JOB-A-THON JAN 2023 Approach

Problem: Predict CLTV of a customer given data about customer and policy.

Approach:

- **Reading Dataset:** first Importing train dataset into pandas data frame

- **Ask Basic Questions:** asking some basic question with that dataset like how big is the data ? how does the data look like ? what are the data types of columns ? are there any missing values ? are there duplicates values in dataset ? how does data look like mathematically ? how is the correlation between columns ? etc.

- **Exploratory Data Analysis:** getting some answer of questions then apply EDA on columns in that some columns are imbalanced data one of categories value high and some is very low on categorical columns and numerical column data is right skewed of claim amount column also there lots of outlier

- **Extract Columns:** extract independent columns and dependent column from train dataset and split training and testing dataset to check model performance

- **Pre-processing:** apply some preprocessing techniques on training dataset like ordinal encoding on ranked columns and one hot encoding on not ranked columns using column transformer

- **Model Selection:** apply various algorithms on training dataset like linear regression, random forest regressor, gradient booting regressor, kneighbors regressor, svm, deep learning model all of this algorithm best perform on training and testing dataset is gradient booting algorithm so select that algorithm

- **Hyperparameter Tuning:** selecting gradient boosting regressor then turn the hyperparameter like max_depth, max_features, leaning rate, n_estimators etc. I get best result of training dataset of from the algorithm