Hotel Tier Prediction

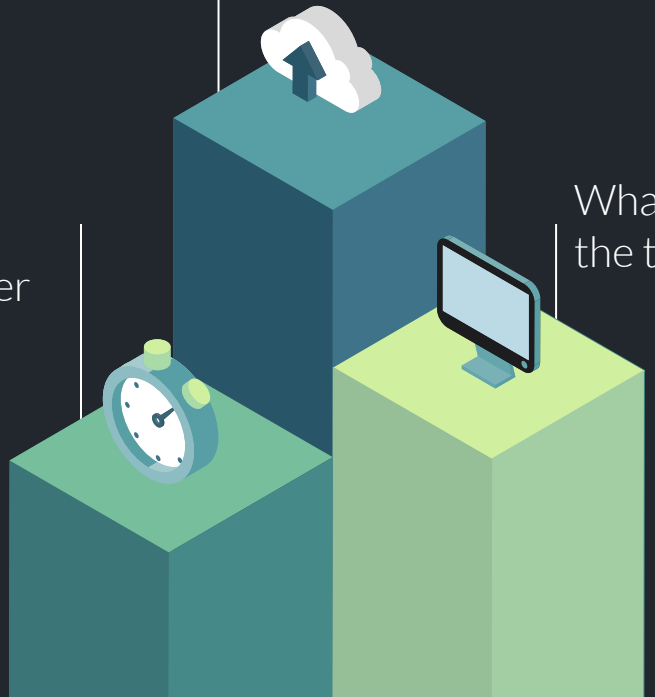
Quan Nguyen

Project Objectives

Predict hotel tiers in New York City area

Is it true that the more expensive is it, the higher the tier of a hotel?

What factors affect the tier of a hotel?



