A

PROJECT REPORT ON

Insurance Marketing by Customer Lifetime Value Analysis

Submitted in partial fulfilment of the requirements for the post graduate program

in

Data Science & Engineering

by

Akash Chaple (SSIDNBTGPH6R25)
Narendra Kumar (SSID:RFYCMJ5ZVP)
Akaanksha Mishra (SSID:5OTNI2Q7IX)
Swinal Meshram (SSID: YH6EGG4XNB)
Santosh Kumar (SSID:OAIJ5B9BRI)

Under the guidance of

Mr. Muppidi Srikar

GREAT LEARNING

Learning for life

HSR LAYOUT:: BANGALORE

ACKNOWLEDGEMENT

We welcome this opportunity to express our heartfelt gratitude and regards to our project guide Mr. Muppidi Srikar , for his unconditional guidance. We would also like to extend our gratitude to Ms. Barkha Patowary, Data Scientist, Great Learning, Bangalore, for encouraging us to undertake the project and help us with the initial stages. We greatly admire and acknowledge the constant support we had from my friends and team members for all the effort and hard work that they have put into completing this project.

ABSTRACT

Customer lifetime value (CLV) is one of the most important metrics to measure at any growing company. If you want your business to acquire and retain highly valuable customers, then it's essential that your team learns what customer lifetime value is. It indicates the total revenue a business can reasonably expect from a single customer account. Businesses use this metric to identify significant customer segments that are the most valuable to the company. CLV tells companies how much revenue they can expect from one customer to generate over the course of the business relationship. The extracted features from the data are then fed to the machine learning regression methods to build a model. Feature selection pre-processing steps are used to enhance the performance and scalability of the regression methods.

greatlearning
Learning for Life

GREAT LAKES
INSTITUTE OF MANAGEMENT, CHENNAI

CONTENTS

1. INTRODUCTION

What is Customer Lifetime Value?

Customer lifetime value (CLV) can be defined as the net present value of cash flow (past and future) attributed to a customer or household for a designated time period. Or, more simply, the difference between the total premium revenue received and total expenses over the course of the relationship. In many cases, this may be greater than 20 years. CLV shows which customers will offer the highest value in the future, which can identify the core attributes insurers should look for in current customers and prospects.

INDUSTRY REVIEW

Insurance companies have historically been concerned about marketing and attracting customers, but there have not been many detailed statistical analyses around marketing and customer attraction. Insurance companies have always been involved in marketing, but it has been targeted mainly at building name recognition for direct companies.

Independent agent companies focused their marketing activities on the independent agent and usually not the end customer. The judgment of most insurers was that customer decisions were driven mainly by price, and so if a company wanted to attract more customers, they would simply decrease the premiums. Insurers are beginning to find out, though, that while price is an important factor, price is not the only factor that drives a customer's purchase decision. There are a number of other considerations that go into purchase choices, including company reputation, customer service, and perceived value. In addition, different customers can place a different level of value on different things. For example, price may be more important for a younger customer, while company reputation may be more important for an older customer. These and many other considerations are now being taken into account in understanding the purchase decisions of a customer.

BACKGROUND AND RELATED WORK

Customers have become the alma mater of any organization, because without them there wouldn´t be incomes, benefits and the resulting market value of the company (Gupta & Lehmann, 2003; Gupta & Zeithaml, 2006). For this reason, identifying the most profitable customers, who will strengthen relationships in the long term, has become a priority for both academics and professionals in marketing. If the latter idea is expressed in more operational terms, the goal is to understand how to effectively manage relationships with customers and how to implement relationship-marketing strategies with the most profitable ones (Kumar, Ramani & Bohling, 2004) in order to retain these customers and increase purchases made by them (Jain & Singh, 2002).

PROBLEM STATEMENT

For an Auto Insurance company, we need to predict the conditions affecting customer lifetime value (CLV). CLV is the total revenue the client will derive from their entire relationship with a customer. we need to predict the customer lifetime value for each customer so as to make sure how much benefit each customer can repay to the company in exchange of the benefits he or she receives. The project will also trace out interdependence amongst the features and will aim at providing high level of interpretability as the readers should be able to comprehend the decisions made by the model given the nature of the domain this project is related with.

