LS

1 page

LS Bigdata School

1주차 프로젝트 결과 보고서

- 제조 데이터 분류 분석

2조

강하연 김지유 박민서 이채은 정재성 정하연

1. 프로젝트 개요 및 팀 구성

3C analysis | 2 page

프로젝트 주제 선정 및 개요

품질 개선을 위한 제조 공정의
결함 원인 분석 및 불량품 예측 모델 구축

데이터 안내

- 실제 제조과정에서 나온 데이터로
  익명처리되어 식별이 불가능
- 제조업 분야에서는 **데이터 불균형**이 자주 발생

분석방법 & 알고리즘

- AUTOVIZ
- HEATMAP
- 의사결정나무
- EXTRA TREES CLASSIFIER
- SHAP
- PYCARET

1. 프로젝트 개요 및 팀 구성

3C analysis | 2-1 page

구분	기간	활동	비고
요구사항 정의 및 분석 기획	5월 14일	문제 정의, 비즈니스 관점 요구사항 정의 등	아이디어 회의
데이터 전처리 및 탐색적 분석(EDA)	5월 14일 ~ 5월 15일	전처리 -> 결측치 / 이상치 확인, 제거기준(상위 하위 10%) 데이터 시각화 (AUTOviz)	산점도, 상자수염그림 등 다양한 시각화 활용
모델 학습 및 선택	5월 16일	PYCARET,SHAP, Feature Importance	
결론도출	5월 16일	의사결정나무 활용	

활용방안

01. 원인분석을 통한 결함 예방:

- 과거 제조 과정 데이터를 분석하여 주요 결함 요인을 식별하고, 생산 과정에서 잠재적인 문제를 사전에 예측합니다.

- 예측된 결함으로 인한 생산 중단 시간을 줄이고 생산 효율성을 높입니다.

02. 불량품 예측 모델 생성:

- 불량품 발생 가능성이 높은 제품을 사전에 감지하여 생산 과정을 조정하고, 불량률을 최소화합니다.

1. 프로젝트 개요 및 팀 구성

3C analysis | 4 page

기대효과

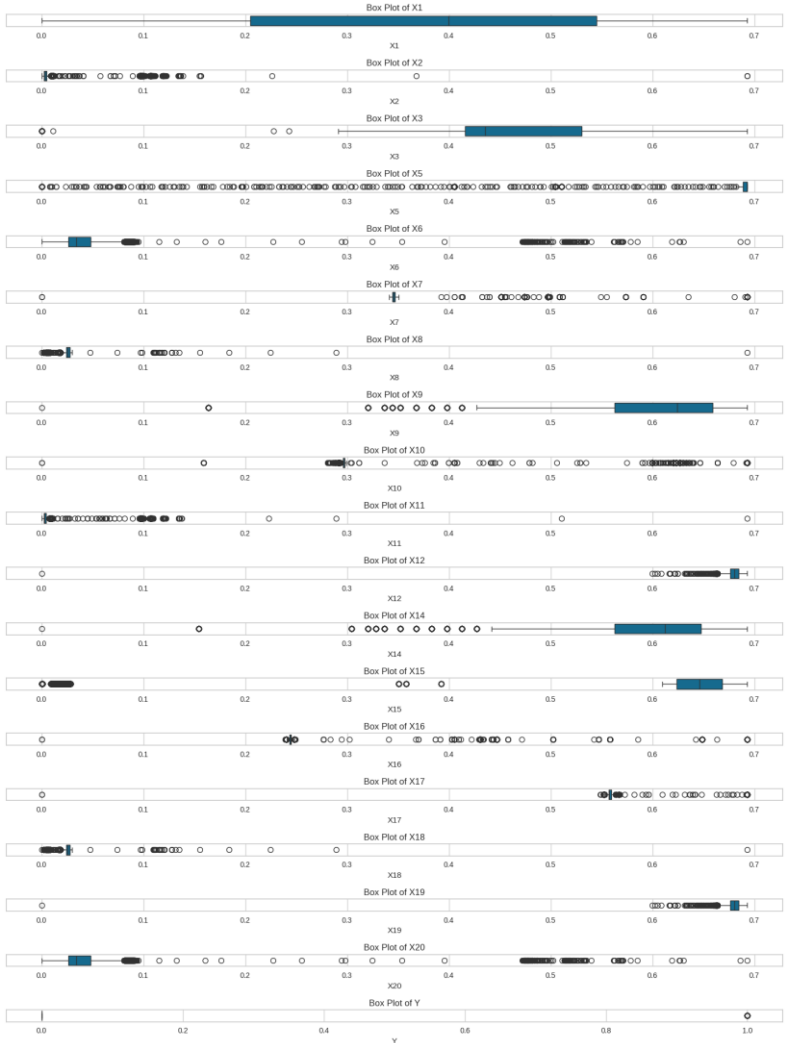
생산성 향상	결함 예방을 통해 생산 과정의 효율성 향상
불량률 감소	불량품 예측 모델을 통해 불량품 발생을 최소화
비용 절감	예방적인 조치로 인한 재작업 및 폐기 비용 절감

