

AIB_13_Section2_Project

당뇨병과 관련된 간단한 설문조사 dataset과 머신러닝
모델을 활용한 당뇨병 진단모델의 학습 및 최적화

AI_13_김강호

목차

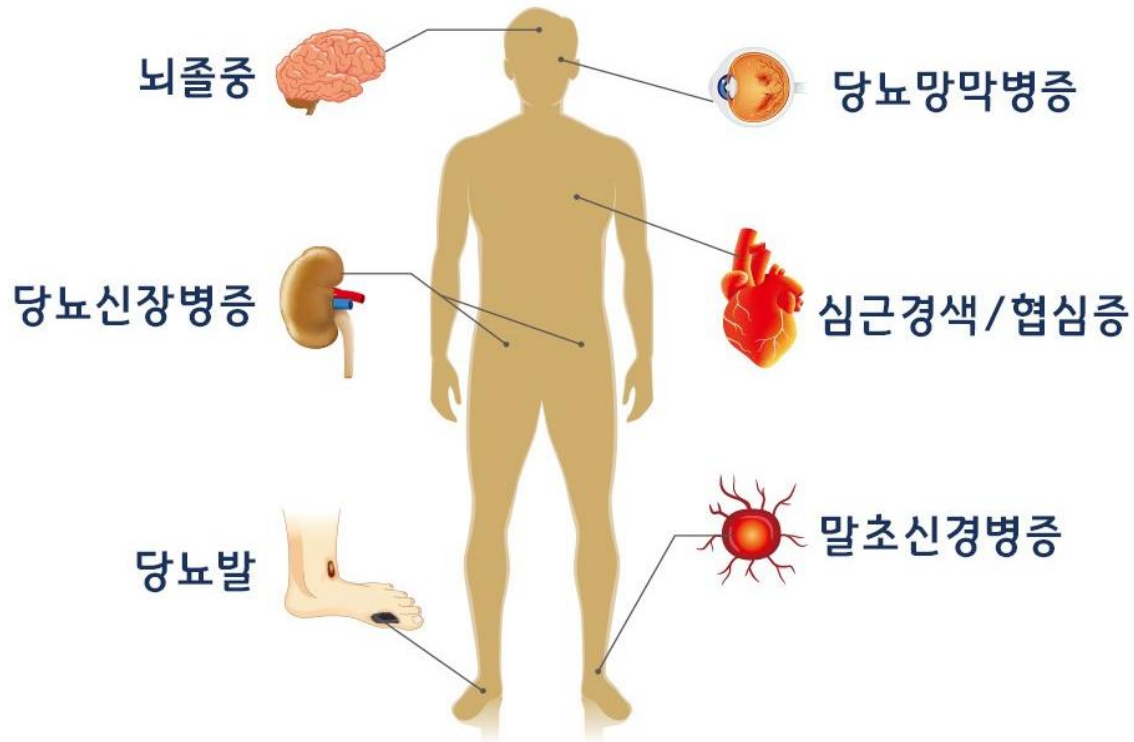
1. 문제정의 및 데이터 선정
2. 데이터 전처리, EDA, 시각화
3. 모델링 및 모델 해석
4. 결론

1. 문제정의 및 데이터 선정

1]. 당뇨병??

[2]

당뇨병 합병증



- 당뇨병(Diabetes) : 높은 혈당 수치가 오랜 기간 지속되는 대사 질환
- 혈당이 높은 상태가 지속이 될 경우, 여러가지 합병증 발생 [1]
 - **뇌 문제** : 뇌졸중
 - **시력 문제** : 당뇨망막병증
 - **신진대사 문제** : 심근경색, 당뇨발, 말초신경병증 등
 - **일상생활에 지장을 줌** >> 치료 필요

[1] <https://ko.wikipedia.org/wiki/%EB%8B%B9%EB%87%A8%EB%B3%91>

[2] https://post-phinf.pstatic.net/MjAxOTA3MjJfMTUx/MDAxNTYzNzYzNTAzMTc0.-ncDM1Jy2txPPTeuyxUzNbZ_OL_byCNf1rQJa5KLiWEg.JltC1CalKE-o1DRqW84oSqPheVk9QZlrfYC7Sg371yEg.JPEG/%EB%8B%B9%EB%87%A8%EB%B3%91%ED%95%A9%EB%B3%91%EC%A6%9D.jpg?type=w1200