2. DATA PRE-PROCESSING

Dataset Description:

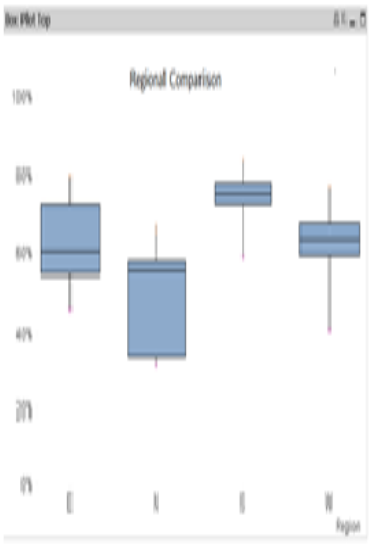
This dataset has 24 columns (features) and 9314 rows (records).

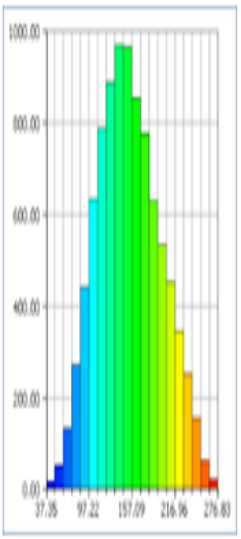
CONTINUOUS FEATURE	FEATURE DESCRIPTION
Income	Customer annual income in USD
Monthly Premium Auto	Monthly Premium for auto insurance
Total Claim Amount	Amount claimed till data
Months Since Last Claimed	Number of months before which the last claim was made
Months Since Policy Inception	Number of months before which the policy commenced
Number of Open Complaints	Numbers of unresolved complaints from the customer
Number of Policies	Number of policies with the current customer
Customer Lifetime value – (Target)	CLV of the customer for the auto insurance company

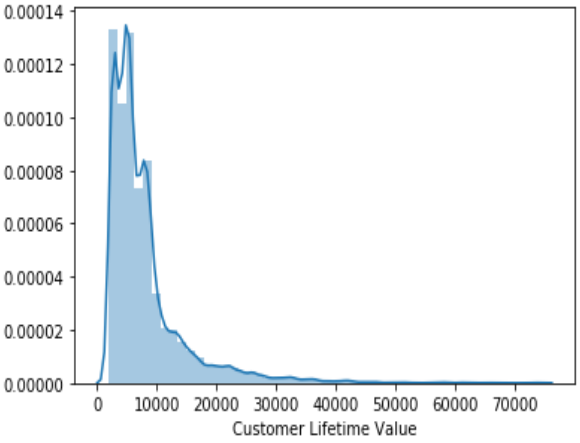
CATEGORICAL FEATURE	CATEGORY DESCRIPTION
State	US province to where the customer belongs
Response	Refers to whether customers have responded to marketing calls or not
Coverage	Nature of insurance coverage
Education	Education level of customer
Employment Status	Current Employment status of the customer
Gender	Gender of the customer
Location Code	Type of location where customer lives
Marital Status	Marital status of the customer
Policy Type	Type of policy
Renew Offer Type	Offer given during renewal
Sales Channel	Channel of sales
Vehicle Class	Type of vehicle
Vehicle Size	Size of vehicle

UNIVARIATE ANALYSIS:

Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	







Distribution of Customer lifetime value does not seem to be following Normal distribution. So Before Further Exploration we need to transform via log to make a bit normally distributed

MISSING VALUE TREATMENT:

In any real-world data set, there are usually few null values. It doesn't really matter whether it is regression, classification or any other kind of problem no model can handle these NULL or NaN values on its own so we need to intervene. Firstly, we need to check whether we have null values in our dataset or not. We can do that using the isnull() method. There are various ways for us to handle this problem. The easiest way to solve this problem is by dropping the rows or columns that contain null values.

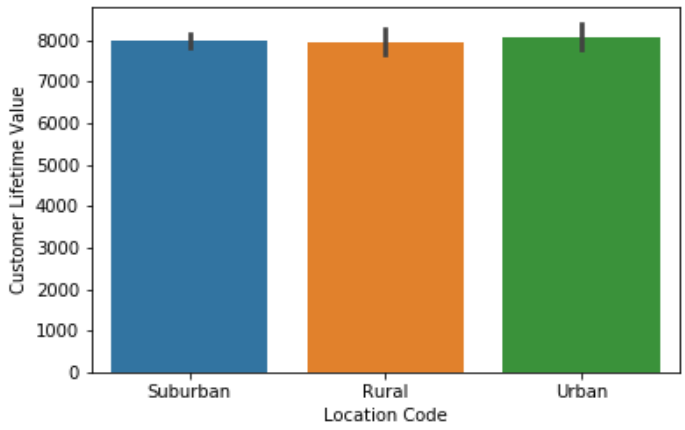
However, it is not the best option to remove the rows and columns from our dataset as it can lead to loss of valuable information. So, if you have 9K data points then removing 2–3 rows won't affect your dataset, whereas if you no one have NaN values for a particular field then you can't simply drop those rows. In real life datasets it can happen quite.

