Predictive Model Plan for Delinquency Prediction

1. Model Logic (Generated with GenAI)

To forecast customer delinquency risk, I propose using a Logistic Regression Model as the primary approach, with Random Forest as a more complex alternative.

Model Workflow: - Data Ingestion: Load customer financial and behavioral data. - Data Preprocessing: Handle missing values using imputation (median for income, mode for categorical variables). Convert categorical features (Employment_Status, Credit_Card_Type) into numerical representations using one-hot encoding. - Feature Selection: Select top predictors: Income, Credit Utilization, Missed Payments, Credit Score, Debt-to-Income Ratio. - Model Training: Train a Logistic Regression model to predict the probability of delinquency (Delinquent_Account = 1). Alternatively, train a Random Forest Classifier for more complex patterns. - Model Evaluation: Use metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC. - Prediction: The model outputs the probability of each customer being delinquent, which can be thresholded into a binary prediction. - Interpretation: Use model coefficients (Logistic) or feature importance (Random Forest) to explain predictions to decision-makers.

Top 5 Features Selected: 1. Missed_Payments – Direct behavioral indicator of past payment issues. 2. Credit_Utilization – High utilization suggests financial stress. 3. Income – Lower income may increase delinquency risk. 4. Debt_to_Income_Ratio – Measures financial burden relative to income. 5. Credit_Score – Established risk predictor in credit modeling.

1. Justification for Model Choice

I recommend using Logistic Regression as the primary model for delinquency prediction due to its: - High interpretability: Logistic Regression clearly shows how each feature influences the probability of delinquency, making it transparent for financial regulators and decision-makers. - Simplicity and speed: It is easy to implement, fast to train, and efficient in making predictions on large datasets. - Alignment with industry norms: Logistic Regression is widely used in financial services due to its robustness and ease of compliance with regulatory requirements. - Actionable outputs: The model provides probability scores, which can be easily converted into risk tiers for customer interventions.

As a secondary option, a Random Forest Classifier can be explored to capture nonlinear patterns or interactions that Logistic Regression may miss, though this comes at the cost of reduced transparency.

Given Geldium's need for fairness, compliance, and operational efficiency, Logistic Regression aligns well with business objectives while offering opportunities to scale towards more complex models later.

1. Evaluation Strategy

To ensure the model is both accurate and fair, the following evaluation strategy will be used:

Performance Metrics: - Accuracy: Measures overall correctness but can be misleading in imbalanced datasets. - Precision & Recall: Precision focuses on minimizing false positives, while Recall ensures true delinquent cases are identified. - F1-Score: Balances Precision and Recall, useful when class distribution is skewed. - AUC-ROC: Measures the model's ability to distinguish between classes across thresholds.

Fairness and Bias Checks: - Apply Demographic Parity checks to ensure equal opportunity across sensitive groups (e.g., employment status, location). - Use Disparate Impact Analysis to identify any groups disproportionately affected by predictions.

Bias Mitigation: - Apply re-sampling techniques (oversampling or undersampling). - Use model explainability tools (e.g., SHAP values) to validate that decisions are not driven by biases. - Continuously monitor model performance across demographic groups.

Ethical Considerations: - Ensure the model does not reinforce existing inequalities. - Maintain transparency and communicate clearly with customers impacted by risk predictions.