What Should A Car Insurance company focus on?

Ziwen Liao [*][†]Xinxin Zhong [‡]Di Ma [§]

April, 2024

Abstract

Following the pandemic's economic impact, auto insurance companies require recovery. To assist companies in understanding their customers better and creating successful strategies, relevant data was collected. This data revealed correlations between customers' lifetime value and 24 influencing factors. Out of these factors, nine were selected as the primary focus of the research. It is hypothesized that income, vehicle class, and driving location are likely to be the most influential factors in customers' lifetime value. To validate this hypothesis, we will use R Studio software to analyze whether a significant correlation exists between customers' lifetime value and these nine independent variables. The analysis methods include the t-test, simple regression model, multiple linear regression model, and logistic regression model. The findings suggest that monthly premiums, marital status, vehicle class, income, coverage, and location may contribute to customers' lifetime value.

Key words: Auto-insurance companies, Marketing strategies, CLV (customers' lifetime value)

1 Introduction

Under the shadow of the pandemic, Americans worked from home instead of commuting to work. Patrick T. Fallon (2023) stated that employees save time, but auto insurance firms suffer significant losses as a result. According to an S&P Global Market Intelligence investigation in

[*] Engineering Department (Industrial Engineering), California State Polytechnic University, Pomona, Los Angeles, 91766, USA

[†] Corresponding author. Email: zliao@cpp.edu

[‡] Engineering Department (Information Management of Business), University College London, London, WC1E 6BT, UK

[§] Department of Mathematics, University of California of San Diego, La Jolla, 92093, USA

2023, the private vehicle insurance market in the United States experienced its poorest underwriting loss in more than 20 years in 2022.

To achieve effective recovery, insurance companies must create efficient strategies that assist their executives in establishing organizational objectives, providing businesses with a competitive edge, and allocating resources. Since the insurance industry is a service sector, these strategies should also address how businesses should interact with various types of clients. Insurance firms should be aware of customers' profitability; in other words, they must understand their priorities in order to develop and justify appropriate marketing initiatives.

To determine the marketing initiatives for an individual customer, there is a useful value called lifetime value. In an article written by Caldwell (2022), it is shown that Customer Lifetime Value (CLV) is a statistic used to determine the amount of money a company can expect to earn from a typical customer throughout the duration of their relationship with the company. Kumar, Ramani, and Bohling (2004) explained that businesses intend to calculate the lifetime value of each customer and use this data to develop distinctive marketing campaigns tailored to each individual. Therefore, the first research question pertains to the traits of customers with higher lifetime values, while the second study question involves finding out how these characteristics might impact the lifetime value of consumers.

RQ1: Which traits of auto insurance customers would affect the lifetime value?

RQ2: Through what mechanisms do these traits affect the lifetime value?

According to the research questions, we formulated hypotheses based on the definition of CLV and other materials:

H1: There is a relationship between the customer's driving location and their lifetime value. The more risky the location, the lower the lifetime value.

The risk of driving is one of the factors presumptively associated with lifetime value. According to SAS (2018), CLV takes into account the difference between total customer revenues and total customer expenses throughout the entire business relationship. Claims represent the costs that clients of auto insurance companies might incur. Therefore, assuming that the premium remains at the same level, the lifetime value of the customer reduces when there is a higher potential for claims while driving. This suggests that there is a negative correlation between lifetime value and driving risks.

Driving in a rural area is significantly riskier than driving in an urban area, according to a study by psychologists Ilan Shrira of Arkansas Tech University and Kenji Noguchi of the University of Southern Mississippi (2016). This study demonstrates how closely the location of driving affects the risks associated with driving. Hence, there is a strong likelihood that the driving location is related to the lifetime value.

H2: There is a positive relationship between income and lifetime value. The higher the customer's income, the higher their lifetime value.

In line with the SAS definition of CLV, income can also influence lifetime value. When income increases, disposable income (net income) also increases correspondingly. Assuming other variables remain constant, the proportion of income that can be spent on insurance also increases, leading to a higher lifetime value for the customer.

H3: The class of the vehicle influences the lifetime value. The higher the vehicle's class, the higher the customer's lifetime value.

It's a well-known fact that a vehicle's value increases with its level of luxury. After examining three different websites providing car insurance quotes, such as comparethemarket, we found that the estimated automobile value is a crucial factor. As the value of the car increases, the predicted premium also increases. When the risk of needing to make claims is reduced, the lifetime value of the client increases due to the higher premium that their car commands. To support our hypothesis and identify more potential variables related to lifetime value, we analyzed a dataset containing information for 9135 auto insurance customers with nine variables.

2 Methodology

This study utilized data from "Kaggle Jenks Natural Breaks and K-means Clustering" (2022) to investigate the characteristics of clients exhibiting a higher lifetime value. We examined a dataset comprising 9135 auto insurance customers across nine variables, establishing comparisons between each variable and the lifetime value. Subsequently, the researchers sought correlations between each variable and lifetime value, assessing the significance of these associations. Data processing was conducted using R Studio software, employing four statistical models to validate the significance of our regression model and confirm our hypotheses.

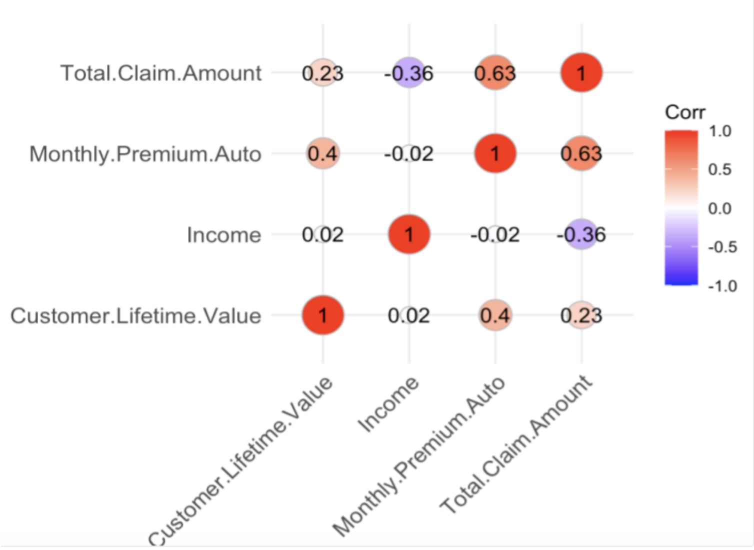

Figure 1: The correlation table

As shown in Figure 1, the final findings suggest a negligible correlation between Lifetime.value and Income. Notably, a positive correlation emerges with Total.Claim.Amount, with a likewise significant positive correlation discernible with Monthly.Premium.Auto. Furthermore, the associations between Lifetime.value and other factor-based variables are visually represented.

