8-1-2010

Application of a Data Mining Process Model: A Case Study- Profiling Internet Banking Users in Jamaica

Gunjan Mansingh
The University of the West Indies, gunjan.mansingh@uwimona.edu.jm

Lila Rao
The University of the West Indies, lila.rao@uwimona.edu.jm

Kweku-Muata Osei-Bryson
Virginia Commonwealth University, KMOsei@vcu.edu

Annette Mills
University of Canterbury, annette.mills@canterbury.ac.nz

Application of a Data Mining Process Model: A Case Study- Profiling Internet Banking Users in Jamaica

Gunjan Mansingh
The University of the West Indies
gunjan.mansingh@uwimona.edu,jm

Lila Rao
The University of the West Indies
lila.rao@uwimona.edu.jm

Kweku-Muata Osei-Bryson
Virginia Commonwealth University
KMOsei@vcu.edu

Annette Mills
University of Canterbury
annette.mills@canterbury.ac.nz

ABSTRACT

Internet banking has become widely available in Jamaica and yet there have been few studies to understand the characteristics of its users. For banks to improve their service and similar services it becomes imperative that they can justify the costs associated with these services. One of the ways these costs can be justified is if their customer base of internet banking users was to increase, this can be achieved if the characteristics (i.e. both demographic and behavioural) of the users can be identified. In this study, survey data of internet banking users and non-users was analyzed to develop predictive models which can be used to improve the uptake of this service. CRISP-DM methodology was used to improve the dependability and quality of the data mining results.

Keywords

Internet banking, Demographic and behavioural profiling, Data mining methodology.

INTRODUCTION

Internet banking involves consumers using the Internet to access their bank account, to undertake banking transactions (Sathye, 1999). For many organizations offering online services there are alternative modes of interaction. For example, in the case of banking the alternatives may include over-the-counter services, ATMs, telephone-banking and internet banking. As firms invest in alternative ways to provide and improve service delivery it is important to understand the factors that motivate an individual to use a particular channel.

Understanding the demographics and behavioural characteristics of customers that use online channels is a key to the successful diffusion of such innovations. Internet banking was introduced in Jamaica in July 2003. Currently all of the commercial banks offer internet banking, as these banks are usually the market leaders in terms of technology other financial institutions are following suit. Profiling those that use internet banking can offer a number of benefits to the banking sector in Jamaica. Knowing the demographics and the behavioural characteristics of those that are likely to use internet banking will help to determine the size of the possible customer base. This will help the business decision makers in targeting those customers who are likely to switch to this mode of banking and therefore more direct marketing can be done which can have considerable cost saving implications.

Although there are a number of existing studies related to understanding the phenomena of internet banking both in developing and developed countries, there is limited research related to the Caribbean and, more importantly, there is limited work in combining both demographic and attitudinal characteristics in a single study. Analysis of such data can be done through the use of data mining methods. More and more organizations are recognising the benefits that can be realized from the use of data mining and are adopting data mining tools in their decision making process. However, as data mining becomes an integral decision making tool it is important that standards are established for this process. Some standards for the data mining process are emerging and their usage will help to ensure that the process of data mining is reliable and repeatable, especially for those persons who are not experts in the field. One such standard is the CRoss Industry Standard Process for Data Mining (CRISP-DM) which is a methodology that provides a structure to data mining efforts by defining a set of steps that should be followed when mining (www.crisp-dm.org). The process begins with understanding the business problem, then capturing and understanding data, applying data mining techniques, interpreting the results, and deploying the knowledge gained.

This paper describes how data mining, using the CRISP-DM methodology, was used in analysing both demographic and behavioural characterises of internet banking users in Jamaica. This will help banks and financial institutions gain a thorough understanding of their customer base and their potential customer base. In the next section, we review the relevant literature in the area. Background information on data mining methodologies and internet banking is presented. This is followed by a section that describes how the CRISP-DM methodology was used to identify the characteristics of the internet banking users. The paper concludes with a summary, outlining the implications of the findings and future directions for research.

