CS5224: Cloud Computing

Assignment 2

Interim Report

CentsTrip:
Beautiful Singapore, Satisfactory Cost

By Group 5

Yu Yingchen (A0159338U)
Ding Wenjia (A0105515R)
Lu Yifan (A0159305E)
Zhu Longwei (A0148601L)
Huang Mengying (A0159188M)

Contents

1. Executive Summary

Over 16 million tourists travelled to Singapore during 2016, which are three times people than local Singaporeans. We can easily understand that Singapore is one of the hottest tourist city in Asia due to the pleasant city view and the open culture. Tourism contributed 4% income out of the whole annual income in Singapore in 2016, which is around 23 billion Singapore dollars. Such a big industry gives Internet product a wonderful opportunity to get into it.

In fact, Singapore is the most expensive city surveyed by EIU, 2016. The cost of travel to most cities in Southeast Asia just half or even less than the cost to Singapore. This is a barrier for people who travel to Southeast Asia with limited budget to travel to Singapore. The survey we have taken to discover the most popular trip advising websites and apps, they treat everyone indistinctly so that people with special needs can hardly find useful traveling information.

In our product, we aimed at the tourist who always wants to enjoy the highest service quality with minimal cost and those who have specific budget to travel. The website we want to build contains two major parts: the cost performance analysis and recommendation system for users. In the first part, we plan to implement on the Bluemix platform to calculate the cost performance in our own algorithm to weighting accommodations and attractions. The data that we collected are from Singapore government website and other sources like Airbnb, which is the most famous company who provide b&b service worldwide. In our own survey, there is no such website or travel agent in Singapore starting business with Airbnb. That means we can offer a competitive price in accommodation to the individual visitors. Then we will let users to give their preferences on the website so that we can recommend suitable plans to them. Finally, everything will be represented by a simple geometric graph, which is very user-friendly and brief to select.

2. Business Case Identification

2.1 Motivation

Looking for the travel-related website in Singapore, the general services they provide are introducing attractions, booking hotels and selling tickets. With the rapid economic growing in Asia, more and more young people beginning to travel aboard. Singapore, as a tourist city, has faced the problem with reduction of visitor arrivals from 2014.

