LG Aimers

스마트 공장 제품 품질 상태 분류 AI 온라인 해커톤

TEAM 틔운돌아와

LG Aimers

목 차

LG Aimers

1.데이터 전처리 과정

- PRODUCT_CODE별 데이터 분리

Col_A	NaN_A	Col_O	NaN_O	Col_T	NaN_T
PRODUCT_ID	249	PRODUCT_ID	6	PRODUCT_ID	343
Y_Class	249	Y_Class	6	Y_Class	343
Y_Quality	249	Y_Quality	6	Y_Quality	343
TIMESTAMP	249	TIMESTAMP	6	TIMESTAMP	343
LINE	249	LINE	6	LINE	343
PRODUCT_CODE	249	PRODUCT_CODE	6	PRODUCT_CODE	343
X_1	0	X_1	6	X_1	343
X_2	0	X_2	6	X_2	343
X_3	0	X_3	6	X_3	343
X_4	0	X_4	6	X_4	343
X_5	0	X_5	6	X_5	343
:	:	:	:	:	:
X_2865	249	X_2865	0	X_2865	0
X_2866	100	X_2866	0	X_2866	0
X_2867	100	X_2867	0	X_2867	0
X_2868	100	X_2868	0	X_2868	0
X_2869	100	X_2869	0	X_2869	0
X_2870	100	X_2870	0	X_2870	0
X_2871	99	X_2871	0	X_2871	0
X_2872	0	X_2872	0	X_2872	0
X_2873	0	X_2873	0	X_2873	0
X_2874	0	X_2874	0	X_2874	0
X_2875	0	X_2875	0	X_2875	0

과정

- A_31, O_31, T_31 별 결측치 확인
- 결측치를 가진 열이 코드별 상이
- O_31과 T_31 유사

결론

A_31과 O_31, T_31 2가지 모델을 나누어 구성

LG Aimers

1.데이터 전처리 과정

- A_31 Preprocessing

과정

- train.csv에서 PRODUCT_CODE == "A_31"인 값
- 결측치 유지 및 모든 열 사용

모델

결측치를 알아서 처리해주는 Tree기반 모델 사용

LG Aimers

1.데이터 전처리 과정

- O_31, T_31 Preprocessing

과정

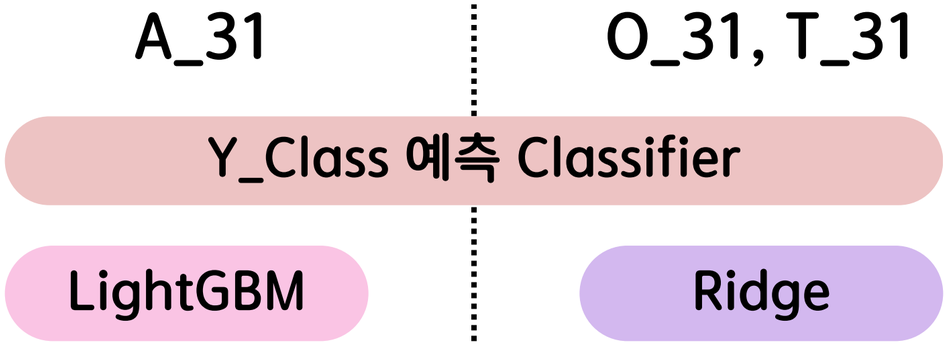
- train.csv에서 PRODUCT_CODE == "O_31" 또는 "T_31"인 값
  - def seperate_code() : Line별로 데이터를 분리 후 모든 값이 같은 열 제거
  - find_duplicate_col() : Line별로 결측치있는 열을 제거 후 교집합하여 사용할 열 결정

모델

총 111개의 열 사용해 모델링

LG Aimers

2.모델 선정



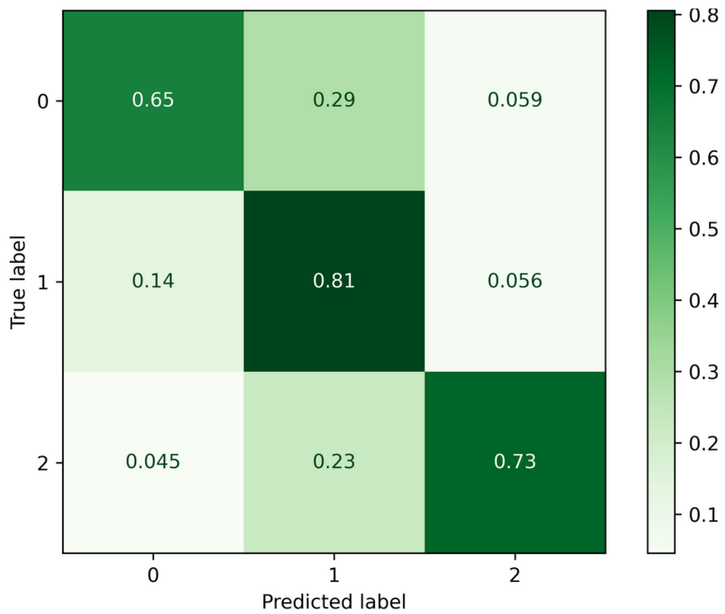
Leaf node를 반복 분할하며 예측 오류 손실을 최소화할 수 있는 LightGBM Classifier사용

