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 Patwari, Senior 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
Global Mindset - Indian Roots

CONTENTS

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

Once insurers have created the data environment, the next step is to use analytics to build out the predictive models to project future customer value. This would include analysis for four key attributes:

Customer segments: Customer segmentation helps insurers identify homogeneous groups within their customer base. It provides a strategic view for identifying patterns and customer behavior so insurers can:

- Price more effectively.
- Focus attention to higher-value segments.
- Develop tactics to improve value segments.

• Retain and serve the customers better.

Customer Loyalty: The most conservative statistics say that it costs five times more to acquire a new customer than retain an existing customer. For the insurance industry, that figure easily jumps to 10 times more to generate new business. So customer retention is critical for insurers. Insurers need to mine the vast amount of customer and policy data available to predict which customers are likely to lapse and – most importantly – design cost-effective strategies to persuade them to remain a customer. Even simple indicators have shown to improve retention rates. For example, customers who pay in full have a higher retention rate compared with those on monthly payment plans.

The main rating factors for auto insurance are:

- Location
- Age
- Gender
- Marital status
- Driving experience
- Driving record
- Claims history
- Credit history
- Previous insurance coverage
- Vehicle type
- Vehicle use
- Miles driven
- Coverages and deductibles

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.

Chapter-2.Data Pre-Processing

Dataset Description:

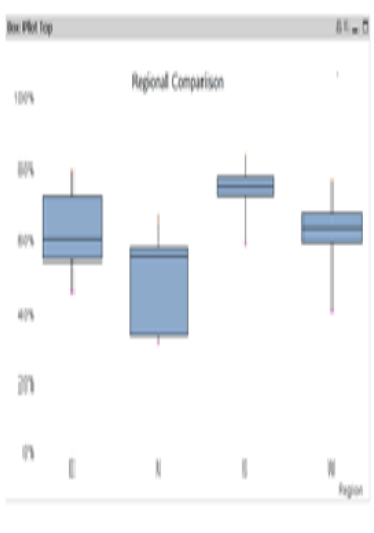
This dataset has 24 columns (features) and rows (records).

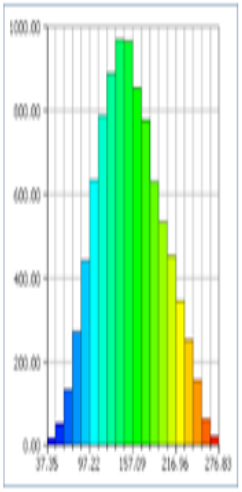
Continuous Feature	Feature Description
Income	Customer annual income in USD
Monthly Premium Auto	Monthly Premium for auto insurance
Total Claim Amount	Amount claimed till data
Months Since Last Claimed	Number of months before which the last claim was made
Months Since Policy Inception	Number of months 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 – (Y)	CLV of the customer for the auto insurance company

Categorical Feature	Category Description
State	US province to where the customer belongs to
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	





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 .

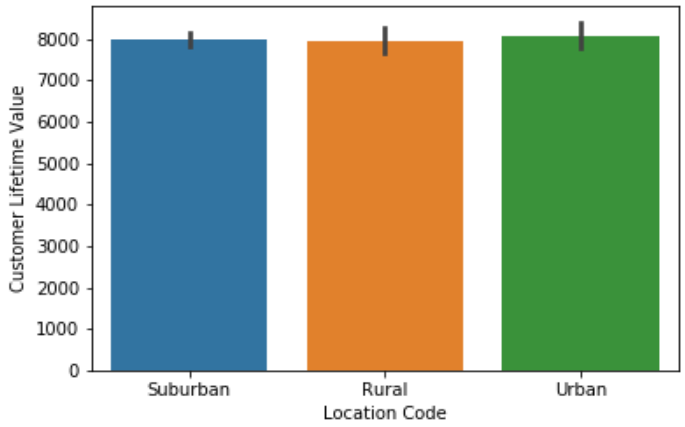
There are NO MISSING VALUES are present in the given Dataset.

```
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
```

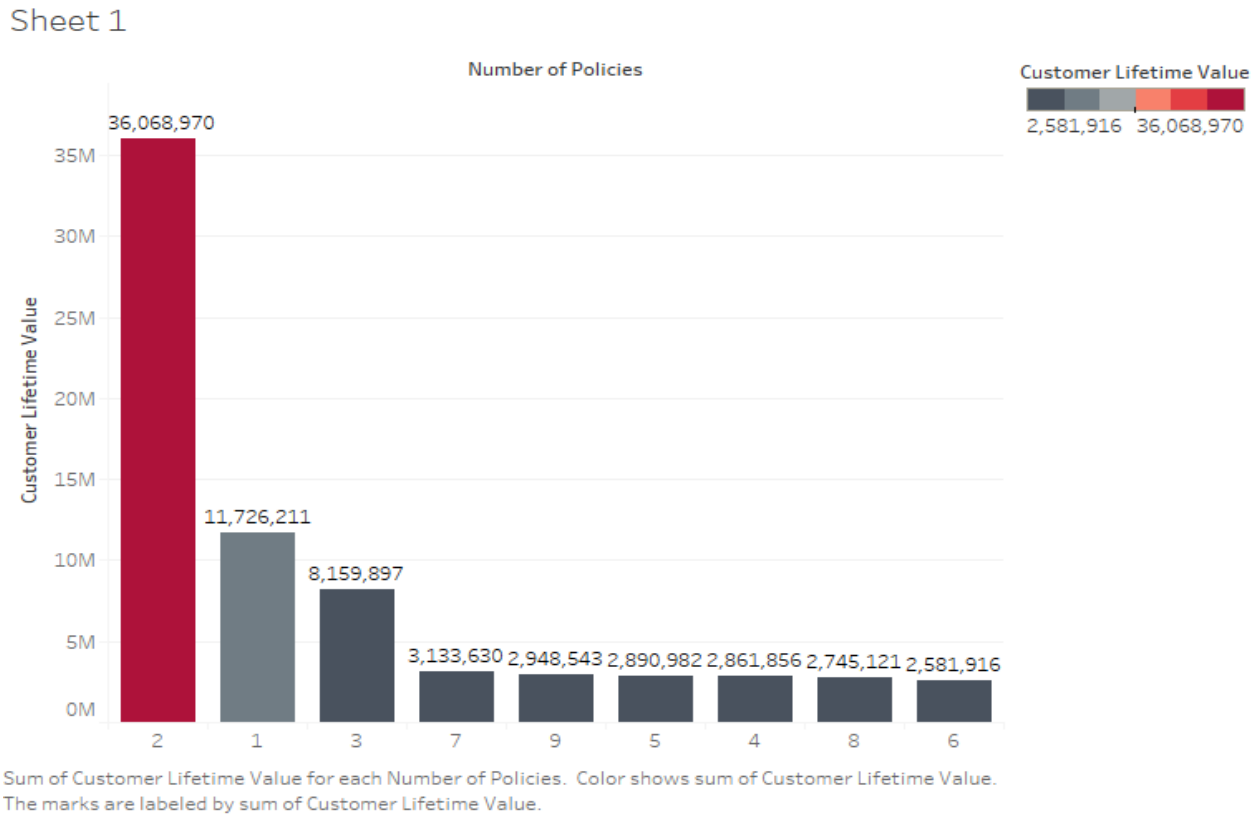
Exploratory Data Analysis:

