GST Analytics Hackathon – Model Performance Evaluation Report

---

Introduction:

This report evaluates the performance of several machine learning models using a variety of metrics, including Accuracy, Precision, Recall, F1 Score, AUC-ROC, Log Loss, and Balanced Accuracy. The models being evaluated include **XGBoost**, **LightGBM**, and ensemble classifiers such as the **VotingClassifier** and **StackingClassifier**. The goal is to assess their effectiveness in handling the dataset, with a focus on predictive accuracy, handling class imbalance, and overall robustness.

(Note: The numbers quoted in this report may vary slightly due to the non-deterministic nature of the predictive algorithms)

Models Evaluated:

1. XGBoost (eXtreme Gradient Boosting)
2. LightGBM (Light Gradient Boosting Machine)
3. VotingClassifier (Ensemble of Models)
4. StackingClassifier (Meta-Learner Ensemble)

Evaluation Metrics:

1. **Accuracy**: The proportion of correctly classified samples to the total number of samples.
2. **Precision**: The proportion of true positives among the predicted positives (a measure of the exactness of the model).
3. **Recall**: The proportion of true positives among all actual positives (a measure of the model's completeness).
4. **F1 Score**: The harmonic mean of Precision and Recall, balancing the two metrics.
5. **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** A performance measurement for classification models at various threshold settings.
6. **Log Loss**: A measure of how far predicted probabilities are from the actual binary labels.

7. **Balanced Accuracy**: The average of recall obtained on each class, useful for imbalanced datasets.

Model Performance Analysis:

1. **XGBoost:**

- **Before optimizing (vanilla model on validation data):**

Accuracy	0.977857
Precision	0.849605
Recall	0.929763
F1 Score	0.887878
AUC-ROC	0.956314
Log-Loss	0.798103
Balanced Accuracy	0.956314

- **After upsampling + optimizing (optimized model on validation data):**

Accuracy	0.977297
Precision	0.831388
Recall	0.952387
F1 Score	0.887784
AUC-ROC	0.966139
Log-Loss	0.818303
Balanced Accuracy	0.966139

- **Final performance on test data**

Accuracy	0.977701
Precision	0.844134
Recall	0.936421
F1 Score	0.887886
AUC-ROC	0.994422
Log-Loss	0.803749
Balanced Accuracy	0.959210

Analysis: XGBoost showed robust performance with a strong AUC-ROC score of ~0.994, indicating excellent ability to differentiate between classes. Its Precision and Recall were relatively balanced, resulting in a solid F1 Score of ~88.79%. Although its Log Loss was slightly higher, the model maintained competitive accuracy across both classes, indicating that it is a good fit for the dataset.

2. LightGBM:

- **Before optimizing (vanilla model on validation data):**

Accuracy	0.972068
Precision	0.772872
Recall	0.996691
F1 Score	0.870627
AUC-ROC	0.983098
Log-Loss	1.006753
Balanced Accuracy	0.983098

- **After upsampling + optimizing (optimized model on validation data):**

Accuracy	0.977469
Precision	0.838469
Recall	0.942662
F1 Score	0.887518
AUC-ROC	0.961877
Log-Loss	0.812105
Balanced Accuracy	0.961877

- **Final performance on test data:**

Accuracy	0.977873
Precision	0.838829
Recall	0.947362
F1 Score	0.889798
AUC-ROC	0.994683
Log-Loss	0.797551
Balanced Accuracy	0.964206

Analysis: LightGBM performed slightly better than XGBoost, with a F1 Score of ~88.98% and an AUC-ROC score of ~0.995. LightGBM's strength lies in its efficiency in handling large datasets. Precision was marginally lower than XGBoost, indicating it caught fewer positive cases as compared to XGBoost. Recall was strong, indicating when the model predicted positive, it was often correct. Log Loss was marginally lower, reflecting slightly more confidence in probability predictions.