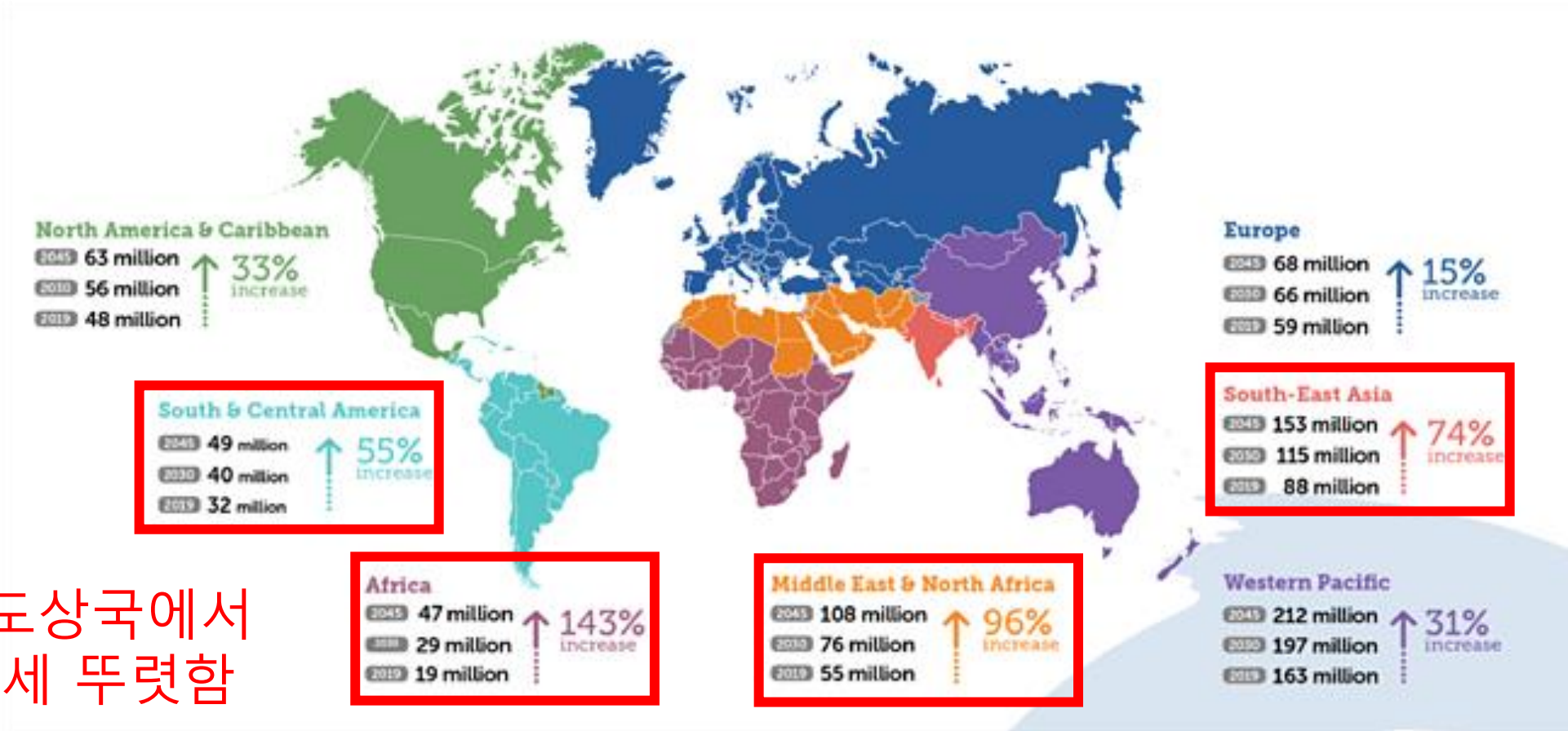
2]. 당뇨병 현황

- 전 세계의 많은 사람들이 당뇨병을 앓고 있음.
 - ✓ 우리나라 30세 이상 성인 7명 중 1명이 당뇨병 (2018년 기준) [1]
 - ✓ 전 세계 성인의 약 10분의 1(5억 3700만명)이 당뇨병 (2021년 기준) [2]
 - ✓ 전 세계의 당뇨병 환자가 증가할 예정 [2]
 - 2045년 경에는 약 7억 8300만명까지 증가 예정
- 당뇨병 환자들 상당수[약 4분의 3]가 개발도상국 사람들 [2]

[1] Diabetes Fact Sheet in Korea, 2020, 대한당뇨병학회

[2] <https://diabetesatlas.org/>

2]. 당뇨병 현황



개발도상국에서
증가세 뚜렷함

각 대륙 별 당뇨병 환자 증가 추세 전망치 [1]

2]. 당뇨병 현황

- 앞으로 당뇨병을 잘 관리하지 못하면(특히 개발도상국)

✓ 의료비와 인력문제와 같은 많은 사회적 비용 발생하고, 그 피해는 우리들이 떠안아야 된다.



∴ 당뇨병을 앓고 있는 사람을 줄여야 함.

3]. 당뇨병 진단

- 당뇨병 환자들을 줄일려면??

- 조기에 진단해서 관리를 하는 것이 중요

- 당뇨병 진단 요구조건 ([1]) :

- ✓ **기구** : 자가혈당측정기, 채혈기, 채혈침(란셋), 시험지(스트립), 혈당 관리수첩
 - ✓ **검사항목** : 혈당, 경구 당 부하, 당화혈색소 등 여러 화학적 수치들
 - ✓ **의료진** : 의사와 간호사가 당뇨병 검사를 진행하고 판단함