Number of policies :



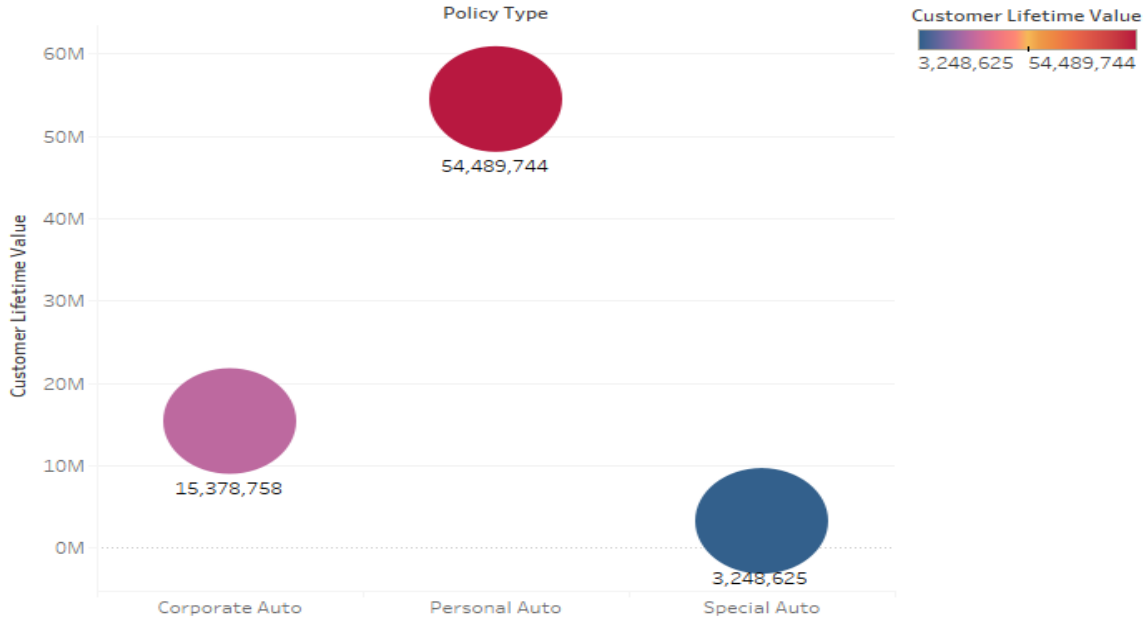
This Bar plot clearly shows that no matter what location code a person is from the average customer lifetime value is same.

This is also proved using ANOVA **statistical test.**



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 comparitively.

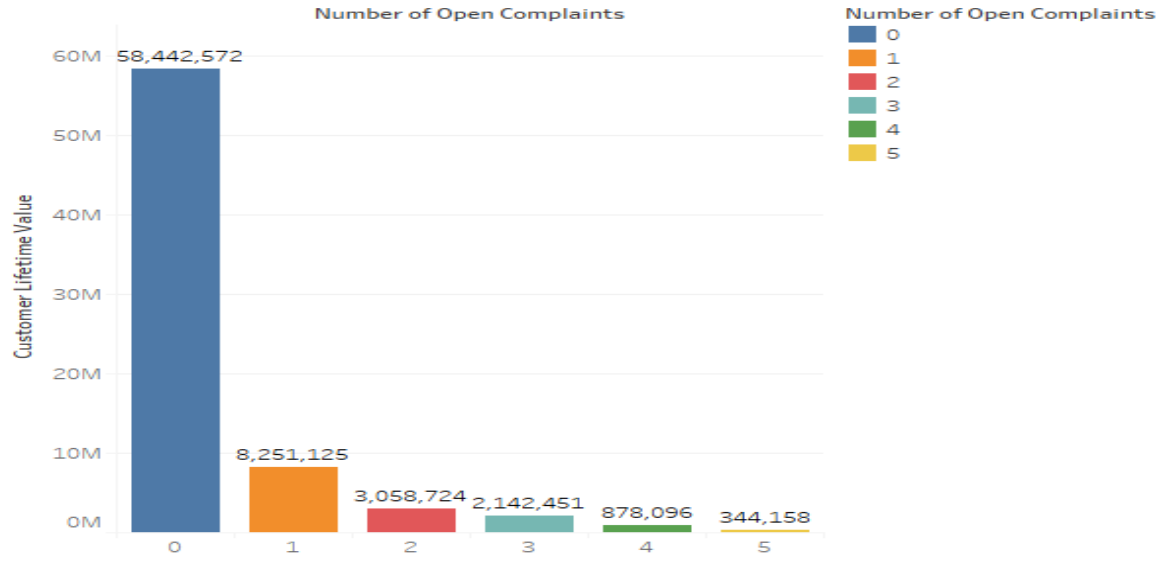
Sheet 2



Sum of Customer Lifetime Value for each Policy Type. Color shows sum of Customer Lifetime Value. The marks are labeled by sum of Customer Lifetime Value.

