Airbnb Pricing and Crime in Seattle

Team members: Chenchen Jiang, Hui Du, Jianjian Liu, Jiayang Liu

Date: April 11, 2024

CONTENT

Introduction

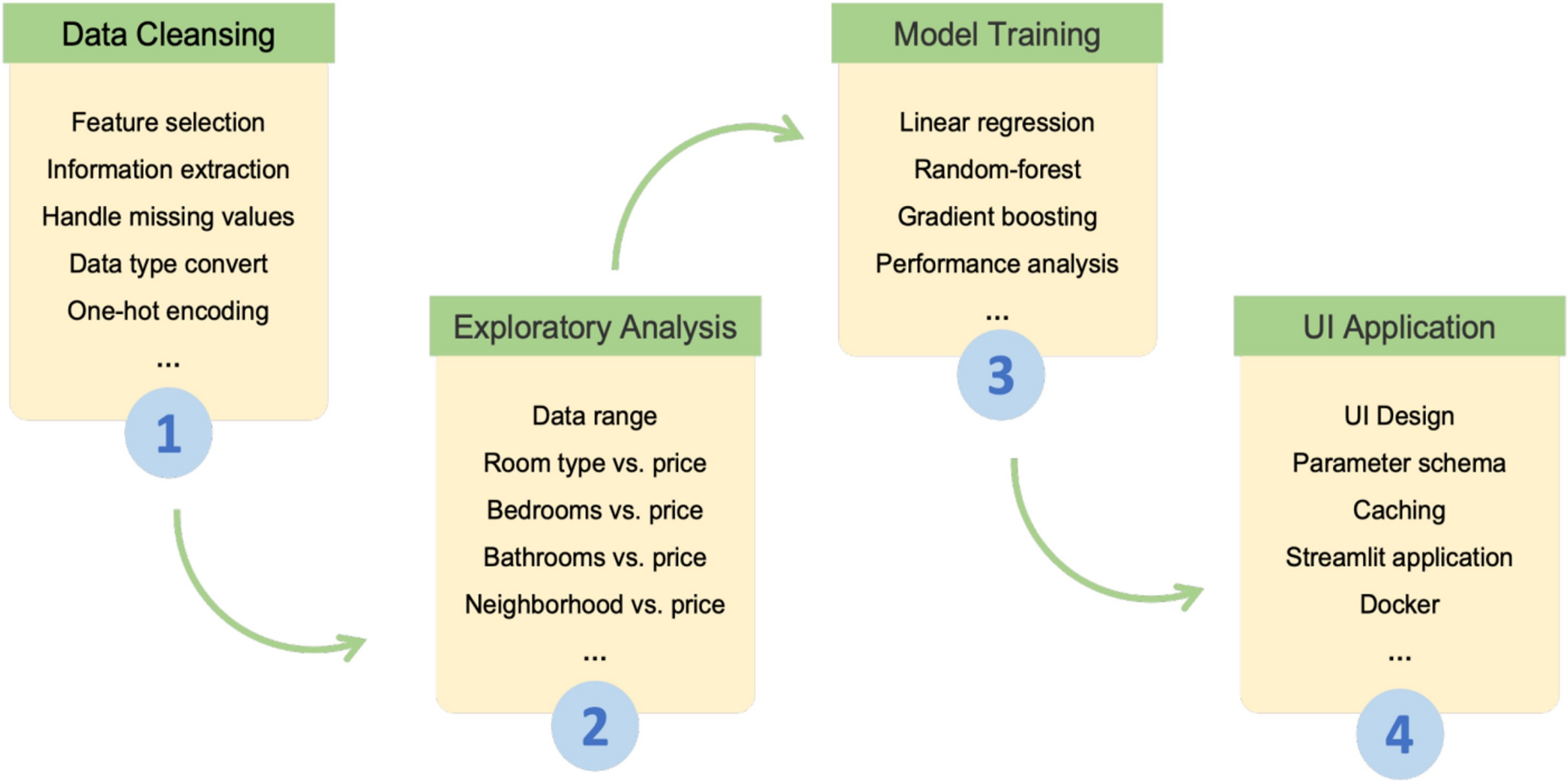
- Main Goal & Background

  As the demand for travel continues to grow, Airbnb is also rapidly expanding, becoming a top choice for those looking for accommodation. To enhance the travel experience in the Seattle area, our project is dedicated to researching the key factors that influence local Airbnb rental prices.

