Hotel Recommendation System. Milestone Report.

Jonas Cuadrado

1. Defining the problem

According to Tripadvisor on their yearly travel report [1], most travelers spend most of the time allocated to book a trip deciding on the flights or lodging, both for vacation or work. While the flight seems to be decided mostly on cost and, in some cases, airline preferences; lodging factors in more parameters, including other people's ratings, and location within the destination area. So far, hotel search engines do not seem to personalize the results based on the user's preferences. Or, at least, they don't encourage users to rate their stays.

In this project I will create a recommendation system for hotels where the users will be encouraged explicitly to rate their stays to improve the recommendations for their next experience. Recommendation systems are quite common in our daily life, including music, tv, or social networks, with very successful outcomes.

2. Clients

Anyone who travels with some frequency can benefit from this project. In reality, the client would be an existing hotel search engine (Hotels.com, Trivago, TripAdvisor, …) who would implement the proposed work on their platform.

3. Data

To develop a recommendation system we need a set of users who rate items, and both parts are of upmost importance: the ratings alone do not give enough information. Unfortunately, no user-rating database is readily available from any public website (most likely due to privacy reasons), so I have simulated one based on academic research.

A dataset containing hotels from different cities in the US, New York, San Francisco, Chicago, and Las Vegas, was obtained from [2]. It contains the average rating for different criteria such as location, room cleanliness, friendliness of staff, etc., as well as an overall rating. These ratings were obtained from analyzing comments on the hotels from the TripAdvisor website on 2011.

To complete this dataset, I used a pythonic scraper (selenium) to obtain the star rating and price of each hotel for a given date, defaulted by the search engine (one month after the search date for Google). Due to the age of the dataset, some hotels are permanently closed, some others do not have their price available easily, so some blanks were stored as well. The total number of hotels is slightly under 1000. To analyze the data and look at correlations reported on the next section, about 40% of the hotels without price or star rating were dropped.

According to [3], the most important attribute to decide on what hotel to book is price. [4] and [5] evidence that other parameters can play a role, but at the end the price is always the major player. From [6] and [7] we can obtain an average amount spent in hotels as a function of household income and age, and use it to generate a virtual user of some hotels. I propose a rating system for those virtual users that mostly depends on the existing mean rating.

