방학세미나

불꽃 학회장팀

정승민 김진혁

INDEX

1

주제 및 팀 소개

1 주제 및 팀 소개

TOPIC

주어진 데이터를 활용하여 성능이 좋은 이진 분류 모델 만들기

DATA

Train : 55000*171

Test : 21000*170

변수명은 확인 불가하며, 자료형만 확인 가능합니다!

bin : Binary / cat : Categorical / 그 외는 Numeric 또는 Ordinal

평가지표

Cost Function

: 클래스마다 다른 비용 발생

$$Cost\ Function = 250 \times \mathrm{FN} + 5 \times FP$$

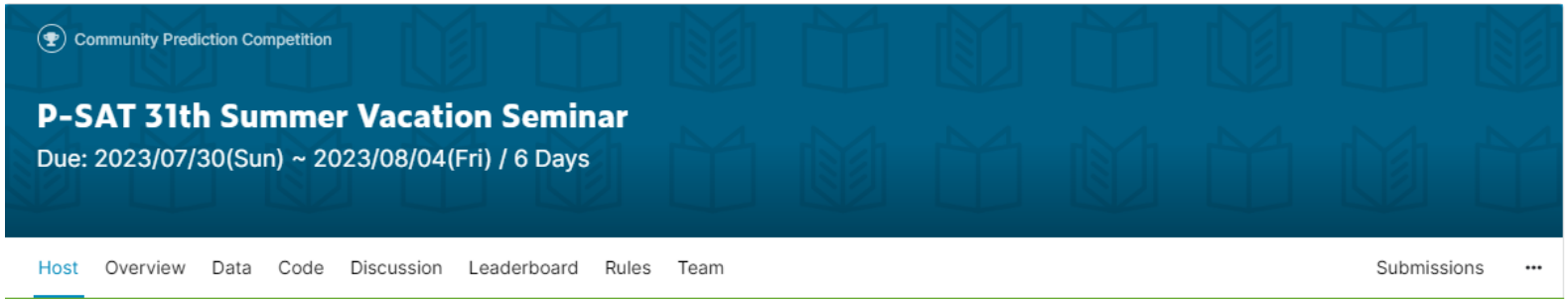
FN(False Negative): 양성임에도 음성으로 예측한 경우
FN(False Positive): 음성임에도 양성으로 예측한 경우

분류 모델의 성능을 평가하는 지표로 사용

1 주제 및 팀 소개

진행방식

kaggle



https://www.kaggle.com/t/1d7dab2d759e44d3bb4c07b30bb9a19b

캐글에서 모델의 성능을 파악할 수 있도록 컴피티션 개최

Public leader board에서 50%의 test set에 대해 채점 가능

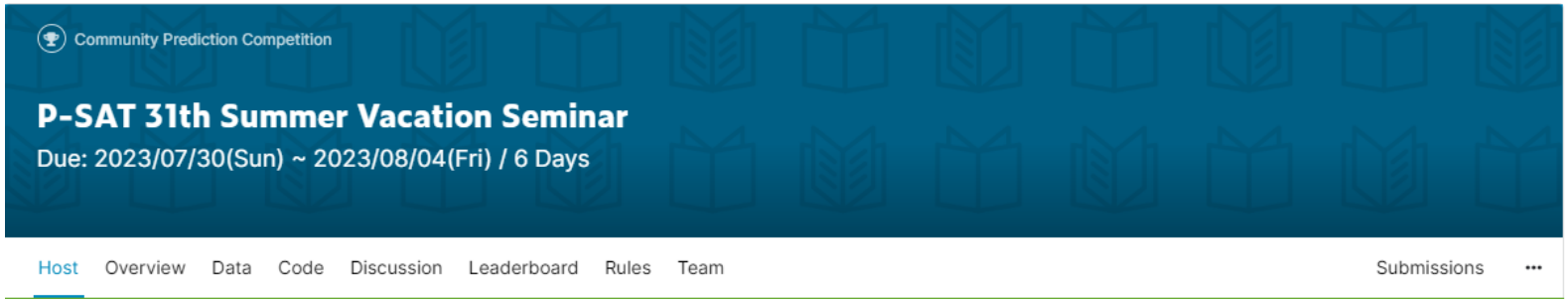
최종 점수는 제출된 submission.csv의 Cost Function로 평가

계정은 1인당 1개만 사용 가능합니다!

