DM Pre-Processing

짱짱조

[RQ1] 중1 학생들의 수면 시간에 큰 영향을 미치는 요소에는 무엇이 있을까?

1. Data Set

한국아동청소년패널조사(KCYPS) 2018의 중1 코호트 2차 조사(2019년) 데이터를 활용하였다. 본 조사는 중학교 1학년 학생 2438명을 대상으로 진행되었으며, 학습시간, 여가시간, 신체증상, 행복감, 친구관계, 현실비행 경험 유무 및 빈도 등과 관련된 문항 총 378개로 구성되어 있다.

	HID	PID	SCLIDw2	WEIGHTA1w2	WEIGHTA2w2	WEIGHTB1w2	WEIGHTB2w2	SURVEY1w2	SURVEY2w2	COHO
0	780	2	20409	214.693708609272	1.260686727224	215.721347394929	1.2667210473997	1	1	
1	1192	2	20912	87.6273251058101	.514549804084321	87.9315523040103	.516336233660278	1	1	
2	1193	2	40920	105.83496049671	.621465486058106	105.583650039588	.619989784884755	1	1	
3	1285	2	20920	281.283010983134	1.65170074538796	280.220791250769	1.64546336504429	1	1	
4	1590	2	20936	83.6442648737278	.491161176625942	83.1256287182711	.488115732389695	1	1	
...	...	...	...	...	...	...	...	...	...	
2585	5157	1	21703	260.64	1.53048447815332	244.89075549328	1.43800452779967	1	1	
2586	5158	1	21703	260.64	1.53048447815332	226.303637303806	1.32886051351716	1	1	
2587	5159	1	21703	260.64	1.53048447815332	244.89075549328	1.43800452779967	1	1	
2588	5160	1	21703	260.64	1.53048447815332	244.89075549328	1.43800452779967	1	1	
2589	5161	1						2	2	

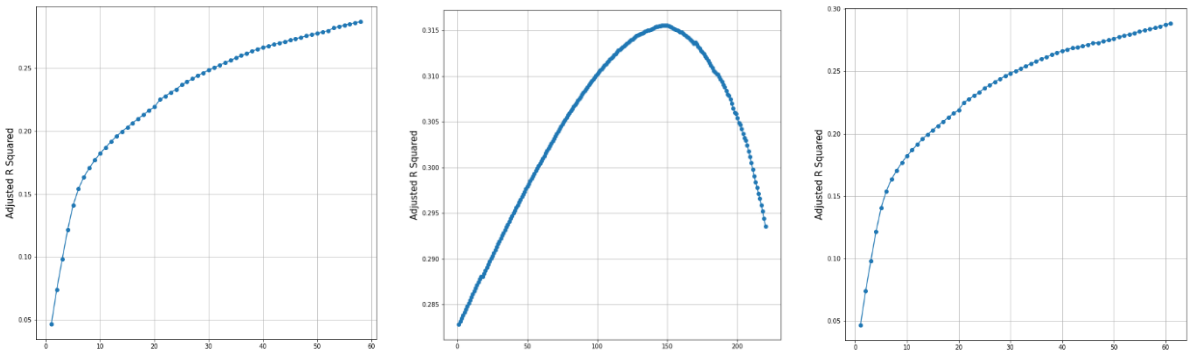
2590 rows × 378 columns

2. Preprocessing X and Y

연구 질문에 따른 종속 변수(Y)를 설정하기 위하여 수면시간과 관련된 문항 8개를 바탕으로 일평균 수면시간(분)을 산출하였다. 독립 변수(X)는 총 378개의 문항 중, 종속 변수로 선택한 8개의 문항, 학교코드, 생년 등 연구 질문과 관련 없는 17개 문항, 그리고 응답에 결측치가 포함된 71개 문항을 제외하고 남은 282개 문항으로 설정하였다. 문항에 대한 응답은 객관식인 경우 리커르트 척도(1: 전혀 그렇지 않다, 4: 매우 그렇다)로 측정되었고, 주관식의 경우 범주가 없는 개방형 숫자로 측정되었다.

