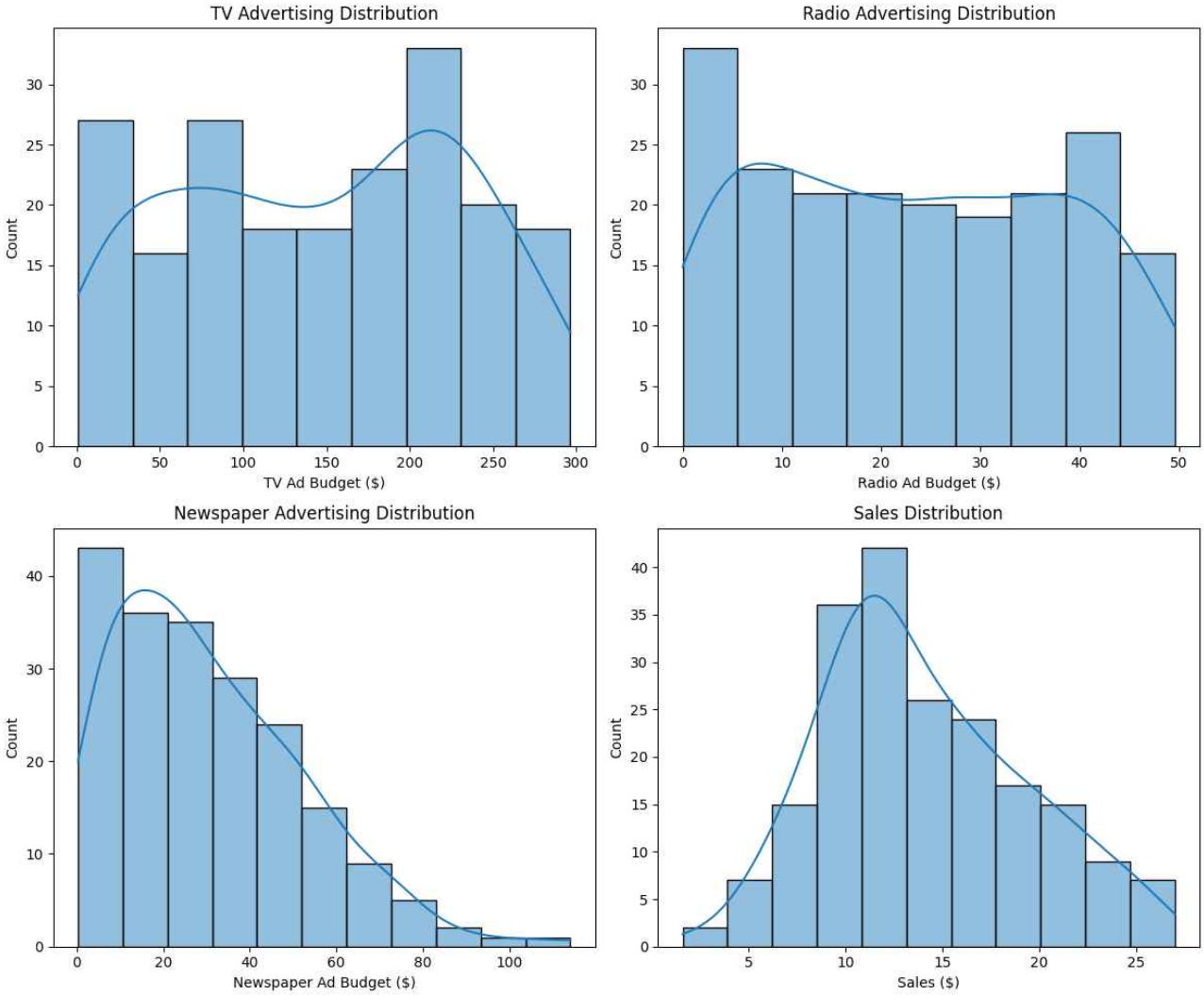
2024년 2학기 기계학습 기말 과제

「Kaggle 데이터를 이용한 머신러닝 예측」

과목명 : 기계학습
담당교수 : 양희정
전공 : 인공지능전공
학번 : 202476627
이름 : 김보겸

1. 데이터 개요

본 프로젝트는 Kaggle의 광고 판매 데이터셋을 활용하여 광고 채널(TV, Radio, Newspaper)별 투자 금액과 매출 간의 관계를 분석함. 데이터는 매출(Sales)을 종속 변수로, 광고 채널별 투자 금액(TV, Radio, Newspaper)을 독립 변수로 설정하여 예측 모델을 구축하는 데 활용됨.



