

13th Global Forum of Tourism Statistics

Theme 3: Using Big Data for Big Statistics

Data Mining in Tourism Data Analysis: Inbound Visitors to Japan

Ms. Valeriya Shapoval, PhD student, Rosen College of Hospitality Management, University of Central Florida

Dr. Morgan C. Wang, College of Graduate Studies, University of Central Florida

Dr. Tadayuki Hara, Rosen College of Hospitality Management, University of Central Florida

Mr. Hideo Shioya, Senior Research Fellow, JTB-Foundation

1. Abstract

Japan is gradually becoming an important touristic destination. Inbound tourism arrivals are dramatically increasing for the past decade. According to the JNTO tourism in 2000 total arrivals to Japan was 4,757,146 people, 2012 total tourist arrivals to Japan was 8,358,105 and estimated 10,363,922 arrivals in 2013 with total increase of 24 % from previous year. As tourism becomes more and more important to a Japanese economy interest is growing in the academia to address questions of motivation, intention to return, positive word of mouth, and satisfaction with Japan as a touristic destination. Nevertheless, research on inbound tourism to Japan is highly limited. According to Uzama (2009) Japanese marketing campaign “Yokoso Japan” although brought some results were mainly unsuccessful in advertising Japan as desirable touristic destination and in spite of the government interest in promoting Japan it did not go beyond of simple promotion. Omura and Fukushige (2010) in their research of international tourists to Japan were looking in to the difference between first time visitors and repeated visitors to the Kansai area of Japan found that first-time visitors are interested in sightseeing, while repeated tourist are more involved and are interested in participating in the events.

However, all limited research done about international tourism to Japan is either qualitative or using classic statistics with the limited sample size. None know to us research is done using big data for understanding tourists behavior that are data driven and using data mining techniques. Only limited research has being done in a tourism field and none known to us was applied to Japan as tourism destination. Hence, the purpose of this research is to contribute to the general body of knowledge in tourism industry of Japan using advance statistical techniques of data mining, which allow us to gauge visitors’ behavioral patterns. As a result of our analysis we found that there are two distinct groups in satisfaction, most important factors are ticket price, experience with Japanese food, shopping and how many times visited Japan ion the past. Furthermore, there are two distinct groups in satisfaction Asian and non-Asian groups with different preferences and purpose of visit. In behavioral intentions desire to experience in

the future some aspects of Japan such as hot spring are motivators for return. Tourists are mainly traveling with family and ticket price is an important factor.

2. Data Mining

Development in technology and computer power made it possible to access immense quantity of data, making use of computer technology in decision making a wide spread phenomena in business and industry. Currently, online companies and financial industry are using data scientist to find actual consumer purchasing and behavioral patterns rather than then asking customer perceptions through surveys, which is also limited by question asked in the survey. Consumer perception of behavior and actual behavior could be quiet different, whereas understanding actual behavior, from extensive consumer data, will narrow this gap and may lead to a better business decisions. With an acceptably accurate learning model one can not only understand, but also predict expected values in the tourism industry. For example, tourism agency may choose to use their visitor's database to predict future arrival and pattern of consumption. Given a new profile agency will be more prepared to a needs of a visitors with better marketing material, establish necessary collaborations with other agencies such as transportation, lodging and other destinations with in a country to improve country's desirability for future touristic visitations, based on actual behavior (Apte and Weiss, 1997).

Managing very large data requires skill that is different from classical statistical tools. Data mining manages such problems by efficient summaries of large amount of data, identifying patterns and relationships of past data and construct predictors for the future. Classic statisticians have well established tools for such things. Many statistical models for explaining relationships and patterns in given data and it is therefore tempting to think of data mining as a branch of statistics. However, data mining has its own merits and working with data sets that a very much larger than regular statistical data set and larger scale. There are differences in approach to modeling, compared to classic statistics data mining pays less attention to the large-sample asymptotic properties of its inferences and more of "learning", including complexity of the modeling and computation required by large data sets. Of cause both classic statistics and data mining are a like in drawing inference from data. Unlike classical statistics data mining is more tolerant toward discreet-valued variables and seeks to minimize a loss function expressed in term of predictor error, where minimization is achieved by cross-validation (Hosking, Pednault & Sudan, 1997).

One of the oldest definition of data mining is "The non-trivial extraction of implicit, previously unknown, and potentially useful information from data (Frawley et al., 1991). Data mining uses machine learning algorithms to find patterns of relationship between data elements in large, noisy, and messy data sets, which can lead to actions to increase benefit is some form (diagnosis, profit, detection, ect.) knowledge discovery in data (Nisbet, Edler and Miner, 2009 p. 17). As data mining evolved new definition was proposed "Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potential useful, and ultimately understandable patterns in data (Fayyad et al., 1996)

