**Project on R**

**Predict Customer Life-time Value for an Auto Insurance Company**

* **Approachs :** In the given dataset Customer Life-time Value is dependent variable and 23 independent variables. According to the problem statement to predict CLV we have to apply the Multiple Linear Regression Model in R.
* **Steps :**
  + First we have to call all the required libraries of Linear Regression Approach
  + We have to set our working directory
  + Now we load the given dataset
  + Checking the structure(data type) of the variables and Converting data types of the variables which are necessary
  + Checking for outliers of the numerical variables. If outlier exists remove these records from the dataset
  + Now check the missing values (if any). If exists remove the records containing missing values
  + Now we start Linear Modeling part. We start it by removing Customer and Effective.To.Date variable as they have no business value.then we run our model and check summary
  + Now we start model validation part. First we check ANOVA it should be less than 0.05 . Then we check p value of the each variables. Every variable should have one star(\*) or more than one star beside them. If any variable did not have any \* beside it then remove the variable from the model but we have to remove one variable at a time. We have to repeat the process until all the variables have \* beside it. Now , we check R squre and Adjusted R squre . Then we Calculate MAPE.
  + Now we start Assumption Diagnostic part. i) There should not any multicollinearity between any two or more independent variables. We test it with VIF, value of GVIF for each variable should be less than 2. ii) For any two observations the residual terms should be uncorrelated, we test it by Durbin-Watson test, should get a high p value , value close to 2 means the residuals are uncorrlated and away from 2 means the residuals are correlated. iii) Next we check Homoscedasticity by Breusch-Pagan test. iv) Then we run Anderson-Darling test for checking residuals are normally distributed or not.
* **Results found and their interpretation** : After doing the above mentioned steps I found the following results
  + After running the linear model first checking ANOVA , I found p value is less than 0.05 , which means that at least one of the co-efficient of not zero, so we can go with this dataset.
  + Next we check each independent variables p value , we kept those independent variables which has p value less than 0.05( Detete one independent variable at a time) , which means co-efficient of that variables are not zero.
  + Next we check **R squre** which is **0.92** and **Adjusted R squre** which is **0.9198**. Both the values close to 1 means better the model.
  + I found the value of **MAPE is 0.1028** , Lesser the MAPE better the model (MAPE ranges between 0 to 1).
  + To check for any two observations the residual terms are uncorrelated or not we use **Durbin-Watson test**. Get the **p value 0.764**, which means residual terms are uncorrelated.
  + Next we check Homoscedasticity by **Breusch-Pagan test** , Get the **p value less than 0.05** , which means it is Heteroscedastic.
  + Next we check Residuals are normally distributed or not by **Anderson-Darling test** , we get the **p value less than 0.05** , which means residuals are not normally distributed.
* **Significance of the variables and their business meaning :**

In the given dataset we have 1 dependent and 23 independent variables. After running Linear Model we found 7 independent variables are significant for the model. Let us discuss about them

* + Variable name – **Education :** It is a dummy variable with base “Bachelor” and from our model we found that if customers education is “Master” then CLV will be more than “Bachelor”. Education “Doctor” Will add more CLV than “Master”. Education “College” and “High school and below” will add more CLV than “Doctor”.
  + Variable name – **Gender :** It is a dummy variable with base “F” and from our model we found that if Gender is Male then CLV will be less than the CLV for Female.
  + Variable name – **Income :** It is a numerical variable. The customers whose income is more , CLV will be more for them.
  + Variable name – **Marital.status :** It is a dummy variable with base “Divorced”. CLV of Married persons will more than Divorced but CVL of Single will be less than the CLV of divorced.
  + Variable name – **Monthly.Premium.Auto :** It is a numerical variable. For more counts CLV will be more.
  + Variable name – **Number of open complaints :** The customers who complaints more, CLV will be less for them.
  + Variable name – **Number of Policies :** The customers who have more than 1 policies, CLV will be more for them.