

# 재무정보를 이용한 ESG 등급 예측

팀명 : Fn Hogarden

팀원 : 김다은  
박준용  
최민서  
최호진

# 목차

1. 분석 배경

2. 데이터 이해

3. 탐색적 데이터  
분석(EDA)

4. 모델링

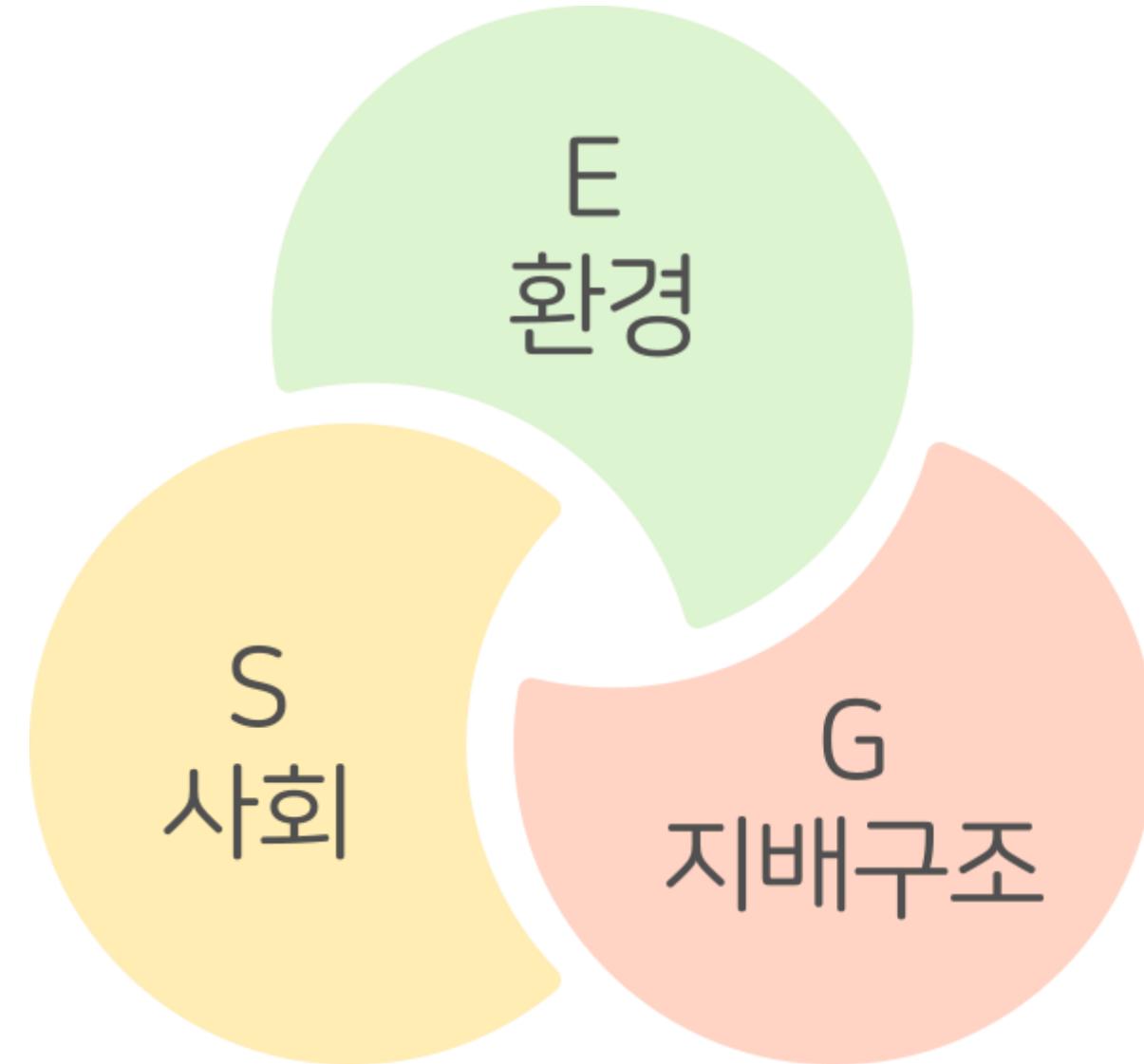
5. 결론

6. 참고문헌

# 1. 분석배경

- ESG 정의
- 아이디어 배경
- 핵심가치
- 유사 아이디어와 차별성

# ESG란?



기업의 중·장기적 가치와 지속 가능성에  
큰 영향을 미칠 수 있는 비재무적 지표

ESG란?

## 환경 (Environmental)

- 기후변화 및 탄소 배출
- 환경오염 및 환경규제
- 생태계 및 생물 다양성

## 사회 (Social)

- 데이터 보호 및 프라이버시
- 인권, 성별 평등 및 다양성
- 지역사회 관계

## 지배구조 (Governance)

- 이사회 및 감사위원회 구성
- 뇌물 및 반부패
- 기업윤리

01

## 아이디어 배경

# EU, 2024년부터 대기업 ESG 공시 의무화 합의

종업원 250명, 연 매출 4000만 유로 초과 기업 대상  
연 매출 1.5억 유로 외국 기업도 대상

**자산 2조 넘는 기업, 2025년 ESG보고서 공개 의무**

## 유럽·캐나다 ESG공시 의무화

국외 공급망 기업들에 대한

인권·환경분야 실사도 검토

美선 탄소배출량 감시 나서

주주·고  
물리적·

3가지 위험에 미리 대비해야

ESG경영에 기업 명운 달렸다

Cover Story 글로벌 보험 중개사 '마쉬'… 제임스 애딩턴 스미스 ESG위원회 의장·이형구 마쉬 코리아 사장 대담

01

## 아이디어 배경



2022

유럽을 비롯한  
ESG 공시 의무화 추진



2025

자산 2조원 이상의  
코스피 상장사의  
지속가능경영 보고서' 공시 의무화



2026

모든 코스피 상장사에  
'기업지배구조 보고서'  
공시 의무화



2030

한국의 모든 상장 회사에  
ESG 공시 의무화 확대

02

## 핵심가치

### ● 정보 제공

ESG 등급평가가 이뤄지지 않은 기업들에 대해 재무 데이터를 바탕으로 ESG등급 분류 예측

### ● 기업 벨류 평가

ESG 등급평가와 재무적인 요소와의 연관성에 대한 검증과 예측을 통해 ESG 공시점수를  
감안한 기업 벨류 평가에 도움

### ● 채권평가

기업의 자금조달의 가장 큰 부분을 차지하는 채권평가에 있어 ESG 평가가 반영된다는 것이  
기정사실화 됨으로써 정보를 선점하여 채권우선평가에 도움

03

## 유사 아이디어와 차별성

- 기존의 선행연구는 ESG 공시 자료를 독립변수로써 재무제표를 종속변수로 하여 ESG등급이 기업에 미치는 영향에 대해 분석
- 본 팀은 ESG등급평가가 완료된 기업을 대상으로 재무제표와 ESG 등급과의 상관성을 학습하여 등급이 없는 기업에 대해 ESG등급을 예측하는 모델 개발

## (1) 사용 데이터



Fn가이드  
재무제표

기업 개요  
QuantWise 재무데이터

기업별  
ESG 등급

CGS기준원

기업별  
세부정보

빅데이터 플랫폼  
통계청 산업코드

## (2) 데이터 전처리

---



### 1. 데이터 병합

Fn가이드 재무 데이터와  
CGS기준원의 ESG 등급 데이터를  
기업코드를 통해 취합



### 2. 결측값 처리

종가, 매출액, EV, 배당금, P/CE,  
임원보수, 직원수정규직, 주권의수,  
지분율, 직원평균근속년수,  
직원1인평균급여액

## (2) 데이터 전처리

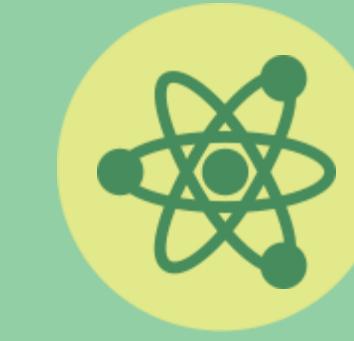
---



### 3. 이상치 제거 및 정규화

이상치의 영향도를 줄이기 위해  
매출액 상위 3% 하위 3%에  
해당하는 값 제거

MinMaxScale을 통해 단위 정규화



### 4. 파생 변수 생성

#### 근속년수 남녀차이

직원평균근속년수(남) - 직원평균근속년수(여)

#### 평균급여액 남녀차이

직원1인평균급여액(남) - 직원1인평균급여액(여)

