CODESTATES_POJECTS2

생애전환기 건강진단

발표 : AI_15_김태수

프로젝트2 순서

소개

데이터
문제 정의
EDA
MODEL
Conclusion

CODESTATES PROJECTS|
2022

소개

"생애전환기 건강진단"이란 무엇인가?

만 40세와 만 66세 의료급여 수급권자를 대상

노화 및 생활습관 등으 원인으로 건강에 큰 변화가 생기는 시기이기도 하며,
만성질환 및 건강위험 요인을 조기에 발견하고 관리하기 위해서 지원하는 것

생애전환기 건강진단으로
1차 건강진단
이학적 검사 : 건강상태 등 혈압, 신체 검사, 진단의학검사(HDL)콜레스테롤 등 12 항목)
구강검진 : 치면 세균막 검사 등
암검사(위함, 유방암, 대장암, 간암, 자궁경부암)

“

데 이 터

”

데이터

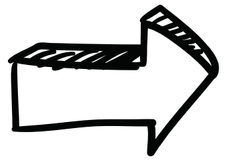
생애전환기 건강진단

건강검진정보란 국민건강보험의 직장가입자와 40세 이상의 피부양자, 세대주인 지역가입자와 40세 이상의 지역가입자의 일반건강검진 결과와 이들 일반건강검진 대상자 중에 만40세와 만66세에 도달한 이들이 받게 되는 생애전환기건강진단 수검이력이 있는 각 연도별 수진자 100만 명에 대한 기본정보(성, 연령대, 시도코드 등)와 검진내역(신장, 체중, 총콜레스테롤, 혈색소 등)으로 구성된 개방데이터입니다.

2020년도 진료 및 건강검진 수진 환자 100만명 무작위 추출

기준년도(HCHK_YEAR) : 해당 정보의 기준년도를 제공함
가입자 일련번호(IDV_ID) : 해당가입자의 부여한 일련번호
시도코드(SIDO) : 해당 수진자 거주지의 시도코드
성별코드(SEX) : 해당 정보 대상자의 성별을 제공함 - 1(남자), 2(여자)
연령대 코드(5단위)(AGE_GROUP) : 기준년도에 수진자의 나이를 5세 단위로 그룹화(범주화)하여 구분한 코드
신장(5cm단위)(HEIGHT) : 검진자의 키
체중(5kg단위)(WEIGHT) : 검진자의 몸무게
허리둘레(WAIST) : 검진자의 허리둘레
시력(좌)(SIGHT_LEFT) : 수검자의 좌측 눈의 시력
시력(우)(SIGHT_RIGHT) : 수검자의 우측 눈의 시력
청력(좌)(HEAR_LEFT) : 수검자의 좌측 귀의 청력
청력(우)(HEAR_RIGHT) : 수검자의 우측 귀의 청력
수축기혈압(BP_HIGH) : 검진자의 최고 혈압으로 심장이 수축해서 강한 힘으로 혈액을 동맥에 보낼 때의 혈관 내압
이완기혈압(BP_LWST) : 검진자의 최저 혈압으로 심장의 완기시의 혈압
식전혈당(공복혈당)(BLDS) : 검진자 식사 전 혈당(혈액 100ml당 함유 되어 있는 포도당의 농도) 수치
총 콜레스테롤(TOT_CHOLE) : 혈청 중의 에스텔형, 비에스테형(유리)콜레스테롤의 합
트리글리세라이드(TRIGLYCERIDE) : 단순지질 혹은 중성지질을 뜻함
HDL 콜레스테롤(HDL_CHOLE) : HDL(고밀도 리포단백질)에 포함되는 콜레스테롤
LDL 콜레스테롤(LDL_CHOLE) : LDL(저밀도 리포단백질)에 함유된 콜레스테롤
혈색소(HMG) : 혈액이나 혈구 속에 존재하는 색소단백으로 글로빈(globin)과 엠(heme)으로 구성되며 혈중의 산소운반체로서의 역할 수행
요단백(OLIG_PROTE_CD) : 소변에 단백질이 섞여 나오는 것
혈청크레아티닌(CREATININE) : 크레아티닌은 크레아틴의 탈수물로 내인성 단백대사의 종말산물로서 신장에서 배설되고 그 증감은 음식물에 관계없이 근육의 발육과 운동에 관계함
(혈청지오티)AST(SGOT_AST) : 간 기능을 나타내는 혈액검사상의 수치, 간세포 이외에 신장, 뇌, 근육 등에도 존재하는 효소로 이러한 세포들이 손상을 받는 경우 농도가 증가함
(혈청지오티)ALT(SGPT_ALT) : 간 기능을 나타내는 혈액검사상의 수치, ALT는 주로 간세포 안에 존재하는 효소로, 간세포가 손상을 받는 경우 농도가 증가함
감마 지티피(GAMMA_GTP) : 간 기능을 나타내는 혈액검사상의 수치, 간 낸의 쓸개관(담관)에 존재하는 효소로 글루타민산을 외부에 펩티드나 아미노산 등으로 옮기는 작용을 함. 쓸게즙(담즙) 배설 장애, 간세포 장애 발생 시 혈중이 증가하게 됨.
흡연상태(SMK_STAT_TYPE_CD) : 해당 수검자의 흡연 상태 여부
음주여부(DRK_YN) : 해당 수검자의 음주 상태 여부
구강검진 수검여부(HCHK_OE_INSPEC_YN) : 해당 검진자가 구강검진을 선택하여 검진하였는지 여부에 대한 항목
치아우식증유무(CRS_YN) : 해당 수검자의 치아우식증 유무에 대한 항목
치석(TTR_YN) : 해당 수검자의 치석 여부
데이터 공개일자(DATA_STD__DT) : 데이터 작성 기준일자

	dtypes	null_count	num_unique_values
HCHK_YEAR	int64	0	1
IDV_ID	int64	0	1000000
SIDO	int64	0	17
SEX	int64	0	2
AGE_GROUP	int64	0	10
HEIGHT	int64	0	14
WEIGHT	int64	0	22
WAIST	float64	108	777
SIGHT_LEFT	float64	257	24
SIGHT_RIGHT	float64	252	24
HEAR_LEFT	float64	222	3
HEAR_RIGHT	float64	230	3
BP_HIGH	float64	7532	174
BP_LWST	float64	7534	126
BLDS	float64	7602	491
TOT_CHOLE	float64	597694	429
TRIGLYCERIDE	float64	597678	1329
HDL_CHOLE	float64	597685	340
LDL_CHOLE	float64	605529	362
HMG	float64	7611	196
OLIG_PROTE_CD	float64	12141	6
CREATININE	float64	7602	188
SGOT_AST	float64	7601	580
SGPT_ALT	float64	7602	637
GAMMA_GTP	float64	7603	971
SMK_STAT_TYPE_CD	float64	343	3
DRK_YN	float64	196	2
HCHK_OE_INSPEC_YN	int64	0	2
CRS_YN	float64	668617	2

