Domain Oriented Case Study
Option III – BFSI Case Study

S K Sahu

Problem Statement

- Most of the BFSI organizations find it hard to give loans to the people due to their insufficient or non-existent credit history.
- The primary objective of the case study is to assist Home Credit in deciding which loan applications should be disbursed, and which should be rejected, based on the applicant’s past behaviour and application information.
- As a business analyst for Home Credit, we first need to gather the information and clean it to make it usable and apply ‘Feature Engineering’ techniques to roll up the information at applicant level, and thereby create manual features for model building.
- Build a classification model to differentiate applicants between approves and rejects.

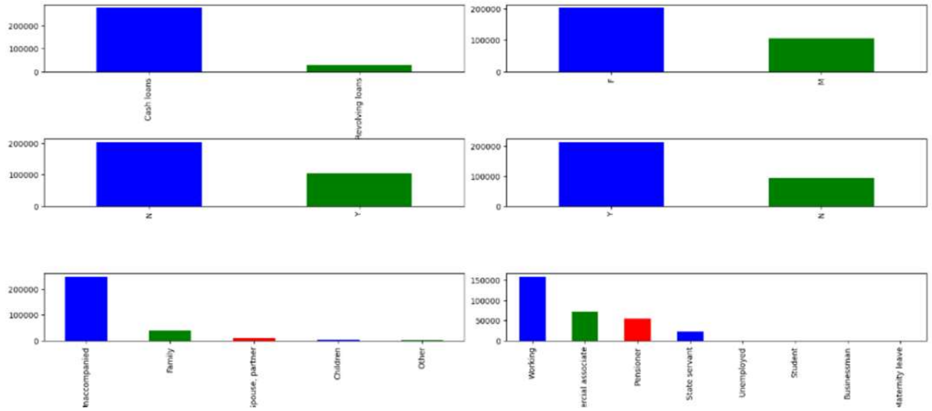
Approach

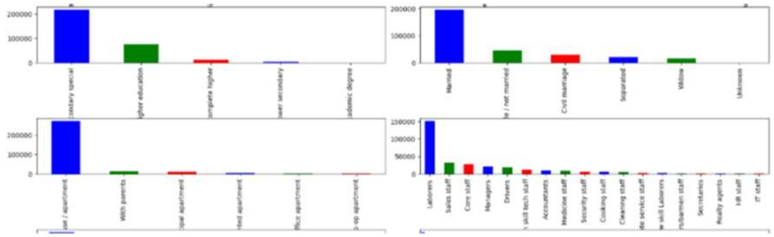
- Data Exploration
- Data Cleaning
  - Missing values treatment
- Exploratory Data Analysis
  - Checking of Outliers
  - Checking of unwanted variables
  - Identifying Variable correlation
- Data Preparation
  - Standardization of data
  - Creating Dummy variables
  - Removal of repeated variables
  - Classification of train and test data
- Feature Engineering
  - Feature Scaling
  - Feature selection using RFE
- Model Building using Logistic Regression
- Model Evaluation using CV

Data Cleaning & Impute missing values

- If the missing values in the column are more than 40%, the column is dropped as the analysis may be not accurate.
- Flagged columns are correlated with the target variables and those columns are dropped as there is no significant correlation. Since, the correlation co-efficient is not in the range of -0.8 to +0.8.
- Columns with description - score from external source doesn’t provide enough information on those columns about the score range and details of the external source. Hence those 2 columns are dropped.
- If the Count of payment column having null value is considered that there is no payment history and those values are replaced with 0.
- Categorical variable missing values are imputed with mode value
- Numerical variable are imputed with median value
- Outliers are treated and capped with range

