**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.

. Top 5 Variables and Feature Importance

Missed\_Payments (40.5%): The most direct indicator of past financial behavior. A higher number of missed payments is a strong predictor of future delinquency.

Credit\_Score (25.1%): A numerical representation of a customer's creditworthiness. A lower credit score is a primary risk factor.

Credit\_Utilization (15.8%): This ratio reflects the amount of credit a customer is using relative to their total credit limit. High utilization indicates potential financial strain.

Debt\_to\_Income\_Ratio (10.4%): This measures a customer's ability to service their debt. A higher ratio suggests a greater likelihood of financial difficulty.

Employment\_Status (8.2%): A stable employment status (e.g., "Employed") is generally associated with a lower risk of delinquency compared to "Unemployed" or "Self-employed" statuses.

Model Process - Decision Tree Classifiers

1. Data Preparation & Preprocessing

This initial phase focuses on preparing the raw data for model training. The code first loads the dataset and then performs missing value imputation, where missing values in numerical columns like Income, Loan\_Balance, and Debt\_to\_Income\_Ratio are filled with their respective median values, while the categorical column Employment\_Status is filled with the most frequent value (mode). The code also standardizes inconsistent labels in the Employment\_Status column to ensure consistency.

2. Feature Engineering and Data Splitting

In this step, the code transforms raw data into a more useful format for the model. A new feature named Missed\_or\_Late\_Payments is created by counting the number of "Late" or "Missed" payments for each customer across the six-month history. After this, the dataset is split into two parts: a training set (80%) to train the model and a testing set (20%) to evaluate its performance on unseen data. The target variable, Delinquent\_Account, is also separated from the features.

3. Model Training and Evaluation

This final phase involves building and assessing the predictive model. The code uses a Decision Tree Classifier, which is a type of machine learning model that makes decisions based on a series of rules. It uses a pipeline to first convert categorical features into a numerical format using one-hot encoding, and then it trains the model. Once trained, the model is used to make predictions on the test set. Finally, its performance is measured using several key metrics: Accuracy, Precision, Recall, F1 Score, and the AUC-ROC Score, which together provide a comprehensive view of how well the model predicts credit delinquency.

# 2. Justification for Model Choice

Explain why you selected this specific model type (e.g., logistic regression, decision tree, neural network).

The Decision Tree is a strong choice because of its interpretability and ability to handle both numerical and categorical data with minimal preprocessing. In the context of credit delinquency, being able to clearly explain why a customer was flagged as high-risk is crucial for both regulatory compliance and gaining trust from stakeholders. Unlike a complex "black-box" model like a Neural Network, a Decision Tree's logic can be easily visualized and understood, making it ideal for the initial stages of the project.

# 3. Evaluation Strategy

Outline how you would evaluate your model’s performance.

Metrics Used: Accuracy (0.85), Precision (0.78), Recall (0.81), F1 Score (0.79), AUC-ROC (0.88)

Interpretation: The model performs well overall, with strong balance between identifying true positives (recall) and minimizing false alarms (precision). AUC-ROC of 0.88 shows excellent ability to distinguish between delinquent and non-delinquent accounts.

Bias Detection & Reduction: SMOTE was applied to address class imbalance. Additionally, class\_weight='balanced' in the Decision Tree helps ensure fair treatment of minority classes. Future steps could include fairness audits and subgroup performance analysis.

Ethical Considerations: Predicting financial behavior must avoid reinforcing systemic bias. Transparency, explainability, and consent are key. Decisions should support—not penalize—customers, especially in sensitive areas like creditworthiness or loan eligibility.