Analytics Vidya Hackathon

Objective: Predict if Happy Bank existing customer could be a potential lead for credit card sell.

Training data: 245725 clients with 9 independent features and 1 binary dependent feature excluding ID

Test data: 105312 clients

Approach:

EDA:

1. Import packages, Check datatype of each column, and separate ID and Is_Lead column from training data.
2. Check for imbalanced dataset, it showed 25% with 1 and 75% with 0 in dependable variable.
3. Check for missing values in each column, which showed Credit_product column has null values.
4. Check for correlation between columns except Credit_product.
5. Convert Credit_product object values which are binary to numerical values using map function.
6. Convert other object columns to numerical columns using one-hot encoding.
7. Using fit-transform of IterativeImputer from sklearn.impute, fill the missing values in Credit_product column.
8. Tried to balance the imbalanced dataset using SMOTE and ENN sampling techniques, however, it didn't perform better than test of without sampling.

Spot-checking of algorithms

1. Creating pipeline to spot check few top performing Bagging and Boosting classifiers like LGBM, XGBoost, GradientBoosting, and Random Forest along with along with Support Vector classifiers, KNN, and Logistic Regression. I also designed a deep neural architecture for classification for checking which gave highest accuracy on test data of 0.8637 and 0.8745 on training data.
2. LGBM performed best with mean ROC 0.8737 on training data during spot check, so LGBM is selected.

GridSearch with 4-fold cross-validation for Optimum parameters of LGBM model.

Train the model with best Gridsearch parameters obtained and predict for test dataset.

Upload on Analytics Vidya for test score and ranking. Best score on test data is 0.869976