스타트업 성공 예측

김영호 (2020215729)

길민준 (202222####)

기업 성공 확률 예측 해커톤:
미래의 성공 기업을 발굴하라!

목적

- 기업의 성공은 투자자나 이해관계자들에게 매우 중요함.

- 기업의 성공 확률 예측 AI 알고리즘 개발

데이터

- 고객수, 직원 수, 분야 등 총 14개의 변수로 구성

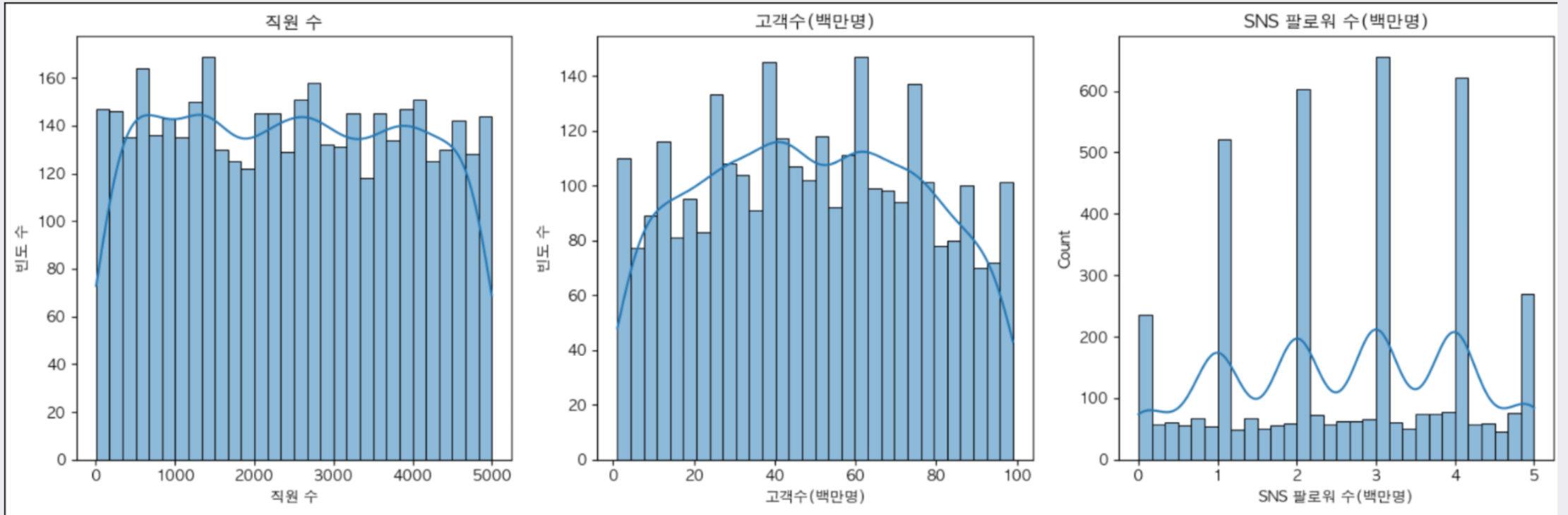
- 총 4,736개의 데이터로 구성

데이터 소개

컬럼명	수치형 컬럼	범주형 컬럼	고유값 목록	결측치 비율
ID		✅	-	-
설립연도	✅		-	-
국가		✅	CT001, CT002, CT003, CT004, CT005, CT006, CT007, CT008, CT009, CT010	-
분야		✅	AI, 게임, 기술, 물류, 에너지, 에듀테크, 이커머스, 핀테크, 푸드테크, 헬스케어	19.50%
투자단계		✅	Seed, Series A, Series B, Series C, IPO	-
직원 수	✅		-	3.90%
인수여부		✅	No, Yes	-
상장여부		✅	No, Yes	-
고객수(백만명)	✅		-	30.10%
총 투자금(억원)	✅		-	-
연매출(억원)	✅		-	-
SNS 팔로워 수(백만명)	✅		-	-
기업가치(백억원)	✅		-	27.80%
성공확률	✅		-	-

