

Project_Section 2

행복한 서울시민을 만들기 위한 정책결정 방향 설정 인사이트

1. 데이터 선정 이유 및 문제 정의

1. 데이터 선정 이유 및 문제 정의

- 근거 기반 정책 수립을 위해 2003년부터 매년 시행되는 서울 서베이 데이터. -> 2020년 데이터
 - 데이터 출처 : <https://data.si.re.kr/node/65072>
- 데이터 분석 목적 : 행복한 서울시민을 위한 정책결정 방향 설정 인사이트
- 데이터 : 40,085명의 각 설문 문항에 대한 응답

1. 데이터 선정 이유 및 문제 정의

- 조사 내용 및 설문지 문항

1. 환경 (4 문항)

2. 주거와 생활 (2 문항)

3. 안전 (1문항)

4. 교통 (3문항)

5. 문화와 여가 (5문항)

6. 건강 (2문항)

7. 사회참여 (2문항)

8. 노후생활 (3문항)

9. 사회 통합 (9문항)

10. 그 외 데모그래픽 정보 (혼인, 나이, 종교, 국적, 장애인여부, 주거형태 등)

1. 데이터 선정 이유 및 문제 정의

• 조사 내용 및 설문지 문항

가구원

- 에너지 절약 시민참여

- 환경보전을 위한 과제

- 미세먼지 저감조치 참여 의향

- 녹지환경 만족도

- 주변환경 만족도

- 물품 구매 경로

- 항목별 도시위험(안전)도

- 교통수단 이용 만족도

- 보행환경 만족도

- 통근/통학 여부

- 문화환경 만족도

- 문화예술 및 스포츠 관람률과 관람비용

- 현재 여가 활용 유형

- 여가활동 동반인

- 여가활동 만족도
- 일상생활 스트레스

- 하루 평균 취침시간

- 자원봉사 참여 여부

- 자원봉사 유형

- 지난 1년간 기부 여부

- 노후생활자금 준비 여부

- 은퇴 후 적정 생활비

- 예상은퇴시기

- 어려울 때 도움 받을 수 있는 사람

- 사람/기관 유형별 신뢰


- 계층이동가능성

- 사회적 차별

- 사회적 약자에 대한 태도

- 가족과 함께 보내는 시간

- 행복지수



승인번호
제 201011 호

2020 서울서베이 (도시정책지표 조사표 - 가구원용) I·SEOUL·U
너와 나의 서울

조사원이 작성하는 란입니다.	일련번호 □□□□□	행정구역코드 □□□-□□□	가구원 코드 □□	출생년도	성명
조사주관	서울특별시 빅데이터담당관		조사기관	(주)케이스탯리서치 02-6188-6011	

■ 환경

문1. 귀하는 아래의 사항들을 얼마나 실천하고 계십니까?

전혀 실천하지 않음	실천하지 않음	보통	다소 실천	항상 실천
1	2	3	4	5

1) 승용차 대신 도보, 자전거 또는 대중교통 이용

2) 재활용, 새활용, 친환경 제품 구매 노력

3) 일회용품 사용하지 않기

4) 배출 요령에 따라 배출

- 비닐류: 깨끗한 상태로 투명비닐에 담아 배출
- 상자류: 테이프, 운송장, 상표 등 제거 후 배출
- 용기류: 내용물 비우고 씻어 배출

문2. 환경보전을 위해 다음 과제들이 얼마나 필요하다고 생각하십니까? 각 항목에 대해 말씀해 주십시오.

불필요	조금 불필요	보통	조금 필요	매우 필요
1	2	3	4	5

1) 시내 녹지공간 확충 및 공원건설

2) 미세먼지 저감

3) 방음벽/방음시설 확충

■ 주거와 생활

문5. 귀하가 살고 있는 동네에 대한 질문입니다. 다음 항목에 대해 어느 정도 동의하십니까?

전혀 동의하지 않음	별로 동의하지 않음	보통	다소 동의	매우 동의
1	2	3	4	5

1) 우리 동네는 달리기나 걷기 같은 운동을 하기 적합하다

2) 우리 동네에는 공공시설(주민자치센터, 도서관, 공원 등)이 충분히 있다

3) 우리 동네는 안전하다

4) 우리 동네 사람들은 내가 도움이 필요할 때 기꺼이 도와주려 한다

문6. 지난 1년간 귀하는 상품을 주로 어떻게 구입하셨습니까?
방문구매 하셨다면, 물품 종류별로 주로 구매한 장소는 어디입니까?

	구매 유형	방문점포 유형 [보기 1]
1) 생활용품 및 식료품	① 구매안함 ② 통신구매 ③ 방문구매 13%	
2) 의류 및 잡화	① 구매안함 ② 통신구매 ③ 방문구매 13%	
3) 내구재 (가구, 가전 등)	① 구매안함 ② 통신구매 ③ 방문구매 13%	

[보기 1]

① 전통시장
③ 기업형슈퍼마켓
⑤ 백화점

② 동네슈퍼
④ 대형할인매장/아웃렛
⑥ 저가점

1. 데이터 선정 이유 및 문제 정의

- 조사 내용 및 설문지 문항

문1. 귀하는 아래의 사항들을 얼마나 실천하고 계십니까?

전혀 실천하지 않음	실천하지 않음	보통	다소 실천	항상 실천
1-----	2-----	3-----	4-----	5-----
1) 승용차 대신 도보, 자전거 또는 대중교통 이용				
2) 재활용, 새활용, 친환경 제품 구매 노력				
3) 일회용품 사용하지 않기				
4) 배출 요령에 따라 배출 - 비닐류: 깨끗한 상태로 투명비닐에 담아 배출 - 상자류: 테이프, 운송장, 상표 등 제거 후 배출 - 용기류: 내용물 비우고 씻어 배출				

• 1) ~4) 의 응답 평균

문6. 지난 1년간 귀하는 상품을 주로 어떻게 구입하셨습니다?
방문구매 하셨다면, 물품 종류별로 주로 구매한 장소는 어디입니까?

	구매 유형	방문점포 유형 [보기 1]
1) 생활용품 및 식료품	① 구매안함 ② 통신구매 ③ 방문구매	
2) 의류 및 잡화	① 구매안함 ② 통신구매 ③ 방문구매	
3) 내구재 (가구, 가전 등)	① 구매안함 ② 통신구매 ③ 방문구매	

[보기 1]

① 전통시장	② 동네슈퍼
③ 기업형슈퍼마켓	④ 대형할인매장/아웃렛
⑤ 백화점	⑥ 전문점
⑦ 편의점	⑧ 기타(구체적으로: _____)

•구매안함, 통신 구매 + 보기 내용 통합

문12. 귀하는 지난 1년 동안 다음과 같은 곳에 몇 번이나 가보셨습니까? 이때 비용은 어느 정도 들었습니까? (본인이 직접 지불한 금액을 기준으로 적어주세요. 없으면 0회 표시) (비용은 음식비 및 교통비는 제외하며, 무료로 관람(입장)한 경우 방문횟수에는 포함되며 비용은 제외됨.)

