

Forage data analytics with GenAI

AI, Machine Learning, and GenAI – What’s the Difference?

Before diving into exploratory data analysis (EDA), it’s important to understand the broader AI landscape and where **Machine Learning (ML)** and **Generative AI (GenAI)** fit into financial risk modeling. Here is a brief run-down:

What is Artificial Intelligence (AI)?

AI refers to any system that can perform tasks requiring human intelligence, such as recognizing patterns, making decisions, or processing language. It is a broad category that includes both traditional rule-based systems and Machine Learning (ML) models that improve based on data.

What is Machine Learning (ML)?

Machine learning is a subset of AI that learns from historical data to predict future outcomes. In financial services, ML models are used for credit scoring, fraud detection, and risk assessment.




- **Example:** A logistic regression model trained on customer repayment history to predict the likelihood of delinquency.

What is Generative AI (GenAI)?

GenAI is a specialized form of AI that can create new content, such as text, images, or structured code, based on training data. Unlike ML, which is designed for structured predictions, GenAI assists with automation, summarization, and data exploration—making it a valuable tool in EDA.

- **Example:** Using GenAI to generate a summary of credit risk trends instead of manually coding an exploratory analysis.

You won’t be building an ML model from scratch, but you will:

-  Explore patterns in financial data that inform risk assessment.
 -  Use GenAI tools to summarize insights and guide predictive modeling.
 -  Understand how AI-driven predictions influence real-world decisions.
-

Introduction to exploratory data analysis (EDA)

Before you start working on your project, it's important to build a solid foundation in the key topics you'll need to understand. This preparation will ensure you have the tools and knowledge necessary to approach your tasks with confidence and skill.

What is exploratory data analysis?

Exploratory data analysis (EDA) is the **first step in understanding a dataset** before applying any predictive models or making business decisions. EDA helps analysts uncover **patterns, trends, inconsistencies, and missing values** to ensure data quality and reliability. In the context of financial services, EDA plays a crucial role in risk assessment, allowing teams to identify key factors contributing to **credit card delinquency** and build stronger prediction models. Here is why EDA matters in predicting delinquency:

- **Ensures data integrity** – Identifies missing values, duplicates, and inconsistencies before analysis.
- **Highlights patterns and anomalies** – Helps detect trends in customer behavior, such as spending patterns before delinquency.
- **Prevents biased models** – Reduces the risk of unfair treatment by ensuring diverse data representation.
- **Supports better decision-making** – Provides Geldium's Collections and Risk teams with clear insights for proactive customer engagement.

Without a thorough EDA process, predictive models can be built on flawed data, leading to inaccurate insights, poor risk management, and potentially unfair decision-making. Later in this task, you will examine a dataset related to delinquency risk, identify missing or inconsistent data points, and generate initial insights using GenAI tools.

Key steps in conducting EDA

EDA consists of four main steps, which can be enhanced with **GenAI tools**:

1 Understanding the dataset

Before jumping into analysis, take time to familiarize yourself with the dataset. Ask yourself:

- What are the **key variables** (e.g., payment history, income levels, credit utilization)?
- Are there **categorical or numerical** data points?
- Are there **missing or inconsistent values**?

Using GenAI: You can use **AI-powered summarization tools** to quickly scan a dataset and generate an overview. Tools such as ChatGPT or Google Gemini can help interpret column headers, suggest relevant features, and summarize key statistics.

💡 *Example prompt: "Analyze this dataset and provide a summary of key columns, including common patterns and missing values."*

2 Identifying missing values and outliers

Incomplete data can lead to poor predictions. Missing values in credit risk datasets may result from human error, system failures, or unreported financial activity.

Here are some common techniques for handling missing data:

- **Statistical imputation (industry standard):** Replace missing values using well-established techniques such as **mean, median, or regression-based imputation**. *(We've provided a resource with further information on imputation below.)*
- **Understanding missingness patterns:** Before filling in missing values, determine whether the data is **missing completely at random (MCAR)**, **missing at random (MAR)**, or **missing not at random (MNAR)** to avoid introducing bias. This [site](https://library.soton.ac.uk/mcar-mar-mnar) includes a great overview document explaining the missing data types.
- **Removing irrelevant data:** If a feature has excessive missing values and cannot be meaningfully imputed without introducing bias, it may be best to exclude it—but only after assessing its impact on model accuracy and fairness..

