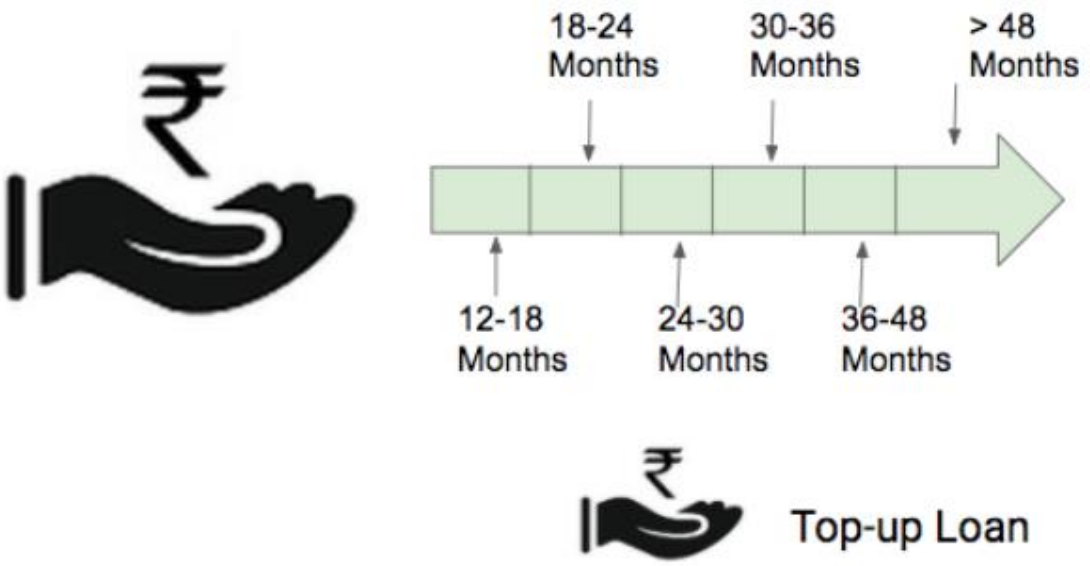
LTFS Top-up loan Up-sell prediction

LTFS provides it's loan services to its customers and is interested in selling more of its Top-up loan services to its existing customers so they have decided to identify when to pitch a Top-up during the original loan tenure. If they correctly identify the most suitable time to offer a top-up, this will ultimately lead to more disbursals and can also help them beat competing offerings from other institutions.



APPROACH

1. Demographic Dataset and Model Trainnig for V1 Model
   - **Handling Missing Values** in Demographics dataset of Train and Test (Replaced most of the features with Median values) , Replace nan values of Zip code with Mode Value based on State.
   - **Categorical Encoding**
     Before categorical encoding , Missing values in Test dataset was also handled and then combined with the Train dataset to do final Categorical Encoding .
     Categorical Encoding Techniques : One Hot Encoding,Mean Categorical Encoding
   - **Model Training V1**
     Model training was done on above dataset and got a score of **0.148** on Leaderboard
   - **Feature Selection on above Dataset**
     - Multivariate correlation was done to find out relationship between Independent and Dependent Feature
     - Correlation between Independent feature was done to identify Highly correlated features in Independent Features
     - Extra Trees Classifier Technique was used to identify Feature Importance
   - **Model Training V2**
     Model training was done on above dataset and got score of **0.152** on Leaderboard

2. Preprocessing of Bureau Dataset
   - **Handling Missing Values** : Missing values of features were first replace by 0
   - **Extraction of Numbers** was done on Installment_Amt Feature
   - **String Manipulation** : Removed commas from No's in AMT_Features
   - **Group By 'ID'** was done with Demographic and Bureau Dataset
   - **Median values** are extracted from below Bureau dataset features
     'DISBURSED-AMT/HIGH CREDIT','INSTALLMENT-AMT','CURRENT-BAL','OVERDUE-AMT','WRITE-OFF-AMT','TENURE'

- **Model Training V3** of Demographic data and Bureau Dataset
  We did all above steps on Test Dataset of Bureau , followed by GroupBy ID with Demographic Dataset and extracted Median Values . Model training was done on above group by 'ID' dataset with median values and got score of 0.157 on Leaderboard
- **Features Creation from Categorical Features of Bureau Dataset**
  Groupby 'ID' was done to extract new features from Categorical Features such as
  Top Value , Top Value Frequency , Unique values , Unique Count from each Categorical Feature
- **Features Creation from Bureau Dataset**
  Preprocessing of DPD_Hist,Cur_Bal_Hist,Amt_Overdue_Hist,Amt_Paid_Hist
  DPD_Hist
  Handled Missing values and replaced string containing 'E' to 0
  Splitting of String at 3rd Digit was Done
  Replace 'DDD' with '000' , 'XXX' with '000' --
  Computed Median value of Each Row of DPD_Hist
  Cur_Bal_Hist,Amt_Overdue_Hist,Amt_Paid_Hist
  Handled Missing Values
  String Manipulation : Splitting at comma
  Compute Median values of Each row of all above Features

  Once Medain Values of each row was computed , I did the GoupBy 'ID' on Bureau dataset to compute median value for Each 'ID' of Demographic Dataset.
  Then I merged the computed Medain values of each id of Bureau Dataset on Demographic Dataset
- **Reducing Cardinality of Categorical Features**
  String Manipulation for Categorical features created above was done to reduce the Cardinality
  Reordering of string was done because after applying GroupBy, unique values of Each categorical feature was not in order and cardinality got increased significantly

  I also calculated MEAN values instead of MEDAIN for all above features but that did not me good results.

3. **Categorical Encoding and Final Model Training**
   Above Preprocessing steps were also carried out on Test Dataset of Bureau to create new features which were then combined with Demographic Dataset

   3 Categorical Encoding techniques were used and 3 Dataset for final Model Training were created
   - Categorical Encoding based on RANK
   - Categorical Encoding based on RANK , One Hot Encoding for TOP 10 categories of Acct_type_unique_values
   - Categorical Encoding based on Value Count

   Model Training was done on Above dataset and I found that Categorical Encoding based on Rank gave me higher score : **0.20** on Leaderboard

While Model Training on above steps I used following algorithms – (Logistic Regression,KNN,Decision Tree,Random Forest,Gradient Boosting,XG Boost,LGBM,Catboost. I also did Hyper parameter tunning in all above algorithms with Random Search,Grid Search,Bayesian Optimization.

At last I also Handle Imbalanced Dataset by using SMOTE and then did model training to improve my score on Leaderboard . But this leads to overfitting on test dataset . So ididnt use this for submission.

Then I avoided use of any information which is available in bureau dataset of training , but not available in test to avoid overfitting.Removed some of the features and trained model on LGBM to get better score.

Final Submission Code is also available on Github : View Code

THANK YOU LTFS and ANALYTICS VIDHYA . I Learned many things in this Competition.