PDCD

관상동맥 질환의 위험인자 파악 및 질병 위험도 예측

한대룡, 백우현

1. 배경 및 방향 제시

과거의 관상동맥 질환과 달리 현재는 그 대상자가 광범위하다. 40대 이상은 물론 심지어 20~30대의 비교적 젊은 연령층도 비만일 경우 발생 위험이 크게 증가하는 것으로 밝혀졌다. 그러나, 문제는 아직까지 이 관상동맥 질환의 뚜렷한 원인이 밝혀지지 않았다는 점이다. 이에 관상동맥 질환자와 비질환자의 건강검진 정보를 바탕으로 그 두 집단의 생체적인 특성의 차이와 관상동맥 질환의 잠재원인을 알아낼 수 있다면 이는 관상동맥 질환의 치료는 물론 예방에도 큰 영향력을 끼칠 것이다.

이번 분석의 목적은 관상동맥질병을 사전정보를 통해 예측하기 위한 것으로 각각의 사람 별로 총 6기간의 코호트 데이터가 주어졌다. 그러나 모든 사람이 6번의 건강검진을 받은 것이 아니었고, 건강검진 기간 역시 상당히 상이하였기에 6기간을 모두 활용하기엔 무리가 있었다. 따라서 한정적인 시점, 즉 한 시점과 두 시점 건강정보만을 활용해 관상동맥 질환을 예측하는 것으로 방향을 잡았으며, 이에 맞게 데이터를 변형할 필요가 있었다.

그리고 무엇보다도 이 데이터에서 가장 큰 걸림돌은 극단적인 unbalanced data라는 점이었다. 건강 검진 데이터의 특성상 질병에 걸린 사람의 비율은 극소수일 가능성이 높은데 우리의 데이터 관상동맥 질환 환자의 비율 역시 3% 이하였기에 이를 고려한 모델링이 필수적이었다. 따라서 우리는 10번의 독립적인 undersampling을 진행하여 10개의 모델을 만들어 냈고, 각 모델에서 softprob로 산출된 최종 10개의 test data에 대한 예측값을 앙상블했다. 이를 통해 unbalanced Data문제를 완화하고 예측력을 높일 수 있었으며, 10개 모델의 Important Matrix를 종합하여 모든 모델에서 Important가 상위 20위 안에 들었던 변수들을 모아 질병 유무 구분에 중요한 변수 리스트를 확인할 수 있었다. 즉, 근 미래에 관상동맥이 걸릴 위험이 높은 사람들을 판별해주는 주요 Feature를 확인할 수 있었다.

2. Data preprocessing

2-1. Data Reshape & NA handling

일반적으로 질병예측을 위한 사전정보는 주어진 데이터처럼 오랜 기간 동안 수집되기 힘들다. 따라서 6개 구간별 변수를 활용하여 모델을 구성한다면 모든 사람들을 대상으로 정기적인 검진을 필수적으로 실시되지 않는 한 그 실용도가 매우 떨어질 것이다. 또한 관상동맥질병이 한 번이라도 걸린 적이 있는 사람, 즉

fuPDCD가 2인 사람들은 223명으로 전체 9468 obs의 2.3% 수준으로 생존분석을 진행하기에는 매우 적은 수라고 판단했다. 따라서 우리는 한정적인 시점(단 시점, 혹은 두 시점)의 데이터를 활용한 분류모델을 생성하기로 했다.

우선 6구간의 변수들을 각각 하나의 단 시점 obs로 분리해주었는데 해당 과정에서 처리해야 하는 NA값들은 두 종류가 있었다. 먼저 'V1' 시점에만 조사되는 V1_N_FAT, V1_N_CA, V1_EDUA와 같은 변수들의 결측치가 있는데 다른 V1시점의 값들에는 값이 입력되어 있기 때문에 해당 시점에 검사는 받았으나 입력을 하지 않거나 해당 종목만 검사받지 않은 것이었다. 또한 V1 시점에만 조사되는 변수들은 단시점 데이터로 구분할 때 모든 obs에 포함되는 중요변수기 때문에 해당 obs삭제 보다는 Imputation 하는 것이 합당했다.

Imputation은 크게 두 과정으로 나뉘었다. 우선 V1시점에만 조사되는 결측값들 중 Numeric 변수들은 'mice' 패키지를 활용한 Multiple Imputation을 사용했다. 그러나 하나의 method만 가지고 다중대체할 경우 결과값에 대한 신뢰도가 떨어지고 분산이 클 수 있기 때문에 선형회귀('norm.predict')를 통해 예측한 값과, pmm를 통해 예측한 값을 구하여 각 결과값의 평균을 최종 대체값으로 사용하였다. Maxit = 25, m (대체값 수) = 5로 설정했는데 해당 파라미터들은 간단한 Grid Search를 통하여 Imputation 결과가 가장 좋았던 값으로 결정했다. (대체 전 후 분포 변화, mice plot을 통한 Convergence 확인) 예를 들어 회귀대체를 통해 나온 5개 값들의 평균이 12이고, pmm이 14라면 최종 대체값은 두 값의 평균인 13을 사용한 것이다.

