**빅데이터 분석 및**

**시각화 개론**

**최종 보고서**

**2017. 12. 15.**

**12조**

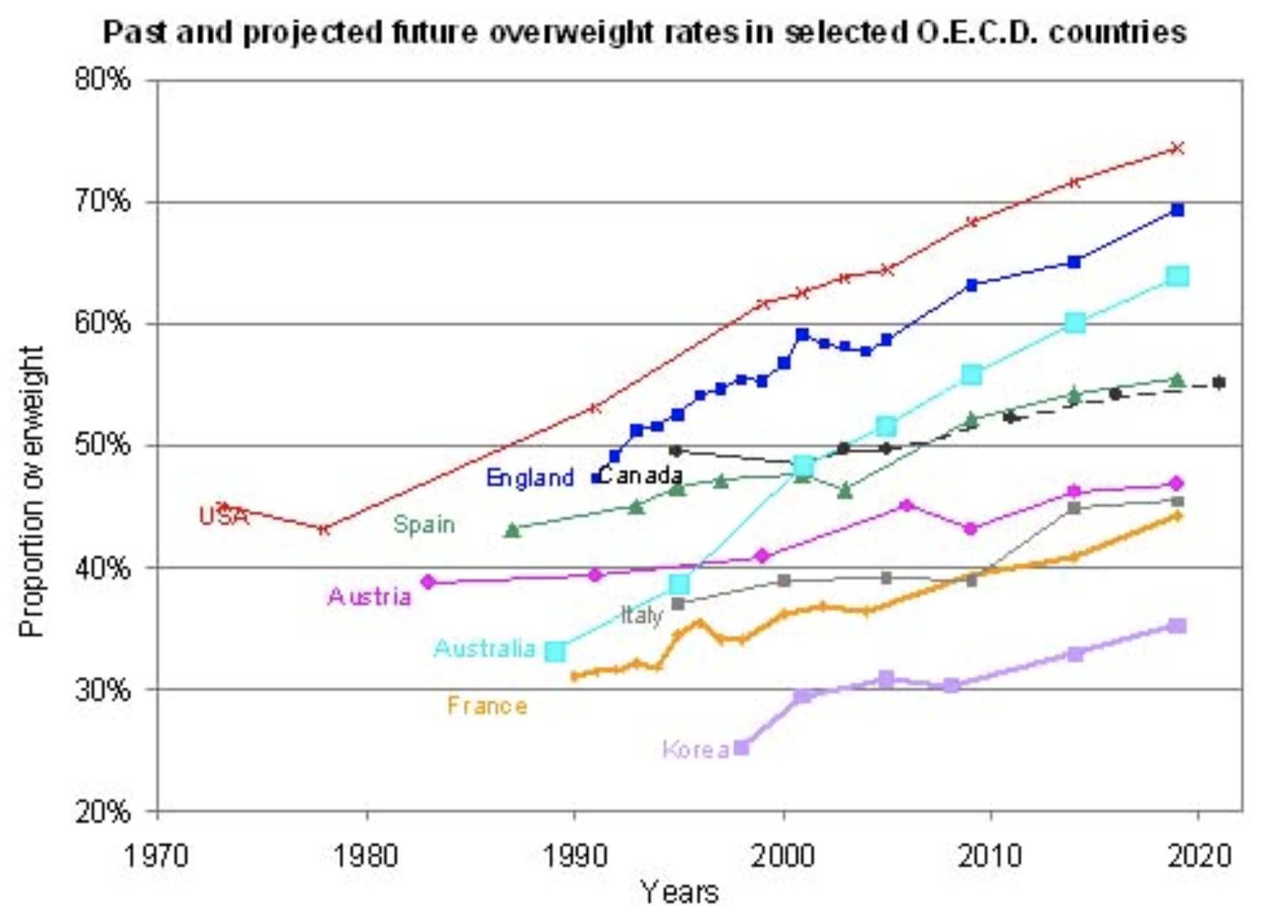
**201411024김성연**

**201411075심성호**

**201411166황세현**

**Contents**

1. **Introduction**
2. **Materials**
3. **Method**
4. **Result**
5. **Discussion**
6. **Appendix**
7. **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.

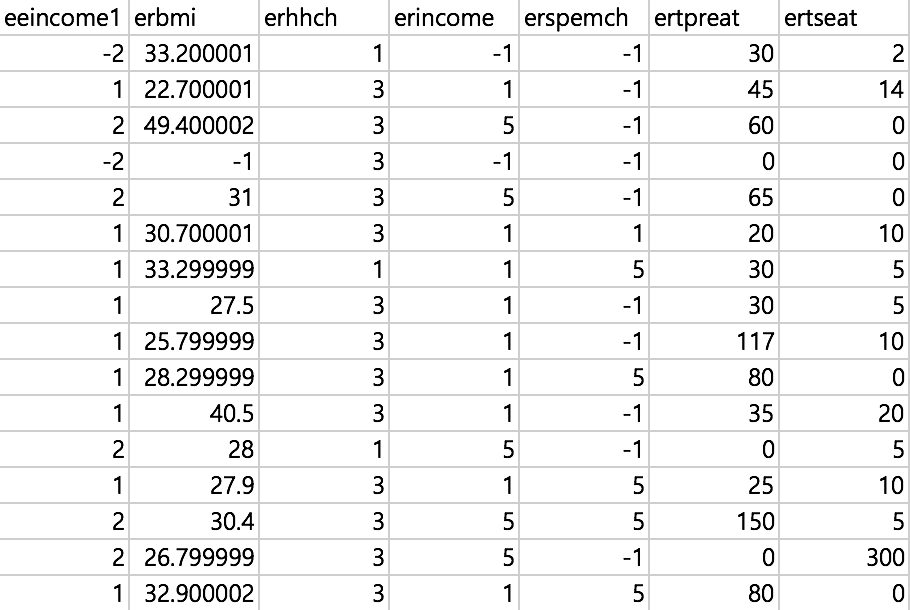
우리는 세계 보건기구(World Health Organization) 입장에서 빅데이터 분석을 통해 점점 증가하는 비만율을 관리하기 위한 정책을 제안하고자 한다.

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 feature를 이용하여 비만 유무를 검사하는 predictor를 만들고 그 성능을 검증하였다.

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

1. **Materials**

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



(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.

1. **Method**
2. Data refinement

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

데이터에 BMI가 30이상인 사람, 즉 비만인 사람의 수가 BMI가 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 filling을 해보았다. 해당 열의 평균과 분산을 구해서 truncated normal distribution을 만들어 이 distribution에서 해당 열에서 Null value인 개수 만큼 sampling하여서 채우는 방식으로 실험해보았다. 하지만, 이렇게 Null value filling을 하였을 때가 결과가 더 좋지 않아서, null value handling은 사용하지 않는 것으로 하였다.

1. Feature selection

Pearson Correlation, Random forest, ANOVA를 통해 BMI와 연관성이 가장 높은 feature 5개를 뽑았다. BMI와 연관성이 높고, ANOVA시 p-value가 0.05 미만인 feature 5개를 선정하였다.

