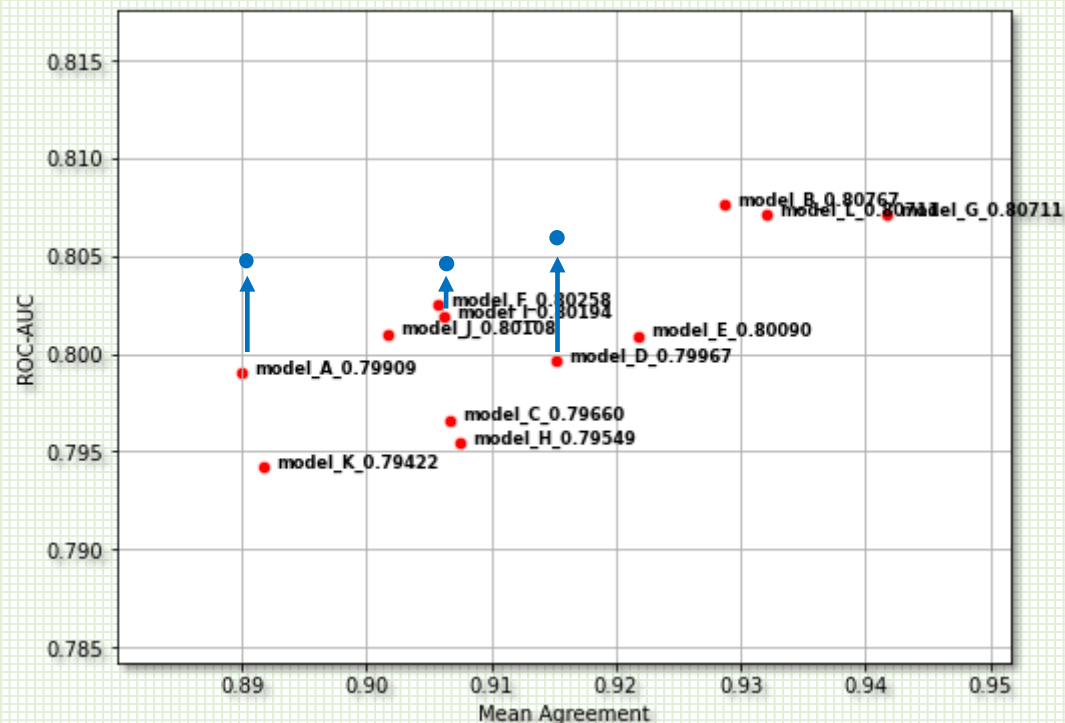


Machine Learning Ensemble Strategy

1st ENSEMBLE STRATEGY

Apply Yourself!



〈그림 1〉
Private 성능 순위를 반영한 상관계수 평균, ROC-AUC 산점도

모델 선택 기준

1. 성능이 높은 모델
2. 전체 모델(A~L)의 상관계수 평균을 넘지 않는 모델

모델 선택 전략

- ◆ Bagging 기법을 활용, 기존 방식에서 채택한 복원추출보다 위 기준에 부합하는 모델을 고려하도록 하였음
- ◆ Private에서도 가장 좋은 성능을 보일 것으로 기대되는 기준 모델 B를 선정
- ◆ 기준 2번에 위배되는 모델 소거
- ◆ A부터 C..L까지를 비교하여 둘 중 더 높은 성능을 보이는 모델 우선 선정
- ◆ 그림1과 같이 Private 기준의 성능 순위를 참고하여 선택

◆ 1차 선정 모델 조합 :

