EXPLORATORY DATA ANALYSIS

STEPS IN EDA

- UNDERSTAND BUSINESS REQUIREMENTS
- DATA UNDERSTANDING
- DATA CLEANING
- ‘DATA ANALYSIS
- GAIN INSIGHTS

UNDERSTAND BUSINESS REQUIREMENTS

- To develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.
- To use EDA to analyze the patterns present in the data before loan sanction.
- 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:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- 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.
- To understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

DATA UNDERSTANDING

- Dataset has 3 files as explained below:

- *'application_data.csv'* contains all the information of the client at the time of application. The data is about whether a **client has payment difficulties.**

- *'previous_application.csv'* contains information about the client’s previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

- *'columns_description.csv'* is data dictionary which describes the meaning of the variables.

DATA CLEANING

- The application data has around 307511 rows and 122 columns.
- Columns with more than 45% are being dropped due to missing values
- The columns below 45% are being imputed with mean, median and mode
- These columns are segregated into categorical and numerical columns to apply imputation.
- Few datatypes were changed and derived like age from Days_Birth.

DATA ANALYSIS

- Low income group has more defaulters followed by high income group.
- GOODS_PRICE and AMT_CREDIT, AMT_ANNUTY and AMT_AMT_CREDIT are highly correlated
- Lower secondary educated clients are the most in number to be defaulted when their previous loans were cancelled or refused
- _Illustrates that irrespective of the income groups, the chances of default decreases as the age of the applicants increases.

GAIN INSIGHTS

- The defaulter percentage is 8% and non defaulter % is 92%.
- There is high chance to be defaulted of the young people. Non-defaulted people are almost equally distributed.

- Data Visualizations play an important role in plotting the Univariate, Bivariate and Multivariate columns and obtain the inferences

- Higher Percentage of loan was obtained by the foll categories as depicted
- Married People
- Unaccompanied People during loan processing
- Working
- For apartment purpose
- People without a car