Project Report

Data Understanding:

In the given dataset there are total 14 feature columns and 50882 observations are given the training dataset. Most of them are don't have any missing values. And also most features are categorical in nature.

There are presence of missing values in following columns:

- Health Indicator – 39191
- Holding Policy Duration – 30631
- Holding policy type - 30631

Datatypes of the following columns need are corrected:

- Region Code
- Recommended Policy Cat
- Holding Policy Type

All the observation are unique in the dataset. Looks like data is **imbalanced**. Proportion of customers who show interest in recommended policy are **12209** out of **50882**. So while building a model this situation needs to be considered.

Missing value Imputation:

As I mentioned above 3 columns whose having missing value presents out of it two of them which are Holding Policy Duration and Holding Policy Type are closely related accordingly means both columns have missing value present for same observation so that might be because customer could be new one so don't have any policy duration and any holding policy. So I impute those columns by zero and -1 value respectively.

For Health Indicator column no analogy with other columns are found and cannot be dropped so introduce new category type as 'unknown'.

Model Building Approach:

After completing data cleaning part on the given dataset. And then encoding of the categorical feature using **label encoder for high cardinal feature** column and mapping function for binary feature columns.

I tried to build a baseline model without any feature engineering. Since it is classification problem having most of the features are of categorical and Catboost can easily handle the categorical columns. But before that splited the train data into X and y and further into train and validation dataset to experiment on. After appropriately giving the categorical columns to the model along with numerical one I get roc auc score 0.7262430500486261/0.5191962748088872 on train and validation data. As you can clearly see the model is overfited. But still on training side score is better without any feature engineering. Then feature importance of a

feature is plotted using in built function to see which feature columns are considered by a model as important. And get

- Reco_Policy_Cat
- City_Code
- Region_Code
- Holding_Policy_Type
- Reco_Policy_Premium

are top 5 are the most important ones.

Then accordingly new feature is engineered using some mathematical functions and using **featuretools;** new but blind feature engineering is done.

As feature engineering is done then again Catboost is used to build a model with a default parameter. I get a roc-auc score 0.7151261400162812 and 0.6705359760632642 for train/val. As you can see the model became more generalizable. But to improve on the score on both side hyper parameter optimization needs to be done so, I used skopt's **BayesianSearchCv** algorithm to get optimum values for the hyper-parameter. Though using the optimization algorithm model score is improved slightly using optimized parameters given by the algorithm.

But I found that for Catboost there is some tweaking of the default parameter with blind feature engineering improved score better and significantly.