HOME CREDIT – BFSI CASE STUDY

ASHISH KUMAR

UPGRAD & IIITB | DATA SCIENCE PROGRAM – DSC68

Introduction – Home Credit BFSI Case Study

Background

All BFSI institutions are faced with a major default problem: not every individual that takes a loan has the willingness, ability and/or integrity to pay it back. Thus, an average of 2-5% default rate is observed across banks for different loan categories like personal loan, education loan, vehicle loan, business loan etc. Given the fact that banks can never get this number to zero, it has to keep it within limits, and rather keep it at the lowest possible levels to be able to retain the money inflow.

In this assignment, you would face real world data of applications and bureau as shared by Home Credit, to practice the end-to-end process of model development in Credit Risk for Banks, Financial Institutions and NBFCs. You would build a bank's internal end-to-end scoring mechanism, based on the application information, clubbed with the raw bureau information.

Objectives

- The primary objective of this 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, you are supposed to first gather the information and clean it to make it usable.
- 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

- **EDA**
  - Data Cleaning
  - Data Analysis
- **Feature Engineering & Model Building**
  - 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

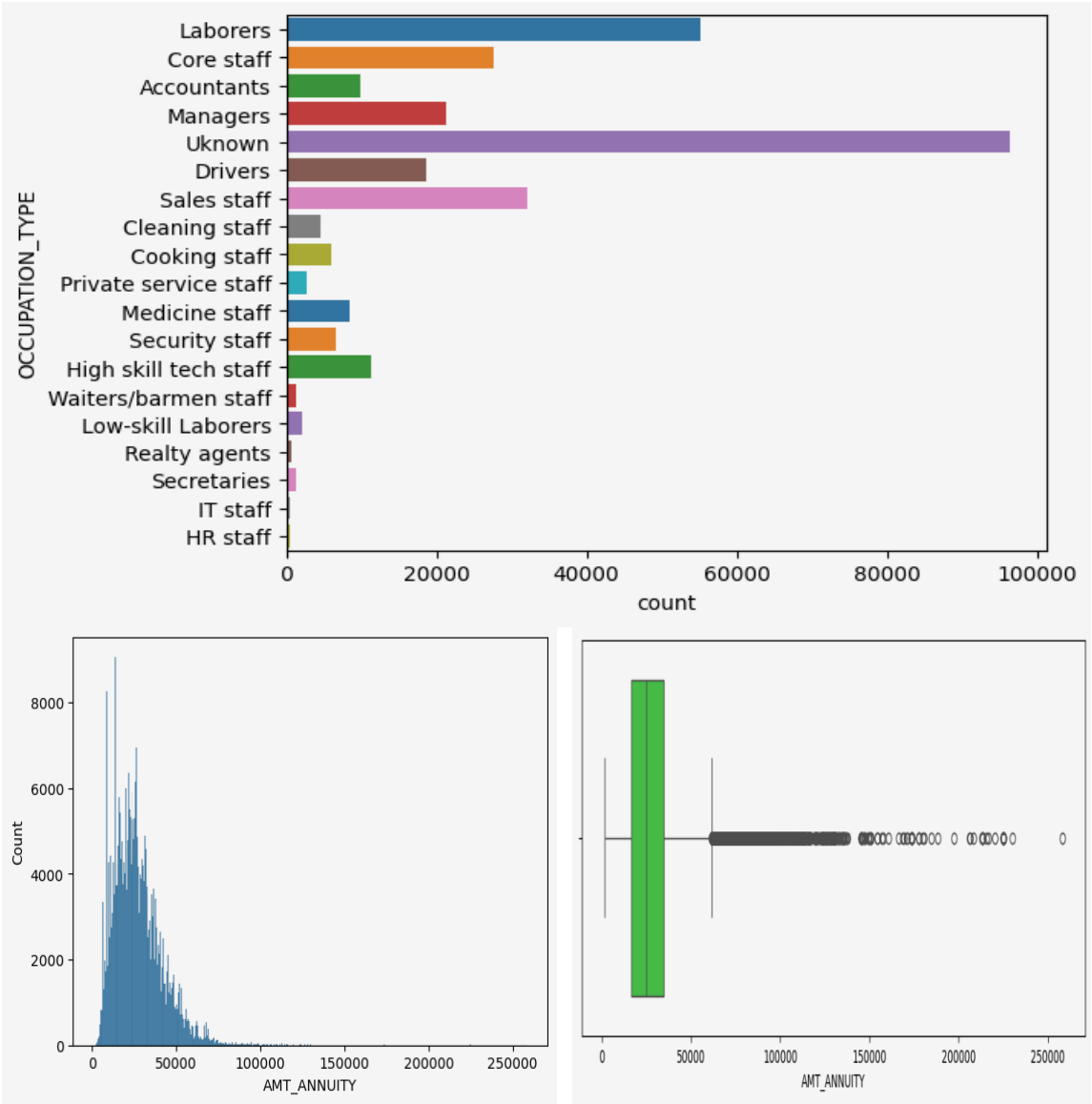
- **Conclusion & Way Forward**
  - Provide useful insights on the basis of analysis and model which can be useful for the Loan business.