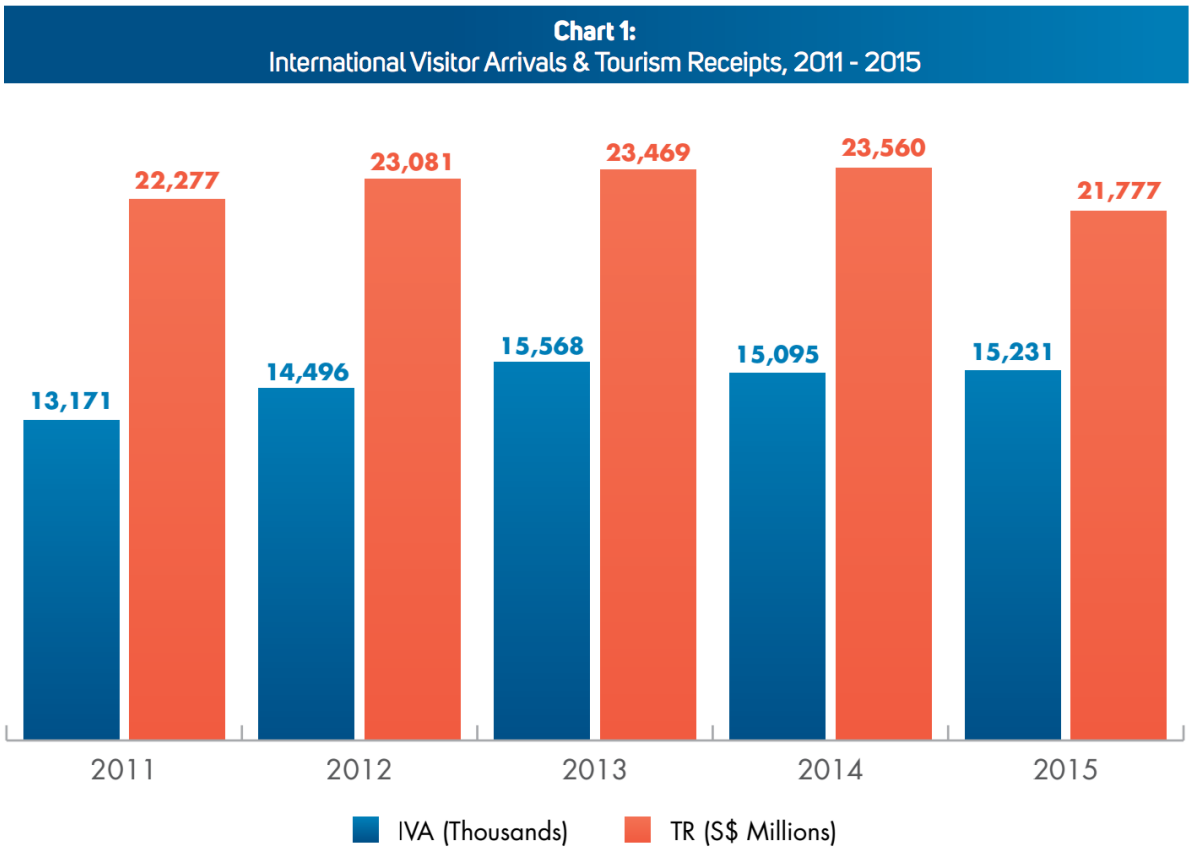

Figure 1: International Visitor Arrivals & Tourism Receipts in Singapore

Looking for the travel-related website in Singapore, the general services they provide are introducing attractions, booking hotels and selling tickets. With the rapid economic growing in Asia, more and more young people beginning to travel aboard. Singapore, as a tourist city, has faced the problem with reduction of visitor arrivals from 2014.

2.2 Target Users

As we mentioned previously, the user that we aimed at is the one who has budget in travel. In our assumption, the young people with budge, for example, young couples and teenagers travel solely, are the potential consumers of our product. They prefer traveling individually, but have clearly budget. These people like to search every message online to find lower travel cost, but also assure quality. For most of the travel advised sites, the great quality is based on the amount of payment. But for our product, we will give some suggestions according to our distribution channels and algorithms, to lower the cost and assure the quality.

3. Business Model

“CentsTrip” is a travel recommendation application for tourist who plans to travel to Singapore. It aims to provide the recommended travel plan with minimal expenditure for users based on their preference and budget.

3.1 Services to Be Developed

Two web-based SaaS services will be designed in this project:

a) Cost Performance Analyzer

It is important to recommend accommodations and attractions in good quality with rational price for users. The first service focuses on analyzing the cost performance of accommodations and attractions. The cost performance analysis of accommodations takes account of five factors, which respectively are rating, price, user-specified budget, distance between accommodations and attractions, and distance between accommodations to transportation junctions like MRT stations. For attractions, the analysis would be based on rating, ticket price, distance between user-selected accommodations to attractions, and user preference. Afterwards, the recommended accommodations and attractions would be determined based on the ranking of cost performance.

b) Recommendation System

The second service is a recommendation system which allows user to input their budget and preference. It will pass the user inputs to first service and exhibit the recommendation results. Users need to input their number of visitors, the duration, budgets of their tourist and their attractions preference. Accommodations recommendation will be provided after the user inputs is submitted. After users selecting their preferred accommodations, attractions recommendation and the detailed information will be listed to provide a more intuitional and comprehensive view for users. The expected cost will also be calculated for users.

3.2 Services Comparison

The existing service providers for traveling recommendation focus more on integrating and exhibiting information, such as discounts and other tourists' review. In this case, they do not offer customized recommendation for target users. Unlike the existing ones, CentsTrip is a unique application that provides users the customized travel plan. The detailed comparison between CentsTrip and well-known traveling service providers are tabulated in Table1.

	Accommodation Search	Attractions Information	Flight Search	Food Information	Customized Travel Plan
TripAdvisor	✓	✓	✓	✓	✗
Booking	✓	✓	✗	✗	✗
Agoda	✓	✗	✓	✗	✗
Ctrip	✓	✓	✓	✗	✗
Airbnb	✓	✗	✗	✗	✗
CentsTrip	✓	✓	✗	✓	✓

Table 1: Services Comparison

3.3 Revenue Model

As a web-based service, CentsTrip will be totally free for users. It attracts potential users by providing a good quality recommendation service. Subsequently, with the website traffic generated by users, customized advertisement such as accommodation discounts or ticket sales of attractions could be exhibited on CentsTrip. Generally, the revenue model of CentsTrip is to attract users and charge the advertisement fee.