2. EDA 및 전처리

3C analysis

5 page

각 변수들의 통계 분석

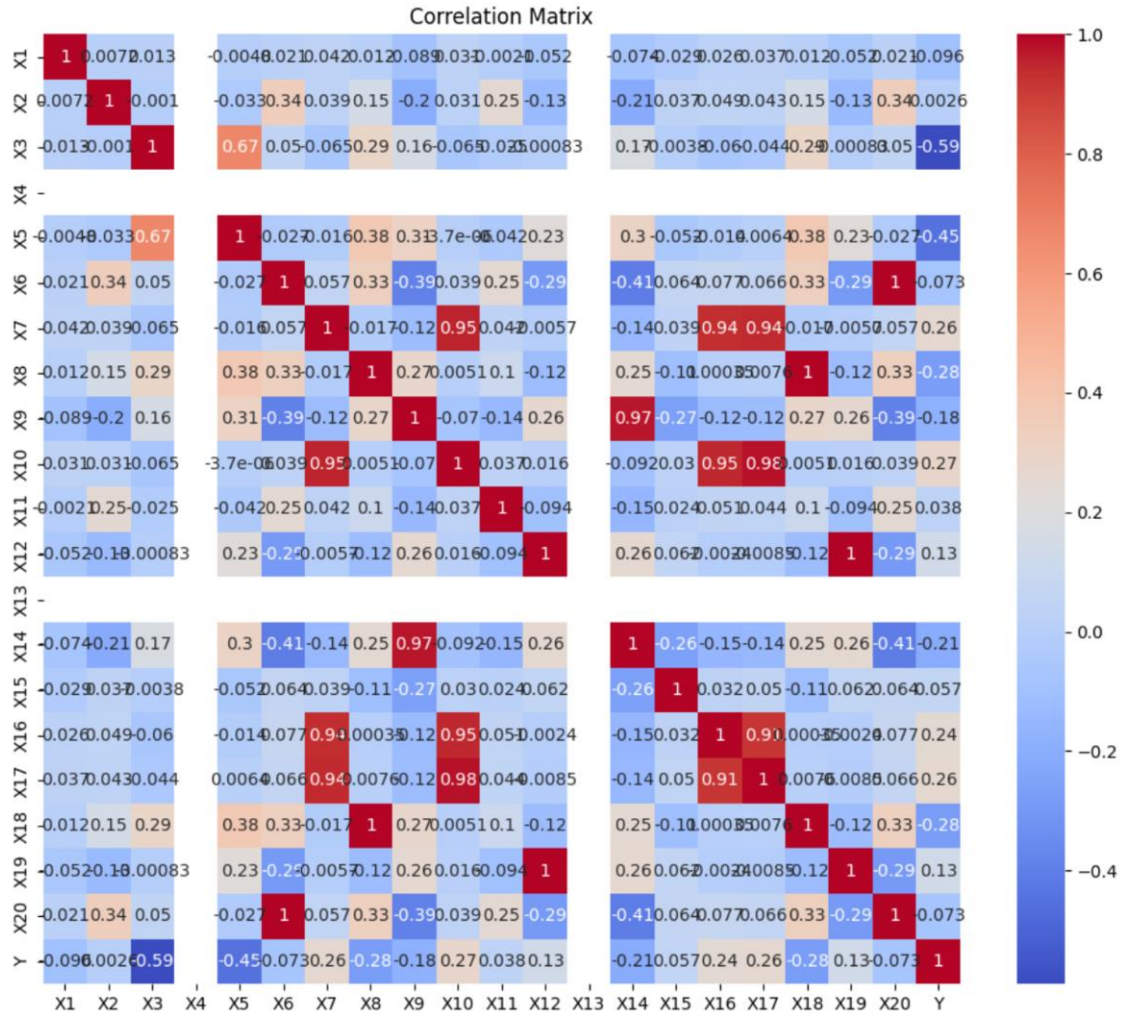


- x1 - x20, y값이 모두 수치형 데이터
- 한 변수당 527,000개의 데이터
- y=0이 양품, y=1이 불량품

2. EDA 및 전처리

3C analysis | 6 page

상관성 분석 (히트맵)



- 상관계수 = 1
  - ▶두 변수 사이에 완벽한 양의 선형 관계

  x6 : x20
  x8 : x18
  x12 : x19

- x4, x13이 비어있음
  - ▶x4, x13은 모두 같은 값 [고정값]