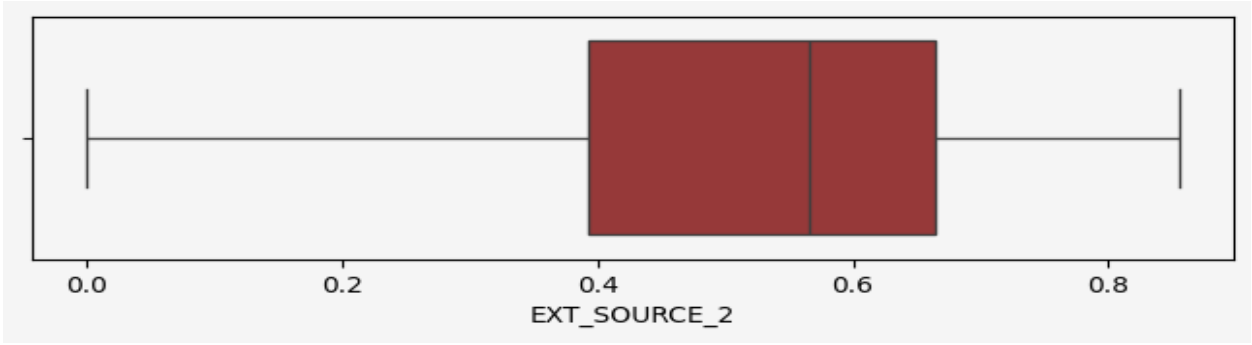
EDA

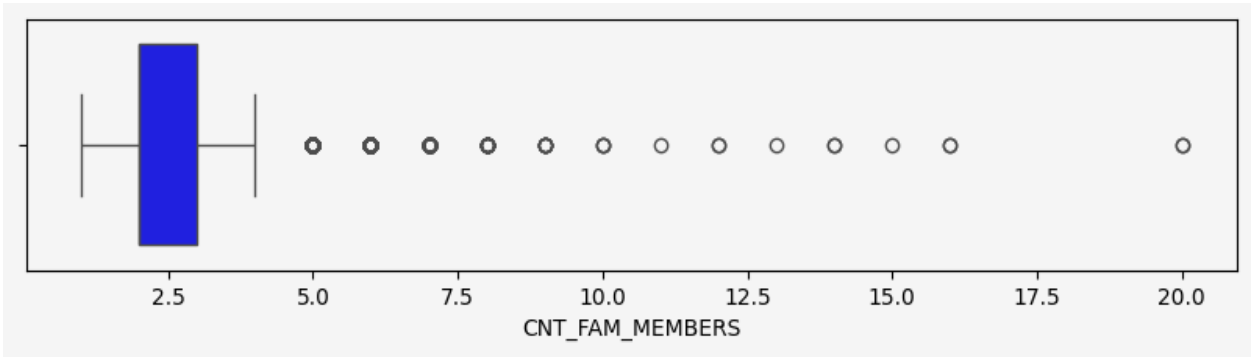
- **Data Cleaning**
  - Drop columns with more than 50% missing values.
  - Identify any other Additional columns which can be dropped as they may not be particularly useful for this analysis.
  - Categorise the columns into 'Numerical' and 'Categorical' bucket on the basis of variable types (integer/float/object)
  - For the remaining data, missing values is imputed using mode for Categorical variables and median/mean for Numerical columns.
- **Data Analysis**
  - Analyse the data by splitting the data into two target categories i.e. Target 1 – Client with payment difficulties and Target 0 – Client without payment difficulties.
  - Perform univariate, segmented univariate, bivariate analysis
  - Find the top correlations.
  - Merge the two datasets and perform further analysis.

