Prediction of Health Insurance Cross-Selling Through Binary Logical Regression and Decision Tree Model

Shenyi Wang*

School of Statistics, Capital University of Economics and Business, Beijing, China

* Corresponding Author Email: 32022120057@cueb.edu.cn

Abstract. Cross-selling is typically a marketing strategy used to identify a customer's various wants and then sell a range of related services or products by attending to those demands. Cross-marketing refers to a variety of tactics and procedures used to encourage cross-selling. The most crucial role in conventional banking and insurance is played by cross-selling. In most nations, auto insurance has traditionally played a significant role in the insurance market. This paper is based on information from all of a medical insurance company's clients and their automobiles throughout the course of a certain year. The binary logistic regression model and decision tree model are used to examine the customer's interest in purchasing auto insurance in the dichotomy, and the two techniques are compared to identify the most accurate prediction method. Through reasonable analysis of independent variables and dependent variables, the prediction accuracy of the binary logical regression model is 87.7% and that of the decision tree model is 87.6%. The accuracy of the two is similar.

Keywords: Binary logical regression, Decision Tree, Cross-selling.

1. Introduction

1.1. Background

Automobile insurance is another name for "vehicle insurance," which is insurance for motor vehicles. It is a type of business insurance that protects automobiles. Vehicle insurance is one type of property insurance that covers losses in terms of personal injury and property damage brought on by accidents and natural disasters. Also, vehicle insurance is a type of property insurance that took a while to create since it was brought about by the development and use of automobiles. the development of life. In most nations, auto insurance has long been a significant component of the insurance market. The market for vehicles is expanding along with the industry's ongoing development and improvement. The current state of the world's four largest vehicle markets. There are four countries: China, the US, Japan, and India. Also, as more and more families purchase cars, the cost of auto insurance is rising. The growth of insurance industry is prospering more and more. Increasingly significant in the insurance industry [1].

Machine learning is also commonly employed in the big data era to solve a variety of categorization and prediction issues. Insurance overlapping Customers' interest in insurance cross-selling items can be viewed as one of two categories in the sales forecasting challenge. Create a quick and precise classification model to identify whether clients are considering purchasing insurance cross-selling items. Interest will considerably shorten the time insurance companies spend looking for clients and assist them in precisely identifying the insured customer group and raising the business's earnings. As a result, the machine learning approach is employed to create an appropriate insurance policy. For insurance businesses, the cross-sales data prediction model has significant practical implications [2].

1.2. Relation research

Foreign vehicle insurance first appeared in Europe in the middle and late 19th centuries. With the emergence and gradual popularization of automobiles at that time, the accidental injuries and property losses caused by traffic accidents are increasing. Some people found business opportunities in it, and auto insurance gradually emerged [3]. Several studies in foreign nations use machine learning and

statistical models to analyze vehicle insurance data. In 2003, Blows and other researchers looked at the relationship between the risk of auto insurance and injury accidents using logistic regression [4]. In 2013, Li and Liu, and other researchers discussed the coverage of bundled auto insurance and the occurrence of claims based on simultaneous equations and probability models, determined the relationship between the two types of auto insurance and the trend of claims, and discovered that there was a certain correlation between the purchase of auto insurance and the occurrence of vehicle accidents. They then proposed that the probability of accidents could be predicted by the data of customer service. supplying pertinent data for insurance companies underwriting guidelines [5].

Compared with foreign countries, China's auto insurance started late, and the insurance law and insurance system are not completely perfect. Cross-selling research has lagged in China, and it only picked up speed with the insurance industry Vehicle Lane after the turn of the twenty-first century. Cross-selling was defined by Guo Guoqing, who also covered its potential in the insurance sector [6]. Cross-selling, according to Cao Yuanhua, has emerged as one of the key drivers of the Chinese insurance industry's potential for rapid expansion after years of development. He also made recommendations regarding cross-selling [7]. Wang Xiaoming selected the cross-selling practice cases of Ping An Insurance to analyze and provide suggestions on the cross-selling mode of China's insurance industry [8].

1.3. Objection

Although machine learning, particularly ensemble learning algorithms, has been widely employed in various industries, according to the state of research both domestically and internationally, no researchers have used ensemble learning algorithms to study insurance cross-selling data. Also, despite the fact that there are several research on insurance cross-selling, they tend to concentrate on theoretical analyses of the model's current state and recommendations for its future development, as well as the dearth of statistical learning techniques in the age of big data. the cross-selling of data [9].

Based on the survey data of all users of a medical insurance company in a certain year, this paper analyzes the impact of various factors on customers' interest in the vehicle insurance provided by the company. Try to help the company plan its communication strategy with customers, optimize its business model, and increase revenue once contact with these customers.

In this study, the dependent variable is whether the customer is interested in vehicle insurance, while the independent variable is the basic information of some customers (age, gender, region, whether there is a driver's license, whether there is vehicle insurance, age of the vehicle, whether the vehicle has been damaged, the insurance premium to be paid in the current year, communication channels with the customer, and the number of days to contact the company). Next, the binary logical regression model and decision tree model are established through SPSS and the data are analyzed respectively. Understand which factors have a significant impact on the dependent variable (whether the customer is interested in vehicle insurance), and analyze which model is more advantageous for data analysis by comparing the accuracy with the real value. Additionally, it assists insurance businesses in precisely identifying the insured consumer groups for cross-selling insurance and assists them in successfully implementing cross-selling.