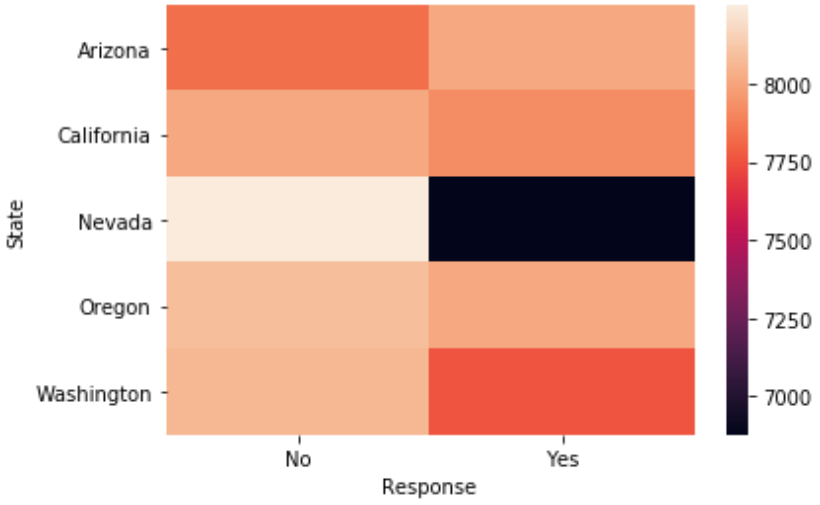
There isn't much difference in the customer lifetime value w.r.t what policy type he has taken, all we need is how much revenue a customer can bring to the company, so it doesnt matter what type of policy he/she has chosen.

Sheet 3



Sum of Customer Lifetime Value for each Number of Open Complaints. Color shows details about Number of Open Complaints. The marks are labeled by sum of Customer Lifetime Value.

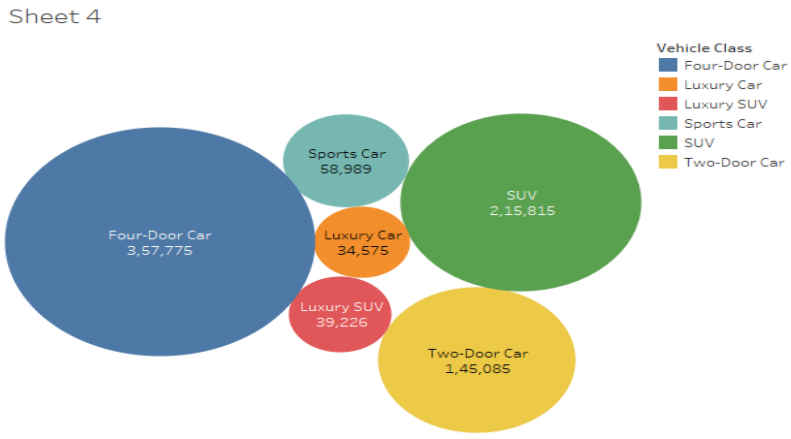
Number of open complaints also show kind of similar trend, where people who have complaints 2 or lesser have a similar pattern but where as >3 do not show any pattern we will have to do statistical test to understand if this feature is really significant or not.



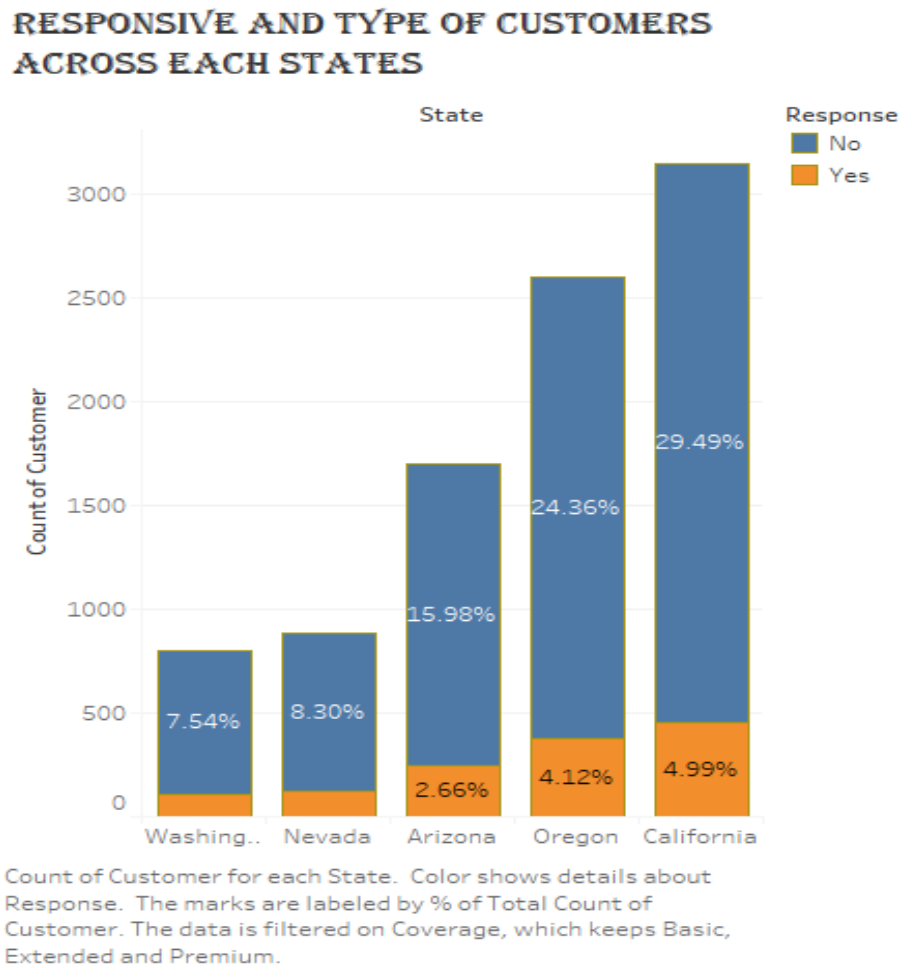
The investment in customers as positive response to generate revenue and profitabillity was more assured in the states of oregon,arizona with the customer lifetime value of above 8000 and the Nevada is having least customer value of 7000 other than this the negative responsers played a crucial role in 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 relevent to that how much premium they are paying based on the category of vehicle its an obvious instinct the person having luxiourious vehicle such as sports car or a two door cars will be paying more premium compare to the other custmers having suv and 4 door cars.