1 주제 및 팀 소개

진행방식

kaggle



https://www.kaggle.com/t/1d7dab2d759e44d3bb4c07b30bb9a19b

제출 가능 횟수 : 1일 3회

Leaderboard를 통해 Score를 실시간으로 확인 가능

가장 Score가 높은 팀원이 속해있는 팀이 Competition 1등

팀 소개

1팀	이정환 성준혁 노정아 김동환 이지원 하희나
2팀	김보현 장다연 천예원 최용원 이상혁
3팀	김다민 김보근 송승현 서유진 조성우 심현구

2

평가 기준

2 평가 기준

평가기준	항목	상세	배점
CODE	성능	Cost Function	5
	재현성	코드가 재현되는가?	3
	가독성	코드가 깔끔히 정리되어 있는가?	3
	시간	Predict 시간	2
Analysis	EDA	시각화를 통한 데이터의 해석 및 인사이트 도출	6
	명확성	분석흐름에 대한 논리와 명확성	5
	참신성	데이터 처리 과정 및 사용 모델의 참신성	3
	Data Leakage	테스트 셋에 대한 정보 사용불가	2
PPT	논리성	분석 흐름에 대한 논리적 명확성	2
	심미성	디자인, 미적구성 (PPT의 가독성과 내용 구성이 적절한가)	2
제출기한	지각	제출시간 준수(기본 2점)	2
총점		35	

2 평가 기준

✓ 성능

Test Set에 대한 평가지표

Cost Function 평가

✓ 재현성

전처리 및 모델링 코드 전반의 재현성

모든 과정이 동일하게 재현되어야 합니다.

랜덤씨드 설정에 유의해주세요

✓ 가독성

코드의 가독성 및 효율성

단계별로 다른 코드 파일을 생성해도 좋으며,

한 코드 파일에 모두 담아도 좋습니다.

✓ 시간

모델 학습 / 모델 예측 시간 측정

코드 파일 내에 아래 2가지가 명시적으로 측정되어야 합니다.

① Train 데이터에 대한 학습 시간

② Test 데이터에 대한 예측 시간

2 평가 기준

✓ EDA

시각화를 통한 변수 특성 파악,
변수와 모델의 해석을 통한
인사이트 도출 과정 전반

✓ 명확성

분석 전개 과정에서
전반적인 흐름의 논리와 방향 평가

✓ 참신성

변수 선택, NA 처리, 모델 선정
등에서의 참신성 평가

✓ Data Leakage

분석 과정에서 Test Data 관련
정보 사용 금지

2 평가 기준

✓ EDA

시각화를 통한 변수 특성 파악,
변수와 모델의 해석을 통한
인사이트 도출 과정 전반

변수명이 익명 처리되어 있으므로
변수의 의미를 파악하기보다는, 아래 3가지에 집중해 주세요!

① EDA가 논리적으로 사용 및 전개되었는지
② 얼마나 다양한 정보를 EDA에 담아낼 수 있는지
③ 새로운 인사이트를 도출해 냈는지

✓ 참신성

변수 선택, NA 처리, 모델 선정
등에서의 참신성 평가

✓ Data Leakage

분석 과정에서 Test Data 관련
정보 사용 금지

2 평가 기준

✓ 명확성

명확성은 아래 2가지를 기준으로 평가할 예정입니다.

① EDA에서 도출된 인사이트가 모델링에서 적절히 활용되었는지
② 데이터의 특성을 반영한 모델을 적절하게 선정했는지

분석 전개 과정에서
전반적인 흐름의 논리와 방향 평가

✓ 참신성

변수 선택, NA 처리, 모델 선정
등에서의 참신성 평가

✓ Data Leakage

분석 과정에서 Test Data 관련
정보 사용 금지

2 평가 기준

✓ EDA

시각화를 통한 변수 특성 파악,
변수와 모델의 해석을 통한
인사이트 도출 과정 전반

✓ 명확성

분석 전개 과정에서
전반적인 흐름의 논리와 방향 평가

✓ 참신성

변수 선택, NA 처리, 모델 선정
등에서의 참신성 평가

1학기 각 팀의 클린업과 주제분석에서 다뤄진
전처리 과정, 모델, 변수 선택법, NA 처리 방법 등을
얼마나 잘 활용했는지를 위주로 평가할 예정입니다.

2 평가 기준

✓ EDA

시각화를 통한 변수 특성 파악,
변수와 모델의 해석을 통한
인사이트 도출 과정 전반

✓ 명확성

분석 전개 과정에서
전반적인 흐름의 논리와 방향 평가

✓ Data Leakage

매우 당연하게도, 전처리부터 모델링 과정 전반에서
Test 데이터와 Test 데이터로부터 도출된 정보를
절대 사용해서는 안 됩니다.
Ex) Scaling, PCA, ...

