Understanding Characteristics of Caravan Insurance Policy Buyer

May 10, 2007

Group 5

Chih-Hau Huang

Masami Mabuchi

Muthita Songchitruksa

Nopakoon Visitrattakul

Executive Summary

This report is intended to understand characteristics of a caravan insurance policy buyer. The dataset we used consists of 9,822 customer records and includes socio-demographic data of the area where a customer lives and product ownership data of the customer. Our aim of this profiling analysis is to acquire managerial insights to create competitive strategies useful for making business decision.

In our analysis, we found that caravan insurance policy buyers show several characteristics which are consistent with our common sense. Intuitively, a caravan policy buyer might own a caravan and a caravan owner might own a car to pull his/her caravan. Our research clearly supports this intuitive profiling. Furthermore, one might suppose that a caravan policy buyer should be rich, leisured or risk averse. Our research also shows correlation between the ownership of caravan policies and indicators of wealth or risk averseness. Overall, our model describes a typical caravan insurance policy buyer as a rich, risk averse person who lives in wealthy area.

Based on the findings from our research, we suggest the following actionable plans to help the company gain deeper understanding of customer characteristics and develop marketing strategies that align with the findings.

Area-Focused Marketing: Marketers could focus on marketing strategies that are most suitable and most likely to induce customers in each specific area. They might place marketing campaigns in areas with high proportion of highly educated people and high proportion of high average incomers.

Advertising Campaign: The insurance company could seek opportunities of cross-selling by knowing who are potential customers (risk-averse and rich).

Bundled Products: The insurance company could offer potential customers a bundle package of car insurance and caravan insurance.

Joint Marketing: Marketers could hold a joint marketing event with caravan makers to target potential customers and arouse their awareness of risks.

Technical Summary

Our analysis is aimed at explaining characteristics of people who are likely to or not to buy caravan policy, based on sampled product usage data, such as contribution and number of insurance policies, and socio-demographical data, such as average household size and average income. Our original dataset, which was originally used in a data mining contest[1], consists of 86 variables[2] and 9,822 customer records, of which are originally 5,822 training data records and 4,000 validation data records. The response variable is CARAVAN_POLICY whose value is either Buyer or Non-buyer of a caravan policy. It is also important to note that the socio-demographic data were derived from zip codes. All customers living in the same zip code are assumed to share the same socio-demographic attributes.

Preprocessing

We preprocessed our dataset in order to understand characteristics of caravan policy buyers clearly. We oversampled the success class we are interested in to train our models since the success class, buyer of a caravan policy, is rare (348/5,822 = 5.98%). We applied the models to the original (nonoversampled) validation set to avoid a biased estimate of model performance. See Exhibit 1 for the framework we used to preprocess our dataset.

We explored the data to see the validity of each variable as a useful predictor. We mainly used bar charts since almost all of the variables in our dataset are categorical. See Exhibit 2 for examples of data visualization. Based on our data analysis, we eliminated 78 variables that do not significantly distinguish people who are likely to buy caravan policy from those who are unlikely to. We also dropped variables which, by its nature, seem to be correlated and kept one in each group of correlated variables in our model; for example, M_RENTED_HOUSE (% of house-renters in a specific zip code area) and M_HOME_OWNERS (% of home owners) which are two sides of a coin.

We transformed some remaining variables to make our analysis more practical based on domain knowledge and insight from data visualization. For example, we created a binary variable, PRIV_3RD, stating whether or not a person buy at least one private third party insurance while we dropped CON_PRIV_3RD (Contribution to private third party insurance) even though the proportion of success class in the records which have higher values of CON_PRIV_3RD tends to be higher. The underlying reason for these transformations is explanatory simplicity. The model with binary variables is better than that with numerical variables because it is easy to interpret as long as the performance of each model does not show significant difference. In this way, we created three derived variables: FIRE_INS (0 indicates that a person does not have any fire policies; 1 otherwise), PRIV_3RD (0 indicates that a person does not have private third party insurance, 1 otherwise), and CAR_INS (0 indicates that a person does not have any car policies; 1 otherwise).