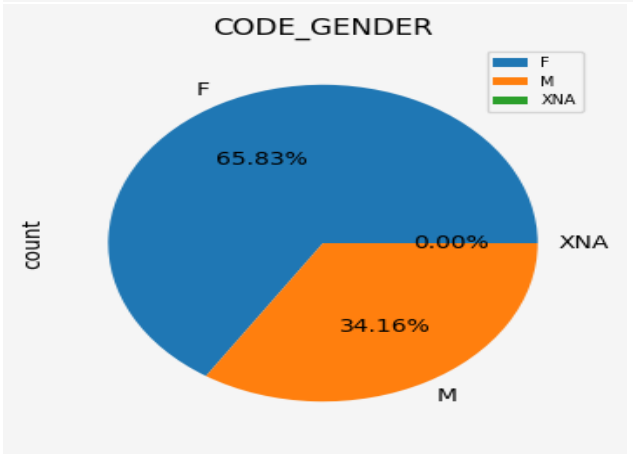
Some of the charts for shown in the following slides for representation. The detailed charts and analysis is present in the python notebook with results.

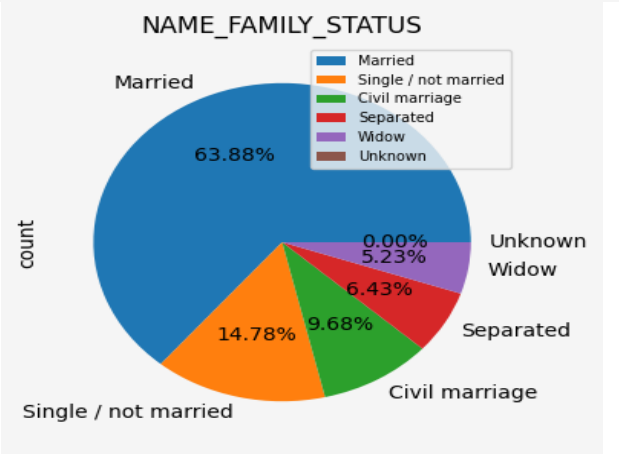
Data Cleaning & Analysis (1/2)



