의사결정나무를 이용한 온라인 자동차 보험 고객 이탈 예측과 전략적 시사점

Customer Churning Forecasting and Strategic Implication in Online Auto Insurance using Decision Tree Algorithms

임세헌 (Sehun Lim) 상지대학교 경영정보학과 전임강사, 교신저자†
허 연 (Yeon Hur) 중앙대학교 상경학부 교수

요 약

본 연구에서는 온라인 자동차보험 고객 이탈 예측에 있어 의사결정나무를 적용하였다. 우리는 본 연구에서 2003년과 2004년 사이에 온라인 자동차 보험을 계약한 고객의 데이터를 이용하여 의사결정나무를 이용해 고객이탈을 예측하였다. 우리는 C5.0 알고리즘에 기반을 둔 의사결정나무의 예측 결과에 대한 비교를 위해 다변량판별분석과 로짓분석을 이용하였다. 분석결과 의사결정나무 알고리즘은 다른 기법보다 예측성과가 매우 뛰어난 것으로 나타났다. 이러한 실증분석 결과는 온라인 자동차 보험에 있어서 마케팅전략 수립에 유용한 가이드라인을 제공해 줄 것이다.

키워드 : 의사결정나무, 온라인 자동차 보험, 고객이탈, 다변량판별분석, 로짓분석

Ⅰ. Introduction

With an expansion of the Internet users, recent auto insurance market in Korea witnesses major changes. Hur (2003) argues that the expansion of the online auto insurers in the market is indebted to the following facts. First, the online insurers suggest lower auto insurance premiums and provide customers other insurance information with lower acquisition cost than offline insurers. Second, the online environment allows customers to switch their insurers easily without complication. Third, insurance premium quote service provided by online aggregators, either independent agents or brokers, makes auto insurance customers to reduce switching cost and information search cost to almost nil. Therefore, the online market becomes more important to the auto insurers and requires them to have more proactive strategic responses on the development to survive in the market and further to maximize their corporate values.

According to the Sigma research (2000), there

† I am appreciative of the valuable suggestions provided by two anonymous reviewers and Prof. Kim Injai at Dongkuk University, who is an editorial board member. I owe thanks, as well, to Prof. Kim Kyoung-jae at Dongkuk University. He suggested valuable solutions in this research.

are several types of e-business models in insurance industry currently: online insurers who sell their own insurance products directly; aggregators (independent agents or brokers); product portals (comprehensive standard websites for financial and/or insurance products); point of sales portals (product marketing through various theme-based web pages); reverse auctions (auctions of insurance demand).

Among the above models, direct online insurer model and aggregators draw our attention to compare. Direct online insurer can minimize the acquisition costs so that it can lower insurance premium.[1] However, its major drawback is that it cannot provide premium quote service of other insurers. On the other hand, aggregators can provide premium quote service of other insurers because they are independent agents or brokers. If we consider a characteristic of the Internet users that they are price sensitive, then ability of premium quote service of aggregators is considered as a major competitive edge to survive in online environment.[2] Therefore, we can argue that maintaining a good customer relationship to achieve customer and brand loyalty is a major task of the online insurance.

It is very scarce, however, to find related researches on the customer churning in online auto insurance market. This article using data mining C5.0 algorithm, proposes a model that can predict customer churning in online auto insurance contracts. Also we suggest strategic implications of the results of our analyses. To compare the results of C5.0 model, we analyze the same data using logistic regression model (LRM), multivariate discriminant analysis (MDA) model as a bench marking. We expect that our results have various practical implications to auto insurance business in Korea.

This study is structured as follows. Section Ⅱ introduces the basic concepts and its application of C5.0. In section Ⅲ, prediction variables and research design and experiment are explained. Section Ⅳ compares the prediction results of C5.0 model with those of LRM and MDA. Also we interpret strategic implications of the results from the customer relation perspective. Finally, conclusions and limitations are presented.

