Classification Performance Comparison
Decision Trees, Random Forest and Boosting Ensemble Methods

Christopher Griffith - @cdgphysics

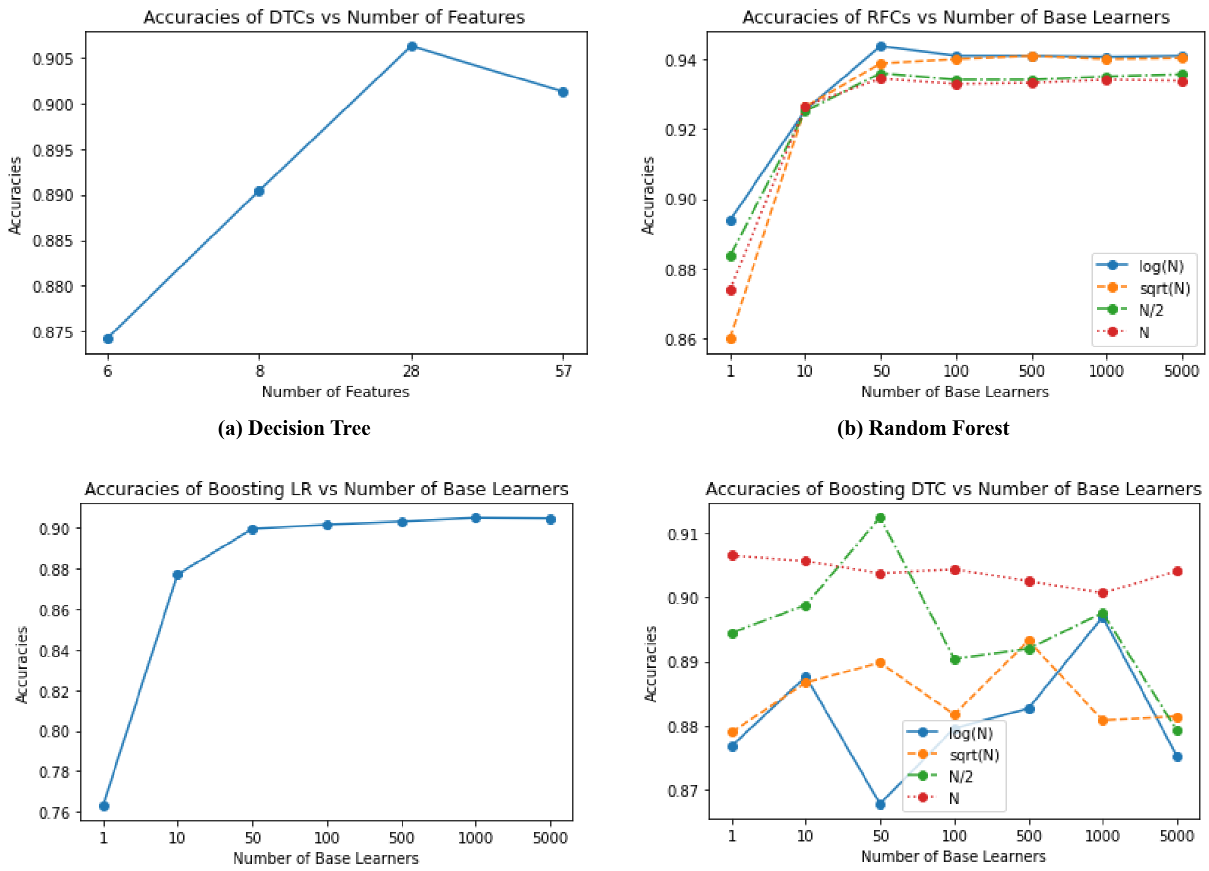
I. INTRODUCTION & MODEL PARAMETERS

This project paper will compare and contrast the performance metrics and statistical inferences of four classification models: *Decision Trees*, *Random Forest*, *Boosting Ensemble with Logistic Regression Classification* and *Boosting Ensemble with Decision Trees Classification*. An email messages data set of 4601 records with 57 attributes highlighting the frequency of specific keywords or characters was provided to evaluate the classification accuracies of the models to predict if a given email message was identified as spam or ham. All four models were trained using 30% of the dataset while 70% of the data was used to evaluate the performance of the trained models. Utilizing stratified random sampling, the class label distribution of the training/testing split datasets maintained the class label distribution of the original dataset (i.e. ~ 60.6 % ham and 39.4% spam). Coupled with the stratified sampling, a static random seed (i.e. random_state = 42) was standardized for the train/test split and for training the models. Therefore, all performance results of the models and dataset splits are both deterministic and reproducible.

