D&A

ML Session 7차시

ML Competition.

2022 / 11 / 08

D&A 학회장 권유진

01. ML Competition

ML Competition

데이터를 통해 주어진 task를 가장 잘 수행하는 기계학습 알고리즘 구축

- 학습 성능 뿐만 아니라 일반화 성능 역시 좋아야 함 (Overfitting 발생 x)
- 해당 task를 수행하는 것에 적합한 feature 생성
- 데이터 형태에 따른 적합한 전처리 수행
- 적절한 모델 및 하이퍼 파라미터 선정

01. ML Competition

주제 설명

신용카드 사용자들의 개인 신상정보 데이터로 사용자의 신용카드 대금 연체 정도를 예측

https://dacon.io/competitions/official/235713/overview/description

- Train (26457, 20), Test: (10000, 19)

index	gender	car	reality	child_num	income_total	income_type	edu_type	family_type	house_type	DAYS_BIRTH	DAYS_EMPLOYED	FLAG_MOBIL	work_phone	phone	email	occyp_type	family_size	begin_month	credit
0	F	0	0	0	202500	Commercial associate	Higher education	Married	Municipal apartment	-13899	-4709	1	0	0	0	None	2	-6	1
1	F	0	1	1	247500	Commercial associate	Secondary / secondary special	Civil marriage	House / apartment	-11380	-1540	1	0	0	1	Laborers	3	-5	1
2	M	1	1	0	450000	Working	Higher education	Married	House / apartment	-19087	-4434	1	0	1	0	Managers	2	-22	2

01. ML Competition

데이터 설명

index: 인덱스 번호

gender: 성별

car: 차량 소유 여부

reality: 부동산 소유 여부

child_num: 자녀 수

income_total: 연간 소득

income_type: 소득 분류 ['Commercial associate', 'Working', 'State servant', 'Pensioner', 'Student']

edu_type: 교육 수준 ['Higher education' ,'Secondary / secondary special', 'Incomplete higher', 'Lower secondary', 'Academic degree']

family_type: 결혼 여부 ['Married', 'Civil marriage', 'Separated', 'Single / not married', 'Widow']

house_type: 생활 방식 ['Municipal apartment', 'House / apartment', 'With parents', 'Co-op apartment', 'Rented apartment', 'Office apartment']

01. ML Competition

데이터 설명

DAYS_BIRTH: 출생일 데이터 수집 당시 (0)부터 역으로 셈, 즉, -1은 데이터 수집일 하루 전에 태어났음을 의미

DAYS_EMPLOYED: 업무 시작일 데이터 수집 당시 (0)부터 역으로 셈, 즉, -1은 데이터 수집일 하루 전부터 일을 시작함을 의미, 양수 값은 고용되지 않은 상태를 의미함

FLAG_MOBIL: 핸드폰 소유 여부

work_phone: 업무용 전화 소유 여부

phone: 전화 소유 여부

email: 이메일 소유 여부

occyp_type: 직업 유형

family_size: 가족 규모

begin_month: 신용카드 발급 월 데이터 수집 당시 (0)부터 역으로 셈, 즉, -1은 데이터 수집일 한 달 전에 신용카드를 발급함을 의미

credit: 사용자의 신용카드 대금 연체를 기준으로 한 신용도 [0, 1, 2]

=> 낮을 수록 높은 신용의 신용카드 사용자를 의미

01. ML Competition

평가 지표

LogLoss (= CrossEntropy)

최종적으로 맞춘 결과만 갖고 성능을 평가할 경우, 얼만큼의 확률로 해당 답을 얻었는지 평가 불가능

높은 확률로 정답을 확신했을 경우 점수를 더욱 높게주는 평가 지표

해당 값이 낮을 수록 좋은 모델임을 의미

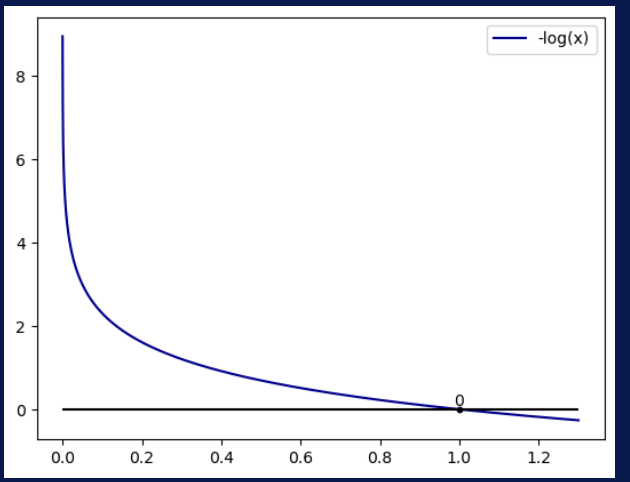
!! 하지만 sklearn은, n으로 나눠주지 않으므로 데이터 수의 영향을 받음!