#### 사외이사유무 (더미변수)

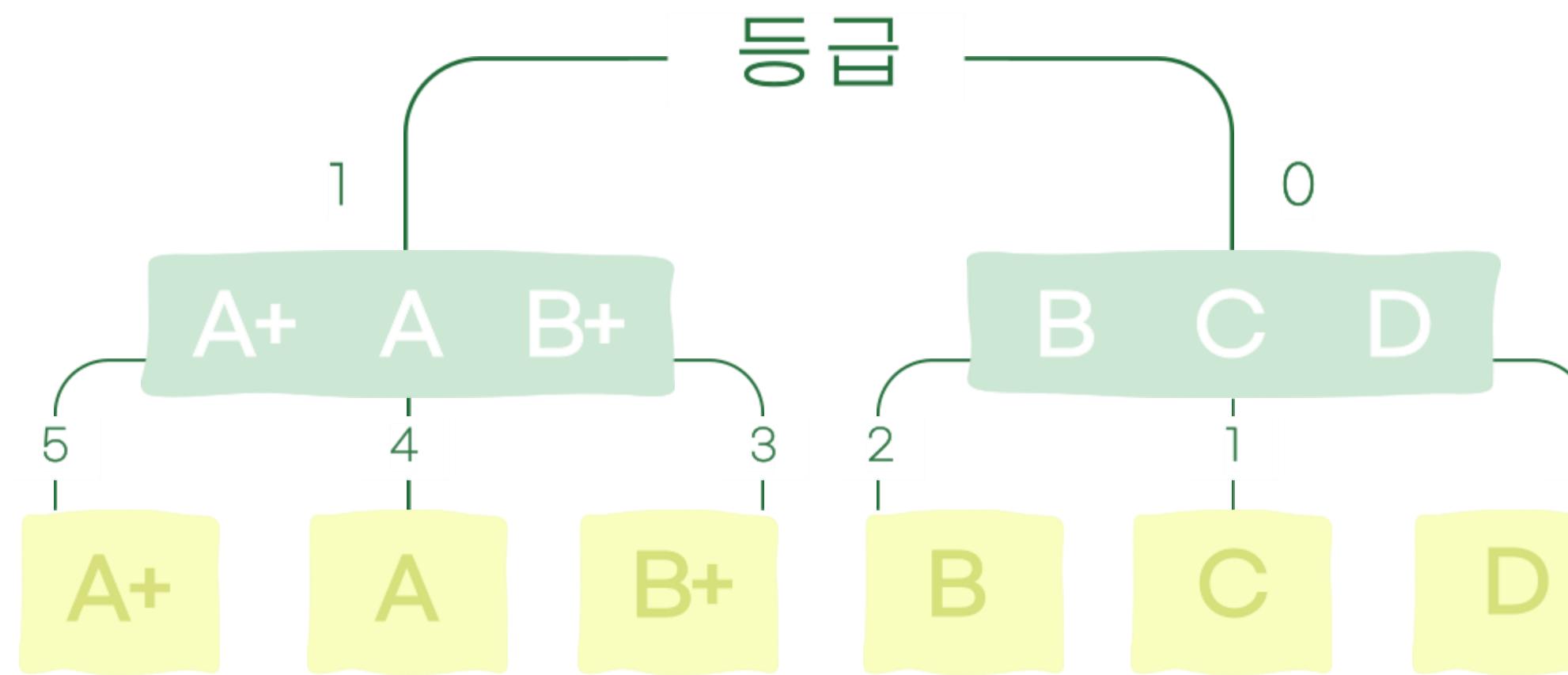
임원보수(사외이사) 값이 0이 아니면 1, 0이면 0

## (2) 데이터 전처리



### 5. 라벨링

범주형인 ESG 등급 문자형 계층화



### 3. EDA

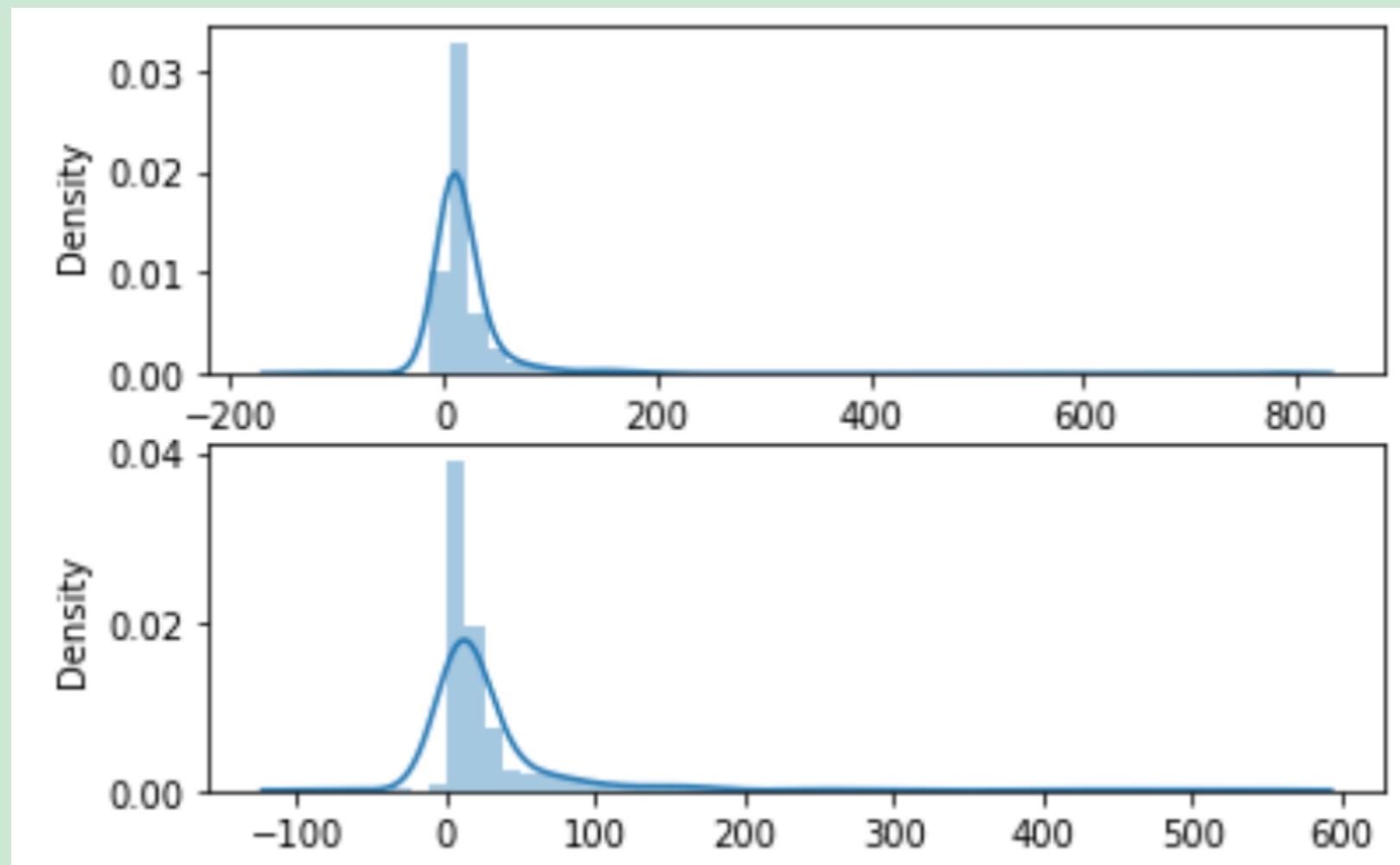
한국 시장상황과 재무정보를 기반으로  
이상치 비중 및 분포도 확인, 유의미한 패턴 탐색