Then, we show the descriptive statistics for all numerical variables. The results are exhibited in Table 1.

Table 1. descriptive statistics

	Number of observations	Mean	Standard deviation	Maximum	Minimum
Customer.Lifetime.Value	9134	8004.94	6870.97	83325.38	1898.01
Income	9134	37657.38	30379.90	99981.00	0.00
Monthly.Premium.Auto	9134	93.22	34.41	298.00	61.00
Total.Claim.Amount	9134	434.09	290.50	2893.24	0.10

From the result, the mean value of the lifetime value is 8004.94, this is much lower than the maximum value (83325.38). It means most of the lifetime value of this variable is lower than its mean value. Regarding the "Income" variable, the average income of the customers is 37657.38. The standard deviation (30379.90) is slightly lower than its mean value. This means that the wide gap between rich and poor customers. In addition, the average monthly premium paid by the customers is 93.22, and the average claim amount made by customers is 434.09.

2.1 Coverge

In the context of this research, the interplay between "Coverage" and "Customer.Lifetime.Value" is probed by employing box plot visualizations. Through the use of the "ggplot2" library, the ascendant trend in Lifetime Value is depicted, progressing from Basic to Extended and culminating in Premium levels of Coverage. To robustly ascertain the influence of Coverage on Lifetime Value, a t-test is conducted. The dataset is divided into three factions predicated on the Coverage level: Basic, Extended, and Premium. As shown in Figure 2, results from the t-test divulge a statistically significant disparity in Lifetime Value between both Extended and Basic Coverage, and Premium and Extended Coverage, with p-values falling below 0.01. This outcome significantly refutes the initial hypothesis and corroborates that both Extended and Premium Coverage tiers exhibit higher Lifetime Values in comparison to Basic Coverage.

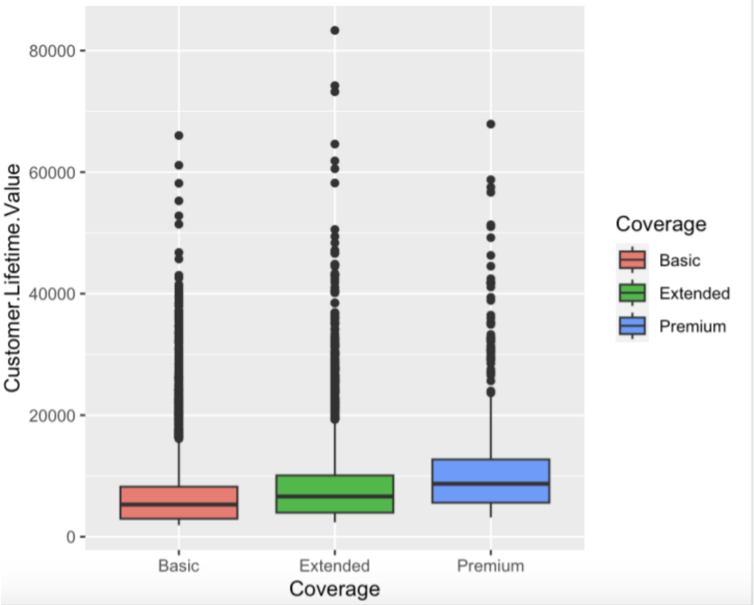

Figure 2: Plot for Customer Lifetime. Value & Coverage

The other quantitative variables are analyzed in a similar way.

2.2 Gender

In the context of this investigation, the prospective association between "Gender" and "Customer Lifetime Value" is scrutinized by leveraging the "ggplot2" library, through box plot visualizations. The diagram elucidates the distribution of "Customer Lifetime Value" across two gender classifications, female (F) and male (M). As shown in Figure 3, from an initial visual inspection, gender does not appear to exert a substantial influence on variations in Lifetime Value. To substantiate this preliminary observation, a t-test was performed on the Lifetime Value of the female and male cohorts. The outcomes display a p-value exceeding 0.1, suggesting that the initial hypothesis proposing a gender effect on lifetime values is statistically untenable. Consequently, a deduction suggests that gender does not appear to be linked to fluctuations in lifetime value within this dataset.

