Finding the Perfect Restaurant

Analyzing the factors that contribute to a restaurant's rating in Bangalore, India

Michael Li, Chaelsey Park, Hemosoo Woo

Background

Location

- Bangalore, India known as India's leading IT exporter
- Features 2000+ types of international cuisines (e.g. American, Asian, European)
- Prominent nightlife with endless bar, pub, and dessert options

Dataset

- "Zomato Bangalore Restaurants" from Kaggle
- Zomato is India's leading restaurant aggregator and food delivery service; similar to Uber Eats

Goal

Objective

- Understand the factors that contribute to a restaurant’s rating
- Hypothesize whether a restaurant's location is significantly correlated to the rating

Value Proposition

- Restaurants can use this knowledge to:
  - improve ratings
  - create better customer experience.
- Tourists can gain a better understanding of areas with higher rated restaurants

Initial Observations

from Explanatory Data Analysis

Correlation between Price and Rating

Observation:

- High Prices correlates more with high rating
- Low Prices has a large range of rating scores

Explanation:

- High end restaurants will tend to have higher ratings
- More affordable restaurants can vary depending on more factors

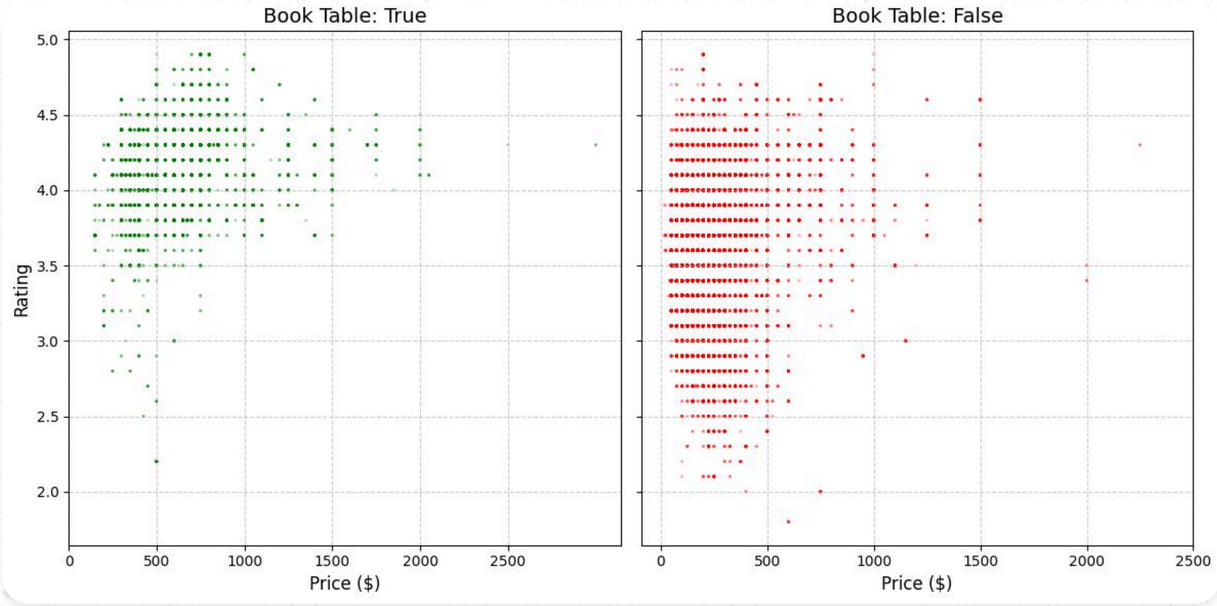
Ability to Reserve a Table

Observation:

- Yes: Ratings are shown to be higher
- No: Ratings have a larger range of ratings

Explanation:

- Restaurants in high demand require a reservation system to handle more traffic.



