Student Intervention judgement

Kosuke Fukui

1. Classification or Regression

In this problem, we'd like to know whether student pass the final exam or not. Therefore, I use classification method. Classification can be applied to the problem of which the answer belongs to particular class like "agree or disagree", "A team or B team or C team" and "pass the exam or not".

2. Evaluating Model Performance

Total number of students: 395

Number of features: 30

Number of students who passed: 265

Number of students who failed: 130

Graduation rate of the class: 0.67%

3. Preparing the data

I used *train_test_split* from *sklearn* for random data assignment for training and test dataset. I adopted *label_binarize* from *sklearn* to binarize the target column from "yes and no" to "0 and 1".

4. Training and Evaluating Models

(1) Decision Tree

Select Reason: The structure is simple and easy to implement.

Application: Buyer behavior analysis.

Strengths: Easy to understand the result and get a suggestion from classification process.

Weaknesses: Low predictability.

Decision Tree	Training data size		
	100	200	300
Training time	0.001	0.001	0.003
Prediction time	0.000	0.000	0.001
F1 for training data	1.000	1.000	1.000
F1 for test data	0.729	0.708	0.695

(2) Random Forest

Select Reason: The structure is simple and calculation speed is high.

Application: This method used for image recognition on medical fields.

Strengths: Training time is short even on big data size.

Weaknesses: The model tends to be complicated because of the number of variables.

Random Forest	Training data size		
	100	200	300
Training time	0.031	0.025	0.025
Prediction time	0.001	0.002	0.001
F1 for training data	0.992	1.000	0.993
F1 for test data	0.782	0.797	0.761

(3) Support Vector Machine

Select Reason: It is known that this method shows high predictability on various problems.

Application: This method often used for text processing like detecting spam mail.

Strengths: High predictability because of high generalization ability.

Weaknesses: Training time is long on big data size.

SVM	Training data size		
	100	200	300
Training time	0.024	0.004	0.027
Prediction time	0.002	0.003	0.006
F1 for training data	0.859	0.869	0.869
F1 for test data	0.784	0.776	0.759

5. Choosing the Best Model

I chose Support Vector Machine as the best model. Decision Tree is simple but that shows low F1 score and it doesn’t have plenty tuning potential. Random Forest shows good F1 score but it tends to take more training time than SVM.

SVM is a process for finding the optimal separating hyperplane which identify the classes. The hyperplane has the largest margin from data points.

In prediction process, the new data classified based on the hyperplane, on which side the data point is.

The final F1 score applied GridSearchCV is as followings:

Best parameter choice: {'kernel': 'rbf', 'C': 100, 'gamma': 0.0001}

Best F1 score: 0.824511363636