* ‘eugenhth’ : 건강 자신감
* ‘euexercise’ : 운동 유무
* ‘erincome’ : 소득 분위
* ‘eusoda’ : 탄산음료 음용 유무
* ‘eusnap’ : 영양 보충 보조 프로그램 수혜 유무

1. Prediction

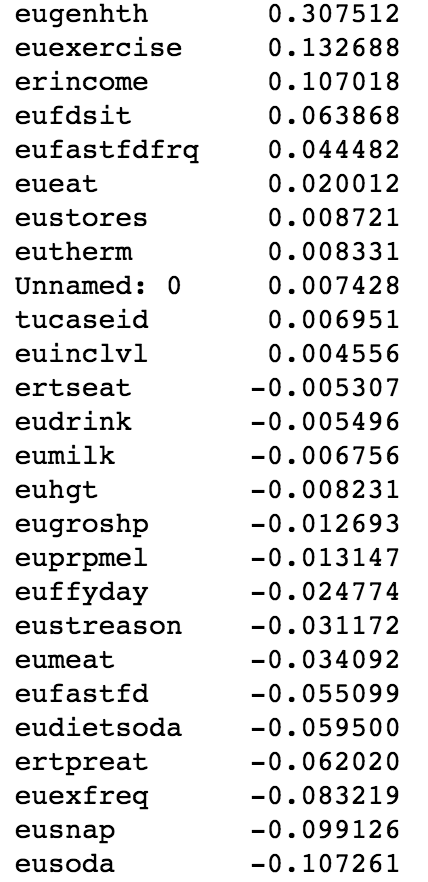
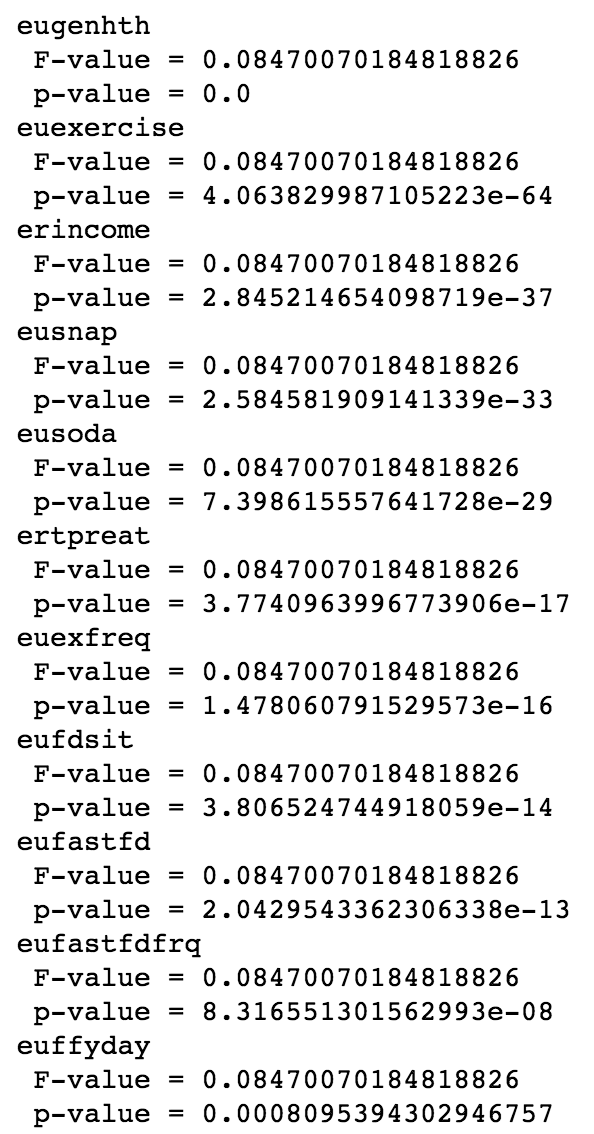
위의 feature selection에서 선정한 5개의 feature들이 실제로 비만과 관련이 있는지를 확인하기 위해서, 위의 feature 5개를 이용하여 비만을 예측하는 모델들을 학습시키고 성능을 확인해보았다. 이때 사용된 학습 데이터는 약 13k이다. 비만인 사람의 수는 그렇지 않은 사람의 수 비해 1/4수준이었다. 따라서 training 데이터가 불균형하여 예측 결과가 한쪽으로 몰리는 현상이 발생했다. 이를 해결하기 위해Test 데이터에 대해서는 앞써 언급한 down sampling 방식을 적용하였고, Training 데이터에 대해서는 Balanced bagging 방법을 확용했다. 이때 Balanced bagging의 커널로 사용한 머신러닝 기법으로는 Random forest, Support vector machine, Multiple layer perceptron이다.