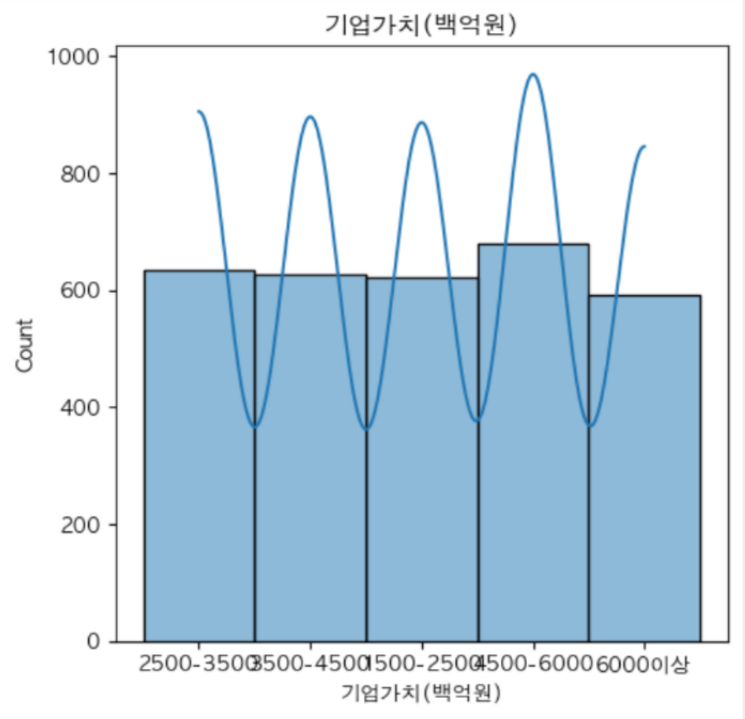
EDA 및 전처리 (수치형)

컬럼명	비고	결측치 비율
설립연도	2001 ~ 2023년까지 모든 연도 포함, 누락 없음	-
총 투자금(억원)	메모리 최적화를 위해 실수형에서 정수형으로 변환	
연매출(억원)		
직원 수	분산을 확인후 평균으로 대체	3.90%
고객수(백만명)	분산을 확인후 중앙값으로 대체	30.10%
SNS 팔로워 수(백만명)	SNS는 기본 홍보 수단이므로, 팔로워 수 0은 결측치로 간주해 최빈값으로 대체	4.18%



EDA 및 전처리 (범주형)

컬럼명	비고	결측치 비율
국가	코드로 구성됨, 원-핫 인코딩 적용	-
분야	누락된 값은 'Unknown'으로 대체	19.50%
투자단계	순서가 있으므로 수치형으로 매핑	-
인수여부	원-핫 인코딩 적용	-
상장여부	원-핫 인코딩 적용	-
기업가치(백억원)	구간형 문자열을 중간값 기준 수치로 매핑, 결측치는 '4500~6000' -> 5250으로 대체	27.80%

