**About**

Customer lifetime value (CLV) is an important metric for insurance companies because it helps them to estimate the profitability of a customer over the course of their relationship with the company. By understanding the CLV of a customer, an insurance company can make more informed decisions about how to allocate resources, such as marketing and customer service efforts, to maximize revenue. Additionally, CLV can also be used to identify and target high-value customers, who may be more likely to purchase additional products or services, or to renew their policies. By focusing on these customers, an insurance company can improve its retention rates and overall financial performance.

**Problem Statement**

VahanBima is one of the leading insurance companies in India. It provides motor vehicle insurances at best prices with 24/7 claim settlement. It offers different types of policies for both personal and commercial vehicles. It has established its brand across different regions in India.

Around 90% of the businesses today use personalized services. The company wants to launch different personalized experience programs for customers of VahanBima. The personalized experience can be dedicated resources for claim settlement, different kinds of services at doorstep, etc. Inorder to do so, they would like to segment the customers into different tiers based on their customer lifetime value (CLTV).

Inorder to do it, they would like to predict the customer lifetime value based on the activity and interaction of the customer with the platform. So, as a part of this challenge, your task at hand is to build a high performance and interpretable machine learning model to predict the CLTV based on the user and policy data.

**About the Dataset**

The sample dataset of the company holding the information of customers and policy such as highest qualification of the user, total income earned by a customer in a year, employee status, policy opted by the user, type of policy and so on and the target variable indicating the total customer lifetime value.

**Approach**

I have started my approach by exploring the train dataset.

Performed various operations on top the train dataset

1. Checking the shape of the dataset.

2. Checking the information of the dataset.

3. Checking null value and duplicate values.

As part of my observation initial observation

* There are no null values in given training dataset.
* There are no duplicate values in the dataset.

As a next step I divided Numerical and categorical features and started Exploratory data analysis.

**Analysis**

**Univariate analysis**

* By exploring the numerical features

***Observation***

* + Vintage and Cltv features are r*ightly skewed.*
  + And I Observed there are many zeros in claim\_Amount feature which effects the model performance heavily and in some cases leads to biasing,
  + And observed some outliers in Target feature.
  + We can convert them into normal distribution by scaling them.
* By exploring the categorical features

***Observation***

* + Features are mostly ordinal variables.
  + Income, num\_policies, policy features are slightly imbalanced.

**Bivariate analysis**

* Bivariate analysis of numerical features with target feature

***Observation***

* + Claim\_amount started gradually decresing with the increase of cltv in a small amount.
  + And the features are very very low correlation.
  + There is no strong Correlation between Dependent and independent features.

**Model Building**

* I tried to train the dataset with different Regression models.
  + LinearRegression, LassoCV,RidgeCV, DecisionTreeRegressor,
  + RandomForestRegressor, GradientBoostingRegressor
* After training all the models GradientBoostingRegressor Performed better than other models.

**Results**

* GradientBoostingRegressor train accuracy score 0.16599203751611236
* GradientBoostingRegressor test accuracy score 0.1623254291092605