2023.11.26

**전기 분해를 통한 특정 소재 생산 공정 데이터 분석 보고서**

**윤혜준**

# <목차>

# [1. Data소개 및 분석 목적](#_<data_소개_및)

# [2. EDA](#_<EDA>)

# [3. Data 전처리](#_<data전처리>)

# [4. 모델링 case 1](#_<모델링_case_1>)

# [5. 모델링 case 2](#_<모델링_Case_2>)

# [6. 총정리 및 비즈니스 인사이트](#_<총정리및_비즈니스_인사이트>)

# **<data 소개 및 분석 목적>**

전기 분해를 통해 특정 소재를 생산하는 공정의 데이터이다. train과 test 파일 각각은 동일한 컬럼을 가지고 있으며 다음과 같은 컬럼으로 이루어져 있다.

## 컬럼소개

### 관리변수

#### A: 제품 생산량 (값이 클수록 좋음)

#### B: 최종 제품 형태 전환율 (값이 클수록 좋음)

#### C: 전류효율 (값이 클수록 좋음)

#### D: 폐기물1 농도 (값이 작을수록 좋음)

#### E: 폐기물2 농도 (값이 작을수록 좋음)

### 조업변수

#### 30개의 조업 변수(ex. 온도,압력,속도 등)

## 분석목적

### 관리 변수 5개를 토대로 품질 및 생산성 향상 및 원가 절감을 위한 모델링 및 비즈니스 인사이트 도출

# **<EDA>**

1. Data 형태

* Train data 형태 : (row 개수 : 43865, column 개수: 36)
* Test data 형태 : (row 개수 : 18798, column 개수: 36)

1. 결측치 여부 확인

* 두 data 모두 결측치 없음

1. Column 정의

* oper\_features
  + X1 ~ X30 :30개의 조업 변수(ex. 온도,압력,속도 등)
* control\_features
  + A: 제품 생산량 (값이 클수록 좋음)
  + B: 최종 제품 형태 전환율 (값이 클수록 좋음)
  + C: 전류효율 (값이 클수록 좋음)
  + D: 폐기물1 농도 (값이 작을수록 좋음)
  + E: 폐기물2 농도 (값이 작을수록 좋음)

1. 이상치 시각화 및 제거

관리변수 (control\_features) 에 대해 이상치 시각화

* Train

도표, 평행, 라인, 텍스트이(가) 표시된 사진

자동 생성된 설명

* Test

도표, 평행, 라인, 텍스트이(가) 표시된 사진

자동 생성된 설명

* 이상치 제거 후 데이터 형태 변화 (전 ->후)
  + Train data : (43865, 36) -> (43367, 36)
  + Test data : (18798, 36) -> (17570, 36)

조업변수 (oper\_features) 에 대해 이상치 시각화

* Train 텍스트, 평행, 번호이(가) 표시된 사진

  자동 생성된 설명
* Test 텍스트, 평행, 번호이(가) 표시된 사진

  자동 생성된 설명
* 이상치 제거 후 데이터 형태 변화 (전 ->후)
  + Train data : (43865, 36) -> (41006, 32)
  + Test data : (18798, 36) -> (17570, 32)
  + 관리변수 개수 그대로, 조업변수 개수 4개 감소