3.4 Cost Analysis

As a SaaS service, CentsTrip has much less costs compared with other on-premise services. Firstly, the service is implemented on cloud platform such as Amazon Web Service or IBM Bluemix, which means up-front cost of setting up the services can be eliminated or significantly reduced. Secondly, due to the elasticity of cloud computing, our SaaS service can rapidly respond to the business cycle, which can help to reduce cost without limiting the capacity. However, those on-premise services do not possess this elasticity, it will take much more costs and time for them to scale up their resources to meet the demand. Lastly, in this competitive market, it is crucial to develop and launch our application fast enough to take the first mover advantage. Using cloud-based platform offers the chance to greatly shorten the development cycle, because its virtualization characteristic allows our SaaS service to be developed without considering setting up hardware, infrastructure and etc.

Above all, the business model of CentsTrip aims to use the minimal cost to provide users with recommendation service and benefit from website traffic and advertisements.

4. SaaS Architecture & Implementation

4.1 Saas Architecture Design

Figure 2 below is the overall architecture design of our SaaS. The application is hosted in IBM Bluemix cloud.

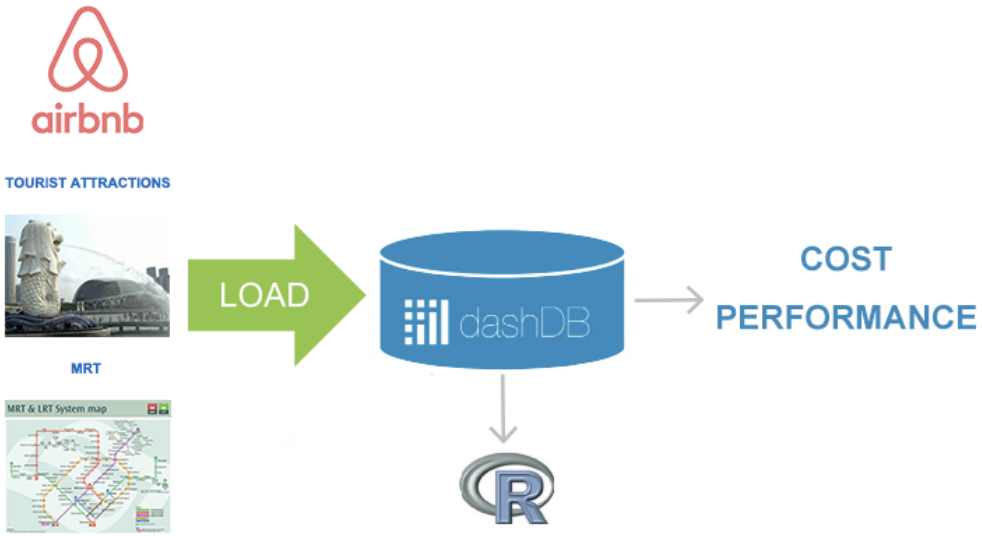

Figure 2: Overview of the SaaS architecture

In this phase, we acquired five sourced data (i.e., hawker center, food court, tourist attractions, Airbnb and MRT data) from https://data.gov.sg/ and other websites. Then several necessary data preprocessing work were done before loading these data into DashDB on IBM Bluemix. R script is the primary tool for us to analyze the cost performance of b&b and attractions. The two parts of our work will be explained in Section 4.2 and 4.3.

4.2 Data Acquisition and Preprocessing

As stated above, two main datasets of CentsTrip are Airbnb and Tourist Attractions, we also collected data source from https://data.gov.sg/ to enhance the value of our SaaS.
(1) We used the unofficial APIs for Airbnb which are represented on http://airbnbapi.org/ to crawl the data that we need. The Python script was used to get the information, including name, location, description, rating and so on. All records were written in csv files for further processing.

(2) We downloaded the tourist attractions data in KML format which was based on XML that could store and show the geographic information in Google Earth related applications. We

imported the data into Google Fusion Table to extract the longitude and latitude of attractions and added other attributes through the web crawling and regular expression extraction.

The following points are the additional fields in the Tourist Attractions table:

- Category: represent the type of attractions, including architecture, heritage, arts, museums, memorials, places-to-see, nature-wildlife and recreation-leisure.
- Ticket-price: the price for an adult in each place.
- Popular Attractions as in Lonely Planet: there are only two values, 1 corresponds to the popular, and 0 is versa.
- Rating: Tourists' review on Google Map.
- Opening_hours: an important attribute for the next recommendation system.
- Description: abstract introduction for attractions.

