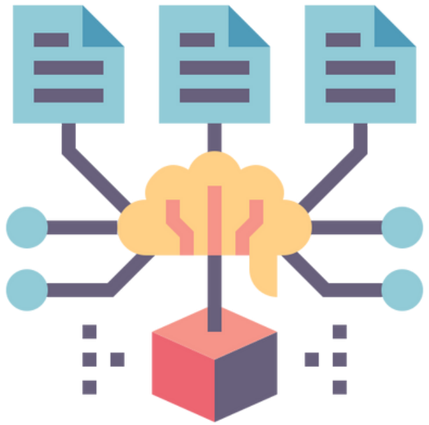
# Predictive Modeling for Pricing on Booking.com

Yenying Chen

# Introduction



**Predictive model for accommodation pricing in Paris on Booking.com**

- **Supervised Learning Methods:** Linear/Non-Linear Regression Models
- **Evaluation metrics:** R^2 and RMSE on test sets
- **XGBoost** showed the strongest predictive ability
- Hyperparameter tuning
- Feature Set Analysis, Feature Importance Analysis, and Feature Ranking Analysis
- **Best model :** XGBoost with top 35 features

Presented by Yenying Chen

# Data Collection

- **Web Scraping:** 20 Paris districts from Booking.com
- **Collected Date:** 1-night stay for 27-28/02, 2025
- **Filters:** 2 adults, property types (hotel & apartment), free cancellation, and free Wi-Fi
- **Data points:** hotel_name, property_type, star_rating, address, room_type, price, reviews, overall_rating, category_ratings, available facilities


Presented by Yenying Chen

# Data Preprocessing

**Handling Missing Values**

- reviews, overall_rate, staff, facilities, cleanliness, comfort, value of money, and location
- Impute **Median** to deal with the skewness

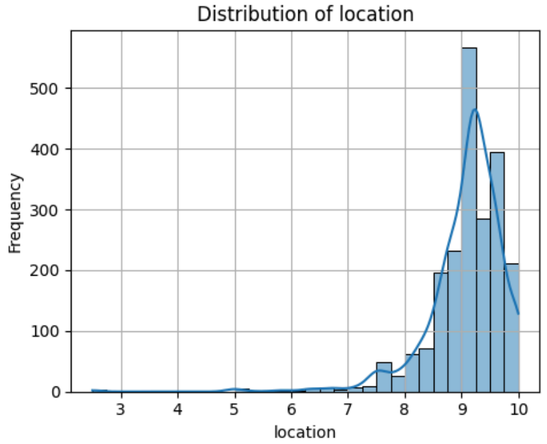
**Categorical Encoding**

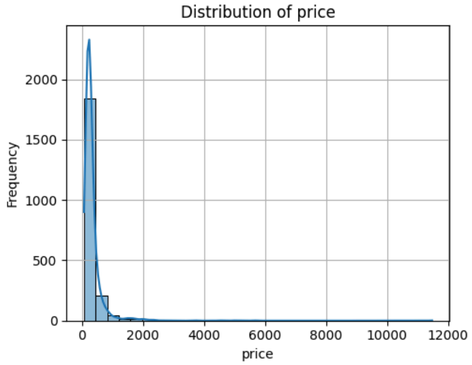
- star_rating, property_type, room_type, and address
- room_type:  double/twin Room, apartment, studio, luxury room (suite/deluxe/premium), and other.

