



제 2회 데이터과학경진대회 데이터분석부문

Data Driven Decision을 자신의 삶에 적용하는 방법: 보건분야 빅데이터를 통한 보험상품의 개인화 전략

▶) 보고서에 음성이 포함되어 있습니다 프레젠테이션 모드로 소리를 켠 상태로 감상해주세요 ☺

☞ 우측 상단에 있는 QR코드를 카메라 앱으로 촬영하면 서비스 프로토타입 웹사이트로 연결됩니다.

길하균, 박지훈, 서형준, 조성민
(주) 유전생명



출처: 국민건강영양조사 분석결과; Plot1

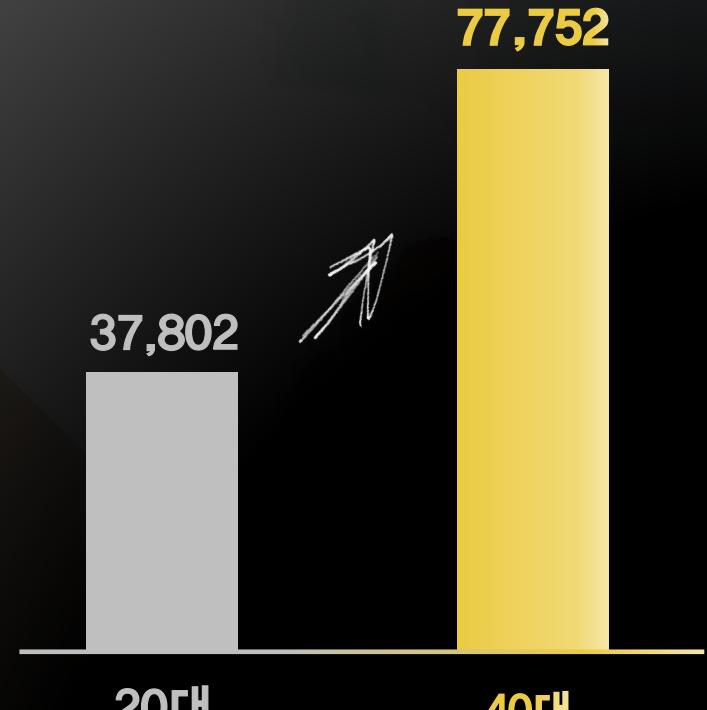


출처: 만성질환 현황과 이슈; 질병관리청(2021)

저는 건강해서 보험 필요 없는데요?

단위: 월 납입금액(원)

지금이 가장 저렴할 때입니다!



연령별 최소 보험료 증가율

출처: 삼성화재 자녀보험 마이 슈퍼스타

일리 있네요. 하지만 지금 당장 돈이 없어요



보장은 적은데 보험료는 많아요

소득이 불균형한 **20대**, 고정적인 비용을 언제 필요할지 모를 의료비 보장을 위한 상품에 투자하는 것은 어렵습니다.

많은 건강보험상품은 자산 증여 목적으로 설계되어 판매되고 있습니다. 때문에 피보험자의 **나이**와 **성별** 외에 **보험비**와 **보장내역**을 판단하는 기준이 없습니다.

보건분야 빅데이터를 통해 보다 **정밀**하고 **개인화**된 **보험설계**를 만들 수 있진 않을까요?

분류기 에서의 질병 예측 확률



자신의 **리스크를 야해하고, 보험을 통해
최소한의 대비는 할 수 있는 세상을 만듭니다.**

– 토스인슈어런스 –

만성질환의 종류와 비용이 다양하기 때문에, 일반적인 생명보험 주계약에서 잘 다뤄지지 않습니다. 때문에 보장을 위해서는 **특약을 통해 보장을 확대해야 합니다**. 저희는 기존의 보험(삼성화재 마이슈퍼스타)에서 보험특약 조건을 **4가지 취약성**에 맞게 분류했습니다.

대사질환

심혈관질환

뇌혈관질환

간질환

정교한 빅데이터를 통한 건강 리스크 평가

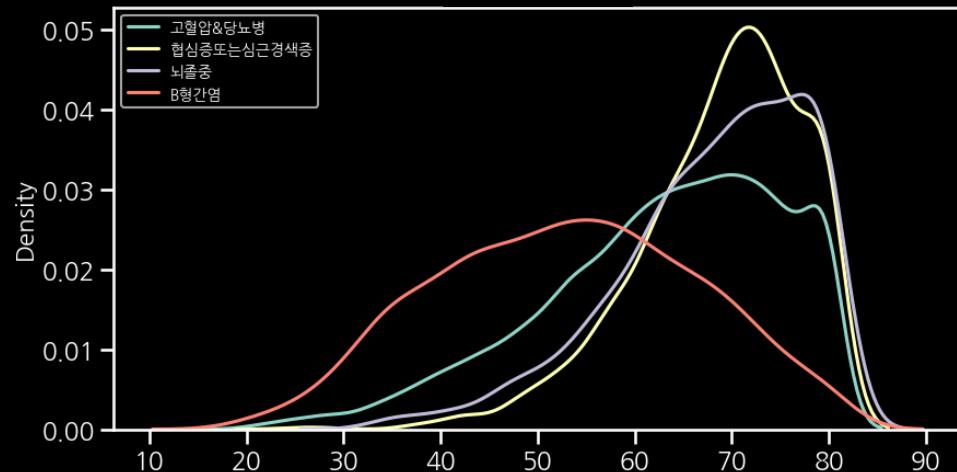
국민건강영양조사 데이터는 20년 동안 축적되어 온 한국인의 영양보건분야의 빅데이터입니다. 조사과정부터 세심하게 설계된 데이터는 1865건이나 되는 논문에 인용되어, 그 정교함을 인정 받았습니다. 분류된 **취약성**에 대응되는 **질병 유병 여부**를 분석목표로 분류모델을 학습 시켰습니다.