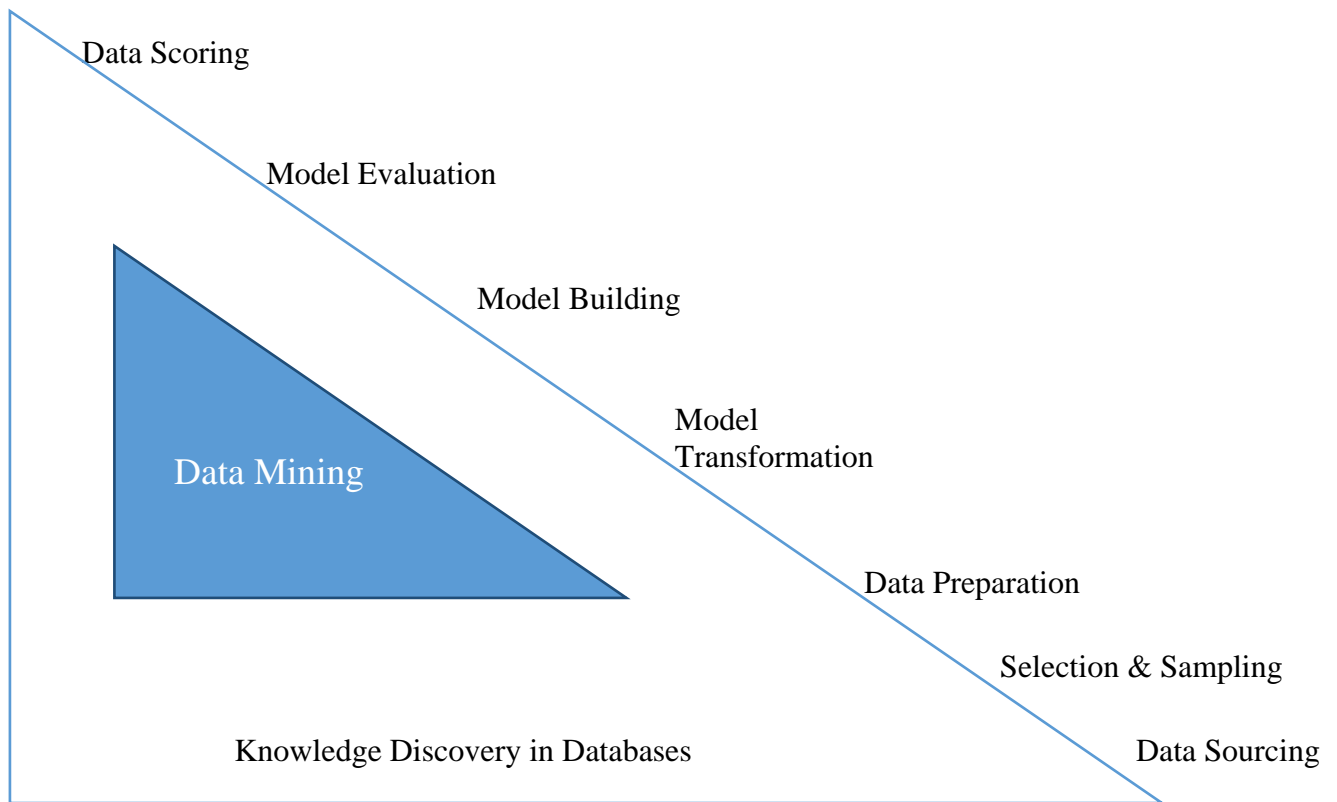


Figure 1 Relationship between data mining and knowledge discovery ((Nisbet, Edler and Miner, 2009)

For development of model-theoretic approach in data mining number of criteria were proposed by Mannila (2000), which include:

1. Model typical data mining tasks
2. Describe data and the inductive generalizations derived from the data
3. Express information from variety of forms of data (relational data, sequences, text etc.,
4. Support interactive and interactive process
5. Express comprehensible relationships
6. Incorporate users in the process
7. Incorporate multiple criteria for defining what is an 'interesting discovery (Nisbet, Edler and Miner, 2009 p.19)

One big difference between classical statistics and data mining is that classical statistics has large subjective component, predictive model is known and main goal is to estimate parameters and/or confirm/reject hypothesis. On the other hand correct model is unknown, in fact goal of the analysis is to discover correct model even if it is not correct. In classical statistics models must be specified, in data mining on the other hand series of competing models will be specified and selected based on data examination. This preference ordering addresses issue of

overfitting. There are many other things can be said about differences between classical statistics and data mining techniques, however this is not a purpose of this paper. In summary, one can say that Statistical learning (Data mining) is much more manageable when there are no restrictions placed on the model for a given data, in other words where analysis are data driven and complexity of given machine learning are dependent on underlying distribution according of which we desire to learn (Hosking, Pednault & Sudan, 1997).

Although there is extensive number of data mining techniques that evolved over the years, decision trees, neural networks, regression analysis, text mining, association rules, clustering and others. Working with large data sets follows certain general steps or rules shown in the Figure 1.