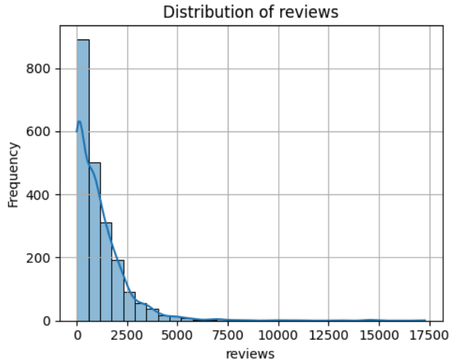
Presented by Yenying Chen

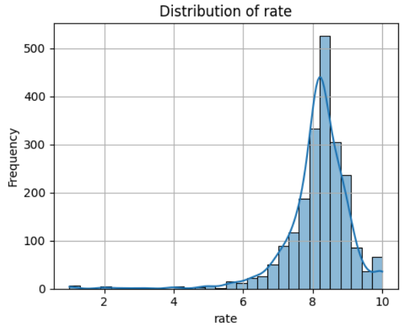
# EDA

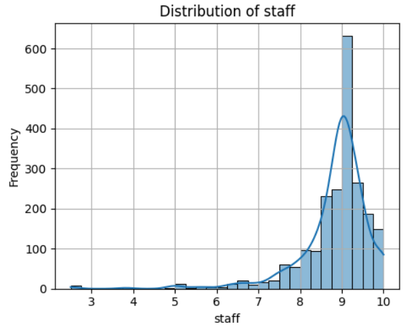
## Numerical Variables

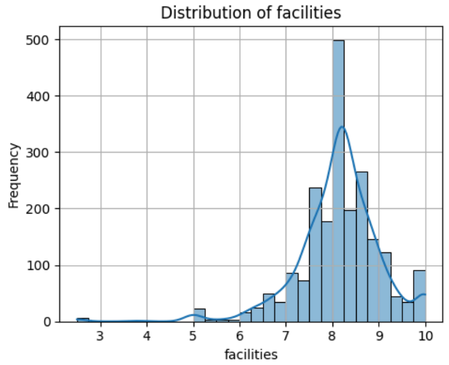


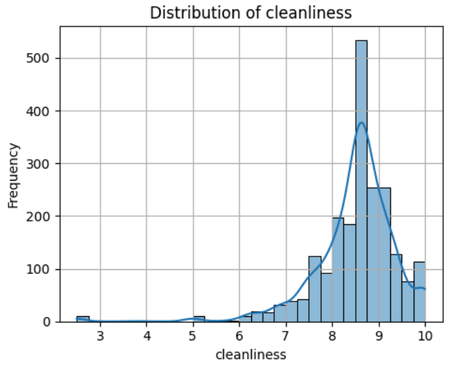



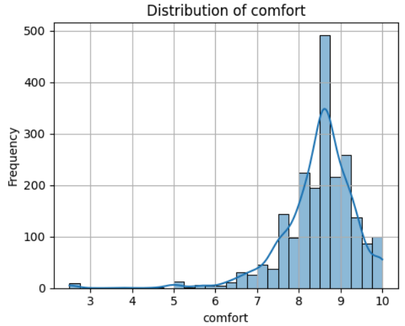



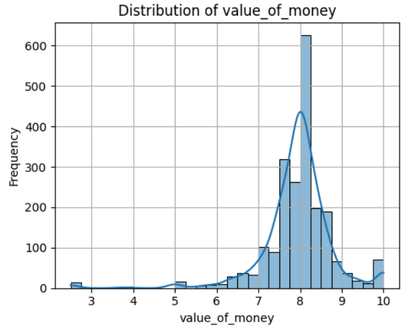



