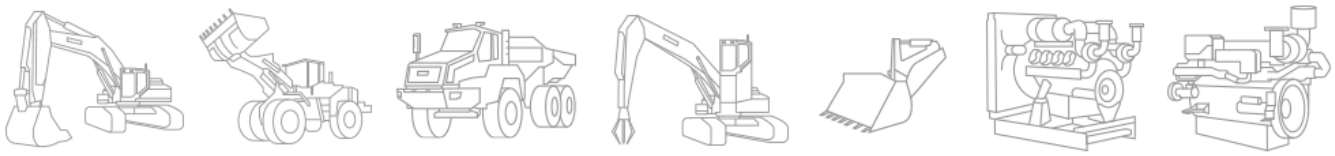
건설기계 오일상태 분류 AI 경진대회

2022. 12. 26

TEAM 새삥_H

김희중, 신웅재, 하태현



0. 팀원 소개

- 현대중공업 건설기계 3사 신입사원 동기로 구성

현대두산인프라코어
융합시스템개발팀
김희중 연구원

현대두산인프라코어
건기제품품질팀
신웅재 매니저

현대제뉴인
DT업무혁신팀
하태현 매니저

1. 데이터 확인

Train set

- Shape : (14095, 54)
- Test 데이터에 없는 Feature에 다수 결측치 존재
  ⋙ 19개의 Feature에 결측치 존재
- 변수의 형태 : 정수형(44개), 연속형(6개), 명목형(2개), 날짜형(1개)

Test set

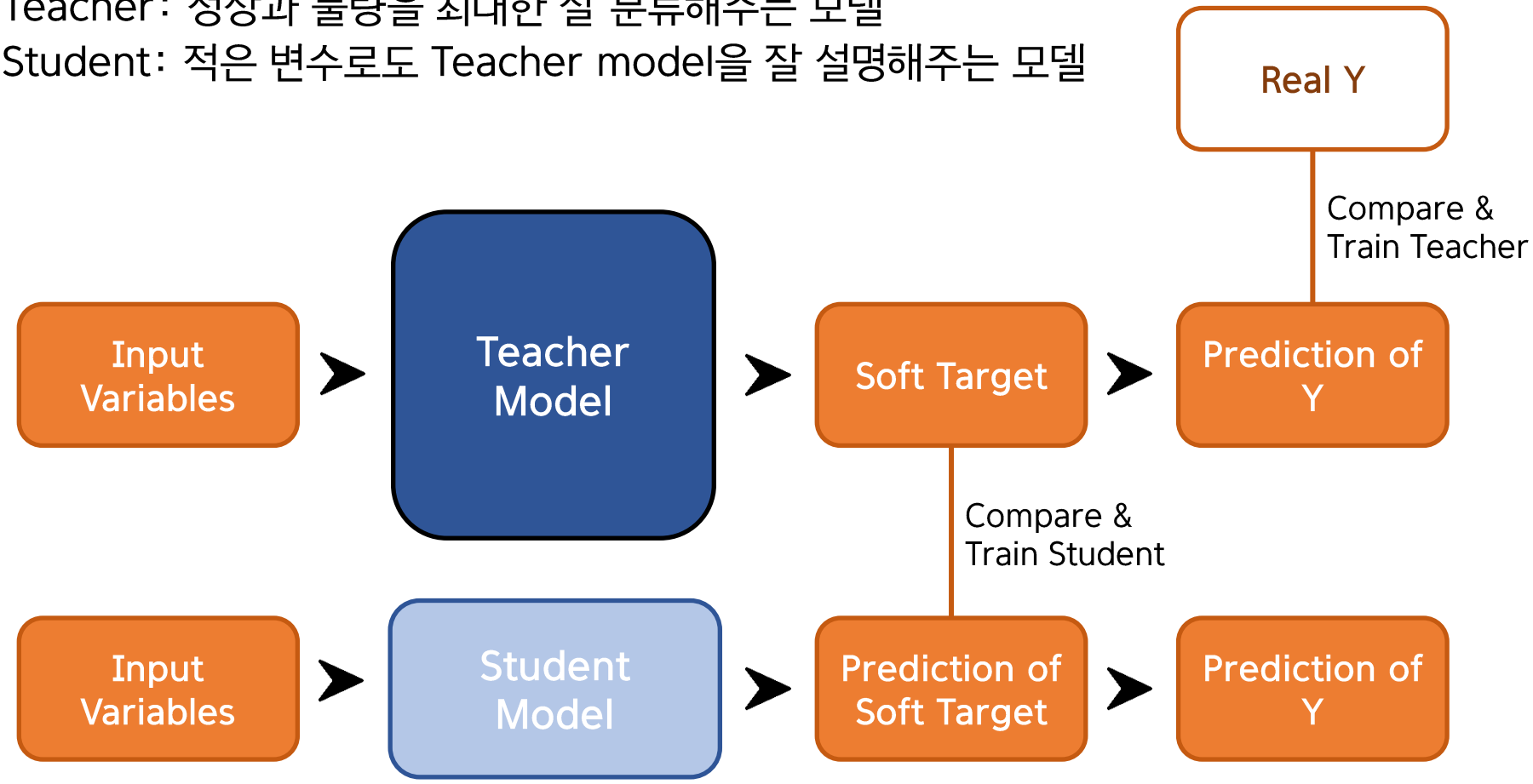
- Shape : (6041, 19)
- 결측치 미존재
- 변수의 형태 : 정수형(14개), 연속형(2개), 명목형(1개), 날짜형(1개)