항목	연간 방문횟수	연간 총 비용
1) 전통예술공연 관람(국악, 민속놀이 등)	회	원
2) 음악 및 무용 발표회 관람 (클래식, 오페라, 발레 등)	회	원
3) 연극공연 관람(뮤지컬 포함)	회	원
4) 극장에서 영화 관람	회	원
5) 전시회 관람 (미술, 사진, 건축, 디자인 등)	회	원
6) 박물관 관람	회	원
7) 대중공연 관람(쇼, 콘서트, 마술 쇼 등)	회	원
8) 운동 경기 관람(e스포츠 관람 포함)	회	원

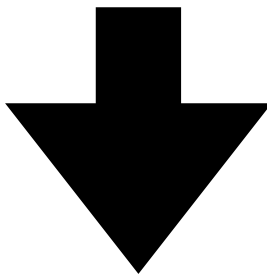
•연간 횟수, 총 비용 합산

1. 데이터 선정 이유 및 문제 정의

	date	ID	GU	NUM_1	FAM15	HOUSE	LIFE	SQ1_1	SQ1_2	SQ1_3	SQ1_4	SQ1_5	SQ1_6	SQ1_7	SQ1_8A	AQ1A1	AQ1A2	AQ1A3	AQ1A4
0	20210326	2	110	4	4	1	1	1	1	1960	1	2	1	2	1	4	3	3	
1	20210326	2	110	4	4	1	1	2	2	1962	1	2	1	2	1	4	2	5	
2	20210326	2	110	4	4	1	1	3	2	1995	2	2	1	2	1	3	4	3	
3	20210326	3	110	5	4	1	3	1	1	1983	1	1	1	2	1	4	3	4	
4	20210326	3	110	5	4	1	3	2	2	1981	1	1	1	2	1	5	3	4	

5 rows × 181 columns

• 181개의 Feature



	HOUSE	LIFE	gender	Age	marriage	religion	Nationality	Disability	Eco_1	Eco_2	Eco_3	Eco_4	life_5	life_6_1	life_6_2
0	1	1	1	61	1	2	1	2	3.50	4.0	4	3	4.25	4.0	7.0
1	1	1	2	59	1	2	1	2	3.75	4.2	4	2	4.25	6.0	6.0
2	1	1	2	26	2	2	1	2	3.50	3.6	3	3	3.75	9.0	2.0
3	1	3	1	38	1	1	1	2	3.75	3.6	4	3	4.00	4.0	2.0
4	1	3	2	40	1	1	1	2	3.75	3.6	4	3	4.25	6.0	2.0

• 46개의 Feature

• 문항 30개 + Demographic Information
(자가여부, 성별, 나이, 결혼여부, 종교, 국적, 장애인 여부, 서울 거주 기간 등)

1. 데이터 선정 이유 및 문제 정의

- Target

문31. 귀하는 요즘 스스로 행복하다고 생각하십니까? 가장 행복한 상태를 10점으로, 가장 불행한 상태를 0점으로 하여 각 영역 별 자신의 행복점수를 표시해 주십시오.

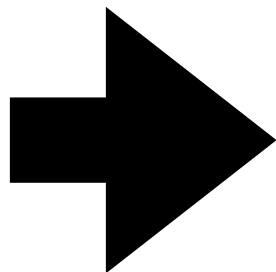
가장 불행한 상태

보통

012345678910가장 행복한 상태

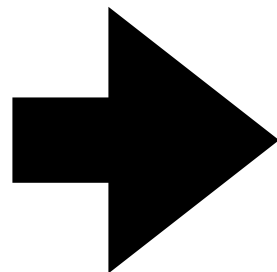
1) 자신의 건강상태	점
2) 자신의 재정상태	점
3) 주위 친지, 친구와의 관계	점
4) 가정생활	점
5) 사회생활(직장, 학교, 종교, 취미, 계모임 등)	점

- 행복 점수



평균
(Cut off)

6.60



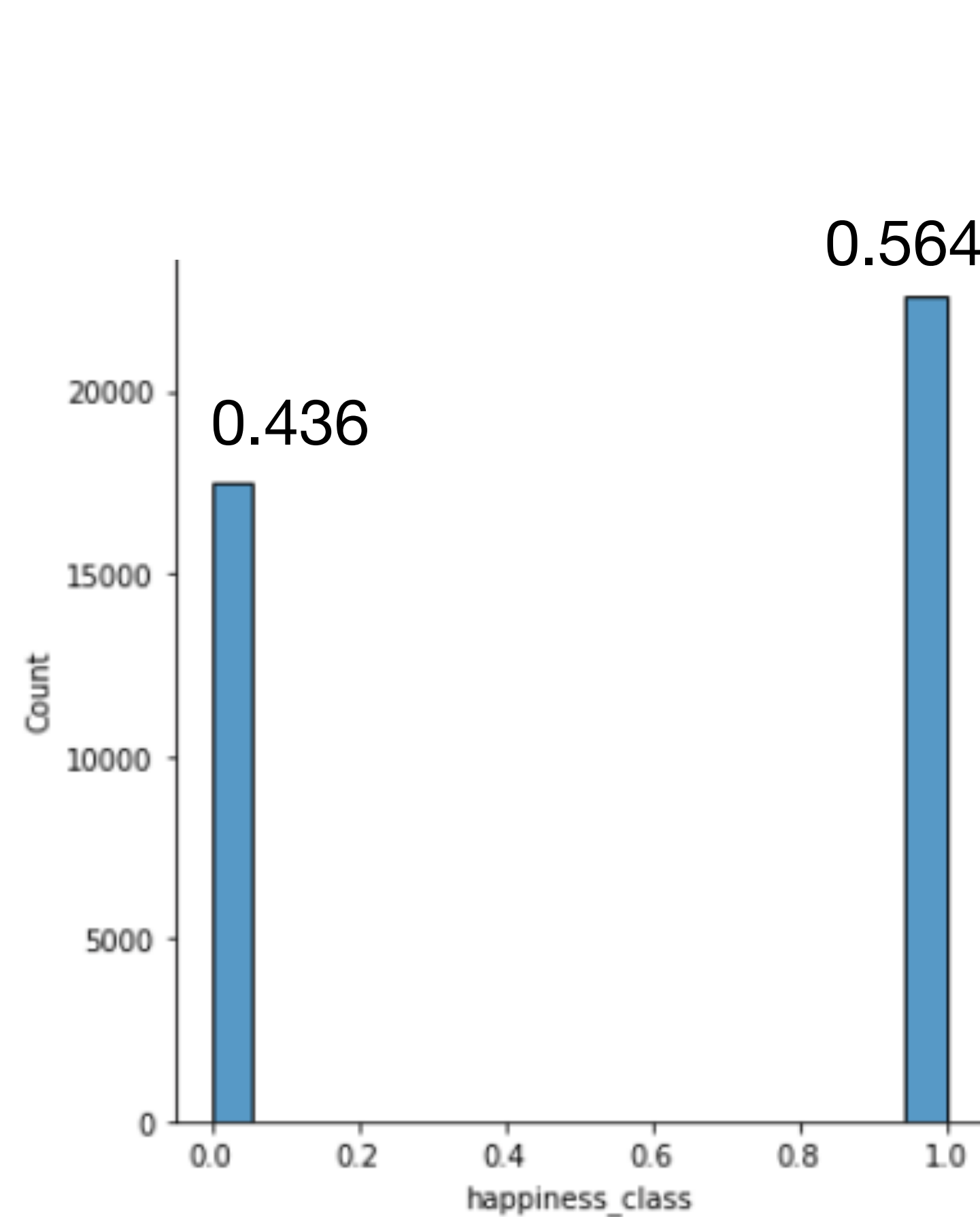
True
False

- 평균보다 행복점수 높음
- 평균보다 행복점수 낮음

2. EDA 와 데이터 전처리

2. EDA와 데이터 전처리

- EDA, Feature Engineering, 데이터의 정규화, 노이즈 제거, 결측치 제거 혹은 대체, 데이터 밸런스, 그외



Weight	Feature
0.0219 ± 0.0023	integration_30
0.0200 ± 0.0033	Age
0.0162 ± 0.0023	health_16
0.0160 ± 0.0043	integration_24_1
0.0098 ± 0.0009	health_17
0.0090 ± 0.0019	integration_25
0.0076 ± 0.0022	retirement_21
0.0076 ± 0.0017	integration_23
0.0075 ± 0.0020	security_7
0.0073 ± 0.0010	life_5
0.0073 ± 0.0025	freetime_15
0.0067 ± 0.0021	living
0.0065 ± 0.0030	transportation_9
0.0061 ± 0.0017	integration_28
0.0055 ± 0.0013	transportation_10
0.0054 ± 0.0014	retirement_20
0.0052 ± 0.0010	freetime_12_2
0.0050 ± 0.0009	integration_29
0.0048 ± 0.0012	life_6_2
0.0047 ± 0.0011	freetime_13_3
0.0042 ± 0.0034	freetime_12_1
0.0040 ± 0.0010	transportation_8
0.0038 ± 0.0004	integration_26
0.0035 ± 0.0014	Eco_1
0.0034 ± 0.0015	life_6_1
0.0032 ± 0.0010	life_6_3
0.0031 ± 0.0008	freetime_13_1
0.0028 ± 0.0006	integration_27_1
0.0027 ± 0.0006	freetime_14
0.0027 ± 0.0008	Eco_2
... 15 more ...	