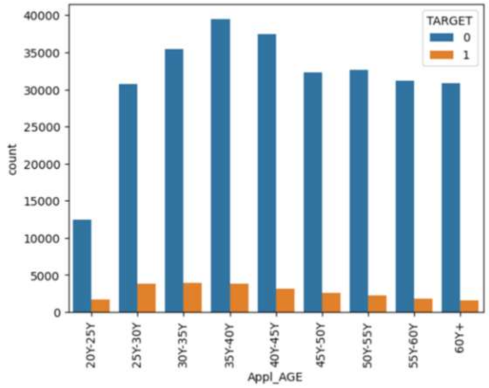
Exploratory Data Analysis - Univariate

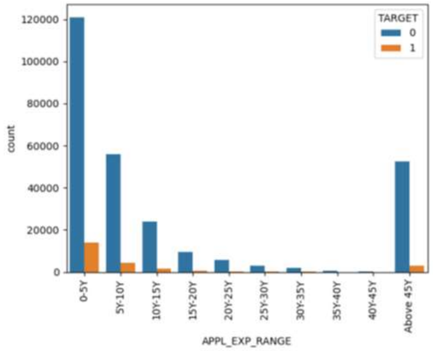


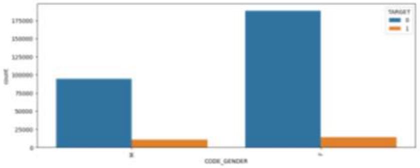


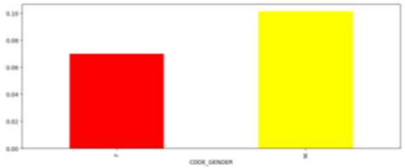
- 90% of applicants applied for cash loans,10% of applicant applied for Revolving loans
- 65% of applicants are Female and working professionals, 35% are male
- 81% are unaccompanied during application of loan
- 52% of applicants belong to working class and 71% of applicants have completed the secondary education.
- 66% of applicants don’t own car, 69% of applicants owns property
- Working professionals, married, laborers, staying in rentals apartments are the most loan applicants.
- Most of the applicants are from Business background

