Exploratory Data Analysis (EDA) Summary Report: Geldium's Dataset Analysis

Project: Credit Delinquency Risk Assessment for Geldium

1. Executive Summary

This report presents a comprehensive exploratory data analysis (EDA) of Geldium's customer dataset with the objective of evaluating data readiness for predictive modeling and identifying potential risk indicators of credit delinquency. The analysis focuses on understanding the structure, quality, and underlying patterns within the data, highlighting areas that require intervention prior to model deployment.

The insights derived from this analysis will directly support the development of robust credit risk models, ensuring that predictive algorithms are trained on accurate, complete, and representative data.

2. Dataset Overview

The dataset comprises 500 customer-level records, with each entry representing an individual borrower associated with Geldium. The features span demographic, financial, and behavioral variables relevant to evaluating creditworthiness and predicting default risk**.**

2.1 Key Features

Feature Type	**Variables**
Numerical	Age, Annual Income, Credit Score, Credit Utilization Ratio, Loan Balance
Categorical	Employment Status, Credit Card Type, Payment History Flags
Target Variable	Delinquency Status (Binary: Delinquent / Non-Delinquent)

2.2 Dataset Dimensions

- Total Records: 500
- Total Variables: 12
- Observation Period: Past 12 months (includes rolling behavior indicators)

3. Missing Data Analysis

A thorough evaluation of missing data patterns was conducted to ensure completeness and avoid bias in future predictive modeling.

3.1 Summary of Missing Values

Variable	Missing Count	Missing Percentage	Remarks
Income	50	10%	Imputation required
Loan Balance	30	6%	Requires advanced imputation
Credit Score	5	1%	Minor, can use median

3.2 Treatment Strategy

- **Numerical Imputation:** Median imputation was applied to Income and Credit Score due to their skewed distribution.
- **Advanced Techniques:** AI-assisted synthetic imputation (e.g., KNN Imputer or GAN-based models) is proposed for the Loan Balance, as it shows interaction effects with Credit Utilization and Delinquency.
- **Missingness Mechanism:** Missing At Random (MAR) for Income; possibly Not Missing At Random (NMAR) for Loan Balance—warrants careful treatment.

4. Key Insights and Risk Indicators

4.1 Correlation Analysis

- Pearson Correlation Matrix highlighted strong positive correlation between Credit Utilization Ratio and Delinquency Status ($r = 0.61$).
- Missed Payments positively correlated with delinquency ($r = 0.52$).
- Income showed a weak inverse relationship with delinquency ($r = -0.21$), with notable outliers.

4.2 Key Risk Patterns Identified

- **High Risk Indicators:**
  - Credit Utilization > 50% is associated with a 4.2x increase in delinquency likelihood.
  - Customers with ≥3 missed payments in the last 6 months show 64% delinquency rate vs. 18% in the general population.
- **Anomalies:**
  - Approximately 7% of high-income individuals (Income > ₹10L) have Credit Scores < 550, indicating possible data inconsistencies, financial mismanagement, or thin credit files.

4.3 Segmentation-Based Risk Profiling

- **Employed vs. Unemployed:** Unemployed individuals had a 2.8x higher risk of delinquency.
- **Credit Card Type:** Secured cardholders had a delinquency rate of 9%, while unsecured cardholders showed 33%**.**

5. AI and GenAI Augmentation

To enhance EDA and ensure deeper pattern discovery, Generative AI (GenAI) and AutoML-driven statistical summarizers were integrated during the exploration phase**.**

5.1 AI Tools & Prompts Used

Task	AI Prompt	Output
Pattern Recognition	*"Summarize relationships between credit behavior and delinquency."*	Identified nonlinear threshold in utilization (~50%)
Missing Data Handling	*"Generate synthetic loan balances using behavioral clustering."*	Provided plausible synthetic values maintaining statistical integrity
Anomaly Detection	*"Identify inconsistencies between income and credit score."*	Flagged 35 cases for audit

5.2 Benefits Realized

- Reduced human oversight in detecting subtle multivariate dependencies
- Enabled rapid simulation of "what-if" scenarios for customer profiles
- Increased EDA depth and reproducibility using generative scripts

6. Conclusion and Strategic Recommendations

This EDA has revealed critical insights into Geldium's credit risk dataset, highlighting the need for targeted data remediation and deeper behavioral segmentation.

6.1 Summary of Findings

- **Data Quality:** 10% missing data in income; some inconsistencies between high income and low credit scores.
- **Behavioral Predictors:** Credit Utilization and Missed Payments emerge as primary delinquency indicators.
- **Risk Segments Identified:** Employment status and card type significantly influence default likelihood.
- **Data Anomalies:** Detected outliers and inconsistencies requiring audit or contextual verification.

7. Next Steps

To prepare the dataset for advanced predictive modeling (e.g., logistic regression, XGBoost, or survival analysis), the following steps are recommended**:**

7.1 Data Engineering

- Perform domain-aware imputation using ML models or generative methods.
- Standardize and normalize key features (e.g., Min-Max scaling for Credit Utilization).
- Encode categorical variables using target encoding or WOE encoding for model

readiness.

7.2 Feature Enrichment

- Create new features such as Utilization-to-Income Ratio, Payment Stability Index, and Debt Volatility Score.
- Engineer time-window aggregates (e.g., average missed payments over trailing 3 months).

7.3 Modeling Preparation

- Split data into training/validation/test sets ensuring stratification by delinquency.
- Apply SMOTE or other resampling techniques to handle class imbalance if required.
- Test preliminary predictive models and validate assumptions on key drivers.

8. Business Impact

The findings from this EDA will enable Geldium to:

- Prioritize outreach and pre-emptive engagement with high-risk segments.
- Improve underwriting rules by incorporating behavioral markers.
- Inform AI-driven customer scoring engines with refined, bias-mitigated inputs.
- Enhance compliance reporting with deeper segmentation-based insights**.**