2. Methodology

2.1. Source of data

The data set used in the study is from Kaggle, and the time span of the data set is one year. The dataset includes 381110 samples. There is 11 basic customer information in each sample, including 10 independent variables and 1 dependent variable which is whether the customer responds to the insurance (whether the customer is interested in insurance).

2.2. Data processing

First of all, in the selection of sample data, the selection of the two models is not the same. Different sample selection methods are used for selecting samples in binary logical regression and decision tree models. Due to the fact that binary logical regression models cannot capture trends and interactive features when dealing with too large samples, the author conducted a random sampling of 50% of the samples [10]. Decision tree models can quickly adapt to data sets, perform well in large data sets, and are extremely fast. Therefore, in order to ensure the integrity of the sample dataset, this article does not randomly select samples in the decision tree model but selects all samples for model construction. And for the binary logical regression model, the number of samples after random sampling is still large and representative. Therefore, the error caused by different data processing methods will not be too large and can be ignored.

There are ten independent variables and one dependent variable in the model. For independent variables, the results are not easy to observe due to a large number of regions, so the region code is used to represent different regions. At the same time, in order to facilitate the analysis, the author expressed the three independent variables of whether the driver's license is available, whether the vehicle insurance is available, and whether the vehicle is damaged in a binary format (1: yes 0: no). In addition, reference variables are set for the convenience of analysis, and the vehicle age is 1-2 years, there is vehicle damage and male are set as reference variables. However, there is no clear distinction between the dependent variable and whether the customer is interested in vehicle insurance. Therefore, whether the customer is interested in this is indicated by whether the customer responds.

Secondly, for the decision tree model, the processing of sample numbers is not the same as that of the binary logical regression model. In order to prevent the loss of characteristic samples from causing the instability of the model and resulting in incorrect results, the author selects all samples in the data set to build the model. The treatment of independent variables and dependent variables is the same as that of the binary logistic regression model except that no reference variables are set.

2.3. Model

2.3.1 Binary logical regression

Logical regression is a typical method of supervised learning, which usually solves the dichotomous problem [11].

Linear regression can be represented by $y=X\theta$ as the prediction function, while logical regression can be represented by y=g(z)

Where $z= X\theta$, the function g maps the range of linear regression values from the global R to the interval $[0,1]$, which usually uses The sigmoid function, namely:

$$g(z)= \frac{1}{1+e^{-z}} \quad (1)$$

Observe the characteristics of the sigmoid function. When $z \to +\infty$, $g(z) \to 1$; When $z \to -\infty$, $g(z) \to 0$. The property of the sigmoid function ensures that the probability predicted by the logistic regression model is between 0 and 1. Then the problem of continuous value is transformed into a classification problem solved by logical regression [2].

The hypothetical function of logical regression is:

$$h_\theta(X) = g(X\theta) = \frac{1}{1+e^{-X\theta}} \quad (2)$$

Where X is the sample data, $h_\theta(X)$ is the result, and θ is the model parameter. If 0.5 is set as the critical value, Then when $h_\theta(X) \geq 0.5$, that is, when $X\theta>0$, y=1; When $h_\theta(X)<0.5$, that is, when $X\theta<0$, y=0. The value of the output result h (X) of the model is within the interval $[0,1]$. The closer $h_\theta(X)$ is to 0, The greater the probability of classification result y=0; The closer $h_\theta(X)$ to 1, the greater the probability of classification result y=1; $h_\theta(X)$ If it is close to 0.5, the Classification result cannot be judged [2].

Logical regression is easy to understand and realize, and the model training speed is fast, and through the mapping of the sigmoid function, it can effectively control the noise of the sample data, and can also eliminate the multicollinearity problem. It is the preferred method to solve the binary classification problem [2]. However, logistic regression also has the problem of under-fitting, and because the probability changes slightly when the sigmoid is close to 0 or 1, but it is more sensitive to the middle of the interval [0, 1], the fluctuation is large, and it is often impossible to judge the critical sample [2].

2.3.2 Decision tree

Decision tree classification is a typical multi-level classification method, which is composed of a series of binary decision trees. As shown in Figure 1, it is a classification method that subdivides the data set from top to bottom layer by layer. It is used to divide the pixels into corresponding categories. Generally, each decision tree divides the pixels in the image into two categories based on the expression of one classification rule, New born categories can continue to be classified to the next level according to other classification rules to achieve the satisfaction degree of the expected classification results [12].

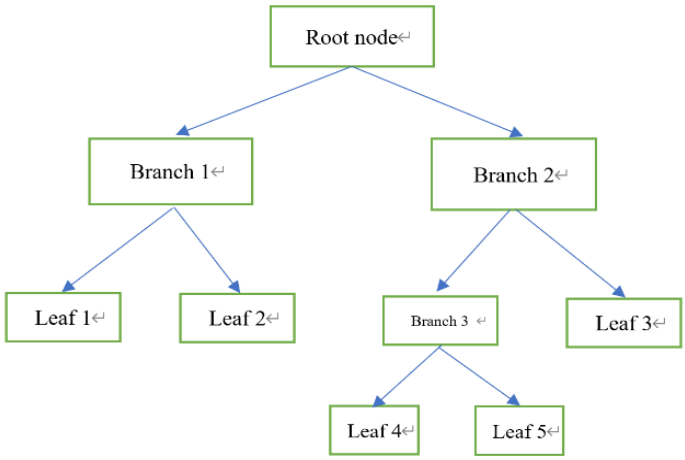


Figure 1. Decision Tree Structure (Photo credit: Original)

Information entropy and Gini value are commonly used to measure sample purity. Assume that the proportion of class k samples in sample set D is $p_k(k=1,2,\ldots\ldots, |y|)$, the information entropy of the sample set is defined as:

$$\text{Ent(D)} = -\sum_{k=1}^{|y|} p_k \log_2 p_k \tag{3}$$

Gini value is:

$$\text{Gini(D)} = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|y|} p_k^2, \tag{4}$$

When the values of formula (3) and formula (4) are smaller, the purity of the sample set will be higher, and the effect of attribute classification will be better. Generally speaking, the criteria for optimal attribute classification include information gain, information gain rate, Gini index, etc [1].