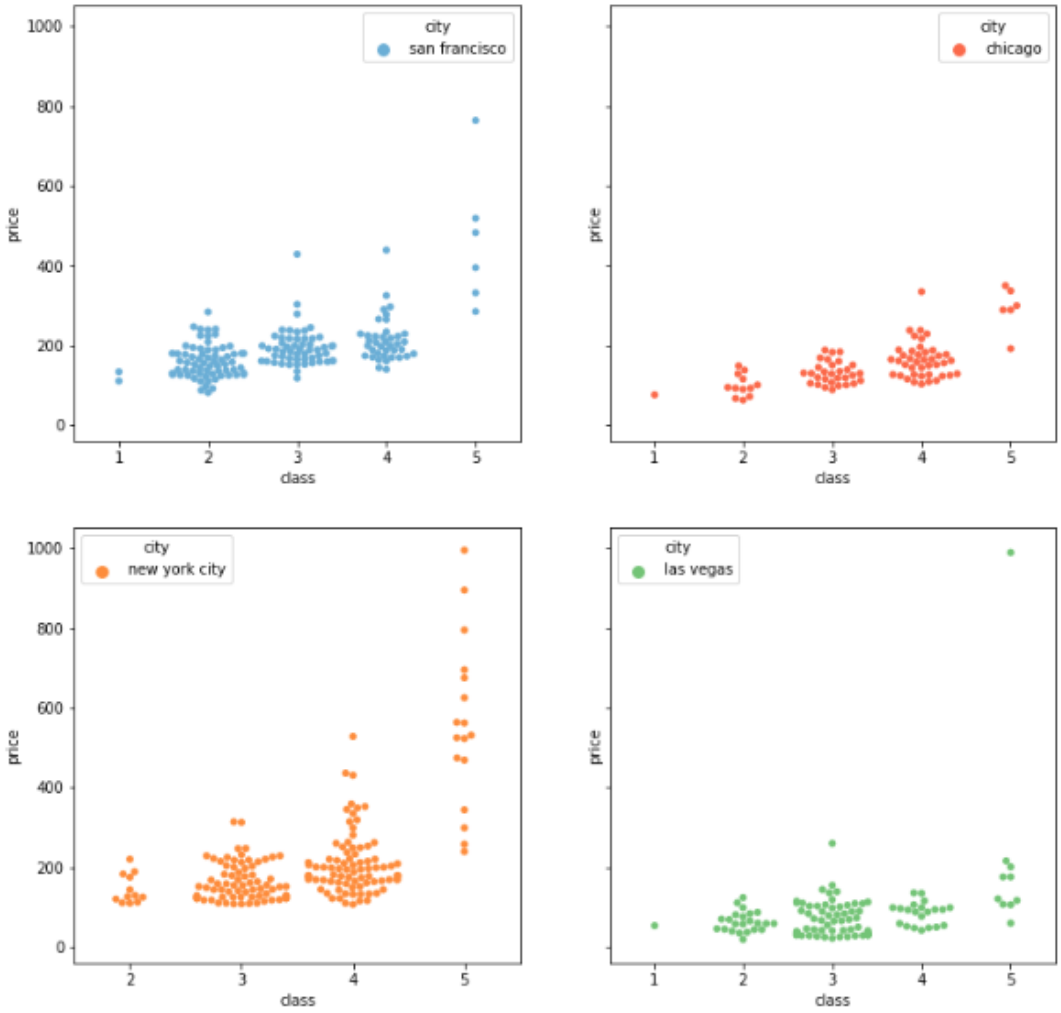
4. *Alternative datasets*

The process of generating users is essentially the only (legit) option we have to generate ratings. There are other hotel lists on which to build:

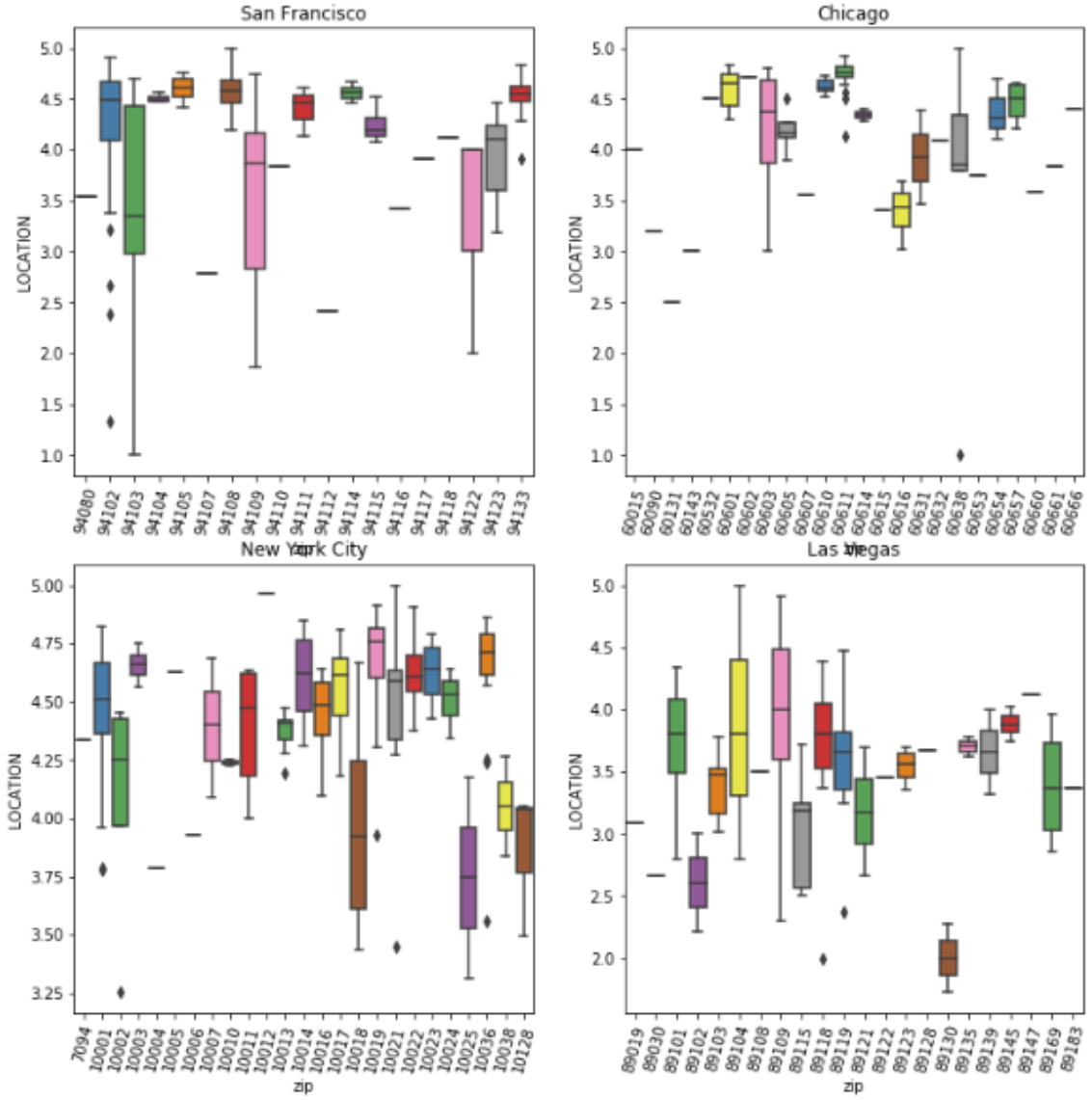
- Kaggle has a large dataset containing hotel reservations, but it doesn't have ratings per se. It may contain comments and reviews from which to extract a numeric rating.
- The US fire department has a list of hotels on which federal employees can stay, with information about their safety standards. No ratings are available whatsoever.

