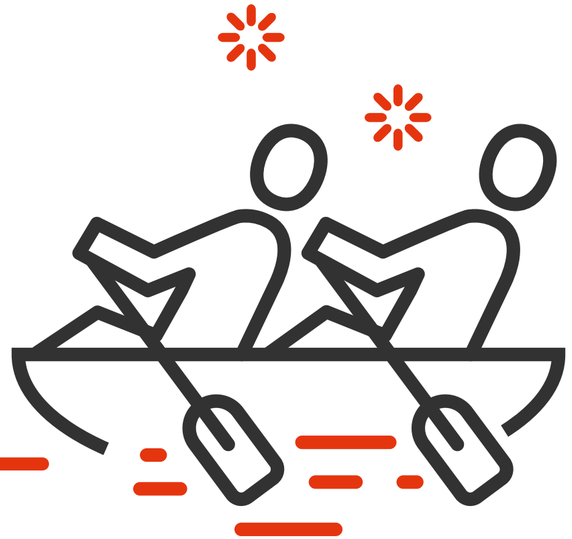
2025. 05. 15. (목)

미래의 성공기업을 발굴하라 !

기업 성공확률 예측 시스템 분석 모델 정의서



양서현 | zzz0x0@gmail.com

1. 분석 과제 정의

분석 과제 소개

과제 목표 및 데이터 소개

- **과제** | 기업 정보 및 경영 데이터 분석을 통한 기업 성공확률 예측 모델 개발
- **과제 목표** |
  - 기업 경영 지표를 기반으로 성공 확률에 영향을 주는 주요 변수 도출
  - 기업의 성공확률을 예측 가능한 모델 개발 및 검증
- **데이터 출처** | Dacon 기업 성공확률 예측 대회
- **데이터 정의** |
  - 다양한 비즈니스 특성을 반영하는 기업 재무 및 경영정보
  - 예측 대상(Label): 기업의 성공 가능성을 0~1 사이 확률로 표현

1. 분석 과제 정의

분석 과제 소개

분석 및 결과 도출 구조도

데이터 불러오기 및 전처리

↓

학습 데이터셋 생성

↓

특성 추출 및 모델 학습

↓

모델 평가 및 예측 결과 도출

↓

웹 기반 예측 시스템 구현 (사용자 입력 기반)

2. 데이터 불러오기 및 전처리 / 학습 데이터셋 생성

입력데이터

기업 정보 목록 (train.csv)

로직

- 컬럼명 재정의
- 결측치 제거
- 데이터 가공 (object→ float)
- 특성 숫자컬럼 추가 (bool→ int)

컬럼수=14, n=4376

결측치 제거
(근거: 결측치 제거 후에도 균등한 분포유지)

컬럼수=14, n=2578

데이터 가공 / 특성 숫자컬럼 추가

기업 가치

[object → float]
' 1500-2500 ' → 2000
' 2500-3500 ' → 3000
' 3500-4500 ' → 4000
' 4500-6000 ' → 5250
' 6000 이상 ' → 6000

투자 단계

[object → 로그끄케일(float)]
' Seed ' → np.log(5)
' Series A ' → np.log(50)
' Series B ' → np.log(175)
' Series C ' → np.log(500)
' IPO ' → np.log(2000)

인수여부 / 상장여부

[object → int]
' Yes ' → 1
' No ' → 0

3. 특성 추출 및 모델 학습

COUNTRY-INDUSTRY 기준 그룹별 회귀 모델 학습

국가(Country) - 산업분야(Industry)를 기준으로 데이터 그룹화

훈련, 학습 (8:2) 데이터셋 분리 (SEED = 10)

[모델 종류]
DecisionTreeRegressor
RandomForestRegressor
LinearRegression

1 groupKey = ["Country", Industry"]
train.groupby(groupKey)

2 특성(features) 추출

label(성공확률) 과 abs(상관계수) > 0.3 인 변수를 features(특성)으로 추출

3 학습 : 훈련 데이터셋 분리

4 회귀 모델 학습

5 예측 및 성능 평가

4. 모델 비교

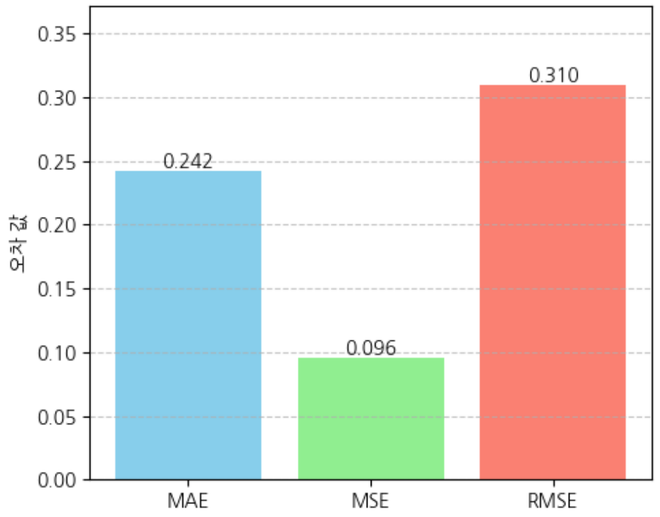
PERFORMANCE

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{(평균 절대 오차)}$$

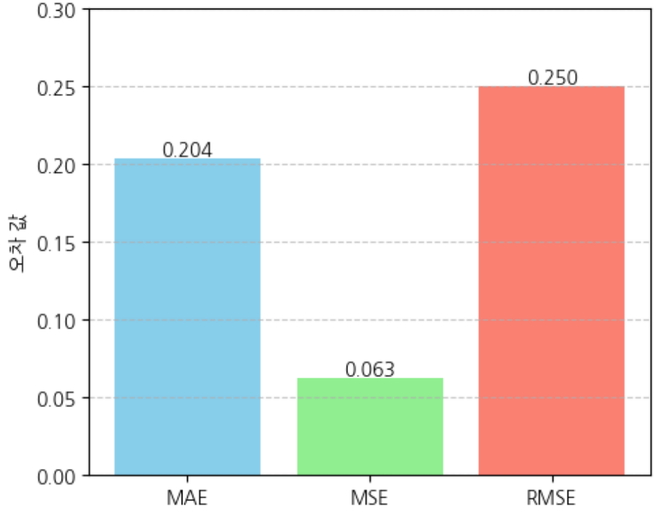
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{(평균 제곱 오차)}$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

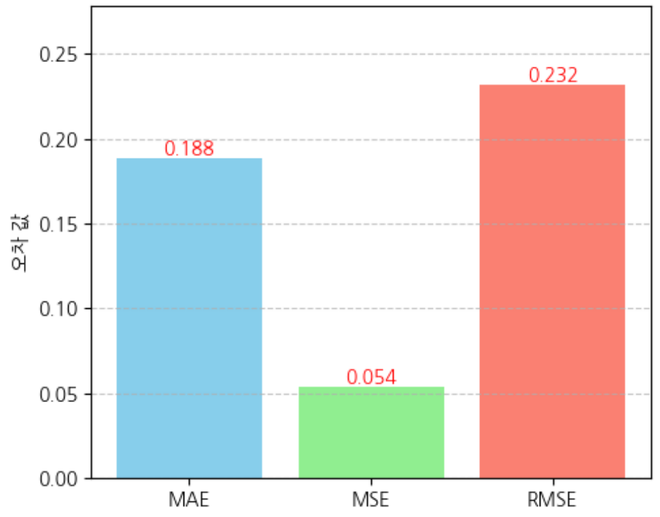
DecisionTreeRegressor



RandomForestRegressor



LinearRegression



THANK YOU.