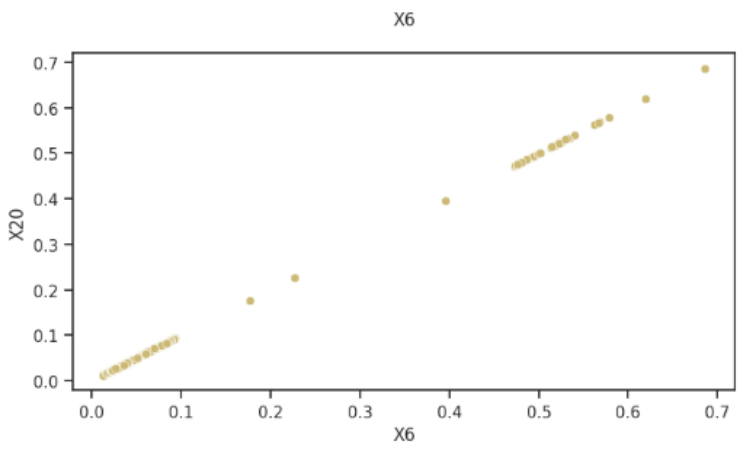
  x4 = 0.015348

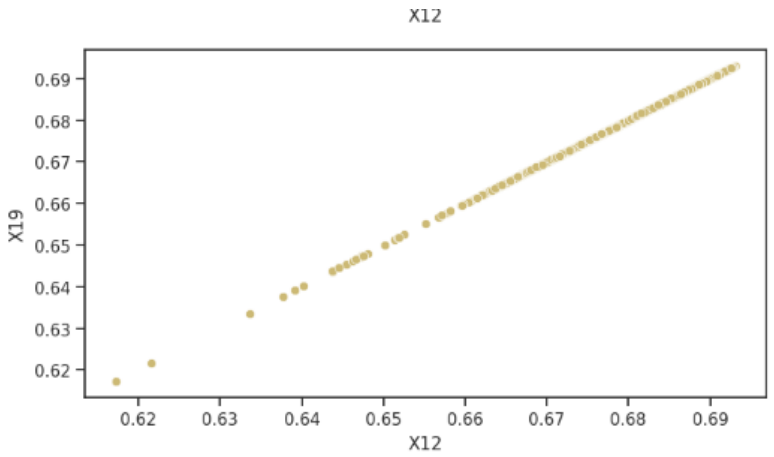
  x13 = 0.249262

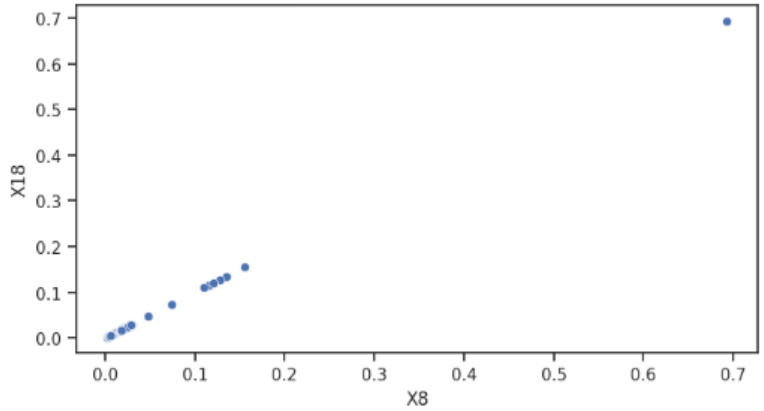
2. EDA 및 전처리

3C analysis | 7 page

경향 분석 (스캐터)







- 오토비즈 시각화로 양의 선형관계 발견
  - x6 : x20
  - x12 : x19
  - x8 : x18

2. EDA 및 전처리

3C analysis | 8 page

코드를 통해 중복된 컬럼 확인

```
# 완전히 같은 데이터인지 확인
identical_columns = []
for col1, col2 in high_correlation_pairs:
    if df[col1].equals(df[col2]):
        identical_columns.append((col1, col2))

print("Identical columns:", identical_columns)
```

```
High correlation (1.0) pairs: [('X6', 'X20'), ('X8', 'X18'), ('X12', 'X19'), ('X18', 'X8'), ('X19', 'X12'), ('X20', 'X6')]
Identical columns: [('X6', 'X20'), ('X8', 'X18'), ('X12', 'X19'), ('X18', 'X8'), ('X19', 'X12'), ('X20', 'X6')]
```

- X20 == X6
- X19 == X12
- X18 == X8

▸ X20, X19, X18 컬럼 제거

```
[47] df['X4'].value_counts()

X4
0.015348    527000
Name: count, dtype: int64

df['X13'].value_counts()

X13
0.249262    527000
Name: count, dtype: int64
```

- X4, X13는 다 같은 값 [고정값]
- X4 = 0.015348
- X13 = 0.249262

▸ X4, X13 컬럼 제거

2. EDA 및 전처리

3C analysis

9 page

T-test 통계적 유의성 검정

```
## T-test 검정
import scipy.stats

t_test = []

for idx, col in enumerate(df.columns):
    t = scipy.stats.ttest_ind(df[df['Y']==0][col],
                              df[df['Y']==1][col])
    t_test.append([col, t[0], t[1]])

t_test_df = pd.DataFrame(t_test, columns=['col', 't', 'p-value'])

## P-value가 0.05보다 작은 변수만 추출
t_test_df = t_test_df[t_test_df['p-value']<0.05]
t_test_df
```

	col	t	p-value
0	X1	70.308079	0.000000e+00
2	X3	530.696338	0.000000e+00
3	X5	363.079979	0.000000e+00
4	X6	52.977659	0.000000e+00
5	X7	-192.741835	0.000000e+00
6	X8	209.101969	0.000000e+00
7	X9	135.787155	0.000000e+00
8	X10	-207.520706	0.000000e+00
9	X11	-27.394397	4.179607e-165
10	X12	-97.017079	0.000000e+00
11	X14	153.808454	0.000000e+00
12	X15	-41.560309	0.000000e+00
13	X16	-181.438928	0.000000e+00
14	X17	-191.516989	0.000000e+00
15	Y	-inf	0.000000e+00

