Paper 1581-2014

Building Gamer Segmentation in the Credit Card Industry Using SAS® Enterprise Guide®

Yang Ge, Lancaster University

ABSTRACT

Credit card companies are competing each other by launching longer and longer 0% balance transfer (BT) period in recent 2 years, which has creating a new segmentation of customer. This is a group of customer who use credit cards as an interest-free loan by transferring their balances between cards during 0% balance transfer periods in order to avoid paying interest. These customers are called gamers. Gamers generate losses for banks due to paying no interest and making no purchases. Since gamers tend to have very good credit histories, it is ineffective to use risk scorecards to identify them according to the definition used by existing risk scorecards. This paper uses Naive Bayes classifier to create gamer scores which allow banks to track the proportion of gamers in population. This result has been achieved using variable selection through logistic regression and discretization through interactive grouping. The procedure is described in detail in this paper.

INTRODUCTION

In the credit card industry, one of the most important promotions for banks these days is to offer 0% balance transfer (BT) period. This promotion allows people to transfer their balances from one credit card to another with a small balance transfer fee, and enjoy a period of paying 0% interest on the transferred balance (called BT period). For banks, this promotion encourages people to move account from other banks to their own (Elzinga and Mills, 1998); therefore their account volume will increase and hence so will total profit. However, there are customers who have exploited this offer. After using credit card, they want to avoid interest payments for as long a period as possible, and therefore they transfer their balance to a BT offer credit card and payback their balance slowly during 0% BT period. During this time, they will rarely use the credit card for purchases (Ying, Iho, Echo, 2007), effectively using the credit card as an interest free loan. When the BT offer ends, they will transfer their unpaid balance to another BT offer account again, until their balance is repaid. These customers are called 'Gamers' in credit card industry. For banks, the money borrowed to customers could be borrowed from other financial institutes, which will generate an interest. This cost of holding this money is called 'cost of fund'. From gamers, banks can earn neither interest nor interchange fee, but will lose cost of fund. Ying, Iho, Echo (2007) mention that lifetime value of gamers is often negative. Therefore identifying gamers and effectively dealing with them is important.

In this article, we consider three different classifiers, namely Naive Bayes Classifiers, logistic regression and decision tree to detect behavior patterns associated with gamers using as input credit history information. In Naïve Bayes classifiers, variables will be selected through logistic regression and will be discretized through interactive grouping. It is considered as the simplest way but also the most performed method after comparison.

RELEVANT VARIABLES

Ying, Iho and Echo (2007) defined Gamers with two criteria: first is 'customers who engage in balance consolidation only', and the second is 'with little or no sales'. The first criterion suggests that gamers do not pay interest and the second one means they rarely use the credit card for purchase. To simplify this problem, the definition of gamer will be no interest and no purchase during 0% BT period and four months after that. The related information includes average amount of card debt (Andrew, Steven and Yiing, 2011), number of credit card holding (Delener and Katzenstein 1994) and credit limit (Kim and DeVaney 2001; Min and Kim 2003). Education is also relevant (Ausubel 1991; Canner et al. 1992; Calem and Mester 1995). In 2001, Chien, Kim and DeVaney showed that better educated card holders have higher tendencies to own card debt, which is a similar behavior with gamers. Age has also been found relevant but only for cardholders who have 5 cards or more (Andrew, Steven and Yiing, 2011).

VARIABLE ANALYSIS

COST MATRIX

This research is based on business project for a company which has been heavily engaged in offering BT cards since the launch of this type of product. Since profits constitute confidential information, no real profit value will be mentioned. In abstract terms, the profit generated from a non-gamer is assumed to be 10 units. After looking at the profit and loss generated by gamers and non-gamers, the loss from each gamer is approximately 1/10 of the profit from a non-gamer, hence is 1 unit. Therefore, if a gamer is correctly identified as a gamer, the bank will have no loss (True Positive=TP). If a gamer is identified as a non-gamer, the bank will lose 1 unit (False Negative=FN). On the

other hand, if a non-gamer is identified as a non-gamer, assuming the bank will then recruit this customer, it can still earn the 10 units from each non-gamer (True Negative=TN). However, if a non-gamer is identified as a gamer, so the bank will reject the application of this customer, the bank will not earn the 10 unit (False Positive=FP). This information is summarized in the cost matrix of Table1.

