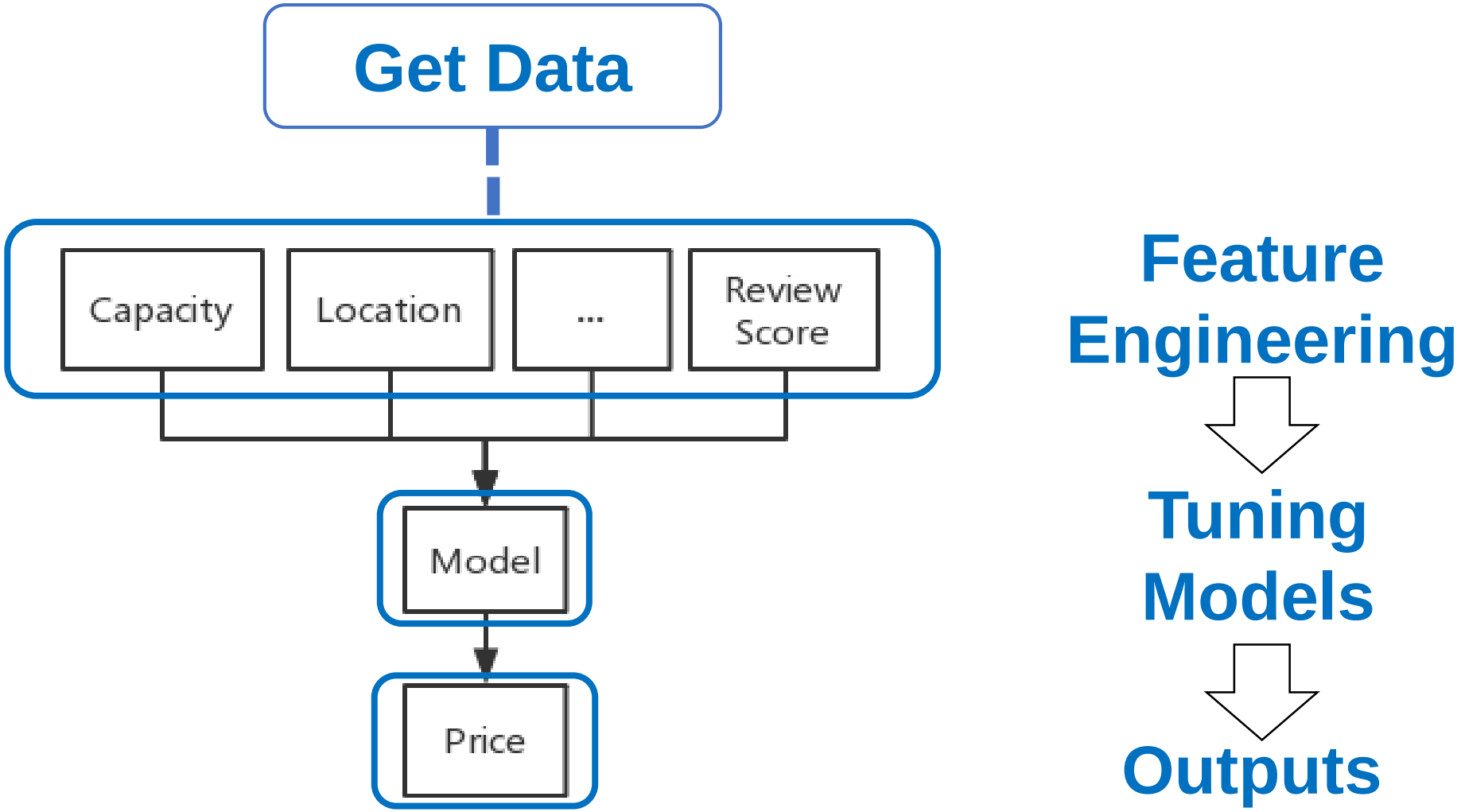


Figure 5 Problem Basic Solving Process

Methodology

Get Data

Raw_Data-Airbnb-Beijing.csv

Total:
19360 Listings

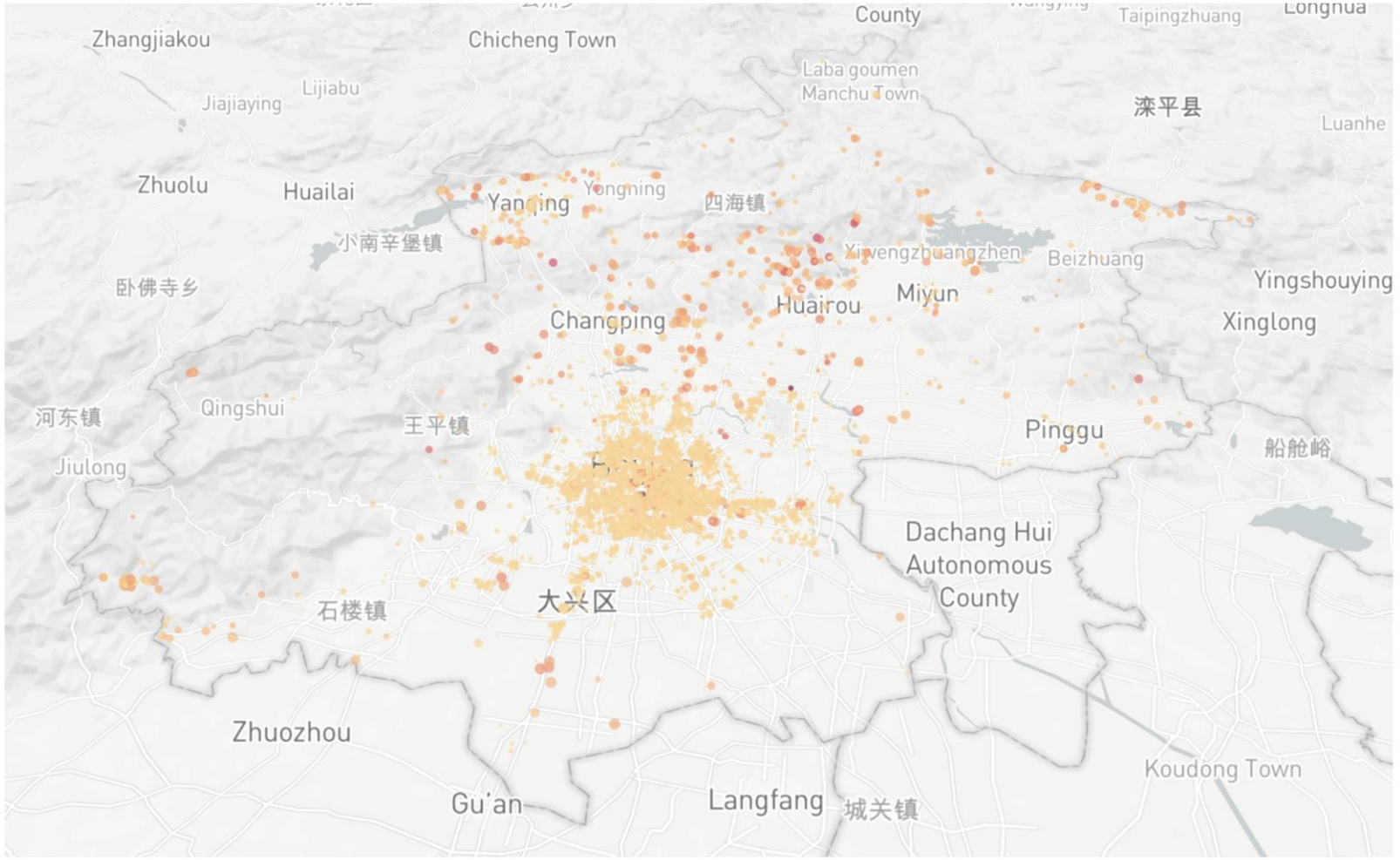


Figure 6 All available Data in Airbnb

Data Pre-Processing

- Drop Properties that lacks too many features (less than 100)
- Drop outliers handpicked via inspection (i.e. Check the Actual ID on www.airbnb.com)
- Drop Properties with unreasonable prices (i.e. the host didn't actually want to Airbnb it)

~~[Price < (50 * accommodates) OR (50 * beds)]~~

~~[Price > 20,000 OR Price = 9999]~~

~~[beds=50 OR bathrooms=101 OR min_nights > 3]~~

- Fill up a few fixable nulls.

