Paper 3449-2015

Applied Analytics in Festival Tourism: A Case Study of Intention-to-Revisit Prediction in an Annual Local Food Festival in Thailand

Thanathorn Vajirakachorn, Ph.D.
Department of Tourism Management, School of Business
The University of the Thai Chamber of Commerce

Jongsawas Chongwatpol, Ph.D.
NIDA Business School, National Institute of Development Administration

ABSTRACT

Improving tourists' satisfaction and intention to revisit the festival is an ongoing area of interest to tourism industry. Many organizers at the festival site strive very hard to attract and retain attendees by investing heavily in their marketing and promotion strategies for the festival. To meet this challenge, the advanced analytical model though data mining approach is proposed to answer the following research question: "What are the most important factors that influence tourists' intention to revisit the festival site?" Cluster analysis, neural network, decision tree, stepwise regression, polynomial regression, and support vector machine are applied in this study. The main goal is to determine what it takes not only to retain the loyalty attendees, but also attract and encourage new attendees to be back at the site.

INTRODUCTION

Improving tourists' satisfaction and intention to revisit the festival is an ongoing area of interest to tourism industry. The amount of regional travelling to special events each year tends to increase. Specifically, the local food festival is one of the interesting tourism destinations often chosen to be visited. However, drastic changes in tourism markets over the last decades have increased the pressures and challenges for the festival industry. Many organizers at the festival site strive very hard to attract and retain attendees by investing heavily in their marketing and promotion strategies for the festival. In fact, tourists' preference is very difficult to understand as it is very subjective due to the complex cognitive and affective aspects each tourist beliefs and feels about the destination choices. Thus, in determining factors influencing tourists' intention to revisit the festival sites, advanced statistical analysis is needed.

However, traditional survey-based study with descriptive statistical analysis and conventional regression modeling may not help festival organizers to gain insights about the tourists and to identify which aspects of festivals and tourists' perception impact the decision to return to the festival sites in the future. This study seeks to fill this gap and answer the following research question "What are the most important factors that influence tourists' intention to revisit the festival site?" The advanced analytical model though data mining approach is proposed. This study differs from others by providing alternatives that have been widely applied in other fields but still be novel in tourism industry to understand the characteristics of tourists visiting the event and attract them with specific profiles to the revisit the event. Cluster analysis, neural network, decision tree, stepwise regression, polynomial regression, and support vector machine are applied in this study. The main goal is to determine what it takes not only to retain the loyalty attendees, but also attract and encourage new attendees to be back at the site.

RESEARCH METHODOLOGY

We follow the CRISP-DM Model, a popular data mining method as a complete blueprint for this study. CRISP-DM breaks down this data mining project into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Shearer, 2000; Wirth & Hipp, 2000).

Business Understanding

The study was conducted at a downtown festival in Pattaya, Thailand, called "The 5th Walk to Remembrance at Naklua Market." The event was held every weekend from Nov 23, 2013 to Jan 12, 2014. The town association organizes the festival in order to promote its local food and historical culture. The estimated average visitors of the festival are approximately 2,000 visitors per evening.

The first task in the CRISP-DM model is to setup the objective of the study, which is to uncover important factors influencing the intention to revisit the festival. Festival organizers are looking for ways not only to satisfy festival visitors so that they become repeat visitors, but also to gain new visitors based on the advertisement, promotion, or even the word-of-mouth from current visitors that positively is conveyed to their family and friends. Specifically, we would like to develop predictive models to determine which aspects of festivals and visitors' perception have the most impact on visitors' intention to return. Figure 1 presents the overall CRISP-DM framework of this study.

Data Understanding

The dataset is derived from the survey questionnaires collected from festival attendees every evening. The visitors were given a questionnaire once they agreed to participate. The survey is developed to construct the attributes related to visitors' demographic information and perception toward the event as follows.

- Demographic Information
- Perceived Value
- Perceived Quality
- Push Factors
- Pull Factors
- Attendee's Overall Satisfaction
- Intention to Revisit is a binary target variable

A total of 3,000 questionnaires were randomly distributed and a completed data records with 30 attributes are used to explain and predict the intention to revisits. The scale of measurement for perception toward the festival is ranged from "strongly disagree" to "strongly agree" on 5-likert scales.

Additionally, cluster analysis is performed to get some senses of what tourists want to accomplish when they visit the festival site. The festival attendees are segmented into sub-clusters that share similar characteristics. Many tourists decide to participate in the event for different reasons. One just wants to take a rest or explore the festival with family and friend. Another might want to learn new culture or meet new people at the festival site. The other intends to visit the festival because of the onstage activities and reasonable price of food and beverage. Thus, understand attendees' characteristics before building predictive models, *k*-Means algorithm is deployed in this clustering process. *k*-Means has been widely used in market segmentation with its ability to identify distinct clusters in a dataset with multiple variables.

