

당뇨병 예측 모형 연구

· 목차

I. 서론

i) 프로젝트 배경

II. 본론

i) 데이터 탐색

- 데이터 수집
- 데이터 설명
- 데이터 전처리

ii) 변수 선택 및 설명

- 카이제곱을 통한 변수 유의성 확인
- 변수변환 및 이상치판단
- 변수 분포 확인

iii) 모델 적합

- 모델 후보군 선정
- 모델 성능 지표

III. 결론

i) 최종 모형 선택에 대한 종합적인 생각

I. 서론

i) 프로젝트 배경

뉴스	
<div><p>세계일보 PICK · 4일 전 · 네이버뉴스</p><p>세계 당뇨병 환자 8억명... 30여년전 대비 4배 증가</p><p>당뇨병 유병률도 7%에서 14%까지 치솟았다. 사진=EPA연합뉴스 WHO는 "1990년 이후 비만 증가와 건강에 해로운 음식의 소비 확대, 신체활동 부족, 경제적 어려움 등 복합적 요인으로 인해 당뇨병 환자가 놀라울 정도..."</p><p>WHO "세계 당뇨병 환자수 8억명 달해...30여년 전의 4배" 조세금융신문 · 4일 전</p><p>WHO "세계 당뇨 환자 8억명...30여년 전의 4배" KBS · 4일 전 · 네이버뉴스</p></div> <div></div> <div><p>[출처] https://www.segye.com/newsView/20241115509911?OutUrl=naver</p></div>	<div><p>탕후루 단짠단짠에 빠진 2030세대 '젊은 당뇨병' 급증... '혈당수치'는 몰라</p><p>서명윤 기자 2023.11.08 17:09 수정 2023.11.10 08:43 댓글 0</p><p>20대 당뇨병자 4년만에 47% 급증 2030세대 10명 중 4명 자신의 혈당수치 몰라 탄수화물, 설탕 등 과다 섭취 피해야</p><div></div><div><p>건강을 위한 건강한 뉴스</p><p>매경헬스</p><p>네이버 스탠드</p><p>구독하기</p></div><p>최신뉴스</p><ul style="list-style-type: none">· 한미합 식음료 '건강' 대체 - CJ, 저당 과자 2종 출시· 배상면주, 느린마음영조장 첫 제주도 매장 열· 환인 다량한 유방 통증, 치료 첫 단추는 '통증 주기 파악'· 작주합착중, 인기 어려울수록 가능한 집 안에서 끌어와<p>시원한 내마 마요 매 11월 11일 01:27</p></div> <div><p>[출처] https://www.mkhealth.co.kr/news/articleView.html?idxno=66426</p></div>

최근 디저트 식품이 큰 인기를 끌면서 당 함량이 높은 음식 섭취가 증가하고, 동시에 당뇨병과 관련된 뉴스와 연구 결과들이 자주 언급되면서 당뇨병이 개인의 건강뿐만 아니라 사회 전반에 미치는 심각한 영향을 다시 한번 생각하게 되었다. 특히, 당뇨병이 단순히 개인의 건강 문제로 끝나는 것이 아니라 의료 비용의 증가, 노동 생산성 저하, 삶의 질 저하 등 사회적으로도 큰 영향을 미친다는 점에서 그 중요성을 실감하게 되었다.

이러한 배경 속에서, 우리는 데이터 분석을 통해 당뇨병의 위험 요인을 파악하고 이를 기반으로 개인별 당뇨병 발병 가능성을 예측할 수 있는 모델을 개발하면 어떨까 하는 아이디어를 떠올리게 되었다. 이러한 모델은 당뇨병 예방과 관리를 위한 실질적인 도구로 활용될 수 있으며, 개인의 생활 습관 개선을 돕고 더 나아가 사회적 비용 절감에도 기여할 수 있을 것이라 기대하며 이 프로젝트를 시작하게 되었다.

– II. 본론

i) 데이터 탐색

• 데이터 수집

데이터 수집 사이트	데이터 유형					
	<p>SAS 파일 형식</p> <table><tr><td>2022</td><td>분과별 상세DB</td><td>영양조사</td><td>식품섭취조사(개인별 24 시회상조사)</td><td>HN22_24RC</td></tr></table> <p>hn22_24rc_240111.zip</p>	2022	분과별 상세DB	영양조사	식품섭취조사(개인별 24 시회상조사)	HN22_24RC
2022	분과별 상세DB	영양조사	식품섭취조사(개인별 24 시회상조사)	HN22_24RC		
• 출처: https://knhanes.kdca.go.kr/knhanes/sub03/sub03_02_05.do [국민건강영양조사]						

– 처음 당뇨병 데이터를 수집하기 위해서 여러가지 다음의 공식 웹사이트를 참고하였다.

- *HIRA 빅데이터 개방 포털: 국민건강보험 데이터를 제공함.
- *DATA.go.kr: 공공데이터 포털로 다양한 건강 관련 데이터를 확인가능.
- *국민건강영양조사: 본 연구에서 활용한 국민건강영양조사 데이터가 제공되는 사이트.

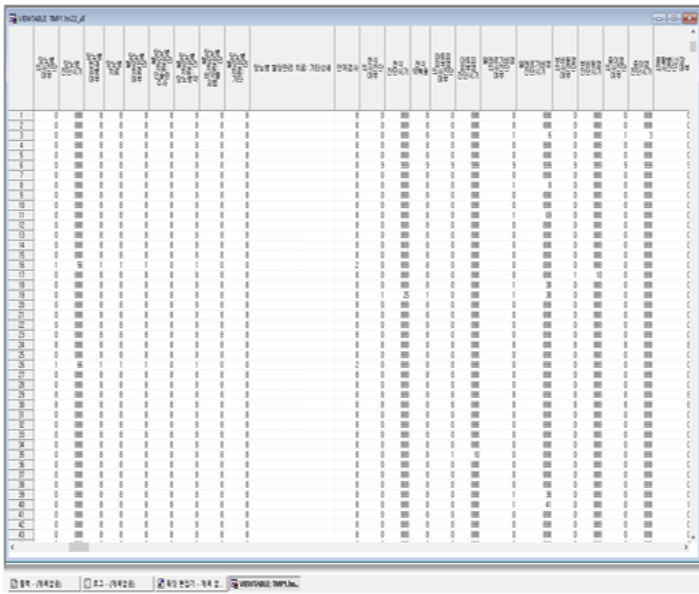
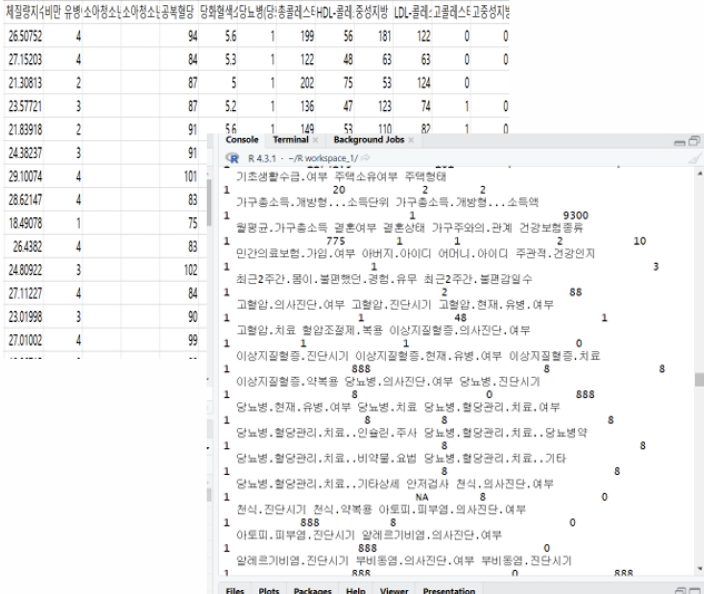
그 중 우리는 국민 건강 영양조사에서 데이터를 수집하는 것이 당뇨병 관련 지표 분석에 적합하다고 판단하였고, SAS파일(식품섭취조사)을 국민 건강 영양조사에서 수집하였다.