Airbnb-listings-Beijing.csv

Table 2 Pre-processed Data

ID	Capacity	Room types	Price	…
44054	9	Entire home	816	…
100213	2	Entire home	1203	…
128496	3	Private room	401	…
161902	2	Private room	387	…
162144	4	Private room	553	…
279078	2	Entire home	401	…
282825	4	Entire home	657	…
287026	3	Entire home	415	…
287511	3	Entire home	415	…
317195	2	Private room	546	…
322292	2	Entire home	436	…
…	…	…	…	…

19360 Listings → **17224 Listings**

Airbnb-listings-Beijing.csv

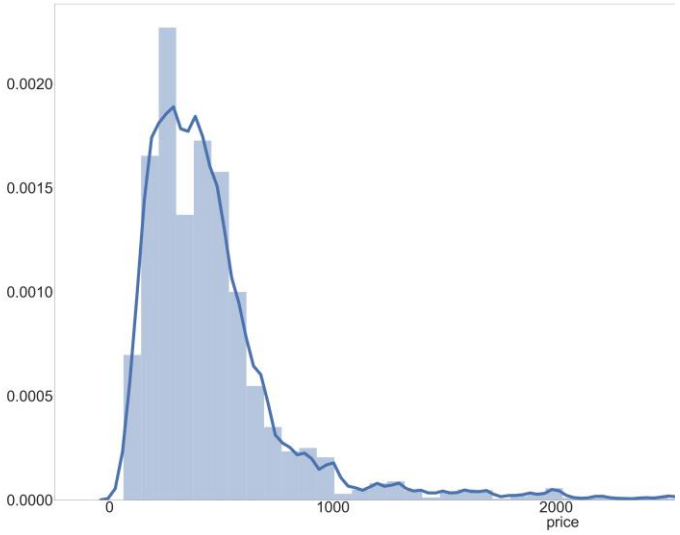


Table 2 Pre-processed Data

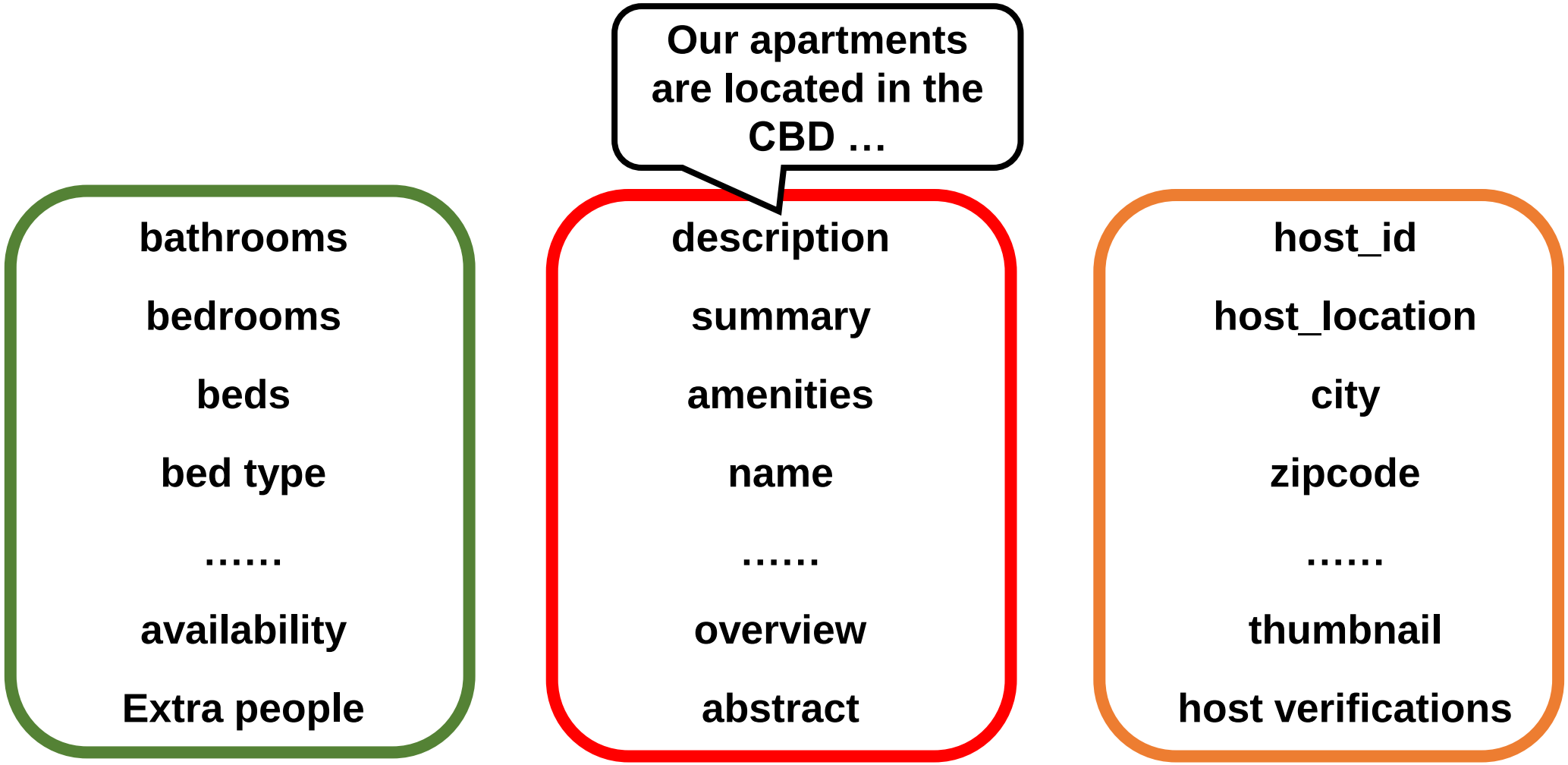
ID	Capacity	Room types	Price	…
44054	9	Entire home	816	…
100213	2	Entire home	1203	…
128496	3	Private room	401	…
161902	2	Private room	387	…
162144	4	Private room	553	…
279078	2	Entire home	401	…
282825	4	Entire home	657	…
287026	3	Entire home	415	…
287511	3	Entire home	415	…
317195	2	Private room	546	…
322292	2	Entire home	436	…
…	…	…	…	…

Total: 17224 Listings

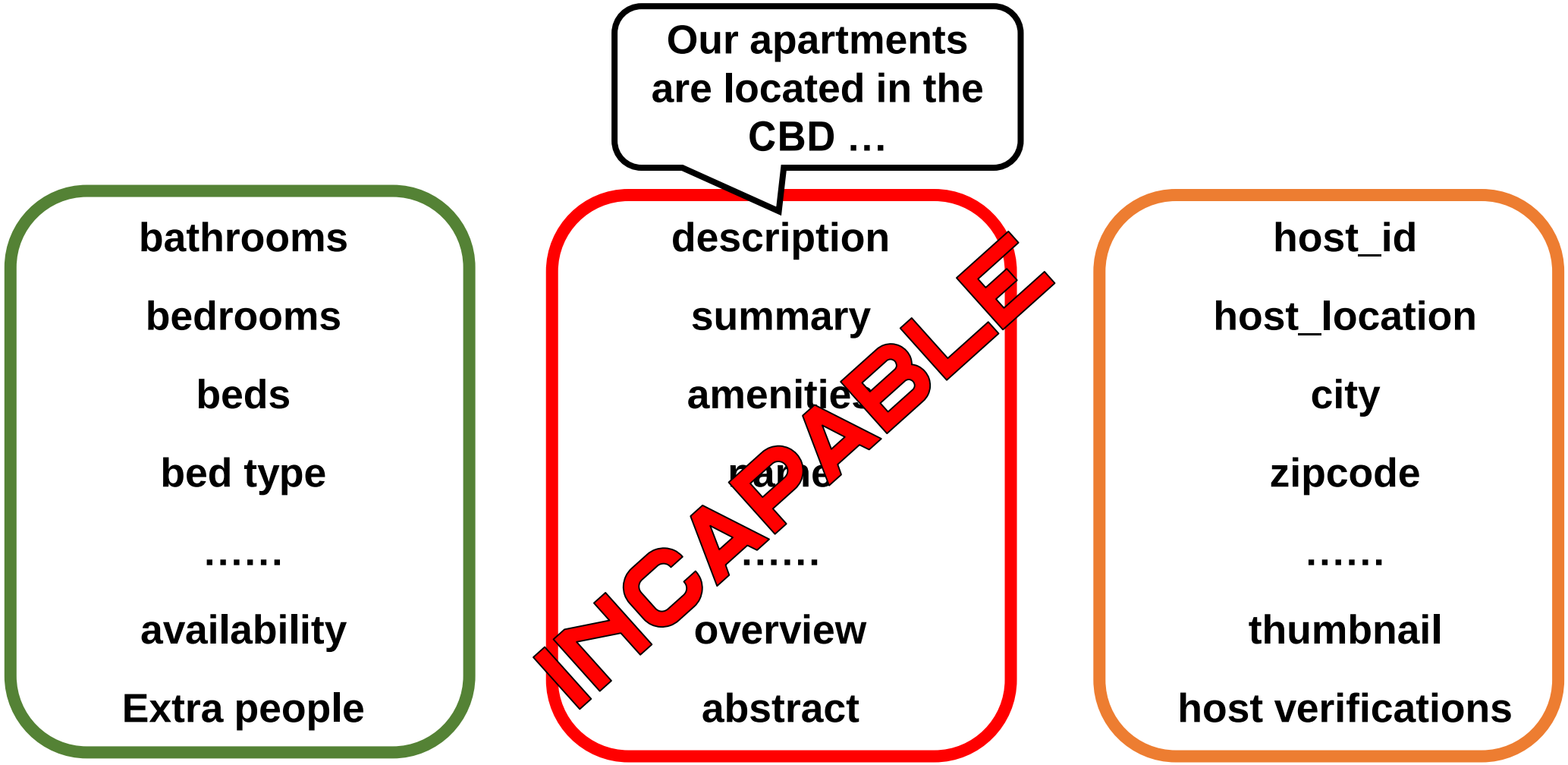
Feature Engineering

bathrooms	description	host_id
bedrooms	summary	host_location
beds	amenities	city
bed type	name	zipcode
……	……	……
availability	overview	thumbnail
Extra people	abstract	host verifications