1. EDA

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

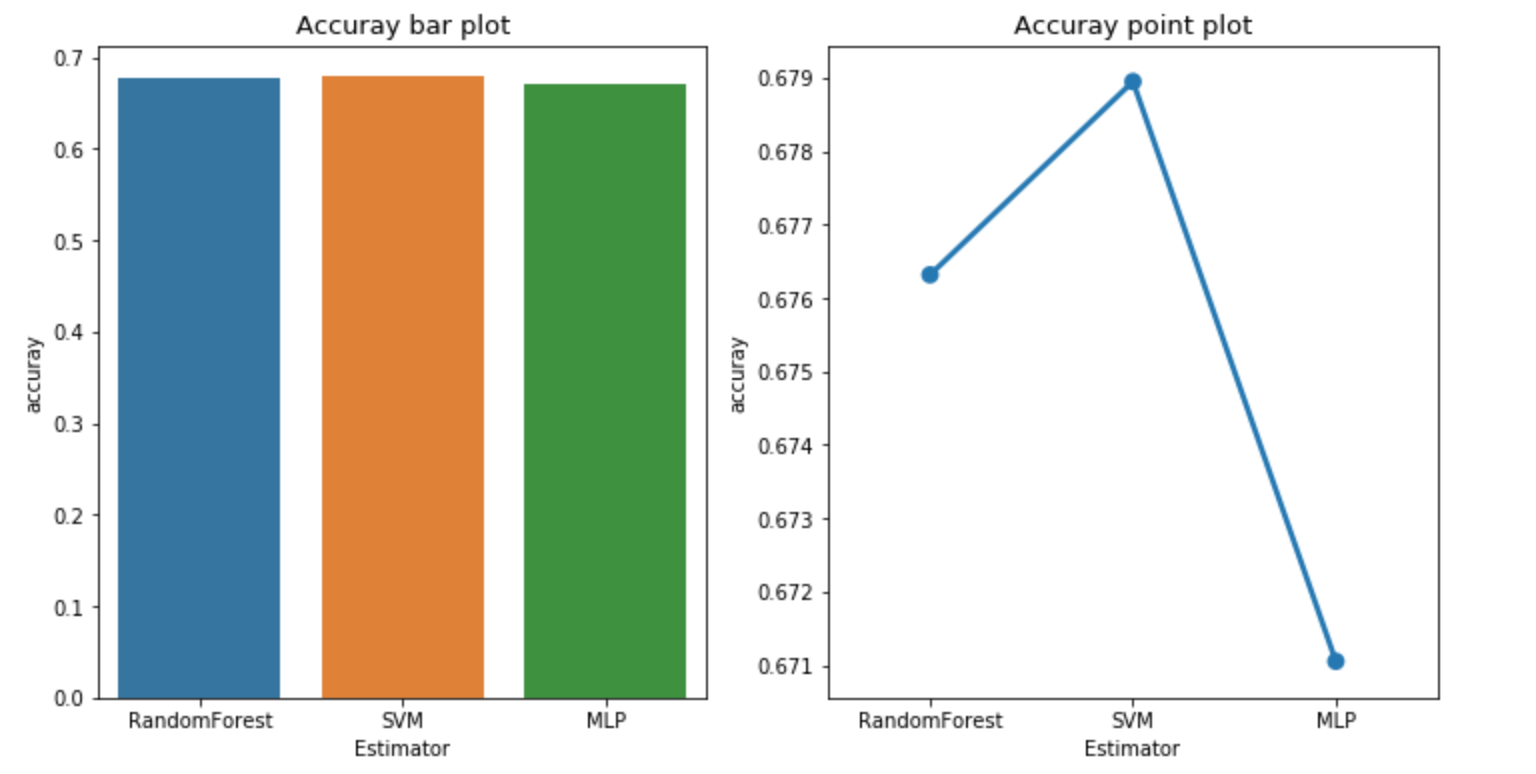
1. **Result**
2. Feature Selection

아래의 왼쪽 그림은 Anova 검정결과 p-value가 0.05 보다 낮은 column중 일부를 나타낸 것이다. 아래의 오른쪽 그림은 correlation의 결과 값의 절대값이 0.09보다 큰 column들을 나타낸 결과이다.

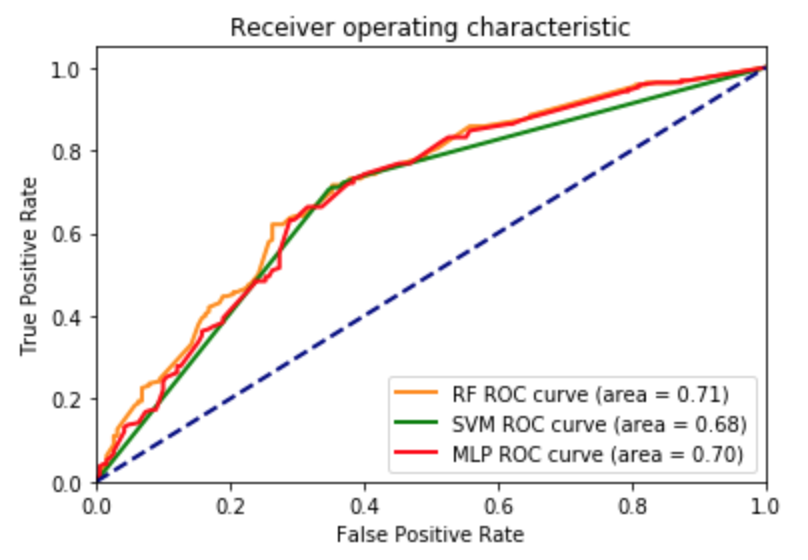


위의 그림에서 볼 수 있듯이, 한 방법에서 bmi와 높은 연관성을 보이는 feature들은 다른 방법에서도 좋은 결과를 나타내는 것을 알 수 있었다. 선택한 5개의 feature들이 실제로 bmi와 관련성을 보인다는 것은 아래의 EDA 항목에서 추가적으로 설명할 것이다.

1. Prediction



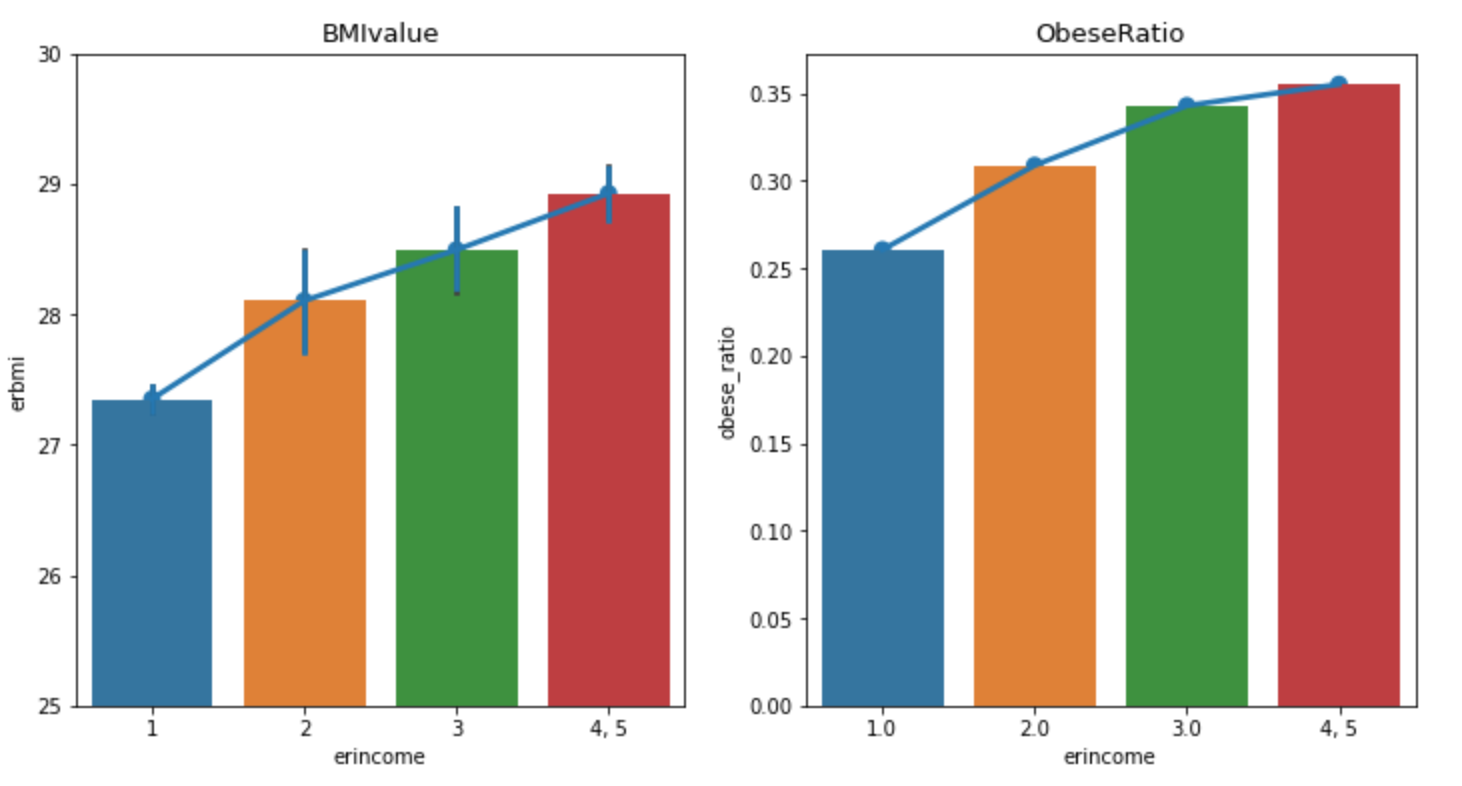
위의 Feature Selection으로 얻은 5개의 feature로 머신러닝 모델을 학습하였다. 사용한 모델은Random Forest, SVM, MLP이다. 세 모델 모두 68%정도의 비슷한 성능을 내고 있는 것을 확인하였다. 비슷한 성능을 내는 이유는 모델을 학습하기에는 데이터가 많지 않기 때문으로 보인다. 모델의 성능을 볼 수 있는 다른 지표로는 ROC Graph의 아래 면적 넓이인 auroc 값을 사용하였다.