- **T-test를 통해 통계적 유의성 검정**
  - **X2 컬럼 제거**

2. EDA 및 전처리

X2, x4, x13, x18, x19, x20 총 6개 컬럼 제거

X2 : 통계적으로 유의하지 않음

X4 : 0.015348 고정값

X13 : 0.249262 고정값

X18 : x8 컬럼과 동일값

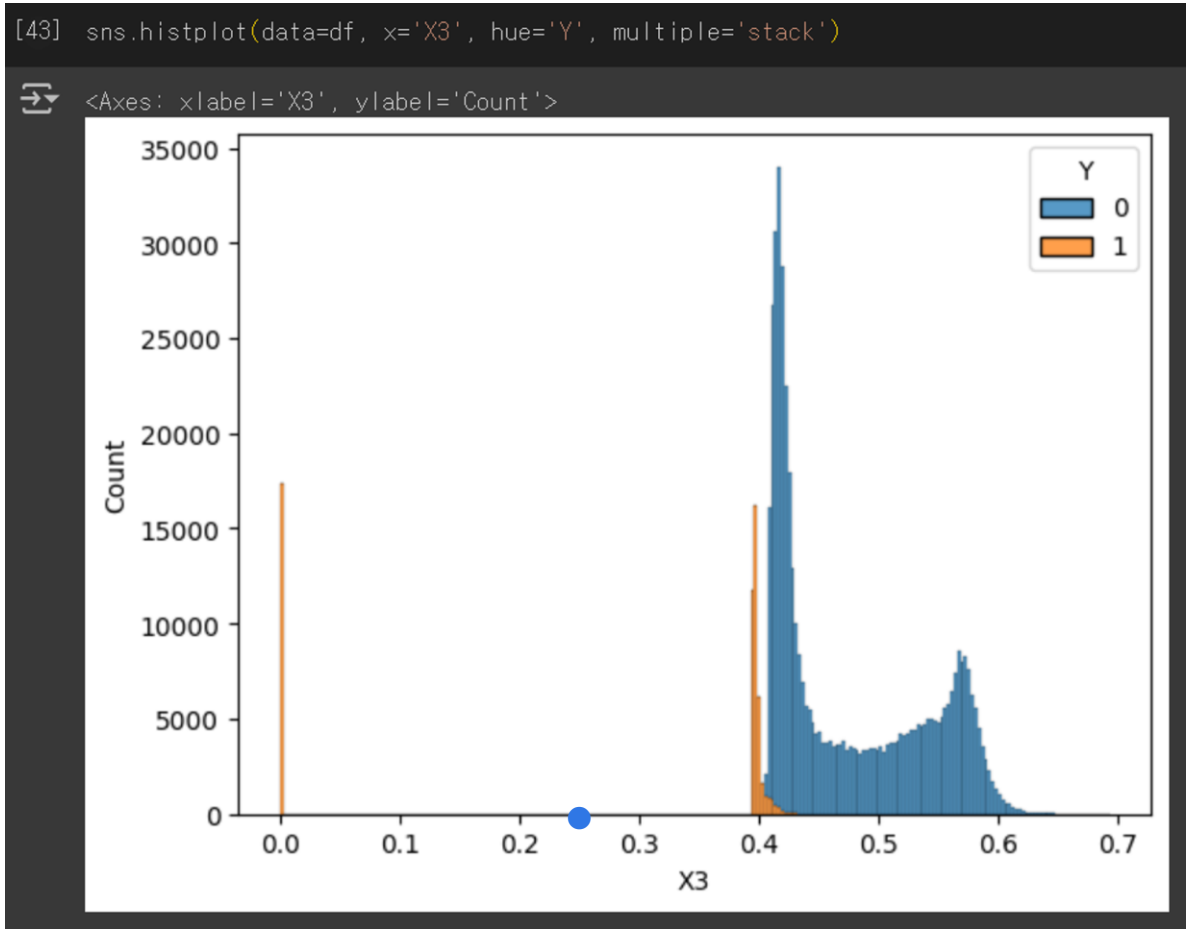
X19 : x12 컬럼과 동일값

X20 : x6 컬럼과 동일값

3. 유효변수 및 임계값 파악

X3 임계값

```
[43] sns.histplot(data=df, x='X3', hue='Y', multiple='stack')
```



- X3<=0.405339 일 때,

모든 Y변수는 1

즉, X3가 0.450339 이하라면

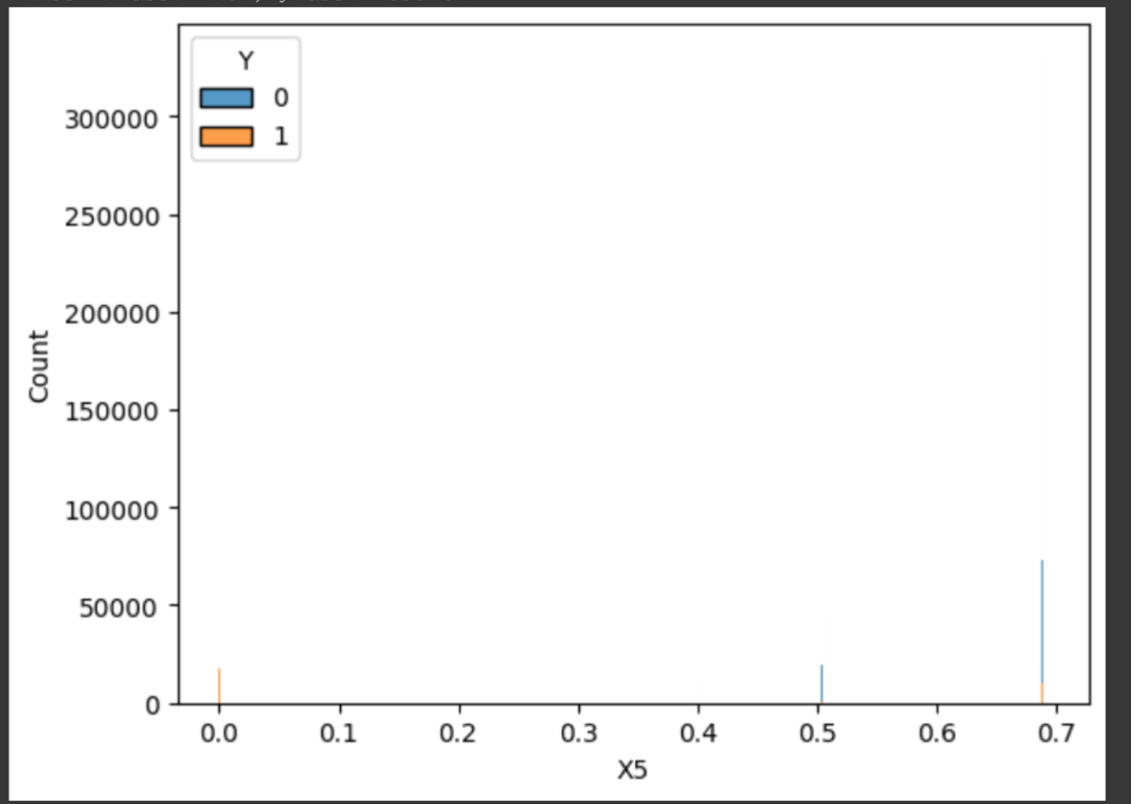
불량발생 (53,352개)

3. 유효변수 및 임계값 파악

X5 임계값

```
sns.histplot(data=df, x='X5', hue='Y', multiple='stack')
```

```
<Axes: xlabel='X5', ylabel='Count'>
```