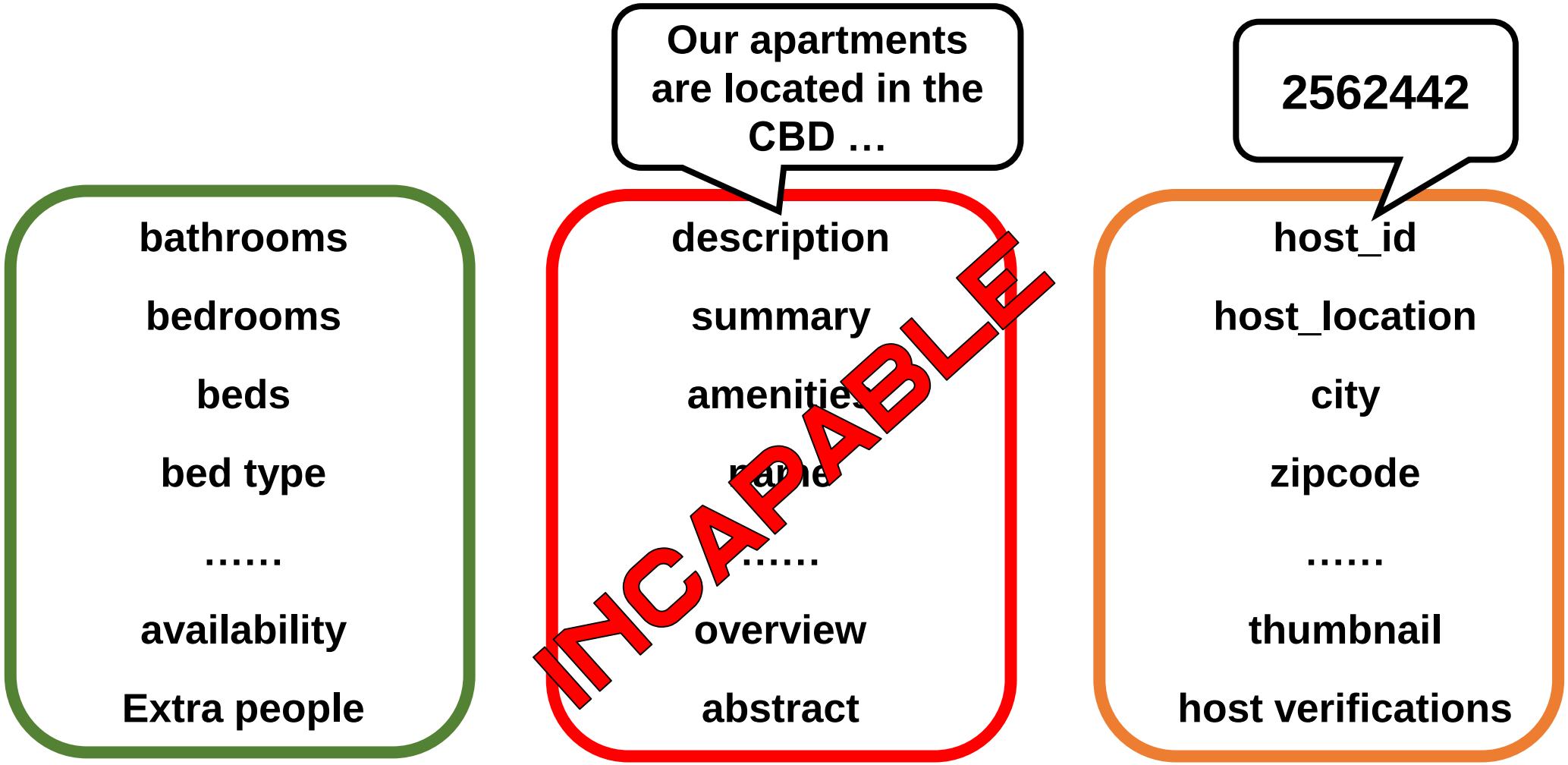
Feature Engineering



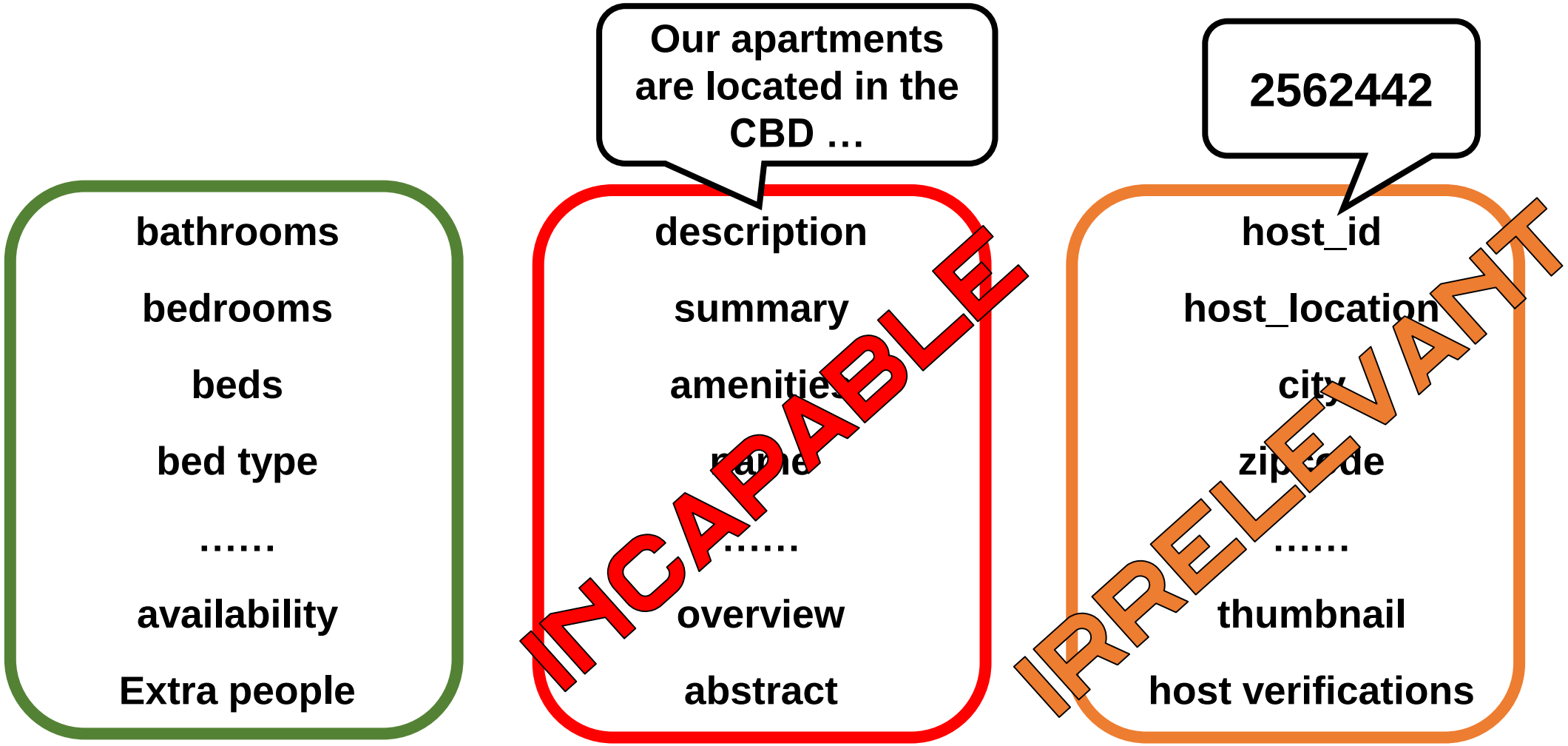
Feature Engineering



Feature Engineering



Feature Engineering



Feature Engineering

Figure 7 Hot-zone Near Beijing Railway Station

Feature Engineering

bathrooms

bedrooms

beds

bed type

……

availability

Extra people

+

Train Station

Airport

Wangjing SOHO

Yanqing IHE

……

Sanlitun

Subway

+

Feature
A / B / C

A	B	C
0	0	0
1	1	1

=

79

Features in total

