1주차 과제

전공: 컴퓨터공학과 학년: 2학년 학번: 20221519 이름: 구현모

1.

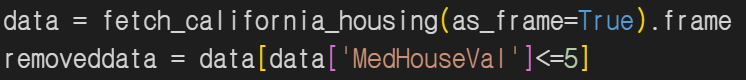


그림 )

그림1)과 같이 원본 데이터에서 5.0이 넘는 데이터를 제외하고 데이터를 입력받는다. 주어진 데이터에서 5.0을 넘는 모든 데이터를 삭제하고 다시 모델을 학습하면 RMSE값이 다르게 출력된다. RMSE값이 변한 이유는 3가지 정도를 찾을 수 있다.

첫번째는 5.0을 넘는 모든 데이터를 삭제하고 모델을 학습시켰기 때문에 데이터의 크기가 감소했기 때문이다. 그림1-1)과 그림2-1)의 상단을 보면 데이터의 크기가 각각 20640, 19628로 나타나 있다. 이처럼 약 1000개의 데이터가 줄어들었다. 이것은 모델이 학습할 수 있는 정보가 적어진 것을 의미한다. 하지만 적어진 데이터의 양이 엄청나게 크지 않고 부자연스러운 정보이기 때문에 모델의 성능이 향상될 수 있다.

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.스크린샷, 텍스트, 도표, 그래프이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

그림-1) 그림-1)

두번째는 데이터의 분포가 변하였기 때문이다. California Housing의 데이터에서 MedHouseVal은 주택의 가격을 가리킨다. 실습에서 사용된 데이터를 기반으로 만들어진 히스토그램(그림1-1))을 살펴보면 5 부근에서 갑작스럽게 데이터가 집중된다. 이것은 데이터들의 부자연스러운 분포를 보인다. 이것을 5.0이 넘는 데이터를 삭제한 히스토그램(그림2-1))을 살펴보면 부자연스럽게 5 근처에 집중되어 있지 않고 자연스러운 모습이 나타난다.

세번째로 이상치를 제거하였기 때문이다. 5.0 초과의 데이터들은 범위를 벗어난 이상치(Outlier)로 볼 수 있다. 하지만 실습에서의 모델은 이러한 이상치들 또한 학습하여 모델에 반영하여 5.0 초과의 데이터를 제거한 모델보다 오차가 커질 수 있다. 부자연스러운 값을 제거하고 학습한 모델이 보다 일반적인 패턴을 학습하게 됨으로 예측 오차인 RMSE값이 변하게 된다. 그림1-2)와 그림2-2)는 각각의 모델이 나타낸 RMSE값이다. 이것을 살펴보면 실습에서 만든 모델보다 과제의 모델이 RMSE값이 작은 것을 알 수 있다. 이는 예측 오차가 더 작은 것을 뜻한다.

텍스트, 폰트, 친필, 타이포그래피이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다. 텍스트, 폰트, 타이포그래피, 그래픽이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

그림 1-2) 그림 2-1)

RMSE값이 감소한 이유를 다시 정리해보면 5.0 초과의 값들을 제거하면서 극단적인 값들을 예측할 필요성이 없어졌으며 데이터들의 분포가 더욱 일관적인 분포임으로 모델이 보다 일반적인 패턴을 학습할 수 있었기 때문이다.

2.

그림d와 같이 모델을 선언하였다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다. 그림

다음은 각각 실습에서의 RMSE값(그림e), Random Forest 모델을 이용한 RMSE값(그림f), 5.0 초과의 값들을 제거한 데이터를 가지고 학습한 Random Forest 모델의 RMSE값(그림g)이다.

텍스트, 폰트, 친필, 타이포그래피이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다. 텍스트, 폰트, 타이포그래피, 디자인이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다. 텍스트, 폰트, 타이포그래피, 친필이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

그림 e 그림 f 그림 g

모두 실습보다 예측 오차가 감소하였으며 그 중 데이터를 전처리 해준 g의 경우가 더 예측오차가 작다. 성능이 더 좋아진 것이다.

실습과 Random Forest 모델을 사용한 것의 성능이 차이나는 이유를 2가지 정도 생각해 볼 수 있다.

먼저 캘리포니아 집값과 여러 관련 데이터들은 단순히 1대1 선형관계를 가지는 것이 아니라 여러 요인들이 섞인 관계를 가지고 있다. 하지만 실습에서 사용한 Linear Regression 모델은 단순한 선형 관계를 가정하여 학습하기 때문에 트리 기반 앙상블로 복잡한 패턴을 학습하는 Random Forest 모델이 데이터의 비선형적인 정보와 제공된 여러 데이터 간의 상호 작용을 보다 잘 적용시켰기 때문이다.

두번째로 제공된 데이터에서 5.0을 초과하는 데이터를 이상치라고 볼 수 있는데 선형 구조인 Linear Regression 모델은 이상치에 상대적으로 민감하고 Random Forest 모델은 트리의 평균화를 통하여 이상치의 영향을 줄일 수 있다. 과제1에서의 결과 중 Test RMSE의 값을 비교해보면 약 13% 감소하였다. 그에 반하여 Random Forest 모델의 Test RMSE의 값을 비교해보면 약 7%감소하였다. 이것으로 Linear Regression 모델이 Random Forest 모델보다 이상치에 민감하다는 것을 알 수 있다. 따라서 이상치에 덜 민감한 Random Forest 모델을 사용한 것의 성능이 더 좋다.

실습에서 Linear Regression 모델을 이용해 구한 Train RMSE와 Test RMSE의 값의 차이가 Random Forest 모델을 이용해 구한 Train RMSE와 Test RMSE의 값의 차이보다 작다. 이것의 이유로는 Random Forest 모델이 과적합되었고 Linear Regression 모델이 보다 일반화되었다는 것을 의미한다. 하지만 과제에서는 Random Forest 모델에 대하여 기본값 Hyperparameter모델을 선언하였기 때문에 Hyperparameter를 적절히 조정한다면 과적합을 해결할 수 있을 것이다. 또한 학습 데이터를 늘리거나 노이즈를 줄이는 방법도 존재한다. 그렇기 때문에 절대적인 Train RMSE와 Test RMSE의 값이 낮고 Hyperparamete를 조절하여 과적합 문제를 해결할 수 있는 Random Forest 모델을 사용하는 것이 더 좋다고 생각한다.