설문지를 활용한 청소년 비행 예측

201611533 통계학과 조근수
esg114@naver.com

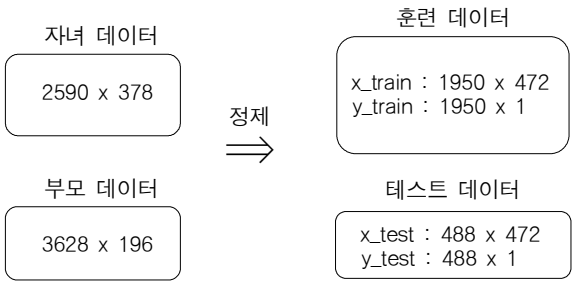
분석 배경 및 목적

최근 연구들에서, 청소년 비행 문제는 단순히 그 양적으로 숫자만 증가하는 것이 아니라 향후 성인범죄로 이어질 수 있는 많은 가능성을 설명하고 있다. 청소년 비행 문제가 사회적 문제로 확대됨에 따라 이를 해결하기 위한 여러 정책들이 제시되어 왔다. 하지만 현재 진행되고 있는 대부분의 청소년 비행 방지를 위한 프로그램은 비행 청소년(이미 비행을 저지른 청소년)을 대상으로 한, 사후약방문적인 측면이 강하다.

따라서 이 분석을 통해, 청소년 비행 유무에 영향을 미치는 **문항들을 선별**하고, 이 선별된 문항들로 설문지를 재구성하여 **청소년 비행 유무를 예측**할 수 있는지 파악하고자 한다. 마지막으로 청소년 비행 방지를 위한 **향후 분석 방향을 제시**하고자 한다.

데이터 구성

NYPI(한국 청소년 정책 연구원)에서 제공하는 2018 한국 아동,청소년 패널조사의 제2차조사(2019년) 데이터[1]로, 자녀 데이터와 보호자 데이터 총2개의 데이터를 사용하였다.



- 자녀 데이터와 부모 데이터를 개인ID, 가구ID를 키(key)로써 병합, 조사에 응하지 않은 데이터는 제거
- 한 자녀에 대응되는 보호자 데이터가 2개인 경우(어머니, 아버지 모두 조사에 참여한 경우) 자녀 데이터를 복사하여 2개의 데이터로 취급
- 결측치는 해당 문항(feature)의 중앙값으로 대체하되, 결측치가 10%넘는 문항은 제거
- **비행관련 문항(흡연여부, 패싸움 등 15개) 중 한 번이라도 경험이 있으면, 반응변수(y)는 1 Positive(+), 모두 경험이 없으면 반응변수(y)는 0 Negative(−)**

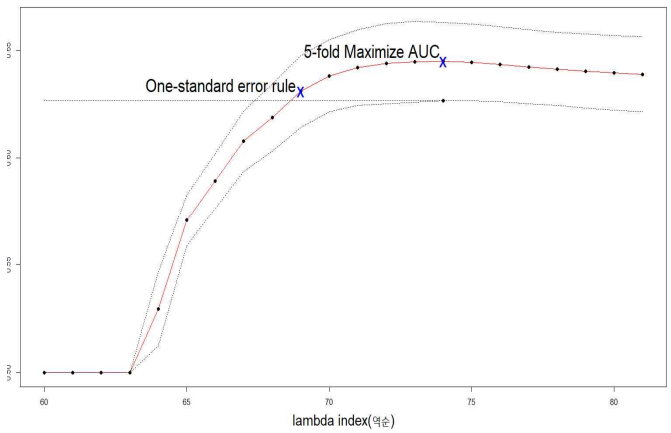
분석 방법 및 과정[2]

분석에는 총 2가지의 알고리즘이 사용되었다. 먼저 문항(특성)들을 선별하기 위해, **LASSO를 활용한 Shrinkage method**를 적용하였다. 다음으로 선별된 문항들로 재구성된 데이터에 **RandomForest를 적용해 비행 유무를 예측**하는 모델을 만들었다. 두 알고리즘 모두 **평가지표로써 AUC를 최대화**하는 모델을 선택하였으며, **5-fold CV**를 사용하였다.

A. LASSO를 활용한 특성 선택 (R-programming)

LASSO는 덜 중요한 특성의 가중치를 제거하려고 하는 특징을 가진다. 이 특징을 이용해 비행유무에 영향을 미치지 못하는 특성을 제거한다. 특히 이 과정에서 **one-standard error rule을 적용**하여, 규제를 증가시킨다.(즉, 더 적은 특성을 선택하기 위해 one-standard error 만큼의 오차는 인정한다.)

<생활변수의 Hyper parameter 튜닝 과정 시각화>



데이터의 문항은 총 12개의 대영역으로 이루어져 있다. 위의 그래프는 12개의 대영역 중 하나인, 생활시간 변수(32개의 문항)를 LASSO와 One-standard error rule을 적용해 변수 선택과정을 시각적으로 나타낸 것이다.

12개의 대영역에 대해 모두 Shrinkage method를 적용하였으며, 각 대영역 별로 1개의 규제 파라미터(Hyper parameter)를 설정하였다. 이 때, 적절한 하이퍼파라미터를 튜닝하기 위해 **Grid Search**를 사용하여, 모델에 적합한 하이퍼파라미터를 구하였다.