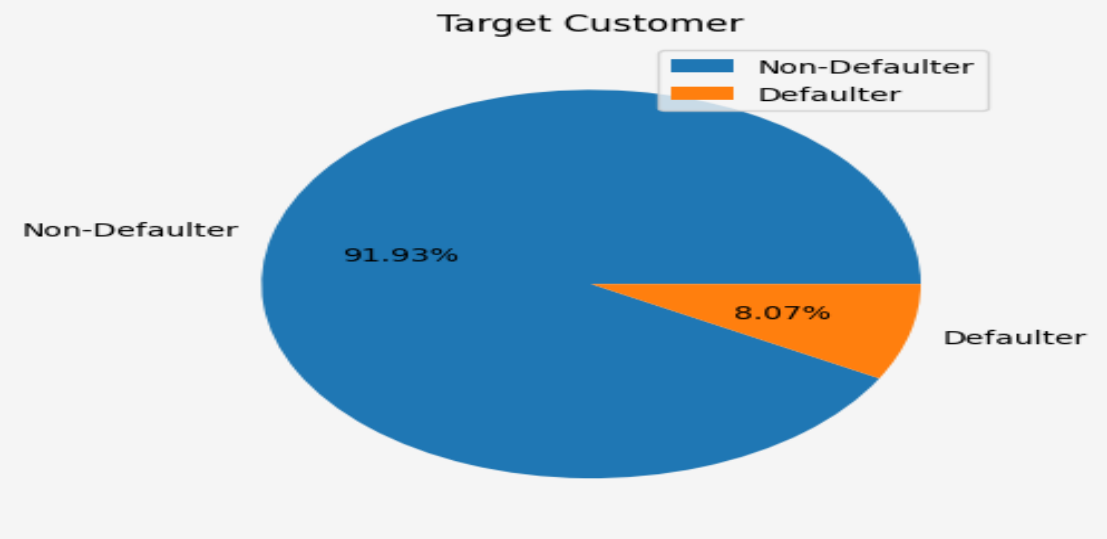


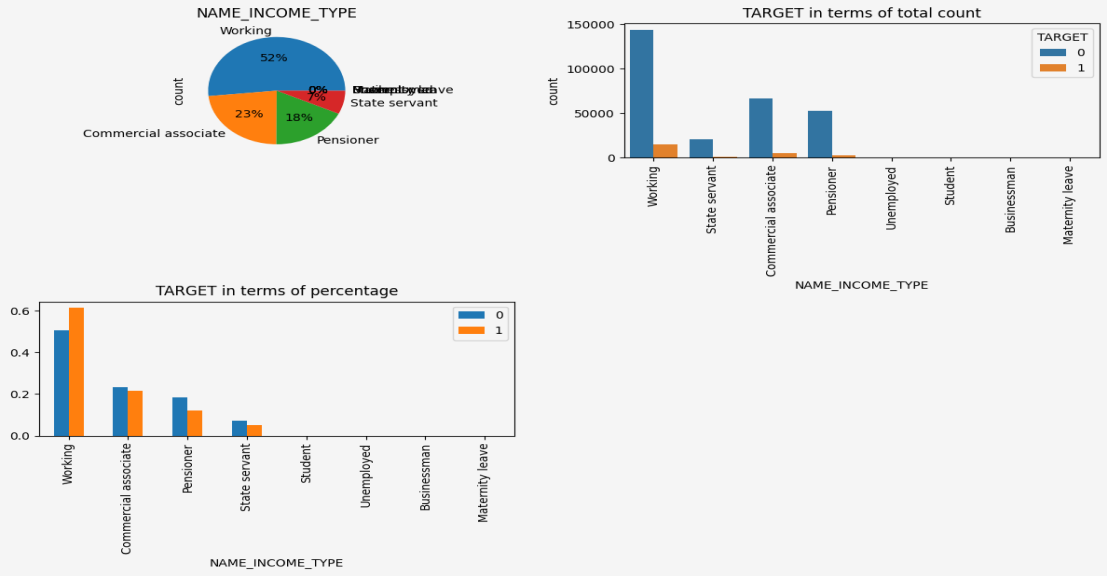