3]. 당뇨병 진단

- 개발도상국 = 당뇨병 진단에 필요한 자원과 인력이 부족
 - ✓ 기존 검사들 보다 더 간단한 절차를 통해 당뇨병 여부를 판단할 필요가 있음



4]. 해결하고자 하는 문제 & 프로젝트 목표

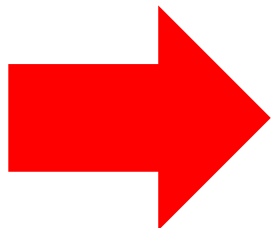
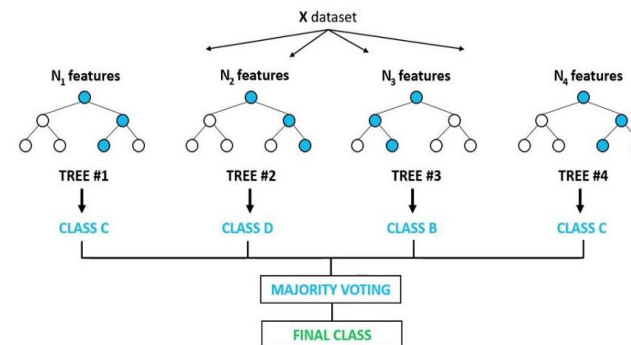
- 당뇨병과 관련된 간단한 설문조사 dataset과 머신러닝 모델을 활용한 당뇨병 진단모델의 학습 및 최적화

당뇨병 관련 설문조사 Dataset

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker
0	0.0	1.0	1.0	1.0	40.0	1.0
1	0.0	0.0	0.0	0.0	25.0	1.0
2	0.0	1.0	1.0	1.0	28.0	0.0
3	0.0	1.0	0.0	1.0	27.0	0.0



ML models



당뇨병 진단모델의 학습 및 최적화

4]. 해결하고자 하는 문제 & 프로젝트 목표

- 당뇨병 진단모델의 기대효과 :
 - ✓ 개발도상국 등 의료여건이 부족한 사람들이 당뇨병 여부를 판단
 - ✓ 당뇨병을 여부를 알게 되면 사람들이 경각심을 가지고 관리를 하게 됨
 - ✓ 당뇨병으로 인한 사회적문제를 줄이는 데 기여

5]. 사용 할 데이터 & 머신러닝 분류모델

- 사용 할 데이터 : 당뇨병 관련 설문조사 dataset

➤ CDC(미국 질병통제예방센터)에서 **25만 3680건의 22개의 항목**을 조사한 당뇨병 관련 설문조사를 진행하여 구성된 데이터

데이터셋 크기 : 253,680 rows × 22 columns

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0

	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1.0	0.0	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
1	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0	1.0
2	0.0	0.0	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0	8.0

5]. 사용 할 데이터 & 머신러닝 분류모델

- 예측해야 할 타겟 변수 : Diabetes_012 - 당뇨병 유무
 - 0 = 당뇨병 없음, 1 = 당뇨병 전 단계, 2 = 당뇨병
 - 1과 2 모두 당뇨병에 대한 케어를 받아야 하는 점에서 모두 1로 치환하고 컬럼명을 Diabetes_01로 바꿈

타겟변수

데이터셋 크기 : 253,680 rows × 22 columns

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0

	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1.0	0.0	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
1	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0	1.0
2	0.0	0.0	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0	8.0

5]. 사용 할 데이터 & 머신러닝 분류모델

• 예측(을 수행 할) 변수

- 고혈압, 고 콜레스테롤 여부, 흡연자 여부, 신체활동 여부, 과일, 야채 섭취여부, 육체 및 정신적으로 건강한지 여부, 성별, 나이, 학벌, 수입 등과 같이 쉽게 응답할 수 있는 정보들로 당뇨병 여부 예측

데이터셋 크기 : 253,680 rows × 22 columns

예측(을 수행할)변수

Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	1.0

	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1.0	0.0	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
1	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0	1.0
2	0.0	0.0	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0	8.0