- Train 데이터에서 결측률이 20% 이상인 변수는 제거
- ID 제거

⋙ Shape : (14095, 36)

1. 데이터 확인

- 이번 대회의 가장 큰 특징 : 변수의 개수가 다른 Train과 Test
- Teacher: 정상과 불량을 최대한 잘 분류해주는 모델
- Student: 적은 변수로도 Teacher model을 잘 설명해주는 모델

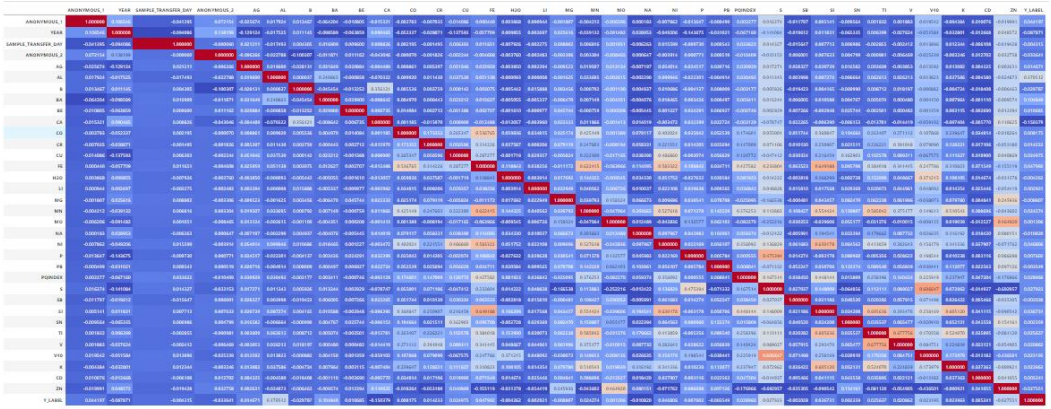


2. 변수 선택

- Train과 Test 데이터에 모두 포함되어 있는 18개의 변수 확인

  ⋙ 최대한 많은 변수를 이용하여 Student model이 예측불량률을 잘 설명하도록 하자!

상관계수 행렬



- 상관계수의 절댓값이 0.7 이상인 관계는 없음, 0.6 이상은 존재
- 18개의 변수 모두 사용하는 것이 좋을 것이라 판단

2. 변수 선택

- Train 데이터에만 존재하는 17개의 변수 확인
- Teacher model 추가 변수는 잠정적으로 2~3개로 채택

  ⋙ 너무 많은 추가변수는 Student model의 병목 현상 유발

변수 추가 기준

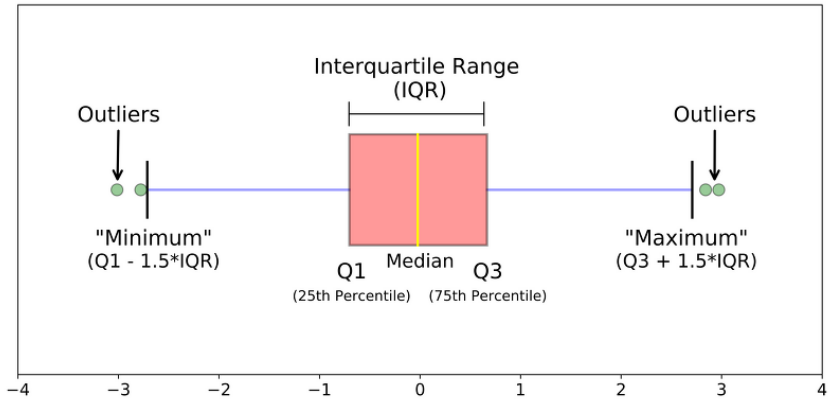
① Wilcoxon 순위합 검정의 P-value

② 이상 범위 데이터의 불량 개수와 불량률

③ 불량 데이터의 중위수가 그 변수의 이상 범위에 존재하는지 유무

2. 변수 선택

① Wilcoxon 순위합 검정의 P-value (유의수준 5%)