Using GenAI (for exploration, not direct imputation):

- AI-powered tools can **automate missing value detection** and **summarize data gaps**, helping analysts assess which variables require attention.
- GenAI can suggest **potential imputation strategies** based on statistical best practices, but **final decisions should be validated with domain expertise**.

<https://library.soton.ac.uk/mcar-mar-mnar>

the above link has info about mcar, mar, mnar

Using GenAI (for exploration, not direct imputation):

- AI-powered tools can **automate missing value detection** and **summarize data gaps**, helping analysts assess which variables require attention.
- GenAI can suggest **potential imputation strategies** based on statistical best practices, but **final decisions should be validated with domain expertise**.
- **Synthetic data generation** may be an option when real data is unavailable, but it should be **validated against real-world distributions** to prevent bias.

💡 *Example prompt: "Identify missing values in this dataset and recommend the best imputation strategy based on industry best practices."*

3 Understanding relationships between variables

EDA also involves examining how different features interact. For example:

- Do customers with high credit utilization rates have a higher risk of delinquency?
- Is there a correlation between income levels and late payments?

Using GenAI:

- AI models can automate correlation analysis, helping identify key risk indicators.
- Instead of manually coding formulas, GenAI can summarize variable relationships in natural language.

💡 *Example prompt: "Analyze the correlation between customer income and delinquency risk, summarizing key findings in simple terms."*

4 Detecting patterns and risk factors

The final step in EDA is to identify patterns that could impact delinquency prediction. These might include:

- Customers who miss one payment often miss multiple.
- Younger customers or those with recently opened accounts may have different risk profiles.

Using GenAI:

- AI models can highlight trends in the data, making it easier to understand how different features contribute to delinquency.
- AI-assisted insights can be refined by asking follow-up questions, ensuring that the analysis remains relevant to Geldium's objectives.

💡 *Example prompt: "Analyze trends in late payments and identify the top 3 risk factors associated with delinquency."*

UNDERSTANDING IMPUTATION: A CHEAT SHEET

WHAT IS IMPUTATION?

Imputation is the process of filling in missing data in a dataset using reasonable or estimated values. Instead of deleting rows with missing values, imputation helps retain valuable information by replacing the missing entries with typical or representative data.

WHAT ARE MEAN, MEDIAN, AND MODE?

MEAN (AVERAGE)

The mean is calculated by adding all the values in a dataset and dividing by the number of values.

Example: For values \$2,000, \$3,000, \$4,000, \$5,000, \$6,000:

Mean = $(2000 + 3000 + 4000 + 5000 + 6000) \div 5 = \$4,000$

MEDIAN (MIDDLE VALUE)

The median is the middle number when the data is ordered from smallest to largest.

Example: For values \$2,000, \$3,000, \$4,000, \$5,000, \$6,000:

Median = \$4,000 (the third number in the sorted list)

MODE (MOST FREQUENT)

The mode is the value that appears most frequently in a dataset.

Example: For values \$2,000, \$2,000, \$3,000, \$4,000, \$5,000:

Mode = \$2,000 (it appears twice)

HOW IS IMPUTATION USED?

When data is missing, analysts use the mean, median, or mode to fill in those blanks depending on the nature of the data:

- Use the MEAN if the data is well-balanced without extreme values.
- Use the MEDIAN if the data contains outliers that could skew the average.
- Use the MODE if a single value is very common and likely represents others.

Techniques to handle missing values and ensure data quality

Now that you understand how EDA helps assess dataset completeness, the next step is addressing missing values to ensure reliable insights. In financial risk analytics, incomplete or inconsistent data can skew predictions and lead to incorrect risk assessments. When handling missing data, it is critical to first analyze **the reason for missingness**—whether it is random, systematic, or indicative of underlying biases. While GenAI models can assist in detecting patterns and suggesting imputation strategies, statistical techniques (e.g., mean, median, or regression-based imputation) remain the industry standard due to their transparency and reproducibility. Blindly applying GenAI-generated imputation without understanding data context can introduce significant bias.