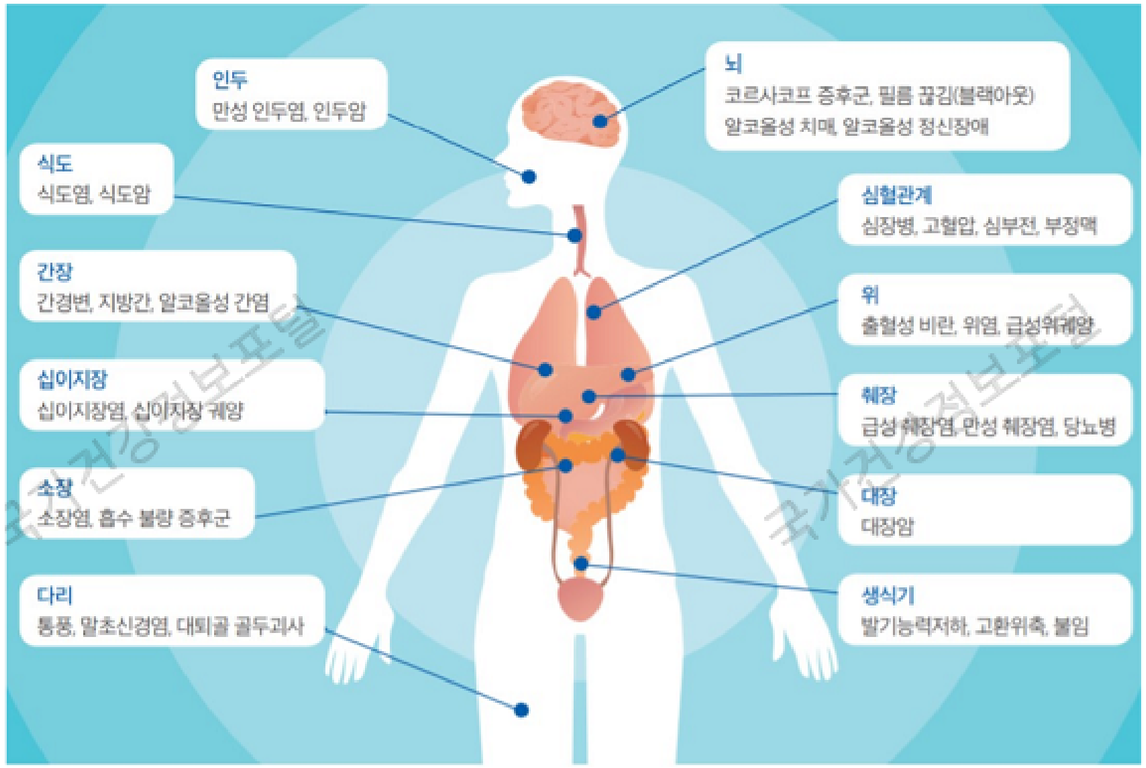


	count	mean	std	min	25%	50%	75%	max
SEX	103662.0	1.472960	0.499271	1.00	1.0	1.0	2.0	2.0
AGE_GROUP	103662.0	10.910758	1.579249	9.00	9.0	11.0	13.0	13.0
BP_HIGH	103662.0	122.032760	13.935062	70.00	112.0	121.0	131.0	260.0
BP_LWST	103662.0	76.323995	9.818133	38.00	70.0	76.0	82.0	144.0
BLDS	103662.0	101.987151	22.262973	50.00	91.0	97.0	106.0	495.0
TOT_CHOLE	103662.0	202.651502	39.499941	54.00	176.0	201.0	228.0	517.0
TRIGLYCERIDE	103662.0	126.183413	70.412024	5.00	75.0	108.0	159.0	399.0
HDL_CHOLE	103662.0	56.503723	14.349932	2.00	46.0	55.0	65.0	545.0
LDL_CHOLE	103662.0	121.234686	36.658592	1.00	96.0	120.0	144.0	1512.0
HMG	103662.0	14.331649	1.526363	4.20	13.3	14.4	15.4	21.9
OLIG_PROTE_CD	103662.0	1.104146	0.451095	1.00	1.0	1.0	1.0	6.0
CREATININE	103662.0	0.855736	0.279579	0.06	0.7	0.8	1.0	24.0
SGOT_AST	103662.0	26.924089	17.002347	1.00	20.0	24.0	29.0	1606.0
SGPT_ALT	103662.0	26.838758	22.223538	1.00	16.0	21.0	31.0	1837.0
GAMMA_GTP	103662.0	36.930611	45.841983	1.00	16.0	24.0	40.0	1731.0
DRK_YN	103662.0	0.672030	0.469476	0.00	0.0	1.0	1.0	1.0
TTR_YN	103662.0	0.608169	0.585356	0.00	0.0	1.0	1.0	2.0

Target

문제 정의

문제 정의



음주가 건강에 미치는 영향

1. 술은 1군 발암물질
2. 우리 몸 전신에 작용하고 200여 종의 질병과 관련
3. 뇌를 손상시키고 중독
4. 고혈압, 부정맥 등 혈관성 질환을 야기
5. 지방간, 간염, 간경화, 간암 등 각종 간 질환 야기
6. 면역체계 파괴

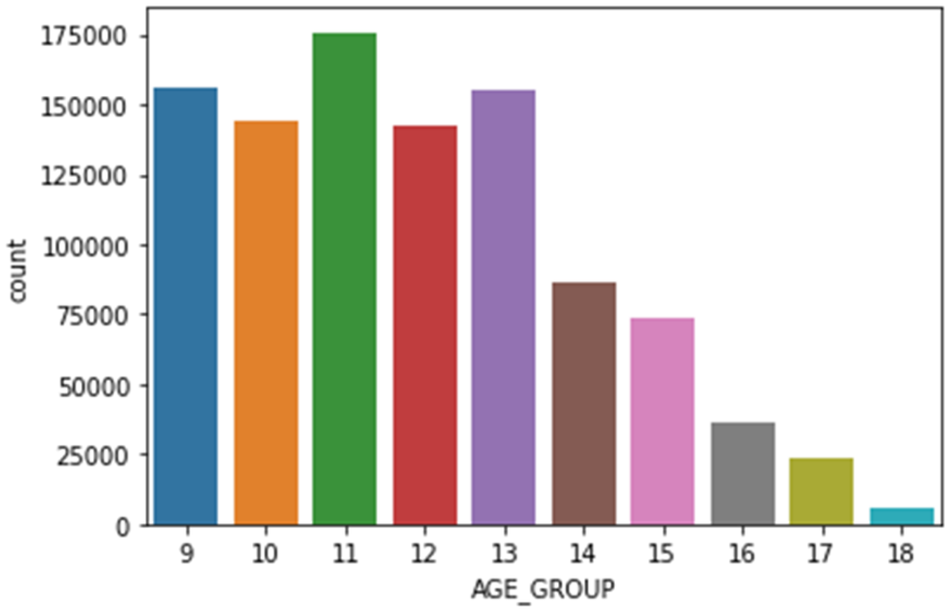
문제 정의

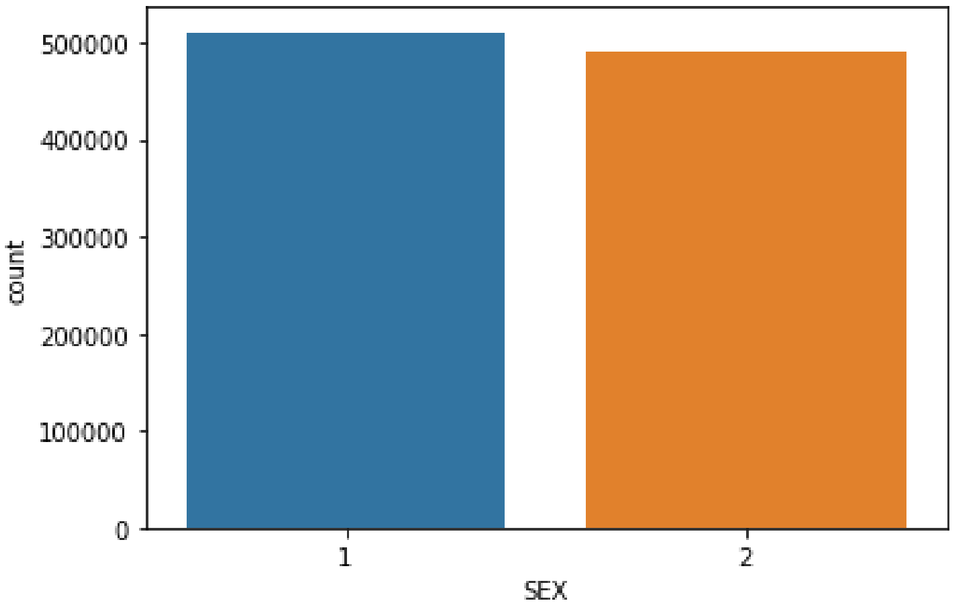
가설 설정

1. 이학적정보를 통해 음주여부를 예상 할 수 있나?