5]. 사용 할 데이터 & 머신러닝 분류모델

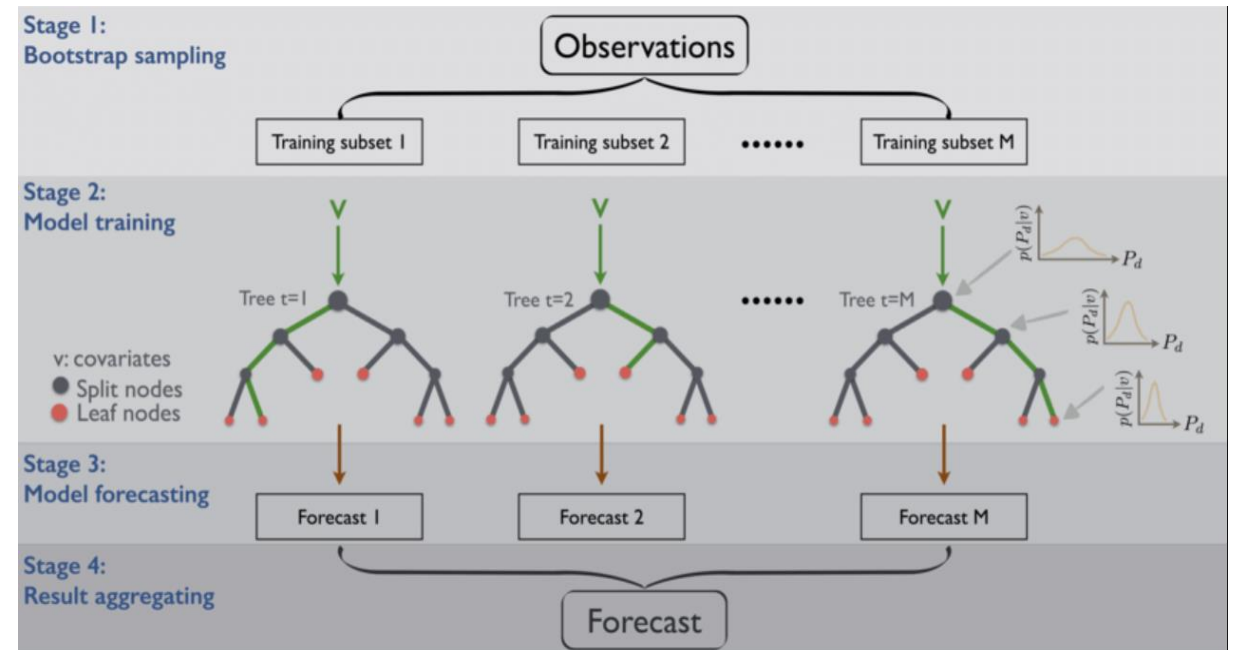
- 머신러닝 분류모델

- 랜덤 포레스트 모델

- ✓ 비교적 높은 정확도, 일반화 성능

- ✓ 다른 머신러닝 모델에 비해 간편하고 빠른 학습 및 테스트 알고리즘

- ✓ 당뇨병 분류모델에 적합



2. EDA, 데이터 전처리, 시각화

1]. 데이터 전처리 과정

- 1). Pandas profiling, EDA(탐험적 데이터 분석)
- 2). 타겟 데이터의 Binary classification
- 3). 특성분석
- 4). 중복행 제거
- 5). Train, test 세트 분리
- 6). 이상치 제거
- 7). 기타 특성공학
 - 1}. 특성드랍
 - 2}. 특성공학 후 생기는 중복 행 제거
- 8). feature, target 데이터셋 분리
- 9). 오버샘플링
- 10). SelectKBest를 이용한 중요특성 선택



머신러닝 모델 학습을 원활하게 하기 위한 준비과정

1]. 데이터 전처리 과정

1). Pandas profiling, EDA(탐험적 데이터 분석)

@ 중점 확인 사항

- 각 피쳐들(예측변수들) 별 데이터타입
- 결측치
- 데이터의 중복여부
- 값의 분포
- 피쳐들 간의 상관관계



- 2). 타겟 데이터의 Binary classification
- 3). 특성분석
- 4). 중복행 제거
- 5) Train, test 세트 분리
- 6). 이상치 제거
- 7). 기타 특성공학
 - 1}. 특성드랍
 - 2}. 특성공학 후 생기는 중복
행 제거
- 8). feature, target 데이터셋 분리
- 9). 오버샘플링
- 10). SelectKBest를 이용한 중요특성 선택

1]. 데이터 전처리 과정

10). SelectKBest를 이용한 중요특성 선택

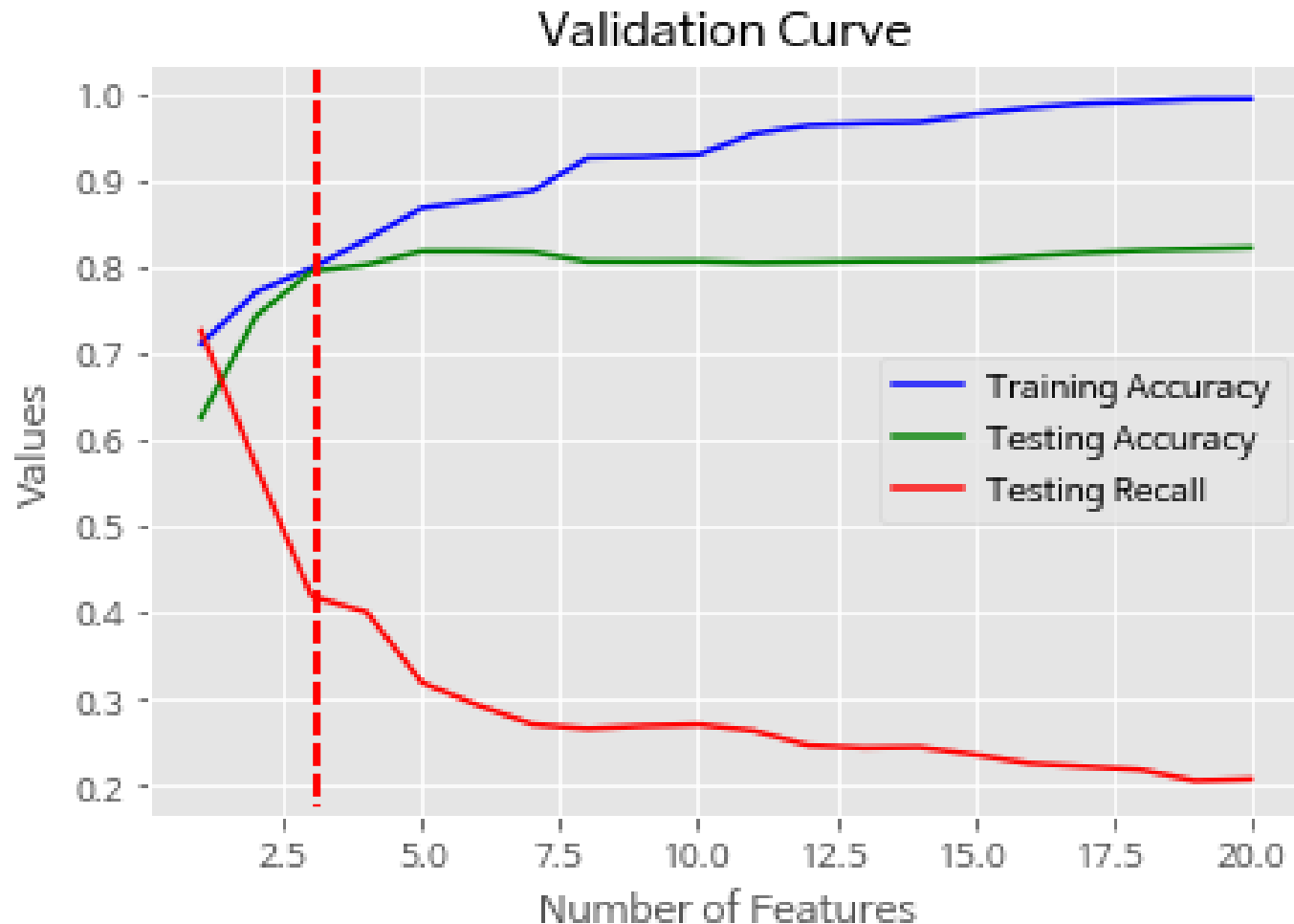
1) HighBP : 혈압이 높은 지 여부
2) HighChol : 콜레스테롤이 높은지 여부
3) CholCheck : 5년 내, 콜레스테롤 검사 여부
4) BMI : 체질량 지수
5) Smoker : 평생 최소 100개비의 담배를 피웠는지 여부
6) Stroke : 뇌졸중 발생여부
7) HeartDiseaseorAttack : 심장질환 발생여부
..... 총 20가지 예측변수 존재