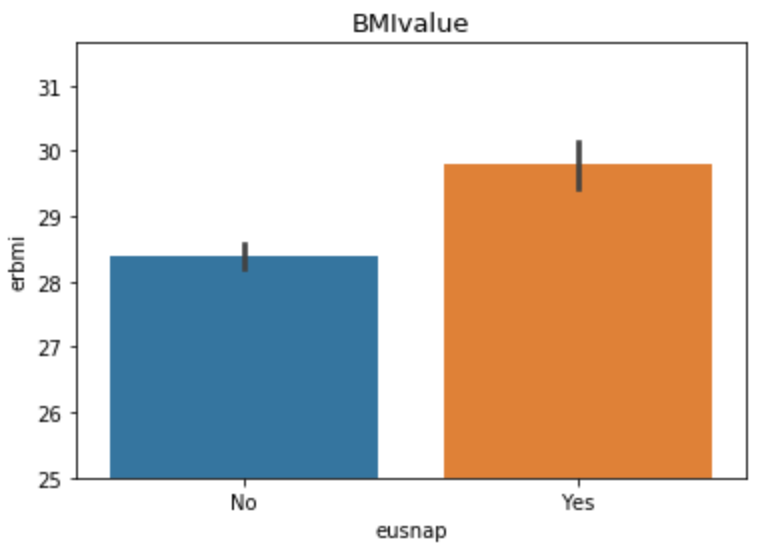
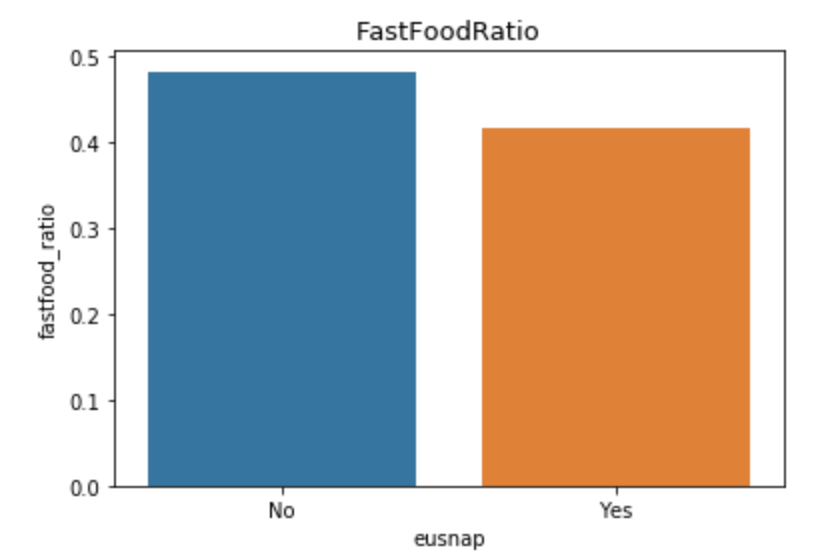
각 모델의 auroc값은 약 0.7정도의 값을 갖고 RF, MLP, SVM 순서로 높은 값을 갖는 것을 알 수 있다. 일반적으로 auroc값이 0.7이상일 때 중등도의 정확한 모델이라고 볼 수 있는데, 우리의 모델들이 약 0.7정도의 성능을 내는 것을 보아 괜찮은 성능을 내는 모델을 얻었다고 할 수 있을 것이다. 이를 통해서 우리가 선택한 5개의 feature들이 비만과 관련성이 높고, 비만을 예측하기에 좋은 feature라는 것을 확인할 수 있었다.

