Promoting sustainability in Auto-Insurance Business

Contents

1.0. Introduction

The increasing number of cars in the last decade has a severe impact on the size of pollution. There are many businesses related to cars, including car insurance. this research focuses on the insurance business of cars.

The motivation of the research is to promote sustainability in auto-insurance business by taking advantage of customer lifetime value (CLV) that is included in the customers data.

The company under study is a car insurance company whose data is available on Kaggle [1]. The company provides insurance services for private and commercial cars with three types of insurance policies:

a- Basic Coverage: This type of car insurance policy has a minimum coverage level, specifically, it only covers any damage that you or your car might cause on other people or cars, including the medical expenses. The downside is that it does not cover the damage you have done to your own car.
b- Extended Coverage: liability insurance as well as some extra coverage, protecting individuals in instances of damage (caused by natural disasters such as earthquakes or floods), fire, as well as theft (excluding vandalism). In addition to the aforementioned, a partial car insurance policy will also cover broken glass, wiring damages.
c- premium Coverage: covers largely anything there is to cover. It is the most substantial level of insurance, and it includes both liability as well as partial insurance, and more (such as acts of vandalism). This type of vehicle insurance covers damages done to your own car, through an accident inflected by yourself. Considering this type of car insurance is the most comprehensive, it is also the most expensive.

1.1. RESEARCH QUESTION

The main research question of this study is:

How to retain the loyal current customers in addition to that increasing the insurance policies for green cars?

Hypothesis: Auto-insurance is a good business, especially since many countries force compulsory insurance. The study includes how to keep the customer while giving priority to the customer who supports sustainability by providing attractive offers known as green discounts. The green discounts are insurance offers for electric and hybrid cars exclusively with higher discounts than discounts for traditional cars. the cars are one of most sources of environmental pollution, which means that if we support the customer who is loyal to the company with an advantage that makes him go more for environmentally friendly cars, this will be good support for the environment, but of course without losing customers to make this principle come true. The priority in the end is profit, while trying to push as much as possible towards sustainability.

1.2. BUSINESS OBJECTIVE

There's no alternative to sustainable development. Even so, many companies are convinced that the more environment-friendly they become, the more the effort will affect their competitiveness. They believe it will add to costs and will not deliver immediate financial benefits [2].

However, in the field of cars, regardless insurance, sale or manufacturing. there is a huge opportunity for investment in insuring electric and hybrid cars as they became more mature and more popular than before. Electric cars specially became a market trend that means it is a good business opportunity to focus on such cars early before competitors. It is worth mentioning that insurance of electric and hybrid cars is more expensive than petrol cars [3].

2.0. Methodology

This section represents the research methodology. The dataset description and the selected features will be discussed in section 2.1. In order to be able to test on the data, it requires pre-processing that will be in section 2.2. and the proposed solution will be discussed in section 2.3.

The CRISP framework [4] has been used to approach the data in this study. This approach will follow several steps i.e., business understanding, data understanding, data preparation, modelling and finally evaluating the model.

2.1. THE USED DATASET

The Auto-Insurance dataset from Kaggle [1] is used in this work. This dataset holds information about auto-insurance customers, social, demographic and insurance policy related details per customer.

I will explain some tables and charts to summarize the data.

The Data is formatted as CSV and contains 9134 observations and the 24 columns as follows:

I will focus on the variables that are relevant to my research. The description of the columns that are relevant to my further analysis is provided by Table 1.

Variable	**Description**
Customer Lifetime Value	contains CLV for each customer
Response	indicates if the customer responds to the marketing campaigns
Coverage	is basic, extended or premium policy
Monthly Premium Auto	the amount paid by customer on a monthly basis
Income	the customer income
Total Claim Amount	the amount paid by the insurance company
Months Since Policy Inception	Numbers of months since coverage started

Table 1: Description of relevant dataset variables of the analysis.

2.2. DATA PREPROCESSING AND EXPLORATION.

The first step in the preprocessing is to remove the null values. After check, the dataset contains no null values. However, the values of the dataset features are different from one feature to another which make the ML model biased toward the large values of the dataset features. Thus, the normalization is the necessarily preprocessing step to make all of the values in the same range between 0 and 1. normalized the numerical features to fit in the used machine learning algorithms in the experiment. I checked for the outliers. In our assumption, I assumed CLV > 50,000 is an outlier. However, the dataset contains only 20 outliers which will have a low impact in the obtained accuracy can be ignored. Furthermore, I converted all categorical features into numerical features to fit in the input of ML models.

