**Predictive Model Plan**

# 1. Model Logic (Generated with GenAI)

Use a GenAI tool (e.g., ChatGPT, Gemini) to generate the logic or structure of your predictive model.  
- You may include pseudo-code, a step-by-step process, or a simplified code snippet.  
- Briefly explain what the model is designed to do.

Paste your GenAI-generated output below or describe the logic in your own words:

**START**

**1. LOAD the dataset from a CSV file**

**2. DROP the 'Customer\_ID' column (not useful for prediction)**

**3. DEFINE 'Delinquent\_Account' as the target variable (y)**

**DEFINE all other columns as features (X)**

**4. HANDLE MISSING VALUES:**

**FOR each numerical column in X:**

**- Fill missing values with the median of that column**

**5. ENCODE CATEGORICAL VARIABLES:**

**FOR each categorical column in X:**

**- Convert categories into numerical labels using label encoding**

**6. SCALE NUMERICAL FEATURES:**

**- Standardize all numerical features to have mean = 0 and standard deviation = 1**

**7. SPLIT the dataset into training and test sets (e.g., 80% train, 20% test)**

**8. INITIALIZE a Random Forest Classifier**

**9. TRAIN the model using the training set (X\_train, y\_train)**

**10. PREDICT outcomes using the test set (X\_test)**

**11. EVALUATE the model:**

**- Calculate Accuracy**

**- Generate a Classification Report (Precision, Recall, F1-score)**

**12. OUTPUT the evaluation results**

**END**

# 2. Justification for Model Choice

Explain why you selected this specific model type (e.g., logistic regression, decision tree, neural network). Consider:  
- Accuracy  
- Transparency  
- Ease of use or implementation  
- Relevance for financial prediction  
- Suitability for Geldium’s business needs

**Model Selection Justification: Random Forest Classifier**

The **Random Forest** model was chosen for predicting loan delinquency due to the following reasons:

* **Accuracy:** It delivers high predictive performance by combining multiple decision trees, reducing overfitting and capturing complex patterns effectively.
* **Transparency:** While not as interpretable as logistic regression, Random Forest still provides feature importance scores, which help explain key drivers of delinquency.
* **Ease of Use:** It’s easy to implement using libraries like scikit-learn, requires minimal preprocessing, and handles both numerical and categorical data efficiently.
* **Financial Relevance:** Random Forest models are widely used in credit risk analysis as they capture non-linear interactions and support decision-making based on diverse financial indicators.
* **Suitability for Geldium’s Needs:** Given Geldium’s goal to reduce risk and maintain explainability, Random Forest offers a balanced trade-off between predictive accuracy and interpretability, making it ideal for practical deployment.

# 3. Evaluation Strategy

Outline how you would evaluate your model’s performance. Include:  
- Which metrics you would use (e.g., accuracy, precision, recall, F1 score, AUC)  
- How you would interpret those metrics  
- Any plans to detect or reduce bias in your model  
- Ethical considerations in making predictions about customer financial behavior

**✅ 1. Evaluation Metrics**

* **Accuracy:** Measures overall correctness of predictions. Useful for general overview, but may be misleading in imbalanced datasets.
* **Precision:** Indicates how many predicted delinquents are truly delinquent. Important to minimize false alarms.
* **Recall (Sensitivity):** Shows how many actual delinquents were correctly identified. Critical in minimizing missed high-risk customers.
* **F1 Score:** Harmonic mean of precision and recall; provides a balanced metric when there is a trade-off between them.
* **AUC-ROC (Area Under Curve - Receiver Operating Characteristic):** Evaluates the model’s ability to distinguish between classes across thresholds. A high AUC indicates strong class separation.

### 🔍 2. Interpretation of Metrics

* High **recall** means the model catches most delinquent customers (fewer defaults slip through).
* High **precision** avoids wrongly flagging good customers (prevents unnecessary actions).
* Balanced **F1 score** ensures neither precision nor recall is overly sacrificed.
* **AUC > 0.85** is generally considered good for financial models.

### ⚖️ 3. Bias Detection and Reduction

* **Stratified sampling** during training ensures class balance is preserved.
* Analyze model performance across subgroups (e.g., gender, age groups, employment status).
* Apply techniques like **SMOTE** for class imbalance and **fairness-aware algorithms** if disparities are detected.
* Regular retraining with updated data reduces long-term model drift or bias.

### 🧠 4. Ethical Considerations

* **Fairness:** Ensure the model does not unintentionally discriminate against vulnerable populations (e.g., unemployed, low-income).
* **Transparency:** Use interpretable models or provide clear explanations for risk scores to customers and regulators.
* **Accountability:** Avoid fully automated decisions—predictions should support, not replace, human judgment in financial actions.
* **Privacy:** Strictly handle customer data under privacy laws (e.g., GDPR), ensuring ethical data use.