1) HighBP : 혈압이 높은 지 여부
2) HighChol : 콜레스테롤이 높은지 여부
3) GenHlth : 건강 상태에 대한 척도

1]. 데이터 전처리 과정

10). SelectKBest를 이용한 중요특성 선택



• 중요특성 개수의 선택 기준

- 일반화가 잘 되는가? (Train 데이터와 Test 데이터의 성능차이 최소화)
- 당뇨병 분류에서 중요한 기준인 재현률이 높은가?
- 하이퍼 파라미터 튜닝을 통한 성능개선여지

∴ 중요특성 :
HighBP, HighChol, GenHlth

3. 모델링 및 모델 해석

1]. 분류문제에서의 기준모델 & 재현율

- 분류문제에서의 기준모델

- 최빈값[0, 당뇨병 아님]의 비율
: 0.834531

- ✓ 학습될 분류모델은 정확도 측면에서
최소한 기준 모델의 비율인 **0.834531**
를 만족해야 한다.



1]. 분류문제에서의 기준모델 & 재현율

- 재현율

- ✓ 분류모델에서 타겟 데이터의 클래스가 불균형한 경우가 존재

- 정확도만으로 모델의 성능을 판단할 수 없다.

- 정밀도, 재현율, f1_score 등의 분류모델 평가지표 사용

- 당뇨병 판단모델에서는 실제 당뇨인데 당뇨가 아니라고 판단하는 것의 risk가 더 큼
 - 실제 당뇨인 사람들 중 당뇨라고 판단되는 비율인 **재현율** 사용

2]. 모델 최적화


- 당뇨병 분류모델의 최적화 방법

- ✓ RandomForestClassifier와 RandomizedSearchCV를 이용해서 하이퍼파라미터를 변화시켜가며 분류모델의 정확도와 재현율 등을 최적화 할 수 있다.

- 조절 하이퍼파라미터 : `max_depth`, `n_estimators`, `min_sample_split`

- 조절방식 : 초기 파라미터의 결과값에 따라 다시 초기 파라미터를 조절해가며 정확도와 재현율 등을 최적화 함

Trial1	max_depth	min_samples_split	n_estimators	accuracy
초기파라미터	randint(1, 50)	randint(1, 50)	randint(50, 500)	recall_test : 0.42
최적파라미터	24	10	455	Acc : 0.7932



Trial2	max_depth	min_samples_split	n_estimators	accuracy
초기파라미터	randint(20, 30)	randint(1, 20)	randint(400, 500)	recall_test : 0.42
최적파라미터	24	2	468	Acc : 0.7932