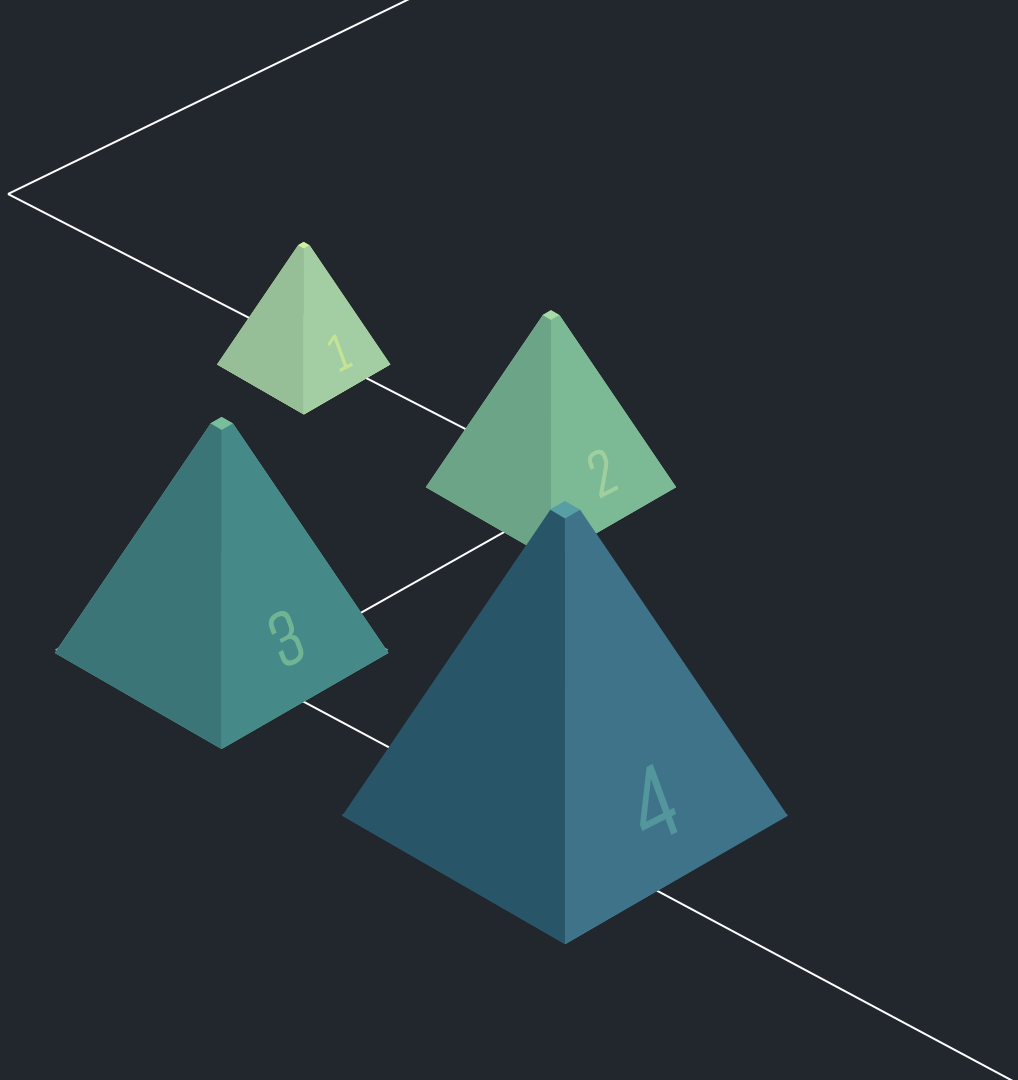
Project Timeline

1. Collecting Data
Scrape **hotels.com**

2.EDA
Visualization, Data Cleaning and Feature Engineering

3.Classification
Tuning and Training Models

4.Insights
Models analysis



Data source

hotels.com: 922 hotels in NYC

Features:

Categorical:

- restaurant
- bar
- fitness
- spa
- pool
- valet parking
- limousine
- rooftop

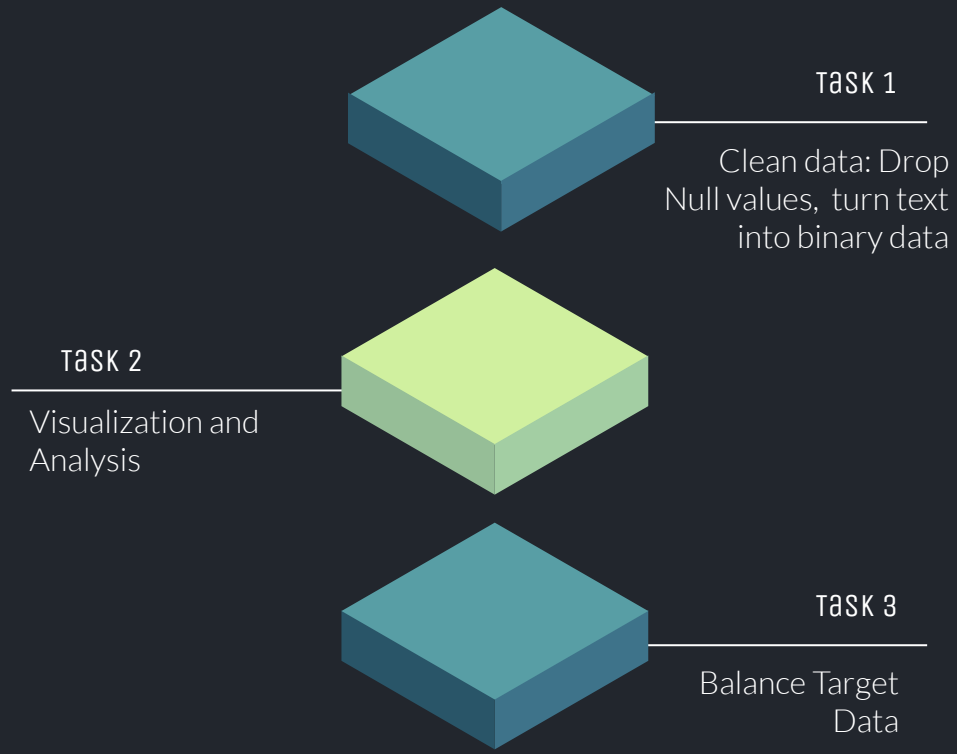
Continuous:

- price
- number of rooms

Target:

- star

EDA



TASK 1

Clean data: Drop Null values, turn text into binary data

TASK 2

Visualization and Analysis

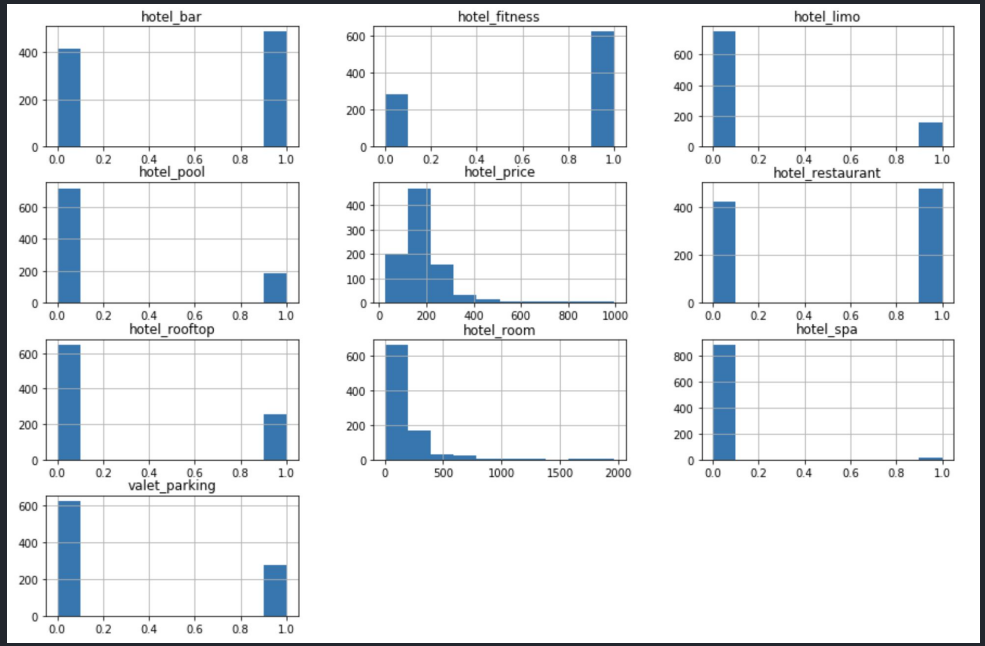
TASK 3

Balance Target Data

EDA

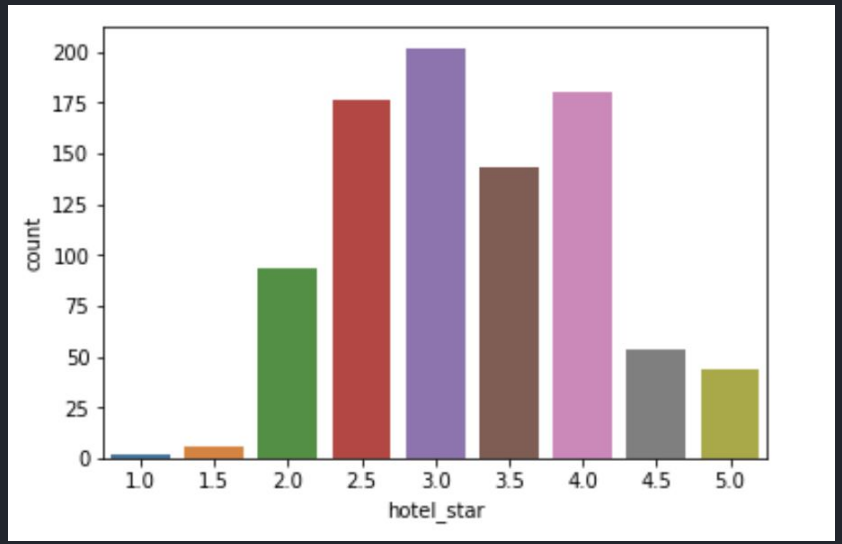
Drop *hotel_price* feature

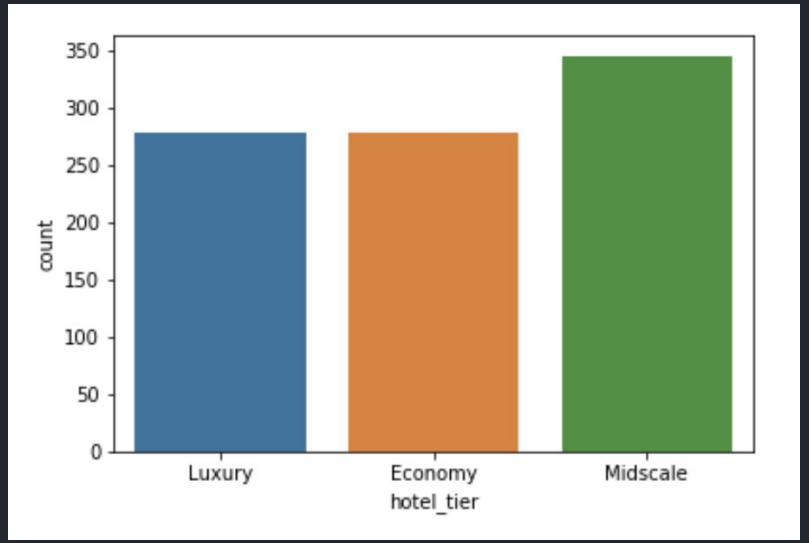
Scale the only continuous feature: *hotel_room*



	hotel_price	hotel_star
hotel_price	1.000000	0.627168
hotel_star	0.627168	1.000000

EDA

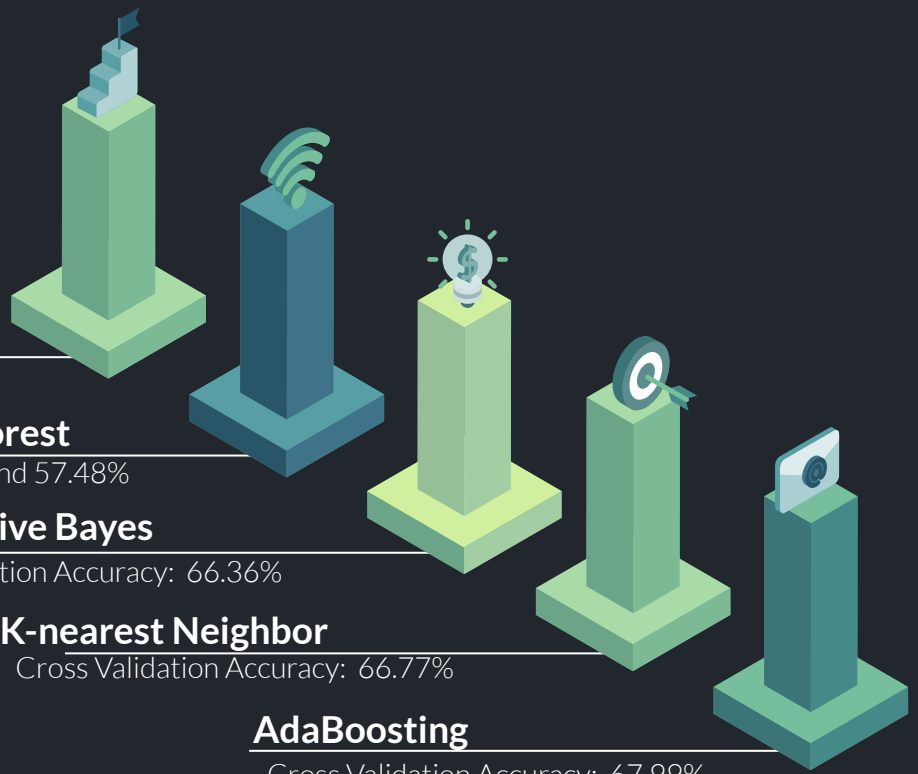




Balance Target Data by combining 10 groups of star into 3 tiers:

- Economy: 1 star - 2.5 star
- Midscale: 3.0 star - 3.5 star
- Luxury: 4 star - 5 star

MODELS



Support Vector Classification
Cross Validation Accuracy: 67.85%

Decision Tree and Random Forest
Cross Validation Accuracy: 57.94% and 57.48%

Bernoulli Naive Bayes
Cross Validation Accuracy: 66.36%

K-nearest Neighbor
Cross Validation Accuracy: 66.77%

AdaBoosting
Cross Validation Accuracy: 67.99%

TUNING MODELS

Random Forest

57.48% → 69.63%

SVC

67.85% → 68.15%

AdaBoosting

67.99% → 67.99%

Bernoulli NB

66.36 → 66.52%

APPLY MODELS TO TEST DATA

Bernoulli Naive Bayes

- Accuracy: 0.633
- Recall [0.86, 0.62, 0.46] *
- Precision [0.74, 0.57, 0.56]

Random Forest:

- Accuracy: 0.659
- Recall [0.89, 0.65 , 0.53]
- Precision [0.74, 0.66, 0.59]

SVC

- Accuracy: 0.65
- Recall [0.82, 0.63, 0.52] *
- Precision [0.72, 0.62, 0.59]

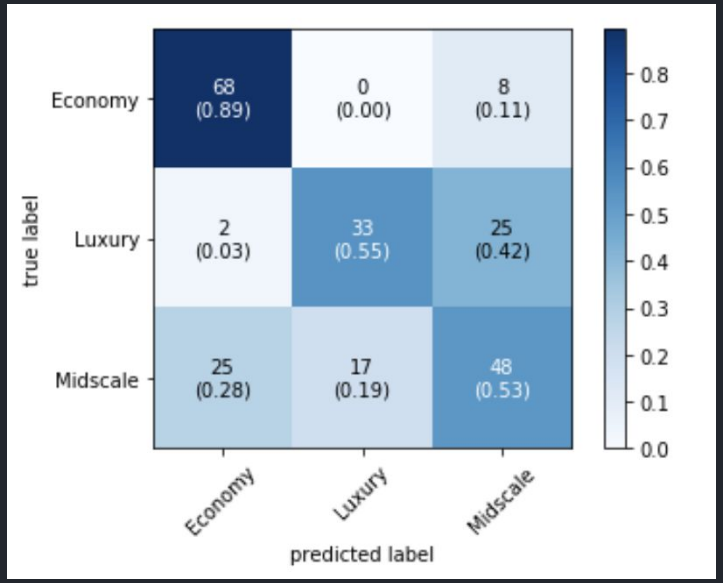
AdaBoosting:

- Accuracy: 0.664
- Recall [0.86, 0.65 , 0.51]
- Precision [0.74, 0.63, 0.61]

*Economy, Luxury, Midscale

winner

Random Forest



Accuracy: 0.575 → 0.659

Recall:

- Economy: 0.89
- Luxury: 0.55
- Midscale: 0.53

Precision:

- Economy: 0.74
- Luxury: 0.66
- Midscale: 0.59

Comparing to a random guess, which is 0.333 , the accuracy is twice as much.

The reason I choose Random Forest is that I want to minimize type I error so I pick a model with high **Precision** score.

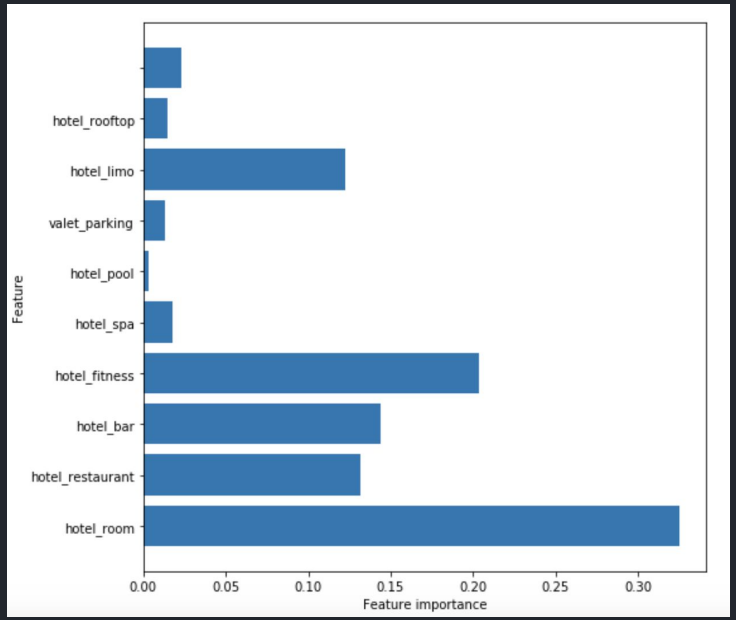
For example, if I booked and paid for a hotel with luxurious amenities, and they give me an Economy class hotel with no pool and no spa. I would be so mad.

Recall is also high for Random Forest, but it would not be as important as Precision because it's better to miss out on some hotels in the tier than to get the wrong tier for a hotel.

PRICE FEATURE

HIGHER THE PRICE, HIGHER THE STAR?