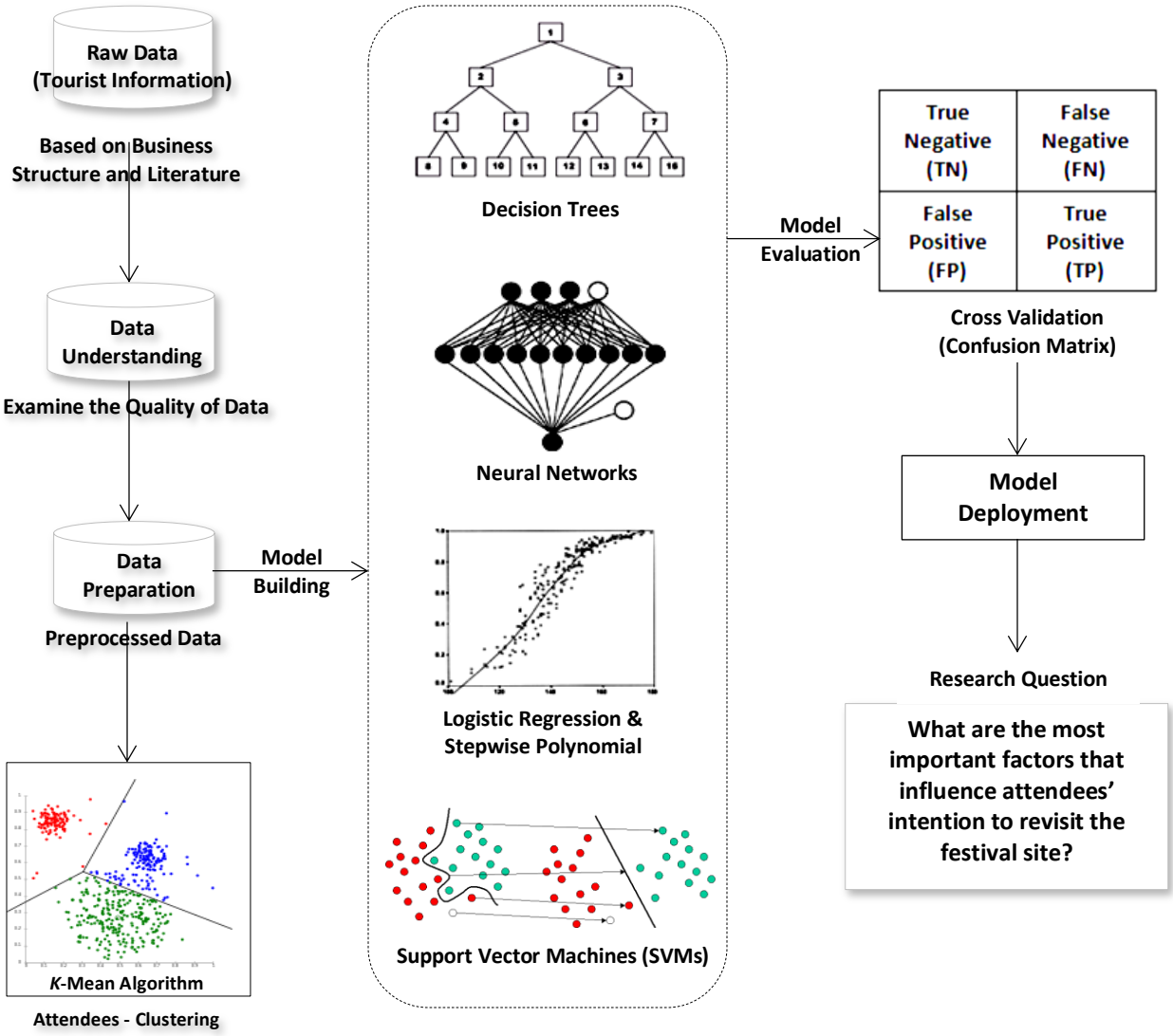


Figure 1. The Overall CRISP-DM Framework

Data Preparation:

Our first task in this step is to get a sense of the dataset for any errors and inconsistencies. Frequency distribution, descriptive statistics, and cross-tab analysis are used in this section. Any outliers such as those who answer "strongly agree or strongly disagree throughout the questionnaires are treated and removed accordingly. The next step is to assess whether the data are complete and which variables are to be included in the model. Because including variables with a high number of missing values can lower the quality of the findings, even after applying a missing value imputation method, an assumption is that the model excludes variables with over 50% of the information missing. Additionally, the dataset is stratified to construct a model set with approximately equal numbers of each target variable. With a 50% adjustment for oversampling, the contrast between the two values is minimized, which makes pattern recognition in the dataset easier and more reliable.

Modeling:

In this study four popular classification platforms (neural networks, stepwise logistic regression, decision trees, and support vector machines) are used in this study to analyze dataset with multiple predictor variables. For a technical summary, including both algorithms and their applications for each method, see (Jackson, 2002 and Turban et al., 2011).

- An artificial neural network (ANN), also called circuit of biological neurons, is a mathematical and computational model for pattern recognition and data classification through a learning process. It is a biologically inspired analytical technique, simulating a biological system or brain nervous systems, where a learning machine algorithm indicates how learning takes place and involves adjustments to the synaptic connections between neurons. ANN has been used significantly in many fields because of its remarkable ability to derive meaningful pattern recognition or data classification from complicated or imprecise data. Data input can be discrete or real-valued; the output is in the form of a vector of values and can be discrete or real-valued as well.

- Stepwise Logistic regression is often used to predict a binary outcome variable or multi-class dependent variables with automatic selection of independent variables. It builds the model to predict the odds of discrete variables (dependent variables) by a mix of continuous and discrete predictors, instead of by point estimate events as in the traditional linear regression model, as the relationship between dependent variables and independent variables is non-linear. The emphasis of stepwise procedure is on choosing predictor variables that best explain a particular predicted variable on the basis of statistical criteria. Stepwise Polynomial regression is an enhanced regression to predict a dependent variable on the basis of *n* independent variables. It is commonly used when the relationship between target variables and the explanatory variables is toward a complicated non-linear phenomenon.