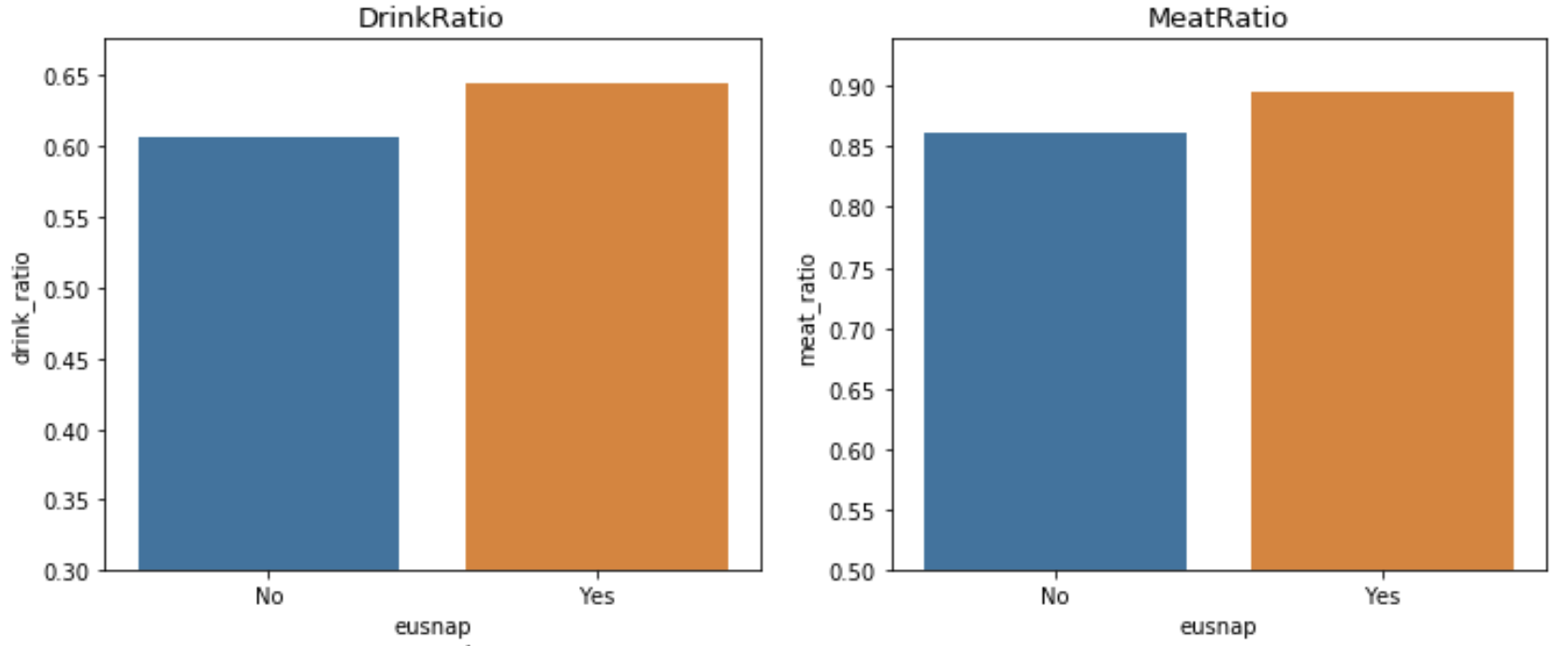
1. EDA

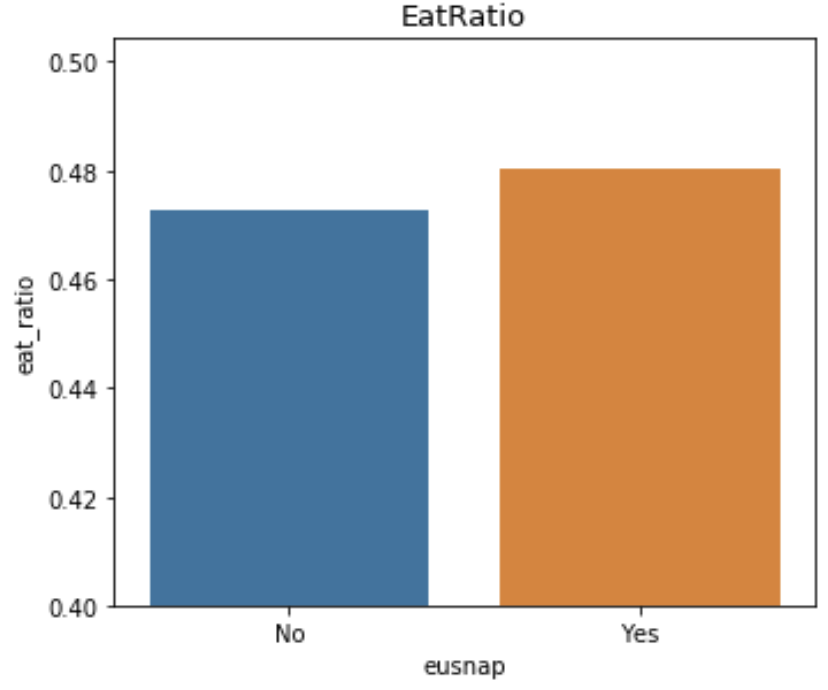
우리는 BMI와 Income이 많은 연관성이 있다는 것을 확인하고 income에 대한 심도 깊은 조사를 해보기로 하였다.

****

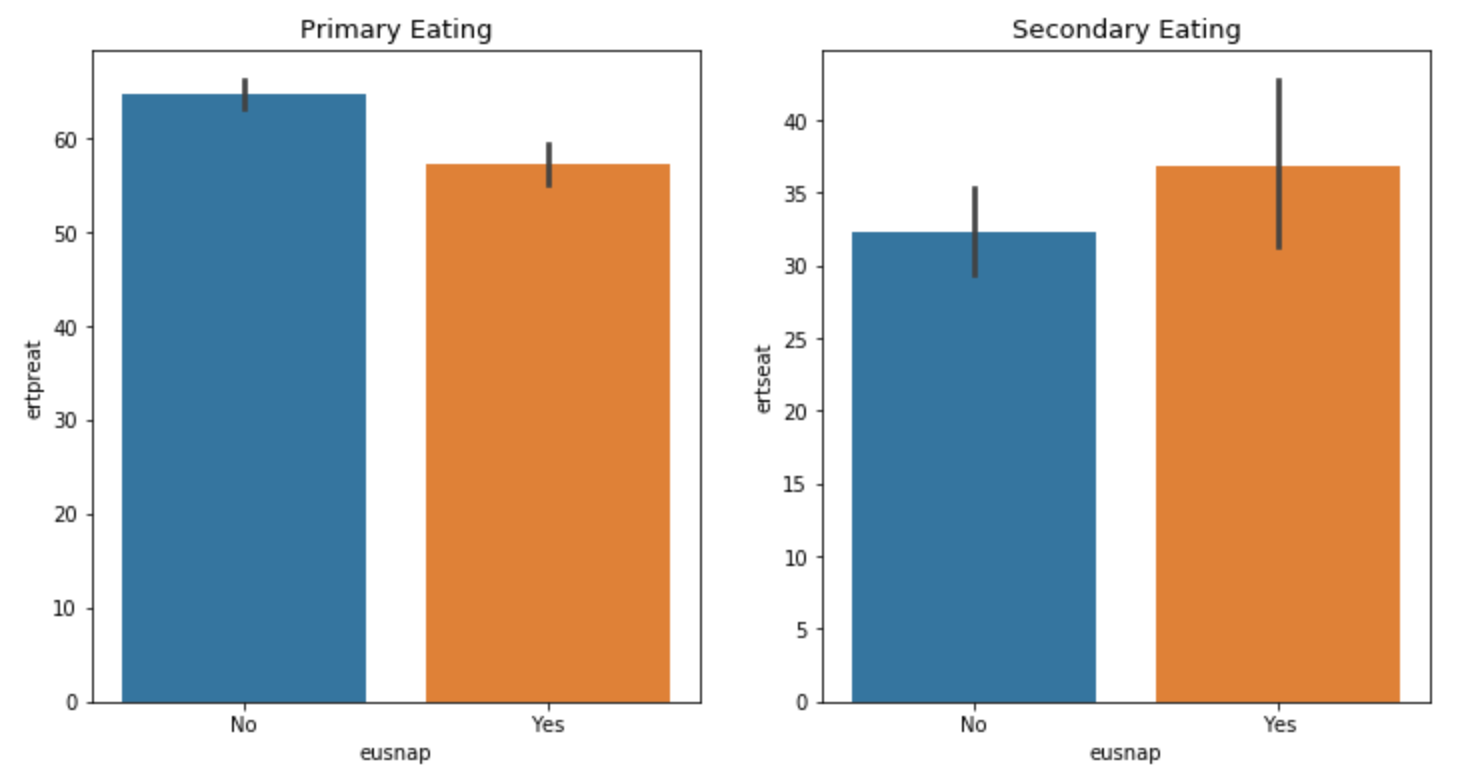
위 그래프는 income과 BMI 지수의 평균과 비만 비율을 나타낸 그래프이다. Income의 값이 커질 수록 소득 분위가 낮은 것이다. 위 그래프에 따르면 소득 분위가 낮을 수록 bmi와 비만의 비율이 커지는 것을 확인할 수 있다. 또한 bmi와 관련 깊은 다른 feature로 SNAP이 있는데, SNAP이란 미국의 저소득층을 대상으로 하는 영양 공급 프로그램으로 쿠폰 형태의 SNAP을 영양적으로 유의미한 식료품 구매에 사용할 수 있도록 한 것이다. SNAP은 저소득층을 대상으로 하기 때문에 수입과 관련이 높다. SNAP은 영양 공급을 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 SNAP을 이용했을 때 BMI가 오히려 증가하는 것을 확인할 수 있었다.



****위와 같은 결과가 나오는 이유를 알아보기 위하여 income과 관련된 feature에 대하여 eda를 진행하였는데, 아래와 같이 snap의 혜택을 받았을 때 전체적인 음식 섭취량 및 간식량이 증가하는 것을 확인할 수 있었다.