5. *Initial findings*

Regarding the hotels, we observe an evident correlation between hotel price and mean user rating, as well as the number of stars the hotel has. In the case of as Vegas, the correlation is weak because hotels make more profit from gambling than the stay itself, but there is enough evidence to accept the correlations. We also observe that the dependence is not necessarily linear, but the lack of points limits how well we can assess its non-linearity.



Regarding the location, some zip codes have a large variability on the users' rating. There are hotels with low values for location in the same zip-area than well-located hotels, which essentially tells us that the area on which a hotel is located needs to be defined to a level smaller than zip-code size.



Regarding the generated reviews, I have only created overall ratings, nothing regarding the other parameters existing on the dataset. To generate them, I first create a user, and assign him/her an age based on a uniform distribution and an economic power based on a gaussian distribution. Based on demographic data, these distributions capture the real trends. From there, I create an Ideal Hotel Price that users want to spend as the average between the chosen rows in the following tables:

Age	*IPH_age*
Under 25	$ 55.20
25-34	$ 118.25
35-44	$ 172.50
35-54	$ 207.00
55-64	$ 218.50
Over 65	$ 138.00

Percentile of household income	*IPH_wealth*
< 20%	$ 100.00
20% - 40%	$ 150.00
40% - 60%	$ 200.00
60% - 80%	$ 400.00
> 80%	$ 800.00

From here, a hotel is selected from a triangular probability distribution around the mean IHP. The rating assigned:

$$overall_{rating} = round(mean(rating) + f(distance\_to\_IHP) + white\ noise)$$

Assume the following:

- If: $distance\_to\_IHP < 50\$, f = 0$
- If: $50\$ < distance\_to\_IHP < 100\$, f = -0.7$
- If: $100\$ < distance\_to\_IHP, f = -1.3$
- $white\ noise < 1.0$

This generates a set of ratings whose mean is very close to the one reported on the original dataset. I have generated (initially) about 9000 reviews from about 2000 users

To recommend the hotels, different algorithms have been considered.

- Singular Value Decomposition performs poorly in time and provides poor results. The dataset is very sparse, so a dimensionality reduction is required to improve the performance.
- NMF behaves very similarly to SVD.
- kNN clustering outperforms all others in all aspects. A Cross-validation grid-search has provided the optimal values to train the model. The fraction of correct predictions is over 52% with such small dataset, most of the other approaches perform as a random recommender.

Here is a summary of the performance of each selected algorithm:

User-based recommenders

Name	RMSE	MAE	FCP	Time
SVD	0.549	0.417	0.523	0:00:03
NMF	0.621	0.461	0.521	0:00:04
SlopeOne	0.671	0.459	0.468	0:00:00
KNNBasic	0.616	0.436	0.491	0:00:00
KNNWithMeans	0.667	0.483	0.48	0:00:00
KNNBaseline	0.538	0.398	0.526	0:00:00
CoClustering	0.771	0.587	0.514	0:00:03
BaselineOnly	0.599	0.448	0.521	0:00:00
NormalPredictor	1.129	0.892	0.497	0:00:00

6. Next steps

There are some open questions. First, it would be interesting to look at the performance on unseen data instead of the cross-validation. If kNN is the selected algorithm, it would be interesting to quantify parameters like the optimal number of neighbors.

The platform is to be deployed on Amazon Web Services using django. It will contain a simple structure with a login- signup, a rating site and a finder site where the recommendations will be displayed both user-based (users similar to you liked …) and item-based (hotels similar to the ones you like are …).

7. References

[1] TripBarometer Travel Trends, (2016). https://www.tripadvisor.com/TripAdvisorInsights/wp-content/uploads/2018/01/Global-Report-US-Travel-Trends-TripBarometer-2016.pdf

[2] Kavita Ganesan and ChengXiang Zhai, (2011) “Opinion-Based Entity Ranking”, Information Retrieval.

[3] Dolnicar, Sara, Otter, T., (2003) “Which Hotel attributes Matter? A review of previous and a framework for future research”. http://ro.uow.edu.au/commpapers/268

[4] Radesh Rao Palakurthi, Sara J. Parks, (2000) “The effect of selected socio-demographic factors on lodging demand in the USA”, International Journal of Contemporary Hospitality Management, Vol. 12 Issue: 2, pp.135-142.

[5] Hokey Min, Hyesung Min, Ahmed Emam, (2002) “A data mining approach to developing the profiles of hotel customers”, International Journal of Contemporary Hospitality Management, Vol. 14 Issue: 6, pp.274-285.

[6] Bureau of Labor Statistics, (2010) “Travel: how much people spend?” https://www.bls.gov/spotlight/2010/travel/

[7] Minnaert, Lynn, (2017) “NYU US Family Travel Survey” https://www.scps.nyu.edu/content/dam/scps/pdf/200/200-4/200-4-16/P1718-0036-2017_Family_Travel_Survey.pdf