LITERATURE

Data analysis techniques can be categorised as confirmatory or exploratory. In confirmatory studies the user provides a hypothesis whose validity is tested against the data. No new knowledge is generated in this method of data analysis. In exploratory or discovery studies new patterns are extracted from data without much guidance from the user. Data mining is an exploratory technique which involves extracting non-trivial, potentially useful patterns from datasets. This technique is robust to categorical and missing data. When using this technique the emphasis is not on verifying previously identified hypothesis but on identifying new patterns.

Prior research on internet banking data has been done using either confirmatory approaches or descriptive statistics. Confirmatory internet banking adoption suggests that factors such as relative advantage, compatibility with values, internet experience, banking needs, trialability, perceived risk, attitude, self-efficacy, and government support may influence individual intentions (Jaruwachirathanakul and D., 2005, Kolodinsky et al., 2004, Sohail and Shanmugham, 2003, Tan and Teo, 2000). Tan and Teo (2000) found that relative advantage, compatibility with values, internet experience, banking needs, trialability, and perceived risk impacted attitude, while intention to adopt internet banking was impacted by attitude and perceived behavioural control. Self-efficacy and government support were also key factors influencing perceived behavioural control.

A number of these studies have also examined the factors that affect internet adoption in particular countries (Furst et al., 2002, Heikki et al., 2002, Jaruwachirathanakul and D., 2005, Kolodinsky et al., 2004, Poon, 2008, Sathye, 1999, Sohail and Shanmugham, 2003, Tan and Teo, 2000). Tan and Teo (2000) sought to identify the attitudinal, social and perceived behavioural control factors that influence the adoption of internet banking in Singapore. In these studies the emphasis has been on understanding the behavioural characteristics that influence adoption in different countries (e.g. Thailand, Malaysia, Finland, U.S.A, Singapore).

Okazaki (2006) points to the lack of research in examining the relationship between consumers' attitude and their demographic characteristics. They address this by applying a two-step cluster analysis in profiling mobile Internet adopters in Japan. In an effort to build user profiles the authors established the need for using both demographic and attitudinal variables. These variables are a combination of continuous numeric and categorical data type, which can be analysed together by using data mining techniques.

There is a need for standards in data mining as it is claimed that they simplify the integration, updating, and maintenance of the applications and systems containing DM models (Grossman et al., 2002). Standards have been developed for defining the business processes used in data mining; one such standard is the CRoss-Industry Standard Process for Data Mining (CRISP-DM) which was designed to capture the data mining process (see www.crisp-dm.org). The model encourages best practices and offers organizations the structure needed to realize better, faster results from data mining (Shearer, 2000). The CRISP-DM process is divided into six phases each of which consists of a number of tasks (Clifton and Thuraisingham, 2001), figure 1 describes the steps and the tasks associated with each of the phases (Shearer, 2000). CRISP-DM methodology has been used in both research and practice for wide-ranging application domains (Kurgan and Musilek, 2006).

CASE STUDY – APPLYING CRISP-DM TO PROFILE INTERNET BANKING USERS

In this study we will use exploratory methods of analysing the behavioural and demographic characteristics of internet banking users. Data mining techniques will be used to extract the relevant characteristics which explain the internet banking users. To ensure that the data mining process is reliable and repeatable the cross-industry standard process model CRISP-DM will be used (Shearer, 2000).

CRISP-DM has six phases, in this paper the focus is on phases 1, 2, 3 and 4 which are business understanding, data understanding, data preparation and modeling phases (see Figure 1). SAS enterprise miner 4.3 was used as a tool for data preparation and modeling, and SPSS was used in the data understanding and data preparation phase.

