

빅데이터 분석 및 시각화 개론

최종 보고서

2017. 12. 15.

12 조

201411024 김성연

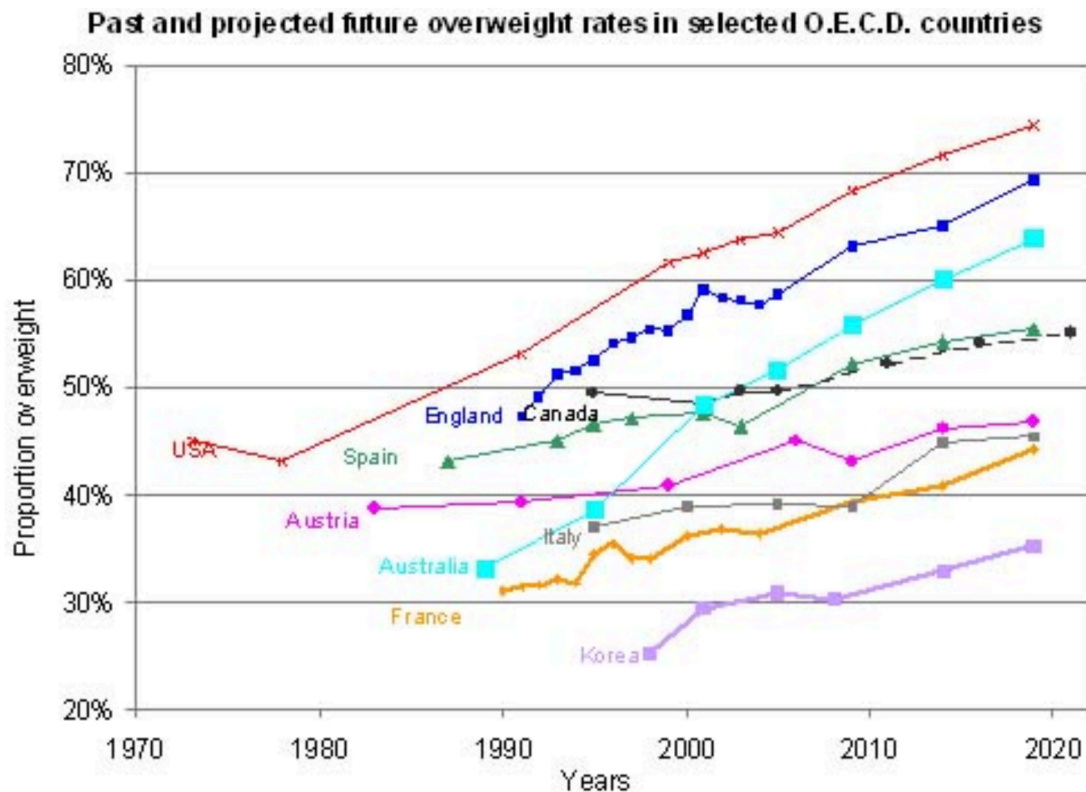
201411075 심성호

201411166 황세현

Contents

- 1. Introduction**
- 2. Materials**
- 3. Method**
- 4. Result**
- 5. Discussion**
- 6. Appendix**

1. Introduction



위 그래프는 OECD 국가들의 년도에 따른 과체중 비율을 나타낸 그래프이다. OECD 모든 국가에서 시간이 지날수록 과체중 비율이 점점 증가하고 있다는 것을 확인 할 수 있다. 과체중과 비만이 각종 성인병의 원인이 된다는 점에서, 이는 심각한 문제라고 할 수 있다.

우리는 세계 보건기구(World Health Organization) 입장에서 빅데이터 분석을 통해 점점 증가하는 비만을 관리하기 위한 정책을 제안하고자 한다.

우리는 미국 농림부 데이터를 분석 대상으로 삼았으며, 국제표준에 따라 BMI 지수가 30 이상인 사람을 비만으로 설정하였다. Correlation, ANOVA, Random forest 등의 방법을 활용하여 비만과 관련이 높은 feature 들을 선정했다. 이때 뽑힌 feature 를 이용하여 비만 유무를 검사하는 predictor 를 만들고 그 성능을 검증하였다.

저소득과 비만의 유의미한 관계를 밝히고 그 원인으로 SNAP 정책의 한계를 지적하고, 이를 해결할 수 있는 정책을 데이터에 기반하여 제시하였다.

2. Materials

우리의 Data set 은 다음과 같다. 미국 농림부 (United States Department of Agriculture)의 'American Time Use Survey(ATUS) eating and health module' 라는 데이터로 2014 년과 2015 년의 미국인들의 식습관과 건강에 대한 데이터를 사용하였다. Row 개수는 21,838 개 이고, Column 개수는 37 개 이다.

eeincome1	erbmi	erhhch	erincome	erspemch	ertpreat	ertseat
-2	33.200001	1	-1	-1	30	2
1	22.700001	3	1	-1	45	14
2	49.400002	3	5	-1	60	0
-2	-1	3	-1	-1	0	0
2	31	3	5	-1	65	0
1	30.700001	3	1	1	20	10
1	33.299999	1	1	5	30	5
1	27.5	3	1	-1	30	5
1	25.799999	3	1	-1	117	10
1	28.299999	3	1	5	80	0
1	40.5	3	1	-1	35	20
2	28	1	5	-1	0	5
1	27.9	3	1	5	25	10
2	30.4	3	5	5	150	5
2	26.799999	3	5	-1	0	300
1	32.900002	3	1	5	80	0

(2014 ATUS data example)

37개 column에 대한 정보는 'American Time Use Survey (ATUS) Data Dictionary: 2014-15 Eating & Health Module Data Variables collected in ATUS Eating and Health Module'를 참고하면 된다. 이 문서에 각 column에 대한 설명이 자세히 나와있다.