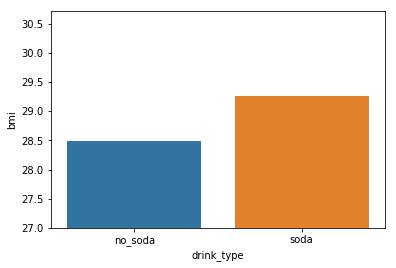
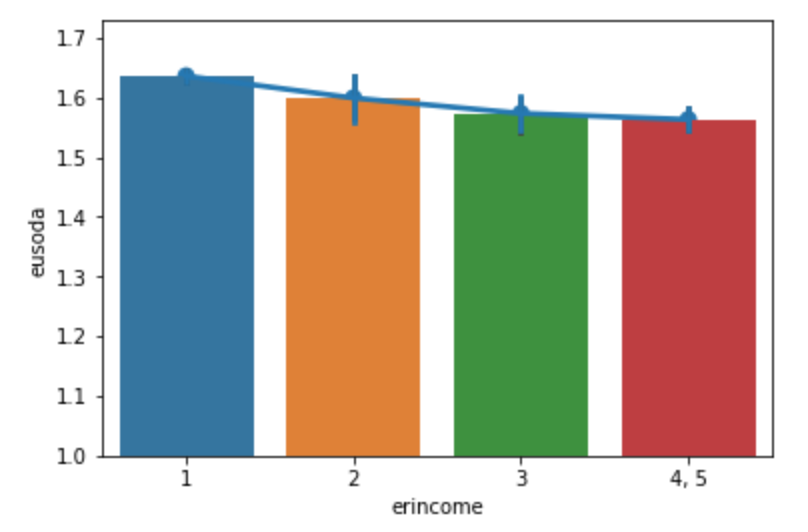
****

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

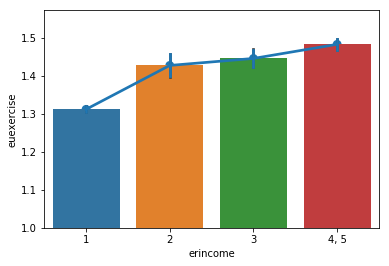
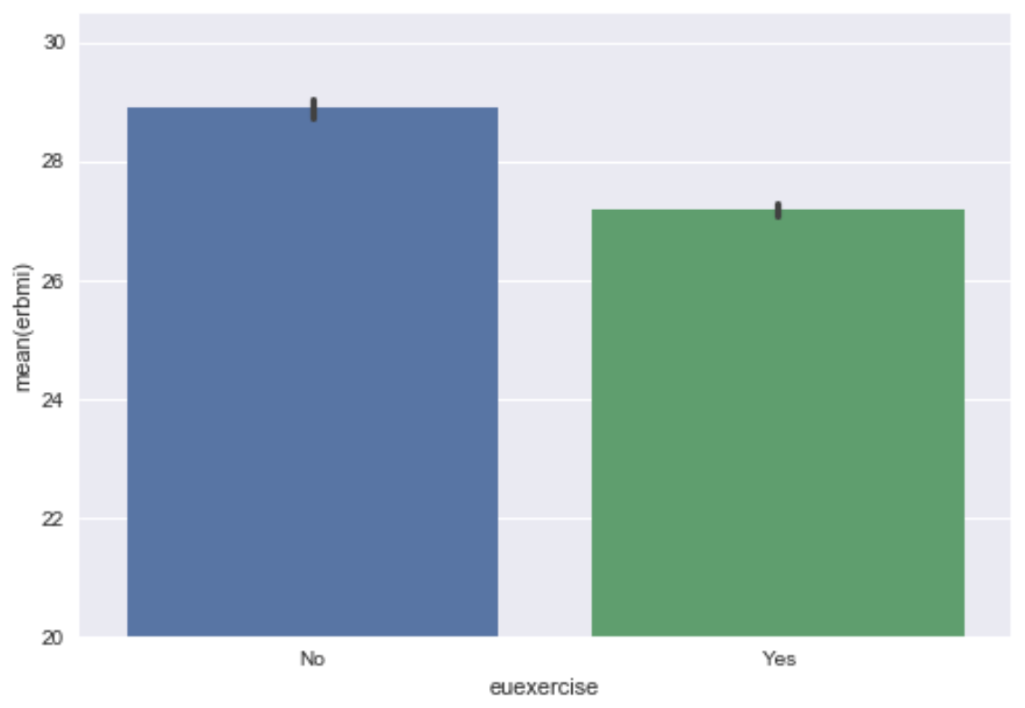
****

우리는 SNAP의 혜택을 받는 사람들이 비만이 되는 이유 때문일 것이라고 생각하였다.

또한 income이 낮은 사람일 수록 soda의 먹는 사람들이 증가하는 것을 알 수 있었다.



위 그래프의 y축은 1이 yes, 2를 no라고 했을 때의 결과의 평균을 구한 것이다. 또한 soda를 마시는 사람들이 soda를 마시지 않는 사람들에 비해 bmi가 높은 것을 보아, income낮은 사람들의 bmi가 높은 이유 중 하나로 soda의 섭취 유무가 관련 있다고 할 수 있다.



또한 income이 낮을 수록 운동을 하지 않는 사람들이 많아지는데, exercise를 하지 않을 수록 bmi가 높아지는 것을 확인할 수 있다. 따라서 income이 낮은 사람들이 운동을 많이 하지 않는다는 사실이 income이 낮은 사람들의 bmi가 높은 이유 중 하나라고 할 수 있다.

1. **Discussion**

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

* Policy#1

SNAP과 비슷한 정책을 실시하되, 지원금으로 구매할 수 있는 Secondary eating량에 제한을 두자.

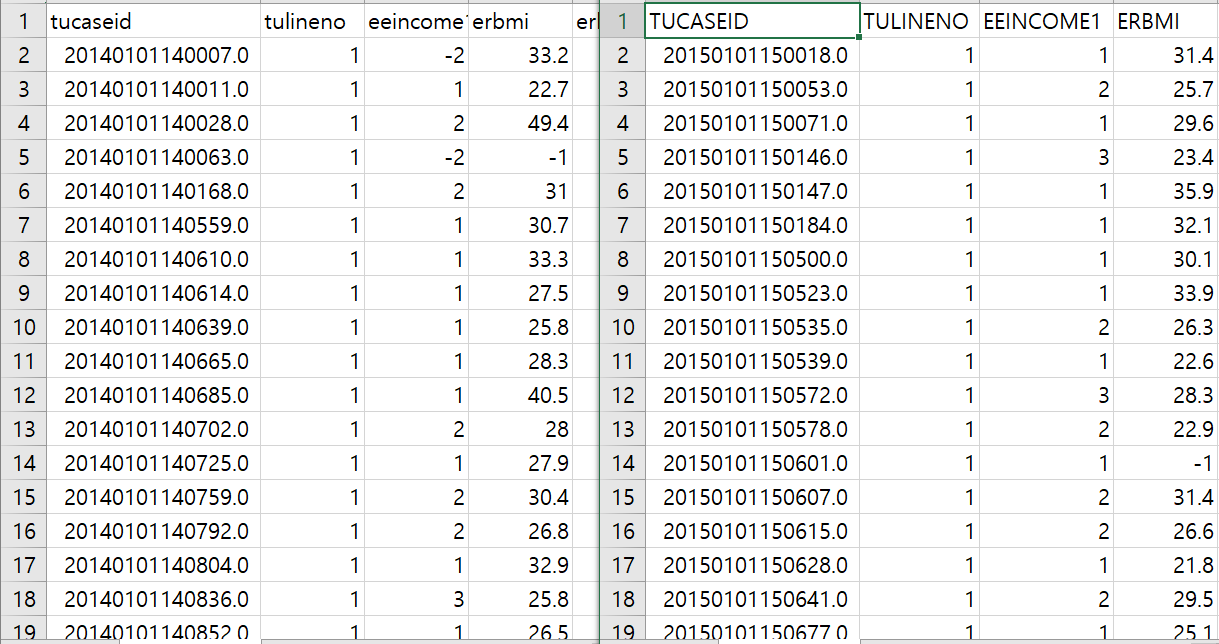
* Policy#2

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

* Policy#3

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

우리 연구의 한계점과 연구의 발전 방향은 다음과 같다. 우리가 2014년과 2015년의 ‘ATUS eating and health module’ 데이터를 사용하였는데, 만약 두 데이터가 동일한 사람들을 2년간 조사한 결과라면 각 feature들의 시간에 따른 변화까지 분석할 수 있어 더 재미있는 분석을 할 수 있었을 것이지만, 밑의 그림에 명시된 것처럼 ID값 자체가 2014년도와 2015년도가 다르게 적용되어서 feature들의 시간에 따른 변화는 분석할 수 없었다.