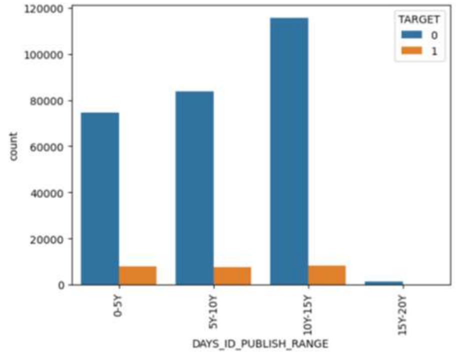
Exploratory Data Analysis - Bivariate

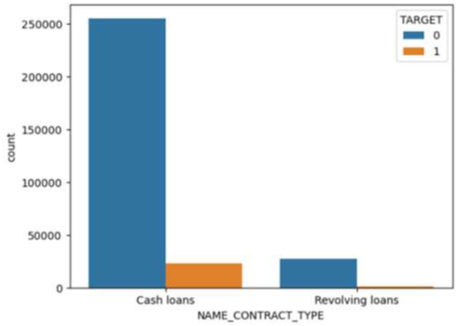






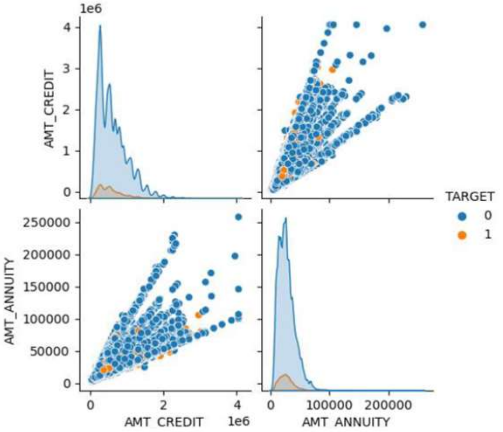


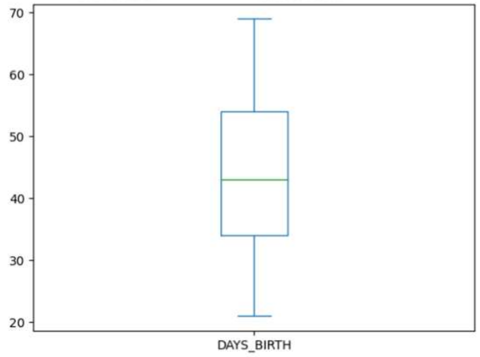


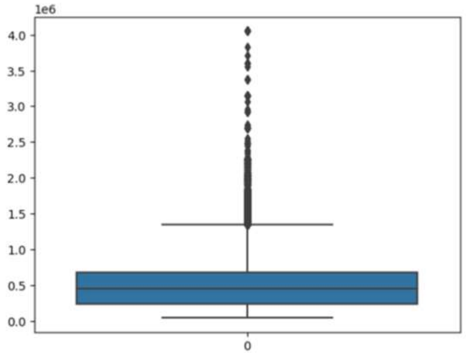


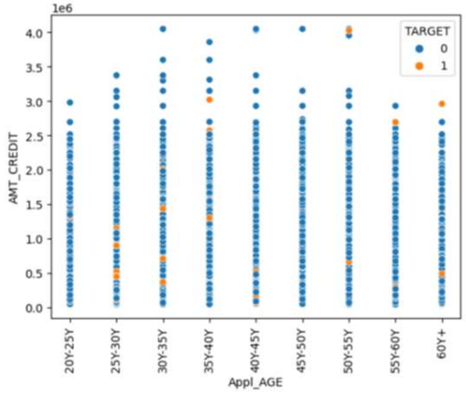
- People of age between 35-40, 40-45 years and 60+ years are less defaulters
- People with 5-10 years of experience are less defaulters and the defaulters are inversely proportionate with the years of experience
- Most of the defaulters are unemployed and people in maternity leave.
- he clients with lower secondary education are with the highest default rate of over 10%.
- Married and widowed clients are the safest one to issue the loan. They are with 6 to 7% default rate
- Most of the clients who apply for the loan are having the income of 0 to 1 million.
- Most of the clients are paying the annuity of 0 to 70k.