	Predicted Class		
Actual Class	C (i, j)	Non-Gamer	Gamer
	Non-Gamer	-10 (TN)	10 (FP)
	Gamer	1 (FN)	0 (TP)

Table 1 Cost matrix for the gamer identification problem

According to Table 1, the loss from falsely identifying a non-gamer is much higher than the loss from falsely identifying a gamer. However, since the measurement metric of accuracy treats each class as equally important, misleading results might be generated if the evaluation of the model relies purely on accuracy rate. Consequently, alternative measures should be considered when building and assessing classifiers. Two performance measurements will be used to evaluate classifiers. One is Receiver Operating Characteristic (ROC) curve, which depicts the trade-off between sensitivity and the specificity. In this application, the priority is to achieve high specificity, which is equivalent to minimizing the misclassification rate of non-gamers. The second performance measure is cost, which is computed using the cost matrix of Table1, in conjunction with the confusion matrix of a classifier.

VARIABLE EXPLORATION

Information of customers is collected from Equifax data, one of the most well-known credit profile checking website. Purchase history data spans the period from September 2011 to June 2013. The model only considers customers obtained from internet channel, including product comparing website, company website and advertisement click through customers. No missing values are present in the dataset. The ratio between gamer and non-gamer is found to be far smaller than 0.5, which can cause misleading results if class imbalance is not taken into consideration. To remove this bias, data can be modified using under-sampling, oversampling, or a hybrid approach, so that prior possibilities of each class become roughly equal. Using under sampling may cause some loss of information, whilst using over sampling runs the risk of replicating noisy data. To ameliorate the disadvantages of under- and over-sampling we adopt a hybrid approach.

In SAS® Enterprise Miner, the importance of variables can be investigated by comparing information value and Gini score, and by looking at the chart of Weight of evidence (WOE). Information value is the weighted sum of WOE over the groups and Gini score measures the discriminative power of variables. Both of these two criteria are positively related with the importance of variables. WOE contains information on being good (non-gamer) or bad (gamer) in data x, which can be explained by the formula below:

WOE(x)=log(p(x|Non-Gamer)/p(x|Gamer))

High negative values of WOE correspond to high risk of being gamer; high positive values correspond to low risk. To solve the business objective problem, a monotone increase or decrease curve of WOE is preferable, since it is more comprehensible. By looking Figure1, utilization, leverage, credit limit and number of card are preferred over other variables.

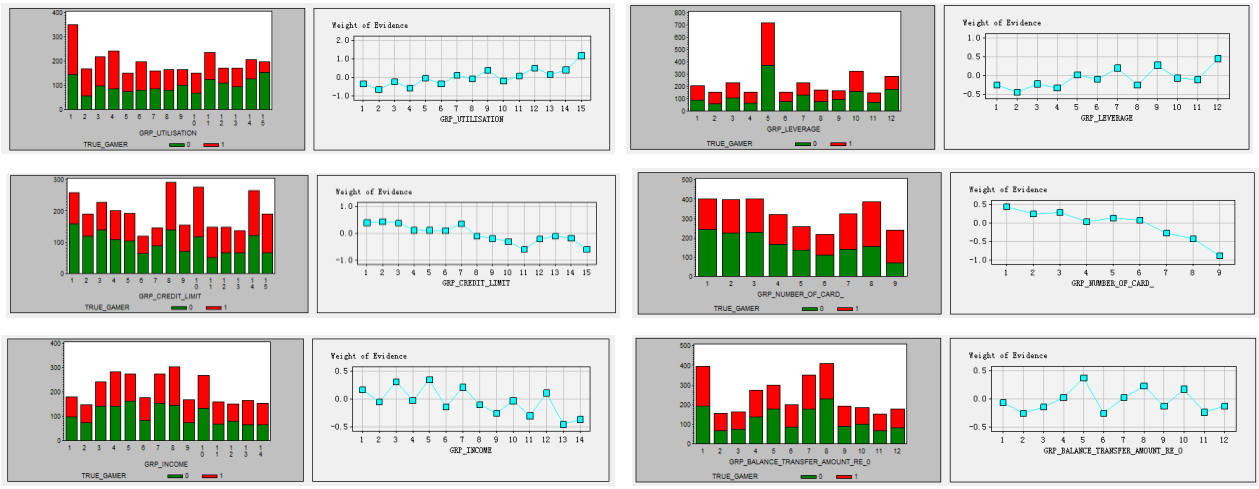

Figure 1 Weight of evidence of variable

METHODOLOGIES STEPHEN.STANDALOST@NATIONWIDE.CO.UK

NAIVE BAYES

Bayesian classification is based on Bayes theorem. Assuming there are two classes, labeled i=1, 0. 'if we define P(i|x) to be the probability that an object with measurement vector x = (x_1, . . . , x_p) belongs to class i, then any monotonic function of P(i |x) would make a suitable score.' Let f (x|i) be the conditional distribution of x for class i objects, and P(i) be the probability that an object will belong to class i if we know nothing further about it (the 'prior' probability of class i), then P(i|x) can be decomposed as proportional to f(x|i)P(i) (Xingdong, Vipin, et al., 2008). The formula is shown as follow:

$P(x,i)=f(x|i)P(i) = P(i|x)P(x)$

,where P(x,i) is the probability of class i and condition of vector x has been satisfied at the same time. In this report, P(1) represents the probability of a customer being gamer, while P(0) represents the probability of a customer being non-gamer. x represents all the variables to be selected, assuming all components of x are independent. Hence P(x,1) means the probability of customers who are gamer and falls into X segments.

To build classification, we need to estimate f(x|i) and P(i). P(i) can be easily estimated from training set if the data set is a random sample. To estimate the f(x| i), $f(x|\ i)=\sum f(x\_\ j|\ i),j=1,\ldots,p$, we should estimates each of the univariate distributions f(x_j |i), j = 1, . . . , p; i = 0, 1, separately. When x_j is continuous, 'then a common strategy is to segment each of them into a small number of intervals' (Xingdong, Vipin, et al., 2008). Since all the variables available in this report are continuous, the following model will be built based on discretized variables of the current ones. According to P(i|x), X can be segmented into three groups, Gamer low, Gamer mid and Gamer high respectively. The gamer proportion can then be estimated accordingly.

To apply the naïve Bayes classifier, we need to select which variables we will use. In this article, we used the stepwise selection in logistic regression. Based on training set data, which is from 2010 to 2012, the result of logistic regression from SAS® Enterprise Guide shows that only two variables have P-values smaller than 0.001, which indicates that only these two variables should be included in the model.

The DMREG Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	0.0123	0.1065	0.01	0.9084		1.012
Number_of_card_	1	0.0640	0.0103	38.93	<.0001	0.1642	1.066
Utilisation	1	-1.2898	0.2168	35.39	<.0001	-0.1667	0.275
Leverage	1	-1.3671	0.3824	12.78	0.0004	-0.1074	0.255
Income	1	-5.69E-7	2.116E-6	0.07	0.7878	-0.00711	1.000
Credit_limit	1	0.000011	5.203E-6	4.24	0.0396	0.0686	1.000
Balance_transfer_amo	1	0.000034	0.000015	5.25	0.0220	0.0577	1.000
Balance	1	7.26E-11	3.67E-10	0.04	0.8430	0.0746	1.000

Figure 2 Result of logistic regression

As discussed before, since all variables are continuous variables, they need to be discretized first. Using interactive grouping node in SAS® Enterprise Miner, discretization can be created as below:

Band	A	B			C			D			E			F
Number of card	0	1	2	3	4	5	6	7	8	9	10	11	12	12+

Band	A	B	C	D	E
Utilization	0%	0%-10%	10%-30%	30%-90%	90%+

Table 2 Discretization of selected variables

Based on training set, proportion of gamers in each combination of these two variables, P(i|x), can be found below:

	Number of card					
Utilization	a.0	b.1-3	c.4-6	d.7-9	e.10-12	d.>12
a.0	15.6%	20.7%	26.5%	35.3%	42.7%	38.3%
b.0-10%		22.3%	27.1%	30.4%	32.5%	38.5%
c.10-30%		18.2%	21.6%	25.6%	29.1%	36.2%
d.30-90%		12.8%	15.3%	18.3%	22.7%	29.1%
e.90%+		4.6%	6.6%	11.3%	25.0%	20.0%

Table 3 Gamer segmentation

According to the proportion of gamers, people falls into these combinations will be classified into three gamer-segments, Gamer-low, mid and high. The way to do it is as follow. To keep Gamer-mid segment has the same proportion of gamer as in blended population, these proportions have been ranked. In this rank, the first quartile of customers with highest gamer proportion will be classified as Gamer-high; second and third quartile of customers will be classified as Gamer-Mid, and the rest of customers will be classified as Gamer-low. As a result, 50% of customers will be in Gamer-mid segment, and 25% of customers will fall in Gamer-low, Gamer-high respectively. The result has been labeled in Table 3 with colors, from which we can see that the population with higher than 25% gamer

proportion will fall into Gamer-High segment, with lower than 16% gamer proportion will fall into Gamer-Low segment, rest will fall into Gamer-Mid segment.

