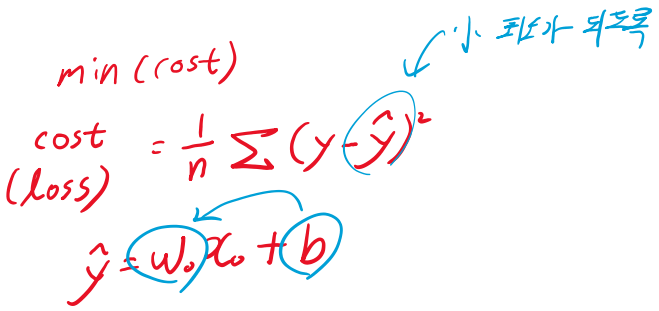
3/13(금) azure ml, 교재 ml이론

2020년 3월 13일 금요일 오전 9:06

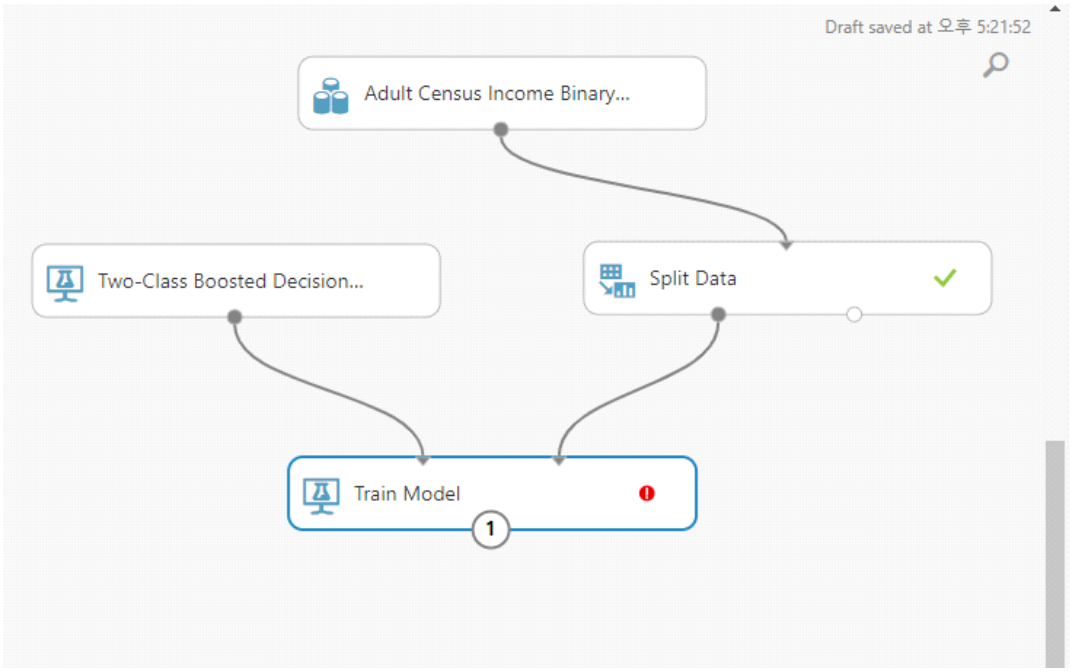
EDA - 처음, 매우 중요. 주로 산점도(scatter plot) 많이 쓰인다.

