스케일러 개요

데이터 모델링에 있어서 스케일링은 필수 과정이다. 스케일링을 통해 다차원의 Feature를 분석하기 쉽게 만들어주며, 자료의 overflow, underflow를 방지하는 효과도 있어 모델링에 있어 overfitting, underfitting을 예방하는 효과를 낼 수 있다. 즉, 최적화 과정에서의 안정성을 얻는다.

회귀 분석에서의 Condition Number

함수의 Condition Number는 민감도로 이해하면 된다. argument 에서 작은 변화에 비율에 대해 함수가 얼마나 변화할 수 있는지에 대한 argument measure이다.

* 주로 변수들의 단위 차이로 인한 스케일 차이는 스케일링으로 해결한다
* 다중 공선성, 상관관계가 큰 독립 변수가 있는 경우 변수 선택이나 PCA를 사용한다.

다루는 데이터가 큰 수치 데이터인 경우 로그를 취하는 경우도 있다.

스케일링의 종류

사이킷 런에는 다양한 종류의 스케일러를 제공하고 있고, 아래는 우리가 사용한 4가지이다.

1. Standard Scaler

평균을 제거하고, 데이터를 단위 분산으로 조정한다.

이상치가 있는 경우 표준편차에 영향을 미쳐 균형 잡힌 스케일링을 보장할 수 없다.

1. MinMax Scaler

모든 Feature 값이 0~1 사이에 있도록 재조정 한다. 이 때도 마찬가지로 아웃라이어의 존재에 있어 매우 민감하다.

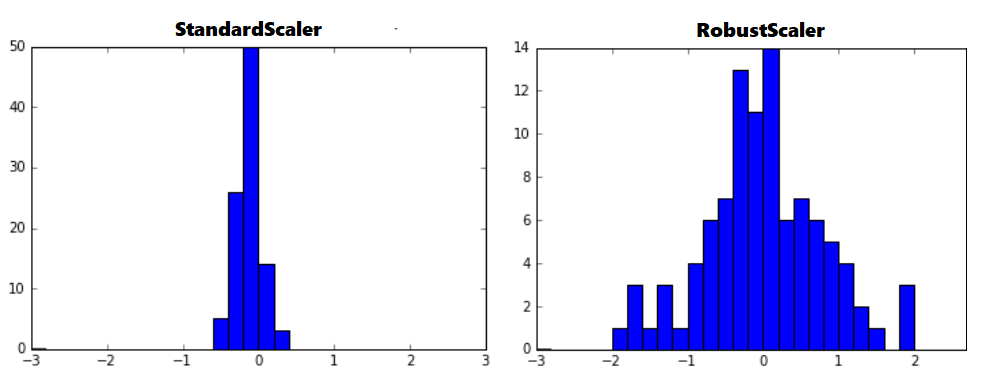
1. MaxAbs Scaler

절대값이 0~1 사이로 (-1~1) 재조정한다. 시그모이드가 생각난다. 양수 데이터로 구성된 데이터 셋에서는 2번 스케일러와 유사하게 동작한다. 2번에 비해 조금 더 이상치의 영향을 덜 받는다