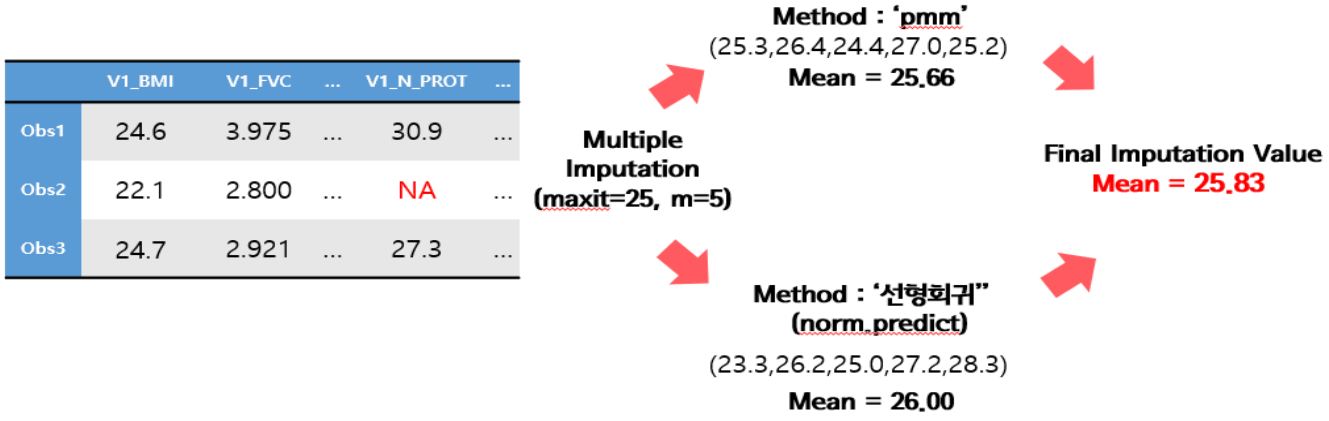

Figure 1 mice를 이용한 NA 대체

그러나 대체할 결측값 중 Factor 변수인 V1_PHYACTL , V1_PHYACTM , V1_PHYACTH, V1_EDUA에 대해서는 mice를 활용한 다중대체가 아닌 다른 방법을 사용했다. 로지스틱 회귀를 이용한 다중 대체를 사용할 수도 있었는데 좀 더 각 변수들의 특성을 세심히 고려한 방법이 더 합리적일 것이라고 생각했기 때문이다. 우선 V1_EDUA(교육수준)은 V1_INCOME(수익)과 밀접한 관련이 있을 것으로 예상했다. 실제로 수익이 높게 분류된 obs들이 대부분 EDUA가 높은 것을 시각화 결과 알 수 있었다. 또한 V1_PHYACTL , V1_PHYACTM , V1_PHYACTH (경동/중동/격동 활동시간)은 obs의 AGE(나이)와 연관성이 있고 일반적을 나이가 들수록 감소하는 것을 시각화에서 확인할 수 있었다. V1_PHYACTL , V1_PHYACTM , V1_PHYACTH에 대해서는 우선 AGE를 기준으로 kmeans (k=3) 방식을 사용하여 군집을 나누어 나이가 많고, 적고, 중간인 그룹으로 나눈다. 이후 이 Age 그룹을 EDUA의 INCOME처럼 활용하여 결측값이 있는 obs의 AGE그룹 내에서 랜덤샘플 1개를 뽑아 대체했다. 이를 활용한 대체방안은 다

음과 같다.