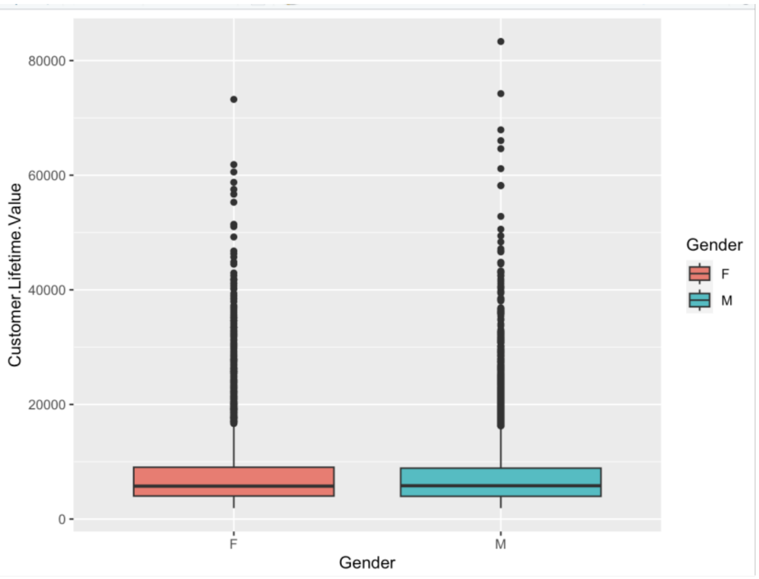

Figure 3: Plot for Customer Lifetime. Value & Gender

2.3 Location.Code

In this study, "ggplot2" box plot is used to elucidate the possible influence of "Location.Code" on "Customer Lifetime Value". The plot demonstrates the dispersion of "Customer Lifetime Value" across distinct Location.Codes, encompassing rural, suburban, and urban areas. As shown in Figure 4, from an initial visual inspection, Location.Code appears to exert minimal effect on the variance in Lifetime Value. To rigorously scrutinize this association, multiple t-tests is implemented to compare lifetime values across the three Location.Code groups. The outcome revealed p-values exceeding 0.1 for all tests, signifying insufficient statistical evidence to refute the initial hypothesis. Thus, the conclusion is that Location.Code does not exert a significant influence on Customer Lifetime Value within this dataset.

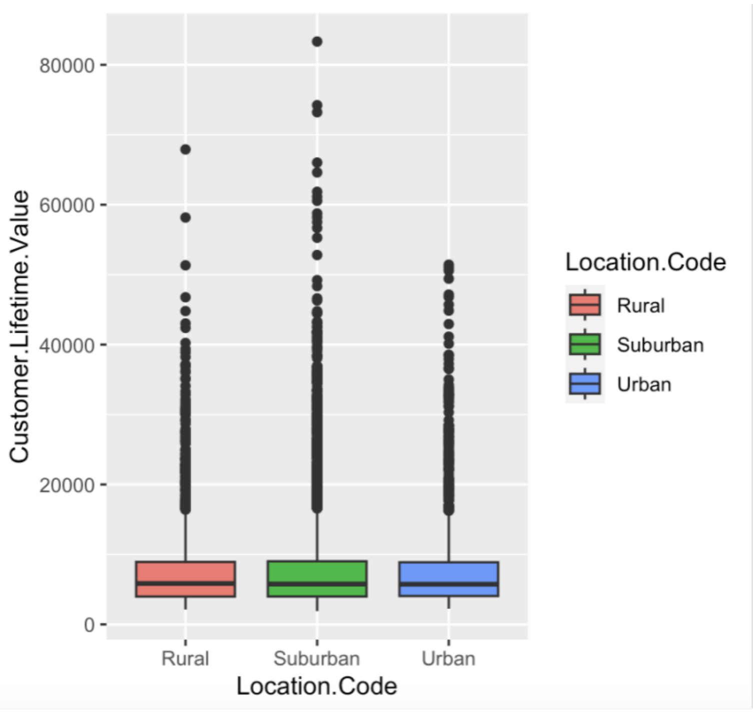

Figure 4: Plot for Customer Lifetime. Value & Location. Code

2.4 Marital.Status

In the course of this investigation, the potential impact of "Marital.Status" on "Customer.Lifetime.Value," is shown utilizing "ggplot2" to generate boxplot visualizations. The plot exhibits the distribution of Lifetime Values across varying Marital Status categories, namely Divorced, Married, and Single. Upon visual examination, it becomes apparent that "Marital.Status" does not significantly sway variations in Lifetime Values. To perform an exhaustive analysis, multiple t-tests is executed, comparing Lifetime Values among the different Marital Status cohorts.

As shown in Figure 5, the results of the t-tests offer intriguing insights: the juxtaposition of Divorced and Married groups returns a p-value surpassing 0.1, suggesting that the original hypothesis of a notable disparity is not upheld. However, the comparison between the divorced and single groups yields a p-value below 0.05, compelling us to reject the original hypothesis and intimating a significant discrepancy in Lifetime Values between these groups. Additionally, the comparison between Married and Single groups yields a p-value less than 0.05, once again indicating a significant difference in Lifetime Values. These findings suggest that Single individuals typically possess lower Lifetime Values compared to those who have been Married at least once. However, no significant difference is observed between the Lifetime Values of the divorced and married groups. This underscores the importance of considering Marital Status as a potential influencer of Customer Lifetime Values within the dataset.

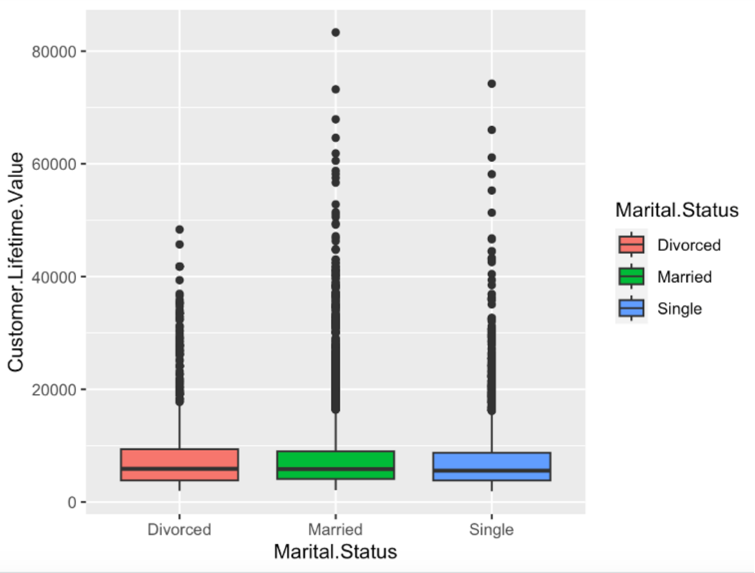

Figure 5: Plot for Customer Lifetime. Value & Marital. Status

2.5 Sales.Channel

This study investigates the potential influence of various sales channels (Agent, Branch, Call Center, and Web) on Customer Lifetime Value (CLV), employing boxplot visualizations constructed via the ggplot2 library. These visualizations depict the distribution of CLV across the sales channel categories, including Agent, Branch, Call Center, and Web.