3. Result and Discussion

3.1. Data visualization

Figures 2-9 show the distribution of customers under different factors. Figure 2 shows the distribution of customers' age, Figure 3 shows the distribution of customers' age and premium amount, Figure 4 and Figure 5 show the distribution of customers' gender, Figure 6 shows the distribution of customers' vehicle insurance, Figure 7 shows the distribution of customers' vehicle age, Figure 8 shows the distribution of customers whether the vehicle is damaged, Figure 9 shows the distribution

of customers' insurance amount, and Figure 10 shows the distribution of the number of days that customers are associated with the company. These distribution charts mainly describe the distribution of samples in different classification methods, showing the distribution of samples more intuitively, and having a preliminary understanding of experimental samples.

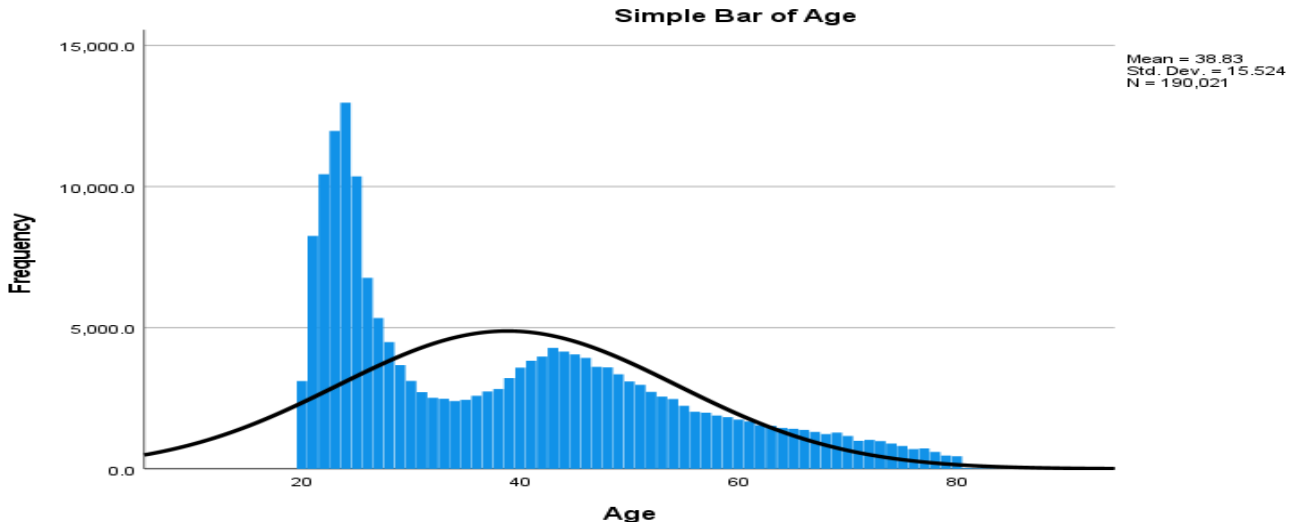


Figure 2. Age Distribution of Customers (Photo credit: Original)

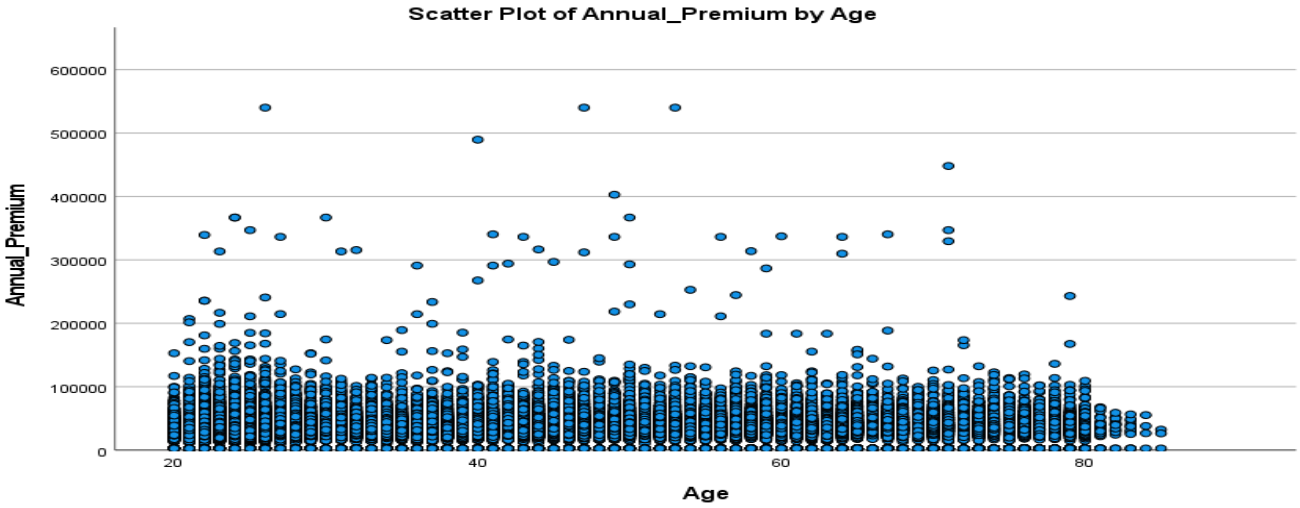


Figure 3. Age Vs Annual premium (Photo credit: Original)