- A decision tree is another data classification and prediction method commonly used because of its intuitive explainability characteristics. A decision tree divides the dataset into multiple groups by evaluating individual data records, which can be described by their attributes. In other words, a decision tree provides unique capabilities to split a dataset into branch-like segments with a root node at the top of the tree. It is also simple and easy to visualize the process of classification where the predicates return discrete values and can be explained by a series of nested or rule-based if-then-else statements.

- Support Vector Machines (SVMs) are among the best supervised learning algorithm that are based on the concept of decision planes defining decision boundaries. SVM is a classier method utilizing the mapping function to construct hyperplanes in a multidimensional space to either categorize the data for the classification tasks or to estimate the numerical value of the desired output in the regression tasks. Various kernel functions such as linear, polynomial, radial basis function (RBF) and sigmoid are used to transform the input data into the higher dimensional feature space in which the algorithms outputs an optimal hyperplane so that input data becomes more separable. The larger minimum distance in the hyperplane indicates the lower error in classifying the input data. For the classification tasks, for instance, the new observations are then mapped into the same space and are predicted into the side of the gap they are belonging to. SVM can handle several thousands of training examples.

Evaluation:

The complicity of the model is controlled by fit statistics calculated on the validation dataset. We measure the performance of our models based on the following three different criteria: false negative, prediction accuracy, and misclassification rate. These performance measures are defined in the classification table in Figure 2. These criteria include False negative (Target = 1 and Outcome = 0) represents the case of an error in the model prediction where model results indicate that diabetes occurrence is not present, when in reality, there is an incident. The false negative value should be as low

as possible. The proportion of cases misclassified is very common in the predictive modeling. However, the observed misclassification rate should be also relatively low for model justification. Lastly, prediction accuracy is evaluated among the three models on the testing dataset. The higher the prediction accuracy rate, the better the model to be selected.

		Target (Actual Value)	
		No (0)	Yes (0)
Outcome (Predicted Value)	No (0)	True Negative (TN)	False Negative (FN)
	Yes (0)	False Positive (FP)	True Positive (TP)
$\textbf{Per} - \textbf{Class Accuracy}$		$\frac{TN}{TN+FP}$	$\frac{TP}{FN+TP}$

$$\textbf{Overall Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\textbf{False Negative} = \frac{FN}{TN+FN}$$

$$\textbf{Misclassification Rate} = \frac{FP+FN}{TP+TN+FP+FN}$$

Figure 2. Classification Table with Performance Measures

RESULTS AND DISCUSSION

Cluster Analysis

Visitors to the festival can be divided into four groups based on the k-means clustering analysis on the factors described in Section 3.2. Figure 3 illustrates the three most influential variables used in classifying visitors in each cluster, which has particular preference to visit the festival site. For instance, Cluster #1 comprises visitors, who favor the variety of products, price, and timing of the festival. Visitors in Cluster #2 are mostly satisfied with traffic management of the festival organizer, value, and facilities at the site. Visitors in Cluster #3 emphasize the value and quality of products offered and they attend the site to find new experience. Visitors in Cluster #4 tend to have their high expectation of the festival and mostly are satisfied with the value and the quality of activities of the festival. These four clusters provide an overview of different clusters' profile so that the event organizer understands visitors' motivation and purpose of site visit. Additionally, further segmentation profile provides valuable insights about visitors' preferences or perception of the festival. For instance, Figure 3 shows a multi-dimensional diagram for evaluating the variables particularly related to push factors for all four clusters. Event organizer can approach the visitors in Cluster #4 by promoting the value and activities the festival is offering rather than emphasizing the experience and culture visitors can learn from. Attending the festival for visitors in Cluster #1 and #2 increase the chance of spending time with their family; thus event organizers may come up with campaign that children can gain new learning experience and culture or promoting the festival site that is easily to access with facilities (rest area or rest room) that are suitable for both adults and children.