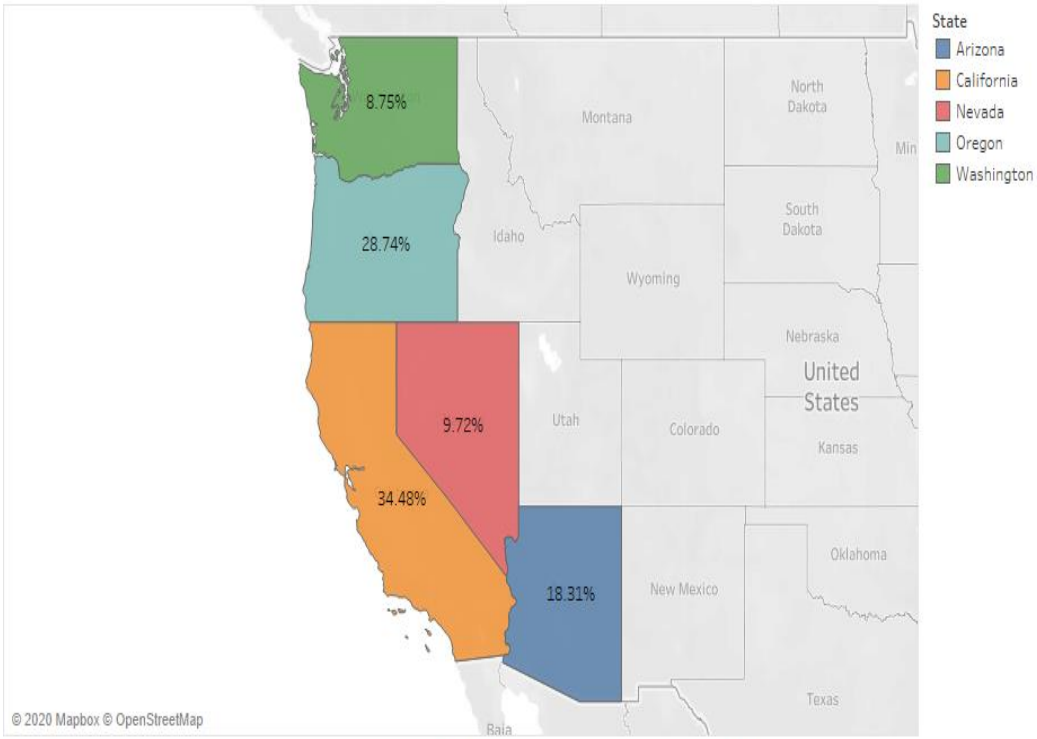


Vehicle Class and sum of Monthly Premium Auto. Color shows details about Vehicle Class. Size shows sum of Monthly Premium Auto. The marks are labeled by Vehicle Class and sum of Monthly Premium Auto.



The graph depicts the count of customer across various states along with the colours represents the Response by the customers and we termed them as responsive customers meaning these customers responded to the approach by the company .The marks labeled shows the percent of customers in each state by keeping the premium coverage as filters. As we can see that the number of responsive customers are more in the state of california i.e 24.9% .

clv across various states according to policy expirtion

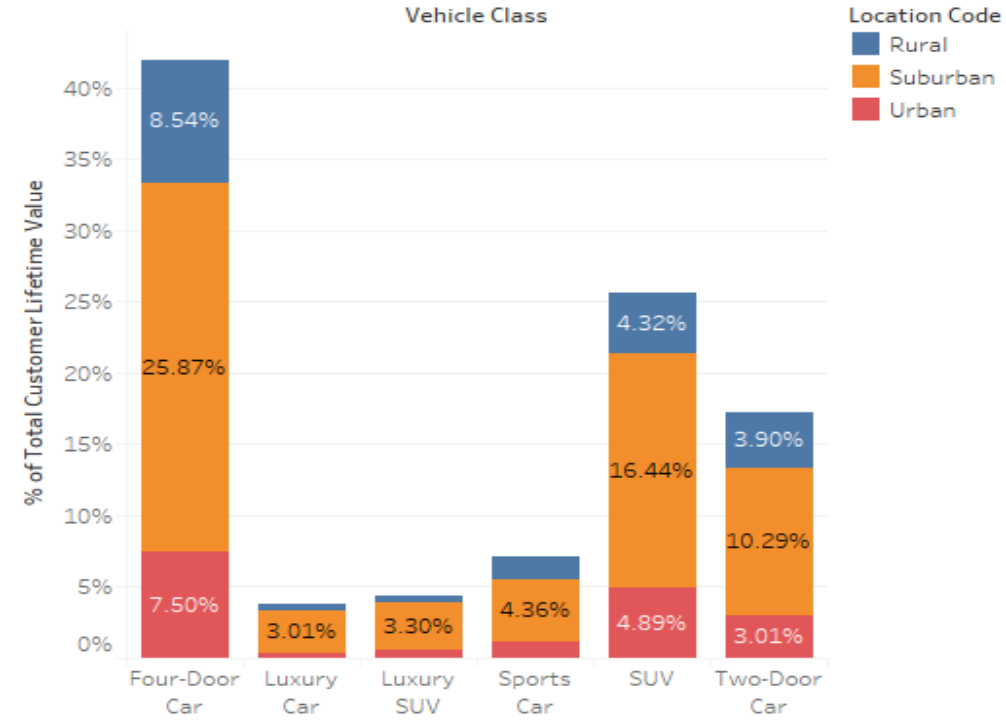


Map based on Longitude (generated) and Latitude (generated). Color shows details about State. The marks are labeled by % of Total Customer Lifetime Value. The data is filtered on Effective To Date Month, which keeps January and February.

If we take a look at the colour section we can find california is having more customer lifetime value than other states and here we have filtered the colun based on the effective to date which means that the customers whose policy is going to be expired in the months of january and february.

Apart from this Arizona is ranked 2nd based on the customer lifetime value .

CUSTOMER LIFETIME VALUE BASED ON THE VEHICLE SIZE OF CUSTOMER ACROSS VARIOUS LOCATIONS



% of Total Customer Lifetime Value for each Vehicle Class. Color shows details about Location Code. The marks are labeled by % of Total Customer Lifetime Value.

The company is getting more profitable customers from The suburban Locations and customers having a regular four door car , If we talk in terms of SUVs the customer lifetime value is highest for sub-urban regions.The people from rural regions are contributing less in terms of other aspects.

