**Studygroup\_A조 (김은총, 박은총, 백건호, 안제인, 구현우)**

1. Classification과 Regression의 차이점은 무엇인가

* 출력 변수(목적변수)가 이산형 : Classification
* 출력 변수(목적변수)가 연속형 : Regression

1. Train Data와 Test Data를 따로 구분하는 이유는?

* Overfitting을 방지하기 위함이다.
* Train data로 학습된 모델을 평가할 때 이미 알고있는 데이터로 평가를 했을 때는 성능이 좋게 나올 수 밖에 없다. 따라서 한번도 보지 못한 데이터(test data)를 구분해서 모델이 일반적인 데이터에서도 높은 성능을 갖는지 확인을 해봐야 한다.

1. 추정 확률로부터 실제 classification을 하는 방법은?

* Threshold를 정해 추정확률이 더 크면 T , 더 작으면 F 로 구분한다.

1. Accuracy는 무엇인가?

* 전체 중 맞춘 것의 비율. (모델이 샘플을 정확히 분류한 비율)
* (TP + TN) / Total

1. Overfitting은 무엇인가?

* 모델이 학습 데이터에는 높은 정확도를 나타내지만, 테스트 데이터나 일반 데이터에서는 성능이 떨어지는 것.
* 학습 샘플의 수가 적을 때, 너무 복잡한 모델일 때 등 발생한다.

1. Accuracy가 충분하지 않은 상황에 대해서 설명하라

* 데이터가 편향되어 있을 때는 Accuracy만 갖고 모델을 판단할 수 없다.

1. Precision과 Recall은 무엇인가?
2. Precision : 모델이 긍정이라고 예측한 것들 중 실제 긍정인 비율

TP / (TP + FP)

1. Recall : 전체 양성인 것 들 중에서 진짜 실제와 예측 모두 양성인 것

TP / (TP + FN)

1. Precision과 Recall 어떤 관계가 있는가?

* 같은 모델에서 trade off 관계를 띈다. Precision이 커지면 Recall이 작아지고, Recall이 커지면 Precision이 작아진다.

1. Precision과 Recall이 각각 어느 때 중요한가?

* Recall : 대상을 얼마나 빠뜨리지 않고 잘 잡아내느냐가 중요할 때
* Precision : 검출된 결과가 얼마나 실제를 잘 포함하고 있는지가 중요할 때 (스팸메일)

1. SV model에서 연속형 변수를 이용해서 예측하는 방법에 대해 설명하라

* Numeric 변수를 적절한 구간을 설정하여 class로 나눠준다.

1. Threshold과 Precision, Recall과의 관계를 설명하라

* Threshold가 낮아지면 precision도 낮아진다 (높아지지는 않는다)

모델은 설정된 Threshold 이상이면 T라고 분류하기 때문에 Threshold가 낮아지면 그만큼 T로 분류한 개수가 많아진다. precision은 모델이 긍정이라고 분류한 샘플 중 실제 긍정인 값의 비율을 말하는 것이므로, T로 분류한 개수가 많아지면 precision의 분모가 커진다. 따라서 precision은 낮아진다.

* Threshold가 낮아지면 recall은 높아진다. (낮아지지는 않는다)

모델은 설정된 Threshold 이상이면 T라고 분류하기 때문에 Threshold가 낮아지면 그만큼 T로 분류한 개수가 많아진다.

recall은 실제로 T인 것들(실제도 T이고 예측도 T + 실제는 T 인데 예측은 F) 중에 실제와 예측 모두 T 인 개수에 대한 비율을 의미하므로, Threshold가 낮아지면 T라고 분류한 개수가 많아 져서 F라고 분류한 값은 작아지고, FN 값이 작아진다.

따라서 전체 recall값은 증가한다.

1. ROC 커브는 무엇이며 어떻게 그려지는가?

* 여러 임계값들을 기준으로 Recall-Fallout의 변화를 시각화한 것이다. FPR 은 실제 False인 data 중에서 모델이 True로 분류한 것이고 Recall은 실제 True인 data 중에서 모델이 True로 분류한 비율을 나타낸 지표로써, 이 두 지표를 각각 x, y의 축으로 놓고 그려지는 그래프를 해석한다.
* TPR(Recall) = TP / (TP + FN)
* FPR= FP / (FP + TN) : FPR값은 높을수록 안 좋다.

1. AUC는 무엇이며 어떤 의미가 있는가?

* ROC 커브 아래의 면적값을 의미한다. FPR에 비해 Recall 값이 클수록 AUC는 1에 가까워지며, 가까워 질수록 좋은 모델이다.