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

1. **Appendix**
2. PPT 발표에서와 달라진 부분에 대한 설명

앞선 결과 보고 발표의 Feature selection 과정에서 ANOVA와 Correlation 그리고 Random forest의 주요 feature 들이 모두 일치되게 구해졌다는 언급이 있었다. ANOVA와 Correlation의 결과가 일치하는 것은 사실이지만 Random forest에서 중요 feature로 뽑힌 것들 중에는 발표된 feature가 해당되지 않는 경우가 1~2개 정도 있다.(exercise, soda) 차이가 나는 feature들은 prediction을 진행하면서 성능이 더 높은 기준으로 판단하였고 그 결과로 얻어진 feature가 발표된 5개이다. 즉, random forest의 주요 feature가 발표된 주요 feature와 어느정도 비슷한 경향을 보이지만 완전히 일치하는 것은 아니며 그 진위는 prediction을 통해 판별했었다. 발표 때에 모든 방법의 feature가 동일하다는 언급은 다양한 case의 실험 과정에서 생긴 오해이다. 그러나 중요한 것은 굳이 세가지 방법의 주요 feature가 일치하지 않더라도 연구의 contribution에 전혀 영향을 끼치지 않는다는 점이다. 중요 feature와 prediction 수치, 방법론, EDA 과정까지 이 사실로 영향을 받는 결과는 없다.

1. 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 dataset의 특징은 음식 습관과 관련이 없는 다양한 column이 존재한다는 것이었다. 100개가 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 feature들은 성별, 나이와 같은 신체 연관 특징이었고, 근소하게 나마 연관성을 보인 것이 교육 수준, 기상 습관, 약속 시간을 지키는 지 유무 등이었다. 그러나 이러한 feature들은 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 13명 정도였다는 점이었다. 이는 슬로바키아와 미국의 국가적 비만도 차이라고 생각된다. 비만 인구 13명으로는 제대로 된 train, test set을 꾸릴 수 없었고, Balanced bagging 방법을 사용하더라도 데이터 적 한계를 넘기 힘들었다. 따라서 최종적으로 Young data를 발표에서 제외하기로 하였다. 그러나 해당 데이터에 대한 분석을 통해 얻은 것은 개인의 영화 취향이나 음악 취향 등과 BMI는 유의미할 정도로 연관성이 없다는 것이었다. 해당 데이터와 분석했던 코드는 <https://github.com/tjddus9597/Team12_BigData_BMI>에서 확인할 수 있다.

1. 생성했던 데이터 파일 목록 첨부



1. 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 – 2차 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 data를 사용하기로 결정

> 11.8 – proposal 생성 모임

: 생성된 데이터셋 정보를 기준으로 proposal을 만듦

> 11.9 – proposal 발표

> 11.27 – 분석 모임

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

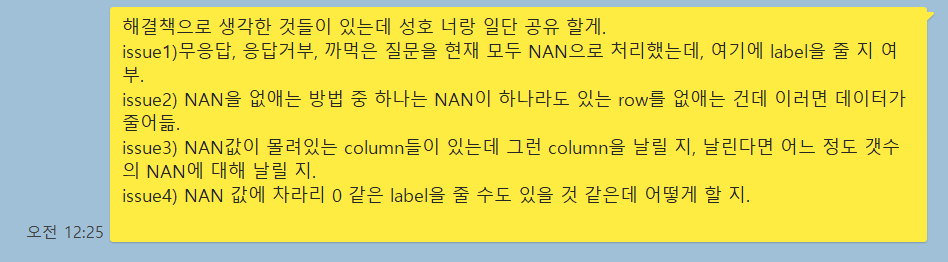
> 11/29

: 추가적인 데이터로 미국 농림부의 2015년 데이터를 Kaggle에서 얻음. Young과 Food 데이터 2개에 대해서 병렬적으로 분석을 진행하기로 함.

첫 데이터 생성 파트를 수행함. 데이터 정제, 모순되는 column을 NAN으로 바꾸고, Nan 값 handling에 대한 토의를 진행

> 11/30

: random forest 방법론과 correlation을 통한 특징 분석을 진행. Young data의 category가 string 형태여서 이를 각각 int형으로 바꾸는 작업을 진행. NAN값이 너무 많다는 한계점이 드러남. 당시 문제 상황에 해당되는 정리는 다음과 같음



>12/2

: Missing 데이터를 handling 하는 방법을 조사함. Distribution을 그리고 거기서 random으로 뽑는 방법이 제한됨.

>12/3

: MLP, ANOVA, forest 방법으로 prediction을 돌림. 처음에는 정확도가 99%에 육박했지만 이는 비만인 사람이 극도로 희귀하여 test set이 biased된 결과라는 것을 확인. 이를 해결하기 위한 방법을 조사함. 이를 통해 balanced bagging을 얻고 이를 예측 방법의 수단을 선정

>12/4

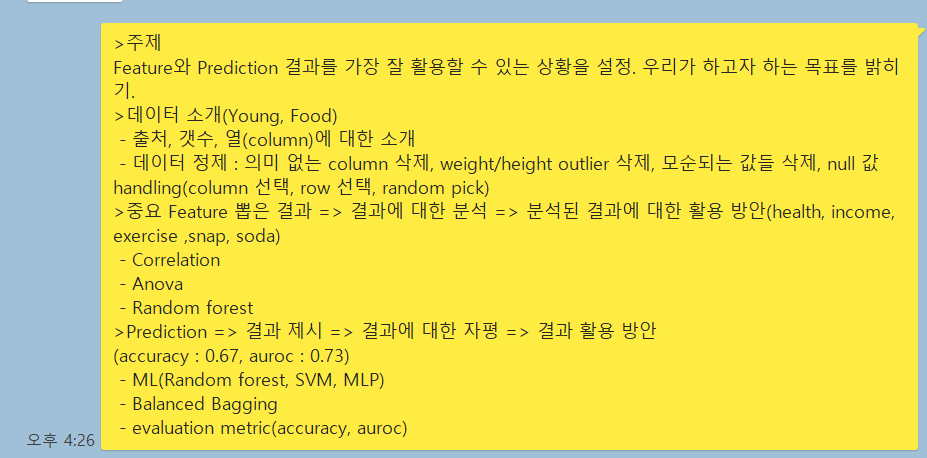
: Nan 값을 distribution에서 random하게 뽑는 방식이 오히려 accuracy를 낮춘다는 사실을 확인함. 어떻게 해야 accuracy를 올릴 수 있을 지를 고민. 또한 accuracy이외에 객관적인 evaluation metric을 고민하고 조사한 결과 auroc 방식을 체택하기로 함.

>12/6

: 결국 우리가 핵심으로 생각하는 column을 뽑고 뽑은 column의 nan을 가진 row만을 제거하는 방식이 가장 높은 성능을 나타냄. Kaggle 데이터 이외에 Kaggle의 source에서 미국 농림부 자료인 것을 확인하고 농림부 홈페이지를 통해 2015년 데이터를 추가적으로 얻음. auroc값이 70을 넘어서 유의미한 수준으로 들어옴.

