**Predict CLTV of a customer**  
  
Can you help the company identify the most potential customers?

**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**  
  
You are provided with 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.  
  
  
**Data Dictionary**  
  
You are provided with 3 files - train.csv, test.csv and sample\_submission.csv  
  
**Train Set**  
  
You are provided with around 90K records containing the attributes of the user and policy and the target variable *cltv* indicating the total customer lifetime value.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| id | Unique identifier of a customer |
| gender | Gender of the customer |
| area | Area of the customer |
| qualification | Highest Qualification of the customer |
| income | Income earned in a year (in rupees) |
| marital\_status | Marital Status of the customer {0:Single, 1: Married} |
| vintage | No. of years since the first policy date |
| claim\_amount | Total Amount Claimed by the customer (in rupees) |
| num\_policies | Total no. of policies issued by the customer |
| policy | Active policy of the customer |
| type\_of\_policy | Type of active policy |
| *cltv* | *Customer life time value (Target Variable)* |

**Test Set**  
  
You are provided with around 60K records containing only the attributes of the user and policy and you need to predict the target variable *cltv* indicating the total customer lifetime value.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| id | Unique identifier of a customer |
| gender | Gender of the customer |
| area | Area of the user |
| qualification | Highest Qualification of the customer |
| income | Income earned in a year (in rupees) |
| marital\_status | Marital Status of the customer {0:Single, 1: Married} |
| Vintage | No. of years since the first policy date |
| claim\_amount | Total Amount Claimed by the customer (in rupees) |
| num\_policies | Total no. of policies issued by the customer |
| policy | Active policy of the customer |
| type\_of\_policy | Type of active policy |

**Submission File Format**  
  
The solution file must contain the format similar to that of sample submission.**sample\_submission.csv**contains 2 variables - id and cltv. 

|  |  |
| --- | --- |
| **Variable** | **Description** |
| id | Unique identifier of a customer |
| cltv | Customer life time value |

**Evaluation metric**  
  
The evaluation metric for this hackathon would be r2\_score.  
  
  
**Public and Private Split**  
  
Test data is further divided into Public (40%) and Private (60%) data.   
  
Your initial responses will be checked and scored on the Public data. The final rankings would be based on your private score which will be published once the competition is over.