The 2022 International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/6/2022165

What Cause RV Insurance? ——A Case Study on TIC Company

Xinzhuo Xu[1,a,*]

[1]*Big Data and Accounting College, Chongqing College of Finance and Economics, Chongqing, 402160, China*
a. xxuu0060@163.com
**corresponding author*

Abstract: With the vigorous development of digital technology, more and more enterprises realize the importance of digital transformation. Data has also become one of the enterprise assets. Through modern information technology, enterprises have improved their ability to collect and integrate data. Then the comprehensive data information is analyzed to contribute to business forecasting and enterprise management. This paper takes the customer data from the TIC insurance company as an example. It uses statistical analysis methods to mine the data, such as descriptive statistics, the Pearson chi-square test, the nonparametric test, K-means clustering, and binary logistic regression prediction. Analyze and forecast customers who may purchase RV insurance.

Keywords: data mining, data analysis, K-means clustering, binary logical regression

1. Introduction

Customer data is essential for enterprise operations. More and more enterprises are beginning to pay attention to big data related to customers. Prepare for improving the company's profitability and occupying more market shares in the next quarter [1]. With the development of science and technology, more and more customer data are stored on computers in databases. The scale of these data is vast and cannot be manually analyzed by the analyst team. Therefore, high-tech data mining technology can help enterprises find desired information [2]. This paper uses insurance company data sets (TIC) to analyze, mine, and predict customers interested in RV insurance.

This paper analyzes and studies the following aspects: In the second part, it analyzes why customers who buy family insurance are more likely to buy RV insurance than customers who buy private insurance. Through variable cross table and Pearson chi-square test, it is found that customers who buy family accident insurance are more likely to buy TIC's RV insurance, and this phenomenon is analyzed. The third part analyzes the impact of the customer's education level on the purchase of RV insurance. The nonparametric test found that the higher the education level, the greater the probability of purchasing RV insurance.

In the fourth part, K-means clustering analysis is carried out for the first variable "MOSTYPE" in the data dictionary. The 41 sub types of the variable are divided into four categories, and the same family characteristics of each category are found. Then, binary logistic regression is conducted for all variables, and it is found that these variables are significant and influential in predicting the purchase of RV insurance.

DOI: 10.54254/2754-1169/6/2022165

Finally, the conclusion summarizes the reasons customers buy RV insurance and puts forward some suggestions for future research.

The data set in this paper belongs to Sentient Machine Research, a Dutch data mining company, and has been used in the CoIL Challenge 2000 data mining competition [3]. In this paper, the variable "CARAVAN: number of mobile home policies" in the dataset is set as the target variable and replaced by RV insurance in the following text.

2. Literature Review

Literature [1] analyzes how much big data can benefit telecom companies. It is mentioned that big data is expected to promote the growth of the entire telecom value chain and improve efficiency and profitability. With the rapid development of the digital era. Not only the telecommunications industry but also all industries are facing digital transformation.

Literature [2] analyzes the essential characteristics of using data mining technology to detect bank direct sales data and collecting customer information is necessary to formulate marketing strategies. For example, in this paper, TIC Insurance Company makes predictions for customers who buy RV insurance by collecting customer data. These customer data are usually stored on computers and are too large to be analyzed manually by the manual analysis team. Therefore, advanced data mining technology can analyze data more conveniently and effectively.

Literature [4] introduces descriptive statistics, which lays a foundation for the descriptive statistics of variables in my paper.

Literature [5] introduced the relevant content of the Pearson chi-square test to help me carry out the Pearson chi-square test of related variables and determine the impact of variables on assumptions.

Literature [6] introduces the knowledge about degrees of freedom and P values in the normal distribution, which helps me to judge whether variables are normally distributed.

Literature [7] analyzed RV to travel in the British leisure market, most of which is family travel. It provides some ideas for this paper to analyze the impact of purchasing family insurance on RV insurance.

Literature [8] describes the plans of most Finns to purchase financial products and analyzes why most users do not like to purchase children's private medical insurance. This article introduces that most customers are loyal to the company they initially chose and preferred to buy all insurance services from the same company. From here to there, TIC Insurance Company can also refer to this point of view and sell RV insurance together with some popular family insurance. Reduce the price of portfolio insurance and increase the overall sales volume.

Literature [9] analyzed the normality test of statistical data, providing me with methods and ideas for the Kolmogorov-Smirnov test of continuous variables.

Literature [10] conclusion is that higher education and more people want to reduce risk, so they think insurance is a product of necessity. This conclusion is helpful for this paper to analyze the impact of education on the purchase of RV insurance.

Literature [11] It provides a new idea for customer classification. The enterprises in this article collect and classify consumer information through blogs. So this paper also attempts to cluster the subclasses and types of TIC insurance company customers by K-means.