3. **VotingClassifier:**

- **After upsampling + optimizing (optimized model on validation data):**

Accuracy	0.978252
Precision	0.850351
Recall	0.933680
F1 Score	0.890069
AUC-ROC	0.958286
Log-Loss	0.783872
Balanced Accuracy	0.958286

- **Final performance on test data:**

Accuracy	0.978568
Precision	0.850263
Recall	0.937880
F1 Score	0.891925
AUC-ROC	0.994991
Log-Loss	0.772486
Balanced Accuracy	0.960342

Analysis: The VotingClassifier was the top performer out of all the models, achieving highest scores in all key metrics (except recall). It outperformed both XGBoost and LightGBM as it combined the predictions of both the models. It achieved the highest Accuracy (~97.9%) and F1 Score (~89.2%), demonstrating that the ensemble model benefits from leveraging the strengths of different classifiers. The AUC-ROC score of ~0.995 also suggests it has superior ability to distinguish between classes compared to the individual models. Log Loss was the

lowest among the models, meaning the VotingClassifier provided the most accurate probability estimates.

4. **StackingClassifier:**

- **After upsampling + optimizing (optimized model on validation data):**

Accuracy	0.975456
Precision	0.846855
Recall	0.903019
F1 Score	0.874036
AUC-ROC	0.943009
Log-Loss	0.884639
Balanced Accuracy	0.943009

- **Final performance on test data:**

Accuracy	0.975958
Precision	0.848907
Recall	0.906354
F1 Score	0.876690
AUC-ROC	0.993055
Log-Loss	0.866551
Balanced Accuracy	0.944779

Analysis: The StackingClassifier was a close runner up behind VotingClassifier, achieving the high scores in all key metrics, including Accuracy (~97.6%), F1 Score (~87.7%), and AUC-ROC (~0.993). This indicates that the stacking approach—where multiple models are combined, and a meta-learner makes the final predictions—can prove to be a good choice for the problem at hand.

Below are the insights and analysis derived from the model's predictions:

1. **Performance of Imputation Strategies:**
    - The **IterativeImputer** proved effective in handling missing data, particularly when combined with the optimal number of neighbors. The F1 score improved when

using ~5 neighbors for imputation, indicating that moderate imputation yielded the best balance between predictive accuracy and model generalizability. Too few or too many neighbors resulted in either insufficient information or overfitting, respectively.

2. **Base Model Performance:**
    - **LightGBM** consistently outperformed XGBoost across key performance metrics, achieving the **highest scores in all metrics** except for Precision. This suggests that LightGBM was able to handle the complexity of the data, including imputed values and class imbalance, more effectively than the other models.
    - **XGBoost** also demonstrated strong performance but fell very slightly short of LightGBM in terms of overall predictive power.
3. **Effect of Scaling and Class Imbalance Handling:**
    - **Scaling the input data** had mixed results. While scaling generally benefits gradient-boosting models by improving their ability to converge during training, the improvement in metrics was not substantial. This suggests that, while important, scaling was not a critical factor for this scenario.
    - The use of **SMOTE** for addressing class imbalance had a significant positive impact. The models' Precision and Recall improved notably after applying SMOTE, particularly for the minority class, **reducing the false negative rate**. This demonstrates the importance of balancing the dataset to improve predictive performance, particularly in datasets with significant class imbalance.
4. **Ensemble Models:**
    - The **VotingClassifier** provided the best overall performance, as it combined the outputs of the base models and leveraged a voting system for final predictions. This model exhibited the highest F1 score and balanced the trade-offs between Precision and Recall better than the individual models. It also gave the **lowest Log-Loss**.
    - The **StackingClassifier** was effective but did not outperform the **VotingClassifier**, because the meta-estimator's hyperparameters were not trained due to a lack of available computational power.

5. **Calibration and Model Confidence:**
    - The **calibration plot** indicated that the **StackingClassifier** and **XGBoost** models had the most well-calibrated predictions, with predicted probabilities aligning closely with actual outcomes. This means these models not only made accurate predictions but were also confident in their predictions, which is crucial for high-stakes decision-making scenarios. **VotingClassifier** performed well across various metrics, but it is trailing closely behind XGBoost in the calibration plot.

Key Insights:

- **LightGBM** was the best individual model, demonstrating superior performance across all metrics except in **Precision**.
- **XGBoost** performed marginally better than LightGBM in **Precision** (~0.01%).
- **VotingClassifier** outperformed all other models by leveraging the strengths of multiple base classifiers, making it the most reliable option.
- **VotingClassifier** also demonstrated the lowest for **Log-Loss**.
- **SMOTE** significantly improved the model's ability to predict the minority class, reducing bias and improving generalization.

These insights highlight the effectiveness of both the chosen models and preprocessing techniques, resulting in high-performance predictions and model confidence across different metrics.