**Exploratory Data Analysis (EDA)**

**Report**

**1. Introduction**

This report summarizes the exploratory analysis conducted on Geldium’s customer dataset. The goal of this study is to review data quality, uncover patterns related to credit risk, and highlight features that could influence the probability of default. The findings will help prepare the dataset for robust predictive modeling and risk assessment.

**2. Dataset Summary**

The dataset contains information on 500 customers, capturing both financial and demographic attributes that may impact repayment behavior. It includes numerical variables like income and outstanding balance, along with categorical variables such as employment type.

Key Details:

* Total records: 500
* Core variables: Age, Income, Credit Score, Credit Utilization, Missed Installments, Debt-to-Income Ratio
* Data types:
  + *Categorical:* Employment Status, Card Type
  + *Numerical:* Annual Income, Loan Balance

**3. Missing Data Evaluation**

Some variables contain incomplete entries, which could bias analysis if not addressed. In particular, missing values are concentrated in income and loan balance fields.

Findings:

* Income: 50 records missing
* Loan Balance: 30 records missing

Proposed Treatment:

* Replace missing values in numerical fields with the median.
* For loan balance, consider AI-based synthetic value generation to maintain distribution consistency.

**4. Key Insights and Risk Factors**

Preliminary exploration shows a strong association between financial behavior and default risk.

Key Observations:

* Customers utilizing more than half of their available credit limit show increased delinquency risk.
* Borrowers with three or more missed installments within six months are significantly more prone to default.
* Anomalous cases exist, such as clients with high incomes but unexpectedly low credit scores, requiring further investigation.

**5. Role of AI & GenAI**

AI-assisted analysis was used to detect missing data patterns, surface anomalies, and validate early risk signals. Generative AI queries helped accelerate data interpretation by providing quick summaries and cross-checks with industry-standard metrics.

Example AI Prompts Used:

* “Highlight variables with missing information and summarize their potential impact.”
* “Evaluate the link between credit utilization, payment delays, and default probability.”

**6. Conclusion & Future Actions**

The exploratory review has provided valuable insight into the dataset’s strengths and weaknesses, highlighting both risks and opportunities for further analysis.

Takeaways:

* Data quality gaps**:** Missing values in income and loan balance could reduce model accuracy if untreated.
* Risk signals: Heavy credit utilization and frequent missed payments appear to be the strongest predictors of delinquency.
* Irregular cases: Customers with mismatched profiles (e.g., high income but poor credit scores) need closer scrutiny.

Recommendations:

* Apply appropriate imputation methods to handle missing data effectively.
* Validate whether identified risk drivers (credit use, missed payments) hold true across different customer groups.
* Conduct deeper checks on inconsistent entries to determine whether they are errors or genuine risk signals.

By addressing these points, Geldium can strengthen its credit risk models and ensure more reliable predictions in future analyses.