There are NO MISSING VALUES are present in the given Dataset.

```
1  data.isnull().sum()
```

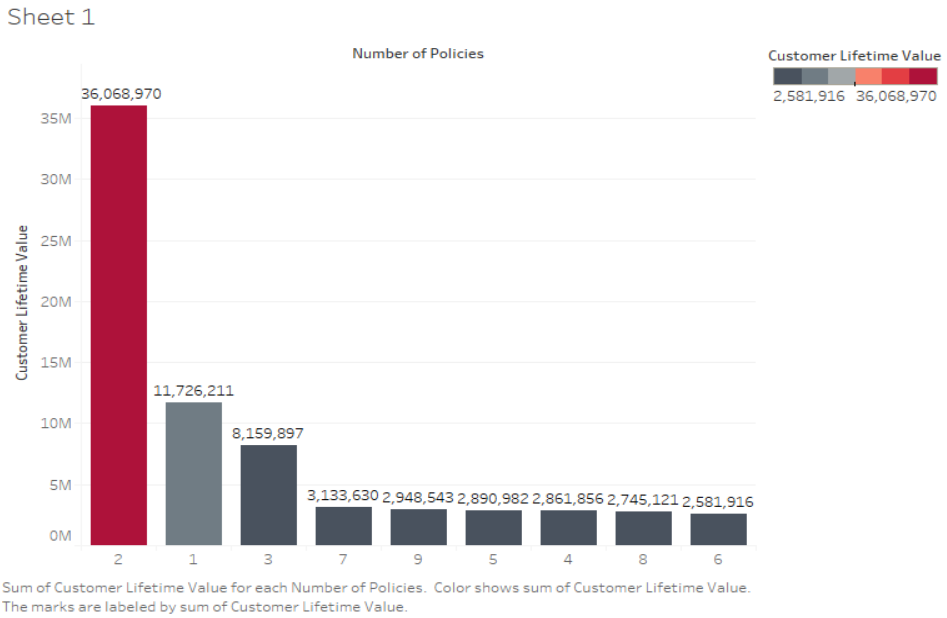
```
State                          0
Customer Lifetime Value        0
Response                       0
Coverage                       0
Education                      0
EmploymentStatus               0
Gender                         0
Income                         0
Location Code                  0
Marital Status                 0
Monthly Premium Auto           0
Months Since Last Claim        0
Months Since Policy Inception  0
Number of Open Complaints      0
Number of Policies             0
Policy Type                    0
Policy                         0
Renew Offer Type               0
Sales Channel                  0
Total Claim Amount             0
Vehicle Class                  0
Vehicle Size                   0
dtype: int64
```

EXPLORATORY DATA ANALYSIS:



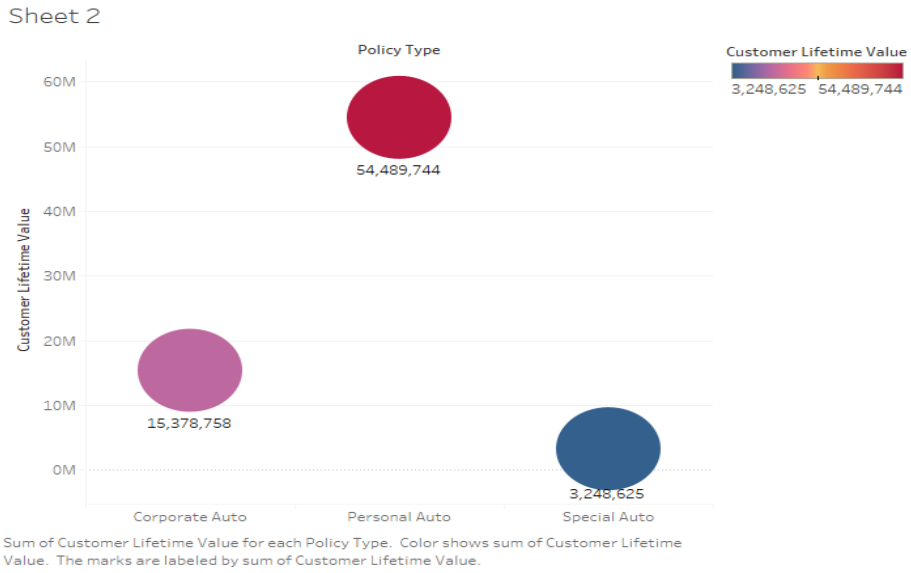
This bar plot clearly shows that no matter what location code a person is from the average customer lifetime value is same. This is also proved using ANOVA statistical test.

