Balanced Random Forest Classifier Model Report

Summary

The Balanced Random Forest Classifier performed reasonably well on this task, with an accuracy of 0.9414577376397725 and an F1-score of 0.942217378487321.

Model Selection

We chose to use a Balanced Random Forest Classifier for this task. This model is an ensemble method that combines the predictions of several base estimators built with a given learning algorithm in order to improve generalizability and robustness over a single estimator. It also handles imbalanced classes, which is a common problem in many machine learning tasks.

Hyperparameter Tuning We used RandomizedSearchCV for hyperparameter tuning. This method performs a random search on hyperparameters, which is more efficient than an exhaustive search like GridSearchCV.

The hyperparameters we tuned were:

‘n_estimators’: The number of trees in the forest. ‘max_depth’: The maximum depth of the tree. ‘min_samples_split’: The minimum number of samples required to split a node. ‘min_samples_leaf’: The minimum number of samples required at a leaf node. ‘bootstrap’: Whether bootstrap samples are used when building trees.

{‘n_estimators’: [50, 100], ‘max_depth’: [None, 5, 10], ‘min_samples_split’: [2, 5], ‘min_samples_leaf’: [1, 2], ‘max_features’: [‘sqrt’]}

Model Performance

The best parameters found by RandomizedSearchCV were:

Best parameters:, {‘n_estimators’: 100, ‘min_samples_split’: 2, ‘min_samples_leaf’: 2, ‘max_features’: ‘sqrt’, ‘max_depth’: None}

With these parameters, the model achieved the following performance metrics: Best cross-validation score: 0.8843703246628051

Accuracy:, 0.9414577376397725 F1-score: 0.942217378487321

Area under Precision-Recall curve: 0.9429039222274846 Area under ROC curve: 0.9806846779110929

Testing Data

Classification report:

	precision	recall	f1-score	support
0	0.97	0.95	0.96	19595447
1	0.85	0.93	0.89	6606154
accuracy			0.94	26201601

macro avg 0.91 0.94 0.92 26201601 weighted avg 0.94 0.94 0.94 26201601

TRAINING DATA Classificatin Report-Confusion Matrix

Training confusion matrix:

[[2103489 73783] [14330 719687]]

Training classification report:

	precision	recall	f1-score	support
0	0.99	0.97	0.98	2177272
1	0.91	0.98	0.94	734017
accuracy			0.97	2911289

macro avg 0.95 0.97 0.96 2911289 weighted avg 0.97 0.97 0.97 2911289

This indicates that the model correctly classified 2103489 instances of class 0 and 719687 instances of class 1,

while misclassifying 73783 instances of class 0 and 14330 instances of class 1.

Area under Precision-Recall curve: 0.9429039222274846 Area under ROC curve: 0.9806846779110929

CV Results: mean_fit_time std_fit_time mean_score_time std_score_time param_n_estimators param_min_samples_split param_min_samples_leaf param_max_features … split0_test_roc_auc split1_test_roc_auc split2_test_roc_auc split3_test_roc_auc split4_test_roc_auc mean_test_roc_auc std_test_roc_auc rank_test_roc_auc 0 173.745271 10.170154 5.048314 0.655296 50 2 1 sqrt … 0.760727 0.762161 0.761910 0.760531 0.761677 0.761401 0.000652 7 1 260.899451 13.120900 7.605988 0.300571 50 2 1 sqrt … 0.842763 0.842145 0.845274 0.843697 0.843999 0.843576 0.001076 6 2 469.964428 15.121271 29.727104 1.066183 50 2 1 sqrt … 0.975202 0.974978 0.975231 0.974953 0.974771 0.975027 0.000171 3 3 248.468495 32.114539 6.705043 0.912500 50 5 1 sqrt … 0.841902 0.844507 0.843219 0.845103 0.843740 0.843694 0.001103 5 4 331.070193 24.053209 7.192338 0.868427 100 5 1 sqrt … 0.760034 0.761901 0.761297 0.760785 0.761598 0.761123 0.000657 8 5 310.626377 31.259569 7.395771 1.227617 100 2 1 sqrt … 0.760153 0.761901 0.761297 0.760610 0.761598 0.761112 0.000643 9 6 330.057937 13.014498 6.784553 0.034583 100 2 2 sqrt … 0.760034 0.761806 0.761296 0.760785 0.761598 0.761104 0.000636 10 7 886.614147 8.807764

51.353684 0.276577 100 2 1 sqrt … 0.976428 0.976381 0.976524 0.976255 0.976109 0.976339 0.000144 2 8 372.536677 9.873788 10.640266 0.069339 100 2 2 sqrt … 0.842954 0.845401 0.844054 0.844437 0.844295 0.844228 0.000785 4 9 688.714733 11.325166 43.841898 1.688386 100 2 2 sqrt … 0.979013 0.978918 0.979037 0.979130 0.978768 0.978973 0.000123 1

[10 rows x 42 columns]