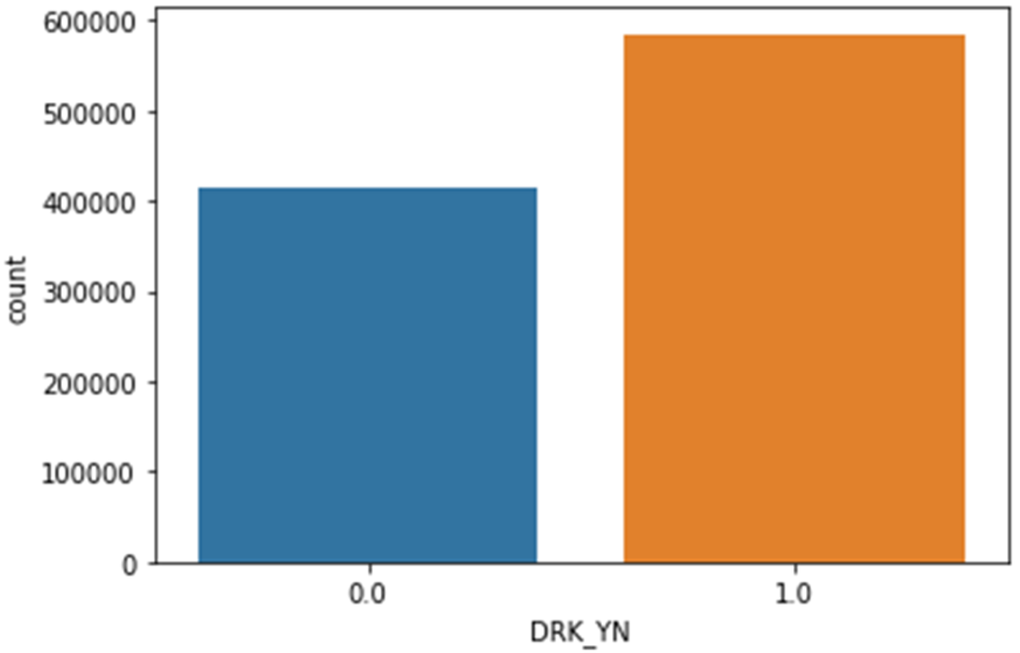
EDA

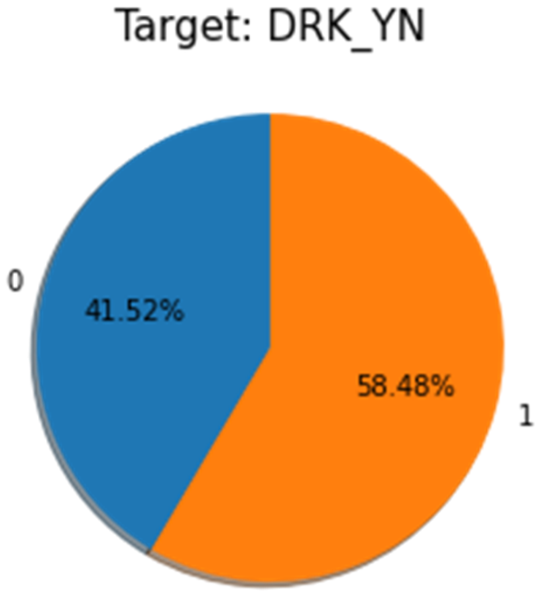
EDA


175000
150000
125000
100000
75000
50000
25000
0
count
9
10
11
12
13
14
15
16
17
18
AGE_GROUP

