**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:

# Step-by-step predictive modeling pipeline (simplified)

1. Load and clean dataset

2. Impute missing values:

- Income → median

- Credit Score → mean

- Loan Balance → median

3. Encode categorical variables (One-Hot Encoding)

4. Scale numerical features (StandardScaler)

5. Split data into training and test sets (e.g., 80:20)

6. Train a Logistic Regression model (baseline)

7. Evaluate model using accuracy, recall, F1-score, and ROC-AUC

8. Compare with Decision Tree and Random Forest for improvement

9. Select best-performing model based on balanced metrics

10. Use SHAP or feature importance to interpret decisions

# 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

**Chosen Models**:

* **Logistic Regression (Baseline)**
* **Random Forest Classifier (Enhanced Model)**

| **Criteria** | **Logistic Regression** | **Random Forest** |
| --- | --- | --- |
| Accuracy | Good for baseline | Higher due to ensemble nature |
| Transparency | High (clear coefficients) | Moderate (can extract feature importances) |
| Ease of use | Easy to implement and interpret | Slightly more complex, but manageable |
| Relevance to Finance | Commonly used in credit scoring models | Handles non-linearity and interactions |
| Suitability for Geldium | Provides explainability and risk scores | More accurate, better suited for production |

**Conclusion**: Start with Logistic Regression to build a benchmark. If accuracy or recall is low, upgrade to Random Forest for better performance, especially with imbalanced data and feature interactions.

# 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

**Metrics Used**:

* **Accuracy**: Overall correctness
* **Precision**: How many predicted delinquents were correct
* **Recall (Sensitivity)**: How many actual delinquents were caught
* **F1 Score**: Harmonic mean of precision and recall
* **AUC-ROC**: Measures separation between classes

**Interpretation**:

* **Recall** is critical → Missing a delinquent customer is costly.
* **AUC > 0.8** indicates good model performance.
* **Precision** ensures the model does not over-warn or flag too many false positives.

**Bias Detection & Mitigation**:

* Analyze performance across different groups (e.g., location, employment status).
* Check for disparities in false negative rates.
* Use SHAP values to ensure no discriminatory features dominate.

**Ethical Considerations**:

* Avoid using sensitive variables like race, gender unless for fairness audits.
* Explain predictions to customers transparently.
* Use decisions for **support**, not rejection (e.g., offer counseling, not denial).