Overfitting을 막기 위해 Ridge Regression을 사용하는 Ridge Classifier사용

LG Aimers

3.모델 결과

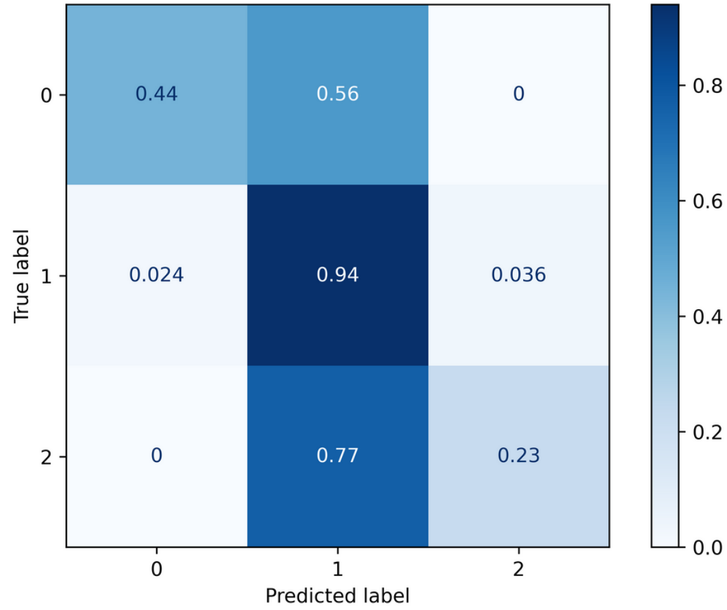
A_31



정확도: 0.747

혼동행렬

O_31 & T_31



정확도: 0.810

LG Aimers

4.시도해본 전처리/모델

- IQR을 이용한 이상치 탐지 및 변경
  - 1. 이상치에 대한 값을 각 열별 제 2사분위수(Q2)로 변경 후 모델링
  - 2. 이상치에 대한 값을 NaN으로 변경 후 Tree기반 모델링
  - 3. 이상치에 대한 개수로 값을 변경 후 분류 모델링

- Y_Class 데이터 불균형 해소
  - 1. SMOTETomek와 TomekLinks를 이용한 언더샘플링 후 모델링
  - 2. SMOTE, ADASYN을 이용한 오버샘플링 후 모델링

LG Aimers

4.시도해본 전처리/모델

- Y_Quality로 Y_Class 분류
  - Y_Quality가 분류하는 Y_Class의 정확도는 0.99로 매우 정확함
  - 회귀로 Y_Quality를 예측한 후, 예측 결과로 Y_Class를 분류하는 모델

- 공정 LINE 종류에 따라 구분해서 학습
  - PRODUCTCODE 뿐만 아니라 LINE 별로 분류해서 학습

LG Aimers

4.시도해본 전처리/모델

- seaborn 그래프 이용
  - hitmap으로 피처의 상관 관계 파악 후 필요없는 피처제거
  - barplot으로 피처와 Y_Class의 상관 관계 파악 후 필요없는 피처제거

- AutoML 오픈소스인 Pycaret이용
  - Pycaret은 빠르게 피처 엔지니어링, 모델 학습, 하이퍼파라미터 튜닝, 예측을 포함하는 전체 프로세스를 실행
  - Pycaret 실행 결과 가장 높은 점수를 얻은 DecisionTreeClassifier, LinearDiscriminantAnalysis, LGBMClassifier, AdaBoostClassifier, GradientBoostingClassifier(+HGB) 을 이용하여 각각 모델링해서 돌려봄

LG Aimers

4.시도해본 전처리/모델

- Feature importance 활용
  - 각 모델마다 Feature importance를 계산하여 연관성이 높은 상위 n개의 feature를 이용해서 모델링

- 결측치 처리
  - IterativeImputer을 이용하여 각 feature들의 상관 계수를 고려하여 예측된 값을 결측된 부분에 채워 넣어줌
  - 참고 : https://www.numpyninja.com/post/how-to-implement-mice-algorithm-using-iterative-imputer-to-handle-missing-values

LG Aimers

감사합니다.

TEAM 틔운돌아와