EDA ASSIGNMENT ON LOAN DATASET

Dhruv Pathak

DS C46

INTRODUCTION

- This case study aims to identify patterns which indicate if an applicant has difficulty in paying his/her installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected and the number of defaulters is also reduced.

- Bank wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The bank can utilize this knowledge for its portfolio and risk assessment.

BUSINESS UNDERSTANDING

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants are capable of repaying the loan are not rejected.

STEPS

- DATAUNDERSTANDING AND SOURCING
- CHECK FOR DATA QUALITY
- CHECK FOR DATA IMBALANCE AND UNIVARIATE,BIVARIATE ANALYSIS
- MERGING OF APPLICATION DATA WITH PREVIOUS APPLICATION DATA
- DATA ANALYSIS BY UNIVARIATE, SEGMENTED UNIVARIATE AND BIVARIATE ANALYSIS
- RECOMMENDATIONS AND RISKS

PREVIOUS APPLICATION

- This dataset is highly imbalanced
- The applicants whose previous loans were approved are more likely to pay current loan in time, than the applicants whose previous loans were rejected. **NAME_CONTRACT_STATUS** is an important feature.
- 7% of the previously approved loan applicants that defaulted in current loan
- 90 % of the previously refused loan applicants that were able to pay current loan
- 'SCO', 'LIMIT' and 'HC' are the most common reason of rejection.

PREVIOUS APPLICATION

- Most of the people did not request insurance during previous loan application.
- For "Cards" defaulter percentage is highest (17%). **'NAME_PORTFOLIO'** is an important feature for analyzing 'TARGET' variable.
- 15% loan application defaulted for AP+ (Cash Loan). **'CHANNEL_TYPE'** is an important feature for analyzing 'TARGET' variable.
- Highest percentage (17%) of default cases is for 'Card Street'. **'PRODUCT_COMBINATION'** is an important driving factor.

APPLICATION DATASET

- -Family Info: (Important driving features : **'CNT_FAM_MEMBERS', 'CNT_CHILDREN'**)
- Most of the applicants are married (and/or) no children (and/or) 2 family members.
- Applicants with relatively more number of children (and/or) family members have higher default percentage. (For some of the cases where count children/family members is high, and the default rate is very high or very low. This cases cannot be considered for analysis as number of applicants having a large family is very low.)
- Education and Occupation Info: (Important driving features :**'NAME_INCOME_TYPE', 'OCCUPATION_TYPE'**)
- Most of the applicants are working.
- Applicants on Maternity Leave and Unemployed has highest percentage of Defaulter
- Businessman have lowest (0) percentage of Defaulter However applicants of income type('Unemployed', 'Student', 'Businessman', 'Maternity leave') are very few in the dataset to contribute in the analysis.

APPLICATION DATASET

- **CODE_GENDER**
- Female applicants are more than male applicants
- Defaulter percentage is higher for male applicants
- **DAYS_BIRTH**
- A derived column 'Age' from this gave useful information.
- People of age 25-35 have higher default rate
- Default cases are less for applicants more than 40 years old.

CONCLUSION

- <u>GROUP OF PEOPLE LESS CHANCE TO BE DEFAULTER</u>
  - CLIENT WITH HIGHER EDUCATION
  - HIGH INCOME CATEGORY
  - WHOSE'S PREVIOUS LOAN WAS APPROVED
  - BUSINESSMAN

<u>RISKY GROUP</u>

- LOWER SECONDARY EDUCATION CLIENTS
- MALE CLIENTS WITH CIVIL MARRIAGE
- PREVIOUSLY REFUSED LOAN
- UNEMPLOYED AND MATERNITY LEAVE