Predictive Model Plan

1. Model Logic (Generated with GenAI)

Goal: Predict whether a customer will become delinquent (Delinquent_Account: 0/1) so the Collections team can prioritize outreach.

Model choice: Logistic Regression (primary, probability-based + explainable). Decision Tree as a secondary option for simple rule explanations.

Pipeline (step-by-step):

1. **Load data** (500 customers, 19 columns). Target = Delinquent_Account.
2. **Handle missing values**
    - Income has 39 missing (~7.8%): impute with median within Employment_Status (fallback: overall median)
    - Credit_Score has 2 missing (~0.4%): impute with **median Credit_Score**.
3. **Preprocess features**
    - Numeric: Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Loan_Balance, Debt_to_Income_Ratio, Account_Tenure.
    - Categorical: Employment_Status, Credit_Card_Type, Location, and payment history Month_1 to Month_6 (On-time/Late/Missed).
    - Encode categorical variables (one-hot encoding). Scale numeric features if required for logistic regression.
4. **Feature engineering (optional but helpful)**
    - Create a **Payment Risk Score** from Month_1…Month_6 (e.g., count of “Missed” + count of “Late”).
5. **Train the model**
    - Use a **stratified split** (keeps delinquent/non-delinquent ratio similar) and train Logistic Regression to output **probability of delinquency**.
6. **Prediction output**
    - Convert probability into risk classes using a threshold (e.g., 0.5 or tuned): **High / Medium / Low risk**.
7. **Key features to focus on (Top 5)**
    - Payment history (Month_1–Month_6) / Payment Risk Score

- Missed_Payments
- Debt_to_Income_Ratio
- Credit_Utilization
- Credit_Score

GenAI recommended an interpretable binary classification approach to predict Delinquent_Account (0/1), using Logistic Regression to generate probability-based risk scores. It suggested median imputation for missing Income and Credit_Score, one-hot encoding for categorical features (Employment_Status, Credit_Card_Type, Location, Month_1–Month_6), and a stratified train/test split due to class imbalance (84% non-delinquent vs 16% delinquent). It also recommended a simple “Payment Risk Score” from recent payment history and prioritizing predictors such as payment history, missed payments, DTI ratio, credit utilization, and credit score.

2. Justification for Model Choice

I selected Logistic Regression because Geldium’s goal is to identify at-risk customers in a way that is both accurate and explainable for business users in the Collections team. Logistic regression outputs a probability score (risk score), which is practical for prioritizing outreach (Probability score + risk tiers (High/Medium/Low) to prioritize outreach) instead of only giving a yes/no answer. It also supports transparency—we can explain directionally how drivers like payment history (Late/Missed), missed payments count, credit utilization, debt-to-income ratio, income, and credit score influence delinquency risk, which is important in financial decision-making. It is relatively easy to implement and monitor, works well with structured tabular data like this dataset (500 rows, mixed numeric + categorical), and provides a strong baseline model that can be improved later (e.g., adding a decision tree as a secondary comparison if needed).

3. Evaluation Strategy

Primary goal metric: Recall for delinquent class, balanced with Precision based on Collections capacity. I would evaluate the model using a stratified train/test split (to preserve the 84% non-delinquent / 16% delinquent ratio). Because delinquency is the minority class, I will not rely only on accuracy. I will track Precision, Recall, and F1-score for the delinquent class (1), plus ROC-AUC to measure how well the model ranks high-risk customers. Business-wise, Recall is critical (missing a truly delinquent customer can increase losses), so I would tune the probability threshold to balance Recall vs Precision based on Collections capacity (how many customers can be contacted). I will also review a confusion matrix to understand false positives/false negatives and compare performance against a simple baseline rule (e.g., “Missed_Payments > 0”).
For fairness and ethics, I will check performance gaps across key groups (e.g., Employment_Status, Location, and age bands) by comparing precision/recall by segment and watching for consistently higher false positives in any group. I would avoid using sensitive attributes directly for decisioning and ensure the model is used to support outreach prioritization, not as the only basis for punitive actions. If bias is detected, I would adjust features (remove proxy variables where needed), rebalance training, and retune thresholds before deployment.