**Final Project**

**Multiple Linear Regression Modelling in R**

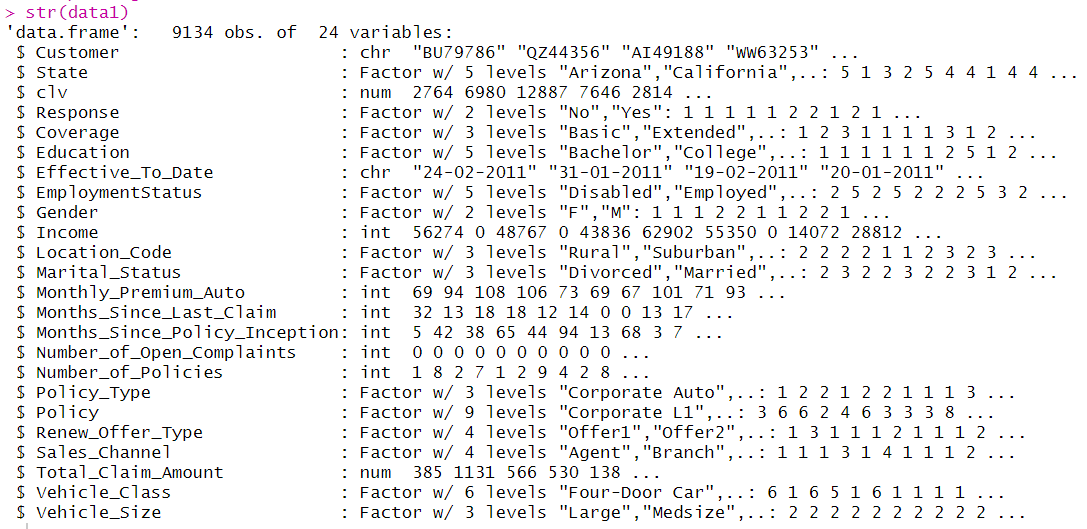
**Submitted by: Mansi**

* **Business Problem:**

Customer Lifetime Value (CLV) is defined as the total revenue that the client will derive from their entire relationship with a customer. The purpose of this project is to forecast the value of CLV based on the significant variables out of total 23 independent variables.

* **About the Data:**

The data provided has 9134 data entries of 24 different variables where 'Customer Lifetime Value' or CLV is the dependent variable and rest are independent variables. Using the str() function in R we can tell which variables are continuous and categorical. The variables of the form chr, int and num are continuous variables and those in factor form are categorical.