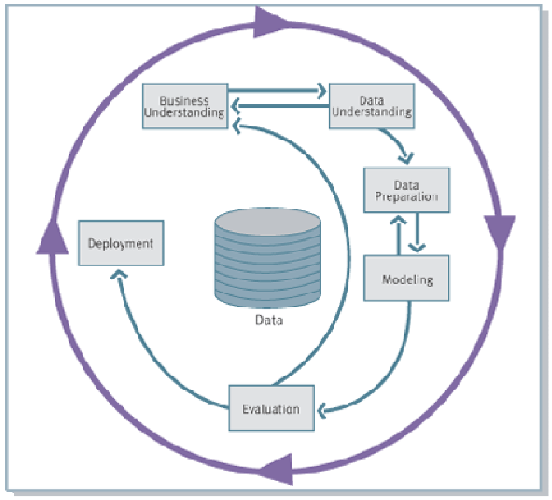


Figure 1. Phases of the CRISP-DM model.

Phase 1 - Business Understanding

The first step of the CRISP-DM methodology involves determining the business objectives. This is sometimes considered to be one of the most important phases of the data mining process. In terms of Internet Banking the banks' primary business goal would be to identify and increase their customer base of internet banking users by identifying both the demographic and attitudinal characteristics of current users. This will ensure that internet banking is financially viable and will identify ways to increase the adoption of this service in Jamaica. To achieve this business goal the data mining goals and their success criteria will also be identified in this phase.

Data Mining Goals and their Success criteria

One of the business objectives is to increase the customer base of internet banking users. Hence the corresponding data mining goal to achieve this business objective will be to build a predictive model using attitudinal and demographic characteristics of both users and non-users to identify characteristics of users. Several predictive models will have to be generated and their applicability and success will be determined by the data mining analyst based on the values of different performance measures. Generating a predictive model for a particular parameter setting is considered to be a relatively easy task but selecting the most appropriate setting requires that the data mining analyst generates several predictive models and analyzes them based on multiple performance measures (Osei-Bryson, 2004). To assess the different predictive models generated several performance criterion have been proposed by Osei-Bryson (2004). We will use accuracy, stability and lift to compare the models and to select the 'best' model(s).

In this phase a value function for each performance measure must be defined. The value functions for accuracy, lift and stability are given in table 1. Accuracy is the proportion of records in the validation data set that are correctly classified by a model. The threshold for accuracy is selected to be 70 meaning that models with an accuracy of greater than or equal to 70 will be selected. Predictive models which fall below this threshold will be rejected. Lift gives the relative improvement of a predictive model over a random guess. It is calculated at a particular decile and the values for the baseline and the exact model should be used. The exact model represents the perfect model and the baseline represents if no model were present. The value of the decile where the lift value is recorded is found using the distribution of the target variable; this will be set after the data preparation phase. The stability of the model will also be calculated by visually inspecting the non-cumulative % response chart. In this chart the posterior probabilities are sorted into deciles. The response rate which is the actual responses in the validation data set is plotted for each decile. In a stable model the earlier deciles should have high responses which decrease in the latter deciles. Hence a model will be considered to be stable if it the graph is a straight line or is gradually decreasing until the chosen decile.

These performance measures then need to be combined to arrive at the overall score, however, the measures may have different levels of importance and therefore could be weighted (see table 1). Analytic Hierarchy Process (AHP) (Saaty, 1980) can be used to determine this overall score as it is suited for situations that require the combining of a number of weighted measures.

Performance Measures	Description	Value function	Weight
Lift	Effectiveness of a predictive model. This can be found using the Cumulative Lift chart.	$Score_{Lift} = \frac{(Tree - Baseline)}{(Exact - Baseline)}$	$w_{Lift} = 0.20$
Accuracy	The proportion of records correctly classified in the validation dataset.	$Score_{Accuracy} = 1 -$ Misclassification Rate $_{validation\ dataset}$	$w_{Accuracy} = 0.65$
Stability	The non-cumulative % Response Lift Chart for a given model is used. The relevant percentile depends on the problem instance. Stability is binary, with 1 indicating a stable model and 0 indicating an unstable model.	$Score_{Stability} = 0$ or 1	$w_{Stability} = 0.15$

Table 1. Description of Performance Measures

Phase 2 – Data Understanding

The data understanding phase of the methodology consists of four steps, including the collection of the initial data, the description of the data, the exploration of the data, and the verification of its quality. Data on internet banking usage was collected using a field survey conducted in two of the parishes of Jamaica (Kingston & St. Andrew and St. Catherine) in 2005-2006. The research instrument used to collect data was a multi-part self-administered questionnaire. This questionnaire consisted of several attitudinal and demographic variables. All variables were measured using multi-item scales adapted from existing sources. Responses were captured using a 7–point Likert scale anchored at either end with "Strongly agree" and "Strongly disagree". Personal demographics such as age, gender, education, occupation, head of household, primary income earner and income group were also collected (see table 2).