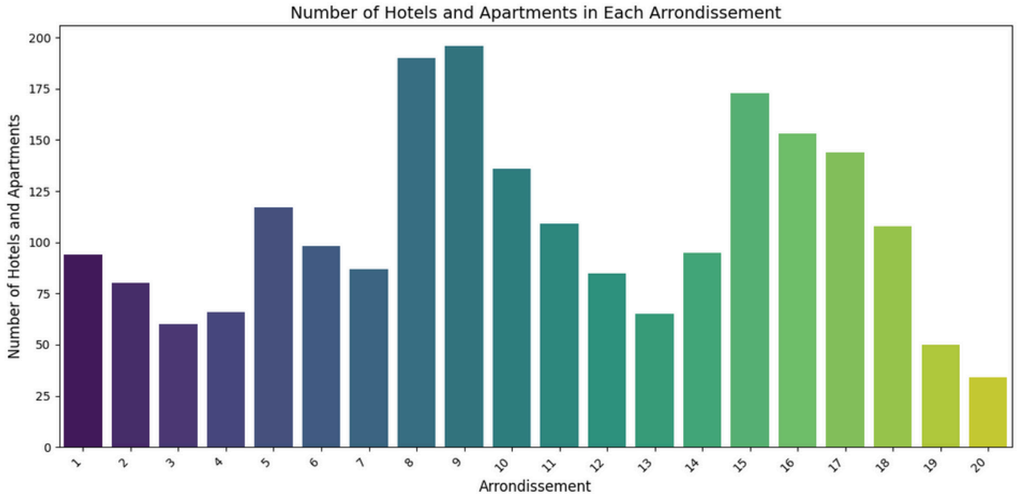



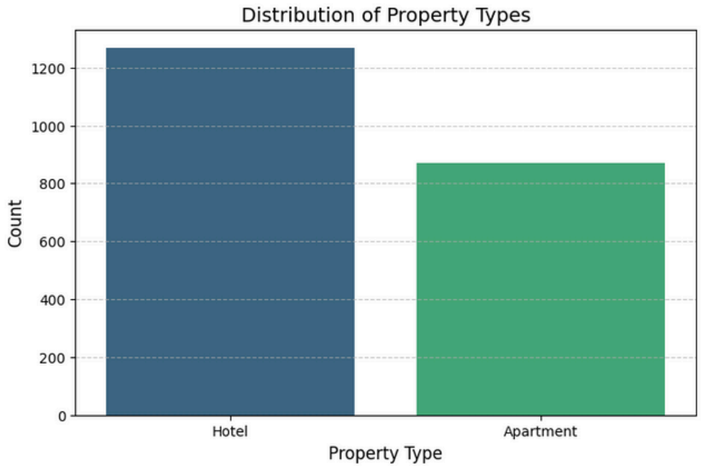



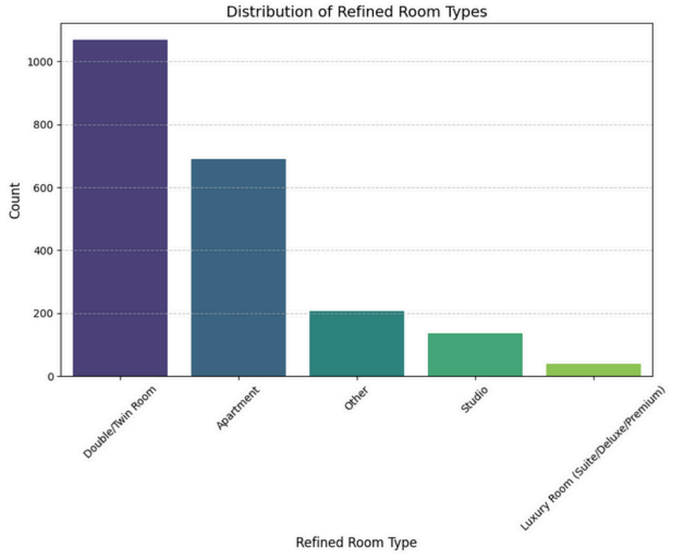

Presented by Yenying Chen

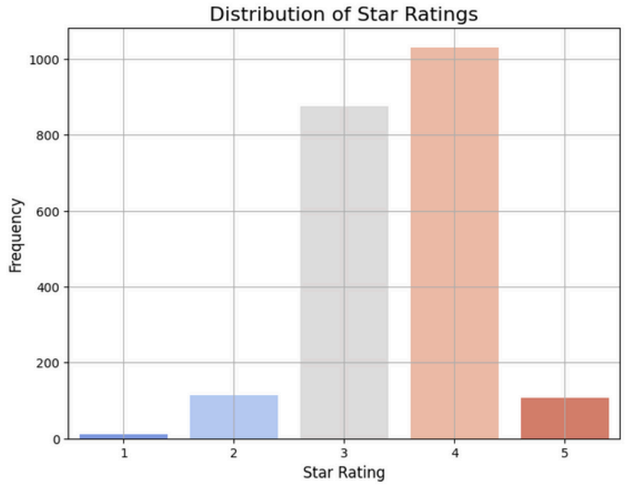
# EDA

## Categorical Variables



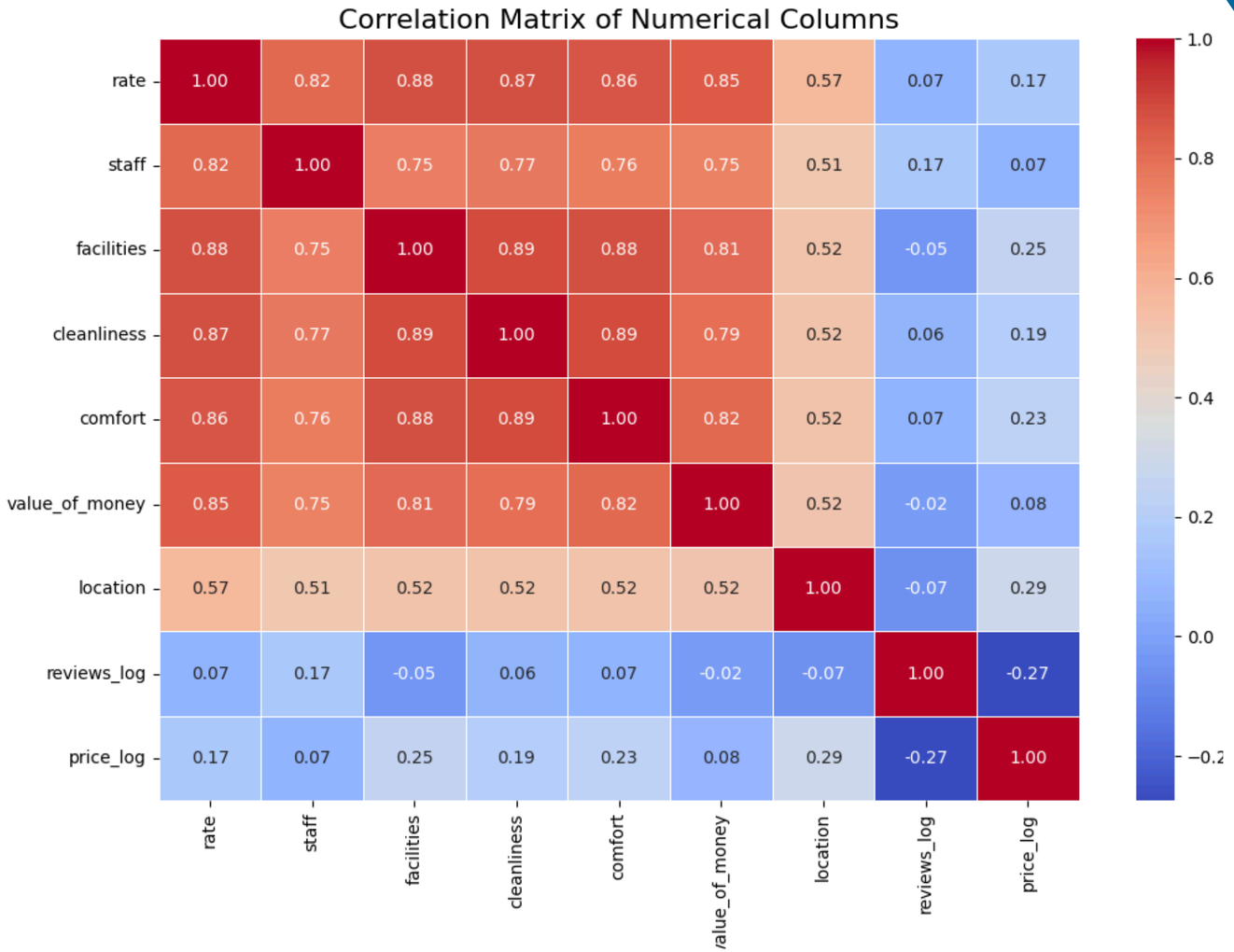




Presented by Yenying Chen