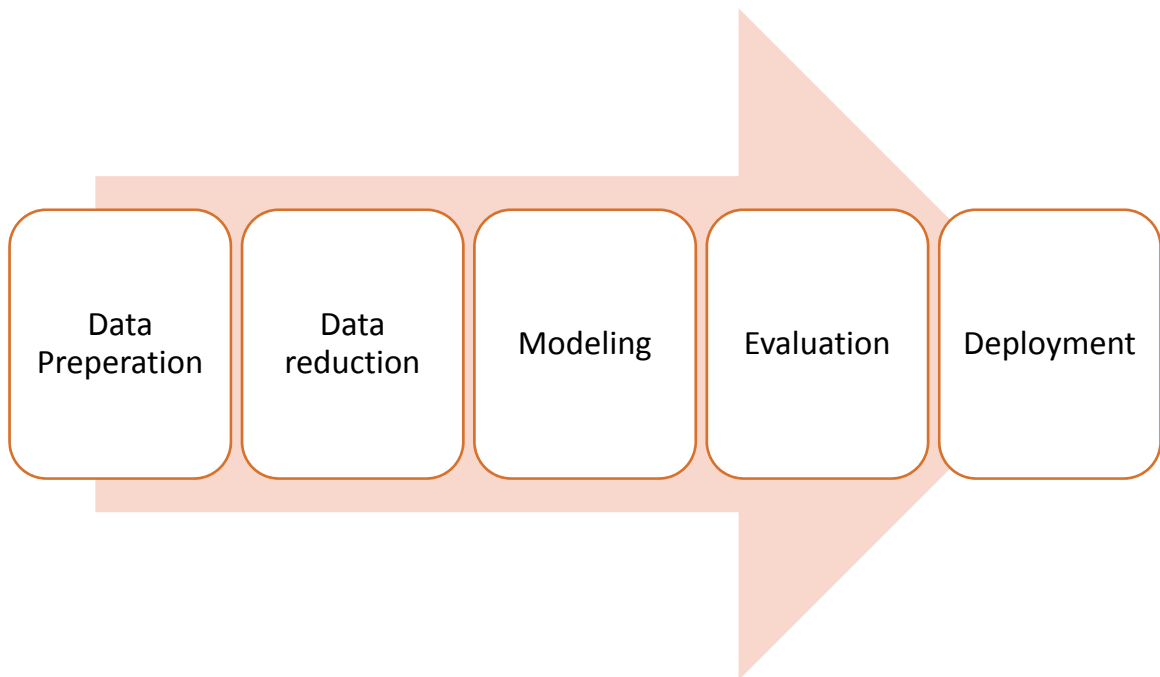


Figure 2

Due to the very large scale of the data, data preparation and reduction are essential steps in data mining. Unlike data sets used in classic statistics, it is impossible to “eyeball” data mining dataset where variables could be counted in hundreds and observations in millions, because just like in the classic statistics data mining quality of the prediction and accuracy of a model depends on quality of data. Furthermore variables should be reduced and manipulated into analytical data set. Some of the modeling specific for data mining techniques will be briefly described below.

2.1 Neural Networks

Neural networks (NN) are capable to generalize and learn from data mimics, which cane be in the way related to a one learning from one’s own experience. Draw back of the technique is results of training NN are weight that are distributed through network and do not provide valid

insight as to why given solution is valid. NN is a good tool for prediction and estimation problems. Training NN builds a model that can be used to estimate target value for unknown examples. The whole process of training is actually process of adjusting weights in the data to arrive at best combination if given weights for making wanted prediction. NN starts with random weights with poor performance, however by processing, adjusting and training it reduces overall error and increasing performance by approximating the target values (Berry and Linoff 2004).

2.2 Decision Trees

Decision Trees (DT) are form of multiple variable analyses.”... it is a structure that can be used to divide up a large collection of records into successfully smaller sets of records by applying a sequence of simple decision rules (Berry and Linoff 2004 p. 6).” Another definition of DT provided by Nisbet, Edler and Miner, (2009) “DT is a hierarchical groups of relationships organized into tree-like structure, starting with one variable (like trunk or an oak tree) called a root node (p. 241). Root node than is split into multiple branches using a split criteria. Each split is defined in terms of impurity measure, reflecting how uniformed resulting cases are. Each split node is referred as parent node and following splits are called child node. Split continues until final or terminal node with minimum number of cases. For example in Figure 3, decision trees are used to indicate pattern of travel behavior based on age, gender and marital status.

