Credit EDA Case Study

BY HARSHALI NIKAM

AND

PRATHYUSHA D

Problem Statement

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
2. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

AIM: This case study aims to identify the variables that can be strong indicators of loan default.

Data Analysis

Null Value Treatment – Previous Application

- Null value % of individual columns have been calculated.
- Columns with more than 50% null values have been dropped.
- Columns with 13% null values were selected. Out of this, 5 relevant columns have been imputed with mean (Continuous) and mode (Categorical) basis the variable type.

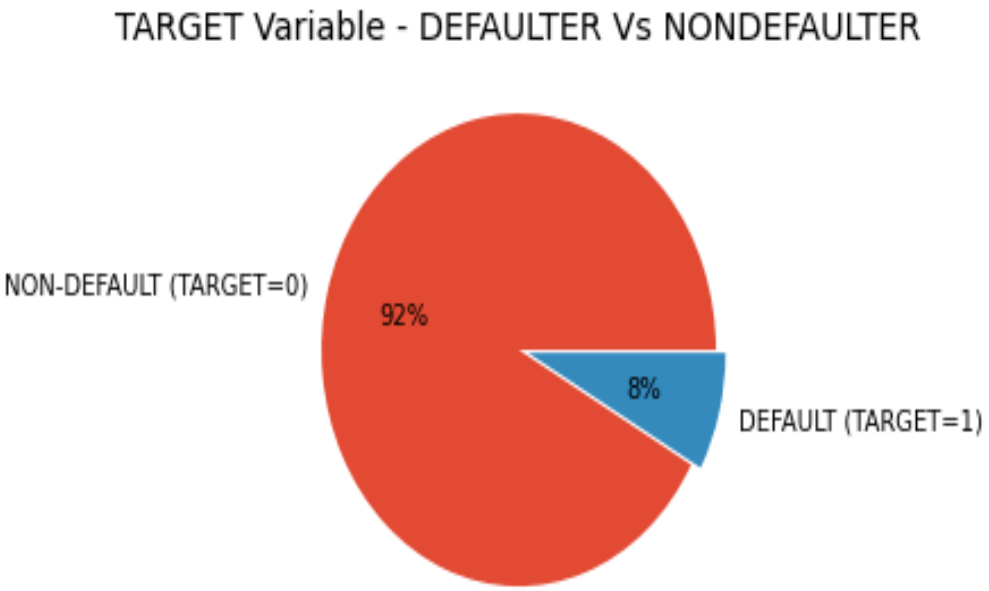
Datatype check – Previous Application

- Selected columns with float64 datatype have been changed to int64.
- Selected columns with datatype ‘object’ have been changed to ‘str’.
- Gender Column was normalised to ‘Male’ and ‘Female’ eliminating the ‘XNA’ (4 in number) value.

Binning

- Customers were categorized into 5 bins of Very Low, Low, Medium, High, Very High basis their income (AMT_INCOME_TOTAL)
- Customers were binned into 5 year age groups to understand default rate at age group level.
- Credit income ratio and % of social circle that have defaulted within 30 and 60 days have been calculated.

Target Variable



- It was noticed that the default rate is lower compared to non-default rate.
- 30 columns, seemingly relevant to the problem statement have been moved into a new dataset.
- Two data frames were further created within this new dataset. This divides the dataset into Target = 1 and Target = 0

Univariate Analysis - Application Data

- Female customers apply for more loans and default lesser, while the male customers apply lesser number of times and default more.
- Most of the customers do not own cars. It was observed that customers that own a car have defaulted lesser.
- Working individuals have a higher default rate, while students and businessmen don't default at all. Customers of type Pensioners, commercial associate and state servant have lower default rate.
- Married people apply for loans more and default comparatively less. Single, separated and people in civil marriages have a higher chances to default.
- People with a house/apartment apply for loans more and default comparatively less. People living with parent's default more when compared with others. This could be due to the added financial expenditure towards the family.

Univariate Analysis - Application Data

- Customers aged under 30 default the most. Customers over 30 don't tend to default.
- The very high-income group tend to default less often as their contribution to defaulters is lesser than the contribution of non defaulters.
- Almost all of the Education categories are equally likely to default. Chances of higher educated ones are less likely to default and secondary educated people are more likely to default.
- More loan applications for 2 rated. More defaulters for people living in better areas(Rating 3). Less defaults for people living in 1 rated areas.
- Family of 3 apply loan more often than the other families.

Bi-Variate Analysis – Application Data

- Customers are likely to default when they have a small family and the credit amount is low. It was observed that larger families with higher credit amount default less.
- Default rate is lower when amount of credit and price of goods is greater than 150,000.

Univariate Analysis – Previous application

- Most applications are for 'Cash loan' and 'Consumer loan'. Cash loans are refused more often than others.
- Clients chose to repay the loan using the 'Cash through the bank'.'Non-Cash from your account' & 'Cashless from the account of the employee' are not popular options of loan repayment.
- Most of the loan applications are from repeat customers, out of the total applications 70% of customers are repeaters. They also get refused most often.
- Annuity is directly proportional to the below factors.
- Annuity of prev_app has a very high and positive influence over: a) Amount on previous credit application b) Final approved credit amount on the previous application. c) Goods price that client asked for on the previous application.

Bi-Variate Analysis – Previous application

- Loan application for people with lower AMT_ANNUITY gets canceled or Unused mostly. Applications with too high AMT ANNUITY got refused more often than others.
- We can infer that when the AMT_CREDIT is too low, get's cancelled/unused most of the time.

Merged Data - Inferences

- Car ownership doesn't have any effect on application approval or rejection. But earlier we saw people that have a car have lesser chances of default. The bank should add more weightage to car ownership while approving a loan amount.
- Similar to car ownership, the bank can add more weightage to female while approving a loan amount.
- People with pre-approved loans defaulted less, where as people who were refused a loan earlier have higher chances of defaulting.

Case Study Summary

LOWER DEFAULT RATE

- Female Customers
- Car owners
- Businessmen/pensioners/commercial associates/state servants.
- Married People
- House/apartment owners
- Aged over 30
- Belonging to high income group
- Family of three. Large Family with high credit amount.
- Higher credit amount and price of goods.

HIGHER DEFAULT RATE

- Working Individuals
- Single/Separated/Civil Marriage
- Customers living with parents
- Secondary educated
- Cash loans
- Previously rejected/cancelled