Scaler 종류에 따른 부도 예측 결과에 대한 영향성 분석

송찬우 안승규 박종현 안현철

국민대학교 국민대학교 국민대학교 국민대학교

비즈니스 IT 전문대학원 비즈니스 IT 전문대학원 비즈니스 IT 전문대학원 비즈니스 IT 전문대학원

cksdnthd2008@kookmin.ac.kr tmdrb0415@kookmin.ac.kr ppjjhh1027@kookmin.ac.kr hcahn@kookmin.ac.kr

Abstract *- 머신러닝 알고리즘은 데이터의 스케일 조정에 매우 민감하며, 통상적으로 전처리 단계에서 각 특성 스케일을 조정해 데이터를 가공한다. 스케일러의 종류는* ***Standard Scaler, Min-Max Scaler, Robust Scaler, Normalizer*** *등이 있는데 부도예측에 있어 어떤 Scaler 를 사용했을 때 가장 우수한 정확도를 보이는지는 알려진 바가 없다.*

이에 본 연구는 부도 예측 데이터를 통해 4 가지의 Scaler 를 적용하여 어떠한 것이 가장 우수한 정확도를 보여주는지 알아보고자 한다.

Key Terms – 부도예측, Standard Scaler, Min-Max Scaler, Robust Scaler, Normalizer

I. 서론

'Data Industry Promotion Strategy – I-KOREA 4.0 Data Field Plan, I-DATA+'(2018)이란 보고서에서는 4 차 산업혁명을 견인하는 핵심 동인인 빅데이터를 통해 사회문제 해결 능력을 강화하는 것을 핵심 과제로 정했다. 또한 2018 년 3 월 4 차산업혁명위원회의 첫 회의에서 문재인 전 대통령은 인공지능, 사물인터넷, 빅데이터를 위한 투자를 확대하여 혁신생태계를 조성할 것임을 밝히게 되면서 금융 분야에서의 빅데이터 분석에 관한 연구의 필요성이 높아지고 있다(차성재와 강정석, 2018).

기업의 부도는 그 기업의 경영자, 노동자 측면에만 국한되는 것이 아니라 그 기업에 직· 간접적으로 연관관계를 가진 이해관계자(투자자, 금융기관, 거래기업 등)에게도 연쇄적인 피해를 양산할 수 있다. 더 나아가 국민경제에도 심각한 타격을 미칠 수 있다(강치형과 신해수, 2015).

따라서 기업의 부도를 예측하는 것은 지속하여 연구되고 있고, 예측 정확도를 보다 높이고자 노력하고 있다.

그러한 데이터 분석 기반 의사결정 지원 시스템을 개발하기 위해서 데이터 전처리가 필요하다. 데이터 전처리 단계에는 Data cleaning, Pruning, Feature selection, scaling 이 존재하지만, 대부분의 부도예측 연구는 이에 집중하지 않고 다른 알고리즘을 적용하며 정확도를 높이고자 하였다(Ahsan, Md Manjurul, et al, 2021).

본 연구에서는 기존 연구된 부도 예측 모델에 4 가지 Scaling 기법을 적용하여 어떠한 영향을 미치는지 평가를 하는 것이 목적이다.

II. 이론적 배경

데이터 전처리란 예측모형 등의 성과를 향상시키기 위해, 입력 데이터에 대해 사전 처리를 수행하는 기법을 의미한다.

데이터 전처리 과정은 중복 항목 제거, 결측값 처리, 정성적 변수 정량화, 이상치 제거, 새로운 파생변수 개발, 데이터의 정규화 또는 표준화, 자료의 구분, 모형에 사용될 후보 입력 변수 선정 등이 있다. 부도예측에서 사용하는 데이터의 경우 독립변수들의 단위와 범위가 다르기 때문에 정규화 또는 표준화가 중요하다.

이러한 Scaling 기법에는 대표적으로 다음과 같은 4 개의 방법이 존재한다.