- 평균 기준 cut-off
 - Balanced class 라고 생각됨
- Permutation Importance
 - Feature Selection 해보았으나 정확도가 더 떨어짐.
 - 모든 Feature를 사용
- 결측치는 0으로 채움

3. 머신러닝 방식 적용 및 교차 검증

3. 머신러닝 방식 적용 및 교차 검증

- 분류 문제 (랜덤포레스트, XGboost 사용)

1. Train set / Validation set / Test set 으로 분리
2. 랜덤포레스트, XGboost 모델 사용
3. XGBoost hyperparameter 조정 (RandomizedSearchCV 사용)

Roc Auc Score	Random Forest	Random Forest	XGBoost	XGBoost
Validation set	0.786	0.739	0.790	0.762
Test set			0.788	

- Baseline : 0.564

3. 머신러닝 방식 적용 및 교차 검증

- 분류 문제 (랜덤포레스트, XGboost 사용)

1. Train set / Validation set / Test set 으로 분리
2. 랜덤포레스트, XGboost 모델 사용
3. XGBoost hyperparameter 조정 (RandomizedSearchCV 사용)

n_estimator = 200
max_depth = 7
learning_rate = 0.2
random_state = 2
n_jobs = -1

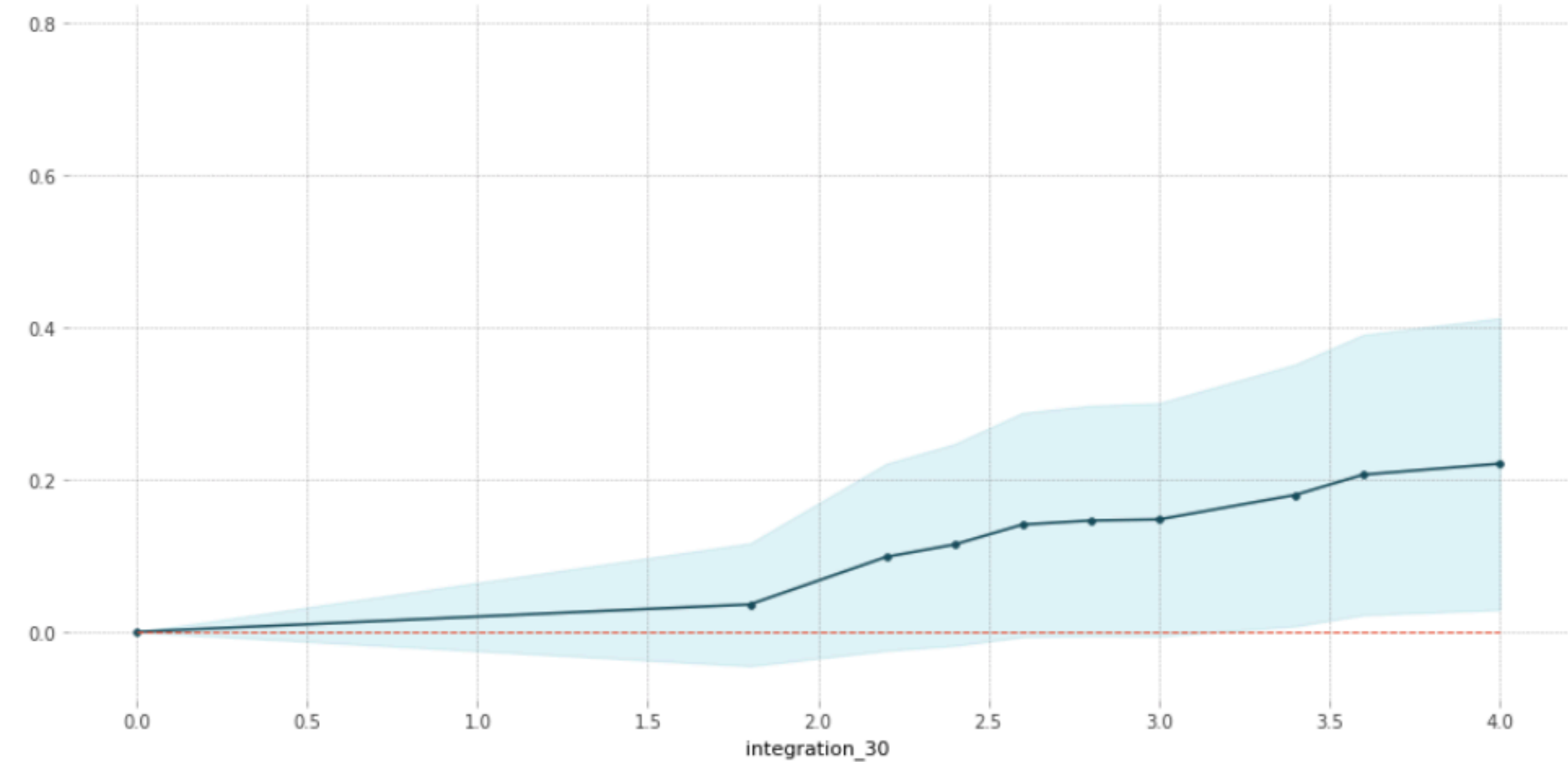
Roc Auc Score	Random Forest	Random Forest	XGBoost	XGBoost
Validation set	0.786	0.739	0.790	0.762
Test set			0.788	