3. Method

1) Data refinement

데이터를 정제하기 위하여 다음과 같은 방법들을 사용하였다. 의미 없는 열들을 제거하고, 값이 이상한 outlier 들을 제거하였다. 예를 들어, 키가 비정상적이게 작거나, 몸무게가 비정상적이게 많이 나가는 데이터를 설문을 잘못된 것으로 판단하고, 이러한 outlier 들을 제거하였다.

데이터에 BMI 가 30 이상인 사람, 즉 비만인 사람의 수가 BMI 가 30 미만인 사람, 즉 비만이 아닌 사람의 수가 훨씬 많아서 이를 맞춰주고자 Imbalanced data handling (down sampling) 작업을 하였다.

또한 데이터 개수를 최대한으로 많이 활용하기 위하여 Null value filling 을 해보았다. 해당 열의 평균과 분산을 구해서 truncated normal distribution 을 만들어 이 distribution 에서 해당 열에서 Null value 인 개수 만큼 sampling 하여서 채우는 방식으로 실험해보았다. 하지만, 이렇게 Null value filling 을 하였을 때가 결과가 더 좋지 않아서, null value handling 은 사용하지 않는 것으로 하였다.

2) Feature selection

Pearson Correlation, Random forest, ANOVA 를 통해 BMI 와 연관성이 가장 높은 feature 5 개를 뽑았다. BMI 와 연관성이 높고, ANOVA 시 p-value 가 0.05 미만인 feature 5 개를 선정하였다.

- 'eugenhth' : 건강 자신감
- 'euexercise' : 운동 유무
- 'erincome' : 소득 분위
- 'eusoda' : 탄산음료 음용 유무
- 'eusnap' : 영양 보충 보조 프로그램 수혜 유무

3) Prediction

위의 feature selection 에서 선정한 5 개의 feature 들이 실제로 비만과 관련이 있는지를 확인하기 위해서, 위의 feature 5 개를 이용하여 비만을 예측하는 모델들을 학습시키고 성능을 확인해보았다. 이때 사용된 학습 데이터는 약 13k 이다. 비만인 사람의 수는 그렇지 않은 사람의 수 비해 1/4 수준이었다. 따라서 training 데이터가 불균형하여 예측 결과가 한쪽으로 몰리는 현상이 발생했다. 이를 해결하기 위해 Test 데이터에 대해서는 앞서 언급한 down sampling 방식을 적용하였고, Training

데이터에 대해서는 Balanced bagging 방법을 활용했다. 이때 Balanced bagging 의 커널로 사용한 머신러닝 기법으로는 Random forest, Support vector machine, Multiple layer perceptron 이다.

4) EDA

탐색적 데이터 분석방법으로 여러 feature 사이의 관계를 찾아보았다. 저소득층의 비만 원인에 대한 탐구를 중심으로 진행하여 soda, exercise SNAP 등을 얻었다. 특히, SNAP 의 한계를 밝히고, 그 원인으로 eat, meat, drink, preat, seat 등을 제시하였다.

4. Result

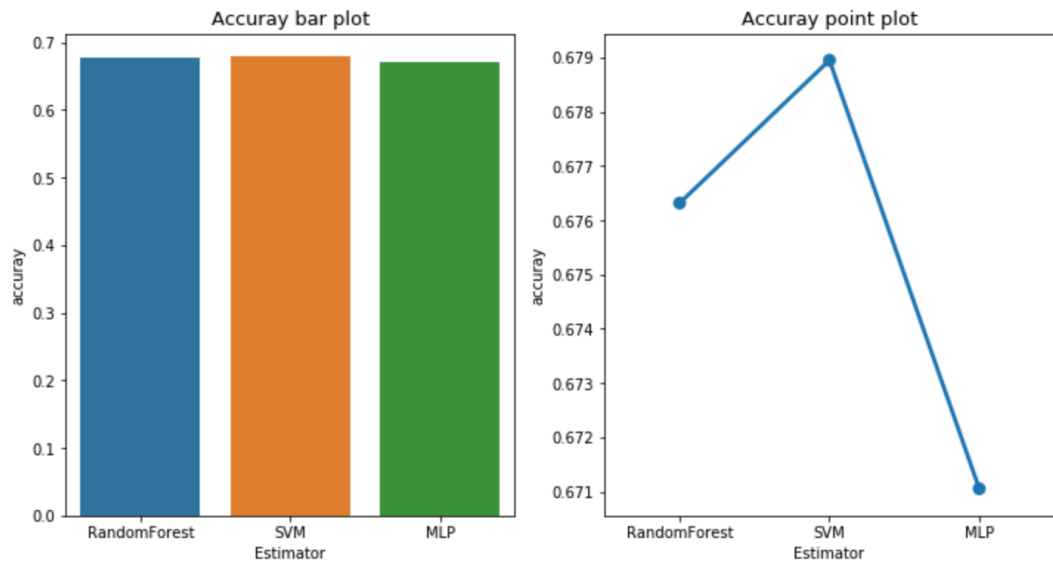
1) Feature Selection

아래의 왼쪽 그림은 Anova 검정결과 p-value 가 0.05 보다 낮은 column 중 일부를 나타낸 것이다. 아래의 오른쪽 그림은 correlation 의 결과 값의 절대값이 0.09 보다 큰 column 들을 나타낸 결과이다.

