Random Forest를 이용한 청소년 자살 위험 예측

<인간 뇌이미징의 데이터사이언스> 기말보고서

송상록 / 2019-12213

취약성-스트레스 모형에 따르면, 아동, 청소년의 자살 사고(suicidal ideation)에는 다양한 생물학적 및 환경적 변인이 영향을 미칠 수 있다. 따라서 각종 변인을 데이터셋으로 정리하고 이를 활용해 자살사고 여부를 예측하는 머신러닝 분류 모형을 훈련하는 것은 자살 위험군 아동 및 청소년을 식별하는 데 유용할 수 있다.

예를 들어, Joo et al.(2022)는 polygenic scores를 이용하여 아동, 청소년의 자살 사고 여부를 예측하는 모형을 훈련하였다. 하지만 실제로 자살에 영향을 미치는 요인은 유전 외에도 다양하므로, 다른 data를 활용했을 때 모형의 성능 역시 확인해야 한다. 따라서 본 보고서에서는 polygenic scores 외에도 청소년의 CBCL, dMRI, sMRI, fMRI 및 인구통계 데이터를 활용하여 머신러닝을 수행하고, 각 모형의 성능을 비교 분석하였다.

실험 1. 데이터 전처리

방 법

보다 효율적으로 데이터를 분석하기 위해 데이터 전처리를 진행했다. 우선 feature가 과도하게 많을 시 머신러닝 과정에서 시간적 비용이 증가할 수 있고, overfitting의 가능성 역시 높아질 수 있다고 판단했다. 따라서 기준을 세워 일부 변수를 분석에서 제외하였다. CBCL 데이터에선 다른 feature의 값을 단순히 합산한, internal, external, total problem scale를 나타내는 feature를 제거하였다. 또한 CBCL 데이터는 각 항목에 대한 raw score, t-score, 결측값 수 및 결측값의 내용으로 구성되어 있었는데, 결측값의 수 및 내용을 나타내는 feature는 결측치가 과도하게 많아서 제외하였다. Raw score 역시 t-score와 제공하는 정보가 크게 다르지 않다고 판단하여 제외하였으며, 결과적으로 t-score만을 데이터 분석에 사용하였다.

dMRI data에선 평균 FA 값이 0.2 이하로 계산된 feature를 제외하였다. 이는 FA 값이 0.2 이하인 feature는, diffusion 모형이 해당 뇌 부위의 tissue structure를 잘 감지하지 못했음을 의미하기 때문이다. 실제로 dMRI를 활용하는 연구에서는 FA 값이 0.2를 넘지 못하는 영역을 분석에서 제외하기도 한다(Fernández-Espejo et al., 2010). 따라서 이러한 영역을 분석에 포함하면 머신러닝 모형의 성능에 악영향을 미칠 수 있다고 판단하였다. 이후 fMRI, sMRI, dMRI 데이터 각각에 대해 tree 기반 모형을 적용하여, feature importance를 분석하였다. fMRI 및 dMRI 데이터에는 XGBoost, sMRI 데이터에는 random forest를 적용했으며, 각 데이터에서 importance가 상위 5%에 해당하는 feature만을 분석에 사용하였다.

최종적으로, 연속형 feature를 대상으로 z-score normalisation을 적용하였고, 결측치는 다른 값들의 중앙값으로 대체하였다. 또한 인구통계 데이터의 education, married, race 등 범주형 feature의 결측치는 다른 값들의 최빈값으로 대체했고, one-hot encoding을 수행했다.

결과 및 논의

전처리 진행 이전 및 이후 각 데이터셋의 feature 수 차이는 아래 [표 1]과 같다. 전처리 결과, feature의 수가 4,679개에서 231개로 크게 감소하여, 과적합 등의 문제가 발생할 가능성이 낮아졌을 것으로 판단된다.

표 1. 전처리 이전 및 이후 데이터셋의 feature 수 차이

	feature 제외 기준	제외 이전	제외 이후
fMRI	XGBoost feature importance 상위 5% 이외 모두 제외	680	34
sMRI	Random Forest feature importance 상위 5% 이외 모두 제외	395	19
dMRI	(1) FA < 0.2인 feature 제외 (2) XGboost feature importance 상위 5% 이외 제외	3,486	115
CBCL	(1) 다른 항목의 합계로 계산된 feature 제외 (2) raw score, 결측치 수 및 내용 feature 제외	84	19
인구통계	(1) categorial variable에 대한 one-hot encoding 진행 (2) CBCL data와 중복된 feature인 sex 제외	8	18
polygenic scores	feature 수에 변동 없음	26	26
합계		4,679	231

실험 2. cross-validation을 통한 모델 선정 및 학습

방 법

[실험 1]에서 전처리된 데이터를 활용하여 자살 사고를 예측하기 위한 분류 모델을 선정하고 학습하였다. 분류 모델 후보군으로는 Logistic Regression + ElasticNet, Support Vector Machine, Random Forest, XGBoost를 선정하였다. 각 모델은 231개의 feature가 전부 포함된 데이터셋으로 학습되었다. 이 과정에서 각 모델에 대해 training 데이터를 대상으로 3-fold cross-validation을 수행하였으며, 각 fold에 대한 training 및 validation AUC, accuracy, recall의 평균값을 계산하였다. 최종적으로 validation AUC에서 제일 높은 성능을 보이는 모델을 선정하였다. 이 과정에서 grid search를 활용하여, 성능을 극대화하는 최적 hyperparameter 조합을 확인하였다.