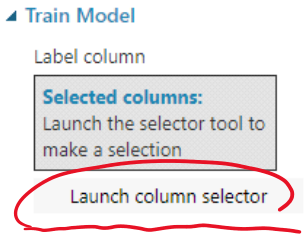


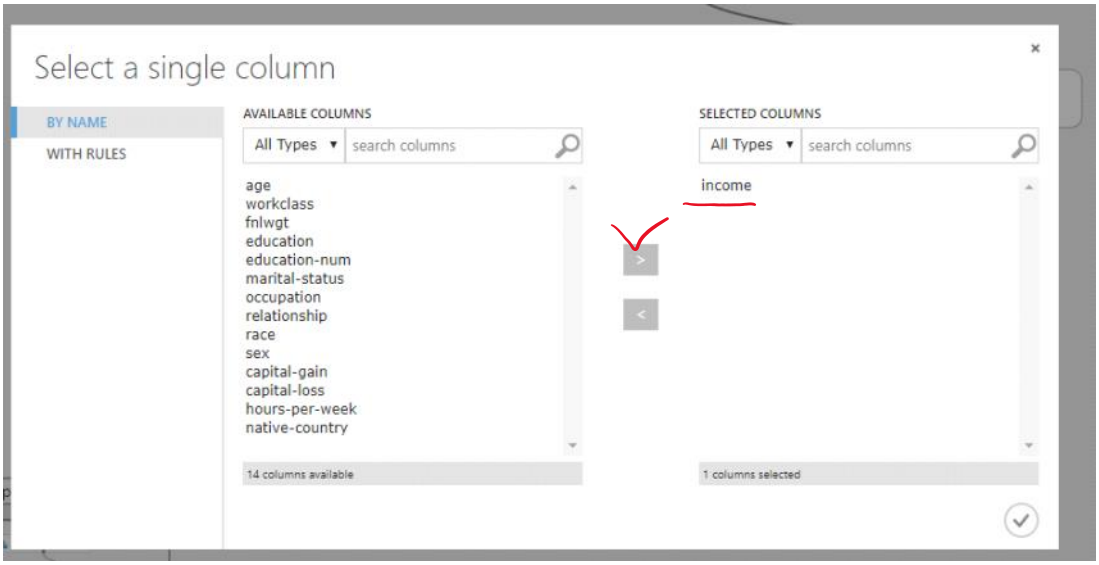
Feature engineering 파생변수 ex)몸무게 키 등으로BMI지수 얻어내는 등
결정계수 - 회귀모델에서 ex)knn 예측의 적합도를 측정한 것 1이 완벽, 0은 훈련 세트의출력 값의
평균으로만 예측, 음수는 예측과 타깃이 상반된경향

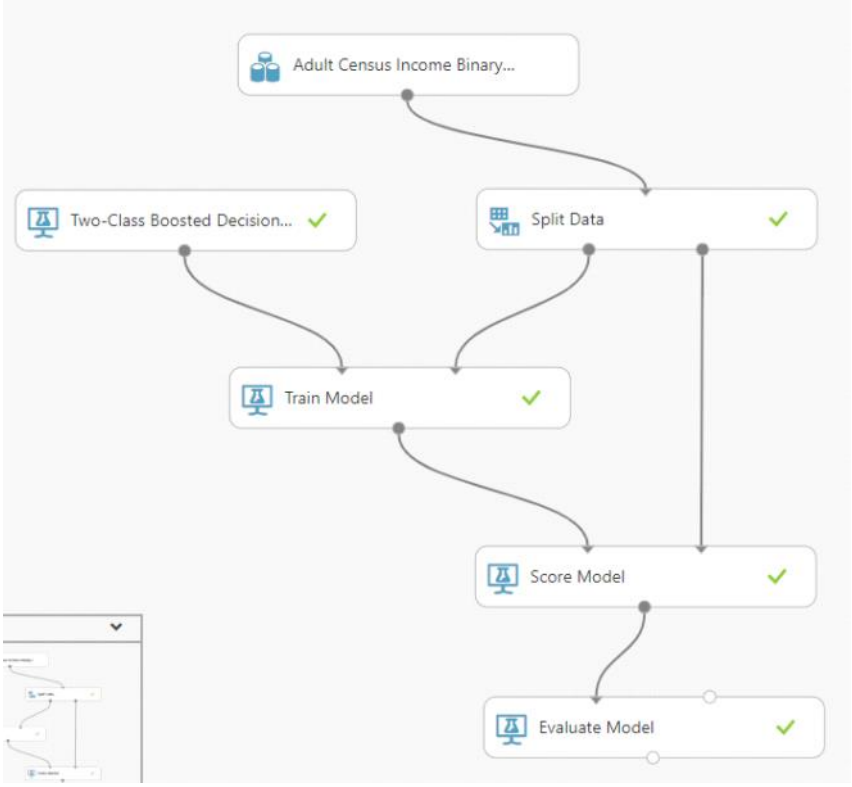
기본 선형 회귀, 과대 적합 문제 발생 -> 리지 회귀로 해결
리지 회귀 - w(가중치 절대값)을 0에 가깝게 하는 것이 핵심 기울기를 작게!
라소 - 리지의 대안, 특정 계수 0 즉 완전히 제외되는 특성이 생김. 과소적합 발생...과소 적합을 줄이기 위해서 alpha값(계수를 얼마나 강하게 0으로보낼지)을 줄임 따라서 max_iter 반복 실행 최대 횟수 늘려야 한다.
의사결정트리 - 전처리 크게 필요 없다. 단위에 영향 x 각 컬럼별 독자적
단점은 과대적합 overfiting 자주 발생... 일반화 시키기 쉽지 않다.
따라서 의사결정트리는 단독으로 잘 쓰이지 않는다. 모델 결합 앙상블 기법으로 사용.
앙상블 장점 - 과대적합이 상당부분 해결이 된다.
단점 - 연산량이 많아짐.
모형결합방법(앙상블)
1)취합 방법론 : 사용할 모형의 집합이 이미 결정되어 있음
-종류 : 다수결, 배깅, 랜덤포레스트
2)부스팅 방법론 : 사용하 모형을 계속해서 늘려가는 방법
-종류 : 에이다부스트(AdaBoost), 그레디언트 부스트(Gradient Boost)