Some samples of the dataset are depicted in Table 2.

Customer	State	Customer Lifetime Value	Response	Coverage	Education	Effective To Date	EmploymentStatus	Gender	Income	...	Months Since Policy Inception	Number of Open Complaints	Number of Policies
GH48715	Arizona	5623.757656	No	Basic	High School or Below	1/31/11	Employed	F	77912	...	21	1	3
GM54859	Oregon	6051.637445	No	Extended	High School or Below	1/21/11	Unemployed	M	0	...	80	0	6
NL41409	Oregon	2606.208503	No	Basic	Bachelor	2/14/11	Employed	F	28519	...	56	0	1
TM23514	Oregon	10272.608200	No	Extended	College	1/1/11	Employed	M	60145	...	28	0	3
OO79691	Arizona	11828.614370	No	Basic	College	2/27/11	Unemployed	F	0	...	8	0	2

Table 2: Sample of the dataset.

Table 3 shows the mean of CLV per customer which is around 8000$ while mean of income per customer is around 37500$, and the monthly premium auto per customer is around 93$.

	Customer Lifetime Value	Income	Monthly Premium Auto	Months Since Last Claim	Months Since Policy Inception	Number of Open Complaints	Number of Policies	Total Claim Amount
count	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000
mean	8004.940475	37657.380009	93.219291	15.097000	48.064594	0.384388	2.966170	434.088794
std	6870.967608	30379.904734	34.407967	10.073257	27.905991	0.910384	2.390182	290.500092
min	1898.007675	0.000000	61.000000	0.000000	0.000000	0.000000	1.000000	0.099007
25%	3994.251794	0.000000	68.000000	6.000000	24.000000	0.000000	1.000000	272.258244
50%	5780.182197	33889.500000	83.000000	14.000000	48.000000	0.000000	2.000000	383.945434
75%	8962.167041	62320.000000	109.000000	23.000000	71.000000	0.000000	4.000000	547.514839
max	83325.381190	99981.000000	298.000000	35.000000	99.000000	5.000000	9.000000	2893.239678

Table 3: summary of the numerical features.

To get intuition of patterns in the variables, univariate analysis has been conducted. Figure 1 shows that how many cars per vehicle class are insured per users. The vast majority of customers is four-door and two-door cars which shows that most users are more into economic cars.

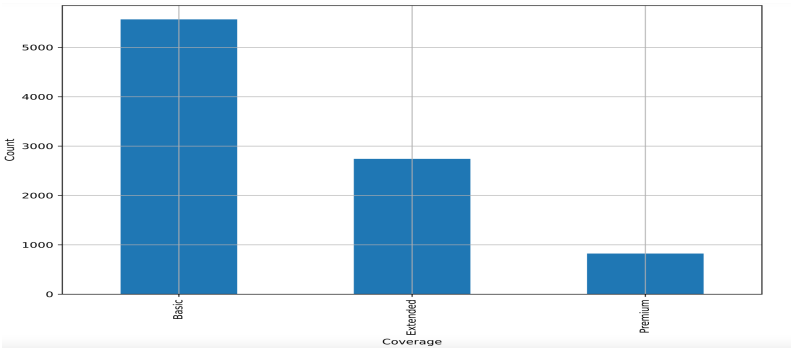

Figure 1: Counts of policy coverage type

I conducted a bivariate analysis to find potential patterns in the data between every pair of variables. In Figure 2, it shows that the marketing personally by agent is effective to customers that other means like web, branch or call center. it means that choosing suitable offers and present that professionally by our marketers will help the company.

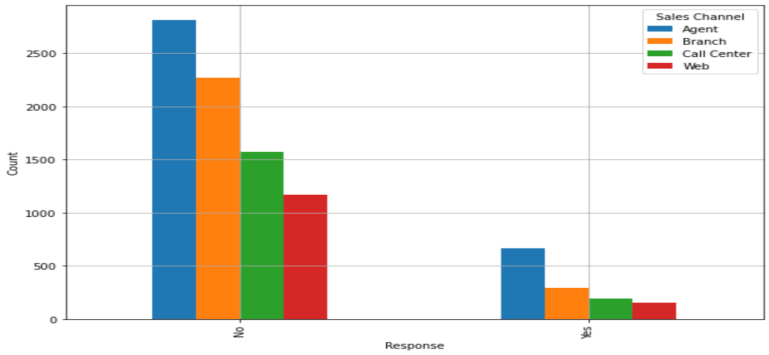

Figure 2: Counts of response per sales channel.