CHAPTER 3: FEATURE ENGINEERING :

- There we have 2 Numerical features, Number of Policies and Number of Open Complaints which has only 6 and 9 level. So we convert it into Categorical
- In Effective To Date, we have 2 months data in which month customer Policy will going to expire.so we split this column into Months column having two level, January and February
- We did capping in Number of policies Column, where we cap the customers who's having more than 2 policies into category of Policy No. 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_anaova 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

Ols Model:

Dep. Variable:	Customer Lifetime Value	**R-squared:**	0.165
Model:	OLS	**Adj. R-squared:**	0.163
Method:	Least Squares	**F-statistic:**	75.19
Date:	Mon, 06 Jan 2020	**Prob (F-statistic):**	0.00
Time:	15:09:03	**Log-Likelihood:**	-92834.
No. Observations:	9134	**AIC:**	1.857e+05
Df Residuals:	9109	**BIC:**	1.859e+05
Df Model:	24		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
State	27.8762	51.126	0.545	0.586	-72.342	128.095
Response	-459.2518	192.806	-2.382	0.017	-837.195	-81.308
Coverage	-184.3730	113.720	-1.621	0.105	-407.289	38.543
Education	92.5407	47.794	1.936	0.053	-1.146	186.228
EmploymentStatus	-106.9485	73.613	-1.453	0.146	-251.246	37.349
Gender	-145.5858	132.648	-1.098	0.272	-405.606	114.435
Income	0.0022	0.003	0.667	0.505	-0.004	0.009
Location Code	100.6493	116.603	0.863	0.388	-127.918	329.216
Marital Status	-240.2018	110.694	-2.170	0.030	-457.186	-23.218

Monthly Premium Auto	82.6863	2.925	28.271	0.000	76.953	88.419
Months Since Last Claim	6.4821	6.550	0.990	0.322	-6.357	19.321
Months Since Policy Inception	-0.6709	2.375	-0.283	0.778	-5.326	3.984
Number of Open Complaints	-249.2682	72.387	-3.444	0.001	-391.162	-107.374
Number of Policies	63.7311	27.666	2.304	0.021	9.499	117.964
Policy Type	447.6573	291.072	1.538	0.124	-122.909	1018.224
Policy	-47.6764	86.267	-0.553	0.581	-216.779	121.426
Renew Offer Type	-347.6404	68.522	-5.073	0.000	-481.959	-213.322
Sales Channel	22.3858	62.304	0.359	0.719	-99.744	144.516
Total Claim Amount	-0.5766	0.362	-1.593	0.111	-1.286	0.133
Vehicle Class	45.5802	32.427	1.406	0.160	-17.985	109.145
Vehicle Size	188.7697	123.409	1.530	0.126	-53.139	430.679
year	0.3090	0.269	1.147	0.252	-0.219	0.837
month	-160.7967	132.756	-1.211	0.226	-421.027	99.434
day	4.6860	7.663	0.612	0.541	-10.334	19.706
day_of_week_number	26.5332	32.128	0.826	0.409	-36.445	89.511

Omnibus:	5663.042	**Durbin-Watson:**	1.996
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	64517.119

Skew:	2.850	**Prob(JB):**	0.00
Kurtosis:	14.706	**Cond. No.**	2.22e+05

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

H0: Feature is not significant
Ha: Feature is significant

But we just cant conclude the significance of the feature's just by base model and also without using any of the feature engineering technique's we have at our disposal. So we will first try to do the statistical test's of the feature for the feature selection, we can also use the forward selection and backward elimination , we will use the Variance inflation factor

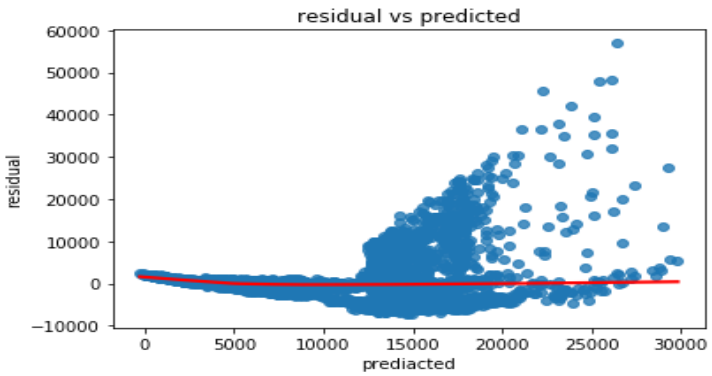
ASSUMPTIONS OF LINEAR REGRESSION.

Linearity
The mean of the residuals is 21.407853224376783

- As mean of residual is not so much large so we can consider that linearity is present
- AS we got p_value greter than aplha(0.05) & we can also see visually that our data is linear

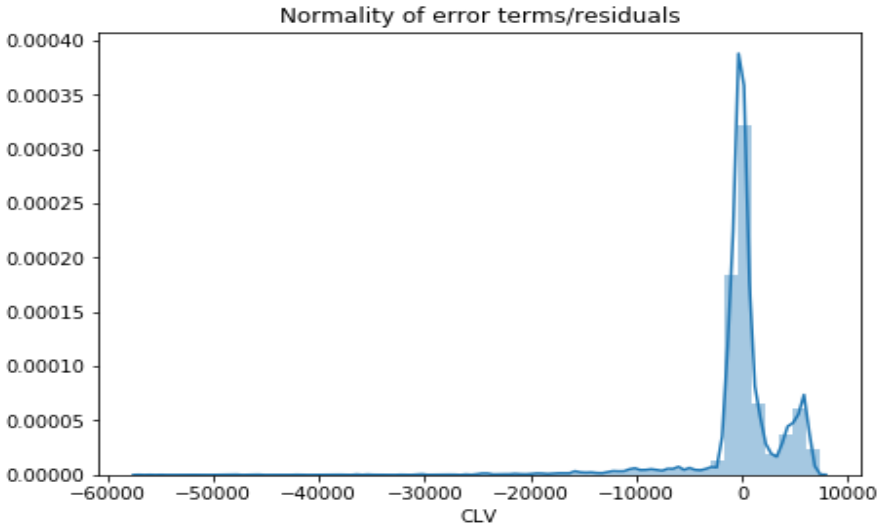
Homoscedasticity_test
We can also use two statistical tests: Breusch-Pagan and Goldfeld-Quandt. In both of them the null hypothesis assumes homoscedasticity and a p-value below a certain level (like 0.05) indicates we should reject the null in favor of heteroscedasticity.
Here, p value is less than 0.05 so, it is homoscedasticity distribution.
$H0{:}\sigma ui\ is\ constant\ across\ the\ range\ of\ data$
$Ha{:}\sigma ui\ is\ not\ constant\ across\ the\ range\ of\ data$

- goldfedquandt test fro homoscadestticity.since prob is much lower than aplha we reject the null hypothesis and accept H1 that residual are hetroscadictic.