1. Feature 별 시각화(각 조업변수와 관리변수 전체 산점도) 후 해석
   * X2 : A 가 750이 넘으면 X2가 일정하게 낮은값 , C 가 50이 넘으면 X2가 일정하게 낮은 값 , E 가 0.04이 넘으면 X2가 일정하게 낮은값
   * X7 : C, E에 대해서 군집화 된 모습
   * X8 : A, B, C, D, E애 대해서 군집화 된 모습
   * X10 : C, E에 대해서 군집화 된 모습
   * X11 : C, E에 대해서 군집화 된 모습
   * X12 : D0.05-0.06 구간에서 급격히 감소후 일정값, E에 대해 일정값 보이다가 0.045부터 급격히 증가
   * X13 : D0.05-0.06 구간에서 급격히 감소후 일정값, E에 대해 일정값 보이다가 0.045부터 급격히 증가
   * X15 : C, E에 대해서 군집화 된 모습, D와는 비례 추세
   * X16 : A, B, C, D, E에 대해서 군집화 된 모습
   * X17 : A, B에 대해 비례상승 추세 , C, D, E에 대해서 군집화 된 모습
   * X18 : A, B에 대해 반비례 추세, D에 대해 비례 추세, C, E에 대해서 군집화 된 모습
   * X19 : A, B, C, D, E에 대해서 군집화 된 모습
   * X20 : A, B, C, E에 대해서 군집화 된 모습, D에 대해 반비례 추세
   * X21 : C, E에 대해서 군집화 된 모습
   * X22 : A, B, C에 대해 일정패턴 보임, D0.05-0.06 구간에서 급격히 감소후 일정값, E에 대해 일정값 보이다가 0.045부터 급격히 증가
   * X23 : C, E에 대해서 군집화 된 모습
   * X24 : C, E에 대해서 군집화 된 모습
   * X25 : A, B, C 에 대해 일정패턴 보임, D0.05-0.06 구간에서 급격히 감소후 일정값, E에 대해 일정값 보이다가 0.045부터 급격히 증가
   * X26 : A, B에 대해 반비례 추세, C, E에 대해서 군집화 된 모습
   * X27 : A, B, C, D, E에 대해서 군집화 된 모습
   * X28 : C, E에 대해서 군집화 된 모습
   * X29 : C, E에 대해서 군집화 된 모습
   * X30 : C, E에 대해서 군집화 된 모습, D0.05-0.06 구간에서 급격히 증가후 일정값, E와 반비례 추세
     + C, E에 대해서 군집화 된 모습 을 띄는 패턴이 많이 보이고, 구간에서 급격히 감소나 증가후 일정 값을 보이는 패턴이 많이 보임
     + 비슷한 조업변수의 특징들로 인해 상관관계 분석으로 상관계수 파악
2. 상관계수 시각화 후 해석