What are the common causes of missing data?

Missing data can occur due to:

- **Random errors** (e.g., a system glitch fails to record a payment).
- **Skewed data collection** (e.g., high-income customers are less likely to disclose salary details).
- **Customer behavior** (e.g., financially distressed individuals may avoid reporting debt).

Understanding the cause helps determine the best approach for handling missing values.

How to handle missing values:

1. **Deleting missing data** – If only a **small percentage** of data is missing, removing incomplete entries may be the best approach. However, this can reduce the sample size.
2. **Imputation (replacing missing values)** –
 - **Mean, median, or mode imputation** fills gaps with typical values.
 - **Forward or backward filling** uses existing data trends to estimate missing entries.
3. **AI-Assisted Imputation** –
 - GenAI models can help **detect patterns** and suggest imputation strategies and statistical techniques (e.g., mean, median, or regression-based imputation).
 - AI tools can also suggest **synthetic data** where needed, provided **data privacy is maintained**.

Beyond missing values, check for duplicates, inconsistent formatting, and logical errors (e.g., high credit scores with multiple missed payments). By maintaining data integrity, you ensure that predictive models generate fair and accurate delinquency assessments.

Example given :

Understanding customer risk factors for delinquency

Now that you've learned how to ensure data quality and handle missing values, the next step is identifying key customer risk factors that influence credit card delinquency. Understanding these factors will help refine predictive models and improve intervention strategies.

Why risk factors matter in financial decision-making:

Lenders assess various financial and behavioral indicators to determine whether a customer is likely to miss payments. By analyzing these factors, Geldium's Collections team can proactively identify customers who may need early intervention, reducing financial losses and improving repayment outcomes.

Key risk factors for delinquency

- **Payment history** – Customers with a history of late or missed payments are more likely to default.
- **Credit utilization rate** – High usage of available credit can indicate financial stress and potential repayment issues.
- **Debt-to-income (DTI) ratio** – A high DTI suggests a customer may struggle to manage their financial obligations.
- **Recent credit activity** – A sudden increase in new credit accounts or loan applications may signal financial instability.
- **Employment and income stability** – Frequent job changes or inconsistent income can contribute to a higher risk of missed payments.
- **Demographic trends** – While AI models must avoid bias, certain patterns (e.g., younger customers with limited credit history) may require additional analysis.

AI-driven models analyze multiple risk factors simultaneously, detecting patterns that may not be immediately obvious. However, ensuring that these models remain **fair and explainable** is crucial to avoiding biases in decision-making.

How to leverage synthetic data generation to enhance datasets

In financial services, incomplete or inconsistent data can make it difficult to build reliable predictive models. If key information is missing—such as payment history or income details—delinquency risk assessments may become inaccurate. When real-world data is limited, sensitive, or incomplete, **synthetic data generation** can help fill gaps, simulate scenarios, and improve dataset quality while maintaining privacy and compliance.

Synthetic data is **artificially generated data** that mimics real-world data patterns. Instead of using actual customer records, synthetic data is created using [statistical models](#) or [AI-driven techniques](#) to supplement missing values or expand datasets for testing. This ensures that no real customer information is exposed while still preserving the integrity of the analysis.

While synthetic data generation can be useful for filling in gaps and expanding datasets, it should be applied with caution in financial risk modeling. Traditional statistical simulation techniques such as [Monte Carlo simulations](#), [bootstrapping](#), and [probabilistic modeling](#) are often preferred due to their explainability, reproducibility, and ability to align with industry regulations.

GenAI-generated synthetic data should be:

- ✓ **Validated against real-world distributions** to ensure accuracy.
- ✓ **Cross-checked with statistical models** to prevent introducing artificial patterns or biases.
- ✓ **Used as a supplementary tool**, not a primary data source, especially in regulatory environments.