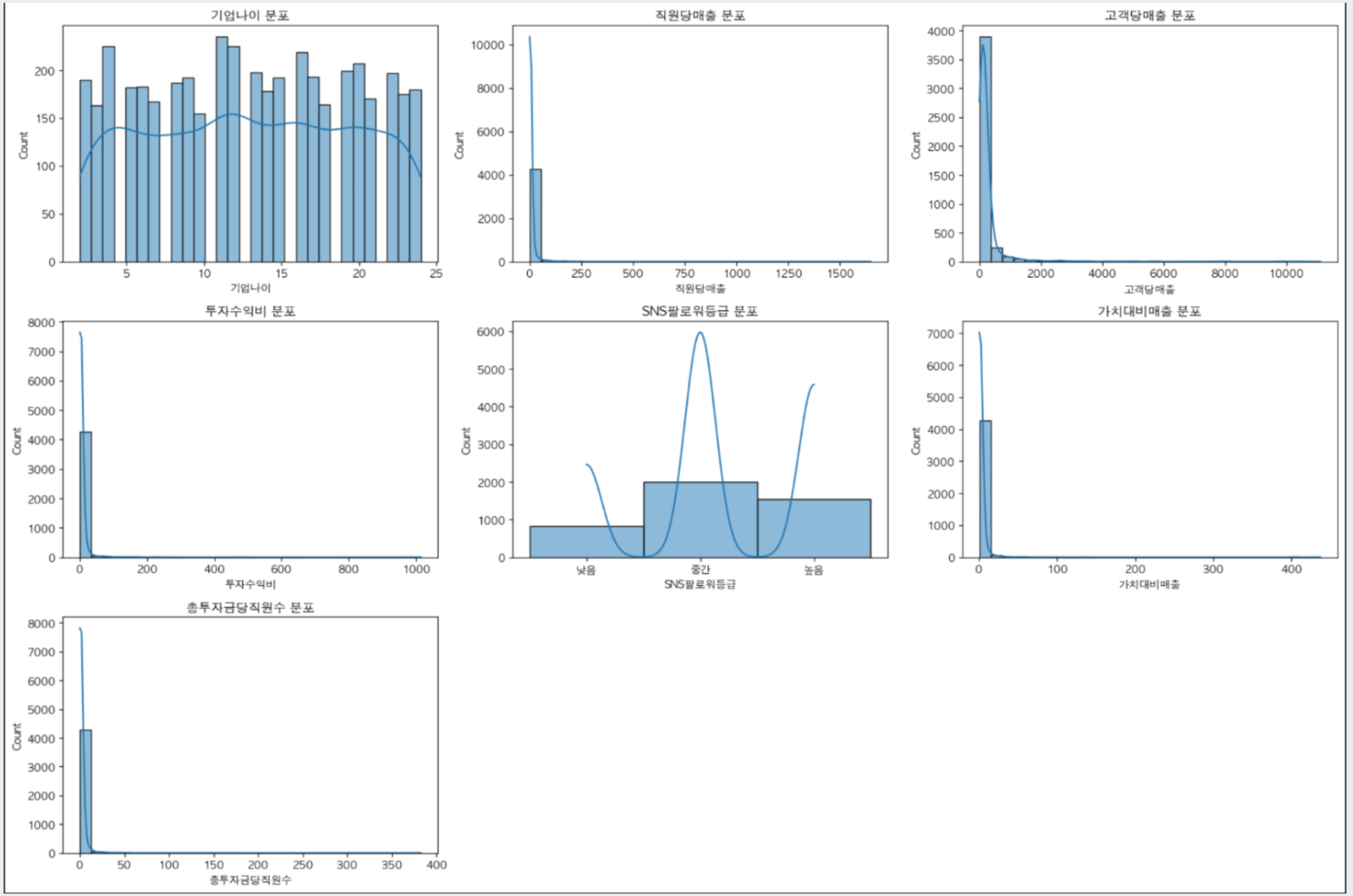


EDA 및 전처리 (파생변수)