- Data Source

  insideairbnb.com

  data.seattle.gov



workflow summary

Data Preprocessing

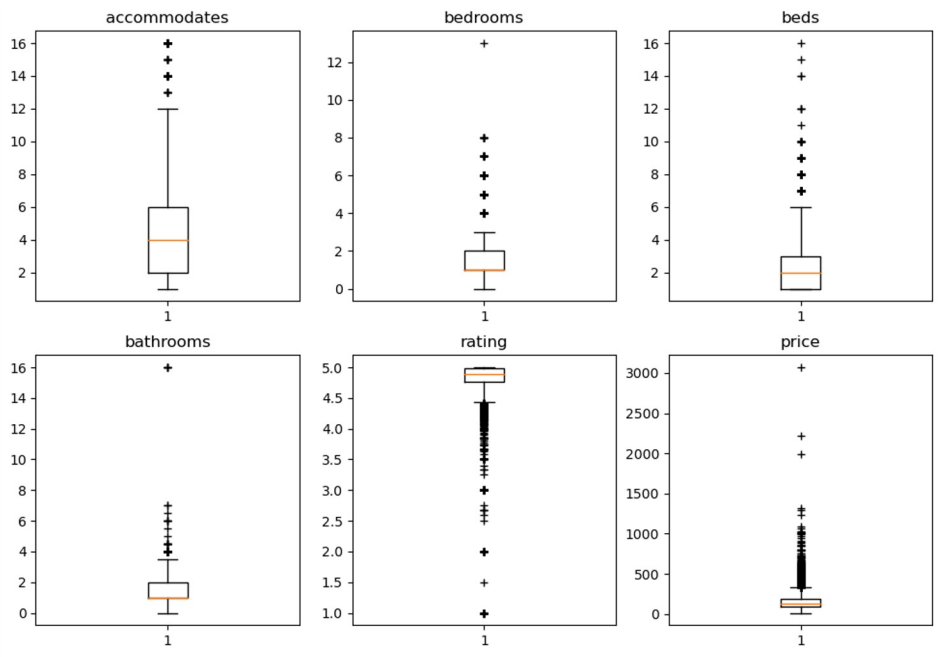
- Data Cleaning
- Handling Missing Values
- Data Transformation
- Feature Scaling
- Dimensionality Reduction
- Handling Outliers
- Splitting Data into Training and Testing Sets

```
missing_percentage = df.isna().sum() / len(df) * 100
print(missing_percentage)
```

```
id                              0.000000
availability in one year        0.000000
name                            0.000000
neighbourhood_cleansed          0.000000
latitude                        0.000000
longitude                       0.000000
property_type                   0.000000
room_type                       0.000000
accommodates                    0.000000
bathrooms_text                  0.011518
bedrooms                        4.192582
beds                            1.128772
price                           0.195807
```