실습 코드 : [Kaggle 데이터 실습](#) (scikit-learn Machine Learning in Python)

2. 모델 성능 비교

분석에서는 선형 회귀, 의사결정 트리, 랜덤 포레스트 세 가지 모델을 학습하였으며, 각 모델의 성능은 R^2(결정계수), MSE(평균 제곱 오차), MAE(평균 절대 오차)로 평가함. 또한 하이퍼파라미터 튜닝을 통해 각 모델의 성능을 개선하고자 함.

모델	R^2 (튜닝 전/후)	MSE (튜닝 전/후)	MAE (튜닝 전/후)
선형회귀	0.907 / 0.907 (릿지)	2.906 / 2.906 (릿지)	1.232 / 1.230 (릿지)
의사결정 트리	0.942 / 0943	1.834 / 1.778	1.005 / 1.019
랜덤 포레스트	0.982 / 0.983	0.579 / 0.535	0.642 / 0.633

3. 결과 분석

1) 선형 회귀

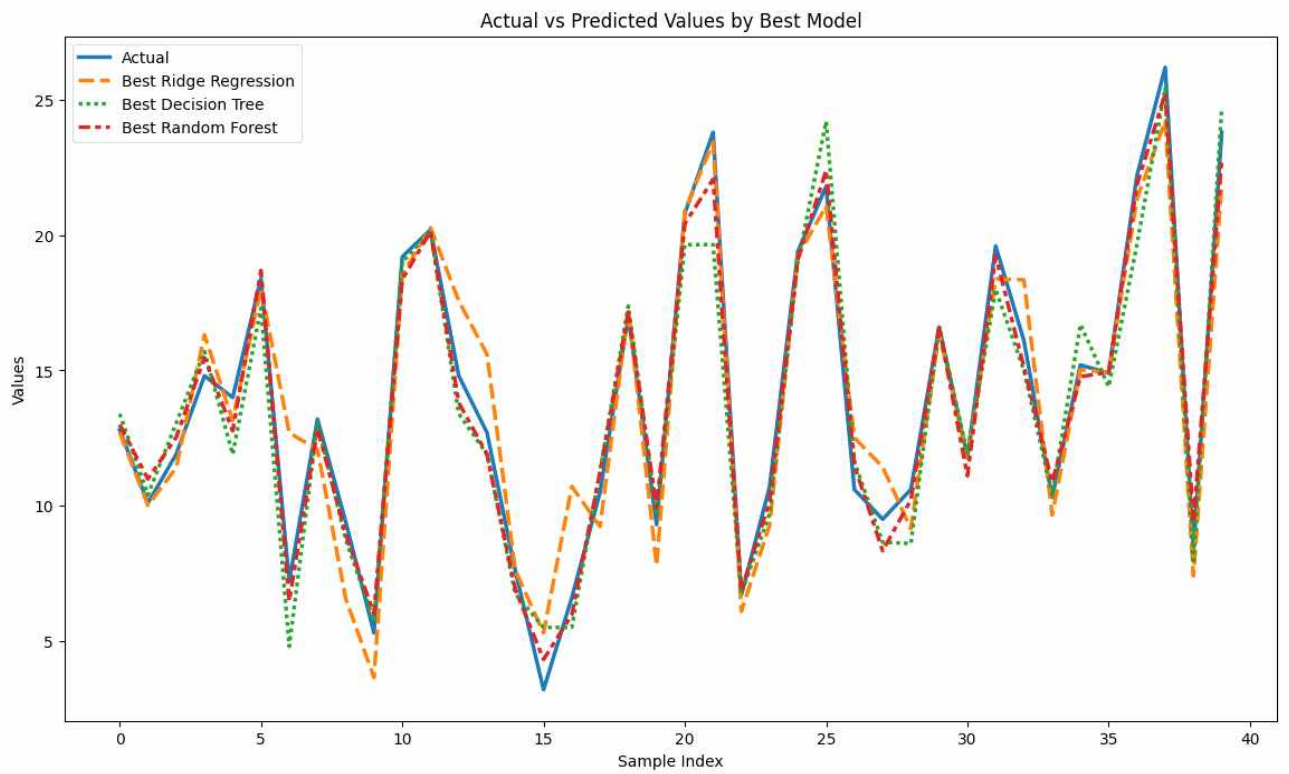
L2 규제를 적용하여 과적합 방지를 시도했으나, 데이터의 비선형 관계를 효과적으로 반영하지 못해 성능 개선이 제한적임을 보여줌.