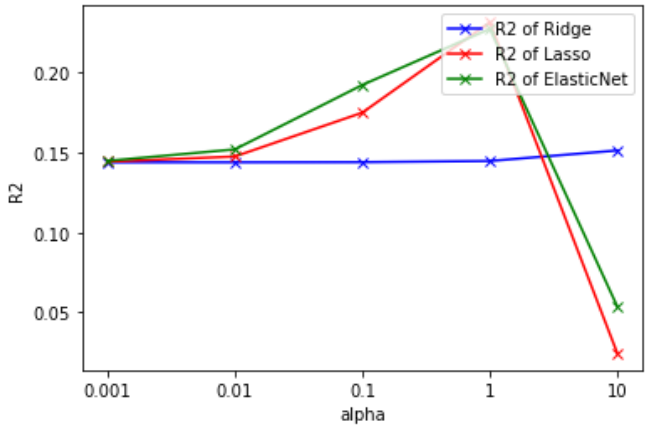
3. Result

1) Forward Selection, Backward Selection, Stepwise Selection 평가



위의 그래프는 Forward selection, Backward Selection, Stepwise Selection을 수행했을 때 step별 r square의 값의 변화를 나타낸다. 각 selection 수행 중 성능이 가장 좋은 모델의 R square값은 각각 0.287, 0.316, 0.288으로 backward selection model이 가장 좋은 성능을 보였다. 하지만 R square 같은 경우는 Adjusted R square라고 하더라도 변수의 수에 어느정도 영향을 받는다. 이를 감안했을 때, 변수 선택적인 측면에서 가장 좋은 결과를 보여주는 변수 축소 방식은 stepwise라고 할 수 있다.

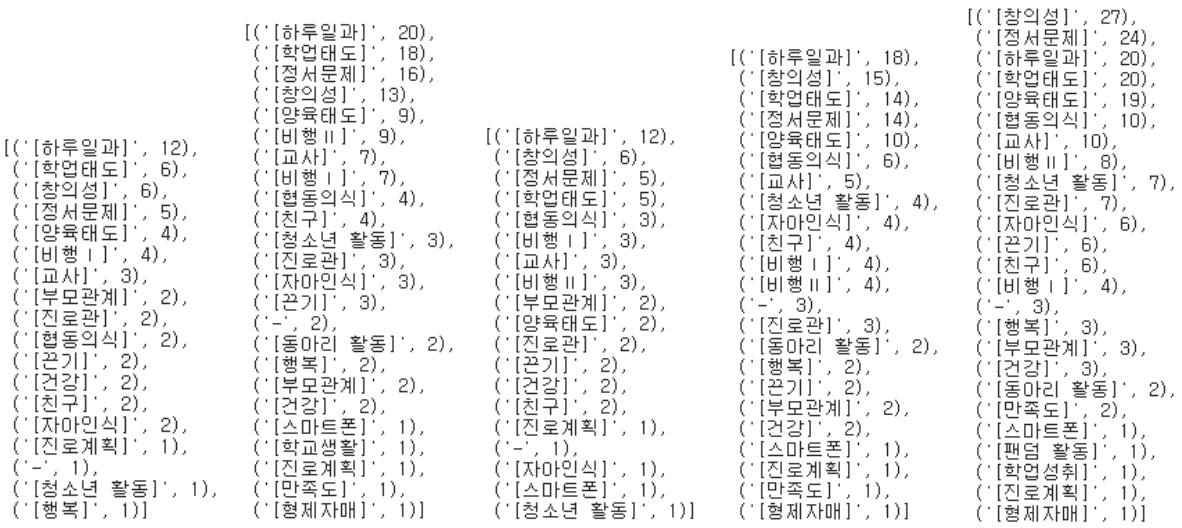
2) Ridge, Lasso, ElasticNet 평가



최적의 성능을 보여주는 모델에서 Ridge, Lasso, ElasticNet은 각각 0.151, 0.232, 0.227의 square값을 가진다. 최적의 모델에서 lasso의 경우 122개의 변수를 ElasticNet의 경우 195개의 변수를 선택했다.

위 그래프에서도 알 수 있듯이 penalty parameter가 너무 커지는 경우를 제외하고는 Lasso와 Elastic Net이 Ridge보다 좋은 성능을 보여주고 있다. Lasso와 Elastic Net은 그래프 상에서는 Elastic Net이 근소하게 좋은 성능을 보여주고 있지만, 최고의 성능을 보여주는 지표에서의 R squared 값은 Lasso가 앞서고 있다. 그렇기에 앞으로 해당 기법을 활용해 변수를 축소선택하면양 모델을 모두 설계한 후 그 결과를 비교하는 형식으로 진행해야 할 것 같다

4. Result Analysis



위의 사진들은 Ridge를 제외한 각 모델들이 채택한 변수들을 카테고리별로 분류한 후 정렬해본 결과이다. 공통적으로 하루일과, 창의성, 학업태도, 정서문제 4가지 종류의 변수들이 순서는 조금씩 다르더라도 상위권에 위치했다. 또한 결정계수가 0.2 ~ 0.3 사이로 생각한 것보다는 낮게 나와서 모델의 문제가 있는 줄 알았으나, Cohen[1]에 의하면 사회과학의 경우 0.3 이상의 결정계수를 추천한다고 한다. 특히, 데이터의 크기에 따라 중간 크기의 데이터는 0.13, 큰 크기의 데이터는 0.3을 기준으로 상관관계가 어느정도 효과가 있음을 시사하고 있다. 이러한 점과 각 모델들이 선택한 변수들의 유형이 어느정도 일치한다는 점으로 보아 본 연구는 중학생의 수면시간에 영향을 끼치는 요소들을 유효하게 식별해냈다고 판단된다.