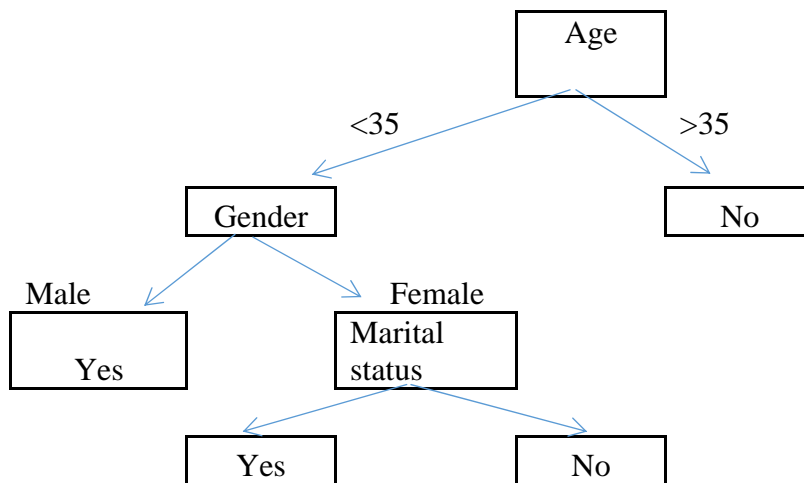


Figure 3

DT are very appealing analysis due to their relative power, ease of use, robustness, ability to handle ordinal data and ease of interpretability. It is a collection of one-cause one-effect relationship presented in the form of tree. DT try to find strong relationship between input and target variable, when set of values is identified that has strong relationship to a target, than all of those values are grouped into the bin that form branches of a DT.

2.3.1 Impurity-based Criteria

In many cases in DT split is done according to the value of single variable. Most common criteria for a split is an impurity based split.

Given random variable x with k discrete values, distribution according to

$$P = (p_1, p_2, \dots, p_k)$$

Is an impurity measure is a function $\phi: [0,1]^k \rightarrow R$ that satisfies the following conditions:

$$\phi(P) \geq 0$$

$\phi(P)$ is minimum if $\exists i$ such that component $p_i = 1$

$\phi(P)$ is maximum if $\forall i, 1 \leq i \leq k, p_i = \frac{1}{k}$

$\phi(P)$ is symmetric with respect to components of P

$\phi(P)$ is smooth (different everywhere) in its range

Probability vector has a component of 1 (variable x gets only one value), than variable is definitely pure. Other the other hand, if all components are equal, the level of impurity reaches maximum. Given training set S , the probability vector of target attribute y is defines as:

$$P_y(S) = \left(\frac{|\sigma_{y=c_1} S|}{|S|} \right), \dots, \left(\frac{|\sigma_{y=c_{|dom(y)|}} S|}{|S|} \right)$$

The goodness-of-split due to discrete attribute a_i is defined as reduction in impurity of the target attribute after partitioning S according to the values $V_{i,j} \in dom(a_i)$

$$\Delta\Phi(a_i, S) = \phi(P_y(S)) - \sum_{j=1}^{|dom(a_i)|} \frac{|\sigma_{a_i=v_{i,j}} S|}{|S|} \cdot \phi(P_y(\sigma_{a_i=v_{i,j}} S))$$

Binary Criteria

The binary criteria are used for creating binary decision trees. These measures are based on division of the input attribute domain into sub-domains. Let $\beta(a_i, dom_1(a_i), dom_2(a_i), S)$ denote the binary creation value for attribute a_i over sample S when $dom_1(a_i)$ and $dom_2(a_i)$ are corresponding sub domains. The value obtained for the optimal division of the attribute domain into two mutually exclusive and exhaustive sub-domains is useful for comparing attributes.

2.3.2 Information Gain

For purposes of this research, Entropy information gain was used. Information gain is impurity based criterion that uses the entropy measure as an impurity measure.

InformationGain (a_i, S)

$$= Entropy(y, S) - \sum_{v_{i,j} \in dom_2(a_i)} - \left| \frac{|\sigma_{y=v_{i,j}} S|}{|S|} \right| . Entropy(y, \sigma_{a_i=v_{i,j}} S)$$

Where:

$$Entropy(y, S) - \sum_{c_j \in dom(y)} - \frac{|\sigma_{y=c_j} S|}{|S|} . \log_2 \frac{|\sigma_{y=c_j} S|}{|S|}$$

Rokach & Miamon 2010

3. Theoretical Background

Tourism is one of the world's major industries that contributes significantly to the global economy and became one of the major sources of wealth for some developing and developed countries. Due to the increasing competition among tourist destinations in the last several decades, destination marketing managers and industry practitioners have become concerned about their destinations' images in the minds of tourists (Wang & Pizam, 2011). Marketability of destination as well as satisfaction with offered services, entertainment, lodging, transportation and shopping leave an impression on visitors, their satisfaction and decision to come back in the future. Japan although well-known country is still largely undiscovered by mass tourism. Mainly known for its industrial power, Japan as a touristic destination is still overshadowed, by its industrial and business image. Limited academic research on Japan as a destination still leaves understanding of its tremendous potential as a touristic destination, generally undiscovered. Through consumption of goods and services tourists benefit local economy. It is widely accepted that purchasing is an essence of a tourism services (Correia, 2002). According to UNWTO Japan had a 23% of positive growth in international tourism receipts, this creates a need in understanding a patterns of consumer expenditures in Japan. Destination marketing organizations need to know how their destination is perceived by potential visitors, so they can better target their market and develop more appropriate tourism products and increase destination attractiveness (Phillips and Back, 2011). Marketers should take consumer behavior into consideration, where cultural differences, extend of planning time before vacation and number of people in the group influences expenditure of tourist (Leasser and Dolnicar, 2012). However, working with small sample size has its limitation and may not represent full degree of consumer behavioral patterns. New modeling approaches for proposed by Law and Au, (2000) by using decision rules in predicting tourists shopping behavior in Hong Kong. Wong, Chen, Chang & Kao (2006) used data mining tools to find valuable travelers (higher profit) by exploring demographics, buying and decision making, destinations visited and destination preferences. Research paper demonstrated with high confidence seven behavioral characteristics that can help travel agencies plan promotion packages and equip travelers with possible marketing and buying information. Such understanding if customer segmentation using actual behavioral patterns can be invaluable to Japan as an immerging touristic destination. In consequence, we similarly will use data mining techniques to understand what attracts tourist to Japan and tourist future intention to return to

Japan as a touristic destination with this paper. Furthermore, we would like to identify what factors contribute to tourists' satisfaction with Japan as a destination.