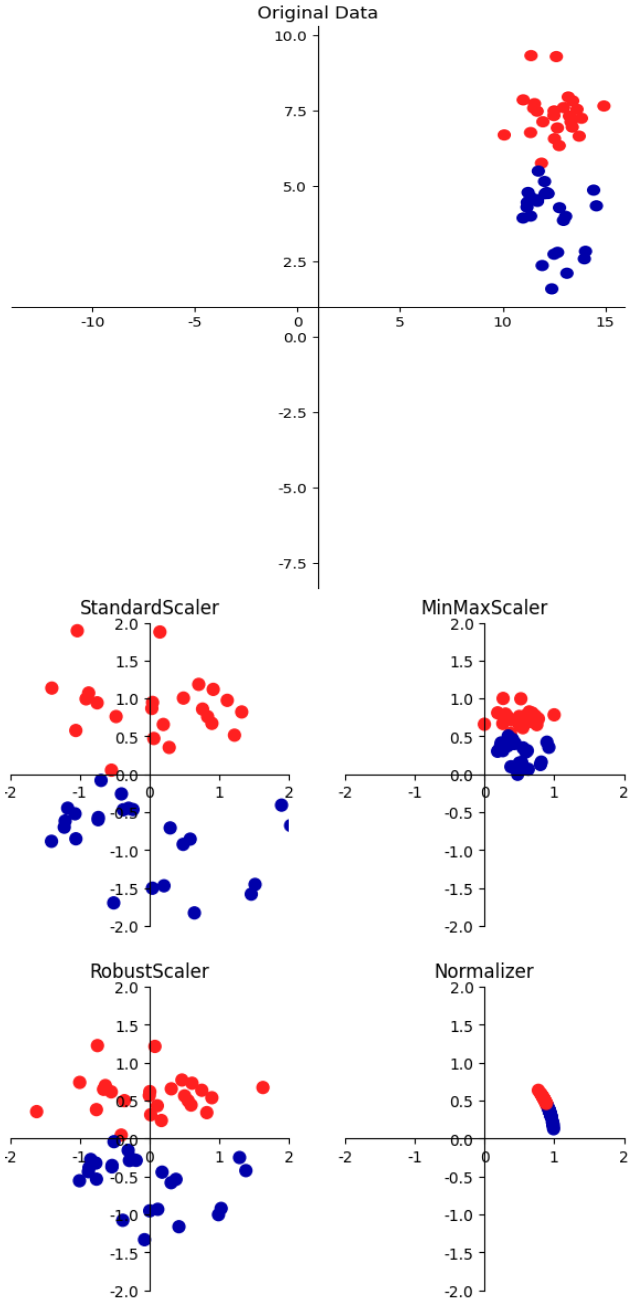
(1) $Standard\ Scaler : z = \frac{x - \mu}{\sigma}$

(2) $Robust\ Scaler : r = \frac{x - median}{IQR} = \frac{x - median}{Q_3 - Q_1}$

(3) $Min - Max\ Scaler : x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$

(4) $Normalizer : |x|_2 = \sqrt{(\sum x_i^2)}$

각 특성 평균을 0, 분산을 1 로 변경해주는 Standard Scaler, 이와 유사하지만, 평균과 분산 대신 중간 값과 사분위 값(IQR)을 사용하는 Robust Scaler, 모든 특성이 정확하게 0 과 1 사이에 위치하도록 변경하는 Min-Max Scaler, 특성 벡터의 Euclidean Distance 가 1 이 되도록 데이터 포인트를 조정하는 Normalizer 가 있다(Andreas C.Muller, Sarah Guido 2016) 이 때, Scaling 을 진행하기 이전 원본 데이터와 각 Scaler 를 적용하였을 경우에 대하여 시각화하면 아래 <그림 1>과 같다.

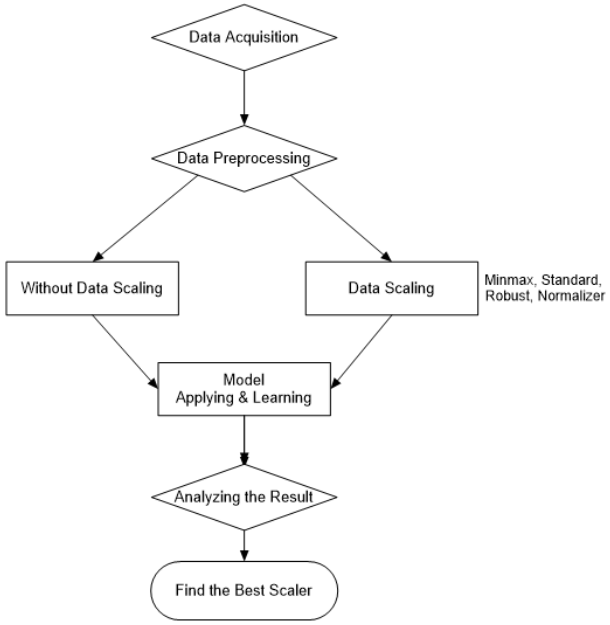


<그림 1> Scaler 별 시각화 형태

III. 제안모형

본 연구에서 제안하는 모형은 아래 <그림 2> 와 같다. 먼저 부도 예측 데이터는 2001~2007 년 제조업 기업의 재무 정보를 활용한다.