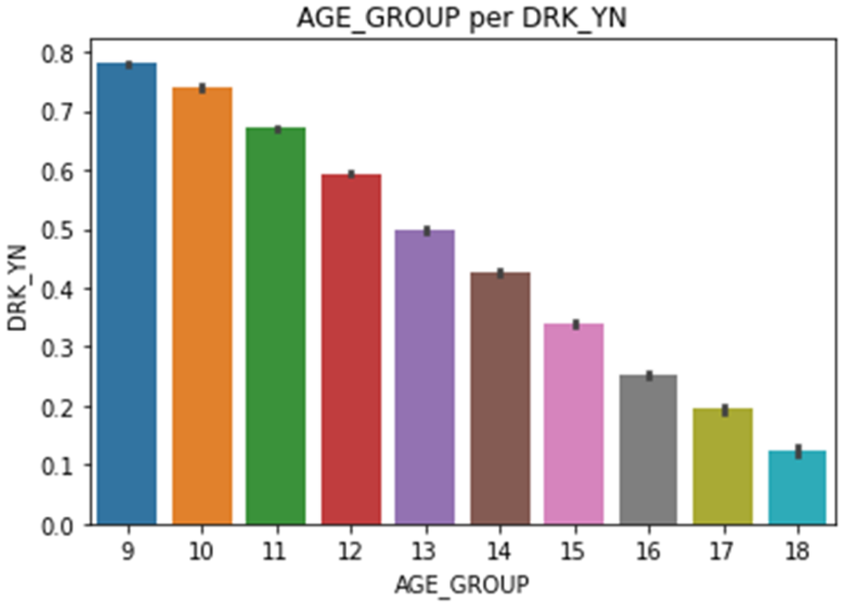

500000
400000
300000
200000
100000
0
count
1
2
SEX

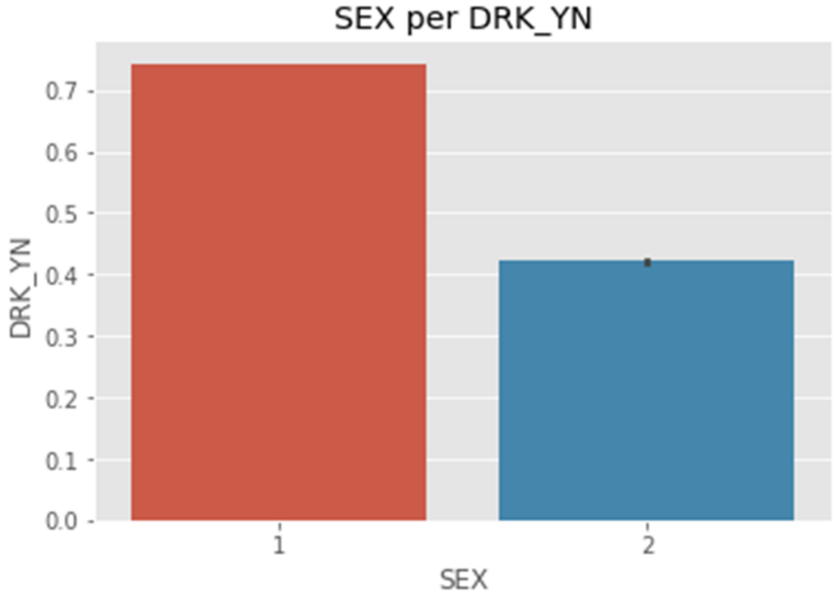
EDA


600000
500000
400000
300000
200000
100000
0
count
0.0
1.0
DRK_YN

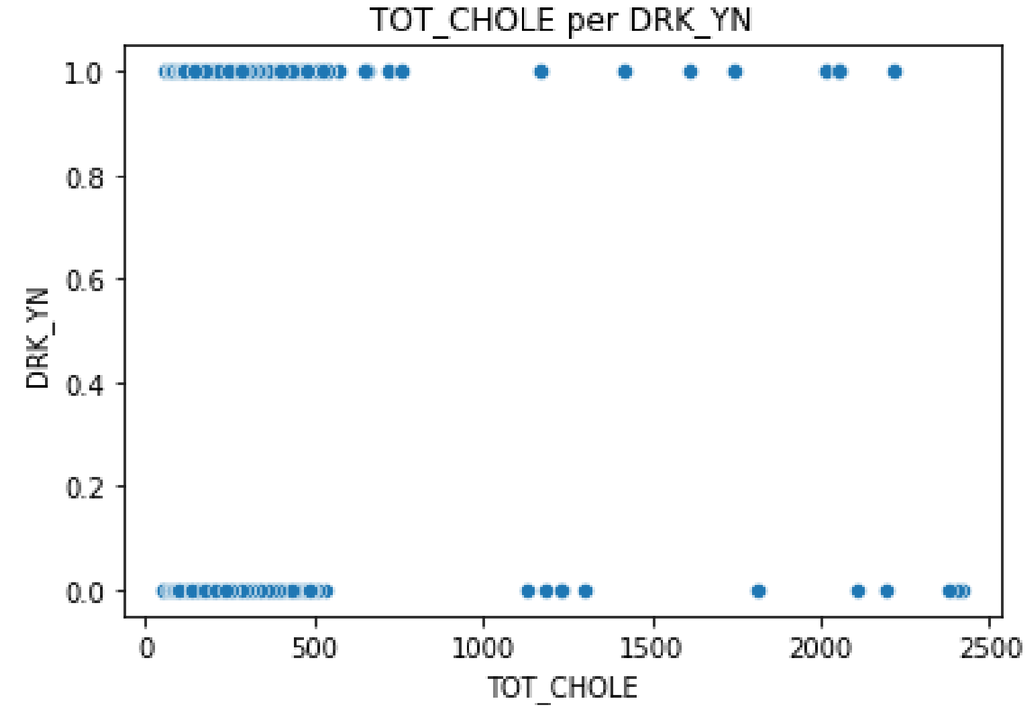

Target: DRK_YN
0
41.52%
58.48%
1

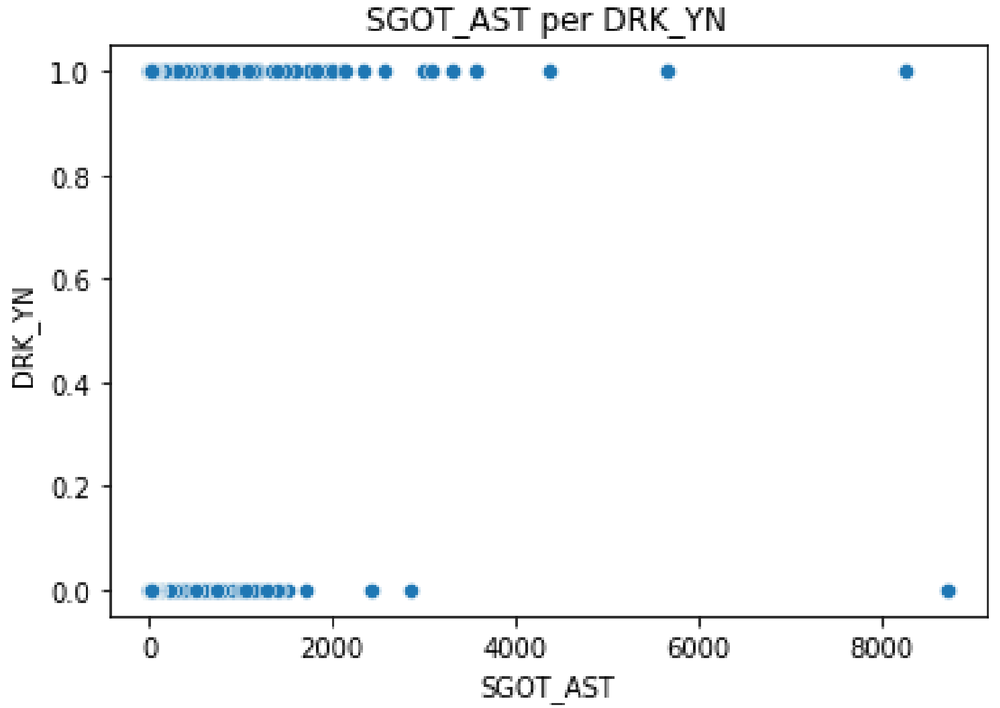
EDA


AGE_GROUP per DRK_YN
DRK_YN
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
9
10
11
12
13
14
15
16
17
18
AGE_GROUP

