PROJECT MILESTONE 3

PACE: Plan Stage

- What are the data columns and variables and which ones are most relevant to your deliverable?

The train dataset contains 18, and the test dataset contains 17 variables. The only difference is the ‘Default’ variable which is not included in the test set because its value should be predicted with this project. Besides ‘Default’ as a target variable, the following predictive variables are detected: ‘Age’, ‘InterestRate’, ‘Income’, ‘MonthsEmployed’ and ‘LoanAmount’.

- What units are your variables in?

The ‘Default’ variable is of ‘int64’ type with only two possible values: loan defaulted (1) or not (0).

Besides ‘Default’, there are additional 7 ‘int64’ variables, 2 ‘float64’ and 8 ‘object’ variables. 7 from 8 ‘object’ variables are categorical with predefined possible values.

- What are your initial presumptions about the data that can inform your EDA, knowing you will need to confirm or deny with your future findings?

Numerical data might follow a normal distribution or another common distribution (e.g., uniform, exponential).

There may be linear or nonlinear relationships between variables.

There might be outliers or anomalies that could impact the analysis.

The dataset might have missing or incomplete data.

Data might have inconsistencies, errors, or duplicates.

Variables might be appropriately encoded and classified (e.g., categorical vs. numerical).

Different groups or categories in the data might have distinct characteristics or behaviors.

Variances might be homogeneous across different groups or conditions.

There might be interactions between variables that affect the outcomes.

- Is there any missing or incomplete data?

No, there are no missing values or incomplete data.

- Are all pieces of this dataset in the same format?

No, the train dataset has 18 variables with three different dtypes: ‘object’ (8), ‘float64’ (2) and ‘int64’ (8), while the test set has 17 variables with three different dtypes: ‘object’ (8), ‘float64’ (2) and ‘int64’ (7).

- Which EDA practices will be required to begin this project?

First, it will be required to load the data and perform the initial inspection.

Second, it will be required to perform descriptive statistics (summary statistics and distribution analysis).

Third, it will be required to perform data cleaning (identify missing values, correct errors, and standardize data).

Forth, it will be required to perform exploratory visualization.

PACE: Analyze Stage

- What steps need to be taken to perform EDA in the most effective way to achieve the project goal?

Data exploration and cleaning.

- Do you need to add more data using the EDA practice of joining? What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?

At this phase of the project, I can't confirm adding more data.

The data is already in a structured format.

It is important to identify outliers if any.

- What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?

Depending on the type of audience, I will choose the right visualization type:

1. For Executives/Stakeholders I will use high-level visualizations that highlight key insights and trends like dashboards, summary charts, and executive summaries with clear, actionable insights with focus on simplicity, clarity, and direct relevance to business goals.

2. For Technical Experts/Data Scientists I will provide detailed and in-depth visualizations that allow for exploration and analysis of complex relationships like scatter plots, heatmaps, box plots and multidimensional plots.

3. For general public I will use intuitive and easy-to-understand visualizations that require minimal explanation like simple bar charts, pie charts and trend lines with focus on clear, straightforward presentations that convey the main message effect.

PACE: Construct Stage

- What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?

  - Box plots will be helpful to determine outliers and where the bulk of the data points reside.
  - Histograms are essential to understand the distribution of variables.
  - Scatter plots will be helpful to visualize relationships between variables.
  - Bar charts are useful for communicating levels and quantities, especially for categorical information.

- What processes need to be performed in order to build the necessary data visualizations?

In order to build the necessary data visualizations, first I will begin with examining the spread and the distribution of the important variables using box plots and histograms.

- Which variables are most applicable for the visualizations in this data project?

The ‘int64’ and the ‘float64’ variables are most applicable for the visualizations in this data project.

- Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

For now, I will leave them, because in this phase I plan to do only EDA analysis.

PACE: Execute Stage

- What key insights emerged from your EDA and visualizations(s)?

  - Analysis revealed that the overall loan default rate is ~12%.
  - I feel that the more deeply I explore the data, the more questions arise. In this case, it's worth asking the company’s data team to provide data about the spending habits and financial literacy (overuse of credit cards or payday loans, or lack of savings or emergency funds) of its clients. Additionally, the industries prone to layoffs or low job security would be a valuable information in predicting loan default.
  - Also, EDA has revealed that clients with bigger interest rates, bigger loan amounts and more number of credit lines are *more* likely to loan default, but clients who are older, with bigger incomes, longer employed, and with bigger credit score are *less* likely to loan default.

- What business and/or organizational recommendations do you propose based on the visualization(s) built?

➔ I recommend to further investigate the relations between key identified predictive variables and the target variable.

➔ Continue to explore user profiles with the greater company’s team; this may glean insights on the reason for the loan default.

➔ Plan to run deeper statistical analyses on the variables in the data to determine their impact on the loan default.

- How might I share these visualizations with different audiences?

- For Executives/Stakeholders I will use the high-level visualizations that highlight key insights and trends (hist plots of the distribution of important variables for churned and not churned users, scatter plots of important variables by churn status).
- 2. For the Data Team I will provide detailed and in-depth visualizations that allow for exploration and analysis of complex relationships (the rest of the charts).