Test of normality of residuals

- As we can see p_value less than aphla that means our Null hypothesis is rejected
- Thus we reject the null hypothesis that the error terms are normally distributed.



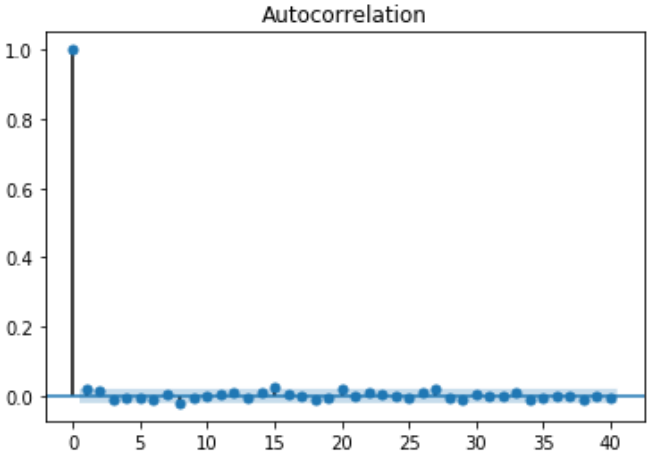
No Autocorrelation:

1) No Auto correlation.

Test needed : Durbin- Watson Test.

It's value ranges from 0-4. If the value of Durbin- Watson is Between 0-2, it's known as Positive Autocorrelation. If the value ranges from 2-4, it is known as Negative autocorrelation. If the value is exactly 2, it means No Autocorrelation. For a good linear model, it should have low or no autocorrelation. we can see here the values of dublin watson test (test for normality): 1.240 (POSITIVE AUTO-CORRELATION)

From the graph below, we can easily see that there is somewhat Positive autocorrelation.



No Multcollinearity:

	vif
State	1.002053
Response	1.054500
Coverage	1.285787
Education	1.008514
EmploymentStatus	2.275529
Gender	1.016600
Income	2.263962
Location Code	1.153212
Marital Status	1.148773
Monthly Premium Auto	2.341140
Months Since Last Claim	1.006257
Months Since Policy Inception	1.015289
Number of Open Complaints	1.003915
Number of Policies	1.010883
Policy Type	4.437370
Policy	4.437098
Renew Offer Type	1.101907
Sales Channel	1.026336
Total Claim Amount	2.557268

A variance inflation factor(VIF) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; it's presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.To check mmulticollinaearity we used Vif approach.

- we can see that there is high VIF in column Policy & policy type
- Also we have seen in statistical apprche these veriable is not significant to predict target
- hence from above 2 conclusion we can simply remove them

RFE (Recusrsive Feature Elimination):

	col	select	rank
0	State_California	True	1
1	State_Nevada	True	1
2	State_Oregon	True	1
3	State_Washington	True	1
4	Response_Yes	True	1
5	Coverage_Extended	True	1
6	Coverage_Premium	True	1
7	Education_College	True	1
8	Education_Doctor	True	1
9	Education_High School or Below	True	1
10	Education_Master	True	1
11	EmploymentStatus_Employed	True	1
12	EmploymentStatus_Medical Leave	True	1
13	EmploymentStatus_Retired	True	1
14	EmploymentStatus_Unemployed	True	1
15	Gender_M	True	1
16	Location Code_Suburban	True	1
17	Location Code_Urban	True	1
18	Marital Status_Married	True	1
19	Marital Status_Single	True	1
20	Number of Open Complaints_1	True	1
21	Number of Open Complaints_2	True	1
22	Number of Open Complaints_3	True	1
23	Number of Open Complaints_4	True	1
24	Number of Open Complaints_5	True	1
25	Number of Policies_2	True	1
26	Number of Policies_3	True	1
27	Policy_Corporate L2	True	1
28	Policy_Corporate L3	False	3
29	Renew Offer Type_Offer2	True	1
30	Renew Offer Type_Offer3	True	1
31	Renew Offer Type_Offer4	True	1
32	Sales Channel_Branch	True	1
33	Sales Channel_Call Center	True	1
34	Sales Channel_Web	True	1
35	Vehicle Class_Luxury Car	True	1
36	Vehicle Class_Luxury SUV	True	1
37	Vehicle Class_SUV	True	1
38	Vehicle Class_Sports Car	True	1
39	Vehicle Class_Two-Door Car	True	1
40	Vehicle Size_Medsize	True	1
41	Vehicle Size_Small	True	1
42	Months_2	True	1
43	Income	False	6
44	Monthly Premium Auto	True	1
45	Months Since Last Claim	False	2
46	Months Since Policy Inception	False	4
47	Total Claim Amount	False	5

Model Building

Before building predictive models we have to transform some of the columns that were object type(categorical) to numerical columns by doing Labelencoder to the columns having binary variables and to the columns having multiple variables such as Vehicle size, Location code , Vehicle type by one hot encoding (dummification) .

We dropped some of the columns which were insignifiatnt afer the statistical test and the variance inflation factor , there were 13 columns that were insignificant and we have build or predictive model based on these columns only.