[1] Cohen, J.(1988), Statistical Power Analysis for the Behavioral Sciences(2nd Ed.), Lawrence Erlbaum Associates, Inc.

[RQ2] 따릉이 대여권 유형(일일권, 정기권 등)에 영향을 미치는 요소에는 무엇이 있을까?

1. Data Set

서울 열린 데이터 광장의 '2021년 공공자전거 일별 이용정보' 데이터와 '공공자전거 대여소 정보' 데이터를 활용하였다. '공공자전거 일별 이용정보'는 대여일자, 대여소번호, 대여소명, 이용자의 정기권유무, 성별, 연령대, 이용건수, 이동거리, 이용시간이 담겨있는 474,631건의 대여기록으로 구성되어 있다. '공공자전거 대여소 정보'는 대여소 번호, 대여소명, 운영방식 등 총 10개의 컬럼으로 구성되어 있다. 두 데이터를 동시에 활용하기 위해 각 데이터의 '대여소 번호' 컬럼을 기준으로 두 테이블을 조인해 분석을 진행했다.

	index	대여일자	대여소번호	대여소	대여구분코드	성별	연령대코드	이용건수	운동량	탄소량	이동거리(M)	이용시간(분)	대여소명	자치구	상세주소	위도	경도	설치시기	거치대수(LCD)	거치대수(QR)	운영방식
5	5	2021-01-01	101	101. (구)합정동 주민센터	정기	M	AGE_003	1	36.23	0.29	1270.63	123	(구)합정동 주민센터	마포구	서울특별시 마포구 동교로8길 58	37.549561	126.905754	2015-09-06	5.0	NaN	LCD
7	14775	2021-01-02	101	101. (구)합정동 주민센터	일일(회원)	₩N	AGE_003	1	32.58	0.38	1645.60	13	(구)합정동 주민센터	마포구	서울특별시 마포구 동교로8길 58	37.549561	126.905754	2015-09-06	5.0	NaN	LCD
8	14776	2021-01-02	101	101. (구)합정동 주민센터	정기	₩N	AGE_002	3	38.36	0.39	1699.35	72	(구)합정동 주민센터	마포구	서울특별시 마포구 동교로8길 58	37.549561	126.905754	2015-09-06	5.0	NaN	LCD
10	14778	2021-01-02	101	101. (구)합정동 주민센터	정기	₩N	AGE_005	3	16.40	0.15	637.23	114	(구)합정동 주민센터	마포구	서울특별시 마포구 동교로8길 58	37.549561	126.905754	2015-09-06	5.0	NaN	LCD
11	14779	2021-01-02	101	101. (구)합정동 주민센터	정기	F	AGE_003	1	23.46	0.21	911.44	11	(구)합정동 주민센터	마포구	서울특별시 마포구 동교로8길 58	37.549561	126.905754	2015-09-06	5.0	NaN	LCD
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
475073	466253	2021-01-31	1627	1627. 수락산역 4번출구	일일(회원)	M	AGE_002	1	154.46	1.13	4875.63	31	수락산역 4번출구	노원구	서울특별시 노원구 동일로 지하 1662	37.676941	127.055099	2020-12-31	NaN	12.0	QR
475075	466255	2021-01-31	1627	1627. 수락산역 4번출구	정기	F	AGE_002	1	72.19	0.70	3038.28	32	수락산역 4번출구	노원구	서울특별시 노원구 동일로 지하 1662	37.676941	127.055099	2020-12-31	NaN	12.0	QR
475076	466256	2021-01-31	1627	1627. 수락산역 4번출구	정기	F	AGE_004	1	100.47	1.01	4374.36	23	수락산역 4번출구	노원구	서울특별시 노원구 동일로 지하 1662	37.676941	127.055099	2020-12-31	NaN	12.0	QR
475077	466257	2021-01-31	1627	1627. 수락산역 4번출구	정기	F	AGE_005	1	74.00	0.62	2669.48	36	수락산역 4번출구	노원구	서울특별시 노원구 동일로 지하 1662	37.676941	127.055099	2020-12-31	NaN	12.0	QR
475078	466258	2021-01-31	1627	1627. 수락산역 4번출구	정기	M	AGE_005	1	94.51	0.64	2743.29	22	수락산역 4번출구	노원구	서울특별시 노원구 동일로 지하 1662	37.676941	127.055099	2020-12-31	NaN	12.0	QR

435404 rows × 21 columns

2. Preprocessing X and Y

이번 분석에서는 총 다섯개의 범주형 변수와 target 변수와의 관계를 확인했다.