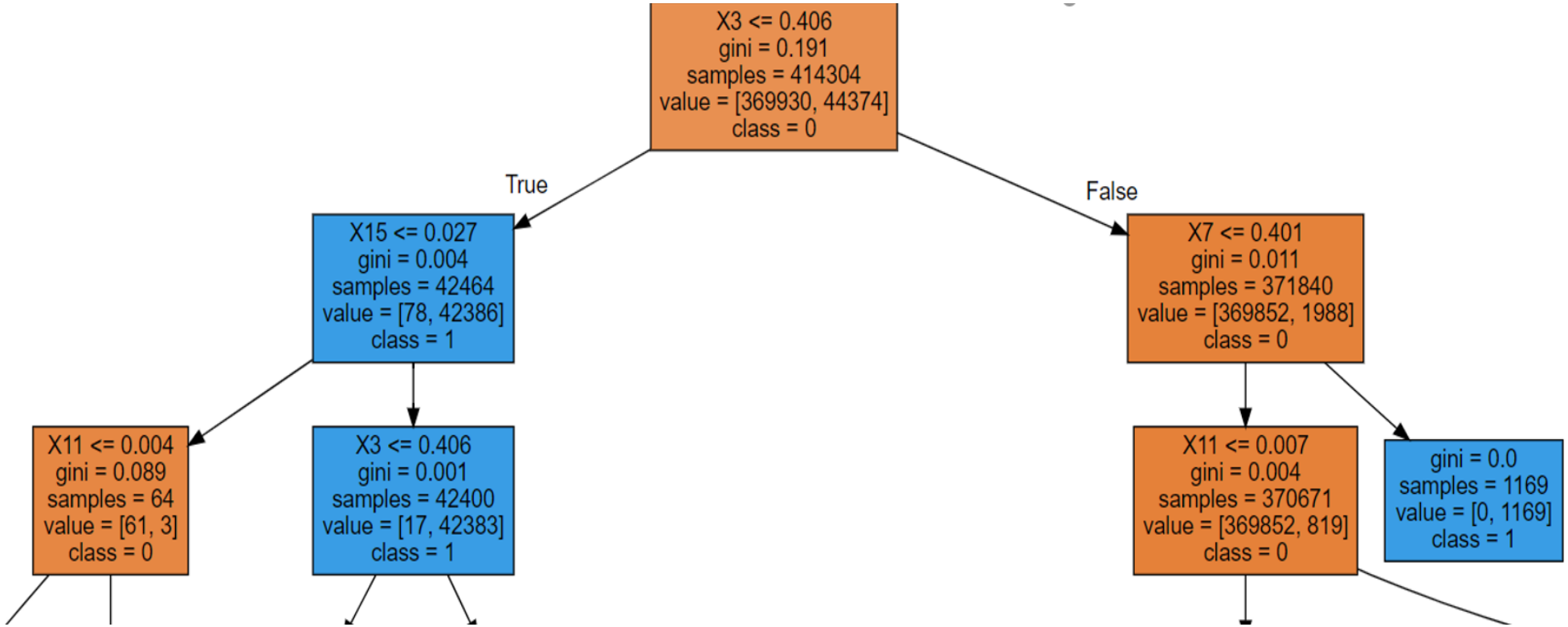
- X5<=0.4 일때

모든 Y변수는 1

즉, X5가 0.4 이하라면 불량발생 (17,496개)

3. 유효변수 및 임계값 파악

의사 결정 나무



- 의사결정 나무를 통해, x3 과 x17이 핵심변수라는 것을 알 수 있다.

4. 솔루션

X3 임계값 이하가 되지 않도록 주의

```
print(df['Y'].value_counts())
print(df['Y'].value_counts(normalize = True))
```

```
Y
0    470000
1     57000
Name: count, dtype: int64
Y
0    0.891841
1    0.108159
Name: proportion, dtype: float64
```

● 전체 데이터 527,000개

이 중 57,000개인 10.8159%가 불량

```
df_X3_0_43=df[df['X3']<=0.405339]
df_X3_0_43['Y'].value_counts()
```

```
Y
1    53352
Name: count, dtype: int64
```

● x3가 0.450339 이하인 모든 데이터가 불량,
총 53,352개 (10.1237%가 불량)

즉, 불량품 중 93.6%는 x3값이 0.450993 이하

X3이 임계값 보다 작아지지 않도록 관리하는 것이 중요

4. 솔루션

X5 임계값 이하가 되지 않도록 주의

```
print(df['Y'].value_counts())
print(df['Y'].value_counts(normalize = True))
```

```
Y
0    470000
1     57000
Name: count, dtype: int64
Y
0    0.891841
1    0.108159
Name: proportion, dtype: float64
```

● 전체 데이터가 527,000개인데,

이 중 57,000개인 10.8159%가 불량

```
df_X5=df[df['X5']<=0.4]
df_X5['Y'].value_counts()
```

```
Y
1    17496
Name: count, dtype: int64
```