- Baseline : 0.564

4. 머신러닝 모델 해석

4. 머신러닝 모델 해석

PDP for feature "integration_30"
Number of unique grid points: 10

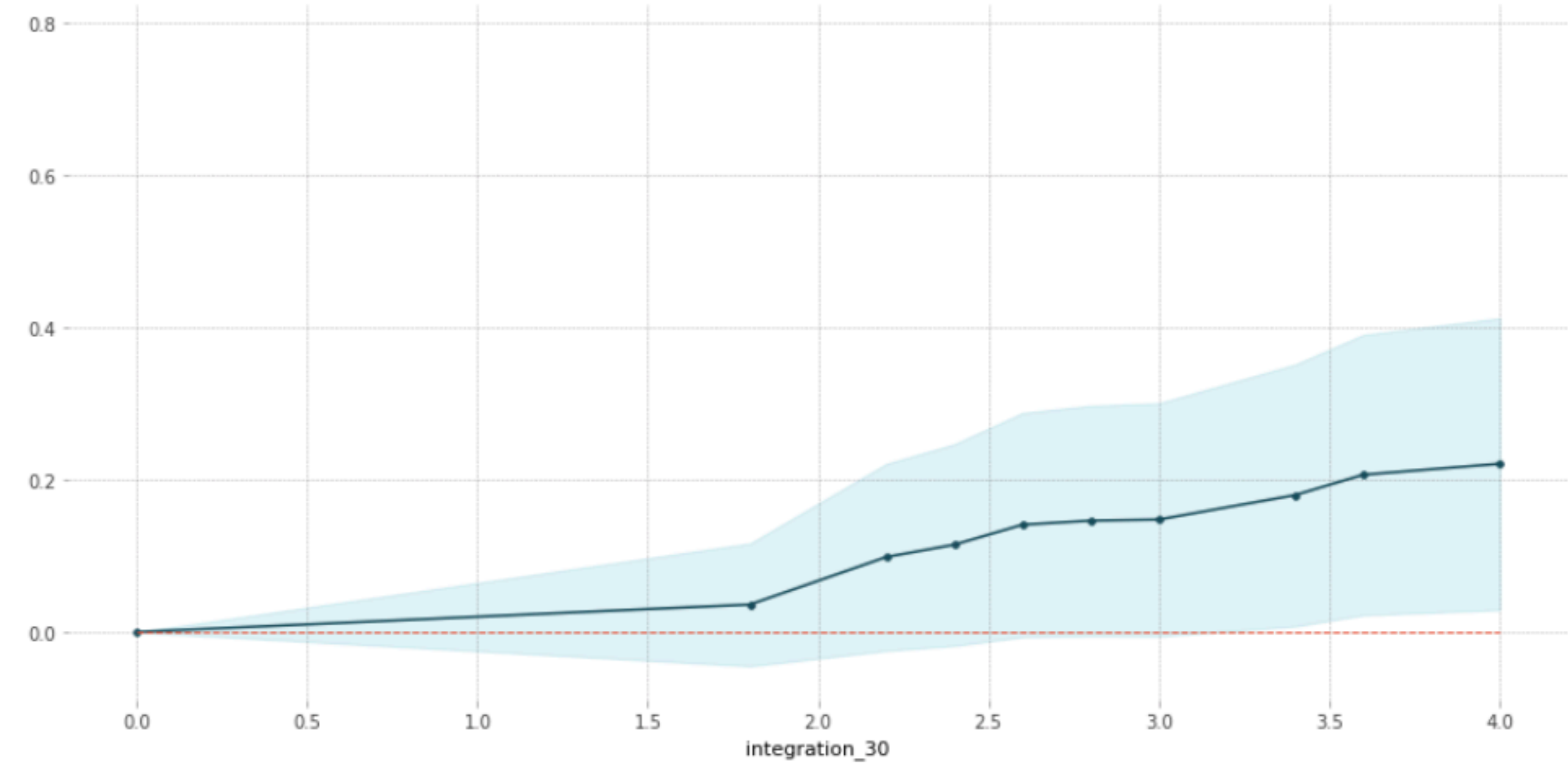


문30. 귀하의 가족 관계에 대하여 다음의 각 항목에 대해 말씀해 주십시오. 같이 살고 있지 않아도 응답해 주십시오.

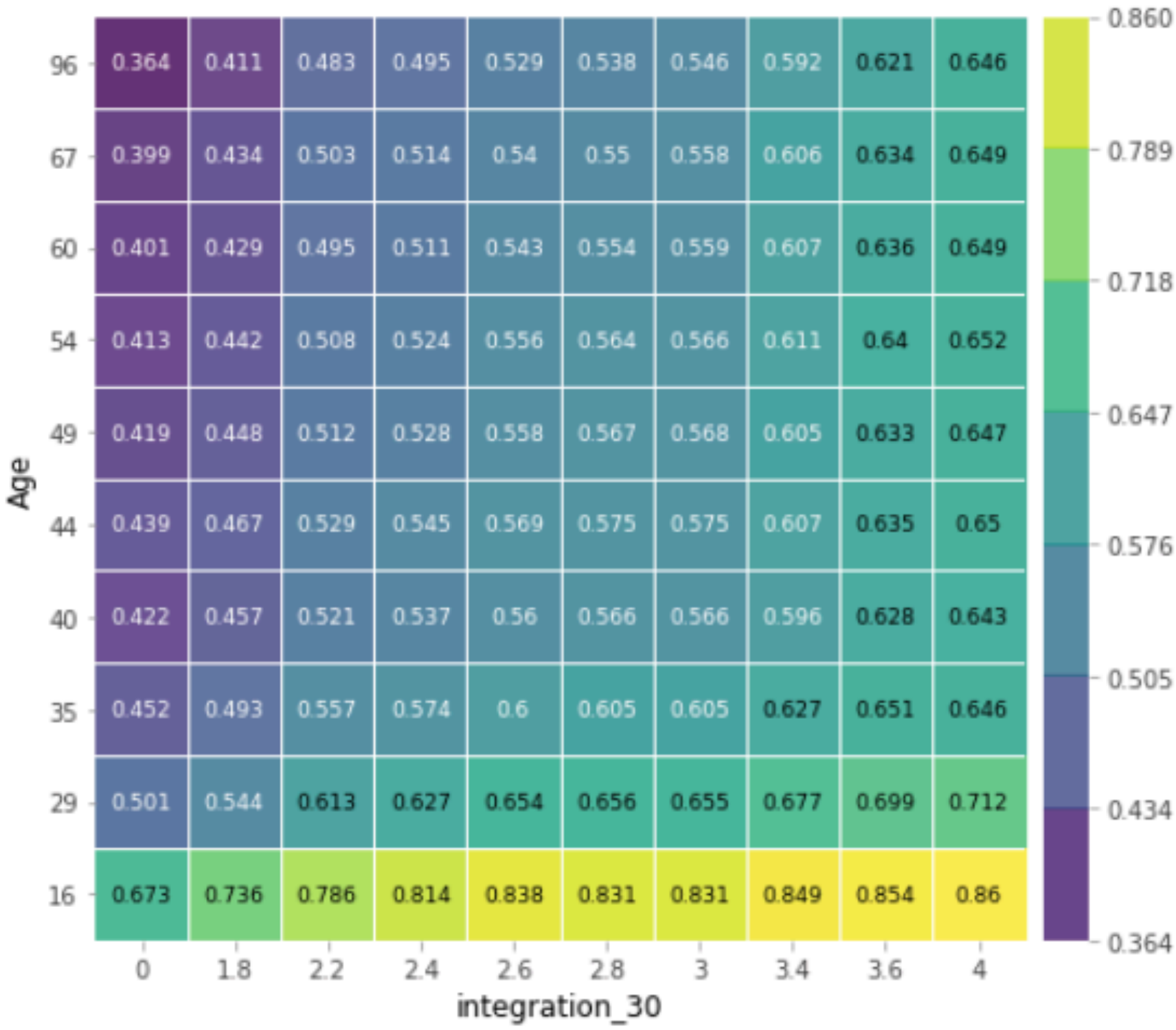
거의 안 한다	가끔 한다	자주 한다	항상 한다	해당사항 없음
1-----	2-----	3-----	4-----	9-----
1) 가족과의 식사				
2) 자녀 또는 부모님과의 대화				
3) 자녀의 배우자 또는 배우자의 부모님과 대화				
4) 부부, 형제, 남매, 자매간 가정 문제 상의				
5) 가족과의 여가 생활				

4. 머신러닝 모델 해석

PDP for feature "integration_30"
Number of unique grid points: 10



PDP interact for "integration_30" and "Age"
Number of unique grid points: (integration_30: 10, Age: 10)