0.00-0.19: 매우 약한 상관관계

0.20-0.39: 약한 상관관계

0.40-0.59: 중간 정도의 상관관계

0.60-0.79: 강한 상관관계

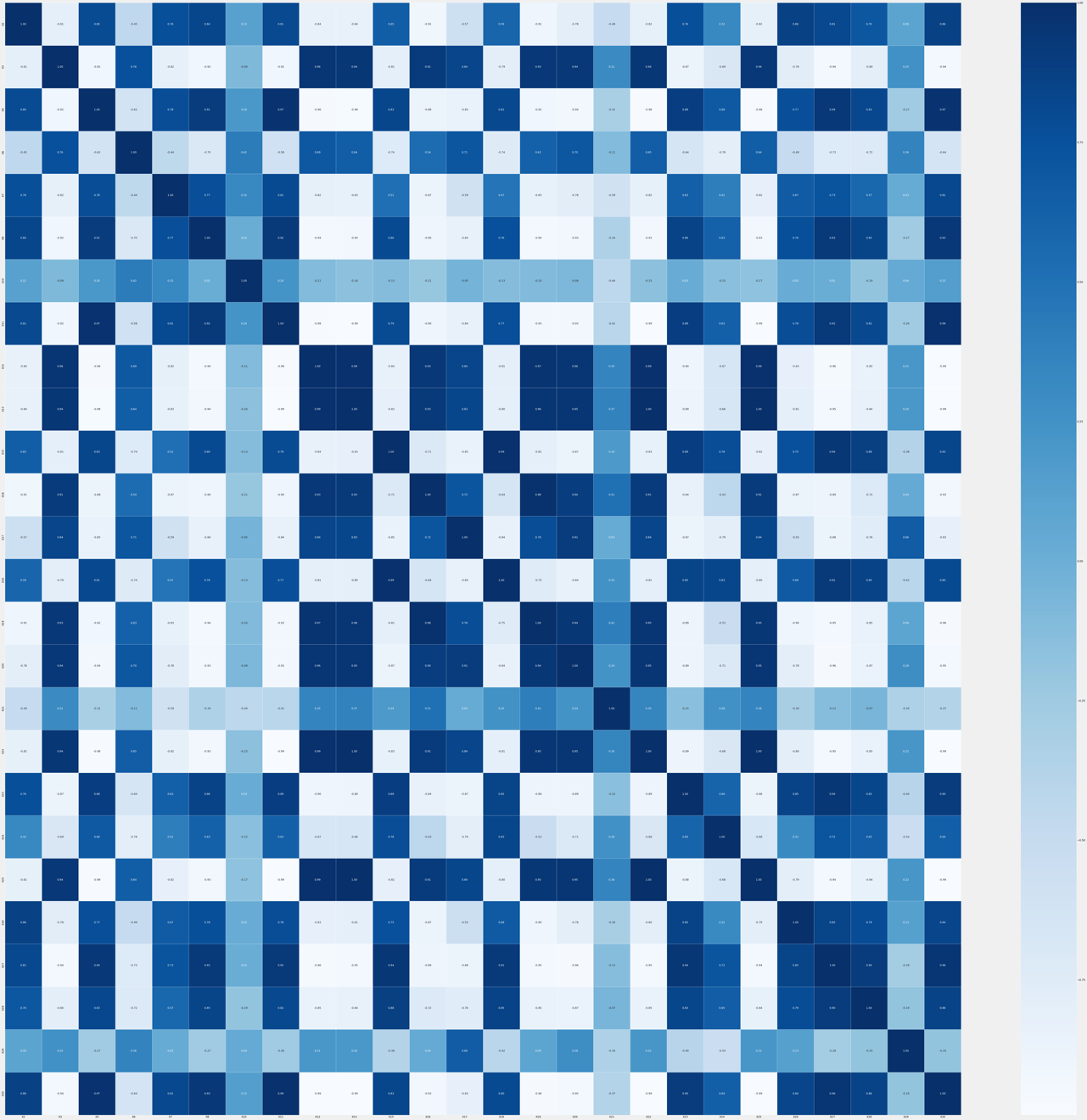
0.80-1.00: 매우 강한 상관관계

1. 관리변수간 상관계수 스크린샷, 직사각형, 사각형, 평행이(가) 표시된 사진

   자동 생성된 설명
2. 관리 변수끼리 대부분이 중간이상의 양/음의 상관관계를 갖고있다.
3. 가장 높은 상관관계는 A, C로 0.93
4. 그 다음 가장 높은 상관관계는 C, E로 0.90
5. 그 다음 높은 상관계수 A, E로 0.79

=> A, B, C는 증가할수록 그리고 D, E는 작아질수록 품질 및 생산성이 좋아지고 원가도 낮아지지만,

A, C 와 C, E처럼 매우 강한 양의 상관관계를 갖고있는 변수들로 보아 A, B, C를 증가시키되 E의 증가또한 염두해야 한다는 것을 알수 있음.

1. 조업변수간 상관계수 
2. 조업 변수끼리 대부분이 중간이상의 양/음의 상관관계를 갖고있다.
3. 예를 들어, 가장 높은 상관관계는 X12, X13으로 0.99의 매우 강한 상관성을 띈다.
4. 산점도에도 나타났듯이 비슷한 패턴을 띄는 조업변수들끼리 강한 상관성이 나타나고 있다.
5. 조업변수와 관리변수 상관계수 히트맵 사각형, 직사각형, 스크린샷, 패턴이(가) 표시된 사진

   자동 생성된 설명
   * 1. 조업 변수와 관리변수 끼리 대부분이 중간이상의 양/음의 상관관계를 갖고있다.
     2. 가장 높은 상관관계는 E, X19으로 0.97의 매우 강한 상관성을 띈다.

