건강 검진 결과를 토대로 흡연 여부 예측하기

팀 : 와글바글

팀원 : 박영원(대표), 서문세완, 홍수호

KU KONKUK UNIVERSITY

* 본 보고서 파일에는 코드를 첨부하지 않았으며,
  따라서 자세한 작업 과정은 담지 않았습니다.
  작업 과정은 함께 첨부된 .ipynb 파일을 참조해주시면 감사하겠습니다.

* 본 보고서에는 프로젝트 도중 각 절차(전처리, 머신러닝 기법 선택)에서
  왜 이러한 선택을 하였는지 구구절절하게 풀어냈습니다.
  따라서 글의 분량이 다소 많은 점 양해 부탁드립니다.

주제 선정 이유

흡연은 인체에 해롭습니다.

수많은 연구와 논문들이 이 사실을 뒷받침하고 있고,
매년 새로운 공익 광고와 금연 캠페인들이 쏟아지는 덕분에
이 사실을 모르는 사람은 아마 없을 겁니다.

하지만 그럼에도 여전히 많은 사람들이 흡연을 하고 있습니다.

그런데 흡연이 그렇게나 위험하다면, 왜 우리는 우리 주변에서 흡연 때문에
일상생활에 지장이 생겼거나, 큰 병을 얻었다는 사람을 볼 수 없는 걸까요?

혹시 겉으로는 드러나지 않더라도,
내 몸의 건강 성적표인 “건강검진 결과”에서는 흡연의 흔적이 나타나지 않을까요?

지금부터 알아보도록 하겠습니다.

프로젝트의 목표

프로젝트의 목표는 머신러닝 기법을 활용하여,
건강검진 결과 데이터를 토대로 흡연 여부 예측 모델을 만들어내는 것입니다.

목표로 하는 모델의 정확도 마지노선을 정해 놓은 것은 아니지만,
이왕이면 프로젝트를 진행하는 만큼 높은 성능의 모델이 만들어졌으면
하는 바람도 있습니다.

모델을 만든 후에는 건강검진 데이터의 어떤 지표(피처)가 흡연 여부를
판단하는데 큰 역할을 했는지 알아볼 계획입니다.

데이터 개요

제공된 데이터셋은 대한민국 국민건강보험공단에서 실시하는 기초건강검진 결과 데이터로, 2009년부터 2020년까지 연간 1만명을 무작위로 샘플링한 데이터입니다.
(통계 최강자전 안내 페이지)

총 120000행 31열로 구성되어 있으며,
포함된 피처(열)들은 '연도', 'ID', '성별', '지역' 등과 같은 인적 사항과, '콜레스테롤', '혈압', '간수치' 등과 같은 건강검진의 결과로 나눌 수 있습니다.

이번 프로젝트의 목적은 '건강검진의 결과'를 토대로 흡연 여부를 예측하는 것이므로,
목적에 적합한 방향으로 전처리를 진행하도록 하겠습니다.

피처 소개

건강검진의 결과

HEIGHT	키
WEIGHT	몸무게
WAIST	허리둘레
SIGHT_LEFT	왼쪽 눈 시력
SIGHT_RIGHT	오른쪽 눈 시력
HEAR_LEFT	왼쪽 청력 이상
HEAR_RIGHT	오른쪽 청력 이상
BP_HIGH	수축기 혈압
BP_LWST	이완기 혈압
BLDS	공복혈당
TOT_CHOLE	총 콜레스테롤
HDL_CHOLE	HDL(콜레스테롤)
LDL_CHOLE	LDL(콜레스테롤)
TRIGLYCERIDE	중성지방
HMG	혈색소
OLIG_PROTE_CD	요단백 수치
CREATININE	크레아티닌
SGOT_AST	혈중 GOT (간 관련)
SGPT_ALT	혈중 GPT (간 관련)
GAMMA_GTP	감마 지티피 (간 관련)

피처 소개

인적 사항 및 그 외

Unnamed: 0	데이터셋을 불러오는 과정에서 생긴 더미 열
YEAR	연도
IDV_ID	식별 번호
SEX	성별
AGE_GROUP	연령 그룹
AREA_CODE	지역 코드
SMK_STAT	흡연 여부
DRK_YN	음주 여부
HCHK_CE_IN	구강 검진 여부
CRS_YN	치아 우식증 여부
TTR_YN	치석 여부

데이터 전처리

1. 타깃 변수(흡연 여부)가 결측값인 행 삭제

주어진 데이터셋에는 피검자의 흡연 여부를 나타내는 피처가 있습니다.
(SMK_STAT)

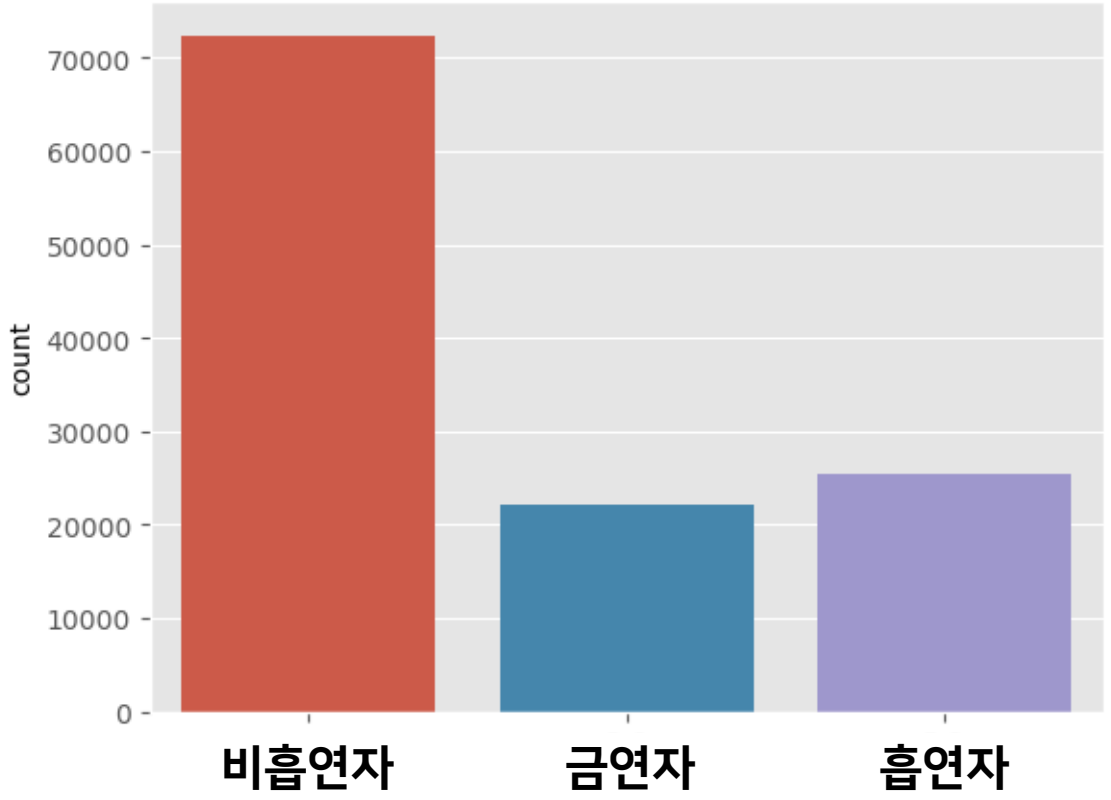
따라서 저희는 이 피처를 타깃 변수로 삼고,
이를 예측하기 위한 모델의 훈련과 테스트를 진행할 예정입니다.

그런데 타깃 변수가 결측되어 있는 행은 이러한 훈련과 테스트가 불가능하므로,
데이터셋에서 잘라내도록 하겠습니다.

흡연 여부(SMK_STAT)가 결측값인 행 삭제

데이터 전처리

2. 타깃 변수(흡연 여부) 확인



'흡연 여부' 의 분포를 확인한 결과,

비흡연자 / 금연자 / 흡연자 총
세 집단으로 구분되는 것을 볼 수 있습니다.

이 프로젝트의 목표는 흡연자와
비흡연자를 구분하는 것인데,

그렇다면 금연자는 어떻게 해야 할까요?

데이터 전처리

금연자들의 경우,
각자 금연 중인 기간은 주 단위부터 년 단위까지 천차만별일 것입니다.

그런데 이제 갓 금연을 시작한 사람과 금연을 시작한지 몇 년이 지난 사람을 같은 범주로 묶는 것은 적절하지 않습니다.

더욱이 데이터셋에는 금연 기간에 대한 아무런 정보도 없기 때문에,
이들을 흡연자 / 비흡연자 집단에 나누어 포함시키는 것도 불가능합니다.

따라서 금연자 데이터는 전부 삭제하도록 하겠습니다.

SMK_STAT = 2(금연자)인 행 삭제

데이터 전처리

3. 프로젝트 목적에 부합하지 않는 피처 삭제 #1

연도, 지역, ID(식별자)와 같은 열들은 흡연 여부와 아무런 관련이 없습니다.

가령, 특정 연도나 특정 지역에서 흡연자 비율이 높게 나타난다 하더라도,
그것은 이번 프로젝트의 목적과 아무런 관련이 없습니다.
(건강검진의 결과로서 얻어진 정보가 아니기 때문에)

따라서 이에 해당하는 피처들은 전부 삭제하도록 하겠습니다.

Unnamed: 0, IDV_ID, YEAR, AREA_CODE 열 삭제

데이터 전처리

4. 결측값 처리 #1

각 피처 별로 결측값 개수를 확인해 본 결과, CRS_YN(치아 우식증 여부),
TTR_YN(치석 여부)의 경우 전체 데이터 행의 절반 이상에서 결측값이 나타납니다.

그 원인은 HCHK_CE_IN(구강검진 여부)가 N인 경우,
구강검진을 하지 않았으니 당연히 그 결과값도 없을 수 밖에 없습니다.

이러한 피처들이 흡연 여부와 관련이 있을지도 모르지만,
결측값 수가 너무 많기 때문에 정상적인 분석이 어려울 것 같습니다.

따라서 이와 관련된 3개 피처는 전부 삭제하도록 하겠습니다.

CRS_YN, TTR_YN, HCHK_CE_IN 열 삭제

데이터 전처리

5. 프로젝트 목적에 부합하지 않는 피처 삭제 #2

음주 여부(DRK_YN)는, 건강검진의 결과로써 나타난 것이 아니라 피검자가 문진표에 자신의 음주 여부를 기록한 것입니다.

따라서 음주 여부와 흡연 여부 간에 상관관계가 있다 하더라도, 그것은 이번 프로젝트의 목적과 아무런 관련이 없습니다.

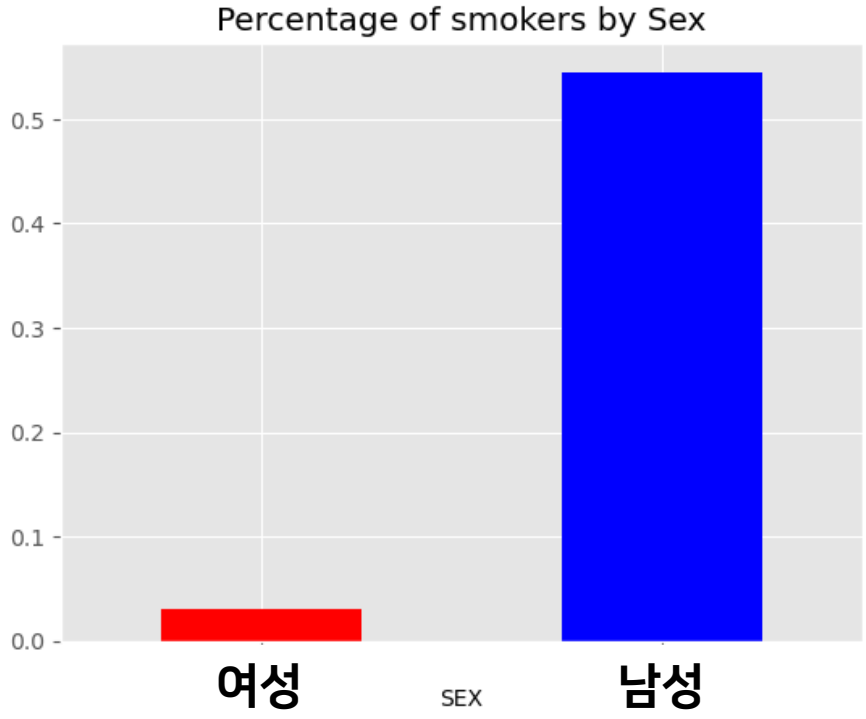
따라서 음주 여부 피처는 삭제하도록 하겠습니다.

음주 여부(DRK_YN) 열 삭제

데이터 전처리

6. 프로젝트 목적에 부합하지 않는 피처 삭제 #3

그 다음으로 주목해 볼 피처는 성별(SEX)입니다.



성별에 따른 흡연자 비율의 분포를 살펴본 결과,

남성의 경우 전체의 절반 이상이 흡연을 하는 반면,

여성의 경우 흡연자 비율이 5%도 되지 않습니다.

이 경우 모델링을 하는데 있어 문제가 발생할 수 있습니다.

데이터 전처리

가령, 모델이 여성 데이터에 한해서 전부 비흡연자라고 판단하면,
(즉, 여성이기만 하면 나머지 피처들을 전부 무시하더라도)
여성 데이터에서 95%가 넘는 정확도를 확보할 수 있습니다.

이러한 상황은 프로젝트의 목적에 부합하지 않습니다.

더욱이 성별은 건강검진의 결과가 아니라,
개인의 인적 사항에 불과합니다.

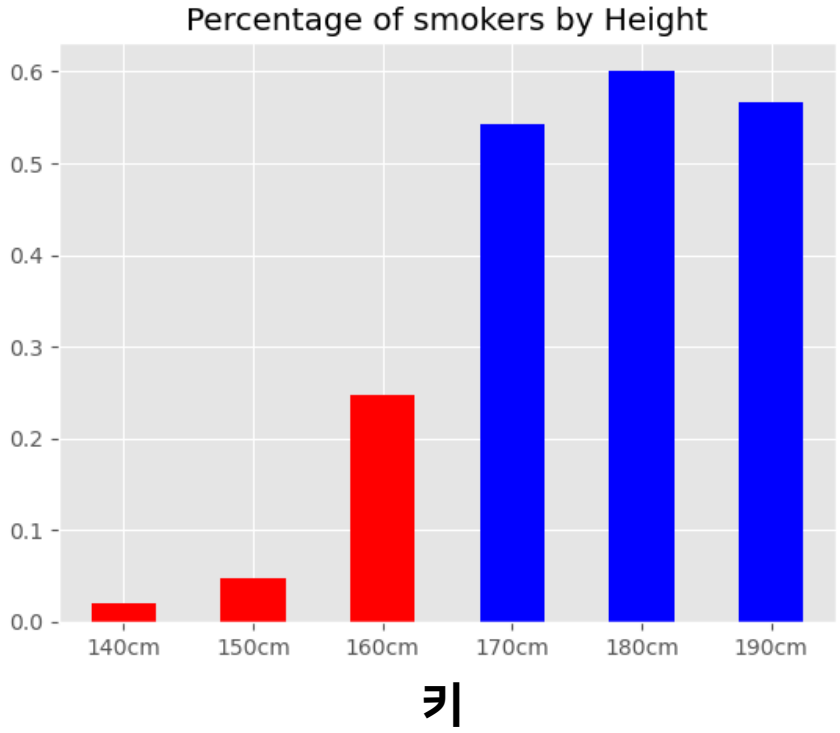
따라서 성별 피처는 삭제하도록 하겠습니다.

성별(SEX) 열 삭제

데이터 전처리

7. 프로젝트 목적에 부합하지 않는 피처 삭제 #4

이번에 주목해 볼 피처는 키(HEIGHT)입니다.



키에 대해10cm 단위로 그룹화하여
그래프를 그려보았습니다.

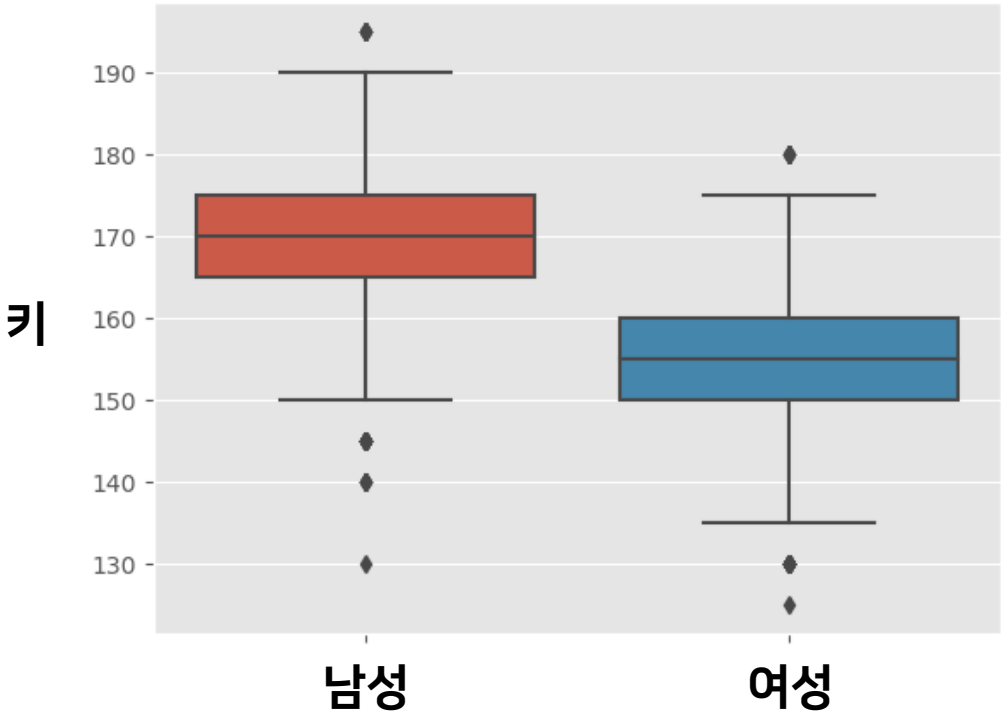
놀랍게도 170cm 이상의 집단에서는
절반 이상이 흡연자인 반면,

그보다 키가 작은 집단에서는
흡연자 비율이 훨씬 적은 것을 볼 수 있습니다.

그렇다면 정말 흡연과 키 사이에 긍정적인
상관관계가 있는 걸까요?

데이터 전처리

아래 그래프를 보면 이러한 상관관계가 나타난 '진짜 이유'를 알 수 있습니다.



왼쪽의 그래프는 성별에 따른 키의 분포를 상자 그림을 나타낸 것입니다.

그래프를 보면,
남성의 중위값 뿐만 아니라, 전반적으로 상자 자체가 여성보다 15cm 정도 높은 곳에서 형성되어 있습니다.
(데이터셋에서 키는 비식별화를 위해 5cm 단위로 반올림 되었으므로, 이 값이 정확한 값은 아니라는 점을 유의해야 합니다.)

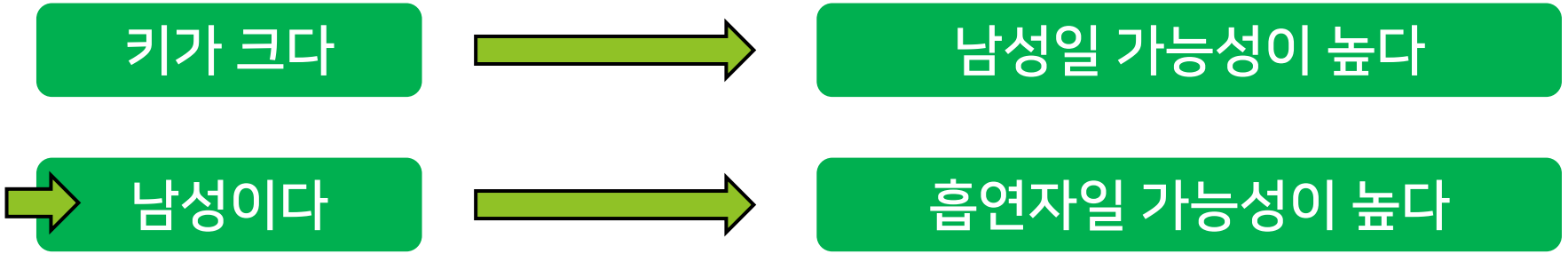
즉, 키가 큰 집단에는 남성이 훨씬 많고,
반대로 키가 작은 집단에는 여성이 훨씬 많습니다.

그렇다면 이게 왜 문제가 되는 걸까요?

데이터 전처리

단순히 키가 높은 집단에서 흡연자 비율이 높다고 해서,
이러한 추측을 하는 것이 옳지 않습니다.

사실은, 아래 추측이 더 신빙성 있습니다.

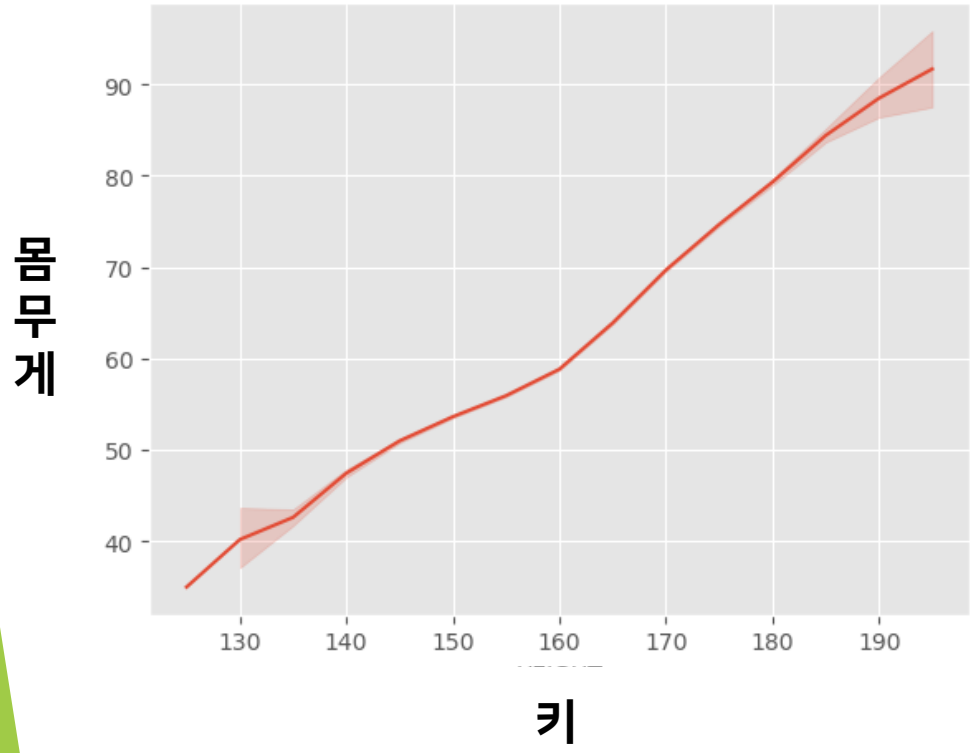


즉, 키가 클수록 남성일 가능성이 높고, 남성 중에는 흡연자가 많으니,
키가 클수록 흡연자일 가능성이 높아 보이는 착각이 발생한 것입니다.

이는 성별이 영향을 준 것이지, 흡연이 키에 영향을 준 것이 아닙니다.

데이터 전처리

그렇다면 몸무게는 어떨까요?
보통 키가 큰 사람의 몸무게가 더 많이 나갑니다.



왼쪽 그래프는 키에 따른 평균 몸무게를 나타낸 그래프입니다.

한눈에 봐도
키가 커질수록 몸무게도 함께 늘어나는 것을 확인할 수 있습니다.

또한 상관계수를 따로 측정해본 결과,

몸무게 – 키 : 0.66
몸무게 – 성별 : 0.56 으로
상관관계가 꽤나 높게 나타났습니다.

데이터 전처리

몸무게가 높다

키가 클 가능성이 높다

남성일 가능성이 높다

흡연자일 가능성이 높다

그렇다면 키의 경우와 마찬가지로,
몸무게가 높을 수록 흡연자일 가능성이 높게
나타날 것입니다.

하지만 이는 궁극적으로 남성일 가능성이 높기
때문인 것이지, 흡연이 몸무게에 영향을 끼쳤다고
보기는 어렵습니다.

그렇다면 몸무게 피처를 어떻게 해야 할까요?

키 피처는 그냥 삭제하게 되더라도,

몸무게 피처가 담고 있는 비만/저체중에 관한 정보를
함께 잃어버리는 것은 큰 손실입니다.

데이터 전처리

저희가 떠올린 방법은, 몸무게 피처를 BMI로 대신하는 것입니다.

BMI 공식 : [몸무게] / [(키 x 0.01)^2]

BMI 지수는 몸무게가 커도 키가 클 수록 이를 상쇄하기 때문에
비만 여부를 알아내는데 유용합니다.

몸무게와의 상관계수
키 : 0.66
성별 : 0.56

BMI와의 상관계수
키 : 0.06
성별 : 0.16

또한 이번에는 BMI 지수와 키, 성별 사이의 상관계수를 측정해보았습니다.

키의 경우 10분의 1, 성별의 경우 3분의 1 아래로 상관계수가 떨어졌습니다.

데이터 전처리

따라서 BMI 지수는 성별의 영향을 거의 받지 않고
비만 및 저체중에 대한 정보를 줄 수 있을 것으로 기대됩니다.

한편 키는 엄밀히 따지면 건강검진의 결과가 맞긴 하지만,
성별과의 강한 상관관계를 해소할 방법이 없습니다.

또한 키는 성장기가 지나면 그 후로 거의 변화가 없는 지표입니다.
즉, 성장기가 이후의 흡연은 키와 큰 관련이 없다고 봐야합니다.

따라서 키 피처는 삭제하도록 하겠습니다.

키(HEIGHT) 열 삭제

몸무게(WEIGHT) -> BMI 변환

데이터 전처리

8. 대체값이 사용된 피처 조정 #1

다음으로 확인할 피처는 왼쪽 시력(SIGHT_LEFT)과 오른쪽 시력(SIGHT_RIGHT)입니다.

통계최강자전의 데이터셋 개요 페이지를 보면,
해당하는 눈이 실명된 경우, 대체값으로 9.9를 사용했다고 적혀있습니다.

이 대체 값을 그대로 사용하는 것은 모델로 하여금 시력이 9.9인 것으로 잘못 인식될 수 있으므로,
-1로 대체하도록 하겠습니다.

0이 아니라 굳이 -1을 택한 이유)
데이터셋이 처음부터 실명자의 시력에 대해 0이 아니라 대체값을 사용한 만큼,
그 의도를 받아들여 저 역시 0과 구분되도록 -1을 사용하였습니다.

시력(SIGHT_LEFT/RIGHT)열의 대체값 9.9를 -1로 조정

데이터 전처리

9. 결측값 처리 #2

현재 데이터셋에는 결측값이 많이 남아있습니다.

그래서 결측값을 많이 포함한 행들의 비율을 확인해 보았습니다.

한 행에 결측값이 5개 이상인 행의 비율 : 0.3%

한 행에 결측값이 4개 이상인 행의 비율 : 14.9%

한 행의 너무 많은 결측값들을 대체하게 되면, 모델 성능에 악영향을 줄 수 있습니다.

따라서 비교적 소수인 결측값 5개 이상인 행은 삭제하고, 4개인 행들을 후에 따로 처리하겠습니다.

한 행에 결측값이 5개 이상인 행 삭제

데이터 전처리

10. 대체값이 사용된 피처 조정 #2

다음으로 살펴 볼 피처는 왼쪽 청력(HEAR_LEFT)과 오른쪽 청력(HEAR_RIGHT)입니다.

데이터셋 개요 페이지에는 이상이 없으면 1, 이상이 있으면 2라고 나와있는데,
설명에는 없는 '3'도 존재합니다.

다만, 그 수가 매우 적은 것으로 보아 사실상 검사 오류나 결측값 정도로 생각됩니다.

또한 정상(1)이 압도적으로 다수이므로, 이 피처에 대해서 결측값('3' 포함)은
최빈값인 정상(1)로 대체하도록 하겠습니다.

청력 이상(HEAR_LEFT/RIGHT)열의 결측값을 최빈값 대체

데이터 전처리

11. 결측값 처리 #3

이번에 살펴 볼 피처들은 꽤나 골치 아픈 문제를 갖고 있습니다.
바로 콜레스테롤 및 중성지방과 관련된 4개 피처입니다.
(TOT_CHOLE, HDL_CHOLE, LDL_CHOLE, TRIGLYCERIDE)

이 피처들은 공통적으로 14000개 가량의 결측값을 가지고 있습니다.
그 비율이 15% 정도에 육박하기 때문에 함부로 결측값 대체를 하는 것은
좋은 선택이 아닌 것 같습니다.

따라서 선택할 수 있는 방안은 아래 2가지입니다.

1. 결측값이 발생한 약 14000개 행을 삭제한다.

2. 해당하는 4개 열(피처)을 삭제한다.

데이터 전처리

이 두 가지 방안 중에 어떤 것을 선택하는 것이 좋을 지 테스트를 위해,
각각의 방안을 채택한 A_Data와 B_Data를 임시로 만들었습니다.

그 후, 두 데이터를 임시로 모델링하여 어떤 데이터를 사용한 모델이 더 좋은 성능을 내는지
한 번 확인해보도록 하겠습니다.

시험에 사용할 모델은 '랜덤포레스트'입니다.
(별도의 튜닝 및 데이터 표준화 없이도 어느 정도의 성능을 기대할 수 있기 때문에 선택)

단, 평가 지표로는 정확도(Accuracy)가 아닌 F1 score를 사용하겠습니다.

왜냐하면 현재 데이터셋에 흡연자보다 비흡연자 수가 훨씬 많은 관계로,
비흡연자에 편향된 예측을 해서 높은 정확도를 달성할 수 있기 때문에,
이러한 문제점이 적은 F1 score를 사용하여 평가하겠습니다.

(F1 score는 정밀도와 재현율의 조화평균을 이용한 지표이며,
정밀도와 재현율에 대해서는 나중에 설명과 함께 한 번 더 다룰 예정입니다.)

데이터 전처리

각 데이터별로 5번씩 테스트를 진행하여 그 평균을 비교하고자 하며,

매 테스트마다 random_state를 달리하여 평가 데이터를 10% 분리한 후,
이 평가 데이터로 점수를 산출했습니다.

A_Data | F1 score
0.6207

VS

B_Data | F1 score
0.6120

정말 근소한 차이로 A_Data의 F1 score가 높게 나왔습니다.

따라서, 콜레스테롤 및 중성지방과 관련하여 결측값이 발생한
약 14000개 행을 삭제하는 방안을 채택하도록 하겠습니다.

**4개 피처(TOT_CHOLE, HDL_CHOLE, LDL_CHOLE, TRIGLYCERIDE)에
결측값이 있는 행 삭제**

데이터 전처리

11. 결측값 처리 #4

이제 마지막으로 남은 소수의 결측값들에 대해,
일괄적으로 (각 피처의) 중앙값으로 대체하도록 하겠습니다.

(각 행의 결측값이 4개를 넘기지 않으며, 그 수도 적기 때문에 일괄적으로 처리)

중앙값을 선택한 이유)
평균값보다는 이상치의 영향을 적게 받는 중앙값을 선택했습니다.

일괄적으로 남은 결측값들에 대해 중앙값으로 대체

전처리를 마치며

구글링을 통해 각종 신문기사와 의학논문을 살펴보면,
그리 어렵지 않게 흡연과 강한 상관관계가 있는 피처들을 추려낼 수 있습니다.

하지만 저희는 특정 논문이나 신문기사를 토대로 특정 피처에 가중치를 부여하거나,
삭제하는 등의 작업을 전혀 하지 않았습니다.

왜냐하면 이 프로젝트의 목표는 오직 건강검진의 결과 그 자체로부터,
흡연 여부를 예측하는 것이 과연 가능할 지를 확인하는 것이기 때문입니다.

그래서 일단 여기서 데이터 전처리를 마치고,
본격적인 모델링에 돌입하도록 하겠습니다.

여러 모델을 테스트해보며 가장 적합한 모델을 선택한 후에는,
추가적인 데이터 핸들링을 통해 모델의 성능을 끌어올릴 방법을 찾아보도록 하겠습니다.

모델링 준비

준비1. 테스트 데이터 / 검증 데이터 분리

본격적인 모델링을 시작하기에 앞서,
원활한 진행을 위해 먼저 테스트 데이터와 검증 데이터를 분리하도록 하겠습니다.

테스트 데이터란, 최종적으로 모델과 관련된 모든 것이 결정된 후에
처음 보는 데이터에 어느 정도의 성능을 기대할 수 있을 지 확인해 볼 데이터입니다.

반면 검증 데이터는, 여러 모델 중 어떤 모델이 가장 적합한 지 선택하는 과정에서
각각의 모델들을 평가하는데 사용할 데이터입니다.

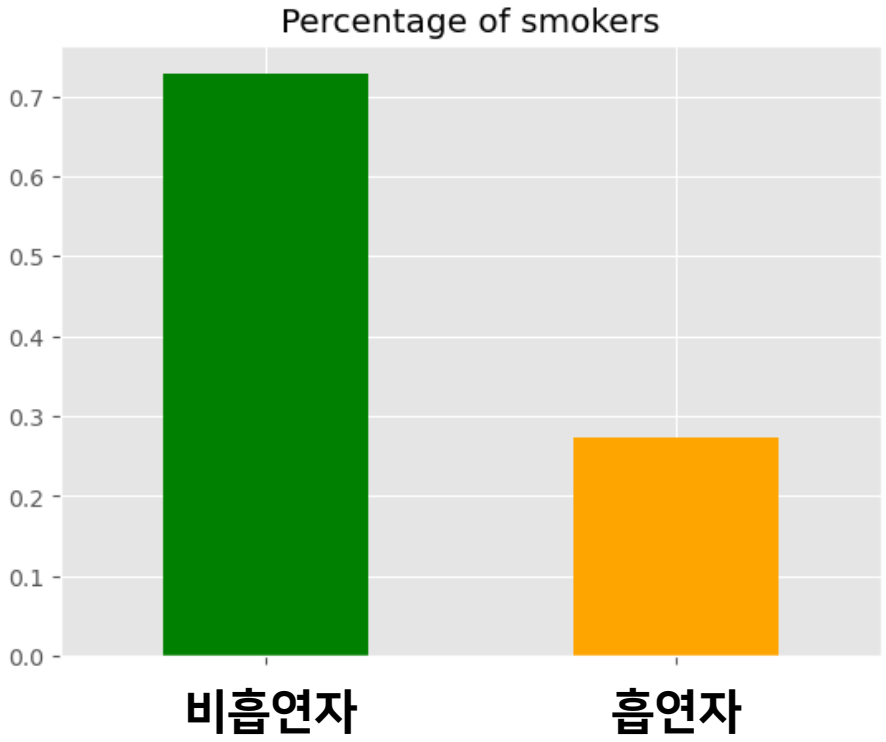
전체 데이터셋에서 10%를 테스트 데이터로 분리하고,
남은 90%에서 또 10%(전체의 9%)를 검증 데이터로 분리하겠습니다.

테스트/검증 데이터 각각 10%씩 분리

모델링 준비

준비2. 흡연자/비흡연자 데이터 불균형 해소

이 데이터셋의 흡연자/비흡연자 비율은 그 차이가 크게 나는 편입니다.



옆의 그래프를 보면
흡연자 수가 비흡연자의 3분의 1 수준
정도밖에 안되는 것을 알 수 있습니다.

이 경우 무엇이 문제가 될까요?

모델은 모든 데이터에 대해 그 어떤 피처도
고려하지 않고 곧바로 '비흡연자'라고
예측을 하더라도,
75% 정도의 정확도를 기대할 수 있습니다.

모델링 준비

그렇다면 정확도 대신 F1 score를 사용하면 어떨까요?

하지만 F1 score을 사용하는 데에는 한 가지 문제점이 있습니다.
바로 F1 score가 “직관적으로 와닿지 않는다는 것”입니다.

정확도와 F1 score 모두 1(100%)에 가까울수록 좋다는 점은 같으나,
정밀도와 재현율의 조화평균으로 얻을 수 있는 F1 score는
정확도와 달리 직관적인 해석이 어렵습니다.

정확도(Accuracy)가 0.7 → 테스트 샘플 중 70%의 예측을 맞췄구나!

F1 score가 0.7 → ……?

모델링 준비

그래서 저희는 앞으로의 분석에서 정확도를 지표로 이용할 것입니다.

물론, 이를 위해서는 데이터의 불균형을 해소해야 하기 때문에
여기서는 흡연자의 데이터를 불리도록 하겠습니다.

훈련 데이터에 대해서, 비흡연자의 데이터 수와 같아지도록
흡연자의 데이터를 리샘플링(복원추출) 하겠습니다.
(비흡연자 데이터는 모두 다른 사람이지만, 흡연자 데이터에는 중복되는 사람이 발생)

이렇게 함으로써, 데이터 손실없이 흡연자/비흡연자 비율을 50:50으로 만들었습니다.

또한, 테스트/검증 데이터는 여전히 기존의 흡연자/비흡연자 비율이 유지됩니다.

훈련 데이터에서 리샘플링을 통해 흡연자 비율 50%로 채우기

모델링 준비

준비3. 데이터 표준화

마지막으로, 데이터의 각 피처들을 '표준화'하도록 하겠습니다.

표준화 절차는 다음과 같습니다.

1)각 피처들의 평균값과 표준편차를 구합니다.

2)각 값들에서 (해당하는 피처의) 평균값을 빼고, 표준편차로 나눕니다.

3)모든 피처들의 평균이 0, 분산이 1로 스케일이 같아집니다.

하지만 위 과정에 대해 직접 코드를 짤 필요가 없으며,
scikit-learn이 제공하는 "StandardScaler"가 이 과정을 대신해줍니다.

StandardScaler로 데이터 표준화

모델링 준비

Q&A

1) 테스트 데이터와 검증 데이터의 비율을 10%씩 추출한 이유

전처리를 마쳤을 때, 데이터 수가 8만 행 가량 남아있었습니다.
전처리 과정에서 데이터를 많이 잘라내긴 했지만,
8만 행의 데이터는 여전히 많은 데이터이기 때문에 각각 10%씩만 잘라내서
검증 / 테스트하더라도 문제가 없다고 판단했습니다.

2) 데이터 표준화를 한 이유

특정 머신러닝 기법을 사용한 모델에서, 다른 피처들보다 스케일이 큰 피처의 경우
모델 및 예측에 더 큰 영향을 끼칠 수 있습니다.
하지만 스케일이 크다고 해서 피처의 중요도가 높은 것은 절대 아니므로,
이러한 문제를 예방하기 위해 표준화를 진행하였습니다.

모델 선택

지금부터는 여러 머신러닝 기법을 이용하여 모델을 만들어보고,
그 결과를 비교하여 최선의 모델(기법)을 선택하도록 하겠습니다.

확인해 볼 머신러닝 기법은 아래 6가지입니다.
([서포트 벡터 머신]은 코드 실행시간이 너무 길어져서 제외)

- 로지스틱 회귀
- 가우시안 나이브 베이즈
- 랜덤포레스트
- XGBoost
- lightGBM
- K-최근접 이웃

모델 선택

모델링 방법

모델링에는 scikit-learn이 제공하는 모델을 사용합니다.

scikit-learn 패키지에서 해당하는 모듈을 import하여 모델 객체를 생성하고,
주요 하이퍼파라미터를 튜닝하여 훈련 데이터를 모델에 학습시킵니다.

그 후, 훈련 데이터와 검증 데이터에 대해 예측을 실시해,
각각의 정확도를 나타내는 “훈련 점수”와 “검증 점수”를 확인합니다.

이렇게 서로 다른 모델들 간의 검증 점수를 비교하여,
검증 정확도(검증 데이터의 예측 정확도)가 가장 높은 모델을 선택하고자 합니다.

모델 선택

하이퍼파라미터 튜닝 #1

하이퍼파라미터란, 모델이 훈련을 시작하기 전에 미리 정해주는 변수로,
훈련을 통해 학습하는 변수가 아닙니다.

이 하이퍼파라미터를 적절하게 세팅해줘야 좋은 예측 성능을 기대할 수 있으며,
적절한 하이퍼파리미터를 찾는 과정이 바로 '하이퍼파라미터 튜닝'입니다.

튜닝을 위해 scikit-learn이 제공하는 서치 모듈을 사용하겠습니다.
이번에 사용할 모듈은 GridSearchCV와 RandomizedSearchCV입니다.

전자는 제시된 선택지의 모든 조합에 대하여 테스트하며,
후자는 주어진 범위에서 랜덤으로 초기값을 잡고, 최적값을 찾아가며 테스트합니다.

적정값을 가늠할 수 없는 연속형 변수를 튜닝할 때 후자를,
그 외의 경우에는 전자를 선택하여 하이퍼파라미터 튜닝을 진행하였습니다.

모델 선택

하이퍼파라미터 튜닝 #2

모든 하이퍼파라미터를 일일이 튜닝할 경우,
각각의 경우에 대해 모두 테스트하는데 시간이 오래 걸립니다.

따라서 이번 튜닝 과정에서는 현실적으로 컴퓨터 성능이 받쳐주는 한에서,
규제 관련 변수에 집중하였습니다.

규제가 너무 약하면 모델이 '과대적합'되며,
처음 보는 데이터셋 상대로 예측성능이 떨어집니다.

하지만 반대로, 규제가 너무 강하면 모델이 '과소적합'되며,
모델 훈련 과정에서 데이터의 구조/패턴을 잘 학습하지 못하게 됩니다.

따라서, 과대적합과 과소적합 사이에서 적절한 균형을 찾아내고자,
규제 관련 변수에 주목하여 튜닝을 진행하였습니다.

모델 선택

로지스틱 회귀

먼저, **로지스틱 회귀** 기법을 이용하여 모델링을 해보겠습니다.

이름에 '회귀'가 들어가지만, 특이하게 '분류'에 사용가능한 기법입니다.
또한, 이진 분류에 특화되어 있기 때문에 이번 프로젝트에 적합하다고 할 수 있습니다.

RandomizedSearchCV를 이용하여 규제 파라미터 'C'(10^(-3)에서 10^3 사이)와,
비용함수 최적화 방법 'solver'에 대해 튜닝을 하였습니다.

훈련 정확도 : 78.615%

검증 정확도 : 78.789%

훈련 정확도와 검증 정확도가 거의 비슷하므로 과대적합되지는 않았으나,
검증 정확도가 훈련 정확도보다 높게 나타나서 과소적합되었을 가능성도 있는 것 같습니다.

모델 선택

가우시안 나이브 베이즈

가우시안 나이브 베이즈 기법을 이용하여 모델링을 해보겠습니다.

주어진 데이터셋에서 많은 피처들이 연속형 변수이기 때문에,
여러 나이브 베이즈 기법 중에서 가우시안 나이브 베이즈를 선택하였습니다.

이 기법의 경우 딱히 유의해서 튜닝할 하이퍼파라미터가 없다고 판단되어,
분산 스무딩 정도만 기본값을 중심으로 RandomizedSearchCV를 하였습니다.

훈련 정확도 : 73.872%

검증 정확도 : 72.453%

모델의 특성상 훈련 속도는 확실히 빨랐지만,
로지스틱 회귀에 비하여 검증 정확도에서 6%p 가량 적게 나타났습니다.

모델 선택

랜덤 포레스트

랜덤 포레스트 기법을 이용하여 모델링을 해보겠습니다.

랜덤 포레스트는 전체 데이터에서 샘플을 복원 추출하여 의사결정나무를 만들고,
최종적으로 각 나무들의 결과를 집계하여 최종 결과를 내놓는 모델입니다.

나무의 최대 깊이를 깊게 조절할수록 모델이 복잡해지며,
이에 따라 모델이 과대적합될 수 있습니다.

훈련 정확도 : 82.64%

검증 정확도 : 78.628%

나무의 최대 깊이를 제한하지 않을 경우, 검증 정확도가 아주 조금 더 높아지지만,
훈련 정확도가 100%로 심각하게 과대적합됩니다.
따라서 나무의 최대 깊이를 9로 제한하여 훈련했습니다.

모델 선택

XGBoost

XGBoost 패키지를 이용하여 모델링을 해보겠습니다.

랜덤포레스트처럼 의사결정나무를 활용한 트리 기반의 모델이지만,

여러 개의 깊이가 얕은 의사결정나무를 순차적으로 만들고,
이전의 오답에 가중치를 부여하면서 더 나은 분류 결과를 만들어 가는
'부스팅' 알고리즘이 사용되었다는 점에서 차이가 있습니다.

훈련 정확도 : 82.117%

검증 정확도 : 78.212%

코드의 실행시간이 너무 길어져서,
코드 파일에는 사전에 튜닝된 값의 근사값을 설정했습니다.
결과에서도 랜덤포레스트와 크게 차이가 나지 않습니다.

모델 선택

lightGBM

lightGBM 패키지를 이용하여 모델링을 해보겠습니다.

XGBoost와 마찬가지로 부스팅을 사용한 트리 기반의 모델이지만,
트리를 생성하는 매커니즘에 차이가 있습니다.

L2 규제의 'reg_lambda'와 L1 규제의 'reg_alpha' 파라미터에 대해
RandomizedSearchCV를 이용하여 튜닝을 진행했습니다.

훈련 정확도 : 86.6%

검증 정확도 : 80.494%

현재까지 테스트한 모델 중에서 검증 정확도가 유일하게 80%를 넘었습니다.
또한, XGBoost보다 월등하게 빠른 속도로 훈련을 마쳤습니다.

모델 선택

K-최근접 이웃

K-최근접 이웃 기법을 이용하여 모델링을 해보겠습니다.

분류하려는 샘플로부터 가장 가까운 k개 이웃 샘플 중,
그 갯수가 가장 많은 집단에 해당 샘플을 할당합니다.

가장 적절한 이웃의 수인 k를 찾기 위해서 GridSearchCV를 활용하였습니다.

훈련 정확도 : 91.55%

검증 정확도 : 72.882%

훈련 정확도와 검증 정확도가 거의 20%p 가량 차이 났습니다.

과대적합이 매우 심각하게 나타나지만, 이 이상 이웃의 수 K를 조절하여
검증 정확도를 올릴 수 없었습니다.

모델 선택

중간점검

지금까지 6개의 모델을 테스트해본 결과,

검증 정확도가 가장 높게 나타난 것은 lightGBM이었고,
그 다음으로 78%대인 랜덤포레스트, 로지스틱 회귀, XGBoost,
72%대인 가우시안 나이브 베이즈, k-최근접 이웃 순으로 나타났습니다.

여기까지만 놓고 봤을 때,
검증 정확도가 가장 높게 나타났고,
모델 훈련 시간이 비교적 짧게 나타난 lightGBM을 채택하는 것이 타당해 보입니다.

하지만, 결정을 내리기 전에 한 번 시도해볼 것이 있습니다.

모델 선택

lightGBM과 다른 모델들 간의 예측 일치율을 살펴보겠습니다.

lightGBM과의 예측 일치율
XGBoost : 93.529%
로지스틱 회귀 : 91.422%
랜덤포레스트 : 93.731%

대체로 타 모델과 90% 이상 예측이 일치하지만,
부분적으로 7~10% 정도 예측이 갈리는 것을 보여줍니다.

만약, 예측이 갈리는 지점에서 서로 다른 모델들이 다수결로 결정하면
어떤 결과가 나올까요?

모델 선택

모델 간 앙상블(보팅)

만약, 한 모델에서 데이터 하나를 잘못 예측하더라도 나머지 모델들에서
올바르게 예측할 경우 다수결은 옳게 됩니다.

이러한 사실에서 착안한 scikit-learn의 VotingClassifier를 활용해보겠습니다.
VotingClassifier는 서로 다른 모델들의 예측을 집계하여 최종 결과를 내놓습니다.

또한, voting 옵션을 'soft'로 두면 각 모델이 내놓은 확률을 집계하여 결과를 내놓습니다.

보팅의 검증 정확도 : 79.876%

lightGBM을 제외한 나머지 모델에 비해서 높은 검증 정확도를 달성했지만,
정작 lightGBM에 비하면 약간 아쉬운 수치입니다.

따라서 최종 기법(모델)으로 **lightGBM**을 선택하도록 하겠습니다.

데이터 핸들링

주성분 분석을 활용한 차원 축소

특성 추출이란 차원 축소의 방법 중 하나로,
원본 데이터셋의 특성(피처)을 추출하여 새로운 특성 공간으로 데이터를 변환합니다.

이때, 새롭게 변환된 피처는 원본의 그것과 달라지며,
그 수는 원본 데이터셋보다 같거나 적어집니다.

피처가 많은 데이터의 경우, 차원의 저주로 인해 모델의 성능이 떨어질 수 있는데,
차원 축소를 통해 이러한 차원의 저주가 다소 완화될 수 있습니다.

이번 프로젝트에서 특성 추출을 위해 사용할 기법은 '주성분 분석(PCA)'입니다.

데이터 핸들링

주성분 분석(PCA)은 행렬의 고유값 분해를 이용하여
기존의 데이터셋을 이전보다 축소된 차원의 부분 공간으로 변환합니다.

이 주성분 분석을 실시하여 모델 성능 향상이 가능한지 확인해보겠습니다.

PCA 적용 이전 검증 정확도
80.494%

PCA(분산 설명률 90%) : 79.353%

PCA(분산 설명률 80%) : 77.769%

주성분 분석을 실시하여 데이터셋의 차원을 축소한 결과,
오히려 예측 성능이 하락한 것을 확인할 수 있습니다.

PCA를 실시하면 결과적으로 원본 데이터셋의 정보가 일부 손실되는 셈이며,
차원의 저주 등으로 인해 성능이 하락한 것이 아니라면 예측 성능 향상을 기대할 수 없습니다.

이번 사례도 여기에 해당하는 것 같습니다.

데이터 핸들링

피처의 정상 범주를 고려한 스케일링

다음으로 시도해볼 것은 '정상 범주 스케일링'입니다.

데이터셋 소개 페이지에서는 각 피처의 정상 범주에 관한 설명이 나와있습니다.
예를 들어 TRIGLYCERIDE(중성지방)의 경우, 정상 범주는 30에서 135 사이라고 합니다.

지금부터 하려는 작업은,
(실제값 - 정상 범주 하한) / (정상 범주 상한 - 정상 범주 하한) 식을 적용하여,
각 피처를 정상 범주에 대해 정규화하는 것입니다.

이렇게 정규화할 경우, 정상 범주에 속한 값은 0과 1 사이에,
정상 범주에 속하지 않은 값은 그 밖에 위 치하게 됩니다.

(아이디어는 '최소-최대 정규화' 의 (실제값 - 최소값) / (최대값 - 최소값) 로부터 착안하였습니다.)

데이터 핸들링

그 후, 이렇게 정규화한 값을 제곱합니다.
(음수의 경우, -1을 곱해서 부호를 유지합니다.)

제곱을 하는 이유는, 정상 범주에 포함된 값은 0에 가까워지는 반면,
정상 범주를 크게 벗어난 값일 수록 그 값이 더 빠르게 커지도록 하여 피처가 주는 변별성을
극대화시켜보기 위함입니다.

단, 모든 피처들에 대하여 스케일링을 진행하는 것이 아니라,
lightGBM이 제공하는 특성 중요도를 참고하여 적당한 피처 3개에 대해서 먼저 적용해보겠습니다.

제가 선택한 3개의 피처는 GAMMA_GTP, CREATININE, HMG 입니다.
특성 중요도가 가장 높은 OLIG_PROTE_CD(요단백)은 1~6 사이의 정수만 나타나는 범주형 변수이므로, 그
다음 순위인 CREATININE(크레아티닌)과 HMG(혈색소)를 선택했습니다.

나머지 하나는 가장 중요도가 낮은 SGPT_ALT(간 관련 수치)를 선택해보았습니다.
그럼 이제 특성들에 대해서 정상 범주를 고려하여 스케일링을 진행해보겠습니다.

데이터 핸들링

3개 피처에 정상 범주 스케일링 후 검증 정확도 : 81.407%

정상 범주 스케일링을 적용한 결과, 검증 정확도가 약 1% 정도 상승했습니다.

이 1%의 상승이 오로지 정상 범주 스케일링의 덕분이라고 단정할 수는 없지만, 적어도 이러한 스케일링이 데이터를 손상시켜 모델의 성능을 저하시키지는 않았다고 판단할 수 있습니다.

그렇다면 나머지 변수들에도 이러한 스케일링을 적용하고 결과를 확인해보겠습니다.

모든 피처에 정상 범주 스케일링 후 검증 정확도 : 80.709%

모든 피처에 정상 범주 스케일링을 적용한 결과, 오히려 검증 정확도가 떨어졌습니다.

따라서 모든 피처에 정상 범주 스케일링을 한 데이터 대신,

이전에 3개의 피처에만 스케일링을 진행한 데이터셋을 사용하도록 하겠습니다.

최종 모델 튜닝 및 테스트

랜덤포레스트를 최종 모델로 결정한 후, 보팅, PCA, 스케일링 등
여러 방법을 시도해본 결과, 일부 피처에 정상 범주 스케일링을 적용했을 때
약간의 성능 향상이 있었습니다.

이 방법 이외에도 더 좋은 방법이 있겠지만,
저희가 알고 있는 선에서 할 수 있는 방 법은 전부 동원해본 것 같습니다.

그렇다면 현재 데이터셋에서 lightGBM 모델을 마지막으로 튜닝하고,
테스트 데이터셋을 이용하여 최종(테스트) 정확도를 측정해보겠습니다.

테스트 정확도 : 79.159%

따라서 저희의 최종 모델은 처음 보는 데이터셋에도
약 79%의 정확도로 흡연 여부를 맞출 수 있을 것으로 기대됩니다.

정밀도와 재현율

정확도 외에도 테스트 결과의 정밀도와 재현율을 살펴보겠습니다.

정밀도) 모델이 양성이라 예측한 것 중에, 실제 양성인 비율
재현율) 실제 양성인 것 중에, 모델이 양성이라 예측한 비율

테스트 정밀도 : 57.945%

테스트 재현율 : 85.955%

재현율이 86%에 가깝게 나왔으므로,
이 모델은 흡연자의 86% 정도에 대해 흡연자로 맞게 예측할 것이라 기대할 수 있습니다.

하지만, 정밀도가 60%가 채 안된다는 것은,
흡연자로 예측한 인원들 중 40% 이상은 실제로는 비흡연자였다는 뜻입니다.

정리하면 이 모델의 흡연자에 대한 판단 기준선이 다소 낮기 때문에,
흡연자를 흡연자로 정확하게 예측할 가능성은 높으나,
비흡연자를 흡연자로 잘못 예측할 가능성 또한 높다고 판단을 내릴 수 있겠습니다.

특성 중요도 분석

lightGBM 모델은 훈련을 마친 후 _feature_importance 메서드를 이용해,
특성 중요도를 확인할 수 있습니다.

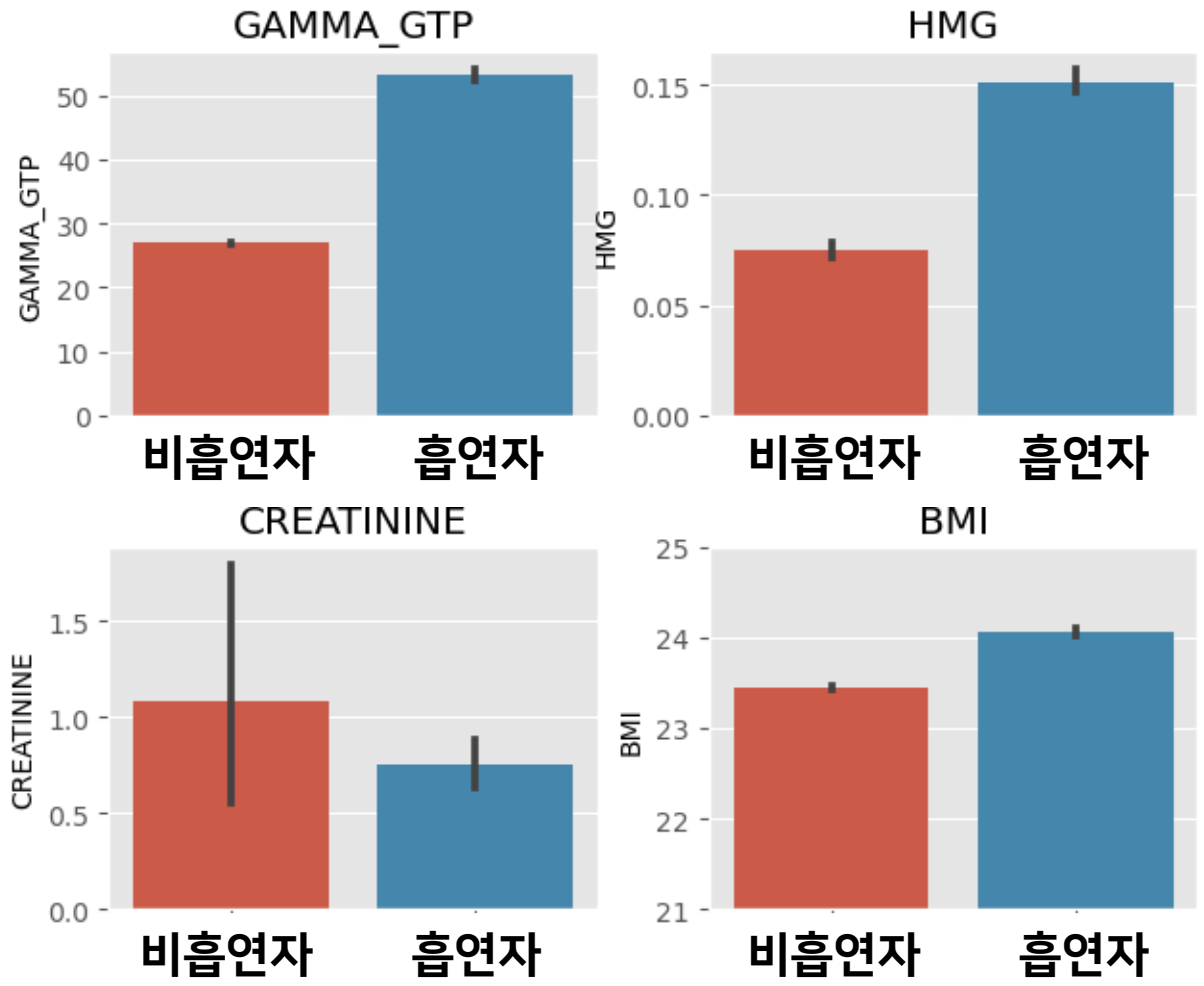
트리형 모델인만큼 나무에서 분기되었을때,
다음 노드에서 불순도가 많이 줄어들수록 더 높은 특성 중요도를 갖게 됩니다.

여기에서는 각 피처의 특성 중요도를 전체 특성 중요도의 합으로 나눠,
전체에 대한 비율을 구한 후, 높은 순으로 4개의 피처를 확인해보았습니다.

피처	특성 중요도 비율 (소수점 셋째자리까지)
감마지티피(GAMMA_GTP)	0.303
혈색소(HMG)	0.229
크레아티닌(CREATININE)	0.145
BMI	0.087

특성 중요도 분석

이렇게 선택된 4개 피처에 대해 흡연자와 비흡연자의 평균 차이를 시각화해 보겠습니다.



특성 중요도 분석

특성 중요도가 가장 높았던 GAMMA_GTP(감마지티피)와 그 다음으로 높았던 HMG(혈색소)의 경우,
흡연자(1)와 비흡연자(0)의 구분이 확실하게 나타납니다.

GAMMA_GTP는 간과 관련된 효소의 일종으로 알코올에 의해 간기능이 저하될 경우 많이 분비되기 때문에,
보통은 알코올 중독과 관련된 지표로 이용됩니다.

하지만, 그래프를 보면 흡연자와 비흡연자 사이의 차이가 뚜렷하게 나타나고 있습니다.
그 원인은 아마도 다음 두 가지로 예상할 수 있습니다.

1. 실제로 흡연이 간 기능에 악영향을 미쳐서 GAMMA_GTP가 높게 나온 것. (인과관계)
2. 흡연자일수록 알코올 섭취량이 많아서, GAMMA_GTP도 높게 나온 것. (선행변수 혹은 매개변수)

그 다음으로 HMG(혈색소)란 헤모글로빈을 말하는 것으로,
체내의 혈액 안에서 산소와 결합하여 산소를 운반하는 물질입니다.

위의 그래프에서는 흡연자와 비흡연자 간에 2배 가까이 차이가 나는 것을 확인할 수 있습니다.

실제로 흡연자일수록 혈색소가 높게 나타나는 것으로 알려져 있으며,
혈색소가 낮으면 빈혈, 높으면 적혈구 과다증(혈색소 과다)이라고 합니다.

특성 중요도 분석

반면 CREATININE(크레아티닌)의 경우, 비흡연자의 그래프에 편차 직선이 길게 나타나 있습니다.

이말은, 비흡연자의 경우 편차 직선을 따라 다양한 수치에 분포되어 있다는 뜻입니다.
더욱이 크레아티닌은 근육에서 생성되는 노폐물로, 수치가 높을수록 신장이 잘 기능하지 못하고 있다는 뜻인데 오히려 비흡연자의 평균이 흡연자의 평균보다 더 높게 나타났습니다.

따라서, 특성 중요도가 비교적 높게 나타나기는 했지만,
앞의 두 피처와 비교하여 유의미한 피처라고 보기 어렵습니다.

마지막으로 BMI의 경우, 흡연자와 비흡연자의 평균 차이가 아주 경미합니다.
직관적으로 생각해보면, 담배보다는 식습관과 운동 여부가 훨씬 더 중요한 요인이 될 것입니다.

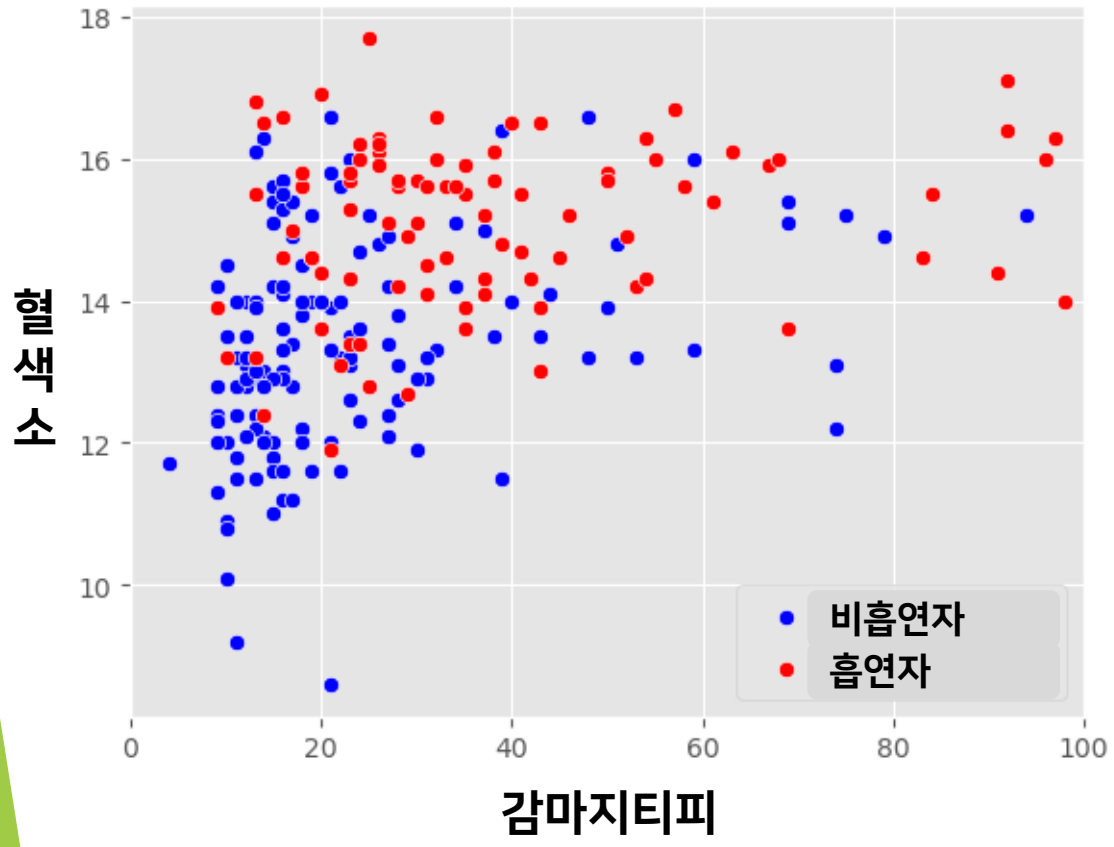
따라서 BMI 역시 앞의 두 피처(감마지티피, 혈색소)에 비하면 덜 유의미한 것 같습니다.

또한, 이보다 특성 중요도가 더 낮은 피처들은 이보다 흡연자 – 비흡연자 간
차이가 더 작게 나타날 것으로 예상할 수 있습니다.

그렇다면 **감마지티피**와 **혈색소**만으로 어느 정도까지 예측이 가능할까요?

특성 중요도 분석

데이터셋에서 샘플을 무작위로 추출하여,
감마지티피와 혈색소의 분포를 산점도로 시각화해 보았습니다.



흡연자와 비흡연자가
각각 미세하게 좀 더 밀집해있는
구역이 있기는 하지만,

이 산점도만으로 흡연자와
비흡연자를 구분하는 것은 사실상
불가능해 보입니다.

따라서, 흡연자 – 비흡연자 구별을
하는데 있어 특정 피처가 핵심
역할을 한다기보다는, 전반적으로
모든 피처가 함께 고려되어야
좋은 예측을 할 수 있을 것 같습니다.

결론

주어진 건강검진 데이터에서 성별/지역/음주 여부 등 프로젝트 목적에 부합하지 않는 일부 피처들을 제외하고 훈련한 결과, 약 80%에 준하는 정확도로 흡연자/비흡연자를 구별해낼 것으로 기대할 수 있습니다.

다만 해당 모델의 정밀도가 꽤나 낮게 나타났으므로, 비흡연자를 흡연자로 잘못 판단할 가능성이 높다는 점에 주의해야 합니다.
데이터셋의 여러 피처 중, 특성 중요도가 높고 두 집단의 평균값에서 유의미한 차이가 난 두 피처는 HMG(혈색소)와 GAMMA_GTP(감마지티피)였으며,
대체로 흡연자들이 비흡연자들에 비해 높게 나타났습니다.

다만 산점도를 그려본 결과, 이 두 피처(HMG, GAMMA_GTP)만으로 흡연자/비흡연자를 구분해내는 것은 거의 불가능에 가까웠으며, 이 이외의 다른 피처들을 종합적으로 고려해야 위에서 언급한 정확도를 기대할 수 있을 것으로 판단됩니다.

또한, 예측 모델과 관련해서 여러 모델 간 앙상블(보팅)과 차원 축소 기법을 통해 모델의 성능 향상을 달성할 수 없었으며, 일부 피처에 대해 정상 범주 정보를 이용하여 스케일링한 결과 아주 약간의 성능 향상이 있었습니다.

결론

최종적으로 선택했던 모델은 lightGBM이며, 여러 개의 깊이가 얕은 의사결정나무를 이용하되, 이전 나무의 오답에 가중치를 부여하며 학습하는 부스팅 알고리즘이 사용된 모델입니다.

이 점은 XGBoost와도 같으나, lightGBM은 트리를 분할하는 과정에서
나무의 균형을 무시하고 비대칭적인 트리를 만든다는 차이점이 있습니다.

XGBoost와 비교하여 확실히 훈련 속도가 빨랐으며, 샘플 수가 적은 데이터셋에서 과적합이 발생하기 쉽다는 문제점이 있으나, 원본 데이터셋이 12만 행이나 되었기 때문에 이와 관 련해서 문제점은 없었습니다.

마지막으로, 이 데이터셋을 토대로 학습한 모델만을 가지고
'흡연이 건강에 명백하게 해롭다/아니다'를 논할 수는 없습니다.
하지만, 흡연자와 비흡연자 간에 아무런 차이가 없었다면 모델의 예측 정확도는 50%에 수렴했을 것입니다.

또한, 여러 부정적인 지표에서 흡연자 집단의 평균값이 근소하게라도 높게 나타난 것은 사실입니다.

이러한 점들을 고려해봤을 때,

"건강검진의 결과지 위에 흡연의 흔적은 분명히 남아있었다"

라고 말씀드릴 수 있을 것 같습니다.

데이터가 가진 한계

모델의 성능을 극대화하는 것이 이번 프로젝트의 전부는 아니었지만,
테스트 정확도가 80%에 미치지 못했던 것은 여러모로 아쉽습니다.

그래서 다시 한 번 데이터를 꼼꼼이 살펴보며,
왜 더 높은 정확도를 달성하지 못했는지 데이터 속에서 원인을 찾아보았습니다.

1. 흡연이 인체에 가장 큰 영향을 미칠 것으로 생각되는 부위는 폐와 기관지를 포함한 "호흡기"입니다.
   하지만, 해당 데이터 내에 호흡기 관련 피처는 포함되어 있지 않습니다.

2. 다 같은 흡연자라 할지라도, 하루에 담배를 얼마나 피우는 지에는 사람마다 큰 차이가 있습니다.
   하루에 반 갑도 안 피우는 사람이 있는 반면, 하루에 두 갑 넘게 피우는 사람도 있기 때문입니다.
   하지만, 해당 데이터 내에 하루 흡연량에 관한 정보는 담겨 있지 않습니다.

3. 건강에 악영향을 끼치는 요인에는 흡연만 있는 게 아닙니다. 건강하지 않은 식습관, 운동 부족, 지나친 음주, 수면 장애 등등 흡연 이외에 건강에 악영향을 끼칠 수 있는 요인은 무궁무진하게 많습니다.
   하지만, 해당 데이터 내에서 이러한 기타 요인들을 확인할 방법이 없습니다.

느낀 점 및 아쉬운 점

사실 지금까지 책으로만 머신러닝에 대해 공부하다가,
직접 데이터 분석을 해보고 싶어서 이번 대회에 참여하게 되었습니다.

그 과정에서 2만 행에 가까운 금연자 데이터를 어떻게 해야 할지 고민하기도 했고,
데이터 불균형을 줄여보려고 리샘플링을 했는데, 전체 데이터에 리샘플링을 하고 거기서 테스트 데이터를 추출하는 실수를 저질렀을 때, 갑자기 모델 점수가 너무 높게 나와서 당황하기도 했습니다.

또한, 더 나은 결측값 처리 방법을 찾지 못하고 결과적으로
전체 데이터의 3분의 1을 날려버린 셈이 된 것도 아쉽습니다.

이러한 실수들을 겪고 나니, 정말 데이터 분석은 경험이 중요하다는 것을 깨닫게 되었습니다.
지금 생각해보면 몇몇은 어처구니 없는 실수들이었지만, 그때만큼은 도저히 그 이유를 몰랐고
당황할 수 밖에 없었기 때문입니다.

그리고 책에서 배웠던 여러 다양한 머신러닝 기법들을 모두 한 번씩 써보고 싶었고,
극한의 모델 튜닝을 통해 최대한의 성능을 끌어내 보고 싶었지만,
컴퓨터 성능의 한계로 인해 많은 것을 포기하기도 했습니다.

느낀 점 및 아쉬운 점

하지만 프로젝트를 마치고 나니, 노력한 보람은 분명히 있었던 것 같습니다.

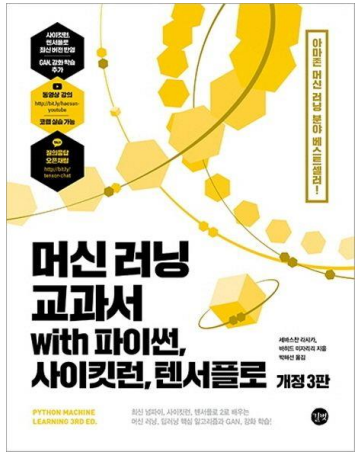
비록 원하는 성능의 모델을 만들어내거나, 완벽한 분석 결과를 내놓지는 못했지만,
그 과정에서 끊임없이 고민하고,
예상치 못한 문제를 맞닥뜨렸을 때 함께 머리를 맞대며 방법을 찾고,
매일 밤에 잠에 들기 직전까지 더 좋은 방법을 떠올리려 노력했던 것들이
다 좋은 경험으로 남은 것 같습니다.

또한 이 프로젝트에 대해 피드백을 받을 수 있다면 이를 발판 삼아 더 많은 대회에 도전하고,
더 많은 것을 탐구하고 분석하며, 앞으로 더 좋은 결과를 낼 수 있을 것 같습니다.

지금까지 이 긴 보고서를 읽어주셔서 감사합니다.

참고 도서

[머신 러닝 교과서
with 파이썬, 사이킷런, 텐서플로 개정판 3판]
(길벗출판사, 2021)



[Must Have 머신러닝·딥러닝 문제해결 전략]
(골든래빗, 2022)