2.3.PROPOSED SOLUTION.

In order to achieve the business objective, I have conducted experiment of two steps.

1- Categorize the customers into groups that shares similar characteristics to facilitate the way we will treat the current loyal groups [3]. I used Kmeans clustering to achieve the customer segmentation. This technique helps to understand the current customers.
2- Predict the CLV for a new customer by training a simple machine learning model our dataset [1] to predict the potential value of the new customer. This technique helps to assess the importance of the new customers, given that The CLV is a profitability metric in terms of a value placed by the company on each customer and can be conceived in two dimensions the customer`s present Value and potential future Value.

By applying the two techniques, we can come to a better understanding of our customers to offer the loyal customers attractive offers to keep them and promote the sustainability.

3.0. THE EXPERIMENT AND RESULTS.

3.1. CLUSTERING THE CUSTOMERS.

I clustered our customer base into segments. I selected all numerical columns for this analysis. I chose Kmeans algorithm for simplicity. I used Silhouette Score To decide on how many clusters I will choose. I figured out that 5 clusters would be accepted. In order to visualize the clusters, I implemented dimensionality reduction to two dimensions to provide a visualization to the clusters. Figure 3 shows the clustering results.

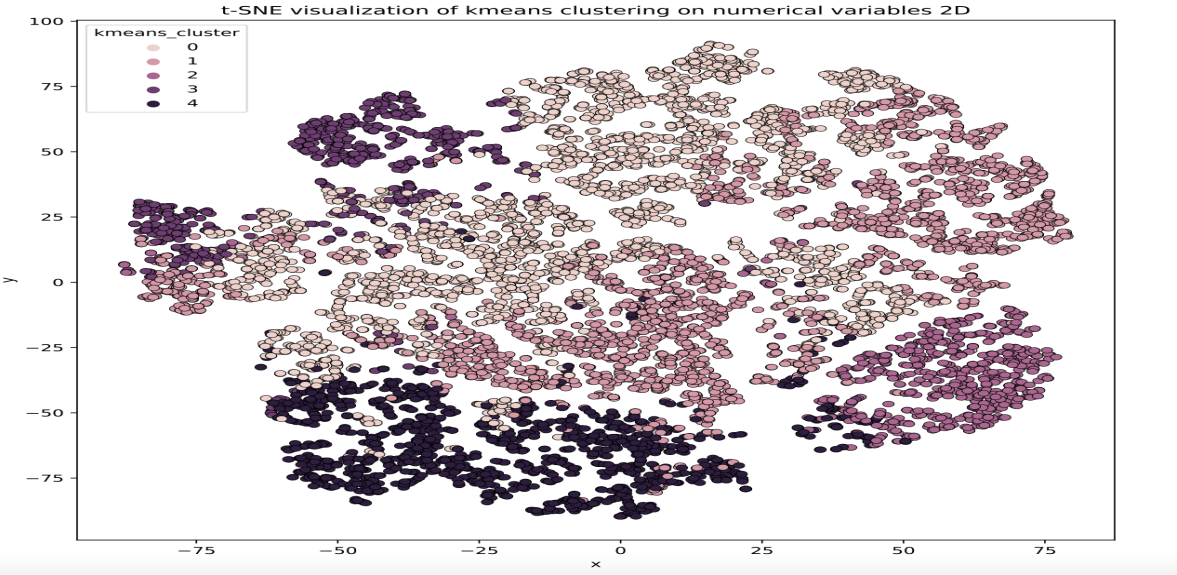

Figure 3: Clusters of the customers after dimensionality reduction to 2 dimensions.

To get some intuition of the clusters, I made some analysis to get the average of the two important variables CLV and Monthly premium auto per cluster. By doing so, I figured out that cluster 3 is our target cluster as it has max mean value of CLV among all cluster with a significant difference as seen in figure 4. In addition to that, cluster 3 has the max mean value of Monthly premium auto among all cluster as seen in figure 5.

	kmeans_cluster	Customer Lifetime Value	PERCENTAGE
0	3	19,888.44	42.113893
1	1	7,388.32	15.644823
2	0	6,703.96	14.195681
3	2	6,626.55	14.031761
4	4	6,618.09	14.013843

Figure 4: Mean value of customer lifetime value per cluster.

	kmeans_cluster	Monthly Premium Auto	PERCENTAGE
0	3	169.74	32.983692
1	0	87.17	16.938848
2	2	86.69	16.845812
3	4	86.04	16.718766
4	1	84.98	16.512882

Figure 5: Mean value of monthly premium auto per cluster.