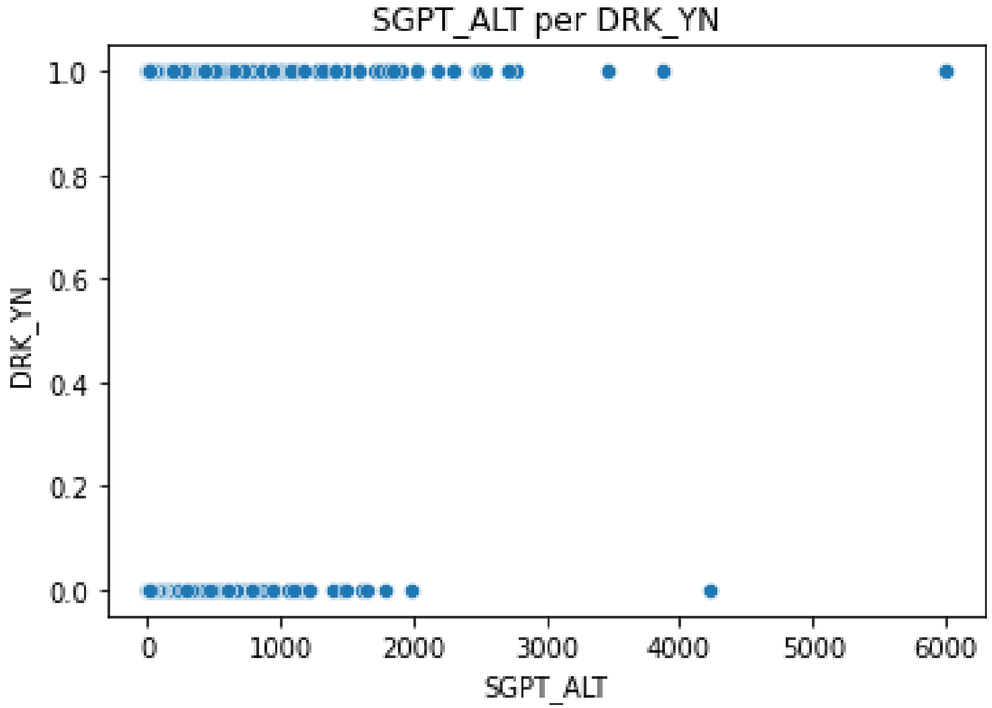

SEX per DRK_YN
DRK_YN
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
SEX

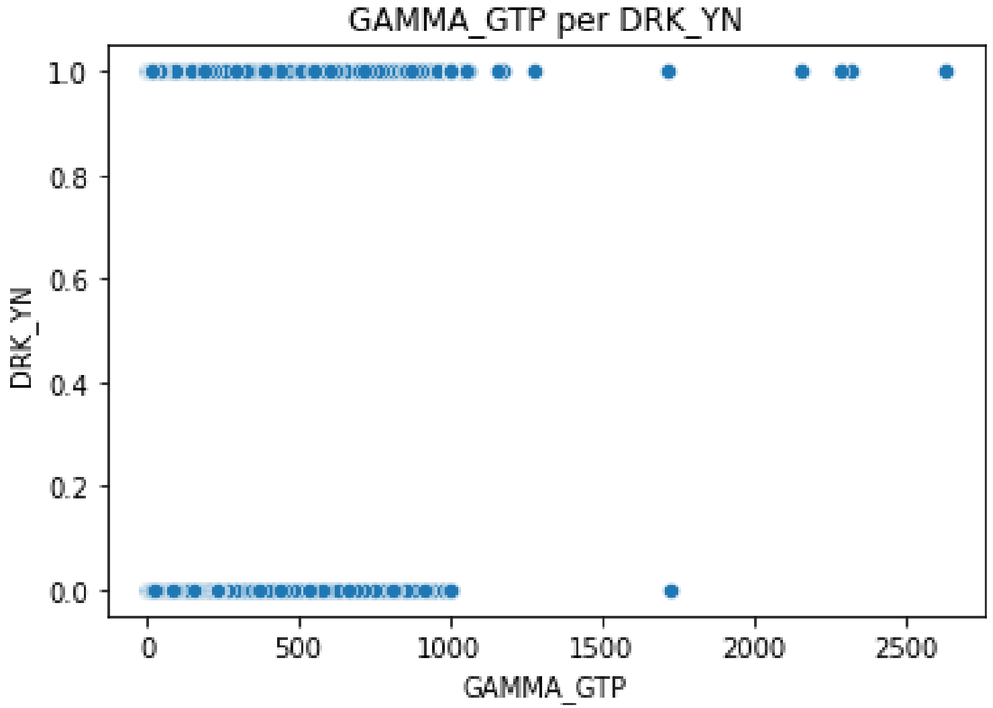
EDA


TOT_CHOLE per DRK_YN
DRK_YN
TOT_CHOLE
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500


SGOT_AST per DRK_YN
DRK_YN
SGOT_AST
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000

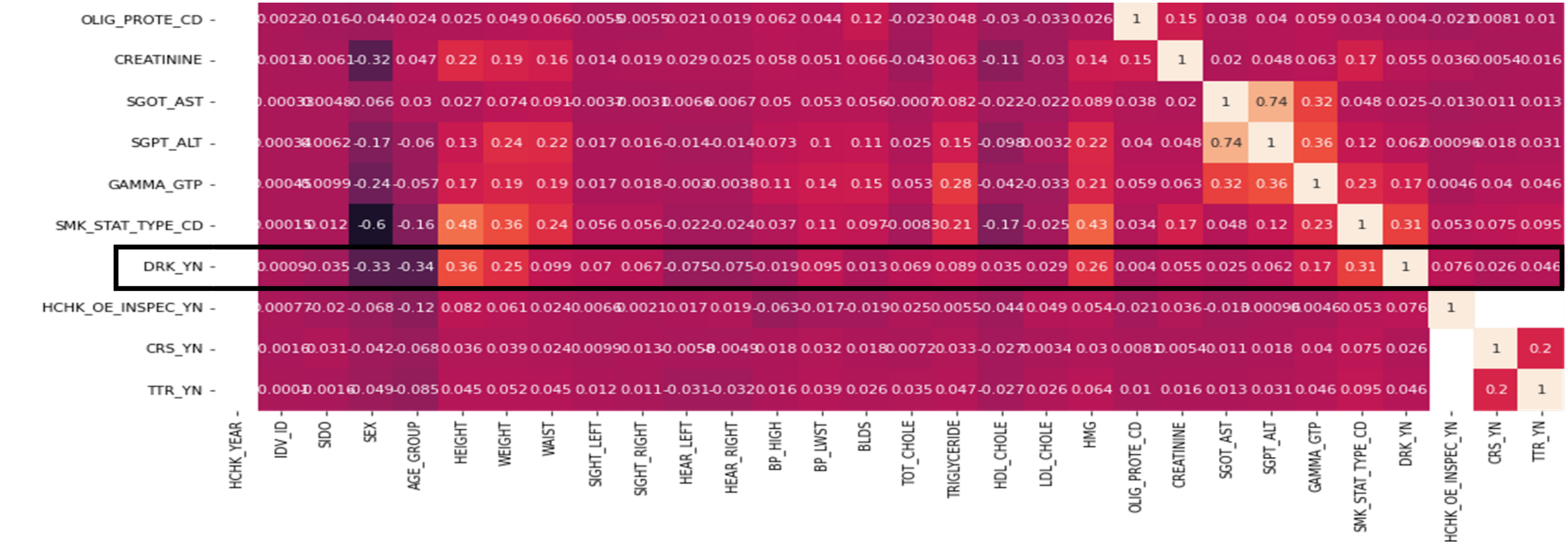
EDA


SGPT_ALT per DRK_YN
DRK_YN
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
5000
6000
SGPT_ALT


GAMMA_GTP per DRK_YN
DRK_YN
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
GAMMA_GTP


HCHK_YEAR
IDV_ID
SIDO
SEX
AGE_GROUP
HEIGHT
WEIGHT
WAIST
SIGHT_LEFT
SIGHT_RIGHT
HEAR_LEFT
HEAR_RIGHT
BP_HIGH
BP_LWST
BLDS
TOT_CHOLE
TRIGLYCERIDE
HDL_CHOLE
LDL_CHOLE
HMG
OLIG_PROTE_CD
CREATININE
SGOT_AST
SGPT_ALT
GAMMA_GTP
SMK_STAT_TYPE_CD
DRK_YN
HCHK_OE_INSPEC_YN
CRS_YN
TTR_YN

EDA


OLIG_PROTE_CD
CREATININE
SGOT_AST
SGPT_ALT
GAMMA_GTP
SMK_STAT_TYPE_CD
DRK_YN
HCHK_OE_INSPEC_YN
CRS_YN
TTR_YN
HCHK_YEAR
IDV_ID
SIDO
SEX
AGE_GROUP
HEIGHT
WEIGHT
WAIST
SIGHT_LEFT
SIGHT_RIGHT
HEAR_LEFT
HEAR_RIGHT
BP_HIGH
BP_LWST
BLDS
TOT_CHOLE
TRIGLYCERIDE
HDL_CHOLE
LDL_CHOLE
HMG
OLIG_PROTE_CD
CREATININE
SGOT_AST
SGPT_ALT
GAMMA_GTP
SMK_STAT_TYPE_CD
DRK_YN
HCHK_OE_INSPEC_YN
CRS_YN
TTR_YN

“

모델

”

모 델

랜덤포레스트 모델링

Baseline

Training accuracy : 0.67

	precision	recall	f1-score	support
0.0	0.64	0.40	0.49	6861
1.0	0.75	0.89	0.81	13872
accuracy			0.73	20733
macro avg	0.69	0.64	0.65	20733
weighted avg	0.71	0.73	0.71	20733