Ⅱ. Research Background: Decision Tree and Their Application Research

Decision tree algorithm is an inductive learning method. It structures decision making rules using a tree figure to solve problems of classification and prediction. We have two types of algorithm that support the decision tree algorithm. One is artificial intelligence technology based algorithms such as ID3 (Interative Dichotomizer 3), C4.5, and C5.0. The other one is based on statistics, i.e., CART (Classification and Regression Tree) and CHAID (Chi-square Automatic Interaction Detection) (Berry and Linoff, 1997). As shown on <Table 1>, Decision tree techniques have some advantages and disadvantage (Sung, Chang, Lee, 1999; Reacock, 1998).

We have several data mining studies employing C5.0 or Decision Tree in the area of finance for example, customer churning prediction of a credit card firm, bankruptcy prediction, and stock price prediction.

[1] According to the Sigma research (2000), e-business can produce cost savings around 10% in claim settlement, 30% in policy administration, 30% in distribution, and 5% in claims payment. Overall, e-business can generate potential savings of 12% of total cost in personal lines (insurance) in U.S.A.

[2] In addition to that, aggregators have an advantage to gather customer database because of frequent users or transactions.

〈Table 1〉 Advantages and Disadvantages of DTA

Feature	Contents
Advantages	(1) able to generate understandable rules (2) able to perform in rule-oriented domains (3) easy of calculation at classification time (4) able to handle continuous and categorical variables (5) able to indicate best fields clearly
Disadvantages	(1) error-prone with too many classes (2) computationally expensive to train (3) trouble with nonrectangular regions

Source: Sung, Chang, Lee (1999)

Braun and Chandler (1987) suggest a decision making support model using ID3 to predict stock price fluctuation for the stock market analysts. Cronen *et al.* (1991) adopt a recursive partitioning method that is an expansion of ID3 algorithm to predict firm bankruptcy and problems of classification of mortgage. In the study, they prove superior predictability of the model. Lee, Jung and Shin (2003) use C5.0 model to analyze customer churning classification in credit card market. They prove that C5.0 outperforms predictability of other methods such as LRM and ANN. Johnson *et al.* (2002) suggest decision tree-based symbolic rule induction system for categorizing text document automatically and verify its usefulness by far in solving practical problems. Monkol *et al.* (2003) also employ decision tree-based algorithm (C4.5 and ID3) to predict fraud in online business transaction under e-business environment and prove its prediction accuracy. On the other hand, Lee and Lee (2003) improves prediction performance by combining ANN method and decision tree algorithm (C4.5) in the customer churning prediction of mobile telecommunication service users. This article adopts a decision tree algorithm (C5.0) to predict customer churning in online auto insurance environment.

Ⅲ. Research Design and Experiments

3.1 Data Description and Measurement of Variables

This study employs total number of 13,200 sample data of auto insurance contracts sold between 2003 and 2004 from an online aggregator (independent agent). This company provides consumers auto insurance premium quote service and related information regarding auto insurance as well as other general insurance products. If a customer provides basic underwriting information and select insurance coverage, then he (she) can compare his (her) auto insurance premium quoted from all insurance companies at the same time. Then he (she) can select an insurer and purchase the coverage chosen through the Internet immediately.

Out of 25 variables which we believe delivers certain meaning to our analysis, this study employs a predefined experiment that can minimize the number of variables to be considered, thereby we can have a parsimonious model and easily interpret the results.

(1) Age: Age of insured

(2) Zip Code[3]: residence zip code (territory)

[3] The zip code represents the territory where the automobile is principally used and garaged in many countries. Depend on the loss experience of each insurer in the region, individual liability premium is determined. As a result, an individual premium depends on the loss experience of each insurer in the region. It also affects customer churning ratio.