The literature [12] introduces binary logic regression prediction, which provides a method for my prediction research.

DOI: 10.54254/2754-1169/6/2022165

3. Variable Cross Table & Pearson Chi-Square Test

Set "CARAVAN: number of mobile home policies" in 86 attributes as a grouping variable. Putting two assumptions: H0: variable does not affect purchasing TIC RV insurance; HA: Variables impact the purchase of TIC RV insurance.

The discrete variable "AGEZONG: Number of family accidents insurance policies" is described and statistically analyzed.

Table 1: AGEZONG and CARAVAN Cross table.

APERSONG*CARAVAN Crosstabulation				
		CARAVAN		
		0	1	Total
APERSONG	0	5442	342	5784
	1	32	6	38
TOTAL		5474	348	5822

Table 2: AGEZONG Chi-Square Tests.

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.553[a]	1	.010		
Continuity Correction[b]	4.913	1	.027		
Likelihoode Ratio	4.646	1	.031		
Fisher's Exact Test				.024	.024
Linear-by-Linear Association	6.552	1	.010		
N of Valid Cases	5822				

Crosstab analysis is a scatter plot of the relationship between two variables, so it is usually used to display the relationship between two variables [4]. The crosstab is first selected to analyze the two variables "AGEZONG" and "CARAVAN." From the crosstab, we can see that 5.91% of the customers who did not purchase family accident insurance bought RV insurance. 15.8% of the customers who bought family accident insurance bought RV insurance, three times as many as the customers who did not buy family accident insurance.

The chi-square test aims to find the correlation hypothesis between two or more groups, populations, or standards [5]. Therefore, the Pearson chi-square test was selected, and the significant improvement was less than 0.05. P value is greater than 0.05, which is highly significant, and the H_0 hypothesis is valid; If the P value is less than 0.05, the significance is low, and the H_0 hypothesis is suspected [6]. Therefore, the H_0 assumption is rejected, and the H_A assumption is accepted. If you buy family accident insurance, you may also buy TIC RV insurance.

Then, the discrete variable "APERSONG Number of private accident insurance policies" is described and statistically analyzed. Table 3 and Table 4 are as follows:

DOI: 10.54254/2754-1169/6/2022165

Table 3: APERSONG and CARAVAN Cross Table.

APERSONG*CARAVAN Crosstabulation				
		CARAVAN		
		0	1	Total
APERSONG	0	5444	347	5791
	1	30	1	31
TOTAL		5474	348	5822

Table 4: APERSONG Chi-Square Test.

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.420[a]	1	.517		
Continuity Correction[b]	.072	1	.789		
Likelihoode Ratio	.499	1	.480		
Fisher's Exact Test				1.000	.439
Linear-by-Linear Association	.420	1	.517		
N of Valid Cases	5822				

The Pearson chi-square test showed that its progressive significance was more significant than 0.05. Therefore, the original assumption H_0 was not rejected. In other words, whether to purchase private accident insurance has no impact on purchasing TIC RV insurance.

Why do customers who have bought family accident insurance prefer to buy RV insurance than those who have bought private accident insurance?

Firstly, from a behavioral economics perspective, private accident insurance belongs to individual behavior, and the beneficiary is an individual. Family accident insurance is a group behavior, and the RV is a group behavior. Therefore, family accident insurance customers are more likely to buy RV insurance than private insurance.

Secondly, RV camping is very popular in the UK market and a popular way to enjoy leisure time with family and friends [7]. RV campers tend to cover all ages and are often family oriented. Adult couples with children are the largest group, so many outdoor reception places focus on families.

Thirdly, customers are loyal to the company they initially chose. They prefer to buy all insurance services from the same company. If customers have purchased family accident insurance in TIC, they will likely continue to purchase RV insurance in TIC [8]. Therefore, TIC Insurance Company can consider selling the combination of family accident insurance and RV insurance. The more insurance policies bundled in the portfolio, the lower the price customers need to pay. This will significantly increase consumers' interest in purchasing RV insurance.

4. Tests of Normality & Hypothesis Test

A group of continuous variables about education level was selected for descriptive statistics and normality tests and got some interesting information. The selected variables are "MOPLHOOG High-level education," "MOPLMIDD Medium-level education," and "MOPLLAAG Lower-level education."

DOI: 10.54254/2754-1169/6/2022165

Kolmogorov Smirnov can check whether a group of variables come from the specified continuous distribution [9].

Table 5: Tests of Normality.