Data Exploration&Visualization

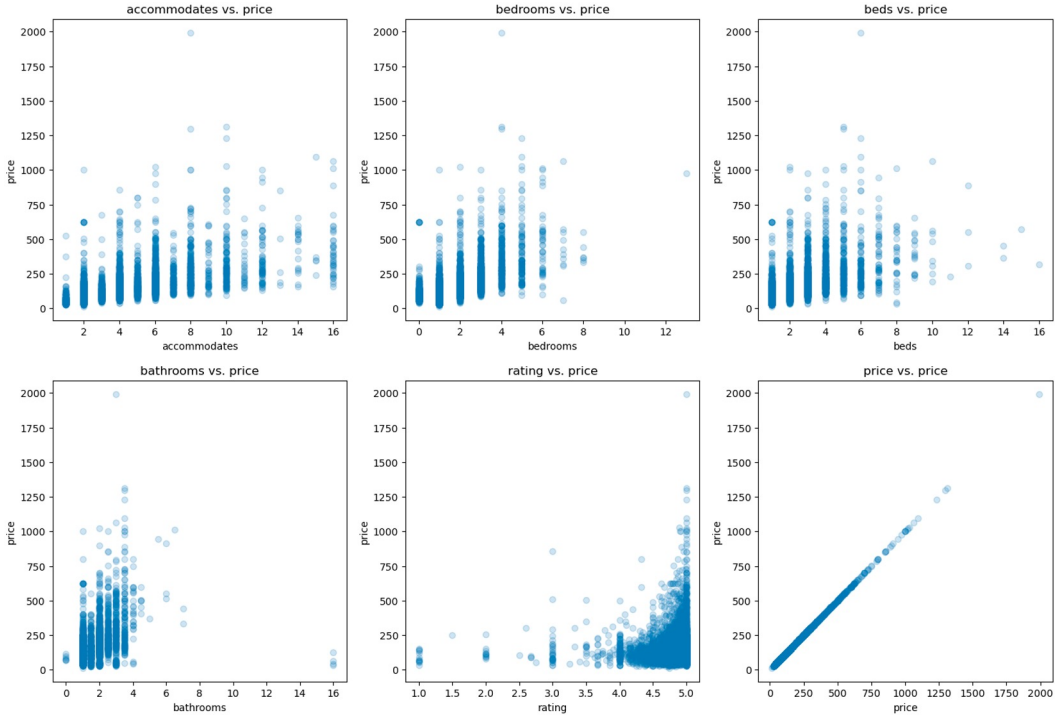
Data Distribution

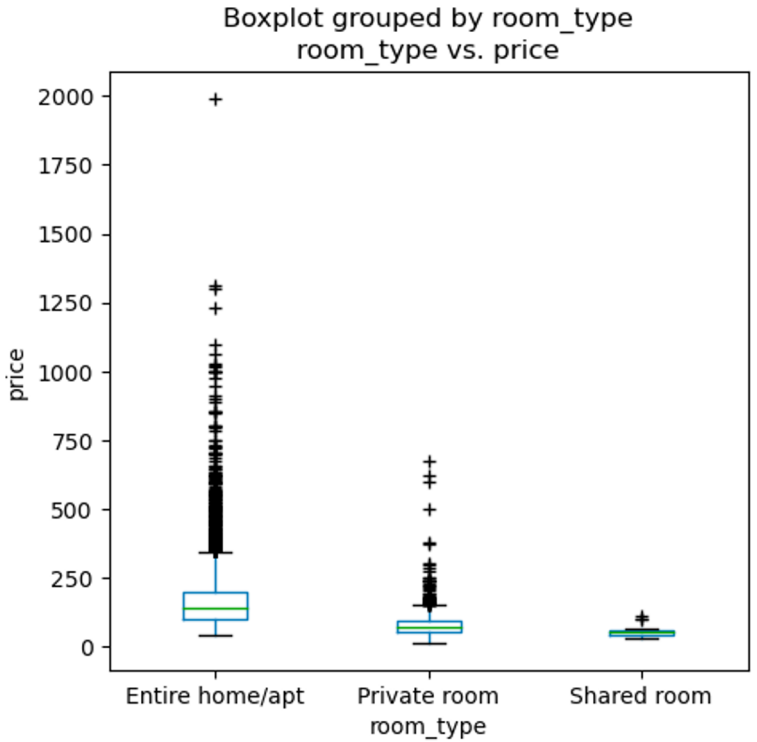


- Majority:
  - 👫accommodate 2-6 people
  - 🚪bedrooms 1-2
  - 🛏️ beds 1-3
  - 🚿bathrooms 1-2
  - 🎖️⭐rating 4.7-4.9,
  - 💰price $100-200
- Outliers:
  - 🎖️⭐rating < 2.0
  - 💰price > $2000