또한, 231개보다 적은 feature를 사용할 때에도 model의 성능이 유지되는지 확인하기 위해, 선정된 최적의 model에 서로 다른 데이터셋을 학습시킨 뒤 성능을 비교하였다. 우선 CBCL feature로만 구성된 데이터셋으로 시작해서, 다른 데이터셋의 feature를 단계적으로 추가하며 학습된 모델들을 대상으로 3-fold cross-validation을 진행했다. 각 모델의 validation AUC, accuracy, recall을 비교하여, 데이터 추가에 따른 성능 향상 정도를 확인한 뒤, 가장 효율적인 feature 구성을 확인하였다.

최종적으로, 선정된 model에 대해 선택 feature로만 구성된 dataset을 학습시킨 뒤, test 데이터셋에서의 AUC, accuracy, recall을 평가하였다.

결과 및 논의

Grid search를 통해 최적의 hyperparameter가 선택된 4개 모델에 대한 cross-validation 결과는 [표 2]와 같다. Random Forest 모델의 validation AUC(0.975), accuracy(0.925), recall(0.880) 모두 다른 모델에 비해 우수한 성능을 보였다. 따라서 최종 모형으로 Random Forest를 선택하였다.

표 2. 각 모델의 Cross-validation 결과

	AUC		Accuracy		Recall	
	train	val	train	val	train	val
Logistic Regression + ElasticNet	0.972	0.970	0.905	0.845	0.815	0.700
Support Vector Machine	0.995	0.953	0.983	0.880	0.965	0.770
Random Forest	0.996	0.975	0.983	0.925	0.965	0.880
XGBoost	1.000	0.959	0.990	0.895	0.980	0.860

Feature의 수를 단계적으로 확장하여 학습된 6개의 Random Forest 모델에 대한 Cross-validation 결과는 [표 3]과 같다. CBCL data로만 학습했을 때도 validation AUC가 0.967으로 높은 편이며, 추가적인 데이터를 포함하더라도 성능이 크게 상승하지 않았다.

표 3. Dataset 구성에 따른 Random Forest 모형의 Cross-validation 결과

	Validation AUC	Validation Accuracy	Validation Recall
CBCL	0.967	0.910	0.889
+ 인구통계	0.964	0.920	0.900
+ polygenic scores	0.966	0.910	0.880
+ fMRI	0.971	0.915	0.880
+ sMRI	0.963	0.930	0.910
+ dMRI	0.961	0.915	0.880

CBCL 데이터만으로도 높은 성능을 보인다고 판단해, 최종 모델은 Random Forest에 CBCL data만을 학습시켜 구성하였다. 이때 test data에 대한 AUC score는 0.960 ([그림 1]),

accuracy는 0.950, recall은 0.950였다.