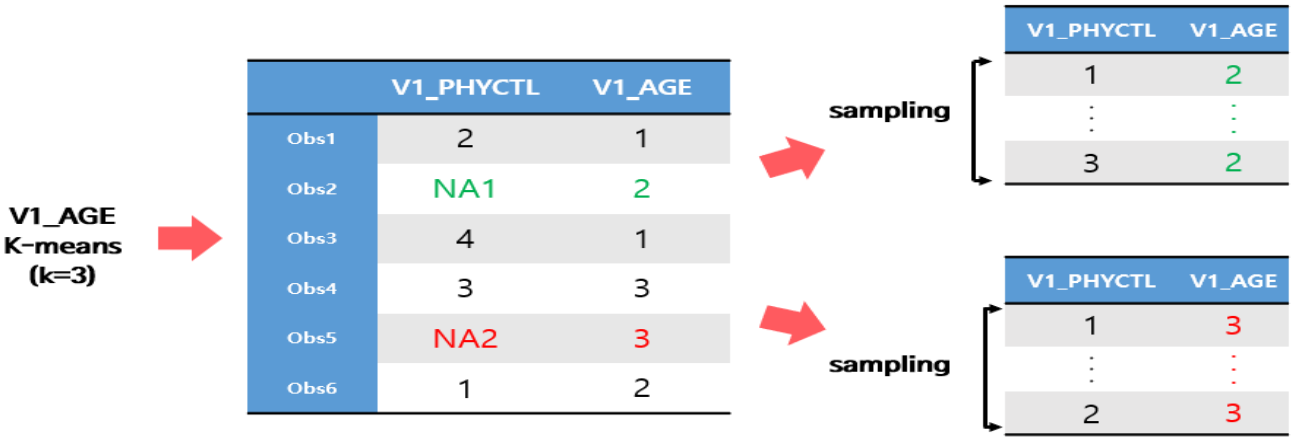

Figure 2 클러스터링을 이용한 NA대체

V1_EDUA도 앞서 설명한 '활동시간' 변수와 비슷하게 처리했는데, 다만 샘플링하는 기준 그룹으로 V1_INCOME을 사용했다. (위의 'V1_AGE'에 해당) 이러한 방식을 통해 INCOME이 1인 obs들의 EDUA 분포에 기반해 대체될 확률이 결정된다.

지금까지는 V1시점에만 있는 변수들 중 Numeric 변수는 Multiple Imputation으로, Factor 변수는 인접변수를 그룹화하여 그 안에서 랜덤추출하는 방식으로 대체했다. 두 번째 NA 유형은 단기간 obs로 나누면서 자연스럽게 조사가 시행되지 않은 경우에 발생하는 '조사받지 않은 시점의 값'이다. 예를 들면 어떤 관측치가 6개 구간의 PDCD hist가 [1 1 1 NA NA 1] 이고 결측값이 있는 4와 5 구간에는 조사를 받지 않은 것이다. 따라서 해당 obs에서는 1 2 3 6시점의 4개 obs만을 추출하고 4와 5구간의 값들은 사용하지 않았다. 앞서 설명한 결측값들과 다르게 이번 NA는 대체해야 할 값이 아닌 처음부터 조사를 받지 않은 구간의 값이기 때문이다.

결측값의 처리가 완료되었으므로 이제 1obs당 6구간으로 되어있는 데이터를 단기간의 여러 obs로 분리해낼 수 있었다. 아래에는 지금까지의 과정을 간단히 나타내는 간단한 예시이다.

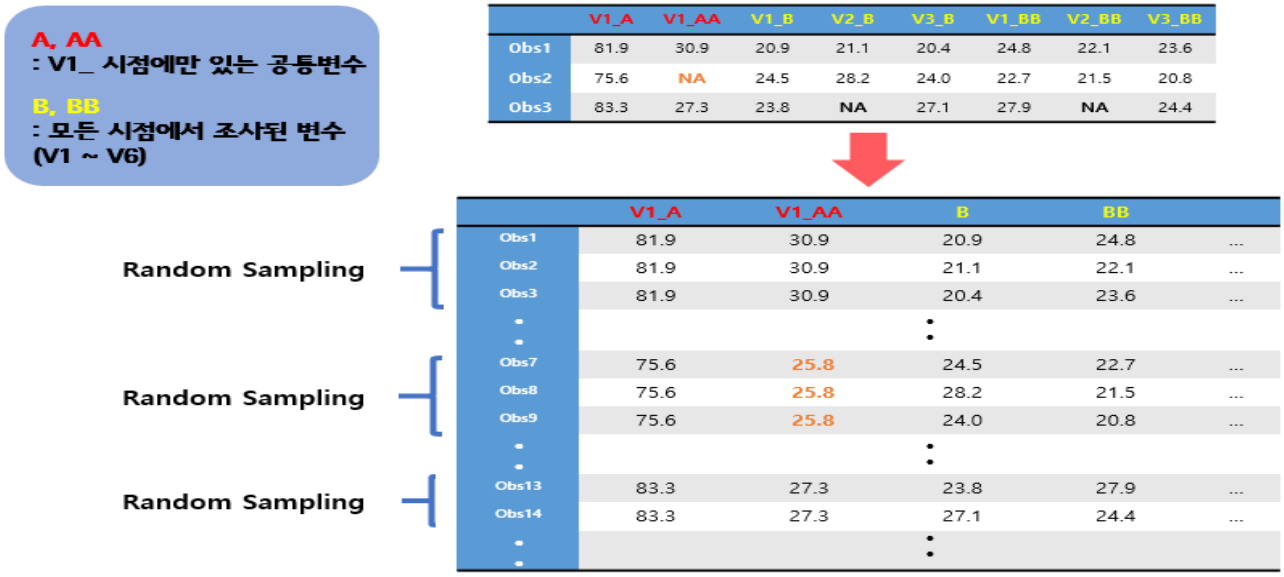

Figure 3 전처리 과정 시각화

그런데 각각의 obs를 전부 사용하지 않고 BAID 별로 한 개씩 랜덤샘플링을 진행하였는데 그 이유는 다음과 같다. 랜덤 샘플링 없이 모든 데이터를 사용할 경우 똑같은 공통변수를 가진 데이터가 보통 3~4개 정도 존재했다. 이 상태로 모델링을 해보니 아니나 다를까 공통변수의 영향력이 굉장히 극대화되었고 이로 인해 올바른 예측 역시 되지 않았다. 따라서 이 문제를 해결하기 위해 BAID 별로 한 개의 관측치를 추출할 필요가 있었다.