NUMBER OF POLICIES vs CLA :



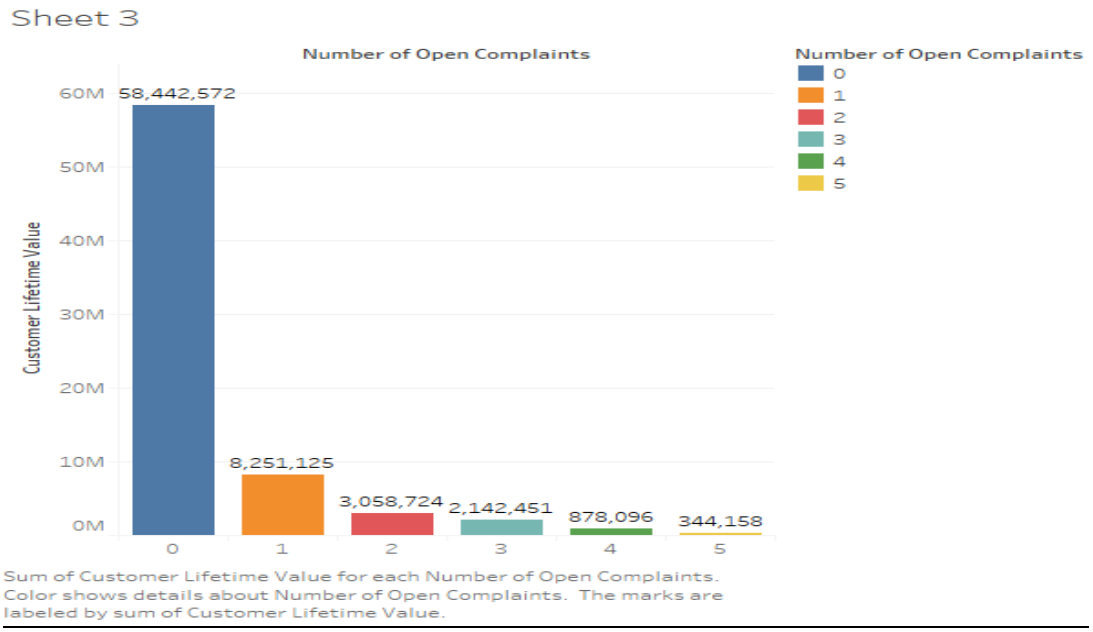
We can see a pattern here, customers who have taken only 1 policy have lower customer lifetime value, and customers who have taken 3 or greater show a similar trend, so we can combine all of them into one bin, and we can also see that the customers who have taken 2 policies have very high customer lifetime value comparatively.

POLICY vs CLA :



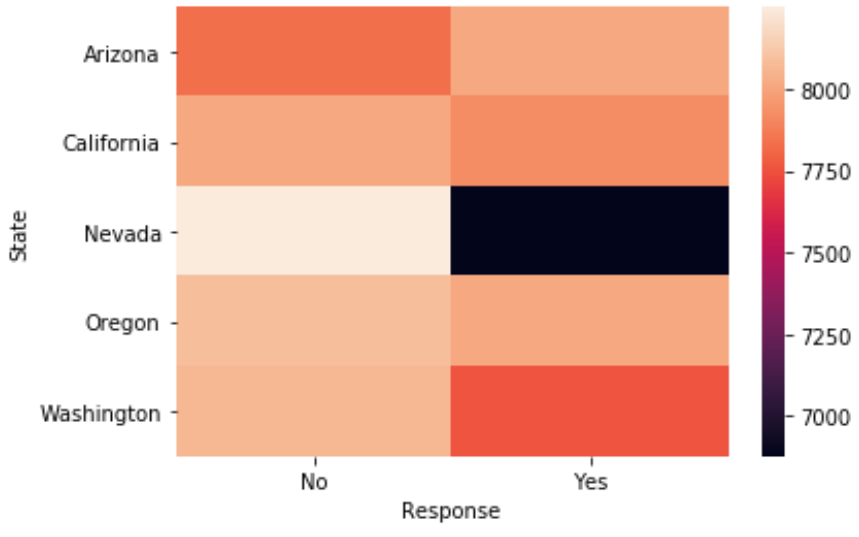
There isn't much difference in the customer lifetime value w.r.t what policy type he has taken, all we need is how much revenue a customer can bring to the company, so it doesn't matter what type of policy he/she has chosen.