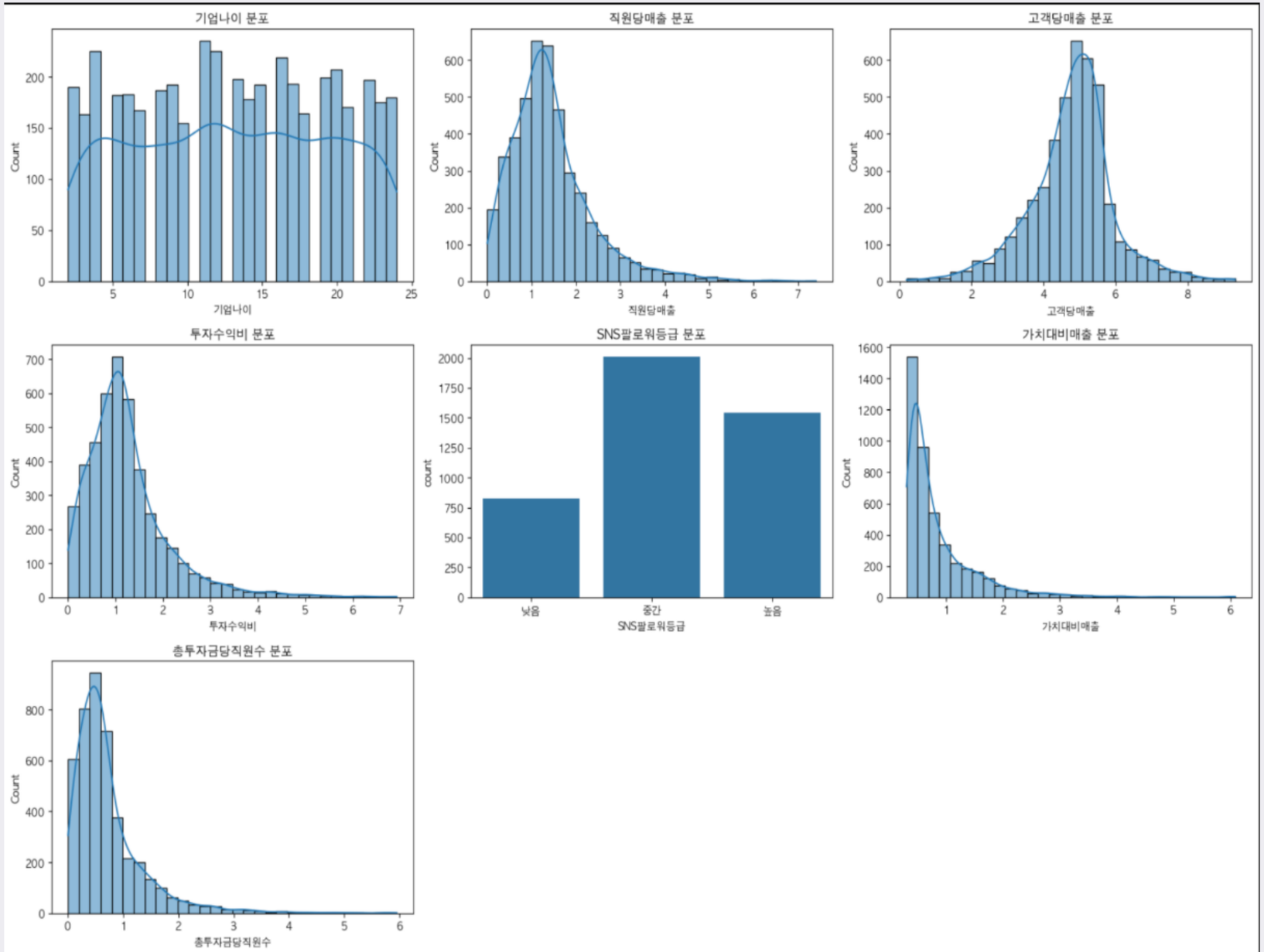
컬럼명	비고	수식
기업나이	설립연도를 기준으로 2025년에서 뺀 값으로 계산	2025 - 설립연도
직원당매출	직원 1인당 연매출	연매출(억원) / 직원 수
고객당 매출	고객 1인당 연매출	연매출(억원) / 고객수(백만명)
투자수익비	투자금 대비 수익	연매출(억원) / 총 투자금(억원)
SNS팔로워 등급	SNS 팔로워 수(백만명) 구간에 따라 '낮음','중간','높음'으로 구간화	
가치대비매출	기업가치 대비 매출	기업가치(백억원) / 연매출(억원)
총투자금당직원수	총 투자금당 직원 수	직원 수 / 총 투자금(억원)