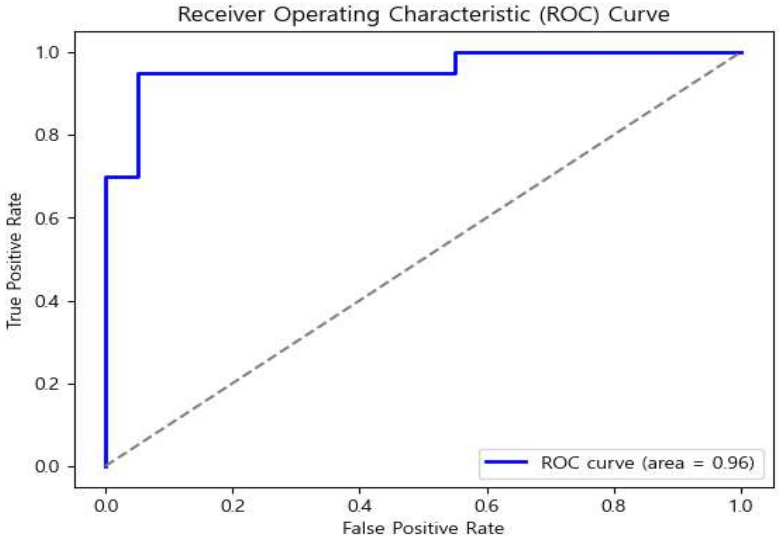


그림 1. CBCL data & Random Forest 모형에 대한 test data AUC Curve

실험 3. SHAP 값 분석을 통한 Feature Importance 및 예측 방향성 확인 방 법

자살 사고의 가능성을 높이는 원인을 파악하기 위해, [실험 2]에서 학습된 Random Forest 모형을 대상으로 feature importance 분석을 진행했다. 이를 통해 양성 및 음성 클래스를 구분하는 데 중요한 영향을 미친 feature를 파악하고자 하였다. 이때 분석에는 [실험 1]에서 전처리된 232개의 CBCL, polygenic scores, 인구통계, fMRI, sMRI, dMRI feature를 모두 포함한 training 데이터를 사용하였다.

Feature importance 분석을 위해, 각 sample 내 feature에 대해 SHAP (SHapley Additive exPlanations) 값을 계산했다. SHAP는 해당 feature가 sample의 label 예측에 기여한 정도를 설명하며, SHAP의 절댓값은 영향을 미친 강도를 나타낸다. SHAP가 양수인 경우 해당 feature가 양성 예측에, 음수인 경우 음성 예측에 기여했음을 뜻한다 (SHAP Documentation, n.d.).

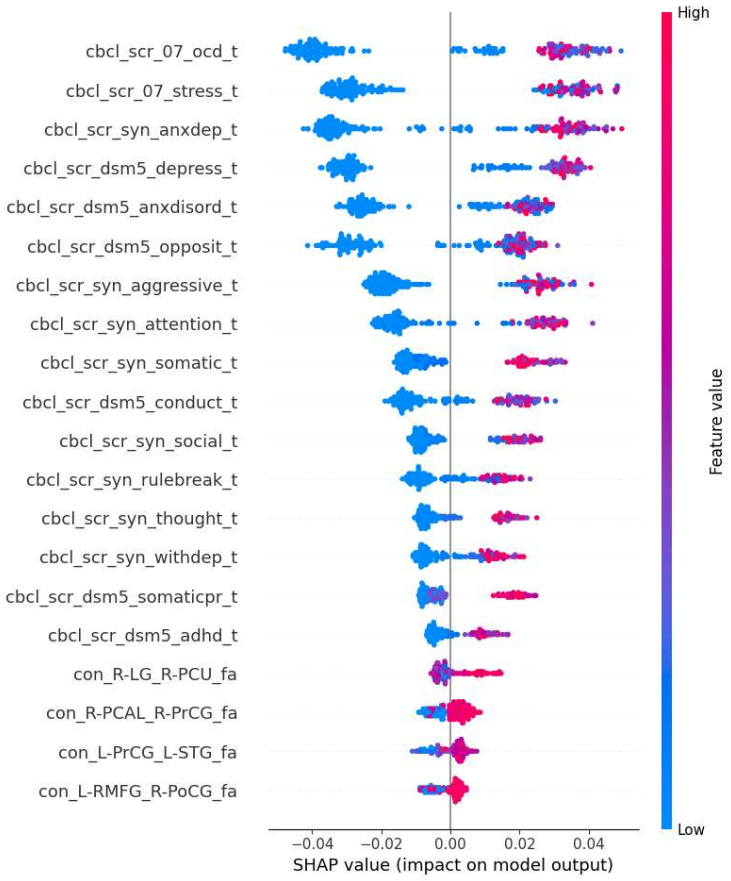
이때 SHAP는 sample별로 계산되기 때문에, 전체 데이터셋에서 특정 feature의 중요도를 파악하기 위해선 SHAP 절대값의 평균을 계산해야 한다. 이때 절대값의 평균이 클수록 보다 중요한 feature라고 판단할 수 있다. 본 분석에서는 232개의 feature 중 중요도가 상위 20개에 포함된 feature가 예측에 보다 결정적일 것이라고 기준을 세웠다.

이후 상위 20개 feature를 대상으로, 해당 feature의 값이 클수록 양성으로 예측될 확률이 높아지는지, 혹은 낮아지는지 파악하는 방향성 분석을 진행했다. Feature의 값이 커질수록 SHAP도 커지는 경우, 해당 feature 값이 증가함에 따라 양성으로 예측될 가능성이 높아짐을 의미한다. 반대로 feature의 값이 커질수록 SHAP가 작아지면, 해당 feature가 클수록 음성으로 예측될 가능성이 높아짐을 의미한다.

이러한 양상은 SHAP의 분포를 시각화한 SHAP summary plot을 통해 분석하였다. 또한 각 feature가 모형의 예측에 미치는 영향을 정밀하게 분석하기 위해, sample별로 feature의 값과 SHAP를 scatterplot 형태로 표시한 partial dependce plot(PDP) 역시 확인하였다. PDP는 독립변수를 제외한 변수는 전체 데이터의 평균값으로 대체하기 때문에, 기존 상관관계 분석에 비해 다른 변수의 영향력을 통제할 수 있다는 장점이 있다.