부도 예측 데이터는 각 재무 비율이 모델에 미치는 영향이 다르기 때문에 데이터 전처리를 통해 편향을 줄여줄 것이다. 다음으로 상기 <그림 1> 에 제시한 Scaler 를 Model 에 적용하여 결과를 비교해보고 최종적으로 부도 예측에 가장 적절한 Scaler 를 도출하고자 한다.



<그림 2> 제안 모형

IV. 실험

아래 <표 1> , <표 2>, <표 3>의 결과를 보면 SVM, 로지스틱 회귀 기법을 사용하는 경우 Scaling 적용 유무에 따라 차이가 명확했다. 하지만 Tree / Boosting 계열 모델의 경우 스케일링을 사용하지 않고 예측하는 것과 안하고 예측하는 것은 유의미한 차이가 없었다.

좀 더 자세히 보면 SVM 의 경우 Scaler 없이는 Recall 0.496 Precision 0.969 를 보였고 Min-Max Scaler 는 Recall 0.867 Precision 0.976, Standard Scaler 는 Recall 0.858 Precision 0.988, Robust Scaler 는 Recall 0.841 Precision 0.987, Normalizer 는 Recall 0.712 Precision 0.686 이라는 결과가 나왔다. 부도예측의 경우 Precision 의 차이를 비교하기보단 Recall 을 비교하는

것이 맞다고 판단하여 Min-Max, Standard, Robust, Normalizer, Non-Scaler 순으로 높은 정확도를 보여주었다.

<표 1> SVC Model 실험 결과

	F1 Score	Recall	Precision	Accuracy
Non-Scaler	0.656	0.496	0.969	0.740
Min-Max	0.918	0.867	0.976	0.923
Standard	0.918	0.858	0.988	0.924
Robust	0.908	0.841	0.987	0.915
Normalize	0.698	0.712	0.686	0.693
Max.	**0.918**	**0.867**	**0.988**	**0.924**
Min.	**0.656**	**0.496**	**0.686**	**0.693**

<표 2> Logistic Regression 실험 결과

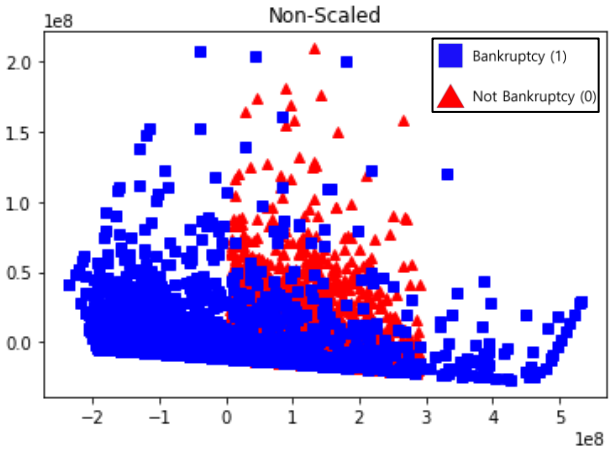
	F1 Score	Recall	Precision	Accuracy
Non-Scaler	0.161	0.088	0.943	0.542
Min-Max	0.903	0.839	0.977	0.910
Standard	0.903	0.840	0.976	0.910
Robust	0.902	0.839	0.974	0.908
Normalize	0.316	0.194	0.848	0.580
Max.	**0.903**	**0.840**	**0.977**	**0.910**
Min.	**0.161**	**0.088**	**0.848**	**0.542**