〈Table 2〉 Input Variable Selection

Input Variable	T-Test		Chi-Squre		Result
	t-value	p-value	Chi-Square value	p-value	
Date	-0.836	0.403	•	•	Not Select
Gender	•	•	0.088	0.767	Not Select
Driver's Age	6.197	0.000	•	•	Select
Zip Code	•	•	23.897	0.001	Select
Type of Car	•	•	27.592	0.000	Select
Year of Car	1.625	0.104	•	•	Not Select
Price of Car	2.278	0.000	•	•	Select
Credit/Debit	•	•	88.190	0.000	Select
Surcharge	•	•	199.568	0.000	Select
Driver Endorsement	•	•	639.477	0.000	Select
Age Endorsement	•	•	243.407	0.000	Select
Number of Airbag	•	•	5.159	0.076	Select
Comprehensive B.I	-1.134	0.257	•	•	Not Select
Property Liability	2,335	0.020	•	•	Select
Medical Expenses	4.337	0.000	•	•	Select
Uninsured Motorist Cover	-0.528	0.597	•	•	Not Select
Deductible(auto)	•	•	33.113	0.000	Select
Last Premium Quote(t)	-3.224	0.001	•	•	Select
Old Insurer's Quote(t+1)	-3.862	0.000	•	•	Select
Liab. Premium(B.I)	-1.183	0.237	•	•	Not Select
Lowest Premium Quoted	0.621	0.534	•	•	Not Select
Highest Premium Quoted	0.616	0.538	•	•	Not Select
Type of Coverage	•	•	22.806	0.000	Select

(3) Car Type: Size of car: smaller (0), small (1), medium (2), large (3), SUV and VAN (4), small truck and others (5)

(4) Price of Car: Car price in the auto market

(5) Credit/Debit: Credit/Debit ratio of the insured: Base credit/debit ratio is 100%; 10% of credit is given to drivers who have no accident record each year up to 6th year (50% of credit); then 5% credit per year of no loss record is given for 2 more years thereafter that is, 45% and 40% credit ratio are applied for 7th and 8th year respectively (maximum up to 60% of discount). For example, 100, 90, 80, 70, 60, 50, 45, 40 credit ratio are applied according to their loss history.

(6) Surcharge: If the driver has loss record in a specific year, the driver is surcharged based on loss type and amount. 0 represents no surcharge, otherwise 1. Generally minor loss (less than 400,000 Won) is not subject to surcharge (varied by insurers).

(7) Driver Endorsement: Driver endorsement represents qualification or category of drivers who drive the car to have the insurance coverage; 1 denotes Family (other than brothers and sisters) drivers only, 2 refers Basic (no limitation), 3 represents Spouse only, 4 indicates the insured (self) only, 5 represents 1+brothers.

(8) Age Endorsement: Age endorsement refers the age limitation of the driver; 1 represents drivers' age 21 or more years old, 2 denotes age 26 or more, 3 represents Basic coverage (no limitation), 4 represents age 24 or more, 5 refers

age 25 or more.

(9) # of Airbag: 10% of credit per airbag is given to the medical expense coverage of the insured (maximum 20%).

(10) Property (3^{rd}) Liability: Property liability denotes the coverage for property damage to the third party.

(11) Medical Expenses: Medical expenses represents payment for bodily injury to the insured person who is at fault.

(12) Deductible (auto): Deductible is applied to the coverage for damage to your auto and represents amount of loss that the insured must pay before the company will pay, up to the limits of the policy. The higher the absolute amout, the lower the premium. 1 refers no coverage, 0 represents zero deductible, 5 denotes 50,000 Won, 10 represents 100,000 Won, 20 refers 200,000 Won, 30 denotes 300,000 Won, and 50 represents 500,000 Won of deductible.

(13) Last Prem (t): Premium amount that the insured paid last year.

(14) Old Insurer's Quote (t+1): Premium quoted this year from the last year's insurer.

(15) Type of Coverage: Type of coverage represents the usage of the car. 1 represents personal passenger car; 2 represents Plus coverage with larger coverage than regular one; 3 represents the coverage for company owned cars, and 5 represents taxi for hire.

3.2 Data Description Information

This study considers several indication variables. First, the output or dependent variable is defined as a binary variable. If a customer switches its insurance company next year, then the dependent variable has a value of "1", otherwise "0" (no switch). Independent variables are composed of two types of variables: indication variables and ratio (numeric) variables.