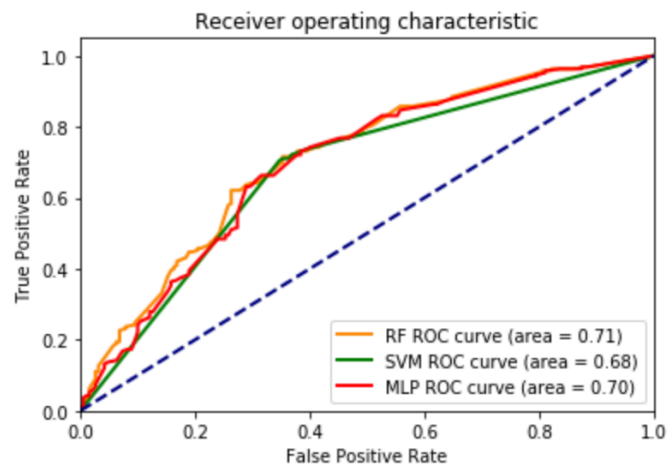
eugenhth	eugenhth	0.307512
F-value = 0.08470070184818826	euexercise	0.132688
p-value = 0.0	erincome	0.107018
euexercise	eufdsit	0.063868
F-value = 0.08470070184818826	eufastfdfrq	0.044482
p-value = 4.063829987105223e-64	eueat	0.020012
erincome	eustores	0.008721
F-value = 0.08470070184818826	eutherm	0.008331
p-value = 2.845214654098719e-37	Unnamed: 0	0.007428
eusnap	tucaseid	0.006951
F-value = 0.08470070184818826	euinclvl	0.004556
p-value = 2.584581909141339e-33	ertseat	-0.005307
eusoda	eudrink	-0.005496
F-value = 0.08470070184818826	eumilk	-0.006756
p-value = 7.398615557641728e-29	euhgt	-0.008231
ertpreat	eugroshp	-0.012693
F-value = 0.08470070184818826	euprpmel	-0.013147
p-value = 3.7740963996773906e-17	euffyday	-0.024774
euexfreq	eustreason	-0.031172
F-value = 0.08470070184818826	eumeat	-0.034092
p-value = 1.478060791529573e-16	eufastfd	-0.055099
eufdsit	eudietsoda	-0.059500
F-value = 0.08470070184818826	ertpreat	-0.062020
p-value = 3.806524744918059e-14	euexfreq	-0.083219
eufastfd	eusnap	-0.099126
F-value = 0.08470070184818826	eusoda	-0.107261
p-value = 2.0429543362306338e-13		
eufastfdfrq		
F-value = 0.08470070184818826		
p-value = 8.316551301562993e-08		
euffyday		
F-value = 0.08470070184818826		
p-value = 0.0008095394302946757		

위의 그림에서 볼 수 있듯이, 한 방법에서 bmi 와 높은 연관성을 보이는 feature 들은 다른 방법에서도 좋은 결과를 나타내는 것을 알 수 있었다. 선택한 5 개의 feature 들이 실제로 bmi 와 관련성을 보인다는 것은 아래의 EDA 항목에서 추가적으로 설명할 것이다.

2) Prediction



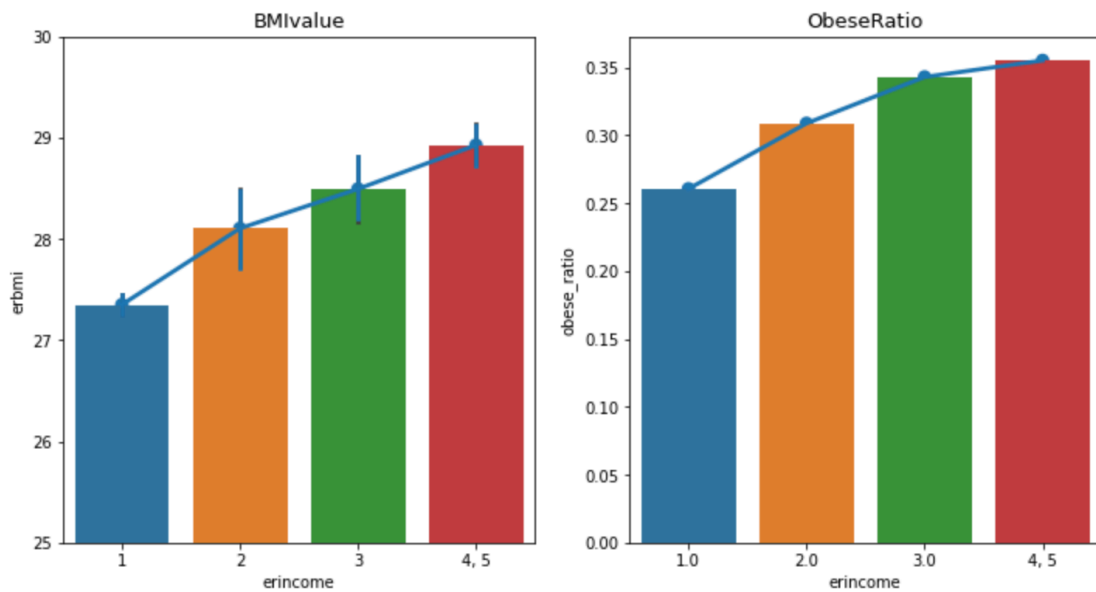
위의 Feature Selection 으로 얻은 5 개의 feature 로 머신러닝 모델을 학습하였다. 사용한 모델은 Random Forest, SVM, MLP 이다. 세 모델 모두 68%정도의 비슷한 성능을 내고 있는 것을 확인하였다. 비슷한 성능을 내는 이유는 모델을 학습하기에는 데이터가 많지 않기 때문으로 보인다. 모델의 성능을 볼 수 있는 다른 지표로는 ROC Graph 의 아래 면적 넓이인 auroc 값을 사용하였다.



각 모델의 auroc 값은 약 0.7 정도의 값을 갖고 RF, MLP, SVM 순서로 높은 값을 갖는 것을 알 수 있다. 일반적으로 auroc 값이 0.7 이상일 때 중등도의 정확한 모델이라고 볼 수 있는데, 우리의 모델들이 약 0.7 정도의 성능을 내는 것을 보아 괜찮은 성능을 내는 모델을 얻었다고 할 수 있을 것이다. 이를 통해서 우리가 선택한 5 개의 feature 들이 비만과 관련성이 높고, 비만을 예측하기에 좋은 feature 라는 것을 확인할 수 있었다.