중앙 정렬

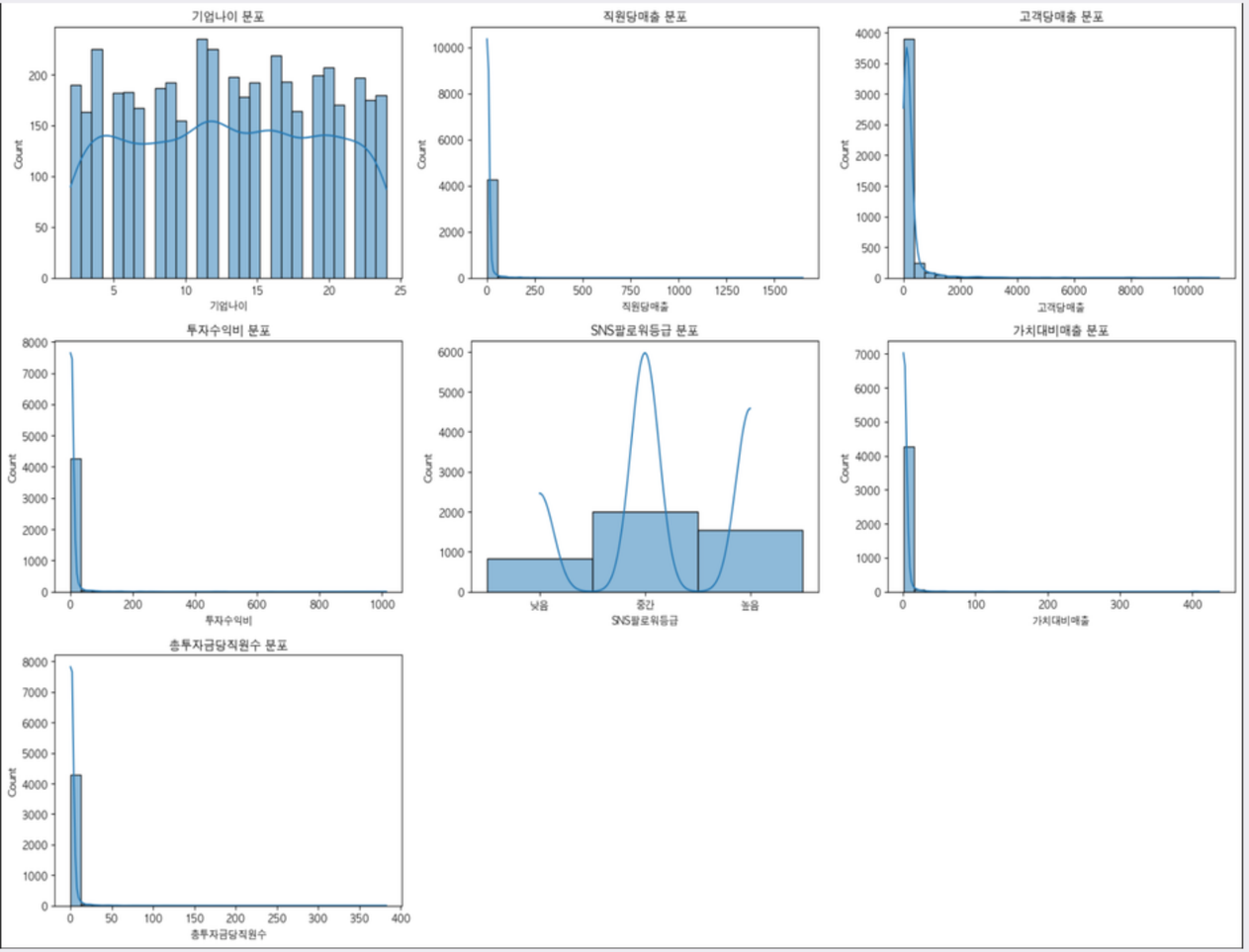
EDA 및 전처리 (파생변수)

