Dietinic Predective Modeling 2023. 07. 14 (tue)

: 머신러닝을 하는 이유는? 자동이라서.
자동화된 process.

알고리즘은 너무 많아.
패키지도 잘 되어있어서 쉽게할 수 있어.

>>
- 좋은 데이터 훈련 (training)
- 좋은 데이터 평가셋 (validation)
- 좋은 데이터 홀드셋 (test)

>>
random_state = 42, 2019, 0 ..
stratify. ['survived'] 에서 0과 1이 test와 train에 고루
분포되게...
~~반드시~~ 필요함. 신뢰성을 위해서.

>> 알고리즘을 공부해야하는 이유는 파라미터의 존재에 대해서 알아야 하기 때문.
파라미터가 어떻게되느냐에 따라 성능이 좌우됨.

피처들을 관리하자.

>> threshold를 0.5로 잡는 것이 아니라 직접 찾아서 넣으면 성능을 더
올릴수 있다.

KNN → K-Nearest Neighbours (KNN)

Neighbours 수가 바뀌면 정확도도 달라진다.

Random forest : 앙상블.. 약한 친구들이 합쳐진.
여러개를 합쳐서 더 강해진다.
앙상블이나 부스팅을 쓰고 그것들로 합친다.. 다양성을 합친다.

Date. No.

[redacted]에서는 다양한 모델이 필요하다.
교차 cost 때문에 앙상블을 많이 한다.
Kaggle에서는 Score 때문에 많이 함.
모델을 몇100개씩 Stacking..

» 훈련 및 테스트 데이터가 변경되면 정확도로 변경된다.
정확도가 증가하거나 감소할 수 있다. 이를 극복하기 위해 교차검증을 사용한다.

christmas Arigatogozaimas

Crossdresser
Yeongsonder

교차검증의 장점은 데이터가 바뀌어도 일반화 운감한가?

Acc1	Acc2	Acc3	Acc4	Acc5
model1	model2	model3	model4	model5
Learn	Learn	Learn	Learn	Predict
Learn	Learn	Learn	Predict	Learn
Learn	Learn	Predict	Learn	Learn
Learn	Predict	Learn	Learn	Learn
Predict	Learn	Learn	Learn	Learn

(acc1 + acc2 + acc3 + acc4 + acc5) / 5

mean_acc = ?

Std_acc = ?

Date. No.

Y는 target값 Survived.

⇒ Confusion Matrix