4. Methodology

Data were collected by JTB-Foundation on behalf of Japan Tourism Agency during year 2010 at the airport and seaport. Inbound tourists to Japan were approached at random by representatives of JTB foundation. Participants were asked to participate in the survey. Data were collected on the likert, binary scale and sample size of 4000 usable observations. Due to the large sample size where classic statistical tools fail data mining, techniques such as decision trees and logistic regression will be used for data analyzing. Specifically, due to the binary and ordinal scale used in the survey decision trees, two step modeling (with two dependent variables) are used to summarize and interpret behavioral and purchasing patterns of tourists in Japan.

In this paper we will use data mining as an exploratory tool and extract hidden knowledge through set of rules that connects a collection of inputs. Decision trees in the sense represent series of question, where answer to a question determines the follow-up question and as such creating a pattern. Decision tree is probably one of the most popular and powerful technique used in data mining (Berry & Linoff, 2000). Decision trees do not have strict assumptions concerning the functional form of the model, have computational efficiency, robust against outliers, resistant to the curse of dimensionality and require less data preparation than other data mining tools.

4.1 Measurements

This study employed casual research design. The survey questionnaire consisted of following major sections: tourist attributes of satisfaction, overall satisfaction, intention to return, and questions that consists of tourists' demographical questions such as country, party size, gender age, and number of children.

4.2 Attribute Satisfaction

Drawing upon the most relevant literature and destination attributes relevant to Japan as a destination. Destination response encompassed information about current trip to Japan, purpose of the visit, expenditures, transportation, stay arrangements, shopping, sources of information, activities at destination, and satisfaction with Japan as a destination and intention to return in the future to Japan. Survey consists of over 150 questions measured on Likert scale and binary response.

4.3 Overall Satisfaction

A single overall measure of satisfaction was used in this study for its ease of use and empirical support. Satisfaction was measured with 7-point Likert scale where 1 being highly dissatisfied and 7 being highly satisfied (Chi and Qu, 2008).

4.4 Behavioral Intentions

A single measure for intention to return was used in this study for its ease of use and empirical support. Intention to return was measured with 7-point Likert scale where 1 being “definitely not return” and 7 being “definitely return” (Chen and Tsai, 2005).