fuPDCD가 1인 경우 즉, 관상동맥 질환에 전혀 걸리지 않은 경우는 앞서 말했듯이 시점마다 obs를 쪼갠 뒤 랜덤샘플링을 하여 변수를 추출했는데, fuPDCD가 2인 경우는 그 방법이 조금 다르다. fuPDCD가 2라는 것은 6시점의 PDCD_HIST가 [1 1 1 2 2 2]처럼 1에서 2로 변하는 것을 의미하는데 여기서 앞의 1 1 1 중 랜덤 추출하는 것보단 1에서 2로 바뀌기 직전의 1을 사용하는 것이 질병의 원인 예측에 효과적일 것이라 판단을 하였기에 가장 마지막에 나오는 1을 사용하기로 하였다.

2-2. Response Variable

fuPDCD는 과거 V1 ~ V6기간 중 단 한 번이라도 관상동맥에 걸렸다면 2로, 그렇지 않으면 1로 표시된다. 그러나 조사기간 동안의 term이 길고, 정확히 어느 시점에서 질병이 걸렸는지 알 수 없기 때문에 반응변수로 부적절하다. 반면 PDCD_HIST는 조사된 시점을 기준으로 과거 기수에 한 번이라도 걸리면 2, 그렇지 않으면 1로 구분된다. 따라서 PDCD_HIST를 이용하여 반응변수를 설정하였는데 구분 방법은 아래의 그림과 같다.

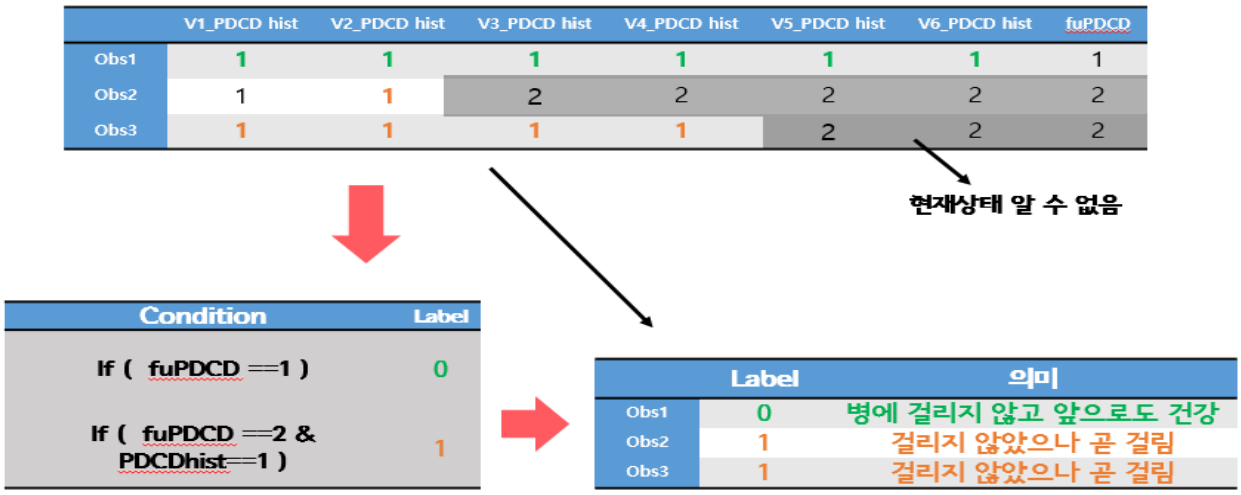

Figure 4 반응변수 설정

위의 그림에서 볼 수 있듯이 우리 모델은 PDCD_HIST가 2인 데이터는 사용하지 않았다. 그 이유에 대해 말하기 앞서 우리의 초기 분석 방향에 대해 말하자면 반응변수 label을 0,1,2 총 3개로 나누어 label0은 병에 걸리지 않고 앞으로도 걸리지 않는 사람, label1은 병에 걸리지 않았으나 미래에 걸릴 사람, label2은 현재 병에 걸린 사람으로 구분을 하는 multiclass classification을 진행함으로써 관상동맥 질환을 예측하고, 원인이 되는 요인까지 밝혀내는 것이 우리의 목적이었다. 계획대로 모델링을 진행하였고, 예측 결과 역시 만족스러운 결과가 나왔으며 LDL과 CHOL가 중요 변수로 나왔다. 그런데 문제는 이 두 변수의 플랏을 그려보니 알 수 있었다.