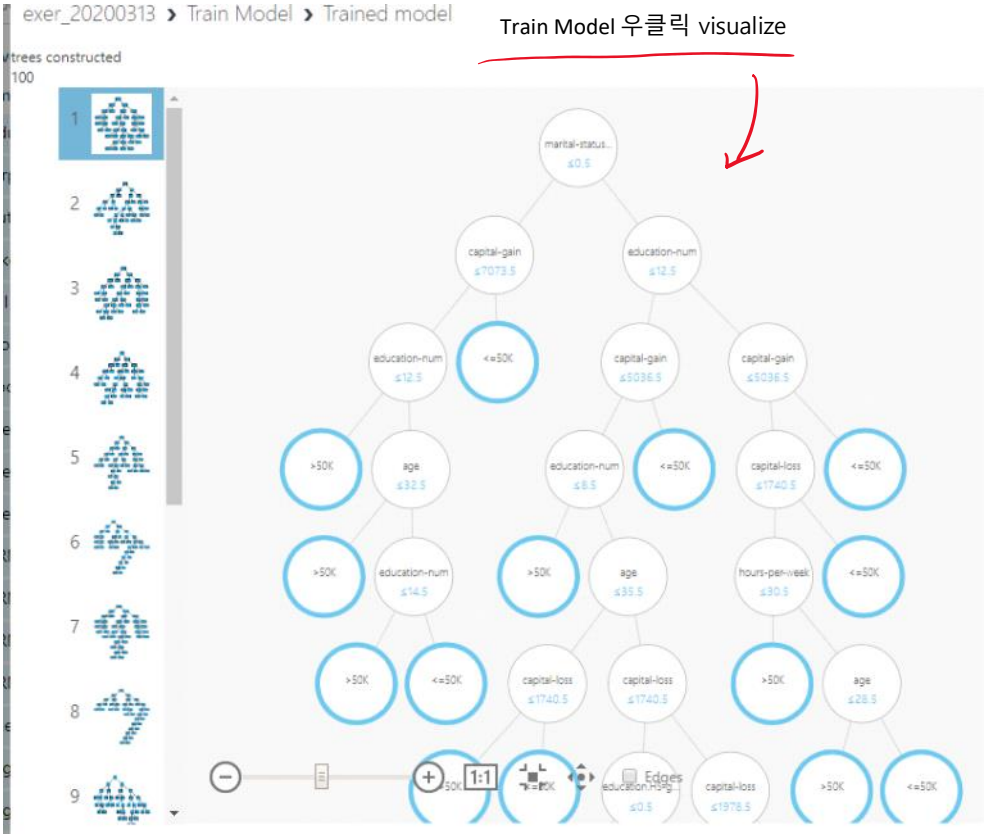
batch-size
epoch


Draft saved at 오후 5:21:52
Adult Census Income Binary...
Two-Class Boosted Decision...
Split Data
Train Model
1

