Predictive Modeling Plan: Delinquency Risk Forecasting for Geldium

Prepared by: Sanket Bansal
Tata iQ

Step 1: Predictive Model Logic

Chosen Model: Logistic Regression (Baseline) with Gradient Boosting (Advanced Option)

We propose two modeling approaches:

- **Baseline Model:** Logistic Regression – for interpretability, ease of deployment, and transparency.
- **Advanced Model:** Gradient Boosted Trees (e.g., XGBoost) – for improved predictive performance and handling non-linear feature interactions.

Given the financial domain's emphasis on interpretability and fairness, logistic regression can serve as a trusted baseline. However, to enhance predictive power, we recommend developing a second version using gradient boosting.

Key Input Features

Based on EDA findings, the top 5 features selected for modeling are:

1. **Credit Utilization Ratio** – High ratios (>90%) strongly correlate with delinquency.
2. **Payment History (3-month window)** – Recency and frequency of missed payments are strong predictors.
3. **Debt-to-Income Ratio** – Indicates financial stress and repayment capacity.
4. **Auto-Pay Enrollment Status** – Customers not enrolled in auto-pay tend to miss payments more often.

5. **Employment Change Indicator** – Sudden employment status changes can impact income stability.

Model Workflow

1. **Data Ingestion:** Cleaned and pre-processed dataset is fed into the pipeline.

2. **Feature Engineering:** Derived ratios, binary indicators, and normalized continuous variables.

3. **Model Training:** Train baseline logistic regression, followed by hyperparameter-tuned gradient boosting.

4. **Prediction Output:** Model outputs a probability score between 0 and 1 indicating delinquency risk.

5. **Thresholding & Classification:** Customers exceeding a risk threshold (e.g., 0.65) are flagged for intervention.

Step 2: Model Justification

We selected **logistic regression** as the primary model due to its **interpretability**, regulatory acceptability, and ease of explanation to business stakeholders. Financial institutions often require transparent models that clearly explain why a decision was made, especially when it affects credit and collection strategies. Logistic regression provides straightforward coefficient interpretation and probability outputs, which are ideal for creating explainable risk scores.

To improve performance, we recommend evaluating a **gradient boosting model** (e.g., XGBoost), which captures complex, non-linear relationships and typically outperforms linear models on structured financial data. While this model is less interpretable by default, techniques like SHAP (SHapley Additive Explanations) can help explain individual predictions, maintaining a level of transparency.

This dual-model approach aligns with **Geldium's goals** of improving prediction accuracy while maintaining fairness and operational feasibility.

Step 3: Evaluation Strategy

Key Metrics for Accuracy and Reliability

- **AUC-ROC (Area Under Curve – Receiver Operating Characteristic):** Measures the model's ability to distinguish between delinquent and non-delinquent customers.
- **F1 Score:** Balances precision and recall, important in handling imbalanced data.
- **Accuracy:** Overall correctness of classification, used cautiously due to potential class imbalance.
- **Confusion Matrix:** To analyze true positives and false negatives.

Fairness & Bias Detection

- **Demographic Parity & Equal Opportunity Checks:** Ensure the model does not discriminate across sensitive groups such as age, gender, or income segments.
- **Disparate Impact Ratio:** Evaluate whether any demographic group is being disproportionately flagged as high risk.
- **SHAP Value Analysis:** Identify which features contribute most to risk predictions, ensuring no hidden bias in feature influence.

Model Monitoring

- Track performance over time using a validation set and monitor model drift.
- Re-evaluate model monthly using updated customer data.
- Establish a feedback loop with the Collections team for misclassified cases.

Conclusion

The proposed predictive modeling framework is designed to meet Geldium’s strategic objectives: improving delinquency detection, guiding intervention strategies, and maintaining ethical AI practices. By combining logistic regression’s transparency with the accuracy of gradient boosting models, we offer a balanced, effective, and responsible approach to credit risk modeling. Evaluation and fairness checks are integrated into the framework to support ongoing model governance and business confidence.