(1) 시각화: 연속형 및 범주형

(2) 시각화: 이상치 확인

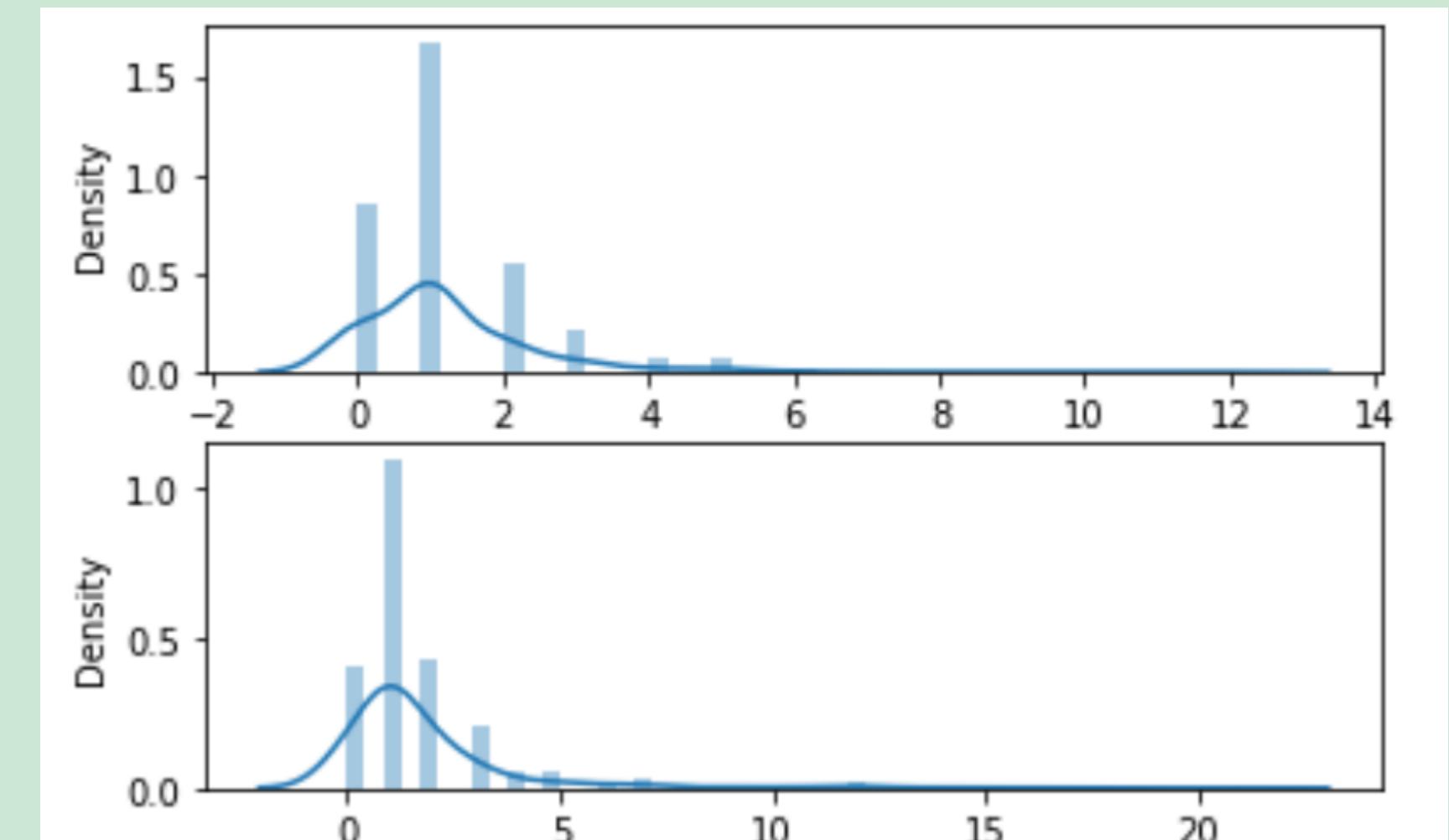
# (1) 시각화: 연속형

- B+ 이상(A+, A, B+) 등급: 위
- B+ 미만(B, C, D) 등급: 아래



S 등급에 따른 PER 분포도

일반적으로 PER의 평균치인 10을 중심으로 분포