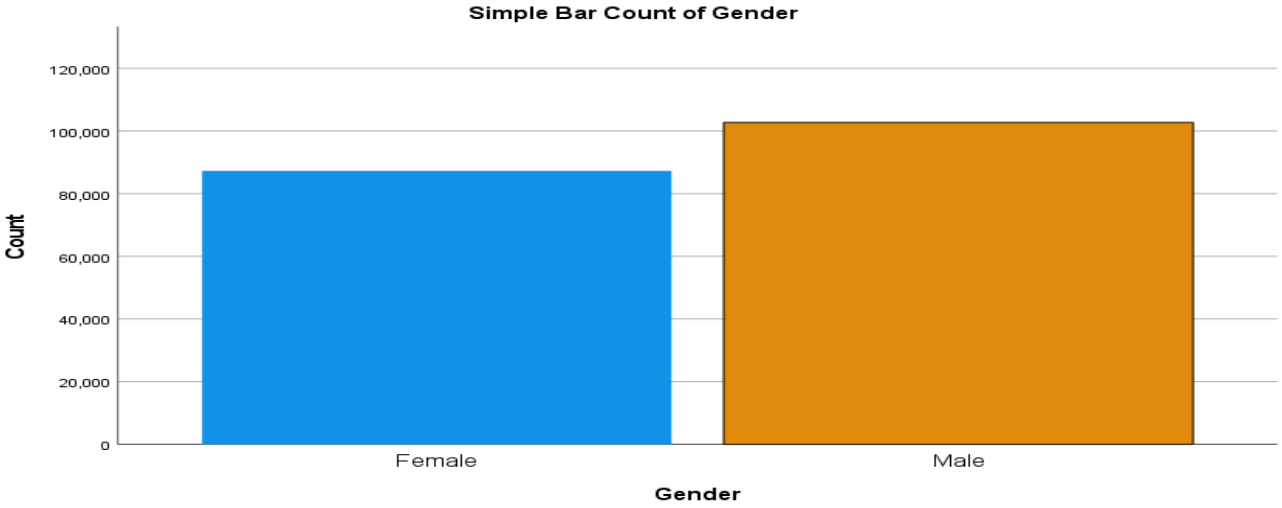


Figure 4. Gender and Response (Photo credit: Original)

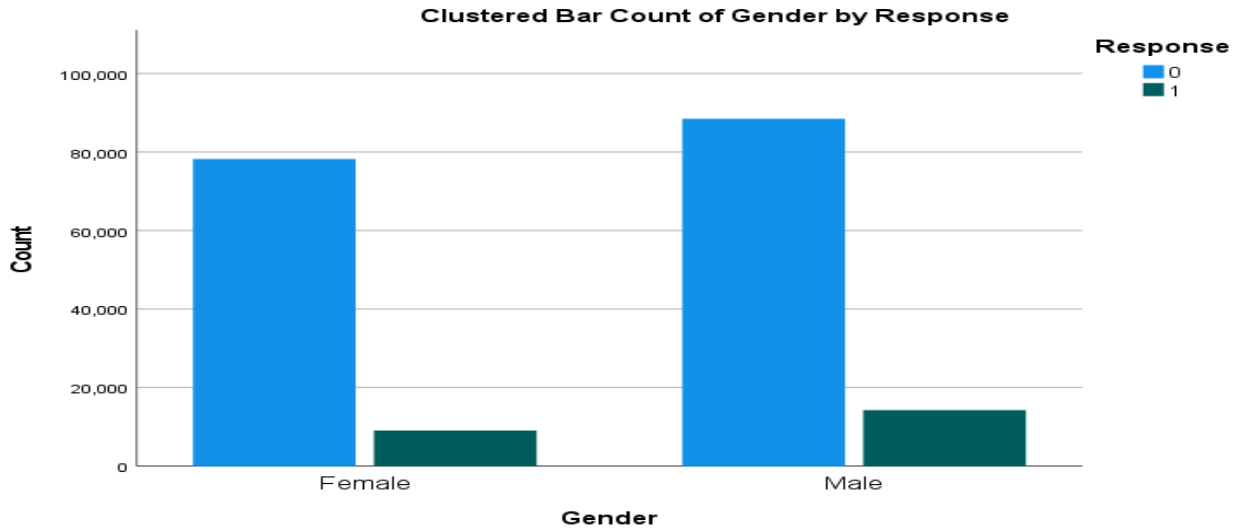


Figure 5. Gender and Response (Photo credit: Original)