Variable	P-value
AL	**0.0000**
BA	**0.0000**
S	**0.0001**
K	**0.0001**
SI	**0.0029**
SN	**0.0056**
SB	**0.0087**
LI	0.1138
PB	0.1590
CD	0.3667
P	0.4350
BE	0.4657
SAMPLE_TRANSFER_DAY	0.7569
MG	0.9294
B	0.9894
NA	0.9995
CA	1.0000

- IQR을 이용한 이상치 판단 함수 생성
- Teacher model에서 중요한 변수

  ⋙ 정상 범위와 이상 범위의 데이터 사이의 차이가 큰 변수

  ⋙ 이상 범위의 데이터의 값이 정상 범위의 데이터보다 클 것이다!!
- Wilcoxon 순위합 검정 결과

  ⋙ 1차 변수 : AL, BA, S, K, SI, SN, SB

2. 변수 선택

② 이상 범위 데이터의 불량 개수와 불량률

변수	정상 범위 데이터 개수	이상 범위 데이터 개수	정상 범위 데이터의 불량 개수	이상 범위 데이터의 불량 개수	정상 범위 데이터의 불량률	이상 범위 데이터의 불량률
AL	12,651	1,444	506	**697**	4.000%	**48.269%**
BA	10,780	3,315	585	**618**	5.427%	**18.643%**
S	14,088	7	1,202	1	8.507%	8.808%
K	10,738	1,058	854	141	7.953%	13.327%
SI	12,246	1,849	1,004	199	8.199%	10.763%
SN	12,778	1,317	1,087	116	8.507%	8.808%
SB	11,128	2,967	902	301	8.106%	10.145%

- 7개의 1차 변수들이 실제로 불량을 잘 분류하는 변수인지 파악
- 이상 범위 데이터의 불량 개수 순위 : AL, BA, SB, SI, K
- 이상 범위 데이터의 불량률 순위 : AL, BA, S, K, SI

⋙ AL, BA는 Teacher model에 넣기로 우선적으로 확정

2. 변수 선택

③ 불량 데이터의 중위수가 그 변수의 이상 범위에 존재하는지 유무

- 나머지 5개 변수들도 중요한 변수인지 좀 더 자세히 파악 필요
- 불량 데이터의 A 변수의 중위값이 전체 데이터의 A 변수의 정상 범위를 벗어난다면
  ⋙ A 변수는 불량을 잘 구분해 낼 변수라 판단

변수	Lower Bound	Median of Defect	Upper Bound
AL	-3.5	**1,444.0**	8.5
BA	0.0	**3,315.0**	0.0
S	-18,523.8	9,422.0	42,714.3
K	-4.5	2.0	7.5
SI	-10.5	7.0	25.5
SN	-1.5	0.0	2.5
SB	0.0	0.0	0.0

📌 Teacher model : 기존 변수 18개 + AL, BA

3. 모델 선택

- AutoML인 Pycaret을 이용하여 Teacher, Student model 채택

Teacher model

- 성능 지표 : F1 score

⋙ CatBoostClassifier의 성능이 가장 좋음

Model	Accuracy	AUC	F1 Score
CatBoost	**0.9539**	**0.8788**	**0.6594**
LightGBM	0.9534	0.8689	0.6558
AdaBoost	0.9536	0.8611	0.6555
GBM	0.9527	0.8768	0.6542
RandomForest	0.9537	0.8591	0.6524

Student model

- 성능 지표 : MAE

⋙ 과대적합 해결과
순서형 변수인 YEAR 변수를 고려하여
CatBoost 선택

Model	MAE	RMSE	R2
Huber	0.0573	0.1776	-0.0566
CatBoost	**0.0823**	**0.1664**	**0.0722**
LightGBM	0.0826	0.1672	0.0631
GBM	0.0828	0.1665	0.0709
KNN	0.0857	0.1803	-0.0896

