Comparative Analysis of Data Mining Classification Techniques for Prediction of Problematic Internet Shopping

XUAN-LAM DUONG[1], SHU-YI LIAW[2*]
[1]Department of Business Administration, Thai Nguyen University of Agriculture and Forestry
Quyet Thang Commune, Thai Nguyen City, Thai Nguyen Province
VIETNAM
[2]Department of Business Administration, National Pingtung University of Sciene and Technology
No. 1, Shuefu Road, Neipu, Pingtung 912301
R.O.C TAIWAN

Abstract: As online shopping has surged, so do disorders on internet purchasing. This study aims to develop and compare predictive models that use data mining methods to predict problematic internet shopping. We used the Artificial Neural Network (ANN), CHAID with bagging, and C5.0 and compared them with traditional logistic regression to construct predictive models on a training cohort of 858 shoppers. Another cohort of 368 buyers was utilized to confirm the accuracy of the predictive model. The accuracy, sensitivity, specificity, and the ROC-AUC were used to assess the predictive performance. The C5.0 algorithm provided better accuracy in predicting PIS than the other models, indicating that C5.0 might be a practical auxiliary algorithm for predicting PIS. Our research findings cater to a comprehensive PIS prediction system, providing timely intervention and appropriate support to individuals with the PIS problem.

Key-Words: problematic internet shopping, machine learning, data mining, online shopping

Received: February 15, 2024. Revised: April 14, 2024. Accepted: May 14, 2024. Published: June 14, 2024.

1 Introduction

Online shopping has become a frequently-used alternative to traditional brick-and-mortar stores by offering substantial convenience and benefits to people's lives. However, for the minority population, the buying-shopping urge becomes uncontrolled or excessive, causing severe financial and psychological consequences in individuals' routines [1]. Extant literature on psychological and consumer behavior has proposed several terms to characterize problematic buying-shopping behavior, including compulsive buying [2], shopping addiction [3, 4], pathological buying [5], and buying-shopping disorder [6]. Research has shown that specific internet attributes such as availability, anonymity, accessibility, and affordability contribute to developing and maintaining an online subtype of buying shopping disorder [5, 7].

Research has scrutinized problematic buying-shopping from different prospects, and so far, no agreed-upon definition has been reached. This broad etiological spectrum adds more complexity to the theoretical explanation and requires further empirical evidence to determine global diagnostic criteria for problematic buying-shopping. Although problematic internet shopping has not been formally included in any monopoly classification of diseases, it has been hypothesized as a behavioral addiction in the literature [1, 3]. Pathological buying online – another derivative of problematic online shopping has been postulated as a sub-type of Internet addiction [5] that might adversely influence one's daily and social routine and economic status [3]. The authors are on the side, supporting that a more neutral term that does not directly imply that the behavior is addictive would be better when referring to uncontrolled and excessive online behavior. Therefore, we use "problematic internet shopping," a more neutral expression in agreement with previous studies [8, 9], to refer to the online version of problematic buying-shopping behavior.

The growing incidence of problematic internet buying/shopping requires a quick and efficient prediction system [10]. However, thus far, no study has employed data mining algorithms and techniques to detect unregulated online buying/shopping behavior. Therefore, the current study sought to develop predictive models and compare the

predictive performance of several classification algorithms to provide a basis for early intervention and proper support for those who might be at risk of problematic internet buying/shopping.

2 Literature review

Big data has significantly influenced the electronic commerce industry and will likely continue act in this way. E-retailers are counting on cloud-based big data analytics to harness the power of big data. Prior research has applied machine learning algorithms and data mining techniques to analyze problematic online behaviors. The methods for predicting whether a person can suffer from problematic online behavior can benefit the medical field and individuals. Nevertheless, a handful of papers have tackled the issue, such as Arora *et al.* [11], who discuss the role of machine learning in assessing the addictive use of various online technologies and its influence on mental and emotional health.