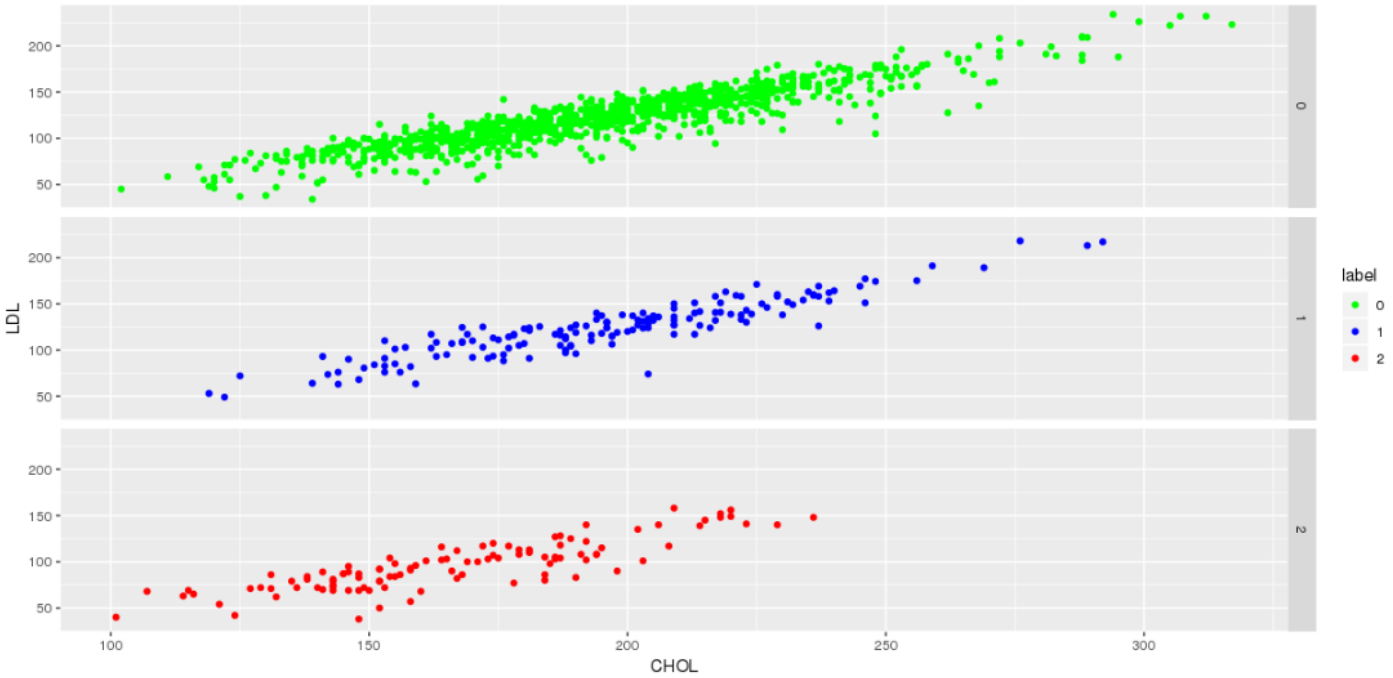

Figure 5 label에 따른 CHOL(x축),LDL(y축) 시각화

위의 그림은 label0, label1, label2에 대하여 x축은 CHOL을, y축은 LDL을 표시한 플랏이다. 플랏에서 볼 수 있듯이 label2의 경우 label0과 label1과 뚜렷한 차이를 보이는데 문제는 LDL과 CHOL이 현저하게 낮다는 점이다. 혹시나 데이터를 추출하는 과정에서 문제가 있었을 수도 있기에 완전 raw data로도 확인을 해봤으나 차이는 없었다. 관상동맥 질환이 동맥이 좁아짐에 따라 발생한다는 점과 LDL의 수치가 높을수록 동맥경화와 심질환의 위험도가 높다는 점을 미루어보았을 때 이는 CHOL과 LDL으로 인해 관상동맥 질환이 발생했다고 보기엔 무리가 있었다. 이 문제는 PDCD_HIST가 2인 지점을 label2로 설정하였는데, PDCD_HIST가 의미하는 것이 현재 관상동맥 질환에 걸렸음이 아닌 관상동맥 질환에 걸렸었다는 것을 의미하기 때문이라고 판단을 하였다. 즉 그 사람이 현재까지도 질환을 앓고 있는지 혹은 완치를 했는지에 대한 아무런 정보가 없었고, 그보단 관상동맥 질환을 치료하면서 LDL과 CHOL을 낮춘 것이라 판단하는 것이 합당했다.