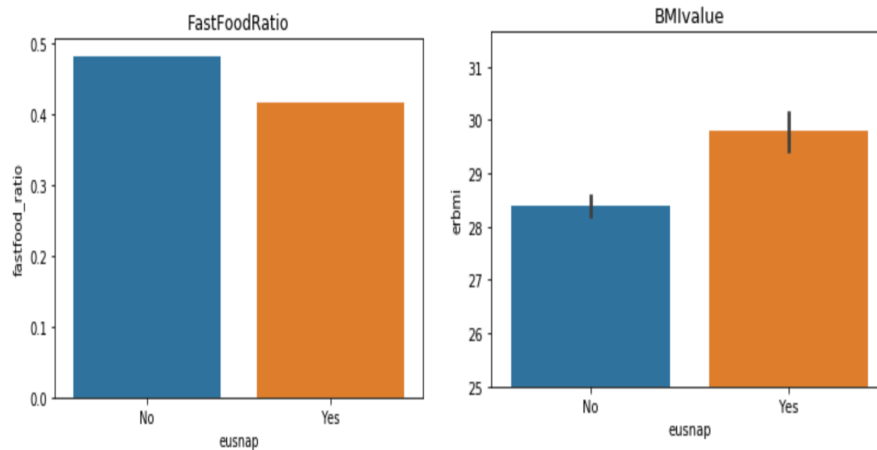
3) EDA

우리는 BMI 와 Income 이 많은 연관성이 있다는 것을 확인하고 income 에 대한 심도 깊은 조사를 해보기로 하였다.

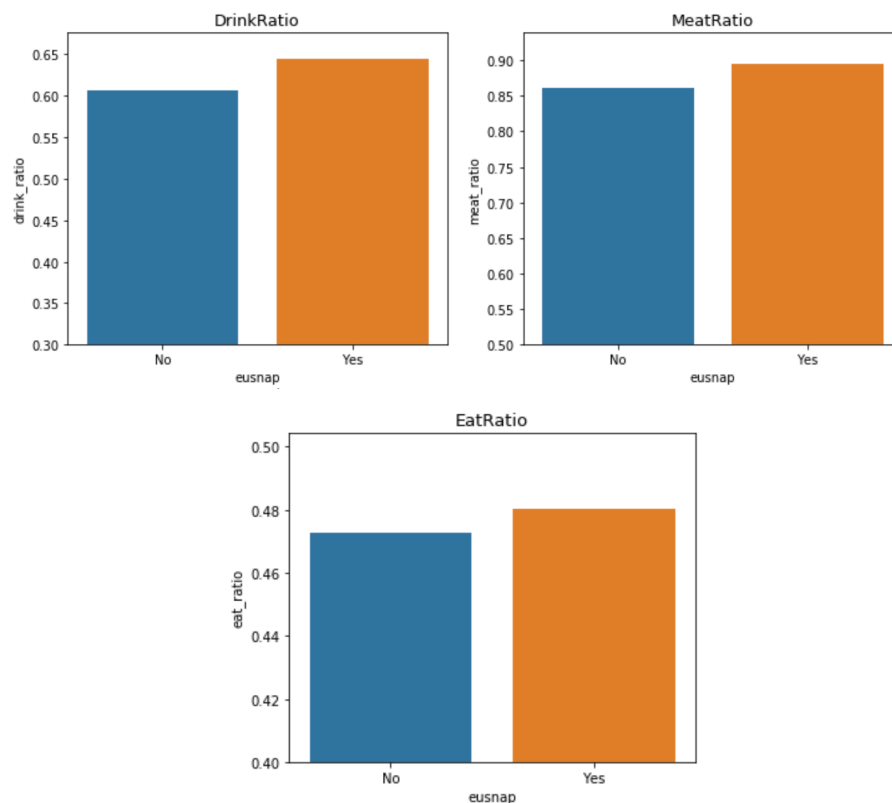


위 그래프는 income 과 BMI 지수의 평균과 비만 비율을 나타낸 그래프이다. Income 의 값이 커질 수록 소득 분위가 낮은 것이다. 위 그래프에 따르면 소득 분위가 낮을 수록 bmi 와 비만의 비율이 커지는 것을 확인할 수 있다. 또한 bmi 와 관련 깊은 다른 feature 로 SNAP 이 있는데, SNAP 이란 미국의 저소득층을 대상으로 하는 영양 공급 프로그램으로 쿠폰 형태의 SNAP 을 영양적으로 유의미한 식료품 구매에 사용할 수 있도록 한 것이다. SNAP 은 저소득층을 대상으로 하기 때문에 수입과 관련이 높다. SNAP 은 영양 공급을 통해 사람들을 건강하게 하는 것이 목표이기 때문에, 패스트 푸드에 대한 지원을 하지 않아 SNAP 특혜를 받았을 때 패스트 푸드의 이용률이 감소하는 것을 알 수

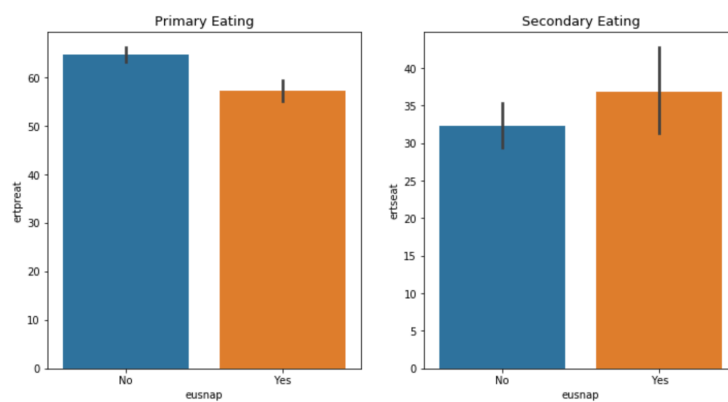
있다. 그러나 직관과 다르게 SNAP 을 이용했을 때 BMI 가 오히려 증가하는 것을 확인할 수 있었다.