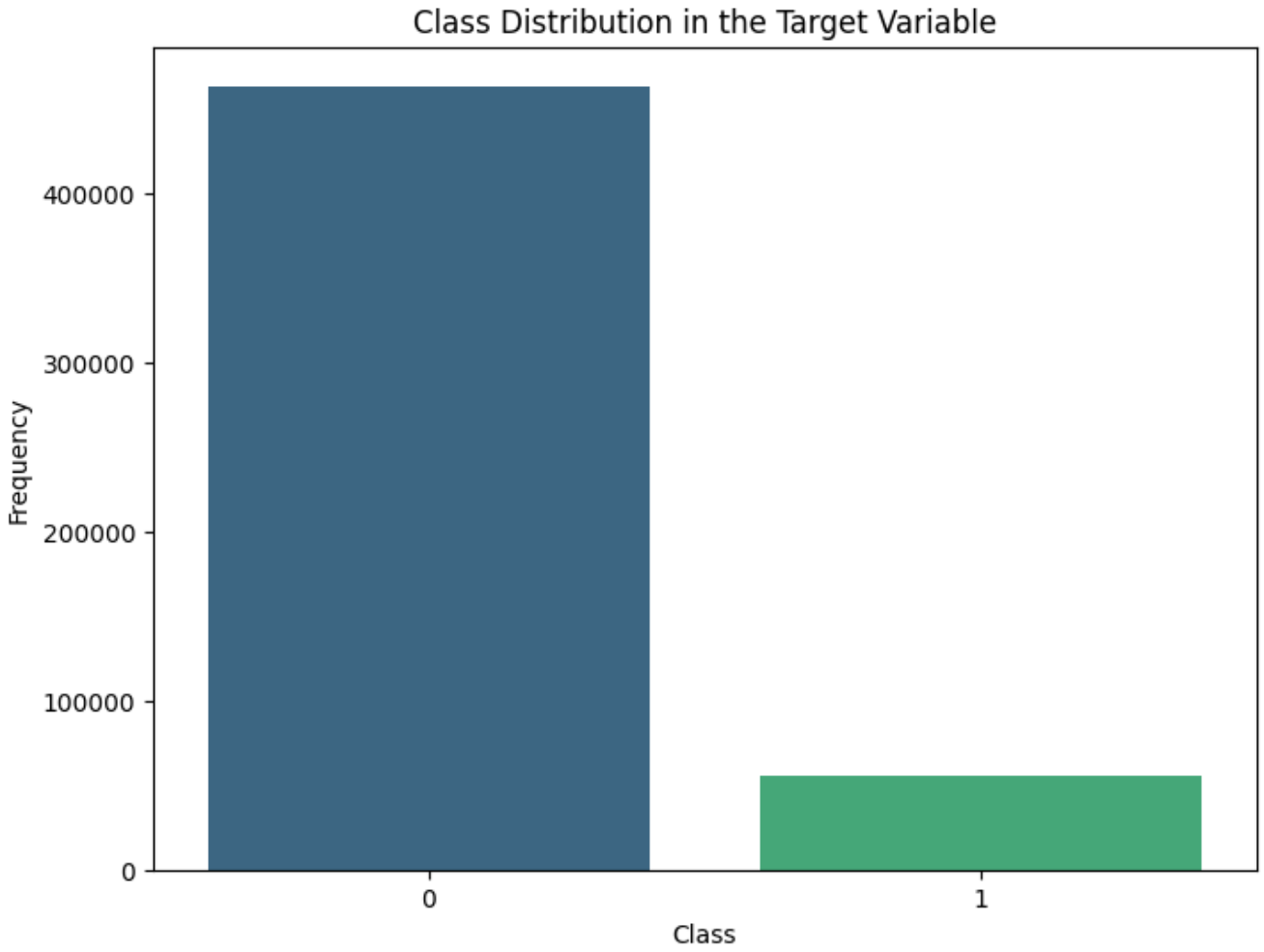
● X5가 0.4 이하인 모든 데이터가 불량이고,
총 17,496개

즉, 전체 불량품 중 30.7%는 X5 값이 0.4 이하

X5가 임계값 보다 작아지지 않도록 관리하는 것이 중요

5. 모델링

데이터 불균형 파악



5. 모델링

3C analysis

17 page

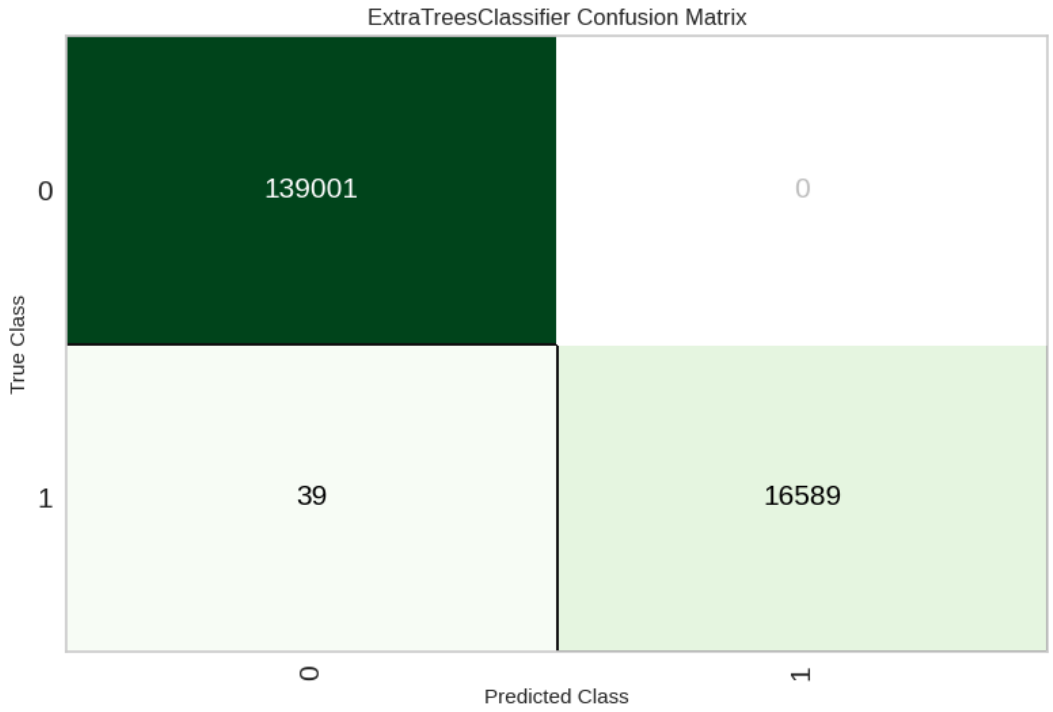
F1 스코어

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
et	Extra Trees Classifier	0.9997	0.9999	0.9976	1.0000	0.9988	0.9987	0.9987
rf	Random Forest Classifier	0.9997	0.9998	0.9975	0.9999	0.9987	0.9985	0.9985
xgboost	Extreme Gradient Boosting	0.9997	0.9998	0.9975	0.9998	0.9987	0.9985	0.9985
knn	K Neighbors Classifier	0.9995	0.9988	0.9959	0.9997	0.9978	0.9975	0.9975
dt	Decision Tree Classifier	0.9994	0.9983	0.9969	0.9975	0.9972	0.9968	0.9968
lightgbm	Light Gradient Boosting Machine	0.9994	0.9993	0.9969	0.9973	0.9971	0.9968	0.9968
gbc	Gradient Boosting Classifier	0.9993	0.9992	0.9940	0.9995	0.9968	0.9964	0.9964
