어떤 피쳐가 있는데, 피쳐 1은 카운터블한데, 인피니트 한경우 (트레인 셋이든, 테스트 셋이든, 커버할 수가 없음)-> 더미를 만든다.

트레인 셋 안에 있는 에프 1에 대한 걸 만들고

테스트 셋에도 만들 것

만든 다음에 데이터를 넣고 모델을 돌려야 함

모델의 제너럴라이제이션

모델이 커버 되는 가?

트레인 셋만 더미를 만들 것

트레인 셋과 트레이닝 셋을 반드시 나눌 것!

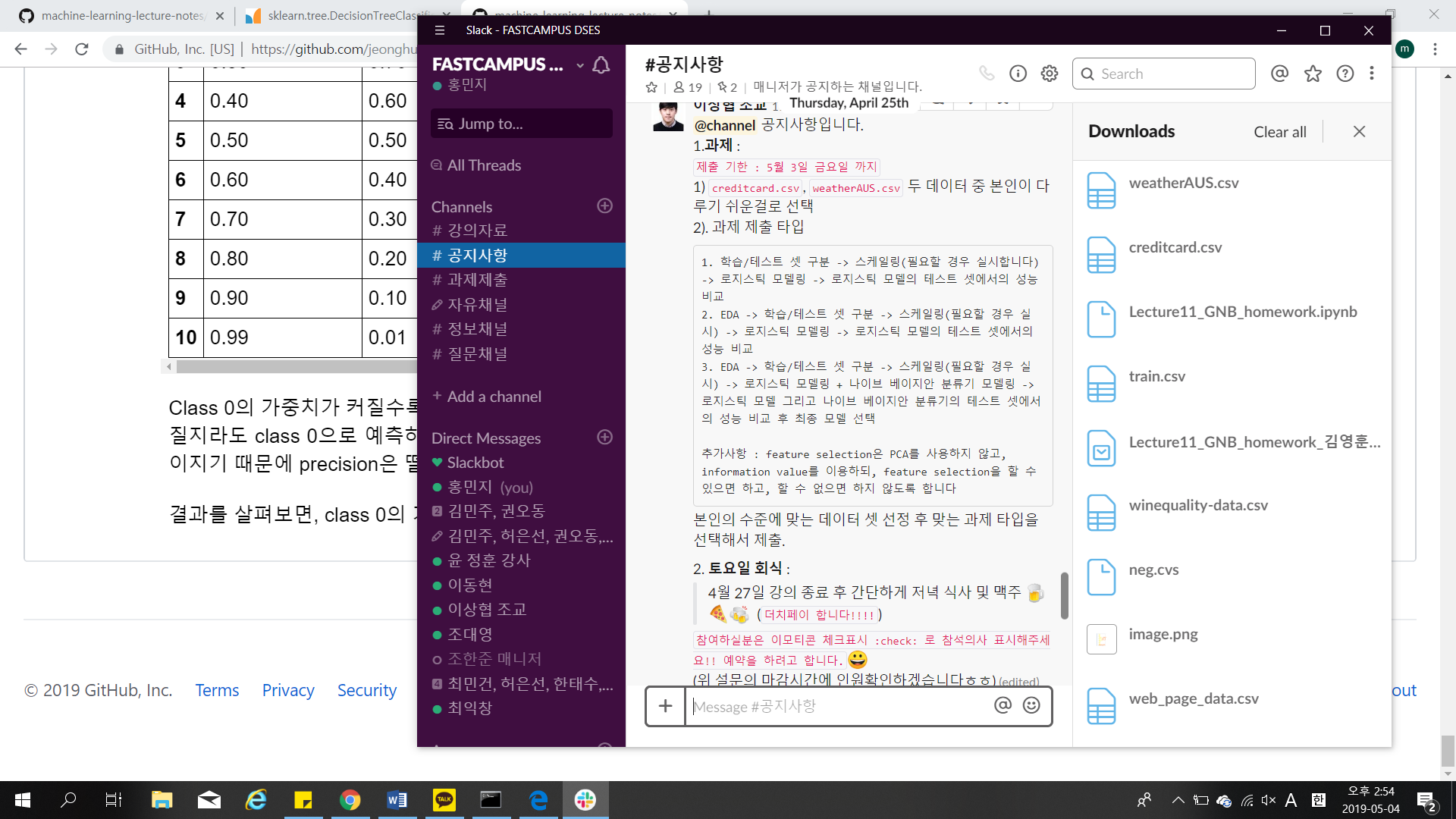
크로스 섹션에 모델 하나를 집어 넣는 다는 것은

모델에 최적화 된 파라미터를 고정 시켰는데, 앞으로는 파라미터를 찾겠다!