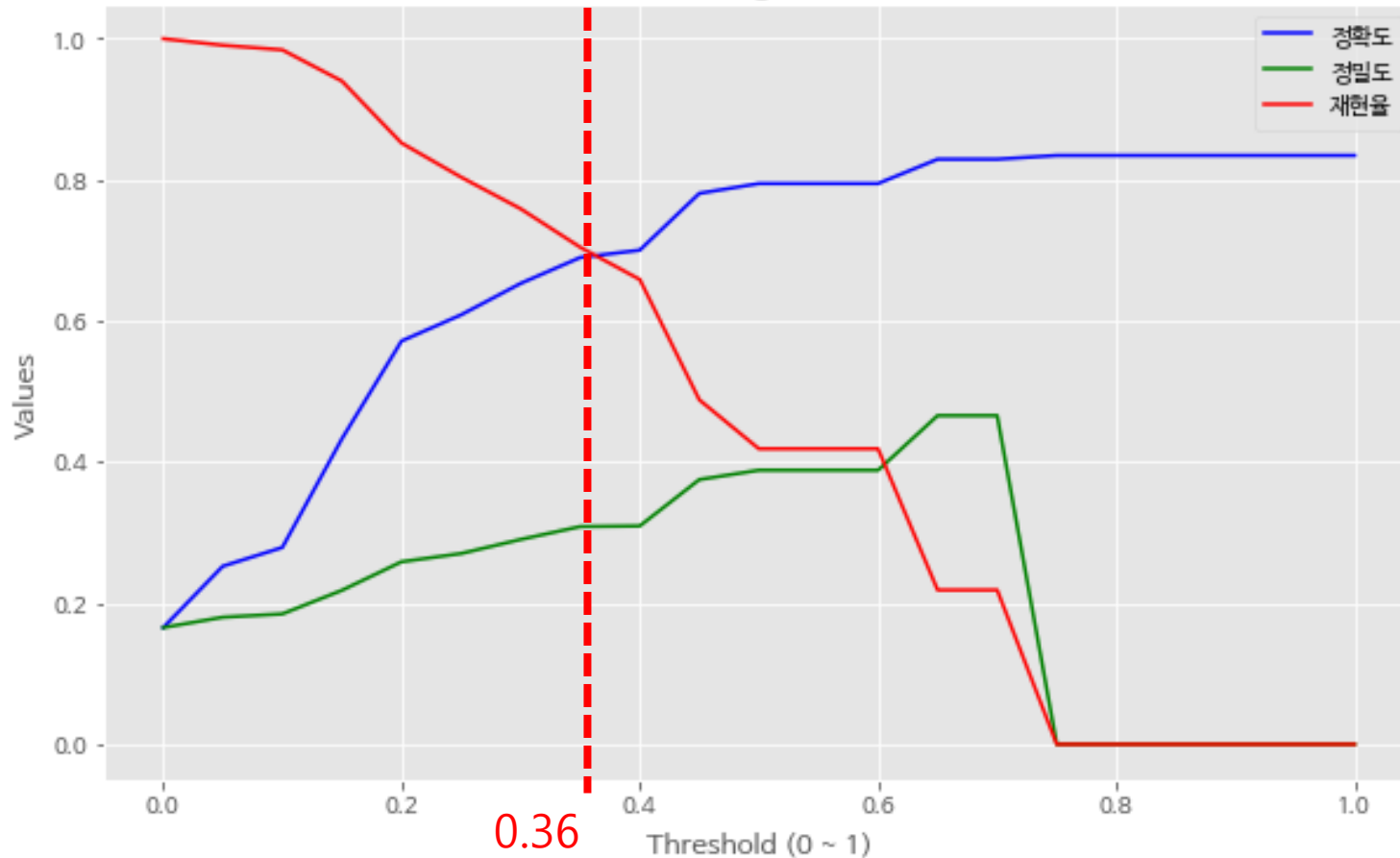
2]. 모델 최적화

- 분류모델의 최적화 결과(정확도의 최적화 실패)
 - ✓ 찾아낸 최적 하이퍼파라미터 : **max_depth = 24, n_estimators = 455, min_sample_split = 10**
 - 결과 : Training_acc : 0.7932, Testing_acc : 0.79(분류모델의 일반화는 성공), Testing_recall : 0.42
 - ✓하이퍼 파라미터들을 조절했음에도 불구하고, 정확도 측면에서 기준모델의 비율인 **0.834531 이상으로** testing 정확도를 끌어내진 못했다.
 - 설문조사 방식의 한계 존재
 - 더 많은 도매인지식 습득 >> 데이터 특성공학

