**Predictive Model Plan**

# 1. Model Logic (Generated with GenAI)

Predictive Model Logic for Delinquency Risk Forecasting

1. Model Overview:

The goal of this predictive model is to classify customers as delinquent (1) or non-delinquent (0) based on historical financial and behavioral data. The model will help Geldium Finance assess credit risk and implement proactive intervention strategies.

2. Model Structure (Step-by-Step Approach)

* Data Preprocessing: Handle missing values using median imputation (for Income, Loan\_Balance) and predictive modeling (for Credit\_Score).
* Standardize numerical variables (Credit\_Utilization, Debt\_to\_Income\_Ratio).
* Encode categorical features (Employment\_Status, Credit\_Card\_Type).
* Convert time-series payment history (Month\_1 to Month\_6) into aggregated features like Average Missed Payments.

3.Feature Selection:

* Use correlation analysis and SHAP values to rank influential predictors.
* Select top features: Credit\_Score, Missed\_Payments, Debt\_to\_Income\_Ratio, Credit\_Utilization, Employment\_Status.

4.Model Training:

* Split data into train (80%) and test (20%) sets.
* Train baseline Logistic Regression model for interpretability.
* Use Random Forest / XGBoost for improved predictive accuracy.
* Apply K-Fold Cross-Validation to ensure robustness.

5.Hyperparameter Optimization

* Use Grid Search / Bayesian Optimization to refine model parameters.
* Optimize decision tree depth, learning rate, and feature importance.

6.Model Evaluation Metrics

* Accuracy → Overall model performance.
* Precision & Recall → Ensure correct delinquency identification.
* F1-Score → Balance precision and recall.
* AUC-ROC Curve → Measure how well the model distinguishes delinquent vs. non-delinquent customers.

7. Pseudo-Code for Model Implementation

python

# Step 1: Import Libraries

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split, GridSearchCV

from sklearn.preprocessing import StandardScaler, LabelEncoder

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score, classification\_report  
# Step 2: Load Cleaned Datasetdf = pd.read\_csv("cleaned\_delinquency\_data.csv")  
# Step 3: Feature Selection  
features = ["Credit\_Score", "Missed\_Payments", "Debt\_to\_Income\_Ratio", "Credit\_Utilization", "Employment\_Status"]

X = df[features]  
y = df["Delinquent\_Account"]

# Step 4: Preprocessing

X = pd.get\_dummies(X) # Encode categorical variables

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

# Step 5: Train-Test Split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_scaled, y, test\_size=0.2, random\_state=42)

# Step 6: Train Model (Random Forest)

model = RandomForestClassifier(n\_estimators=100, max\_depth=5, random\_state=42)

model.fit(X\_train, y\_train)

# Step 7: Evaluate Model

y\_pred = model.predict(X\_test)

print("Model Accuracy:", accuracy\_score(y\_test, y\_pred))

print("Classification Report:\n", classification\_report(y\_test, y\_pred))

8. Conclusion & Next Steps

The predictive model is designed to forecast delinquency risk using a combination of financial, behavioral, and categorical features.

Feature engineering enhances accuracy by summarizing payment history trends.

Random Forest and XGBoost are ideal due to their ability to capture complex relationships.  
Next Steps: Fine-tune hyperparameters, perform bias checks, and deploy the model in real-world scenarios.

# 2. Justification for Model Choice

For Geldium Finance’s delinquency risk prediction, XGBoost was selected as the primary model due to its high accuracy, robustness, and ability to handle complex financial data. Unlike traditional logistic regression, which assumes linear relationships between variables, XGBoost efficiently captures nonlinear interactions, making it ideal for financial risk assessment where multiple factors—such as debt-to-income ratio, missed payments, and credit utilization—contribute to delinquency. Additionally, XGBoost supports feature importance analysis using SHAP values, ensuring transparency in decision-making, which is crucial for regulatory compliance in financial services. While neural networks could offer deeper learning, they pose challenges in explainability and computational efficiency, making XGBoost a better fit for operational needs like scalability, speed, and ease of monitoring. Furthermore, XGBoost’s capability to handle imbalanced datasets through boosting techniques ensures that delinquent cases are correctly identified without overfitting. By leveraging this model, Geldium can provide accurate, explainable, and actionable predictions, supporting responsible financial decision-making while optimizing intervention strategies for at-risk customers.

# 3. Evaluation Strategy

1.To ensure the model predicts delinquency risk accurately and fairly, we will assess performance using multiple evaluation metrics:

* Accuracy – Measures overall correctness of predictions.
* Precision – Ensures customers flagged as delinquent are correctly identified, minimizing false positives.
* Recall – Captures most actual delinquent customers, reducing false negatives.
* F1-Score – Balances precision and recall, useful for imbalanced datasets.
* AUC-ROC – Evaluates the model's ability to distinguish between delinquent and non-delinquent customers.
* Fairness Metrics – Checks for biases across demographic groups, such as demographic parity and equalized odds.

2. Interpretation of Metrics

* High Precision: The model avoids mistakenly identifying low-risk customers as delinquent.
* High Recall: The model captures most actual delinquent customers, ensuring intervention happens on time.
* Balanced F1-Score: Ensures trade-offs between precision and recall to improve effectiveness.
* High AUC-ROC (>0.85): Indicates strong discrimination ability between delinquent and non-delinquent customers.
* Fairness Checks: If certain demographic groups (e.g., employment status, income levels) have disproportionate false positives, further bias mitigation is needed.

3. Bias Detection & Mitigation

* To reduce unfair treatment of customer groups:
* Demographic Parity: Ensure loan approvals and delinquency predictions do not disproportionately target certain income levels, employment statuses, or age groups.
* Equalized Odds: Check that recall rates for different customer demographics are consistent, preventing unequal treatment.
* Feature Importance Analysis (SHAP Values): Use SHAP values to understand how features influence predictions, ensuring transparency.
* Adversarial Debiasing: Adjust model predictions by introducing fairness constraints in training.

4. Ethical Considerations

* Regulatory Compliance: The model should align with fair lending laws and ethical AI guidelines in financial decision-making.
* Explainability & Transparency: Predictions must be interpretable, ensuring customers understand their credit risk.
* Avoiding Discriminatory Practices: Decisions should be based on financial behavior, not inherently biased attributes like location or age.

Next Steps:

* Perform initial model evaluation using the selected metrics.
* Run fairness audits and adjust the model if bias exists.
* Improve predictions while ensuring responsible AI practices in financial services.