CAPSTONE PROJECT SUMMARY

Market Anomaly Analysis

Veronika Palotai

Project Sponsor: **Andras Kelemen**
Data Science Consultant
Archipelago International

CEU eTD Collection

Central European University
2020

Contents

CEU eTD Collection

1 Business Understanding

1.1 Company Background

Archipelago International is Indonesia's leading international hotel group and it is also the largest privately owned and independent hotel operator in Southeast Asia. They are the market leader in Bali, one of the world's most popular travel destinations [1].

In terms of numbers, Archipelago International is operating 136 hotels under 9 core brands in all major destinations across Indonesia, Malaysia and the Philippines with a further 100 new properties under development across Indonesia, the Caribbean, the Philippines, Saudi Arabia and Malaysia [2].

1.2 Project Objectives

1.2.1 Statement of the Problem to Be Solved

Occupancy and revenue trends are often difficult to explain. Unexpected drops and spikes happen from time to time. The client realized that it is not enough to rely on internal analysis, it is also important to observe their competitors to better understand the market and identify external factors that might contribute to out of the ordinary effects in the data.

Therefore, the goal of this Capstone Project is to identify the direct competitors of a given hotel and compare them to the hotel in question. The results have to be presented in the form of a Shiny dashboard tab so that they are in alignment with the analyses conducted by the company's data scientists. This Shiny based analytics dashboard should be capable of analyzing and visualizing trends and anomalies regarding the competitors of the chosen hotel.

1.2.2 The Project's Value to the Client

By making use of competitor data, the company will be able to better understand the trends in their key metrics which will allow for the optimization of their pricing and marketing strategy.

CEU eTD Collection

2 Data Collection

2.1 Identifying Competitor Hotels

A hotel in Bali which the project was going to focus on was selected. The first task was to define its competitors. Two metrics were proposed as a basis of identifying such hotels; proximity and similarity.

As for proximity, Google Places API was used to locate accommodations in the region. After supplying a search term, an API key and a place type, the API returned hotels in the specified region in batches of 20. 3 consecutive requests were allowed thus data was gathered on 60 hotels in the region.

Regarding similarity, Trivago's suggestions were scraped as its dynamic algorithm shows a range of relevant offers when a search is made for a particular hotel. Data was obtained on 125 properties, however, Airbnb listings were excluded as there is no public API to access Airbnb data and scraping the website proved to be much more difficult than ordinary travel websites. As a result, 47 observations remained.

Next, the two datasets were merged, duplicates were filtered out and IDs were generated for all the 102 hotels that remained. These IDs and the hotel names were stored in a table that was going to be the *fact* table of the star schema of the database that would be assembled out of the data gathered on these hotels via web scraping.

2.2 Data Enrichment with RSelenium

The next task was to obtain sufficient data on these hotels to identify the direct competitors of the hotel in question. Travel websites and platforms with data on several aspects of the properties were identified as proper data sources to do so. The technique used to gather this data is web scraping, the tool which made it possible is called Selenium.

The initial approach was to open the Selenium standalone server from a Docker container, however, it soon turned out that debugging is rather difficult this way as it was not possible to see what the browser was doing. Therefore, the server was run locally even though the ideal option for a company would be Docker.

Functions were written to scrape different hotel features from *booking.com* and *agoda.com*. Data was obtained on the reviews, ratings, prices, room types and amenities of hotels. Validation was done using the Levenshtein distance which is a string metric that helped to determine how close the scraped hotel name and the previously obtained hotel name were as it had to be made sure that the scraped data actually belonged to the hotel of interest.

In case of both travel platforms, the same set of features were obtained which allowed for the construction of a relational database with **PRICE**, **REVIEW** and **AMENITIES** dimensions.

3 Identifying Direct Competitors

The technique used to identify the direct competitors is clustering. Clustering is a great way to do so as it aims at finding homogeneous subgroups within the data such that data points in each cluster are as similar as possible. There are a few different clustering algorithms, in this project k-means clustering and hierarchical clustering were considered.