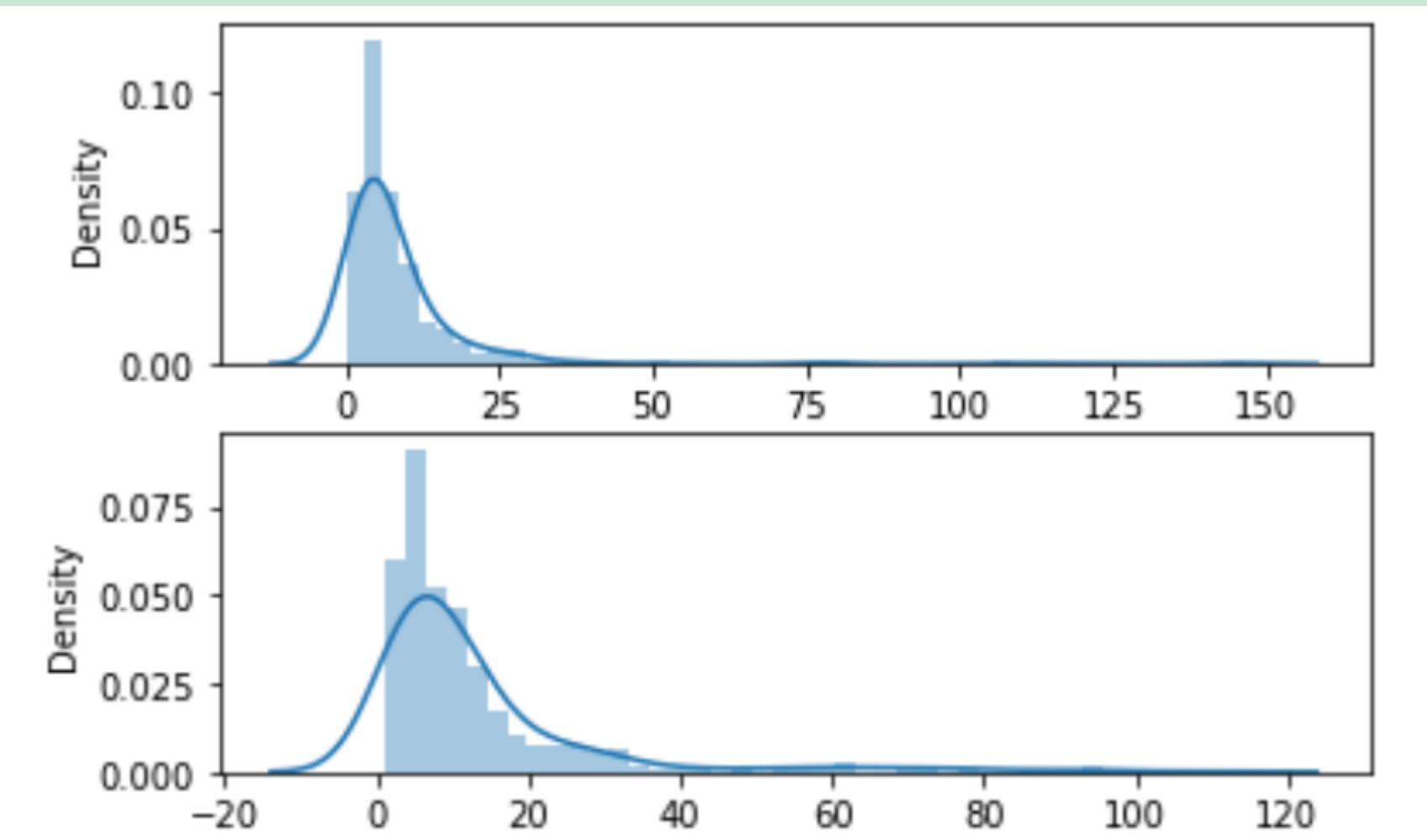


S 등급에 따른 PBR 분포도

일반적으로 PBR의 평균치인 1을 중심으로 분포

# (1) 시각화: 연속형

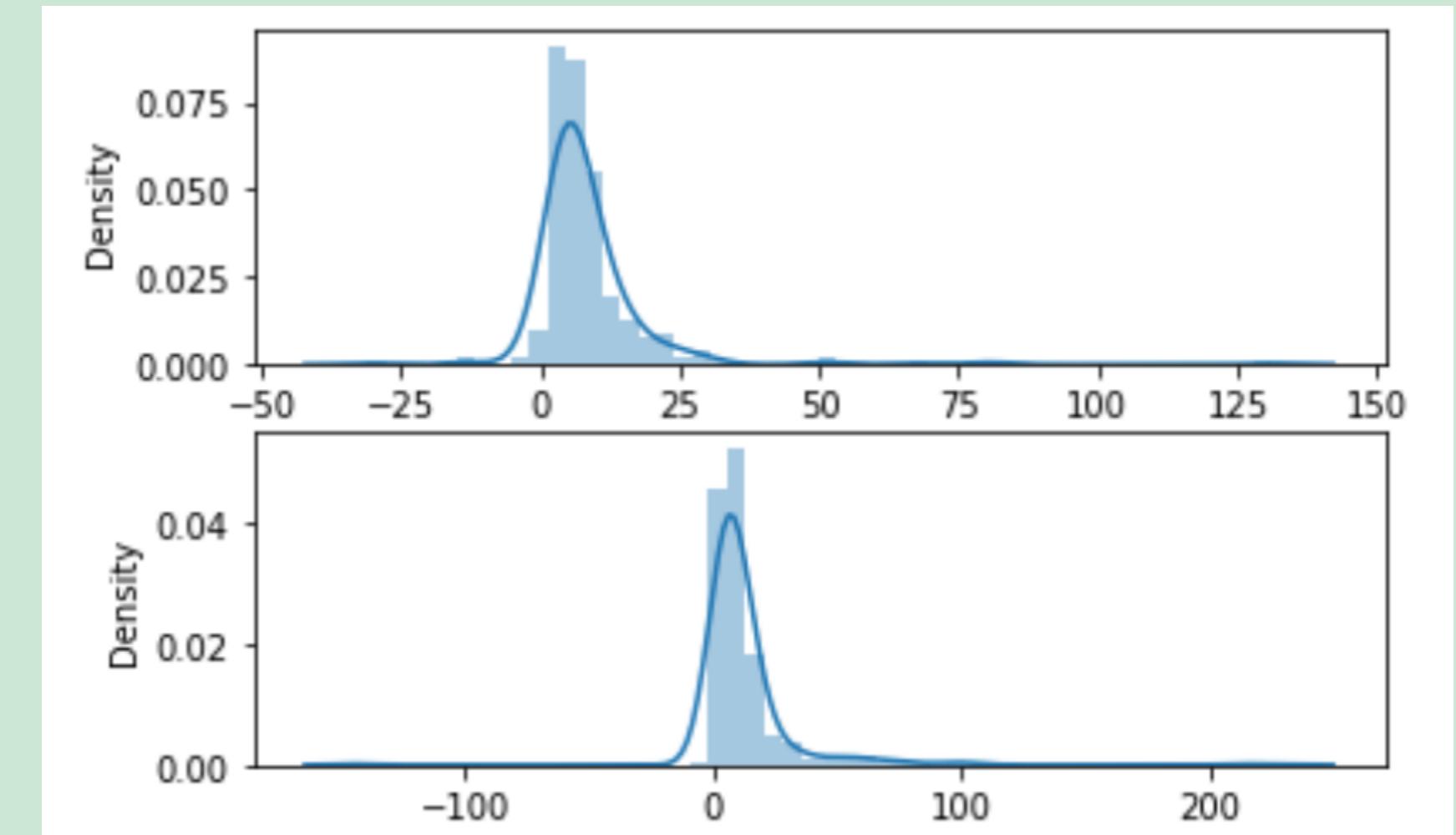
- B+ 이상(A+, A, B+) 등급: 위
- B+ 미만(B, C, D) 등급: 아래