Having had the model, 'gamer scores' can be created to approximate the proportion of gamer in a population. The proportion of gamers in Gamer low, mid and high segments has been known (P(i|x)). Therefore, when applying the gamer model into validation set (in practice, it will be another group of customers or another time period), the proportion of customers falls into each gamer segment P(X) can be got. The proportion of gamers in the whole population can be simulated as below,

P(Gamer)= P(i=1)= Σ (P(X)*P(i=1 | X) = Σ (P(Gamer segment=low/mid/high)*P(Gamer | Gamer segment=low/mid/high))

Gamer scores can be used to predict the proportion of gamers in a population month by month, according to the moving customer structure. This will tell the bank if a longer BT product should be launched or not.

Assuming that the card applicants who are in Gamer-high segment will be rejected by bank, based on test set which is the data from 2012 to 2013, the confusion matrix will be as below:

	Predicted Class using Naive Bayes classifier		
Actual Class	D (i, j)	Non-Gamer	Gamer
	Non-Gamer	62%	18%
	Gamer	12%	8%

Table 4 Confusion matrix of Naive Bayes classifier

Some criteria that will be used to compare models, which include:

Sensitivity=(true positive)/(true positive + false negative)=40%,

Specificity=(true negative)/(true negative + false positive)=77.5% and

Precision=(true positive)/(true positive + false positive) =70%.

According to the cost matrix, the total profit or loss can be approximate as total cost= C (i ,j)* D (i ,j), hence cost of the model will then be -4.27, which stands for 4.27 profits.

LOGISTIC REGRESSION

Logistic regression 'models the probability of some event occurring as a linear function of a set of predictor variables.' (Jiawei, Micheline, 2001). To assess the performance of logistic regression, response curve (lift chart), ROC curve and cost matrix should be reviewed. ROC curve and cost matrix has both been explained before. The response curve ranks the data according to the probability of people being gamer generated by the model from the highest to the lowest and divide them into 10 parts. The higher response rates in the first several deciles, the better is the model.

Since all the independent variables to be considered are continuous, and all are highly skewed, before building this model, transformations of variables should be considered. After trying different transformations for each variable, it turns out that when variable utilization, leverage, credit limit and balance required are transformed into logarithm, number of cards be standardized and income be transformed into square root, the model has the best performance. The two response curves and ROC curves below show that using transformed variables has higher response rate at the first decile and very similar specificity. Therefore the model using transformed variables is better than the model using original values.

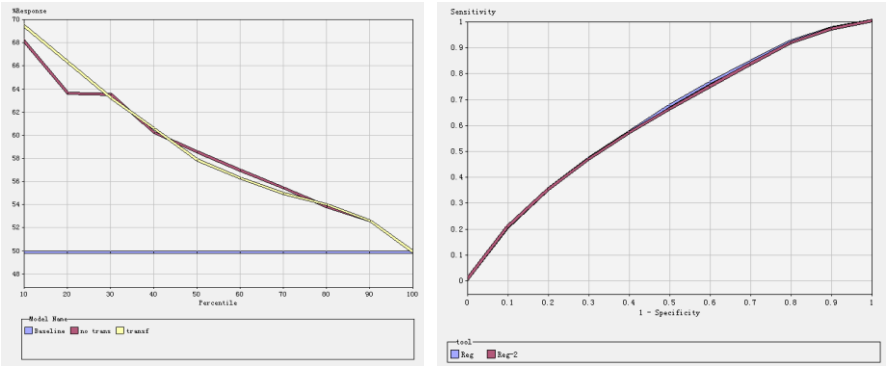

Figure 3 Lift chart (a) and ROC curve (b) to compare models with and without transformation

There are two methods to build this model. First method is to use the transformed value directly; the other method is to use weight of evidence (WOE) from interactive grouping node. For each method, variables should be selected. Forward, backward and stepwise selections are carried out to build three different models for each method. AIC (to find the lowest Akaike's Information Criteria statistic), SBC (to find the lowest Schwarz's Bayesian Criteria statistic), validation error (to find the highest log-likelihood) and validation misclassification (to find the lowest misclassification rate) has all been selected as assessment criteria to select variables. In the end, it turns out that for the first method, only the transformed variables of utilization, income and number of card should be included in the model (Table 5); and for the second method, WOE of all variables should be selected (Table 6).