K-means clustering is a very well-known clustering algorithm, it is fast and easy to implement. However, the number of clusters had to be specified. The Elbow and the Silhouette methods were used to do so. 13 was identified as the ideal number of clusters.

A bottom-up type of hierarchical clustering algorithm was used. This type of algorithm does not require the user to specify the number of clusters, however, this makes it more complex and therefore less efficient. The resulting hierarchy of clusters is represented as a dendogram. This method also resulted in 13 clusters.

In order to decide which algorithm to work with, their goodness had to be evaluated.

CEU eTD Collection

Silhouette plots were chosen to do so. A Silhouette plot displays the assigned silhouette coefficient for each observation. The silhouette coefficient (S_i) measures how similar an observation i is to the other observations in its own cluster versus those in the neighbour cluster. S_i values range from 1 to -1. If the value of S_i for an observation is close to 1, it indicates that the observation is well clustered. In the other words, observation i is similar to the other observations in its group.

For hierarchical clustering the average silhouette width had a value of 0.56 which is quite good. However, several observations in clusters 1, 3, 4, 7, and 8 had a negative silhouette coefficient which meant that they were not in the right cluster.

As for k-means clustering, the average silhouette width was 0.55 which is basically the same as it was for hierarchical clustering. However, unlike previously, this time there were no observations with negative silhouette coefficients in any clusters which means that the k-means clustering algorithm did a better job at clustering the data. Therefore, the results of this algorithm were chosen.

Observations that belong to the hotel in question, were found in three clusters out of the thirteen. As the price of the different room types varied a lot, it was not surprising to find observations belonging to the same hotel in different clusters despite review score and amenities being the same for every record of a particular hotel. With this in mind, it was also unsurprising to find that the main difference between the observations in the three clusters in focus was room types and consequently prices. There was a cluster with low prices (7 USD/night - 37 USD/night) and simple rooms like *studio room* or *standard double room*, one with somewhat higher prices (38 USD/night - 63 USD/night) and more equipped rooms (e.g. deluxe rooms or rooms with pool access) or rooms that can accommodate more people and another one with high prices (64 USD/night - 95 USD/night) and luxurious rooms like *deluxe king room* or entire apartments or villas.

4 Shiny Dashboard Development

The final deliverable of the project was a Shiny dashboard where an analysis on the direct competitors is displayed. A dashboard with two tabs was designed, one of them provides a general overview on the direct competitors, the other one enables the user to select a competitor hotel and compare it to the hotel in question. Both tabs allow for filtering the data, contain six KPIs and four different plots with insight on the direct competitors.

The first tab, **ALL COMPETITORS** has five filters. Users can filter to the travel website and price category of their interest. Moreover, they can select specific room types and amenities and it is possible to filter the data according to review categories as well. The second tab, **1-ON-1 COMPARISON** has three filters. Users can filter to the travel website and price category of their interest. Furthermore, from a drop-down menu they can select the hotel to which they would like to compare the hotel in question to.

5 Future Improvement

Although the project reached its goal and the final deliverable is ready, there is plenty of room for improvement.

CEU eTD Collection

- **Code structure**

  The ultimate goal of the client is to integrate this project into their existing dashboard and automatically pull data into this dashboard which means that any code must be structured in such a way that it can be easily applied on any hotel of interest.

  It should also be ensured that the possible changes in the structure of websites do not affect the data collection process.

- **Airbnb listings**

  As it was mentioned, around two-thirds of the similar properties that Trivago suggested were listed on Airbnb. In this project these were disregarded, however, should a workaround be found to tackle the issue of scraping this valuable data, the quality of the analysis would be significantly improved.

- **Other travel platforms**

  Another possibility to gather more data is to scrape other travel platforms which are not necessarily Airbnb. These could be TripAdvisor or Exepedia for example. Doing so would also contribute to having a more insightful analysis.

References

[1] https://www.archipelagointernational.com/en, Official Website of Archipelago International

[2] https://www.linkedin.com/company/archipelago-international-hotels-resorts-and-residences/, Official LinkedIn Page of Archipelago International

CEU eTD Collection