**1. Objective**

The lifetime value of a customer, or customer lifetime value (CLV), represents the total amount of money a customer is expected to spend in business, or on products, during their lifetime. This is an important figure to know because it helps company to make decisions about how much money to invest in acquiring new customers and retaining existing ones.

**2. Data Understanding**

1. There are 9134 Observations of 24 Variable
2. There are mix of categorical and continous DataType.
3. Dependent Variable is Customer Life Time Value as we have to predict the CLV.
4. Independent Variables are: Customer, StateCustomerLifetimeValue, Response, Coverage, Education, EffectiveToDate, EmploymentStatus, Gender, Income, LocationCode, MaritalStatus, MonthlyPremiumAuto, MonthsSinceLastClaim, MonthsSincePolicyInception, NumberofOpenComplaints, NumberofPoliciesPolicyType, Policy, RenewOfferType, SalesChannel, TotalClaimAmountVehicleClass, VehicleSize
5. Continues Independed Variables are : CustomerLifetimeValue, Income,MonthlyPremiumAuto, MonthsSinceLastClaim, MonthsSincePolicyInception, NumberofOpenComplaints, NumberofPolicies, TotalClaimAmount
6. There are no null values, so no further action required to replace missing or null values.
7. “Customer” column is serial number so it is insignificat for analysis and removed from the dataset.

**3. Exploratory Data Analysis (EDA)**

**1. Descriptive Analysis of CustomerLifetimeValue**

1. Maximum CLV is 83325.381𝑎𝑛𝑑𝑡ℎ𝑒𝑚𝑖𝑛𝑖𝑚𝑢𝑚𝐶𝐿𝑉𝑖𝑠83325.381andtheminimumCLVis1898.008.
2. Mean of CLV is 8005𝑎𝑛𝑑𝑡ℎ𝑒𝑀𝑒𝑑𝑖𝑎𝑛𝑖𝑠8005andtheMedianis5780.
3. The Variance in CLV is 47210196 and the Standard Deviation is 6870.968.
4. Skewness is 4.031284. CLV is positive skewed and most values are concentrated on the left of
5. the mean value, yet all the extreme values are on the right of the mean value.
6. Kurtosis is 13.81163. Since kurtosis > 3, means distribution has thicker tails than normal
7. distribution and have more outliers (extreme values).
8. This means that the distribution of CLV is positively skewed (as expected) and is heavily
9. Leptokurtic.
10. These results indicate a distribution that is heavily skewed with a very large tail.
11. There are a LOT of Customers with low CLV. Very few customers with high CLV.
12. This can be visually understood using the Histogram.

**2. Descriptive Analysis of Monthly Premium Auto(MPA)**

1. Maximum MPA is 298 and the minimum MPA is 61
2. Mean of MPA is 93.21929 and the Median is 84.00
3. The Variance in MPA is 1183.908 and the Standard Deviation is 34.40797
4. Skewness is 2.122849. MPA is positive skewed and most values are concentrated on the left of
5. the mean value, yet all the extreme values are on the right of the mean value.
6. Kurtosis is 6.187546. Since kurtosis > 3, means distribution has thicker tails than normal
7. distribution and have more outliers (extreme values).
8. There is a Positive Corelation of 39.62 % of MPA with CLV. From scatter plot, it is clearly
9. visible that on MPA, CLV is also Increasing.7.
10. Monthly premiums follow a trend similar to CLV although the distribution is NOT as skewed
11. or as long tailed as CLV. This can be visually seen in the Histogram.

**3. Descriptive Analysis of TotalClaimAmount (TCA)**

1. Maximum TCA is 0.099007𝑎𝑛𝑑𝑡ℎ𝑒𝑚𝑖𝑛𝑖𝑚𝑢𝑚𝑇𝐶𝐴𝑖𝑠0.099007andtheminimumTCAis2893.239678
2. Mean of TCA is 434.0888𝑎𝑛𝑑𝑡ℎ𝑒𝑀𝑒𝑑𝑖𝑎𝑛𝑖𝑠434.0888andtheMedianis383.945
3. The Variance in TCA is 84390.3 and the Standard Deviation is 290.5001
4. Skewness is 1.714403. TCA is positive skewed and most values are concentrated on the left of
5. the mean value, yet all the extreme values are on the right of the mean value.
6. Kurtosis is 5.973506. Since kurtosis > 3, means TCA distribution has thicker tails than normal
7. distribution and have more outliers (extreme values).
8. There is a Positive Corelation of 22.65 % of TCA with CLV. From scatter plot, it is clearly
9. visible that on TCA, CLV is also Increasing.
10. Total Claim amounts also follow a trend similar to CLV and MPA although the distribution is
11. NOT as skewed or as long tailed as MPA. This can be visually seen in the Histogram.