또한, 이 데이터를 이용하여 예측을 하더라도 지도학습이기에 그것은 관상동맥 질환을 예측한 것이 아닌 과거에 관상동맥 질환을 앓았는지를 예측하는 것이기에 의미가 없다고 판단을 한 것이다. 따라서 우리는 label0(현재 질환에 걸리지 않았고, 앞으로도 걸리지 않을 사람)과 label1(현재 질환에 걸리지는 않았으나 후에 걸릴 사람)을 이용하여 binary classification을 진행하였고, 한 사람이 질병에 걸릴 확률과 그 요인을 파악하고자 하였다.

2-3. 변수 생성 및 통합

조사한 바에 따르면 V1_FVC 변수의 수치는 그 자체 변수만으로 건강함의 정도를 판단하는 것이 아니라 FEV1과 같이 고려함으로써 건강의 상태를 나타내기에 이 두 변수를 통합하여 fvc라는 새로운 변수를 생성하였다. 추가적으로 맥압을 나타내는 SBP(수축기혈압)과 DBP(이완기혈압)의 차이 역시 계산을 하여 새로운 변수로 추가하였다.

관상동맥 질환은 물론 그 외 여러 질환의 가족력을 나타내는 변수가 존재했는데 각각의 질환마다 형제자매가족력, 모가족력, 부가족력 이렇게 세 항목으로 나뉘어 있었다. 그런데 유전이라는 것이 무조건 발생하는 것이 아니며 잠재적인 유전으로 인해 그 질환에 걸리지 않더라도 질환의 잠재적인 요소를 갖고 있을 것이라 판단하였다. 실제로 형제자매, 부, 모의 가족력을 각각 나누어 모델링은 한 결과 비중이 상대적으로 약하여 모델링 과정에 전혀 고려되지 않아 이 세 변수를 하나로 통합하여 가족 중 질환을 앓은 사람이 있는 경우를 2, 없는 경우를 1로 변수를 통합하여 이용하였다.

	FMHTREL3	FMHTREL2	FMHTREL1
Obs1	2	1	2
Obs2	2	1	1
Obs3	1	1	1

	FMHTREL
Obs1	2
Obs2	2
Obs3	1

Figure 6 변수통합

3. Unbalanced Data Handling & Modeling

위의 전처리 과정을 끝나고 남/여 데이터를 구분하여 별도의 데이터로 만들었다. 이후 모델링은 크게 두 가지 방향이 존재하는데 처음 건강검진을 받는 사람의 관상동맥 질환을 예측하는 모델과 만일 그 사람이 이전의 건강검진 정보가 존재한다면 그것과 현재의 건강검진 정보 둘 다를 이용하여 좀 더 정확한 질환 예측이 가능 모델이다.

3-1. 단시점 모델링

우선 모델 검정을 위해 전체 Data의 20%를 Test set으로 분리하고, 모델링이 끝날 때까지 train set으로부터 test set으로 정보가 흘러 들어가지 않도록 했다. 80%의 train set을 활용하여 모델링을 시작하기 전 반응변수 label의 table을 다음과 같았다.

성별/반응변수	Label0	Label1
남자	3071	80
여자	4243	76

Figure 7 반응변수 비율

한 눈에 봐도 unbalanced data의 형태임을 알 수 있다. 이 상황에서 바로 분류 모델을 생성할 경우 트리가 생성될 때 수가 적인 label '1' 분류로의 Gain이 매우 작게 계산될 것이고, 따라서 모델이 거의 모든 test data를 label '0'으로 만 예측하게 될 것으로 예상되었다. 따라서 Unbalanced Data를 해결하기 위한 여러 방안을 모색했는데, 우선 비대칭의 정보가 매우 심하기 때문에 Oversampling은 기각하였다. 비율에 맞게 Label '1'를 복제할 경우 과적합이 발생할 것을 우려했기 때문이다. 마찬가지의 이유로 인공 데이터를 생산해서 oversampling 하는 기법인 Smote 역시 기각되었다. 실제로 Smote를 사용해본 결과 80개의 label이 3000, 4000개가 되어 버려서 예측결과가 망가져버렸다. Label '1' Class에 가중치(weight)를 주는 방식도 사용했으나 역시 비대칭의 정도가 너무 힘하여 가중치만으로는 결과가 크게 바뀌지 않았다.

