

# 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:

.....  
[Insert GenAI model logic here]  
.....

## 1. Model Logic (Generated with GenAI)

**Model Objective:** The model is designed to predict whether a customer is likely to become delinquent (i.e., miss payments) based on financial and behavioral attributes. It uses logistic regression to estimate the probability of delinquency and classify customers into risk categories.

### Predictive Modeling Workflow (Step-by-Step)

#### 1. Data Ingestion

- Load Geldium's dataset containing 500 customer records.

#### 2. Feature Selection

- Select key predictors based on EDA insights:
  - Credit\_Utilization
  - Missed\_Payments
  - Debt\_to\_Income\_Ratio
  - Income
  - Employment\_Status

#### 3. Preprocessing

- Normalize numerical features (e.g., income, utilization)
- Encode categorical variables (e.g., employment status)

- Handle missing values using imputation strategies

#### 4. Model Training

- Apply logistic regression to learn the relationship between input features and the target variable Delinquent\_Account
- Split data into training and test sets (e.g., 80/20)

#### 5. Prediction Output

- Generate probability scores (0 to 1) for each customer
- Apply a threshold (e.g., 0.5) to classify customers as delinquent or non-delinquent

#### 6. Evaluation

- Use metrics such as accuracy, precision, recall, F1 score, and AUC to assess performance
- Visualize results with a confusion matrix
- Conduct fairness checks across demographic groups

#### Simplified Pseudo-Code (GenAI Output)

---

*python*

```
# Step 1: Load data
data = pd.read_csv("geldium_dataset.csv")

# Step 2: Select features
features = ["Credit_Utilization", "Missed_Payments", "Debt_to_Income_Ratio", "Income",
"Employment_Status"]
X = data[features]
y = data["Delinquent_Account"]

# Step 3: Preprocess
X = preprocess_features(X) # Normalize & encode
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# Step 4: Train model
model = LogisticRegression()
model.fit(X_train, y_train)

# Step 5: Predict
y_pred = model.predict(X_test)
y_prob = model.predict_proba(X_test)[:,1]

# Step 6: Evaluate
evaluate_model(y_test, y_pred, y_prob)
```

This model logic was generated and refined using GenAI tools to ensure clarity, structure, and alignment with Geldium's business goals. It provides a transparent and scalable framework for predicting customer delinquency risk.

## 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

*For this task, I selected **logistic regression** as the predictive model for forecasting customer delinquency. This choice is grounded in its strong alignment with Geldium's operational and regulatory needs.*

Logistic regression is highly effective for **binary classification problems**, making it ideal for predicting whether a customer will become delinquent (Yes/No). It offers **transparent decision-making**, with interpretable coefficients that clearly show how each input feature—such as credit utilization or missed payments—impacts the likelihood of delinquency. This transparency is crucial in financial services, where decisions must be explainable to stakeholders, auditors, and regulators.

From an implementation standpoint, logistic regression is **easy to deploy**, computationally efficient, and well-suited to structured datasets like Geldium's. It handles numerical and categorical variables with minimal preprocessing and performs reliably even with moderate data volumes. While more complex models like neural networks may offer marginal gains in accuracy, they sacrifice interpretability and require more resources to train and monitor.

Ultimately, logistic regression strikes the right balance between **accuracy, simplicity, and accountability**, making it the most appropriate choice for Geldium's goal of building a fair, explainable, and actionable delinquency risk model.

## 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 predictive model for delinquency risk is both effective and responsible, I will use a multi-dimensional evaluation strategy that balances performance metrics with fairness and ethical oversight.

### Performance Metrics

Metric	Purpose
Accuracy	Measures overall correctness of predictions across all classes

Metric	Purpose
<b>Precision</b>	Evaluates how many predicted delinquents are truly delinquent
<b>Recall</b>	Assesses how many actual delinquents were correctly identified
<b>F1 Score</b>	Balances precision and recall, useful when both false positives and negatives are costly
<b>AUC-ROC</b>	Measures the model's ability to rank customers by risk level; a score close to 1 indicates strong discriminatory power

#### Interpretation:

- High **recall** is critical in financial risk modeling to avoid missing high-risk customers.
- High **precision** ensures that flagged customers are truly at risk, minimizing unnecessary interventions.
- **F1 score** helps balance these priorities when trade-offs exist.
- **AUC-ROC** provides a broader view of model effectiveness across thresholds.

#### 👑 Bias Detection and Mitigation

To ensure fairness, I will:

- Conduct **demographic parity checks** to verify that no group (e.g., by age, income, employment status) is disproportionately flagged as high-risk
- Use **disparate impact analysis** to compare prediction rates across protected attributes
- Apply **SHAP** values to interpret feature influence and detect proxy bias (e.g., ZIP code acting as a proxy for socioeconomic status)
- Rebalance the dataset if class imbalance skews predictions (e.g., oversample delinquent cases)

If bias is detected:

- I will adjust feature selection to remove problematic variables
- Reassess model thresholds to reduce unfair classification
- Consider alternative models (e.g., decision trees) if logistic regression shows persistent bias

#### ⊗ Ethical Considerations

Predicting financial behavior carries real-world consequences. To uphold ethical standards:

- I will ensure **transparency** in how predictions are made, using interpretable models and explainability tools

- Avoid using features that may act as proxies for protected characteristics (e.g., race, gender)
- Flag predictions with **confidence scores** to guide human review, not automate decisions blindly
- Ensure that model outputs are used to **support customers**, not penalize them—e.g., offering financial counseling or flexible repayment options

This evaluation strategy ensures that the model is not only accurate but also fair, explainable, and aligned with Geldium's commitment to responsible financial decision-making.