**This means that variation in data is CLV > MPA > TCA**

**4. Descriptive Analysis of other variables.**

The positive correlation values close to zero show that that there is no strong relationship of Income, MonthsSinceLastClaim, NumberofPolicies etc with CLV.

**4. Inferential Statistics**

**1. Effect of Insurance Coverage on Customer Life Time Value (CLV)**

Customers who have taken Basic Insurance for their vehicals are more valuable then Extended or Premium Insurance Policy holders.

**2. Effect of Education on Customer Life Time Value (CLV)**

Educated customers (with a bachelors or equivalent degree) are more valuable than others.

**3. Effect of Employment Status on Customer Life Time Value (CLV)**

Employed customers are more valuable than others as compared to Retired, Unemployed or Disabled Customers.

**4. Effect of Gender on Customer Life Time Value (CLV)**

Gender has no role to play in determining the value of a customer. Both Male and Female looks valuable.

**5. Effect of Location on Customer Life Time Value (CLV)**

Rural customers are LESS valuable than Urban customers.

**6. Effect of Marital Status on Customer Life Time Value (CLV)**

Married customers are buying more auto insurance and adding more value to company.

**7. Effect of Policy Type on Customer Life Time Value (CLV)**

Customers having their own Personal Policy are more valuable to company then Corporate and Special Insurance policy holder.

1. **Effect of Renew Offer Type on Customer Life Time Value (CLV)**

Offers 1 and Offer 2 attracts more customers.

**9. Effect of Sales Channel on Customer Life Time Value (CLV)**

Call Center is not performing well compared to other channels throughout the country (in terms of high value customers)

**10.Effect of Vehicle Class on Customer Life Time Value (CLV)**

Customers having Four-Door car and SUV are more valuable.

**11.Effect of Vehicle Size on Customer Life Time Value (CLV)**

Customers having Mid Size vehicles are adding more value to Insurance company.

**12.Effect of States on Customer Life Time Value (CLV)**

California and Oregon customers are more valuable.

**13.Effect of Policy on Customer Life Time Value (CLV)**

Personal L3 Policy is adding more value to company.

**5. Regression Analysis with Continuous Variables**

1. Dependent Variable CLV is continuous and we have seen that independent variables are mostly depending linearly with dependent Variable, So Linear Regression algorithm is best for this type of this Data.
2. The goal of the Linear regression is to find the best fit line that can accurately predict the output for the continuous dependent variable.
3. Removing qualitative variables because Linear Regression works best when variables are quantitative/numeric in nature. We have only 8 continuous independent variables.
4. **Model Interpretation**

1. Split the Data in Training and Testing Set.

2. Training Dataset is for building Model and Testing Dataset to test the model on unlabeled data.

3. Build Linear Regression Model using all the continoues independent variables.

4. Analysis the significance of the independent variable and if require rerun the model.

**Null Hypothesis** - None of the independed variables are significant for CLV.

**Alternate Hypothesis** - At least one of the independent variables are significant and can effect the CLV.

1. p-value of model is less than 0.05, so atleast one of the independent variables are significant.
2. p-value of MonthlyPremiumAuto, NumberofOpenComplaints and NumberofPolicies are less then 0.05, so rejecting the null hypothesis. So atleast one of them independent variables are significant and can effect the CLV.
3. However R squared is very low, only 16.02% of the variance found in the CLV can be explained by Income, MPA, MonthsSinceLastClaim, MonthsSincePolicyInception, NumberofOpenComplaints, NumberofPolicies, TCA.
4. Adjusted R squared is 0.1537 which is less than R squared.
5. Residual standard error is 6322 which is very very high, so it means the actual CLV will deviate from the true regression line by approximately 6322 on an average. The smaller the standard error, the less the spread and the more likely it is that any sample mean is close to the population mean. A small standard error is thus a Good Thing.
6. Gap between R-squared and Adjusted R-squared is 1.4% only, which is good. Typically the more non-significant variables you add into the model, the gap between two increases.
7. F-statistic: 6.958 - The lower the F-statistic, the closer to a non-significant model. So F-statistic is low means it is not very significant model.

Rerun Model

There are more than one insignificant variables in the model, so need to run the model again with only significant variables.

**The estimated regression line equation can be written as follow:**

**CLV = 486.82 + 82.22 MPA – 247.02 NoOC + 80.63 NoP - 0.665 TCA**

New Model Interpretation

Null Hypothesis - None of the independed variables are significant for CLV.

Alternate Hypothesis – At least one of the independent variables are significant and can effect the CLV.