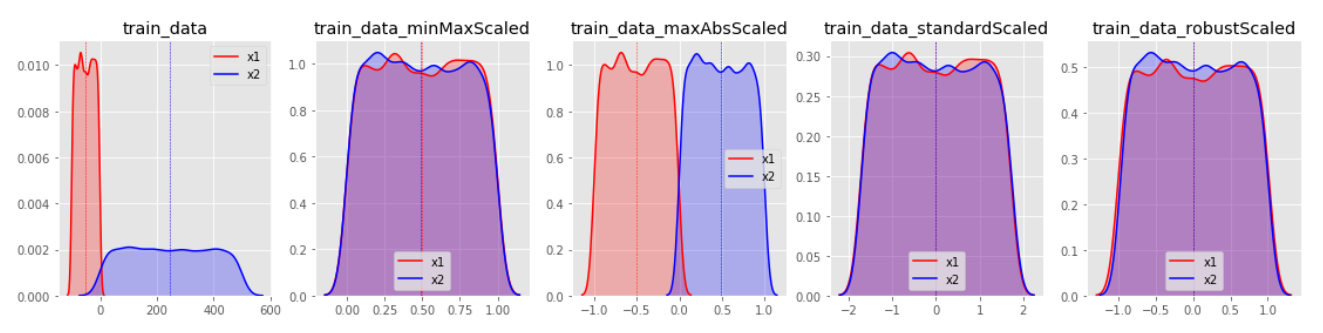
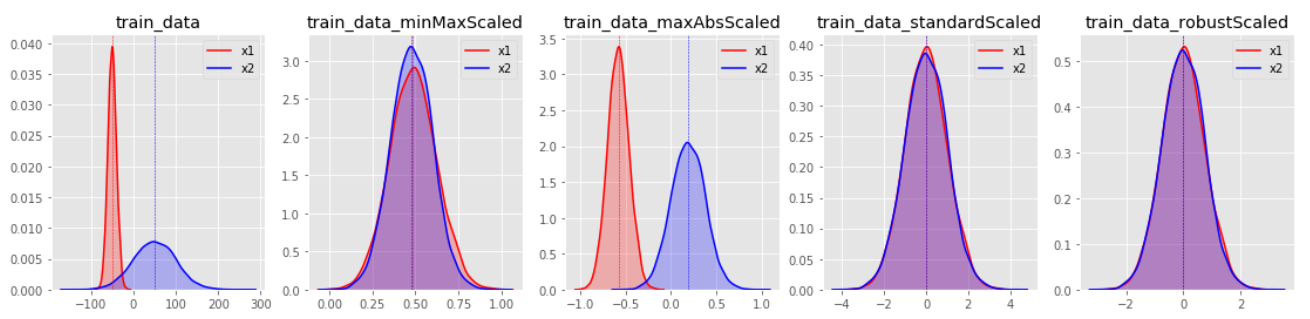
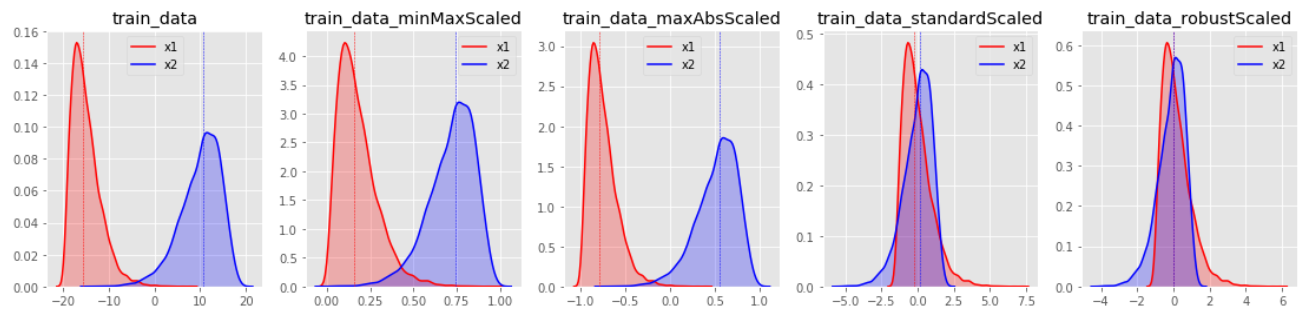
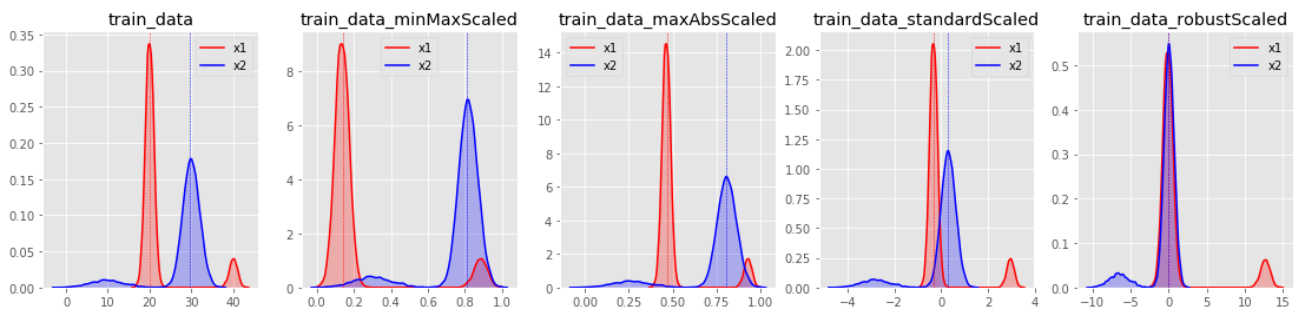
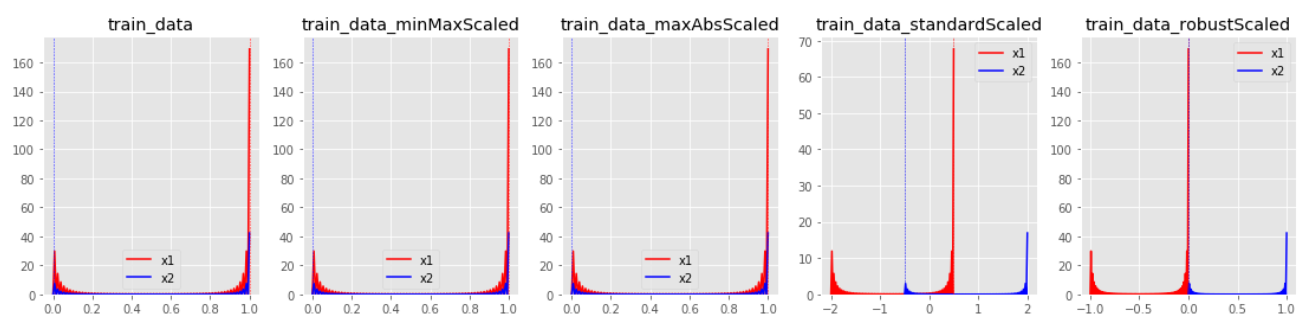
1. Robust Scaler

이상치의 영향력을 최소화 하기 위해, 중앙값과 InterQuartileRange를 사용한다.

1번에 비해 표준화 후 동일한 값을 더 넓게 분포 시킨다.



아래는 스케일러와 데이터 분포에 따른 변환 결과이다.

Standard Scaler와 Robust Scaler는 표준화 된 느낌이 있는데

MinMax, MaxAbs Scaler는 데이터의 쏠림 현상이나 표준화 된 느낌이… 아마 아웃라이어 때문일 것 같은데

<https://datascienceschool.net/view-notebook/afb99de8cc0d407ba32079590b25180d>

<https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html>

이건 실습 예제가 같이 있는 자료

<https://m.blog.naver.com/PostView.naver?blogId=demian7607&logNo=222009975984&proxyReferer=https:%2F%2Fwww.google.com%2F>

이건 시각화가 잘 된 자료

<https://sosoeasy.tistory.com/385>

Normalizer 설명이 있는 자료

https://wooono.tistory.com/96