방학세미나

즐거운 학회장팀
권남택
박서영

INDEX

1

주제 및 팀 소개

주제 소개

데이터 : Truck Air Pressure System Failures Prediction

문제 : 주어진 변수들을 통해 비용을 최소화하는 예측모델 만들기

- 미리 고장을 예측함으로써, 문제 상황으로부터 발생하는 비용을 줄이는 것이 목표.
- 데이터 불균형 상황. 비고장(neg)/고장(pos)의 비율은 약 98:2
- 데이터 변수들은 모두 익명처리 되어 있음.
- 데이터에는 NA가 매우 많음.

데이터 소개

Train : 40000 * 171 (row / col), Masked Variables

Neg 39333, pos 667

Test : 8000 * 171 (row / col), Masked Variables

엄청난 NA 데이터

기존 60000/16000 데이터였는데 조금 사이즈 줄임!

https://drive.google.com/drive/folders/1H5PIrQFQh9UXwFCvyk6LJTGzk1VXhXkD?usp=sharing

변수 소개

변수들의 의미는 아무도 몰라요! ㅎㅎㅎ

Numeric variable들만 존재

평가 metric

총 비용의 최소화

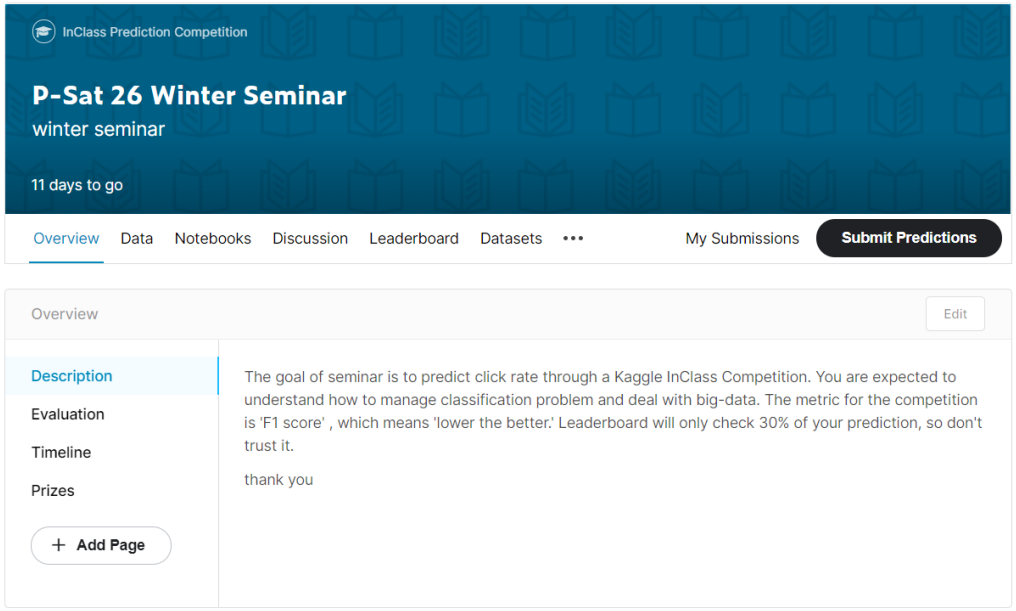
	True pos	True neg
Predicted pos	Zero cost	Cost1 : 10
Predicted neg	Cost2 : 500	Zero cost

총 비용 = 10 *해당 개수 + 500*해당 개수

고장이 안난다고 예측했는데 실제 고장이 날 경우 훨씬 더 큰 페널티 부여

주제 및 팀 소개

평가 metric



하지만 캐글에서 비용을 평가지표로 할 수 없기 때문에
F1-score로 30% test set에 리더보드 채점
https://www.kaggle.com/c/psat26-winter-seminar

최종 코드 제출시에 총비용 계산할 예정!

팀 소개

Team 1	Team 2	Team 3
심은주	황정현	이지연
황유나	염예빈	이수경
김지민	한유진	진효주
문서영	진수정	김서윤

공모전/인턴 등의 일정이 있는 학회원들을 고려해서 배정했으니
서로 부담을 줄이면서 했으면 좋겠어요!

2

평가 요소

2 평가 요소

평가 기준	항목	상세	배점
Code	성능	총 비용	5
	재현성	코드 재현이 되는가??	3
	시간	Predict 시간 등수	5
	가독성	정리된 코드 가독성	3
Analysis	EDA	변수 의미 파악 및 해석 정도	2
	NA 처리	적절한 NA 대체	3
	클래스불균형 처리	클래스 불균형/비용 고려한 모델링	3
	No Data Leakage	테스트셋에 대한 정보 사용 불가	3
	감성 점수	참신한 전처리/모델링 등등 재밌으면 추가점수	3
PPT	스토리텔링	분석 흐름에 대한 논리적 명확성	5

총점 : 35

2 평가 요소 - Code

성능 – 총 비용의 최소화

Test Set에 대한 총 비용 계산

1등 : 5점
2등 : 4점
3등 : 3점

cf) 리더보드는 30% 테스트 데이터에 대한 F1 Score
총 비용에 대한 리더보드 제작이 불가능했음

2 평가 요소 - Code

코드 재현성

전처리 및 모델링 코드 재현성

재현 가능 : 3점
재현 불가능 : 0점

서영이가 preprocess/modeling/predict 코드 제출 예시를 올릴 예정!