This questionnaire was developed and pre-tested by using a convenient sample of graduate students at a tertiary educational institution. The feedback from the pre-test was used to modify the questionnaire. This modified questionnaire was then administered in a trial study, in which 50 forms were returned of which 40 were suitable for determining the reliability of the items on the questionnaire. In the final study 506 questionnaires were administered of which 480 were used based on the number of questions that were answered. It is the results of this final dataset that are being reported in this paper.

The internet banking dataset has 47 variables which include demographic characteristics, attitudinal characteristics and the banking practices of individuals. The demographic variables are all categorical with no multiple responses allowed. The attitudinal variables were continuous and were associated with general perceptions of the internet banking platform, willingness to access internet banking, compatibility, confidence and security issues in using internet banking. The banking practices variables were both categorical and continuous.

In the data collected approximately 34% of the respondents were internet banking users (see Figure 2). A total of 42.9% were male respondents and the majority (approximately 45%) of respondents were in the income range of JA $40,000-$99,999 gross monthly income. Approximately 70% of the respondents were under 35 years of age. Details of the demographic profiles of the users and non-users are presented in table 1. Details of the variable Occupation were not reported in this table as it had 147 values. This was addressed in the data preparation phase.

Phase 3: Data Preparation

In the data preparation phase all the tasks are focused on creating the final data set that will be used in the modeling phase. The CRISP-DM model identifies five stages in this phase: selection, cleansing, construction, integration and formatting of data.

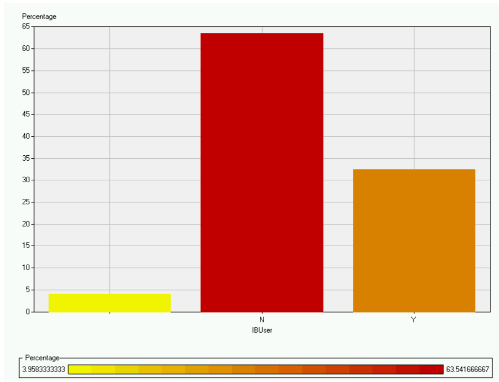


Figure 2. Internet banking users and non-users (IBUSER).

Prediction accuracy of models can be improved by excluding irrelevant variables, several techniques have been proposed for variable selection (Reunanen, 2003, Osei-Bryson et al., 2003). In the selection process some of the variables in the dataset were subjectively removed while others were selected for removal by SAS enterprise miner. Missing values for the behavioural variables were replaced as part of the cleansing process. To ensure the adequacy of the selected variables testing for the absence of multicollinearity has been recommended (Hair et al., 1998). This was investigated by examining the significance values in the correlation matrix.

	Users		Non Users	
Variables	**Freq.**	**%**	**Freq.**	**%**
Gender				
Male	83	54	110	37.2
Female	71	46	186	62.8
Age				
< 25	29	18.6	105	34.4
25-34	72	46.2	107	35.1
35-44	37	23.7	56	18.4
> 44	16	10.2	29	9.5
Income (gross monthly)				
< 40,000	19	13.4	101	41.1
40,000-99,999	61	43	113	45.9
100,000-149,999	30	21	15	6.1
> 150,000	32	22.6	17	6.9
Education Level				
Secondary education & less	15	9.6	73	23.9
Graduate or Professional degree	41	26.3	31	10.2
Other	2	1.3	14	4.6
Head of Household				
Yes	89	58.9	134	46.5
No	62	41.1	154	53.5
Primary Income Earner				
Yes	83	53.2	146	46.2
No	67	43	141	47.9

Table 2. Demographic profile.