NUMBER OF OPEN COMPLAINTS vs CLA :



Number of open complaints also show kind of similar trend, where people who have complaints 2 or lesser have a similar pattern but whereas >3 do not show any pattern we will have to do statistical test to understand if this feature is really significant or not.

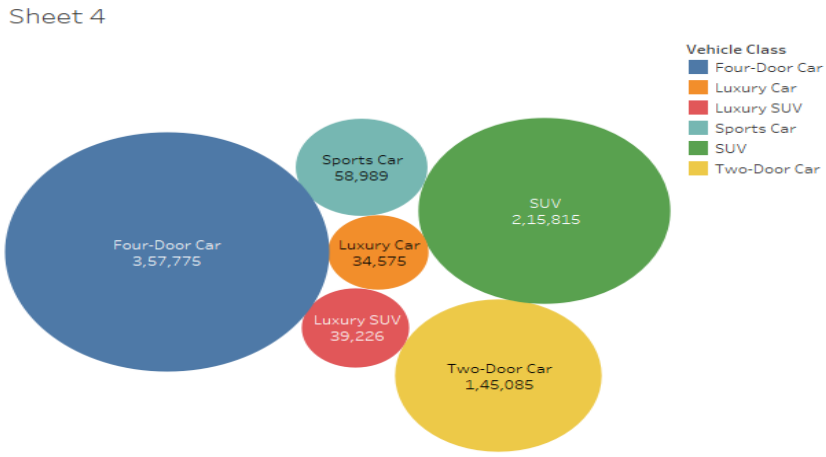
STATEWISE DISTRIBUTION OF CLV BASED ON RESPONSE:



The investment in customers as positive response to generate revenue and profitability was more assured in the states of Oregon, Arizona with the customer lifetime value of above 8000 and the Nevada is having least customer value of 7000 other than this the negative responsers played a crucial role in gaining customer lifetime value of Baove 8000 which is a good instinct to play with. Nevada had a negative response with the highest customer lifetime value above 8500.

VEHICLE CLASS VS MONTHLY PREMIUM AUTO

The bubble graph shows the customers having different vehicle class. And the relevant to that how much premium they are paying based on the category of vehicle it's an obvious instinct the person having luxurious vehicle such as sports car or a two door cars will be paying more premium compare to the other customers having suv and 4 door cars.



3. STATISTICAL ANALYSIS

Statistical testing was done to look at the evidence for a particular hypothesis being true. This helped us to accomplish decisions. Hypothesis will be performed to statistically validate the impact of respective independent variables on the outcome. Since all the variables are discrete, we will perform t_test and 1_way_annova to test feature significance.

- **Hypothesis**

Ho: *The feature is not significant predictor of Target.*
Ha: *The feature is significant predictor, i.e., it has high association with Target.*

Categorical Columns:

	columns	P_vlaue	status
0	State	2.862454e-01	not significant
1	Response	2.608203e-01	not significant
2	Coverage	9.775384e-110	significant
3	Education	1.679997e-02	significant
4	EmploymentStatus	1.381445e-08	significant
5	Gender	2.379516e-01	not significant
6	Location Code	2.854097e-01	not significant
7	Marital Status	3.052088e-05	significant
8	Policy Type	9.956656e-02	not significant
9	Policy	4.363039e-01	not significant
10	Renew Offer Type	2.766519e-36	significant
11	Sales Channel	2.155648e-01	not significant
12	Vehicle Class	1.072399e-280	significant
13	Vehicle Size	8.406723e-03	significant

Numerical Columns

	Numerical_column	P_value	Status
0	Income	0.017	Significant
1	Monthly Premium Auto	0.000	Significant
2	Months Since Last Claim	0.187	Not_Significant
3	Months Since Policy Inception	0.612	Not_Significant
4	Number of Open Complaints	0.001	Significant
5	Number of Policies	0.001	Significant
6	Total Claim Amount	0.049	Significant

After Looking at the base model and the p-value of the feature's, we know that the

Hypothesis for the features is:-
H0: Feature is not significant
H1: Feature is significant

But we just can't conclude the significance of the features by base model and also without using any of the feature engineering techniques we have at our disposal. So we will first try to do the statistical tests of the feature for the feature selection, we can also use the forward selection and backward elimination, we will use the Variance inflation factor