Vehicle Loan Default Prediction

Objective:

To build a model that is capable of predicting whether a person will default on his/her first EMI payment of vehicle loan.

Client/End User:

The end user Larsen & Toubro Financial Services (LTFS) will have a deeper insight on decision making on sanctioning loan for a particular applicant. Such as the result given by the model will give more confidence when deciding.

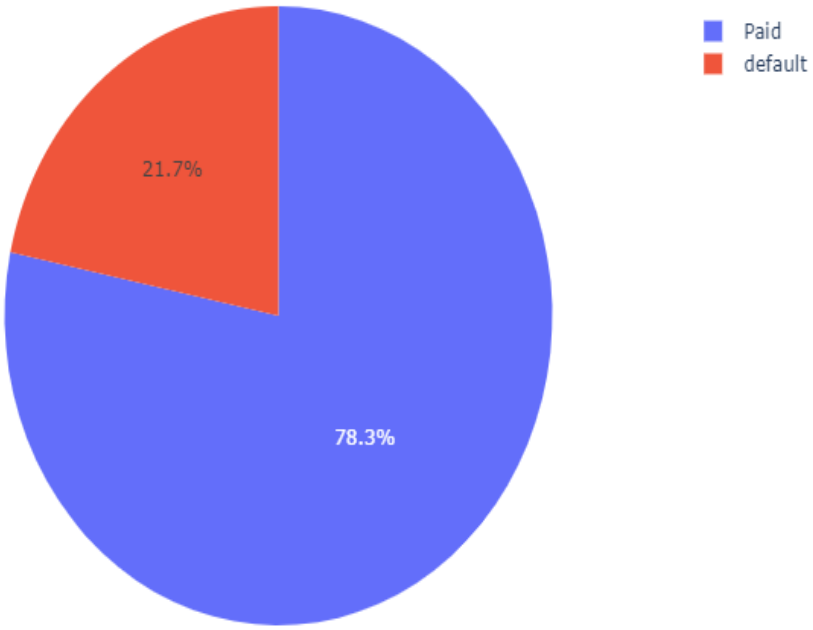
Data:

We use LTFS's own data which they used to conduct a competition in 2019.The dataset contains about 53000 entries with features such as loan amount, asset cost, customer employment status etc. The dataset contains 40 independent columns with 1 dependent column which 0 or 1 as value. (0-Paid, 1-Default).

Workflow outline:

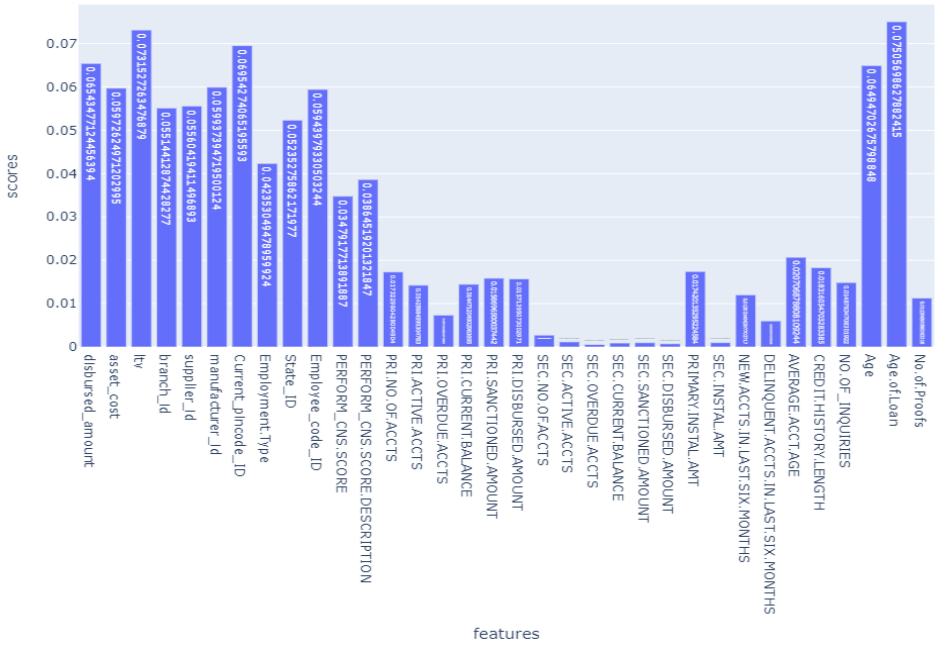
- **Data Wrangling**:
    - Transformed data type for columns like DOB, DisbursalDate into proper date strings.
    - Handled missing values in Employment Type column. NaN were replaced with 'unknown' string.
    - Employment type, PERFORM_CNS.SCORE.DESCRIPTION columns had string values which were mapped to integers.
    - AVERAGE.ACCT.AGE,CREDIT.HISTORY.LENGTH columns had values like '2 yrs 4months',replaced them with float values like 2.4 with a custom function yrscalc.
    - lables column was created out of loan_default column for easy human interpretation.

- **Feature Engineering/Selection:**
    - Proofs columns were combined (summed up) to create a new feature called No.of.Proofs.
    - The dataset faced imbalance class issue with about 78%- Paid and 22% Default class. To handle it SMOTE from imblearn was used to overcome it.

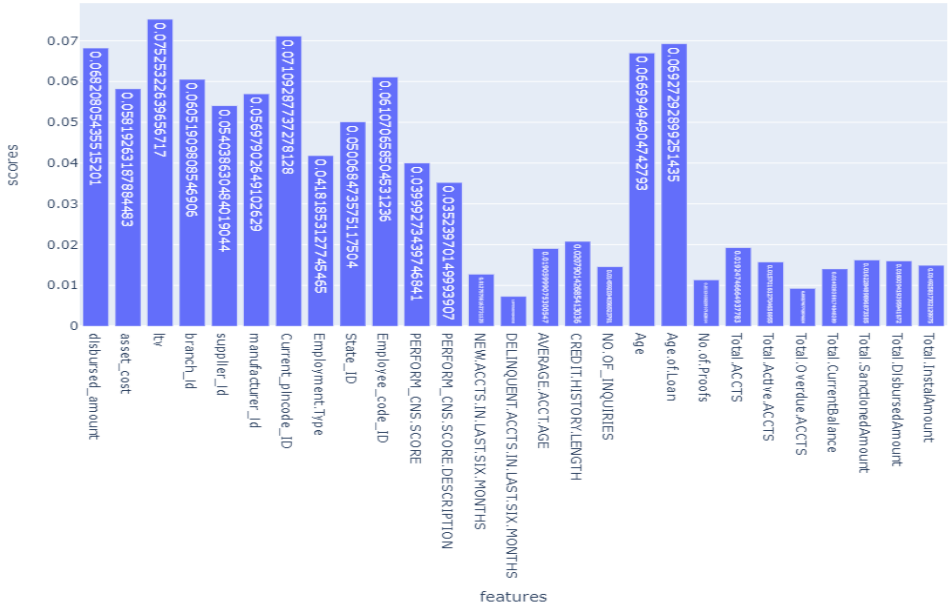


- Using ExtraTressClassifer to obtain feature importance scores it was found that secondary account related features did not have much significance and hence primary and secondary account features were clubbed.

Before clubbing



After clubbing



- **Model building(work in progress):**
  - Explore possible model such as Random forest, Gradient boosting etc and choose a model that performs well or a combination of models.
  - Precession and Recall metric to be used to evaluate model performance.
  - Publish reports.