Data Exploration&Visualization

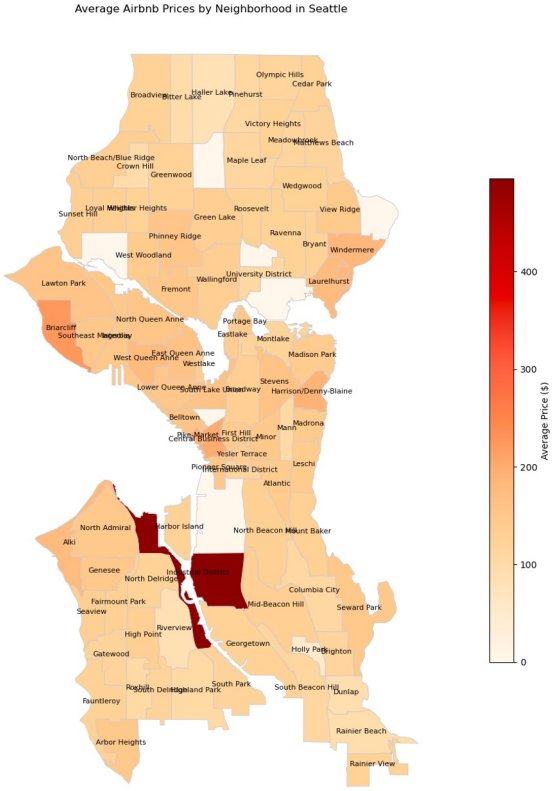
Data Correlation

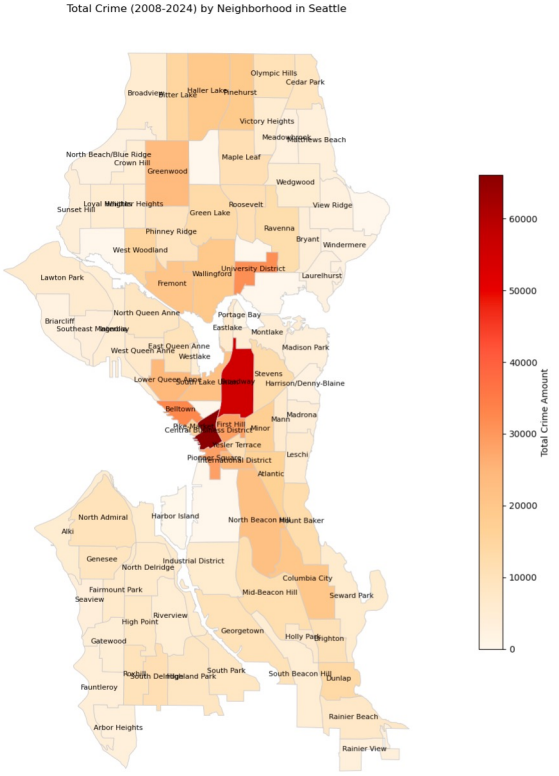




Data Exploration&Visualization

Choropleth Map





Modeling & Analysis

Independent variables: area, property type, room type, accommodates, bedrooms, beds, total crime.

Dependent variable: price of each night for different property.

Applying different models for prediction. And evaluate the results by using RMSE, MAE and R-squared.