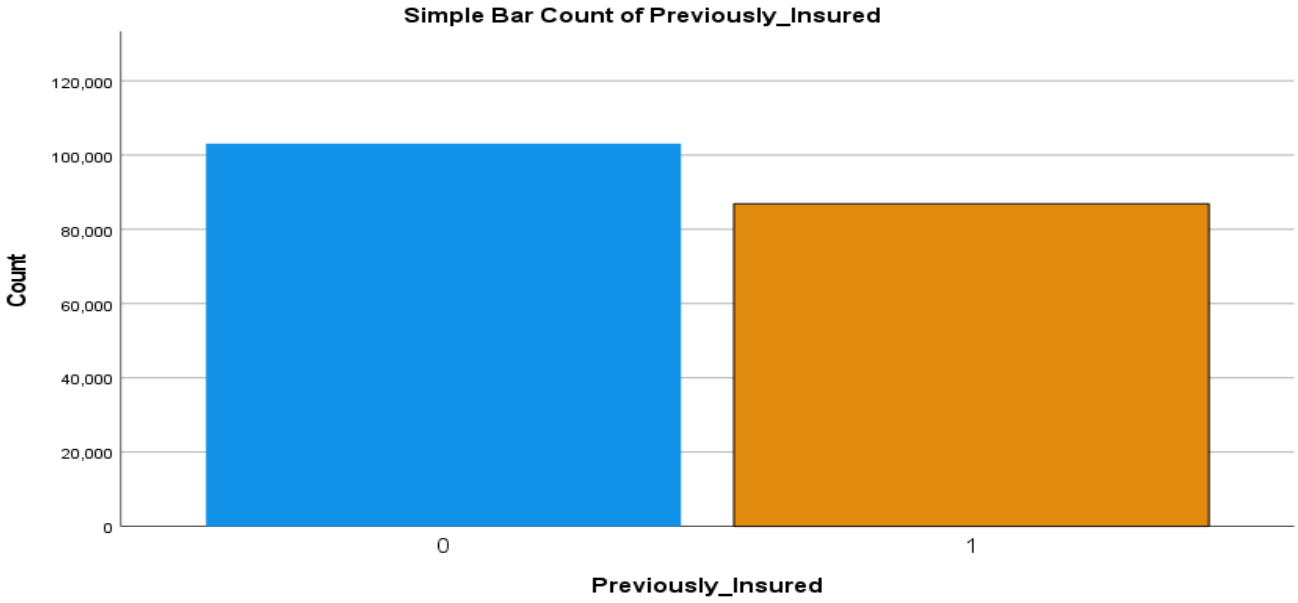


Figure 6. Customers having Vehicle insurance already (Photo credit: Original)

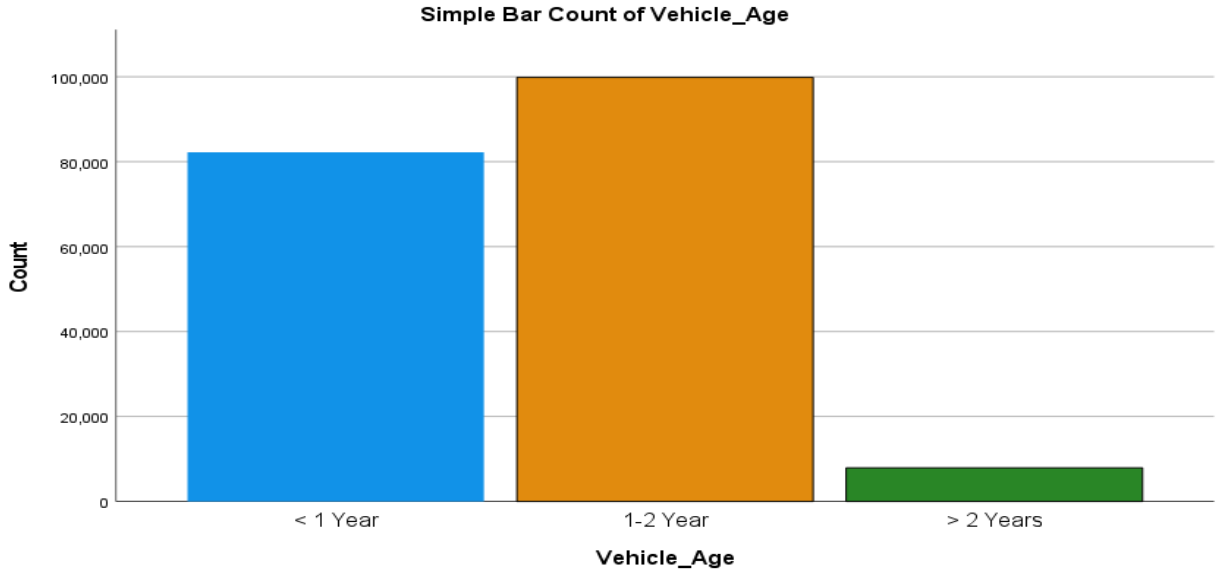


Figure 7. Customers having damaged vehicle (Photo credit: Original)

