Loan Default Prediction: An EDA Approach

This presentation outlines an Exploratory Data Analysis (EDA) approach to predict loan defaults. The goal is to identify patterns that indicate a client's difficulty in repaying loans, enabling informed decisions like loan denial or adjusted interest rates. This ensures that creditworthy applicants are not unjustly rejected, optimizing the lending process.

by Ramesh Sunkara

Business Understanding and Objectives

Problem Statement

Loan companies struggle to assess credit risk due to insufficient credit history, leading to potential defaults and financial losses.

Business Objective

Identify key indicators of loan default to improve risk assessment and portfolio management, ensuring deserving applicants are approved.

Goal

Minimize financial losses by accurately predicting loan defaults and optimizing lending strategies.

Data Overview and Understanding

Application Data

Contains client information at the time of loan application, indicating payment difficulties.

Previous Application Data

Includes data on previous loan applications, detailing outcomes like approval, cancellation, refusal, or unused offers.

Columns Description

Provides a data dictionary explaining the meaning of each variable in the datasets.

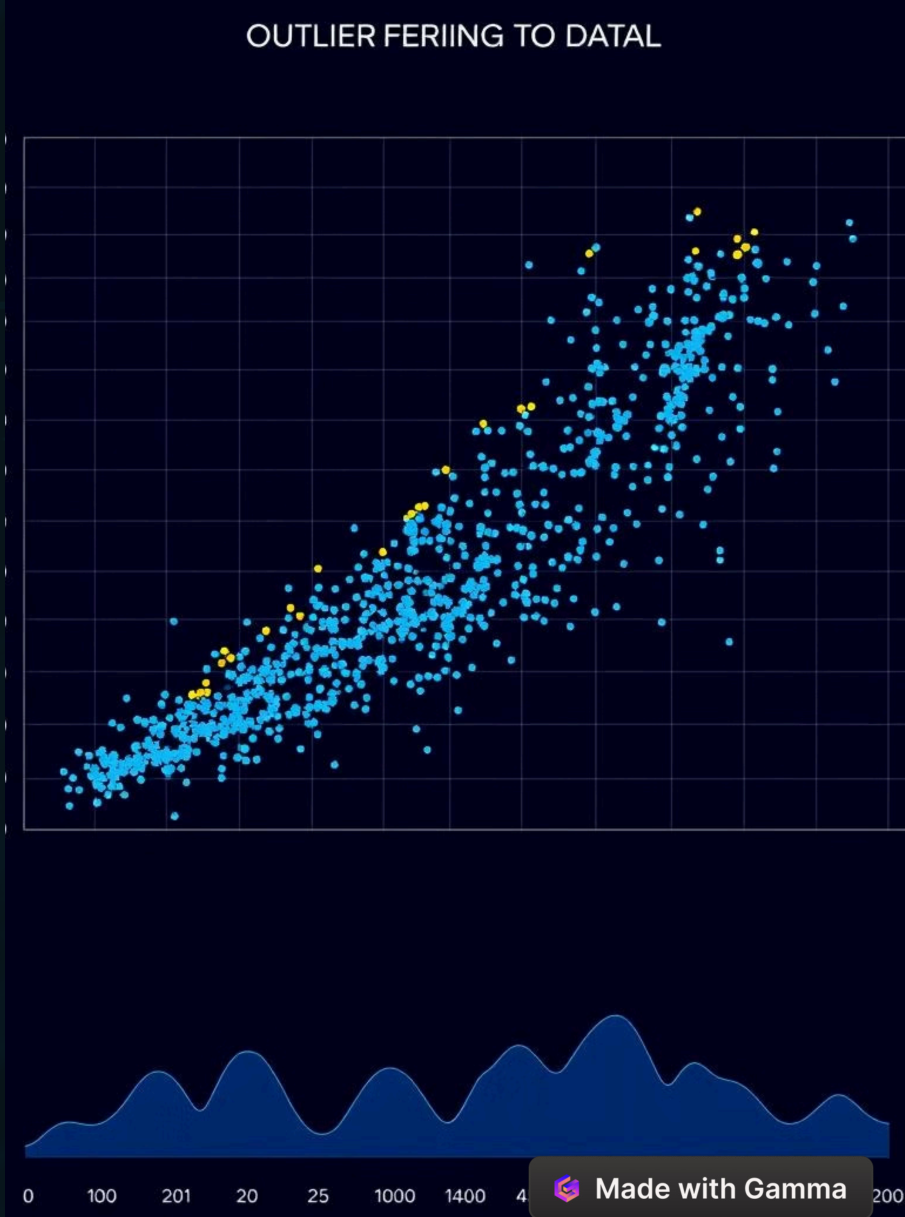
Handling Missing Data

Missing data is identified and addressed using appropriate methods. Columns with excessive missing values may be removed, while others are imputed with suitable values. The approach is clearly documented to maintain transparency and ensure data integrity.

Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.

Outlier Identification

Outliers are identified within the dataset, with justifications for their classification. While removal isn't necessary for this exercise, understanding their impact is crucial. Outliers can skew analysis and affect model performance, so their presence is carefully considered.

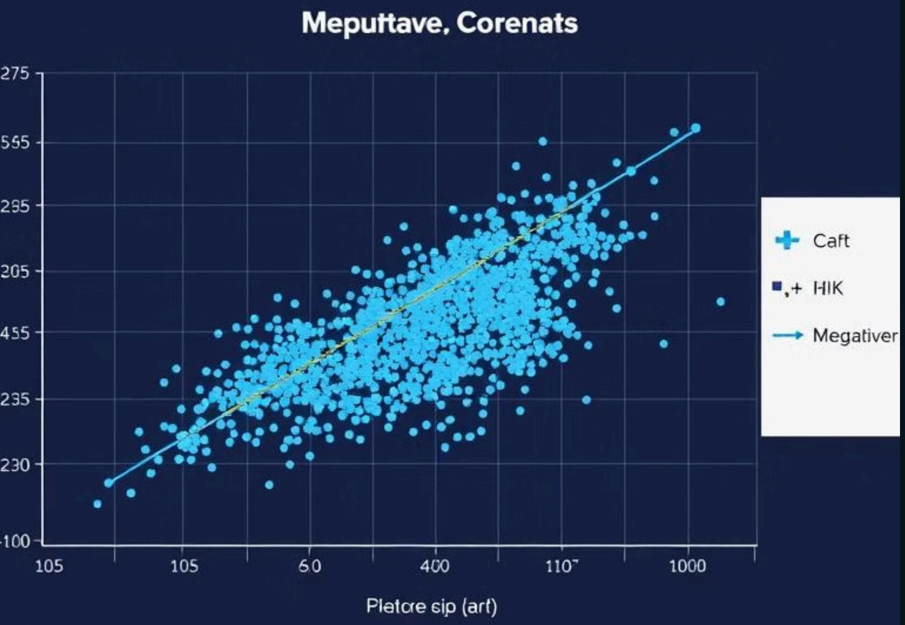


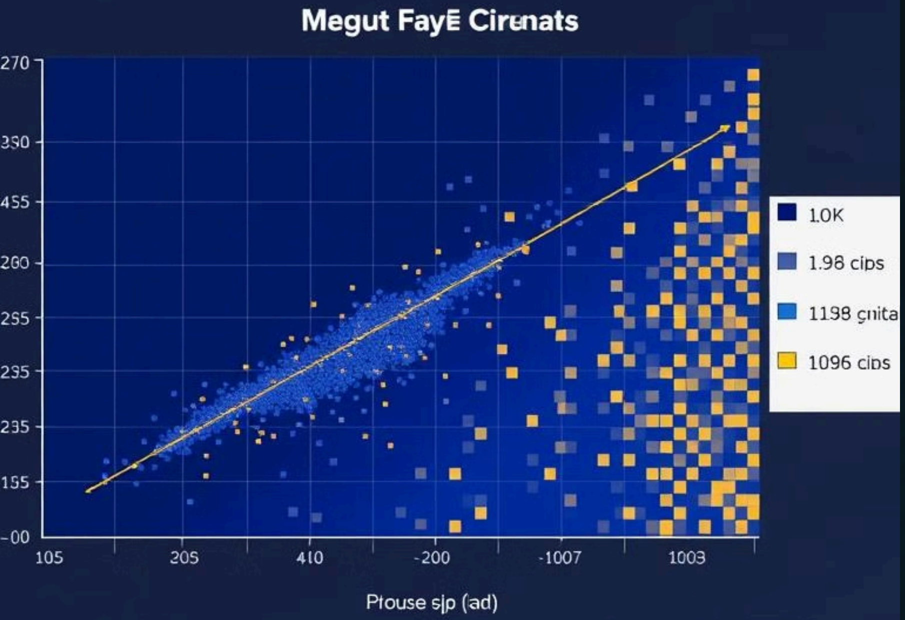
Data Imbalance Analysis

Data imbalance is assessed by examining the ratio of clients with payment difficulties versus all other cases. Visualizations, such as plots with varying scales (percentage or absolute value), are used to analyze different aspects of this imbalance. This analysis focuses on the 'Target variable' to understand its distribution.