고혈압&
당뇨병

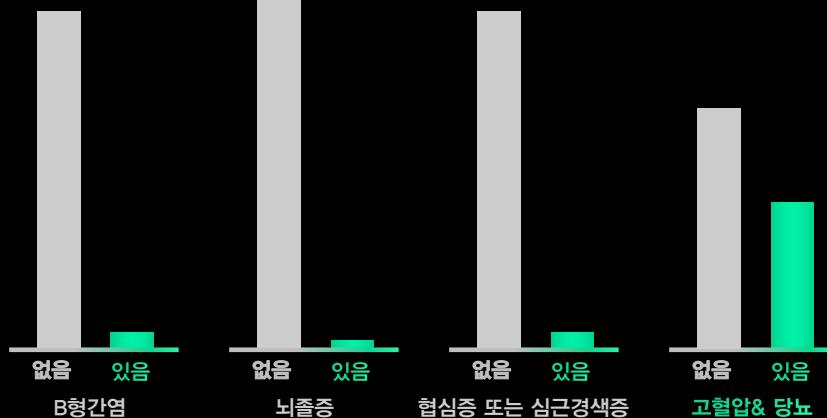
협심증 or
심근경색증

뇌졸증

B형간염



B형 간염을 제외한 질병들이 중장년기(▲)를 기점으로 가파르게 증가하는 경향을 띠고 있다는 것을 확인할 수 있다.



Target 질병의 Label 상황

출처: 국민건강영양조사 분석결과; Plot3



우리나라의 평균 연령은 43.5세(통계청; 2021)로 중년층(40.4%)이 전체 인구집단에서 지배적인 역할을 하고 있습니다.

때문에 질병과 연령의 명확한 상관성을 증명하기 위해 Target 질병에 대한 연령별 분포를 통해 KDE(커널밀도 추정)를 실시하였습니다. (국민건강영양조사분석결과; plot2)



고혈압과 당뇨병 유병자간의 유사도를 자카드 지수로 측정하였을 때, 0.24로 다른 질병보다 더 높았습니다. (국민건강영양조사분석결과; plot4)

이와 같은 당뇨병과 고혈압 간의 상호인과 관계는 이전부터 많은 기반연구(오태섭; 2022)에서 다뤄진 만큼, 변수들의 합집합으로 병 합해 주었습니다.

입력변수에 대한 신뢰도 검정

RandomForest 모델의 Feature Importance

불순도 기반 Feature Importance는
High Cardinality 변수에 대해 편향 될 수 있음



Binary

HE_HPdr: 검진당일 고혈압 약 복용 여부

0.0 : 30966 | 아니요
1.0 : 236 | 예

Categorical

educ: 교육수준: 학력

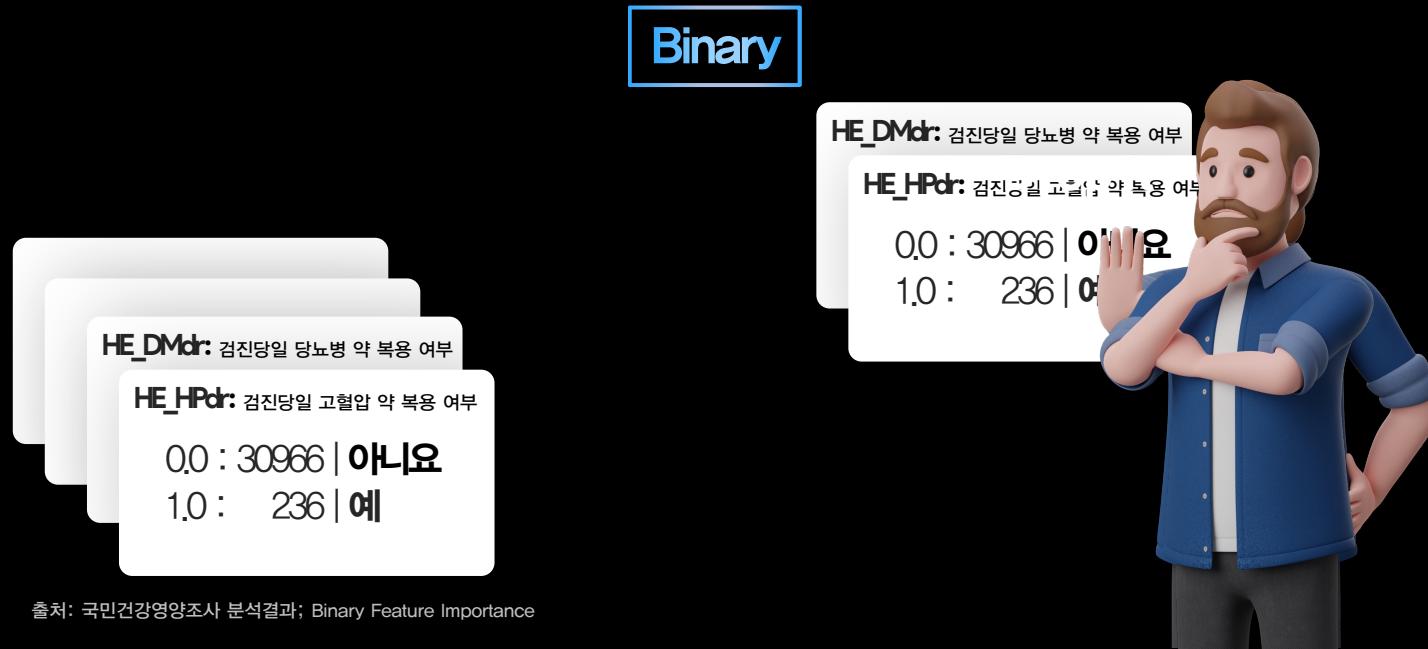
5.0 : 10193 | 고등학교
7.0 : 7125 | 4년제 대학
⋮

Continuous

HE_sbp: 최종 수축기 혈압(2,3차)

110.0 : 887 114.0 821
120.0 : 831 118.0 820
115.0 : 827 ⋮

다양한 방법들이 있지만, 한번에 통계량을 측정하기는 해한 디자인의 데이터에는 모델학습에 유의한 변수를 기준으로 인사이트를 확대하는 전략이 유효했습니다.
변인 통제를 위해 입력 변수를 대별로 분리하여 추출된 상위 10개의 중요특성을 기준으로 변수에 대한 신뢰성을 확보했습니다.



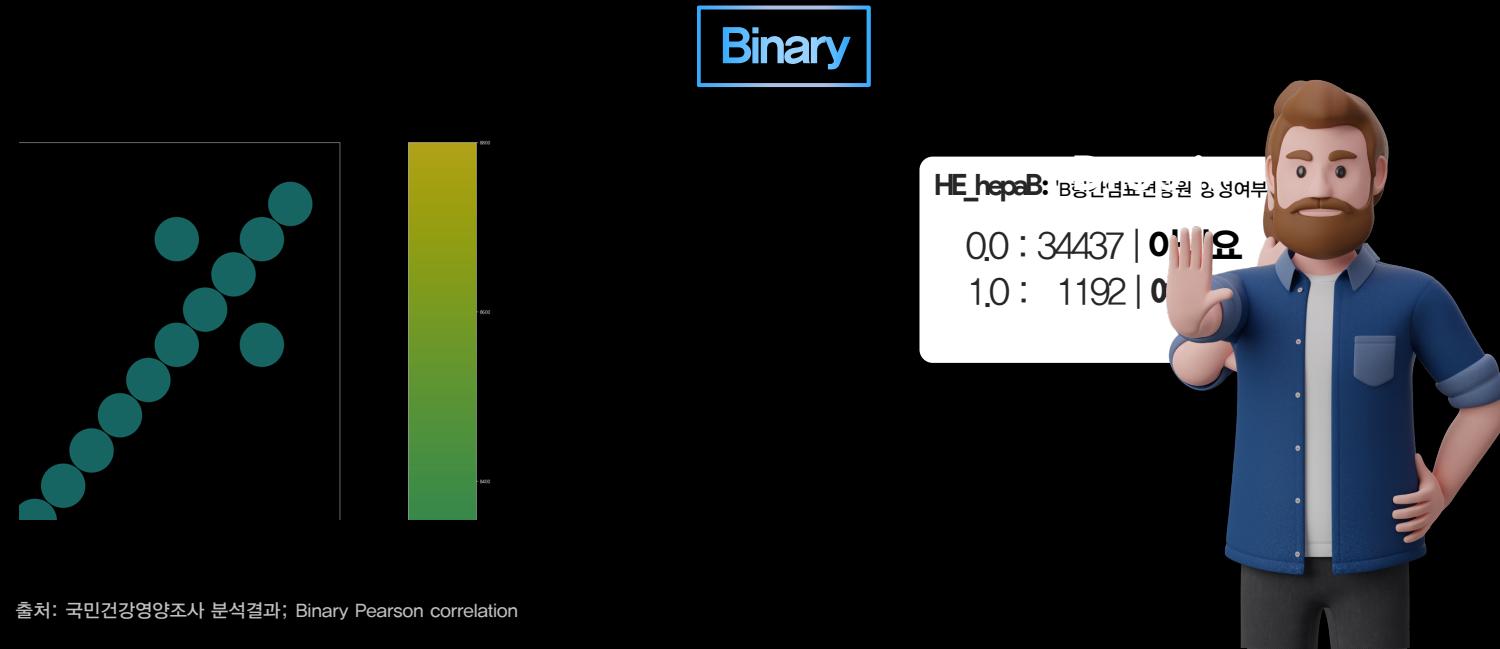
모델학습에 있어 지배적인 영향을 끼친 변수가 있는가?

YES

있다면 질병 리스크 예측 문제의 설명변수로 적합한가?

NO

이전데이터만으로 학습된 RF모델에서 추출한 상위 10개의 입력변수에는 직접적인 약물복용 여부에 대한 정보를 담은 변수가 포함되어 있었습니다.
입력도지 못하는 정보를 두고 과적합이 일어날 우려가 있으므로 설명변수에서 제외해 주었습니다.



독립변수 간 상관계수가 ±0.4를 초과하는 강한 상관관계를 가지는 변수가 있는가?

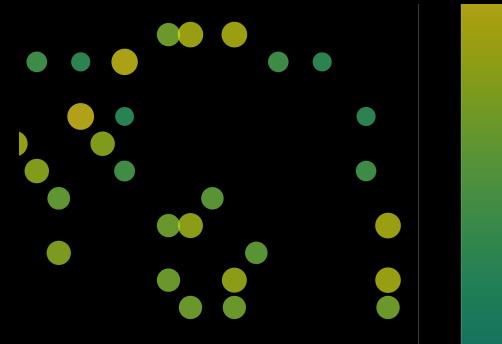
YES

있다면 질병 리스크 예측 문제의 설명변수로 적합한가?

NO

이전형 데이터들에 대한 명확한 인사이트로 확보하기 위해 결측값 처리한 음수 값들을 부정(0)으로 대치 시켜 상관분석을 시행했습니다.
이를 두고 독립변수간의 상관관계가 보통의 상관성을 지닌 변수들을 확인했을 때, B형간염 표면 항원 양성 여부가 정확히 B형간염 유병 여부에 대응되는 변수였기에 설명변수에서 제외하였습니다.
실제로 멀티클래스 분류 문제에서 B형간염 예측정확도가 지나치게 높은 문제가 이를 통해 해결되었습니다.

Categorical



Choose it!



출처: 국민건강영양조사 분석결과; categorical Pearson correlation

출처: 국민건강영양조사 분석결과; categorical Pearson correlation

모델학습에 있어 지배적인 영향을 끼친 변수들 중 강한 상관관계를 가지는 변수가 있는가?

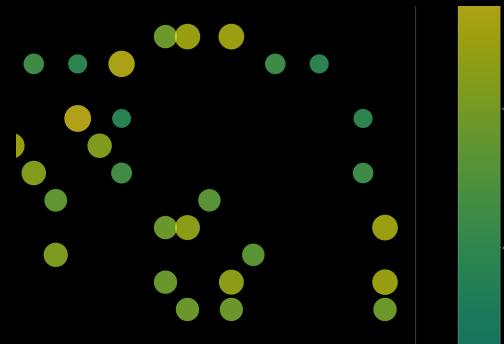
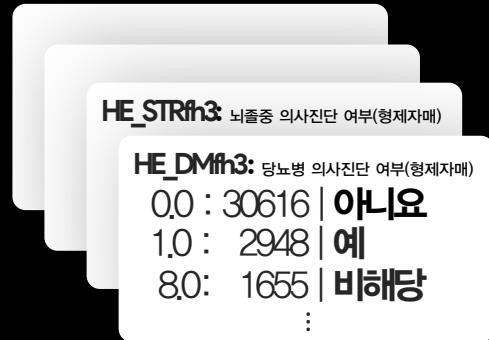
YES

있다면, 그 변수에 대한 강한 상관관계가 종속관계로 설명되는가?

YES

범주형 데이터의 치원을 축소하기 위해 유료한 전략은 특징중요도로 정렬된 리스트에서 종속관계를 가진 독립변수가 있는지를 확인하는 것입니다.
저희는 소득분위에 대한 동일한 정보가 4개의(개인, 가구, 4분위, 5분위) 독립변수에 개별적으로 저장되어 있음을 확인했고, 가장 중요도 높은 변수를 선택하여 조정했습니다.

Categorical



Merge it!



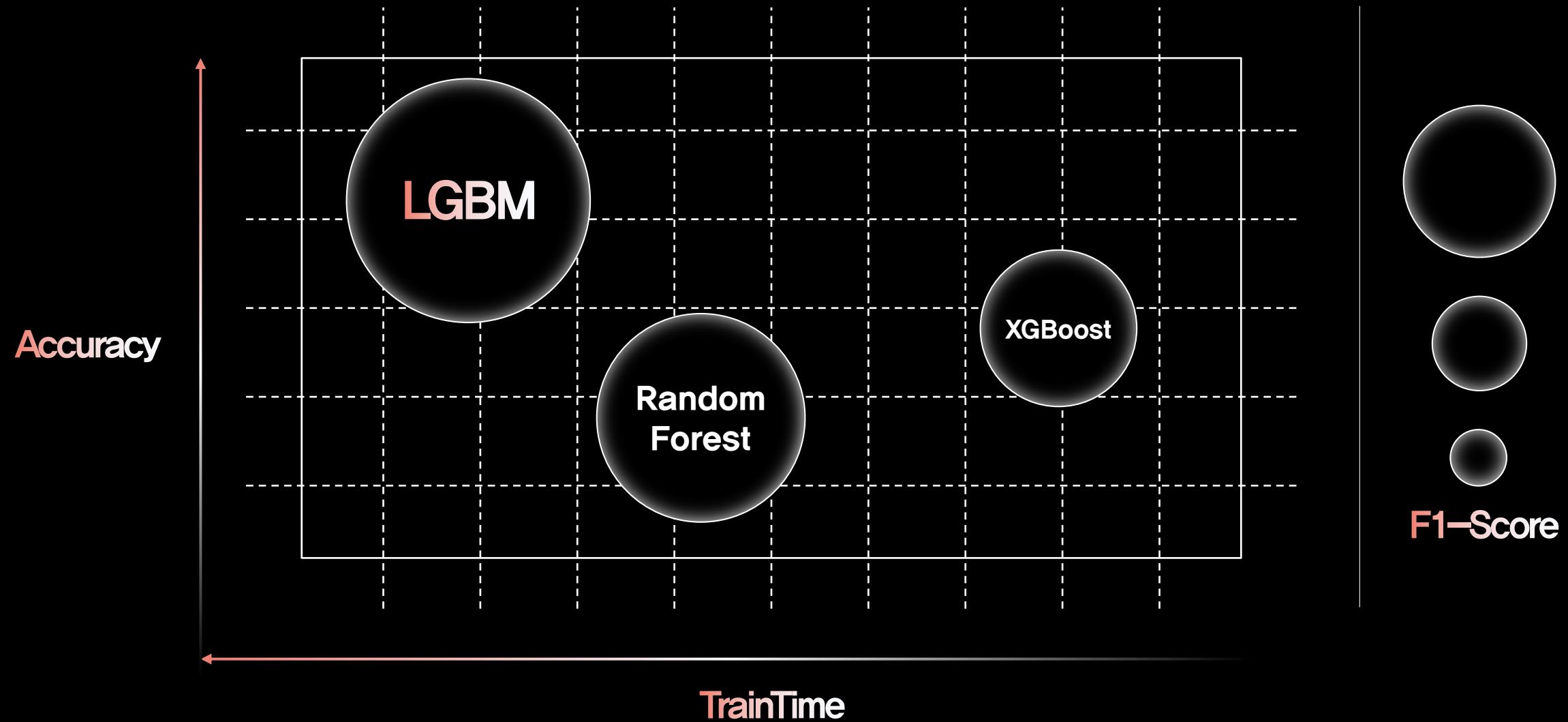
모델학습에 있어 지배적인 영향을 끼친 변수들 중 강한 상관관계를 가지는 변수가 있는가?

YES

있다면, 그 변수에 대한 강한 상관관계가 종속관계로 설명되는가?

NO

마찬가지로 모델에는 강한 영향을 주면서, 독립변수간의 상관관계가 강한 입력변수들을 확인할 수 있었습니다.
 다만, 주관적인 영역에서 가족력에 대한 변수는 정보에 있어 보완적인 측면이 있다고 판단하여, 대치가 아닌 변수를 병합하여 편향을 줄이는 전략을 사용했습니다.



모델 선택에 있어 3종류의 CART기반의 알고리즘으로 실험을 설계한 결과 정확도, 학습시간, F1Score 등의 지표에서 LGBM이 모두 좋은 모습을 보여주어 선택하게 되었습니다.

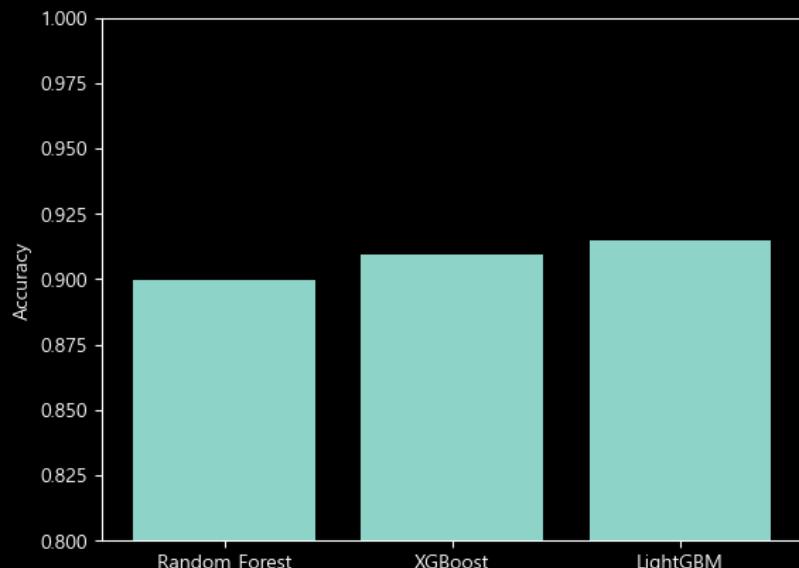
데이터 Inbalance 문제의 해결

데이터 불균형
문제

실험설계

가중치
&샘플링

결론



모델 학습 결과 모든 알고리즘이 90%가 넘는 정확도로 꽤나 신뢰할 만한 모델을 완성했다고 생각할 수 있다.

하지만, 클래스별 F1-score를 적용하면, 현재 Target은 뇌졸증과 심장질환에서 나타나는 target의 비대칭성이 모델 학습 단계에서 극복하지 못한 것으로 보인다.

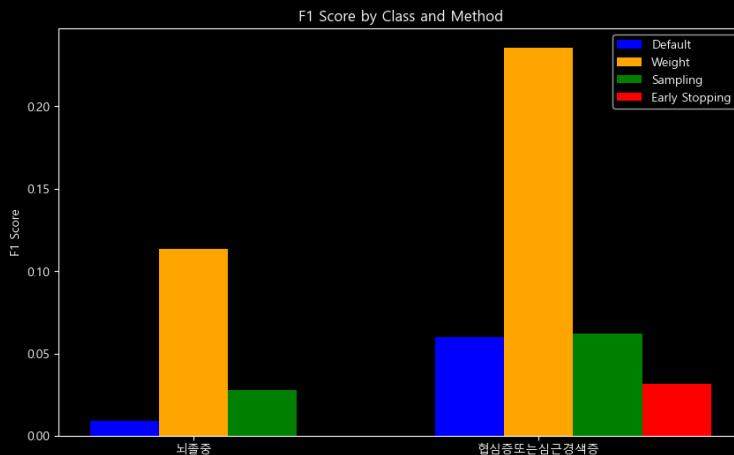
데이터 Inbalance 문제의 해결

데이터 불균형
문제

실험 설계

가중치
&샘플링

결론



데이터의 비대칭성을 극복하기 위해서 범용적으로 사용되는 3가지의 방법을 통해 실험 해보았다.

실험 결과 이전에는 좋은 효과를 보기 힘들었던 UP-sampling에서 차이를 보였고, 고전적으로 사용되는 종속변수에 가중치를 추가하는 방법은 여전히 유효했다.

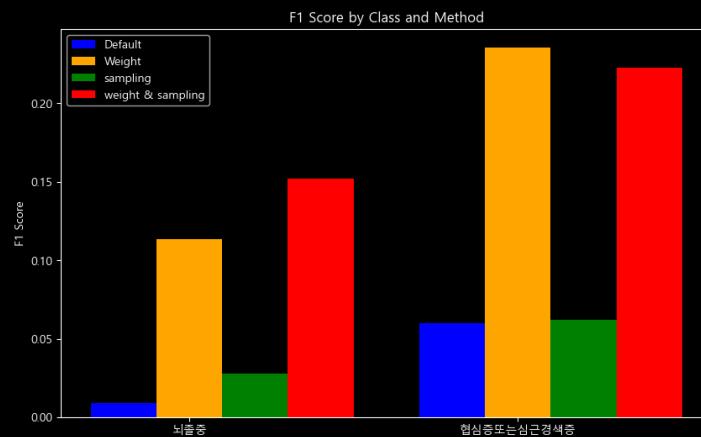
데이터 Inbalance 문제의 해결

데이터 불균형
문제

실험설계

가중치
&샘플링

결론



실험에서 성능이 좋았던 두 방법론을 동시에 적용해서 비교해보았다. 병합한 방법론에 대해 유의한 효과가 발생했다.

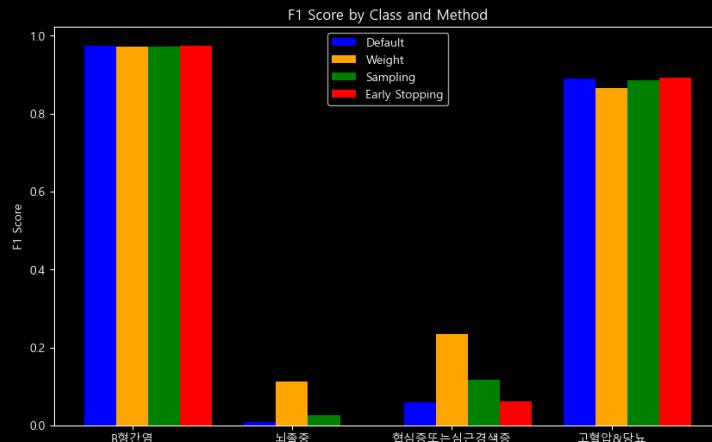
데이터 Inbalance 문제의 해결

데이터 불균형
문제

실험설계

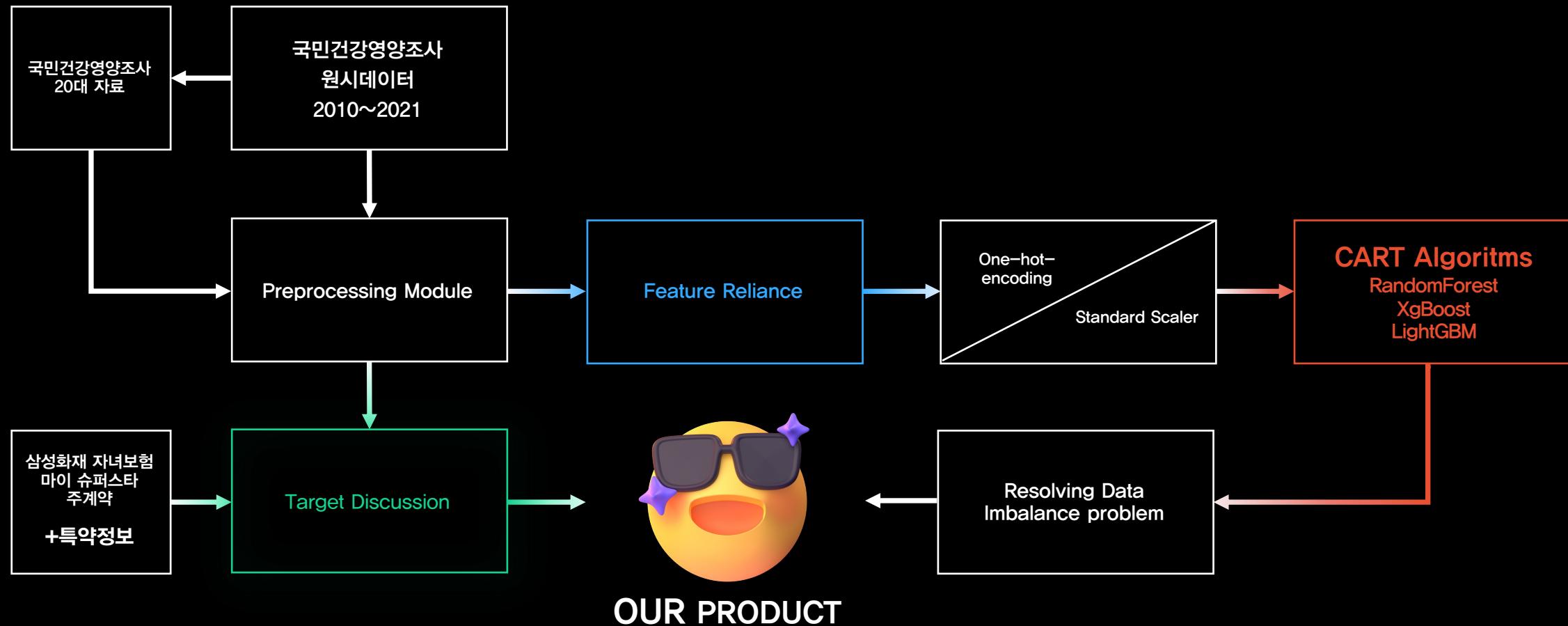
가중치
&샘플링

결론



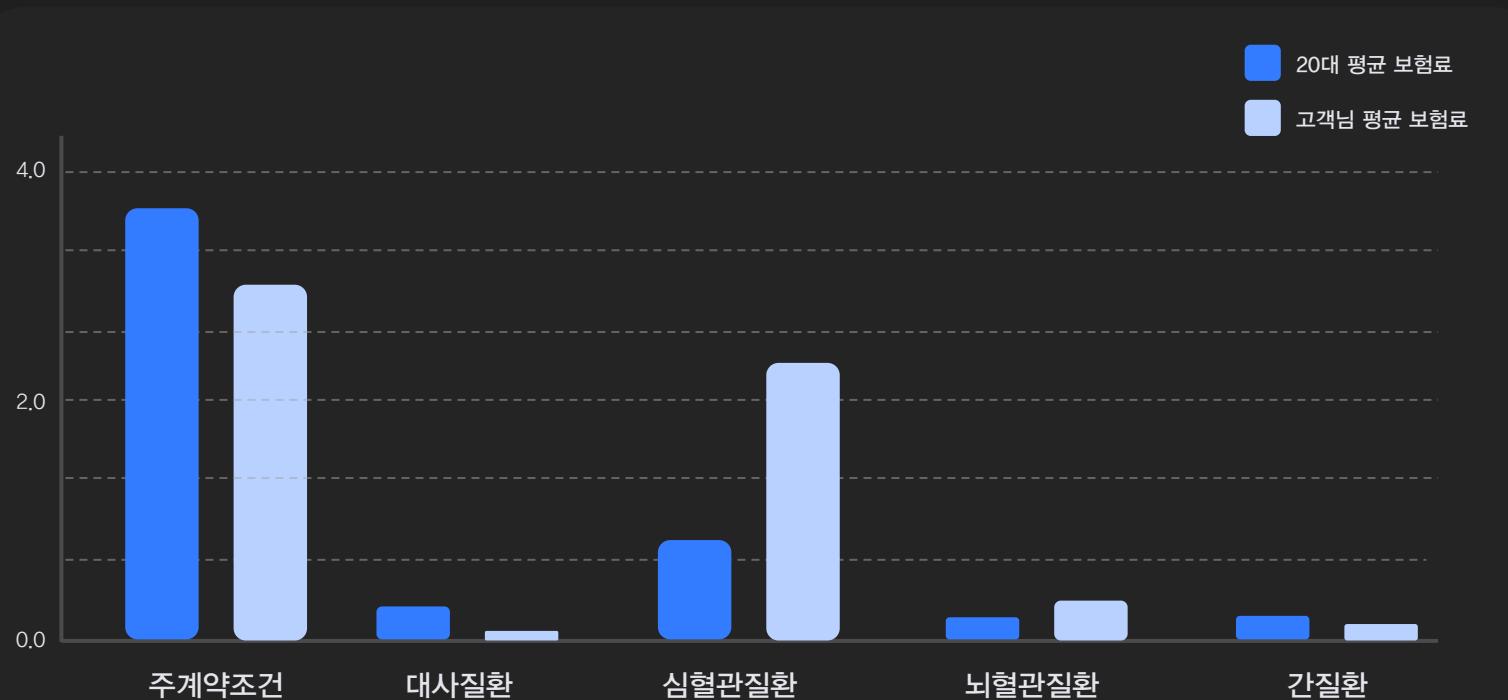
실험을 통해 소숫점 두자리 수 아래의 f1-score를 한자리 수로 만드는데는 성공했지만, 멀티클래스 분류문제에서 모델의 이런 약점은 합리적이라고 보기에는 힘들다. 뇌졸증과 협심증에 대한 다른 데이터에서 분석을 실행하여 적용하는 방법도 고려해봐야 할 것으로 보인다.

Full process of Project



DNA Insurance

보험료 추천 알인보험료 Beta



DNA Insurance
적합 보험료 계산 공식

$$IP = \left(\begin{bmatrix} 19000 \\ 162 \\ \vdots \\ 18256 \\ 2000 \end{bmatrix} - \begin{bmatrix} 380 \\ 162 \\ \vdots \\ 2608 \\ 40 \end{bmatrix} \right) \odot \left(\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.345 \\ 0.75 \\ 0.21 \\ 0.12 \\ 0.3 \end{bmatrix} \right) + \begin{bmatrix} 380 \\ 162 \\ \vdots \\ 2608 \\ 40 \end{bmatrix}$$

$$\text{Total} = \sum_{i=1}^n IP_{i1}$$

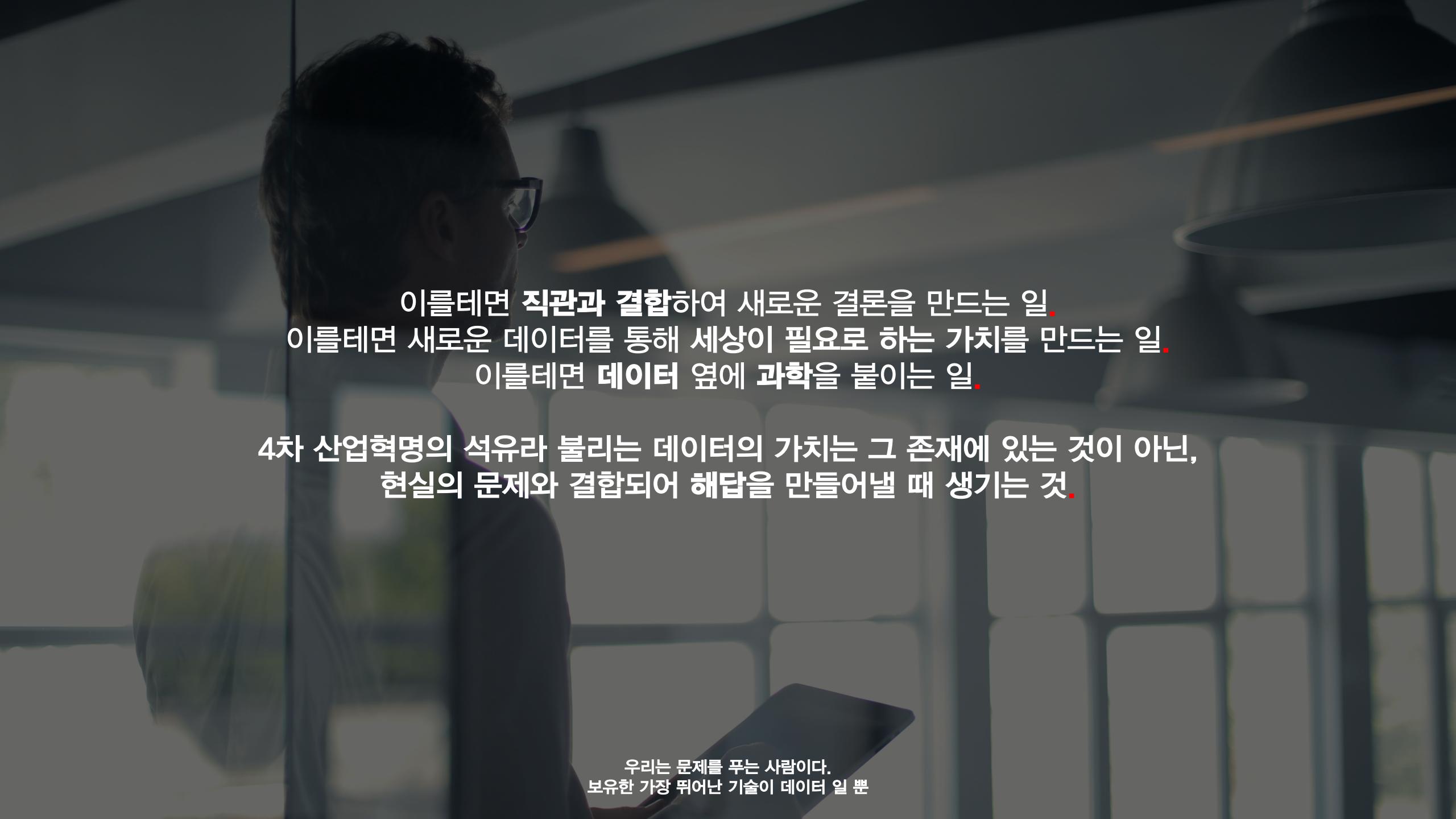
지금 상담신청하기
GO!





“우리가 무슨 일을 할 수 있는지 보여주자.”

어느덧 우리의 목표가 되어버린
나의 객기(客氣)



이를테면 **직관과 결합하여 새로운 결론을 만드는 일.**
이를테면 **새로운 데이터를 통해 세상이 필요로 하는 가치를 만드는 일.**
이를테면 **데이터 옆에 과학을 붙이는 일.**

4차 산업혁명의 석유라 불리는 데이터의 가치는 그 존재에 있는 것이 아닌,
현실의 문제와 결합되어 해답을 만들어낼 때 생기는 것.

우리는 문제를 푸는 사람이다.
보유한 가장 뛰어난 기술이 데이터 일 뿐