	Model	RMSE	MAE	R2
0	Linear Regression	68.688701	46.481987	0.533314
1	Random Forest	70.291394	44.894421	0.511282
2	Gradient Boosting	68.072180	45.279054	0.541654

Modeling & Analysis

Applying Mutual-Information, select the columns with high MI scores to see any possible improvement from datasets.

accommodates	0.324199
bedrooms	0.270951
beds	0.234636
bathroom_count	0.212345
room_type_Entire home/apt	0.179387
room_type_Private room	0.173234
Total crime	0.147735
property_type_Private room in home	0.129895
bathroom_type_standard	0.112234
bathroom_type_shared	0.101257

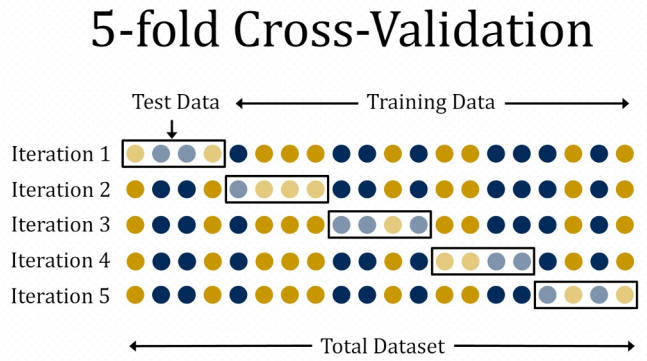
	Model	RMSE	MAE	
0	Linear Regression	68.688701	46.481987	0.5333
1	Random Forest	70.291394	44.894421	0.5112
2	Gradient Boosting	68.072180	45.279054	0.5416

	Model	RMSE	MAE	R2
0	Linear Regression	73.185889	49.468502	0.470204
1	Random Forest	73.076608	47.598107	0.471785
2	Gradient Boosting	68.983334	46.367501	0.529302

Modeling & Analysis

Optimal Random Forest Model:

1. Define parameters: number of estimators(trees), max of depth(depth of the tree), and minimum number of samples required to split a node.
2. Random search: for each iteration, RS pick a set of parameters for model.
3. 5-fold Cross validation: the model is trained and tested five times.
4. Then we evaluate and find the best selection.



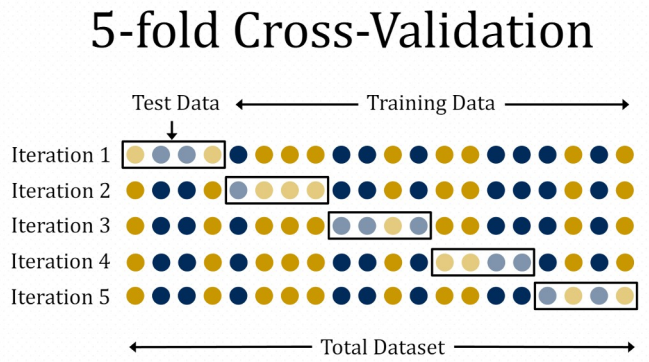

```
Fitting 5 folds for each of 10 candidates, totalling 50 fits
Best parameters: {'max_depth': 10, 'min_samples_split': 13, 'n_estimators': 444}
RMSE: 68.72959052576944
MAE: 45.34864180618583
R-squared: 0.5327583825577196
```

	Model	RMSE	MAE	R2
0	Random Forest Before Optimization	70.291394	44.894421	0.511282
1	Random Forest After Optimization	68.729591	45.348642	0.532758

Modeling & Analysis

Optimal Random Gradient Boosting:

1. Define parameters: number of estimators(trees), max of depth(depth of the tree), learning rate, min of split node and min of split leaf.
2. Random search: for each iteration, RS pick a set of parameters for model.
3. 5-fold Cross validation: the model is trained and tested five times.
4. Then we evaluate and find the best selection.




```
Fitting 5 folds for each of 10 candidates, totalling 50 fits
Best parameters: {'learning_rate': 0.05, 'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 8, 'n_estimators': 343}
RMSE: 67.43080705405146
MAE: 44.17175857869453
R-squared: 0.550250467180026
```

	Model	RMSE	MAE	R2
0	Gradient Boosting Before Optimization	68.072180	45.279054	0.541654
1	Gradient Boosting After Optimization	67.430807	44.171759	0.550250

Modeling & Analysis

Further improvement: Stacking model of MLR, RF(trained), and GB(trained) with 5-fold Cross validation.