The maximum number of features N = 57 is based on the 57 attributes that identifies an email message. Such attributes include the number of times the words like technology or conference appears in a given email message. The float values of these attributes range from zero to 15,841 which, if not normalized, could cause inadequate performance for the models. Nonetheless, the models were trained using these features without normalization.

II. CLASSIFICATION ACCURACIES & CONFUSION MATRICES REVIEW

Three of the four classification models were trained using (both or either) tunable number of feature parameters and number of base learners parameters for ensemble models. The number of features for each classification model, except for the Boosting Ensemble with Logistic Regression Classification, ranged between four parameters (smallest to largest): *log(N), sqrt(N), N/2 and N*. The Logistic Regression Classifier was tuned to reach a maximum of 10,000 iterations for convergence. Based on these parameters settings, Figure 1 illustrates the accuracies (i.e. # correctly classified/total # of records presented for classification) for each of the four models with the tunable appropriate parameters.



(a) Decision Tree

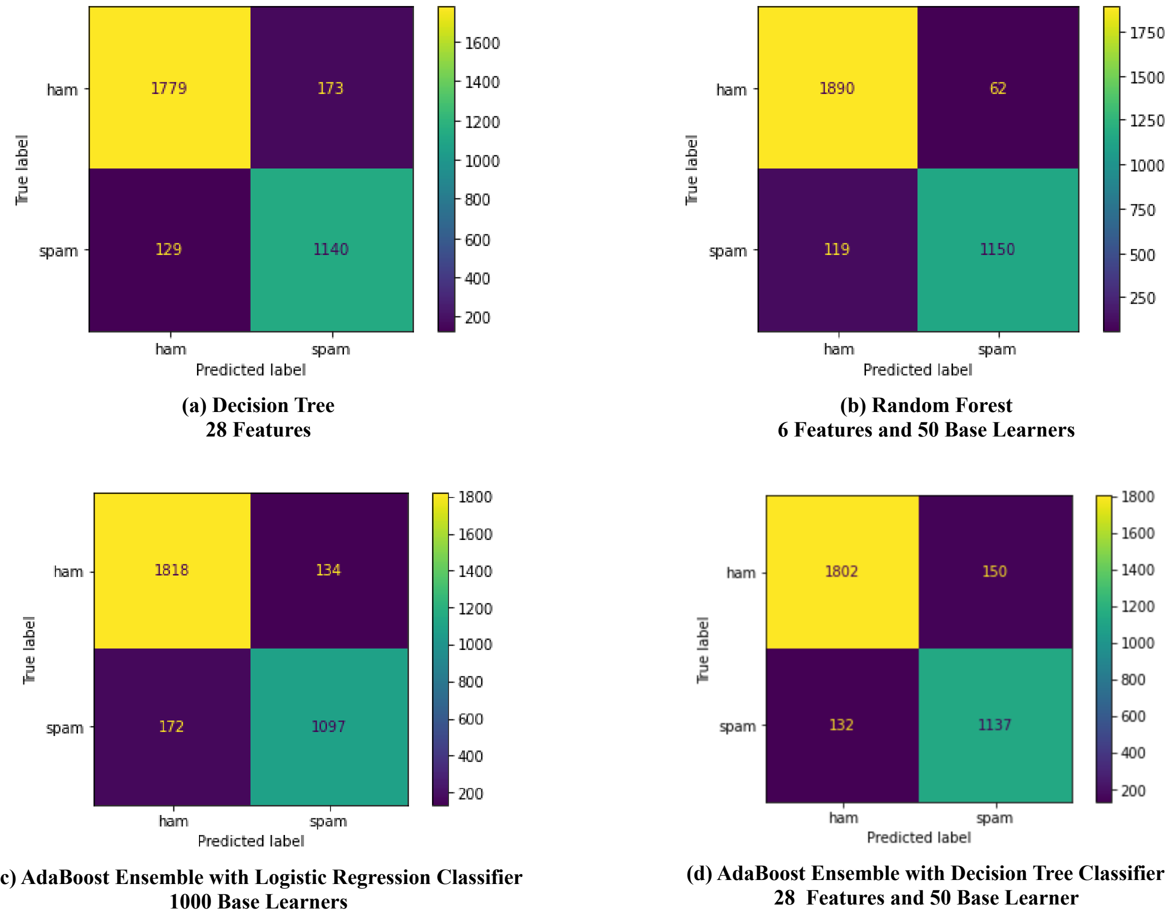
(b) Random Forest

(c) AdaBoost Ensemble with Logistic Regression Classifier

(d) AdaBoost Ensemble with Decision Tree Classifier

Figure 1: Classification Accuracies per Classifier based on Features and Base Learners Parameters

In Figure 1a, the Decision Tree models' peak accuracy was reached when utilizing only half of the maximum features (i.e. 28 features). The corresponding confusion matrix in Figure 2a provides clarity on how well the best performing Decision Tree model compares against the other three models. It is possible the Decision Tree Models could have performed better if the features were normalized before training. The remaining ensemble models fairly well despite the performance influence from the non-normalized features. As the number of base learners increased, both the Random Forest and the AdaBoost Ensemble with Logistic Regression Classifier Model increased in accuracy. The smallest amount of available features (i.e. 6 features) produced the maximum accuracy for the Random Forest model**.** The Random Forest and the Linear Regression AdaBoost models' accuracy curves appeared to peak and plateau relatively at specific base learner parameters. After reaching 50 to 100 base learners for Random Forest (Figure 1b) and 1000 base learners for the Linear Regression AdaBoost (Figure 1c), there appears to be no significant benefit in increasing base learners to increase classification accuracy. Conversely, this is not the case for the AdaBoost Ensemble with Decision Tree Classifier. Figure 1d clearly indicates that there is a significant difference in how parameters influence the classification accuracy. Similar to the Decision Tree model, the AdaBoost Ensemble with Decision Tree Classifier produced a maximum accuracy utilizing only half of the maximum features (i.e. 28 features). The Boosting Ensemble model was able to increase the accuracy of the Decision Tree Classifier by exploiting 50 base learners. The confusion matrices for the tree and ensemble models (Figure 2) examine how well the models performed per class accuracy while probing the probability distribution and confidence of the models' classification predictions. In the next section, we will investigate the confidence of the models and shed light on which model performed the best.



(a) Decision Tree
28 Features

(b) Random Forest
6 Features and 50 Base Learners

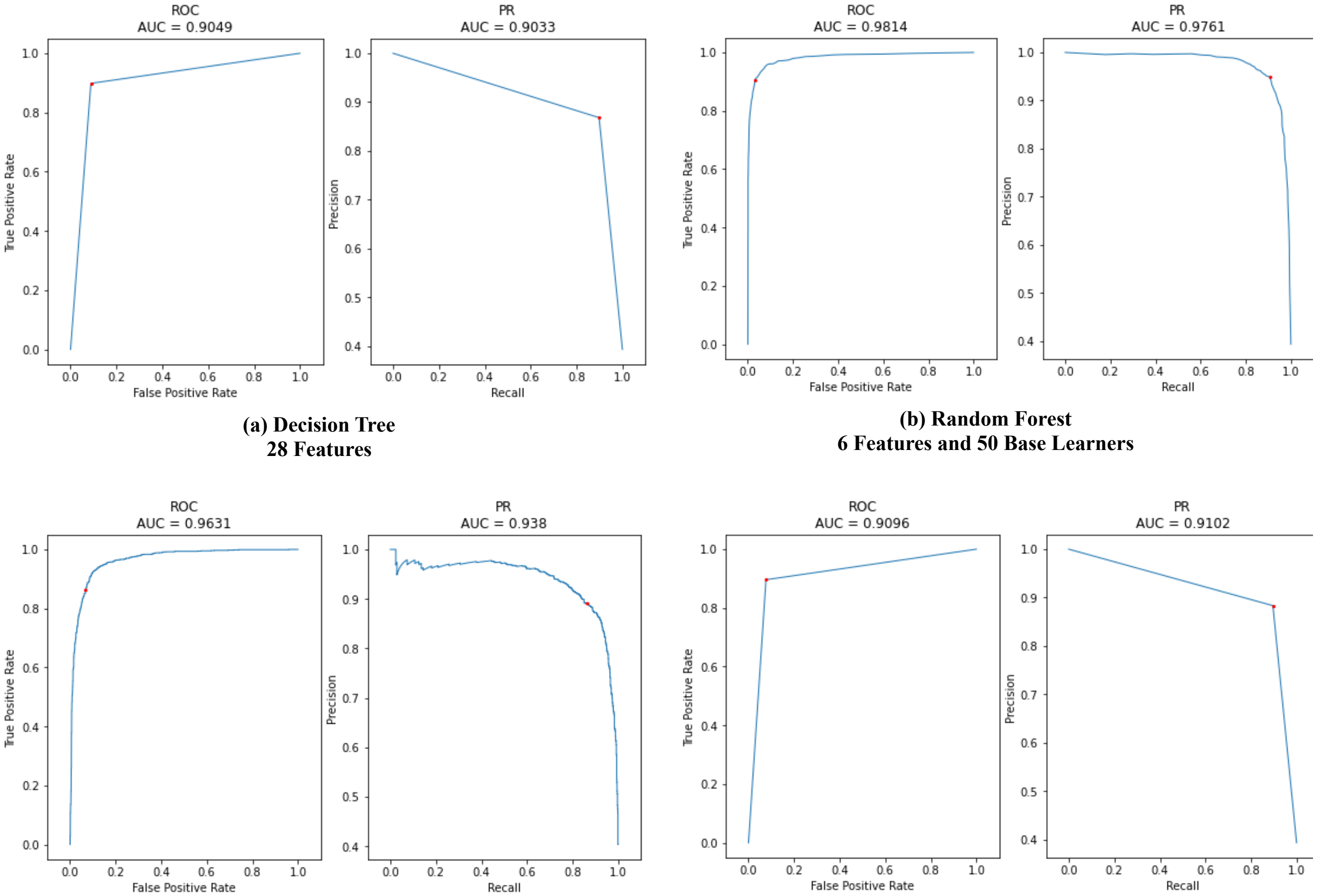
(c) AdaBoost Ensemble with Logistic Regression Classifier
1000 Base Learners

(d) AdaBoost Ensemble with Decision Tree Classifier
28 Features and 50 Base Learner

Figure 2: Confusion Matrix for Maximum Accuracy per Classifier

III. ANALYSIS OF ROC & PR CURVES FOR BEST PERFORMANCE METRICS

The Receiver Operating Characteristics (ROC) and the Precision-Recall (PR) curves for the best performance metrics are illustrated in Figure 3. The red dot on each curve represents the True Positive Rate: TP/(TP+FN), False Positive Rate: FP/(FP+TN), Precision: TP/(TP+FP) and Recall: TP/(TP+FN), all generated from the confusion matrices in Figure 2. Overall, each model reached an AUC of at least 0.9 for both ROC and PR curves. The highest AUC of all of the models for both curves was reached by the Random Forest Classification with 6 Features and 50 Base Learners. Table 1 summarizes the confidence and performance of the models and reinforces the evidence that the Random Forest Model with 6 Features and 50 Base Learners performed better than the rest of the models.



(a) Decision Tree
28 Features

(b) Random Forest
6 Features and 50 Base Learners

(c) AdaBoost Ensemble with Logistic Regression Classifier
1000 Base Learners

(d) AdaBoost Ensemble with Decision Tree Classifier
28 Features and 50 Base Learner

Figure 3: ROC and PR Curves of Maximum Accuracy Models per Classifier with Corresponding Performance Metrics (red plot point)

Classifier	**# of Features**	**# of Base Learners**	**Accuracy**	**True Positive Rate**	**True Negative Rate**	**Precision**
Decision Tree	$\frac{N}{2} \approx 28$	1	90.6240 %	89.8345 %	91.1373 %	86.8241 %
Random Forest	$log_2 N \approx 6$	**50**	**94.3806 %**	**90.6225 %**	**96.8238 %**	**94.8845 %**
AdaBoost Ensemble w/ Logistic Regression Classifier	$N = 57$	1000	90.4998 %	86.446 %	93.1352 %	89.1145 %
AdaBoost Ensemble w/ Decision Tree Classifier	$\frac{N}{2} \approx 28$	50	91.2450 %	89.5981 %	92.3156 %	88.345 %

Table 1: Maximum Accuracies with Performance Metrics per Classifier