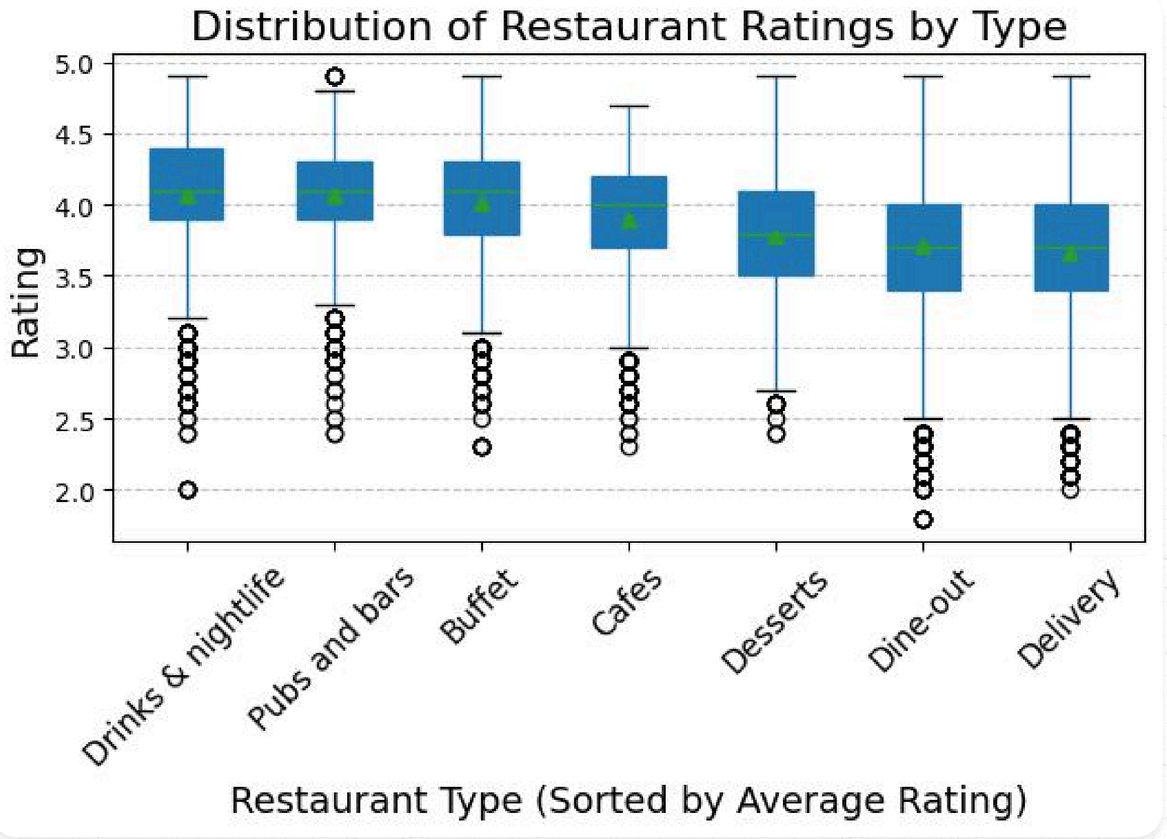
Correlation between Type and Rating

Observation:

- Drinks & Nightlife restaurants have the highest average rating
- The top rated types have +1 rating score than the bottom rated types

Explanation:

- Bangalore has a strong nightlife culture, explaining why Drinks & nightlife + Pubs and Bars is rated favorably



