**Exploratory Data Analysis (EDA) Summary Report : Credit Card Delinquency Risk Prediction**

1. Introduction:

This report aims to analyze Geldium’s financial dataset to assess its suitability for delinquency risk prediction. Through exploratory data analysis (EDA), we will identify data quality issues, detect early risk indicators, and ensure that the dataset is clean and structured for effective modeling. The findings will inform Tata iQ’s analytics team on key patterns that can improve delinquency risk intervention strategies.

2. Dataset Overview:

This dataset contains financial and behavioral attributes of customers, designed to predict delinquency risk based on credit history, financial stability, and employment status. Below is a summary of the dataset’s key attributes:

Key dataset attributes:

* Number of records: 500 (customers)
* Key variables:

- Customer\_ID – Unique identifier for each customer (Categorical)

- Age – Customer’s age in years (Numerical)

- Income – Annual income in USD (Numerical, contains missing values)

- Credit\_Score – Credit score ranging from 300 to 850 (Numerical)

- Credit\_Utilization – Percentage of available credit in use (Numerical)

- Missed\_Payments – Total missed payments in the past year (Numerical)

- Delinquent\_Account – Indicator of delinquency status (Binary: 0 = No, 1 = Yes)

- Loan\_Balance – Outstanding loan balance (Numerical, contains missing values)

- Debt\_to\_Income\_Ratio – Debt compared to income percentage (Numerical)

- Employment\_Status – Current employment status (Categorical, contains inconsistencies)

- Account\_Tenure – Number of years an account has been active (Numerical)

- Credit\_Card\_Type – Type of credit card held (Categorical)

- Location – City or region of residence (Categorical)

- Month\_1 to Month\_6 – Payment history across six months (Categorical)

* Data Types:

1. Categorical Variables: `Customer\_ID`, `Employment\_Status`, `Credit\_Card\_Type`, `Location`, `Month\_1` to `Month\_6`
2. Numerical Variables: `Age`, `Income`, `Credit\_Score`, `Credit\_Utilization`, `Missed\_Payments`, `Loan\_Balance`, `Debt\_to\_Income\_Ratio`, `Account\_Tenure`
3. Binary Variable: `Delinquent\_Account`

* Notable Anomalies and Inconsistencies:

1. Missing values: `Income`, `Loan\_Balance`, `Credit\_Score`
2. Inconsistent data: `Employment\_Status` includes non-standard representations (e.g., "EMP" vs "Employed").
3. Outliers: `Credit\_Utilization` values exceeding 1 indicate potential errors.
4. Unusual values: Some customers report missed payments but have no delinquency status.

# 3. Missing Data Analysis:

Identifying and addressing missing data is critical to ensuring model accuracy. This section outlines missing values in the dataset, the approach taken to handle them, and justifications for the chosen method.

* Key Missing Data Findings:

Variables with missing values:

1. Income (39 missing values)
2. Loan\_Balance (2 missing values)
3. Credit\_Score (29 missing values)

* Missing Data Treatment Strategy:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Handling Method** | **Justification** |
| Income | Imputation (Median) | Preserves realistic distribution and avoids skewing due to extreme values. |
| Loan\_Balance | Imputation (Mean) | Ensures consistency in debt calculations while maintaining overall dataset integrity. |
| Credit\_Score | Predictive Imputation | Uses regression modeling to estimate scores based on other financial indicators like income and missed payments |

Additionally:

* Employment\_Status was standardized ("EMP" → "Employed", "Self-employed" → "Self-Employed").
* Month\_1 to Month\_6 were encoded (On-time=0, Late=1, Missed=2) for predictive modeling.

### **Key Outliers & Treatment:**

* Credit\_Utilization **capped at 1.0** (values exceeding 100% were invalid).
* Debt\_to\_Income\_Ratio **capped at 0.552** (excessively high values adjusted).
* Income & Loan\_Balance outliers removed using the **Interquartile Range (IQR) method**.

By addressing these missing values systematically, we ensure that the dataset remains robust and reliable for delinquency risk modeling.

# 4. Key Findings and Risk Indicators:

1. Credit Score & Delinquency Trends

* The median credit score remains similar for both delinquent and non-delinquent customers (~600).
* Customers with lower credit scores (<400) exhibit a slightly higher delinquency rate.
* Outliers are minimal, indicating a well-structured dataset without extreme values.

2. Debt-to-Income Ratio & Financial Stability

* Customers with a higher Debt-to-Income Ratio (>0.4) show a greater likelihood of delinquency.
* Many delinquent accounts belong to individuals with a Debt-to-Income Ratio near the dataset's maximum (0.55).
* Employment status affects financial stability: Unemployed individuals tend to have a higher Debt-to-Income Ratio and missed payments.

3. Payment History Impact

* Customers with consistent missed payments over six months show a strong correlation with delinquency.
* Late payments vs. missed payments: Missed payments (coded as 2) have a higher impact on delinquency risk than late payments (1).
* Payment behavior trends can serve as early warning indicators for financial stress.

4. Account Tenure & Delinquency

* New accounts (<3 years tenure) exhibit higher delinquency rates, possibly due to lower financial stability.
* Customers with longer account tenure (>10 years) tend to maintain better payment history.

5. Credit Utilization & Risk Indicators

* Higher credit utilization (>0.7) is a strong predictor of delinquency.
* Customers exceeding 100% credit utilization initially showed anomalies, but were corrected during cleaning.

# 5. AI & GenAI Usage

Generative AI tools assisted in summarizing patterns, imputing missing data, and detecting risk factors.

Example AI Prompts Used:

* "Summarize key patterns in the dataset and identify anomalies."
* "Suggest an imputation strategy for missing income values based on industry best practices."
* "Detect correlations between delinquency risk and financial attributes."

# 6. Conclusion & Next Steps

Summary of Key Findings:

* Top delinquency predictors include missed payments, high debt-to-income ratios, and low credit scores.
* Younger accounts (<3 years tenure) show higher delinquency risk.
* Unemployed individuals and customers with high credit utilization are more likely to default.

Recommended Next Steps:

1. Refine predictive models by selecting the most relevant delinquency risk factors.
2. Optimize intervention strategies for at-risk customers using AI-driven insights.
3. Develop ethical AI-based policies ensuring fairness and transparency in delinquency risk assessment.

The dataset is now fully prepared for predictive modeling, allowing advanced delinquency risk assessments.