Supervised and Unsupervised Learning Applications in Hotel Pricing Analysis

Sijie Xiang, Yijian Liu, Huiyi Gao

BOSTON UNIVERSITY

Introduction

Since its launch in 2000, TripAdvisor has become one of the most widely used travel and restaurant sites across the internet. Whether it is providing rental reviews, accommodation bookings, or other travel-related content, it has become a great platform for people to share their evaluations about hotels. This project uses hotel data scraped from TripAdvisor and seeks to analyze different trends among features trying to develop business insights using supervised learning and unsupervised learning techniques.

Goals

We have three main goals for this project:

1) Clustered five features of each hotel via K-Means, Hierarchical Clustering, GMM and see how clusters change over seasons
2) Compared 6-month hotel price trends via time series analysis and various distance functions.
3) Tried various methods and decided Utilize Gradient Boosted Regression Tree to be our optimal model to forecast future hotel price

The Data

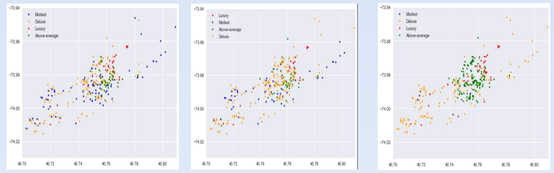
- All our data are scraped from TripAdvisor
- Part One:
  Six variables we initially scraped from TripAdvisor included Hotel Name, Hotel Price, Hotel Score, Hotel Star, Latitude and Longitude. However, since Hotel Name is not a numerical value and has no influence in clustering hotels, we exclude this feature in the data preprocessing part.

Index	Hotel Name	Hotel Price	Hotel Score	Hotel Star	Latitude	Longitude
0	Viceroy Central Par…	232	4	5	40.7646	-73.9785
1	New York Hilton Midt…	185	3.5	4	40.7624	-73.9796
2	Room Mate Grace	143	4.5	3.5	40.7574	-73.9838
3	Sofitel New York	299	4.5	4	40.7558	-73.9818
4	NYLO New York City	141	4	4	40.7822	-73.9804
5	Doubletree Hotel Metro…	99	3.5	4	40.7569	-73.9719

- Part Two & Three:
  Web scraped 20 hotels' prices for 6 months with corresponding dates(month/day/year). In part three, splitting data into training and testing sets, with first 4-month as training set and remaining 2 month as testing set

1) Clustering hotels in the City of NY and LA

- Assuming hotel price normally do not have drastic price movement in a season, we collected this data in March 3st, June 2st, September 1st, and December 2st, with each date represents a season.
- Initial assumption is most hotels to which districts classify should remain unchanged because hotels are very likely to react alike to the very same situations.
- Hotels clustering in December 2st in NY using K-means, Hierarchical, and GMM



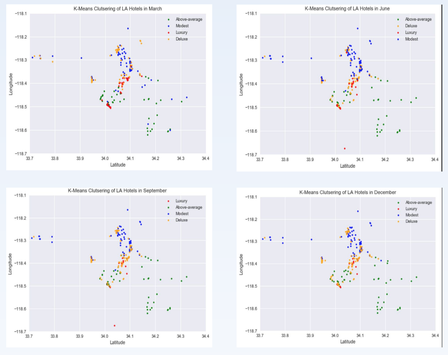
NY_Kmeans NY_Hierachical NY_GMM

how to manually label these clusters?

- Carried out a summary statistics in each cluster

	Price	Score	Star
0	141.60	4.04	3.86
1	127.35	3.79	2.90
2	199.58	4.30	4.16
3	699.73	4.55	4.89

- Hotels clustering in LA using just K-Means classification over seasons.

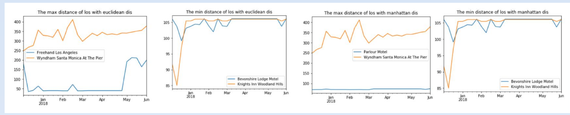


- As we can see from the graphs above, the shape and distribution of clusters mostly remain unchanged.
- Because shape and distribution of clusters mostly remain unchanged, it validates my hypothesis that most hotels to which districts classify stay the same over seasons.
- Inspired by hotel clustering, our groups would like to analyze trends of hotel prices in New York and Los Angeles at a period of time using times series analysis, which leads to part two of our project.

2) Time series for hotel prices in NY and LA

- Got 6 months prices of 20 hotels in NY and LA
- Used Manhattan distance and Euclidean distance to measure the similarity between hotels
- Made time series plots to illustrate the most alike and different hotels

The plot of LA are like:

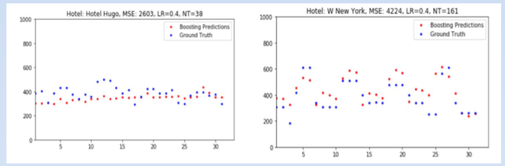


The plot of NY are like:

- When measuring the most alike hotels in LA and the least similar hotels in NY, we got the same results using different distance metrics. In other cases, results are different.
- When measuring the most similar hotels in NY using Manhattan distance, two hotels had the exactly same trend.
- When measuring the most similar hotels in LA, two series measured by Euclidean distance had very similar fluctuations while two series measured by Manhattan distance were different.
- No matter they were most similar or least similar, most hotels still followed the same trend in a six month period

3) Prediction for hotel prices in NY and LA

- Utilized the data from part 2 and construct 39 one-hot vectors feature(12month, 5 weeks, 7 days, 15 holidays) vectors based on the date.
- Tried CNN(Convolutional Neural Network) with LSTM(Long short-term memory) but didn't work out well.
- Tried pure Fully Connected Layer, result was not ideal.
- Used Gradient Boosting Regression with decision trees to train the model.
- Leveraged least absolute deviation as loss function, set mini sample split to 2 and maximum tree depth to 3.
- Adjusted many possible parameters such as learning rate and number of trees and chose the set that lead to minimum MSE



- As we can see from the figure, two pictures on the top are two hotels from NYC and the bottom are two hotels from LA.
- Hotel prices in New York had a larger fluctuation, which make it more difficult to predict. The Mean Square Errors are large.
- As hotel prices fluctuated over seasons, which means acquiring hotel price for all year would help improving the prediction accuracy.
- Our training data only contained date, but factors such as sports events, musical events could also impact the price.
- Predictive accuracy was decent, although only 6 month data was provided. Our Model can be further refined if more data were acquired

Conclusion

In this project we leveraged a widely variety of machine learning techniques to explore different topics in hotel business. While these three topics may seem independent, they are indeed closely related and their applications can be informed in practice to generate actionable business insights. For instance, part one result suggested that a business owner could gain comparative advantage in the market share by opening their new stores in "Luxury" districts to target on high-profile clients. Part two result suggested even though not all hotels react alike to the very same situations, most of them follow a 6-month trend on pricing. Part three gave us a statistical machine learning tool to forecast future hotel prices, later comparing forecasts with actual data to refine our model in the future.

Future Work

- Include more features that would impact hotel prices
- Scrape a year-long data to better capture hotel prices fluctuation
- try and test a better model for hotel price prediction