**Predictive Model Plan – Student Template**

Use this template to structure your submission. You can copy and paste content from GenAI tools and build around it with your own analysis.

# 1. Model Logic (Generated with GenAI)

Use a GenAI tool (e.g., ChatGPT, Gemini) to generate the logic or structure of your predictive model.  
- You may include pseudo-code, a step-by-step process, or a simplified code snippet.  
- Briefly explain what the model is designed to do.

Paste your GenAI-generated output below or describe the logic in your own words:

The predictive model is designed to forecast whether a customer will become delinquent (Delinquent\_Account = 1) using demographic, financial, and behavioral features.

**Step-by-step process:**

1. **Data Ingestion** – Load dataset with features: Age, Income, Credit Score, Credit Utilization, Missed Payments, Debt-to-Income Ratio, Loan Balance, Employment Status, Account Tenure, Credit Card Type, Location, and Payment History (Months 1–6).
2. **Data Preprocessing** –
   * Handle missing values (e.g., impute Income, Credit Score, Loan Balance).
   * Normalize continuous variables (Income, Credit Utilization, Debt-to-Income Ratio).
   * Encode categorical variables (Employment Status, Credit Card Type, Location, Month\_1–Month\_6).
3. **Feature Selection** – Focus on top predictors:
   * Missed Payments
   * Credit Utilization
   * Debt-to-Income Ratio
   * Income
   * Credit Score
4. **Model Training** –
   * Use **Logistic Regression** to output delinquency probability (0–1).
   * Alternative: **Decision Tree** for transparent segmentation of customer risk.
5. **Prediction Output** – Generate risk score; customers above threshold (e.g., 0.5) flagged as high-risk.
6. **Evaluation** – Assess performance using Accuracy, Precision, Recall, F1-score, AUC-ROC, and fairness checks across demographics.

💡 GenAI Prompt used:  
"Outline a predictive modeling pipeline to forecast credit delinquency, from feature selection to model evaluation."

# 2. Justification for Model Choice

Explain why you selected this specific model type (e.g., logistic regression, decision tree, neural network). Consider:  
- Accuracy  
- Transparency  
- Ease of use or implementation  
- Relevance for financial prediction  
- Suitability for Geldium’s business needs

I selected **Logistic Regression** as the primary model for predicting delinquency because it balances accuracy, transparency, and practical applicability in financial risk prediction. Logistic regression provides clear probability scores that help Geldium classify customers as delinquent or non-delinquent, making it straightforward to interpret and explain. This is especially important in financial services, where **regulatory compliance and transparency** are critical.

Compared to more complex models like neural networks, logistic regression is easier to implement, faster to train, and avoids the “black box” problem. While decision trees are also transparent, they are prone to overfitting, whereas logistic regression generalizes better across new data. Overall, logistic regression offers the best fit for Geldium’s business needs: it is accurate enough for reliable forecasting, interpretable for stakeholder trust, and efficient for deployment in real-world financial decision-making.

# 3. Evaluation Strategy

Outline how you would evaluate your model’s performance. Include:  
- Which metrics you would use (e.g., accuracy, precision, recall, F1 score, AUC)  
- How you would interpret those metrics  
- Any plans to detect or reduce bias in your model  
- Ethical considerations in making predictions about customer financial behavior

To ensure the model is accurate, fair, and reliable, I will evaluate performance using a combination of metrics and fairness checks:

* **Metrics:**
  + **Accuracy** – Measures overall correctness of predictions.
  + **Precision** – Of the customers flagged as delinquent, how many truly are delinquent.
  + **Recall (Sensitivity)** – Of all actual delinquent customers, how many were correctly identified.
  + **F1 Score** – Balances precision and recall, useful when both false positives and false negatives carry costs.
  + **AUC-ROC** – Evaluates how well the model distinguishes between delinquent and non-delinquent customers across thresholds.
  + **Confusion Matrix** – Provides a breakdown of true positives, false positives, true negatives, and false negatives to diagnose errors.
* **Interpretation of Metrics:**
  + High recall ensures that fewer risky customers are missed, minimizing potential financial losses.
  + Precision ensures outreach is focused on true high-risk cases, avoiding unnecessary interventions.
  + A balanced F1 score reflects a trade-off between precision and recall.
  + AUC-ROC close to 1 indicates the model ranks risk effectively.
* **Bias Detection & Mitigation:**
  + Test model outputs across demographic groups (e.g., income levels, employment status, locations).
  + Apply fairness checks such as demographic parity or disparate impact ratio.
  + If bias is detected, rebalance the dataset (e.g., oversample delinquent cases, adjust feature selection to avoid proxy variables).
* **Ethical Considerations:**
  + Predictions should be used to **support proactive assistance** (e.g., offering repayment plans) rather than penalize customers unfairly.
  + Maintain **transparency** in model decisions for regulatory compliance.
  + Ensure **human oversight** before final action is taken on high-risk classifications.

This strategy ensures the model not only achieves high predictive accuracy but also remains fair, transparent, and aligned with responsible financial decision-making.