분석 과정에서 Test Data 관련
정보 사용 금지

2 평가 기준

✓ 논리성

분석 흐름 소개 과정의
논리적 명확성

✓ 심미성

PPT의 가독성 및 구성

✓ 지각

제출 기한 엄수

3

제출

제출 방법

1) Leader board

- 1일 최대 제출 가능 횟수 : 3회

- 제출 양식 : id(1:178564)와 target 예측값

id	target
1	0
2	0
3	1
...	...
178562	0
178563	1
178564	0

2) 학회장 제출

- 팀원 중 1인이 대표로 학회장에게 카톡으로 파일 제출

<R, Python 다 가능>

- 제출 파일 : PREPROCESSING/MODELLING/PREDICT 코드 파일
  + 최종예측결과 csv 파일

제출 방법

1) Leader board

- 1일 최대 제출 가능 횟수 : 3회

- 제출 양식 : id(1:178564)와 target 예측값

id	target
1	0
2	0
3	1
...	...
178562	0
178563	1
178564	0

2) 학회장 제출

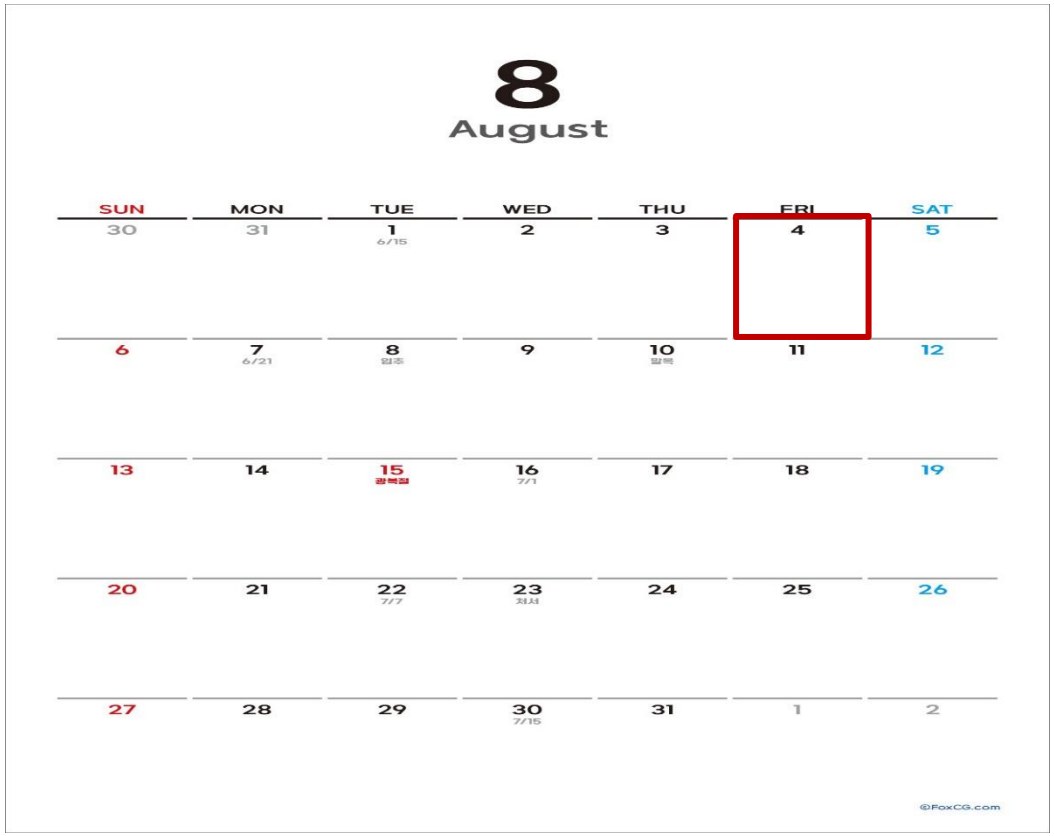
- 팀원 중 1인이 대표로 학회장에게 카톡으로 파일 제출

<R, Python 다 가능>

- 제출 파일 : PREPROCESSING/MODELLING/PREDICT 코드 파일
  + 최종예측결과 csv 파일

하나의 zip파일로 제출

제출 기한



2023년 8월 4일 (금), 05시 59분까지 제출

<상대적으로 짧은 방세기간 고려…>

4

발표

4 발표

세미나

1) 일시

- 2023년 8월 4일(금) 오후 4시 오프라인 진행
- 인문관 31608호로 시간에 맞춰서 와주시면 됩니다.
- 오후 2시부터 대여를 해 놨으니 미리 오셔서 준비하셔도 됩니다.

2) 발표자 선정

- 발표자 및 발표 순서는 3시에 랜덤으로 선정 후 공지
- 발표시간 : 팀당 10분

3) 결과 발표 및 시상

- 2023년 8월 4일 : 각 팀 발표 종료 후 시상
- 1등 팀은 소소한 상품과 방세 1등 팀이라는 큰 명예를 얻을 수 있습니다!

THANK YOU