# EDA

## Correlation Matrix of Numerical Variables




Presented by Yenying Chen

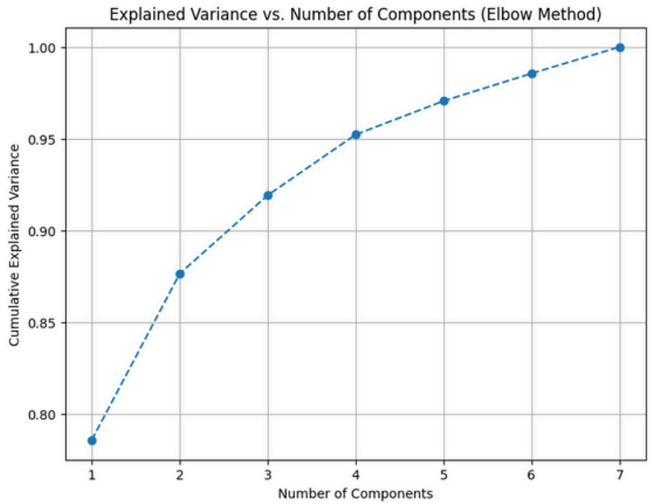
# Methodology

- Dataset: training(80%)/ test(20%)
- 5-fold cross-validation
- Linear Regression Models (+ PCA)
  - baseline
  - baseline + RFE
  - Ridge
  - Lasso
- Non-linear Regression Models
  - Random Forest
  - Gradient Boosting
  - XGBoost