OLS Regression Results

Dep. Variable:	CLV	**R-squared (uncentered):**	0.845
Model:	OLS	**Adj. R-squared (uncentered):**	0.844
Method:	Least Squares	**F-statistic:**	919.3
Date:	Wed, 15 Jan 2020	**Prob (F-statistic):**	0.00
Time:	11:59:10	**Log-Likelihood:**	-89051.
No. Observations:	9134	**AIC:**	1.782e+05
Df Residuals:	9080	**BIC:**	1.786e+05
Df Model:	54		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
State_California	-148.6878	124.482	-1.194	0.232	-392.700	95.324
State_Nevada	-6.3325	172.300	-0.037	0.971	-344.080	331.415
State_Oregon	-113.5449	129.096	-0.880	0.379	-366.602	139.512
State_Washington	140.1306	178.345	0.786	0.432	-209.466	489.727
Response_Yes	-230.0450	137.626	-1.672	0.095	-499.823	39.733
Coverage_Extended	758.2759	138.272	5.484	0.000	487.231	1029.321

The approach was made by the regresssin algorithms such as Linear regression, Decision tree regressor, Random forest regressor and below are the accuracies were achieved alng with the RMSe(root mean square error value) .

	Models	R2_score	RMSE
0	LinearRegression numerical features only	0.156	6627.47
1	DecisionTreeRegressor numerical features only	0.457	5316.67
2	RandomForestRegressor numerical features only	0.680	4079.66
3	LinearRegression base model	0.154	6634.77
4	DecisionTreeRegressor base model	0.470	5252.48
5	RandomForestRegressor base model	0.692	4004.95

Accuracies with final score:

```
----- LinearRegression with final change -----
r2_score :  0.6259891695519411
RMSE :  4411.719261860843


----- DecisionTreeRegressor with final change -----
r2_score :  0.4903915705911982
RMSE :  5149.725507545851


----- RandomForestRegressor with final change -----
r2_score :  0.675270692023871
RMSE :  4110.801655329676
```

	Models	R2_score	RMSE
0	LinearRegression numerical features only	0.156	6627.47
1	DecisionTreeRegressor numerical features only	0.457	5316.67
2	RandomForestRegressor numerical features only	0.680	4079.66
3	LinearRegression base model	0.154	6634.77
4	DecisionTreeRegressor base model	0.470	5252.48
5	RandomForestRegressor base model	0.692	4004.95
6	LinearRegression after feature engineering	0.159	6615.04
7	DecisionTreeRegressor after feature engineering	0.501	5094.34
8	RandomForestRegressor after feature engineering	0.677	4097.53
9	LinearRegression with final change	0.626	4411.72
10	DecisionTreeRegressor with final change	0.490	5149.73
11	RandomForestRegressor with final change	0.675	4110.80

Predictions based on train and test data :

	Original	Prediction
708	4222.631209	5024.781134
47	5514.344018	4835.969433
3995	3808.122147	3771.965352
1513	7914.823110	7316.552156
3686	7931.722181	13368.093083
...	...	...
4271	4335.353131	4497.301616
7923	9031.214859	9530.591112
5633	5522.524223	7319.479615
8432	5093.479191	5780.405689
681	29194.366390	18133.118483

As we can see the prediction across train and test data is pretty well as it has predicted the original and predicted values pretty much similar ,but there are sme predictions that are predicted much less that are been affected by the outliers as our data is havng a to outlier which are affecting the prediction.

Modelling after Log Transformation :

```
----- LinearRegression with log transformation -----
r2_score :  0.894627472514081
RMSE :  0.21706477770749683

----- DecisionTreeRegressor with log transformation -----
r2_score :  0.8451192259545816
RMSE :  0.2631627490162356


----- RandomForestRegressor with log transformation -----
r2_score :  0.9047960916855896
RMSE :  0.20632556730955307
```

Prediction After Modelling

```
----- LinearRegression -----
-----Log----
r2_score :  0.894627472514081
RMSE :  0.21706477770749683


After Anti Log
RMSE :  4233.170562194246
r2_score :  0.6556500880774305
```

	Original_log	Pred_log	Original_anti_log	Prediction_anti_log
5317	10.217793	9.717615	27386.150240	16607.587519
8196	8.912464	8.947201	7423.929008	7686.344600
5669	8.947419	8.851561	7688.023506	6985.283180
6135	8.105361	8.090104	3312.176781	3262.025808
6244	8.574383	8.543660	5294.283790	5134.102561
6463	8.672166	8.631244	5838.133617	5604.046232
1197	8.471593	8.545351	4777.121596	5142.790086
4689	8.337728	8.327393	4178.586494	4135.622972

Decision Tree with log and Anti-log transformation

```
----- DecisionTreeRegressor -----
-----Log----
r2_score :  0.847390487507917
RMSE :  0.26122603688828266


After Anti Log
RMSE :  4984.1498950610185
r2_score :  0.5226349365172018
```

	Original_log	Pred_log	Original_anti_log	Prediction_anti_log
708	8.348214	8.351182	4222.631209	4235.183999
47	8.615108	8.461121	5514.344018	4727.354548
3995	8.244891	8.244891	3808.122147	3808.122147
1513	8.976493	8.967297	7914.823110	7842.377593
3686	8.978625	9.286972	7931.722181	10796.448320
...	...	...	...	...
4271	8.374558	8.373412	4335.353131	4330.386020
7923	9.108442	9.108442	9031.214859	9031.214859
5633	8.616590	8.601295	5522.524223	5438.696611
8432	8.535716	8.655951	5093.479191	5744.229745

Random Forest Regressor:

Random forest is like bootstrapping algorithm with Decision tree (CART) model. Say, we have 1000 observation in the complete population with 10 variables. Random forest tries to build multiple CART models with different samples and different initial variables. For instance, it will take a random sample of 100 observation and 5 randomly chosen initial variables to build a CART model. It will repeat the process (say) 10 times and then make a final prediction on each observation. Final prediction is a function of each prediction. This final prediction can simply be the mean of each prediction.

```
----- RandomForestRegressor -----
-----Log----
r2_score :  0.9047273682734901
RMSE :  0.20640002244115468


After Anti Log
RMSE :  3977.5611180934407
r2_score :  0.6959799929255935
```

	Original_log	Pred_log	Original_anti_log	Prediction_anti_log
708	8.348214	8.327838	4222.631209	4137.462316
47	8.615108	8.520183	5514.344018	5014.969049
3995	8.244891	8.244891	3808.122147	3808.122147
1513	8.976493	8.964725	7914.823110	7822.227477
3686	8.978625	8.941815	7931.722181	7645.063104
...	...	...	...	...
4271	8.374558	8.338687	4335.353131	4182.593627
7923	9.108442	9.108442	9031.214859	9031.214859
5633	8.616590	8.580008	5522.524223	5324.148114
8432	8.535716	8.622053	5093.479191	5552.776215
681	10.281731	9.997983	29194.366390	21982.073793