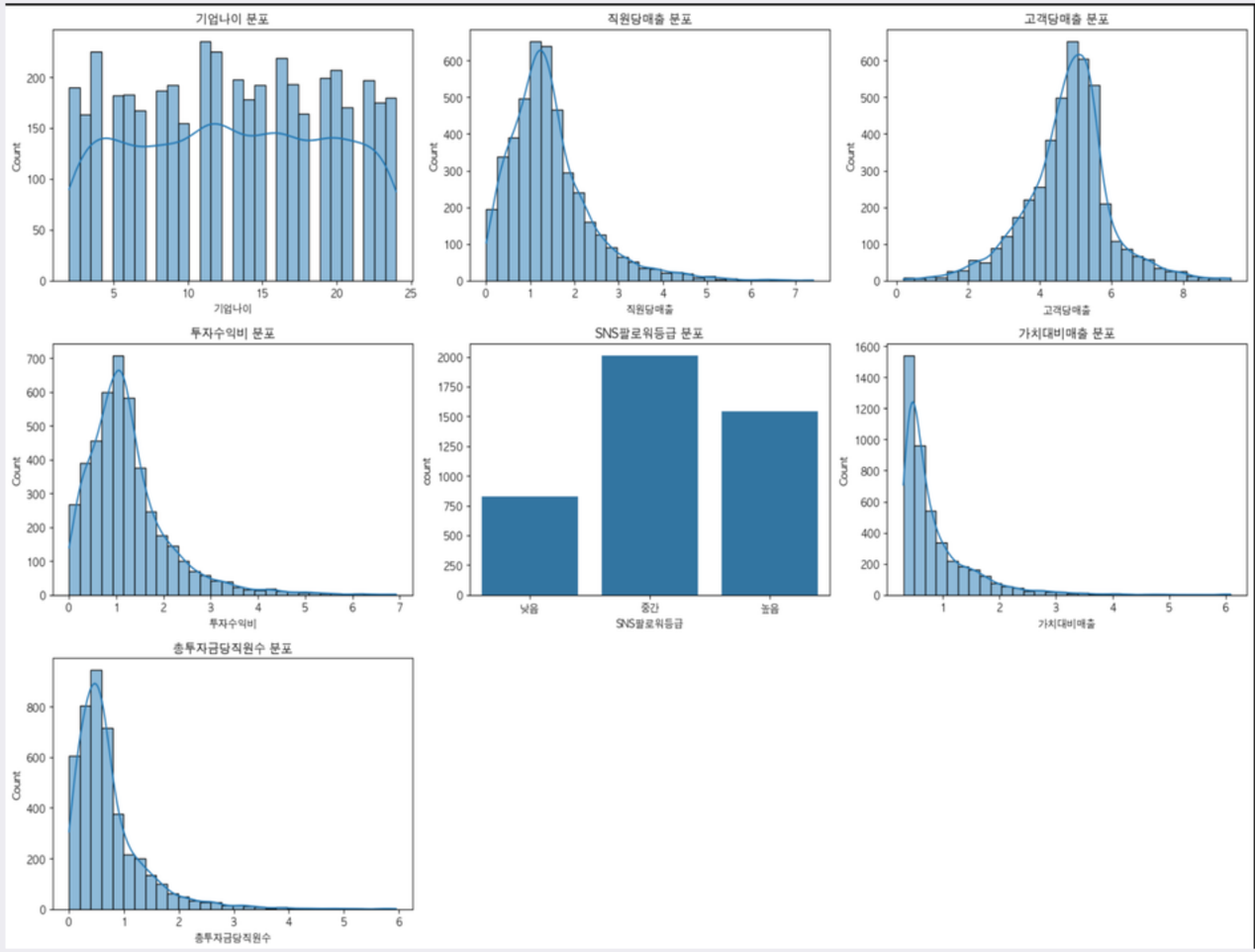


EDA 및 전처리 (파생변수)