4. Teacher Model

- Input Features: 18개 변수 + AL, BA을 포함한 20개 변수
- 사용모델 : CatBoostClassifier
- Optuna를 이용한 Hyper parmeter 최적화 사용
- 데이터 분할 시 test_size = 0.3, Stratify = y_train 사용
- StratifiedKFold를 통해 불균형 데이터 교차 검증(Train set 불량률 : 약 8.5%)

초모수	Range
Learning Rate	[0.001, 1]
N Estimators	[100, 1000]
Max Depth	[3, 16]

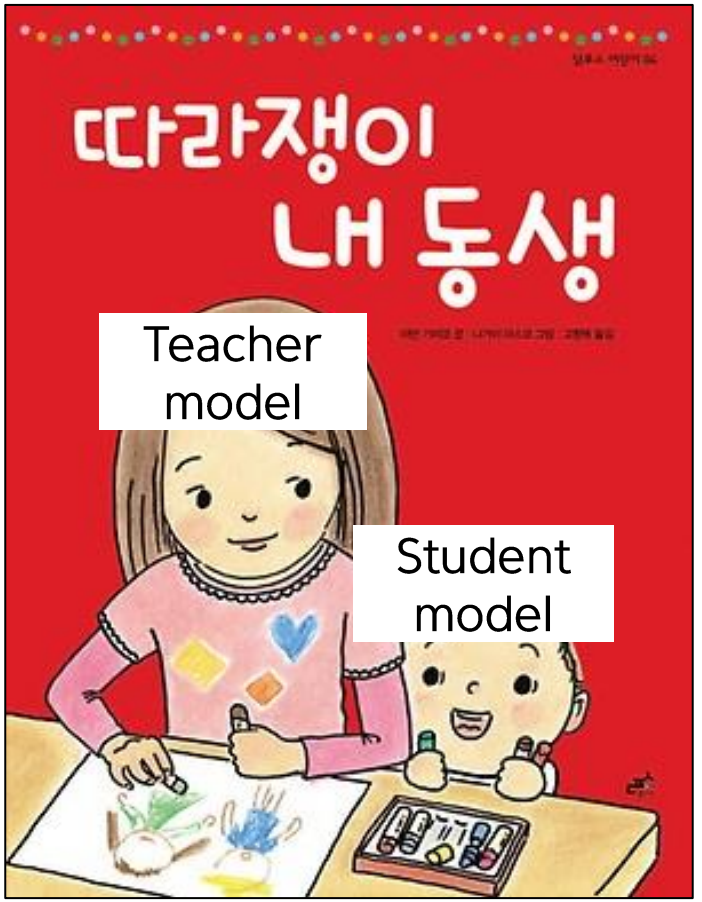
초모수	Value
Learning Rate	**0.0314234**
N Estimators	**513**
Max Depth	**6**

5. Student Model

- Input Features: 18개의 변수
- 사용모델: CatBoostRegressor
- Optuna를 이용한 Hyper parmeter 최적화 사용
- K-Fold로 교차 검증

OPTUNA

초모수	Value
Learning Rate	0.0131004
N Estimators	848
Max Depth	9



6. Submit

- Train set 예측값을 바탕으로 Threshold를 0.10 ~ 0.16 범위로 잡음
- 경험적으로 Test set의 불량갯수가 500~600개 일때 Score가 가장 높았음

Threshold	Score
0.00	0.157409
0.02	0.160011
0.04	0.176768
0.06	0.203085
0.08	0.228134
0.10	**0.246596**
0.12	**0.260094**
0.14	**0.257133**
0.16	**0.230072**

⋙ Threshold로 0.15 선택

	#	Score
Public	8th	0.60260
Private	12th	0.57764

7. 한계점

1. 모델의 안정성
   - Public score에 집중하여 모델링 ⋙ 과대적합의 위험
   - Stacking Ensemble을 Teacher 모델에 적용했으나, 제출 누락

2. 모델의 동작속도
   - CatBoost는 다른 Boosting 모델에 비하여 상대적으로 연산 속도가 느림
   - 작동 오일의 분류 모델이 얼마나 실시간성을 필요로 하는지 정보의 부족
   - LightGBM을 적용한 Student 경량 모델 시도 부족

3. 변수 선정
   - Train set과 Test set의 18개 변수 중 불필요한 변수 제거 과정 부족
   - Teacher model에 추가되는 변수에 대해 초점을 맞춤

Thank You