The Customer Lifetime Value is the net present value of a customer. It considers the difference between the total amount of revenues from a customer and the companies` expenses for this customer during the whole duration of relationship.

3.2. PREDICTING THE CLV OF A NEW CUSTOMER.

I implemented linear regression model to predict the CLV. Referring to Figure 7, we see that not all features are highly correlated.

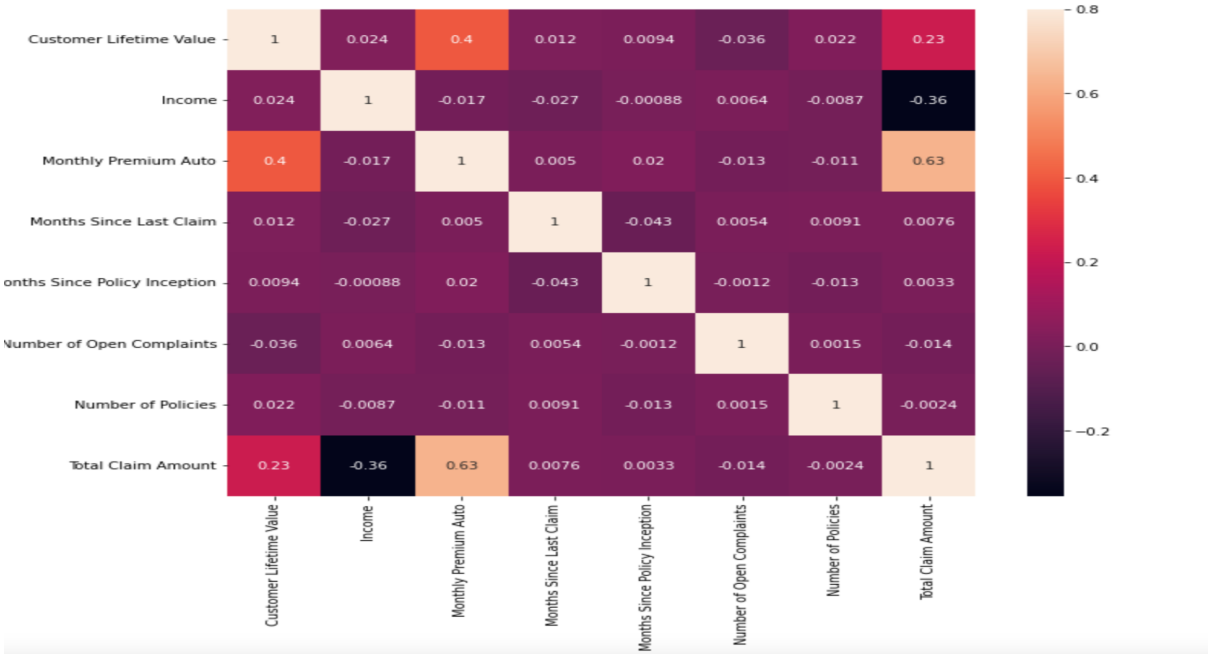

Figure 7: Correlations between numerical variables.

I decided to take into account the most significant feature based on the correlation values. I used a code to r ank them in a descending order and took the first four variables into account (Income, Total Claim Amount, Months Since Policy Inception, Monthly Premium Auto). The data has been splitted into training and test data. I assessed the model using R squared values, Root Mean Square Error, Mean Absolute Error, Median Ab solute Error, R^2 and Adjusted R^2 .at first, inaccurate results have been noticed. The normality of the CLV ta rget variable was checked afterwards. I figured out it is not suitable for linear regression robust model as it is skewed (I applied shapiro test to check normality). I converted CLV to the logarithmic representation. By repe ating the experiment with the new modified CLV, I could get accurate results. *the steps are explained in the at tached python script*. As seen in Figure 8, the predictions are very close to the actual data. We now can gener alize this to new data.