2]. 모델 최적화

- 임계값 변화를 통한 최적화

임계값의 변화에 따른 Testing 정확도, 정밀도, 재현율 추이



- 임계값 = 0.36

- ✓ Testing 정확도 : 0.6925

- ✓ Testing 재현율 : 0.6924

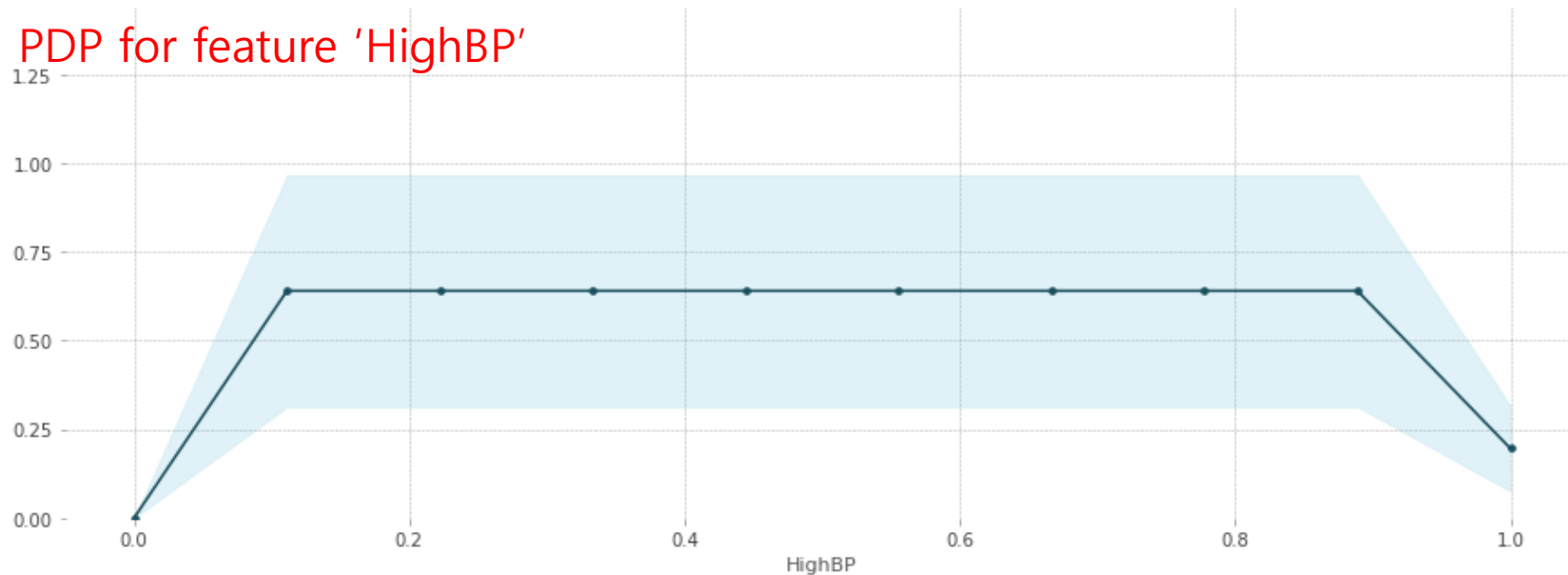
3]. 최종모델 설명 - 순열 중요도

- 순열 중요도 순위

- ✓ 순열중요도 : 타겟변수(당뇨병 여부) 에 영향을 많이 미치는 정도
- ✓ 아래 순열중요도 표를 통해, GenHlth, HighChol, HighBP 순으로 당뇨병 여부를 판단하는 데 영향을 많이 미친다는 것을 알 수 있다.

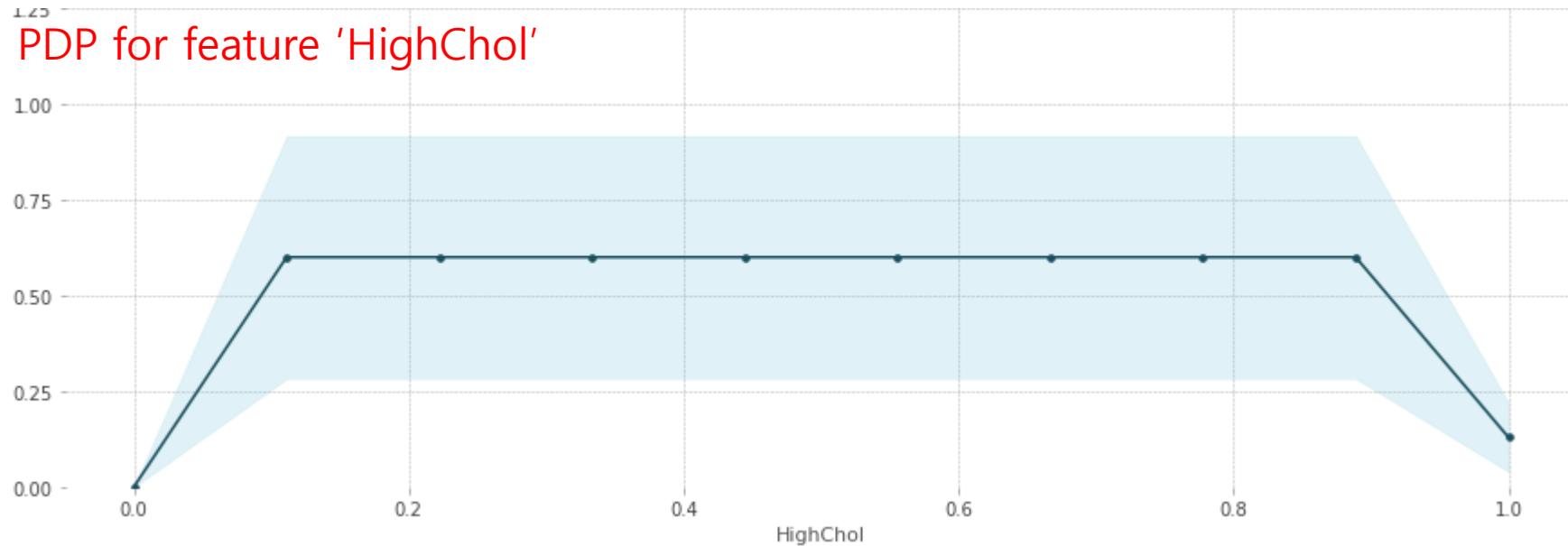
Weight	Feature
0.0079 ± 0.0018	GenHlth
0.0022 ± 0.0018	HighChol
-0.0009 ± 0.0012	HighBP

3]. 최종모델 설명 – PDP plot



- PDP plot (HighBP ~ Diabetes_01)
 - 특정 Feature의 값이 변할 때 타겟특성이 어떻게 변하는 지 알 수 있다.
 - 그래프 해석 :
 - ✓ HighBP[고혈압 여부]가 0에서 1로 변할 때, 당뇨일 확률을 약 0.2 정도 올려준다.