Adding Features

One-Hot Encoding

accommodates

bathrooms

bedrooms

beds

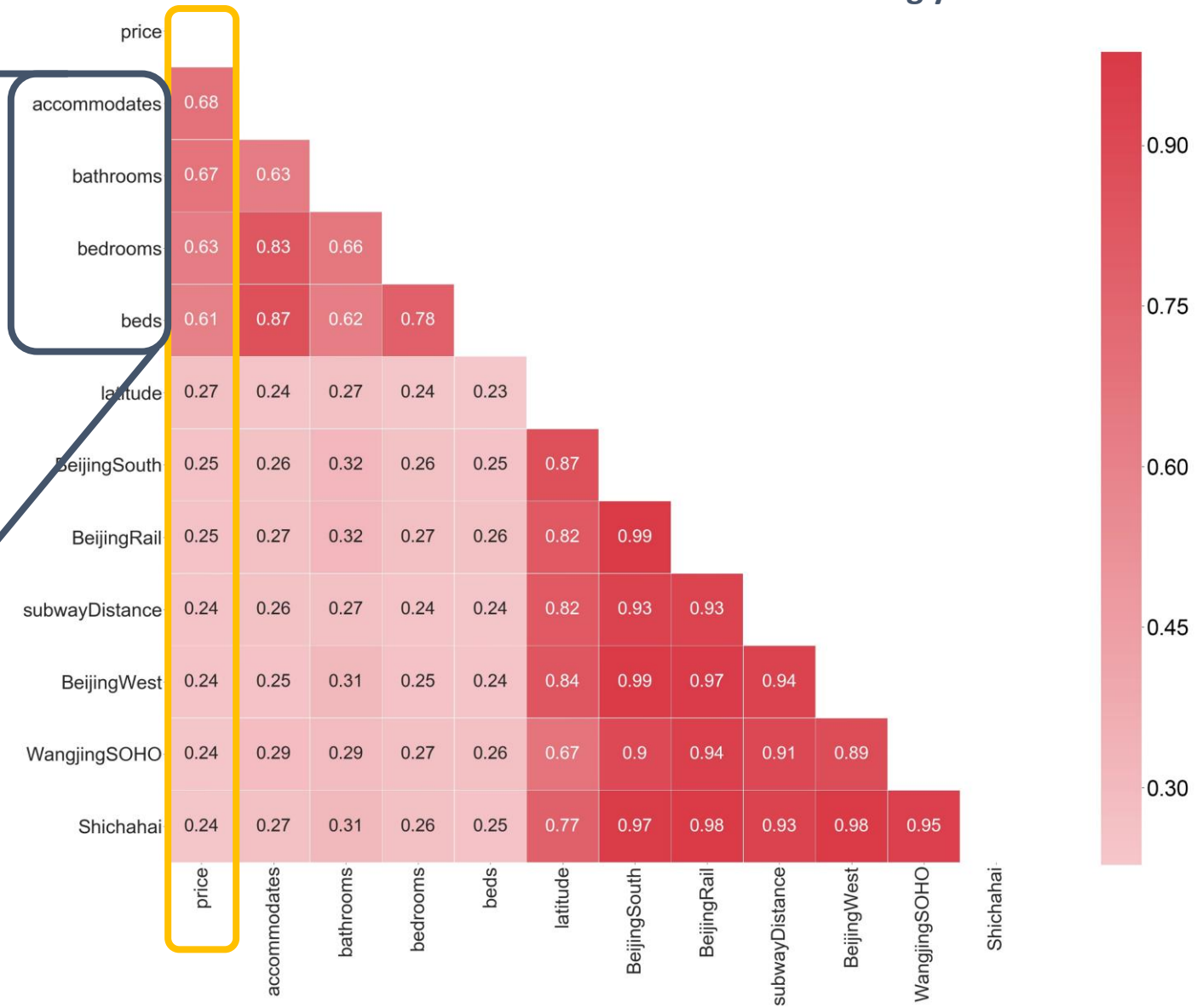


Figure 8 Correlation Coefficient Heatmap

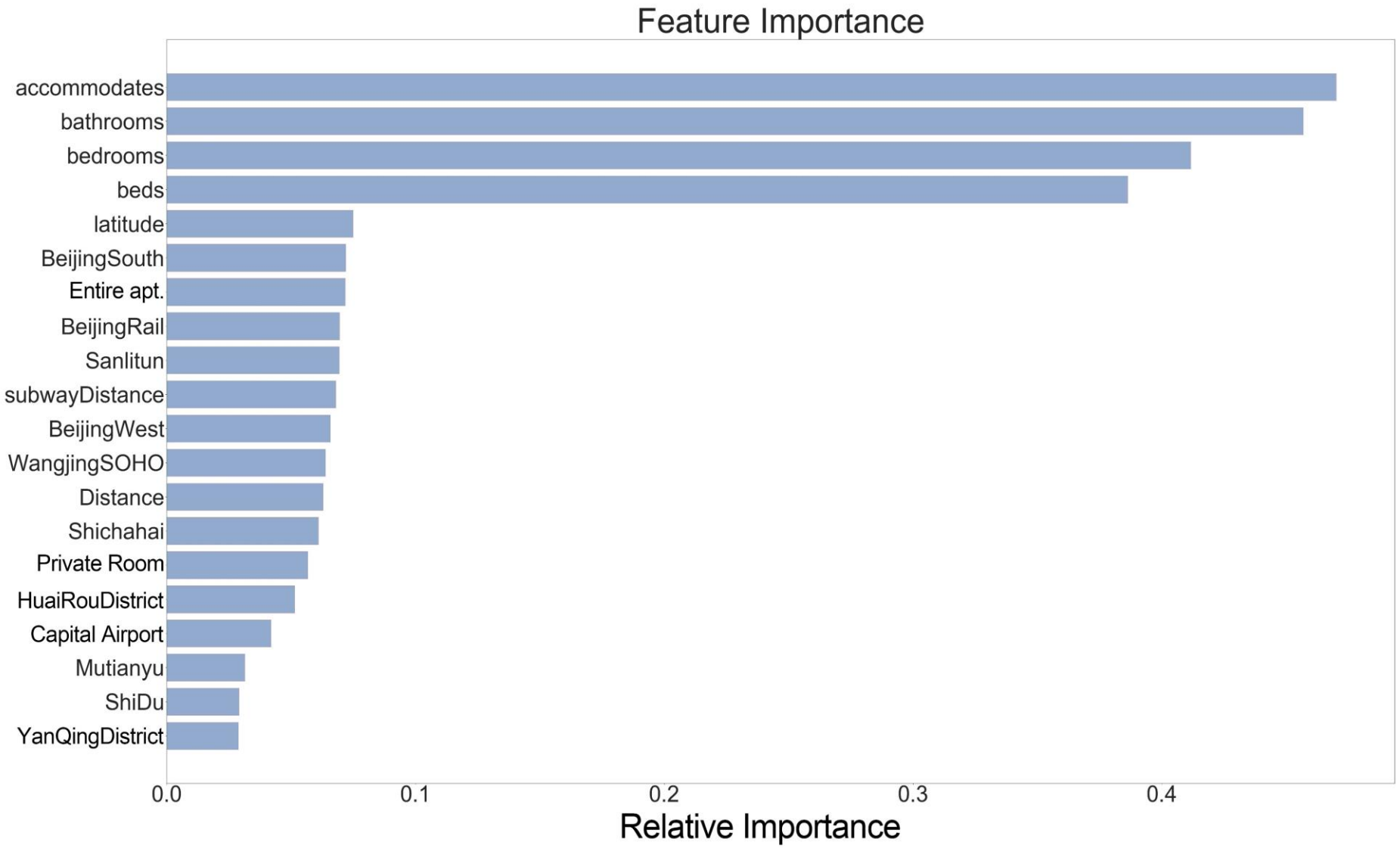


Figure 9 Feature Importance comparison

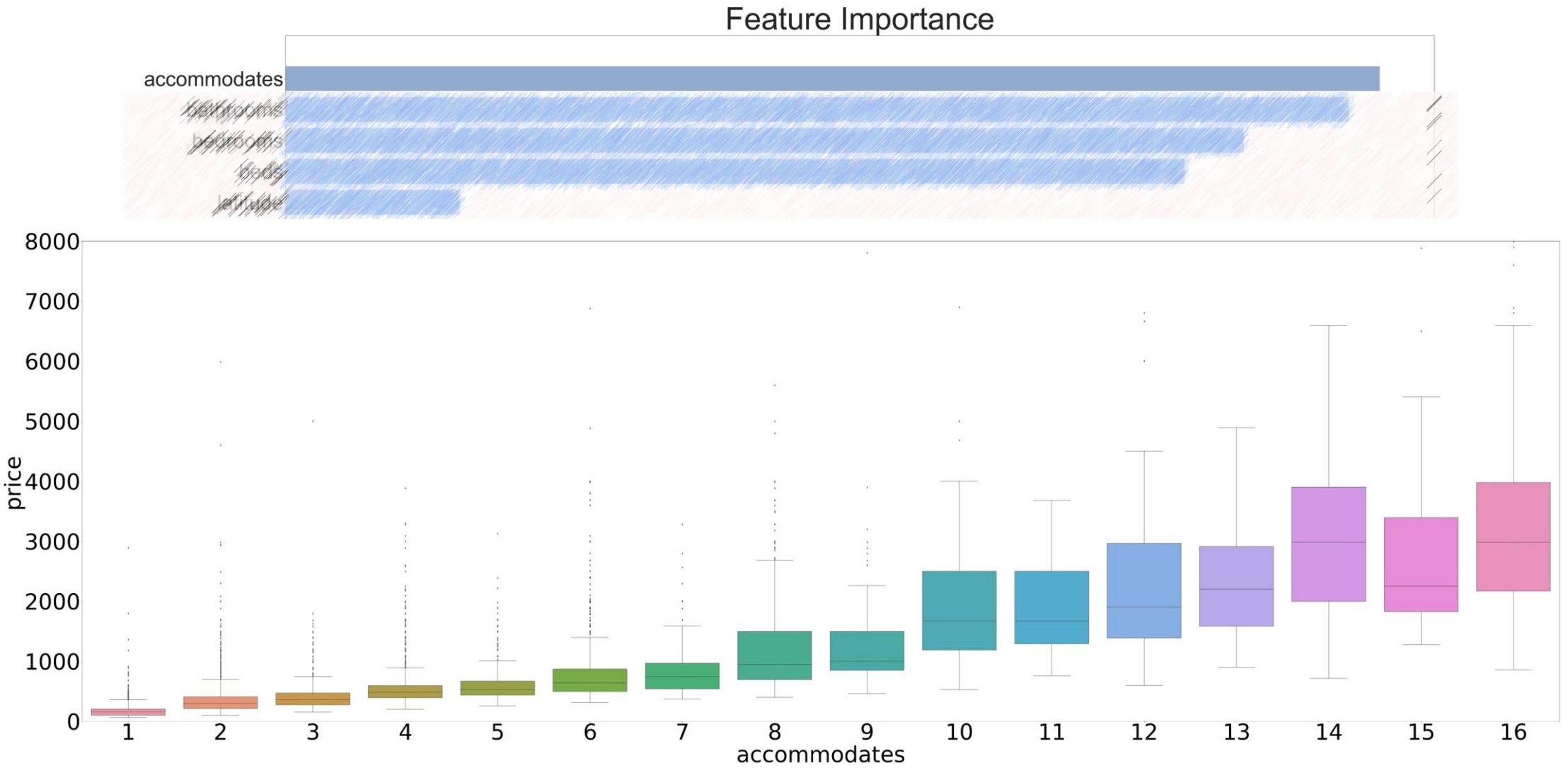


Figure 10 Accommodates-Price Box-plot

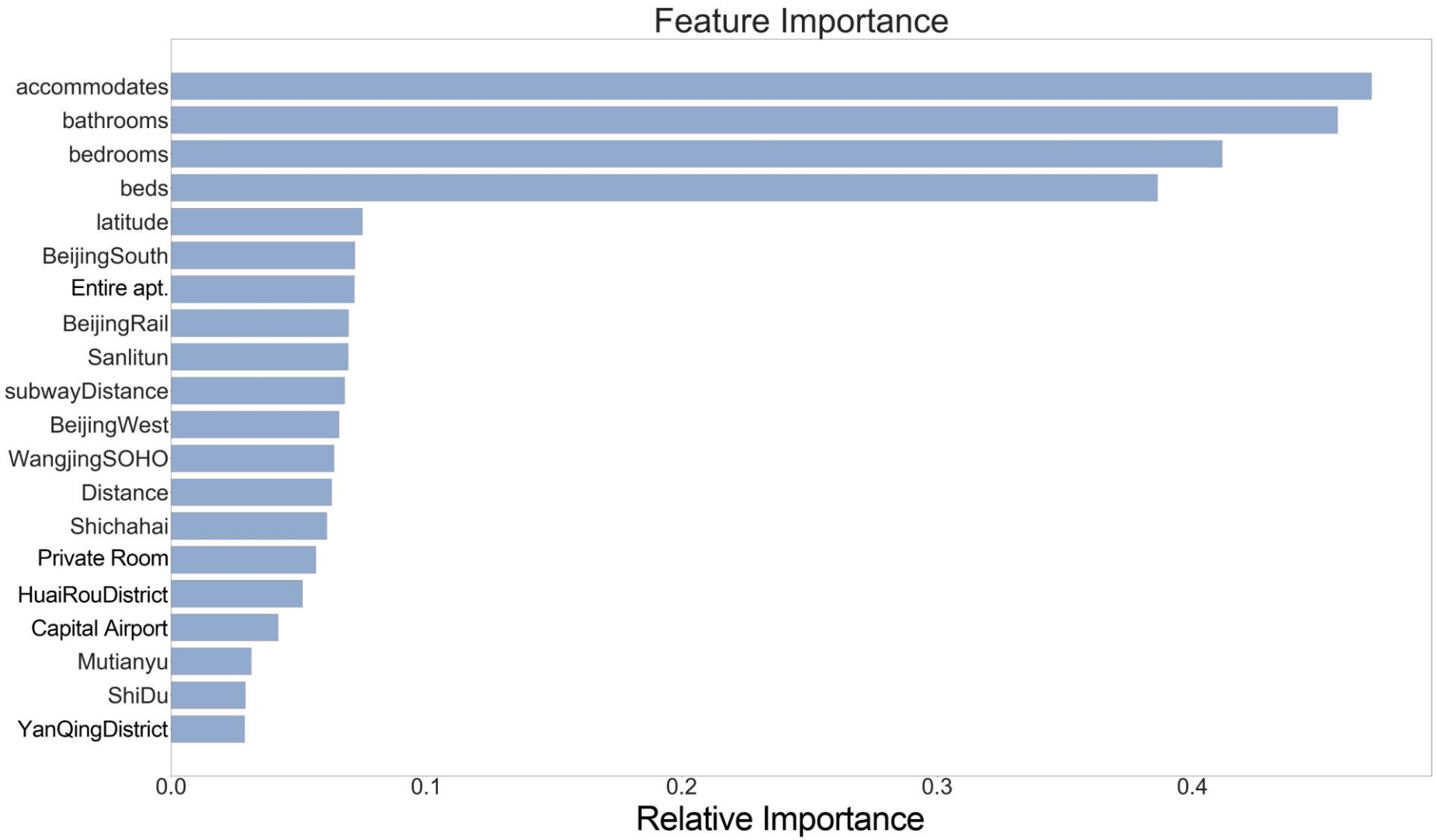


Figure 9 Feature Importance comparison

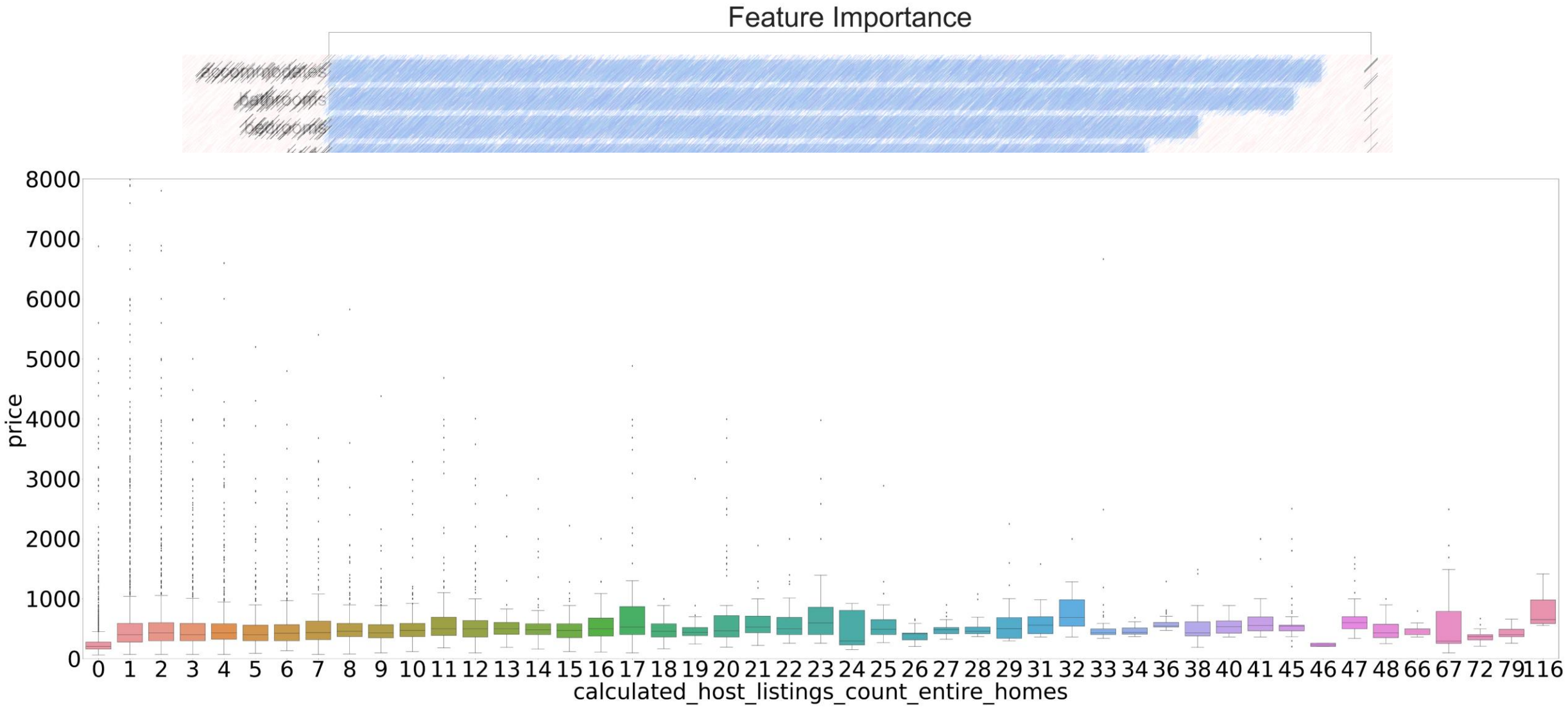
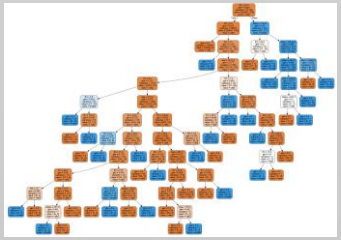