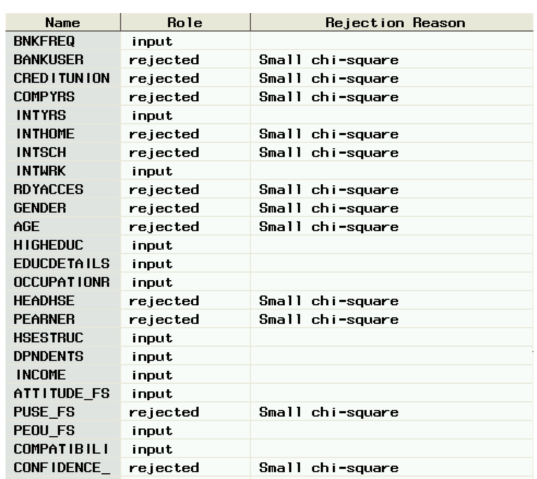
Confirmatory factor analysis was then performed, in SPSS, on the behavioural variables and all items loadings were greater than 0.70 (see Table 3). The factor score created was included in the data set; this score is a numerical value that indicates an individual's relative standing on a latent factor. For example, for a factor measuring security, an individual with a factor score of 1.32 is more concerned with security than an individual with a score of -0.85.

From the data understanding stage it was recognized that the variable *Occupation* had more than 147 values therefore as a part of the data construction process the *occupational classification* used in Jamaica was applied to this variable to reduce the number of possible occupation categories to 12. From the dataset the target variable (IBUSER) was identified and all input variables that directly predict the target variable were rejected by the data mining analyst (e.g. number of years of internet banking usage and satisfaction with internet banking).

Attribute	**Items**	**Variance Extracted**
Attitude	4	76.5
Perceived Use	5	73.8
Perceived Ease of Use	4	75.2
Compatibility	4	76.7
Confidence	3	91.6
Innovativeness	3	74.1
Perceived behavioural control	3	73.7
Subjective norm	3	88.2
Human Interaction	3	86.7
Perceived Cost	5	76.2
Switching	3	78.2
Security	3	81.4
Intention to use	3	88.0
Preference to use	3	92.6
Infusion	3	91.6

Table 3. Item loadings of the behavioural variables.

Additional variables were also rejected by using the variable selection node in SAS enterprise miner (see Figure 3). This node helps in reducing the number of input variables as it rejects variables that are not related to the target and identifies variables that are potential predictors. In this node chi-square method was selected as it is recommended in situations when the target variable in binary. In the transform variables node derived variables were created by combining security and confidence as it assumed that security is a part of confidence.



Name	Role	Rejection Reason
BNKFREQ	input	
BANKUSER	rejected	Small chi-square
CREDITUNION	rejected	Small chi-square
COMPYRS	rejected	Small chi-square
INTYRS	input	
INTHOME	rejected	Small chi-square
INTSCH	rejected	Small chi-square
INTWRK	input	
RDYACCES	rejected	Small chi-square
GENDER	rejected	Small chi-square
AGE	rejected	Small chi-square
HIGHEDUC	input	
EDUCDETAILS	input	
OCCUPATIONR	input	
HEADHSE	rejected	Small chi-square
PEARNER	rejected	Small chi-square
HSESTRUC	input	
DPNDENTS	input	
INCOME	input	
ATTITUDE_FS	input	
PUSE_FS	rejected	Small chi-square
PEOU_FS	input	
COMPATIBILI	input	
CONFIDENCE_	rejected	Small chi-square

Figure 3. Output of variable selection node – using chi-square.

Phase 4: Modeling

This phase involves selecting the modeling techniques, building multiple models by applying the modeling techniques to the prepared dataset and then assessing the generated models. In this study the objective was to understand the characteristics of users. Since the dataset consists of both users and non-users (i.e. a target variable exists) inferences can be performed on this data to make predictions. Logistic regression and decision trees were the models selected to classify the discrete target variable and develop profiles of internet banking users.

