Predictive Model Plan – Student Template

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)

Use a GenAI tool (e.g., ChatGPT, Gemini) to generate the logic or structure of your predictive model.
- You may include pseudo-code, a step-by-step process, or a simplified code snippet.
- Briefly explain what the model is designed to do.

Paste your GenAI-generated output below or describe the logic in your own words:

```
import xgboost as xgb

from sklearn.model_selection import
train_test_split

from sklearn.metrics import roc_auc_score


# 1. Load data with relevant features

# Features: income, credit_utilization,
missed_payments_12m, debt_to_income,
age

X = data[['income', 'credit_utilization',
'missed_payments_12m', 'debt_to_income',
'age']]

y = data['is_delinquent'] # Target variable (1 =
default, 0 = paid)


# 2. Split into Training and Testing sets

X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2,
random_state=42)


# 3. Initialize and Train the Complex Model
(XGBoost)
```

```
model = xgb.XGBClassifier(

n_estimators=500,

max_depth=5,

learning_rate=0.05,

objective='binary:logistic'

)

model.fit(X_train, y_train)


# 4. Predict and Evaluate

probs = model.predict_proba(X_test)[:, 1]

print(f"Model AUC Performance:
{roc_auc_score(y_test, probs)}")
```

[

Summary: XGBoost Classifier

This model uses a gradient boosting framework to build an ensemble of decision trees, iteratively correcting the errors of previous trees to achieve high precision in risk detection. It is ideal for modern fintech applications were maximizing the "Area Under the Curve" (AUC) directly translates to millions in saved capital.

Top 5 Input Features:

1. **Credit Utilization Ratio:** (Total Balance / Total Limits)
2. **Payment History:** Number of 30/60/90 days-past-due instances.
3. **Debt-to-Income (DTI):** Monthly debt obligations vs. gross income.
4. **Length of Credit History:** Age of the oldest active account.
5. **Recent Inquiries:** Number of hard credits pulls in the last 6 months.

]

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
[

Given Geldium’s objective to transition from historical trend analysis to a proactive, AI-driven system while maintaining strict regulatory compliance and transparency, Logistic Regression is the most justifiable primary choice. While the project explored Neural Networks and Decision Trees, Logistic Regression directly supports Geldium’s need for "actionable insights" and "ethical decision-making" by providing a clear, statistically grounded reason for every risk score. This ensures that as Geldium automates its collections and outreach, the underlying decisions remain fair, auditable, and easily explainable to both regulators and customers.

]

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
[

Our strategy prioritizes **Reliability** and **Equity** to ensure Geldium's AI meets both financial and ethical standards.

- **Primary Metric: Precision-Recall AUC (PR-AUC).** Because our data is imbalanced (84% non-delinquent), PR-AUC provides a truer picture of performance than standard accuracy. We interpret a high PR-AUC as the model successfully identifying high-risk individuals without flooding the collections team with "false alarms."

- **Fairness Audit: Group-wise Recall Analysis.** We will calculate Recall across Employment_Status and Location. If the model is significantly better at catching delinquents in one group than another, it suggests the model is "learning" a bias rather than a financial pattern.

- **Interpretation Strategy:** We use **SHAP (SHapley Additive exPlanations)** to ensure the top drivers of a "High Risk" score are justifiable variables (like Credit_Utilization or Missed_Payments) rather than sensitive proxies (like Location or Age).

- **Success Threshold:** A model is considered "Production Ready" only if it achieves a **Gini > 0.40** and maintains a **Disparate Impact Ratio > 0.85** across all protected classes.]