Credit Card Lead Prediction Problem

Presented by: Ashutosh Kumar

Exploratory Data Analysis

1. Shape and data types in train and test datasets
2. Checked given problem is balanced or imbalanced problem
3. Treated it as balanced problem given enough responders
4. Corelation and summary statistics for numeric features
5. Distinct categories across train and test datasets for object features
6. Found missing values in Credit_Product feature
7. Did 80:20 startify split of train dataset for feature engineering

Feature Enginnering

1. Checked distribution of numerical features
2. Avg_Account_Balance was highly rightly (+ve) skewed so took the log transformation
3. Region_code coulmn had multiple categories so replaced the value using frequency encoding
4. Scaled all the numerical features using Standard Scaler (Also tried out Min max scaler but it didn’t add any incremental value so deleted this step)
5. Missing values in Credit_Product feature were treated as a separate category for dummy encoding (Also tried out mode imputation which was deleted in later step)
6. Label encoding was performed on Gender, Is_Active and, Credit_Product features
7. Duumy encoding was performed on Occupation, Channel_Code, Credit_Product and Credit_Product_mode_impute features
8. Co-relation across all the columns were checked, if any two columns had +/- 0.7 corelations coefficients then one of those features were deleted after checking the corelation with target variables
9. All the preprocessing steps were performed X_test and test dataframes

Feature Selection

1. Logistic model was built using all the features
2. On the basis of p values, features were selected

Model Building

Used 2 methods

1. Logistic Regression
   This was more of baseline model which gave 0.85504 score
2. XGBoost Model
   This was champion model which was trained using same set of features but with hyper-parameter tuning and 3 fold cross validation technique which gave much better score (0.87096)

THANKS