To select meaningful variables, this study did Chi-Square test for indication variable data and t-test for numeric variable data. Based on the two tests, we select 15 meaningful variables to be considered in our model. The summary statistics are presented in <Table 3>.

〈Table 3〉 Summary Statistics

Feature name	Mean	Std. Deviation	Minimum	Maximum
Age of Driver	39.14621	9.11668	19	95
Zip Code	3.637727	1.891552	1	7
Car Type	2.112273	1.404095	0	5
Price of Car	607.2974	550.3214	1	9600
Credit/Debit	66.69394	24.31643	40	200
Surcharge	2.347197	6.212417	0	50
Driver Endorsement	2.920076	1.080609	1	5
Age Endorsement	1.981894	0.13334	1	2
Number of Airbag	0.434394	0.656068	0	2
Property Liability	3258.258	2628.364	0	100000
Medical Expense	5396.856	5945.052	0	20000
Deductible(auto)	4.092197	8.421139	-1	50
Last Premium(t)	145764.1	245559.1	0	2005720
Old Insurer's Quote(t+1)	442528.5	245609.1	0	2504970
Type of Coverage	1.216439	0.481858	1	3

This study uses total number of 13,200 sample data that are divided into two groups based on company switching results. One group of data is consisted of 6,600 observations that switched their insurance company when they renewed their insurance policy next year. The other group is also consisted of the same number of observations who renewed their contracts with the same insurers. Also the sample data are grouped into training data set and holdout data set to test predictability of the model. The composition of the training data and holdout data are 80:20 or 10,560 and 2,640 observations respectively in LRM, MDA and C5.0.

Ⅳ. Results and Discussion

4.1 Experimental Results

Using the sample of online auto insurance customers data we test how decision tree-based model (C5.0) works on the prediction of customer churning. We compare the result of C5.0 with those of LRM and MDA model.

4.1.1 C5.0

First, in the decision tree analysis, we use the Clementine 8.1. We control the pruning level at 75% of C5.0 in our experiment. In addition, we select minimum record 2 of a branch on decision tree. The predict performances using C5.0 shows 67.39% in the test data and 68.71% in the holdout data. The results of customer churning analysis show 11 depth of decision tree, 58 rules for the data of staying with the same auto insurer (non-switching customer), and 65 rules for the data of switching insureds (who switch their insurers)next policy year.

4.1.2 LRM

Second, in the LRM, we analyze the categorical variables with the cutoff point at 0.5. We use SPSS 11.1 software. Our results show 65.3% of accuracy with the training data and 60.0% with the holdout data.

4.1.3 MDA

Third, we analyze the data using MDA after standardized those categorical variables and serial variables with Z-score. We use SPSS 11.1 software. The results indicate 59.7% of accuracy with the training data and 59.4% with the holdout data.

〈Table 4〉 Prediction Accuracy of LRM, MDA and C5.0

Data set	LRM	MDA	C5.0
Training data	65.3	59.7	67.39
Holdout data	60.0	59.4	68.71

In sum, C5.0 outperforms by far to compare with LRM and MDA. Based on the result, C5.0 algorithms suggests a way of setting marketing strategy and of developing online auto insurance business.

4.2 Discussion

As we discussed C5.0 shows superior predictability in the customer churning analysis. In an application of decision tree rules, analyzing switched customers who changed the insurers to others is more important than that of unswitched customers who show some loyalty to the insurers.

The reason is that analyzing those switched insureds provides insurance firms important strategic implications to hold potentially switching customers. Our analysis using C5.0 results in 65 customer churning rules. <Table 5> shows exemplary rules