According to a recent systematic review, a bulk of studies in addiction research employed machine learning to predict substance addiction [12], leaving a handful of research that has tackled the issue of internet-related addictive behaviors, such as internet addiction [13, 14] or problematic smartphone use severity [15]. Efforts have been made to implement machine learning in diagnosing and detecting problematic buying or shopping using different methods or combining it with several algorithms to enhance accuracy [16]. For example, Prashar *et al.* [17] employ multiple machine learning classifiers to predict impulsive buying behavior. Their findings suggest the superiority of logistic regression regarding predictive power to other techniques. In contrast, Prashar and Mitra [18] provide statistical evidence that SVM surpasses logistic regression, linear discriminant analysis, quadratic discriminant analysis, and *k*-Nearest Neighbor methods in predicting power. Problematic internet buying or shopping is becoming prevalent in our consumer society. Surprisingly, less is known about the predictivity of problematic internet shopping using data mining algorithms.

3 Methodology

3.1 Data collection and sample

An online questionnaire was designed and administered to acquire demographic information, internet usage habits, and perceived online shopping benefits and risks. The authors developed an online questionnaire that might take approximately 20 minutes to comprehend. The first part of the questionnaire, aiming at collecting general demographic characteristics and everyday internet use statistics, is followed by a battery of validated scales (*see* Table 1). After pre-testing to mitigate any ambiguous wording or administering problems, the revised questionnaire was disseminated online to internet users, those who had shopped online over their last twelve months via communication applications and social networks, such as Facebook, Line, and WeChat. The cover of the questionnaire contains information about the purposes of the study. Respondents are asked to provide written informed consent before proceeding to the body of the questionnaire. Data were collected anonymously and treated confidently. The author put numerous efforts to improve the quality of the data. The online survey management system enables us to identify duplicate responses and records that contain unusual patterns (i.e., straight-lining or answers completed in an abnormally instant manner), enhancing the accuracy and appropriateness of the data. There is no missing data in our dataset. Finally, the data screening and outliers removal procedure resulted in 1,226 eligible respondents available in our dataset.

3.2 Measures

We adopt validated scales measuring the perceived benefits and risks of online shopping. All measures were assessed on a seven-point Likert-type scale where 1 = strongly disagree, 7 = strongly agree. Cronbach's alpha and McDonald's omega are used to evaluate the scale's reliability, whereas the latter is used to overcome the deficiencies of alpha. The analysis results showed that the Cronbach's α of all variables ranged from 0.759 to 0.897 while the McDonald's' ω ranged from 0.770 to 0.898. The minimal differences between the two measures demonstrate that the scales have adequate reliability [19].

The Online Shopping Addiction Scale [22, 23] was adapted to measure problematic internet shopping severity. The 18-item Likert-type scale measures problematic internet shopping based on six core components of addictive behaviors – i.e., salience, mood modification, tolerance, withdrawal symptoms, conflict, and relapse [24]. The mean scores and their corresponding standard deviations were used to group respondents into two categories - i.e., regular and problematic internet shoppers using the cut-off score of one standard deviation above the mean.

Table 1. List of measures

Variables	Number of items	Cronbach's α	McDonald's ω	References
Information search	4	0.852	0.854	[20]
Recommendation system	4	0.816	0.821	[20]
Dynamic pricing	4	0.782	0.786	[20]
Customer service	4	0.883	0.883	[20]
Privacy	4	0.767	0.770	[20]
Security	4	0.759	0.765	[20]
Group influence	4	0.897	0.898	[20]
Deception	4	0.866	0.867	[21]

Afterward, the authors submitted eight variables measuring the benefits and risks of online shopping (Table 1) and respondents' demographic characteristics and treated them as input variables in the predictive models. This information includes age, gender, marital status, education level, internet experience, daily internet usage, daily internet shopping usage, and monthly budget for internet shopping. All the continuous variables were standardized to enhance the interpretation capability (e.g., age, internet experience, internet usage, and internet shopping usage). Categorical variables were transformed into numerical values. Data mining prediction models were constructed using SPSS Modeler version 18.0 (SPSS Inc., Chicago, IL, USA).

