**Predictive Model Plan – Student Template**

# 1. Model Logic (Generated with GenAI)

**Purpose:**  
Produce a per-customer probability of becoming delinquent, then flag high-risk accounts for proactive outreach.

**Step-by-Step Process:**

1. **Data Ingestion**  
   • Load cleaned data with features: Missed\_Payments, Credit\_Utilization, Debt\_to\_Income\_Ratio, Credit\_Score, Income, plus one-hot–encoded Employment\_Status and Credit\_Card\_Type.
2. **Preprocessing**  
   • Cap Credit\_Utilization at 1.0.  
   • Impute missing Loan\_Balance and Credit\_Utilization by median per card type.
3. **Feature Engineering**  
   • High\_Util = 1 if Credit\_Utilization > 0.8 else 0.  
   • High\_DTI = 1 if Debt\_to\_Income\_Ratio > 0.5 else 0.
4. **Model Definition & Tuning**  
   • Use DecisionTreeClassifier(max\_depth=5, min\_samples\_leaf=20) as the estimator.  
   • Perform stratified 5-fold CV grid search over max\_depth ∈ {3,5,7} and min\_samples\_leaf ∈ {10,20,50}.
5. **Training**  
   • Fit on features and the binary target Delinquent\_Account.
6. **Prediction**  
   • Compute p\_delinquent = model.predict\_proba(X)[:,1].  
   • Assign predicted\_flag = 1 if p\_delinquent > 0.5, else 0.

from sklearn.tree import DecisionTreeClassifier

from sklearn.model\_selection import StratifiedKFold, GridSearchCV

# Preprocessing & feature engineering omitted for brevity...

X, y = df[FEATURE\_COLUMNS], df["Delinquent\_Account"]

param\_grid = {"max\_depth":[3,5,7], "min\_samples\_leaf":[10,20,50]}

cv = StratifiedKFold(n\_splits=5, shuffle=True, random\_state=42)

grid = GridSearchCV(DecisionTreeClassifier(random\_state=42),

param\_grid, cv=cv, scoring="roc\_auc")

grid.fit(X, y)

best\_tree = grid.best\_estimator\_

df["p\_delinquent"] = best\_tree.predict\_proba(X)[:,1]

df["predicted\_flag"] = (df["p\_delinquent"] > 0.5).astype(int)

# 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

** Accuracy:** Decision trees capture non-linear interactions (e.g., interplay between utilization and DTI) often missed by linear models.

* **Transparency:** Tree structures are easily visualized and explainable—critical for regulatory review in finance.
* **Ease of Implementation:** Minimal feature scaling; handles mixed data types and missing values natively.
* **Financial Relevance:** Pruned trees guard against overfitting on historical outliers, aligning well with credit-risk stability requirements.
* **Business Suitability:** Provides clear decision rules (e.g., “if Missed\_Payments > 4 and High\_Util = 1 then high risk”), enabling straightforward integration into Geldium’s risk workflows.

3. Evaluation Strategy

**Metrics:**

* **AUC-ROC:** Measures discrimination ability across thresholds.
* **Precision & Recall:** Prioritize recall on the delinquent class to minimize missed high-risk cases; monitor precision to contain false alarms.
* **F1 Score:** Balances precision/recall for overall model effectiveness.

**Interpretation:**

* AUC-ROC ≥ 0.75 indicates solid separation of delinquent vs. non-delinquent.
* Recall ≥ 0.80 ensures most future delinquencies are flagged.
* Precision ≥ 0.60 keeps outreach resources focused on true positives.

**Bias Detection & Mitigation:**

* **Subgroup Analysis:** Compare performance (e.g., recall, precision) across demographics (age bands, card types) to detect disparate outcomes.
* **Rebalancing:** If under-prediction exists for any group, introduce sample-weight adjustments or fairness-constrained optimization.

**Ethical Considerations:**

* Avoid unjust discrimination: ensure features used (e.g., Location) don’t proxy for protected attributes.
* Maintain transparency: provide clear explanations for each credit-risk decision to customers and regulators.
* Privacy: handle all customer data in compliance with data-protection policies.

# 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

**Metrics:**

* **AUC-ROC:** Measures discrimination ability across thresholds.
* **Precision & Recall:** Prioritize recall on the delinquent class to minimize missed high-risk cases; monitor precision to contain false alarms.
* **F1 Score:** Balances precision/recall for overall model effectiveness.

**Interpretation:**

* AUC-ROC ≥ 0.75 indicates solid separation of delinquent vs. non-delinquent.
* Recall ≥ 0.80 ensures most future delinquencies are flagged.
* Precision ≥ 0.60 keeps outreach resources focused on true positives.

**Bias Detection & Mitigation:**

* **Subgroup Analysis:** Compare performance (e.g., recall, precision) across demographics (age bands, card types) to detect disparate outcomes.
* **Rebalancing:** If under-prediction exists for any group, introduce sample-weight adjustments or fairness-constrained optimization.

**Ethical Considerations:**

* Avoid unjust discrimination: ensure features used (e.g., Location) don’t proxy for protected attributes.
* Maintain transparency: provide clear explanations for each credit-risk decision to customers and regulators.
* Privacy: handle all customer data in compliance with data-protection policies.