	AIC			SBC			Validation Error			Validation Misclassification		
	Backward	Forward	Stepwise	Backward	Forward	Stepwise	Backward	Forward	Stepwise	Backward	Forward	Stepwise
Utilisation	√	√	√	√	√	√	√	√	√	√	√	√
Number of card	√	√	√	√	√	√	√	√	√	√	√	√
Income	√	√	√	√			√	√	√	√	√	√
Balance_required	√											
Credit limit												
Leverage												

Table 5 Variable selection for transformed variables

	AIC			SBC			Validation Error			Validation Misclassification		
	Backward	Forward	Stepwise	Backward	Forward	Stepwise	Backward	Forward	Stepwise	Backward	Forward	Stepwise
WOE_Utilisation	√	√	√	√	√	√	√	√	√	√	√	√
WOE_Number of card	√	√	√	√	√	√	√			√	√	√
WOE_Income	√	√	√	√	√	√	√			√	√	√
WOE_Balance_required	√	√	√	√	√	√	√			√	√	√
WOE_Credit limit	√	√	√	√	√	√	√			√	√	√
WOE_Leverage	√	√	√	√	√	√	√			√	√	√

Table 6 Variable selection for grouped variables

After deciding the subsets of the two models, the effectiveness of two models can be compared in validation set by looking at the lift chart and response rate (Figure 4).

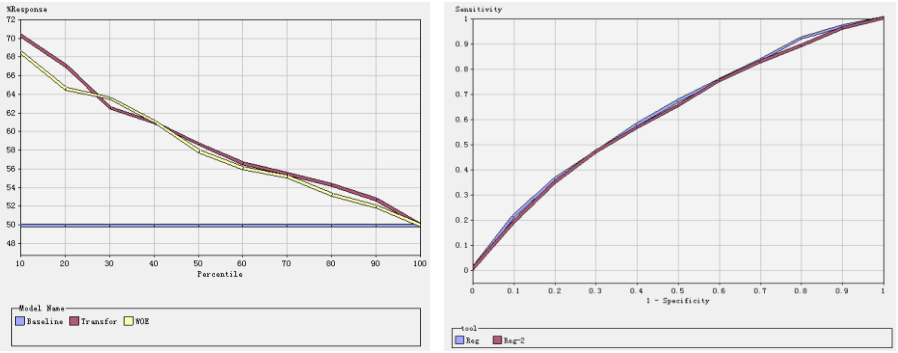

Figure 4 Lift chart (a) and ROC chart (b) of models using transformed variables and WOE

It is clear that the model using transformed variables has 2% higher response rate at the first decile than the model using WOE but very similar sensitivity and specificity. When taking the cost into consideration, the WOE model can generate profits of 1.19 whilst the model using transformed variables can only generate 0.74 profits. Therefore WOE model should be selected to be the best model for further comparison. The coefficients show that number of card, utilization and credit limit are the three most influential variables

Since the model is built based on the training sample with 1:1 odd ratio, when comparing result with other methods, the result should be adjusted to odd ratio (assume to be 1:4). To do this, classification rates of actual gamer and non-gamer in the original confusion matrix should be multiplied by 2/5 and 8/5 respectively. The confusion matrix in test set is shown in Table 7.

	Predicted Class using logistic regression		
Actual Class	D (i, j)	Non-Gamer	Gamer
	Non-Gamer	51%	29%
	Gamer	6%	14%

Table 7 Confusion Matrix of logistic regression

The profit generated by this model is 2.1. The sensitivity of the model is 70%, specificity is 63.75% and accuracy of the model is 65%.

DECISION TREE

A decision tree is 'a hierarchical group of relationships organized into a tree-like structure, starting with one variable called the root node' (Robert, John, Gary, 2009). The root node can be split into two or more branches. If the root node is continuous, the branches will represent classes of the node; if the root node is categorical, the branches will represent specific ranges along the scale of the node. At each split, a question is asked, which has an answer in terms of the classes or range of the variable being split. The questions are defined in terms of some impurity measure, reflecting how uniform the resulting cases must be in the splits. Each branch is split further using the classes or ranges of other variables. At each split, the node that is split is named parent node, and the nodes which split into are called the child nodes. This process continues until some stopping rule is satisfied or splitting is impossible.

To build decision tree, there is no need to transform the values because tree node will group continuous variables into bins automatically.

First, the splitting criteria should be decided from Chi-square, Entropy reduction and Gini reduction. It can be found that the response rate has the best value when using Gini reduction. The lift chart of cumulative response rate and ROC curve is shown below (Figure 5). About 65% of gamers are found in the first 10% decile using Gini reduction and both specificity and sensitivity of using Gini reduction (purple line) is higher than other two methods. After checking their confusion matrix, the profit generated by Gini reduction is also higher than the other two criteria.