The logistic regression and decision tree (DT) models were built using SAS Enterprise Miner (EM) 4.3. The process flow can be seen in Figure 4. There are several options for parameter values, therefore, different parameter settings were explored to determine which were most suitable. Different splitting methods such as Entropy, Gini and Chi were used to build the predictive models. Predictive modeling requires partitioning the dataset into learning and assessment subsets. In SAS EM in the *data partition* node the dataset was split into 60% for learning and 40% for assessment. Stratified sampling based on the target variable was used in partitioning the dataset. Different decision trees were generated with different splitting criterion (e.g. Chi, Entropy and Gini). Separate trees were generated with and without replacing the missing values. Also trees were generated with and without variable selection node. In the logistic regression nodes stepwise selection method was selected.

After the models were generated the output was assessed using the performance measures and their corresponding value functions set is business understanding phase (see table 1). The value functions for stability and lift were based on figures 5 and 6. The number of records correctly classified was taken from the output of each model. For lift and stability the 3rd decile was selected for calculating the value function. This decile was selected by examining the distribution of the target variable (see figure 2).

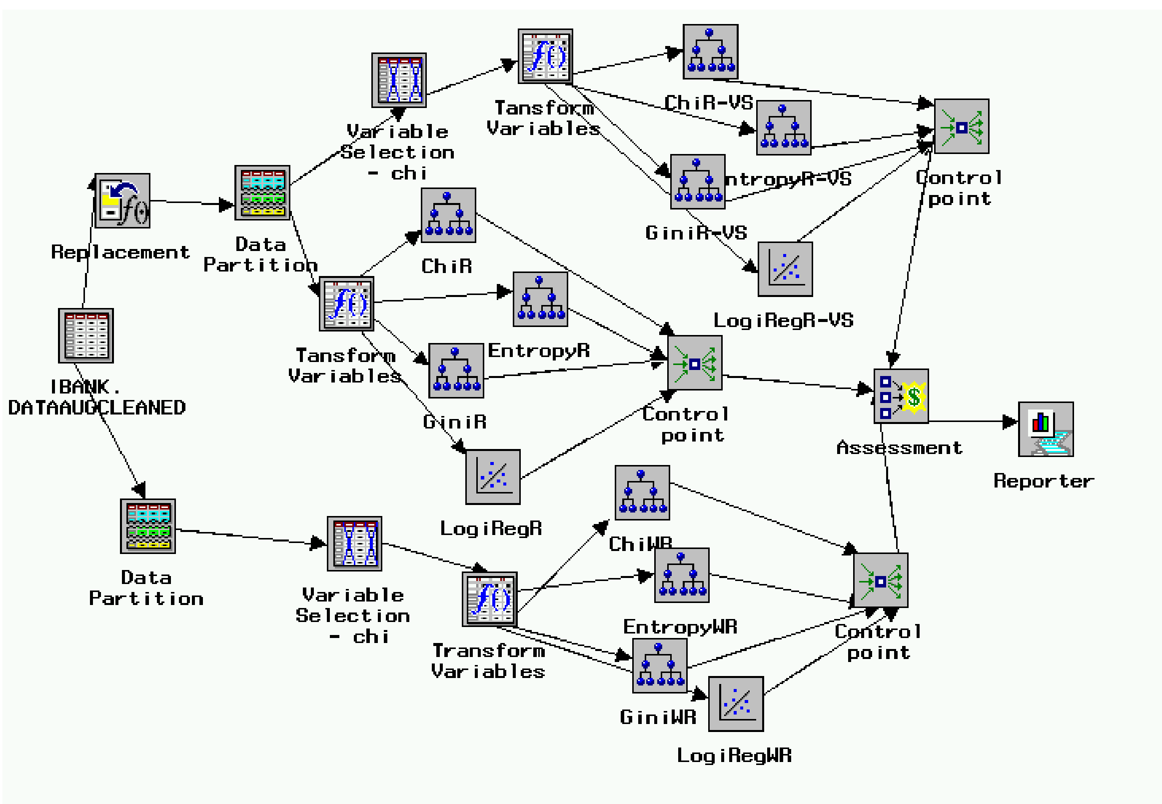


Figure 4. Process flow for building IBUSER models in SAS EM.

