Predictive Model Plan for Credit Delinquency Risk Assessment

1. Model Logic (Generated with GenAI)

Using ChatGPT, I generated a predictive model using logistic regression to estimate the likelihood of a customer becoming delinquent. The model uses key features such as Credit_Utilization, Missed_Payments, Income, Debt_to_Income_Ratio, and Account_Tenure to predict a binary outcome: 1 if the customer is likely to become delinquent, and 0 otherwise.

1.1 Pseudo-code for Model Workflow:

```
1. Load dataset
2. Select features: ['Credit_Utilization', 'Missed_Payments', 'Income', 'Debt_to_Income_Ratio', 'Account_Tenure']
3. Define target variable: 'Delinquent_Account'
4. Split data into training and testing sets
5. Fit logistic regression model
6. Predict and evaluate using classification metrics
```

```
# Step 1: Load Dataset
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix, roc_auc_score

# Step 2: Select Features
features = ['Credit_Utilization', 'Missed_Payments', 'Income', 'Debt_to_Income_Ratio', 'Account_Tenure']
target = 'Delinquent_Account'

# Step 3: Define Target Variable
df = pd.read_csv('customer_data.csv')
X = df[features]
y = df[target]

# Step 4: Split Data into Training and Testing Sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Step 5: Fit Logistic Regression Model
model = LogisticRegression()
model.fit(X_train, y_train)

# Step 6: Predict and Evaluate
y_pred = model.predict(X_test)
print(classification_report(y_test, y_pred))
print(f"AUC: {roc_auc_score(y_test, model.predict_proba(X_test)[:,1])}")
```

2. Justification for Model Choice

2.1 Why Logistic Regression?

I selected **logistic regression** for the following reasons:

- **Simplicity & Interpretability**: Logistic regression is a linear model that offers clear and understandable coefficient outputs. For each predictor, we can assess its impact on the probability of delinquency, making it highly interpretable—critical in financial services where regulatory compliance and transparency are essential.
- **Strong Baseline Performance**: Logistic regression serves as an excellent baseline model for binary classification problems. Even though it may not always be the most complex or accurate model, its simplicity enables fast iteration and testing. It works particularly well when the relationship between the dependent and independent variables is approximately linear.
- **Low Computational Overhead**: Logistic regression does not require extensive computational resources, making it suitable for real-time predictions in production environments where computational efficiency is crucial.
- **Scalability**: While more complex models may improve performance, logistic regression remains a strong model for scalable use cases in risk assessment, where predictability and execution speed are key requirements.

2.2 Business Fit

In the context of **Geldium**, logistic regression helps align the **credit risk assessment process** with business goals by:

- **Risk Transparency**: Clear understanding of how each feature (e.g., income, credit utilization) influences delinquency prediction allows business stakeholders to validate and communicate the model’s decisions to customers.
- **Model Speed & Simplicity**: In a fast-paced financial environment, the simplicity of logistic regression supports quick decision-making processes, such as flagging high-risk customers for targeted outreach or interventions.
- **Regulatory Compliance**: Financial institutions must ensure that the decision-making process is transparent, justifiable, and auditable. Logistic regression’s coefficients are easy to explain in layman’s terms, which facilitates regulatory compliance.

3. Evaluation Strategy

3.1 Performance Metrics

Given that credit delinquency is a critical business issue, the model's evaluation requires a multi-faceted approach. The following metrics will be employed to assess performance:

- **Accuracy**: Measures the overall correctness of the model. While useful, it may not be sufficient for imbalanced datasets (e.g., if most customers do not become delinquent).
- **Precision**: Focuses on the proportion of true positive predictions out of all positive predictions. Precision is crucial for minimizing false positives, as unnecessary interventions or outreach for low-risk customers can be costly.
- **Recall (Sensitivity)**: Measures the ability of the model to identify actual delinquents. Recall is critical for ensuring that high-risk customers are not overlooked.
- **F1 Score**: This is the harmonic mean of precision and recall. It balances the importance of both metrics and is crucial for evaluating models where the costs of false positives and false negatives are both significant.

- **AUC-ROC**: The area under the ROC curve gives an indication of how well the model distinguishes between the delinquent and non-delinquent classes across various thresholds. A high AUC suggests the model has a good capability of differentiating between the two classes.

3.2 Cross-Validation

- **K-fold Cross-Validation** will be used to ensure the model generalizes well across different subsets of the data, minimizing overfitting and providing a more robust performance estimate.
- **Stratified Sampling**: For a more balanced evaluation, stratified sampling will be used to maintain the proportion of delinquent and non-delinquent instances across training and testing splits.

3. Bias Detection and Model Fairness

4.1 Bias Detection

Ensuring fairness and transparency is paramount, especially in financial services, where discriminatory models could have legal and reputational implications. The model evaluation will include:

- **Demographic Group Analysis**: The model's predictions will be examined for potential disparities across **Employment Status**, **Location**, and other demographic features. If the model shows significant bias (e.g., lower accuracy for unemployed individuals), it will prompt a rebalancing or modification of the model's features or training process.
- **Feature Sensitivity**: An in-depth sensitivity analysis will be conducted to verify whether any features (such as **Income** or **Credit Score**) disproportionately affect certain demographic groups in an unfair way.

4.2 Ethical Considerations

- **Proxy Bias**: Special care will be taken to avoid using features that could introduce indirect bias (e.g., using race or gender as proxies for creditworthiness).
- **Transparent Decision-Making**: Clear documentation will be maintained for all model decisions and outputs. A transparent **explainability framework** will be put in place (e.g., **LIME** or **SHAP**) to explain individual predictions in business terms.
- **Fairness Regulations**: We will adhere to financial regulations such as the **Equal Credit Opportunity Act (ECOA)** to ensure that no one is discriminated against based on race, religion, national origin, or other prohibited categories.

5. Model Refinement and Next Steps

5.1 Hyperparameter Tuning

To further optimize the logistic regression model:

- **Regularization**: L1 or L2 regularization will be applied to prevent overfitting and improve model generalization.
- **Grid Search**: A grid search approach will be used to identify the best regularization strength and solver parameters for the logistic regression model.

5.2 Model Update Strategy

- **Continuous Monitoring**: Post-deployment, the model will be actively monitored for changes in performance, particularly as the customer base evolves over time.
- **Recalibration**: The model's decision threshold can be adjusted based on evolving business requirements or as new data becomes available (e.g., adjusting the threshold to favor recall in high-risk periods).

5.3 Model Deployment

Once the model is validated and tuned, it will be integrated into the credit risk decisioning process within Geldium. A deployment pipeline will be set up, ensuring that the model is automatically retrained on fresh data, providing timely insights to the business.

6. Business Impact and Stakeholder Communication

This predictive model will directly influence Geldium's ability to proactively manage credit risk, reducing the number of delinquent accounts, improving customer retention, and optimizing collections efforts.

- **Operational Efficiency**: By accurately identifying high-risk customers, resources can be directed towards those most likely to default, improving operational efficiency.
- **Customer Engagement**: The model will enable personalized outreach strategies, helping to engage customers early and offer tailored solutions, reducing the likelihood of delinquency.
- **Regulatory and Ethical Compliance**: Ensuring model transparency and fairness will not only mitigate legal risks but will also enhance customer trust in Geldium's services.

Conclusion

This predictive model, based on logistic regression, is designed to provide actionable insights into the risk of customer delinquency. By using a simple yet powerful model, Geldium can identify at-risk customers early, enhance decision-making, and take proactive steps to mitigate risk. The evaluation and ethical considerations laid out in this plan will ensure the model's robustness, fairness, and transparency, aligning it with industry standards and business goals.