==> 특히 A, C, E는 조업변수와 상관계수가 매우크다

1. 다중공선성 해석  
   다중공선성 확인 : 일반적으로 VIF지수 10이상을 다중공선성 확률 높은 변수로 취급
2. 조업변수 다중공선성 확인

텍스트, 스크린샷, 메뉴, 폰트이(가) 표시된 사진

자동 생성된 설명

1. 관리변수에 대해 다중공선성 확인텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진

   자동 생성된 설명
2. **EDA 결론**

#### 관리 변수끼리 대부분이 중간이상의 양/음의 상관관계를 갖고있다.

##### 가장 높은 상관관계는 A, C로 0.93

##### 그다음 가장 높은 상관관계는 C, E로 0.90

##### 그다음 높은 상관계수 A, E로 0.79

##### A, B, C는 증가할수록 그리고 D, E는 작아질수록 품질 및 생산성이 좋아지고 원가도 낮아지지만 A, C 와 C, E처럼 매우 강한 양의 상관관계를 갖고있는 변수들로 보아 A, B, C를 증가시키되 E의 증가또한 염두해야 한다는 것을 알수 있음.

#### 조업 변수끼리 대부분이 중간이상의 양/음의 상관관계를 갖고있다.

##### 가장 높은 상관관계는 X12, X13으로 0.99의 매우 강한 상관성을 띈다.

##### 산점도에도 나타났듯이 비슷한 패턴을 띄는 조업변수들끼리 강한 상관성이 나타나고 있다.

#### 조업 변수와 관리변수 끼리 대부분이 중간이상의 양/음의 상관관계를 갖고있다.

##### 가장 높은 상관관계는 E와 X19으로 0.97의 매우 강한 상관성을 띈다. ==> 특히 A, C, E는 조업변수와 상관계수가 매우크다.

#### 변수들끼리의 다중공선성이 나타나고, 중요도를 파악하여 인사이트를 얻기 위해 때문에 앙상블 모델링 진행.

#### train, test data의 outlier에 대해 처리를 진행했지만 분포가 다른 변수들이 많은것으로 보아 train에 대한 model의 과적합 예상된다.

#### 변수별 중요도를 앙상블 모델들에서 추출 후 가장 많이 나타나는 변수들이 관리변수를 최적화 하는데 중요한 요인으로 파악될것으로 보인다.

#### Case1(개별 관리변수에 대한 모델링)과 Case2(관리변수 통합하여 새로운 관리변수에 대한 모델링)를 나누어 진행하여 개별 관리변수에 대한 모델의 설명력 뿐만 아니라, 통합 관리변수 모델링의 설명력 또한 확인한다.

# **<data전처리>**

## Train 데이터에 대해 8:2로 train, val 데이터 split

## Train, Test데이터에 대해 조업변수 X, 관리변수 y로 설정

## Normalization 진행 ( 평균 0 , 표준편차 1)

### Train 데이터 X에 대해 표준화를 적용하고Test 데이터의 X에 대해 서도 표준화를 적용

## Y(관리변수)변수 변형시키기

### 모델링2를 위해서 관리변수들을 normalization 후 new\_y로 변형

### 각 관리 변수의 중요도가 같다고 가정

### new\_y = A + B + C – D – E

# **<모델링 case 1>**

## 관리변수 (A, B, C, D, E)를 각각 결과변수로 모델링 진행 ex) y\_train\_A, y\_val\_A = y\_train['A'], y\_val['A']

## Light GBM Regression모델 사용 LightGBM Regression은 그래디언트 부스팅 알고리즘을 활용하여 회귀 문제를 해결하는 모델이며 여러 결정 트리를 순차적으로 학습하여 예측 성능을 향상시킨다.

