BANK LOAN CASE STUDY

Project Description

The Bank Loan Case Study aims to analyse loan application data to identify key factors affecting loan defaults. The study involves handling missing data, detecting outliers, analysing data imbalance, performing statistical analysis, and identifying the most significant correlations that influence loan repayment behaviour.

Approach

1. **Handling Missing Data**
    1. Identified missing values using COUNTIF function.
    2. Approach for handling missing values is explained.
    3. Created a bar chart to visualize missing values across variables.
2. **Detecting Outliers**
    1. Used QUARTILE and IQR methods to detect outliers in numerical variables like AMT_INCOME_TOTAL, DAYS_EMPLOYED and others
    2. Created scatter plots to visualize outlier distribution.
    3. Removed extreme outliers whereever necessary for accurate analysis.
3. **Data Imbalance Analysis**
    1. Checked imbalance in the TARGET variable using COUNTIF.
    2. Calculated the proportion of TARGET = 1 (loan defaults) and TARGET = 0 (non-defaults).
    3. Created a pie chart to visualize data imbalance.
4. **Univariate, Segmented Univariate, and Bivariate Analysis**
    1. Performed univariate analysis on key financial and demographic factors using AVERAGE, MEDIAN, and STDEV. Also, prepared histograms, box plots and scatter plot for different variables for visualization.

    2. Conducted segmented univariate analysis by calculating the descriptive statistics based on Targett variable. Created charts for visualization.

    3. Performed bivariate analysis by creating a correlation matrix on major factors influencing the Target variable. Also created a scatter plot for better visualization.

5. **Correlation Analysis & Key Indicators of Default**

    1. Used the CORREL function to compute correlations for each segment.

    2. Identified top correlated features influencing loan defaults.

    3. Created heatmap to compare top correlations between both groups.

Tech Stack Used

- Microsoft Excel (Data cleaning, statistical functions, charts)
- Data Visualization (Bar charts, Box plots, Scatter plots, Heatmaps)

Insights and Results

1. Missing Data: Significant missing values in some financial attributes required imputation.

2. Outliers: Extreme outliers were detected in CNT_CHILDREN, AMT_INCOME_TOTAL and DAYS_EMPLOYED.

3. Imbalance Analysis: The dataset was highly imbalanced, with a small proportion of TARGET = 1 cases.

4. Loan Default Factors: Key financial indicators, such as AMT_CREDIT, AMT_ANNUITY, and DAYS_CREDIT, showed strong correlations with loan defaults.

5. Segmented Correlations: The correlation between AMT_CREDIT and AMT_ANNUITY was stronger in defaulters, indicating potential financial distress as a predictor.

Hyperlink:
https://docs.google.com/spreadsheets/d/1YkkY4CqwNNqvfudXDj1oeV9ICKogiYH8/edit?usp=sharing&ouid=111719312717552042778&rtpof=true&sd=true