Phase 1: Preprocessing

The preprocessing done for this challenge was minimal given that the datasets doesn't have any missing values. In this sense, I just extend it with interaction features of secon order.

Given that I'm trying two types of machine learning models: Linear based (Logistic Regression and Support Vector Machine) and Tree based (Random Forest, XGBoost, CatBoost and GradientBoosting), I performed two forms of preprocessing, besides adding the second order interaction terms.

For Linear Based:

After the dataset expantion: Continuous Feature Normalization throught Standard Scaller following PCA analysis, also just applied to continuous features.

For the Tree Based classifiers, just the dataset expantion was done due to Tree Based Models do not suffer from multicollinearity and are invariant to linear and non-linear transformations.

Phase 2: Training and Hyperparameter tunning

I experiment training and hyperparamenter tunning with and without interaction features, and concluded that the expanded dataset lead to more accurate results.

In a second stage of training and hyperparameter tunning, done after the addition of the data augmention step to the preprocessed pipeline, the best performing classifiers in the validation set were the CatBoost and RandomForest, with respectively, 0.846 and 0.821 of f1-score metric.

Phase 3: Output and Error Analysis

After model training and hyperparameter optimization, output analysis was carried out.

The significant observation made was that in the validation set, 25% of instances belonging to the targets' less observed class were misclassified while only 2.5% os the instances belonging to the targets' most observed class were misclassified.

Due to this observation, I later add to the preprocessing pipeline a data augmention proccedure throught a variant of SMOTE.

To end this project report, the selected model (CatBoost) got a f1_score on the test set of 0.846