Model Performance Report

Development of a Mathematical Model for GSTN Problem

1. Introduction

The objective of this project was to develop a robust mathematical model using LightGBM to handle a given dataset with class imbalance. In such cases, it is crucial to minimize false negatives by ensuring that all instances of the minority class (labeled as '1') are correctly identified, even if it leads to some increase in false positives. To achieve this, a very low threshold (0.001) was selected after evaluating model performance across different thresholds.

2. Evaluation Metrics

The model's performance was evaluated using the following key metrics:

- **Accuracy**: The proportion of correctly classified instances (both true positives and true negatives) out of the total instances.
- **Precision**: The proportion of true positive instances out of the instances predicted as positive.
- **Recall (Sensitivity)**: The proportion of true positive instances out of the actual positive instances. Given the class imbalance, recall was prioritized to minimize false negatives.
- **F1 Score**: The harmonic mean of precision and recall.
- **AUC-ROC**: The Area Under the Receiver Operating Characteristic Curve, which measures the ability of the model to distinguish between classes.
- **Confusion Matrix**: A table that breaks down true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), providing insight into the classification accuracy.
- **Feature Importance**: Analysis of the contribution of each feature to the model.

3. Model Performance Metrics

After training the LightGBM model, the following performance metrics were achieved using the threshold of **0.001**:

- **Accuracy**: 96.56%
- **Mean F1 Score**: 91.32%
- **Recall**: 100% (1.0)
- **Precision**: 73.2%
- **ROC AUC Score**: 0.98

4. Insights from the Predictions

- **Recall Maximization**: The primary focus of this problem was to ensure that all instances of the minority class (label '1') were correctly identified, which was achieved by setting a threshold of 0.001. This resulted in a **recall of 1**, meaning no cases of the minority class were missed.
- **Trade-off Between Precision and Recall**: While the recall was maximized, precision dropped to 73.2%. This means that while we correctly identified all instances of the minority class, there were also some false positives where majority class instances (label '0') were incorrectly classified as the minority class. This trade-off is acceptable given the objective of minimizing false negatives.
- **High Accuracy**: Despite the lower precision, the overall accuracy of the model is still very high at **96.56%**, indicating that the model is making correct predictions in a large majority of cases.
- **AUC-ROC**: The ROC AUC score of **0.98** indicates that the model has an excellent ability to distinguish between the two classes. The ROC curve was plotted to visualize the trade-off between the true positive rate (Recall) and the false positive rate.