5. Results

5.1 Important variables

5.1.1 Variables are listed in order of importance for future intention to return

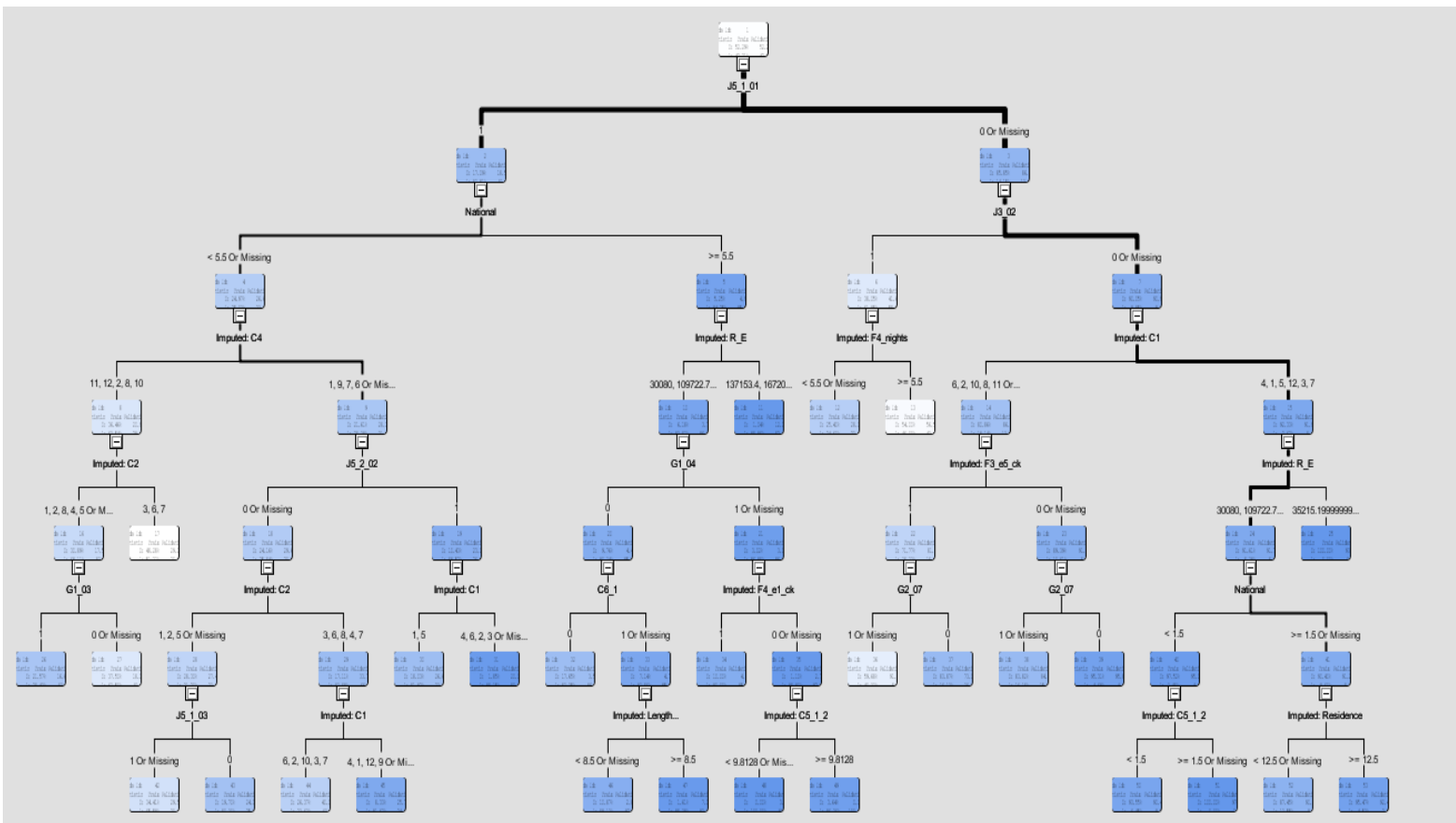
Variable	Description
J5_1_01	Experienced Japanese Food
J5_1_06	Shopping
J3_02	Transportation
J1_01	Lonely Planet as a major source of information about Japan prior to visit
C1	Which airport did you land in Japan
C2	How many time have you visited Japan including this visit
C5_1_1Area	Main area (destination) in Japan visited
J2_06	Internet as a main helpful source in obtaining information while in Japan
J5_2_04	Desire to experience nature/scenery sightseeing next visit
R_E	Flight cost
Resident	Country of residency
J5_2_05	Want to walk around downtown in the future
F4_b_ck	Catering cost
F3_e5	Cosmetics and pharmacy expenditure
National	Nationality
G2_07	Credit Cards as a method of payment in Japan
Age	Age
Residents of China	Residents of China