결론적으로 우리는 Bootstrap UnderSampling과 모델 앙상블(Ensemble)기법을 사용했다. 그 과정은 다음과 같다. 우선 Train Set에서 Label이 '0'인 obs를 300개 랜덤 추출한다. label '0'과 label'1'의 비율이 적절히 감소된 상태에서 Xgboost 파라미터를 튜닝하고 xgb.cv를 통해 최적의 nround값을 뽑아낸다. Optimized 된 Xgboost 모델을 생성하여 최초에 분리해 두었던 Test set의 값들을 Predict한다. 단 이때 결과값은 SoftProb(확률값)으로 추출한다. 이 전체 과정을 10번 반복하는 것이다. 그러면 Undersampling을 10번하여 10개의 모델로 각각 Test data를 예측하게 된다. 그러면 예측 확률값 10개를 평균내어 최종 예측값을 계산한다. 예를 들어 Test data의 어떤 obs의 예측값이 (0.1, 0.1, 0.2, 0.2, 0.2, 0.2, 0.2, 0.3, 0.4, 0.8) 이라면 이 obs의 최종 예측값은 0.24가 된다. 이후 Cutoff를 통해 해당 obs의 label을 최종 결정하게된다. (Cutoff는 trian set의 label 비율인 0.3 수준으로 결정) 다음은 해당 과정을 간단히 표현한 그림이다. 10개의 모델을 차례대로 진행한다고 이해하면 보면 될 것이다.

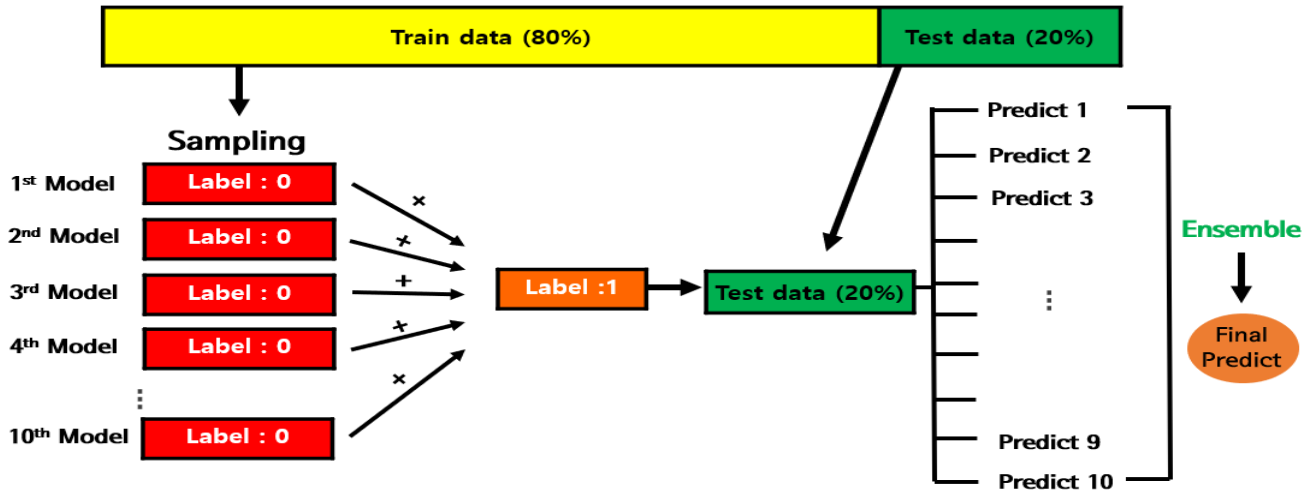

Figure 8 unbalanced data 처리

Xgboost 모델의 특성상 Factor 변수를 더미화 시켜서 Numeric 꼴로 만들어 주어야 하기 때문에 'dummies '패키지를 활용하여 더미화 시킨 후 위의 모든 모델링 과정을 진행했다.

모델링을 진행하고, 모델을 평가하는 과정에서 주 Metric으로 F2-measure를 이용하여 모델을 평가하였다. F2-measure란 일반적으로 이용하는 F1-measure보다 recall의 비중을 높인 metric이다. 그 이유는 질병 예측의 특성상 위험군(label : 1)을 정상(label : 0)으로 분류하는 오류를 줄이는 것이 그 반대의 경우보다 더욱 중요하다고 판단했기 때문이다. 즉 Recall이 Precision보다 더욱 비중있게 고려되어야 한다는 것이다. F2-measure의 식은 아래 그림과 같으며, 보다시피 Precision 한 단위 증가에 따른 F2-measure의 증가량보다 Recall의 한 단위 증가에 따른 F2-measure의 증가량이 훨씬 큰 것을 알 수 있다. 다시 말해 Recall에 더 sensitive하게 지표가 변하는 것이다.

$$F_2 = 5 \cdot \frac{precision \cdot recall}{4 \cdot precision + recall}$$

10개의 모델을 앙상블 한 결과 계산된 평가지표(metric)값은 다음과 같다.