Models	Misclassification Rate	$Score_{Accuracy}$	Lift of Model	$Score_{Lift}$	$Score_{Stability}$	Overall Score
ChiWR-VS	0.254	0.746	58	0.482	1	0.731
ChiR-VS	0.227	0.773	60	0.517	1	**0.755**
EntropyWR-VS	0.286	0.714	52	0.379	1	0.689
EntropyR-VS	0.286	0.714	54	0.413	1	0.696
GiniWR-VS	0.286	0.714	52	0.379	1	0.689
GiniR-VS	0.286	0.714	54	0.413	1	0.695
RegWR-VS	0.345	0.655	46	0.275	0	0.055
RegR-VS	0.227	0.773	63	0.568	0	0.616
ChiR	0.227	0.773	60	0.517	1	**0.755**
EntropyR	0.286	0.714	54	0.413	1	0.696
Gini-R	0.286	0.714	54	0.413	1	0.696
RegR	0.281	0,719	57	0.465	1	0.710

Table 3. Performance Scores.

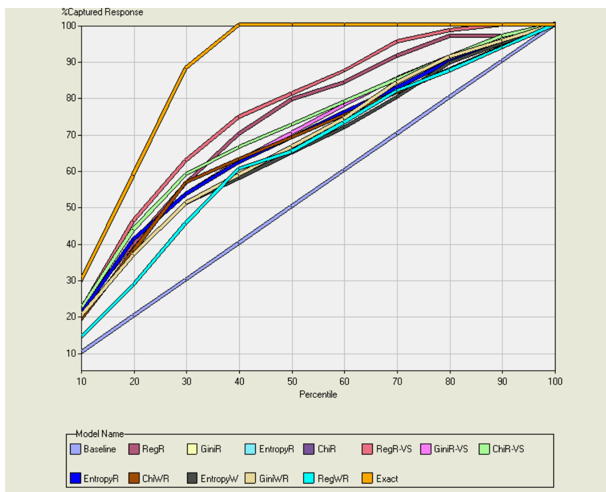


Figure 5. Lift.

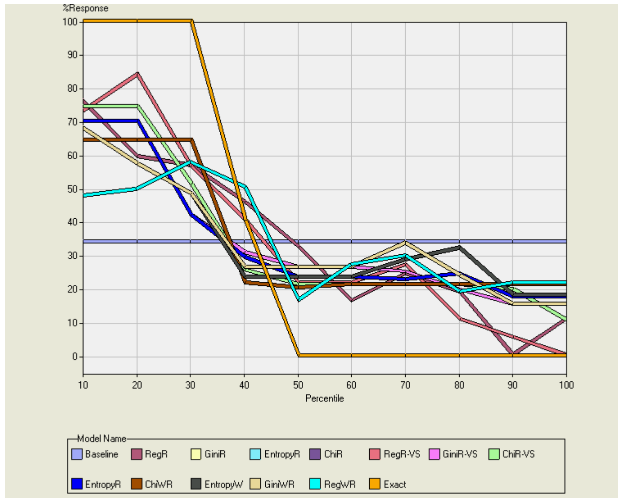


Figure 6. Stability.

After accuracy, lift and stability were calculated they were combined using the weights assigned to each performance measure to determine the overall score. The weights that were set in phase 1 (see table 1). From table 3 ChiR-VS and ChiR were considered to be the ‘best’ models based on the selected performance measures. They outperformed the other models on all measures. The rules generated from ChiR-VS and Chi-R are similar and are displayed in table 4, each rule represents the posterior probabilities for predicting the target variable (e.g. rule 4 states that if the frequency of performing banking transactions is high and an individual has access to the internet at work and has a positive attitude towards internet banking then there is a 87.3% probability that they are likely to be internet banking users). The results of these models and other models such as ChiWR-VS and RegR will be explored in future work.