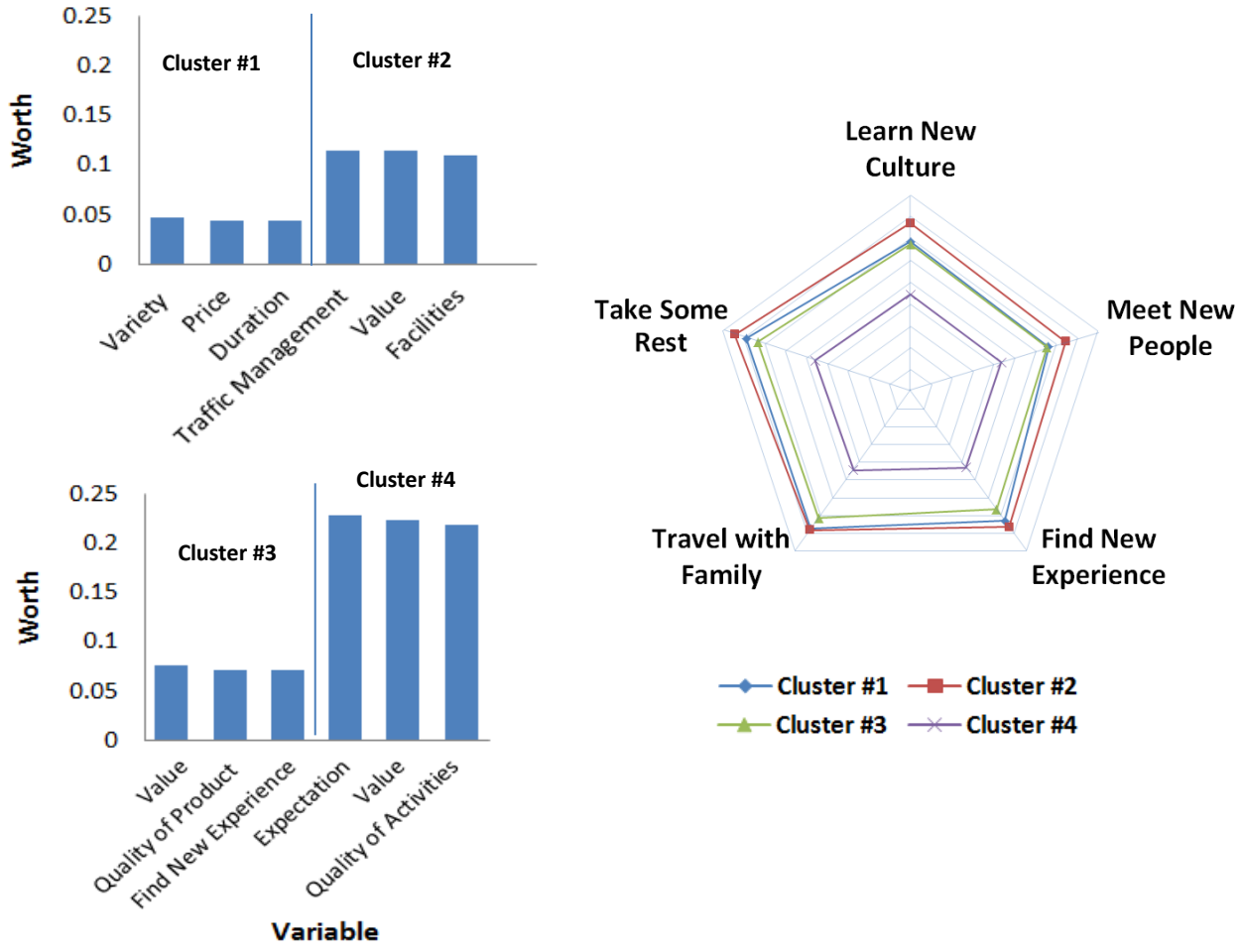


Figure 3. Cluster Analysis – Variable Importance and Segmentation Profile

Model Comparison

After we get a sense of visitors' data, SAS® Enterprise Miner™ 12.1 are used for model development and comparison. The performance measures for all five predictive models are presented in Figure 4. Neural network produces the best results with overall misclassification rate of 4.42%, followed by the support vector machines and polynomial regressions with misclassification rates of 12.30% and 12.62%, respectively. Neural network also has the lowest false negative rate of 2.35% and decision tree comes out as a runner up with false negative rate of 8.49%, followed by support vector machine (11.81%) and polynomial regression (14.18%).

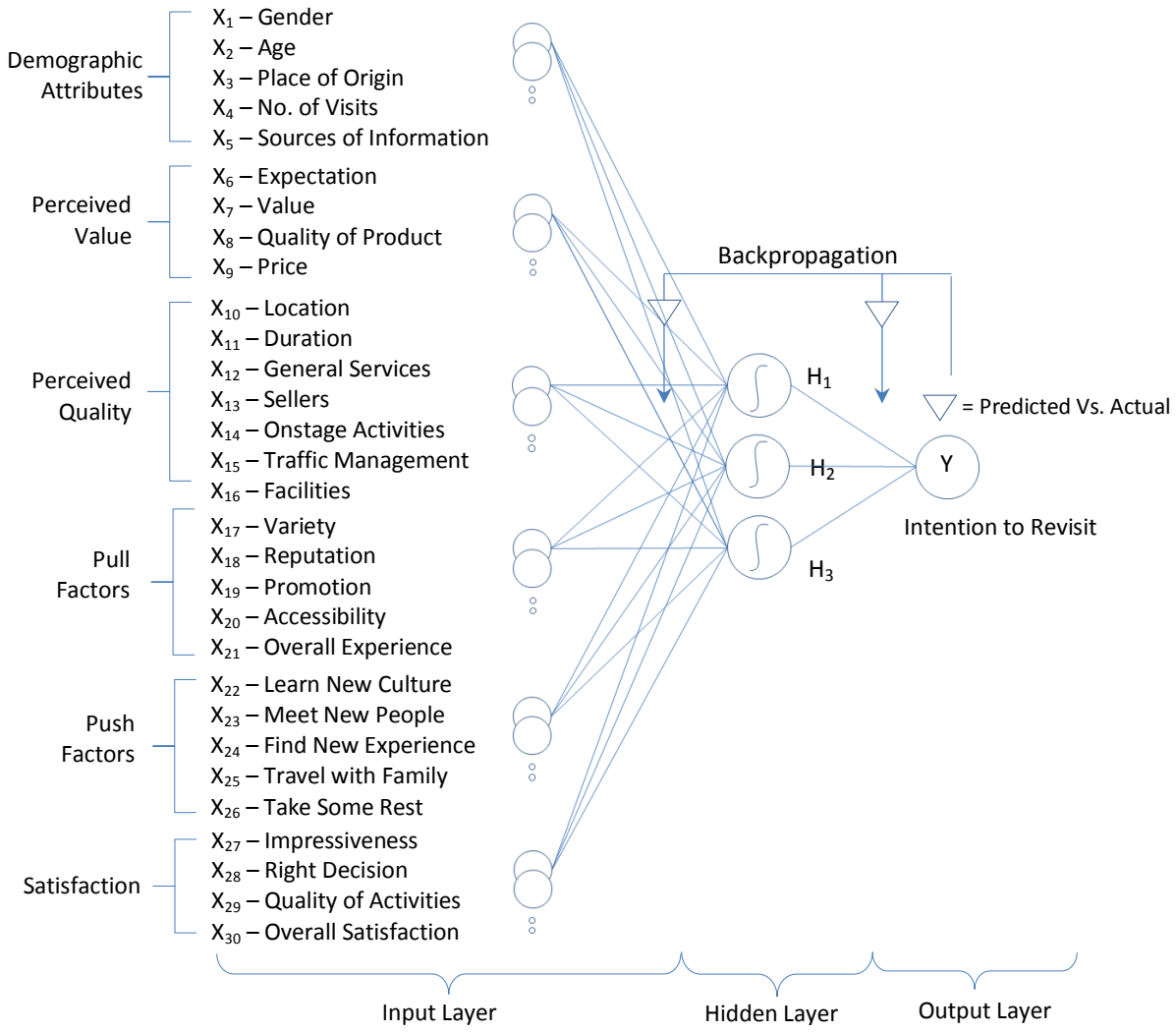
		Logistic Regression Target (Actual Value)		Polynomial Regression Target (Actual Value)		Decision Trees Target (Actual Value)	
		No (0)	Yes (0)	No (0)	Yes (0)	No (0)	Yes (0)
Outcome (Predicted Value)	No (0)	106	24	115	19	97	9
	Yes (0)	30	157	21	162	39	172
	Total	136	181	136	181	136	181
Per-class Accuracy		77.94%	86.74%	84.56%	89.50%	71.32%	95.03%
Overall Accuracy		82.97%		87.38%		84.86%	
False Negative		18.46%		14.18%		8.49%	
Misclassification Rate		17.03%		12.62%		15.14%	

