Q1. Training Set, Testing Set의 점수가 유사하다는 것은 무엇을 의미할까요?

Training set score 값과 Testing set score 값이 낮으면서 유사한 경우는 학습된 모델이 **'과소적합 (Underfitting)'** 되었다는 것을 의미함.

sci-kit learn 라이브러리의 Linear Regression이 사용하는 Score 값은 R^2 (Coefficient of determination, 결정계수)로 학습된 모델이 얼마나 Dataset의 분포를 유사하게 흉내 내는지를 보여주는 척도임. R^2는 평균 오차의 제곱값과 분산의 비율을 1로 뺀 것으로 구성되어있음. **일종의 Prediction에 대한 오차 분포 비율**이라 생각할 수 있음. 그러므로 R^2은 오차 (Error)가 작을 수록 높은 값을 가지며, Dataset에 대한 Prediction과 정답 데이터 사이에 오차가 0인 경우 R^2은 최대 1이 될 수 있음. **R^2가 높을수록 회귀분석(Regression)의 결과 모델이 Dataset을 더 잘 표현하는 것이라 할 수 있으며, Prediction의 정확도(Accuracy)가 높다고 할 수 있음.**

Training set score R^2값과 Testing set score R^2값이 거의 유사하다는 것은 학습된 모델이 Train과 Test시 동일한 Feature에 대해 반응하여 학습한다는 것을 의미함. 그러나 **Training set score R^2값과 Testing set score R^2값이 낮으면서 서로 유사한 경우에는 모델이 Train과 Test시 반응하는 동일 Feature가 Dataset의 전체를 표현하는 데에 부족하거나 부적합하다는 것임.**

사용하는 Train Dataset, Test Dataset의 개수에 비해 학습에 사용하는 Feature의 개수가 적으면 편중된 (Highly Biased) 학습인 과소적합이 발생함.

그러므로 Training set score 값과 Testing set score 값이 낮으면서 유사한 경우는 모델이 특정 소수의 Feature에 대해서만 High Bias를 가지고 학습된 현상이며, 이때 과소적합이 발생한다고 할 수 있음.

Q2. Training set의 점수는 높으나 Testing set의 점수는 낮다는 것은 무엇을 의미할까요?

Training set score 값이 높으면서 Testing set score가 낮은 경우는 학습된 모델이 **'과대적합 (Overfitting)'** 되었다는 것을 의미함.

sci-kit learn 라이브러리의 Linear Regression이 사용하는 Score 값은 R^2 (Coefficient of determination, 결정계수)로 학습된 모델이 얼마나 Dataset의 분포를 유사하게 흉내내는지를 보여주는 척도임. **R^2가 높을수록 회귀분석(Regression)의 결과 모델이 Dataset을 더 잘 표현하는 것이라 할 수 있으며, Prediction의 정확도(Accuracy)가 높다고 할 수 있음.**

Training set score 값이 높으면서 Testing set score가 낮다는 것은 모델이 **Train Dataset의 Feature에 대해 과도하게 최적화되어 Train 이후에 들어오는 새로운 데이터인 Test Dataset에 대해서 유연하게 반응하지 못한다는 것을 의미함.** Train 단계에서 모델이 **너무 많은 Feature를 학습할 경우 모든 Feature에 대해 오차를 최소화하려는 경향 때문에 Train Dataset의 Feature에 딱 맞는 Weight값을 가짐.** 이로 인해 **여러 Feature에 대해 Prediction 결과가 배치되기 때문에 전반적으로 Feature vs Prediction의 분산(Variance)이 커짐.**

사용하는 Train Dataset, Test Dataset의 개수에 비해 학습에 사용하는 Feature의 개수가 많으면 모든 Feature를 모두 학습하려는 분산된 (High Variance) 학습인 과대적합이 발생함.

그러므로 Training set score값이 높으면서 Testing set score값이 낮은 경우는 모델이 너무 많은 Feature에 대해서 모두 학습하면서 분산된 학습을 하여 Train Dataset에만 최적화되는 현상이며, 이때 과대적합이 발생한다고 할 수 있음.