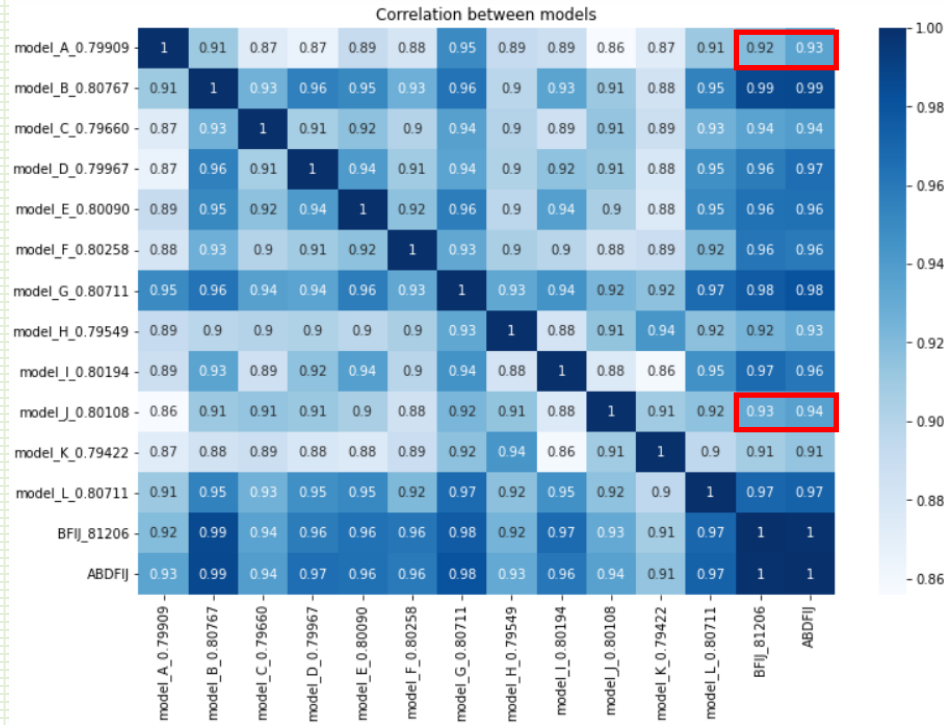
(1) BFIJ (2) ABDFIJ

- Private LB 기준 모델별 성능 순위: model-D > model-A > model-I > model-F
- 나머지 모델은 Public LB와 Private LB 간 성능 순위 차이가 적음

〈그림 2〉
모델 선정 참고 자료

2nd ENSEMBLE STRATEGY

Apply Yourself!



〈그림 2〉
1차 모델(BFij, ABDFij)가 반영된 Heatmap

2차 앙상블 전략

◆ 앙상블이 된 1차 모델과 타 모델과의 상관관계 볼 때, 전체적인 상관계수 평균이 증가한 것을 볼 수 있었다. 유사도와 성능 관점에서 바라보았을 때, 가장 이질적이면서 좋은 성능을 나타내는 A, J를 후보 모델로 선정하였다

추가 모델 투입

◆ 2차 앙상블에 대한 수 차례 실험 결과, A와 J를 투입하여 얻는 성능에 한계가 있다고 판단하여 계수가 높지만 성능도 탁월한 모델 L에 낮은 가중치를 부여해 private 성능 향상을 기대하였다.

AVERAGING STRATEGY

	순위	가중치
B	4th	$4/10 = 0.4$
F	3rd	$3/10 = 0.3$
I	2nd	$2/10 = 0.2$
J	1st	$1/10 = 0.1$

가중 평균 전략

◆수 차례 실험 결과 모델 구성 뿐만 아니라 예측 확률에 대한 평균법에도 성능에 상당한 결과를 미친다는 것을 포착했다
 ◆Random하게 모델을 선택하였다면 산술평균을 쓰는 것이 좋아 보이지만 성능을 고려한 평균화를 해야 했기 때문에 성능에 따른 가중 평균 전략을 사용하여 성능 향상을 기대했다

가중치 설정 전략

◆ 1차, 2차 모델 모두 가중 평균을 이용하였다. 다만, 가중치를 너무 높게 두었을 경우, Overfitting 현상이 발생할 수 있다고 판단하여 제출마다 가중치를 조율하며 최적의 가중치를 조율함
 ◆특히 BFIJ 모델을 구성할 때, 순위 평균법에 근거하여 가중치를 설정하였다

〈그림 2〉
가중치 평균 전략에 사용된 순위 평균 산정 방식

FINAL & CONCLUSION

- ◆ 본 과제는 일반적인 **배깅** 앙상블 기법에서 **가중 평균**과 **가중치**를 지속적으로 변화시키는 **부스팅** 원리를 차용했다는 점에서 의미가 있다.
 - ◆ 1차 제출물은 public을 기준으로 좋은 모델을 선정
 - ◆ 2차 제출물은 private에 특화된 방향으로 모델을 선정하여 결과물을 제출하였다
 - ◆ 1차 모델 선택은 기존 public 에서 제출했던 모델 조합과 같이 B, F, I, J 를 선정 후 추가로 private에서 좋은 성능을 보일 것이라고 기대했던 A, J 를 추가하여 앙상블 하였다
 - ◆ 위에서 앙상블한 모델을 중심으로 상관계수가 낮은 A, J를 앙상블 목록에 추가하였고, 상관계수는 높은 편이지만 성능이 좋았던 L 모델을 추가하여 최종으로 앙상블 하였다
 - ◆ 1차 제출물 같은 경우는 Kaggle에 직접 제출해보며 가중치를 조율할 수 있었지만, 2차 제출물은 기회가 얼마 남지 않아 순위 평균 사용(순위를 정하고 일정 비율로 가중치 설정) 가중치 (1차 모델 / A / J / L : 0.4 / 0.3 / 0.2 / 0.1)
 - ◆ 1차 제출물 **SCORE** : 0.81307(public 기준)
 - ◆ 2차 제출물 **SCORE** : 0.81284 (public 기준)
- 상대적으로 낮은 점수이지만 private을 목표로 한 모델이기에 더 좋은 최종 성능을 보일 것으로 기대하고 있다