**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:

We propose two options for forecasting customer delinquency risk:  
  
1. \*\*Simple Model\*\*: Logistic Regression — Ideal for binary classification (delinquent vs. non-delinquent). It offers interpretability and is widely accepted in finance.  
2. \*\*Complex Model\*\*: Random Forest Classifier — An ensemble of decision trees, better at handling non-linear interactions and missing data, with robust performance.  
  
Recommended: Start with Logistic Regression for transparency and regulatory compliance, then scale to Random Forest if better accuracy is required.  
  
\*\*Model Workflow:\*\*  
- Step 1: Data ingestion and cleaning  
- Step 2: Feature selection (top variables: Income, Credit\_Utilization, Missed\_Payments, Credit\_Score, Debt\_to\_Income\_Ratio)  
- Step 3: Train-test split (e.g., 80-20)  
- Step 4: Model training using Logistic Regression  
- Step 5: Predict probability of delinquency (output range: 0 to 1)  
- Step 6: Threshold tuning (e.g., 0.5) to classify risk  
- Step 7: Evaluate and iterate  
  
Pseudocode (simplified):  
```  
X = dataset[['Income', 'Credit\_Utilization', 'Missed\_Payments', 'Credit\_Score', 'Debt\_to\_Income\_Ratio']]  
y = dataset['Delinquent\_Account']  
model = LogisticRegression()  
model.fit(X\_train, y\_train)  
predictions = model.predict(X\_test)  
```

# 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

Logistic Regression is chosen for its simplicity, interpretability, and wide acceptance in financial applications. It outputs probability scores, which are valuable for risk stratification. Regulators often require explainable models, and logistic regression coefficients offer direct insights into how each feature impacts delinquency risk.  
  
While tree-based models like Random Forests offer higher accuracy, they lack transparency and may be harder to justify in audits. Starting with a logistic model ensures clarity, faster deployment, and alignment with Geldium’s needs for compliance, fairness, and operational readiness.

# 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

The model will be evaluated using the following metrics:  
- \*\*Accuracy\*\*: To measure overall correctness.  
- \*\*Precision & Recall\*\*: Important to balance false positives and false negatives.  
- \*\*F1 Score\*\*: Especially useful if the dataset is imbalanced.  
- \*\*ROC-AUC\*\*: To evaluate how well the model discriminates between classes.  
  
Bias detection methods like disparate impact analysis and subgroup performance metrics (e.g., across employment types or locations) will be applied. If performance gaps exist, techniques like reweighting, fairness-aware preprocessing, or equalized odds post-processing may be used.  
  
Ethical considerations include avoiding over-reliance on sensitive features (e.g., location, employment status) and ensuring predictions do not unfairly disadvantage protected groups.