3.3 Data analysis and model building

The authors employed three supervised machine learning algorithms, namely Artificial Neural Network (MLP) with bagging, CHAID, and C5.0, to construct the predictive models using the holdout testing method. The performance was then used to compare with the traditional Logistic regression. The original dataset was arbitrarily split into two sets, with the training dataset comprising about 70% of the respondents (n = 858) and the testing dataset including 30% of the participants (n = 368). The PIS score was treated as a target variable, whereas the sum score of eight factors of internet purchasing, demographic characteristics and internet use patterns were treated as input variables in predictive models. We consult the overall accuracy and additional metrics such as ROC curves, AUC, and Gini to assess the predictive performance.

4 Results

4.1 Descriptive statistic

The sample comprises 519 (42.33%) males and 707 females (57.67%). The mean age was 31.28 (SD=9.81), with a median of 29, ranging from 17 to 70. Regarding marital status, 587 (47.88%) were single or without stable partners; 639 (52.12%) reported being in a relationship. From the educational level point of view, 403 (32.87%) attained high school diplomas, 627 (51.14%) obtained a university/college degree, and 196 (15.98%) owned a graduate's degree. On average, respondents have more than 13 years of internet experience. They spent over 6.6 hours on internet use but only consumed approximately 1.7 hours per day for online shopping-related activities. More than 85% of the respondents reported spending less than $1,000 for online shopping every month. This sample is deemed appropriate because younger people, females, and those with higher online shopping frequency are more prone to manifest problematic internet shopping.

4.2 Predictive performance

Table 2 shows the results from the predictive models. ANN C5.0, and CHAID outperform logistic regression in all performance evaluation statistics in the training dataset.

The Artificial Neuron Network (MLP) achieved a classification accuracy of 77.51% with a sensitivity of 80.46% and a specificity of 74.46%; the CHAID with bagging achieved a classification accuracy of 84.73%, with a sensitivity of 84.37% and a specificity of 85.11%. However, the C5.0 classifier performed best among the four evaluated models. The C5.0 model had a classification accuracy of 86.06%, with a sensitivity of 85.16 and a specificity of 88.09%. The AUC values across four models ranged from 0.769 (LR) to 0.942 (C5.0), and the Gini ranged from 0.537 to 0.885, respectively.

Similar patterns can be observed in the testing dataset, where the decision tree (C5.0) outperforms other models in all respective performance indicators.

Table 2. The performance of predictive models

Dataset	Logistic regression		Artificial Neural Network (MLP)		CHAID		Decision tree (C5.0)	
	C^+	C^-	C^+	C^-	C^+	C^-	C^+	C^-
Training dataset								
True Positive	308	127	350	85	367	68	373	65
True Negative	134	289	108	315	63	360	50	370
Accuracy (%)	69.58		77.51		84.73		86.60	
Sensitivity (%)	70.80		80.46		84.37		85.16	
Specificity (%)	68.32		74.46		85.11		88.09	
AUC	0.769		0.849		0.925		0.942	
Gini	0.537		0.699		0.851		0.885	
Testing dataset								
True Positive	128	58	141	45	124	62	148	32
True Negative	57	125	66	116	62	120	34	154
Accuracy (%)	68.75		69.84		66.30		82.07	
Sensitivity (%)	68.82		75.81		66.67		82.22	
Specificity (%)	68.68		63.74		65.93		81.91	
AUC	0.731		0.749		0.740		0.921	
Gini	0.462		0.498		0.480		0.843	
Total dataset								
True Positive	436	185	491	130	491	130	524	97
True Negative	191	414	174	431	125	480	84	521
Accuracy (%)	69.33		75.20		79.20		85.24	
Sensitivity (%)	70.21		79.06		79.06		84.38	
Specificity (%)	68.43		71.24		79.34		86.11	
AUC	0.758		0.819		0.873		0.936	
Gini	0.515		0.638		0.746		0.873	