4. 머신러닝 모델 해석

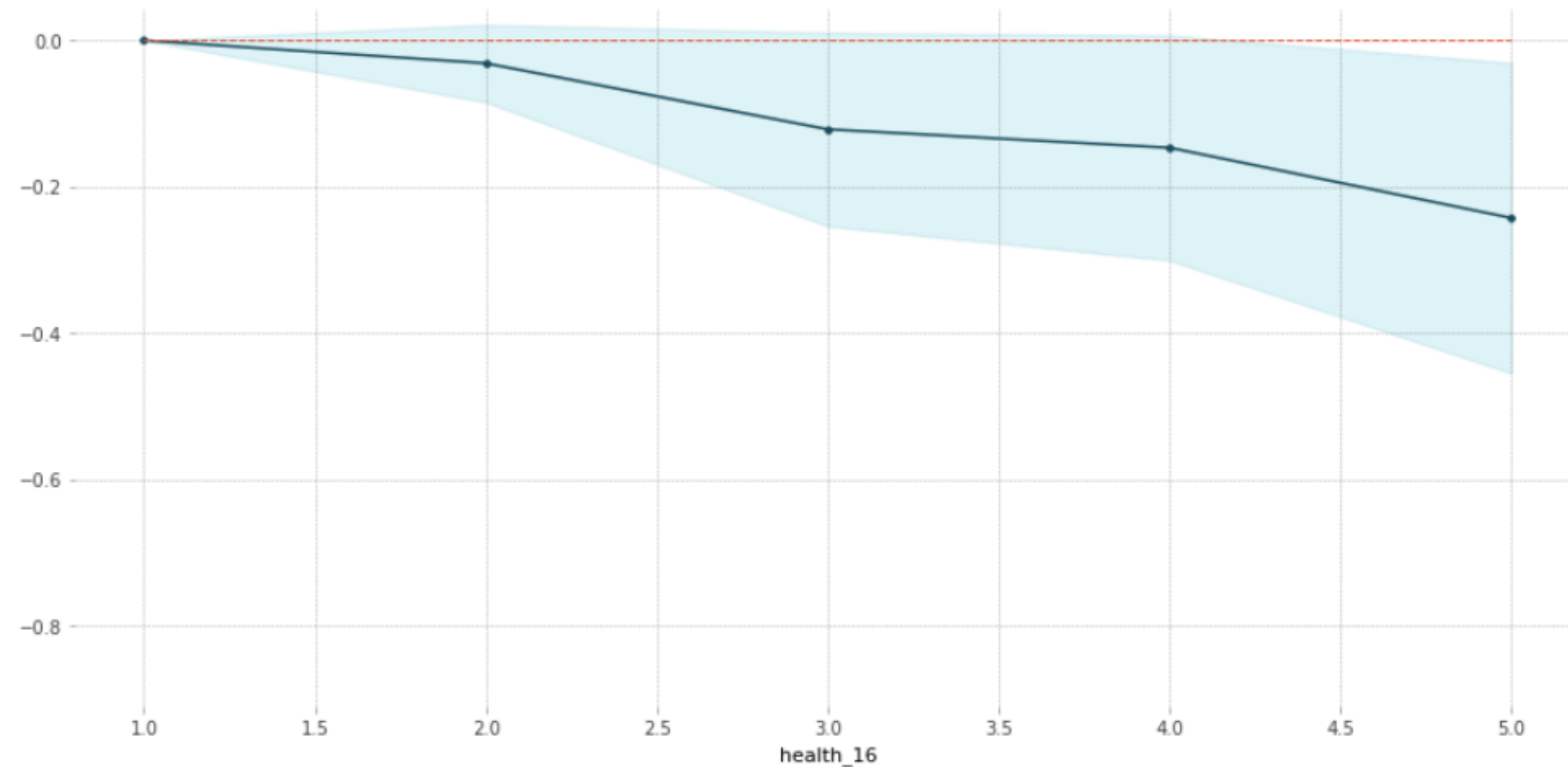
- PDP, SHAP

문16. 귀하는 지난 2주일 동안 일상생활에서 전반적으로 스트레스를 어느 정도 느꼈습니까?

- ① 전혀 느끼지 않았다 ② 느끼지 않은 편이다
③ 보통이다
④ 느낀 편이다 ⑤ 매우 많이 느꼈다

PDP for feature "health_16"

Number of unique grid points: 5

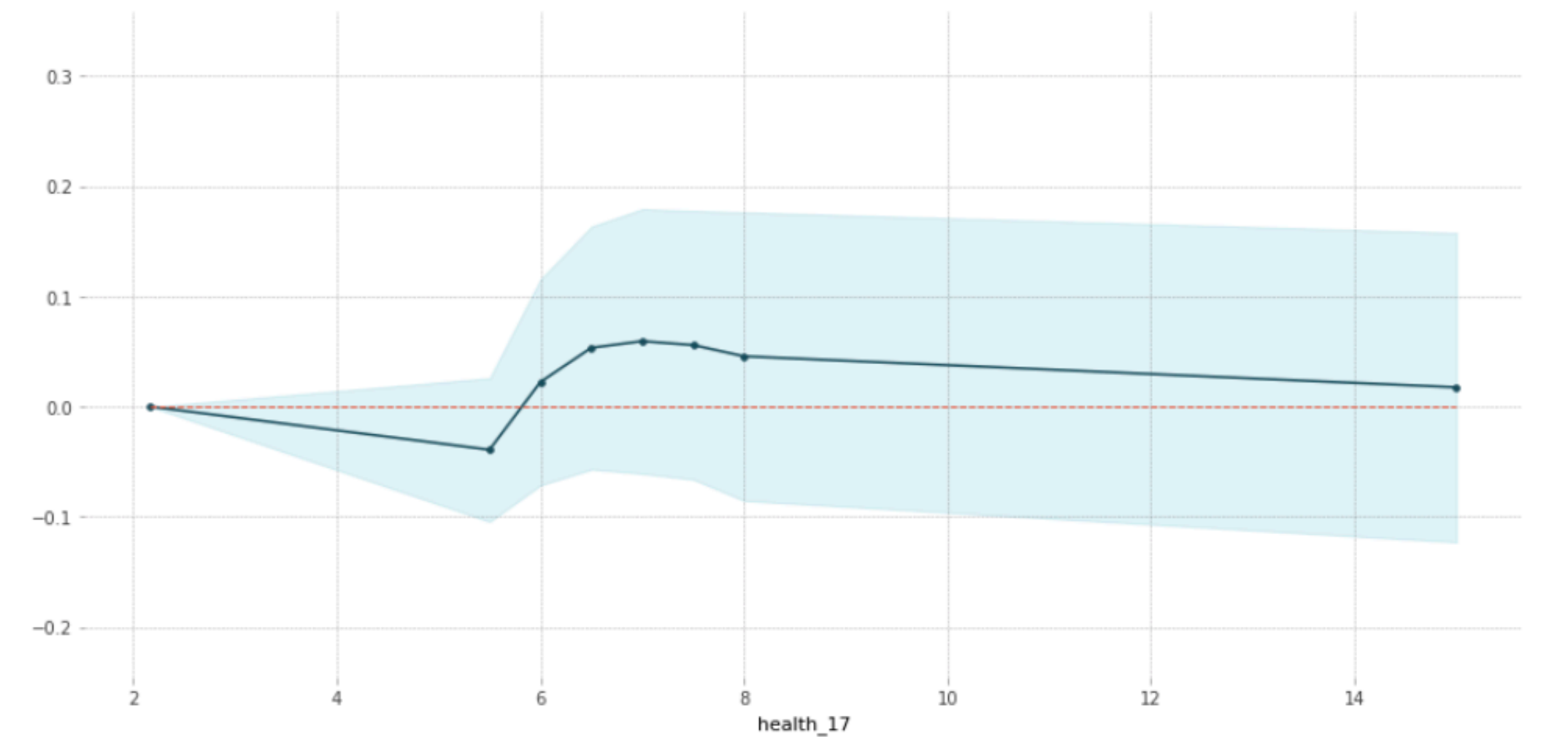


문17. 귀하는 지난 한 주 동안 하루에 평균 몇 시간을 주무십니까?

