**Predictive Model Plan**

## **1. Model Logic (Generated with GenAI)**

I will use a **Gradient Boosted Trees model (LightGBM/XGBoost)** to predict whether a customer will become delinquent (Delinquent\_Account = 1) based on financial and behavioral features.

The model takes numeric inputs like Credit\_Score, Credit\_Utilization, Debt\_to\_Income\_Ratio, and categorical attributes like Employment\_Status and Credit\_Card\_Type. It also learns from past payment behavior in Month\_1 to Month\_6 to capture trends in missed or late payments.

**Pseudocode (conceptual):**

## 1. Load dataset (customer\_id, features, delinquent flag)

## 2. Handle missing values (median for numeric, mode for categorical)

## 3. Encode categorical columns:

## - Employment\_Status

## - Credit\_Card\_Type

## - Location

## 4. Create behavior-based features:

## - payment\_pattern\_score = count("Missed"/"Late") across Month\_1–Month\_6

## - credit\_risk\_index = Credit\_Utilization \* Debt\_to\_Income\_Ratio

## 5. Split data into train/test sets (80/20) with stratified sampling.

## 6. Train models:

## - Logistic Regression (baseline)

## - Gradient Boosted Trees (primary)

## 7. Evaluate using AUC, F1, Precision, Recall.

## 8. Use SHAP values to explain top risk factors for each prediction.

## 9. Save model and deploy as risk scoring API or batch process.

## **2. Justification for Model Choice**

The **Gradient Boosted Trees model (LightGBM/XGBoost)** is chosen because it performs exceptionally well on structured financial data like this, where numeric and categorical features interact in complex ways (e.g., how Credit\_Utilization and Debt\_to\_Income\_Ratio together indicate repayment risk).

This model automatically handles missing values, nonlinear effects, and feature importance — while SHAP explainability ensures transparency for regulators and analysts.

A **Logistic Regression** model will also be used as a baseline for comparison, offering simplicity and interpretability. Together, they provide a balance between **accuracy**, **fairness**, and **clarity**, aligning with Geldium’s focus on responsible credit assessment.

## **3. Evaluation Strategy**

**Evaluation metrics:**

* **AUC-ROC** and **AUC-PR:** measure how well the model distinguishes between delinquent and non-delinquent customers.
* **Precision, Recall, and F1-score:** ensure a balance between catching risky customers and avoiding false alarms.
* **Confusion Matrix:** shows true/false positives and negatives for threshold decisions.
* **Brier Score and Calibration Curve:** check if predicted probabilities match real delinquency rates.

**Fairness and bias checks:**

* Compare True Positive Rate (TPR) and False Positive Rate (FPR) across Location and Employment\_Status groups.
* Ensure model predictions are not biased toward any demographic or employment type.
* Use SHAP value plots to confirm that sensitive attributes do not drive predictions.

**Bias mitigation techniques:**

* Reweigh samples if one group is underrepresented.
* Add fairness constraints during model training.
* Adjust classification thresholds per subgroup if needed.

**Ethical considerations:**  
Predictions will support (not replace) human decision-making. Geldium should use results for early risk management and fair treatment of all customers, ensuring transparency and compliance with financial regulations.

\