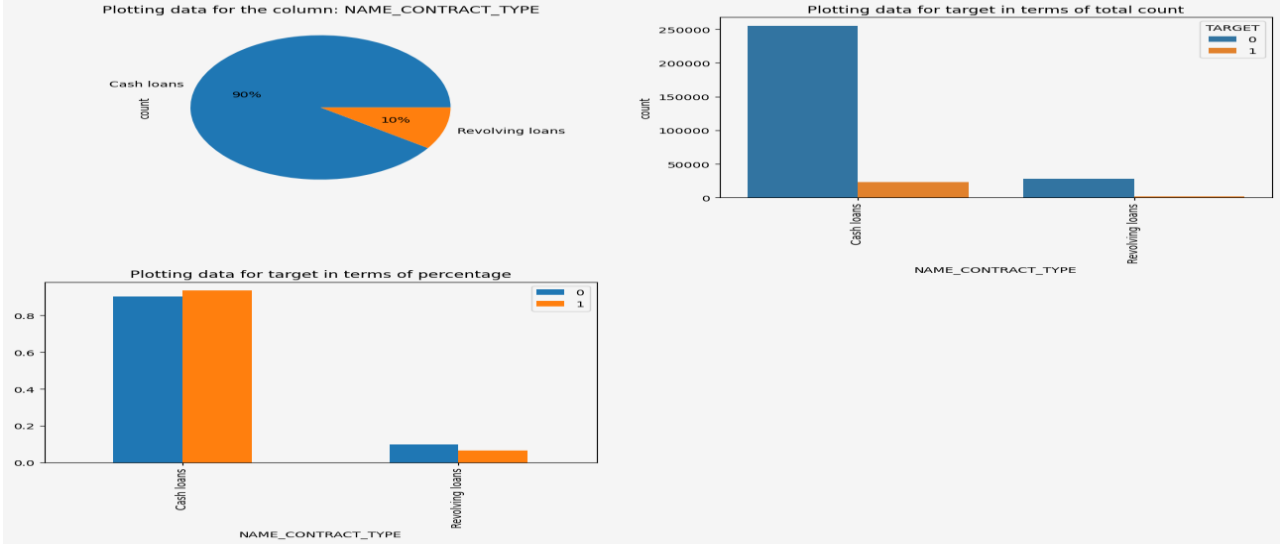


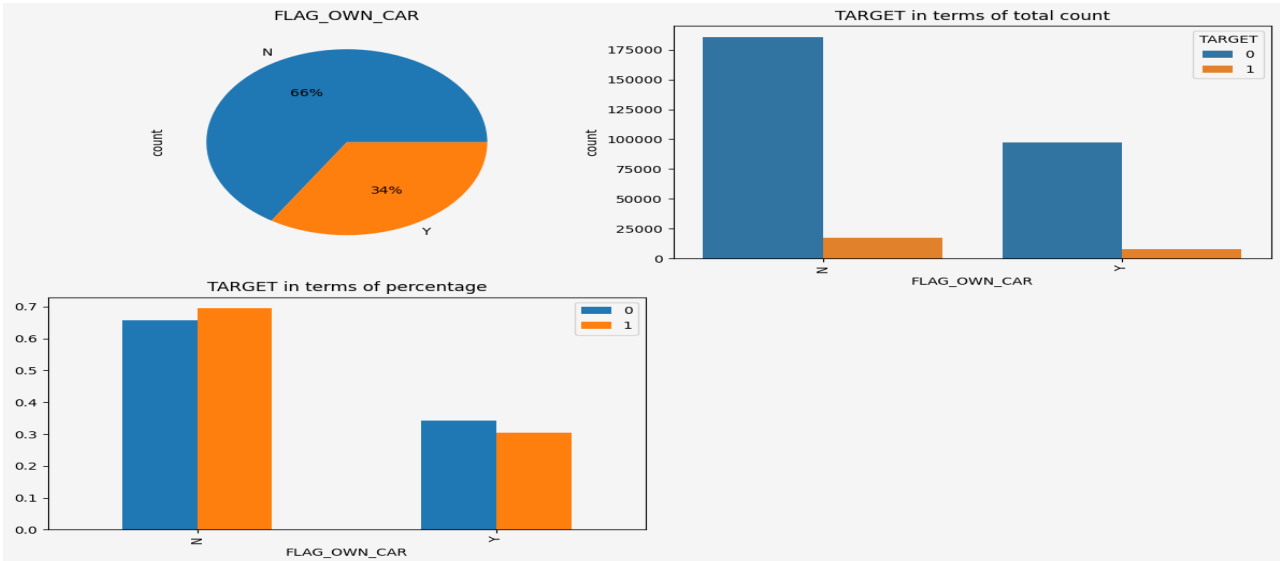


Data Cleaning & Analysis (1/2)