1. p-value of MonthlyPremiumAuto, NumberofOpenComplaints, NumberofPolicies and TotalClaimAmount is less than 0.05 so they are significantly impact the CLV.
2. Coefficients of Independent Variables :-

i. MonthlyPremiumAuto : 86.4478. One unit increase in MonthlyPremiumAuto will increase CLV by 86.4478

ii. NumberofOpenComplaints : -199.3526. One unit increase in NumberofOpenComplaints will decrease CLV by 199.3526

iii. NumberofPolicies : 76.3861. One unit increase in NumberofPolicies will increase CLV by 76.3861

iv. TotalClaimAmount : -1.0445. one unit increase in TotalClaimAmount will decrease by 1.0445

1. So the customers having more number of policies with high monthly premium will add more value to company.
2. On the other hand, customer's Open Complaints and More Claim Amount will decrease the CLV.
3. R squared is 0.1656 which means 16.56% of dependent variable is explained by independ variable.
4. Adjusted R squared is 0.1652 which is less than R squared.

**Predict the value of CLV for all observations based on the above calculated regression model.**

Calculate Error; Difference between actual CLV and predicted CLV

Average error rate of model is 60.43%, which is high and we can say that model is not so good.

**2. Residuals Analysis**

Check the normality of Error/Residual Term (Linear Regression assumes that error are normally distributed.)

Null Hypotheses - Errors are normally distributed.

Alt Hypothese - Errors are not normally distributed.

p-value(0.00837) < 0.05, Null Hypotheses get rejected, and so the errors are not normally distributed.

**6. Assumption Testing of Linear Regression Analysis**

**1. Detecting multicollinearity**

In our model, only those independent variable should exist which are not correlated with each other. This is done using Correlation Matrix.

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.

If there is high correlation between two independed variables (high multicollinearity), then you will not be able to seperate out the impact of individual independed variable on depended variable.

Due to multicolinearity we can't define the complete impact of only one independed variable on the depended variable.

**MonthlyPremiumAuto**

1.67728411765984

**NumberofOpenComplaints**

1.00023669427589

**NumberofPolicies**

1.0002991479061

**TotalClaimAmount**

1.67692535683477

**2. Detecting Homoscedasticity**

Null Hypothesis - Homoscedasticity is present in Residuals.

Alternate hypothesis - Heteroskedasticity is present in residuals.

This is done by Breusch-Pagan test.

p-value < 0.05, so it rejects that errors are homoscedasticity. So errors terms are heteroscedasticity and does not have constant variance which is not good for model.

**3. Detecting Autocorrelation**

This is done Durbin-Watson Test If D-W Statistic is around 2, then we have autocorrelation in model. and away from 2 means no autocorrelation.

Here D-W Statistic is 2.01194, so there is autocorrelation in the model.

**4. Detecting MAPE**

MAPE computes the average absolute percent difference between two numeric vectors.

Average error rate of model is 60.43%, which is high and we can say that model is not so good.

**7. Prediction Curve**

**8. Summary**

1. There are a lot of Customers with low CLV. Very few customers with high CLV.
2. Customers who have taken Basic Insurance for their vehicle are more valuable then Extended or Premium Insurance Policy holders.
3. Educated Employed customers (with a bachelors or equivalent degree) are more valuable than Retired, Unemployed or Disabled Customers.
4. Gender has no role to play in determining the value of a customer. Both Male and Female looks valuable.
5. Marital customers are buying more auto insurance and adding more value to company.
6. Rural customers are LESS valuable than Urban customers.
7. Customers having their own Personal Policy are more valuable to company then Corporate and Special Insurance policy holder.
8. Offers 1 and Offer 2 attracts more customers.
9. Call Center is not performing well compared to other channels throughout the country (in terms of high value customers)
10. Customers having Mid Size vehicles, Four-Door car or SUV are more valuable.
11. California customers are adding more value to the company.
12. Personal L3 Policy is adding more value to company.xiii.
13. The customers having more number of policies with high monthly premium will add more value to company. On the other hand, customer's Open Complaints and More Claim Amount will decrease the CLV.

**9. Business Recommendation**

This report represents my analysis for the XYZ Insurance company. It is my opinion that based on the data provided, targeting appropriate customers could increase the Customer Lifetime Value. The two proposed changes are as follows:

A) Insurance company should target educated married employed customers from Urban areas having Mid Size vehicles to increase the Customer Lifetime Value (CLV) increase.

B) On the other hand, if customer's Open Complaints would not be resolved soon and claim amount would not bring down, then both could decrease the Customer Lifetime Value (CLV).

C) About 38% value was added by the agents to the company whereas call centers added only 20% value. So agents should be preferred over call centers while selling the auto insurance to customers.

D) Factors which are responsible for increasing the CLV are Monthly Premium and Number of Policies, however Open Complaints and Claim Amount can decrease the CLV.