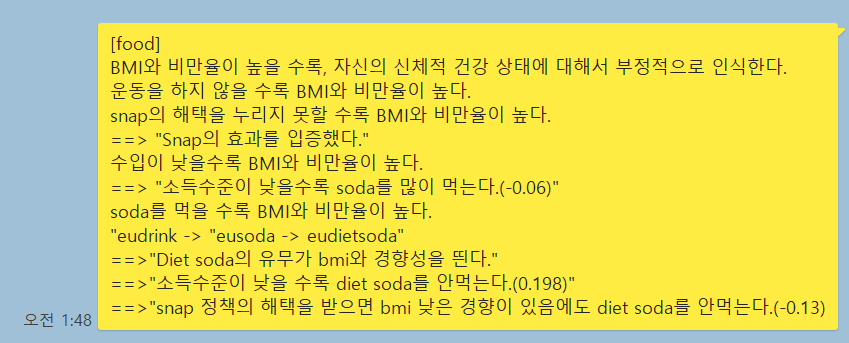
>12/7

: 분석한 데이터로 얻은 인사이트를 정리하는 시간을 가짐. 발표 자료를 형태와 구성을 선정. 회의 요약은 다음과 같음



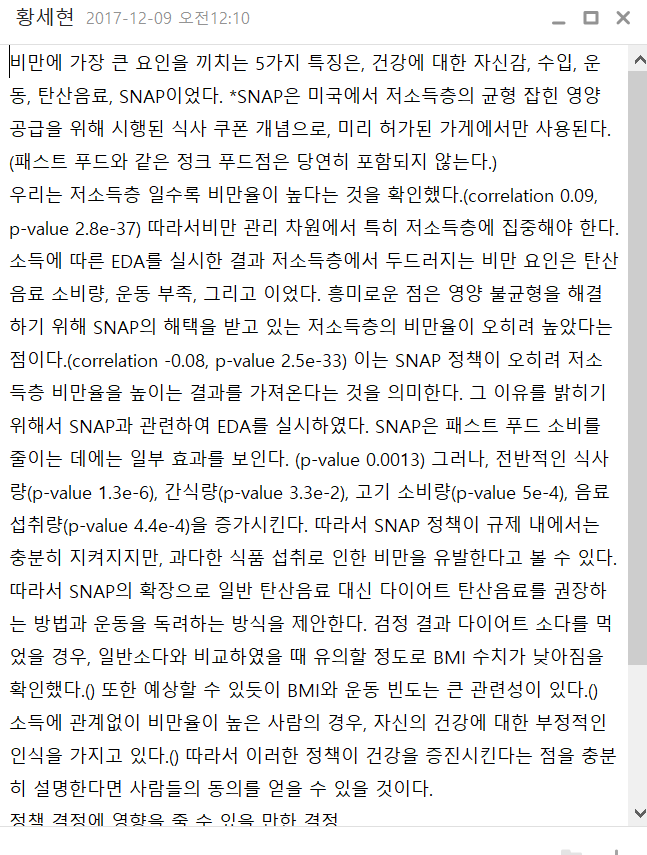
>12/8

: 주로 얻은 인사이트 등을 공유하는 시간을 가짐. 정리본.



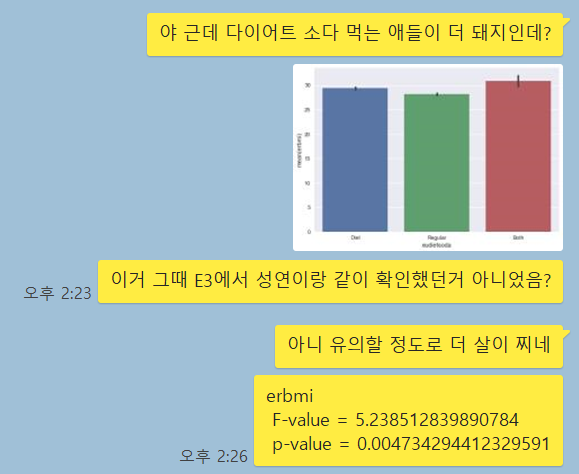
>12/9

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



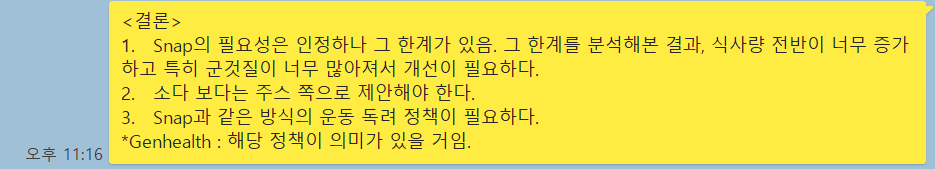
>12/9

: 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.



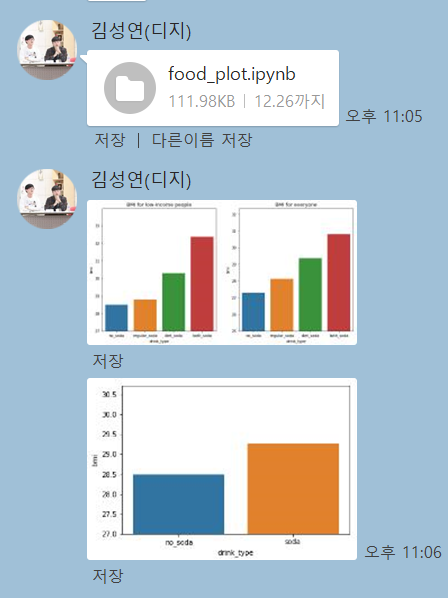
>12/11

: 최종 결론 생성



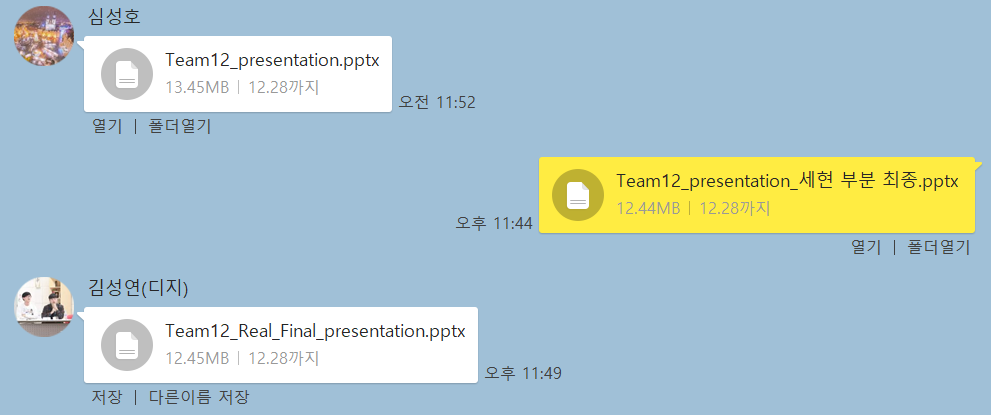
>12/12

최종 발표에 필요한 visualization 추가 생성



>12/14

최종 PPT 완성 및 발표 준비용 script 작성



>12/15

최종 발표