- Select the best predictive model
  - Hyperparameter Tuning
  - Feature Set Analysis
  - Feature Importance Analysis
  - Feature Ranking Analysis
- Evaluation Metrics
  - $R^2$ and RMSE on test set
- Log Transformation for price & reviews

Presented by Yenying Chen

# Implementation




| Principal Component | Variance Explained (%) |
|---|---|
| PC1 | 78.57 |
| PC2 | 9.06 |
| PC3 | 4.27 |
| PC4 | 3.30 |
| PC5 | 1.83 |
| PC6 | 1.50 |
| PC7 | 1.44 |

| Variable | VIF after PCA |
|---|---|
| reviews_log | 1.009831 |
| PC1 | 1.000262 |
| PC2 | 1.001763 |
| PC3 | 1.007806 |

Presented by Yenying Chen

# Result

**Linear**

| Evaluation Metrics | Baseline | Baseline with RFE | Ridge | Lasso |
|---|---|---|---|---|
| $R^2_{\text{test}}$ | 0.4532 | 0.4539 | 0.4544 | 0.4538 |
| $\text{RMSE}_{\text{test}}$ | 0.4379 | 0.4392 | 0.4376 | 0.4379 |

**Non-Linear**

| Evaluation Metrics | Random Forest | Gradient Boosting | XGBoost |
|---|---|---|---|
| $R^2_{\text{test}}$ | 0.5399 | 0.5291 | 0.5611 |
| $\text{RMSE}_{\text{test}}$ | 0.4019 | 0.4066 | 0.3925 |

# Result

## Hyperparameter Tuning

- The best parameters for the second round exhibited ‘n_estimators’: 700, ’learning_rate’: 0.01, ’max_depth’: 5, ’colsample_bytree’: 0.6, ’subsample’: 0.7

| Evaluation Metrics | Original | Round 1 | Round 2 | Round 3 |
|---|---|---|---|---|
| $R^2_{\text{test}}$ | 0.5611 | 0.5676 | 0.5711 | 0.5690 |
| $\text{RMSE}_{\text{test}}$ | 0.3925 | 0.3896 | 0.3880 | 0.3890 |

Presented by Yenying Chen

# Result

## Feature Set Analysis

| Feature Set | Component | $R^2_{\text{test}}$ | $\text{RMSE}_{\text{test}}$ |
|---|---|---|---|
| 1 | ratings + property_specific + facilities + ['reviews_log'] | 0.5711 | 0.3880 |
| 2 | property_specific | 0.4522 | 0.4385 |
| 3 | ratings | 0.4140 | 0.4536 |
| 4 | facilities | 0.1325 | 0.5519 |
| 5 | ratings + property_specific | 0.5600 | 0.3930 |
| 6 | property_specific + facilities | 0.4756 | 0.4291 |
| 7 | ratings + facilities | 0.4156 | 0.4529 |
| 8 | ratings + property_specific + facilities | 0.5554 | 0.3951 |

# Result

## Feature Importance Analysis

| Feature Group | Feature Category | Feature Importance | Individual Feature (Top5) |
|---|---|---|---|
| Property-Specific | address | 0.1754 | address_16: 0.0370<br>address_8: 0.0219<br>address_18: 0.0186<br>address_12: 0.0170<br>address_20: 0.0168 |
| | star_rating | 0.3007 | star_5: 0.1194<br>star_3: 0.0858<br>star_4: 0.0606<br>star_2: 0.0194<br>star_1: 0.0076 |
| | refined_room_type | 0.1327 | apartment: 0.0576<br>double / twin room: 0.0304<br>studio: 0.0216<br>luxury room: 0.0174<br>other: 0.0058 |
| | property_type | 0.0313 | apartment: 0.0214<br>hotel: 0.0099 |
| Ratings | | 0.1058 | location: 0.0426<br>comfort: 0.0177<br>value_of_money: 0.0107<br>facilities: 0.0102<br>rate: 0.0082<br>cleanliness: 0.0081<br>staff: 0.0080 |
| Facilities | | 0.1666 | air conditioning: 0.0249<br>heating: 0.0186<br>room service: 0.0184<br>breakfast: 0.0141<br>daily housekeeping: 0.0139 |
| | Others | | reviews_log: 0.0349 |