As shown in Figure 6, the preliminary visual analysis, however, reveals no significant correlation between the choice of "Sales.Channel" and the variability in Lifetime Values. This initial observation necessitates a deeper examination, hence a sequence of t-tests, contrasting the Lifetime Values among diverse sales channel groups. Remarkably, the t-tests systematically present p-values exceeding 0.1 for all pairings, implying insufficient statistical evidence to challenge the primary assumption. Consequently, it is inferred that the selection of a sales channel has no detectable impact on the Customer.Lifetime.Value in this data set.

These insights underscore the notion that the type of sales channel, be it Agent, Branch, Call Center, or Web, does not significantly influence the determination of CLV. It reinforces the premise of the sales channel's independence in relation to CLV variability, suggesting that other parameters may exercise a more pronounced effect in shaping customer behaviors within this dataset's context.

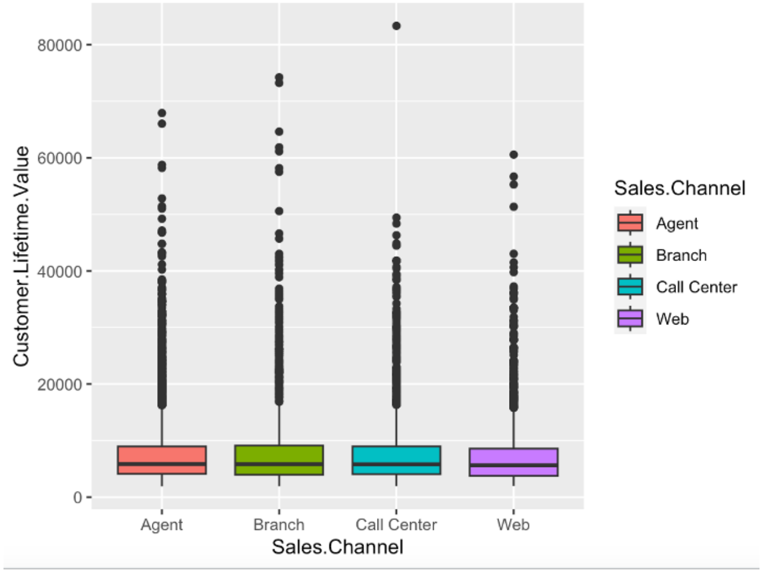

Figure 6: Plot for Customer Lifetime. Value & Sales. Channel

2.6 Vehicle.Class

This investigation explores the effect of vehicle classes (Two-Door Car, Four-Door Car, Sports Car, SUV, Luxury Car, and Luxury SUV) on the CLV, using boxplot visualizations generated through ggplot2.

As shown in Figure 7, upon visual inspection, a distinct correlation emerges between vehicle class and CLV variance. To validate this observation, a series of t-tests are executed comparing the CLV among different vehicle classes.

The t-tests expose fascinating results: While the comparisons between Two-Door Car, Four-Door Car, and SUV with other vehicle classes manifest p-values greater than 0.1, contrasts between Two-Door Car, Four-Door Car, and classes such as Sports Car, Luxury Car, and Luxury SUV yield p-values less than 0.01. These results compel the rejection of the initial hypothesis.

Conclusively, vehicle class appears to segregate CLV into three distinct tiers: the lowest tier correlates with Two-Door Car and Four-Door Car categories, followed by Sports Car and SUV. The highest tier is dominated by Luxury Car and Luxury SUV categories, implying a higher CLV among customers driving these vehicles. This data highlights the significant role vehicle class plays in analyzing CLV, as various vehicle classes exert diverse effects on customer lifetime values.

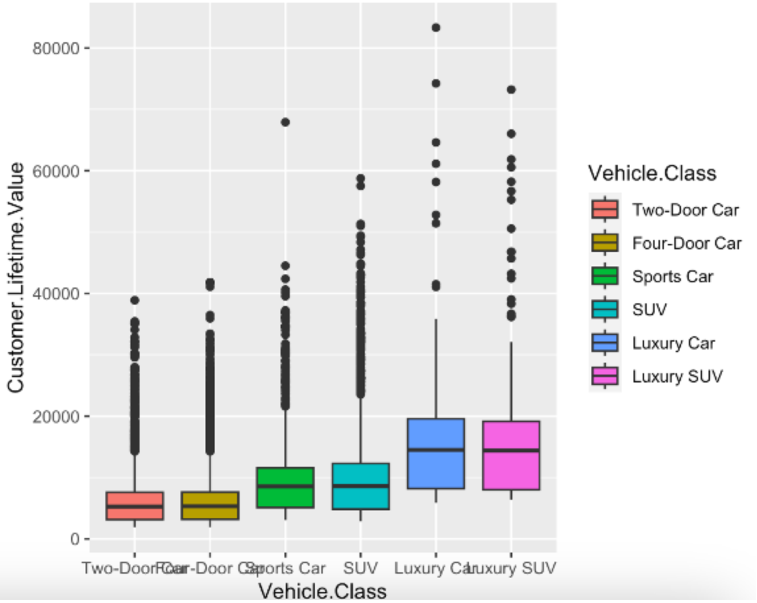

Figure 7: Plot for Customer Lifetime. Value & Vehicle. Class

Under T test:

In addition to the box plots above, we additionally applied the t-test to determine whether Custome Lifetime Value is dependent on these variables. The t test results are reported in Table 2.