1) 대여권 유형(Y) – 연령대, 성별, 대여소 운영방식(X)의 관계

연령대코드	AGE_001	AGE_002	AGE_003	AGE_004	AGE_005	AGE_006	AGE_007	AGE_008
대여구분코드								
단체	588.0	725.0	305.0	1030.0	175.0	17.0	2.0	42.0
일일(비회원)	0.0	1.0	4.0	0.0	0.0	0.0	0.0	2200.0
일일(회원)	9023.0	43222.0	21870.0	11293.0	4327.0	852.0	167.0	1464.0
정기	11416.0	98799.0	81926.0	66735.0	50795.0	19829.0	3107.0	4951.0

대여권 유형 – 연령대 Dataframe

성별	F	M
대여구분코드		
단체	678	835
일일(비회원)	32	19
일일(회원)	22335	28707
정기	67957	121254

대여권 유형 – 성별 Dataframe

운영방식	LCD	QR
대여구분코드		
단체	2236	648
일일(비회원)	1704	501
일일(회원)	69930	22288
정기	256618	80940

대여권 유형 – 운영방식 Dataframe

'연령대코드' 컬럼, '성별' 컬럼, '대여소 운영방식' 컬럼의 데이터는 범주형으로 별도의 전처리 없이 분석이 가능하다. 카이제곱검정을 위해 대여권-연령대 분석에서는 대여권 유형이 담긴 '대여구분코드'와 '연령대코드'를 각각 축으로 하고, 대여권-성별 분석에서는 '대여권구분코드'와 '성별'을 각각 축으로 해 Dataframe으로 재구성하였다.

2) 대여권 유형(Y) – 이동거리, 이동시간(X)의 관계

Distance_cut	~1km	~2km	~3km	~4km	~5km	~6km	~7km	~8km	~9km	~10km	10km~
대여구분코드											
단체	70.0	187.0	215.0	209.0	210.0	203.0	192.0	154.0	156.0	136.0	1152.0
일일(비회원)	138.0	263.0	300.0	232.0	184.0	146.0	142.0	117.0	83.0	66.0	534.0
일일(회원)	9645.0	16901.0	12621.0	9314.0	6994.0	5412.0	4455.0	3668.0	3066.0	2490.0	17652.0
정기	60173.0	74800.0	49007.0	33358.0	24505.0	17693.0	13446.0	10948.0	8571.0	6909.0	38148.0

Time_cut	~6분	~9분	~13분	~19분	~26분	~36분	~49분	~68분	~105분	105분~
대여구분코드										
단체	19	19	54	79	123	185	253	405	672	1075
일일(비회원)	61	57	113	149	184	245	293	350	346	407
일일(회원)	5960	5868	7552	9134	8391	9606	10529	11320	10733	13125
정기	45740	30971	34127	38819	33111	33222	31203	30191	31710	28464

이동거리와 이동시간 변수는 모두 연속형 변수로 전처리가 필요하다. 먼저, 이동거리는 1km

단위로 범주화시켜 총 11개의 범주로 구분했다. 이동시간은 범주에 속하는 행의 수가 일정하도록 10개의 범주로 구분해 범주형 변수로 변환시켰다.
마찬가지로 카이제곱검정을 위해 대여권-이동거리 분석에서는 '대여구분코드'와 '이동거리_범주'를 각각 축으로 하고, 대여권-이동거리 분석에서는 '대여권구분코드'와 '이동시간_범주'을 각각 축으로 해 Dataframe으로 재구성하였다.

3. Result

본 분석에서는 Y와 X가 연관성이 없다(독립적이다)는 것을 귀무가설(H0)으로 두고 Y와 X가 연관성이 있다(독립적이지 않다)는 것을 대립가설로 두고 이를 카이제곱검정을 수행하고자 한다.

Y변수가 '대여이용권'이고 X변수가 '연령대', '이동거리', '이동시간', '성별', '운영방식'일 때 카이제곱통계량은 각각 136653.531, 10033.340, 12718.605, 1100.478, 7.212이고 pvalue는 0, 0, 0, 0, 0.065가 나왔다. 이 중 '대여소 운영방식'을 제외하고 pvalue가 유의수준인 0.05보다 작으므로 귀무가설이 기각된다. 다시 말해, 대여 이용권과 연령대, 이동거리, 이동시간, 성별은 모두 연관성을 지닌다. 이에 반해, '대여소 운영방식'은 pvalue가 0.065로 유의수준보다 크므로 귀무가설이 검 큰 X변수는 '대여소 운영방식'일 뿐으로 이므로 귀무가설이 진실임을 알 수 있다. 따라서 '대여 이용권'과 '대여소 운영방식'은 독립적이다.