$$LogLoss = \frac{-\Sigma_j \Sigma_i \, y_{i,j} \log p_{i,j}}{n}$$

n: 데이터 수 $(1 < i < n)$

$p_{i,j}$: i 번째 데이터의 j 클래스 예측 확률

$y_{i,j}$: i 번째 데이터의 j 클래스 정답 여부 (정답 시 1, 아닐 시 0)



01. ML Competition

평가 방법

성능 평가 (50%) + 분석 평가 (50%)

성능 평가

1등 성능의 점수를 50점, baseline 성능의 점수를 10점으로 두고 점수 계산

Baseline 성능: 0.91282

분석 평가

상호 평가 (20%) + 운영진 평가 (30%)

타당한 feature 생성, 분석 기법, 독창적인 기법을 사용했는지 평가

영역	심사기준	점수
모델 성능	Public, private 점수 평균	50
데이터활용능력	데이터를 적절하게 활용하였는가?	10
분석 기법	해당 목적에 맞는 데이터 분석 기법을 적용하였는가?	10
독창성	분석 결과를 적용하기 위한 적절한 방안을 제시하였는가?	10
코드 정리	코드가 오류 없이 잘 정리되었는가?	10
종합		100

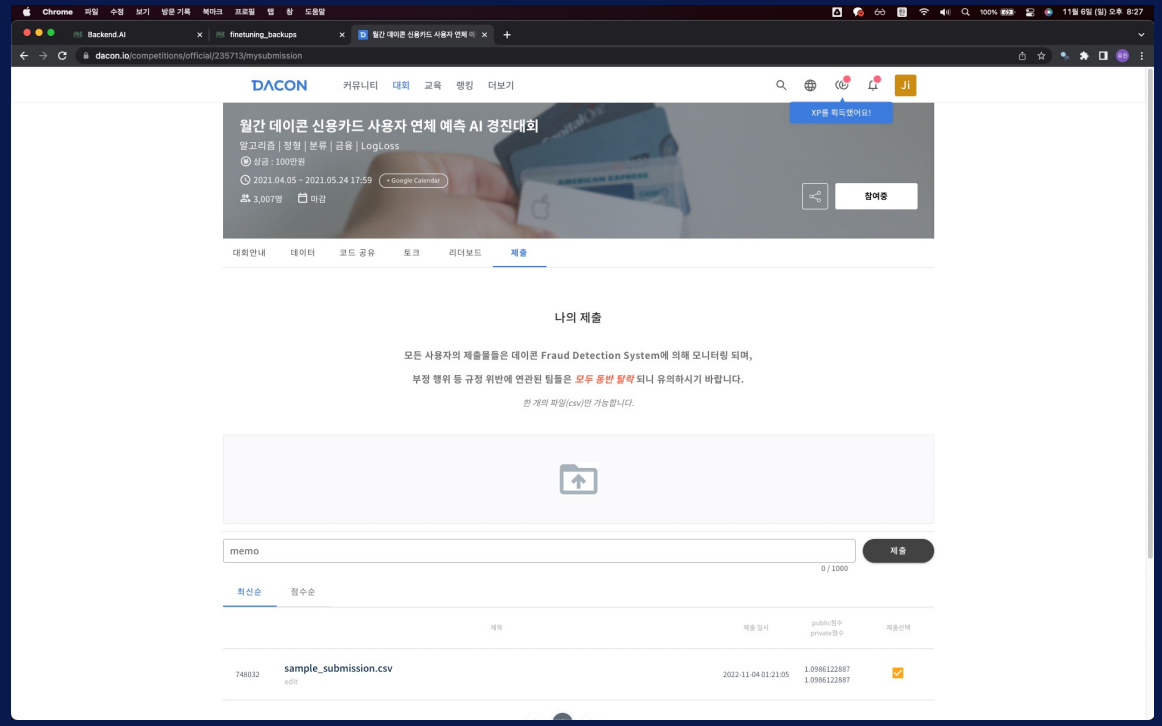
01. ML Competition

진행 방식

11/08 (화) ~ 11/24 (목)

11/24 23:59까지 kyja4639@naver.com으로 아래 파일 전송

- "조이름_이름"으로 닉네임 변경 (계정관리 > 닉네임변경)
- 성능 캡쳐 사진
- 코드
- submission 파일



11/29(화): ML Competition 발표회

→ 각 조당 7분씩 발표

11/28(월)까지 발표 ppt 제출

→ feature 생성 과정, 전처리 방법, 모델 등

D&A

ML Session 7차시 ML Competition

Thank You.

2022 / 11 / 08

D&A 학회장 권유진