Table 2. T test result

Variable	Null hypothesis	T Statistic
Coverage	mean(Y|Basic) = mean(Y|Extended)	-9.75^{***}
	mean(Y|Extend) = mean(Y|Premium)	-6.49^{***}
Gender	mean(Y|Male) = mean(Y|Male)	1.30
Location.Code	mean(Y|Rural) = mean(Y|Suburban)	-0.28
	mean(Y|Suburban) = mean(Y|Urban)	-0.31
	mean(Y|Rural) = mean(Y|Urban)	-0.47
Marital.Status	mean(Y|Divorced) = mean(Y|Married)	0.77
	mean(Y|Divorced) = mean(Y|Single)	2.27^{**}
	mean(Y|Married) = mean(Y|Single)	2.20^{**}
Sales.Channel	mean(Y|Agent) = mean(Y|Branch)	-0.90
	mean(Y|Agent) = mean(Y|Call Center)	-0.70
	mean(Y|Agent) = mean(Y|Web)	0.82
	mean(Y|Branch) = mean(Y|Call Center)	0.09
	mean(Y|Branch) = mean(Y|Web)	1.46
	mean(Y|Call Center) = mean(Y|Web)	1.27
Vehicle.Class	mean(Y|Two-Door Car) = mean(Y|Four-Door Car)	0.28
	mean(Y|Two-Door Car) = mean(Y|Sports Car)	-10.13^{***}
	mean(Y|Two-Door Car) = mean(Y|SUV)	-17.00^{***}
	mean(Y|Two-Door Car) = mean(Y|Luxury Car)	-10.49^{***}
	mean(Y|Two-Door Car) = mean(Y|Luxury SUV)	-11.10^{***}
	mean(Y|Four-Door Car) = mean(Y|Sports Car)	-10.51^{***}
	mean(Y|Four-Door Car) = mean(Y|SUV)	-18.85^{***}
	mean(Y|Four-Door Car) = mean(Y|Luxury Car)	-10.58^{***}
	mean(Y|Four-Door Car) = mean(Y|Luxury SUV)	-11.19^{***}
	mean(Y|Sports Car) = mean(Y|SUV)	0.72
	mean(Y|Sports Car) = mean(Y|Luxury Car)	-5.97^{***}
	mean(Y|Sports Car) = mean(Y|Luxury SUV)	-6.31^{***}
	mean(Y|SUV) = mean(Y|Luxury Car)	-6.61^{***}
	mean(Y|SUV) = mean(Y|Luxury SUV)	-7.01^{***}
	mean(Y|Luxury Car) = mean(Y|Luxury SUV)	-0.05

Note: * represent 10% significance level, ** represent 5% significance level; *** represent 1% significance level. Y is the ‘Custome Lifetime Value’. When t statistic > 0, it means that mean(Y|situation 1) > mean(Y|situation 2).

In the Table 1 above, when the t statistics is significant, we reject the corresponding null hypothesis and then the variable is going to impact the Custome.Lifetime.Value. Here, we find that the Custome Lifetime Value has nothing to do with variables “Gender”, “Location.Code”, “Sales.Channel”. It depends on “Coverage”, “Marital.Status” and “Vehicle.Class”. In addition, we can get several interesting results. First, people who have a higher degree of Coverage possess higher lifetime value; Second, people who are married or have been married before have higher lifetime values than those who are single. At last, Vehicle.Class categorizes Lifetime Value into three classes. "Two-Door Car" and "Four-Door Car" have the lowest lifetime value, followed by "Sports Car" and "SUV", and the highest lifetime value are "Luxury Car" and "Luxury SUV".

3 Simple regression

Simple regression analyses were undertaken to elucidate the relationship between Customer Lifetime Value (CLV) and several independent variables. The regression result is shown in Table 3.

Table 3: Simple regression result

	Customer.Lifetime.Value								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Coverage	1,598.971***								
(Extended)	(158.021)								
Coverage	3,704.897***								
(Premium)	(252.816)								
Gender		-187.051							
(Male)		(143.809)							
Income			0.006**						
			(0.002)						
Location.Code				50.758					
(Suburban)									

				(186.557)					
Location.Code				110.434					
(Urban)				(237.656)					
Marital.Status					-162.272				
(Married)					(208.264)				
Marital.Status					-526.402^{**}				
(Single)					(231.505)				
Monthly.Premium.Auto						79.130^{***}			
						(1.919)			
Sales.Channel							162.003		
(Branch)							(178.802)		
Sales.Channel							142.376		
(Call Center)							(200.817)		
Sales.Channel							-177.921		
(Web)							(221.834)		
Total.Claim.Amount								5.356^{***}	
								(0.241)	
Vehicle.Class									$10,421.620^{***}$
(Luxury Car)									(511.580)
Vehicle.Class(Luxury									$10,491.270^{***}$
SUV)									(482.558)
Vehicle.Class									$4,119.263^{***}$
(Sports Car)									(306.681)
Vehicle.Class									$3,811.785^{***}$
(SUV)									(178.495)
Vehicle.Class									39.304
(Two-Door Car)									(175.401)
Constant	$7,190.706^{***}$	$8,096.602^{***}$	$7,797.421^{***}$	$7,953.699^{***}$	$8,241.239^{***}$	628.500^{***}	$7,957.709^{***}$	$5,679.933^{***}$	$6,631.727^{***}$
	(90.771)	(100.670)	(114.475)	(163.195)	(185.655)	(190.643)	(116.526)	(125.919)	(94.430)
Observations	9,134	9,134	9,134	9,134	9,134	9,134	9,134	9,134	9,134

From the Table 3, we can get the following information. The analytical observations revealed that among the examined variables, Coverage emerged as a significant predictor of the dependent variable. Conversely, Gender, as an independent variable, demonstrated consistent insignificance concerning the CLV. Both Income and Monthly Premium Auto were confirmed as significant predictors, whereas Location Code and Sales Channel failed to achieve any significance in their respective regressions. The regression involving Marital Status suggested that the Single category was significantly associated with the dependent variable. Similarly, Vehicle Class was discerned as a significant predictor. In summary, Income, Marital Status, Monthly Premium Auto, and Vehicle Class were observed to share significant relationships with CLV, in contrast to Gender, Location Code, and Sales Channel, which exhibited no such correlations.