〈Table 5〉 Strategy Planning for Customer Churning using C5.0 Algorithms

Rule 1 if AGE_ENDO = 2 and DRIVER_E = 2 and AIRBAG = 0 and SURCHARG <= 5 and ZIP_CODE = 1 and MEDIC._E <= 1500 then 1	Rule 4 if AGE_ENDO = 2 and DRIVER_E = 2 and AIRBAG = 0 and SURCHARG <= 5 and ZIP_CODE = 7 then 1	Rule 7 if AGE_ENDO = 2 and DRIVER_E = 3 and PROPERTY <= 3000 and DEDUCT = -1 and ZIP_CODE = 1 and LAST_PRE <= 232110 then 1
Rule 2 if AGE_ENDO = 2 and DRIVER_E = 2 and AIRBAG = 0 and SURCHARG <= 5 and ZIP_CODE = 1 and MEDIC._E > 1500 and CREDIT <= 65 then 1	Rule 5 if AGE_ENDO = 2 and DRIVER_E = 2 and AIRBAG = 1 and SURCHARG <= 1 and DEDUCT in [1] then 1	Rule 8 if AGE_ENDO = 2 and DRIVER_E = 3 and PROPERTY <= 3000 and DEDUCT = -1 and ZIP_CODE = 2 and SURCHARG <= 1 then 1
Rule 3 if AGE_ENDO = 2 and DRIVER_E = 2 and AIRBAG = 0 and SURCHARG <= 5 and ZIP_CODE = 4 and AGE > 44 then 1	Rule 6 if AGE_ENDO = 2 and DRIVER_E = 2 and AIRBAG = 1 and SURCHARG > 1 then 1	Rule 9 if AGE_ENDO = 2 and DRIVER_E = 3 and PROPERTY <= 3000 and DEDUCT = -1 and ZIP_CODE = 4 and PROPERTY <= 2000 and SURCHARG <= 5 and CREDIT > 85 then 1

(rule 1 and 2) out of 15 rules that determine customer churning. More detailed explanation on the rule 1 and rule 2 is as follows.

In rule 1, the customers who have the following facts are likely to switch their insurance companies next year: who choose (1) driver's age endorsement = 2, which denotes driver's age should be more than 21 years old, (2) driver endorsement = 2, any driver can drive the car. There is no specification about driver. Driver endorsement defines a relationship of the driver to the policyholder, for example, family or spouse or myself, (3) the car has no airbag (airbag = 0), (4) resident in zip code 1 (1 = Seoul), (5) medical expenses coverage less than 1.5 million Won for the insured.

In rule 2, if the customer choose (1) driver's age endorsement = 2, (2) driver endorsement = 2, (3) the car has no airbag (airbag = 0), (4) who experienced a loss and surcharged less than 5%, (5) resident in zip code 1 (= Seoul), and finally (6) if the credit/debit ratio is less than 65% or more than 35% discount in the premium payment, the insured is likely to switch insurance company next year.

Based on the results of the customer churning analysis, the insurance company can set its marketing strategy effectively to minimize customer churning and can improve its operation performance.

V. Conclusion

This article analyzes predictability of customer churning in online auto insurance case using C5.0 algorithm. The results suggest strategic implications for marketing strategy in practice. This study argues that the prediction performance of C5.0 is better than those of other models such as LRM and MDA models.

There are many factors and reasons that determine switching their insurers. If insurance firms are able to analyze the factors and reasons of customer churning and to hold the potential switching customers with the company, they can enhance their management performance.

References

Berry, M. J. A. and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*, Wiley Computer Publishing, 1997.

Braun, H. I. and J. Chandler, "Predicting stock market behavior through Rule Induction: An Application of the Learning-From-Example Approach", *Decision Science*, Vol.18, 1987, pp. 415-429.

Breiman, L., J. J. Friedman, R. A. Olshen, and C. J. Stine, *Classification and Regression Trees*, Beltmont, Wadsworth, 1984.

Cronan, T., L. Glorfeld, and L. Perry, "Production System Development for Expert Systems using a Recursive partitioning Induction Approach: An Application to Mortgage, Commercial and Consumer Lending", *Decision Science*, Vol. 22, 1991, pp. 812-845.

Johnon, D. E., F. J. Oles, T. Zhang, and T. Goetz, "A Decision Tree Based Symbolic Rule Induction System for Text Categorization", *IBM System Journal*, Vol.41, No.3, 2002, pp. 428-437.

