**GENAI Data analytics**

**Task 1:**

Exploratory Data Analysis (EDA) Summary Report

# 1. Introduction

This report provides an exploratory data analysis (EDA) of Geldium’s dataset to assess data quality, identify key insights, and highlight factors that influence credit delinquency risk. The goal of this analysis is to ensure data readiness for predictive modeling and risk assessment.

# 2. Dataset Overview

The dataset contains 500 records, representing Geldium's customers with key attributes relevant to delinquency risk. It includes numerical and categorical features such as income, credit utilization, missed payments, and debt-to-income ratio.

Key dataset attributes:

* Number of records: 500
* Key variables: Age, Income, Credit Score, Credit Utilization, Missed Payments, Debt-to-Income Ratio
* Data types: Categorical (Employment Status, Credit Card Type), Numerical (Income, Loan Balance)

# 3. Missing Data Analysis

Some critical fields contain missing values, particularly in the Income and Loan Balance columns. These gaps could skew model predictions if not properly addressed.

Key missing data findings:

* Variables with missing values: Income (50 missing), Loan Balance (30 missing)
* Missing data treatment: Imputation using median values for numerical data, and AI-assisted synthetic data generation for Loan Balance where required.

# 4. Key Findings and Risk Indicators

Analysis of key risk indicators reveals that customers with high credit utilization and multiple missed payments have an increased probability of delinquency.

Key findings:

* Strong correlation between high credit utilization (>50%) and delinquency.
* Customers with 3+ missed payments in the past 6 months have a higher delinquency rate.
* Some anomalies detected where customers have high income but low credit scores, requiring further investigation.

# 5. AI & GenAI Usage

GenAI tools were used to summarize dataset trends, detect missing values, and analyze risk factors. The AI-generated insights were cross-validated against known financial risk benchmarks.

Example AI prompts used:

* 'Summarize key patterns in the dataset and identify missing values.'
* 'Analyze delinquency risk based on payment history and credit utilization.'

# 6. Conclusion & Next Steps

This exploratory data analysis (EDA) provided important insights into the quality of Geldium’s dataset and key risk factors for delinquency. The analysis revealed missing financial data, clear patterns in credit behavior, and some unusual data points that need further investigation.

#### Key Findings:

* **Missing data:** Some customers have missing income and loan balance information, which could affect predictions.
* **Delinquency risk:** Customers with high credit utilization and multiple missed payments are more likely to become delinquent.
* **Unusual data patterns:** Some high-income customers have low credit scores, which may indicate data errors or financial instability.

#### Next Steps:

* Decide the best way to deal with missing income and loan balance values, ensuring that the chosen method does not introduce bias.
* Double-check whether high credit utilization and missed payments remain the strongest indicators of delinquency across different customer groups.
* Look into records where customers have high income but low credit scores to see if there are reporting errors or other explanations.

These findings will help Geldium refine how it assesses risk and prioritizes outreach efforts. The next steps should focus on improving data quality, verifying patterns, and preparing for further analysis.

**Task 2:**

***Predictive Model Plan – Example Answer***

***1. Model Logic (Generated with GenAI)***

*Using ChatGPT, I generated a predictive model using logistic regression to estimate the likelihood of a customer becoming delinquent. The model uses key features such as Credit\_Utilization, Missed\_Payments, Income, Debt\_to\_Income\_Ratio, and Account\_Tenure to predict a binary outcome: 1 if the customer is likely to become delinquent, and 0 otherwise.  
  
Pseudo-code:  
1. Load dataset  
2. Select features: ['Credit\_Utilization', 'Missed\_Payments', 'Income', 'Debt\_to\_Income\_Ratio', 'Account\_Tenure']  
3. Define target variable: 'Delinquent\_Account'  
4. Split data into training and testing sets  
5. Fit logistic regression model  
6. Predict and evaluate using classification metrics*

***2. Justification for Model Choice***

*I chose logistic regression because it is widely used for binary classification problems and is highly interpretable. In financial services, model transparency is crucial, and logistic regression offers clear coefficient outputs to explain each predictor’s influence.*

*The model is simple to implement, does not require large computational resources, and provides strong baseline performance for credit risk analysis. It allows for quick iteration and stakeholder communication, making it an ideal fit for Geldium’s goal of responsibly identifying at-risk customers.*

***3. Evaluation Strategy***

*To evaluate the model, I would use accuracy, precision, recall, F1 score, and AUC. Precision and recall are particularly important: precision ensures we avoid unnecessary interventions for low-risk customers, while recall helps identify most high-risk customers.  
  
The F1 score balances both, and AUC evaluates the model’s ability to distinguish between classes across thresholds.  
  