ada	Ada Boost Classifier	0.9990	0.9995	0.9919	0.9984	0.9951	0.9945	0.9945
lr	Logistic Regression	0.9984	0.9985	0.9868	0.9983	0.9925	0.9916	0.9916
svm	SVM - Linear Kernel	0.9982	0.9983	0.9847	0.9980	0.9913	0.9903	0.9903
qda	Quadratic Discriminant Analysis	0.9940	0.9980	0.9444	0.9999	0.9713	0.9680	0.9685
nb	Naive Bayes	0.9342	0.9389	0.3838	0.9999	0.5546	0.5266	0.5978
ridge	Ridge Classifier	0.9331	0.9803	0.3778	0.9899	0.5468	0.5184	0.5893
lda	Linear Discriminant Analysis	0.9329	0.9803	0.3778	0.9838	0.5459	0.5173	0.5872
dummy	Dummy Classifier	0.8932	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

- pycaret을 통해 모델탐색
- F1 스코어 : 정밀도와 재현율을 조화 평균, 불균형 데이터에서의 모델의 성능을 평가하는데 유용

5. 모델링

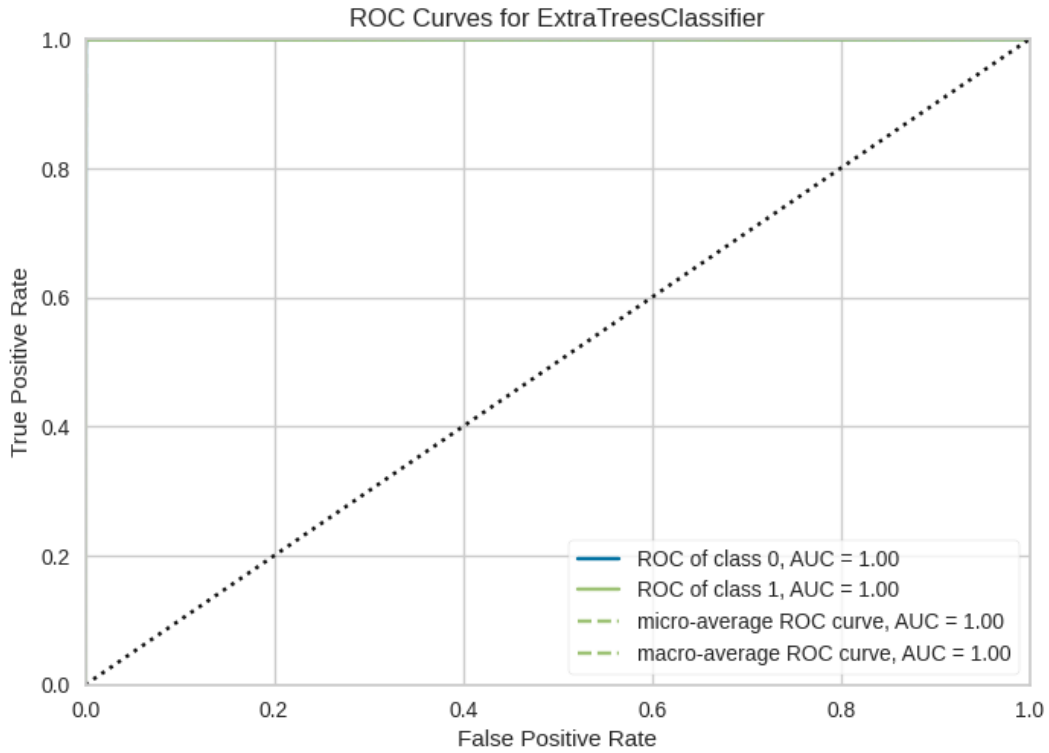
오차행렬



● 오차행렬로 FP(실제 클래스가 0인데, 모델이 1로 예측)와 FN(실제 클래스가 1인데, 모델이 0으로 예측)이 각각 0과 39로, 잘못 분류한 경우가 적다