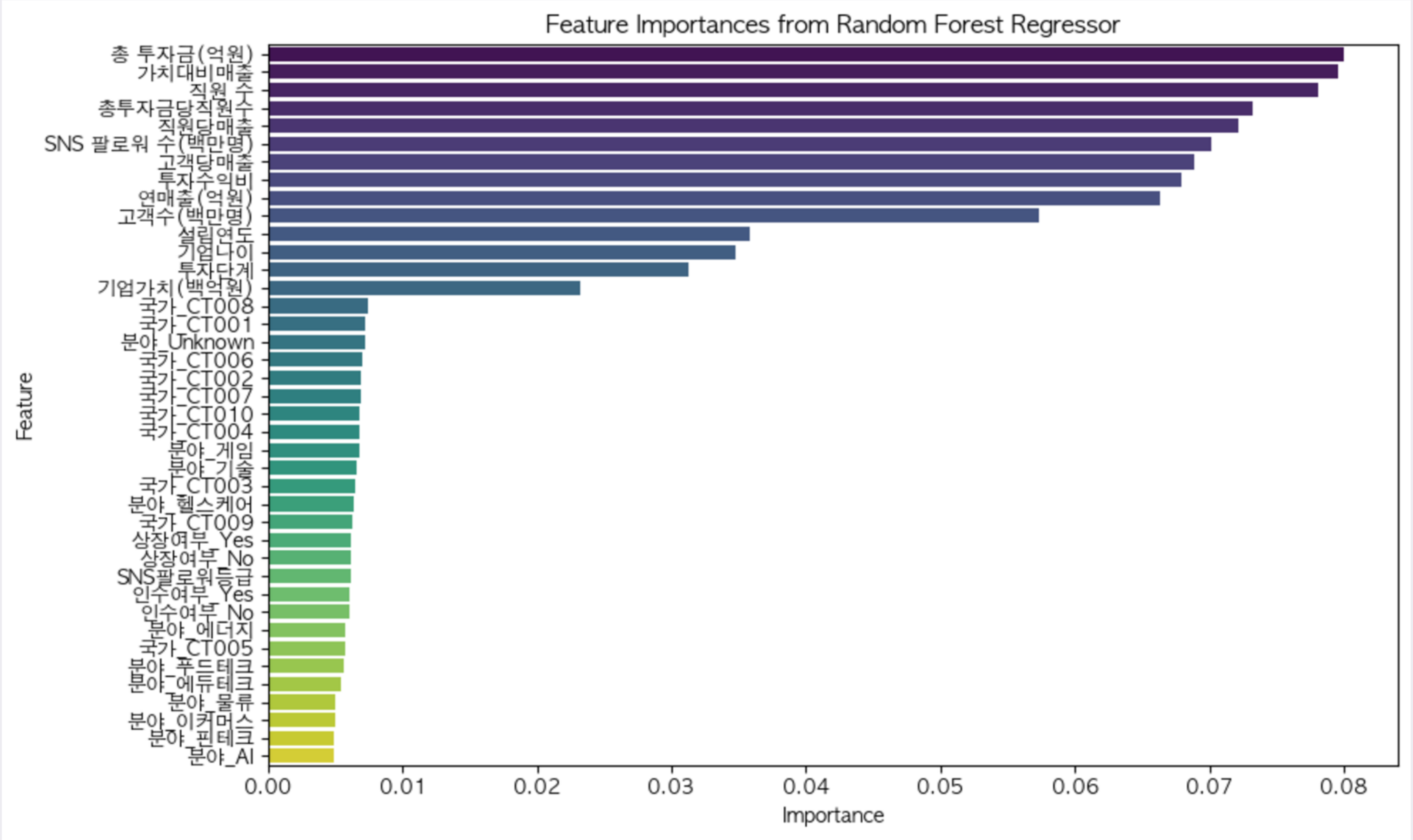


EDA 및 전처리 (파생변수)





EDA 및 전처리 (주요 변수 추출)



전처리 결과 (결과)