Note: C^+ denotes the count of predictive positive; C^- denotes the count of predictive negative

The CHAID model appeared to be the worst prediction model, with a classification accuracy of 66.30%, lower than the 68.75% obtained from Logistic regression. Also, there are minimal differences between the sensitivity and specificity of the CHAID and the logistic regression models. The C5.0 predictive decision tree achieved the highest accuracy of 82.07%, with 82.22% sensitivity and 81.91% specificity. The AUC across models ranges from 0.731 to 0.921, while the Gini ranges from 0.462 to 0.843. In the whole sample, the accuracy, sensitivity, and specificity of the logistic regression model were 69.33%, 70.21%, and 68.43%, respectively. While predictive performance differences between the Neural Network and the CHAID are minimal yet, the C5.0 achieved the highest accuracy (85.24%), sensitivity (84.38%), and specificity (86.11%). Also, the AUC and Gini of C5.0 were superior comparing to the others.

As shown in Figure 1 for the training sample, the four predictive models manifested differently, except for the substantial convergence between CHAID and the C5.0. Also, the CHAID and the C5.0 achieved the highest predictive performance, whereas the logistic regression showed the worst predictive capability.

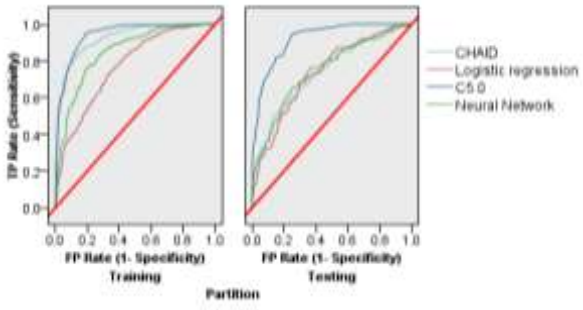

Figure 1. ROC curve of four classifiers

While the C5.0 maintains its strong forecasting capability for the testing sample, the other three models exhibit a minimal distinction.

4.3 Predictor importance

The sensitivity analysis was performed where each variable is placed in order of its relative importance, giving the objective is to determine the relative importance of each of the 16 independent variables within different models. The results from the sensitivity analysis are presented in Table 3.

Table 3. The importance of the input variables in four models

Order[a]	Logistic regression	Neural network (MLP)	C5.0	CHAID (bagging)
1	Shopping usage	Shopping usage	Shopping usage	Security
2	Internet experience	Security	Education	Shopping usage
3	Recommendation system	Recommendation system	Group influence	Recommendation system
4	Security	Internet experience	Age	Dynamic pricing
5	Privacy	Information search	Information search	Information search
6	Monthly budget	Customer service	Security	Privacy
7	Deception	Dynamic pricing	Deception	Customer service
8	Dynamic pricing	Privacy	Privacy	Group influence
9	Education	Age	Customer services	Internet experience
10	Information search	Group influence	Monthly budget	Deception
11	Gender	Internet usage	Marital status	Monthly budget
12	Internet usage	Education	Gender	Gender
13	Marital status	Deception	Dynamic pricing	Internet usage
14	Age	Monthly budget	Internet usage	Education
15	Group influence	Marital status	RS	Marital status
16	Customer service	Gender	Internet experience	Age

[a]The order according to importance, from the most to the least important.