Exploratory Data Analysis



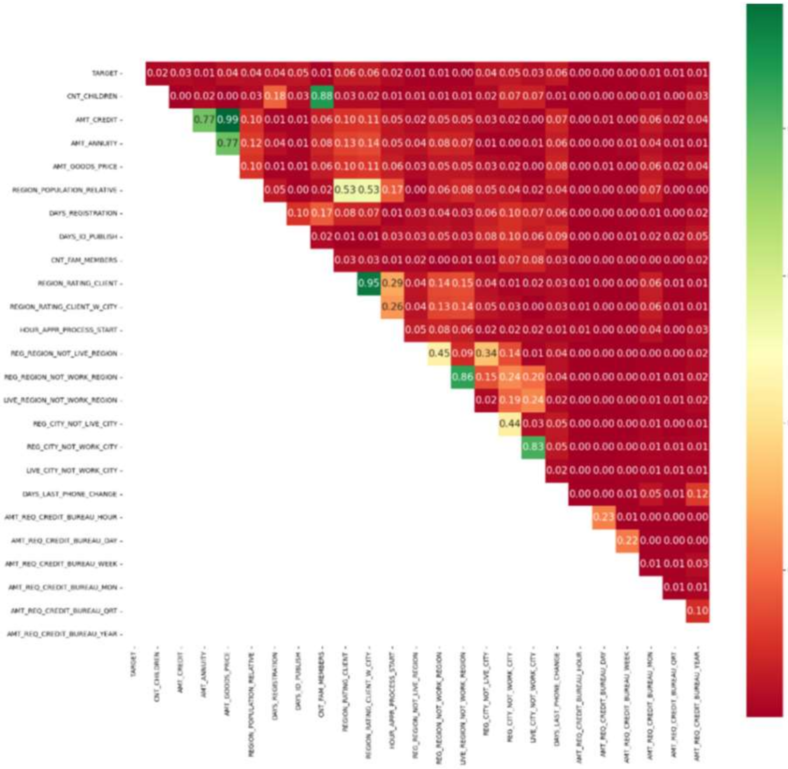


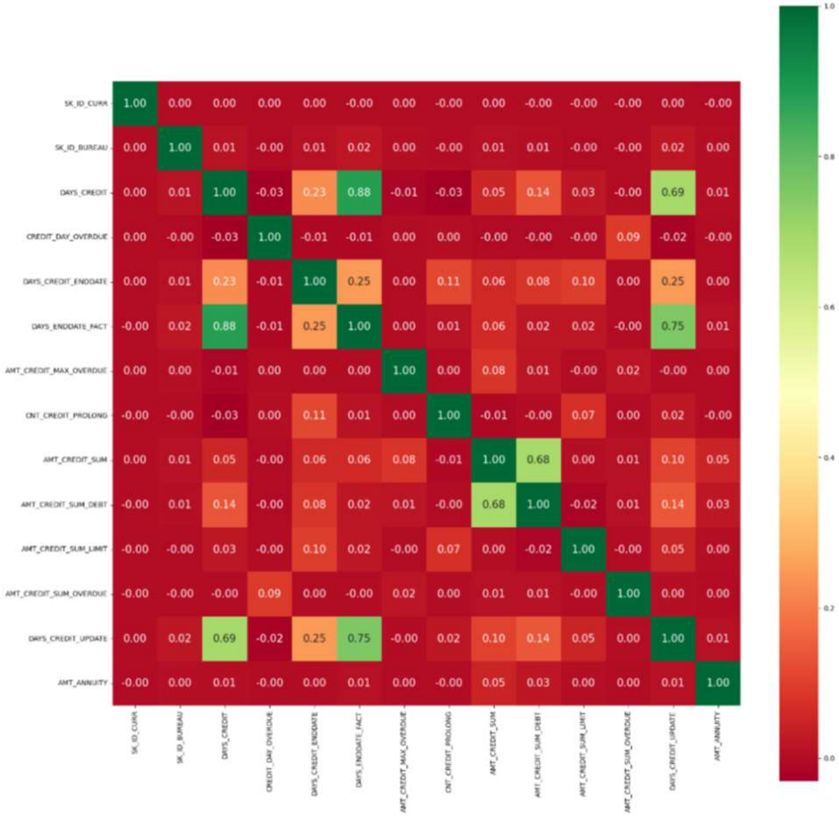




- The clients having income less than 1 million are likely to take loans with goods price between 1 to 3 million and are repayers.
- The amount credit is in linear relationship with goods price. When the goods price increases, the credit amount also increases for the credit amount issued between2 to 3 million and have no defaulters.
- The clients having children count of 1 to 3 are the most safest for giving credit upto 2.5 million, since there is no defaulters in the category.
- 5+ children's count clients are likely more to be defaulters
- When family members count are between 2 and 5 are taking more loans and are less defaulters.

Correlation Matrix





- 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'CNT_FAM_MEMBERS', 'REGION_RATING_CLIENT_W_CITY', 'LIVE_REGION_NOT_WORK_REGION', 'LIVE_CITY_NOT_WORK_CITY‘ are highly correlated with each other.
- Region population relative is inversely correlated with Region Ratings.

Data Preparation

- One-hot encoding for categorical variables
- Aggregating with mean/median of trade level data to applicant level
- Merging the two data sets
- 70% of data are train data set and 30% of data are test data set
- Dropped repeated, unwanted and highly correlated variables

