Credit Delinquency Prediction Plan

Step 1: Model Logic (Conceptual Framework)

Chosen Model Type: For this task, I propose using a Logistic Regression model as the primary approach. Logistic regression is widely used in financial services because it provides clear interpretability and outputs a probability score for delinquency risk.
Alternative Model: A Random Forest (Decision Tree ensemble) could also be considered as a more complex option, offering higher predictive power but reduced interpretability.

Top 5 Input Features:

- Credit Utilization Ratio – Strong indicator of financial stress; higher utilization often correlates with higher delinquency risk.
- Missed Payments (History) – Direct behavioral measure of repayment reliability.
- Debt-to-Income Ratio – Reflects repayment capacity relative to income.
- Credit Score – Captures historical creditworthiness.
- Employment Status – Employment stability often impacts repayment ability.

Workflow Overview:

- Data Ingestion: Customer profile data is collected (income, credit utilization, repayment history, etc.).
- Feature Processing: Normalize ratios (utilization, debt-to-income), encode categorical variables (employment, location), and handle missing values.
- Model Training: Logistic regression is trained on historical delinquency outcomes.
- Prediction: The model outputs a probability score (0–1), which is then thresholded (e.g., >0.5 = high risk).
- Output: Risk score is delivered to business systems for decision-making (loan approval, monitoring, etc.).

Step 2: Model Justification

I selected Logistic Regression because it balances accuracy, interpretability, and regulatory compliance — all essential in financial services. Unlike complex models (e.g., neural networks), logistic regression provides transparent coefficients that regulators, auditors, and risk managers can interpret easily. This makes it suitable for explaining why a customer is classified as high risk, which supports fairness and compliance requirements. While decision trees or random forests could capture non-linear relationships for potentially better accuracy, they sacrifice explainability and may be harder to monitor for bias. Given Geldium's need for trust, transparency, and compliance, logistic regression is the most appropriate first-choice model, with more complex models as secondary options for internal benchmarking.

Step 3: Evaluation Strategy

Key Metrics:

- Accuracy: Overall proportion of correct predictions.

- Precision & Recall: To assess balance between false positives (wrongly labeling customers as risky) and false negatives (missing actual risky customers).
- F1 Score: Harmonic mean of precision and recall — useful when delinquency cases are imbalanced.
- AUC-ROC: Evaluates the model's ability to distinguish between high- and low-risk customers.
- Fairness Checks: Test for disparate impact across groups (e.g., age, gender, location) to ensure compliance and fairness.

Bias Mitigation Techniques:

- Monitor subgroup performance (e.g., false negative rates by demographic group).
- Use reweighting or resampling methods if imbalances are found.
- Regular retraining to reflect changing economic conditions and customer profiles.

Interpretation & Improvement:

- If AUC falls below target (e.g., 0.75), or fairness checks reveal bias, the model will be retrained with adjusted features or thresholds.
- Continuous monitoring will ensure both business performance and ethical standards are met.

Step 4: Final Deliverable (Summary)

- Model Logic: Logistic Regression model, using key features (credit utilization, missed payments, debt-to-income, credit score, employment).
- Justification: Chosen for interpretability, transparency, and compliance in financial decision-making; Random Forest as a secondary benchmark.
- Evaluation Strategy: Multi-metric approach (Accuracy, F1, AUC, fairness checks), with regular monitoring and bias mitigation techniques.