### Validation data 에 대한 성능텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진 자동 생성된 설명

### Test data에 대한 성능

### 텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진 자동 생성된 설명

## RandomForest 모델 사용

### Validation data 에 대한 성능 텍스트, 폰트, 스크린샷, 블랙이(가) 표시된 사진 자동 생성된 설명

### Test data에 대한 성능

### 텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진 자동 생성된 설명

### LGBM, RandomForest 성능비교

### 사용한 지표 : MAE : 실제 값과 예측 값 간의 절대적인 차이를 평균한 값

### R-squared : R²는 회귀 모델의 설명력을 나타내는 지표로서 종속 변수의 총 변동 중 모델에 의해 설명되는 비율이다. R² 값이 1에 가까울수록 모델이 데이터를 잘 설명한다고 볼 수 있다.

### MSE : 평균 제곱 오차를 나타내는 지표로, 실제 값과 예측 값 간의 제곱된 차이를 평균한 값, 값이 작을수록 모델의 예측이 정확하다고 평가

### R-squared : R²는 회귀 모델의 설명력을 나타내는 지표로서 종속 변수의 총 변동 중 모델에 의해 설명되는 비율이다. R² 값이 1에 가까울수록 모델이 데이터를 잘 설명한다고 볼 수 있다.

### test data에 대하여 두가지모델을 MAE, R^2, MSE, RMSE 를 지표로 사용하여 비교했을 때, 전반적으로 LGBM이 모델을 더 잘 설명하고 있음을 알 수 있다.

### 특히 C, D, E에 대해서는 설명력(R-squared)도 매우 높으며 예측성능 또한 매우 높다.

### 더 성능이 높은 LGBM모델에 대해서 중요성 높은 조업변수 찾는 과정 진행

## LGBM 모델에서 중요도 높은 조업변수 탐색

### 관리변수 A를 결과변수로 사용한 모델텍스트, 그래프, 스크린샷, 라인이(가) 표시된 사진 자동 생성된 설명

#### 중요도가 가장 높은 변수 5개 선택 후 산포도 시각화 라인이(가) 표시된 사진 자동 생성된 설명

#### 선택된 특성들과 A 변수 간의 상관 계수 계산 X11 X10 X12 X17 X7

#### -0.762789 0.021186 0.813873 0.707558 -0.653501

### 관리변수 B를 결과변수로 사용한 모델 텍스트, 그래프, 라인, 스크린샷이(가) 표시된 사진 자동 생성된 설명

#### 중요도가 가장 높은 변수 5개 선택 후 산포도 시각화 도표, 라인, 폰트, 지도이(가) 표시된 사진 자동 생성된 설명

#### 선택된 특성들과 A 변수 간의 상관 계수 계산 X10 X11 X7 X13 X4

#### 0.192683 -0.436761 -0.189201 0.452635 -0.460702

### 관리변수 C를 결과변수로 사용한 모델 텍스트, 그래프, 라인, 스크린샷이(가) 표시된 사진 자동 생성된 설명

#### 중요도가 가장 높은 변수 5개 선택 후 산포도 시각화 라인, 그래프, 폰트이(가) 표시된 사진 자동 생성된 설명

#### 선택된 특성들과 A 변수 간의 상관 계수 계산 X10 X12 X11 X7 X21

### -0.012691 0.941277 -0.899078 -0.734221 0.183341

### 관리변수 D를 결과변수로 사용한 모델 텍스트, 그래프, 라인, 스크린샷이(가) 표시된 사진 자동 생성된 설명

#### 중요도가 가장 높은 변수 5개 선택 후 산포도 시각화 라인, 폰트, 텍스트이(가) 표시된 사진 자동 생성된 설명

#### 선택된 특성들과 A 변수 간의 상관 계수 계산 X12 X10 X11 X18 X7

### -0.689538 -0.146162 0.631502 0.917352 0.390517

### 관리변수 E를 결과변수로 사용한 모델 텍스트, 그래프, 라인, 스크린샷이(가) 표시된 사진 자동 생성된 설명

#### 중요도가 가장 높은 변수 5개 선택 후 산포도 시각화 라인, 폰트이(가) 표시된 사진 자동 생성된 설명

#### 선택된 특성들과 A 변수 간의 상관 계수 계산 X10 X12 X11 X7 X8

### -0.189024 0.939882 -0.936909 -0.818233 -0.903058

## **모델링 CASE 1 - Insight 정리**

### A, B, C에 대하여

#### A, B, C는 클수록 공정에서 좋은 결과를 가져오는 관리변수이기때문에, 양의상관관계를 갖고있는 변수들을 증가시키고, 음의 상관관계를 갖는 변수들을 감소시킬 필요가 있다.

#### A와 각 변수에 대한 상관계수는 X11, X10, X12, X17, X7에 대하여 -0.762789 0.021186 0.813873 0.707558 -0.653501 이고, 따라서 X11, X7을 감소시키고, X12, X17를 증가하는 방향으로 변화 시키면 A값이 커지는 방향으로 변화될것으로 예상할 수 있다.