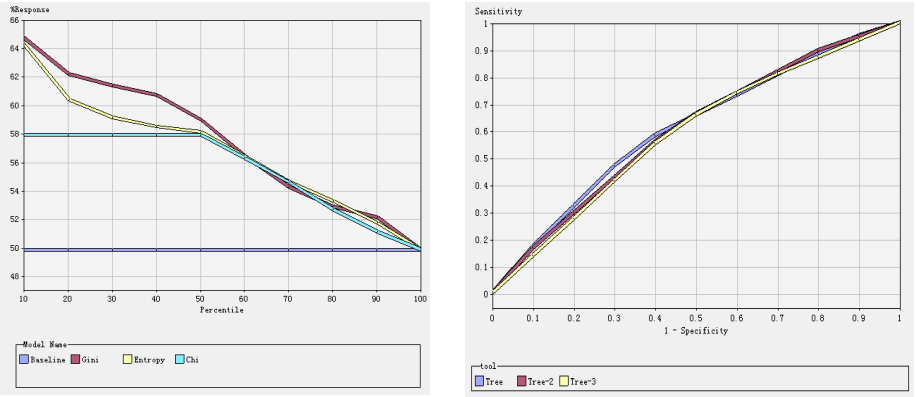

Figure 5 Lift chart (a) and ROC chart (b) to choose split criteria

Similarly, the model assessment measure is compared. It turns out that using the proportion misclassified measure could contribute the best cumulative response curve, ROC curve and the highest profit. Therefore, the proportion misclassified is chosen to be the assessment measure for the model.

Following this, efforts are made to adjust other pre-pruning settings in tree node. To avoid over fitting, the maximum number of branch is remained at 2 and the minimum observations and observations required for a split search are also remained. The maximum depth of tree is adjusted in order to achieve a better result. After some trials, it is found that performance has only little improvement in terms of cost and ROC curve when the depth increases, but the response rates get worse when the depth increase to 8 (Figure 6a). Therefore, there is no need to increase the depth of the tree. The depth will be set as 6 (default).

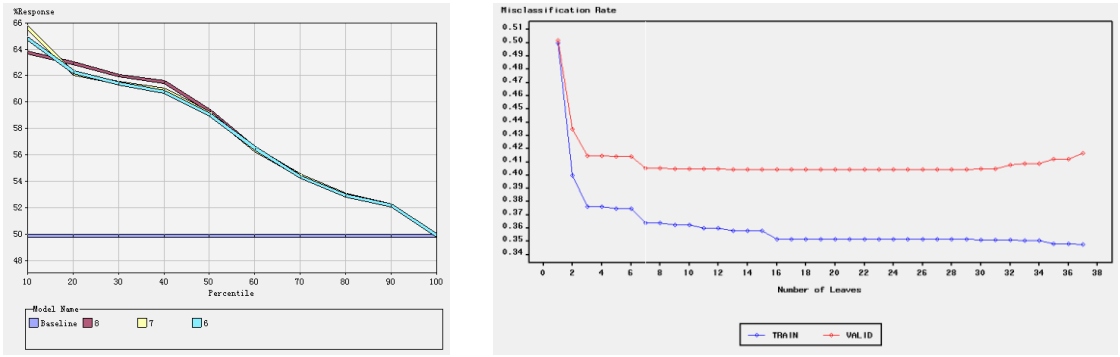

Figure 6 Lift chart to choose tree depth (a) and assessment plot to choose number of leaves (b)

The assessment plot (Figure 6b) displays the training and validation assessment values for several candidate sub trees. The misclassification rate has a big drop at leaf 7 and has only small decreases afterwards. Hence the sub tree with 7 leaves is selected. However, it is very clear that misclassification rate in validation set is much higher than training set, which indicates that the model performs very poorly.

As a result, the detail of the tree splitting is as follow:

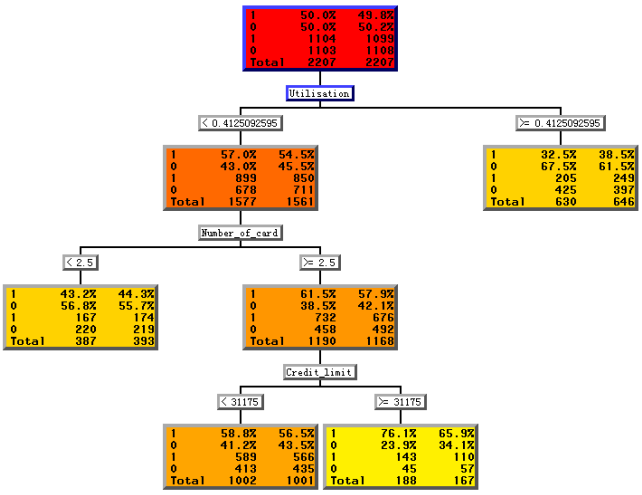

Figure 7 Tree splits

Same conclusion has come out as from logistic regression and variable exploration: utilization and number of card holding contains the most information about people being a gamer. Similar with logistic regression model, since this decision tree is also built based on the training set with 1:1 odd ratio, the confusion matrix should be adjusted as well. Based on the result from validation set, the confusion matrix after adjustment is as follow.