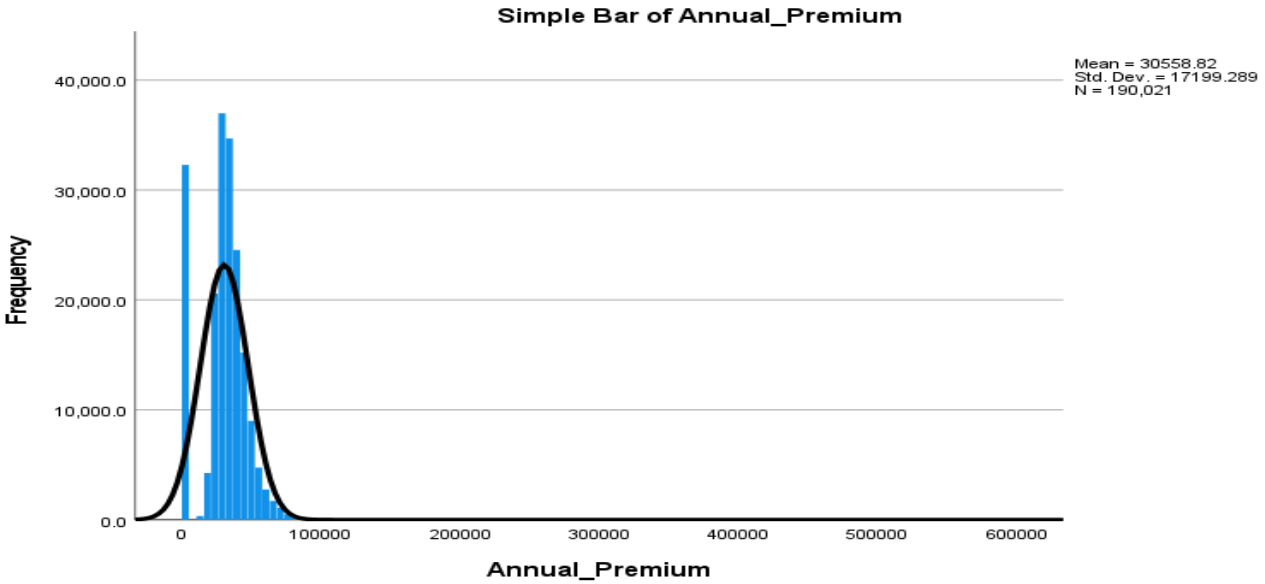


Figure 8. Annual Premium Distribution (Photo credit: Original)

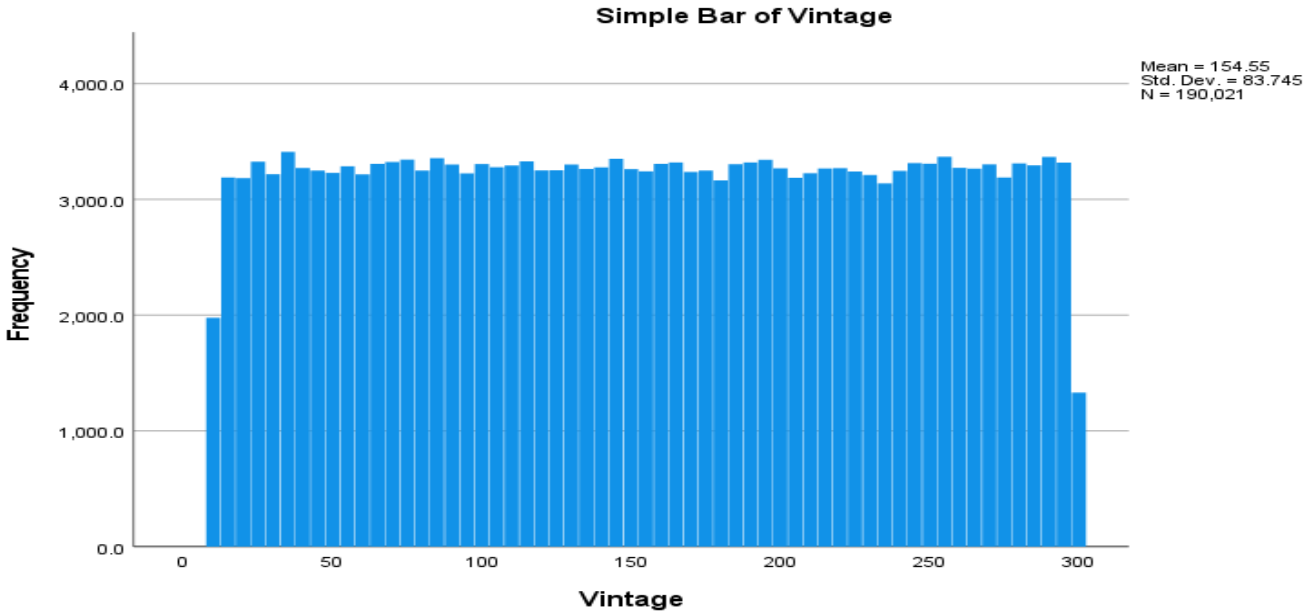


Figure 9. Vintage(Photo credit: Original)

3.2. Binary Logistic Regression:

3.2.1Model building steps

First, the dataset collected from Kaggle is substituted into SPSS, and then a binary logical regression model is selected, with independent variables and dependent variables added separately. At the same time, 50% random sampling is selected for the selection of sample data. Subsequently, model analysis was conducted to obtain the following table.

Table 1. Cases Processing Summary

Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	190021	100.0
	Missing cases	0	.0
	Total	190021	100.0
Unselected Cases		0	.0
Total		190021	100.0

Table 2. Categorical Variables Codings

			Parameter coding	
		Frequency	(1)	(2)
Vehicle_Age	<1 Year	82242	1.000	.000
	>2 Year	7904	.000	1.000
	1-2 Year	99875	.000	.000
Vehicle_Damage	No	94012	1.000	
	Yes	96009	.000	
Gender	Female	87289	1.000	
	Male	102732	.000	

Table 3. Classification Table

				Predicted	
			Response	Percentage	correct
	Observed		0	1	
Step1	Response	0	166733	3	100.0
		1	23281	4	.0
	Overall Percentage				87.7

Table 4. Model Summary

Step	**-2Log likelihood**	**Cox & Snell R Square**	**Nagelkerke R Square**
1	104628.181	.176	.335