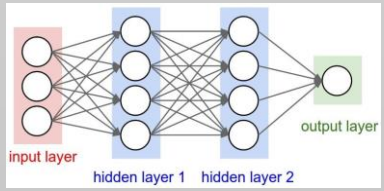


Figure 11 Host_listings-Price Box-plot

Model Building

Model





Splitted Data

Train Dataset
80 %

Test Data set
20%

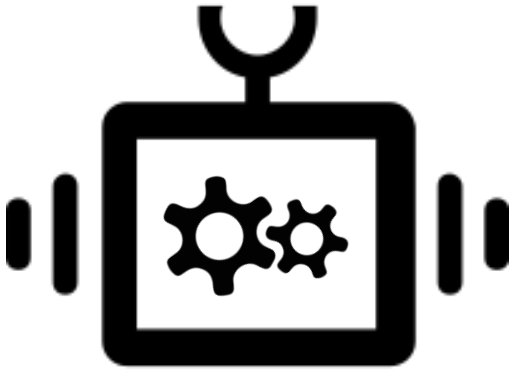
Cross Validation
Grid Search

PIPELINE

Scores

Visuals

Evaluation



Model

Train Dataset
80 %

Test Dataset
20%

Table 3 Model Results

Model Name	R^2 Score
Random Forest	0.705
Artificial Neural Network	0.712
Gradient Boosting	0.745
eXtreme Gradient Boosting	0.746

