**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. In order to do so, they would like to segment the customers into different tiers based on their customer lifetime value (CLTV). In order 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.

The dataset is divided into two files train and test:

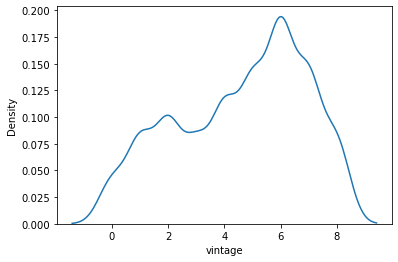
Independent Variables:

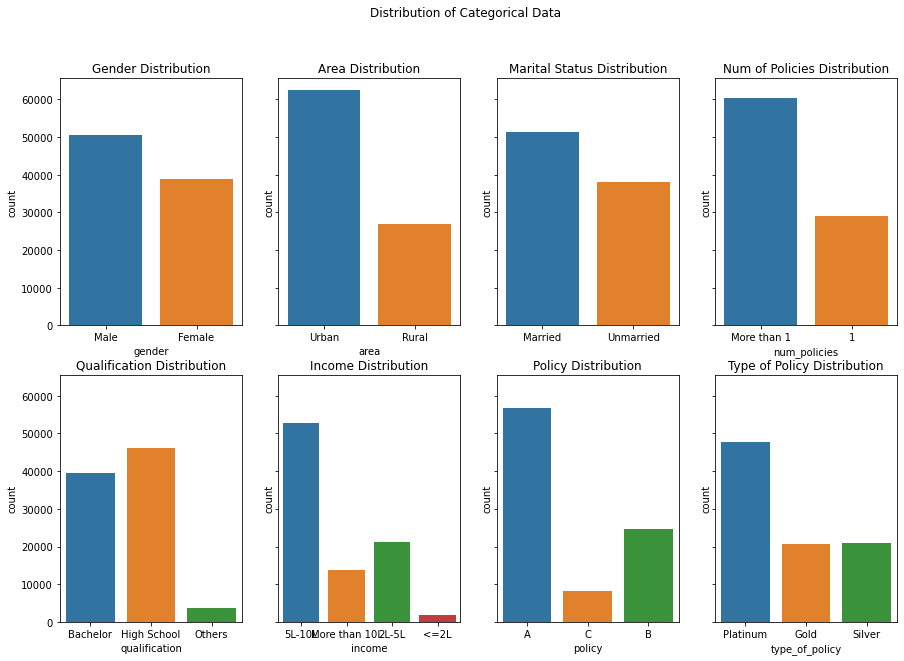
'gender', 'area', 'qualification', 'income', 'marital\_status', 'vintage','claim\_amount','num\_policies', 'policy', 'type\_of\_policy

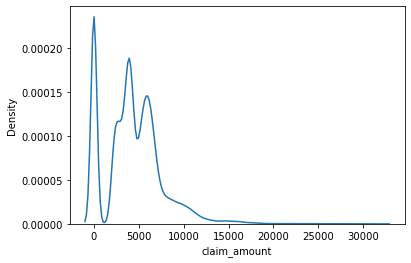
Dependent Variable: ‘cltv’

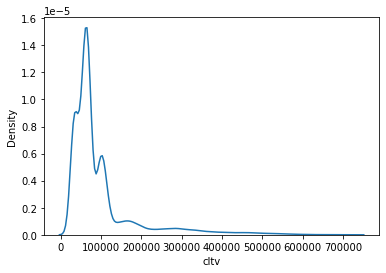
Approach:

1. Imported required modules and read the dataset into a pandas dataframe
2. Checked data for consistency, variable datatypes, central tendency, spread, density and unique values
3. Performed EDA on these variables to have a better understanding of the dataset

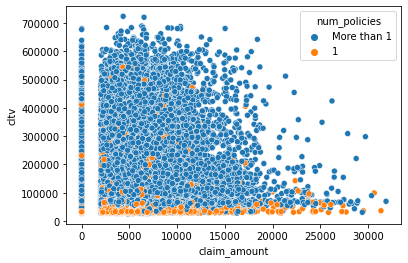








1. Found out some interesting factors for e,g(num of policy held is very strongly correlated to the cltv[target variable])



1. Manually Mapped **Encoded ordinal column** values with our desired values.
2. Encoded the rest of the non numerical nominal columns using pandas get\_dummy
3. Trained the dataset on various simple regression models like linear regression, KNN, etc and recorded their results.
4. Trained the data again on Ensemble models like Random Forest, ADA Boost and XG Boost and found out XG Boost gave the best result among all the alternatives.
5. Used Grid Search CV to find out the best parameters and Finalized the model
6. Trained the model with the best parameters on the complete training dataset and predicted ‘cltv’ for the test dataset with the trained model.

Steps tried but rejected because they were adversely affecting performance of the model:

1. Transforming CLTV to logarithmic value while training and then doing exponential on the result to bring it back to the same level.
2. Using Neural networks