분류를 통한 삶의 질 점수 추정

자료분석 및 실습 기말 프로젝트

최동혁

2023-05-25

1 주제 소개 및 연구 목적

KNHANES (국민건강 영양조사) 중 **EQ_5D변수는 삶의 질 점수**이다. 0에서 1사이의 연숙형 변수로 기록되어 있으며, 질병관리 본부의 삶의 질 가중치 연구 보고서에 의해 연구된 점수이다.

이는 운동능력(mobility), 자가간호 능력(self-care), 일상활동 능력(usual activity), 통증/불편감(pain/discomfort), 불안/우울(anxiety/depression)의 5개 문항으로 구성되어 있다. 앞으로 이들을 '문항 변수' 라고 하겠다.

이 문항 변수들은 각 3개의 선택지로 구성되어 있으며 모두 1, 2, 3의 카테고리형 변수로 기록되어 있다. 해당 문항 변수들로 EQ-5D 변수를 추정하는 식은 Figure1.에 첨부해두었다.

본 연구에서는 이 문항 변수를 다른 나머지 변수(이하 세부 변수)를 통해 추정하여 최종적으로 EQ_5D의 값을 추정하고자 한다. 각 문항 변수가 1, 2, 3으로 구성되어있기 때문에 세부 변수를 이용하여 classification을 진행하고자 한다. 그래서 각 문항 변수들을 classification을 하여 각각의 prediction을 만들고 이를 통해 EQ_5D의 prediction을 만들어 실제 값과 통계적 유의성을 파악해본다.

이를 통해서 우리 삶에 대해서 삶의 질을 높일 수 있는 적절한 변수가 무엇인지에 대한 적절한 탐색과 각 문항 변수에 대한 세부 변수 분석도 가능 할 것이다. 또한 어떤 변수가 문항 변수에 importance을 보고 우리의 삶의 질을 개선하는 수치적 접근도 가능하리라 생각된다.

2 변수 소개

각 문항 변수는 다음과 같다.

LQ_1EQL: 운동능력, LQ_2EQL: 자가간호 능력, LQ_3EQL: 일상활동 능력, LQ_4EQL:통증/불편감, LQ_5EQL:불안/우울 다섯가지 항목이다.
문항 변수의 설문 응답은 1은 지장없음, 2은 다소 지장있음, 그리고 3은 지장이 큼 형태로 구성되어있다.

LQ_1EQL을 예시로 들면
1. 나는 걷는데 지장이 없음 2. 나는 걷는데 다소 지장이 있음 3.나는 종일 누워있어야 함 처럼 구성되어 있다.

세부 변수는 크게 공통 변수와 개별 변수로 구성되어있다. 공통 변수는 LQ_1EQL에 대한 사전 데이터 분석으로 문항 변수에 반영할 수 있는 개인 변수를 선정하였다. 개별 변수는 각 문항 변수의 특성에 따라 영향이 클 것으로 추정이 되는 변수들을 선택하였다. 대부분 건강에 관련된 변수들로 이루어져있으며 세부 변수는 공통 변수와 개별 변수를 합쳐 모두 14개의 변수로 이루어져있다. 변수들에 대한 정보는 **Figure2.**에 첨부해두었다.

(*세부 변수는 개인에게 해당하는 개인 변수와 건강에 관련된 건강 변수롤도 구분 할 수 있으나 marri_2(결혼여부)와 D_1_1(주관적 건강 인지)을 제외하고는 공통 변수와 개별 변수로 구분 가능하기 때문에 앞으로는 공통 변수와 개별 변수로 구분하도록 하겠다.)

3 데이터 전처리

hn19 데이터를 연구의 목적에 맞게 전처리하기 위한 작업은 다음과 같다. 공통 변수와 개별 변수를 살펴보면 다음과 같다.

1) 건강 변수를 살펴보면 경우 성인의 기준을 대상으로 진행하는 경우가 많았다. 그래서 age(나이)을 19세 이상의 성인으로 취급했으며 80세 이상은 모두 80으로 처리하기 때문에, age 변수의 범위를 19~79세의 성인으로 전처리하였다.
2) ainc(월 평균 가구소득) 변수도 17만원 이하는 17만원, 1500만원 이상은 1500만원으로 취급하였기에 이 부분을 필터링하여 17~1500만원 사이의 값으로 전처리하였다. 3) educ, sex, EC1_1, incm의 경우에는 정보를 알 수 없는 설문 응답을 제외하였다.

1) 모든 건강 변수(ex. DM1_dg, AC1_yr 등)의 미응답 및 해당 없음에 대한 부분을 제거하였다.
2) DF2_pr (우울증 현재 유병 여부) 변수의 경우에는 우울증 없음:0 과 의사에게 진단받지 않음:8 을 없음:0 으로 바꾸어주었다. 이는 의사에게 진단받지 않았을 경우를 우울증의 binary에 포함시켜야하는데 없음에 더 가깝다고 판단하였다.

