Predictive Model Plan – Student Template

1. Model Logic (Generated with GenAI)

Step-by-Step Model Workflow:

1. **Data Ingestion**
   Load structured customer data containing demographic and financial attributes.

2. **Preprocessing**
   - Handle missing values: Impute Income (median), Loan_Balance (mean), Credit_Score (KNN or mean).
   - Encode categorical variables like Employment_Status, Credit_Card_Type.
   - Normalize continuous variables such as Credit_Utilization.

3. **Feature Selection**
   - Use domain knowledge and correlation analysis to select top predictors:
     - Missed_Payments
     - Credit_Utilization
     - Debt_to_Income_Ratio
     - Income
     - Credit_Score

4. **Model Training**
   - Train a **Logistic Regression** model using the selected features and target variable (Delinquent_Account).

5. **Prediction**
   - For new inputs, the model generates a **delinquency risk probability** (e.g., 0.83).
   - Classify the output as **delinquent (1)** if risk ≥ 0.5, else **not delinquent (0)**.

6. **Evaluation**
   - Assess performance using metrics like **F1 Score**, **AUC**, **Accuracy**, and **fairness checks**.

Pseudocode Example:

```
# Load and preprocess data
data = load_data()
data = preprocess(data)  # Impute, encode, scale

# Define features and label
X = data[["Missed_Payments", "Credit_Utilization", "Debt_to_Income_Ratio", "Income",
"Credit_Score"]]
y = data["Delinquent_Account"]

# Split data and train model
X_train, X_test, y_train, y_test = train_test_split(X, y)
model = LogisticRegression()
model.fit(X_train, y_train)

# Predict and evaluate
predictions = model.predict(X_test)
evaluate_model(predictions, y_test)
```

2. Justification for Model Choice

I selected **Logistic Regression** as the primary model for forecasting credit delinquency because it offers the ideal balance of **accuracy**, **transparency**, and **practicality**, particularly in financial services settings like Geldium's.

- **Accuracy:** While logistic regression may not outperform complex models like random forests in all cases, it still achieves strong baseline accuracy when used with well-engineered features such as `Missed_Payments`, `Credit_Utilization`, and `Debt_to_Income_Ratio`.

- **Transparency:** One of its biggest strengths is interpretability. Each model coefficient clearly shows how a feature influences delinquency risk. This is crucial for gaining internal stakeholder trust and satisfying external regulatory demands.

- **Ease of Implementation:** Logistic regression is easy to implement, computationally efficient, and scales well even with large datasets. It's also less sensitive to overfitting than deeper models if regularized appropriately.

- **Relevance for Financial Prediction:** The model's probabilistic output (a risk score between 0 and 1) aligns perfectly with credit risk applications, where cutoffs can be adjusted based on business policy (e.g., 0.6 for higher-risk clients).

- **Suitability for Geldium:** For a company like Geldium, which needs explainable and reliable credit risk predictions, logistic regression provides both compliance-aligned decision-making and actionable insights, without the black-box tradeoffs of neural networks or ensemble models.

3. Evaluation Strategy

Model Evaluation Plan

To ensure the predictive model is both **accurate** and **fair**, I would evaluate it using a combination of **performance metrics**, **bias checks**, and **ethical safeguards**.

1. Metrics Chosen

- **F1 Score** – Handles class imbalance; balances false positives and negatives.
- **AUC-ROC** – Evaluates how well the model distinguishes delinquent vs. non-delinquent customers.
- **Equal Opportunity Difference** – Measures fairness in identifying delinquents across groups (e.g., Employment_Status).
- **Disparate Impact Ratio** – Ensures that favorable outcomes (not flagged delinquent) are equitably distributed.
- **Calibration by Group** – Verifies predicted probabilities match actual delinquency risk by segment (e.g., income levels).

2. Metric Interpretation

- **F1 Score > 0.7** = good balance of recall & precision.
- **AUC > 0.8** = strong ability to rank customers by risk.
- **Equal Opportunity ≈ 0 & Disparate Impact ≈ 1.0** = minimal demographic bias.
- **Poor calibration** may suggest the model is over/underestimating risk for certain customer types.

3. Bias Detection & Reduction (on Geldium data)

- Evaluate false positive/negative rates across:
    - `Employment_Status` (e.g., Self-employed vs. Employed)
    - `Income` brackets
    - `Credit_Card_Type`

- If bias is found:
    - Apply **reweighting**, **resampling**, or **fairness constraints**.
    - Retrain using balanced or group-aware techniques.

4. Ethical Considerations

- Avoid using **proxy variables** that encode bias (e.g., location, if strongly tied to economic disparity).
- Provide **explainable risk scores** for customer transparency.
- Enable **human review** for high-risk classifications to prevent wrongful rejection.
- Ensure compliance with **regulatory expectations** (e.g., explainability, fairness in lending).