		Neural Network Target (Actual Value)		Support Vector Machines Target (Actual Value)	
		No (0)	Yes (0)	No (0)	Yes (0)
Outcome (Predicted Value)	No (0)	125	3	112	15
	Yes (0)	11	178	24	166
	Total	136	181	136	181
Per-class Accuracy		91.91%	98.34%	82.35%	91.71%
Overall Accuracy		95.58%		87.70%	
False Negative		2.34%		11.81%	
Misclassification Rate		4.42%		12.30%	

Figure 4: Model Comparison

Neural Network

Figure 4 present the pictorial representation of the neural network architecture developed to explain and predict the visitors' intention to revisit the festival site. This neural network model consists of one hidden layer and three hidden units with various input variables. The result of the prediction model is presented where each hidden unit (H_1 to H_3) has its own function and weights. The transfer function ($\int$), also called the *tanh* function, combines all the inputs into a single value for each hidden unit and then calculates the output value of Intention to Revisit (Y). The following seven variables with high coefficient estimates in the neural network model such as Accessibility (H_{20}), Onstage Activities (H_{14}), Overall Satisfaction (H_{30}), Quality of Product (H_8),, Age (40 to 59) (H_2),, Travel with Family (H_{25}),, and Traffic Management (H_{15}), are among the important variables influencing the decision to return to the festival.



$$\text{Logit}(Y=1) = -10.138 - 20.902H_1 - 20.872H_2 - 11.706H_3 \quad [\text{where Intention to Revisit (Y): Yes (Y = 1) and No (Y=0)}]$$

$$H_1 = \tanh(-10.713 - 20.412X_{20} - 3.596X_6 + 5.787X_{16} + 2.99X_{12} + 1.91X_{11} - 2.68X_{24} + \ldots\ldots\ldots\ldots + 1.16X_9)$$

$$H_2 = \tanh(1.226 - 3.181X_{20} + 2.638X_6 - 2.119X_{16} - 3.035X_{12} - 5.762X_{11} - 3.708X_{24} + \ldots\ldots\ldots\ldots + 1.927X_9)$$

$$H_3 = \tanh(-5.897 + 3.55X_{20} - 9.808X_6 - 7.462X_{16} + 5.742X_{12} + 4.733X_{11} - 2.823X_{24} + \ldots\ldots\ldots\ldots + 3.279X_9)$$

Figure 5. Neural Network Model

Stepwise Logistic and Polynomial Regression

The stepwise logistic regression equation is presented in Equation #1. Of the 30 variables used in the intention-to-revisit prediction model, only 5 variables are selected in the model. This finding helps the festival organizer prioritize the important factors associated with the decision to revisit the sites. The organizer definitely starts by maintaining the festival reputation (X_{18}) on both products and services so that the overall satisfaction (X_{30}) of visitors is always at excellence level. Interestingly, the other three reasons that influence the decision to revisit the festival include Promotion (X_{19}), Accessibility (X_{20}), and Lean New Culture (X_{22}). Promoting the cultural themes of the festival along with the various promotions

on food and products at the site while facilitating the accessibility to the festival site can increase the likelihood to revisit the site in the future.

However, conventional regression analysis may not be able to explain the visitor' perception and behavior, which is a sophisticated and dynamic-nonlinear phenomenon. Thus, the more complex polynomial logistic regression is then applied to improve the accuracy of the predicted outcomes and to test whether the interaction effects among variables impact the intention to revisit the site. The results are quite reasonable as the misclassification rate decreases from 17.03% to 12.62%. Besides the variables specified in the stepwise logistic regression in Equation #1, the polynomial regression equation (as presented in Figure 6) indicates that visitors, who attend the site to find new experience (X_{24}) with high expectation of the festival (X_{21}) and whose the overall experience after visiting the festival (X_6) is very high, are likely to attend the festival again in the future. Location (X_{10}) of the festival site is still one of the important factors for the tourists' selection of Festival choices.