English Rules of ChiR
IF ATTITUDE_FS < 0.438875 THEN IBUSER = {N: 87.3%, Y: 12.7%}
IF INTWRK EQUALS N AND 0.438875 <= ATTITUDE_FS THEN IBUSER = {N: 88%, Y: 12%}
IF Group: BNKFREQ EQUALS 1 2 3 AND INTWRK EQUALS Y AND 0.438875 <= ATTITUDE_FS THEN IBUSER = {N: 51.6%, Y: 48.4%}
IF Group: BNKFREQ EQUALS 4 5 6 7 AND INTWRK EQUALS Y AND 0.438875 <= ATTITUDE_FS THEN IBUSER = {N: 12.7%, Y: 87.3%}

Table 4. English rules –All rules for Decision Tree ChiR.

CONCLUSION AND FUTURE WORK

This results of the study showed that both demographic and behavioural characteristics are important in understanding the internet banking users. This is important for decision makers as it not only helps in identifying the characteristics of existing users and thus those non-users that are likely to adopt, but the important behavioural variables can also be used to focus the efforts in improving the service.

In the study the emphasis was on building predictive models, the methods used for building these models split the selected dataset into a training dataset (i.e. data used for building the model) and a validation dataset (i.e. data used for assessing model). Traditional confirmatory data analysis methods use all data for training thus risking the generation of an overfitted model. These methods also don’t check the stability of a model. Hence by following the CRISP-DM methodology and using a set of performance measures to evaluate multiple models we are seeking to improve the dependability and quality of the data mining results.

This study focused on only one of the business objectives identified in phase 1 of CRISP-DM, however, other objectives identified and will be explored using various data mining techniques. The results of the other good models (i.e. those models with an overall score close to the score of the best model) will be explored in future work to determine how the data mining goals satisfy the business objectives in the evaluation phase. The results of this study will be compared with other studies to determine the factors that are unique to Jamaica, the Caribbean and developing countries and those factors that are general to internet banking.

REFERENCES

1. Clifton, C. & Thuraisingham, B. (2001) *Computer Standards & Interfaces,* **23,** 187-193.
2. Furst, K., Lang, W. W. & Nolle, D. E. (2002) Program on Information Resources Policy, Center for Information Policy Research Harvard University.
3. Grossman, R. L., Hornick, M. F. & Meyer, G. (2002) Communications of the ACM. SPECIAL ISSUE: Evolving data mining into solutions for insights **45,** 59-61.
4. Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W. (1998) *Multivariate Analysis,* Prentice Hall.
5. Heikki, K., Minna, M. & Tapio, P. (2002) *Financial Services Marketing,* **6,** 346-361.
6. Jaruwachirathanakul, B. & D., F. (2005) *Internet Research,* **15,** 295.
7. Kolodinsky, J. M., Hogarth, J. M. & Hilgert, M. A. (2004) *The International Journal of Bank Marketing,* **22,** 238-259.
8. Kurgan, L. A. & Musilek, P. (2006) *The Knowledge Engineering Review,* **21,** 1-24.
9. Osei-Bryson, K.-M. (2004) *Computer & Operations Research,* **31,** 1933-1945.
10. Osei-Bryson, K.-M., Giles, K. & Kositanurit, B. (2003) *Journal of the Operational Research Society* **54,** 790-797.
11. Poon, W.-C. (2008) Journal of Business & Industrial Marketing **23,** 59-69.
12. Reunanen, J. (2003) Journal of Machine Learning Research, **3,** 1371-1382.
13. Sathye, M. (1999) International Journal of Bank Marketing, **17,** 324-334.
14. Shearer, C. (2000) Journal of Data Warehousing, **5**.
15. Sohail, M. S. & Shanmugham, B. (2003) *Information Sciences,* **150,** 207-217.
16. Tan, M. & Teo, T. S. H. (2000) Journal of the Association of Information Systems, **1,** 1-44.
17. www.crisp-dm.org Cross Industry Standard Process for Data Mining (CRISP-DM). Vol. 2009, pp.