S 등급에 따른 P/CE 분포도



P/CE의 중심값인 3을 중심으로 분포



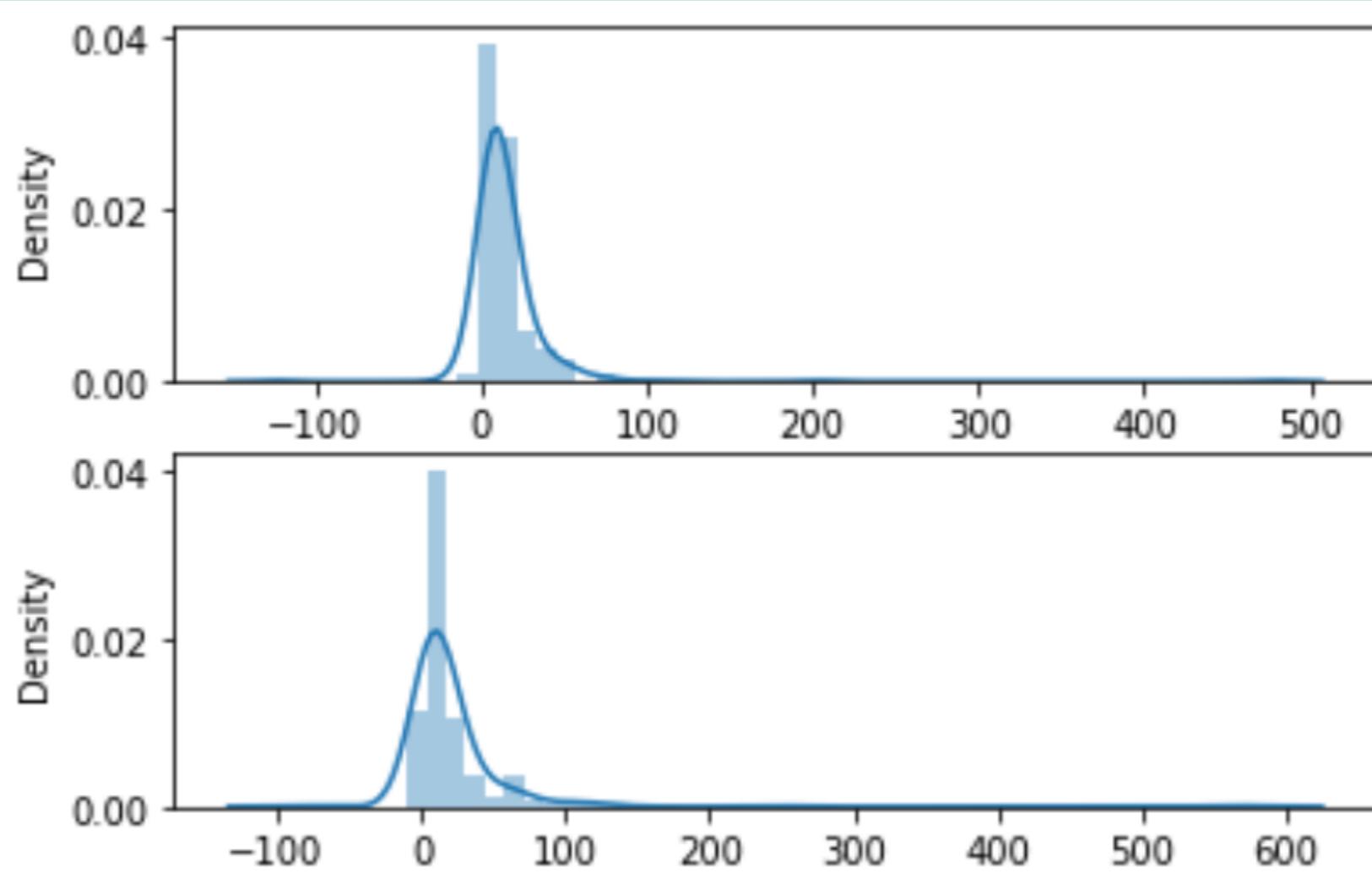
S 등급에 따른 EV/EVITDA 분포도



EV/EVITDA의 중심값인 6을 중심으로 분포

# (1) 시각화: 연속형

- B+ 이상(A+, A, B+) 등급: 위
- B+ 미만(B, C, D) 등급: 아래



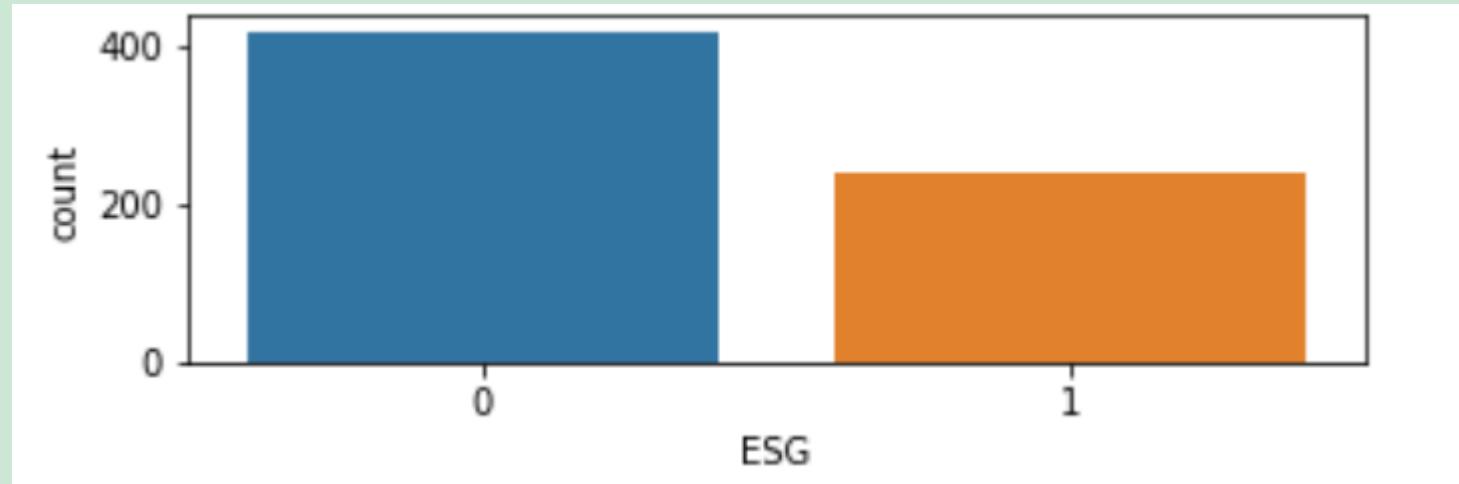
S 등급에 따른 EV/EVIT 분포

EV/EVIT의 중심값인 8을 중심으로 분포

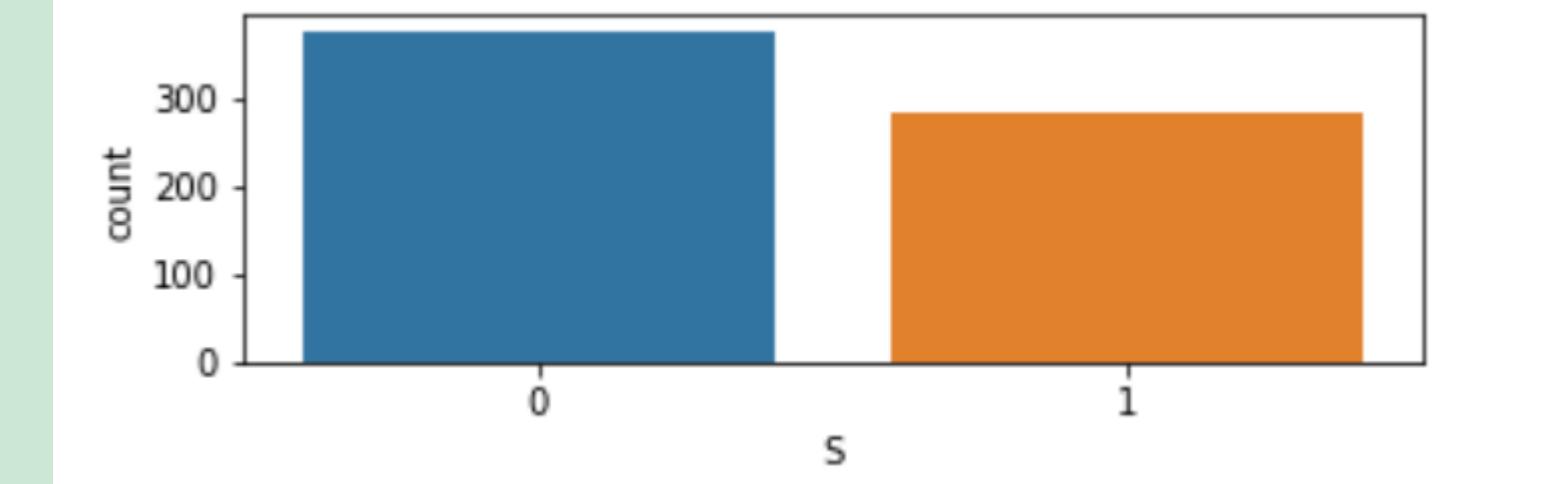
- B+ 이상(A+, A, B+) 등급: 위
- B+ 미만(B, C, D) 등급: 아래

## (2) 시각화: 범주형

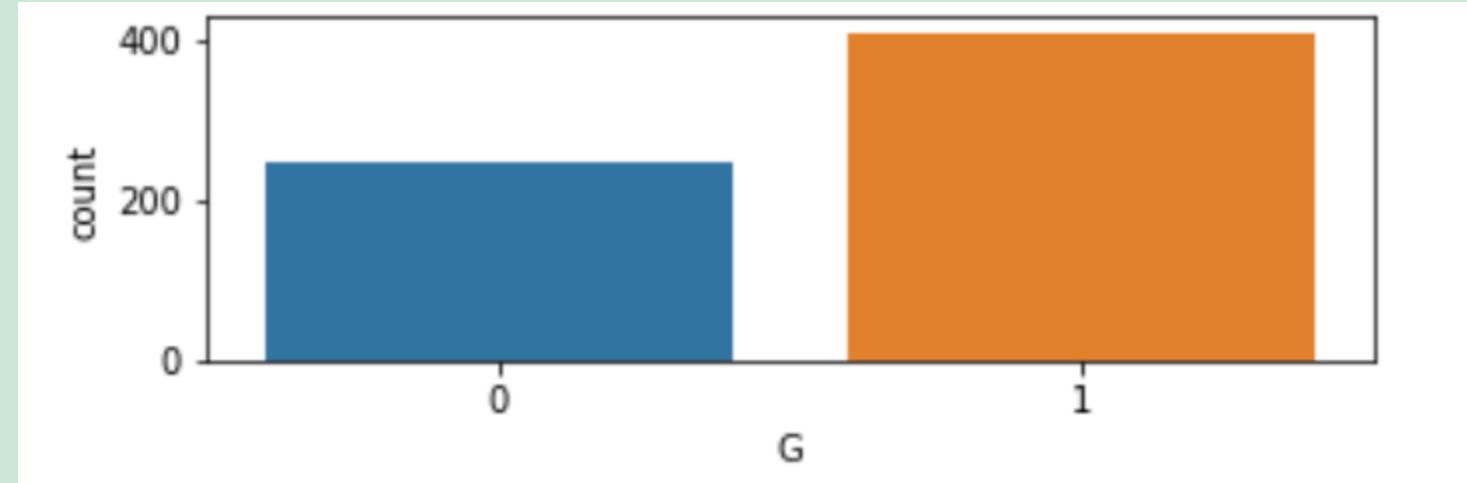
- B+ 미만(B, C, D) 등급: 0
- B+ 이상(A+, A, B+) 등급: 1