위와 같은 결과가 나오는 이유를 알아보기 위하여 income 과 관련된 feature 에 대하여 eda 를 진행하였는데, 아래와 같이 snap 의 혜택을 받았을 때 전체적인 음식 섭취량 및 간식량이 증가하는 것을 확인할 수 있었다.

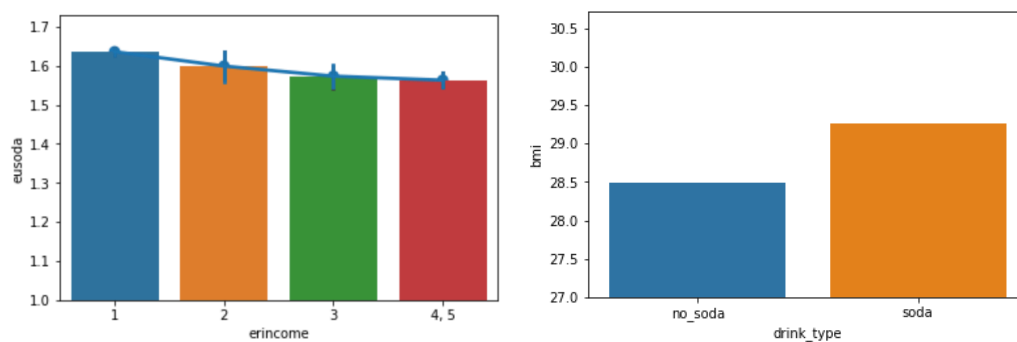


또한 SNAP 의 혜택을 받는 사람들의 식사 시간이 줄어들고 간식 시간이 증가하는 것을 확인할 수 있다. 이는 SNAP 혜택으로 받은 금액을 식사가 아닌 간식에 사용하고 있다는 의미로 SNAP 의 취지와 달리 건강하지 못한 음식 섭취가 증가하였다고 볼 수 있다.

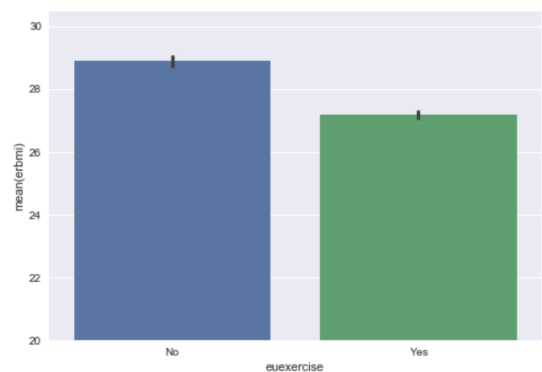
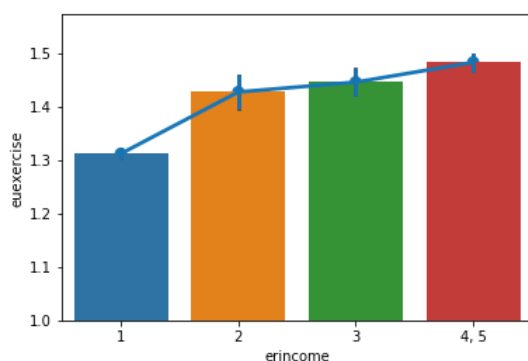


우리는 SNAP 의 혜택을 받는 사람들이 비만이 되는 이유 때문일 것이라고 생각하였다.

또한 income 이 낮은 사람일 수록 soda 의 먹는 사람들이 증가하는 것을 알 수 있었다.



위 그래프의 y 축은 1 이 yes, 2 를 no 라고 했을 때의 결과의 평균을 구한 것이다. 또한 soda 를 마시는 사람들이 soda 를 마시지 않는 사람들에 비해 bmi 가 높은 것을 보아, income 낮은 사람들의 bmi 가 높은 이유 중 하나로 soda 의 섭취 유무가 관련 있다고 할 수 있다.



또한 income 이 낮을 수록 운동을 하지 않는 사람들이 많아지는데, exercise 를 하지 않을 수록 bmi 가 높아지는 것을 확인할 수 있다. 따라서 income 이 낮은 사람들이 운동을 많이 하지 않는다는 사실이 income 이 낮은 사람들의 bmi 가 높은 이유 중 하나라고 할 수 있다.

5. Discussion

위의 결과로 우리는 세계보건기구(WHO) 입장에서 저소득층의 비만을 문제를 해결하기 위한 정책을 다음과 같이 제안한다.

- Policy#1

SNAP 과 비슷한 정책을 실시하되, 지원금으로 구매할 수 있는 Secondary eating 량에 제한을 두자.

- Policy#2

SNAP 과 비슷한 정책을 실시하되, 지원금으로 구매할 수 있는 탄산음료 구매량에 제한을 두자.

- Policy#3

SNAP 과 비슷한 정책을 실시하되, 지원금으로 헬스장을 다닐 수 있도록 하는 등의 운동 활동을 지원해주자.

우리 연구의 한계점과 연구의 발전 방향은 다음과 같다. 우리가 2014 년과 2015 년의 'ATUS eating and health module' 데이터를 사용하였는데, 만약 두 데이터가 동일한 사람들을 2 년간 조사한 결과라면 각 feature 들의 시간에 따른 변화까지

분석할 수 있어 더 재미있는 분석을 할 수 있었을 것이지만, 밑의 그림에 명시된 것처럼 ID 값 자체가 2014 년도와 2015 년도가 다르게 적용되어서 feature 들의 시간에 따른 변화는 분석할 수 없었다.

