**Predictive Model Plan**

# 1. Model Logic (Generated with GenAI)

The model is designed to **predict whether a customer is likely to become delinquent on their payments**, based on historical financial, demographic, and behavioral data. The prediction is binary:

* 0 = Not Delinquent
* 1 = Delinquent

**Step-by-Step Logic:**

- Step 1: Load and clean data

- Import dataset into Python (e.g., using pandas)

- Handle missing values in Income, Loan\_Balance, etc.

- Convert monthly payment history (Month\_1 to Month\_6) to numeric scores

(e.g., "On-time" = 0, "Late" = 1, "Missed" = 2)

- Step 2: Feature Engineering

- Encode categorical variables (Employment\_Status, Location, etc.)

- Normalize numerical features (e.g., Credit\_Score, Income, Loan\_Balance)

- Step 3: Model Setup

- Define target: Delinquent\_Account

- Split dataset into training and testing sets (e.g., 70/30)

- Step 4: Model Training

- Train Logistic Regression or Decision Tree Classifier

- Step 5: Evaluation

- Evaluate model using Accuracy, F1 Score, Precision, Recall, AUC-ROC

- Step 6: Interpretation

- Use feature importance or SHAP to explain top risk factors

- Step 7: Deployment (if needed)

- Save model and serve via Flask or another interface for real-time prediction

**What the Model Does:**

This predictive model uses historical payment behavior, credit utilization, income levels, and other customer features to **identify those at high risk of missing future payments**. It helps financial institutions **take early action**, such as customer outreach or restructuring, to reduce delinquency rates.

# 2. Justification for Model Choice

**Model Selected**: Logistic Regression

**Reasons for Selection:**

**Binary Classification Suitability**: Logistic Regression is ideal for predicting binary outcomes such as Delinquent (1) or Not Delinquent (0), aligning directly with the business problem.

**Transparency & Interpretability**: It provides clear insight into the influence of each variable (via coefficients), which is crucial for financial stakeholders and compliance teams to understand the basis of predictions.

**Ease of Implementation**: It is straightforward to implement, requires minimal computational power, and is easy to maintain or update as new customer data becomes available.

**Proven Accuracy**: While more complex models may offer slight performance improvements, logistic regression performs competitively on structured financial data with well-engineered features.

**Business Relevance**: In financial services, decision-making often requires explainable models. Logistic Regression enables Geldium’s risk analysts to justify credit decisions and comply with audit or regulatory requirements.

**Scalability**: The model can be scaled or integrated with dashboards, and supports real-time decision systems, helping Geldium reduce delinquency risks proactively.

# 3. Evaluation Strategy

**Evaluation Metrics:**

* **Accuracy**: Measures the overall percentage of correct predictions. Useful for getting a general idea of model performance but not sufficient alone.
* **Precision**: Indicates how many predicted delinquents are actually delinquent. High precision reduces false alarms.
* **Recall (Sensitivity)**: Captures how many actual delinquents the model correctly identified. High recall ensures at-risk customers are not missed.
* **F1 Score**: Harmonic mean of precision and recall. Useful when both false positives and false negatives matter.
* **AUC-ROC**: Shows how well the model distinguishes between delinquent and non-delinquent customers across different thresholds.

**Metric Interpretation:**

* A high **recall** is critical in this context, as it ensures the business can intervene early with potentially delinquent customers.
* **Precision** helps ensure that interventions are cost-effective by minimizing false positives.
* The **F1 score** balances both, giving a holistic measure in imbalanced data scenarios.

**Bias Detection and Mitigation:**

* Monitor model performance across subgroups (e.g., gender, employment status, location) to check for fairness and bias.
* Use techniques like **stratified sampling**, **re-weighting**, or **fairness-aware modeling** to mitigate bias.
* Ensure features used do not directly or indirectly encode sensitive attributes unless legally justified.

**Ethical Considerations:**

* Avoid over-reliance on automated predictions—ensure human review for edge cases.
* Maintain **data privacy** and **customer consent**, especially when using financial and behavioral data.
* Be cautious in labeling customers as "risky" to prevent discrimination or unfair treatment.