Accordingly, the predictor performance indicates that daily online shopping time, security, and recommendation systems were the most critical predictors explaining PIS in our study samples. In contrast, gender, age, marital status, and daily internet usage manifested a modest role in predicting problematic internet shopping. The excessive time spent on internet shopping appears to be the most critical indication of problematic internet shopping, followed by the internet experience and the effects of the recommender systems, which could influence consumers' online shopping. Conversely, gender, age, and marital status show a minimal explanation for detecting individuals who might be at risk of PIS. Prior studies have employed predictive models for different purposes of interest. The results of this study indicate that the C5.0 decision tree is the best classifier, with 85.24% accuracy on the whole dataset. The CHAID model came out second, with 79.20% accuracy, leaving the ANN model the poorest performance among the three models with 75.20% accuracy. Overall, the predictive capability of C5.0, CHAID, and ANN are much higher than the logistic regression. These findings align with prior investigations [25].

5 Conclusion

Our study constructed and compared three data mining algorithms to predict the problematic internet shopping behavior and found that decision trees – i.e., C5.0 and CHAID outperformed Neural Network and Logistic Regression in classifying consumers in the 'at risk' group. Predicting those at risk of PIS is crucial in management decision-making since timely diagnosis is coupled with more advantageous treatment outcomes. We expect data mining methods such as C5.0 and CHAID could serve as effective alternatives to conventional logistic regression in identifying the critical variables more accurately and timely. Nevertheless, it is worth scrutinizing more complex machine learning algorithms (*i.e.*, Random Forest, Xgboost, etc.), albeit the predictive performance of C5.0 and CHAID in the current study prevails over other machine learning algorithms.

There are several limitations to our study. On the one hand, the sample was collected from a comparatively sizeable non-clinical sample with a minimal possibility of obtaining information from those diagnosed with PIS. The authors contemplate that the strength of employing data mining algorithms to specify the diagnostic criteria of PIS may be more fully verified in more extensive or even more diverse populations. A further limitation was that the present study employs a scale measuring problematic internet shopping from an addiction perspective, a more adverse and stringent condition that might result in fewer problematic internet shoppers being detected than it would have been capable of. We believe that a new scale that explicitly measures problematic buying/shopping behavior might be beneficial in detecting relevant observations.

References:

[1] Lejoyeux, M., and Weinstein, A, "Shopping Addiction," *Principles of Addiction*, 847–853. (2013).

[2] Weinstein, A., Maraz A, M., Griffiths, M. D., Lejoyeux, M., and Demetrovics, Z, "Compulsive buying – features and characteristics of addiction. In: *Neuropathology of Drug Addictions and Substance Misuse*, **3**, pp.993-1007. Academic Press, 2016.

[3] Andreassen, C.S., Griffiths, M. D., Pallesen, S., Bilder, R. M., Torsheim, T., and Aboujaoude, E, "The Bergen Shopping Addiction Scale: reliability and validity of a brief screening test," *Frontiers in Psychology*, **6**:1374 (2015).

[4] Rose, S., and Dhandayudham, A, "Towards an understanding of Internet-based problem shopping behavior: The concept of online shopping addiction and its proposed predictors," *Journal of Behavioral Addictions*, **3**(2), 83-89 (2014).

[5] Trotzke, P., Starcke, K., Müller, A., and Brand, M, "Pathological buying online as a specific form of Internet addiction: a model-based experimental investigation," *PloS ONE*, **10**(10: e0140296 (2015).

[6] Müller, A., Steins-Loeber, S., Trotzke, P., Vogel, B., Georgiadou, E., and de Zwaan, M, "Online shopping in treatment-seeking patients with buying-shopping disorder," *Comprehensive Psychiatry*, **94**: 152120 (2019).

[7] Dittmar, H., Long, K, and Bond, R, "Association between materialistic values, emotional and identity-related buying motives, and compulsive buying tendency online," *Journal of Social and Clinical Psychology*, **26**(3), 334-361 (2007).

[8] Ko, Y. M., Roh, S. W., and Lee, T. K, "The association of problematic internet shopping with dissociation among South Korean internet users," *International Journal of Environmental Research and Public Health*, **17**: 3235 (2020).

