Predictive Model Plan

1. Model Logic (Generated with GenAI)

For predicting customer delinquency, I propose using a **logistic regression** model. This model is ideal for binary classification problems such as delinquency (yes/no). It estimates the probability of a customer becoming delinquent based on input features.

Top 5 input features selected:

- **Income:** Lower income may increase delinquency risk.
- **Credit Utilization:** High utilization indicates potential financial stress.
- **Missed Payments:** Strong historical predictor of future delinquency.
- **Credit Score:** Lower scores are often associated with higher risk.
- **Loan Balance:** High balances may indicate difficulty in managing debt.

Pseudocode Outline:

1. Load and preprocess the data (handle missing values, normalize numeric fields).
2. Select key features: income, credit utilization, credit score, missed payments, loan balance.
3. Split the dataset into training and testing sets.
4. Fit a logistic regression model on the training set.
5. Predict delinquency outcomes on the test set.
6. Evaluate model performance using metrics like accuracy, F1 score, AUC.
7. Refine the model as needed based on results.

2. Justification for Model Choice

For the task of predicting customer delinquency, logistic regression has been selected as the most appropriate model due to its strong alignment with financial industry standards and business needs. Logistic regression is specifically designed for binary classification problems, such as predicting whether a customer will be delinquent or not. It provides interpretable results by quantifying the effect of each input variable on the outcome, which supports transparency and regulatory compliance—critical for financial institutions like Geldium. While models like decision trees offer interpretability too, logistic regression is less prone to overfitting and is more efficient to train and deploy at scale. Additionally, logistic regression handles structured, tabular financial data well and produces probability scores that help the Collections team prioritize intervention efforts. Overall, it offers a reliable balance between accuracy, explainability, and operational simplicity—making it ideal for responsible, data-driven decision-making in credit risk assessment.

3. Evaluation Strategy

To ensure the reliability and fairness of the logistic regression model for predicting delinquency, a multi-metric evaluation strategy will be adopted. The **accuracy** metric will provide an overall view of how often the model predicts correctly, but this alone is insufficient in imbalanced datasets. Therefore, **precision** and **recall** will be used to evaluate how well the model identifies true delinquent cases while minimizing false positives. The **F1 score**, a balance between precision and recall, will serve as a holistic measure when both false positives and false negatives are critical. Additionally, the **AUC-ROC curve** will assess the model's ability to distinguish between delinquent and non-delinquent customers, with a higher area under the curve indicating better performance.

To address fairness, the model will be evaluated for **bias across demographic groups** using statistical fairness metrics such as **demographic parity** and **equal opportunity**. If disparities are identified, bias mitigation techniques such as **re-sampling**, **reweighting**, or **removal of proxy variables** will be considered. Finally, regular review of the **confusion matrix** will help diagnose specific error types and guide necessary model adjustments. Together, these steps will help maintain model transparency, ensure regulatory alignment, and support responsible AI usage in financial services.