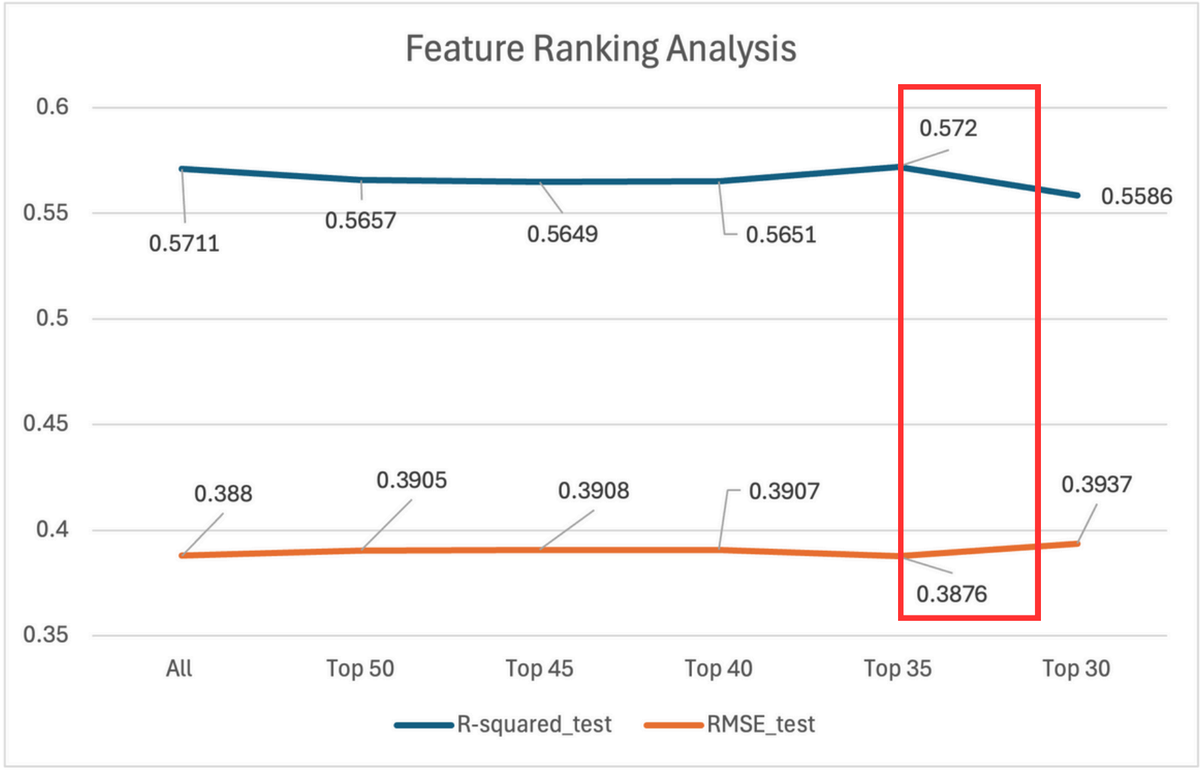
Presented by Yenying Chen

# Result

| Feature Set | Features | $R^2_{\text{test}}$ | $\text{RMSE}_{\text{test}}$ |
|---|---|---|---|
| 9 | top 50 | 0.5657 | 0.3905 |
| 10 | top 45 | 0.5649 | 0.3908 |
| 11 | top 40 | 0.5651 | 0.3907 |
| 12 | top 35 | 0.5720 | 0.3876 |
| 13 | top 30 | 0.5586 | 0.3937 |

- RMSE of 0.3876 on log scale. Exponentiated RMSE 47% average deviation from actual prices.

## Feature Ranking Analysis

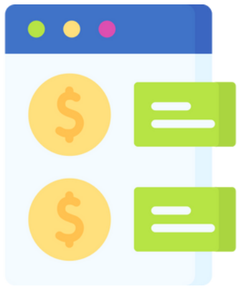



Presented by Yenying Chen

# Conclusion

XGBoost
with top 35 features

Feature Selection
Importance



Insights for Pricing
Strategies

Presented by Yenying Chen

# Thank You

Presented by Yenying Chen