Table 5. Variables in the Equation

Variables in the Equation									
								95%C.I.forEXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1	Gender(1)	-.098	.016	38.798	1	<.001	.907	.879	.935
	Age	-.025	.001	1090.923	1	<.001	.975	.973	.976
	Driving License	1.218	.234	27.105	1	<.001	3.382	2.138	5.350
	Region Code	.000	.001	.145	1	.704	1.000	.999	1.001
	Previously Insured	-3.908	.113	1195.078	1	<.001	.020	.016	.025
	Vehicle Age			2024.348	2	.000			
	Vehicle Age(1)	-1.179	.027	1965.408	1	.000	.308	.292	.324
	Vehicle Age(2)	.223	.027	67.649	1	<.001	1.250	1.185	1.318
	Vehicle Damage(1)	-1.994	.048	1734.760	1	.000	.136	.124	.150
	Annual Premium	.000	.000	32.030	1	<.001	1.000	1.000	1.000
	Policy Sales Channel	-.002	.000	232.838	1	<.001	.998	.997	.998
	Vintage	.000	.000	.050	1	.822	1.000	1.000	1.000
	Constant	-.788	.240	10.742	1	.001	.455		

3.2.2 model analysis

Table 1 is the Cases Processing Summary. It can be seen from the table that the total number of samples selected in the analysis is 190021, including 19021 available samples and 0 missing samples. Therefore, no missing samples are found in the binary logical regression model, and the sample selection rate is 100%.

Table 2 is Categorical Variables Codings. This table shows the selection of reference variables in the independent variables in the analysis. It can be seen from the table that for vehicle age classification, there are 82242 samples less than one year, which are set to (1.000, 0.000). There are 7904 samples larger than two years, which are set to (0.000, 1.000). There are 99875 samples from one year to two years. Set it to (0.000, 0.000). It can be seen that for the vehicle age, the vehicle age is 1 to 2 years as the reference variable

For whether the vehicle is damaged, there are 94012 undamaged samples, which are set to (1.000). The number of damaged samples is 96009. Set it to (0.000). It can be seen that whether the vehicle is damaged or not is the reference variable. For gender, the female sample has 87289, which is set to (1.000). The male sample has 102732, which is set to (0.000). It can be seen that male is the reference variable for the gender of customers. Therefore, when describing the above three variables in the following table, the results are obtained by comparing non-reference variables with reference variables. For example, when describing the impact of gender on the dependent variables, Table 5 shows the impact of women compared with men.

Table 3 is the Classification Table. This table mainly describes the accuracy of the results obtained from the model analysis of the experimental data compared with the real data. Therefore, the accuracy of the binary logical regression model is 87.7%

Table 4 is the Model Summary, which mainly shows the fitting degree between the model and the real data. For the binary logistic regression model, when the values of Cox&Snell R Square and Nagelkerke R Square are between 0.1-0.3, the model can be considered reasonable and has a high degree of fit. It can be seen from the table that the values of Cox&Snell R Square and Nagelkerke R Square of the constructed model are 0.176 and 0.335 respectively, about 0.1-0.3, so the model is considered reasonable.

Table 5 is Variables in the Equation. The focus of this table is on the column that is circled. This column is significant, because which mainly shows the influence of different independent variables on dependent variables. When the significance of an independent variable is less than 0.01, the independent variable can be considered to have a significant impact on the dependent variable. Therefore, this paper can conclude that the result of the logistic regression model is that all independent variables have a significant impact on the dependent variables except for the region and vintage. That is, except for the region and vintage, other factors have a significant impact on whether the customer responds (whether the customer is interested in vehicle insurance).

3.3. Decision Tree

3.3.1 Model building steps

For the decision tree model. First, the dataset collected from Kaggle is imported into SPSS, and then a decision tree model is selected. In the validation, 70% of the training samples and 30% of the detection samples are selected. Secondly, in order to facilitate observation, select the number of terminal nodes, prediction values, prediction probabilities, and sample allocation during storage. Finally, to prevent inaccurate results due to too few node cases, click the "Condition" button to set the minimum case tree parent node to 400 and the child node to 200 [13].

Table 6. Model Summary

Model Summary		
Specifications	Growing Method	CHAID
	Dependent Variable	Response
	Independent Variables	Gender, Age, Driving_License, Region_Code, Previously_Insured, Vehicle_Age, Vehicle_Damage, Annual_Premium, Policy_Sales_Channel, Vintage
	Validation	Split Sample
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	400
	Minimum Cases in Child Node	200
Results	Independent Variables Included	Vehicle_Damage, Age, Policy_Sales_Channel, Previously_Insured
	Number of Nodes	51
	Number of Terminal Nodes	37
	Depth	3

Table 7. Risk Table

Risk		
Sample	Estimate	Std. Error
Training	.122	.001
Test	.124	.001

Growing Method: CHAID
Dependent Variable: Response

Table 8. Classification table

Classification				
Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	234445	0	100.0%
	1	32621	0	0.0%
	Overall Percentage	100.0%	0.0%	87.8%
Test	0	99954	0	100.0%
	1	14089	0	0.0%
	Overall Percentage	100.0%	0.0%	87.6%

Growing Method: CHAID
Dependent Variable: Response

3.3.2 model analysis

Table 6 is the Model Summary, which shows the growth method, independent variables, and results of the model. The independent variables shown in the results are those that have a significant influence on the dependent variables. Therefore, it can be seen from the table that in the decision tree model, Vehicle_Damage, age, Previously Insured, and Policy Sales Channel have a significant impact on the dependent variables Sales Channel, whether there is vehicle insurance.