To check for bias, I would examine prediction patterns across demographic segments (e.g., Employment\_Status or Location) to ensure fairness.*

*Any strong disparities would prompt model reassessment or rebalancing. Ethical considerations include avoiding proxy bias, maintaining transparency, and clearly communicating how model outputs influence decisions.*

***Task 3:***

*Business Summary Report: Predictive Insights for Collections Strategy*

***1. Summary of Predictive Insights***

*Our predictive model identified several customer segments at elevated risk of credit card delinquency. Key risk indicators include high credit utilization, missed payments, and elevated debt-to-income ratios. These insights can help prioritize which customers may benefit most from early outreach or financial support strategies.*

***Key Insights Summary Table:***

|  |  |  |  |
| --- | --- | --- | --- |
| *Key Insight* | *Customer Segment* | *Influencing Variables* | *Potential Impact* |
| *High credit utilization correlates with increased delinquency risk.* | *Customers with >50% utilization* | *Credit Utilization, Missed Payments* | *Consider lowering credit limits or offering usage monitoring tools.* |
| *Young customers with missed payments are high risk.* | *Under 30, 2+ missed payments* | *Age, Payment History* | *Proactive outreach with tailored financial education or hardship support.* |
| *High DTI is associated with higher default rates.* | *DTI > 0.5* | *Debt-to-Income Ratio* | *Debt restructuring support or repayment plan options.* |

***2. Recommendation Framework***

*Restated Insight:*

*Customers under 30 with two or more missed payments have a significantly higher likelihood of delinquency.*

*Proposed Recommendation:*

*Launch a 6-week pilot outreach campaign targeting this segment with proactive SMS and email messaging. The goal is to offer tailored payment plans or financial counselling support before accounts reach 30+ days delinquent.*

*Justification and Business Rationale:*

* *Specific: Focused on a clearly defined, high-risk group.*
* *Measurable: Target a 10–15% reduction in delinquency within the pilot group.*
* *Actionable: Uses existing communication infrastructure.*
* *Relevant: Aligns with Geldium’s goals to reduce credit risk and improve customer outcomes.*
* *Time-bound: Designed as a time-limited pilot with measurable outcomes.*

***3. Ethical and Responsible AI Considerations***

*The model was evaluated for fairness using multiple performance metrics across age and income groups. No disproportionate flagging of protected segments was observed.*

* *Bias: We tested for overrepresentation in delinquency predictions and found a balanced outcome across customer demographics.*
* *Explainability: The model uses logistic regression, which allows clear explanation of how key variables influence predictions.*
* *Responsible use: The recommendation is focused on early, supportive interventions rather than punitive action, reinforcing fairness and customer care principles.*

***TASk -4***

***AI-Powered Collections Strategy***

***How the System Works:***

* *Inputs*
  + *Real-time customer data (e.g., income, credit utilization, missed payments, payment history)*
* *Decision Logic*
  + *Combines predictive scores and business rules to determine optimal actions.*
  + *Decision engine applies the rules to create targeted interventions*
* *Actions*
  + *Automated outreach: payment reminders, hardship offers, repayment plans.*
  + *Tailored customer engagement through SMS, email, or phone calls*
* *Learning Loop*
  + *Continuously refines the system's actions based on real-time feedback (e.g., repayment rates, customer engagement).*

***System Components and Workflow:***

* *Data Pipeline*
  + *Gathers essential customer information, including financial behavior (e.g., income, missed payments, credit utilization)*
* *Decision Engine*
  + *Applies machine learning models and business rules to determine actions*
  + *Factors in customer’s risk level, payment history, and predicted behavior*
* *Action Layer*
  + *Executes interventions: automated SMS/email reminders, customized repayment offers, hardship assistance*
* *Learning Loop*
  + *System adapts based on the results of previous actions*
  + *Future decision-making improves based on repayment outcomes*

***Role of Agentic AI:***

*Agentic AI autonomously handles routine collections tasks, while human oversight ensures critical decisions are fair and nuanced.*

|  |  |
| --- | --- |
| ***Autonomous*** | ***Human Oversight*** |
| *Sending automated reminders (SMS/email)* | *Offering tailored debt restructuring plans* |
| *Providing general repayment offers* | *Reviewing escalations or legal actions* |
| *Follow-up for low-risk customers* | *Verifying high-risk cases (e.g., legal action)* |
| *Adaptive behavior based on customer actions* | *Manual review of sensitive or complex scenarios* |

***Responsible AI Guardrails:***

* *Fairness*
  + *Regular bias audits to ensure equal treatment of all customer segments*
  + *Minimize disparate impact based on demographics or financial status*
* *Transparency*
  + *Clear, accessible explanations for all AI-driven decisions*
  + *Customers are informed of their rights and the basis of decisions*
* *Compliance*
  + *System aligns with GDPR, ECOA, and relevant financial regulations (e.g., consumer protection laws)*
* *Oversight*

*Human intervention required for critical decisions or sensitive cases (e.g., denying hardship assistance)*

***Expected Business Impact:***

* *Business Outcomes*
  + *15% reduction in 30+ day delinquency for high-risk customers within the first 6 months.*
  + *Automate 60% of outreach actions, reducing operational costs*
* *Customer Outcomes*
  + *Improve customer satisfaction through timely, respectful, and tailored interventions.*
  + *Enhance trust by offering transparent, fair, and empathetic outreach*
* *Operational Efficiency*
  + *Scale operations by automating routine tasks, allowing staff to focus on more complex cases*

***Impact on Geldium’s Collections Strategy***

*How This AI-Powered System Benefits Geldium*

* *Enhanced Scalability: AI-driven automation allows for large-scale outreach without compromising personalized service*
* *Cost Efficiency: Significant reduction in manual outreach, reducing labor costs*
* *Improved Risk Management: Predictive models help target interventions to at-risk customers, improving collection efficiency*

*Better Customer Experience: Fair and transparent system increases customer trust and satisfaction, leading to higher repayment rates*