_____ 시간 _____ 분

PDP for feature "health_17"

Number of unique grid points: 8



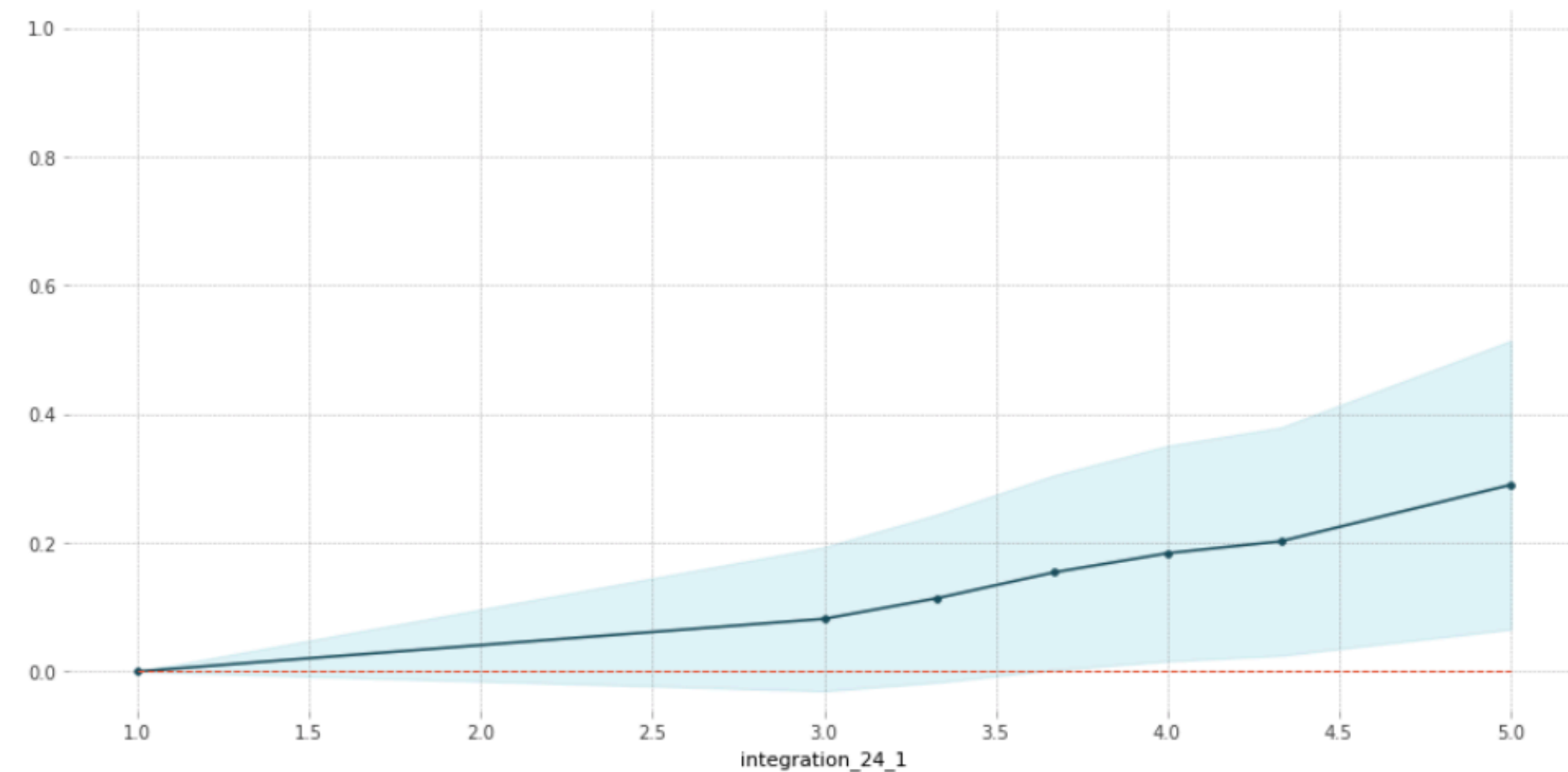
4. 머신러닝 모델 해석

문24. 귀하는 다음 각 사람 또는 기관을 얼마나 신뢰하십니까?

	전혀 신뢰 안함	별로 신뢰 안함	보통	다소 신뢰	매우 신뢰
	1-----	2-----	3-----	4-----	5
1) 가족					
2) 친구					
3) 이웃					
4) 처음 만난 낯선 사람					
5) 국내 거주 외국인					
6) 공공 기관(서울시, 구청 등)					

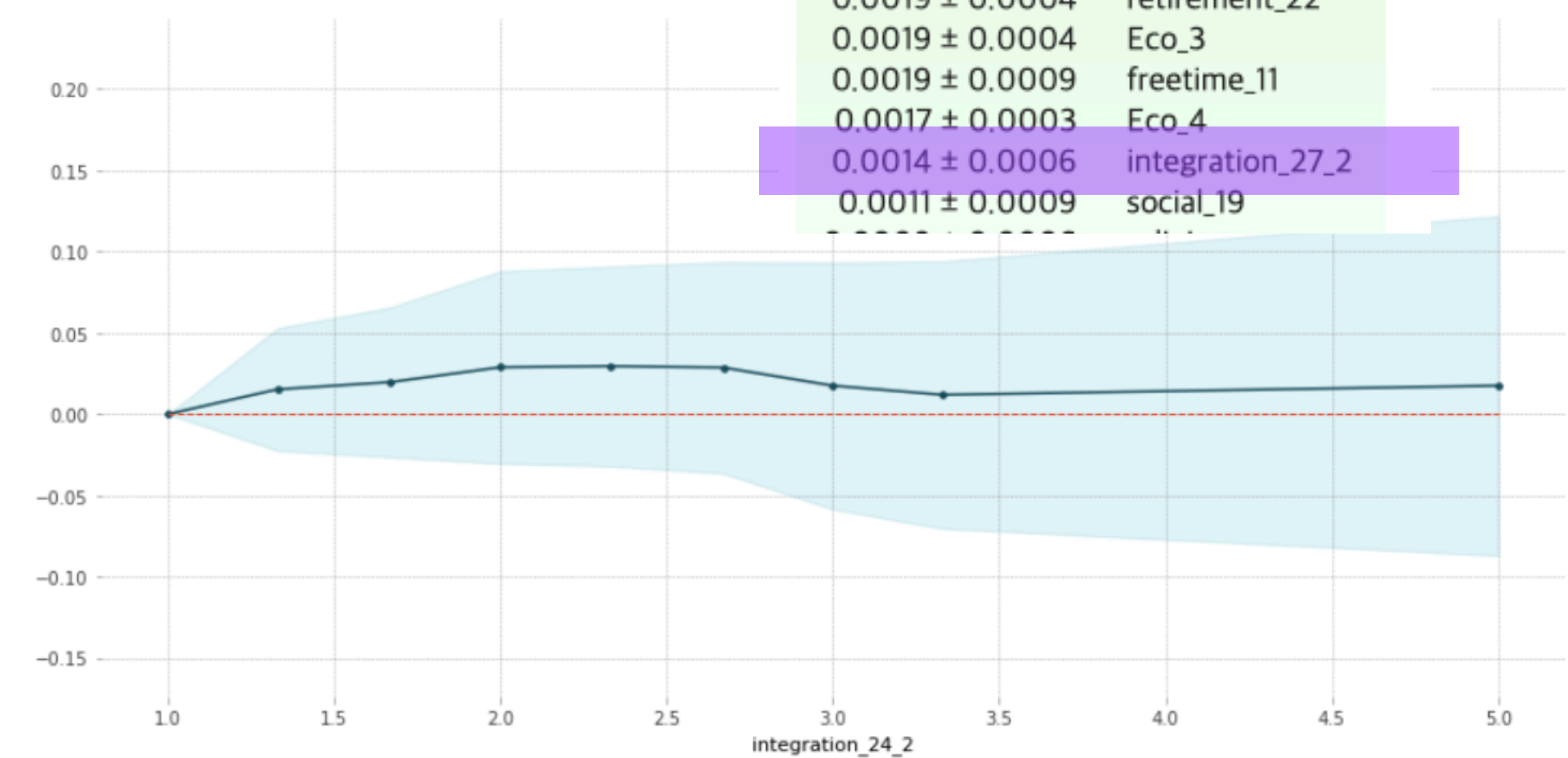
PDP for feature "integration_24_1"