(3) For the other three datasets, Google Map API was adopted to acquire the longitude and latitude of MRT. And we wrote Python script to crawl and convert the information of the popular food courts in Singapore and combine it with the hawker center data for the food recommendation.

4.3 Data Analysis

In the first application of CentsTrip, we analyzed the cost performance of b&bs and tried to understand which attributes will be the important ones to be considered in the recommendation phase. In case that several different factors will contribute to the popularity of b&b, we calculated the total distance between b&b and several famous attracted places and the distance with the nearest MRT station. The rating of b&b was also an essential factor that was concerned by us.

Before fitting data into our model, we adopted statistic method to filter outliers. In statistic, if a value is larger than third quartile (152) adds 1.5 times interquartile range (IQR=Q3-Q1) or less than first quartile (60) deducts 1.5 times IQR, it will be treated as an outlier. So we chose the price falled between 0 and 290. In our analysis, we performed a weighted least square estimator in our regression model. Statistically, we have no need to use the regularized regression in our model because datasets have much more observations than the number of predictors (n > p). Besides, we have no idea whether the price of a b&b is reasonable or not, so it may cause problem if we use the simple linear regression model and put each observation with same weight as those unreasonable observations may lead to bias.

In general, we first fitted a linear regression model and find out the residuals in each observation. Then we used $\frac{1}{|residual|^2}$ as the weight for each observation because this is a general way in statistic for selecting weights. We find that p-value of the distance with nearest MRT is much larger than 0.05 meaning it not significantly different from 0, and the estimator for this predictor is larger than 0 which violates our common sense. So we decided to remove this attribute from our model and perform regression again.

After we get our regression model by weighted least square estimator, we could predict a price by fitting attributes into the model. And we call the predicted price the intrinsic value of the b&b. By subtract intrinsic value from price, we could get the result whether a b&b excess its value by a certain margin or it's cheaper than its value. After we sort the cost performance of b&b, we could perform our recommendation. The below figures show the comparison between the real and predicted price, the second one was the detailed information of b&b sorted by the value of cost performance in decreased order.