랜덤 시드 고정 필수

평가 요소 - Code

Predict 시간

모델 부르기 및 Predict 시간 등수

1등 : 5점
2등 : 4점
3등 : 3점

전체 preprocess/modeling/predict 중
Predict 부분의 시간만 평가!

평가 요소 - Code

코드 가독성

코드 가독성 및 효율성

만점 : 3점
(3점부터 감점)

코드를 읽었을 때 가독성 및 깔끔함

주석만 잘 달려있어도 읽기 편해요!

EDA : 변수 파악 및 해석

EDA : 변수 파악 및 해석

만점 : 2점
(2점부터 감점)

변수간의 관계 및 의미 파악 & 시각화 & 해석

EDA의 한계가 존재할 수 있다는 점 고려해
많은 배점 부여하지 않음.

평가 요소 - Analysis

NA 처리

수많은 NA에 대한 합리적 처리

만점 : 3점
(3점부터 감점)

Test set에도 NA 존재함을 고려!

NA Imputation 코드가 돌아가는 시간이
상당히 오래 걸릴 수 있으니 참고!

Imputed set을 저장해놓고 사용하길 추천

평가 요소 - Analysis

클래스 불균형

클래스 불균형 모델링

만점 : 3점
(3점부터 감점)

Train/test의 클래스 비율은 거의 비슷!

class imbalance를 위한 적절한 방법을 제시했는지 평가!

2 평가 요소 - Analysis

No Data Leakage

Test 데이터에 대한 정보유출 X

만점 : 3점
(3점부터 감점)

Test data는 predict 단계에서만 등장해야함

Ex) 전처리/모델링 과정에서 train/test를 합쳐서 전체 데이터에 동일한 PCA X
Ex) 전처리/모델링 과정에서 train/test를 합쳐서 전체 데이터에 NA 대체 X

2 평가 요소 - Analysis

감성 점수

특별한 데이터 전처리/모델링 기법 사용

만점 : 3점
(0점 부터 가점)

추가 점수를 주기 위함. 논리가 더 중요

특별함의 기준은 매우 낮으니 무언가를 엄청 찾으려고 노력하지 마세요!

2 평가 요소 – PPT

PPT

분석 흐름에 대한 논리적 명확성

만점 : 5점
(5점 부터 감점)

전체적인 과정이 논리적으로 깔끔하도록 구성!

PPT의 부담을 줄이기 위해

1. PPT 템플릿은 P-Sat 템플릿으로
2. 절대 PPT 30장이 넘지 않도록 구성
3. 딱 필요한 내용 위주로!

2 평가 요소

평가요소를 보고…

평가 기준을 제시하는 것은
어떤 문제이고/어디에 집중해야 할지 방향을 알려드리는 거니까

평가 요소를 보고 너무 어렵다고 생각하지 않으셔도 됩니다!!

언제든 어렵거나 궁금한 부분이 있으면 물어보세요!

3

제출 방법

제출 방법

Leader Board

리더보드

팀	점수
1팀	0.2406
2팀	0.2507
3팀	0.2888

제출 가능 횟수 : 2회 / 1일

제출 양식 : id(1:8000)와 pos/neg 여부

제출 방법

학회장/부학회장 카톡 제출!

Preprocess/modeling/predict 코드 + 최종 예측 결과 csv + PPT

3 제출 방법

Leader Board

리더보드

팀	점수
	0.2406
	0.2507
	0.2888

id	class
1	pos
2	pos
3	pos
4	pos
5	pos
6	pos
7	pos
8	pos

제출 양식 : Submission.csv 형식

다음 날 랜덤 시간 중으로 업데이트

Id와 pos/neg 여부

제출

학회장의 ' 구글 드라이브 ' 에 파일 제출

Ex) 08_41_1팀_1회.csv

<예시>

제출 방법

최종 제출 및 결과 발표

최종 제출

2월 7일 일요일 23시59분까지

Preprocess/modeling/predict 코드 + 최종 예측 결과 csv + PPT

결과 발표

2021년 2월 8일, 월요일 오후 6시 30분 세미나 때 발표

4

상품

상품

1등

험난한 채점을 거쳐 1등을 한 팀에게는 5만원의 회식비 지원을 해드립니다!

열심상

등수와 관계 없이 소정의 상품이 있을 예정입니다!

THANK YOU