Table 7 is the risk table, which mainly shows the effect of model construction. Therefore, it can be seen from the table that the training estimate is 0.122, that is, 12.2% of the 70% of training samples are incorrectly classified. The test estimate is 0.124, that is, 12.4% of the 30% of the test samples are incorrectly classified.

Table 8 is a classification table, which also shows the effect of the model. It can be seen from the table that there are 32621 samples responding in the training set, 234445 samples not responding, and 87.7% of the training set indicates that the correct rate of the model is 87.7%. The test set indicates that there are 14089 samples responding, 14089 samples not responding, and 87.6% of the test set indicates that the data of the test set is being tested with the model trained in the training set, and the correct rate is 87.6%.

4. Discussion

Comparing the results of the binary logical regression model and decision tree model, the common independent variable with significant influence is Vehicle_ Damage, age, Previously Insured, Policy Sales Channel. Therefore, the conclusion is that these factors (Vehicle_Damage, Age, Previously Insured, Policy Sales Channel) have a significant impact on whether the dependent variable customers respond (whether customers are interested in vehicle insurance). And the accuracy of the two models is between 87.6% and 87.7%, and the accuracy of the two models is similar.

Therefore, vehicle insurance companies can pay more attention to the above factors to help them improve the selection, communication, and strategy of customers, so as to further improve the sales structure and improve the model quality and income.

It should be noted that since the data used in this study is only the income of an insurance company in a certain year, the independent variables in this paper are not extensive and diversified enough, which may lead to errors in the results. And the results obtained by using the two models are not exactly the same, which may lead to some factors not being identified as key factors with significant impact, thus causing some errors. At the same time, in the analysis of the decision tree model, some samples in the training set and the test set are incorrectly classified, which may cause the loss of some decisive samples and cause unnecessary errors. Secondly, this paper only uses two models to analyze the problem. The model has limitations, leading to some variables that have not been proven to have a significant impact.

5. Conclusion

Research shows that Vehicle_ Damage, age, Previously Insured, and Policy Sales Channel have a significant impact on their interest in purchasing vehicle insurance. The accuracy of the constructed binary logical regression model is similar to that of the decision tree model.

Insurance providers ought to concentrate on Vehicles_ Create more precise marketing strategies for automobile insurance based on variables like damage, age, previous insurance, and policy sales channel. For instance, insurance companies can improve communication with customers who have never purchased auto insurance, customers who have had damage to their cars, and customers whose cars are older than two years, as well as consistently publish and push content promoting auto insurance. And use some unique strategies for various consumer age groups, such as employing new media for young customers.

This paper presents instances of binary logical regression models and decision tree models, enriching the models for forecasting automobile insurance. It can be applied by later researchers to help in model selection.

In order to forecast the interest in purchasing automobile insurance, the binary logical regression model and decision tree model are utilized in this work. Both models have shown rather accurate classification results. The binary logical regression and decision tree models used in current research are both simple and fast. However, there are still some gaps between the accuracy and some complex models, which can lead to some objective errors that cannot be avoided. Subsequent researchers may consider using other models for further research and selectively changing some model settings.

References

[1] Hu Jialin Research on prediction of vehicle insurance purchase interest based on LightGBM model [D]. Dalian University of Technology, 2021. DOI: 10.26991/d.cnki.gdlu.2021.001409.

[2] Wu Peisen Insurance cross-selling forecast based on Ada-Voting integrated model [D]. Huazhong Agricultural University, 2022. DOI: 10.27158/d. cnki. ghznu.2022.001583.

[3] Cohen A. Siegelman P. Testing for Adverse Selection in Insurance Markets [J] Journal of Risk & Insurance. 2010.77(1): 39-84.

[4] Blows S, Ivers R, Connor J, et al. Car insurance and the risk of car crashinjury [J]. Accident Analysis& Prevention, 2003, 35(6): 987-990.

[5] Li C S, Liu C C, Peng S C. Bundled automobile insurance coverage and accidents [J] Accident; analysis and prevention, 2013, 50(2): 64-72.

[6] Guo Guoqing, Wu Jianfeng, Qian Minghui. Research on the application prospect of cross-selling in China's financial industry: the experience and enlightenment of the United States. Nankai Management Review, 2005 (01): 47-51.

[7] Cao Yuanhua. On Six Combinations to Help Insurance Cross-selling. Financial Horizon, 2013 (09): 93-98.

[8] Wang Xiaoming. Analysis and Improvement of Cross-selling Model in Insurance Industry I - Take Ping an Insurance as an Example, Times Finance, 2019 (25): 73-74.

[9] Wang Sijing Insurance Cross Sales Forecast Based on Unbalanced Data [D]. Heilongjiang University, 2022. DOI: 10.27123/d.cnki.ghlju.2022.000814.

[10] Li Fuxiao, Chen Zhanshou. Structural Change Point Estimation for Multiple Logistic Regression Models [J]. Mathematical Practice and Understanding, 2020, 50 (09): 122-131.

[11] Li Yunxia, Liu Weitang. Changing Point Test of Logistic Regression Model Based on Empirical Likelihood [J]. Journal of College Applied Mathematics A, 2015,30 (03): 367-378. DOI: 10.13299/j.cnki. amjcu.001882.

[12] Liu Yonghong, Niu Zheng, Wang Changyao. Research and application of decision tree classification method based on MODIS data [J]. Journal of Remote Sensing, 2005 (04): 405-412.

[13] Yang Haiqin. Analysis of influencing factors of hospitalization expenses in AMI based on quantile regression and decision tree model [J]. China Medical Records, 2023,24 (02): 67-72.