Predicting school closings from the DOE College Scorecard: Model Metrics

Kirsten Regier

1 Decision Tree Model

1.1 Hyperparameters

The hyperparameters for depth and node purity method were determined by a grid search, as shown in Figure 1. The optimal depth is 22 and node purity method is gini.

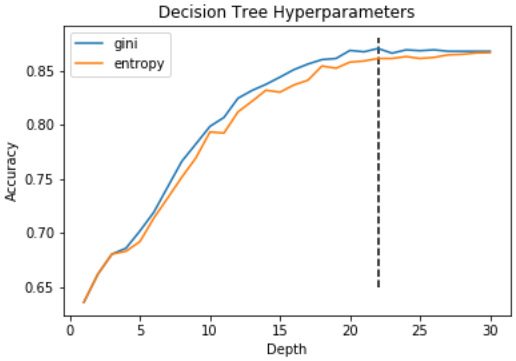

Figure 1: Grid search results for **depth** and **node purity** method hyperparameters.

1.2 Model parameters

DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion=’gini’, max_depth=22, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=’deprecated’, random_state=21, splitter=’best’)

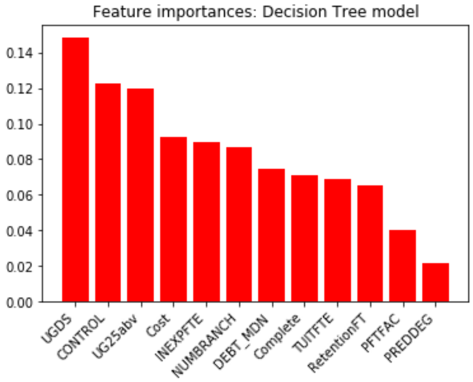

Figure 2: Feature importance levels based on the Decision Tree model.

1.3 Model evaluation

Table 1: Decision Tree - Confusion matrix and Evaluation metrics

Currently operating		Predicted		Recall
		No	Yes	
Actual	No	1533	123	0.93
	Yes	380	1852	0.86
Precision		0.80	0.94	
Accuracy				0.87

2 AdaBoost Model

2.1 Hyperparameters

The hyperparameter for number of estimators was determined by a grid search using the previous Decision Tree model as a base (orange) and without providing a base model (blue), as shown in Figure 3. The optimal number of estimators is 122.

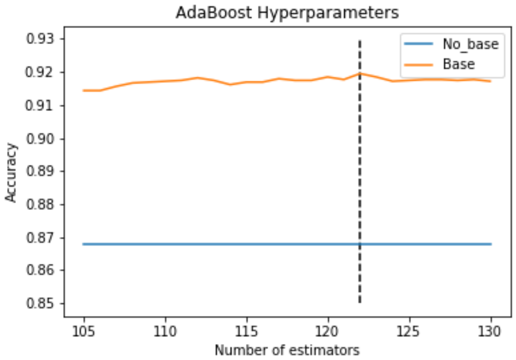

Figure 3: Grid search results for **number of estimators** hyperparameter.

2.2 Model parameters

AdaBoostClassifier(algorithm=’SAMME.R’, base_estimator=DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion=’gini’, max_depth=22, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=’deprecated’, random_state=21, splitter=’best’), learning_rate=1.0, n_estimators=122, random_state=21)

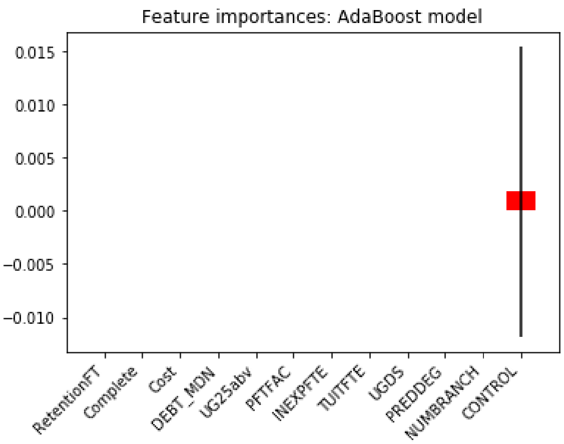

Figure 4: Feature importance levels based on the AdaBoost model.

2.3 Model evaluation

Table 2: AdaBoost model - Confusion matrices and Evaluation metrics

Table 3: Resampled data

Currently operating		Predicted		Recall
		No	Yes	
Actual	No	1538	118	0.93
	Yes	195	2037	0.91
Precision		0.89	0.95	
Accuracy				0.92

Table 4: Without duplicate data

Currently operating		Predicted		Recall
		No	Yes	
Actual	No	1164	93	0.93
	Yes	195	2037	.91
Precision		.86	.96	
Accuracy				.92

2.4 Threshold calculation

Various thresholds for assigning schools to currently operating or not were tried for both the final Decision Tree model and the final AdaBoost model, as shown in Figure 5. Initial models on the unbalanced dataset and the Decision Tree model showed that adjusting the threshold could improve model performance. However, after resampling the minority class to balance the data, and the using the AdaBoost model which focusses on misclassified data, improved the model performance to the point where the 50% threshold produced the best scores for both accuracy and balanced accuracy.

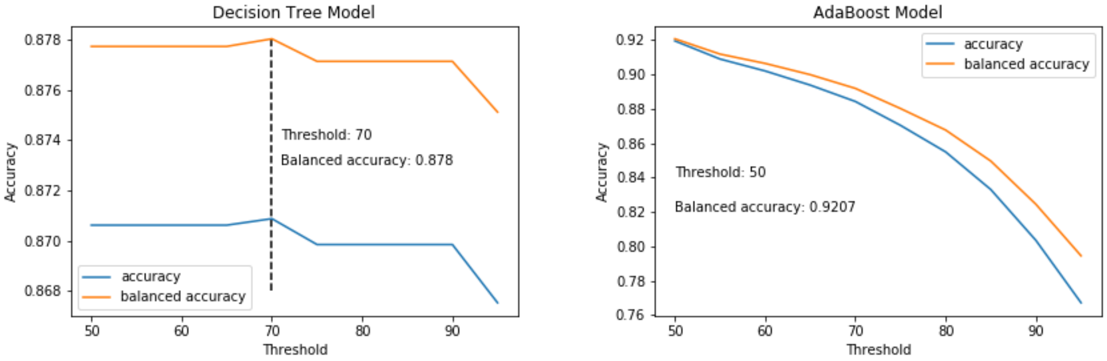

Figure 5: Accuracy and balanced accuracy scores at various decision thresholds for the Decision Tree model (left) and AdaBoost model (right).