When to use synthetic data in financial services:

- **Enhancing small datasets** – If real customer data is limited, synthetic records can supplement training data for AI models.
- **Filling in missing data** – When real data is incomplete, synthetic data can approximate realistic values based on existing patterns.
- **Testing AI models** – Before deploying AI risk models, synthetic data can be used to simulate different delinquency scenarios.
- **Ensuring privacy compliance** – Financial data must be handled responsibly. Synthetic data allows analysis without exposing personal customer details.

As you analyze Geldium's dataset, you may find missing values that could impact delinquency predictions. If removing or imputing data isn't viable, synthetic data generation can supplement datasets; however, it must be strictly validated to ensure it accurately reflects real-world trends and does not introduce bias. GenAI-assisted synthetic data generation should be used with

<https://www.ibm.com/think/topics/monte-carlo-simulation>

<https://www.datacamp.com/tutorial/bootstrapping>

<https://www.sciencedirect.com/topics/computer-science/probabilistic-modeling>

As you analyze Geldium's dataset, you may find missing values that could impact delinquency predictions. If removing or imputing data isn't viable, synthetic data generation can supplement datasets; however, it must be strictly validated to ensure it accurately reflects real-world trends and does not introduce bias. GenAI-assisted synthetic data generation should be used with caution, as improper constraints can lead to unrealistic outputs that misrepresent risk factors.

💡 *Example GenAI prompt: "Generate synthetic payment history data for customers with missing records while ensuring that distributions align with historical patterns observed in the dataset (e.g., standard deviations, typical payment behaviors)."*

As you explore Geldium's dataset, consider whether synthetic data could be useful for filling gaps or testing delinquency risk models. Your approach should balance realism, fairness, and privacy to ensure high-quality AI-driven insights.

Ethical considerations:

While synthetic data can enhance dataset completeness, it must be used carefully to avoid:

- **Introducing bias** – Ensure synthetic records reflect realistic patterns.
- **Misrepresenting risk factors** – Avoid generating overly optimistic or pessimistic data.
- **Compromising compliance** – Validate that synthetic data aligns with industry regulations.

Step 1:

Based on my analysis of the delinquency prediction dataset, here are the key observations:

The dataset contains 500 customer records with a 16% delinquency rate. Data quality issues include missing values in Income (39 records), Credit_Score (2 records), and Loan_Balance (29 records), which may need imputation or removal. No duplicate records were found, indicating good data integrity.

Key patterns in delinquent accounts show they have slightly lower average credit scores (591 vs 575) and higher debt-to-income ratios (0.306 vs 0.297) compared to non-delinquent customers. Interestingly, delinquent customers have slightly fewer missed payments on average (2.85 vs 2.99), which may indicate that missed payments alone aren't a strong predictor of delinquency in this dataset.

Notable risk indicators include payment history patterns, where delinquent customers show different distributions in "Late" and "Missed" payment statuses across the six-month period. Certain segments appear higher risk: Business and Student credit card holders, Self-employed and Unemployed individuals, and customers in Los Angeles, Houston, and

Phoenix. The early indicators of delinquency risk seem to be more correlated with credit score, debt-to-income ratio, and employment/credit card type rather than just payment history.

Step 2:

Missing Data Issue	Handling Method	Justification
Income (39 missing values, 7.8%)	KNN Imputation	Preserves relationships with other financial variables while filling moderate missing data
Credit_Score (2 missing values, 0.4%)	Median Imputation	Minimal missing data makes simple imputation sufficient and robust
Loan_Balance (29 missing values, 5.8%)	KNN Imputation	Captures relationships with other variables for more accurate imputation than mean/median

Critical Feature Gaps:

1. **Income:** 39 missing values (7.8% of data)
2. **Credit_Score:** 2 missing values (0.4% of data)
3. **Loan_Balance:** 29 missing values (5.8% of data)
4. **Credit_Utilization:** 0 missing values
5. **Missed_Payments:** 0 missing values

Payment History: No missing values found in any of the Month_1 through Month_6 columns.