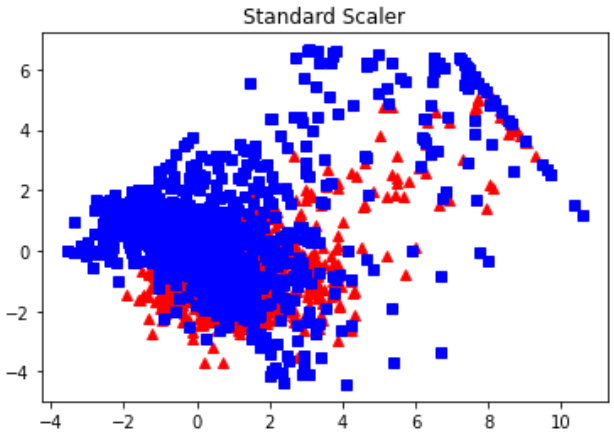
<표 3> LightGBM 실험 결과

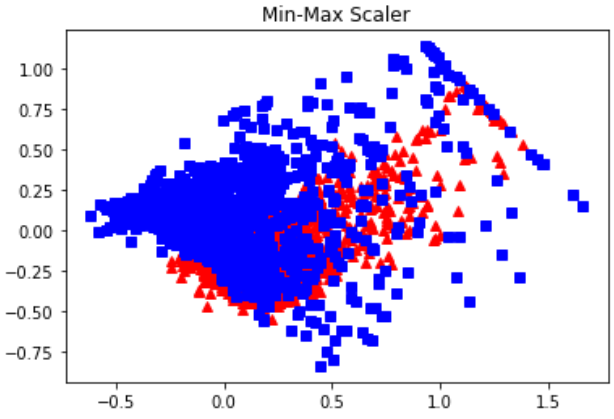
	F1 Score	Recall	Precision	Accuracy
Non-Scaler	0.919	0.870	0.973	0.923
Min-Max	0.920	0.870	0.977	0.925
Standard	0.923	0.871	0.982	0.928
Robust	0.923	0.871	0.981	0.927
Normalize	0.814	0.752	0.887	0.828
Max.	**0.923**	**0.871**	**0.982**	**0.928**
Min.	**0.814**	**0.752**	**0.887**	**0.828**

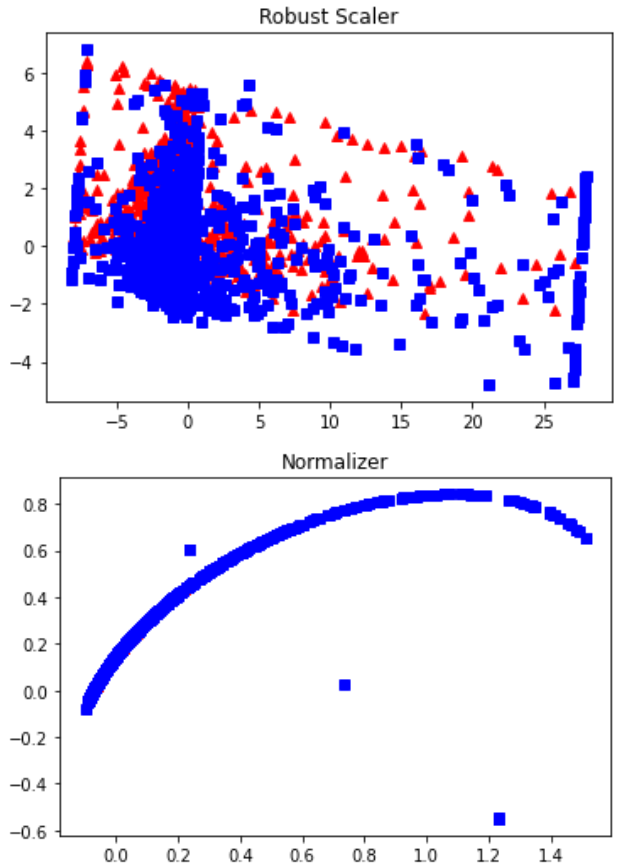
Logistic Regression 의 경우 Scaler 없이는 Recall 0.088 Precision 0.943 를 보였고 Min-Max Scaler 는 Recall 0.839 Precision 0.977, Standard Scaler 는 Recall 0.840 Precision 0.976, Robust Scaler 는 Recall 0.839 Precision 0.974, Normalizer 는 Recall 0.194 Precision 0.848 이라는 결과가 나왔다. 위와 마찬가지로 Recall 을 기준으로 나열하되 같은 경우 Precision 을 고려하여 나열하자면 Standard, Min-Max, Robust, Normalizer, Non-Scaler 순으로 높은 정확도를 보였다.

위 결과에서 시사할 점은 Normalizer 의 경우 부도예측에 있어 올바른 스케일링 기법이 아니라는 것이다.