We also created another two dummy variables since some characteristics of zip code areas distinguish well people who buy caravan insurance from those who do not: M_MAIN=DRV_GRW (1 indicates a person whose customer main type is “driven growers”; 0 otherwise), and M_SUB=MID_CLS (1 indicates a person whose customer subtype is “middle-class families”; 0 otherwise).

Analysis

We constructed models using the two algorithms, logistics regression and classification tree to analyze our data. We didn't use discriminant analysis because almost all of our variables are

[1] The CoIL Challenge 2000, http://www.liacs.nl/~putten/library/cc2000/

[2] We revised names of variables used in the original contest to make them more understandable in this report

categorical and some of them are dummies which violate the assumption of discriminant analysis. See Exhibit 4 for the outputs from models.

In our logistics regression analysis, we eliminated useless or insignificant variables based on our domain knowledge and p-values for each variable. After all, our final model had four explanatory variables;

- M_EDU_HIGH (binned proportion of people who have high education in a specific area)
- M_AVG_INCOME (binned level of average income of a specific area)
- PRIV_3RD_INS
- CAR_INS

Of these four variables, CAR_INS had the lowest p-value, thus its significance contributed the most to the model. Based on our analysis, we found that all these four variables were attributable to two main customer characteristics, wealth and risk aversion. In our classification tree analysis, CAR_INS, M_EDU_HIGH, and M_AVG_INCOME played an important role in explaining characteristics of caravan insurance policy buyer. Although these two models yielded some similar results; that is, some variables in both models had explanatory power, the logistics regression model seemed to perform better than the classification tree. In the logistics regression, the percentage errors of Buyer (success class) in training and validation sets are (28.74% and 34.87% respectively) lower than those of the classification tree (37.36% and 39.50% respectively). We thus selected logistics regression model to explain behavior of caravan insurance policy buyers.

The implications drawn from the logistics regression model are as follows:

1. We found that caravan insurance buyers are likely to live in wealthy area. The underlying logic of this finding might be as follows: Based on our common sense, caravan owners would be rich and wealthy people would live in wealthy area.

2. Residents living in area which has high proportion of highly educated people are more likely to be high incomers, thus they tend to be a caravan policy insurance buyers.

3. A caravan owner is likely to own a car. In addition, private third-party insurance policy is required for car owners. The ownership of car insurance and private third-party insurance is a good indicator of car ownership.

4. If a person is risk averse, he/she would buy insurances. The ownership of car insurance and private third-party insurance is a good indicator of risk averseness. A risk-averse caravan owner is likely to buy a caravan policy.

Recommendations

The implications of the four explanatory variables show that wealth and risk aversion are two critical attributes of a caravan buyer. These characteristics give us some hints for further actions:

Area-Focused Marketing: Based on socio-demographic predictor such as M_EDU_HIGH and M_AVG_INCOME, marketers could place marketing campaigns in areas with high proportion of highly educated people and high proportion of high average incomers.

Advertising Campaign: Based on customers' past transactions, the insurance company could seek opportunities of cross-selling by knowing who are likely to be potential customers.

Bundled Products: The company could bundle a car insurance and a caravan insurance to target the potential customers.

Joint Marketing: Marketers could join caravan makers in launching promotion campaigns to focus target potential customers and to arouse their awareness of risks.