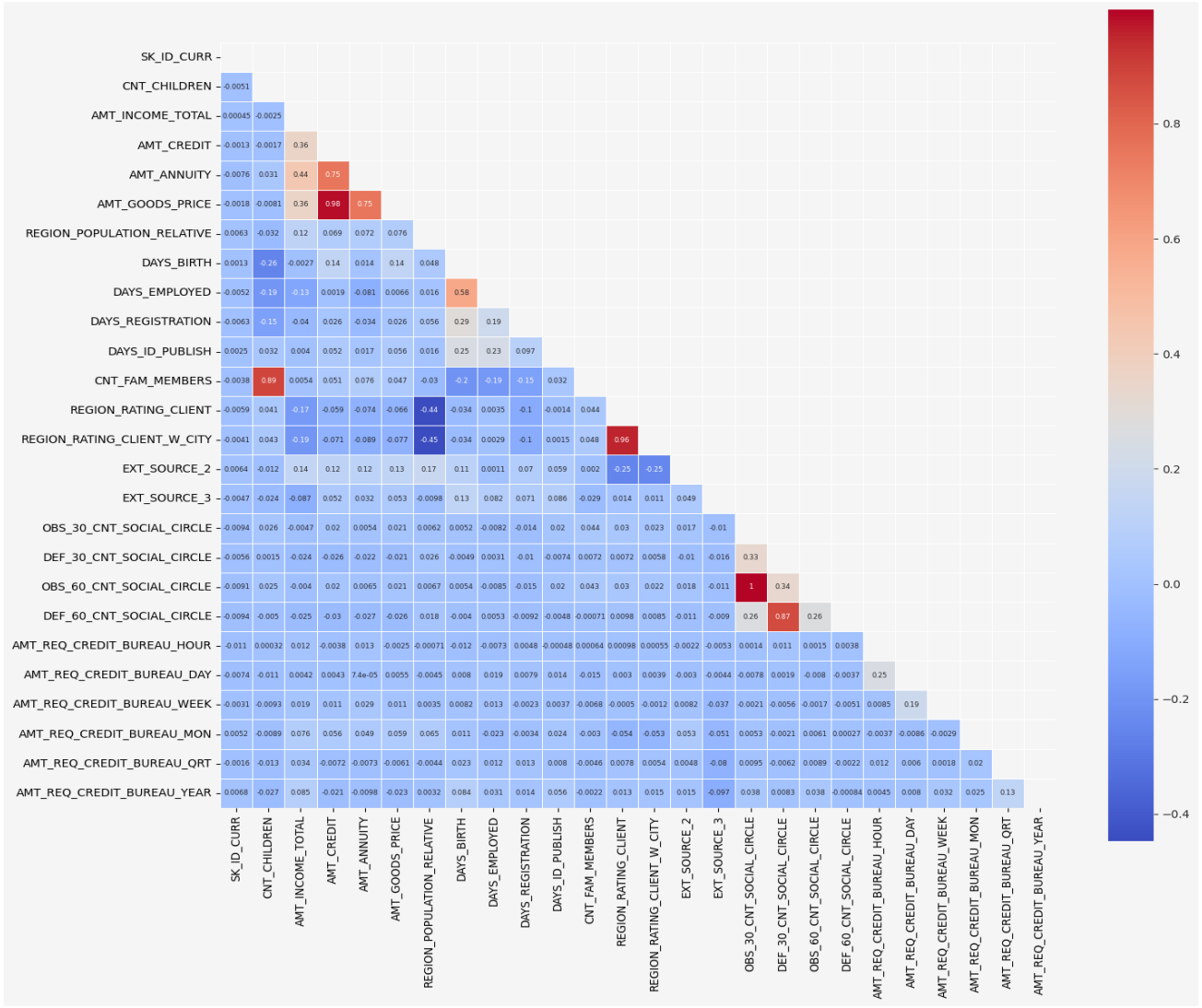




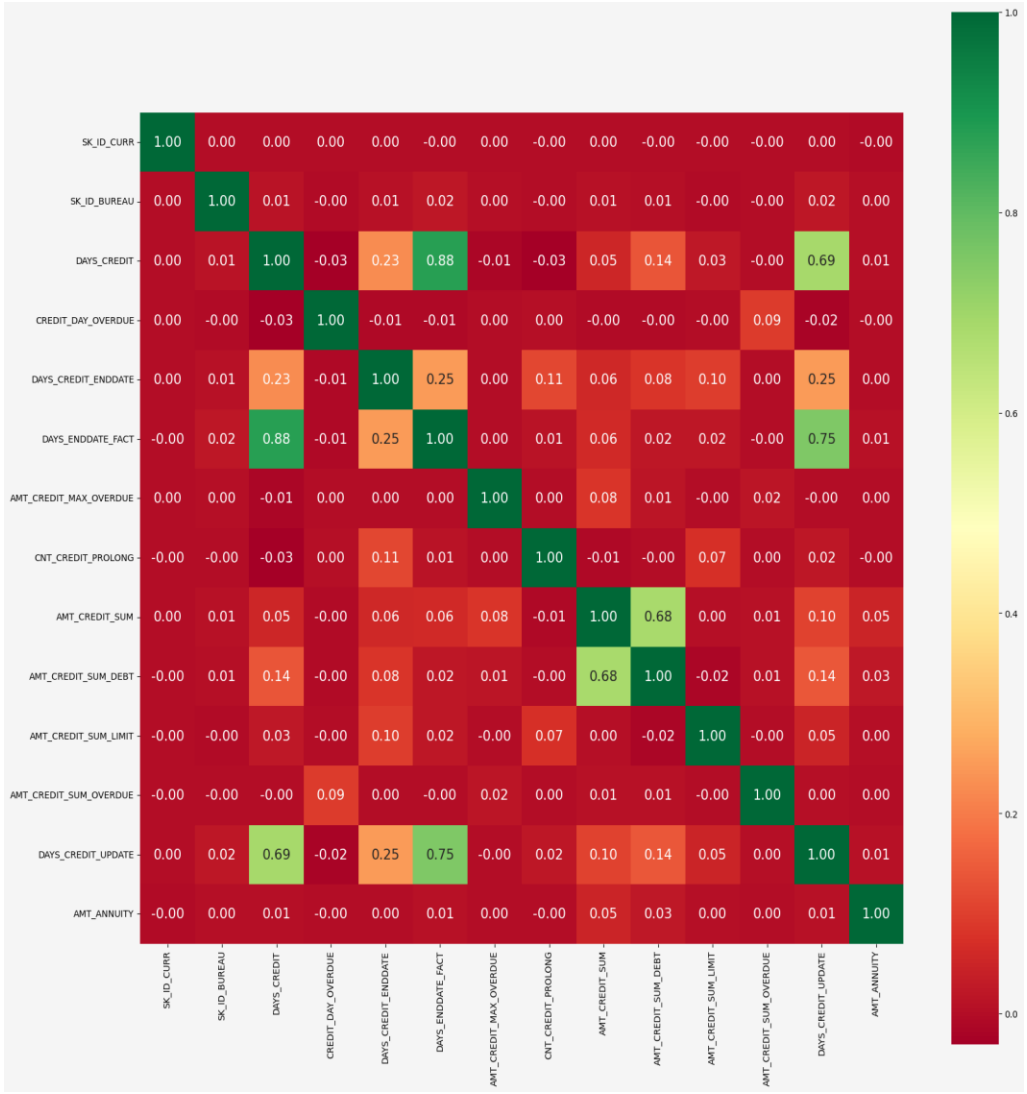


Correlation Analysis

Correlation - Defaulters



Correlation - Bureau



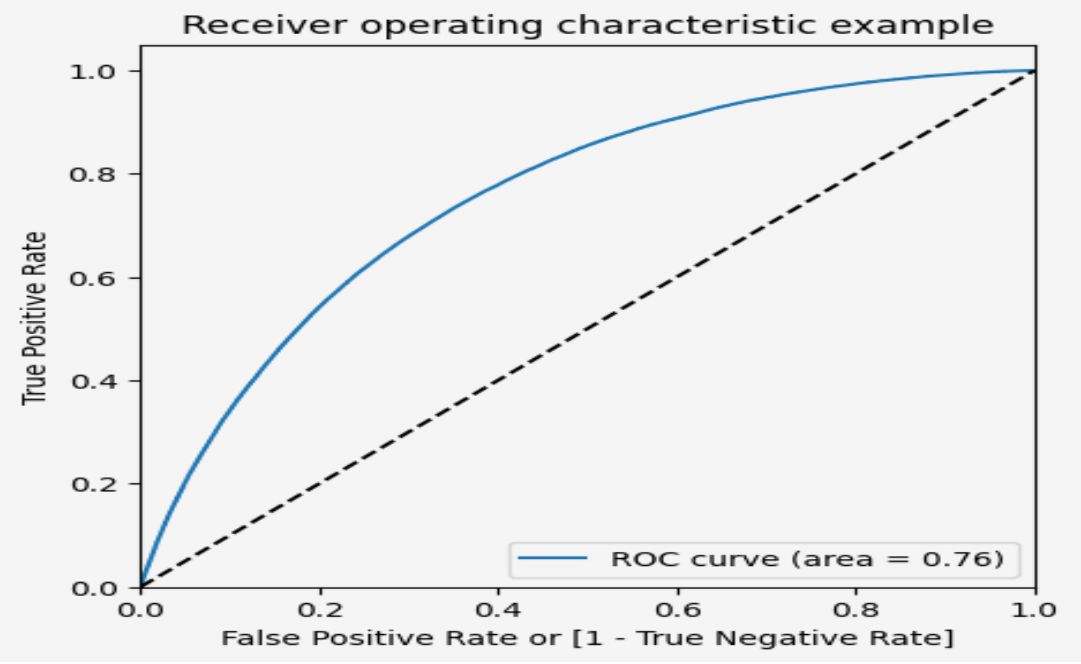
Feature Engineering & Model Building

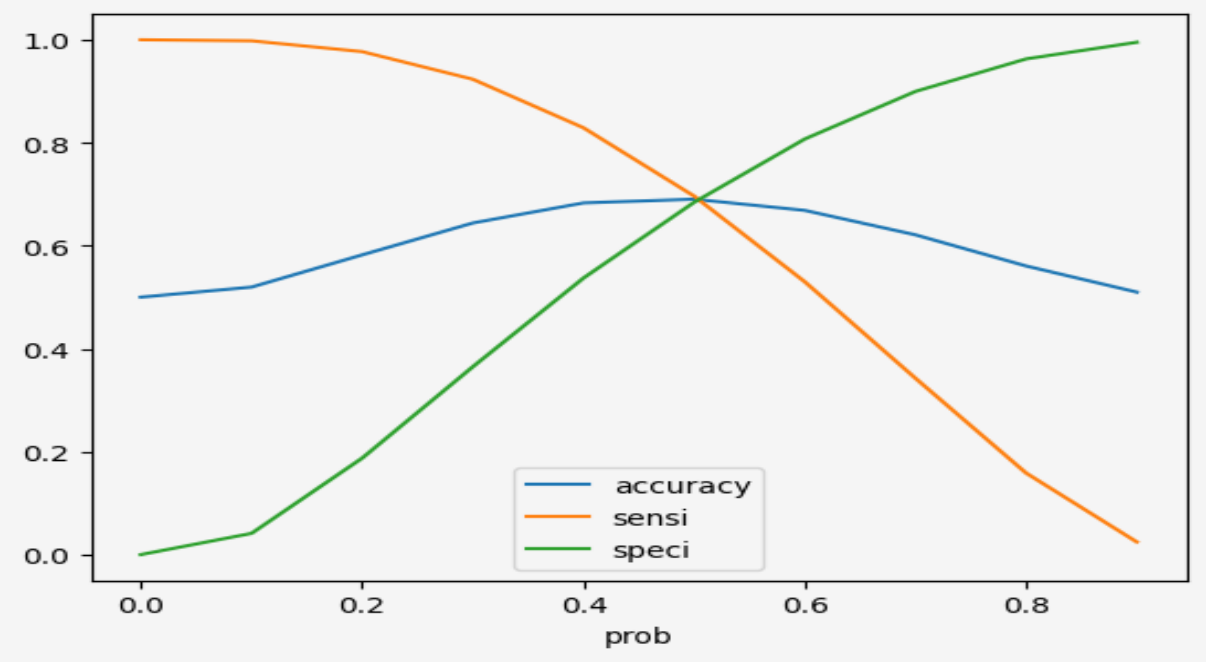
- One-hot encoding for categorical variables
- Aggregating with mean/median of trade level data to applicant level
- Merging the two cleaned up data sets i.e. Application data & Bureau data
- Dropped repeated, unwanted and highly correlated variables from the dataset
- Split the train and test data - using train_test_split library
- 70% of data are train data set and 30% of data are test data set
- Feature scaling using standard scaler
    - Creation of dummy variables for both the application data and bureau data set variables
- Handling class imbalance using SMOTE & TOMEK technique
- Feature selection using RFE
- Model building using Logistic Regression
    - Iterative model run is performed to check for the auto correlation and insignificant variables using the p-value and VIF
- Model evaluation & Cross validation

Model Evaluation

- Finding the optimal cutoff point on the curve

Confusion Matrix	Train Data	Test Data
Accuracy	69.11%	65.73%
Sensitivity	69.62%	68.66%
Specificity	68.43%	65.47%





Conclusion & Way forward

- The firm should target clients with higher education who are less likely to default
- Clients who own car are less likely to default
- There is a -ve correlation between count of children and total income for Defaulters. The firm should target clients with low count of children.
- Pensioners default rate is relatively less than other income types and can be targeted more. State servants and commercial associate can also be targeted given they are negatively correlated.
- Home Credit should target Female client more as the compared to Male clients as they are less likely to default.

THANK YOU