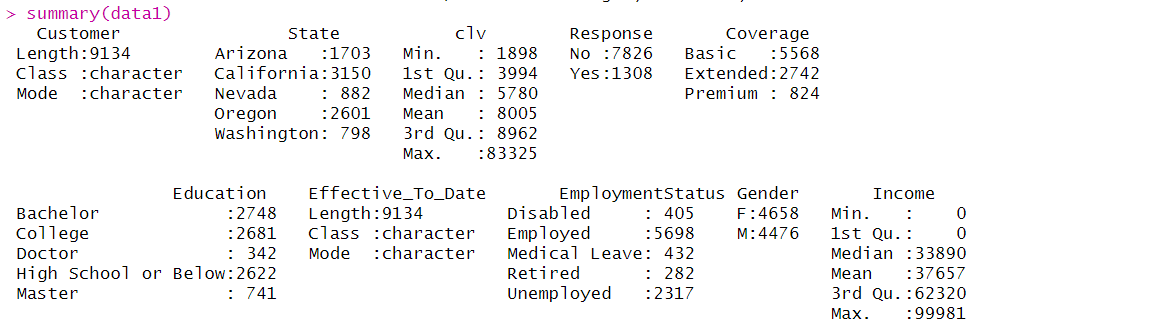


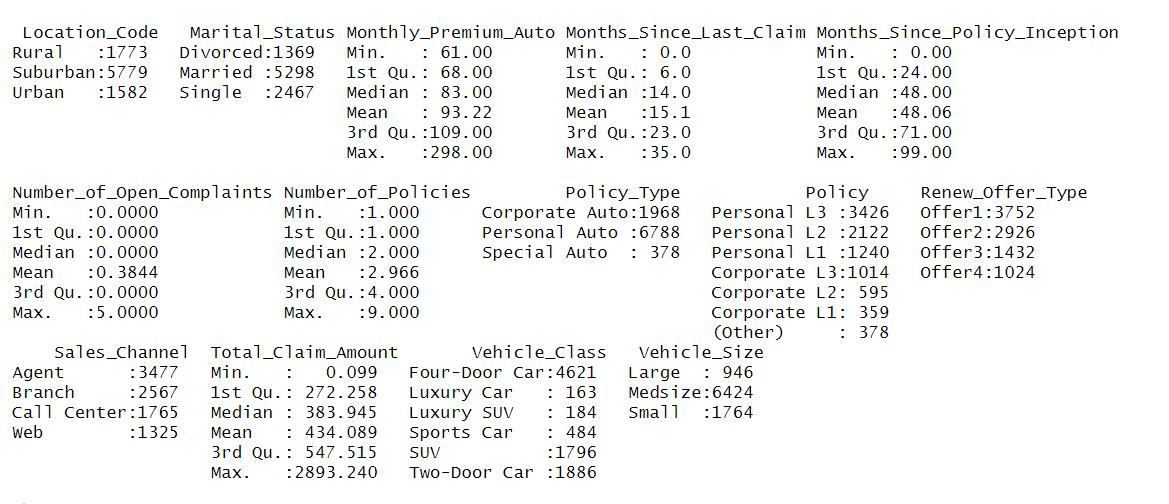
* **Independent Variables**: Customer, StateCustomerLifetimeValue, Response, Coverage, Education, EffectiveToDate, EmploymentStatus, Gender, Income, LocationCode, MaritalStatus, MonthlyPremiumAuto, MonthsSinceLastClaim, MonthsSincePolicyInception, NumberofOpenComplaints, NumberofPoliciesPolicyType, Policy, RenewOfferType, SalesChannel, TotalClaimAmountVehicleClass, VehicleSize
* **Continous Independent Variables**: CustomerLifetimeValue, Income,MonthlyPremiumAuto, MonthsSinceLastClaim, MonthsSincePolicyInception, NumberofOpenComplaints, NumberofPolicies, TotalClaimAmount

Using is.null() function we find out that there are no null values, so no further action required to replace missing or null values.

Summary() function in R gives us the fundamental statistic values of each variable can be found out. For categorical variables, the count of each level is displayed, whereas for continuous variables, mean, median, mode, maximum value, minimum values are obtained.

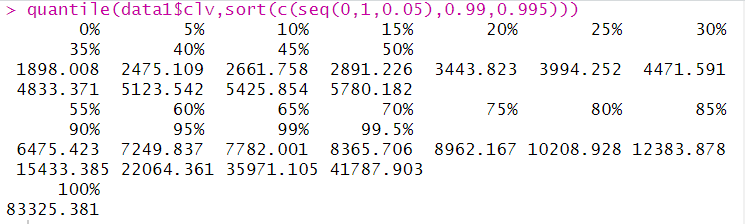
* From this, one can infer that the company has least customers in Washington and it is doing good in states like California and Oregon.
* Also, there is not a wide gap between the number of male and female customers.
* It can also be observed that the buying insurance cover is quite common among students of college or Bachelor as compared to Masters or Doctors students.
* The mean and median of variables such as income and total claim amount are not much varying, thus their means can be relied upon.



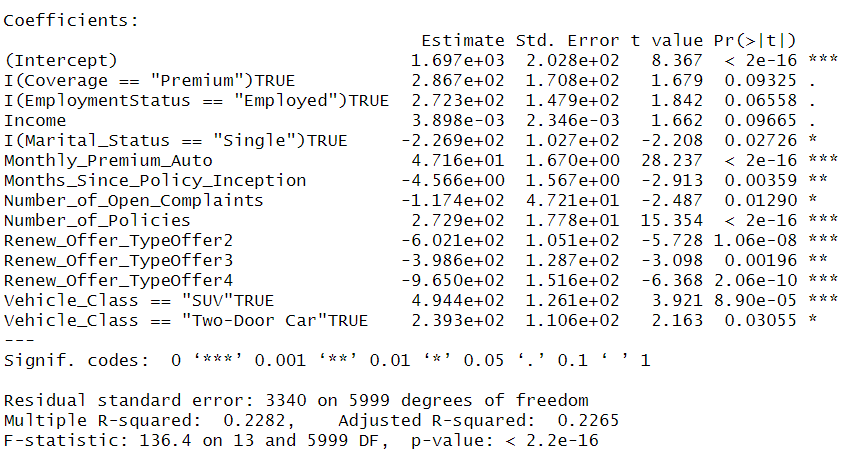


* **Outlier Treatment:**

The “clv” variable has very distinguished mean and median, which points out to the fact that there are outliers in the data. So, the data requires outlier treatment. Here, outliers are treated by capping them after identifying them from quantiles using quantile() function in R.



* **Regression Modelling Methodology**
* The data is first divided into training and validate data in the ratio of 70:30.
* Model is first prepared on the training data with all variables.
* Multiple iterations are performed iterations to remove the insignificant variables at 90% confidence level i.e. p=0.10
* The insignificant variables are removed from the model until the model has all significant variables.
* Various diagnostic tests are performed on the final model to check its explanatory power and efficiency.
* Final model on train data looks like the one given below.



* **Model Interpretation**

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

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

For a test to come out productive, there should be some relation established between the dependent and independent variable. Thus, we need to reject the Null Hypothesis. This can be achieved by a low p value.

On looking at the p values obtained and performing iterations, finally, 10 out of 23 variables with particular levels come out to be significant from the train data.

