Comparative Analysis of Algorithms on the Training Dataset of Data Science Challenge organized as a part of the ML4Crypto 2024 workshop

Name: Soumyadeep Nag
Date: 14th November 2024

Introduction

This document summarizes the comparative performance of the three classification algorithms used—RandomForest, XGBoost, and GradientBoosting—on the given bitstream dataset(TrainingData.csv).Here, Cross-validation accuracy and validation accuracy were used to identify the best-performing algorithm.

Methodology

Data Preparation

The training dataset includes sequences of bitstreams with class labels. Here I did preprocessing for converting the bitstreams to integer sequences and then padding them for uniformity. Cross-validation was also conducted using 5 folds to assess the model's robustness.

Algorithms Evaluated

1. **RandomForest Classifier**: An ensemble learning technique which leverages multiple decision trees.
2. **XGBoost Classifier**: A gradient-boosting classification model which is optimized for classification and regression.
3. **GradientBoosting Classifier**: A boosting model which works by sequentially building an ensemble of decision trees, where each tree is made to correct the errors of the previous ones, thus by improving the overall prediction accuracy.

Results

Algorithm	Cross-Validation Accuracy (Mean)	Validation Accuracy
RandomForest	0.50	**0.52**
XGBoost	0.50	0.50
GradientBoosting	0.50	0.50

The **Cross Validation scores** for different models are:

1. RandomForest: [0.4925 0.4775 0.5225 0.5125 0.4950]
2. XGBoost: [0.5075 0.4975 0.4900 0.4600 0.5325]
3. GradientBoosting: [0.5125 0.5038 0.4819 0.5063 0.4969]

Discussion

Despite getting identical cross-validation accuracy across the different models, the validation accuracy differs, with RandomForest Classifier achieving the highest at 0.52. This suggests that RandomForest is generalizing slightly better on unseen data as compared to the other two i.e. XGBoost (0.50) and GradientBoosting (0.50).

Best Performing Algorithm: RandomForest

Given the higher validation accuracy of the RandomForest Classifier, I selected the RandomForest Classifier as the optimal algorithm.

Conclusion

RandomForest has demonstrated the best performance on the validation dataset. So according to me, it is the most suitable model.