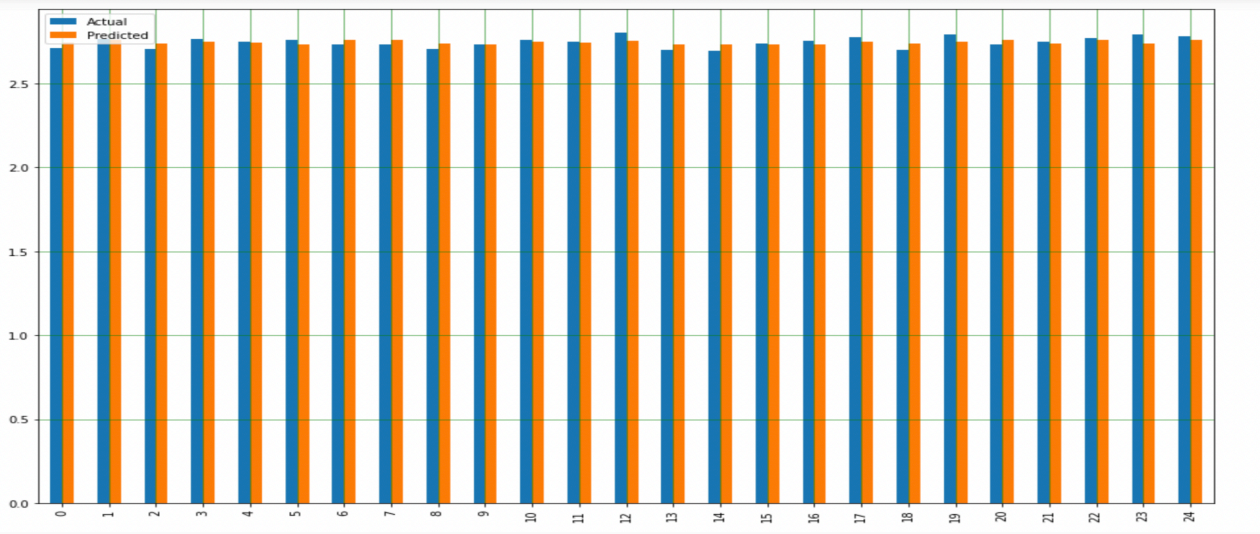

Figure 8: Actual vs predicted CLV values

For a new customer, if the CLV prediction is more than the mean value of cluster 3 CLV value (our target cluster) then we can promote green discounts for him/her even though assigning a new customer to any cluster is not validated yet and requires further analysis. But, if the CLV is very high we better to take care of such customer from the beginning.

4.0. Discussion and Future Work.

The basic idea behind my analysis is that Targeting the loyal customers is a key point for promoting the sustainability as those customers would be with the company more than the other customers. Offering the loyal customers green discounts would help in getting more customers in insuring green vehicles. Moreover, electric and hybrid in a rise generally as getting less expensive as the technology becoming more mature. Business wise, targeting electric and hybrid cars is important as they are the next generation technology. That means that it is not just promoting sustainability but also making more money by increasing our market share in the emerging car market.

It is worth mentioning that offering attractive offers for the loyal customers is important even for non-green cars. we could offer less attractive offers but we should stick to competitive compared to our competitors offers in order to keep the business stable as well.

As a future work, the experiment could be extended by linking the two steps together. This is could be achieved by classifying a new customer to the relevant cluster, given that we could predict his/her CLV. our target cluster for loyal customers was rated by CLV as I showed in step 1 of the experiment.

In terms of data management, understanding the data in the best possible way is crucial, preprocessing, data cleansing, choosing the relevant variables for further analysis, exploratory data analysis and choosing the right algorithms are vital to take advantage of the data.

5.0. REFERENCES

[1] the case study dataset: https://www.kaggle.com/code/tariqmuneer/wa-fn-usec-marketing-customer-lm/data

[2] Why sustainability is a driver for innovation: https://hbr.org/2009/09/why-sustainability-is-now-the-key-driver-of-innovation

[3] Do electric vehicles cost more to insure than gasoline-powered cars?
https://www.insurancebusinessmag.com/us/news/commercial-auto/do-electric-vehicles-cost-more-to-insure-than-gasolinepowered-cars-425631.aspx

[4] DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model
https://doi.org/10.1016/j.procir.2019.02.106

[5] Customer segmentation and strategy development based on customer lifetime value: A case study:
https://doi.org/10.1016/j.eswa.2005.09.004