Predictive Model Plan – Delinquency Prediction

1. Model Logic (Generated with GenAI)

Model Objective:
The goal is to predict whether a customer will become delinquent on their financial obligations based on features such as income, credit score, past payment behavior, and account tenure.

Model Logic (Pseudocode):
1. Load and preprocess the dataset.
2. Handle missing values:
  - Use median imputation for Income, Credit Score, and Loan Balance.
3. Encode categorical variables:
  - Apply one-hot encoding to features like Employment_Status, Credit_Card_Type, and Location.
4. Convert monthly payment behavior into numerical features (e.g., count of 'Late' and 'Missed').
5. Split the data into training and testing sets (e.g., 80/20 split).
6. Train a Logistic Regression model (or alternative like Random Forest for comparison).
7. Evaluate using metrics such as accuracy, precision, recall, and AUC.
8. Optimize using grid search and cross-validation.
9. Deploy the final model for inference on new customer data.

2. Justification for Model Choice

We selected Logistic Regression as the initial model due to its interpretability and suitability for binary classification problems such as delinquency prediction. This model is especially beneficial for financial institutions that require explainable models.

- Accuracy: Logistic Regression offers high accuracy on structured data, especially with engineered features.
- Transparency: Provides interpretable coefficients for regulatory compliance.
- Ease of Use: Easy to deploy, maintain, and scale.
- Relevance: Produces probability scores that help in risk segmentation.
- Business Suitability: Aligns well with Geldium's need to assess risk and prioritize customer management strategies.

For benchmarking, we also consider tree-based models like Random Forest or Gradient Boosting Machines (e.g., XGBoost) for potentially improved performance.

3. Evaluation Strategy

Performance Metrics:
- Accuracy: To check overall correctness of predictions.
- Precision: To avoid falsely labeling non-delinquent customers as high-risk.
- Recall: To capture most delinquent customers and minimize risk.
- F1 Score: To balance between precision and recall.
- AUC-ROC: To evaluate the model's ability to distinguish between classes.

Bias & Ethics:
- Sensitive attributes like race or gender are not included, but categorical features such as Employment_Status and Location will be monitored for fairness.
- Bias will be tested using fairness metrics and model outputs will be reviewed by compliance teams.
- We aim to build responsible AI that does not unjustly impact underrepresented groups.

Validation:
- k-fold Cross Validation will be used to ensure model generalizability.
- Confusion matrix and classification report will guide error analysis.
- Monthly performance data (Months 1-6) will be aggregated into numerical summaries to improve signal detection.