Problem Solving Approach for VahanBima

Harisri M Thulasi

January 22, 2023

Abstract

This document gives a detailed description of the approach towards solving the VahanBima problem. VahanBima is an insurance company. The customer lifetime value (cltv) for a customer is to be predicted using based on the activity and interaction of a customer with the help of machine learning.

1 Given data

The data given by the company includes a train set and a test set of data. The sample dataset of the company holding the information of customers and policy consists of the following values:

- Unique identifier of a customer
- Gender of the customer
- Area of the customer
- Highest Qualification of the customer
- Income earned in a year (in rupees)
- Marital Status of the customer {0:Single, 1: Married}
- Number of years since the first policy date
- Total Amount Claimed by the customer (in rupees)
- Total number of policies issued by the customer
- Active policy of the customer
- Type of active policy
- Customer life time value

Customer life time value is the target variable which is to be predicted for the values corresponding to the test set of data.

1.1 Procedure

The machine learning code is written in Jupyter notebook (python 3) and the file can be found with this document. The detailed approach for attaining the solution is as follows:

- The train data set was imported. Check for any null value in the dataset is done as part of data pre-processing. As no null value was present, data cleaning is not required.
- Explanatory analysis is carried out in the given dataset to find out the importance of different independent variables. The number of high-school graduates more compared to Bachelors and there are very few customers with qualification as 'Others'. The number of male customers is high than the number of female customers and more number customers are married. Customers are mainly from urban areas compared to rural areas. More proportion of customers have taken more than one policies. Policy-A is preferred by many customers and policy-C is least preferred. Mostly, customers take platinum policy and the number of customers having gold and silver policy are almost same. Most customers had their first policy since six years.
- There are 89,392 values in the train dataset. Mean of the cltv values is 97,952.828978 with an minimum value of 24,828 and maximum value of 7,24,068. The standard deviation of the cltv data given is 90,613.814793.
- Non parametric tests are carried out since the dependent variable is not normally distributed
- Categorical data are separated and Kruskal test and Manwhiteneyy test are carried out on the variables and p-values were found out and found to be desirable.
- Simple regression was carried out using OLS model. The condition number, $3.83e+05$, is large. This indicates that there are strong multicollinearity or other numerical problems. Hence it is necessary to check for the assumptions while doing a linear regression.
- Auto correlation was tested using Durbin-Watson test; Normality of Residuals was tested using Jarque–Bera test; Linearity of residuals was tested using Rainbow test; Homoscedasticity was tested using Goldfeld test; and Multi collinearity was tested from the VIF values.
- The basic linear regression model had low R^2 score. Hence, to improve the model, Random Forest Regressor was used.
- Once the test data split from the training dataset is found to be having an improved R^2 score, the regression fitting was done for the test data and the necessary cltv values are found out corresponding to customer id in the test data set. The values are then saved to *Harisri_Full_VahanBima.csv* file.

2 Summary

The final result obtained from test dataset can be found with this document.