DATA 622

Assignment 2

Erick Hadi

The objective of the Portuguese Bank Marketing Dataset is to be able to predict whether client would subscribe to a term deposit. Since the target is a binary, this problem falls under supervised binary classification. From exploratory data analysis there is a mixture of categorical and numerical variable, non-linear feature interactions, and class imbalance. These properties make tree-based ensemble methods particularly effective. Four experiments were performed: a Decision Tree baseline and tuned, Random Forest baseline and tuned, AdaBoost model baseline and tuned, and a baseline XGBoost model. The models in this experiment were evaluated using five key metrics accuracy, precision, recall, F1-score, and AUC-ROC. These metrics are important for evaluating how well a model is performing. To ensure consistency all the models were trained on the same 80-20 split and hyperparameters tuning was conducted through GridSearchCV, optimizing for F1-score.

The decision tree baseline model provided an interpretable and highly flexible approach. However, as a single tree it had high variance, capturing noise from the training data which leads to overfitting. Tree depth and split were tuned to reduce tree complexity introducing bias to lower variance. This allowed an increase in f1-score from 0.51 to 0.61. The random forest model should reduce variance by averaging predictions from a group of trees. The baseline model had a f1 score of 0.58 and auc-roc score of 0.9487 which improved to f1 score of 0.59 and auc-roc of 0.9498. This indicates a slight improvement but not as good as a tuned decision forest. This may be because the model was not trained with deeper or more trees. However, during experimentation long model training times led to a cap to the parameters. AdaBoost takes a boosting approach and sequentially on misclassified examples. The model didn't perform as well as the others even with tuning. This may be because AdaBoost variance increases when there is noise and class imbalance in the data. Lastly the XGBoost model outperformed all of the models achieving the highest AUC 0.9548 and the best F1-score 0.6132. Overall Decision tree had higher variance, Random Forest achieved the lowest variance with moderate bias, AdaBoost had high variance and overfitting issues, and XGBoost had the best bias-variance tradeoff.

What all of this means is that for the Portuguese Bank Marketing data, XGBoost algorithms should be prioritized. They combine the strengths of decision trees and ensemble methods while addressing their weaknesses through regularization and efficient optimization. Also, proper hyperparameter tuning remains essential to achieving the best bias-variance tradeoff. The bank can integrate the XGBoost model in its marketing workflow to predict which clients are most likely to subscribe. Also, from the previous assignment of exploratory data analysis focusing on the most important features such as call duration and previous campaign outcome can help tailor communications strategies, improve timing, and personalize customer engagement for better business.