1	tucoseid	tulineno	eeincome	erbmi	er	1	TUCASEID	TULINENO	EEINCOME1	ERBMI
2	20140101140007.0	1	-2	33.2	2	2	20150101150018.0	1	1	31.4
3	20140101140011.0	1	1	22.7	3	3	20150101150053.0	1	2	25.7
4	20140101140028.0	1	2	49.4	4	4	20150101150071.0	1	1	29.6
5	20140101140063.0	1	-2	-1	5	5	20150101150146.0	1	3	23.4
6	20140101140168.0	1	2	31	6	6	20150101150147.0	1	1	35.9
7	20140101140559.0	1	1	30.7	7	7	20150101150184.0	1	1	32.1
8	20140101140610.0	1	1	33.3	8	8	20150101150500.0	1	1	30.1
9	20140101140614.0	1	1	27.5	9	9	20150101150523.0	1	1	33.9
10	20140101140639.0	1	1	25.8	10	10	20150101150535.0	1	2	26.3
11	20140101140665.0	1	1	28.3	11	11	20150101150539.0	1	1	22.6
12	20140101140685.0	1	1	40.5	12	12	20150101150572.0	1	3	28.3
13	20140101140702.0	1	2	28	13	13	20150101150578.0	1	2	22.9
14	20140101140725.0	1	1	27.9	14	14	20150101150601.0	1	1	-1
15	20140101140759.0	1	2	30.4	15	15	20150101150607.0	1	2	31.4
16	20140101140792.0	1	2	26.8	16	16	20150101150615.0	1	2	26.6
17	20140101140804.0	1	1	32.9	17	17	20150101150628.0	1	1	21.8
18	20140101140836.0	1	3	25.8	18	18	20150101150641.0	1	2	29.5
19	20140101140852.0	1	1	26.5	19	19	20150101150677.0	1	1	25.1

이러한 시간에 따른 데이터 변화 분석을 할 수 없기 때문에 causality 를 분석하는데 어려움이 있었다. 예를 들어, BMI 가 높을 수록 다이어트 소다를 많이 먹는다는 데이터 분석을 하였을 때, BMI 가 높아서 다이어트 소다를 많이 먹는지, 아니면 다이어트 소다를 많이 먹어서 BMI 가 높아진 것인지 원인과 결과를 분석하는데 있어 어려움이 있다. 따라서, 시간에 따른 변화 데이터가 있다면 더 명확한 데이터 분석을 할 수 있을 것으로

기대된다.

6. Appendix

1. PPT 발표에서와 달라진 부분에 대한 설명

앞선 결과 보고 발표의 Feature selection 과정에서 ANOVA 와 Correlation 그리고 Random forest 의 주요 feature 들이 모두 일치되게 구해졌다는 언급이 있었다. ANOVA 와 Correlation 의 결과가 일치하는 것은 사실이지만 Random forest 에서 중요 feature 로 뽑힌 것들 중에는 발표된 feature 가 해당되지 않는 경우가 1~2 개 정도 있다.(exercise, soda) 차이가 나는 feature 들은 prediction 을 진행하면서 성능이 더 높은 기준으로 판단하였고 그 결과로 얻어진 feature 가 발표된 5 개이다. 즉, random forest 의 주요 feature 가 발표된 주요 feature 와 어느정도 비슷한 경향을 보이지만 완전히 일치하는 것은 아니며 그 진위는 prediction 을 통해 판별했었다. 발표 때에 모든 방법의

feature 가 동일하다는 언급은 다양한 case 의 실험 과정에서 생긴 오해이다. 그러나 중요한 것은 굳이 세가지 방법의 주요 feature 가 일치하지 않더라도 연구의 contribution 에 전혀 영향을 끼치지 않는다는 점이다. 중요 feature 와 prediction 수치, 방법론, EDA 과정까지 이 사실로 영향을 받는 결과는 없다.

2. Young Data(GITHUB appendix 폴더에 포함)

본래 우리는 현재의 미국 농림부 데이터 이외에 슬로바키아 통계학 수업에서 수집된 "YOUNG PEOPLE DATASET"을 동시에 활용하였다. 그 과정에서 Categorical 데이터에 대한 변환, 정제, feature 분석, 예측, EDA 등을 모두 진행하였다. Young people dataset 의 특징은 음식 습관과 관련이 없는 다양한 column 이 존재한다는 것이었다. 100 개가 넘는 column 중에는 음악 취향, 영화 취향, 심리검사 자료, 공포증 자료, 소비 습관, 신체 지수 정보, 출신지 정보 등이 포함된다. 그러나 저희가 발견한 것은 이러한 취향 정보가 BMI 수치나 비만율과는 크게

관계가 없다는 사실이었다. 대부분 큰 관계성을 보인 feature 들은 성별, 나이와 같은 신체 연관 특징이었고, 근소하게나마 연관성을 보인 것이 교육 수준, 기상 습관, 약속 시간을 지키는 지 유무 등이었다. 그러나 이러한 feature 들은 비만 유무를 예측하는데 불충분하여 63% 정도의 정확도만을 나타내었다. 무엇보다 큰 문제는 해당 데이터의 총 크기가 1K 인데 비해서 비만(BMI>30)인 사람의 수는 오직 13 명 정도였다는 점이었다. 이는 슬로바키아와 미국의 국가적 비만도 차이라고 생각된다. 비만 인구 13 명으로는 제대로 된 train, test set 을 꾸릴 수 없었고, Balanced bagging 방법을 사용하더라도 데이터 적 한계를 넘기 힘들었다. 따라서 최종적으로 Young data 를 발표에서 제외하기로 하였다. 그러나 해당 데이터에 대한 분석을 통해 얻은 것은 개인의 영화 취향이나 음악 취향 등과 BMI 는 유의미할 정도로 연관성이 없다는 것이었다. 해당 데이터와 분석했던 코드는

https://github.com/tjddus9597/Team12_BigData_BMI 에서

확인할 수 있다.