Number of unique grid points: 7



PDP for feature "integration_24_2"

Number of unique grid points: 9



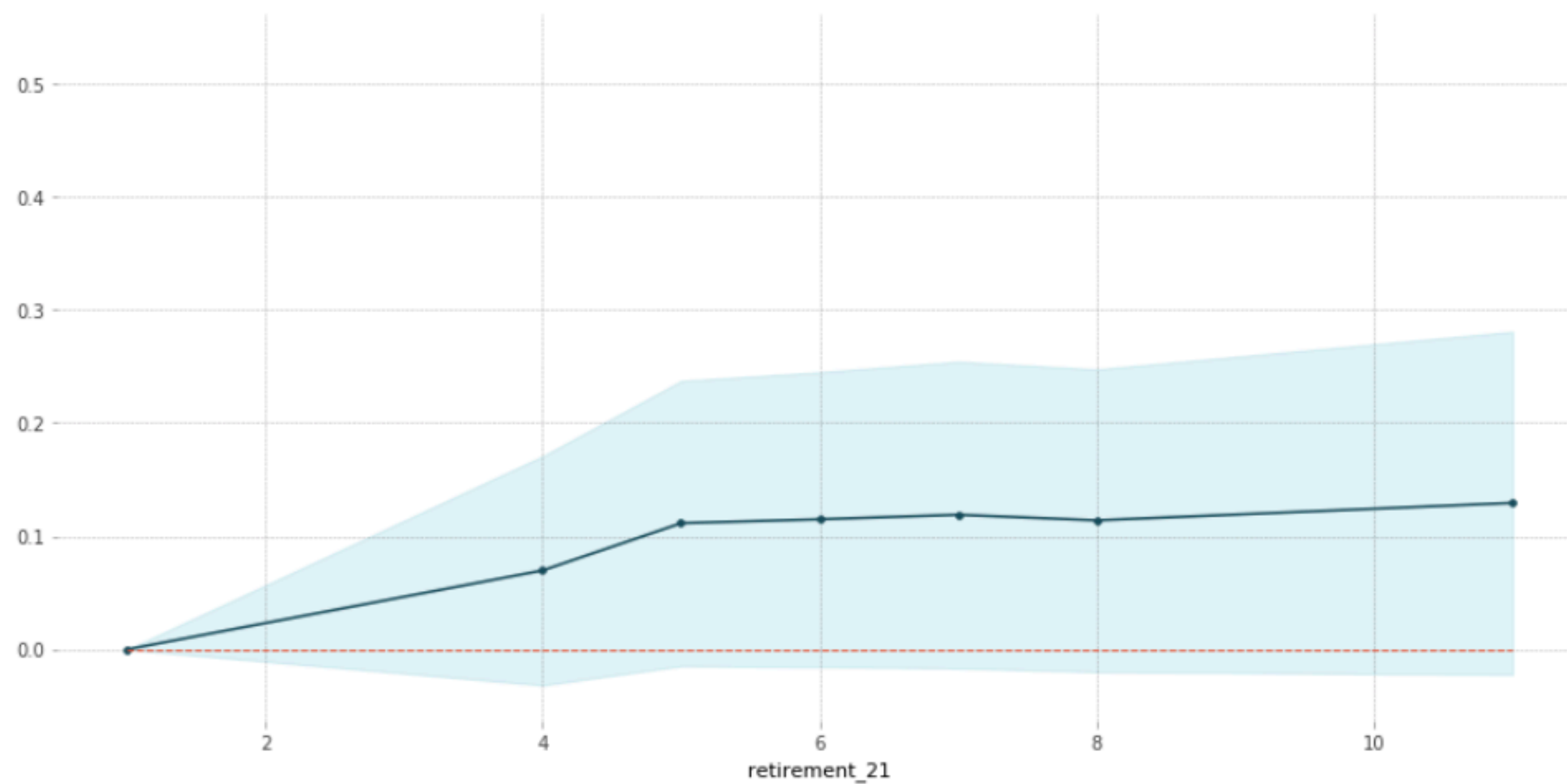
Weight Feature

0.0219 ± 0.0023	integration_30
0.0200 ± 0.0033	Age
0.0162 ± 0.0023	health_16
0.0160 ± 0.0043	integration_24_1
0.0098 ± 0.0009	health_17
0.0090 ± 0.0019	integration_25
0.0076 ± 0.0022	retirement_21
0.0076 ± 0.0017	integration_23
0.0075 ± 0.0020	security_7
0.0073 ± 0.0010	life_5
0.0073 ± 0.0025	freetime_15
0.0067 ± 0.0021	living
0.0065 ± 0.0030	transportation_9
0.0061 ± 0.0017	integration_28
0.0055 ± 0.0013	transportation_10
0.0054 ± 0.0014	retirement_20
0.0052 ± 0.0010	freetime_12_2
0.0050 ± 0.0009	integration_29
0.0048 ± 0.0012	life_6_2
0.0047 ± 0.0011	freetime_13_3
0.0042 ± 0.0034	freetime_12_1
0.0040 ± 0.0010	transportation_8
0.0038 ± 0.0004	integration_26
0.0035 ± 0.0014	Eco_1
0.0034 ± 0.0015	life_6_1
0.0032 ± 0.0010	life_6_3
0.0031 ± 0.0008	freetime_13_1
0.0028 ± 0.0006	integration_27_1
0.0027 ± 0.0006	freetime_14
0.0027 ± 0.0008	Eco_2
0.0026 ± 0.0007	integration_24_2
0.0022 ± 0.0007	LIFE
0.0019 ± 0.0004	retirement_22
0.0019 ± 0.0004	Eco_3
0.0019 ± 0.0009	freetime_11
0.0017 ± 0.0003	Eco_4
0.0014 ± 0.0006	integration_27_2
0.0011 ± 0.0009	social_19

4. 머신러닝 모델 해석

PDP for feature "retirement_21"