#### B와 각 변수에 대한 상관계수는 X10, X11, X7, X13, X4에 대하여 0.192683 -0.436761 -0.189201 0.452635 -0.460702 이고, 따라서 X11, X7, X4을 감소시키고, X10, X13를 증가하는 방향으로 변화 시키면 B값이 커지는 방향으로 변화될것으로 예상할 수 있다.

#### C와 각 변수에 대한 상관계수는 X10, X12, X11, X7, X21에 대하여

### -0.012691 0.941277 -0.899078 -0.734221 0.183341 이고,

### 따라서 X11, X7을 감소시키고, X12, X21를 증가하는 방향으로 변화 시키면 C값이 커지는 방향으로 변화될것으로 예상할 수 있다.

### D, E에 대하여

### D, E는 작을수록 공정에서 좋은 결과를 가져오는 관리변수이기때문에, 음의 상관관계를 갖고있는 변수들을 증가시키고, 양의 상관관계를 갖는 변수들을 감소시킬 필요가 있다.

### D 와 각 변수에 대한 상관계수는

### X12, X10, X11, X18, X7에 대하여

### -0.689538 -0.146162 0.631502 0.917352 0.390517 이고,

### 따라서 X11, X18, X7을 감소시키고, X12, X10를 증가하는 방향으로 변화 시키면 D값이 작아지는 방향으로 변화될것으로 예상할 수 있다.

### E 와 각 변수에 대한 상관계수는X10, X12, X11, X7, X8 에 대하여

### -0.189024 0.939882 -0.936909 -0.818233 -0.903058 이고,

### 따라서 X12 을 감소시키고, X10, X11, X7, X8를 증가하는 방향으로 변화 시키면 E값이 커지는 방향으로 변화될것으로 예상할 수 있다.

# **<모델링 Case 2>**

## new\_Y (A + B + C – D - E)에 대해 모델링 진행

## Light GBM Regression모델 사용 LightGBM Regression은 그래디언트 부스팅 알고리즘을 활용하여 회귀 문제를 해결하는 모델이며 여러 결정 트리를 순차적으로 학습하여 예측 성능을 향상시킨다.

### Validation data 에 대한 성능텍스트, 폰트, 스크린샷이(가) 표시된 사진 자동 생성된 설명

### Test data에 대한 성능텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진 자동 생성된 설명

## RandomForest 모델 사용

### Validation data 에 대한 성능 텍스트, 폰트, 스크린샷이(가) 표시된 사진 자동 생성된 설명

### Test data에 대한 성능

### 텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진 자동 생성된 설명

### LGBM, RandomForest 성능비교

### 사용한 지표 : MAE : 실제 값과 예측 값 간의 절대적인 차이를 평균한 값

### R-squared : R²는 회귀 모델의 설명력을 나타내는 지표로서 종속 변수의 총 변동 중 모델에 의해 설명되는 비율이다. R² 값이 1에 가까울수록 모델이 데이터를 잘 설명한다고 볼 수 있다.

### MSE : 평균 제곱 오차를 나타내는 지표로, 실제 값과 예측 값 간의 제곱된 차이를 평균한 값, 값이 작을수록 모델의 예측이 정확하다고 평가

### R-squared : R²는 회귀 모델의 설명력을 나타내는 지표로서 종속 변수의 총 변동 중 모델에 의해 설명되는 비율이다. R² 값이 1에 가까울수록 모델이 데이터를 잘 설명한다고 볼 수 있다.

### test data에 대하여 두가지모델을 MAE, R^2, MSE, RMSE 를 지표로 사용하여 비교했을 때, 전반적으로 LGBM이 모델을 더 잘 설명하고 있음을 알 수 있다. 특히 모델의 설명력이 70%이상으로 매우 높다.