On the other hand, Promotion (X_{19}) advertised as a part of marketing campaigns to persuade a target group of tourists is commonly considered a pull motive of attendees at the festival. Based on both logistic and polynomial regression in Figure 6, it implies that the higher the products and services at the festival site are promised, the lower the chance attendees intent to comeback in the future. The finding is reasonable especially when the festival is exaggeratedly promoted and the quality of products or services attendees experienced below their expectation.

Support Vector Machines (SVMs)

Figure 7 presents the basic architecture of SVM, which produces a binary classifier of attendees, who do and do not plan to revisit the festival site in the future. The overall accuracy of the model prediction is 87.70% with false negative of 11.81%. The following examples are the characteristics of two attendees with the highest possibility (greater than 80%) to return to the site.

- Attendee #1: a local female resident with an age of between 40 and 59 year old. She visits the festival more than once. The main purpose of her visit is to experience new culture with her family. She receives the information of the festival from cable TV. Although the price is a little high based on her perceived value she experienced and the variety of products is below her expectation, she is still satisfied with the quality of products and services offered at the site. The overall experience and impressiveness of the festival overcome the minor issues of traffic management and accessibility to the site.
- Attendee #2: a female visitor who lives in the province nearby with an age of between 20 and 39 year old. She also visits the festival more than once. She really likes to take a break from her routine job with her family. She feels that the facilities such as the rest area and restrooms are not appropriate and need significant improvement. However, the on-stage activities and the variety of products and services far exceed her expectation. She also notes that the location and the timing of the festival are mediocre and the advertisement is not widely promoted.

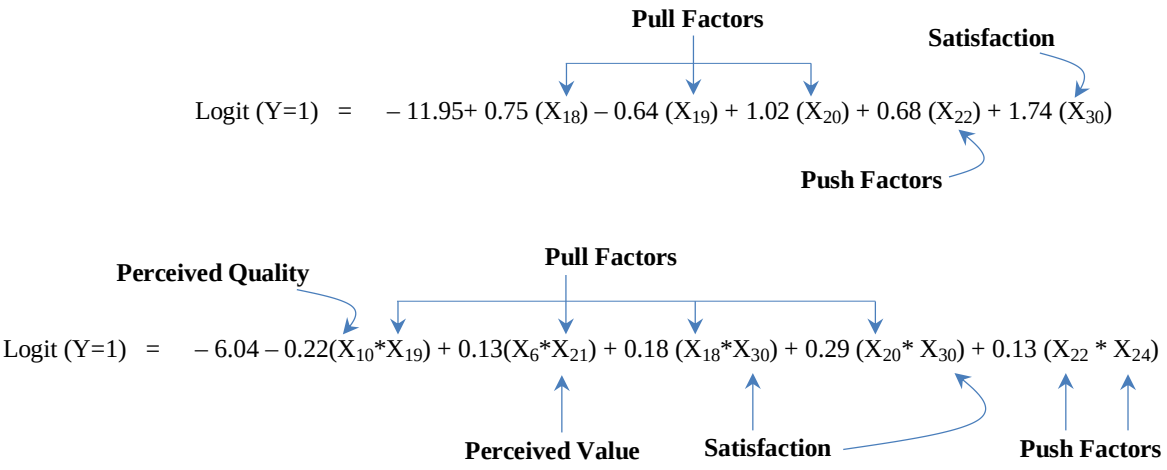
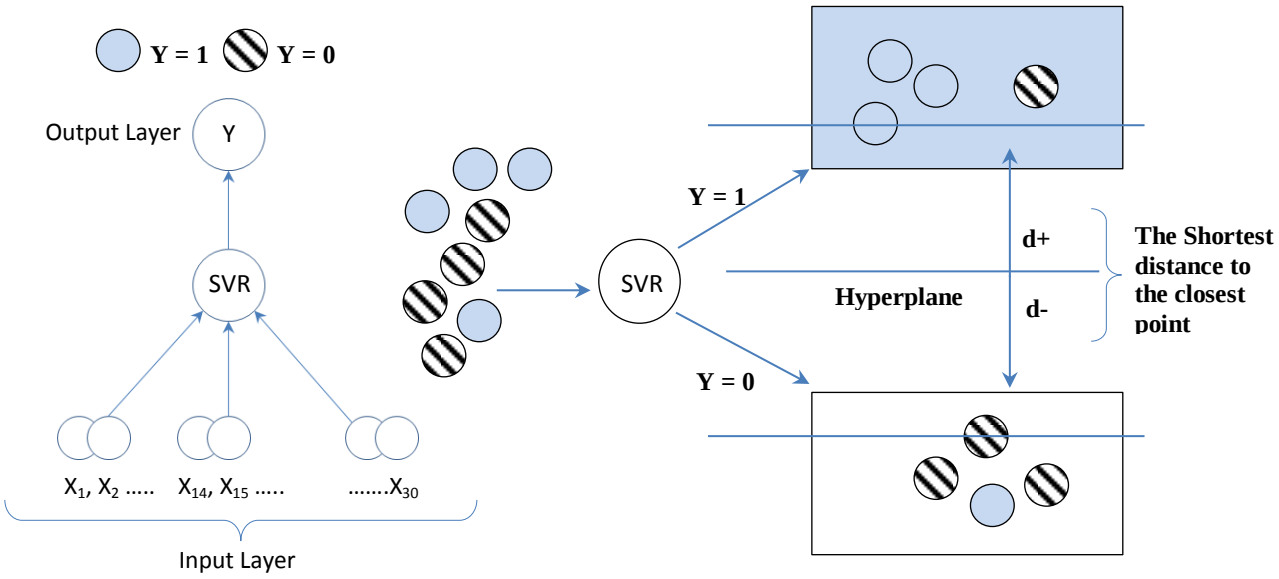