이렇게 공통 변수와 세부 변수를 정하고 필터링을 통해서 데이터 클리닝을 통해 의미있는 데이터 셋을 만들었다. 그 다음에는 결측치가 존재하는 데이터를 제거해서 분석에 의미있는 데이터를 만들었다. 그래서 나온 데이터는 **4473 obs. 30 variables**로 만들어졌다.

전처리의 마지막 단계로 classification을 진행해야하기 때문에 data을 train과 test로 split해야 한다. split의 비율은 다양하지만 이번 프로젝트에서는 train과 test set의 비율을 2대1로 split해주었다.

해당 전체 데이터 셋의 example은

앞서 언급한 것 처럼 전체 4239 개의 데이터가 있으며, train은 4239 개, test는 1413개로 split이 된다.

4 모델 선정

Classification을 위해 선정한 모델은 4개로 다음과 같다. 모두 classification에 특화된 모델들로 선정했으며 decision tree를 기반으로 더 ensemble을 기반으로 더 발전시킨 모델들이다.

1) randomForest: 여러개의 decision tree을 모아서 bagging 앙상블한 모델이다. 전체의 feature가 100개가 있다면 100개 모두를 사용했을 경우 over-fitting이 나기때문에, 10~20개 정도를 선택해서 만든 decision tree 여러개를 만들어 Forest를 만드는 방식이다.

2) xgBoost: 또한 decision tree을 ensemble한 알고리즘입니다. 하지만 RandomForest와 달리 한개의 예측 모델에 대한 error를 줄이는 boosting ensemble로 구현된 모델이다.

3) lightGBM: 트리를 생성할 때 Leaf-wise방법으로 트리를 만들어 나가 적은 메모리를 사용하게 되고, 빠른 모델 생성을 한다.

4) catBoost: lightGBM의 over-fitting 문제를 해결하고 categorical data에 특화된 모델이다.

이렇게 4가지 모델을 사용하여 5개의 문항 변수에 대해 각각 accuracy가 가장 높게 나오는 모델의 prediction을 사용할 예정이다.

5 분석 과정

문항 변수에 대한 classification 과정은 다음과 같다. 앞선 변수 서정, 데이터 전처리, 모델 선정까지 진행되었다고 가정한다. 분석과정은 다음과 같다.

1) 삶의 질 변수 추정을 위한 세부 변수 선택과 이에 따라 데이터 전처리

2) 각 모델 별로 input할 형태의 데이터로 전처리 진행(e.g. label 값 변경, input 형태 함수 사용)

3) k-fold cross validation과 hyper parameter tuning을 통해서 모델의 최적의 parameter들을 도출

4) 3)에서 도출된 최적의 parameter와 세부 변수들을 이용해 문항 변수 classification을 진행하고accuracy와 confusion matrix 확인

5) Variable importance을 보고 14개중 상위 10~12개 importance을 가진 variable selection을 진행

6) 5)에서 고른 변수를 통해서 5, 6 번의 과정을 반복해서 classification 진행 후 accuracy 확인

7) 해당 4개의 모델에 대해서 2)~6)의 과정을 반복하여 문항 변수에 대해 가장 높은 accuracy을 가지는 model과 예측 값(prediction)을 저장

8) 각각의 삶의 질 문항 변수에 대해 추정된 prediction을 가지고 삶의 질 점수를 추정하는 가중치 수식에 대입해서 삶의 질 점수를 추정

9) 삶의 질 점수 추정치와 실제 삶의 질 점수를 통계적 유의성 분석 진행

6 Classification accuracy 결과

Classification은 앞서 분석 과정을 통해 20번의 과정을 진행하였다. 이에 따라 accuracy을 모두 산출하였고 5개의 문항변수에 모두 xgBoost가 가장 좋은 성능을 내었다. 각 문항 변수와 모델에 대한 accuracy는 Table 2.에서 확인할 수 있다.

이 classification 결과는 물론 모델도 중요했지만 그만큼 중요했던것이 2가지라고 생각하는데 CV와 hyper parameter tuning을 했던 것이 첫번째이고 importance가 높은 variable을 뽑은 것이 두번째라고 생각한다.

1) over-fitting을 막기 위해서 CV을 진행함과 동시에 parameter까지 찾는 iteration도 진행을 했기 때문에 시간이 오래 걸리더라도 더 좋은 accuracy을 가질 수 있는 parameter을 찾을 수 있었다.

2) variable 중에 개인적 주관에 의해 중요하다고 생각했던 변수도 importance을 막상 뽑아보면 거의 영향을 주지 않는 것도 있었다. 이런 것들은 오히려 제거를 해주는것이 중요했다. 또한 실제로 처음에 variable을 많이 쓰지 않았기 때문에 꼭 필요한 것을 제외하고는 제거하는 과정을 진행했다.

