Predictive Model Plan - Student Template

1. Model Logic (Generated with GenAI)

Model Objective:

The goal is to predict whether a customer is at risk of credit delinquency using financial and behavioral attributes.

Chosen Model Type: Logistic Regression

This is a binary classification model that predicts the probability of a customer being delinquent (1) or not (0) based on key financial features.

Modeling Workflow:

1. Data Ingestion: Load the customer dataset, focusing on features like Income, Credit_Score, Credit_Utilization, Missed_Payments, Loan_Balance, and Debt_to_Income_Ratio.

2. Preprocessing:

   - Handle missing values.
   - Encode categorical variables like Employment_Status.
   - Normalize/scale numerical features.

3. Feature Selection:

   - Select top predictors using correlation analysis and domain knowledge.
   - Final selected features:
     - Credit_Score
     - Missed_Payments
     - Credit_Utilization
     - Debt_to_Income_Ratio
     - Income

4. Model Training:

   - Train a logistic regression model on 80% of the data.

- Use 20% for validation/testing.

5. Prediction:

   - The model outputs a probability score from 0 to 1.

   - Customers with probabilities > threshold (e.g., 0.5) are flagged as high-risk.

6. Interpretation:

   - Coefficients of the model indicate the importance and direction of influence of each feature.

2. Justification for Model Choice

I chose Logistic Regression because it offers a strong balance between interpretability and predictive power, especially suited for financial services.

- Transparency & Explainability: In financial domains, it's essential to understand why a customer was classified as high risk. Logistic regression provides clear coefficient values that explain the weight of each input feature, which supports regulatory compliance and transparency.
- Accuracy & Relevance: While more complex models like Random Forests may offer slightly better performance, logistic regression performs competitively when paired with strong feature engineering.
- Ease of Implementation: It is fast, scalable, and easy to deploy and monitor in production, which aligns with Geldium's operational needs.
- Business Fit: Financial institutions often prioritize models that are interpretable and easy to audit - logistic regression meets this requirement while still offering solid predictive accuracy.

3. Evaluation Strategy

To ensure the model is both accurate and fair, I would use a combination of performance metrics and bias detection techniques.

Performance Metrics:

- Accuracy: Measures overall correctness of predictions.

- Precision: Helps evaluate how many predicted "delinquents" are actually at risk - useful to avoid false alarms.
- Recall: Measures how many real delinquents we catch - important to prevent missed risk cases.
- F1 Score: Balances precision and recall, especially valuable in imbalanced datasets.
- AUC-ROC: Evaluates how well the model distinguishes between delinquent and non-delinquent customers across different thresholds.

Fairness and Bias Detection:

- Assess model predictions across customer subgroups (e.g., income levels, employment status) using:
  - Disparate Impact Ratio
  - Equal Opportunity Difference
  - Group-wise Precision/Recall
- If bias is found, mitigation strategies like:
  - Feature re-balancing
  - Fairness-aware algorithms
  - Removing proxy variables will be applied.

Ethical Considerations:

- Clearly communicate that this model is risk-based, not a final credit decision tool.
- Ensure that no single feature unfairly determines risk (e.g., employment status alone).
- Monitor post-deployment for performance drift or new biases over time.