Figure 6: Stepwise Logistic and Polynomial Regression



The SVM Procedure

Support Vector Classification	C_CLAS
Kernel Function	Linear
Estimation Method	DQP
Maximum QP Size	100
Number of Observations (Train)	317
Number of Effects	41
Regularization Parameter C	0.550000
Classification Error (Training)	39.000000
Objective Function	-55.396654
L1 Loss	1.207368E-14
Inf. Norm of Gradient	1.931788E-14

Squared Euclidean Norm of w	6.462817
Geometric Margin	0.786718
Number of Support Vectors	125
Number of S. Vector on Margin	93
Norm of Longest Vector	6.403124
Radius of Sphere around SV	6.020797
Estimated VC Dim of Classifier	235.277105
Linear Kernel Constant (Fit)	-1.173039
Linear Kernel Constant (PCE)	-1.173039
Number of Kernel Calls	773644

Figure 7. Basic Architecture of SVM Model

Decision Trees

To illustrate the implication of the decision tree model (see Figure 8), the following four rule-based "if-then" algorithms are presented. The variables selected in this decision three model are quite in line with those in the neural network model. Overall Satisfaction (X_{30}), Accessibility (X_{20}), Place of Origin (X_3), Variety (X_{17}), and Reputation (X_{18}) are of concern in the rule-based algorithms. Thus, closer attention to these factors can be promptly initiated. For instance, to increase the probability to return to the festival, the organizer should primarily focus on increasing the variety of products and services along with maintaining its reputation and attendees' overall satisfaction

IF "Overall Satisfaction" is greater than 3.5 ***AND*** "Accessibility" is greater than 3.5 ***THEN*** the probability to return to the festival is 90.5%

IF "Overall Satisfaction" is greater than 3.5 ***AND*** "Accessibility" is less than 3.5 ***AND*** "Place of Origin" is not resident ***THEN*** the probability to return to the festival is 37.5%

IF "Overall Satisfaction" is less than 3.5 ***AND*** "Reputation" is greater than 3.5 ***AND*** "Variety" is less than 3.5 ***THEN*** the probability to return to the festival is 6.1%

IF "Overall Satisfaction" is less than 3.5 ***AND*** "Reputation" is greater than 3.5 ***AND*** "Variety" is greater than 3.5 ***THEN*** the probability to return to the festival is 55.6%

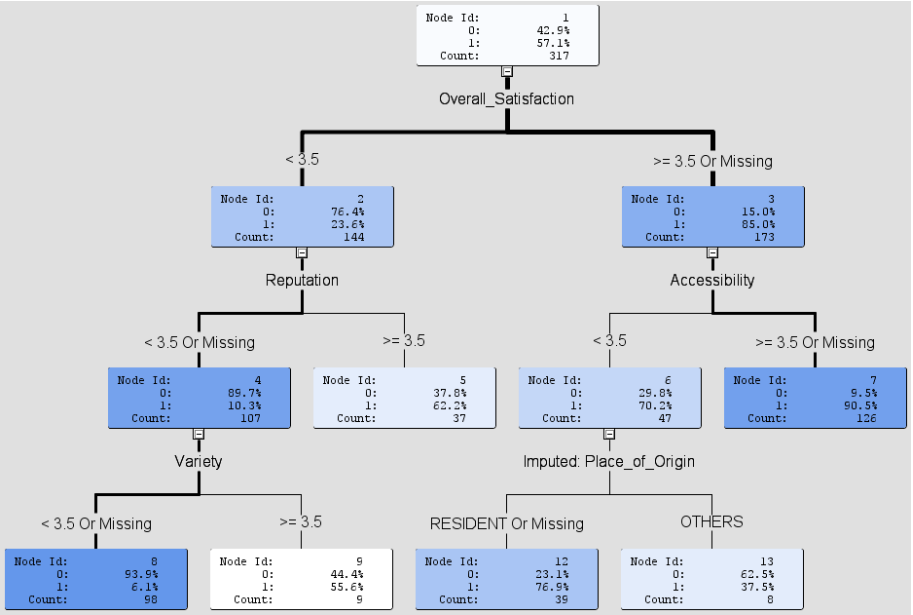


Figure 8: Decision Tree Model

CONCLUSION

Our results show that data mining approaches are capable of predicting attendees' intention to revisit the festival sites, given sufficient data with the proper input variables. Among the five different prediction platforms, a neural network performs the best with the lowest misclassification rate, followed by polynomial regression and support vector machine. Successful festival organizers require careful planning and comprehensive analysis of data and information obtained from tourists who plan to revisit the events and those who do not. The ability to identify and understand attendees' characteristics through data mining approach has become a necessity in order to attract and retain most valuable tourists for the event.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Thanathorn Vajirakachorn
Enterprise: Department of Tourism Management, School of Business at the University of the Thai Chamber of Commerce
Address: 126/1 Vibhavadee-Rangsit Road, Dindaeng, Bangkok, 10400, Thailand
Email: thorn_utcc@yahoo.com, p_thorn_p@yahoo.com

