Tree Based Methods

Eun Soo Choi

Outline

- Base model: Logistic Regression
- Basic vs. Hyperparameter tuned Tree based methods
    1. Decision Tree
    2. Random Forest
    3. Adaptive Boosting
    4. Gradient Boosting

Results

Basic

Decision Tree

max_depth = 5

	precision	recall	f1-score
No	0.86	0.79	0.83
Yes	0.39	0.51	0.45
accuracy			0.73

Random Forest

bootstrap=True
max_depth=7
max_features=5
n_estimators=200

	precision	recall	f1-score
No	0.86	0.89	0.88
Yes	0.53	0.47	0.50
accuracy			0.80

Adaptive Boosting

	precision	recall	f1-score
No	0.88	0.90	0.89
Yes	0.60	0.54	0.57
accuracy			0.83

Gradient Boosting

max_depth=3

	precision	recall	f1-score
No	0.87	0.90	0.89
Yes	0.57	0.50	0.54
accuracy			0.82

Hyperparameter Tuning

Decision Tree

	precision	recall	f1-score
No	0.86	0.89	0.87
Yes	0.52	0.44	0.47
accuracy			0.80

Random Forest

	precision	recall	f1-score
No	0.87	0.92	0.89
Yes	0.61	0.49	0.54
accuracy			0.83

Logistic Regression

	precision	recall	f1-score
No	0.87	0.91	0.89
Yes	0.60	0.50	0.54
accuracy			0.83

Imbalanced Classification

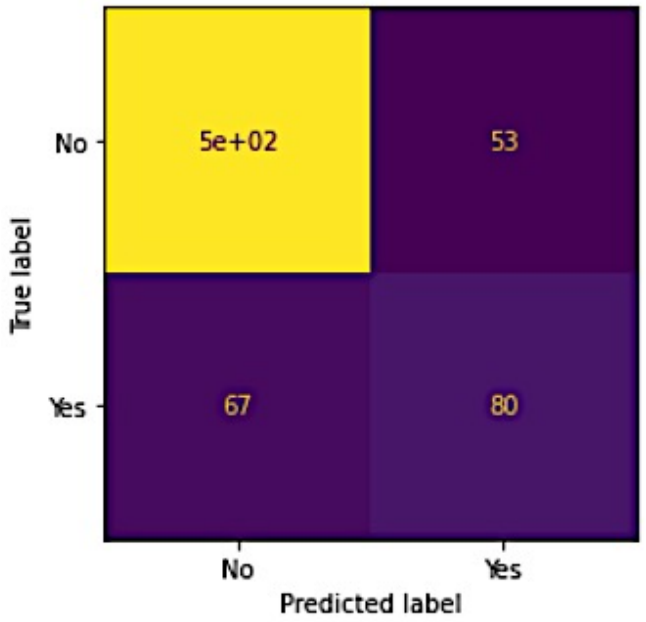
Eun Soo Choi, Ph.D.

Outline

- Classification metrics
- Imbalanced classification cases
- Challenges of imbalanced classification

Classification Metrics

	precision	recall	f1-score
No	0.88	0.90	0.89
Yes	0.60	0.54	0.57
accuracy			0.83



		Predicted label			
		Negative 0 H_0' Reject	**Positive** 1 H_1' Accept	**Recall**	
True label	**Negative** 0 H_0 Null Hypothesis	TN	FP Type 1 error	Specificity 1-α: confidence level TN/ (TN + FP)	Type 1 error rate (α) FP/ (TN + FP)
	Positive 1 H_1 Alternative Hypothesis	FN Type 2 error (Bad!!)	TP	Sensitivity 1-β: power of test TP/ (FN + TP)	Type 2 error rate (β) FN/ (FN + TP)
	Prevalence	**Precision**		**F1-score**	**Accuracy**
	(TP + FN)/ (TP + FN + TN + FP)	NPV(Precision) TN/ (TN + FN)	PPV TP/ (FP + TP)	2*Precision*Recall/ (Precision + Recall)	(TP + TN) /(TP+FN+FP+TN)

Imbalanced classification cases

Small portion of

- Medical conditions
- Credit card fraud
- Churn(Cancellation of service)

…

Challenges in Imbalance Classification

➢ Resampling: Undersampling from major class + Oversampling from minor class

1. Model parameters: Learned by major class (neglect small portion of minor class)

➢ Weighted loss: n_samples / (n_classes * np.bincount(y))
   ex) weight_no = 704/(2*557), weight_yes = 704/(2*147)

2. Model performance evaluation:
   - Accuracy is not an appropriate metric because it depends on Prevalence.

   ex1) 704 = 557 vs. 147 (79% vs. 21%)
      Predict all as No(0) = 557 vs. 147, Accuracy = 557/704 (79%), Recall = 0/127 = 0

   ex2) 10,000 = 9,900 vs. 100 (99% vs. 1%)
      Predict all as 0 = 10,000 vs. 0, Accuracy = 9,900/10,000 (99%), Recall = 0/100 = 0

   - Recalls does not depend on Prevalence
   - Diagnostically, Precisions are more helpful.

➢ F1-score(Precision, Recall), Precision-Recall Curve, AUC ROC (using different thresholds)

Base model: Logistic Regression

	Best without class_weight				Best with class_weight		
Logistic Regression	precision	recall	f1-score		precision	recall	f1-score
No	0.87	0.91	0.89	No	0.91	0.74	0.82
Yes	0.60	0.50	0.54	Yes	0.43	0.73	0.54
accuracy			0.83	accuracy			0.74

Tree based methods

	Best without class_weight				Best with class_weight			
Decision Tree		precision	recall	f1-score		precision	recall	f1-score
max_depth = 5	No	0.86	0.89	0.87	0	0.91	0.72	0.81
	Yes	0.52	0.44	0.47	1	0.41	0.74	0.53
	accuracy			0.80	accuracy			0.73
Random Forest bootstrap=True max_depth=7 max_features=5 n_estimators=200		precision	recall	f1-score		precision	recall	f1-score
	No	0.87	0.92	0.89	0	0.91	0.77	0.83
	Yes	0.61	0.49	0.54	1	0.45	0.73	0.56
	accuracy			0.83	accuracy			0.76
Adaptive Boosting		precision	recall	f1-score		precision	recall	f1-score
	No	0.88	0.90	0.89	0	0.92	0.72	0.81
	Yes	0.60	0.54	0.57	1	0.42	0.77	0.55
	accuracy			0.83	accuracy			0.73
Gradient Boosting		precision	recall	f1-score				
	No	0.87	0.90	0.89				
	Yes	0.57	0.50	0.54				
	accuracy			0.82				