3. 생성했던 데이터 파일 목록 첨부



4. 회의록 첨부

> 10.31 – Kick off meeting

: 주제 탐색. 범죄 데이터, 경마 데이터, 대학 데이터, 연봉 데이터 등이 후보군으로 선정됨. 대부분의 한국 공공데이터가 정리된 형태여서 raw 데이터를 얻기가 힘들다는 단점이 있었음.

> 11.1 – 2 차 주제 선정 미팅

: 최종적으로 경마, 범죄자 교육 정보, 학생 피해자, 이미지, 날씨 데이터 등이 후보군으로 오름. 관련하여 교수님께 상담을 요청하기로 함.

> 11.3 – 교수님 상담

: 해외 데이터를 사용하거나 국내 데이터를 합치는 방식을 활용하라는 조언을 얻음

> 11.3 – 최종 데이터 선정 모임

: 최종적으로 Young people data 를 사용하기로 결정

> 11.8 – proposal 생성 모임

: 생성된 데이터셋 정보를 기준으로 proposal 을 만듦

> 11.9 – proposal 발표

> 11.27 – 분석 모임

: 데이터 분석 방식을 생성 – 특징 분석 – 예측 – 시각화 순으로 나누어서 진행하기로 함. 각 파트에 2 명씩 번갈아가면서 일을 분배함.

> 11/29

: 추가적인 데이터로 미국 농림부의 2015 년 데이터를 Kaggle 에서 얻음. Young 과 Food 데이터 2 개에 대해서 병렬적으로 분석을 진행하기로 함.

첫 데이터 생성 파트를 수행함. 데이터 정제, 모순되는 column 을 NAN 으로 바꾸고, Nan 값 handling 에 대한

토의를 진행

> 11/30

: random forest 방법론과 correlation 을 통한 특징 분석을 진행. Young data 의 category 가 string 형태여서 이를 각각 int 형으로 바꾸는 작업을 진행. NAN 값이 너무 많다는 한계점이 드러남. 당시 문제 상황에 해당되는 정리는 다음과 같음

해결책으로 생각한 것들이 있는데 성호 너랑 일단 공유 할게.

issue1) 무응답, 응답거부, 까먹은 질문을 현재 모두 NAN으로 처리했는데, 여기에 label을 줄 지 여부.

issue2) NAN을 없애는 방법 중 하나는 NAN이 하나라도 있는 row를 없애는 건데 이러면 데이터가 줄어듦.

issue3) NAN값이 몰려있는 column들이 있는데 그런 column을 날릴 지, 날린다면 어느 정도 갯수의 NAN에 대해 날릴 지.

issue4) NAN 값에 차라리 0 같은 label을 줄 수도 있을 것 같은데 어떻게 할 지.

오전 12:25

>12/2

: Missing 데이터를 handling 하는 방법을 조사함.
Distribution 을 그리고 거기서 random 으로 뽑는 방법이 제한됨.

>12/3

: MLP, ANOVA, forest 방법으로 prediction 을 돌림.
처음에는 정확도가 99%에 육박했지만 이는 비만인

사람이 극도로 희귀하여 test set 이 biased 된 결과라는 것을 확인. 이를 해결하기 위한 방법을 조사함. 이를 통해 balanced bagging 을 얻고 이를 예측 방법의 수단을 선정

>12/4

: Nan 값을 distribution 에서 random 하게 뽑는 방식이 오히려 accuracy 를 낮춘다는 사실을 확인함. 어떻게 해야 accuracy 를 올릴 수 있을 지를 고민. 또한 accuracy 이외에 객관적인 evaluation metric 을 고민하고 조사한 결과 auroc 방식을 채택하기로 함.

>12/6

: 결국 우리가 핵심으로 생각하는 column 을 뽑고 뽑은 column 의 nan 을 가진 row 만을 제거하는 방식이 가장 높은 성능을 나타냄. Kaggle 데이터 이외에 Kaggle 의 source 에서 미국 농림부 자료인 것을 확인하고 농림부 홈페이지를 통해 2015 년 데이터를 추가적으로 얻음. auroc 값이 70 을 넘어서 유의미한 수준으로 들어옴.

>12/7

: 분석한 데이터로 얻은 인사이트를 정리하는 시간을 가짐.

발표 자료를 형태와 구성을 선정. 회의 요약은 다음과 같음

>주제
Feature와 Prediction 결과를 가장 잘 활용할 수 있는 상황을 설정. 우리가 하고자 하는 목표를 밝히기.
>데이터 소개(Young, Food)
- 출처, 갯수, 열(column)에 대한 소개
- 데이터 정제 : 의미 없는 column 삭제, weight/height outlier 삭제, 모순되는 값들 삭제, null 값 handling(column 선택, row 선택, random pick)
>중요 Feature 뽑은 결과 => 결과에 대한 분석 => 분석된 결과에 대한 활용 방안(health, income, exercise, snap, soda)
- Correlation
- Anova
- Random forest
>Prediction => 결과 제시 => 결과에 대한 자평 => 결과 활용 방안
(accuracy : 0.67, auROC : 0.73)
- ML(Random forest, SVM, MLP)
- Balanced Bagging
- evaluation metric(accuracy, auROC)

오후 4:26

>12/8

: 주로 얻은 인사이트 등을 공유하는 시간을 가짐. 정리본.