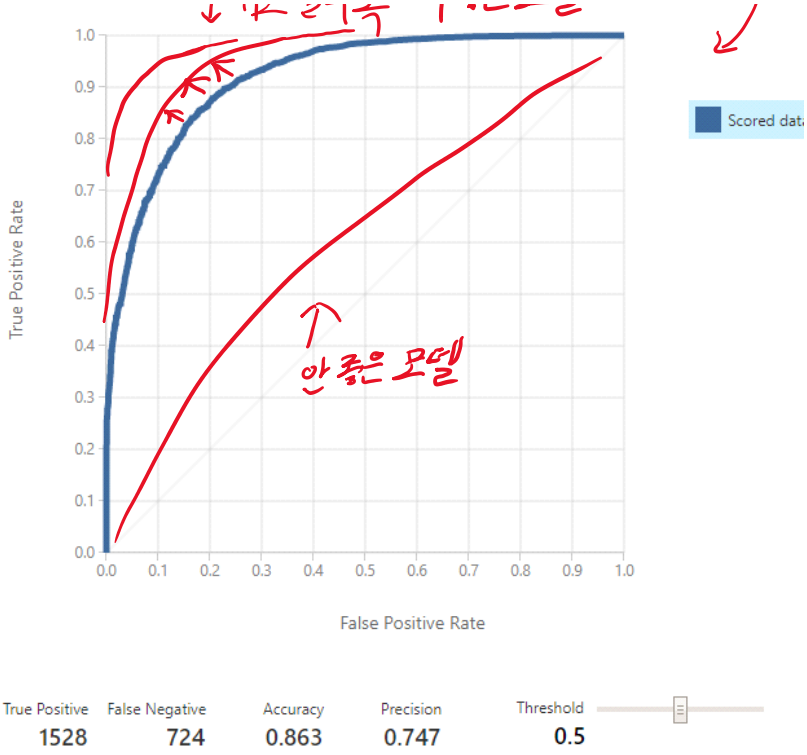

Train Model
Label column
Selected columns:
Launch the selector tool to make a selection
Launch column selector


Select a single column
BY NAME
WITH RULES
AVAILABLE COLUMNS
All Types
search columns
age
workclass
fnlwgt
education
education-num
marital-status
occupation
relationship
race
sex
capital-gain
capital-loss
hours-per-week
native-country
14 columns available
SELECTED COLUMNS
All Types
search columns
income
1 columns selected


Adult Census Income Binary...
Two-Class Boosted Decision...
Split Data
Train Model
Score Model
Evaluate Model


exer_20200313 › Train Model › Trained model
Train Model 우클릭 visualize
trees constructed
100


exer_20200313 › Evaluate Model › Evaluation results
우클릭 visualize
ROC PRECISION/RECALL LIFT
↓ 위로 갈수록 더 좋은 모델
1.0
0.9



True Positive	False Negative	Accuracy	Precision	Threshold
1528	724	0.863	0.747	0.5

ROC곡선은 정밀도와 재현율 대신 진짜 양성 비율(TPR)에 대한 거짓 양성 비율(FPR) 나타냄
FPR = FP / (FP+TN)
-> 전체 음성 중에서 실제는 음성인데 양성이라고 예측한 경우

FPR(병에 걸리지 않은 사람을 양성으로 판정)
FPR은 0이 가장 이상적
-TPR(병에 걸린 사람을 양성으로 판정)
=> 가장 이상적인 모델(TPR=1, FPR=0)