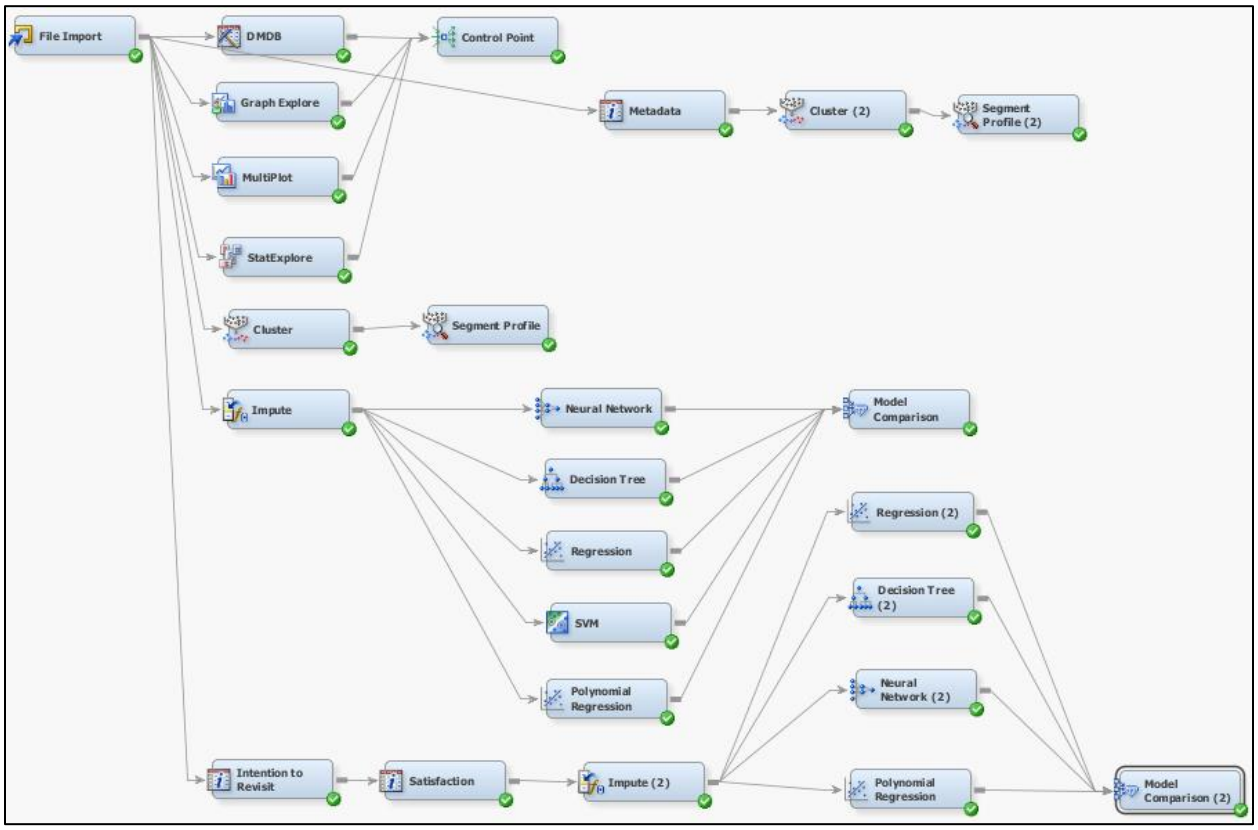
Name: Jongsawas Chongwatpol, Ph.D.
Enterprise: NIDA Business School, National Institute of Development Administration
Address: 118 Seri Thai Road, Bangkapi, Bangkok, 10240 Thailand
Email: jongsawas.c@ics.nida.ac.th, jong_tn@hotmail.com

Thanathorn Vajirakachorn is a lecturer in the Department of Tourism Management, School of Business at the University of the Thai Chamber of Commerce, Bangkok, Thailand. She received her M.S. degree in Hospitality and Tourism from University of Wisconsin-Stout, and Ph.D. in Tourism from Texas A&M University. Her major research interests include community-based tourism, sustainable tourism, and gastronomic tourism.

Jongsawas Chongwatpol is a lecturer in NIDA Business School at National Institute of Development Administration. He received his BE in industrial engineering from Thammasat University, Bangkok, Thailand, and two MS degrees (in risk control management and management technology) from University of Wisconsin - Stout, and PhD in management science and information systems from Oklahoma State University. His research has recently been published in major journals such as Decision Support Systems, Decision Sciences, European Journal of Operational Research, Energy, Industrial Management and Data Systems, and Journal of Business Ethics. His major research interests include decision support systems, RFID, manufacturing management, data mining, and supply chain management.

APPENDIX

Appendix A – Process Diagram



Appendix B – Polynomial Stepwise Regression

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-6.0379	0.7528	64.34	<.0001		0.002
Accessibility*Overall_Satisfaction	1	0.2873	0.0641	20.08	<.0001		1.333
Expection*Overall_Experience	1	0.1319	0.0504	6.84	0.0089		1.141
IMP_Find_New_Experience*Learn_New_Culture	1	0.1323	0.0387	11.67	0.0006		1.141
Location*Promotion	1	-0.2197	0.0587	14.02	0.0002		0.803
Overall_Satisfaction*Reputation	1	0.1802	0.0649	7.72	0.0055		1.197