4 Multiple Linear Regression

A multivariate linear regression was executed to evaluate the collective impact of diverse variables on the Customer Lifetime Value. The result is shown in Table 4.

Table 4: Multiple Linear Regression result

	Customer.Lifetime.Value					
	(1)	(2)	(3)	(4)	(5)	(6)
Coverage	121.491	121.573				
(Extended)	(249.561)	(249.543)				
Coverage	179.727	168.693				
(Premium)	(527.771)	(527.686)				
Gender	-177.346	-185.203	-184.452	-180.969	-196.145	
(Male)	(132.859)	(132.712)	(132.663)	(132.626)	(132.107)	
Income	0.004	0.005*	0.005*	0.005*	0.006**	0.006**
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Location.Code	-69.115					
(Suburban)	(257.762)					
Location.Code	185.358					
(Urban)	(241.095)					
Marital.Status	-259.392	-248.558	-247.778	-242.889	-237.397	-235.909
(Married)	(191.718)	(191.511)	(191.443)	(191.410)	(191.369)	(191.379)
Marital.Status	−490.744**	−483.825**	−482.776**	−483.563**	−524.387**	−532.486**
(Single)	(220.046)	(218.916)	(218.843)	(218.821)	(216.525)	(216.470)
Monthly.Premium.Auto	70.588***	70.991***	74.459***	74.526***	72.284***	72.432***
	(10.026)	(9.906)	(4.346)	(4.345)	(3.982)	(3.981)
Sales.Channel	184.866	185.077	184.564			
(Branch)						

	(164.222)	(164.200)	(164.177)			
Sales.Channel	220.589	219.732	217.555			
(Call Center)	(184.439)	(184.430)	(184.360)			
Sales.Channel	-126.205	-124.866	-127.474			
(Web)	(203.713)	(203.687)	(203.558)			
Total.Claim.Amount	-0.320	-0.428	-0.433	-0.435		
	(0.468)	(0.338)	(0.338)	(0.338)		
Vehicle.Class	1,209.726	1,215.593	759.430	769.504	736.242	705.240
(Luxury Car)	(1,386.213)	(1,386.148)	(735.873)	(735.653)	(735.227)	(734.979)
Vehicle.Class	1,199.025	1,205.723	752.218	717.400	704.756	671.409
(Luxury SUV)	(1,373.969)	(1,373.930)	(719.286)	(719.046)	(719.005)	(718.701)
Vehicle.Class	1, 083.182**	1, 081.758**	928.901***	919.029***	932.540***	922.018***
(Sports Car)	(526.230)	(526.210)	(349.540)	(349.494)	(349.349)	(349.300)
Vehicle.Class	871.506*	876.100*	726.432***	730.618***	732.586***	724.670***
(Sports Car)	(456.911)	(456.852)	(244.430)	(244.409)	(244.413)	(244.371)
Vehicle.Class	76.736	74.752	73.677	82.432	80.711	75.375
(Two-Door Car)	(172.418)	(172.407)	(172.370)	(172.311)	(172.312)	(172.286)
Constant	1, 381.393*	1, 347.179*	1, 134.658***	1, 199.153***	1, 188.087***	1, 083.659***
	(726.348)	(693.978)	(386.883)	(376.936)	(376.852)	(370.255)
Observations	9,134	9,134	9,134	9,134	9,134	9,134
Adjusted R^2	0.159	0.159	0.159	0.159	0.159	0.159

From the Table 4, the preliminary regression embraced all variables, yielding an Adjusted R-squared value of 0.159. This metric offers an estimate of the model's capability to replicate observed outcomes, where a value ranging from 0 to 1 denotes the fraction of total variation 'explained' by the model.

To further optimize the model, a stepwise regression technique was employed. This procedure entails the fitting of regression models by systematically adding or eliminating predictors based on their statistical significance. Firstly, the variable "Location.Code", possessing the highest p-value of 0.7886, was excluded from the model. This action was underpinned by the assumption that an elevated p-value signifies a potential lack of significance in the variable, particularly within the context of the other variables. Secondly, following the elimination of "Location.Code", "Coverage" was identified as the variable with the subsequent highest p-value (0.7492). This indicated the variable's prospective insignificance, and thus, it was extricated from the regression model. Thirdly, after the removal of "Coverage", all p-values linked to "Sales Channel" exhibited insignificance. Consequently, "Sales Channel" was the ensuing variable to be purged. Fourthly, "Total.Claim.Amount" was determined as an insignificant variable and, hence, was excised from the regression equation. Fifthly: "Gender" surfaced as the variable with the highest p-value in the updated model, hinting that it might not constitute a significant predictor in the presence of other variables. Thus, it was successively removed.

After the stepwise elimination process, the variables retained in the regression model were "Income", "Marital.Status", "Monthly.Premium.Auto", and "Vehicle.Class". These four variables were suggested to be significant predictors of Customer Lifetime Value within the context of the existing dataset. This optimized model offers a more parsimonious and potentially comprehensible interpretation of the relationship between the predictors and the outcome.

5 constructing logistic regression

Within the analysis, logistic regression is leveraged to predict a binary outcome on the basis of multiple predictors. The binary outcome in this scenario signifies whether the Customer Lifetime Value surpasses or falls below its median value.

Initial Preprocessing:

- The dataset is imported, and a novel variable, "LifeT", is established. This variable is assigned the value "1" if a customer's Lifetime Value equals or exceeds the median value, and "0" otherwise.
- Two columns, "Customer.Lifetime.Value" and "Customer", are removed from the dataset to evade redundancy and potential multicollinearity.

Then, we perform the logistic regression model. The result is reported in Table 5.

Univariate Logistic Regression:

Prior to progressing with a comprehensive model, each predictor was individually evaluated against the dependent variable "LifeT" to comprehend their individual significance. The result is reported in Table 5.

