피씨에이를 하기 전에 정규화를 해야함

정규화가 된 상황에서

엑스는 0 임

(=

Xxt은 공분산 이다-> 정규화된 경우

공분산이 뭔지

코벨리언스가 뭔지 -> 꼭알아야 함

SVD 는 아이겐 벡터로 쪼개는 것

PCA, SVD의 연결 고리

아이겐 페이스

픽셀이 0~255개 임.

너무 많아서 255개의 픽셀을 다 쓰지 않고 정말 필요한 것만 가져다 씀

피씨에이를 써서 피쳐가 변형된 것

디멘셔널 리덕션!

차원의 저주: 차원이 늘어나면 늘어날수록 데이터가 기하급수적으로 늘어남!

* 피쳐를 다쓴다고 해서 피쳐가 많다고 해서 데이터를 잘 설명해주는 것은 아님.
* 공분산 행렬에서 이어지는 부분
* 분산을 최대화 하는 방법
* 분산: 어떤 데이터에서 그 놈들의 평균을 빼고
* 분산에서 미치는 영향을스 조절하기 위해서 정규화를 해야함-> 스탠다트 스케일러

\*\* pca 할때는 시각화 하는 방법을 알아야 한다.

모델의 해석력 구할 때 알 스퀘어 구함

아이겐 벡터의 가장 기본적인 전제가 오소고날(직교)

두번째 피씨에이를 구할때는 직교를 구할 수 있

독립적이고

엣 모스트

피씨에이는 차원의 저주

차원의 개수를 줄여서 피씨에이의 설명력만 따진다?-> 놉 -> 왜? 설명력이 떨어지는 피쳐라도 다른 피쳐랑 합쳐져서 설명력을 높여줄 수 있음

* 차원을 줄일 때 쓸 수 있는 방법

1. 피씨에이
2. 에스브이디

\*\* 둘다 결과 값은 같은 것이 함정임!

피쳐 수가 많다고 차원의 저주가 걸리지는 않음->피쳐수가 많으면 데이터 수가 많으면 차원의 저주가 걸리는 것이 아님 ->

몇 개의 피쳐에 몇 개의 데이터가 적절한 지에 대한 공식은 없음.

비쥬얼라이제이션해서 감각으로 쳐내야 할 것 이 많음

구분이 안된다

추측할 수 있다

피쳐 임포턴스-> 트리를

이디아이를 통해서 각각의 카테고리컬 변수가 어떻게 분포되어 있는지를 알아야 함.

데이터의 분포를 알아야 데이터를 설명할 수 있음

아웃라이어가 있는지, 데이터가 잘 분포 되어 있는 지를 볼 수 있어야 함

박스 플롯, 스트립 플롯 등이 있음

* 히트 맵과 바 플롯(카운트)가 합쳐져야 좋은 모델임
* 바 플롯을 통해서 어떤 데이터가 좋은 지 인사이트를 가져야 함!
* 여러 메모리를 다 저장할 필요는 없고 / 해당 인덱스(번호)만 알고 있으면 됨.
* Feature importance 를 제공하는 라이브러리는 1) 트리 (지니 임퓨러티\_지니 임퓨러티가 가장 떨어진 놈을 찾는 것임)-> 피쳐별로 지니 임퓨러티의 평균을 냄 -> 이게 피쳐 임포턴스를 구하는 방법임 -> 피쳐 임포턴스는 되게 유의한 피쳐 라고 볼 수는 없음. 다만, 이 피쳐로 인해서 분기가 되었을 때, 지니 임퓨러티가 가장 높음, 그래서 분기가 정말 잘 된 것임-> 그렇다고 1순위를 쓰는 것은 아님
* 2) 그레디언트 부스팅(에서 슬플릿의 기준은 엠에스이\_ 엠에스 이는 가장 낮은 거)