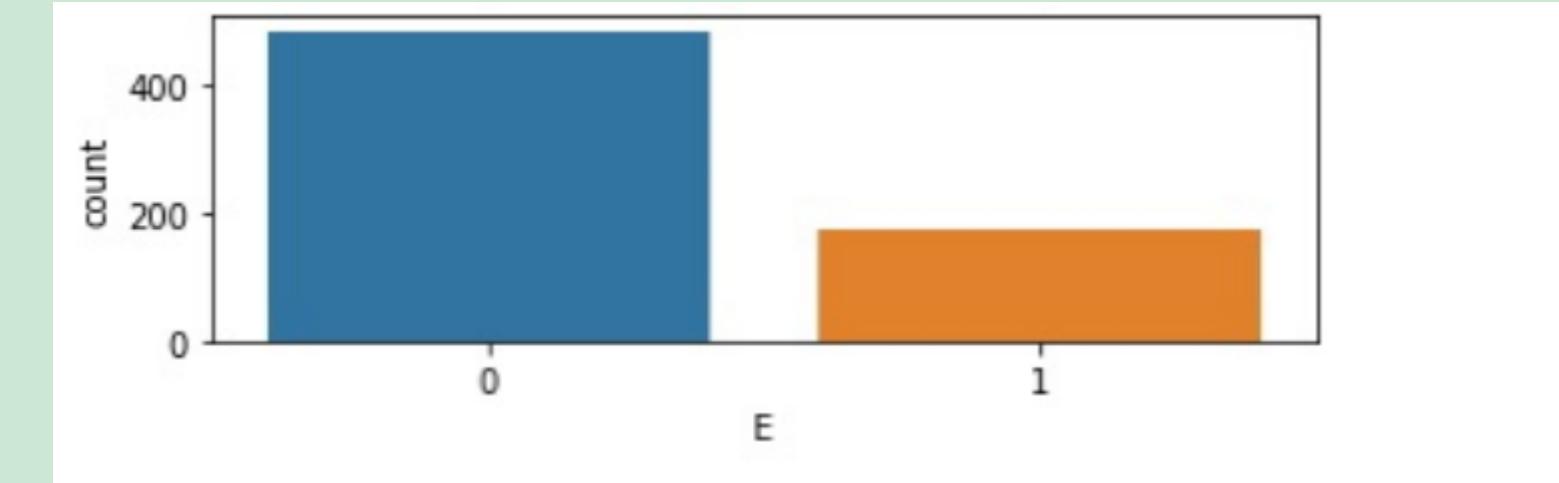
ESG에 대한 기업 수



S등급에 대한 기업 수



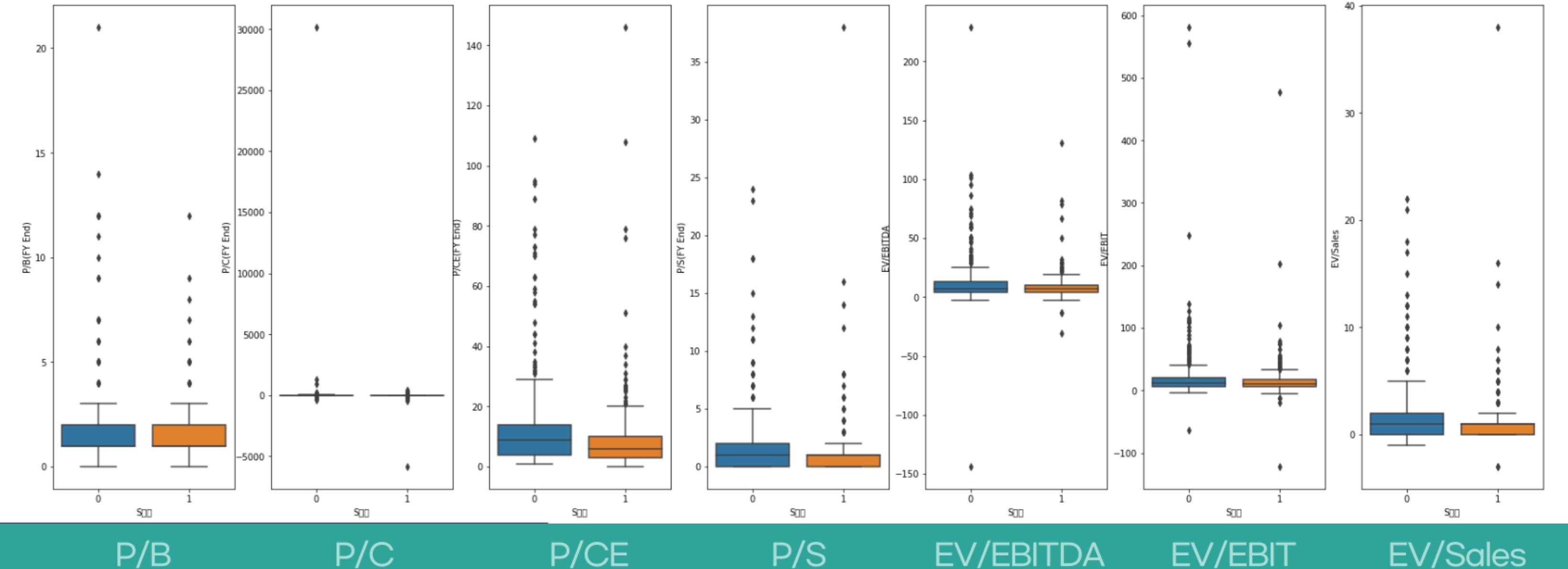
G등급에 대한 기업 수



E등급에 따른 기업 수

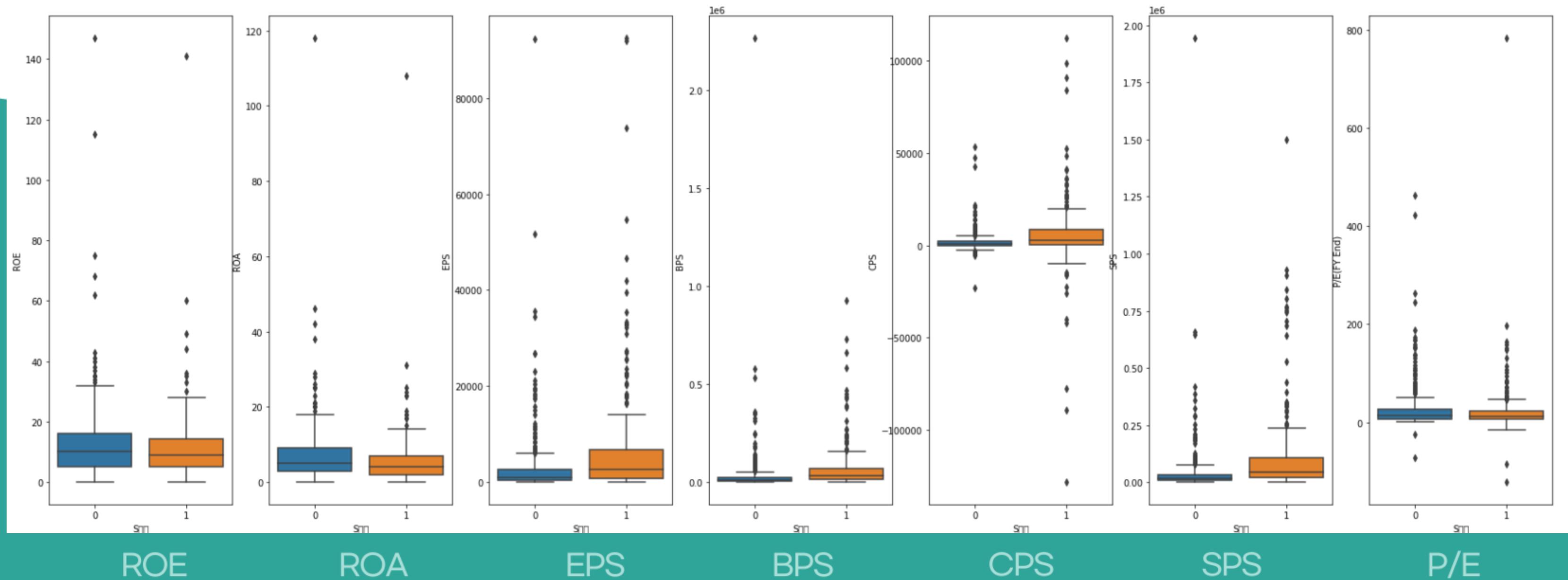
## (2) 시각화: 이상치 확인

- B+ 미만(B, C, D) 등급: 0
- B+ 이상(A+, A, B+) 등급: 1



## (2) 시각화: 이상치 확인

- B+ 미만(B, C, D) 등급: 0
- B+ 이상(A+, A, B+) 등급: 1



## 4. 모델링

- (1) 모델 탐색: AutoML
- (2) 모델 선정: 직접 비교
- (3) 학습 및 예측

# (1) 모델 탐색: AutoML

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ada	Ada Boost Classifier	0.4388	0.4093	0.3173	0.2658	0.3162	0.2309	0.2880	0.194
et	Extra Trees Classifier	0.4355	0.5069	0.3482	0.4041	0.4121	0.2345	0.2386	1.124
rf	Random Forest Classifier	0.4116	0.5120	0.2877	0.3881	0.3916	0.2028	0.2062	1.096
lightgbm	Light Gradient Boosting Machine	0.4083	0.4949	0.2991	0.3916	0.3938	0.1992	0.2014	0.731
gbc	Gradient Boosting Classifier	0.4064	0.4979	0.3247	0.4010	0.3958	0.2044	0.2068	2.592
ridge	Ridge Classifier	0.3913	0.0000	0.2625	0.4134	0.3656	0.1562	0.1628	0.018
dt	Decision Tree Classifier	0.3877	0.4240	0.2856	0.3986	0.3870	0.1857	0.1879	0.029
nb	Naive Bayes	0.3759	0.4658	0.2677	0.3186	0.3080	0.1477	0.1734	0.016
lda	Linear Discriminant Analysis	0.3673	0.4667	0.3015	0.3977	0.3504	0.1343	0.1414	0.029
knn	K Neighbors Classifier	0.3624	0.4592	0.2596	0.3380	0.3414	0.1397	0.1419	0.179
dummy	Dummy Classifier	0.3027	0.3500	0.1767	0.0917	0.1408	0.0000	0.0000	0.011
qda	Quadratic Discriminant Analysis	0.2058	0.0000	0.1767	0.0424	0.0703	0.0000	0.0000	0.017
lr	Logistic Regression	0.1973	0.4713	0.2006	0.0744	0.0827	-0.0051	-0.0124	0.578
svm	SVM - Linear Kernel	0.1580	0.0000	0.2064	0.0817	0.0948	-0.0008	-0.0007	0.115

## Automated Machine Learning

- 다양한 머신러닝 모델을 자동 학습, 예측
- 모델 간의 정확도를 비교 및 성능 확인  
의사결정에 도움을 주는 프로세스
  - PyCaret

데이터에 적합한  
분류 알고리즘 모델을  
선정하기 위해  
AutoML을 통해  
탐색하는 과정

## (2) 모델 선정: 직접 비교

### 비교 모델

- Adaboost
- XGboost

- Logistic Regression

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.01, 'n_estimators': 100}
```