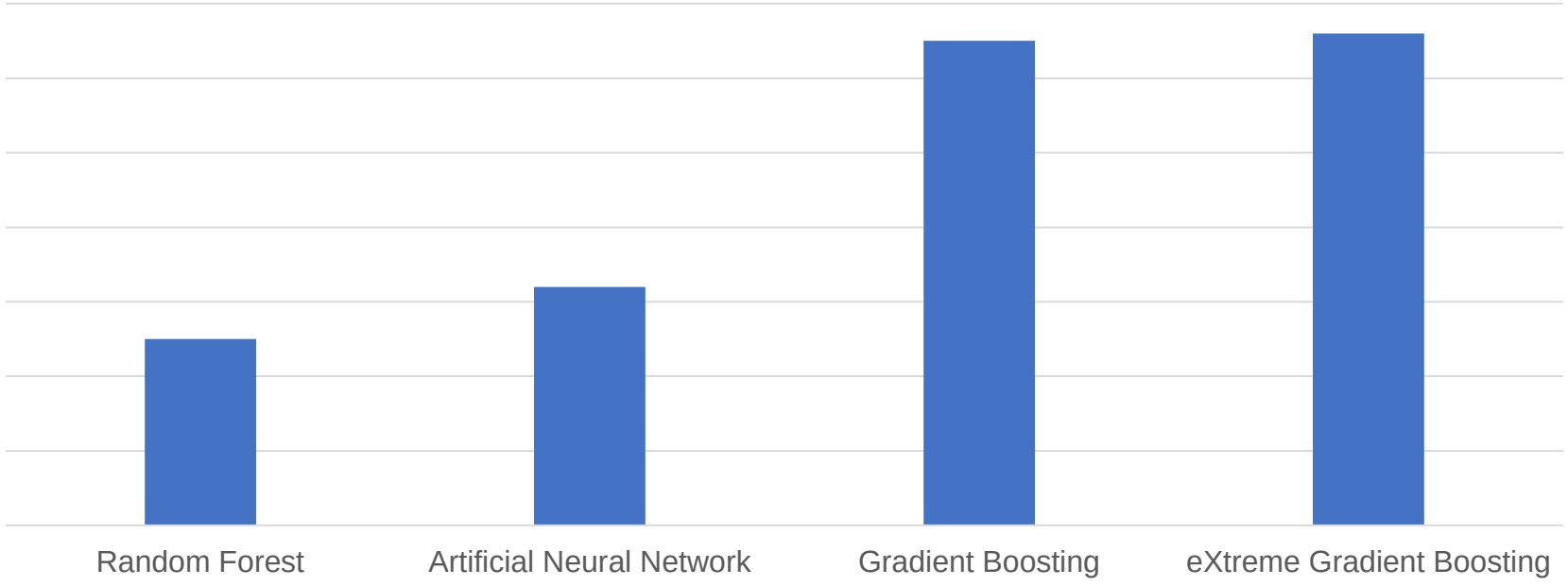
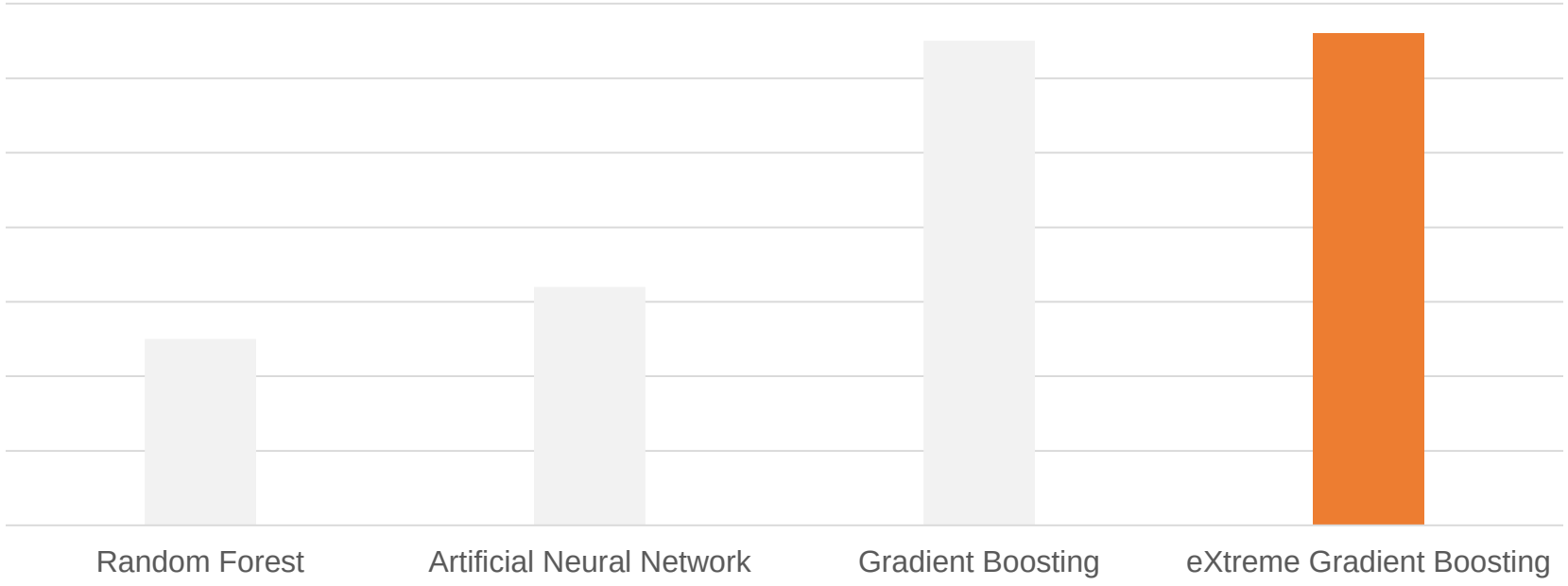


Table 3 Model Results

Model Name	R^2 Score
Random Forest	0.705
Artificial Neural Network	0.712
Gradient Boosting	0.745
eXtreme Gradient Boosting	**0.746**