### 더 성능이 높은 LGBM모델에 대해서 중요성 높은 조업변수 찾는 과정 진행

## LGBM 모델에서 중요도 높은 조업변수 탐색

### new\_y에 대한 LGBM모델내의 조업변수 중요도텍스트, 스크린샷, 폰트이(가) 표시된 사진 자동 생성된 설명

### 중요도가 가장 높은 변수 5개 선택 및 산포도 시각화 라인, 그래프, 폰트이(가) 표시된 사진 자동 생성된 설명

### 새로운 y와 중요도가 높은 변수들 간의 상관계수 계산

X10 X11 X7 X12 X4

0.187824 -0.627949 -0.402429 0.698011 -0.661533

## **모델링 CASE 2 - Insight 정리**

### 새로 조합한 new\_y = A + B + C - D – E는 클수록 공정에서 좋은 결과를 가져오는 관리변수이기때문에, 양의상관관계를 갖고있는 변수들을 증가시키고, 음의 상관관계를 갖는 변수들을 감소시킬 필요가 있다.

### new\_y 는 X10, X11, X7, X12, X4에 대하여 다음과 같은 상관계수를 보인다.

### 0.187824 -0.627949 -0.402429 0.698011 -0.661533

### - 따라서, new\_y와 양의 상관계수를 갖고있는 X10, X12를 증가시키고, 음의상관계수를 갖고있는 X11, X7, X4를 감소하는 방향으로 변화시킨다면 new\_y를증가하는 방향으로 변화시킬 수 있을것이다.

# **<총정리 및 비즈니스 인사이트>**

## 모델링 case 1 – 관리변수별 모델링

### A (제품 생산량) 의 값을 증가시키기 위해 : X11, X7을 감소시키고, X12, X17를 증가시킨다.

### B (최종 제품 형태 전환율) 의 값을 증가시키기 위해 : X11, X7, X4을 감소시키고, X10, X13를 증가시킨다.

### C (전류효율) 의 값을 증가시키기 위해 : X11, X7을 감소시키고, X12, X21를 증가시킨다.

### D (폐기물1 농도) 의 값을 감소시키기 위해 : X11, X18, X7을 감소시키고, X12, X10를 증가시킨다.

### E (폐기물2 농도) 의 값을 감소시키기 위해 : X12 을 감소시키고, X10, X11, X7, X8를 증가시킨다.

|  |  |  |
| --- | --- | --- |
| 관리 변수 | 감소시킬 조업 변수 | 증가시킬 조업 변수 |
| A(제품 생산량) | X11, X7 | X12, X17 |
| B(최종 제품 형태 전환율) | X11, X7, X4 | X10, X13 |
| C(전류효율) | X11, X7 | X12, X21 |
| D(폐기물1 농도) | X11, X18, X7 | X12, X10 |
| E(폐기물2 농도) | X12 | X10, X11, X7, X8 |

## 모델링 case 2 – 관리변수 통합 후 모델링

### New\_y = A+B+C-D-E

|  |  |  |
| --- | --- | --- |
| 관리 변수 | 감소시킬 조업 변수 | 증가시킬 조업 변수 |
| New\_y | X11, X7, X4 | X12, X10 |

### X11, X7, X4를 감소시키고, X12, X10을 증가시킨다.

# 

## **최종Insight**

### case 1의 결과를 종합해 보면, case2가 각 관리변수들 간의 상호작용을 고려한 통합적인 모델이며 설명력이 뛰어나고 신뢰도가 높은 모델임을 알 수 있다.

### 개별 관리변수의 관리를 위해서는 case1의 모델의 insight를 활용하여 A,B,C의 값을 크게 만들고 D, E의 값을 작게 만들기 위해 조업변수들을 조절할 수 있을 것 이다.

### 관리 변수 5개는 밀접한 상호작용을 갖고 있는 변수들이기 때문에 case 2의 를 사용하여 전반적인 품질향상과 생산성향상에 기여할 수 있을 것이며, 원가 절감에도 기여할 수 있을 것이다.

### 관리 변수 E의 경우 중요도 높은 조업변수의 경향이 나머지 관리 변수와 다른 것이 존재하기 때문에 이점을 유의하여 조업변수의 변화가 필요하다.