Table 5: Univariate Logistic Regression result

	LifeT								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Coverage	0.894***								
(Extended)	(0.048)								
Coverage	1.408***								
(Premium)	(0.084)								
Gender		0.038							
(Male)		(0.042)							
Income			0.000*						
			(0.000)						
Location.Code				-0.033					
(Suburban)				(0.054)					
Location.Code				-0.043					
(Urban)				(0.069)					
Marital.Status					-0.021				
(Married)					(0.061)				
Marital.Status					−0.125*				
(Single)					(0.067)				
Monthly.Premium.Auto						0.032***			
						(0.001)			
Sales.Channel							-0.012		
(Branch)							(0.052)		
Sales.Channel							-0.021		
(Call Center)							(0.058)		
Sales.Channel							-0.087		
(Web)							(0.065)		
Total.Claim.Amount								0.002***	
								(0.0001)	
Vehicle.Class									16.914
(Luxury Car)									(187.947)
Vehicle.Class									16.914
(Luxury SUV)									(176.897)
Vehicle.Class									1.177***
(Sports Car)									

									(0.103)
Vehicle.Class (SUV)									1.035^{***} (0.058)
Vehicle.Class (Two-Door Car)									-0.002 (0.055)
Constant	-0.385^{***}	-0.016	-0.048	0.030	0.048	-2.860^{***}	0.022	-0.673^{***}	-0.348^{***}
	(0.027)	(0.029)	(0.033)	(0.048)	(0.054)	(0.090)	(0.034)	(0.041)	(0.030)
Observations	9,134	9,134	9,134	9,134	9,134	9,134	9,134	9,134	9,134

From the Table 5, we find two interesting conclusions:

- "Coverage", "Income", "Marital.Status", "Monthly.Premium.Auto", "Total.Claim.Amount", and "Vehicle.Class" were all identified as significant predictors.
- In their individual regressions, "Gender", "Location.Code", and "Sales.Channel" did not display significant impacts.

Multiple Logistic Regression:

Then, we use a similar way as the multiple OLS regression to perform multiple logistic regression here. The result is shown in Table 6.

Table 6: Multiple Logistic Regression result

	LifeT					
	(1)	(2)	(3)	(4)	(5)	(6)
Coverage (Extended)	0.494^{***} (0.096)	0.495^{***} (0.095)	0.397^{***} (0.053)	0.397^{***} (0.053)	0.396^{***} (0.053)	0.394^{***} (0.053)
Coverage (Premium)	0.441^{**} (0.198)	0.441^{**} (0.198)	0.226^{**} (0.096)	0.226^{**} (0.096)	0.226^{**} (0.096)	0.230^{**} (0.096)
Gender (Male)	0.053 (0.046)	0.054 (0.046)	0.058 (0.046)	0.058 (0.046)		
Income	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	
Location.Code (Suburban)	-0.095	-0.093	-0.110	-0.118^{*}	-0.112^{*}	-0.142^{**}

	(0.094)	(0.094)	(0.093)	(0.064)	(0.064)	(0.060)
Location.Code	-0.025	-0.024	-0.035	-0.040	-0.040	-0.041
(Urban)	(0.084)	(0.084)	(0.084)	(0.075)	(0.075)	(0.075)
Marital.Status	-0.060	-0.060	-0.059	-0.059	-0.059	-0.058
(Married)	(0.066)	(0.066)	(0.066)	(0.066)	(0.066)	(0.066)
Marital.Status	−0.152**	−0.151**	−0.149*	−0.150**	−0.148**	−0.164**
(Single)	(0.076)	(0.076)	(0.076)	(0.075)	(0.075)	(0.074)
Monthly.Premium.Auto	0.024***	0.024***	0.029***	0.029***	0.029***	0.029***
	(0.004)	(0.004)	(0.001)	(0.001)	(0.001)	(0.001)
Sales.Channel	-0.015					
(Branch)	(0.057)					
Sales.Channel	0.002					
(Call Center)	(0.064)					
Sales.Channel	-0.066					
(Web)	(0.071)					
Total.Claim.Amount	-0.0001	-0.0001	-0.00002			
	(0.0002)	(0.0002)	(0.0002)			
Vehicle.Class	13.891	13.887				
(Luxury Car)	(172.954)	(173.026)				
Vehicle.Class	13.809	13.806				
(Luxury SUV)	(162.969)	(162.969)				
Vehicle.Class	0.224	0.222				
(Sports Car)	(0.205)	(0.205)				
Vehicle.Class	0.111	0.111				
(SUV)	(0.182)	(0.182)				
Vehicle.Class	0.019	0.020				
(Two-Door Car)	(0.058)	(0.058)				
Constant	−2.336***	−2.353***	−2.693***	−2.687***	−2.663***	−2.595***
	(0.296)	(0.294)	(0.137)	(0.126)	(0.125)	(0.114)
Observations	9,134	9,134	9,134	9,134	9,134	9,134

From the Table 6, we can get 7 interesting results.

1. Commencing with a comprehensive logistic regression inclusive of all predictors, the derived model was statistically significant. The ensuing stepwise refinement aimed to eradicate potentially extraneous predictors.

2. Initially, the "Sales.Channel" variable was pruned, ascribed to its highest p-value suggesting the least substantive predictor among the assemblage.

3. Following this, "Vehicle.Class", with a diminished level of significance, was eliminated.

4. Subsequently, the "Total.Claim.Amount" was abandoned, thereby streamlining the model further.

5. Following it, "Gender" showed the highest p-value and was excluded from the model.

6. Ultimately, the "Income" variable, despite its significance in individual regression, manifested diminished importance in the multivariate context and was therefore excised.

7. After this systematic elimination, the refined model retained "Coverage", "Location.Code", "Marital.Status", and "Monthly.Premium.Auto" as its integral variables.