결과 및 논의

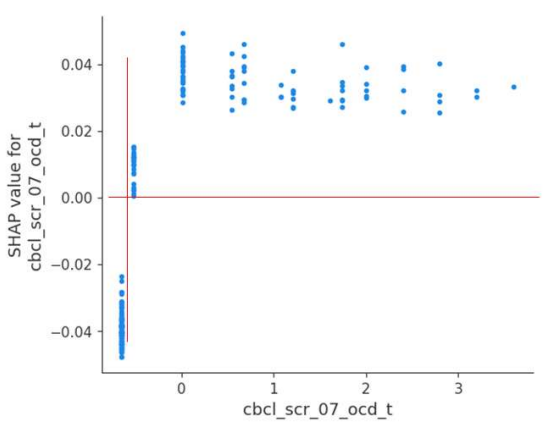
각 feature에 대한 SHAP summary plot은 [그림 2-3]와 같다. 상위 20개 feature를 확인했을 때, CBCL feature가 16개, dMRI feature가 4개로 구성되었다. CBCL feature 중에서도 OCD(0.033), stress(0.031), anxiety & depression(0.031)의 SHAP 절댓값 평균이 높았다. 상위 20개의 feature에서는 전반적으로 feature 값이 증가할수록 SHAP 값도 커지는 경향성이 나타났다. 이를 통해 상위 20개의 feature, 특히 CBCL feature는 값이 클수록 모형에서 양성 예측을 할 가능성이 높아진다고 해석할 수 있다.

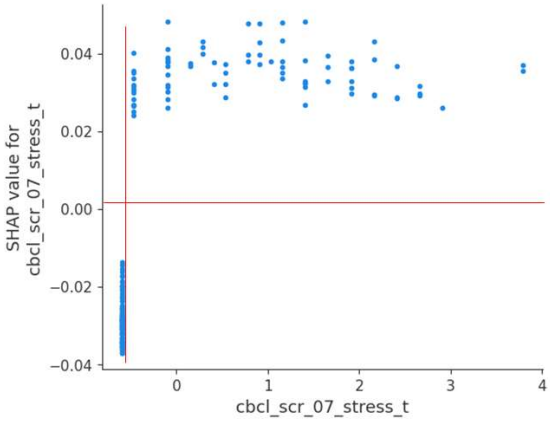


[그림 2] Training data에 대한 SHAP summary plot

이후 상위 20개의 feature에 대해 PDP을 확인하였다. 지면의 한계상 일부 중요 feature에 대한 분석 결과만 수행하였다. OCD나 stress 등 CBCL feature에 대한 PDP([그림 4-5])에서 점수가 상승할수록 SHAP 값이 커지는 경향이 관찰되었다. 그러나 이러한 관계는 선형

적이지 않으며, 점수가 특정 임계점(빨간 선으로 표시)을 초과할 때 SHAP 값이 음에서 양으로 변화하는 양상을 보였다.





[그림 4 (좌)] CBCL에서 측정된 obsessive-compulsive problems에 대한 partial dependence plot
[그림 5 (우)] CBCL에서 측정된 stress에 대한 partiial dependence plot

종합논의

MRI, polygenic 및 인구통계 feature를 제외하고 CBCL feature만을 사용하여 Random Forest 모형을 학습시킨 결과, test data 중 95%를 정확히 분류하였으며, AUC 역시 0.960으로 매우 높았다. 또한, SHAP value를 활용한 feature importance 분석에서도 CBCL feature가 다른 feature에 비해 예측에 더 큰 기여를 하는 것으로 나타났다.

이러한 결과는 아동, 청소년의 자살 사고와 가장 밀접한 변인은, CBCL에서 측정되는 각종 증후군에 대한 위험 수준임을 시사한다. 따라서, CBCL에서 확인된 위험 수준을 낮출 수 있는 대책을 세우는 것이 아동, 청소년 자살 예방에 필수적이라는 결론을 도출할 수 있다.

참고문헌

Fernández-Espejo, D., Junque, C., Cruse, D., et al. (2010). Combination of diffusion tensor and functional magnetic resonance imaging during recovery from the vegetative state. *BMC Neurology, 10*, 77. https://doi.org/10.1186/1471-2377-10-77

Joo, Y. Y., Moon, S. Y., Wang, H. H., Kim, H., Lee, E. J., Kim, J. H., ... & Cha, J. (2022). Association of genome-wide polygenic scores for multiple psychiatric and common traits in preadolescent youths at risk of suicide. JAMA network open, 5(2), e2148585-e2148585.

SHAP Documentation. (n.d.). *An introduction to explainable AI with Shapley values*. SHAP. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html