**Predictive Model Plan – Student Template**

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

We propose using **logistic regression** as our primary predictive model to assess customer delinquency risk. Logistic regression is well-suited for this binary classification task because it estimates the probability that a customer will become delinquent (where Delinquent\_Account = 1). The model will compute a weighted sum of input features and apply the sigmoid (logistic) function to generate risk probabilities between 0 and 1.

**Top 5 input features:**

* Credit\_Score: Lower scores indicate higher delinquency risk.
* Missed\_Payments: A strong predictor based on recent payment behavior.
* Credit\_Utilization: High utilization suggests financial stress.
* Debt\_to\_Income\_Ratio: Higher values indicate debt burden relative to income.
* Income: Lower income may reduce repayment capacity.

**Model pipeline:**

1. **Data ingestion & preprocessing**:
   * Load customer data.
   * Impute missing Income values (e.g., using the median).
   * Encode categorical features (Employment\_Status, Credit\_Card\_Type) via one-hot encoding.
   * Scale numeric features for model stability.
2. **Training**:
   * Fit logistic regression using Delinquent\_Account as the target.
   * The model learns coefficients for each feature that map to the log-odds of delinquency.
3. **Prediction**:
   * For new data, compute risk probabilities.
   * Output a score (e.g., 0.72 = 72% risk).
   * Flag customers above a risk threshold (e.g., 0.5) for intervention.

An alternative option considered was a **Gradient Boosted Tree (e.g., XGBoost)**, which could capture nonlinear interactions but would sacrifice interpretability.

**2. Justification for Model Choice**

We selected **logistic regression** because it provides a strong balance between predictive performance and interpretability. In financial services, it is essential that models are transparent so that predictions can be explained to regulators, auditors, and customers. Logistic regression allows us to clearly quantify how each feature contributes to risk. The model is simple to implement, easy to monitor, and fast to deploy. Additionally, logistic regression aligns with Geldium’s need for explainable, auditable, and compliant risk models. While complex models like XGBoost could potentially improve predictive accuracy, they add complexity and reduce transparency, which may conflict with regulatory requirements.

**3. Evaluation Strategy**

We will assess model performance using the following metrics:

* **AUC-ROC**: To measure the model’s ability to discriminate between delinquent and non-delinquent customers.
* **F1 score**: To balance precision and recall, especially if the dataset is imbalanced.
* **Accuracy**: For an overall measure of correct classifications.
* **Precision and recall**: To understand the balance between false positives and false negatives.

For fairness checks:

* We will evaluate model metrics across subgroups (e.g., Employment\_Status, Location) to identify potential bias.
* Apply fairness metrics (e.g., demographic parity or equal opportunity) to ensure no group is unfairly disadvantaged.
* If bias is detected, mitigation strategies such as reweighting or post-processing threshold adjustments will be considered.

Ethically, our model will ensure decisions are explainable and do not unfairly disadvantage any protected groups, supporting responsible lending practices.