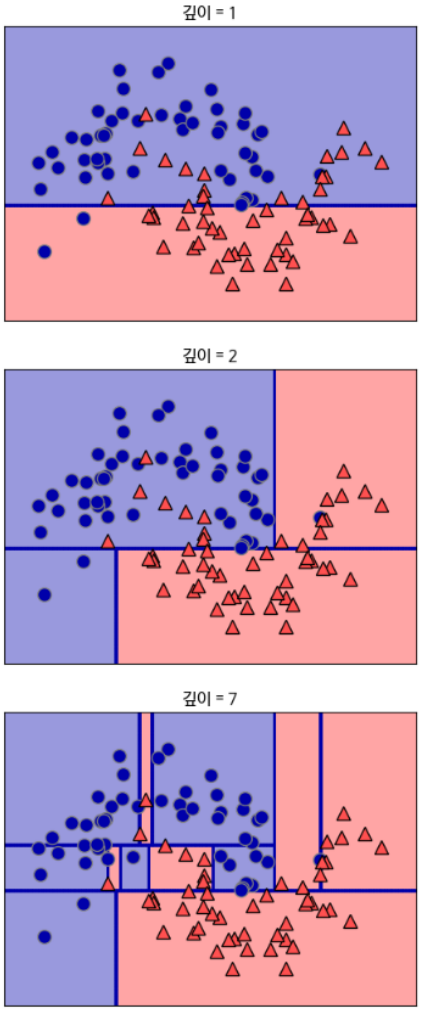


<그림 3> Scaler 에 따른 데이터 분포

Scaler 의 종류에 따라 데이터의 분포는 <그림 3>와 같다. 이 중 Normalizer 에서는 데이터의 분포가 부도인 경우와 아닌 경우가 일직선상에 놓여 있어 예측을 제대로 못하는 것이라 추측된다.

Tree / Boosting 계열 모델 중 LightGBM 의 경우 Scaler 없이는 Recall 0.870 Precision 0.973 를 보였고 Min-Max Scaler 는 Recall 0.870 Precision 0.977, Standard Scaler 는 Recall 0.871 Precision 0.982, Robust Scaler 는 Recall 0.871 Precision 0.981, Normalizer 는 Recall 0.752 Precision 0.887 이라는 결과가 나왔다. Normalizer 를 제외한 Non-Scaler, Min-Max, Standard, Robust 는 결과의 큰 차이가 없었다.

위 결과에서 시사할 점은 Tree / Boosting 계열 모델에서도 Normalizer 는 똑같이 성능이 떨어졌다는 점과 Tree / Boosting 계열 모델에서는 스케일링이 따로 필요하지 않다는 점이다.



<그림 4> Tree/Boosting 계열

그 이유로는 <그림 4>를 통해 추측할 수 있다. Tree / Boosting 계열의 경우 리프(Leaf) 수에 따라 데이터를 나누기 때문에 스케일링의 유무와 관련없이 비슷한 정확도를 보인다고 생각된다. 하지만 Tree/Boosting 계열의 경우 느린 수행 시간과 과적합이 빈번하게 발생한다는 문제점이 있다. 그런 경우 SVM, Logistic Regression 을 사용해 규제를 주어 그러한 문제점을 해결할 수 있기에 Scaling 은 우리가 예측하는데 있어 중요하다.

V. 결론

본 연구는 데이터 전처리 과정에서 Scaler 가 부도 예측 모델에 어떠한 영향을 미치는지에 대해 실험을 진행하였다..

Tree / Boosting 계열을 제외한 예측 모델에서는 Standard, Min-Max, Robust, Normalizer 중 Normalizer 를 제외한 나머지에서는 비슷한 성능을 보였다. 반면 Tree /Boosting 계열에서는 Data

Scaling 의 적용 유/무에 관련 없이 비슷한 성능을 보였다. 따라서 어떠한 알고리즘 모델을 적용하는가에 따라 알맞은 방법의 Scaling 이 필요하다고 생각된다. 부도예측에 있어 SVM, 로지스틱을 사용할 경우 Standard Scaler 나 Min-Max Scaler 를 사용하는 것이 비교적 낫다고 판단된다.

다만, 본 실험에서는 한 개의 Data Set 을 가지고 SVM, Logistic, LightGBM 을 사용해서 진행하였기 때문에 일반화할 수 없다. 따라서 후속 연구에서는 여러 부도 데이터를 가지고 알고리즘을 늘려 분석 결과의 타당성을 높이는 것이 좋다고 생각된다.

참고문헌

차성재, 강정석. (2018). 딥러닝 시계열 알고리즘 적용한 기업부도예측모형 유용성 검증. 지능정보연구,24(4),1-32.

강치형, 신해수. (2015). 회원제 골프장기업의 부도예측 모형개발. 관광연구논총, 27(4), 241-269.

Ahsan, Md Manjurul, et al. (2021) "Effect of data scaling methods on machine learning algorithms and model performance." *Technologies* 9.3: 52.

Andreas C. Muller & Sarah Guido, “Introduction to Machine Learning with Python”