7 Classification 결과로 EQ-5D 유의성 분석

가장 accuracy가 좋은 xgBoost로 예측된 값을 통해서 Figure 1.에 나와있는 가중치 평균식을 통해 삶의 질 예측값으로 바꾸어주었다. 이 과정에서는 2가지 작업이 필요하다.

1) xgBoost로 예측된 것은 precition의 class가 0, 1, 2로 예측되기 때문에 이를 class을 1, 2, 3으로 변경해준다.

2) 그 후에 Figure 1.에 있는 식을 함수로 구현하여 이값을 대입해준다.
그 다음 예측값과 실제 값을 통계적으로 비교하기 위해 RMSE와 mean의 차이를 구하면

- RMSE (평균 제곱근 편차) : 0.091
- mean_diff: 0.0193(prediction이 더 크다)

이 됨을 알 수 있다.

그 다음 데이터의 분포를 살펴보기 위해 figure 4.를 그려보았는데 여기서 알 수 있는 점은 데이터가 굉장히 descrete하고 skewe된 데이터라는 점을 알 수 있다. 그래서 실제값과 예측값을 통계적 유의성을 분석할때 쉽게 쓰는 t-test를 쓸 수 없다. 왜냐함녀 정규성 분포 가정이 어렵기 떄문이다.

그래서 사용하는 것이 Wilcoxon-test을 사용한다. 그 test 결과는 다음과 같다.

wilcoxon-test: alternative hypothsis. not equal.

귀무가설이 기각되고 다른 것으로 나온다. 아쉽게도 예측값이 실제값과 통계적 유의성을 가지지는 못하는 것으로 알 수 있다.
마지막으로 이들의 denstiy distribution을 figure 5.에서 확인 할 수있다. 실제로 값이 0에서 1사이를 가지는 값들이기 때문에 그리고 0.7 이상에 skewed된 데이터 이기 때문에 신뢰구간안에서 유의성을 갖추기는 매우 어려워 보인다.

8 결론 및 분석

본 연구는 EQ5D 변수를 LQ_1EQL, LQ_2EQL, LQ_3EQL, LQ_4EQL, LQ_5EQL를 예측한 뒤 함수를 통해 생성하는 과정을 통해 예측했다. 가중치 식에 따르면 EQ-5D는 세 질문을 모두 최상의 선택지를 고르면 1을 출력하는 비선형 모델이기에 오차가 예측결과 이상으로 발생할 수 있음을 알 수 있었다. 이에 따라 향후 연구할 수 있는 방안으로는 다음과 같다.

1) 더 많고 다양한 변수 추가로 연관성 있는 변수를 최대한 많이 찾아내고 importance가 높은 변수를 selection하여 accuracy을 높일 수 있다.
2) 문항 변수의 1, 2, 3의 뷸균형이 생각보다 심하다는 것을 table.에서 알 수 있다. 이처럼 imbalanced data 문제 해결를 해결하는 것 또한 accuracy을 높히는 방법이 될 것이다. 진행과정에서 연구해본 결과 R에서는 xgBoost가 multiclass에서 imbalanced data을 해결하기 위한 weight을 찾는 logic을 제공하는 module이 없다.

3) grid_search 과정에서 보다 많은 iter와 다양한 parameter에 대한 tuning으로 더 정확한 hyper parameter tuning을 통해 더 정확한 예측 모델 생성 가능 할 것이다.

그리고 해당 프로젝트를 진행하면서 생긴 질문들과 예상되는 추론을 간단히 담아보았다.

Q1. 왜 LQ_4EQL의 정확도만 낮은 정확도를 보였을까?

각 문항 변수의 1, 2, 3 분포를 살펴보았을 때 4번 문항 변수가 2의 분포가 다른 문항 변수보다 많았다. 이에 대한 분포와 전체 테이블에서 2번 문항의 분포의 비율은 Table 3.에서 살펴볼 수 있다.
실제로 confusion matrix을 살펴보았을 때 모든 모델에서 2가 실제 값이었을 때 1로 잘못 예측하는 확률이 높았다. 이 때문에 2번 문항이 차지하는 비율이 낮을수록 accuracy가 높은 경향이 있고, LQ_4EQL은 높기 때문에 accuracy가 낮게 나온 것으로 추정된다.

Q2. 왜 xgBoost가 가장 좋은 성능을 냈을까?

1.lightGBM이 xgBoost 보다 속도와 정확도 측면에서 우수하지만 약 10000건 미만의 데이터에서는 overfitting이 일어나기 쉽다는 연구가 있다.
2.Catboost 모델에서 hyperparameter tuning을 하기위한 grid_search 방식이 너무 단순했다.
3.randomForest의 bagging 앙상블 방식보다 error을 줄이는 boosting 앙상블이 이번 다중 분류에 더 적합했던 것 같다.