[9] Lam, L., and Lam, M. K, "The association between financial literacy and Problematic Internet Shopping in a multinational sample," *Addictive Behaviors Reports*, **6**, 123-127 (2017).

[10] Gori, A., Topino, E., and Casale, S, "Assessment of online compulsive buying: Psychometric properties of the Italian compulsive online shopping scale (COSS)", *Addictive Behaviors*, **129**, 107274 (2022)

[11] Arora, A., Chakraborty, P., and Bhatia, M. P. S, "Problematic Use of Digital Technologies and Its Impact on Mental Health During COVID-19 Pandemic: Assessment Using Machine Learning," in I. Arpaci, M. Al-Emran, M. A. Al-Sharafi, and G. Marques (Eds.), *Emerging Technologies During the Era of COVID-19 Pandemic,* Springer International Publishing, 197-221 (2021).

[12] Mak, K.K., Lee, K., and Park, C, "Applications of machine learning in addiction studies: a systematic review," *Psychiatry Research*, **275**, 53-60 (2019).

[13] Di, Z.L., Gong, X.L., Shi, J.Y., Ahmed, H.O.A., and Nandi, A.K, "Internet addiction disorder detection of Chinese college students using several personality questionnaire data and support vector machine," *Addictive Behaviors Reports*, **10**: 100200 (2019).

[14] Ioannidis, K., Chamberlain, S.R., Treder, M.S., Kiraly, F., Leppink, E.W., Redden, S.A., Stein, D.J., Lochner, C., and Grant, J.E, "Problematic internet use (PIU): Associations with the impulsive-compulsive spectrum. An application of machine learning in psychiatry," *Journal of Psychiatry Research*, **83**, 94-102 (2016).

[15] Elhai, J.D., Yang, H.B., Rozgonjuk, D., and Montag, C, "Using machine learning to model problematic smartphone use severity: The significant role of fear of missing out," *Addictive Behaviors*, **103**: 106261 (2020).

[16] Nawodya, A. G. S., and Kumara, B. T. G. S, "Machine learning approach to detect online shopping addiction, *2nd International Conference on Advanced Research in Computing (ICARC)*, Belihuloya, Sri Lanka (2022).

[17] Prashar, S., Parsad, C., and Vijay, T. S, "Predicting impulsive buyers: A comparative study of binary classifiers' discriminative ability," *International Journal of Strategic Decision Sciences (IJSDS)*, **7**(2), 1-17 (2016).

[18] Prashar, S., and Mitra, S. K, "Forecasting impulsive buying behavior: a comparative study of select five statistical methods," *International Journal of Business Forecasting and Marketing Intelligence*, **3**(3), 289-308 (2017).

[19] Hair, J, F., Babin, B, J., Anderson, R. E., and Black, W. W, *Multivariate Data Analysis 8th Edition*. CENGAGE (2019).

[20] Le, T.M and Liaw, S.-Y, "Effects of Pros and Cons of Applying Big Data Analytics to Consumers' Responses in an E-Commerce Context," *Sustainability, 9*, 798 (2017).

[21] Román, S, "The ethics of online retailing: A scale development and validation from the consumers' perspective," *Journal of Business Ethics*, **72**, 131-148 (2007).

[22] Zhao, H-Y., Tian, W., and Xin, T, "The development and validation of the Online

Shopping Addiction Scale," *Frontiers in Psychology*, **8**:375 (2017).
[23] Duong, X-L., and Liaw, S-Y, "Psychometric evaluation of Online Shopping Addiction Scale (OSAS)", Journal of Human Behavior in the Social Environment, **32** (5), 618-628 (2023)
[24] Griffiths, M, "A 'components' model of addiction within a biopsychosocial framework," *Journal of Substance Use,* **10**(4), 191-197 (2005).
[25] Meng, X-H., Huang, Y-X., Rao, D-P., Zhang, Q., and Liu, Q, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *Kaohsiung Journal of Medical Sciences*, **29**, 93-99 (2013).

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.