	Predicted Class using logistic regression		
Actual Class	D (i, j)	Non-Gamer	Gamer
	Non-Gamer	50.4%	29.6%
	Gamer	8.8%	11.2 %

Table 8 Confusion matrix of decision tree

The profit generated by this model will be 2. The sensitivity of the model is 56%, specificity is 63% and accuracy of the model is 62%.

COMPARISON

After achieving the four models, cost of them can be compared based on test set to choose a best final model. Four of the confusion matrixes have been listed below:

	Predicted class		
	D (i, j)	Non-Gamer	Gamer
Naïve Bayes Classifier			
Actual Class	Non-Gamer	62%	18%
	Gamer	12%	8%
Logistic Regression			
Actual Class	Non-Gamer	45%	34%
	Gamer	9%	11%
Decision Tree			
Actual Class	Non-Gamer	48%	32%
	Gamer	9%	11%

Table 9 Confusion matrix comparison

What to be noted is that both results of logistic regression and decision tree has been adjusted from odd ratio of 1:1 to odd ratio of 1:4. Consequently, the profit generated from using ever active at 3 is 3.7, from Naive Bayes Classifier (NBC) is 4.27, from logistic regression is 1.02, and from Decision tree is 1.53. This means that using NBC can generate more profit than other three models. This result is mainly driven by the high specificity of Naive Bayes Classifier, which is 78%, while the specificity of logistic regression and Decision Tree are both only less than 60%. Since the objective of building these models is to identify gamers hence reduce the loss of the bank, it is recommended to use Naive Bayes Classifier to find gamers.

CONCLUSION

The losses generated by each gamer is one tenth of the profit generated by a normal customer. In this report, three methods have been used to identify gamers, including Naive Bayes classifier, Logistic regression and decision tree. It has been found that variable 'utilization of cards' and 'number of card holdings' are the two most influential variables when identifying gamers. The lower utilization or the more cards held, the higher possibility of a person being a gamer. In logistic regression, weight of evidence of all six variables, including utilization of cards, number of cards, income, credit limit, leverage and transferred balance required, have been selected, and the profit can be generated is 1.02 in test set. In decision tree, only utilization, number of card and credit limit has been selected, and the profit can be generated is 1.53. The profit using ever active rate is 3.7, whilst the highest profit is generate by using Naive Bayes classifier, which is 4.27 unit of money. This is because the model built with Naive Bayes classifier has the highest specificity among all models. In the end, the simplest method is 'not so stupid after all' (David J. Hand and Kerning Yu, 2001).

The advantage of using Naive Bayes classifier is not only that it is more accurate, but also that banks can know the propensity of gamer at the start of launching a product instead of few months later. Therefore, if a product is attracting too many gamers, banks can stop launching the product immediately instead of waiting for another few months. Also, Naïve Bayes classification is easy to build and allow users to update the model easily. On the other hand, it is also easy to interpret, which is very important in industrial practice.

However, this method is only built based on internet channel and on the assumption that base rate will not change, since it will influence the cost matrix, hence the choice of the final model. Moreover, this model should only be used within the same country since behavior of people may vary among different countries. Finally, this model highly replies on the stable financial environment. For example, new functions like 'Transfer to bank' will change the balance level of customers, hence change the model completely.

The suggested model can generate 4.27 profits if it is used for underwriting and targeting. However, in practice, rejecting gamer-like customers will make the bank taking risk of having lower market share and losing potential profits. Another way to deal with them is to lower their credit limit instead of rejecting their application, so that the losses generated by them will be reduced. At the same time, if the bank can encourage them to purchase, such as by rewarding their purchase, these people may generate profit for the bank.

REFERENCES

Elzinga, K. G. and Mills, D. E, 1998, 'Switching Costs in the Wholesale Distribution of Cigarettes', Southern Economic Journal, 65(2), pp. 282-293

Ying Lei, Iho Chen and Echo Liang, 2007, 'Methods and systems for managing transaction card customer accounts', United states patent application publication

Andrew K.G. Tan, Steven T. Yen and Yiing Jia Loke, 2011, 'Credit card holders, convenience users and revolvers: a tobit model with binary selection and ordinal treatment', Journal of Applied Economics. 14(2): 225-255

Delener, Nejdet, and Herbert Katzenstein, 1994, 'Credit card possession and other payment systems: Use patterns among Asian and Hispanic consumers', International Journal of Bank Marketing 12: 13–25.