Final improvement: Applying bagging method to improve the stacking model.

```
Stacking Model Evaluation Metrics:
RMSE: 67.27960349946956
MAE: 44.13780615812178
R-squared: 0.5522651986322878
```

```
Bagging Stacked Model Evaluation Metrics:
RMSE: 66.38752719905041
MAE: 44.00577740079706
R-squared: 0.5640597284581361
```

Modeling & Analysis

1. Crime affect the price.
2. Potential risk.

accommodates	0.324199
bedrooms	0.270951
beds	0.234636
bathroom_count	0.212345
room_type_Entire home/apt	0.179387
room_type_Private room	0.173234
Total crime	0.147735

T1: accuracy and value

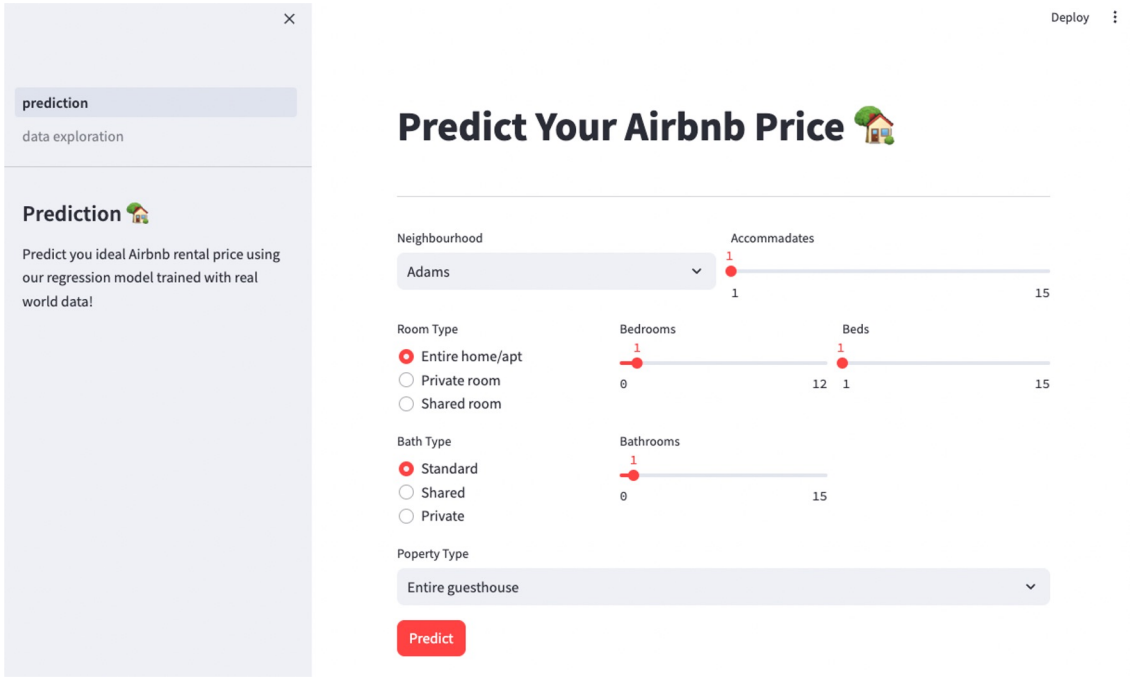
T2: cleanliness and communication

T3: checkin and location

	Rating	accuracy	cleanliness	checkin	communication	location	value
Rating	1.00	0.85	0.77	0.67	0.77	0.62	0.84
accuracy	0.85	1.00	0.70	0.60	0.69	0.58	0.78
cleanliness	0.77	0.70	1.00	0.50	0.59	0.47	0.66
checkin	0.67	0.60	0.50	1.00	0.68	0.49	0.61
communication	0.77	0.69	0.59	0.68	1.00	0.49	0.69
location	0.62	0.58	0.47	0.49	0.49	1.00	0.60
value	0.84	0.78	0.66	0.61	0.69	0.60	1.00

Application Demo

http://localhost:8501



Thank You!