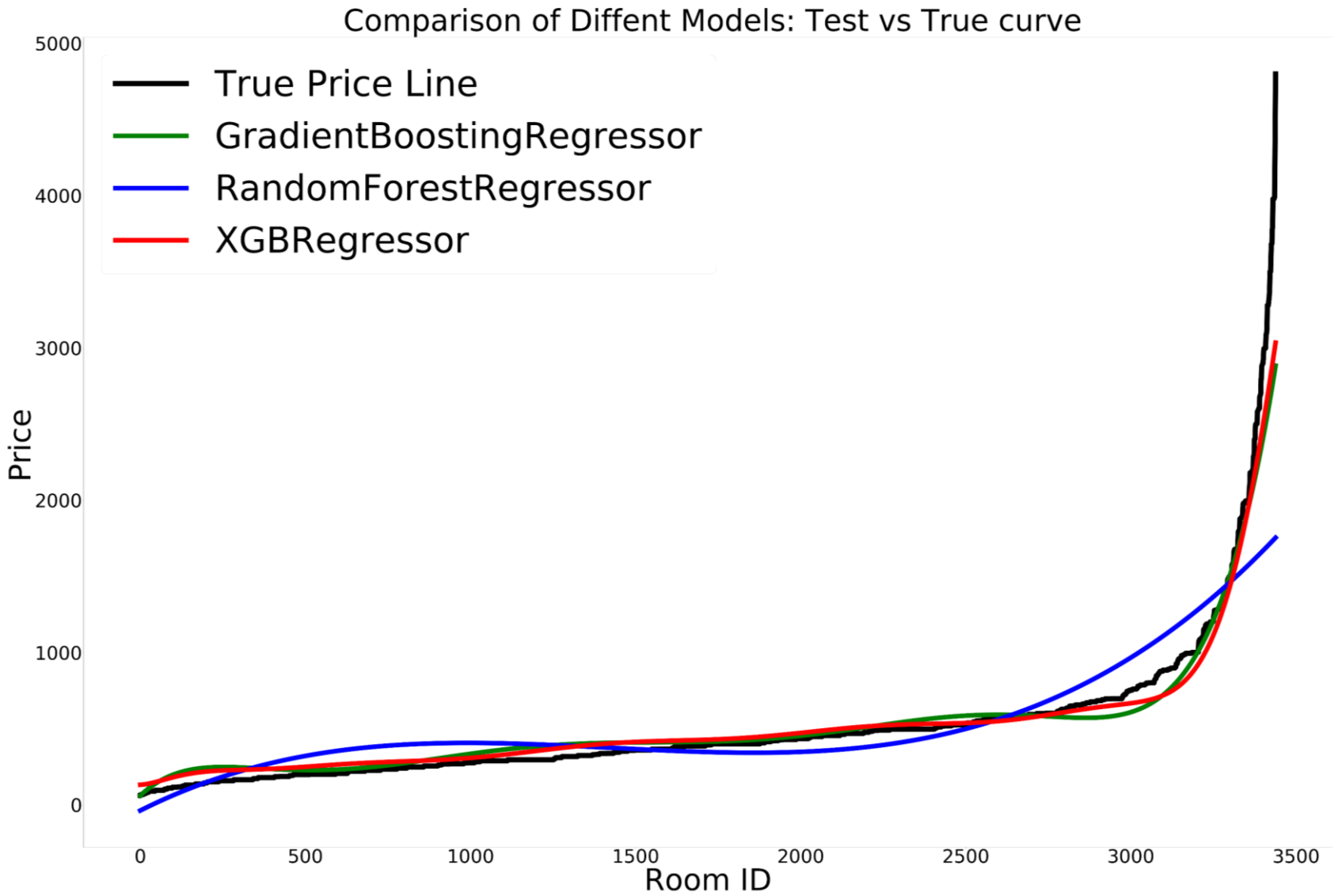


Figure 12 Comparison of Diffent Models

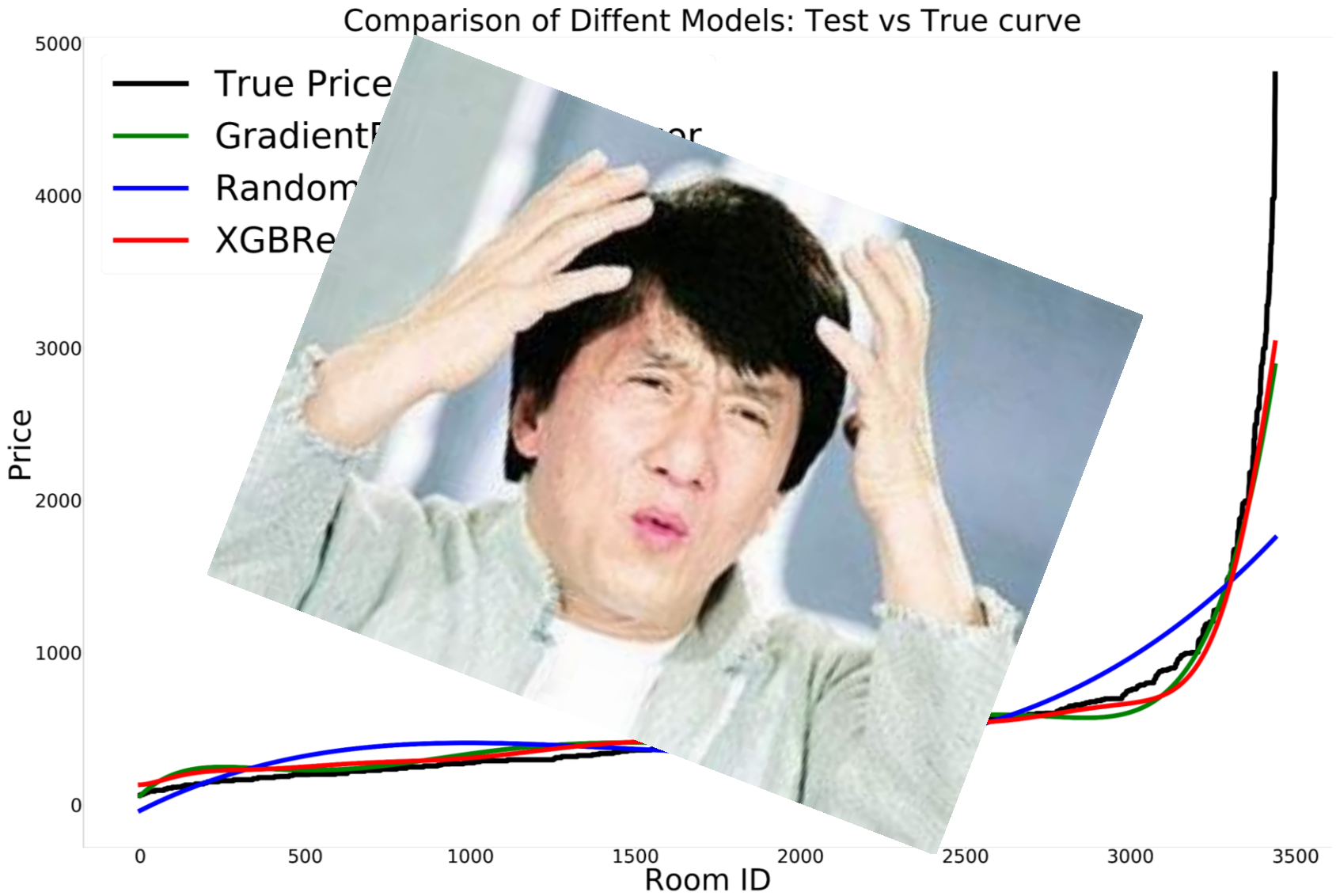


Figure 12 Comparison of Diffent Models

Table 4 Actual Performance

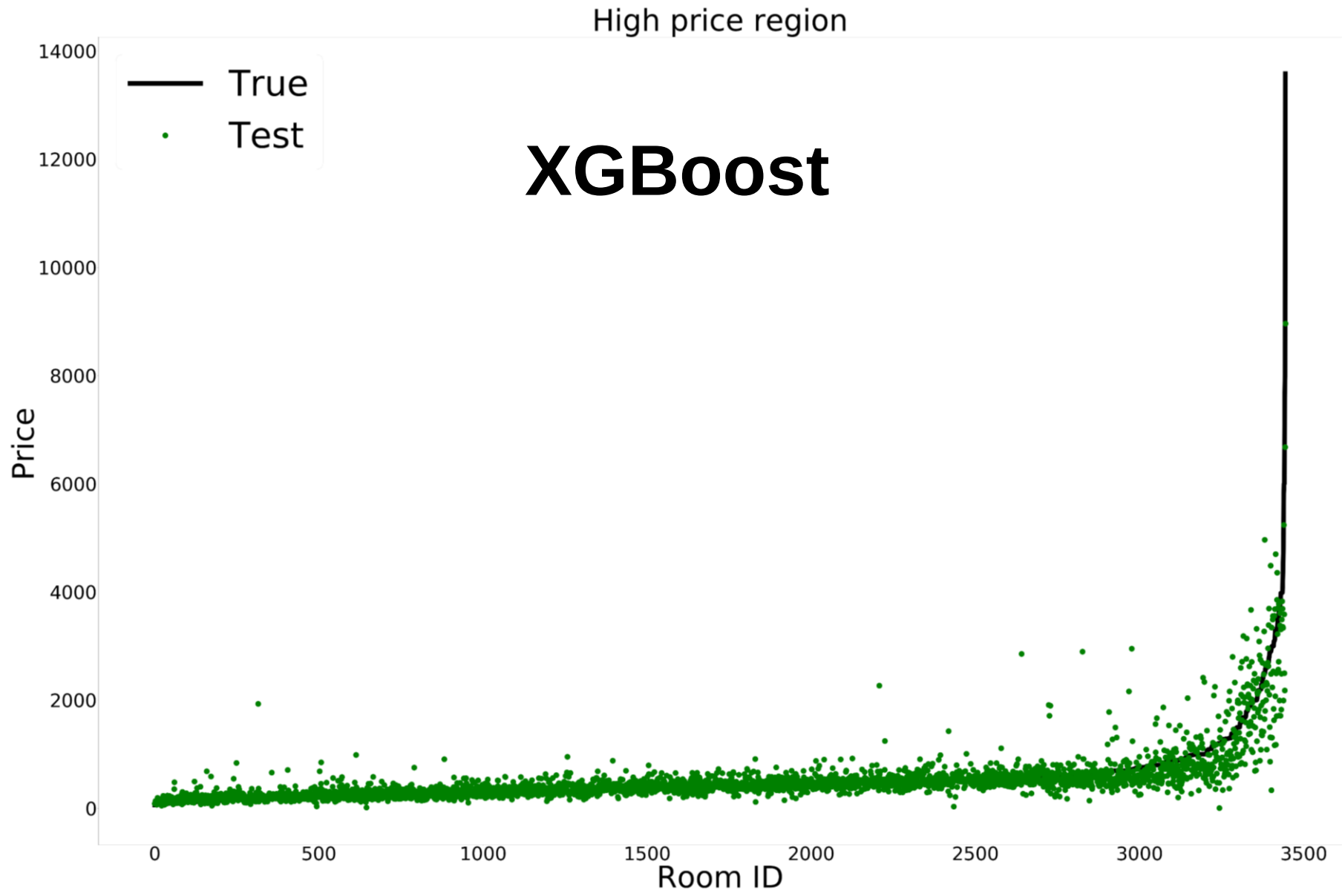
Model Name	R^2 Score	Prediction
Random Forest	0.705	¥ 1798.06
Artificial Neural Network	0.712	¥ 1820.95
Gradient Boosting	0.745	¥ 1882.24
eXtreme Gradient Boosting	**0.746**	**¥ 1887.61**