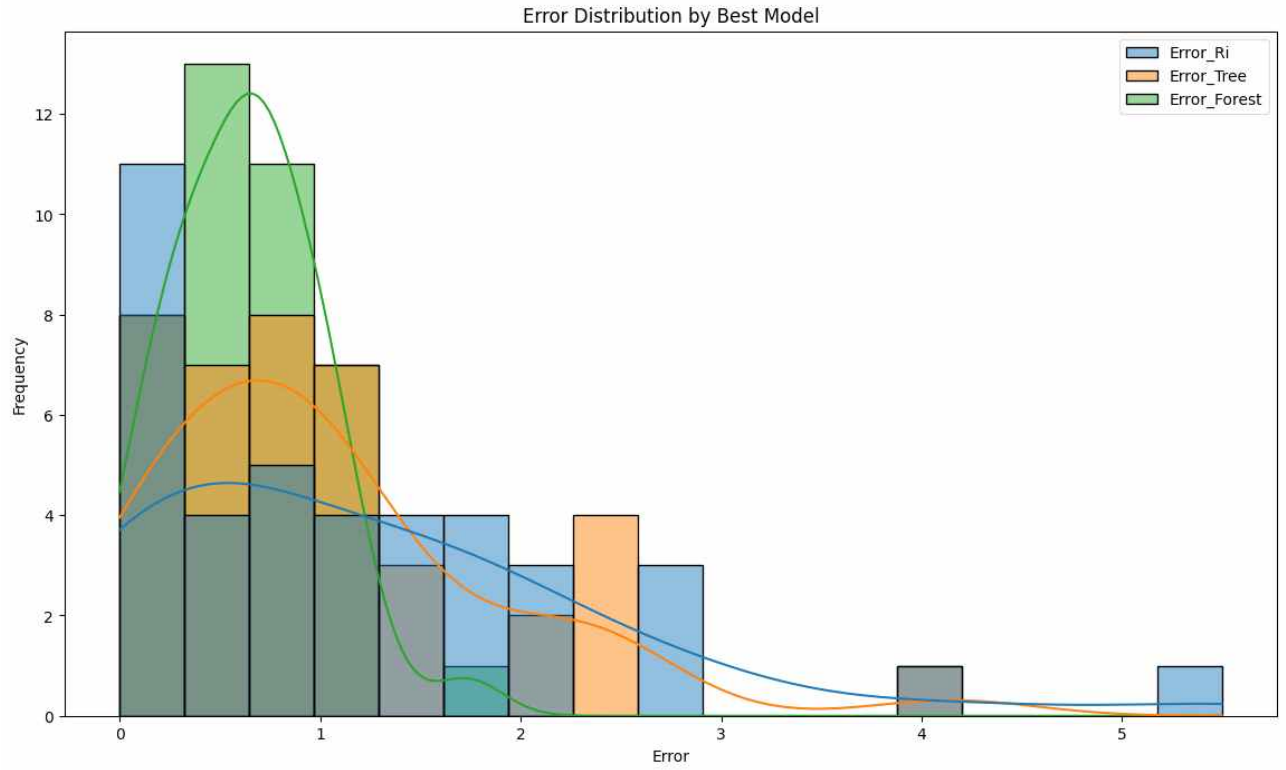
2) 의사결정 트리

트리의 깊이(max_depth), 최소 샘플 수(min_samples_split) 등을 조정하여 과적합을 제한하였지만, 랜덤 포레스트 모델과 비교해 성능이 제한적임을 알 수 있음.

3) 랜덤 포레스트

다수의 트리를 앙상블 하여 의사결정 트리의 단점을 보완함과 동시에 과적합을 방지하는 뛰어난 성능을 보임.





4. 결론 및 제언

1) 결론

랜덤 포레스트 모델은 튜닝 전후 모두 가장 높은 R^2 값을 기록하였으며, MSE와 MAE에서도 가장 낮은 값을 보임. 이는 랜덤 포레스트가 데이터의 비선형 관계를 반영하고 예측 성능이 높은 적합한 모델임이 확인됨.

2) 제언

- 추가 모델 : 본 데이터 분석에 사용된 모델 이외에 다른 모델 검토 필요함. (Elastic Net, XGboost 등)
- 추가 변수 : 추가적인 독립 변수(예: 계절성, 시장 트렌드 등)를 고려하여 모델 성능 개선 필요 요망됨.

5. 참고문헌

- 파이썬 라이브러리를 활용한 머신러닝 개정 2판
- *Kaggle Dataset: Advertising Sales Dataset*