**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)

The predictive model for credit delinquency will be a binary classification model. The goal is to predict the `Delinquent\_Account` (1 for delinquent, 0 for not delinquent) based on various customer features.  
  
  
**Step-by-Step Pipeline**

* **Data Ingestion**  
  Load the customer dataset containing demographic, financial, and behavioral attributes such as: Age, Income, Credit Score, Credit Utilization, Missed Payments, Loan Balance, Debt-to-Income Ratio, Employment Status, Account Tenure, Credit Card Type, Location, and Month\_1–Month\_6 payment status.
* **Data Preprocessing**
  + Address missing values (median for continuous features; domain-guided approaches for Loan Balance).
  + Encode categorical fields:
    - One-Hot Encoding for Employment Status, Credit Card Type, and Location.
    - Ordinal encoding for Month\_1–Month\_6 payment histories (On-time=0, Late=1, Missed=2).
* **Feature Engineering**
  + Create derived metrics, e.g., total missed payments across 6 months and ratio of late vs. total payments.
  + Build risk-related aggregates such as utilization bands or tenure categories.
* **Feature Scaling**  
  Normalize continuous features using **StandardScaler** to improve model stability.
* **Data Partitioning**  
  Split into **80% training** and **20% testing** sets, applying stratification to preserve delinquency distribution.
* **Model Selection**  
  Initial baseline: **Logistic Regression**.
  + Provides probabilistic outputs.
  + Transparent coefficients for interpretation.
  + Serves as a strong benchmark before moving to more complex models if needed.
* **Model Training**  
  Fit the logistic regression model, learning optimal coefficients for each predictor.
* **Prediction**  
  Generate delinquency probabilities and classify as delinquent if probability > 0.5 (threshold tuning may be applied later).
* **Evaluation**  
  Assess model with precision, recall, F1 score, and AUC-ROC. Conduct subgroup fairness checks to confirm unbiased performance across demographic groups.

**Intended Outcome:**  
The model segments customers into two groups **at risk of delinquency** vs **not at risk** supporting proactive intervention strategies for the collections team.

# 2. Justification for Model Choice

**Logistic Regression Why?**

* **Suitability for Binary Tasks**: Well-established in classification problems where the target variable is dichotomous.
* **Interpretability**: Coefficients map directly to feature influence, which is essential for explaining risk drivers.
* **Regulatory Alignment**: Transparent predictions meet compliance standards and foster trust among financial stakeholders.
* **Computational Efficiency**: Fast training, low overhead, and minimal hyperparameter tuning make it practical for production pipelines.
* **Industry Proven**: Widely adopted in credit risk scoring due to balance of accuracy, interpretability, and explainability.

**Why not alternatives?**

* **Decision Trees**: Can overfit small datasets without careful pruning.
* **Neural Networks**: Offer high accuracy but lack interpretability, creating “black box” challenges and fairness concerns.

Thus, logistic regression is both technically sound and business-aligned for Geldium’s needs.

# 3. Evaluation Strategy

To ensure robust and ethical performance of the model, the following evaluation strategy will be implemented: **Key Metrics**

* **Precision** – Share of correctly identified delinquent customers out of all delinquent predictions (avoids unnecessary interventions).
* **Recall (Sensitivity)** – Fraction of true delinquents captured by the model (reduces missed high-risk accounts).
* **F1 Score** – Balanced measure combining precision and recall; valuable in imbalanced datasets.
* **AUC-ROC** – Global measure of model’s discriminatory power across thresholds.

**Fairness & Bias Analysis**

* **Representation Check**: Review whether certain subgroups (e.g., by employment status, location, or age band) are under/overrepresented.
* **Error Distribution**: Assess disparities in false positives/negatives across demographic cohorts.
* **Equal Opportunity Audit**: Ensure similar true positive rates across sensitive attributes.

**Bias Mitigation (if needed)**

* **Preprocessing**: Oversampling, undersampling, or reweighting for balanced representation.
* **Training-time Adjustments**: Fairness-aware optimization objectives.
* **Postprocessing**: Threshold adjustments to equalize subgroup outcomes.

# Ethical & Governance Principles

 **Transparency**: Provide interpretable coefficients and clear explanations for each prediction.

 **Fairness**: Avoid proxy discrimination; regularly audit for unintended bias.

 **Privacy Compliance**: Ensure adherence to GDPR and local financial data regulations.

 **Human Oversight**: Analysts and risk managers remain involved in reviewing flagged cases.

 **Customer Impact**: Monitor false predictions to prevent reputational or financial harm.

 **Model Lifecycle Management**: Regularly track data drift, recalibrate thresholds, and retrain models as patterns evolve.