RandomForest

훈련 정확도 : 1.0
검증 정확도 : 0.7251296273965996

Get Accuracy Score on Test Set

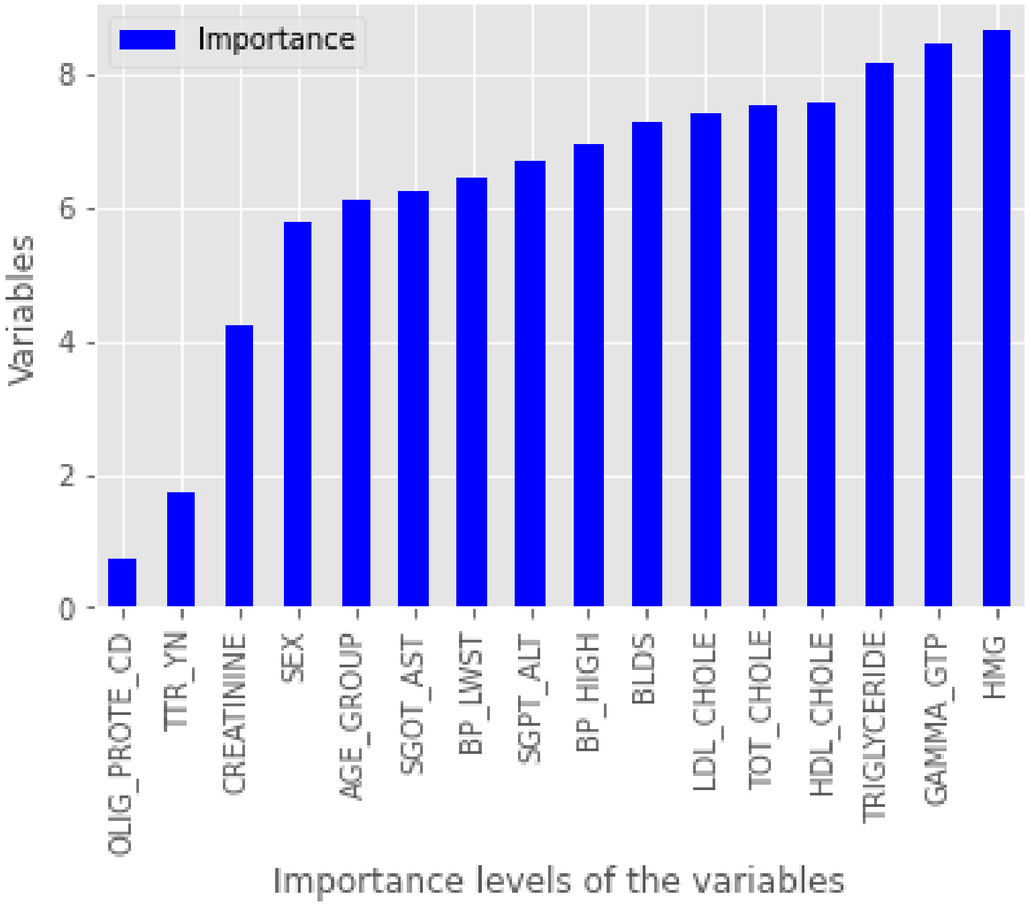
일반화 성능 : 0.7261370761587806

RandomForest(max_depth = 20)

훈련 정확도 : 0.972039250561476
검증 정확도 : 0.724406125648137

모 델

랜덤포레스트 모델링



모 델

XGBoost -> Tunning

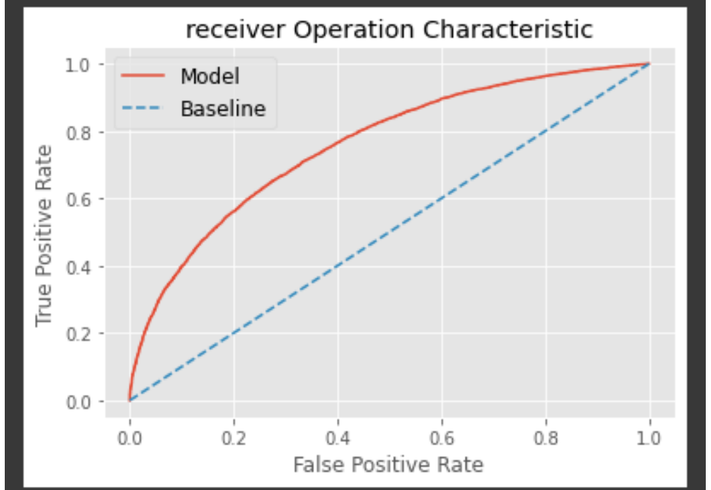
검증 정확도 0.7244061256481 37
일반화 성능:

	precision	recall	f1-score	support
0.0	0.63	0.41	0.49	6861
1.0	0.75	0.88	0.81	13872
accuracy			0.72	20733
macro avg	0.69	0.64	0.65	20733
weighted avg	0.71	0.72	0.71	20733

튜닝

일반화 성능:

	precision	recall	f1-score	support
0.0	0.65	0.40	0.50	6861
1.0	0.75	0.89	0.82	13872
accuracy			0.73	20733
macro avg	0.70	0.65	0.66	20733
weighted avg	0.72	0.73	0.71	20733



Get Accuracy Score on Test Set

일반화 성능 : 0.7588637455204996

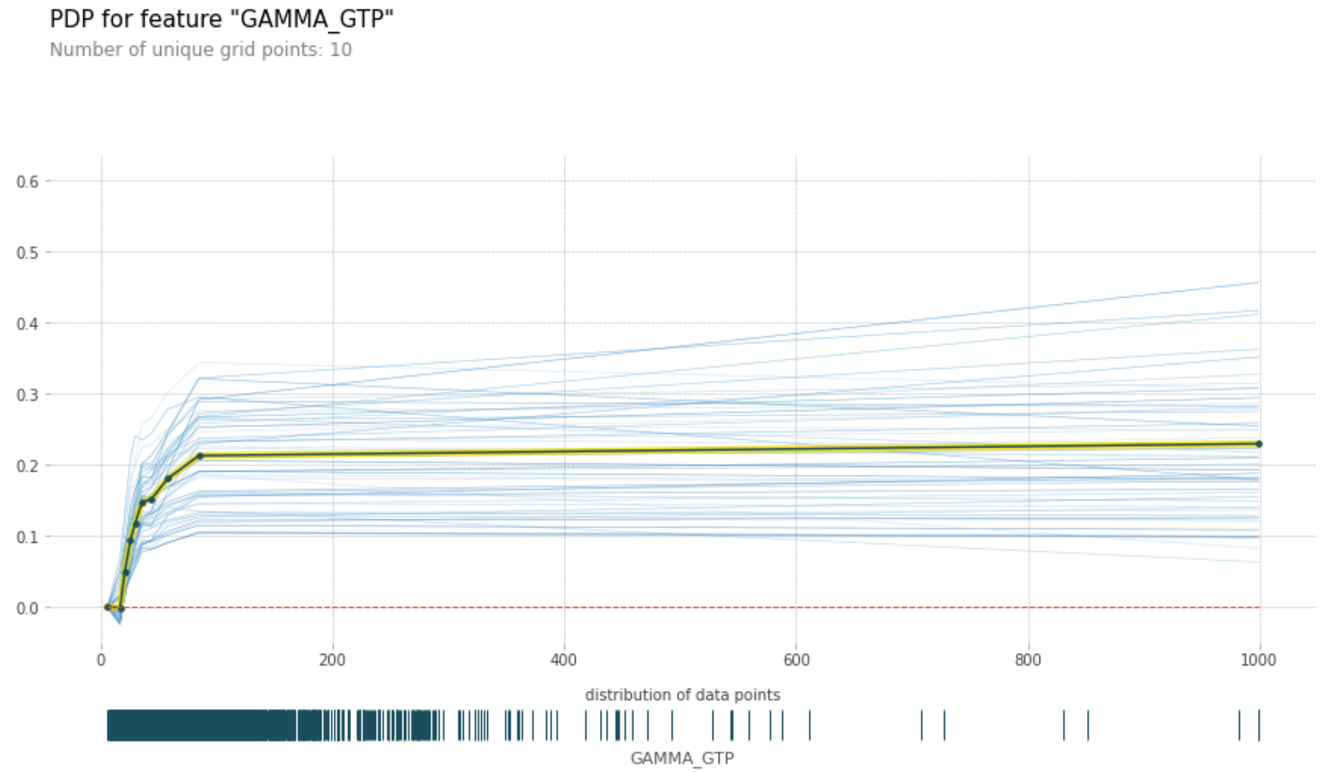
모 델

XGBoost - Feature Important

Features	Feature Importance
GAMMA_GTP	0.10344672217033471
AGE_GROUP	0.04737434117802075
SGPT_ALT	0.044500981145426464
HDL_CHOLE	0.03644419660125431
SGOT_AST	0.019822663519400808
TRIGLYCERIDE	0.006118629236184447
LDL_CHOLE	0.002929410634419116
HMG	0.0025075930372850273
CREATININE	0.002077250897307725
BP_LWST	0.0020278969129805978