엑스 쥐 부스팅은 그래디언트 부스팅의 가장 빠른 버전 임

\

트리에서는 노말라이제이션 필요 없지

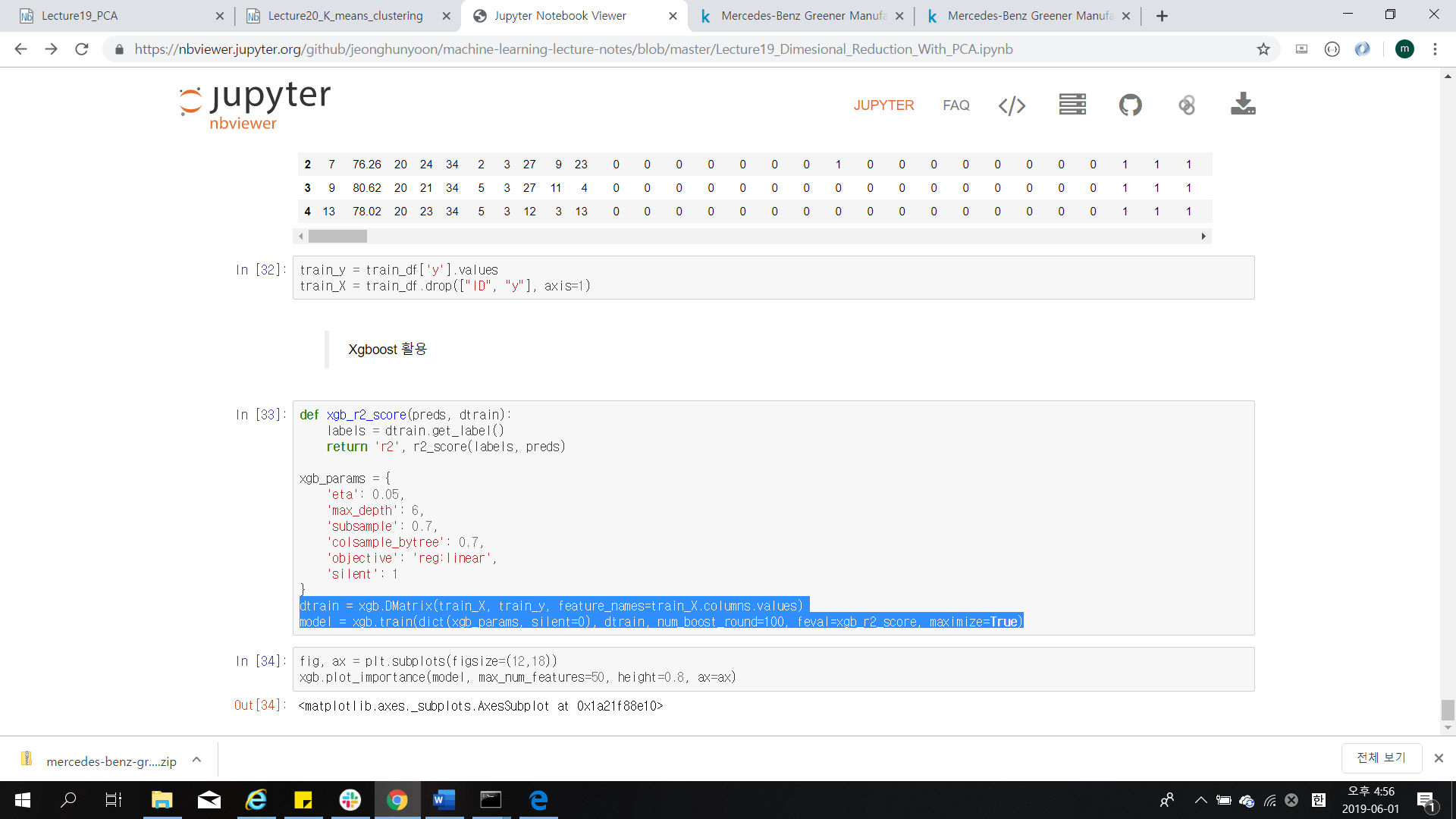
카데고리 컬 데이터를 더미 변수로 바꿀 필요도 없음(트리의 경우)

지니 임퓨러티 감소 값의 평균-> 피쳐 임포턴스

기존의 피쳐 수를 줄이고 싶은때, 피쳐 임포턴스를 사용하면 됨.

어떤 모델을 쓰느냐에 따라 피쳐 임포턴스가 달라짐 -> 따라서 피쳐 임포턴스만 가지고 판단 하면 안됨.!

엑스 쥐 부스트는 넘파이 어레이를 받지 못함



엑스 쥐 부트스는 피쳐 임포턴스를 플로팅 해줌.

\*데이터를 통해 인사이트를 가져오는 것은 진짜 중요함

그것 이외의 피쳐 임포턴스는 다른 방법 이 있음

피쳐를 기존에 있던 피쳐 셋에서 디멘젼을 줄일 때임

주성분 분석에서 가장 중요한 것은 아이겐 벡터라서 알 수가 없음

그래서 피씨에이는 피쳐 엔지니어링을 할 때, 우리 모델에서 가장 영향력을 주는 변수는 이거 이다 라는 것을 알 수가 없음!

피씨에이를 쓸 때, 스케일 맞춰 야함/ 뉴머릭한 피쳐는 0,1 밖에 없음 카테고리 컬 변수는 0~39개 까지 있음 / 라벨 인코더는 트리에서만 쓰고 다른 데에는 절대 쓰면 안됨(라벨 인코더는 오더이기 때문에, 의미 없는 오더 임 근데 다른 모델에 씀으로써 그 오더가 의미가 생겨짐)

/ 따라서 피씨에이를 쓸 때는 더미 변수로 변환 해야함

피씨에이를 쓴 후에 분류 모델을 적용 할 것

피씨에이를 통해서 회귀 모형을 만드는 것

결정 계수의 최대 치는 1 임

이론이랑 어플리케이션의 차이가 없다

차원의 저주- 피쳐 수에 비해 데이터의 수가 많지 않다

80프로를 설명하는 것 40 몇 개 였지만

기존의 데이터를 바꾸고 싶지 않다면, 피쳐 임포턴스나 이디아이를 쓰면됨.

엠엘에프는 좀 씀-> 토픽 모델링을 할 때, 엘 에스 에이, 피엘에스 에이 라고 해서 더불어서 많이 씀.

\*\* 피씨에이만 완벽하게 쓸 것