5.1.2. Variables are listed in order of importance for satisfaction with Japan as a destination

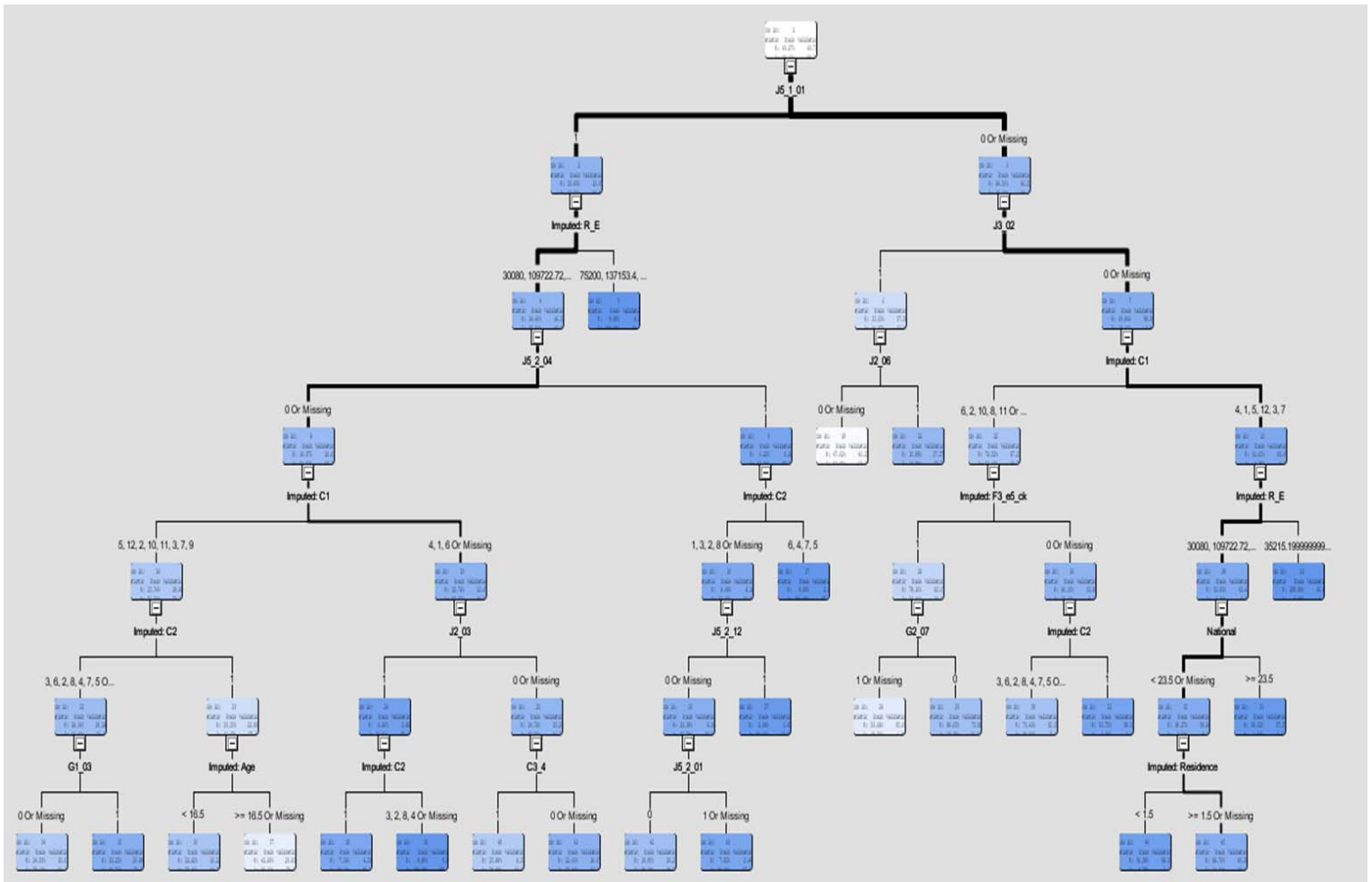
Variable	Description
J5_1_01	Japanese food
J5_1_06	Shopping
J3_02	Availability of Information on transportation
Residence	Country of residence
National	Nationality
C1	Airport
C5_1_1	Main area (destination) in Japan visited
C4	Main purpose of the visit
C5_1_2	Secondary destination visited in Japan
F4	Main place where tourist stayed in Japan
C2	Prior visit to Japan
J5_1	Business trip
F3_e5	Cosmetics and Pharmacy expenditure
G2_07	Credit Cards as a method of payment in Japan
Length of stay	Length of stay
J2_02	Would like to stay in Japanese style inn next time/appeal of Japanese hospitality
J1_03	Hot spring experience
J5_2_04	Desire to experience nature/scenery sightseeing next visit
C7	Organized tour

5.3.3 Decision tree rules

Satisfaction



Intention to return



5.3.4 Demographics

Majority of tourists came from Asia (62%) such as Korea (19.51%), Taiwan (18.10%), Main Land China (14.16%). Second largest visitors are from USA (10.65%). From Main Land China two largest groups are coming from Beijing and Shanghai. Gender was rather evenly distributed with man (56%) and woman (43%). Average age was 23 years with standard deviation of 13 years. Majority of the tourists arrived in Narita (53.88%), Kansai (17.63%), and New Chitose (Sapporo) (6.212%). Majority of the respondents visited Japan only once or several times. 42% of respondents visited Japan for the first time, 15% visited for the second time and 10% for the third time. General distribution of group travelers are: alone (17%), family (21%), work colleague (19%), and friends (19%). 57.9% of respondents travel for tourism and leisure (57.9%), and business training, conference or trade fair (25%).

5.3.5 Decision trees node rules

Satisfaction

For purposes of better classification with decision trees variable satisfaction was recoded into binary variable where 1 is highly satisfied and satisfied and 0 is everything else. Binary response was equally distributed.

Rule 1:

Log Odds Ratio of tourists being satisfied are is higher by 1.39 if they are from non-Asian country, experienced Japanese food, came for business purposes or visit friend, and shopped at local department store.

Rule 2:

Log Odds Ratio of tourists being satisfied are is higher by 1.64 if they are from neighboring Asian country (Korea, China, Taiwan, Hong Kong and Thailand), stayed at Japanese style inn, experience Japanese food, came for tourism/leisure, Incentive travel, Study, or International conference, and came through two main airports (Narita/Haneda)

Rule 3:

Log Odds Ratio of tourists being satisfied are is higher by 1.64 I they are mainly non-Asian countries, experienced Japanese food, paid between 300 and 1500 for air fare, and used accommodations other than western-style hotels

Rule 4:

Log Odds Ratio of tourists being satisfied is higher by 2.32 if tourists are mainly from non-Asian countries, had experience with Japanese food, paid between 300 and 1500 for air fare, purchased Japanese fruits, and shopped at supermarket.

Rule 5

Log Odds Ratio of tourists being satisfied is higher by 1.51 if tourists are from neighboring Asian country (Korea, China, Taiwan, Hong Kong and Thailand), experienced Japanese food, came for tourism or exhibition/conference/company meeting, and being to Japan more than once or twice.

Rule 6:

Log Odds Ratio of tourists being satisfied is higher by 2.21 if tourists are mainly from non-Asian countries, paid between 300 and 1500 for air fare, experienced Japanese food, stayed less than 8 days, stayed at western style hotel.