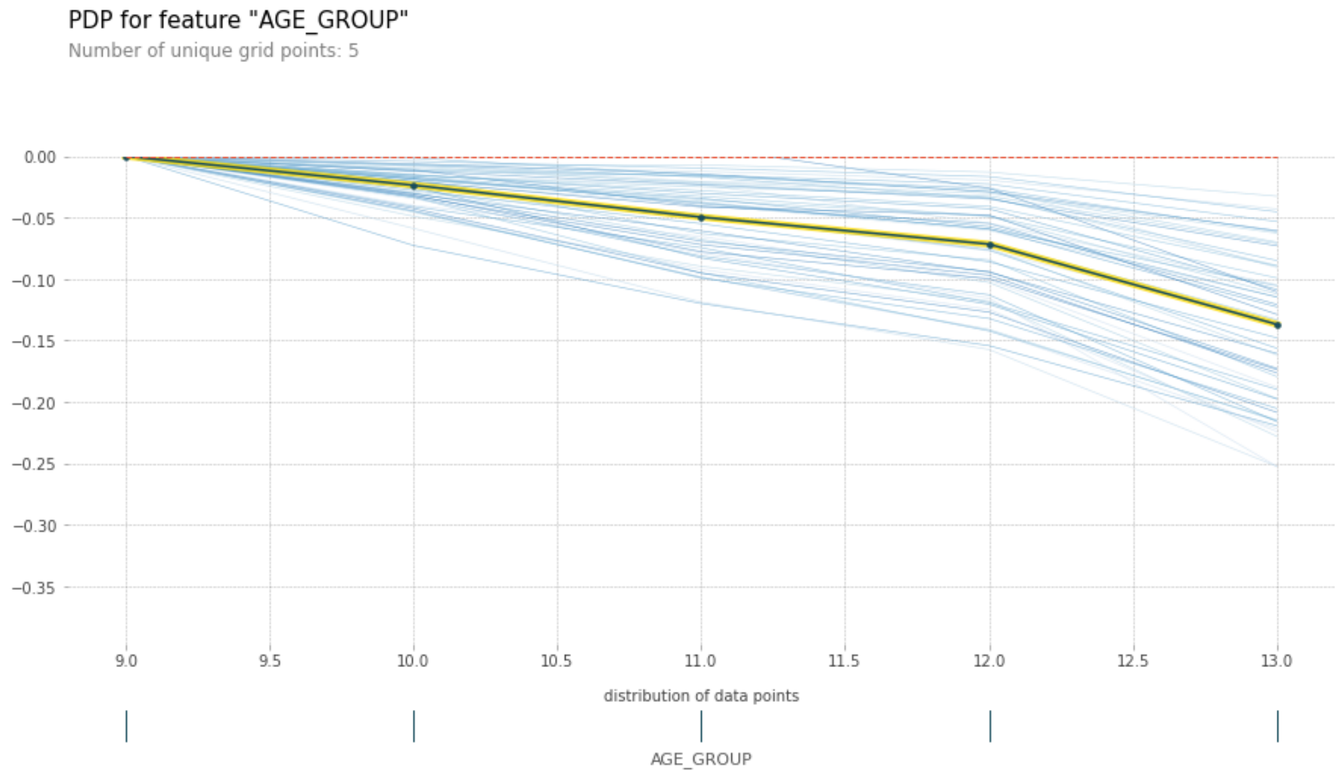
모 델

PDP Plot


PDP for feature "GAMMA_GTP"
Number of unique grid points: 10
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
200
400
600
800
1000
distribution of data points
GAMMA_GTP

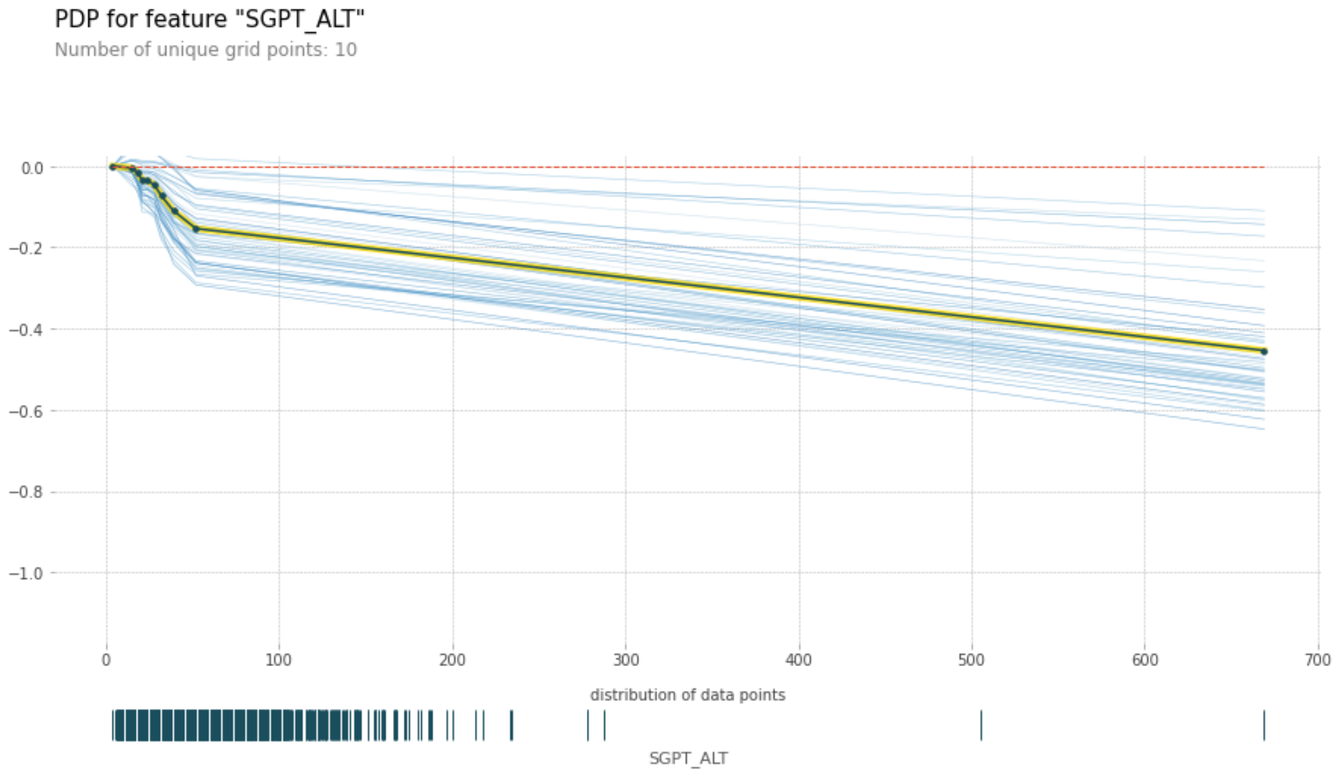
모 델

PDP Plot


PDP for feature "AGE_GROUP"
Number of unique grid points: 5
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
−0.30
−0.35
9.0
9.5
10.0
10.5
11.0
11.5
12.0
12.5
13.0
distribution of data points
AGE_GROUP

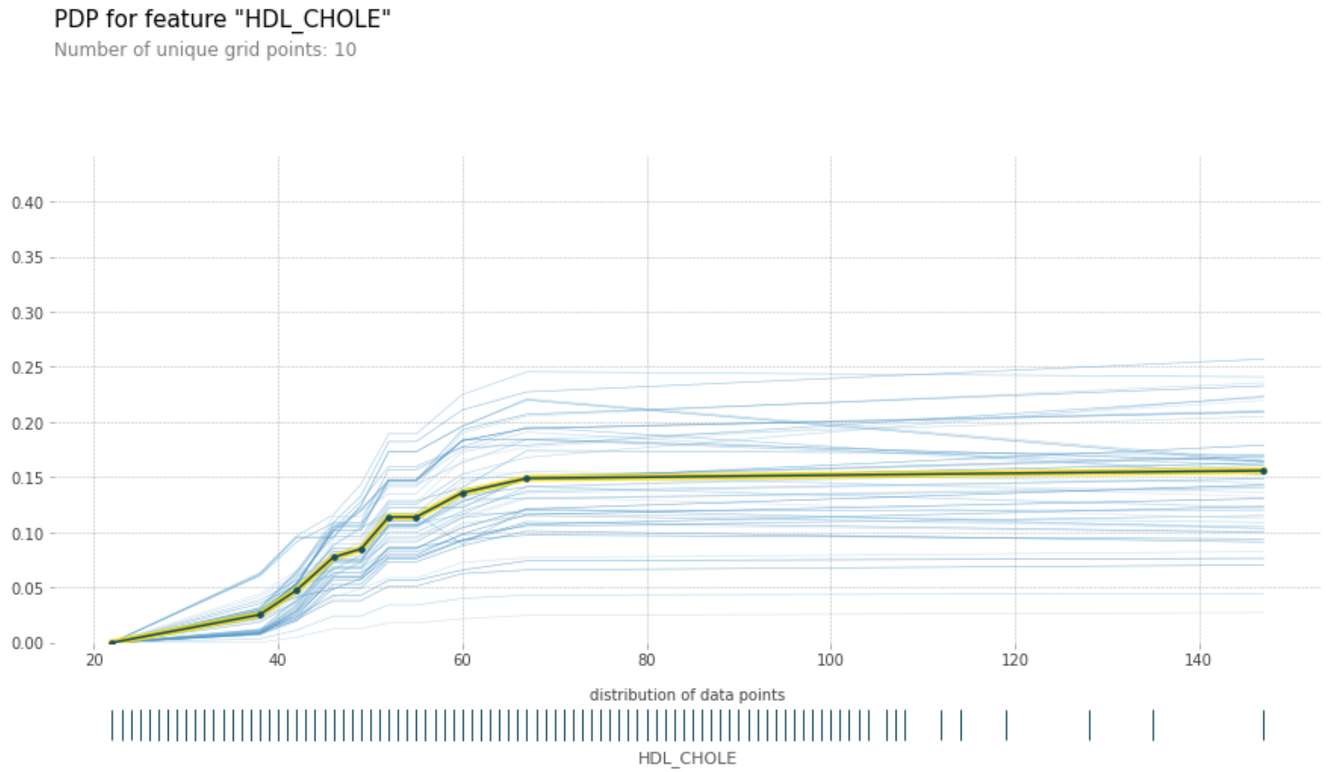
모 델

PDP Plot


PDP for feature "SGPT_ALT"
Number of unique grid points: 10
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
0
100
200
300
400
500
600
700
distribution of data points
SGPT_ALT

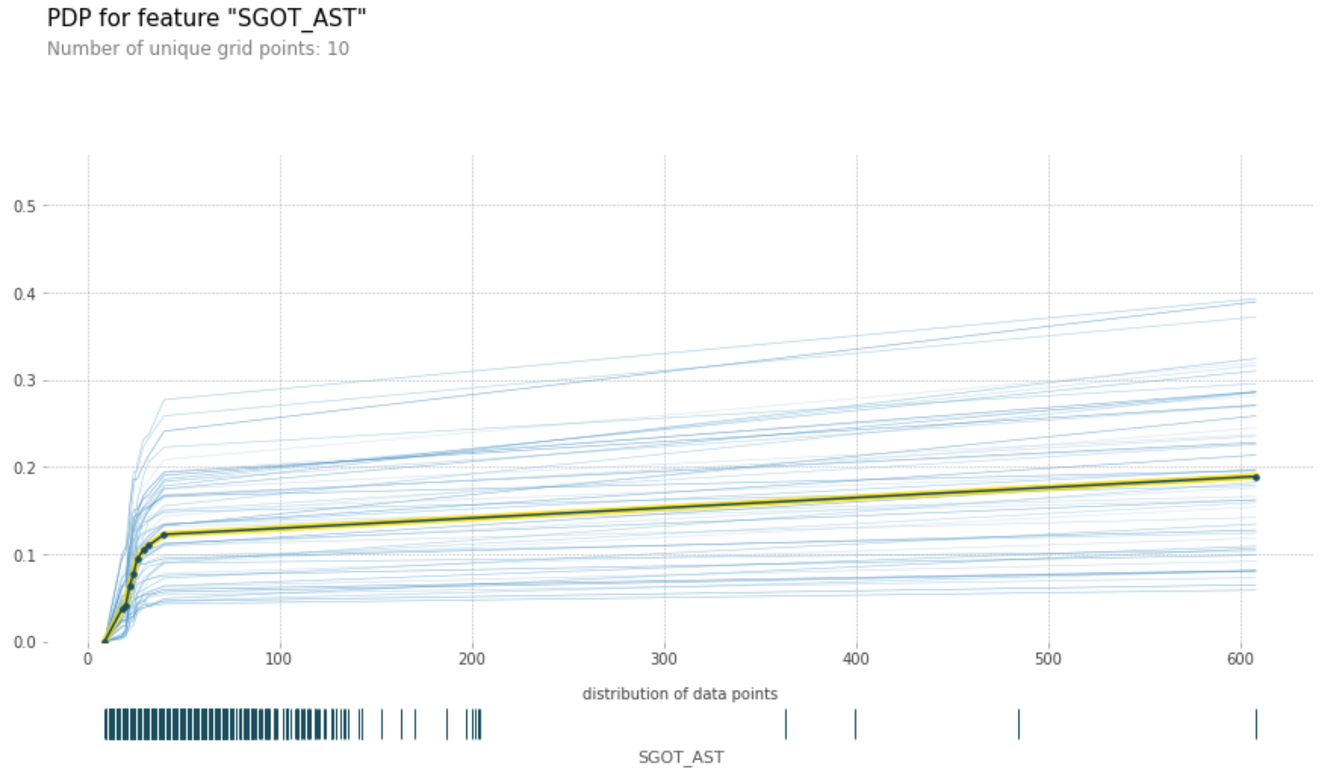
모 델

PDP Plot


PDP for feature "HDL_CHOLE"
Number of unique grid points: 10
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
20
40
60
80
100
120
140
distribution of data points
HDL_CHOLE

모 델

PDP Plot


PDP for feature "SGOT_AST"
Number of unique grid points: 10
0.5
0.4
0.3
0.2
0.1
0.0
0
100
200
300
400
500
600
distribution of data points
SGOT_AST

결론

CONCLUSION

생애전환기건강검진 데이터를 통해 음주여부를 예상 할 수 있나?

생애전환기건강검진 데이터를 머신러닝으로 훈련하여 음주여부를 예측하는 모델을 만들었습니다.
어느정도의 예측성능이 나왔고, 2차 검진 데이터를 추가 한다면 더 성능 좋은 머신러닝 모델을 만들수 있을 것입니다.