SAS 파일은 각 개인의 영양 섭취량과 건강 상태를 기록하고 있으며, 주요 변수는 다음과 같다

- *개인 ID: 설문에 참여한 각 개인을 식별하는 고유 번호
- *혈당 관련 지표: 공복 혈당(Fasting Glucose), 당화혈색소(HbA1c) 등
- *영양소 섭취량: 탄수화물, 지방, 단백질 섭취량
- *건강 상태: 당뇨병 진단 여부 및 기타 관련 변수

당뇨병 분석에 있어서 각 개인의 영양 섭취량과 건강 상태에 대한 정보가 핵심 요소라는 판단을 했기에, 본 보고서에서는 국민건강영양조사의 데이터를 활용하여 당뇨병 관련 지표를 분석하기 위한 데이터 수집 및 처리 과정을 정리하였다.

• 데이터 설명

데이터 원본 (변환 전)	데이터 원본 (변환 후)
	
[36816 rows x 180 columns]	

본 연구에서 사용된 데이터는 SAS 전용 형식의 데이터로, Python과 R을 이용한 분석을 위해 CSV 파일로 변환하는 과정이 필요하였다. 데이터셋에는 주요 지표로 공복혈당(fasting blood glucose)과 당화혈색소(glycated hemoglobin, HbA1c)가 포함되어 있다. 당뇨병 진단 기준은 공복혈당이 126 mg/dL 이상이거나 당화혈색소가 6.5% 이상일 때로 정의된다. 따라서, 본 연구에서는 이 두 생리적 지표를 반응변수로 설정하여 당뇨병의 발병 여부를 예측하는 분석을 진행하였다.

이 데이터는 국민건강보험공단에서 제공한 2022년 국민건강영양조사 데이터를 기반으로 한다. 2022년 데이터는 총 6,265개의 행과 623개의 설명변수로 구성되어 있었으며, 데이터 전처리 과정에서 발생한 결측값 처리 및 기타 데이터 손실을 보완하기 위해 2018년부터 2022년까지의 기간에 해당하는 데이터를 통합하는 작업이 진행되었다. 이를 통해 최종적으로 180개의 설명변수를 확보하였고, 이 과정에서 443개의 변수는 손실되었다.

그럼에도 불구하고, 의학적 지식과 당뇨병 예측의 핵심 요소를 고려한 결과, 주요한 변수들이 포함되어 있어 설명변수의 손실에 대한 우려는 최소화되었음. 본 연구에서는 2022년부터 2018년까지의 통합 데이터를 바탕으로 총 36,816개의 샘플을 확보하게 되었으며, 이는 기존 데이터에 비해 30,551명의 추가 데이터를 포함한 결과이다. 이러한 확장된 데이터셋을 통해 보다 정확하고 신뢰할 수 있는 당뇨병 예측 모델을 구축할 수 있는 기반을 마련할 수 있었다.

특히, 당뇨병 예측을 위한 분석은 공복혈당과 당화혈색소 외에도 다양한 생활습관, 유전적 요인, 식습관 및 체질량지수(BMI) 등의 설명변수들이 중요한 역할을 할 수 있음을 염두에 두고, 종합적인 데이터 분석을 통해 더욱 정밀한 예측이 가능하도록 하였다. 이러한 다각적인 접근은 당뇨병의 예방 및 관리에 중요한 기여를 할 수 있을 것으로 기대가 된다.

• 데이터 전처리

데이터 전처리

데이터 전처리 코드

```

# 만나이구간 생성
def categorize_age(age):
    if age < 40:
        return 0
    elif 40 <= age < 64:
        return 1
    else: # 65세 이상
        return 2

data['만나이구간'] = data['만나이'].apply(categorize_age)

# BMI구간 생성
def categorize_bmi(bmi):
    if bmi < 18.5:
        return 0
    elif 18.5 <= bmi < 22.9:
        return 1
    elif 23.0 <= bmi < 24.9:
        return 2
    elif 25.0 <= bmi < 29.9:
        return 3
    else: # 30 이상
        return 4

data['BMI구간'] = data['BMI'].apply(categorize_bmi)

# 결과 출력
print(data)

```

당뇨병유발	만나이	성별	BMI	근력운동	당뇨병부도	현재흡연율	아버지당뇨병	어머니당뇨병	두분다유전	#
0	0	56	2	26.507517	0.0	1.0	0.0	0.0	0.0	
1	0	30	1	27.152029	NaN	2.0	0.0	1.0	0.0	
2	0	25	2	21.308131	NaN	2.0	0.0	1.0	0.0	
3	0	66	1	29.577937	0.0	1.0	0.0	0.0	0.0	

데이터 전처리 결과

당뇨병유발	만나이구간	BMI구간	성별	근력운동	당뇨병부도	현재흡연율	아버지당뇨병	어머니당뇨병	두분다유전
0	1	3	2	0	1	0	0	0	0
0	2	2	1	0	1	0	0	0	0
0	1	1	2	1	1	0	0	0	0
0	1	3	2	0	1	0	0	0	0
0	0	3	2	0	1	0	0	0	0
0	1	0	2	1	1	0	0	0	0
0	2	2	2	0	1	0	0	0	0
0	1	3	2	0	1	0	0	0	0
0	1	2	2	0	1	0	0	0	0
0	1	3	1	0	3	0	0	1	0
0	1	1	2	0	2	0	1	0	0
1	2	2	1	1	1	0	0	0	0
0	1	2	2	0	1	0	0	0	0

본 프로젝트에서는 로지스틱 회귀 모델을 활용하여 당뇨병 예측을 진행할 예정이며, 이에 따라 반응변수의 이진화가 필요하다. 앞서 설명한 바와 같이, 당뇨병을 판단하는 주요 지표는 공복혈당(fasting blood glucose)과 당화혈색소(HbA1c)라는 연속형 변수이다. 우리는 이 두 변수의 값을 기반으로 이진 변수를 생성하는 과정에 착수하였다.

두 지표 모두 결측값이 존재할 수 있으며, 한 항목에만 결측값이 있을 수도 있다. 이를 처리하기 위해 다음과 같은 기준을 적용하였다. 만약 두 변수 모두 결측값이 아닌 완전한 값으로 존재한다면, 두 지표 중 하나라도 당뇨병 진단 기준을 만족하는 경우, 즉 공복혈당이 126 mg/dL 이상이거나 당화혈색소가 6.5% 이상이라면 새로운 컬럼인 ‘당뇨병 유발’을 생성하여 1(양성) 값을 할당하였다.

만약 한 항목에만 결측값이 있고 다른 항목이 완전한 값으로 제공되었을 경우, 그 값이 당뇨병 기준을 충족한다면 당뇨병 유발 컬럼에 1을 할당하고, 그렇지 않으면 해당 샘플을 제거하는 방식으로 처리했다. 결측값을 단순히 ‘음성’으로 처리할 수 없었던 이유는, 결측값이 실제로 당뇨병 기준을 만족할 가능성을 배제할 수 없기 때문이다. 즉, 한 항목의 결측값이 당뇨병을 음성으로 판단할 수 없다는 것을 고려했다.

마지막으로, 두 변수 모두 결측값인 경우에는 해당 행을 제거하여 데이터 품질을 확보했다. 이와 같은 데이터 전처리 과정을 거친 후, 최종적으로 28,350개의 행이 남았으며, 8,446명의 데이터 손실이 발생하였다. 비록 상당한 데이터 손실이 있었지만, 모델의 예측 성능을 극대화하기 위해 이러한 손실을 감수했다. 데이터 손실을 최소화하면서도 정확한 예측을 위한 데이터 정제는 머신러닝 모델의 신뢰성을 높이는 중요한 과정임을 인식하고, 모델 성능 향상을 위한 최적의 선택이라 판단하였다.

ii) 변수 선택 및 설명

• 카이제곱 유의성 확인

카이제곱 유의성 확인

코드 및 결과

```
[172]: import numpy as np
import pandas as pd
from scipy.stats import chi2_contingency

# 데이터 설정
data = np.array([[1905, 311], # 남자: 당뇨병유발 == 0, 당뇨병유발 == 1
                 [2612, 269]]) # 여자: 당뇨병유발 == 0, 당뇨병유발 == 1

# 카이제곱 검정 수행
chi2, p, dof, expected = chi2_contingency(data)

# 결과 출력
print(f'Chi-squared statistic: {chi2}')
print(f'P-value: {p}')
print(f'Degrees of freedom: {dof}')
print(f'Expected frequencies:\n{expected}')

# 유의수준 설정 (예: 0.05)
alpha = 0.05

# 결과 해석
if p < alpha:
    print("남자와 여자의 당뇨병 유발 비율에 차이가 있습니다.")
else:
    print("남자와 여자의 당뇨병 유발 비율에 차이가 없습니다.")

Chi-squared statistic: 26.9417515896075
P-value: 2.0968006799482097e-07
Degrees of freedom: 1
Expected frequencies:
[[1963.83598195  252.16401805]
 [2553.16401805  327.83598195]]
남자와 여자의 당뇨병 유발 비율에 차이가 있습니다.
```

*카이제곱 검정 (Chi-square test)**은 두 범주형 변수 간의 관계 또는 독립성을 평가하는 통계적 방법이다. 주로 범주형 데이터에서 사용되며, 변수들 간에 상관 관계가 존재하는지 또는 독립적인지를 확인하는 데 활용된다. 카이제곱 검정은 두 가지 주요 유형으로 나눌 수 있다. 본 연구에서는 카이제곱 검정을 통해 두 변수 간 동질성의 차이가 유의미한지 여부를 확인하고, 만약 동질성 차이가 유의미하다면 해당 변수가 반응변수를 설명할 수 있는 중요한 변수로 활용될 수 있음을 확인하였고, 이를 통해 5개 주요 변수에 대한 유의미성을 평가할 수 있었다.

*성별: 남성과 여성 간의 당뇨병 발병률 차이는 통계적으로 유의미한 차이가 있었다. 분석 결과, 남성이 여성보다 당뇨병 발병 비율이 더 높다는 것을 확인할 수 있었다.

*만나이: 만나이와 공복혈당 사이에는 양의 선형 관계가 존재하였으며, 이는 나이가 많을수록 공복혈당 수치가 높아지는 경향이 있다는 것을 의미한다. 로지스틱 회귀 모델의 보편화를 위해, 연속형 변수인 만나이를 3개의 구간(40세 미만, 40세 이상 64세 이하, 65세 이상)으로 구간화하여 가변수화하였다. 이 구간화된 변수 간의 동질성 차이는

통계적으로 유의미하며, 나이가 많을수록 당뇨병 발병률이 증가하는 경향이 확인되었다.

BMI: 본 데이터에는 BMI 변수는 없으나, 키와 체중 데이터를 바탕으로 BMI 값을 새롭게 계산하여 활용하였다. BMI와 공복혈당 사이에도 양의 선형 관계가 존재하였고, 로지스틱 회귀 모델을 최적화하기 위해 BMI 값을 4개의 구간(30 이상, 25.0 이상 29.9 이하, 23.0 이상 24.9 이하, 18.5 이상 22.9 이하, 18.5 이하)으로 구간화하였다. 각 구간 간의 동질성 차이는 통계적으로 유의미했으며, BMI가 높을수록 당뇨병 발병률이 증가함을 확인할 수 있었다.

흡연: 흡연자와 비흡연자 간의 당뇨병 발병률 차이는 통계적으로 유의미한 차이를 보였으며, 흡연자가 비흡연자보다 당뇨병 발병율이 유의미하게 높다는 결과를 얻었다.

운동: 운동을 자주 하는 사람과 그렇지 않은 사람 간의 동질성을 비교한 결과, 운동을 자주 하지 않는 사람들이 운동을 자주 하는 사람들보다 당뇨병 발병률이 유의미하게 높다는 결과가 도출되었다.

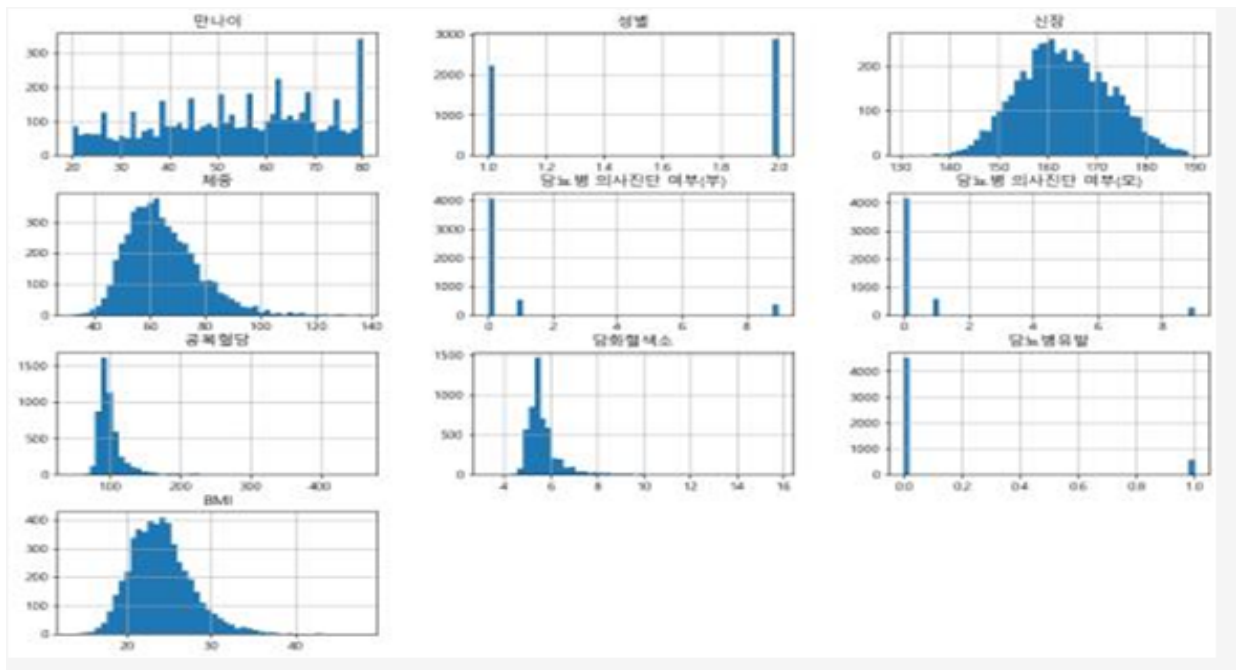
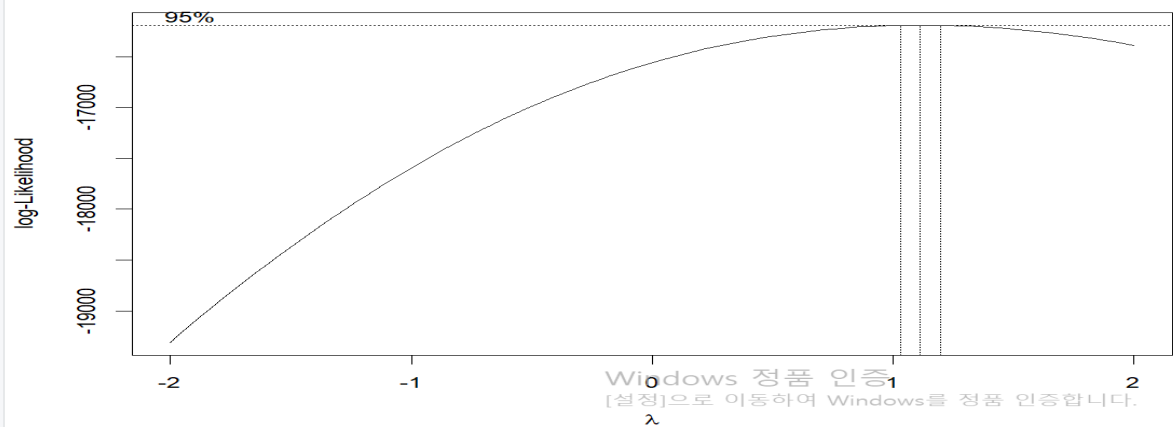
유전: 부모님이 당뇨병이 있을 때의 발병률을 분석한 컬럼입니다. 부모님 중 아버지만 당뇨병일 때, 어머니만 당뇨병일 때, 부모님 모두 당뇨병일 때, 부모님 모두 당뇨병이 아닐 때 총 4개의 범주가 존재합니다. 특히, 어머니만 당뇨병일 때가 아버지만 당뇨병일 때보다 통계적으로 유의미하게 높은 발병률을 보였으며, 이는 당뇨병의 유전성이 모계 유전이 더 강하다는 기존 의학적 지식과 일치하는 결과입니다. 이 변수는 하나의 변수로만 사용될 경우 각 집단 간 차이가 동일하게 나타날 수 있어, 세분화하여 3개의 변수로 분할하여 분석하였다.

이처럼, 카이제곱 검정을 통해 통계적으로 유의미한 동질성 차이가 발견된 변수들을 설명변수로 포함시키는 것은 당뇨병 예측 모델의 설명력을 높이는 데 중요한 요소입니다. 그러나 변수의 수가 많아질수록 모델의 복잡성이 증가하고, 모델의 해석 가능성이나 실제 적용 가능성에 문제를 일으킬 수 있다. 또한, 데이터에 대한 충분한 의학적 이해가 없으면, 특정 변수에 대해 잘못된 해석이나 과도한 변수 선택이 발생할 수 있다.

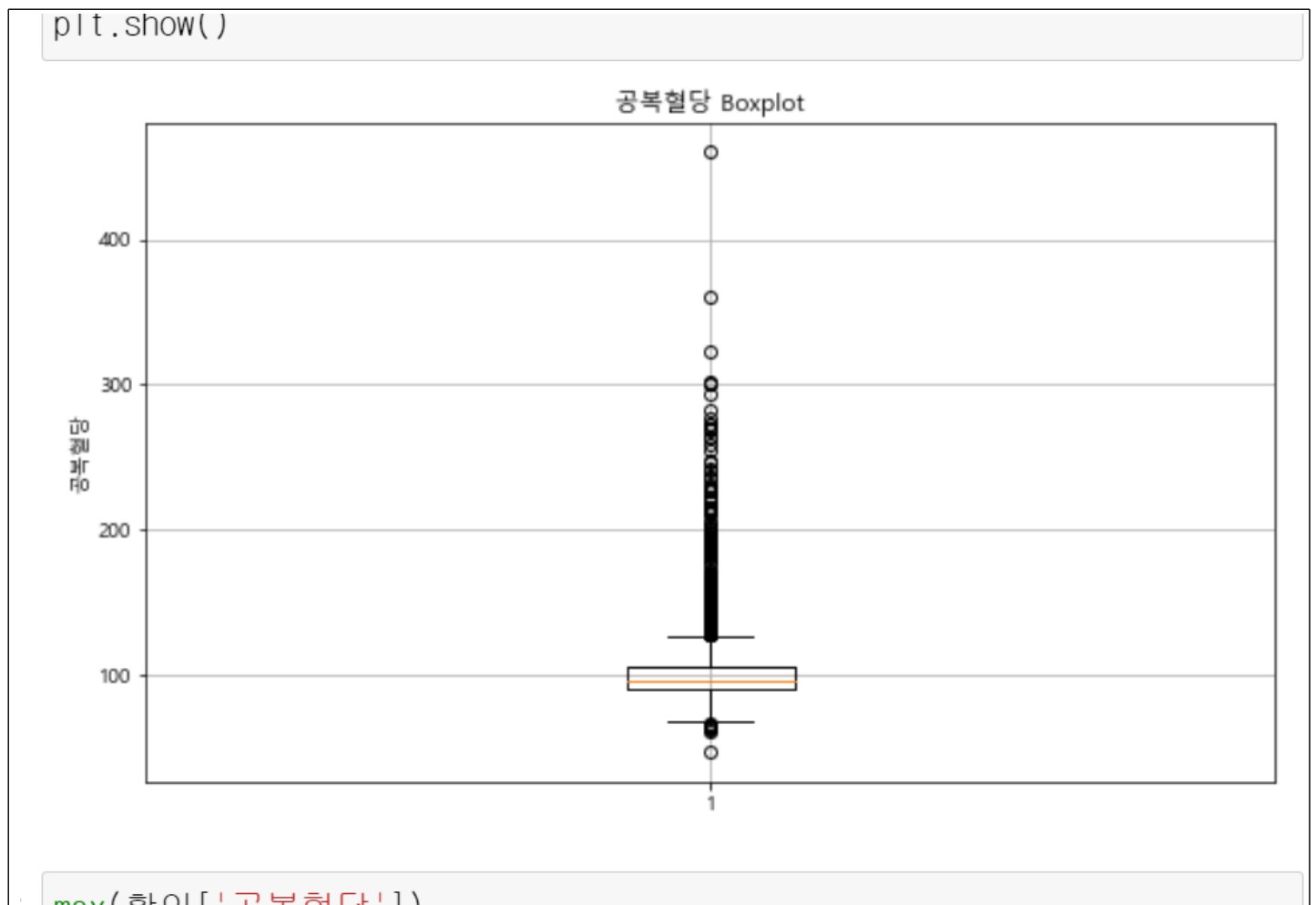
예를 들어, 음주에 대한 통계 분석 결과, 음주를 하지 않는 사람들이 음주를 하는 사람들보다 당뇨병 발병율이 더 높게 나타났다. 하지만 이는 기존 연구와 상반되는 결과로, 기존의 연구에서는 음주가 적당히 이루어졌을 때 당뇨병 발병율에 긍정적인 영향을 줄 수 있다는 것이 알려져 있다. 따라서, 통계적으로 유의미하더라도 기존 연구의 방향성을 바탕으로 이 변수를 설명변수에서 제외하기로 결정하였다.

• 변수변환 및 이상치 판단

변수 변환



이상치 판단



데이터의 비정규성은 통계 분석, 특히 회귀 분석에서 모델의 정확도와 신뢰성에 부정적인 영향을 미칠 수 있다. 이러한 이유로, 우리는 변수 변환을 통해 데이터의 분포를 정규 분포에 가깝게 조정하여 분석의 정확도를 높이려고 했다. 변수 변환 방법에는 여러 가지가 있지만, 그중에서 대표적인 방법들을 살펴보자.

로그 변환 (log transformation): 주로 양의 값을 가지며, 오른쪽으로 치우친 분포를 정규화하는 데 효과적이다.

제곱근 변환 (square root transformation): 분포가 비대칭적인 경우, 특히 작은 값들이 많을 때 유용하다.

Box-Cox 변환: 최적의 λ 값을 찾아 적용하는 변환 방법으로, 모든 양의 연속형 변수에 적용 가능하다.

변수 변환의 필요성 및 시각화

우리는 각 변수의 분포를 시각화하여 적합한 변환 방법을 결정하기 위해 히스토그램을 활용했다. 이를 통해 각 변수의 분포 상태를 분석하였다.

첫 번째 행: 연령, 성별, 신장

연령: 연속형 변수로, 특정 연령대(중년층)에 데이터가 집중되어 있음을 확인했다. 이는

모집단의 특성상 중년층이 많다는 것을 반영하며, 이 분포는 정규성을 갖추기 어려운 특징을 보인다.

성별: 이진형 변수(0, 1)로, 남성과 여성이 각각 나타낸다. 데이터가 균등하게 분포되지 않음을 알 수 있다. 이는 성별이 두 집단으로 나뉘어 있는 특성상 자연스러운 현상이다.

신장: 신장은 대체로 정규 분포에 가까운 분포를 보였지만, 오른쪽 꼬리가 약간 나타나는 경향을 보였다. 이에 대해 변환이 필요하지 않다고 판단하였다.

두 번째 행: 체중, 공복 혈당, 당화혈색소

체중: 오른쪽으로 치우친 분포를 보였으며, 이는 분포가 비대칭적이라는 특성을 나타낸다. 이를 해결하기 위해 로그 변환이나 제곱근 변환을 고려할 수 있다.

공복 혈당: 특정 구간(정상 범위)에 데이터가 집중되어 있으며, 일부 높은 혈당 값을 가진 데이터가 꼬리를 형성하고 있다. 이 경우 Box-Cox 변환이 유용할 수 있다.

당화혈색소 (HbA1c): 당뇨병 진단의 핵심 지표인 당화혈색소는 오른쪽으로 치우친 분포를 보였다. 따라서 로그 변환이 효과적일 수 있다.

세 번째 행: 운동량, 총 에너지 섭취량

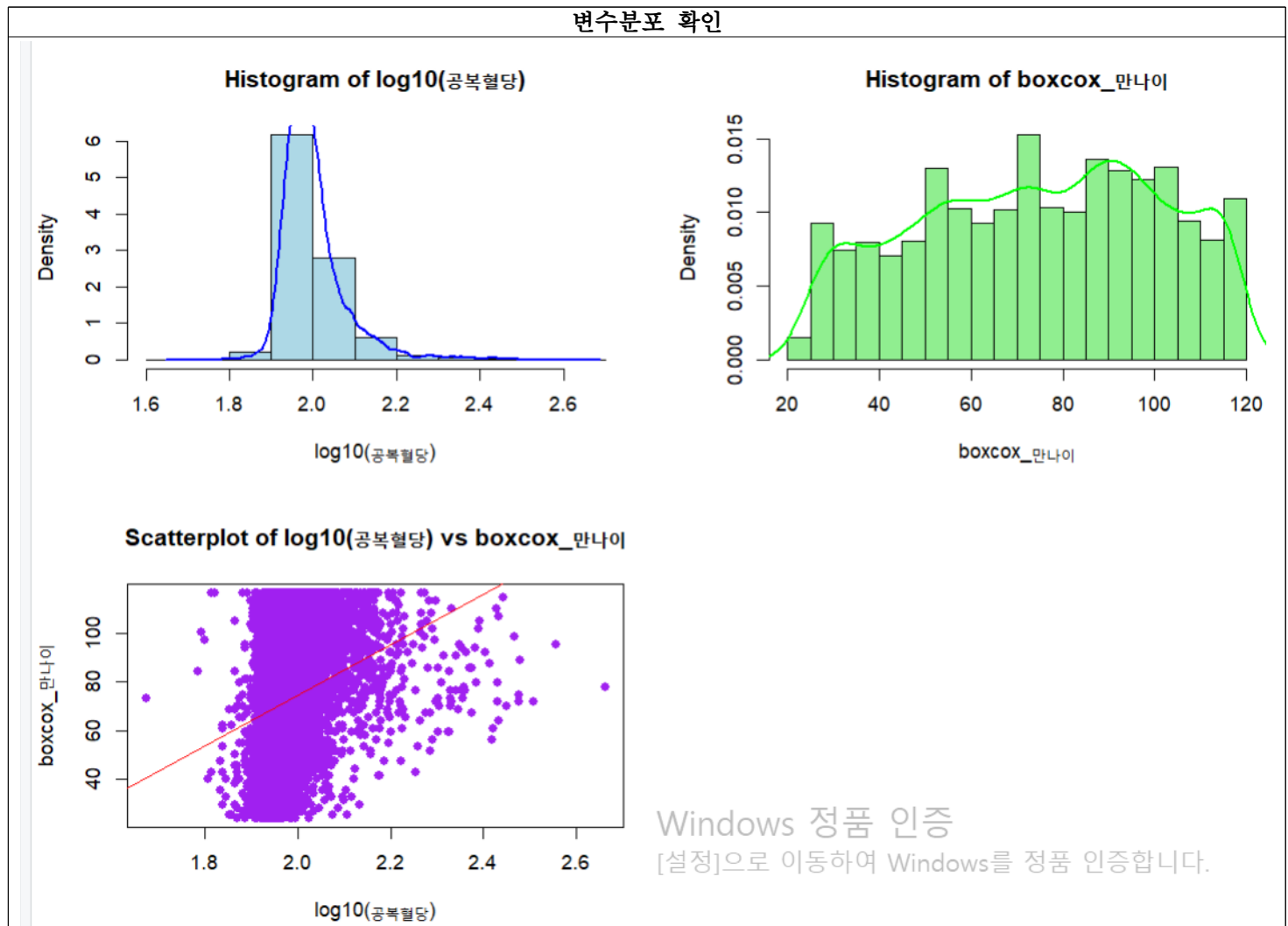
운동량: 많은 데이터가 낮은 값에 몰려 있으며, 일부 극단적인 값들이 꼬리를 형성하고 있다. 이 변수에 대해 Box-Cox 변환을 적용하여 최적의 λ 값을 찾는 것이 유효할 수 있다.

총 에너지 섭취량: 상대적으로 대칭적인 분포를 보였지만, 여전히 극단값이 존재할 수 있다. 따라서 극단값 제거 또는 변환을 고려해야 할 수 있다.

Box-Cox 변환을 통한 최적 λ 값 찾기

Box-Cox 변환에서는 최적의 λ 값을 찾아 적용하는 것이 중요하다. 로지스틱 회귀 모델에서는 정규성 가정이 필수적이기 때문에, 설명변수의 분포를 확인하고 이를 정규화하는 과정이 필요하다. 그러나 최적 λ 값을 찾더라도 완전한 정규 분포를 얻기는 어렵다는 점을 알게 되었다. 이러한 결과는 정규성에 대한 허용 범위를 만족시키지 못할 가능성이 크다는 판단을 하게 만들었다.

- 변수분포 확인



정규화 대신 가변수화 적용

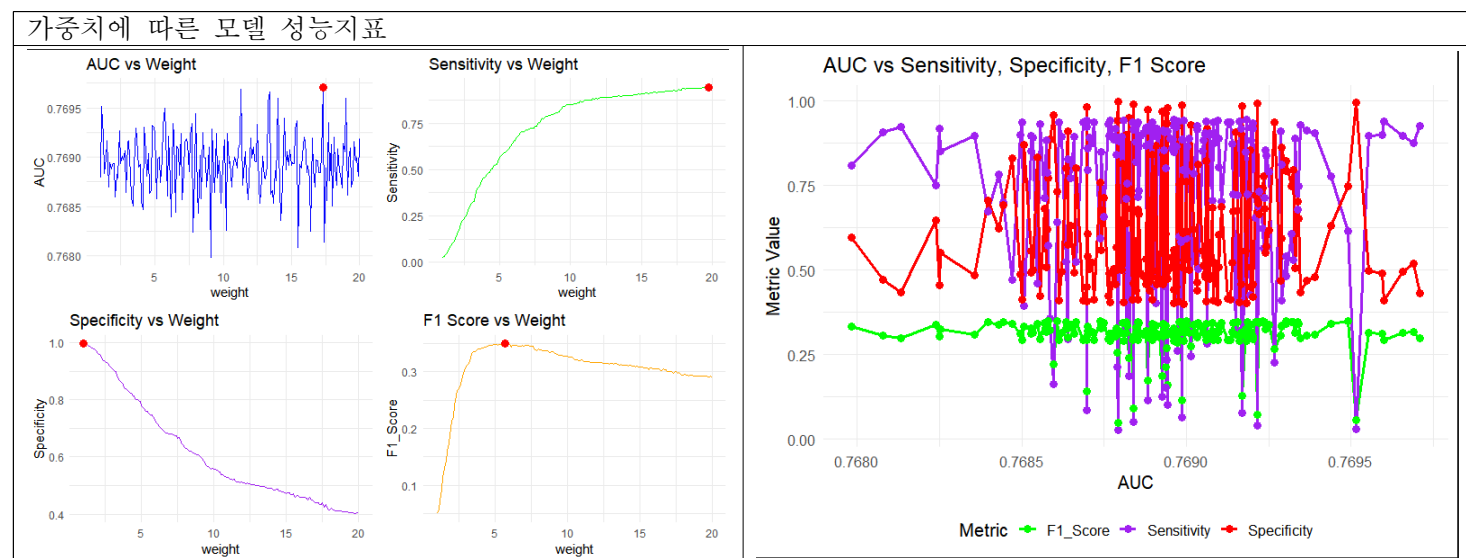
결국, 우리는 변수의 정규화를 진행하는 대신 가변수화 방법을 채택하였다. 연속형 변수들을 가변수화하면 로지스틱 회귀 모델에서 변수의 분포를 고려할 필요가 없으며, 모델의 과적합(overfitting)이나 편향(bias) 문제를 완화할 수 있다. 또한, 가변수화된 변수는 회귀 분석에 더 적합한 형태로 변환되어 모델의 성능을 향상시킬 수 있다.

또한 가변수화는 구간마다 모두 동일한 비중을 부여하므로 이상값에 대해 덜 민감하기에 이상값을 제거하지 않아도 되는 장점이 있다. 이상값을 제거하지 않아도 되어 데이터의 손실을 최소화 할 수 있다는 장점 또한 존재한다.

따라서, 우리는 연속형 변수들의 정규화 작업을 생략하고, 대신 해당 변수들을 가변수화하여 분석을 진행하였다. 이 접근법은 모델의 해석력을 높이며, 의학적 해석에도 더 유용하게 작용할 수 있다.

결론적으로, 변수 변환을 통해 데이터의 분포를 정규화하는 것만으로는 항상 최적의 결과를 얻을 수 없다는 것을 확인했다. 대신, 변수들을 가변수화하는 방법이 로지스틱 회귀 모델에서 더욱 유효하게 작용했으며, 이는 모델의 성능 향상뿐만 아니라 데이터의 해석 가능성 또한 높이는 결과를 가져왔다.

iii) 모델 적합



• 모델 후보군 선정

본 팀이 보유한 당뇨병 데이터의 양성 사례와 음성 사례 비율은 약 1.5:8.5로, 모델이 양성 클래스에 대해 제대로 학습하기 어려운 경향이 있다. 이에 따라, 우리는 양성 클래스에 가중치를 부여할 필요가 있었고, 1부터 20까지 가중치를 0.1씩 증가시키면서 총 200개의 모델을 만들었다. 이 모델들 중에서 AUC, 민감도, 특이도, F1 스코어가 각각 최고인 모델을 선정해야 했다.

가중치를 증가시킬수록 민감도는 계속해서 증가하고, 특이도는 감소하는 경향이 나타났다. 따라서 민감도나 특이도가 최상의 모델은 최종 후보군에서 제외되었다. 결국, 우리는 AUC 또는 F1 스코어가 최상의 모델을 최종 모델로 선정하기로 했다.

AUC는 가중치 값에 따라 일정하지 않게 변동했으며, 200개의 모델 중 AUC의 최대값과 최소값의 차이가 0.01 이하로 매우 작은 경향을 보였다. AUC가 최대인 모델을 분석한 결과, 민감도가 매우 낮은 것으로 나타났다. 즉, AUC 값을 일부 손실하더라도 민감도를 높이는 것이 필요하다고 판단되었다. AUC의 손실이 미미하므로, 이 손실을 감수하는 것은 큰 문제가 되지 않는다고 생각했다.

F1 스코어는 최종 후보군 4개 모델 중에서 민감도와 특이도가 가장 균형 잡힌 모델을 나타냈다. 민감도가 0.5, 특이도가 0.75인 모델이었지만, 민감도가 0.5로 낮아 최종 모델로 선정하는 데 고민이 있었다. 이에 우리는 추가적인 두 가지 제약 조건을 설정하여, 모델을 재구성했다.

- 1.민감도가 60 이상, 특이도가 70 이상이어야 한다.
- 2.민감도가 70 이상, 특이도가 60 이상이어야 한다.

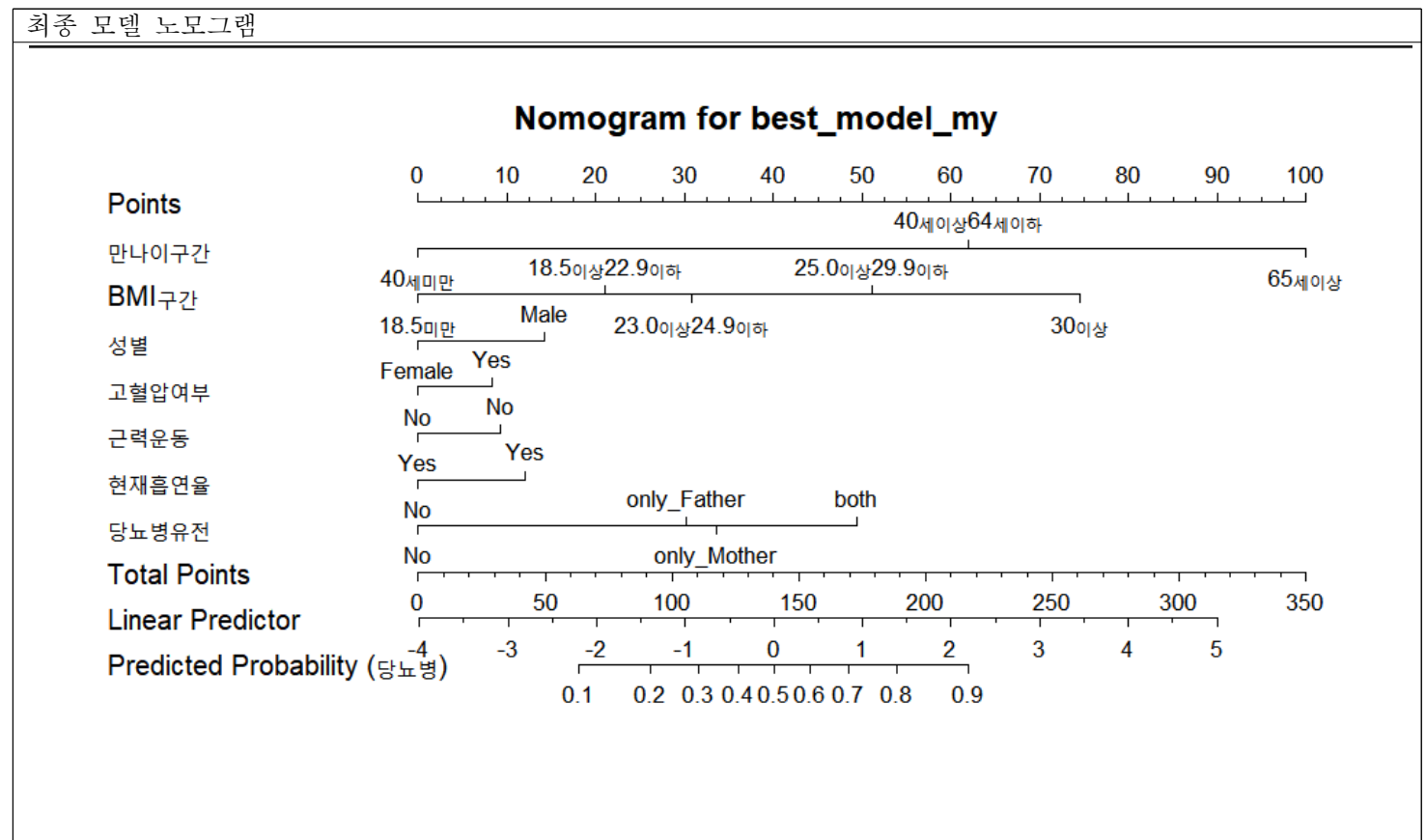
위 두 제약 조건을 모두 만족하는 모델들 중, AUC가 최상의 모델을 선택했다. 그 결과, 첫 번째 제약 조건에서는 민감도 0.6, 특이도 0.75를 가진 모델이 선정되었고, 두 번째 제약 조건에서는 민감도 0.7, 특이도 0.66을 기록한 모델이 선정되었다. 따라서 이 두 모델이 최종 모델 후보군으로 결정되었다.

• 모델 성능 지표

모델 성능 지표	
<pre> Reference Prediction No Yes No 2897 199 Yes 955 309 Accuracy : 0.7353 95% CI : (0.722, 0.7484) No Information Rate : 0.8835 P-Value [Acc > NIR] : 1 Kappa : 0.2189 McNemar's Test P-Value : <2e-16 Sensitivity : 0.60827 Specificity : 0.75208 Pos Pred Value : 0.24446 Neg Pred Value : 0.93572 Prevalence : 0.11651 Detection Rate : 0.07087 Detection Prevalence : 0.28991 Balanced Accuracy : 0.68017 'Positive' Class : Yes > > cat("best_weight_my2 : ",best_weight_my,"\n") best_weight_my2 : 5.8 > cat("best_auc_my2 : ",best_auc_my,"\n") best_auc_my2 : 0.7691117 > print("best_cm_my2") [1] "best_cm_my2" > print(best_cm_my2) Confusion Matrix and Statistics Reference Prediction No Yes No 2578 152 Yes 1274 356 Accuracy : 0.6729 95% CI : (0.6588, 0.6869) No Information Rate : 0.8835 P-Value [Acc > NIR] : 1 Kappa : 0.1889 McNemar's Test P-Value : <2e-16 Sensitivity : 0.70079 Specificity : 0.66926 </pre>	<pre> > summary(best_model_my) Call: NULL Coefficients: Estimate Std. Error z value Pr(> z) (Intercept) -3.34076 0.12030 -27.769 < 2e-16 *** 만나이구간40세이상64세이하 1.77560 0.04833 36.741 < 2e-16 *** 만나이구간65세이상 2.86232 0.05302 53.984 < 2e-16 *** BMI구간18.5이상22.9이하 0.60031 0.11277 5.323 1.02e-07 *** BMI구간23.0이상24.9이하 0.88336 0.11364 7.773 7.65e-15 *** BMI구간25.0이상29.9이하 1.46431 0.11216 13.055 < 2e-16 *** BMI구간30이상 2.13276 0.11748 18.154 < 2e-16 *** 성별Female -0.40705 0.03090 -13.171 < 2e-16 *** 고혈압여부Yes 0.23663 0.03560 6.647 3.00e-11 *** 근력운동Yes -0.26474 0.03882 -6.820 9.11e-12 *** 현재흡연율Yes 0.34402 0.03910 8.798 < 2e-16 *** 당뇨병유전only_Father 0.86503 0.04763 18.163 < 2e-16 *** 당뇨병유전only_Mother 0.96122 0.04256 22.586 < 2e-16 *** 당뇨병유전both 1.41264 0.09323 15.152 < 2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for binomial family taken to be 1) Null deviance: 37214 on 17439 degrees of freedom Residual deviance: 30368 on 17426 degrees of freedom AIC: 30929 Number of Fisher Scoring iterations: 5 </pre>

위의 그림은 본 팀이 만든 모델이 성능 지표를 요약한 내용이다. 각 모델에 따른 정확도 민감도 특이도 또한 확인이 가능하다. 각 변수의 회귀 계수(Esimatae) 값을 통해 변수들이 당뇨병 발병에 미치는 영향을 확인할 수 있다. p값이 모두 0.05보다 작아 모든 회귀계수가 유의함을 알 수 있다. 먼저, 나이가 많아질수록 당뇨병 발병 확률이 유의미하게 높아지는 것을 확인할 수 있으며, BMI 역시 증가할수록 발병 확률에 긍정적인 영향을 미친다. 특히, BMI가 30 이상인 경우가 계수가 가장 높아 당뇨병 위험도가 크다는 점을 보여준다.L 성별에서는 남성이 여성보다 발병 위험이 높고, 고혈압 여부와 운동 유무 역시 중요한 요인으로 작용하고 있다. 마지막으로, 유전적 요인에서는 어머니가 당뇨병일 경우, 아버지가 당뇨병인 경우 보다 더 높은 영향을 미치는 것으로 나타난다. 이와 같이, 우리 모델은 주요 변수들과 당뇨병 발병 간의 관계를 잘 설명하고 있으며, 높은 신뢰성을 갖추고 있다고 판단한다.

III. 결론 (모형 선택)



i) 최종 모형 선택에 대한 종합적인 생각

이제 우리는 두 모형 중에서 민감도에 더 높은 우선순위를 두어야 할지, 아니면 특이도에 더 높은 우선순위를 두어야 할지 결정해야 합니다. 일반적으로 의학적 예측 모형에서는 민감도가 더 중요한 의미를 가지지만, 본 연구에서는 첫 번째 모형이 더 높은 정확도를 보이기 때문에 민감도를 일부 포기하더라도 더 높은 정확도를 보장하는 방향으로 선택하였다. 따라서 최종 모형은 민감도 0.6, 특이도 0.75를 보이는 모형로 결정되었다.