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	CARAVAN	Statistic	df	Sig.	Statistic	df	Sig.
MOPLHOOG	0	.209	5474	.000			
	1	.176	348	.000	.887	348	.000
MOPLMIDD	0	.128	5474	.000			
	1	.141	348	.000	.961	348	.000
MOPLLAAG	0	.105	5474	.000			
	1	.125	348	.000	.961	348	.000

a. Lilliefors Significance Correction

Table 5 MOPLHOOG MOPLMIDD and MOPLLAAG Tests of Normality.

So, those three variables join normality tests. From Table 5, the significance of the Kolmogorov-Smirnov test is 0.00, less than 0.05. Low significance indicates that these three variables do not conform to the normality assumption, so it is necessary to continue the nonparametric test.

Through the nonparametric test, the significance of the three variables is less than 0.05. Therefore, the original hypothesis H_ 0 was rejected. In other words, the level of education affects the purchase of TIC RV insurance. In Table 5, descriptive statistics show that the higher the education level, the greater the possibility of purchasing RV insurance.

What causes consumers with a higher level of education to be more likely to buy RV insurance?

Firstly, people with a high degree of education have a stronger sense of safety. This potential customer is more likely to buy the insurance and use it to ensure the high safety of their medical care, health care, and property [10].

Secondly, people with high education levels primarily work in excellent jobs, such as teachers. As a teacher, there are many holidays and free time, so there is more time to travel by RV.

5. K-Means Clustering

The purpose of enterprise analysis on its customers is to deeply understand consumers' thinking processes, predict consumer behavior, and create a market segment [11]. K-means clustering can be used for user and market segmentation, so it can be used to classify the variable "MOSTYPE customer subtype." The cluster attributes of 41 subtypes are found through 10 iterations, and four categories are formed.

Then, the newly generated case clustering factor column is described and statistically analyzed to get Table 6. It can be seen from Table 6 that the progressive significance of the Pearson Chi-square experiment is 0.000 less than 0.05, and the P value is less than 0.05, so it is significant.

Table 6: Cluster Number of Case Chi-Square Test.

Chi-Square Tests			
	Value	df	Asymptotic Significance(2-sided)
Pearson Chi-Square	7354.963[a]	117	.000
Likelihood Ratio	7121.186	117	.000
Linear-by-Linear	2171.861	1	.000
N of Valid Cases	5822		

DOI: 10.54254/2754-1169/6/2022165

Table 7: Subtype I.

Cluster Number of Case	% Within MOSTYPE	Customer Subtype
1	64.00%	Residential elderly
1	52.30%	Porch less seniors: no front yard
1	75.30%	Lower class large families
1	74.70%	Large family, employed child
1	61.70%	Village families
1	52.40%	Couple with teens “married with children”
1	72.70%	Mixed small town dwellers
1	49.00%	Traditional families
1	59.50%	Large religious families
1	80.30%	Large family farms
1	95.60%	Mixed rurales

From the cross table of cluster number of cases and "MOSTYPE", we can get the following characteristics of the four subsets according to the percentage of 41 customer subtypes in different categories.

The subtypes of customers belonging to the first category are shown in Table 7.

Most of these families live in small towns or rural areas, and the income of large families is not high. The second is the elderly who live alone, and the house is not spacious. This type of customer has a low income, so the probability of purchasing RV insurance is low. This type of customer has a low income, so the probability of purchasing RV insurance is low.

The subtypes of customers belonging to the second category are shown in Table 8.

Table 8: Subtype II.

Cluster number of case	% Within MOSTYPE	Customer Subtype
2	66.70%	Suburban youth
2	72.00%	Ethnically diverse
2	58.90%	Young, low educated
2	91.70%	Own home elderly
2	56.00%	Seniors in apartments
2	66.90%	Religious elderly singles
2	88.80%	Low income Catholics
2	49.60%	Mixed seniors

This type of family has the following characteristics. Most are elderly, Catholics, and young people with low educational backgrounds. Most of them live in the suburbs or rural areas, and their income is low. Therefore, the probability of purchasing RV insurance is small.

The subtypes of customers belonging to the third category are shown in Table 9.

DOI: 10.54254/2754-1169/6/2022165

These families have the following characteristics: most live in central provinces and cities, and their income is above average. Family members are young and modern, willing to invest more in their children. Therefore, the probability of participating in family RV travel or RV camping is high, and the probability of purchasing RV insurance is high.

Table 9: Subtype III.

Cluster number of case	% Within MOSTYPE	Customer Subtype
3	96.80%	High income, expensive child
3	73.20%	Very important Provincials
3	68.30%	High status seniors
3	89.90%	Career and childcare
3	86.40%	Dinki's (double income no kids)
3	90.60%	Middle class family
3	75.80%	Stable family
3	50.30%	Family starters
3	70.30%	Affluent young families
3	50.80%	Young all American family

The subtypes of customers belonging to the fourth category are shown in Table 10.