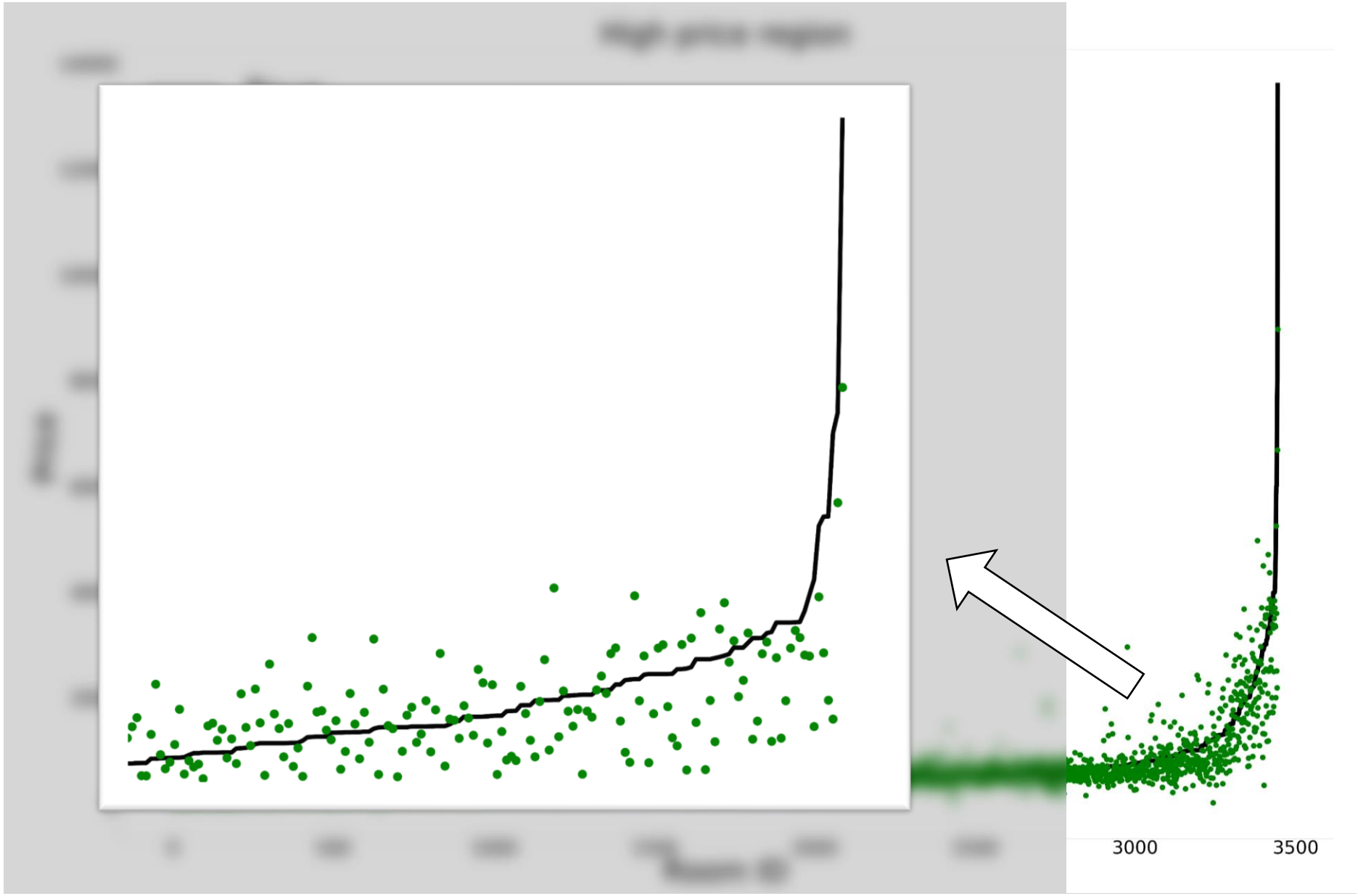
Beijing

Entire house

12 guests 6 bedrooms 6 beds 2 baths

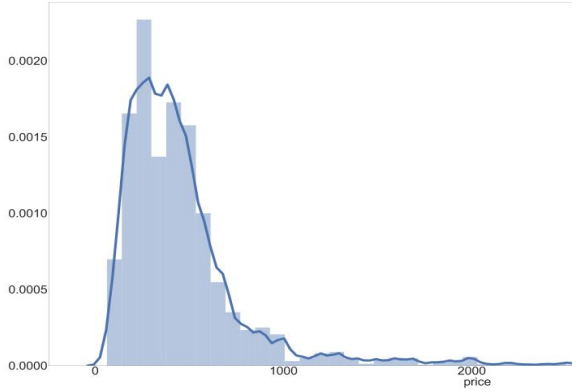
¥ 1,888 per night


High price region
True
Test
XGBoost
Price
Room ID
14000
12000
10000
8000
6000
4000
2000
0
0
500
1000
1500
2000
2500
3000
3500


3000
3500

Error Analysis

- Expensive listings have **less** data than cheap listings

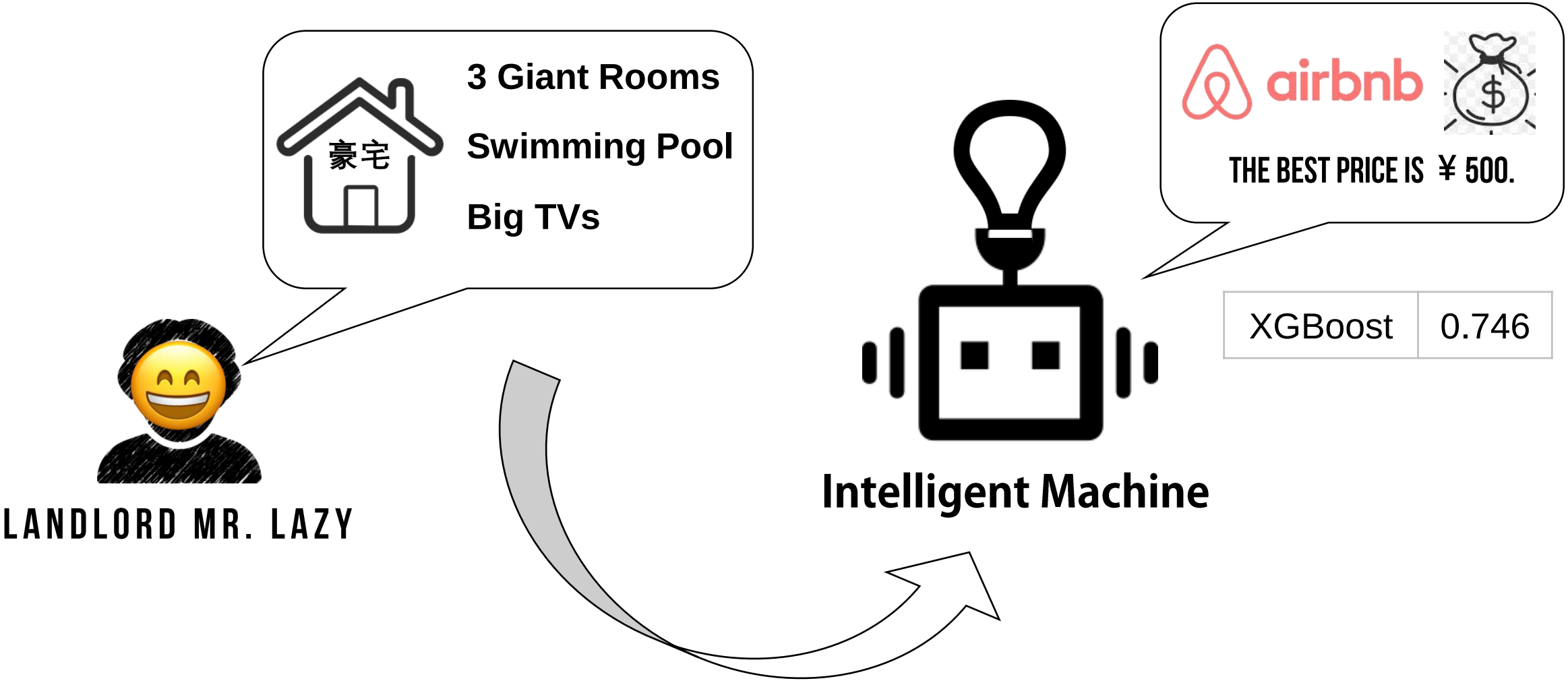


- **Feature isn't everything for price.**

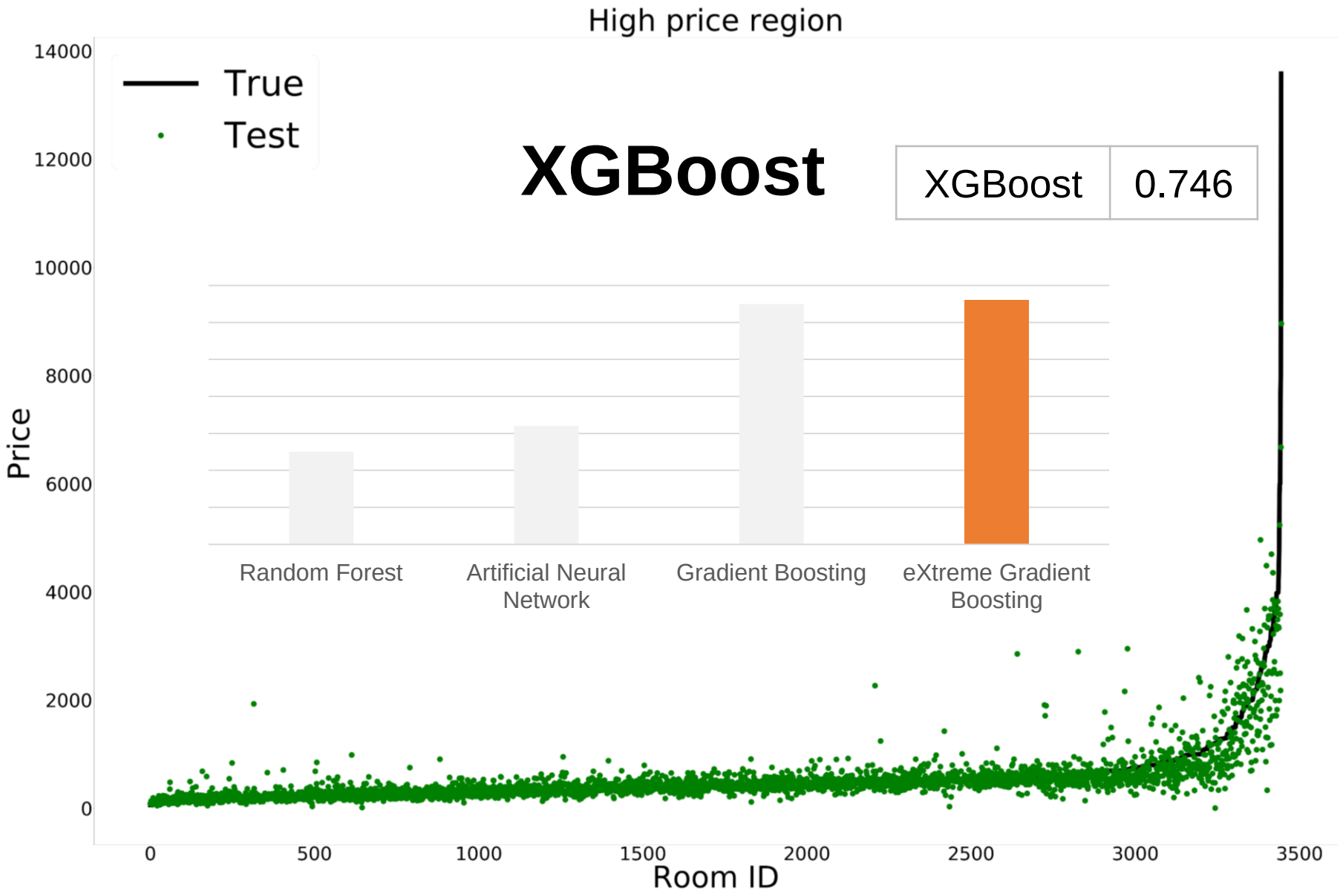
¥ 200

¥ 2348

To wrap up,



XGBoost	0.746


High price region
True
Test
XGBoost
XGBoost 0.746
Random Forest
Artificial Neural Network
Gradient Boosting
eXtreme Gradient Boosting
Price
14000
12000
10000
8000
6000
4000
2000
0
0
500
1000
1500
2000
2500
3000
3500
Room ID

Thank you!

Thank you!

AI HUB 😘

Q & A