Kim, Haejong, and Sharon A. DeVaney, 2001, 'The determinants of outstanding balances among credit card revolvers', Financial Counseling and Planning 12: 67–78.

Min, Insik, and Jong-Ho Kim, 2003, 'Modeling credit card borrowing: A comparison of type I and type II Tobit approaches', Southern Economic Journal 70: 128–143.

Ausubel, Lawrence M. (1991), 'The failure of competition in the credit card market', American Economic Review 81: 50–81.

Canner, Glenn B., and Charles A. Luckett, 1992, 'Developments in the pricing of credit card services', Federal Reserve Bulletin 78: 652–666.

Chien, Yi-Wen, and Sharon A. DeVaney, 2001, 'The effects of credit attitude and socio economic factors on credit card and instalment debt', Journal of Consumer Affairs 35: 162–179.

Jiawei Han, Micheline Kamber, 2001, 'Data mining: Concepts and Techniques', Simon Fraser University, Morgan Kaufmann Publishers, pp.296

XindongWu , Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh , Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach , David J. Hand , Dan Steinberg, 2008, 'Top 10 algorithms in data mining', Springer, Knowl Inf Syst, 14:1–37

Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 2006, 'Introduction to Data mining', Addison-Wesley, Pearson International Edition. pp:231

Max Bramer, 2007, 'Principles of Data Mining', School of Computing University of Portsmouth , pp.29

Daniel T.Larose, 2005, 'Data mining methods and models', Central connecticut state university, John Wiley&Sons Inc, pp.215

Jiawei Han, Micheline Kamber, 2001, 'Data mining: Concepts and Techniques', Simon Fraser University, Morgan Kaufmann Publishers, pp.322

Daniel T.Larose, 2005, 'Data mining methods and models', Central connecticut state university, John Wiley&Sons Inc.pp.155

Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, J. Christopher Westland, 2011, 'Data mining for credit card fraud: A comparative study', Decision Support Systems 50: 602–613.

George Fernandez, 2002, 'Data mining using SAS applications', Chapman&Hall CRC,pp.163

Robert Nisbet, John Elder, Gary Miner, 2009, 'Handbook of statistical Analysis & data mining applications', Elsevier Inc,pp.241

Jiawei Han, Micheline Kamber, Jian Pei, 2001, 'Data mining: Concepts and Techniques;, Simon Fraser University, Morgan Kaufmann Publishers, Third Edition, pp.341

Kasteridis, Panagiotis P., Murat K. Munkin, and Steven T. Yen, 2010, 'A binary-ordered probit model of cigarette demand', Applied Economics 42: 413–426.

Clare Francis, 'What is a good credit score?' Money supermarket credit card,

Available at: http://www.moneysupermarket.com/credit-cards/what-is-a-good-credit-score.

David J. Hand and Kerning Yu, 2001, 'Idiot's Bayes-Not So Stupid After All?', Netherlands: International statistical review 69 (3): 385-398.

ACKNOWLEDGMENTS

Firstly, I would like to express my greatest thanks to my supervisor, Doctor Nicos Pavlidis. Thank him for leading me into the world of data mining and using his sophisticated knowledge guiding me with this project. I benefit so much from what he has taught me. Without his great support and sharp opinions, I could not finish my paper and presentation with confidence.

Secondly, I would like to thank my supervisor in Barclaycard, Sam Vaidya. He gave me great help during this whole project and taught me a lot. I really appreciate his prompt reply for my questions and his great patience. It would be impossible for me to come up all the ideas without him. I enjoyed working with him very much and would like to wish him all the best in the future.

My line manager in Barclaycard, Karishma Jaitly, has also given me great support for both my work and life in London. I would like to express my sincere thanks to her for commenting my work and leading me to better and better models.

I would like to thank David Worthington, Nicos kourentzes and Orville D'Silva for their effort of giving me this wonderful opportunity to work in this top-ranked bank.

Also thanks to Christopher Kirkbride for teaching me how to use SAS.

Besides, I would like to thank management science department of Lancaster University for giving me such a great studying experience this year and this great chance to apply my knowledge into practice.

I also want to thank all my friends, for supporting me and accompanying me during my touch time I experienced this summer.

Finally, I would like to owe my greatest gratitude to my beloved parents, for their great support, time engagement and selfless love.

CONTACT INFORMATION

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

Name: Yang Ge
Organization: Barclaycard
City, State ZIP: London
Email: efflor.gy@gmail.com
sascommunity.org: