Data 622 Machine Learning and Big Data_HW2

By: Jiaxin Zheng

Introduction:

The Bank Marketing dataset collect information from a marketing campaign of a Portuguese bank. The goal is to predict whether a client will subscribe to a term deposit. This is a binary classification analysis with both categorical variables (e.g., job, education, marital, housing loan) and numeric variables (e.g., age, balance, campaign, pdays, previous).

Missing and Duplicated Values:

There are a large number of "unknown" values in the categorical variables. In this experiment, I chose not to remove them. Instead, I treated "unknown" as a separate category, as it might contain useful information or patterns related to the target variable. Additionally, both tree-based models and AdaBoost can effectively handle categorical variables, making this approach reasonable for this dataset.

Decision Tree:

1. **Decision Tree Model 1:**
   In the first Decision Tree model, I used the default settings of the Decision Tree algorithm. This baseline model was easy to interpret but demonstrated limited recall for the minority class ("yes"). Although it successfully captured key patterns in the data, it also showed signs of overfitting. The model achieved an accuracy of **0.8978**, precision of **0.6173**, recall of **0.3311**, F1-score of **0.431**, and an AUC of **0.7227**.

2. **Decision Tree Model 2:**
   In the second Decision Tree model, I applied 5-fold cross-validation and tuned the complexity parameter (cp) to control pruning and test the model's generalization ability. Smaller cp values (e.g., 0.000–0.005) allowed the tree to grow deeper, testing for overfitting, while larger cp values (up to 0.05) pruned the tree more aggressively, testing for underfitting. The goal was to find a balance that improved variance and overall performance. However, the results remained the same as Model 1, with an accuracy of **0.8978**, precision of **0.6173**, recall of **0.3311**, F1-score of **0.431**, and AUC of **0.7227**, indicating no significant improvement.

Random Forest:

1. **Random Forest Model 1:**
   The baseline Random Forest model, built with 100 trees, outperformed the Decision Tree models. It achieved an accuracy of **0.9066**, precision of **0.6330**, recall of **0.4797**, F1-score of **0.5457**, and an AUC of **0.9276**.

   In **Model 2**, I increased the number of trees to 500, expecting a potential improvement in performance. However, the results remained identical to Model 1, suggesting that increasing the number of trees beyond 100 did not provide additional benefits for this dataset.

2. **Random Forest Model 3:**
   In this model, I kept the number of trees at 500 but tuned the mtry parameter, which determines the number of features randomly selected at each split. Adjusting mtry helps control model randomness and bias-variance trade-off—smaller values can reduce correlation and variance but increase bias, while larger values can do the opposite.
   This tuning resulted in an accuracy of **0.9055**, precision of **0.6293**, recall of **0.4674**, F1-score of **0.5364**, and an AUC of **0.9276**. The adjustment of mtry helped produce a more stable model with a balanced trade-off between precision and recall.

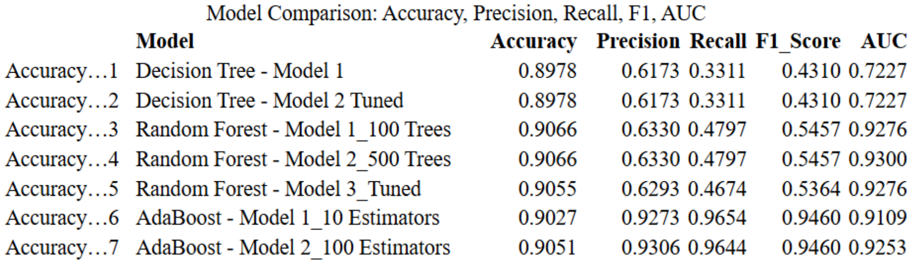
AdaBoost:

The AdaBoost models prioritized correcting misclassified observations, resulting in strong recall and F1-scores. Performance improved notably after increasing the number of boosting iterations (mfinal = 100) and allowing deeper base learners.

Conclusion:

Among all tested algorithms, **AdaBoost Model 2** with 100 estimators delivered the best overall performance, striking an excellent balance between precision, recall, and AUC.

Random Forest models were a strong second choice.



Model Comparison: Accuracy, Precision, Recall, F1, AUC

	Model	Accuracy	Precision	Recall	F1_Score	AUC
Accuracy…1	Decision Tree - Model 1	0.8978	0.6173	0.3311	0.4310	0.7227
Accuracy…2	Decision Tree - Model 2 Tuned	0.8978	0.6173	0.3311	0.4310	0.7227
Accuracy…3	Random Forest - Model 1_100 Trees	0.9066	0.6330	0.4797	0.5457	0.9276
Accuracy…4	Random Forest - Model 2_500 Trees	0.9066	0.6330	0.4797	0.5457	0.9300
Accuracy…5	Random Forest - Model 3_Tuned	0.9055	0.6293	0.4674	0.5364	0.9276
Accuracy…6	AdaBoost - Model 1_10 Estimators	0.9027	0.9273	0.9654	0.9460	0.9109
Accuracy…7	AdaBoost - Model 2_100 Estimators	0.9051	0.9306	0.9644	0.9460	0.9253