Features Importance Order	
1) hotel_price	0.325076
2) hotel_bar	0.203393
3) hotel_restaurant	0.144000
4) hotel_room	0.131979
5) valet_parking	0.122698
6) hotel_rooftop	0.023183
7) hotel_fitness	0.017802
8) hotel_limo	0.014957
9) hotel_pool	0.013514
10) hotel_spa	0.003397

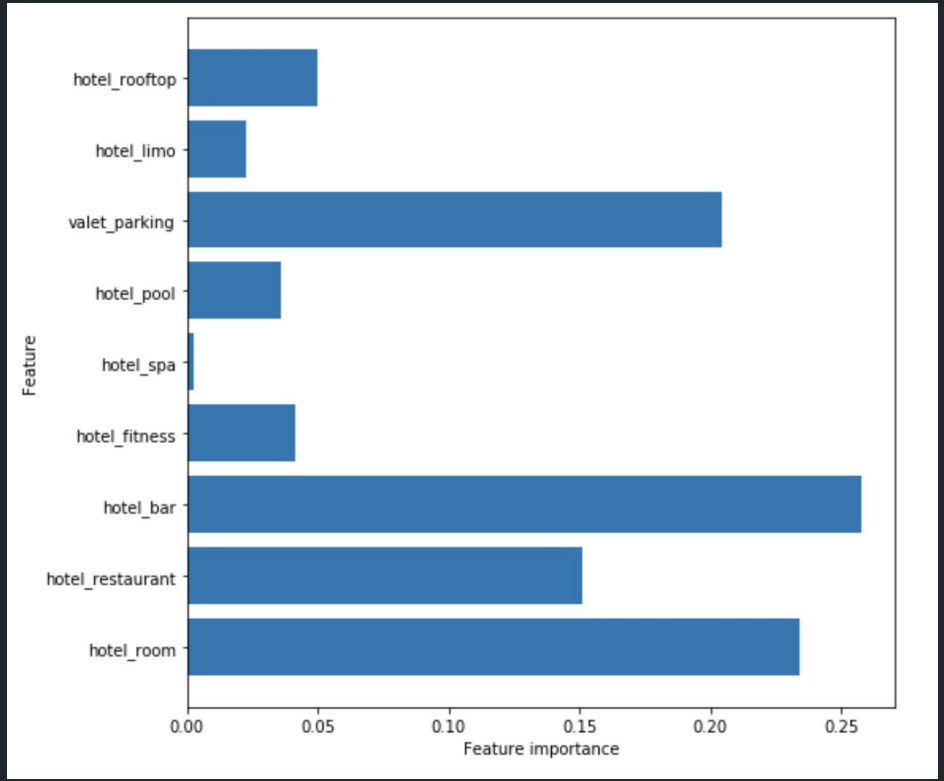


- Accuracy increases at least 4% for all models
- Precision increases at least 3% and at most 15%
- Recall increases at least 3% and at most 12%

→ SHORT ANSWER: YES!

Price Feature

```
Features Importance Order
 1) hotel_bar                  0.257722
 2) hotel_room                 0.234376
 3) valet_parking              0.204636
 4) hotel_restaurant           0.150918
 5) hotel_rooftop              0.049903
 6) hotel_fitness              0.041591
 7) hotel_pool                 0.035827
 8) hotel_limo                 0.022243
 9) hotel_spa                  0.002784
```



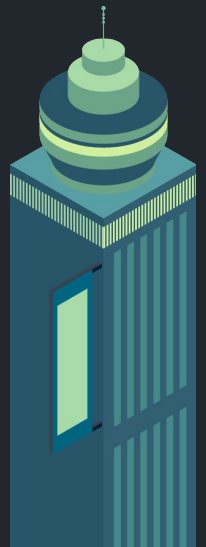
Without *price*, *valet_parking*, *hotel_bar*, and *hotel_room* are the main features affecting the tier of a hotel

conclusion

- Without the price feature, size of the hotel, aka the number of rooms a hotel has, and whether a hotel has valet parking, and bar are important factors in deciding which tier the hotel belongs to.
- I have better predictive power using the features with price. However, I think it is common sense to see that the more expensive it is, the higher the tier of the hotel.

what's next?

- More features to improve accuracy and precision
- Look into Location of a hotel:
  - New York vs. other cities
  - Urban vs. Suburban area



thanks!

facebook.com/Freepik

company/freepik-company

@Freepik_Vectors