3]. 최종모델 설명 – PDP plot

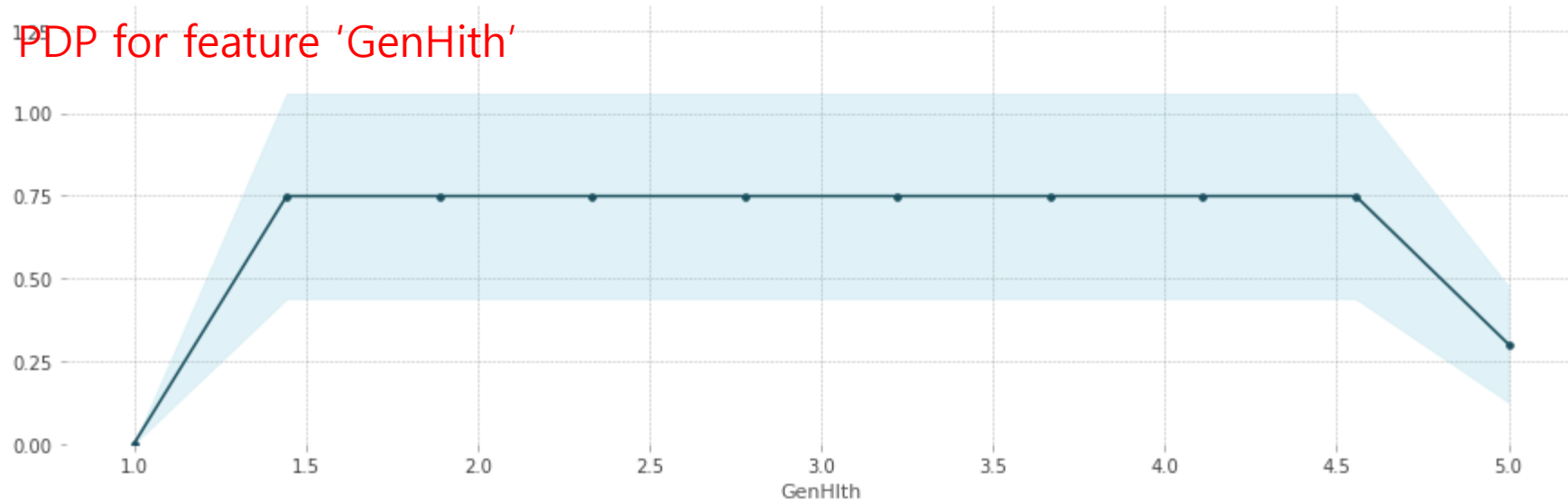


- PDP plot(HighChol ~ Diabetes_01)

- 그래프 해석 :

- ✓ HighChol[고콜레스테롤 여부]가 0에서 1로 변할 때, 당뇨일 확률을 약 0.13 정도 올려준다.

3]. 최종모델 설명 – PDP plot



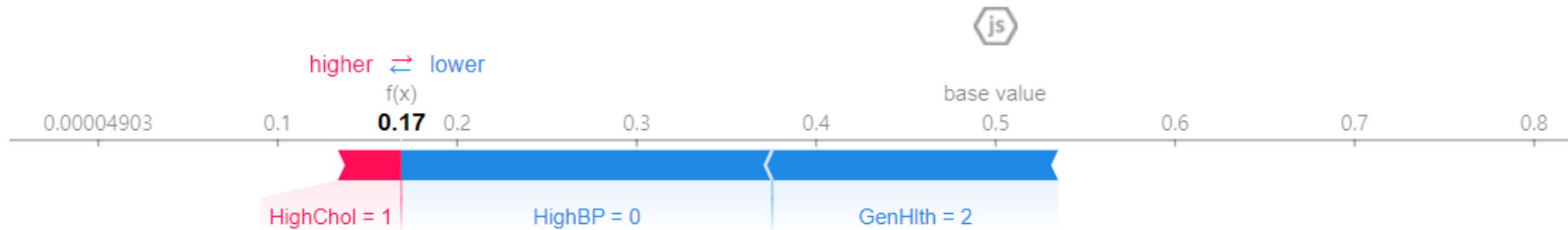
- PDP plot(HighChol ~ Diabetes_01)

- 그래프 해석 :

- ✓ GenHith[건강의 척도]가 0에서 5로 변할 때[건강이 나빠질 때], 당뇨병 확률을 약 0.3 정도 올려준다.

3]. 최종모델 설명 – Shap value

```
Force_plot(X_test_selected.iloc[[5400]])
```



- Shap Value의 Force plot

- 그래프 해석 :

- ✓ Test 샘플 중 하나에서 그 사람이 당뇨병 확률을 0.17이라고 예측했는데, 당뇨병이라고 예측하게 하는 확률을 높인 요소는 콜레스테롤 관련요소이고, 당뇨병확률을 낮게 한 요소는 고혈압[HighBP], 건강척도[GenHlth] 관련 요소이다.

4. 결론

- 당뇨병 진단모델의 학습 및 최적화에 있어.....
 - ✓ 일반화 성능은 괜찮았지만, 정확도를 기준모델 만큼 올리는데 실패
 - ✓ 대안으로 임계값을 0.36정도로 올려서 재현율을 올리는 것은 가능
 - ✓ 도메인 지식 + 순열중요도, PDP value, Shap Value 더 잘 해석
 - 당뇨병 진단모델을 더욱 최적화하자!!