6 Results

The results concluded from the data analysis indicate that marriage status, monthly premium amount, and vehicle class are significantly correlated with the lifetime value of auto insurance customers in all analysis models. However, income is significant in most models except in logistic regression. As for coverage and location code, they emerge as significant only in the logistic regression. The reasons for these outcomes warrant further research. First and foremost, the monthly premium amount has the highest correlation with lifetime value. This is because customer lifetime value (CLV) is closely related to the value that a customer brings to the organization. In the insurance industry, the value the customer provides is the premium. Premium, which is an indicator of customer value, is also included in the traditional formula for calculating CLV, as mentioned by Caldwell (2022):

$$CLV = \text{Customer Value} \times \text{Average Customer Lifespan} \tag{1}$$

Next, the consistent significance of marital status across models indicates that marital status plays a crucial role in determining CLV. This could be because marital status might be associated with financial stability, purchasing patterns, or risk behavior, which, in turn, affects insurance premiums or claims. There is also a possibility that this data may be biased and not randomly chosen, as the company investigated may be more focused on young clients.

Vehicle class can be a reflection of lifestyle, financial status, and even risky behavior. More directly, it is an indicator of the value of the vehicle, which can impact insurance quotes, as mentioned in the introduction. The significant association suggests varying CLVs for different vehicle categories, possibly due to differences in premiums, claim frequency, or claim amounts.

In the majority of models, except for logistic regression, income is considered important because it is linked to other variables. Income is connected to lifetime value when tested on its own. However, the results of logistic regression would be invalid if income and other variables related to income were tested alongside lifetime value. Income is a key factor in determining a person's purchasing power and financial behavior, and variables like vehicle class have a positive link with income, according to Team, T.I. (2023). The higher the income, the higher the class of car.

As for coverage and location code, which only show significance in logistic regression, it's possible that when modeling the probability of CLV being above or below a median (as is the case in logistic regression), the type of coverage a customer has and their location become pivotal determinants. Perhaps certain coverages or locations are associated with significantly higher or lower lifetime values.

In summary, the conducted analysis illuminates how a combination of socio-economic factors (such as marital status and income), product-specific variables (like monthly premium and coverage), and demographic factors (such as vehicle class and location) influence the determination of Customer Lifetime Value. The consistent emergence of specific variables highlights their strategic value, guiding enterprises to concentrate their efforts on these areas when strategizing for customer retention and value optimization.

7 Discussion

After analyzing the results, it is evident that six out of nine variables are related to lifetime value, which can be divided into three different groups for better optimizing CLV. From a socio-economic perspective, the company should focus more on single customers. When developing marketing strategies, the company should consider their emotional state and needs. For instance, a single person might be willing to pay extra for auto insurance because they feel uneasy without a secure financial status. The business can leverage this uncertainty and highlight how their solution helps people secure funding following an accident. A call to action and personalized messaging should be used in the company's direct marketing efforts, especially for the high-income segment of clients, as they have a higher potential lifetime value, as indicated by Ugenti (2023). Phone interviews could be employed in direct marketing to gather feedback and make customers feel valued.

From a product-specific perspective, the business needs to concentrate more on customers with premium coverage who pay a higher monthly premium. If consumers with higher incomes require more regular stimulation, premium customers are the ones the business needs to retain for longer. Regular calls, emails, and direct marketing are useful for keeping them engaged. Customer feedback can be used as an opportunity to interact, educate, and build trust with clients. Care should be taken not to increase premiums too drastically, as it might lead to a loss of interest from customers whose willingness to pay is lower.

From a demographic standpoint, the business should pay more attention to clients who drive high-class vehicles (luxury SUVs or luxury cars) or live in rural areas. In less developed rural areas, businesses may promote themselves more on TV or billboards along the road through advertisements. For customers with high-end vehicles, auto insurance companies can use direct marketing to inform customers about the value of auto insurance products and the maintenance costs of luxury cars without insurance. This can be done on a regular but not too frequent basis.

There are numerous strategies to increase the company's revenues, such as increasing insurance prices or reducing the number of claims, but each carries the risk of losing clients. Vehicle insurance businesses must always be aware of their clients' demands, allocate more resources to clients with higher lifetime values, and employ reasonable marketing strategies to maximize clients' potential value and attract new clients.

8 Conclusion

In summary, the analysis of the dataset concludes that marriage status, monthly premium amount, and vehicle class are strongly associated with the customers' lifetime value of the chosen auto insurance. During the process of generating ideas and conducting modeling, some challenges were encountered. One challenge was the inability to determine whether the data pertains to all auto insurance companies or a specific one. Consequently, it is assumed that it belongs to a specific company for the purpose of this article. In the future, researchers could continue to investigate datasets from other auto insurance companies to enhance the study's accuracy, collecting different indicators to see if similar conclusions are reached. Furthermore, new possibilities, such as conducting surveys via questionnaires with companies, could help explore the factors that genuinely influence customers' lifetime value. With these insights, companies can implement data-driven strategies for client retention, personalized products, and optimized CLV, ultimately supporting long-term growth and profitability in the auto insurance sector.

References

[1] Patrick t. fallon/Agence F.-P. (2023) Car insurance rates are going up again, The Wall Street Journal.

[2] Admin (2023) Auto insurance results 'historically bad' in 2022: S&P, Insurance Journal.

[3] Caldwell, A. (2022) How to calculate customer lifetime value, Oracle NetSuite.

[4] Kumar, V., Ramani, G. and Bohling, T. (2004) 'Customer lifetime value approaches and best practice applications', Journal of Interactive Marketing, 18(3), pp. 60–72. doi:10.1002/dir.20014.

[5] Seyerle, M. (2018) Customer Lifetime Value and its determination using the SAS Enterprise Miner and the SAS OROS-Software, Papers and Presentations.

[6] Shrira, I., & Noguchi, K. (2016). Traffic fatalities of drivers who visit urban and rural areas: An exploratory study. Transportation Research Part F, 41, 74–79. doi: 10.1016/j.trf.2016.05.003

[7] Sanngrahi, M. (2022) ‘Jenks Natural Breaks and K-means Clustering’.

[8] Team, T.I. (2023) What is the income effect? its meaning and example, Investopedia.

[9] Ugenti, M. (2023) Council post: The evolution of direct marketing, Forbes.