5. 모델링

ROC 커브



- ROC 커브

y축은 TPR(참 양성 비율),

x축에는 FPR(거짓 양성 비율)에 1이 가까울수록 성능이 좋음

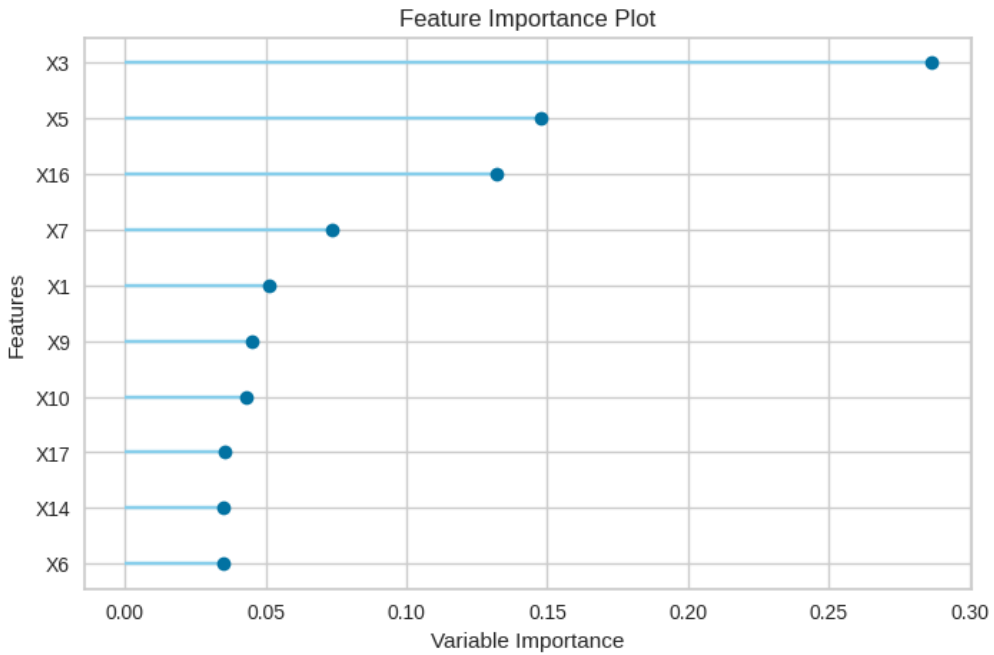
-> 여기서 모든 AUC가 1임

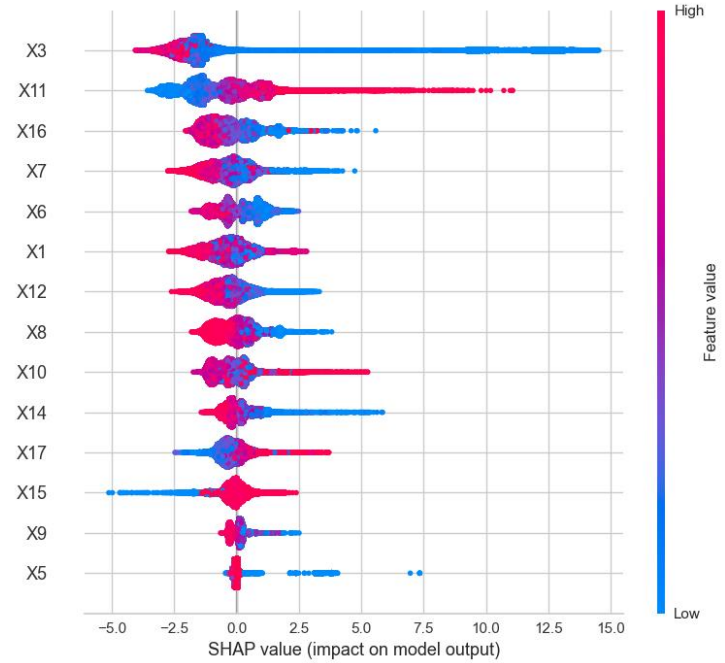
5. 모델링

3C analysis

20 page

변수 중요도 (SHAP)





- x3: 가장 중요한 피처, 높은 값은 모델의 예측을 증가시키고, 낮은 값은 예측을 감소시킴

- x11: 높은 피처 값이 모델의 예측을 감소시킴

6. 과제 요약 및 자체 평가

변수 6개 제거

- X18, X19, X20 [데이터 중복이라 제거]

  X8==X18

  X12==X19

  X6==X20

- X4, X13 다 같은 값 [고정값이라 제거]

  X4==0.015348

  X13==0.249262

- X2 t검정했을 때 통계적으로 유의하지 않아 제거

6. 과제 요약 및 자체 평가

솔루션

- X3, X5 특정값 이하면 불량 (꽤 많은 개수)

다른 변수도 특정값 이하면 불량인 변수들이 많지만, 개수가 적다.

- 불량품 중 93.6%는 X3 값이 0.450993 이하이다.

X30이 임계값보다 작아지지 않도록 관리하는 것이 중요하다.

- 전체 불량품 중 30.7%는 X5 값이 0.4 이하이다.

X5가 임계값보다 작아지지 않도록 관리하는 것이 중요하다.

역할 분담

훈련생	주 역할	담당 업무
강하연	데이터 분석	EDA 및 모델링
김지유	x	x
박민서	PPT 제작 및 발표	EDA 및 발표 자료 제작
이채은	PPT 제작 및 발표	시각화 자료 생성, 발표 자료 제작
정재성	x	x
정하연	데이터 분석	EDA 및 모델링, 발표

6. 과제 요약 및 자체 평가

훈련생 자체평가

평가지문	답변
사전 기획의 관점에서 프로젝트 결과물에 대한 완성도 평가	9점 (10점 만점)
우리 팀의 잘한 부분과 아쉬운 점	•잘한 점 : 조원 사이 배려하는 태도 •아쉬운 점 : 준비 시간 부족
프로젝트 결과물의 추후 개선점이나 보완할 점	빅데이터로 모델을 만들 때 시간이 오래 걸림, 샘플링 적극 활용 필요
프로젝트를 수행하면서 느낀 점이나 경험한 성과(경력 계획 등과 연관)	두 달 간 학습한 다양한 코드를 프로젝트에 활용하며 실무 능력을 강화할 수 있는 계기가 됨 . 팀 내 협업을 통해 복잡한 문제를 해결하는 경험을 쌓고, 대규모 데이터를 효과적으로 처리하고 분석할 수 있는 능력을 키울 수 있었음.

End Of Documents
Thank you