Number of unique grid points: 7

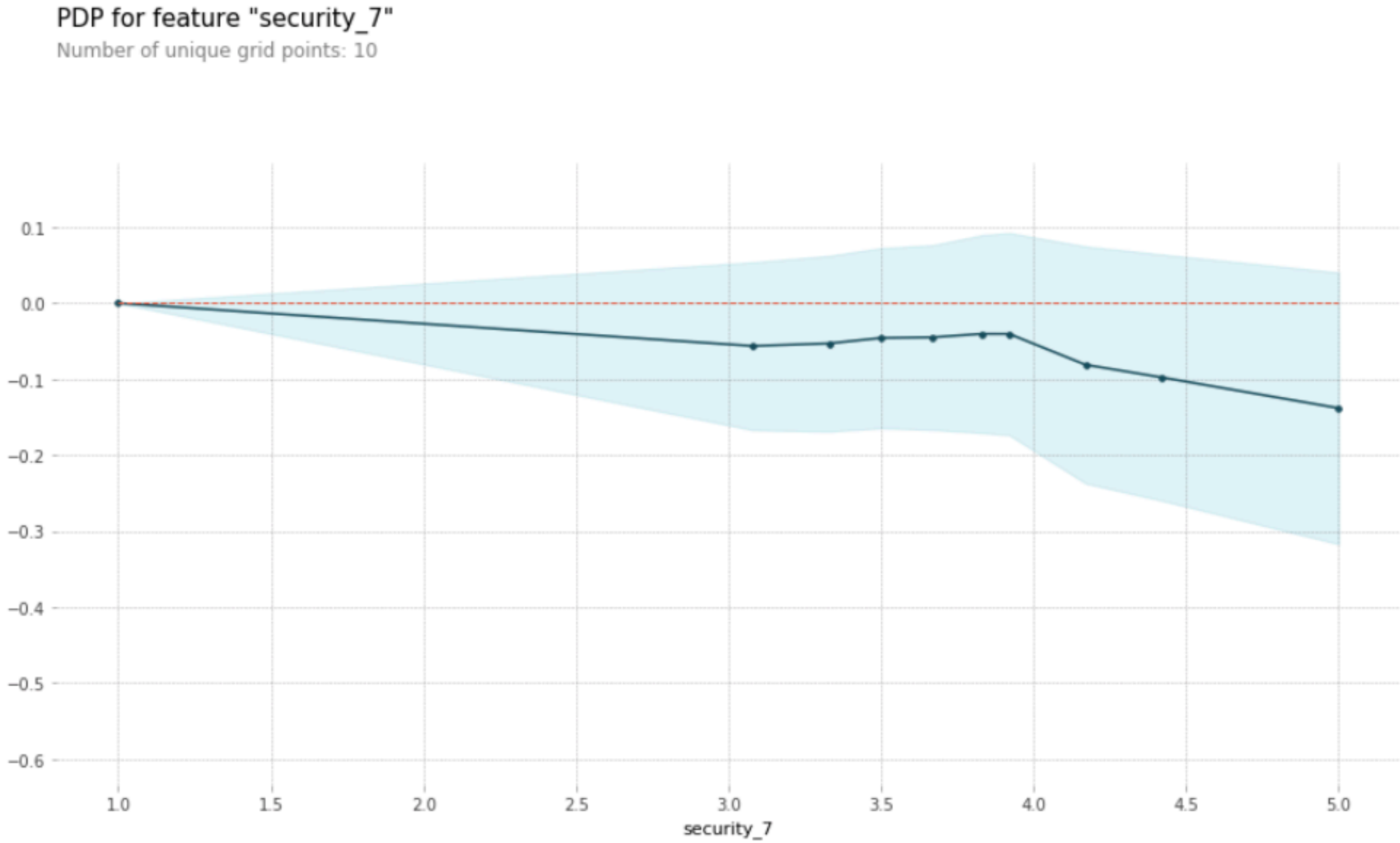


문21. 귀하는 은퇴 후의 월 생활비는 얼마가 적절하다고 생각하십니까? 현재의 물가를 기준으로 말씀해 주십시오.

* 가구 기준으로 응답

- | | |
|------------------|------------------|
| ① 50만원 미만 | ⑫ 50 ~ 100만원 미만 |
| ③ 100 ~ 150만원 미만 | ④ 150 ~ 200만원 미만 |
| ⑤ 200 ~ 250만원 미만 | ⑥ 250 ~ 300만원 미만 |
| ⑦ 300 ~ 350만원 미만 | ⑧ 350 ~ 400만원 미만 |
| ⑨ 400 ~ 450만원 미만 | ⑩ 450 ~ 500만원 미만 |
| ⑪ 500만원 이상 | |

4. 머신러닝 모델 해석



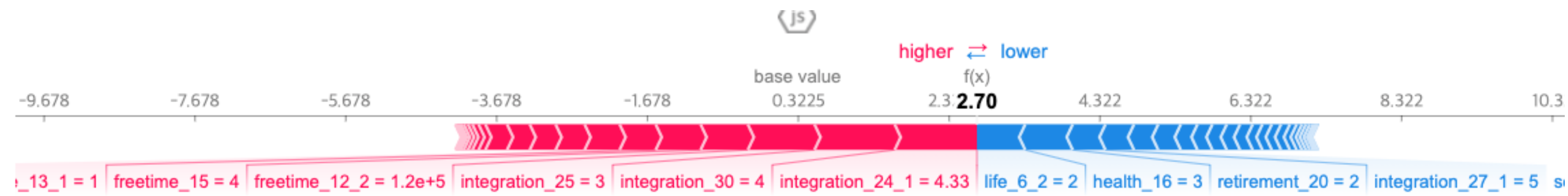
문7. 다음의 각 항목에 대해 우리 사회가 어느 정도 심각하다고 생각하십니까?

전혀 심각하지 않다 1	별로 심각하지 않다 2	보통 이다 3	약간 심각하다 4	매우 심각하다 5
1) 자연재해(태풍, 지진, 홍수 등)				
2) 핵폐기물, 방사능 사고				
3) 감염병(코로나, 사스, 결핵, 콜레라, 장티푸스, 메르스 등)				
4) 안전사고(선박사고, 항공기사고, 싱크홀 등)				
5) 안보(전쟁, 국제분쟁, 북한과 대치 등)				
6) 경제위기(금융위기 등)				
7) 실업				
8) 부정부패				
9) 폭력 범죄(성폭력, 학교폭력, 강도, 유괴, 폭행 및 살해 등)				
10) 사회 갈등(빈부격차, 불평등, 세대갈등 등)				
11) 컴퓨터 바이러스, 해킹 등으로 인한 사이버 보안, 개인정보유출 문제				
12) 인터넷상의 괴롭힘(모욕, 따돌림, 협박, 명예훼손 등)				

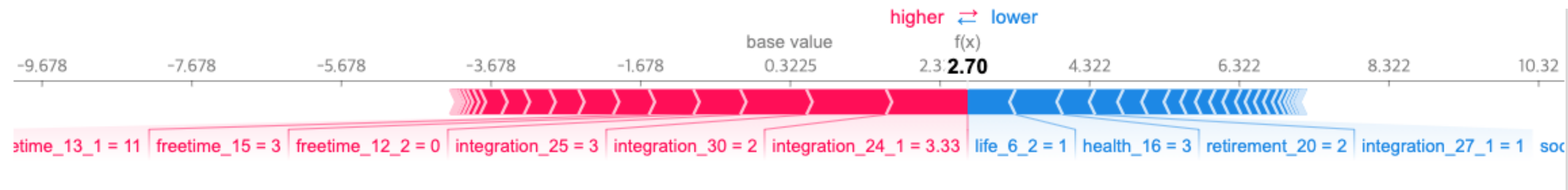
4. 머신러닝 모델 해석

- PDP, SHAP

- 52세 한국 국적 등록장애인 여성 : 행복점수 평균 이상



- 62세 한국 국적 등록장애인 남성 : 행복점수 평균 이하



5. 결론

5. 결론

* 정책 반영 참고 요소

- 1) 가족, 친구, 이웃을 많이 신뢰한다고 응답한 사람일 수록 행복점수가 평균보다 높을 가능성이 올라감
- 2) 가족과의 식사, 대화, 여가생활을 자주 한다고 응답한 사람일수록 행복점수가 평균보다 높을 가능성이 올라감
- 3) 은퇴 이후 최소 생활비는 200만원을 기점으로 행복점수가 평균보다 높을 가능성이 올라가지는 않음
- 4) 수면 시간이 7시간 정도일 때 행복 점수가 평균보다 높을 가능성이 제일 높음

5. 결론

* 향후 개선 방안

1) 각 부서가 담당하는 업무에 맞는 분석 필요

ex. 미혼모 지원 부서, 외국인 지원 부서, 장애인 지원 부서 등은 결혼 여부, 외국인 여부, 등록 장애인 여부에 따른 예측모델 필요. 맞춤형 지원 정책 기획에 도움
- 데이터 수가 현저히 적다는 한계. 샘플을 고르게 수집할 필요성

2) 서울 거주 기간, 자가 여부, 주택 형태 등의 특성에 따른 예측모델 세분화 필요성

- Feature의 수를 조절하여 다양한 시나리오를 토대로 행복점수 예측

감사합니다