Model with Hyper Parameter Tuning:

After doing Hyperparameter tuning by using grid search cv we were able to get a best parameter of

RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=30,max_features='auto', max_leaf_nodes=None,min_impurity_decrease=0.0, min_impurity_split=None,min_samples_leaf=1, min_samples_split=2,min_weight_fraction_leaf=0.0, n_estimators=200,n_jobs=None, oob_score=False, random_state=None,verbose=0, warm_start=False)

```
----- Random forest With Hyperparamter Tuning -----
-----Log----
r2_score :  0.9115126206420755
RMSE :  0.19891444657751878


After Anti Log
RMSE :  3892.5590218925845
r2_score :  0.7088352107348126
```

	Original_log	Pred_log	Original_anti_log	Prediction_anti_log
708	8.348214	8.329271	4222.631209	4143.396451
47	8.615108	8.552959	5514.344018	5182.065762
3995	8.244891	8.243721	3808.122147	3803.668477
1513	8.976493	8.958671	7914.823110	7775.015897
3686	8.978625	9.054293	7931.722181	8555.186725
...	...	...	...	...
4271	8.374558	8.346719	4335.353131	4216.322966
7923	9.108442	9.105928	9031.214859	9008.536291
5633	8.616590	8.572350	5522.524223	5283.530529
8432	8.535716	8.640029	5093.479191	5653.492708
681	10.281731	9.892384	29194.366390	19779.159561

Regularisation Techniques

In order to create less complex (parsimonious) model when you have a large number of features in your dataset, some of the Regularization techniques used to address over-fitting and feature selection are:

1. **L1 Regularization**

Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds "*absolute value of magnitude*" of coefficient as penalty term to the loss function.

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

2. **L2 Regularization**

Ridge regression adds "*squared magnitude*" of coefficient as penalty term to the loss function. Here the *highlighted* part represents L2 regularization element.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

The **key difference** between these techniques is that Lasso shrinks the less important feature's coefficient to zero thus, removing some feature altogether. So, this works well for **feature selection** in case we have a huge number of features.

```
----- Lasso Regression -----
-----Log----
r2_score :  0.1959012872686573
RMSE :  0.5996258438178986


After Anti Log
RMSE :  6916.048219381348
r2_score :  0.08085407383799703
```

:

	Original_log	Pred_log	Original_anti_log	Prediction_anti_log
708	8.348214	8.844981	4222.631209	6939.472987
47	8.615108	8.585159	5514.344018	5351.642139
3995	8.244891	8.755899	3808.122147	6348.025852
1513	8.976493	8.793017	7914.823110	6588.076061
3686	8.978625	8.570312	7931.722181	5272.773277
...	...	...	...	...
4271	8.374558	8.859828	4335.353131	7043.271938
7923	9.108442	8.963757	9031.214859	7814.662503
5633	8.616590	9.067686	5522.524223	8670.537013
8432	8.535716	8.585159	5093.479191	5351.642139
681	10.281731	8.993451	29194.366390	8050.190559

```
----- Ridge Regression -----
-----Log----
r2_score :  0.8945851162897012
RMSE :  0.2171083997070356


After Anti Log
RMSE :  4235.84145521244
r2_score :  0.6552154200627774
```

	Original_log	Pred_log	Original_anti_log	Prediction_anti_log
708	8.348214	8.335108	4222.631209	4167.650397
47	8.615108	8.548413	5514.344018	5158.559472
3995	8.244891	8.168925	3808.122147	3529.548031
1513	8.976493	8.860777	7914.823110	7049.959363
3686	8.978625	9.254619	7931.722181	10452.730857
...	...	...	...	...
4271	8.374558	8.297456	4335.353131	4013.647302
7923	9.108442	9.139431	9031.214859	9315.460710
5633	8.616590	8.610560	5522.524223	5489.322622
8432	8.535716	8.644717	5093.479191	5680.061569
681	10.281731	9.878903	29194.366390	19514.308895

Final Prediction :

	Models	R2_score	RMSE
0	LinearRegression numerical features only	0.156	6627.47
1	DecisionTreeRegressor numerical features only	0.457	5316.67
2	RandomForestRegressor numerical features only	0.680	4079.66
3	LinearRegression base model	0.154	6634.77
4	DecisionTreeRegressor base model	0.470	5252.48
5	RandomForestRegressor base model	0.692	4004.95
6	LinearRegression after feature engineering	0.159	6615.04
7	DecisionTreeRegressor after feature engineering	0.501	5094.34
8	RandomForestRegressor after feature engineering	0.677	4097.53
9	LinearRegression with final change	0.626	4411.72
10	DecisionTreeRegressor with final change	0.490	5149.73
11	RandomForestRegressor with final change	0.675	4110.80
12	LinearRegression	0.656	4233.17
13	DecisionTreeRegressor	0.523	4984.15
14	RandomForestRegressor	0.696	3977.56
15	Random forest With Hyperparamter Tuning	0.709	3892.56
16	Lasso Regression	0.081	6916.05
17	Ridge Regression	0.655	4235.84

CONCLUSION:

- In this study with the purpose if determining customer lifetime value based on the benefit segmentation, CLV was calculated for each group of customers of an insurance company for the first time, and finally the gold customers who had the highest profit rate for the organization were determined.

- CLV approach enables us to realize the goals of market segmentation, since development of close relationships with those customers who are profitable or potentially profitable for the organization is a key goals of customer relationship management in market segmentation, that can be realized through the concept of customer lifetime value.

- Based on this segmentation and identification of the customers, the company can develop appropriate programs to retain existing customers and to identify and attract new customers. Therefore, the guaranteed high performance of the organizations must be sought in a successful customer relationship management, and for

Achieving these objectives of the organization the task should begin with those who are associated directly with customers. Finally, an appropriate marketing programs, price discounts, and rewards for profitable customer are strategies to attract and retain customers.

REFERENCES:

Indian Journal of Science and Technology, *Vol 9(1), DOI: 10.17485/ijst/2016/v9i1/72307, January 2016*

Elham Farzanfar and Narges Delafrooz*
Department of Business Management, Rasht Branch, Islamic Azad University, Rasht, Iran