```
adaboost 정확도: 0.6105610561056105
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.03, 'max_depth': 1, 'n_estimators': 100}
```

```
xgboost 정확도: 0.6204620462046204
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 100.0, 'penalty': 'l2'}
```

```
Logistic Regression 정확도: 0.5907590759075907
```

가장 정확도가 높은 모델로 선정

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.03, 'max_depth': 1, 'n_estimators': 100}
```

~~~~~ xgboost 정확도: 0.6204620462046204

예측 대상에 따라  
정확도를 비교해  
적합한 모델 선정

## (2) 모델 선정: 직접 비교

- Adaboost

정확도가 높고 학습 데이터에 과적합 현상이 적게 발생  
노이즈 데이터 및 이상치에 민감하고, XGB에 비해 느림

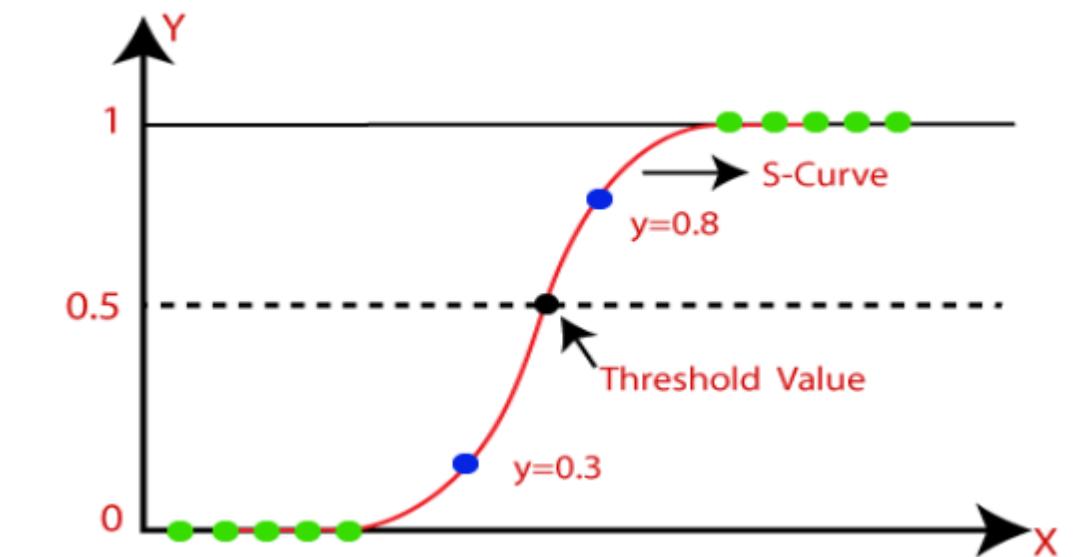
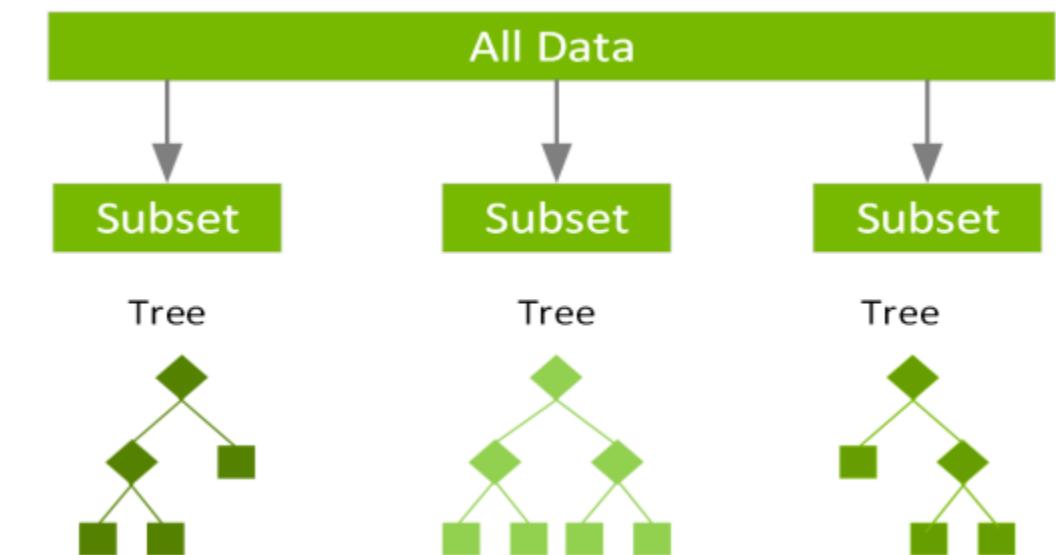
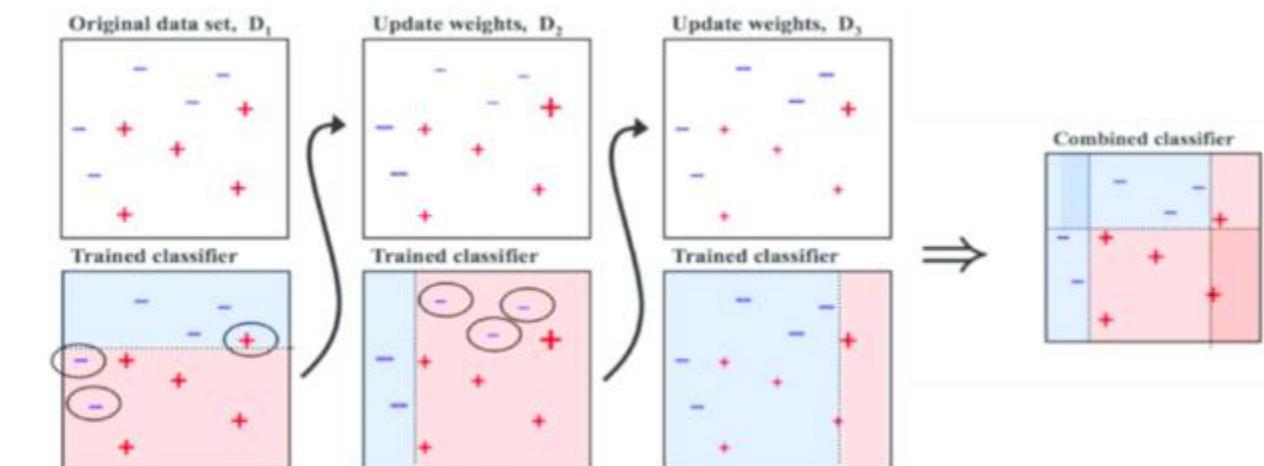
- XGboost

병렬 처리로 학습하여 효율적이고 뛰어난 예측 성능 발휘  
튜닝할 파라미터가 다수, 시간비용이 다소 큼

- Logistic Regression

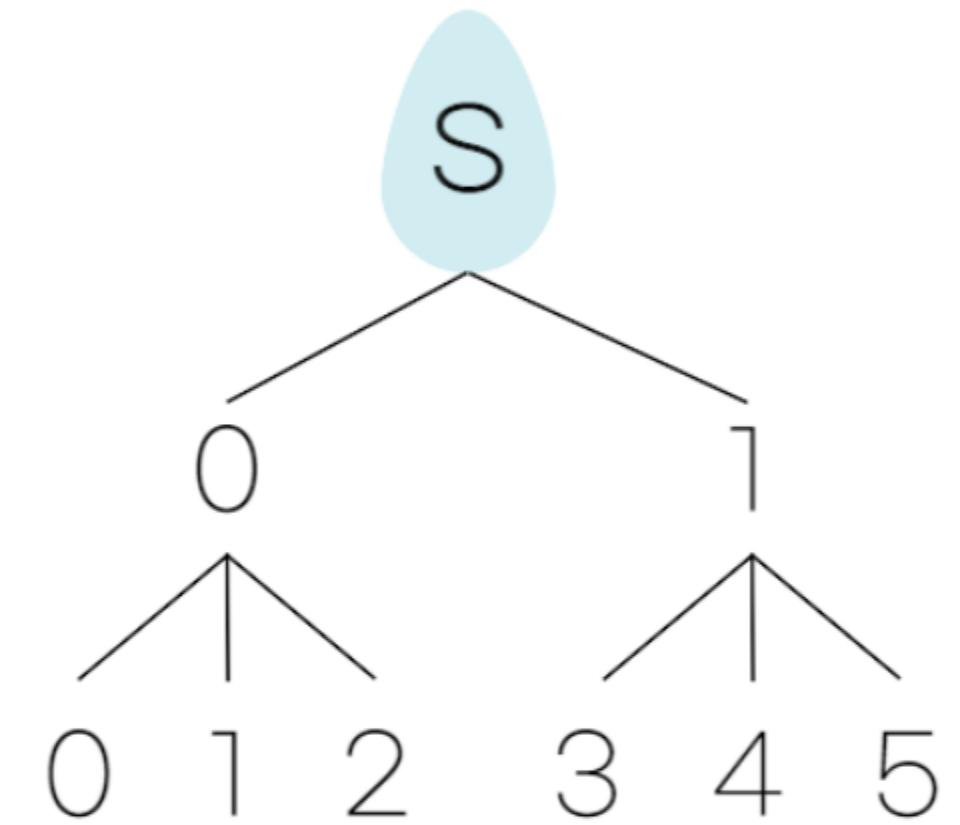
쉽게 해석할 수 있으며 정규화하기 쉬움  
비선형 문제를 해결하는 데 사용하기 어려움

Grid Search 방식을 이용해 최적화



### (3) 1차 학습 및 예측: S등급

- E, S, G 등급을 모두 고려해 분류 예측을 실시했을 때, S등급에 대한 정확도가 가장 높음을 확인
  - S등급과 모델 변수들 사이에 상관관계 높음
- S등급을 우선적으로 2개의 class로 분류 예측
  - 1: A+, A, B+ 등급
  - 0: B, C, D 등급
- 0과 1로 분류된 값을 다시 각각 3개의 class로 재분류
  - 5: A+ 등급
  - 4: A 등급
  - 3: B+ 등급
  - 2: B 등급
  - 1: C 등급
  - 0: D 등급



나눠서 예측했을 때  
정확도가 더 높음

## (4) 2차 학습 및 예측: E등급, G등급, ESG 등급

- 예측된 S등급 값을 이용해 ESG등급을 예측 실시
- S등급과 동일한 예측 과정으로 재학습 및 예측
- 예측된 ESG등급을 반영해 E등급과 G등급도 마찬가지로 동일한 예측 과정으로 재학습 및 예측

S등급 추정값



E, G, ESG등급 예측

# 5. 결론

## 01-1. S등급 예측(0, 1)

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.17, 'n_estimators': 100}  
adaboost 정확도: 0.804263565891473
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.33, 'max_depth': 10, 'n_estimators': 100}  
xgboost 정확도: 0.8023255813953488
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 100.0, 'penalty': 'l2'}  
Logistic Regression 정확도: 0.7596899224806202
```