* 앞으로는 학습을 할 때, 하이퍼 파라미터를 찾는 그리드 리서치를 할 것!
* 트리의 최대 단점은 베리언스 임!(오버 피팅)\_
* 10개 이하
* 그리드 리서치는 크로스 벨리데이션이 기본 장착 되어 있는 것임.
* n\_jobs는 내 씨피유를 다 쓴것
* 학습할 필요 없이 베스트를 리턴해줌!
* 크로스 벨리데이션이 3개 이니까, 1,2개는 0
* 모델을 비즈니스 쪽에서 학습할 때, 예측력이 좋은 것은 중요함.
* 프리시젼(내가 1로 예측했는 데, 진짜 1인 거) 과 리콜(진짜 1인 것 중에 내가 1로 예측 할것
* *Confusion Matrix이 모델의 실제 성능을 알아볼 수 있음!()*
* )모델을 모르는 기획자에게 설명해줄 수 있음
* 미니멈 리브 값이 1라서 오버 피팅 됨
* 가중값을 준다는 것은 그걸 틀리면 로스 펑션이 커지는 것. 따라서/ 퇴직자에 대해서 예측하는 횟수가 커짐.
* 모델을 만들때는 예측 수치를 높이는 것도 중요하지만, 클래스 웨이트를 조절하는 것도 중요하다.
* 1. 예측력이 좋은 하이퍼 파라미터를 찾기
* 그 파라미터를 넣고 돌리면 됨!(클래스 웨이트에)
* 클래스 웨이트를 두고 돌려서 최상의 클래스 웨이트를 찾는 것이 중요함!

그 데이터가 임발란스 할 때, 12만 : 12만

1:1을 뽑아내는

ㅍ 

다 맞을 수는 없다!

베리언스(머신러닝에서의 베리언스는 트레이닝 셋이 변할 때, 그 모델이 얼마나 센스티브하게 변하는 가\_ 수학에서의 분산과는 다르다!)

앙상블을 통해서 베리언스를 줄일 수 있다.

데이터 과학에서 말하는 부트 스트랩이랑은 다른 개념임

그때의 부트 스트랩은 가설 검정(예측)을 할 때, 표본 평균을 예측을 할 때\_모델에서 사용하는 파라미터라든지, 통계쩍으로 검정할 때 부트 스트랩을 씀.

앙상블에서 부스 스트랩은 그것과 다름 (가설 검정을 하기 위해서 하는 것은 아님)

바이어스-베리언스 트레이드 오프를 통하여, 충분히 큰 샘플의 수가 중요하다는 것을 알 수 있다는 것이 중요함.

부트 스트랩에서는 두번 이상 뽑히는 경우, 한번도 안뽑히는 경우가 발생한다가 중요함!

문제와 유사한 여러 하위 문제들

문제에서 샘플링을 하는 것

엘레멘트오브 스태티컬 러닝\_ 책\_강사님 추천 책!

다이버시티(다양성)을 늘리는 것이 앙상블의 목적이다(분산을 줄이는 것)

데이터 전체를 리샘플링 한다는 것과 칼럼 조차 리샘플링 한다는 것의 차이는?

앙상블을 구성하는 분류기: 베이스 러너-> 그 결과를 최종 결과로!

이게 다 다양성 때문임!

* 앙상블 결과를 어그리게이션 한다는 것은?

성능이 좋은 분류기에 웨이트를 주겠다: 다중 다수 투표

Behavior knowledge class: 아까 웨이트 보팅과는 다른 컨셉임.

다수결 원칙이 성능을 고도화 할 때 쓰는 방법이 Behavior knowledge class 이다.

소프트 맥스를 최종값으로 많이 쓰기도 함.

클래스 랭킹은 참고용으로 알아 둘 것.

분류기의 결과가 실수 값이 나왔을 때 -> 소프트 맥스를 쓰거나,

소프트 맥스의 경우, 결과 값이 실수 값인데, 그 결과 값이 평균을 결과값으로 주거나, 결과 값을 확률값으로 줘야할 수도 있음.(근데 각각의 베이스 러너의 결과 값이 실수 인데, 최종 결과 값을 확률 값으로 줘야 할 때->소프트 맥스를 사용 하면 됨-> 소프트 맥스를 통해서 앙상블을 할 수 있음

앙상블의 목적은 다양성/

배깅에서 평균을 쓰는 이유는 다양성 때문임