JOB-A-THON JAN 2023 Approach

Name: Mahesh Chandra Duddu
Email: duddumaheshchandra@gmail.com
Phno: 9440642368

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

Approach:

- Duplicates: Training Data has duplicated rows with different values only at dependent feature and id. Same rows can be seen on test data, surprisingly. Tried handling them reducing duplicates using median/mean of duplicated data and imputing on them test data, followed by training and predicting on remaining data. When checked with scores on leaderboard, leaving the duplicates worked well relatively. So, left them untouched.

- Outliers: “claim_amount”, “cltv” features are highly skewed. Flooring/Capping improved the score locally but not on leaderboard. Log transformation or Square root transformation did make the predictions worse both locally and on leaderboard. So, Left them untouched.

- Feature Engineering: Using categorical features, new dense features are created using both One-Hot Encoding and LabelEncoder. Also, features are handled using CatBoost categorical features argument.

- Model Building: CatBoost along with 10 Fold Cross validation Strategy is used, by specifying cat_features. Mean of the predictions across folds were taken to be final predictions.

- Feature Selection: Features are selected based on average of feature scores using catboost feature importance from all the folds.

- HyperParameter Tuning: Used Optuna and Automl both gave similar score.