[food]
BMI와 비만율이 높을 수록, 자신의 신체적 건강 상태에 대해서 부정적으로 인식한다.
운동을 하지 않을 수록 BMI와 비만율이 높다.
snap의 혜택을 누리지 못할 수록 BMI와 비만율이 높다.
==> "Snap의 효과를 입증했다."
수입이 낮을수록 BMI와 비만율이 높다.
==> "소득수준이 낮을수록 soda를 많이 먹는다.(-0.06)"
soda를 먹을 수록 BMI와 비만율이 높다.
"eudrink -> "eusoda -> eudietsoda"
==> "Diet soda의 유무가 bmi와 경향성을 띈다."
==> "소득수준이 낮을 수록 diet soda를 안먹는다.(0.198)"
==> "snap 정책의 혜택을 받으면 bmi 낮은 경향이 있음에도 diet soda를 안먹는다.(-0.13)

오전 1:48

>12/9

12/8 의 회의가 이어진 새벽에, 몇몇 인사이트에 있어서 true(1)와 false(2)를 착각하는 문제가 발생하여 이를 수정한 결론을 생성함.

비만에 가장 큰 요인을 끼치는 5가지 특징은, 건강에 대한 자신감, 수입, 운동, 탄산음료, SNAP이었다. *SNAP은 미국에서 저소득층의 균형 잡힌 영양 공급을 위해 시행된 식사 쿠폰 개념으로, 미리 허가된 가게에서만 사용된다. (패스트 푸드와 같은 정크 푸드점은 당연히 포함되지 않는다.)

우리는 저소득층 일수록 비만율이 높다는 것을 확인했다.(correlation 0.09, p-value 2.8e-37) 따라서 비만 관리 차원에서 특히 저소득층에 집중해야 한다. 소득에 따른 EDA를 실시한 결과 저소득층에서 두드러지는 비만 요인은 탄산음료 소비량, 운동 부족, 그리고 이었다. 흥미로운 점은 영양 불균형을 해결하기 위해 SNAP의 혜택을 받고 있는 저소득층의 비만율이 오히려 높았다는 점이다.(correlation -0.08, p-value 2.5e-33) 이는 SNAP 정책이 오히려 저소득층 비만율을 높이는 결과를 가져온다는 것을 의미한다. 그 이유를 밝히기 위해서 SNAP과 관련하여 EDA를 실시하였다. SNAP은 패스트 푸드 소비를 줄이는 데에는 일부 효과를 보인다. (p-value 0.0013) 그러나, 전반적인 식사량(p-value 1.3e-6), 간식량(p-value 3.3e-2), 고기 소비량(p-value 5e-4), 음료 섭취량(p-value 4.4e-4)을 증가시킨다. 따라서 SNAP 정책이 규제 내에서는 충분히 지켜지지만, 과다한 식품 섭취로 인한 비만을 유발한다고 볼 수 있다. 따라서 SNAP의 확장으로 일반 탄산음료 대신 다이어트 탄산음료를 권장하는 방법과 운동을 독려하는 방식을 제안한다. 검정 결과 다이어트 소다를 먹었을 경우, 일반소다와 비교하였을 때 유의할 정도로 BMI 수치가 낮아짐을 확인했다. 또한 예상할 수 있듯이 BMI와 운동 빈도는 큰 관련성이 있다. 소득에 관계없이 비만율이 높은 사람의 경우, 자신의 건강에 대한 부정적인 인식을 가지고 있다. 따라서 이러한 정책이 건강을 증진시킨다는 점을 충분히 설명한다면 사람들의 동의를 얻을 수 있을 것이다.

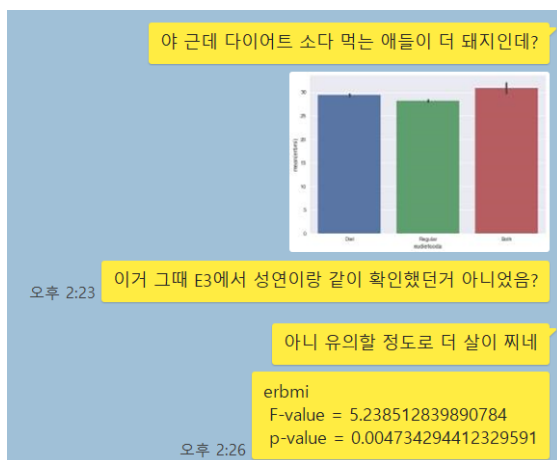
정책 격자에 영향력 축 수 인력 마하 격정

>12/9

: 다이어트 소다 섭취에 따른 비만 원인에 대한 논쟁이 벌어짐.

다이어트 소다를 먹을 수록 비만이 되는 원인을 설명할 수 있는

근거나 관련 EDA 결과를 얻지 못해서 곤란을 겪음.



>12/11

: 최종 결론 생성

오후 11:16

<결론>

1. Snap의 필요성은 인정하나 그 한계가 있음. 그 한계를 분석해본 결과, 식사량 전반이 너무 증가하고 특히 군것질이 너무 많아져서 개선이 필요하다.
2. 소다 보다는 주스 쪽으로 제안해야 한다.
3. Snap과 같은 방식의 운동 독려 정책이 필요하다.

*Genhealth : 해당 정책이 의미가 있을 거임.

>12/12

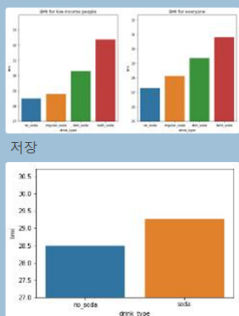
최종 발표에 필요한 visualization 추가 생성

김성연(디지)

food_plot.ipynb
111.98KB | 12.26까지
오후 11:05

저장 | 다른이름 저장

김성연(디지)



저장

오후 11:06

저장

>12/14

최종 PPT 완성 및 발표 준비용 script 작성

심성호

Team12_presentation.pptx
13.45MB | 12.28까지
오후 11:52

열기 | 풀더열기

Team12_presentation_세현 부분 최종.pptx
12.44MB | 12.28까지
오후 11:44

열기 | 풀더열기

김성연(디지)

Team12_Real_Final_presentation.pptx
12.45MB | 12.28까지
오후 11:49

저장 | 다른이름 저장

>12/15

최종 발표