Predictive Model Plan

Use this template to structure your submission. You can copy and paste content from GenAI tools and build around it with your own analysis.

1. Model Logic (Generated with GenAI)

- This predictive model is designed to classify whether a customer is at risk of becoming delinquent (i.e., failing to meet credit or loan obligations).

- Based on behavioral and financial data, the model outputs a binary prediction (0 = not delinquent, 1 = delinquent) to support credit risk assessment and early intervention strategies.

Modelling Pipeline Overview

Step 1: Feature Selection

❖ **Top features identified through EDA:**

- **Missed_Payments** – Direct behavioural risk indicator.
- **Credit_Score** – Captures historical credit responsibility.
- **Credit_Utilization** – Measures financial stress.
- **Debt_to_Income_Ratio** – Reflects repayment capacity.
- **Income** – Provides context for financial stability.

Step 2: Data Preprocessing:

- Impute missing values (e.g., synthetic for Income, median for Loan_Balance).
- Normalize or scale numerical features. Encode categorical features like Employment_Status using one- hot encoding.

Step 3: Model Options:

I. **Simple Model**: Logistic Regression
A highly interpretable linear model that estimates the probability of delinquency using a weighted combination of input features. It's ideal when transparency and explainability are essential in credit decision-making.

II. **Complex Model:** Neural Network (Multilayer Perceptron)
A more advanced, non-linear model capable of capturing complex interactions between variables. With multiple hidden layers and non-linear activation functions, neural networks can improve predictive performance in high-dimensional data**.**

```
[Insert GenAI model logic here]
```

1. **Input:** New customer data (Income, Credit Score, Missed Payments, etc.)

2. **Preprocessing:** Handle missing values, normalize inputs, encode categories

3. **Model Inference:**

- **Logistic Regression:** Apply linear weights to compute risk score

- **Neural Network:** Pass inputs through multiple layers to compute prediction

4. **Output:** Return delinquency probability (e.g., 0.82)

5. **Classification:** Apply threshold (e.g., 0.5) to determine risk label

6. **Interpretability (for logistic):** Use feature coefficients to explain predictions

Summary:

The recommended model depends on the use case:

- I prefer **Logistic Regression** for its clarity and reliability.
- I prefer **Neural Networks** for higher accuracy and adaptability when working with large datasets and complex interactions.

❖ Top 5 Input Features:

1. Missed_Payments
2. Credit_Score
3. Credit_Utilization
4. Debt_to_Income_Ratio
5. Income

2. Justification for Model Choice

I recommend **logistic regression** as the most appropriate model for predicting credit delinquency in alignment with Geldium's business priorities. This model is not only accurate and efficient for binary classification tasks but also excels in **interpretability**, a critical factor in the financial domain. With logistic regression, stakeholders can clearly trace how features like **missed payments**, **credit score**, and **credit utilization** influence the risk prediction, enabling fully transparent and **auditable decision-making**. This is especially important for ensuring **regulatory compliance** and gaining stakeholder trust. While advanced models like neural networks may offer marginally higher accuracy, they often lack explainability and require more complex infrastructure. In contrast, logistic regression strikes the right balance between performance, **ease of deployment**, and operational clarity—making it the ideal choice for Geldium's risk management objectives and long-term model governance.

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

To evaluate the performance of the delinquency prediction model, I would use a combination of metrics that balance accuracy, reliability, and fairness. The **primary evaluation metrics** include:

- **Accuracy**: Measures the overall correctness of the model, but can be misleading in imbalanced datasets.
- **Precision**: Indicates how many customers predicted as delinquent were actually delinquent, which is crucial for minimizing false alarms.
- **Recall**: Measures how many actual delinquents the model correctly identified — important for capturing risky customers.

- **F1 Score**: The harmonic mean of precision and recall, providing a balanced view of the model's performance.
- **AUC-ROC (Area Under the Curve)**: Evaluates the model's ability to discriminate between delinquent and non-delinquent customers across all thresholds.

These metrics will be interpreted as follows: **A high F1 score** would signal the model is effectively identifying delinquency without over-flagging, while a **high AUC score** would indicate strong separation between classes. In addition, **confusion matrices** will be reviewed to understand false positives and false negatives.

To ensure fairness, I would conduct **bias and fairness audits**, examining model performance across subgroups such as age, income level, or employment status. If disparate impact is detected, techniques like **reweighting**, **group-specific thresholds**, or **fairness-aware training** would be applied to mitigate bias.

Finally, ethical considerations are paramount when modeling customer financial behavior. Transparency in model decision-making and limiting over-reliance on sensitive features help protect customer rights. Regular reviews will be built into the deployment process to ensure the model continues to operate fairly and responsibly.