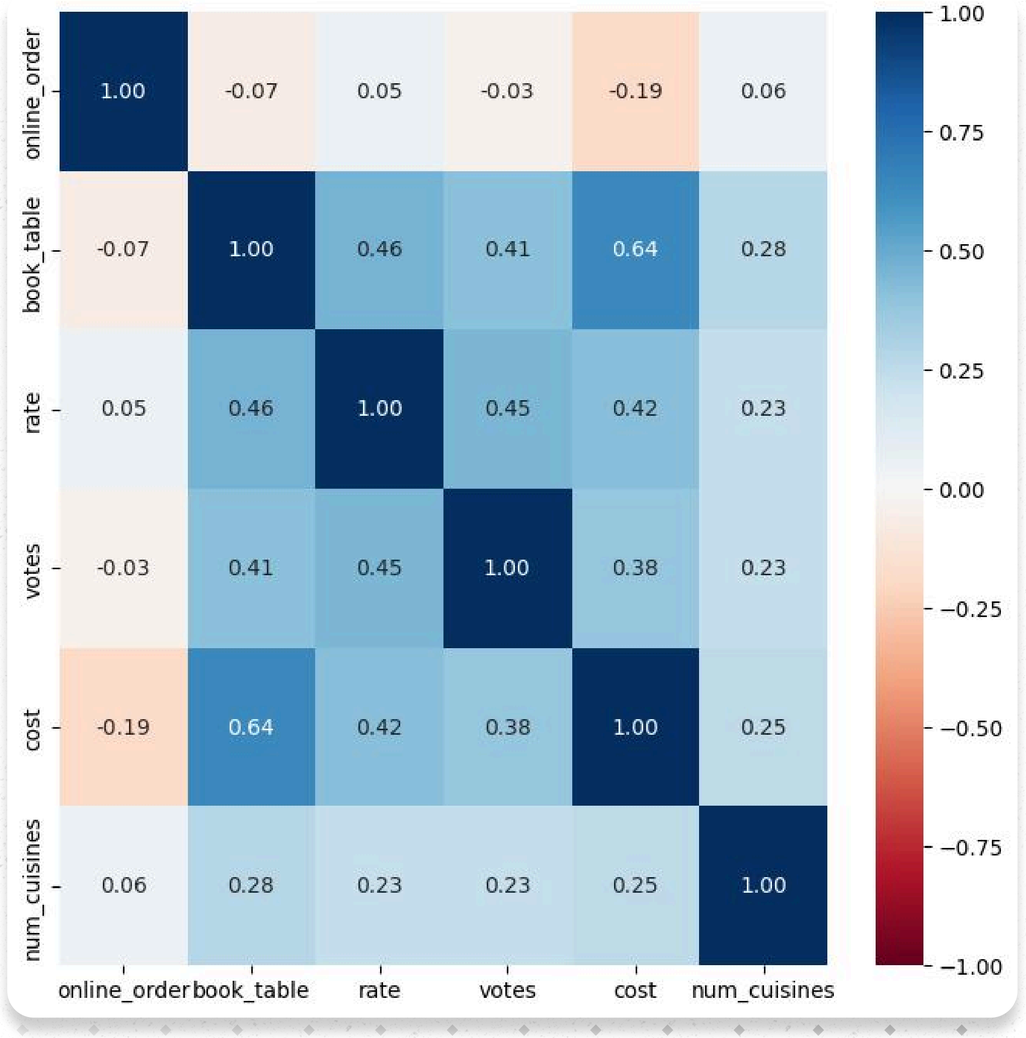
Feature Correlation Heatmap

Observation:

- Cost and book_table seem slightly correlated
- online_order seems to have the lowest correlation with every other feature

Explanation:

- Being able to book a table is inherently associated with a higher quality restaurant
- Online ordering is offered by all restaurants whether good or bad quality



Performance Review

from Modeling + Hyperparameter Tuning

Model Performance

Random Forest Regressor

- More accurate and effective
- Ensemble of different decision trees and averages results
- Greater R^2 than Linear Regression meaning model is more accurate

```
Test Set R^2 value: 0.6185873097529937
Train Set R^2 value: 0.6417312876641263
Mean Squared Error: 0.3842656148756201
```

Linear Regression

- Unable to create an accurate model
- Data does not have a very linear relationship

```
Test Set R^2 value: 0.39810754640899326
Training Set R^2 value: 0.39963281294499664
Mean Squared Error: 0.6063945424006751
```