총 67개 문항이 LASSO를 통해 선택되었다. 65개 문항은 자녀 데이터로부터 온 문항이였고, 2개의 문항만이 부모 데이터로부터 온 문항(건강 관련, 사회 정서 관련 문항)이였다. 반응 변수(비행유무)와 상관관계가 높은 상위 5개 문항은 다음과 같다.

문항	상관계수
누군가에게 욕설을 직접 보낸 적이 있다.	0.370
일부러 시비를 걸어 상대방의 성격이 문제가 있어보이게 유도한 적이 있다.	0.274
남이 하는 일을 방해할 때가 있다.	0.183
별 것 아닌 일로 싸우곤 한다.	0.158
입맛이 없을 때가 있다.	0.154

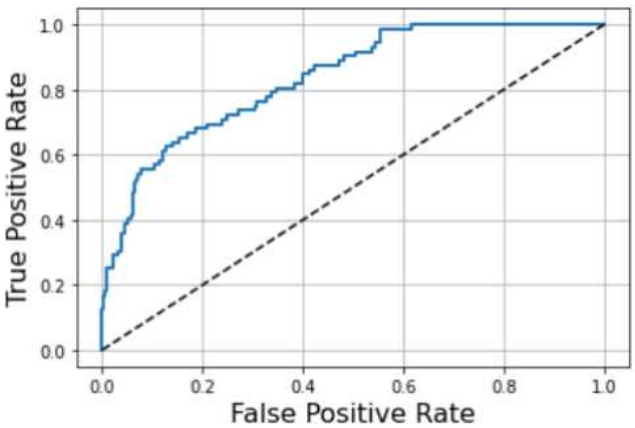
프로젝트엔 생략되었지만 상,하위 10개 문항을 제외하고 대체로 0.8~0.15의 상관계수 값을 가졌다.

B. RandomForest를 활용한 비행 예측(Google Colab)

앞서 LASSO를 통해 선택된 67개 문항을 사용하여, 비행유무를 예측하는 RandomForest 모델을 만들었다. 최적의 예측 모델을 만들기위해, 총 6개의 하이퍼파리미터를 튜닝 하였다. 이 때, 사용된 매개변수는 'bootstrap', 'max_depth', 'max_features', 'min_samples_leaf', 'min_samples_split', 'n_estimators'이다. 이 때, 매개변수를 효율적으로 조절하기 위해 **Randomized Search**를 사용하여, 최적의 하이퍼파라미터를 구했다. 최적 모델의 매개변수는 다음과 같다.

Hyperparameter	values
bootstrap	True, False
max_depth	10, 20, 30, 50, 70, None
max_features	'auto', 'sqrt'
min_samples_leaf	1, 2, 4, 6
min_samples_split	2, 4, 6
n_estimators	100, 150, 200, 250, 300, 400, 500, 600, 800

최종적으로 만들어진 모델에 test set를 적합하여 청소년 비행 유무를 잘 예측하는지 확인한다. ROC curve와 AUC는 다음과 같다.



AUC : 0.8355

결론 및 회고

NYPI에서 제공하는 청소년 패널조사 데이터를 활용해, 비행유무에 영향을 미치는 문항들을 선택하였다. 그 후 선택된 문항들로 비행 유무를 예측하는 모델을 만들었다. AUC=0.84로 상대적으로 적은 샘플 수(대략 2500)로도 청소년 비행을 예측하는데 꽤 괜찮은 성능을 보였다.

원 데이터의 경우 설문지 구성 및 표본 수집과정[3]이 복잡하여 비용이 많이 들 수밖에 없다. 이 분석을 통해, 선택된 새로운 문항들로 **청소년 비행 방지를 위한 설문지를 재구성**함으로써, 설문지 구성 및 표본 수집과정에서 드는 **비용을 절감**할 수 있을 것으로 기대된다.

또한 축약된 설문지를 사용함으로써, 더 많은 표본을 수집하는데 용이할 것으로 기대된다. 이는 예측 모델에 필요한 충분한 데이터의 양을 확보할 수 있게 한다. 따라서 **더 많은 표본을 확보**한다면, 현재 분석에서 진행된 결과보다 더 **성능이 좋은 청소년 비행 예측 모델**을 만들 수 있을 것으로 기대된다.

1-1) 데이터 출처 :
https://www.nypi.re.kr/archive/mps/program/examinDataCode/view?menuId=MENU00226&pageNum=1&titleId=24&schType=0&schText=&firstCategory=%ED%8C%A8%EB%84%90%EC%A1%B0%EC%82%AC&secondCategory=%ED%95%9C%EA%B5%AD%EC%95%84%EB%8F%99%C2%B7%EC%B2%AD%EC%86%8C%EB%85%84%ED%8C%A8%EB%84%90%EC%A1%B0%EC%82%AC%202018

1-2) 데이터 코드북 :
https://www.nypi.re.kr/archive/mps/program/examinDataCode/view?menuId=MENU00226&pageNum=1&titleId=22&schType=0&schText=&firstCategory=%ED%8C%A8%EB%84%90%EC%A1%B0%EC%82%AC&secondCategory=%ED%95%9C%EA%B5%AD%EC%95%84%EB%8F%99%C2%B7%EC%B2%AD%EC%86%8C%EB%85%84%ED%8C%A8%EB%84%90%EC%A1%B0%EC%82%AC%202018

2) https://github.com/geunsu-jo/KCYPS_project

3) https://www.nypi.re.kr/archive/board?menuId=MENU00220