Recommended Treatment Approaches:

1. **Income (7.8% missing):**
 - a. **Approach:** Impute with KNN (K-Nearest Neighbors)
 - b. **Rationale:** While the percentage is moderate, Income has relationships with other variables that KNN can leverage for more accurate imputation than simple mean/median.
2. **Credit_Score (0.4% missing):**
 - a. **Approach:** Impute with median

- b. **Rationale:** Very few missing values make complex imputation unnecessary; median is robust and simple.
- 3. **Loan_Balance (5.8% missing):**
 - a. **Approach:** Impute with KNN
 - b. **Rationale:** Moderate missing percentage with some relationships to other financial variables that KNN can capture.
- 4. **Credit_Utilization & Missed_Payments:**
 - a. **Approach:** No treatment needed
 - b. **Rationale:** No missing values found.
- 5. **Payment History (Month_1 to Month_6):**
 - a. **Approach:** No treatment needed
 - b. **Rationale:** Complete data with no missing values.

No columns require removal as none exceed 50% missing data. **No synthetic data generation is needed** since the missing data percentages are manageable with imputation techniques.

Step 3:

Key Risk Indicators and Insights from Delinquency Analysis

- 1. Obvious Patterns:
- 2. • Higher missed payments strongly correlate with delinquency (3+ missed payments show significantly higher delinquency rates)
- 3. • High credit utilization (60%+) customers have elevated delinquency risk
- 4. • Customers with low credit scores (<500) show higher delinquency rates than those with excellent scores
- 5. • High debt-to-income ratio (>40%) customers are more likely to become delinquent
- 6. Unexpected Findings:
 - Credit utilization has very weak correlation with missed payments (0.02), contrary to expectations
 - Higher credit score customers surprisingly have similar delinquency rates to lower score customers
 - High-income customers don't show significantly lower delinquency rates than lower-income customers
- 7. Key Variables for Prediction:
 - Missed_Payments (strongest predictor)

- Debt_to_Income_Ratio
- Credit_Score
- Account_Tenure (negative correlation - longer accounts have lower delinquency)

8. High-Risk Segments:

- Business credit card holders (21% delinquency rate)
- Unemployed individuals (19% delinquency rate)
- Los Angeles residents (20% delinquency rate)

9. Recommendations for Further Investigation:

- Investigate why high credit score customers still show high delinquency rates
- Examine the weak relationship between credit utilization and missed payments
- Analyze why income level doesn't significantly impact delinquency risk

Action:

High-Risk Indicator	Explanation	Impact on Delinquency Prediction
Missed Payments (3+)	Customers with 3 or more missed payments show significantly higher delinquency rates, making this the strongest predictor.	Should be weighted heavily in any predictive model as it directly reflects payment behavior.
High Credit Utilization (>60%)	Customers using over 60% of their available credit have elevated delinquency risk, indicating potential financial stress.	Important for early identification of at-risk customers before they miss payments.
Low Credit Score (<500)	Customers with poor credit scores show higher delinquency rates, reflecting historical credit challenges.	Useful for segmenting customers but may be less predictive than payment behavior metrics.
High Debt-to-Income Ratio (>40%)	Customers with high debt relative to income are more likely to become delinquent due to financial strain.	Essential for assessing a customer's capacity to manage additional debt.
Business Credit Card Holders	Business card users show a 21% delinquency rate, the highest among card types, possibly due to business financial volatility.	Important demographic segment that may require different risk models than personal cards.

Unemployed Status	Unemployed individuals have a 19% delinquency rate, likely due to lack of steady income.	Critical employment variable for assessing payment capacity and default risk.
New Accounts (<5 years)	Newer accounts show higher delinquency rates, possibly due to less established credit history.	Account tenure is negatively correlated with delinquency, suggesting loyalty programs could reduce risk.
Los Angeles Residents	Customers in Los Angeles show elevated delinquency rates (20%), potentially due to regional economic factors.	Geographic risk factors may indicate need for location-based risk adjustments.
Combined Risk Factors	Customers with both high credit utilization AND high missed payments show extremely high delinquency rates.	Compound risk factors are more predictive than single indicators and should trigger enhanced monitoring.

Report is made in other document