Lee, G. C., N. H. Jung, and K. S. Shin, "Customer Churning Analysis by Using Data Mining in Credit Card Market", *Korea Journal of Intelligent Information System*, Vol.8, No.2, 2003, pp. 15-35.

Lee K.-K. and H.-C. Lee, "A Study on the Combined Decision Tree (C4.5) and Neural Network Algorithms for Classification of Mobile Telecommunication Customer", *Korea Journal of Intelligence Information System*, Vol.9, No.1, 2003, pp. 139-155.

Hair *et al.*, *Multivariate Data Analysis*, Macmillan publishing co. New York, 1992.

Hur, Y., "Decision Factors of Insurance Company Selection in Online Auto Insurance", *Risk Management Journa*l, Vol.14, No.1, 2003, pp. 23-45.

Monkol, L., Benjamin Anandarajah, Narciso Cerpa and Rodger Jamieson, "Data Mining Prototype for Detecting E-Commerce Fraud", *The 9th European Conference on Information Systems*, Vol.9, 2002, pp. 160-165.

Quinlan, J. R. and J. Quinlan, *C4.5 Programming for Machine Learning*, Morgan Kaufman Publishers, 1997.

Reacock, R. P., "Data Mining in Marketing Part 1", *Marketing Management*, Vol.6, No.4, 1998, pp. 17-28.

Sigma Research (www.swissre.com), *The impact of e-business on the insurance industry: Pressure to adapt-chance to reinvent,* Sigma, Swissre, 2000.

Sung, T. K., N. S. Chang, and G. H. Lee, "Dynamics of Modeling in Data Mining: Interpretive Approach to Bankrupcy Prediction", *Journal of Management Information Systems*, Vol.16, No. 1, 1999, pp. 63-86.

Information Systems Review
Volume 8 Number 3
December 2006

Customer Churning Forecasting and Strategic Implication in Online Auto Insurance using Decision Tree Algorithms

Sehun Lim* · Yeon Hur**

Abstract

This article adopts a decision tree algorithm (C5.0) to predict customer churning in online auto insurance environment. Using a sample of on-line auto insurance customers contracts sold between 2003 and 2004, we test how decision tree-based model (C5.0) works on the prediction of customer churning. We compare the result of C5.0 with those of logistic regression model (LRM), multivariate discriminant analysis (MDA) model. The result shows C5.0 outperforms other models in the predictability. Based on the result, this study suggests a way of setting marketing strategy and of developing online auto insurance business.

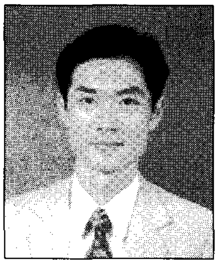
Keywords: Decision Tree (DT), Online Auto Insurance, Customer Churning, Logistic Regression Model (LRM), Multivariate Discriminant Analysis (MDA)

* Dept. of Management Information Systems, Sangji University
** Dept. of Business Administration, Chung-Ang University

◐ 저 자 소 개 ◑

Lim Se Hun (slimit@sangji.ac.kr)

He received his Ph.D. degrees in MS/MIS at the Chung-Ang University. He was a post-doctoral researcher of the department of information system and decision sciences at the University of North Texas. He is currently the full-time lecturer of the department of MIS at the Sangji University. His research interests include an application of machine learning algorithms in business area, data-mining, computational intelligence and e-business (ERP, CRM, SCM and RFID) etc. Dr. Lim has published in many international and domestic journals.



Hur Yeon (yeonhur@cau.ac.kr)

He earned his MBA from the College of Insurance (N.Y) and Ph.D. in Risk Management & Insurance from Temple University. He is a professor of the department of business in Chung-Ang University. His research area includes property insurance overall, insurance consumer behavior, corporate insurance, risk management, insurance company strategy, and on-line insurance issues. He has many papers published in many international & domestic journals. He is actively involved in insurance industry works as well.

논문접수일 : 2005년 07월 21일　　게재확정일 : 2006년 10월 20일
1차 수정일 : 2006년 01월 13일