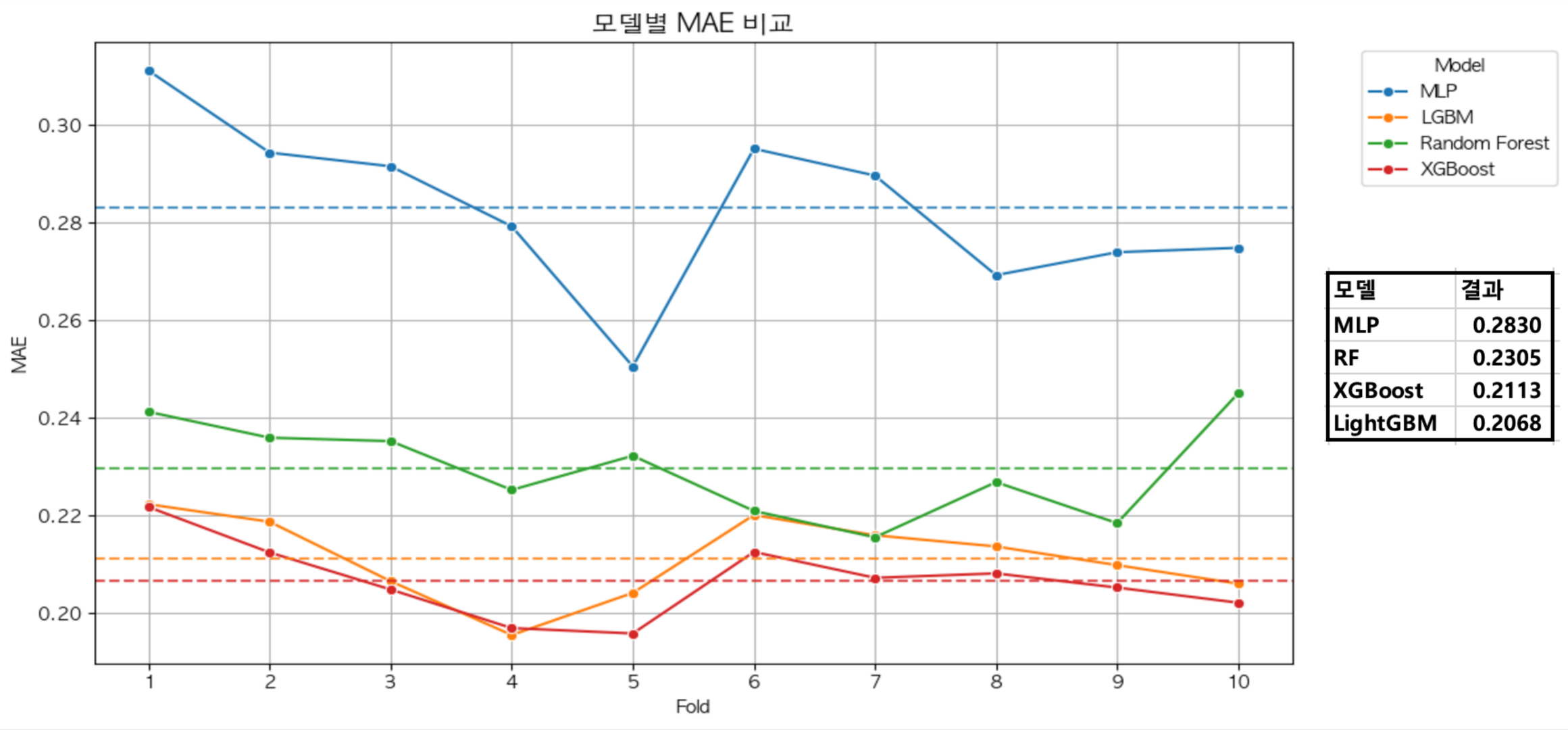
- 학습 데이터

	ID	직원 수	직원당매출	가치대비매출	SNS 팔로워 수(백만명)	연매출(억원)	고객당매출	총 투자금(억원)	투자수익비	투자단계	성공확률
0	TRAIN_3959	0.407170	0.001406	0.001694	0.908	0.318636	0.035461	0.073081	0.008842	0.75	0.1
1	TRAIN_0733	0.609053	0.000414	0.006809	0.400	0.140176	0.002390	0.233858	0.001237	0.75	0.1
2	TRAIN_3824	0.493956	0.003112	0.000438	0.596	0.855976	0.023184	0.586491	0.003014	0.00	0.1
3	TRAIN_4168	0.363509	0.002390	0.001011	0.854	0.483930	0.013109	0.168157	0.005907	0.25	0.1
4	TRAIN_3771	0.333267	0.000830	0.001688	0.352	0.153815	0.004168	0.684972	0.000464	1.00	0.1

- 검증 데이터

	ID	직원 수	직원당매출	가치대비매출	SNS 팔로워 수(백만명)	연매출(억원)	고객당매출	총 투자금(억원)	투자수익비	투자단계
0	TEST_0000	0.653028	0.001019	0.000492	0.400	0.476184	0.012858	0.706330	0.001752	0.75
1	TEST_0001	0.742479	0.000624	0.002381	0.840	0.331500	0.005749	0.224165	0.003833	0.75
2	TEST_0002	0.046330	0.019614	0.001794	0.200	0.662501	0.009035	0.662343	0.002600	1.00
3	TEST_0003	0.126755	0.005479	0.000447	1.000	0.499393	0.035722	0.300860	0.004306	0.00
4	TEST_0004	0.986161	0.000767	0.001273	0.872	0.541384	0.009668	0.702665	0.002003	0.00

실험 결과



모델	결과
MLP	0.2830
RF	0.2305
XGBoost	0.2113
LightGBM	0.2068