Exhibit 1: Framework for data pre-processing

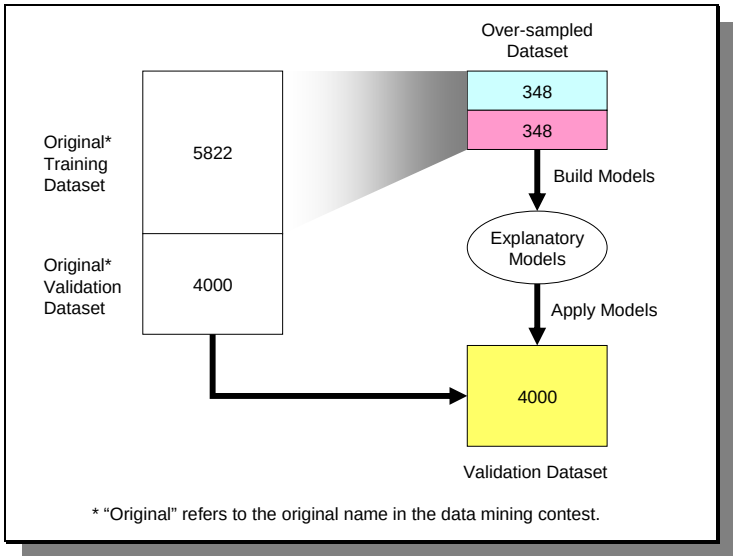
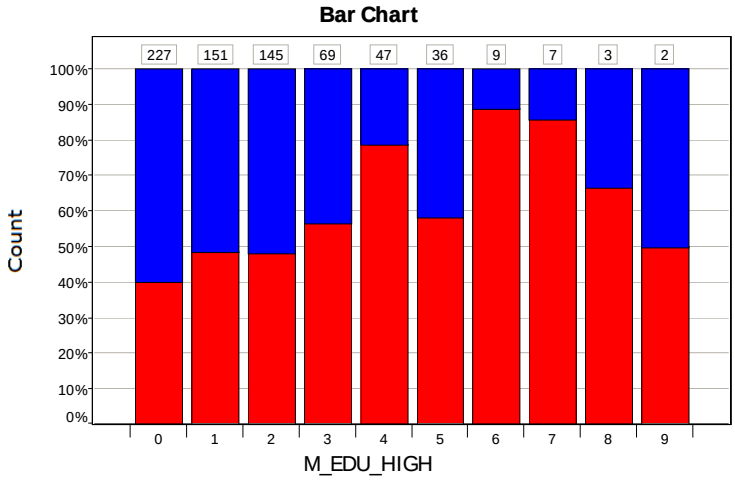


Exhibit 2: Bar Charts for selected variables



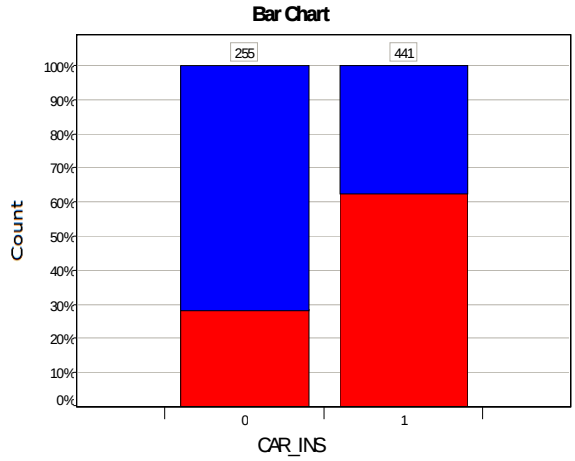


Exhibit 3: Logistics Regression Model

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.05754209	0.31352952	0	*
M_EDU_HIGH	0.17362288	0.05885206	0.00317612	1.1896069
M_AVG_INCOME	0.17481972	0.07740599	0.02391586	1.19103146
PRIV_3RD_INS	0.4663696	0.17033134	0.00618115	1.59419608
CAR_INS	1.32716846	0.18058152	0	3.77035236

Residual df	691
Residual Dev.	847.1951904
% Success in training data	50
# Iterations used	8
Multiple R-squared	0.12195091

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Buyer	Non-Buyer
Buyer	248	100
Non-Buyer	131	217

Error Report			
Class	# Cases	# Errors	% Error
Buyer	348	100	28.74
Non-Buyer	348	131	37.64
Overall	696	231	33.19

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Buyer	Non-Buyer
Buyer	155	83
Non-Buyer	1455	2307

Error Report			
Class	# Cases	# Errors	% Error
Buyer	238	83	34.87
Non-Buyer	3762	1455	38.68
Overall	4000	1538	38.45

Exhibit 4: Classification Tree

Training Data scoring - Summary Report (Using Full Tree)

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Buyer	Non-Buyer
Buyer	218	130
Non-Buyer	96	252

Error Report			
Class	# Cases	# Errors	% Error
Buyer	348	130	37.36
Non-Buyer	348	96	27.59
Overall	696	226	32.47

Validation Data scoring - Summary Report (Using Best Pruned Tree)

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Buyer	Non-Buyer
Buyer	144	94
Non-Buyer	1191	2571

Error Report			
Class	# Cases	# Errors	% Error
Buyer	238	94	39.50
Non-Buyer	3762	1191	31.66
Overall	4000	1285	32.13