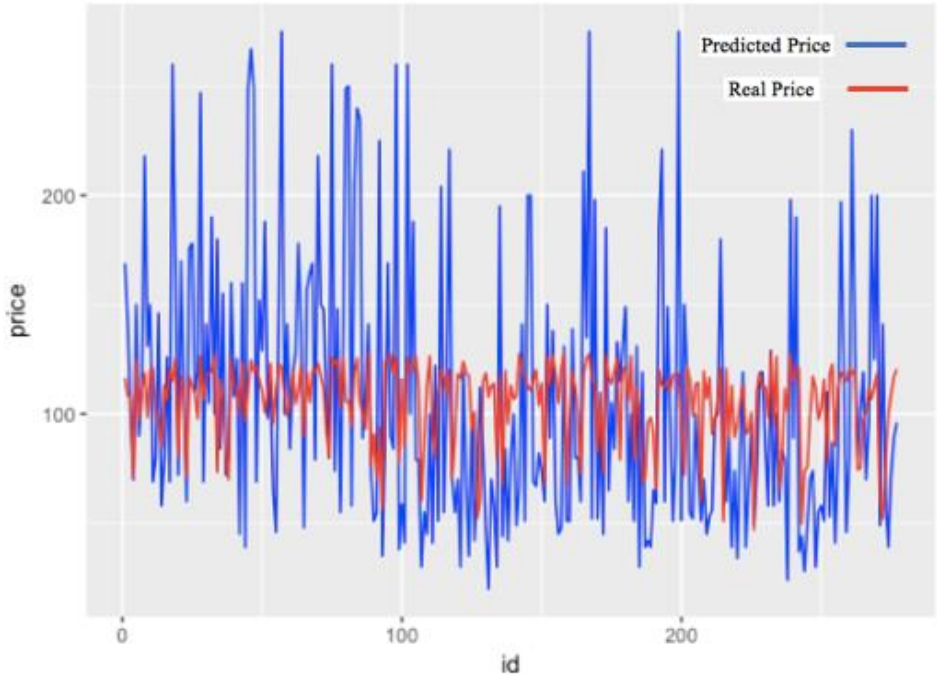

Figure 3: Price vs Predicted Price of Accommodations

	ROOMID	PRICE	LATITUDE	LONGITUDE	costPerformance
11	10030097	20	1.311816262	103.8617381	-4.298338e+00
12	16199811	24	1.31846307	103.9084893	-3.274425e+00
15	12367388	30	1.308921359	103.8593252	-2.797774e+00
18	8932274	30	1.288307414	103.7922988	-2.760737e+00
17	17426064	30	1.35235977	103.8806396	-2.455660e+00
21	16865763	35	1.287004695	103.8469107	-2.292332e+00
22	10818230	37	1.316419517	103.8573384	-2.279018e+00
27	16608371	39	1.312948255	103.8845583	-1.917970e+00
19	7388897	34	1.314806669	103.8918444	-1.781380e+00
13	11933991	28	1.335725446	103.9357015	-1.668248e+00
16	10377877	30	1.312840424	103.9398309	-1.635040e+00
48	16407748	49	1.321182042	103.8678657	-1.577801e+00
39	16538491	45	1.305942455	103.8859324	-1.541428e+00
195	14792764	53	1.293150052	103.8268237	-1.522229e+00
29	9963524	39	1.347407128	103.849093	-1.521164e+00
26	16761767	39	1.312552822	103.8869599	-1.487504e+00
47	9635192	48	1.284709572	103.8435968	-1.465558e+00
40	13564613	45	1.317535743	103.8914566	-1.462219e+00
44	11591661	46	1.315534812	103.8313379	-1.450619e+00
24	7843168	39	1.314217248	103.9048094	-1.449579e+00

Figure 4: Rank of Accommodation Cost Performance

5. Economic Factors

5.1 Economic Benefits and Key Consideration

There are many economic benefits carried out by our application. Since the application supports users to search for accommodations and tourist attractions as the travelling plan, it contributes to the tourism activity in Singapore associated to economy and improves the flow of spending in the region. We analyze the related data from government statistics, and model the plan according to the users' preferences and requirements. In this measurement process, the population size, the distance between destinations, the budget quantity, and the duration of time are the key factors that change our estimations and plan. When the users access our website, there will be profit generating from the website due to the website traffic. The placement of advertisements on the website could also gain economic benefits, such as discount information and attractions service, so that we can generate income and revenue from them.

Depending on the various terms in a year, the number of users' transactions may be different, for example, summer holiday is a peak season for people to travel and relax. Therefore, there will be a request fluctuation for the website, and we need to deal with the increase of demand. Accordingly, the price of different facilities may also fluctuate, the more people coming in, the more expensive the price may be. The availability of the travel agencies is also a consideration of our system. Since different attractions' and food courts' opening hours are not the same, we will need to analyze them and promote a rational travel plan for better performance.

5.2 Tradeoff between Cost and SLA

At current stage, we provide our service for free. Different from other on-premise services, we reduce the cost to buy the IT resources for building the website and run the application. We guarantee our service quality using IBM Bluemix as the platform to develop, which saves time and money spending on preliminary stage setup and machine supply. We use dashDB Entry plans which cost zero dollar to fit the user's needs. The availability of our service can reach 99.95% on the basis of IBM Bluemix platform. In terms of response time, we adopt advanced searching algorithm to analyze the function and retrieve data from database, so that the website can display the results in a fast and robust way. Furthermore, since we already obtain the preference of accommodations according to their cost performance ranking in our initial service, we are able to quickly achieve the accommodation information with daily plans for users for the recommendation process. Additionally, we collect data from reliable sources, mostly from government's statistics, which guarantees the reliability of our services and results. Nevertheless, in the latter stages of development, we may want to charge a considerably amount of fee from users for the cost of updating the data storage, and providing multiple deployment of the application in order to increase the availability and reliability based on SLA.

6. Conclusion

In the interim stage, we have successfully preprocessed all the data for our project. Some datasets that we found are from Singapore government data website, for example, the hawker center information and Singapore attractions information. Others are crawled by ourselves on various websites. Then we implemented the algorithm to evaluate the cost performance of accommodations in Singapore on Bluemix. We used DashDB to store the data and run our R script. Finally, every accommodation was given a rank representing their cost performance.

In the next stage, we will start to establish cost performance analysis model for Singapore attractions based on rating, ticket price, distance between user-selected accommodations to attractions, and user preference. Afterwards, a web-based SaaS will be implemented on cloud platforms as a user interface to acquire user's inputs and display the visualized recommendation results.