5.3.6 Intention to return

For purposes of better classification with decision trees variable satisfaction was recoded into binary variable where 1 is highly satisfied and satisfied and 0 is everything else. Binary response was equally distributed.

Rule 1:

Log Odds Ratio of tourists having intention to return is 3.9 if they experienced Japanese food, are not interested experience Japanese nature/scenery sightseeing, paid between 300 and 1670 for air fare, visited Japan for the first time and came through airports such as Narita, New Chitose (Sapporo), or Fukuoka.

Rule 2:

Log Odds Ratio of tourists having intention to return is 3.9 if tourists experienced festival/event, Nature/scenery/sightseeing, Japanese food, paid between 300 and 1670 for air fare, and visited Japan several times.

Rule 3:

Log Odds Ratio of tourists having intention to return is 1.30 if tourists experienced Japanese food, but not Nature/scenery/sightseeing, paid between 300 and 1670 for air fare, first time visitors, and young age.

Rule 4:

Log Odds Ratio of tourists having intention to return is 1.13 if tourists experienced Japanese food, but not Nature/scenery/sightseeing, want to experience Japanese hot spring in the future trip, paid between 300 and 1670 for air fare, and came with family, spouse or friends.

Rule 5:

Log Odds Ratio of tourists having intention to return is 1.94 if tourists experienced Japanese food, but not Nature/scenery/sightseeing, want to experience Japanese hot spring in the future trip, and came with family, spouse or friends.

Rule 6:

Log Odds Ratio of tourists having intention to return is 1.49 if tourists want to experience in the future nature/scenery/sightseeing, experienced Japanese food, and paid between 300 and 1670 for air fare.

6. Discussion

Results of the study indicate that tourist satisfaction differs between two distinct groups which are Asian and non-Asian tourists with different preferences to achieve high level of satisfaction. Main influencing factors of satisfaction for non-Asian tourists would include experience with Japanese food, shopping at department store, stayed at western style hotel, came on business or visit friend and air fare cost. For Asian tourists those factors would be experience with Japanese

food, stay at Japanese style inn, attending an event such as conference, incentive travel or study, previous visit to Japan and importantly they have no preferences for air fare cost.

On the other hand, nationality plays no role in decision for future return, and unlike “satisfaction”, tourists are more family oriented or non-business without previous visits. However, main drive for a future return is not experiences people had, but rather experiences people want to have in the future such as Japanese hot spring or nature. Experience with a Japanese food remains universally attractive with all combinations.

This study has a fair share of limitations, including sampling errors. For future research, data collected in multiple years and seasons can be explored to see if the tourists’ behavioral patterns changed over a major event such as Great East Japan Earthquake in March 11, 2011.

References

- Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future generation computer systems*, 13(2), 197-210.
- Au, N., & Law, R. (2002). Categorical classification of tourism dining. *Annals of Tourism Research*, 29(3), 819-833.
- Berry, M. J., & Linoff, G. (2000). Mastering data mining.
- Chen, C. F., & Tsai, D. (2007). How destination image and evaluative factors affect behavioral intentions?. *Tourism management*, 28(4), 1115-1122
- Chi, C. G. Q., & Qu, H. (2008). Examining the structural relationships of destination image, tourist satisfaction and destination loyalty: An integrated approach. *Tourism management*, 29(4), 624-636.
- Correia, A. (2002). How do tourists choose? A conceptual framework. *Tourism (Zagreb)*, 50(1), 21-29.
- Hosking, J. R., Pednault, E. P., & Sudan, M. (1997). A statistical perspective on data mining. *Future Generation Computer Systems*, 13(2), 117-134.
- Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining.
- Laesser, C., & Dolnicar, S. (2012). Impulse purchasing in tourism—learnings from a study in a matured market. *Anatolia*, 23(2), 268-286.
- Nisbet, R., Elder IV, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press.
- Okamura, K., & Fukushige, M. (2010). Differences in travel objectives between first-time and repeat tourists: An empirical analysis for the Kansai area in Japan. *International Journal of Tourism Research*, 12(6), 647-664.

Uzama, A. (2009). Marketing Japan's travel and tourism industry to international tourists. *International Journal of Contemporary Hospitality Management*, 21(3), 356-365.

Maimon, O. Z., & Rokach, L. (Eds.). (2005). *Data mining and knowledge discovery handbook* (Vol. 1). New York: Springer.

Wang, Y., & Pizam, A. (Eds.). (2011). *Destination marketing and management: theories and applications*. CABI.

Wong, J. Y., Chen, H. J., Chung, P. H., & Kao, N. C. (2006). Identifying valuable travelers and their next foreign destination by the application of data mining techniques. *Asia Pacific Journal of Tourism Research*, 11(4), 355-373.

UNWTO retrieved from <http://media.unwto.org/press-release/2013-12-12/international-tourism-engine-economic-recovery>

Corresponding author: Valeriya Shapoval

Email: Valeriya.Shapoval@ucf.edu or valeriya.shapoval@gmail.com