Univariate Analysis Results

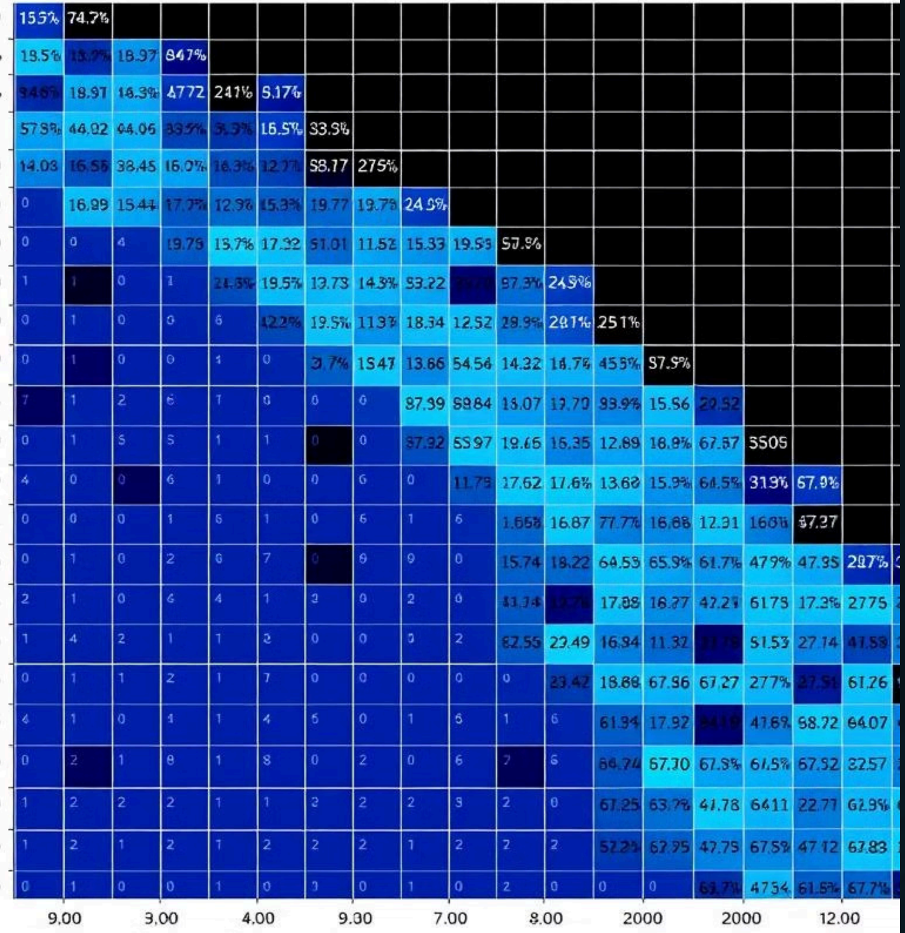
Univariate analysis results are explained in business terms, providing insights into individual variables. This includes understanding the distribution, central tendency, and spread of each variable. These insights help in identifying key characteristics of the client population and potential risk factors.





Bivariate Analysis Results

Bivariate analysis results are interpreted in business terms, highlighting relationships between variables. This includes segmented univariate analysis to understand how different groups behave. These analyses reveal potential drivers of loan default and inform targeted interventions.



Top Correlations with Target Variable

The top 10 correlations for clients with payment difficulties and all other cases are identified. This involves segmenting the data frame with respect to the target variable and finding the top correlations for each segment. Insights are derived by comparing these correlations to understand key differences.

Key Takeaways and Next Steps

Insights

Identified key variables driving loan default, enabling targeted risk mitigation strategies.

Actions

Implement stricter lending criteria, adjust interest rates, and refine portfolio management.

Future Work

Develop predictive models to automate loan approval processes and minimize financial losses.