Business Objectives

The objective of this project is to identify factors that contribute to credit card defaults, which can help the credit card company to reduce losses, mitigate risks, and improve customer satisfaction. By gaining insights into the data and understanding the relationships between variables, the credit card company can make informed decisions to improve its business operations, marketing efforts, and customer targeting.

Problem Statement

As a data analyst for a credit card company, you have been tasked with analyzing customer data to identify factors that contribute to credit card defaults. Your goal is to clean and prepare the data, perform exploratory data analysis to identify relationships between variables, and use inferential statistics and hypothesis testing to draw meaningful conclusions about the population based on a sample. Through this project, you will gain experience in data cleaning, exploratory data analysis, and statistical analysis, and provide recommendations based on the analysis to improve the credit card company's business operations.

Dataset

The columns/features in the given dataset are as follows:

- ID: ID of each client
- AMT: Amount of given credit in dollars (includes individual and family/supplementary credit)
- GENDER: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown)
- MARITAL STATUS: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years

• REPAY_SEP: Repayment status in September, 2005 (0=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)

• REPAY_AUG: Repayment status in August, 2005 (scale same as above)

• REPAY_JUL: Repayment status in July, 2005 (scale same as above)

• REPAY_JUN: Repayment status in June, 2005 (scale same as above)

• REPAY_MAY: Repayment status in May, 2005 (scale same as above)

• REPAY_APR: Repayment status in April, 2005 (scale same as above)

• AMTBILL_SEP: Amount of bill statement in September, 2005

• AMTBILL_AUG: Amount of bill statement in August, 2005

• AMTBILL_JUL: Amount of bill statement in July, 2005

• AMTBILL_JUN: Amount of bill statement in June, 2005

• AMTBILL_MAY: Amount of bill statement in May, 2005

• AMTBILL_APR: Amount of bill statement in April, 2005

• PRE_SEP: Amount of previous payment in September, 2005

• PRE_AUG: Amount of previous payment in August, 2005

• PRE_JUL: Amount of previous payment in July, 2005

• PRE_JUN: Amount of previous payment in June, 2005

• PRE_MAY: Amount of previous payment in May, 2005

• PRE_APR: Amount of previous payment in April, 2005

• DEF_AMT: Default payment (1=yes, 0=no)

Task - Data Cleaning and Exploratory Data Analysis

EDA Python

Data cleaning and exploratory data analysis are critical components of any data analytics project. In this task, you will clean and prepare the dataset, identify and handle any data quality issues, and create visualizations to identify relationships between variables.

- Load the dataset into a dataframe and handle any missing values, outliers, or duplicates. You can use libraries such as pandas and NumPy to handle these data quality issues.
- Calculate summary statistics such as the mean, median, mode, and standard deviation for the numerical variables. You can use the describe() function in pandas to calculate these statistics.
- Create visualizations such as histograms, scatter plots, and box plots to identify the distribution of the data and relationships between variables. You can use libraries such as matplotlib and seaborn to create these visualizations.
- Identify and handle any data quality issues that may arise. For example, if there are any missing values, you can either drop the rows or fill in the missing values with appropriate values.

Task - Inferential Statistics and Hypothesis Testing

Statistics

Inferential statistics and hypothesis testing are used to make inferences about a population based on a sample. In this task, you will use probability distributions to model the data, calculate confidence intervals, and conduct hypothesis tests to identify factors that contribute to credit card defaults.

- Use probability distributions such as the normal, Poisson, and binomial distributions to model the data. You can use libraries such as scipy.stats to model the data.
- Calculate confidence intervals to estimate population parameters. You can use the t-distribution and z-distribution to calculate confidence intervals for the mean and proportion, respectively.
- Conduct hypothesis tests using the t-test, F-test, ANOVA-test, and Chi-Square test to identify factors that contribute to credit card defaults. You can use libraries such as scipy.stats and statsmodels to conduct these tests.
- Interpret the results and provide recommendations based on the analysis. For example, if the hypothesis test shows that there is a significant difference in credit card defaults between different age groups, you can recommend that the credit card company target specific age groups with different marketing strategies.