Table 10: Subtype IV.

Cluster number of case	% Within MOSTYPE	Customer Subtype
4	51.90%	Affluent senior apartments
4	66.70%	Mixed seniors
4	65.80%	Modern, complete families
4	100.00%	Senior cosmopolitans
4	93.80%	Students in apartment
4	100.00%	Fresh masters in the city
4	89.50%	Single youth
4	53.30%	Young urban have-mots
4	52.00%	Mixed apartment dwellers
4	82.90%	Young and rising
4	59.80%	Young seniors in the city

This type of family has the following characteristics: most live in big cities with high incomes. Most of them are young people who have not married yet, and some are older people who are willing to accept new things. This kind of person has a higher income and accepts new things quickly, so the probability of purchasing RV insurance will be higher.

DOI: 10.54254/2754-1169/6/2022165

6. Binary Logistic Regression

In binary logistic regression, a single dependent variable can be predicted by multiple independent variables [12]. The dependent variable "CARAVAN Number of mobile home policies" is indicated, so the other 85 variables are set as independent variables.

Input all variables through the enter method of binary logical regression. And set classification cutoff to 0.05 to get the classification table 11.

In table 11, the predicted number of people who do not purchase RV insurance accounts for 64.2% of the actual number. The predicted number of people who buy RV insurance accounts for 76.7% of the actual number. This table shows that the binary logic regression prediction is effective.

Table 11: CARAVAN Binary logistic regression prediction.

Classification $Table^a$					
Predicted					
			CARAVAN		
	Observed		0	1	Percentage Correct
Step 1	CARAVAN	0	3514	1960	64.2
		1	81	267	76.7
	Overall Percentage				64.9

7. Conclusion

In general, through the customer data mining and analysis of TIC Insurance Company. This paper found that customers who buy family insurance are more likely to buy RV insurance than private insurance. Customers' education level also has a certain impact on purchasing RV insurance. The higher the education level, the greater the probability of purchasing RV insurance. In addition, a K-means clustering analysis was conducted for the first variable, "MOSTYPE" in the data dictionary. The 41 Customer Subtype are clustered into four categories, and the characteristics of each category are found out, and the probability of purchasing RV insurance for each category is analyzed. Finally, we made binary logistic regression for all independent variables and found that these variables are significant and influential for predicting the dependent variable – the purchase of RV insurance.

TIC insurance companies' customer data can also explore more interesting information, which can be well studied through data mining.

References

[1] *Acker, O., Blockus, A., & Pötscher, F. (2013). Benefiting from big data: A new approach for the telecom industry. Strategy&, Analysis Report.*

[2] *Parlar, T., & Acaravci, S. K. (2017). Using data mining techniques for detecting the important features of the bank direct marketing data. International journal of economics and financial issues, 7(2), 692-696.*

[3] *(2022). The Insurance Company (TIC) Benchmark. Kaggle. https://www.kaggle.com/datasets/kushshah95/the-insurance-company-tic-benchmark?select=dictionary.txt*

[4] *Sarstedt, M., & Mooi, E. (2019). Descriptive Statistics. In A Concise Guide to Market Research (pp. 91-150). Springer, Berlin, Heidelberg.*

[5] *Nihan, S. T. (2020). Karl Pearsonâ€™s chi-square tests. Educational Research and Reviews, 15(9), 575-580.*

[6] *Rana, R., & Singhal, R. (2015). Chi-square test and its application in hypothesis testing. Journal of the Practice of Cardiovascular Sciences, 1(1), 69.*

[7] *Lashley, C. (2015). Researching snails on holiday: An agenda for caravanning and caravanners?. Research in Hospitality Management, 5(2), 115-122.*

[8] *Lehtonen, T. K. (2017). Domesticating insurance, financializing family lives: The case of private health insurance for children in Finland. Cultural Studies, 31(5), 685-711.*

DOI: 10.54254/2754-1169/6/2022165

[9] *Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. Annals of cardiac anaesthesia, 22(1), 67.*

[10] *Outreville, J. F. (2015). The relationship between relative risk aversion and the level of education: a survey and implications for the demand for life insurance. Journal of economic surveys, 29(1), 97-111.*

[11] *Ahuja, V., & Medury, Y. (2011). Corporate blogs as tools for consumer segmentation-using cluster analysis for consumer profiling. Journal of Targeting, Measurement and Analysis for Marketing, 19(3), 173-182.*

[12] *Sreejesh, S., Mohapatra, S., & Anusree, M. R. (2014). Binary logistic regression. In Business Research Methods (pp. 245-258). Springer, Cham*