Happy Bank - Credit Card Lead

30.05.2021

Akkash K N R
Business Analyst, Landmark Groups
aka7h.sathya@hotmail.com
+91-8056111831

Overview

Happy Customer Bank wants to cross sell its credit cards to its existing customers. The bank has identified a set of customers that are eligible for taking these credit cards.

Goals

1. Identity Customers who would show higher intent towards a recommended credit card

A Brief Approach

Data processing One hot Encoding, Freq. Encoding, Binning and Log transformation for Categorical and Numerical variables.

SMOTE, ROSE, DOWN and UP Sampling to fix Imbalance class.

Bagged Imputation of missing values.

The above mentioned are used in both Catboost and H2O AutoML to evaluate the model perforation on a validation set.

Ensemble of Multiple variations of Catboost and H2O AutoML is used

Catboost model with manual Hyperparameter tuning and log transformation on Avg_Account_Balance and filling -999 on missing values is used as the final model with a score of 0.87301 in the public leaderboard.

Feature Engineering

- Log Transformation on Avg_Account_Balance was used across all models though it showed no variations in the model.
- Bagged Imputation / missForest is used for missing variables and showed no improvement when used with H2O and Catboost ,
- One Hot Encoding of Categorical variable, Near Zero variance feature removal and SMOTE, ROSE, Down and Up sampling for imbalance class handling.
- Variations of these data processing gave a very low AUC score less than 0.7961* in H2O AutoML and 0.800* for Catboost (Max AUC across all variations used in both models).
- Frequency Encoding on Region and Binning of Age and Vintage into 4 groups ran with -999 imputation of Missing values on Catboost improved the score to 0.868
- Credit_Product, Occupation, Age and Vintage are considered the most important features from the Catboost model mentioned above.
- Log Transformation, Frequency Encoding, SMOTE/ROSE, One Hot encoding showed low importance and ROC Score in models but also showed overfit between train/valid vs public leaderboard.
- Splitting Train data to train and validation set to find best iterations which is to be used to evaluate the model and run on full dataset.
- Ensemble of top 3 catboost model was used which gave an AUC of 0.872699

Final Approach

- Since multiple fancy feature engineering failed to show improvement in AUC both in validation set and public leaderboard.
- Implemented the simplest feature engineering approach. Convert Character to Factor, missing impute -999. This gave a 0.8721 AUC in public leaderboard
- Catboost was used as a final model with 2957 iterations, with tree depth between 5 to 8, rsm between 0.5 to 0.8, l2_leaf_reg = 8 and auto_class_weight as “balanced”
- The above model gave 0.873 AUC for the validation set. The model gave a best score of 0.87301