XGBoost Regressor

- Most accurate
- Sequences of decision trees created based of issues of past decision trees

```
Test Set R^2 value: 0.7679686143334198
Train Set R^2 value: 0.8104900897230702
Mean Squared Error: 0.23376695470166106
```

Hyperparameter Tuning

Random Forest Regressor

```
Test Set R^2 value: 0.6185873097529937
Train Set R^2 value: 0.6417312876641263
Mean Squared Error: 0.3842656148756201
```

```
{'max_depth': 40, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
0.8979921446975772
```

XGBoost Regressor

```
Test Set R^2 value: 0.7679686143334198
Train Set R^2 value: 0.8104900897230702
Mean Squared Error: 0.23376695470166106
```

```
Best Parameters: {'learning_rate': 0.1, 'max_depth': 18, 'n_estimators': 500}
Best R^2: 0.908362413443699
```

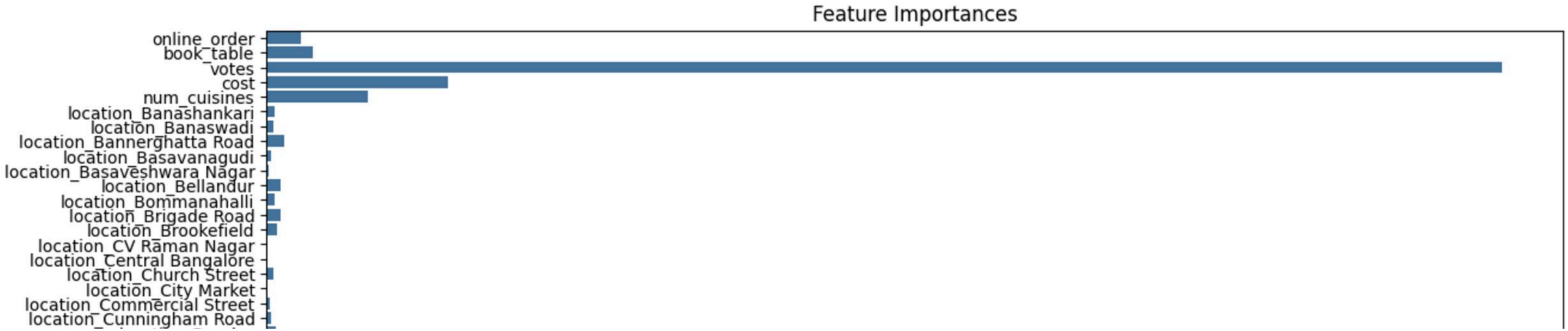
Clearly, Hyperparameter Fine Tuning greatly improved our models performances, reaching an R^2 very close to 1.

Implications, Insights

from Hypothesis Testing + Conclusion

Feature Importance

- For the Random Forest Model
- Number of votes demonstrates the highest importance
- Cost and the number of cuisines next, although with a much lower impact.
- Makes sense because in the extremes of restaurants being either extremely good or bad, customers are more influenced to post a rating in response to their very positive or negative experience.



Hypothesis Testing

Observation:

- How much does it help to know the location of the restaurant?
- Null Hypothesis: The “location” column has no meaningful impact on the rating (“rate” column).

Explanation:

- p_value = 0.025, p < 0.05, we reject the Null Hypothesis
- The “location” column has a meaningful impact on the rating of a restaurant.
- Our observed_r2 without the “location” column was not due to randomness, but because the “location” column does in fact have an impact.

```
[ ]  print(one_zero_p_value)
     print(observed_r2)
```

```
0.025
0.18574600345831338
```

Conclusions

Challenges

- What columns do we actually need?
- Some columns appear useful later became clear later that other columns had a much greater impact
- Difficult to perfect the first time

Limitations

- Limited to the specific location Bengaluru, India,
- Location and cuisine proved to be a major factors in models
- Models may not be accurate in other locations

Future Plans

- Either
  - Attain a larger dataset for the entire world
  - Joining different datasets to account for all countries and places
- Create a model that achieves the same accuracy for any restaurant in the world

Thank You!