## 01-2. S등급 중 1로 예측된 것(B+, A, A+)을 다시 세 개로 분류

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.48, 'n_estimators': 500}  
adaboost 정확도: 0.5569416498993963
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.03, 'max_depth': 1, 'n_estimators': 100}  
xgboost 정확도: 0.5519114688128773
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 1.0, 'penalty': 'l2'}  
Logistic Regression 정확도: 0.5425217974513749
```

## 01-3. S등급 중 0으로 예측된 것(D, C, B)을 다시 세 개로 분류

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.01, 'n_estimators': 100}  
adaboost 정확도: 0.6105610561056105
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.03, 'max_depth': 1, 'n_estimators': 100}  
xgboost 정확도: 0.6204620462046204
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 100.0, 'penalty': 'l2'}  
Logistic Regression 정확도: 0.5907590759075907
```

01-1 정확도: 0.8043

01-2 정확도: 0.5569

01-3 정확도: 0.6204

# 5. 결론

## 02-1. ESG등급 예측

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.02, 'n_estimators': 100}  
adaboost 정확도: 0.8953488372093025
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.03, 'max_depth': 1, 'n_estimators': 100}  
xgboost 정확도: 0.8934108527131782
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 0.1, 'penalty': 'l2'}  
Logistic Regression 정확도: 0.8953488372093025
```

## 02-2. ESG등급 중 1로 예측된 것(B+, A, A+)을 다시 세 개로 분류

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.14, 'n_estimators': 100}  
adaboost 정확도: 0.7388888888888889
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.06, 'max_depth': 1, 'n_estimators': 100}  
xgboost 정확도: 0.8277777777777778
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 0.001, 'penalty': 'l2'}  
Logistic Regression 정확도: 0.8166666666666666
```

## 02-3. ESG등급 중 0으로 예측된 것(D, C, B)을 다시 세 개로 분류

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.49, 'n_estimators': 500}  
adaboost 정확도: 0.6160714285714285
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.01, 'max_depth': 1, 'n_estimators': 500}  
xgboost 정확도: 0.7142857142857143
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 0.01, 'penalty': 'l2'}  
Logistic Regression 정확도: 0.7113095238095238
```

02-1 정확도: 0.8953

02-2 정확도: 0.8278

02-3 정확도: 0.7143

# 5. 결론

## 03-1. E등급 예측

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.01, 'n_estimators': 500}  
adaboost 정확도: 0.8643410852713179
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.26, 'max_depth': 4, 'n_estimators': 100}  
xgboost 정확도: 0.8740310077519379
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 1.0, 'penalty': 'l2'}  
Logistic Regression 정확도: 0.881782945736434
```

## 03-2. E등급 중 1로 예측된 것(B+, A, A+)을 다시 세 개로 분류

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.35, 'n_estimators': 500}  
adaboost 정확도: 0.626984126984127
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.42, 'max_depth': 5, 'n_estimators': 500}  
xgboost 정확도: 0.7142857142857143
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 0.1, 'penalty': 'l2'}  
Logistic Regression 정확도: 0.746031746031746
```

## 03-3. E등급 중 0으로 예측된 것(D, C, B)을 다시 세 개로 분류

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.04, 'n_estimators': 100}  
adaboost 정확도: 0.6118067978533095
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.15, 'max_depth': 9, 'n_estimators': 100}  
xgboost 정확도: 0.6143907771814748
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 10.0, 'penalty': 'l2'}  
Logistic Regression 정확도: 0.6068177300735439
```

03-1 정확도: 0.8818

03-2 정확도: 0.7460

03-3 정확도: 0.6144

# 5. 결론

## 04-1. G등급 예측

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.01, 'n_estimators': 500}  
adaboost 정확도: 0.686046511627907
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.36, 'max_depth': 5, 'n_estimators': 100}  
xgboost 정확도: 0.7073643410852712
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 1.0, 'penalty': 'l2'}  
Logistic Regression 정확도: 0.6976744186046512
```

## 04-2. G등급 중 1로 예측된 것(B+, A, A+)을 다시 세 개로 분류

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.01, 'n_estimators': 100}  
adaboost 정확도: 0.8184981684981686
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.01, 'max_depth': 1, 'n_estimators': 100}  
xgboost 정확도: 0.8886752136752136
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 0.1, 'penalty': 'l2'}  
Logistic Regression 정확도: 0.8791514041514041
```

## 04-3. G등급 중 0으로 예측된 것(D, C, B)을 다시 세 개로 분류

```
adaboost 최적 하이퍼 파라미터: {'learning_rate': 0.07, 'n_estimators': 100}  
adaboost 정확도: 0.8159203980099502
```

```
xgboost 최적 하이퍼 파라미터: {'learning_rate': 0.02, 'max_depth': 5, 'n_estimators': 100}  
xgboost 정확도: 0.8258706467661692
```

```
Logistic Regression 최적 하이퍼 파라미터: {'C': 1.0, 'penalty': 'l2'}  
Logistic Regression 정확도: 0.8308457711442786
```

04-1 정확도: 0.7074

04-2 정확도: 0.8887

04-3 정확도: 0.8308

## 6. 기대효과 및 한계점

---

**기대효과** ESG 평가가 이루어지지 않는 기업에 대해 기업 평가를 진행할 때 예측 모형을 통해 ESG공시점수를 감안한 기업 벨류평가에 도움이 될 수 있다.

### 한계점

본팀은 재무적인 요소로 ESG등급을 평가하였기 때문에 등급평가에 있어 괴리가 발생할 수 있다.

## 6. 참고문헌

양진수. "자본시장 변동 시 기업의 ESG 경영 수준이 기업가치에 미치는 영향." 국내석사학위논문 한 성대학교 지식서비스&컨설팅대학원, 2022. 서울

김영환,허정하, and 송동엽. "기업의 ESG활동과 자율공시가 기업가치에 미치는 영향." **財務管理研究** 39.1 (2022): 121-144.

김세희,선우희연,이우종, and 정아름. "ESG 활동과 기업가치의 상관관계와 인과관계." 회계저널 31.3 (2022): 31-60.

김진배, and 김주일. "빅데이터를 활용한 금융권 ESG 이슈 분석." 인문사회 21 12.6 (2021): 555-570.

감사합니다