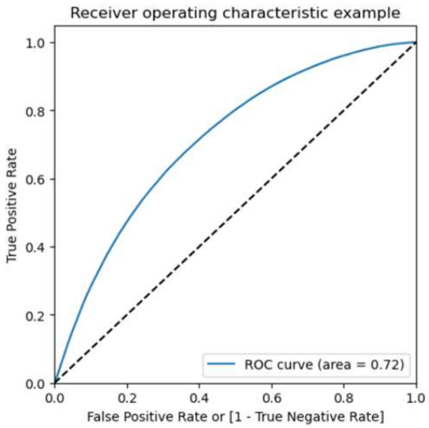
Model Building

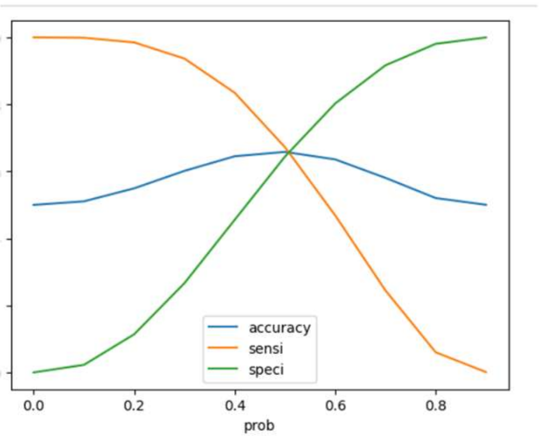
- Split the train and test data using train_test_split library
- Feature scaling using standard scaler
- Simple logistic regression technique gives 90+% accuracy due to data imbalance.
- Handling class imbalance using SMOTE & TOMEK technique
- Feature selection using RFE
- Model building using Logistic Regression
- Model evaluation
- Cross validation

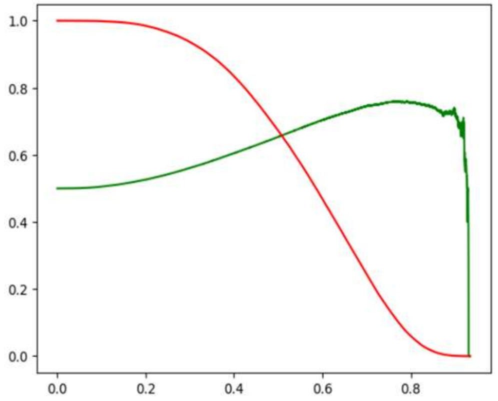
Model Evaluation

Comparing the values obtained for Train & Test data:

- Train Data:
  - Accuracy : 65.81%
  - Sensitivity : 67.22%
  - Specificity : 64.40%
- Test Data:
  - Accuracy : 60.91%
  - Sensitivity : 63.90%
  - Specificity : 60.65%
- Precision: 65.38%
- Recall: 67.22%







Recommendation

The Home Credit should target the clients -

- Clients who are unaccompanied
- Female clients as they are least defaulters
- Pensioner, commercial associate & state servants with 5-6% percent of default
- Clients who have completed higher education
- Married and widowed clients
- Clients staying in own house/apartment, municipal apartments
- Clients who are Accountants, core_staff, Managers, Labours and HR staff
- Clients with 0-5 years of working experience

Recommendation

The following variables are the good predictors for separating good and bad borrowers

- DAYS_REGISTRATION
- DAYS_ID_PUBLISH
- REGION_RATING_CLIENT
- APPL_EXP_RANGE_5Y-10Y
- NAME_INCOME_TYPE_State servant

The following variables are inversely proportional to the defaulters or less exposed defaulters

- APPL_EXP_RANGE_10Y-15Y
- Appl_AGE_50Y-55Y
- Appl_AGE_55Y-60Y
- NAME_CONTRACT_TYPE_Revolving loans
- NAME_EDUCATION_TYPE_Higher Education

Answers

1. How to leverage trade level information for Credit Bureaus by aggregating trade level information to applicant level in order to capture their payment behavior?

Answer:

Using mean to aggregate trade level information to applicant level in order to capture the payment behaviour.

2. Which application or payment behavior factors significantly influence borrower's behavioron any new disbursed loan?

Answer:

Below variables highly influence borrowers behavior

- DAYS_REGISTRATION
- DAYS_ID_PUBLISH
- REGION_RATING_CLIENT
- APPL_EXP_RANGE_5Y-10Y
- NAME_INCOME_TYPE_State servant

3. After identifying these factors, how to leverage them in the form of a model which can be used for decisioning?

Answer:

Using RFE can help to identify the variables. The data should be standardized using standard scaler and target variable should be balanced using SMOTE TOMEK.