Model Parameters:

* p-value of model is less than 0.05, so at least one of the independent variables are significant.
* R squared value of the model is very low, only 22.82% of the variance found in the CLV can be explained by the significant variables.
* Adjusted R squared is 0.2265 which is less than R squared.
* Gap between R-squared and Adjusted R-squared is low, which is good for the model.
* F-statistic value is 136.4 implies that the model is a good fit.

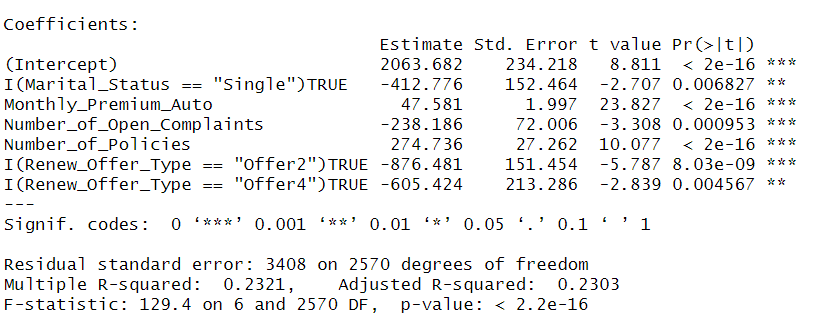
Variable parameters:

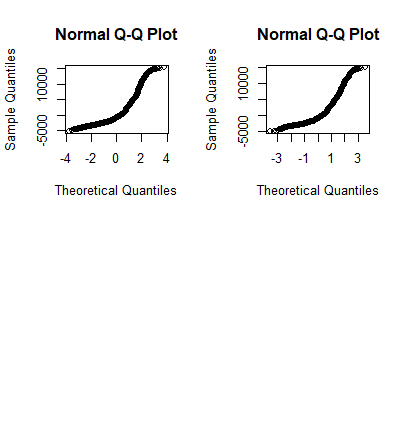
**Null Hypothesis** - None of the independent variables are significant for CLV.

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

* Variables marked with 3 stars are significant at 100% confidence level, implied by their very low p value.
* MonthlyPremiumAuto: One unit increase in MonthlyPremiumAuto will increase CLV by 47.1
* NumberofOpenComplaints : One unit increase in NumberofOpenComplaints will decrease CLV by 117.4
* NumberofPolicies: One unit increase in NumberofPolicies will increase CLV by 272.9
* So, 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 will decrease the CLV.
* **Model Validation**

Rerunning the model on validation data provides us the following parameters.





Comparison of plots on train and validation data respectively.

* **Testing the efficiency of model**

1. Test for multi collinearity: In our model, only those independent variables should exist which are not correlated with each other. This is done using Correlation Matrix or VIF. 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 independent variables (high multicollinearity), then you will not be able to separate out the impact of individual independed variable on depended variable.

VIF of our model is less than 2.5 therefore no multi collinearity.

1. Test for Homoscedasticity: This is measured by Breusch-Pagan test.

**Null Hypothesis** - Homoscedasticity is present in Residuals.

**Alternate hypothesis** - Heteroskedasticity is present in residuals.

If p-value < 0.05, we reject that errors are homoscedasticity. So, errors terms are heteroscedasticity and do not have constant variance which is not good for model.

1. Test for Autocorrelation**:** This is found using Durbin-Watson Test. If D-W Statistic is around 2, then we have autocorrelation in model. and away from 2 means no autocorrelation. In our model, we have obtained a value of 0.316 which implies that it is a good model.
2. **Detecting MAPE**: Mean Absolute Percentage Error Loss. It computes the average absolute percent difference between two numeric vectors. MAPE value is 0.4277 which is pretty low signifying the good fit of the model

* **Summary:**

1. Customers who took Basic Insurance for their vehicle seem more valuable than Extended or Premium Insurance Policy holders.
2. Marital customers have bought more auto insurance and adding more value to company.
3. Educated Employed customers (with a bachelors or college degree) seem more valuable than Retired, Unemployed or Disabled Customers.
4. Gender played no play in determining the value of a customer. Both Male and Female looks valuable.
5. Rural customers seem to be less valuable for the company than Urban customers.
6. Customers with their own Personal Policy are more valuable to company than Corporate and Special Insurance policy holder.
7. Customers having Mid-Size vehicles, Four-Door car or SUV are more valuable.
8. California customers are adding more value to the company.
9. The customers having a greater number of policies with high monthly premium will add more value to company. On the other hand, customer's Open will decrease the CLV.

* **Business Recommendations:**

The proposed recommendations are as follows:

1. Insurance company should target educated and married employed customers from Urban areas having Mid-Size vehicles to increase the Customer Lifetime Value (CLV) increase.
2. On the other hand, if customer's Open Complaints would not be resolved soon, then both could decrease the Customer Lifetime Value (CLV).
3. Agents should be preferred over call centers while selling the auto insurance to customers.
4. Factors which are responsible for increasing the CLV are Monthly Premium and Number of Policies, however Open Complaints can decrease the CLV.