남자(Male)

```
> re<-CM$byClass[6]
> pr<-CM$byClass[5]
> f2<-(5*pr*re)/(4*pr+re)
> names(f2)<-'F2-Measure'
> f2
F2-Measure
 0.6563136
```

```
   Recall
0.60477002

       F1
0.75251722
```

Test data	
실제 Label 1	모델예측 Label 1
16	10

여자(Female)

```
Recall      > re<-f_CM$byClass[6]
0.697350993 > pr<-f_CM$byClass[5]
            > f2<-(5*pr*re)/(4*pr+re)
F1          > names(f2)<-'F2-Measure'
0.820572764 > f2
            F2-Measure
              0.741915
```

Test data	
실제 Label 1	모델예측 Label 1
16	11

3-2. 단 시점 모델 Feature Importance

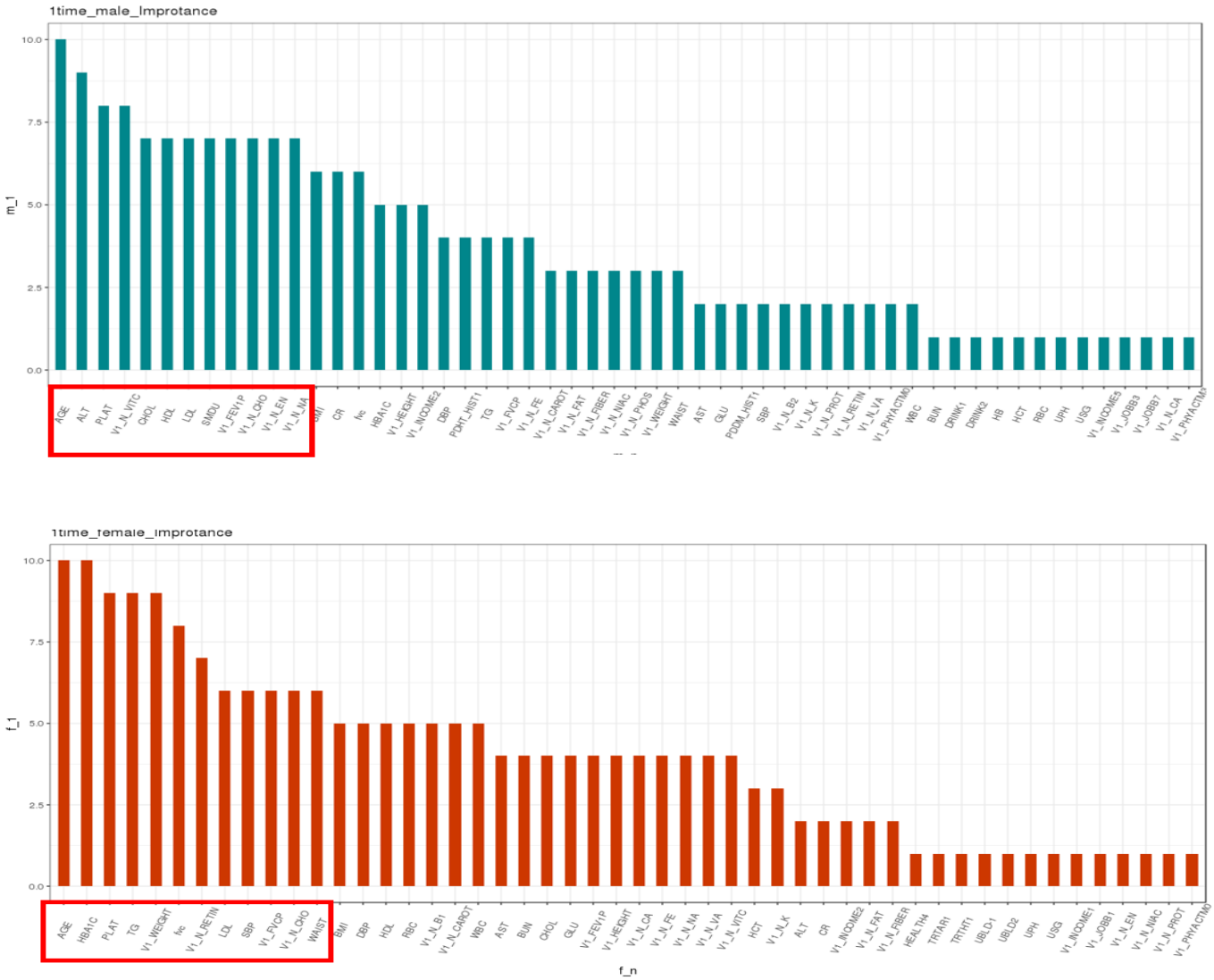

Figure 9 단 시점 주요 feature

10개의 모델을 생성하는 과정에서 각 모델마다 Feature Importance를 확인할 수 있었다. Feature Importance를 통해 Xgboost 모델이 트리를 생성하는 과정에서 더 많이 사용한 변수가 무엇인지 파악할 수 있다. 노드가 나뉠 때 더 Gain이 높은 쪽으로 obs들을 나누게 되므로 Feature Importance는 Label 0 과 Label 1을 구분

짓는 중요한 구분자를 파악하는 지표로 활용할 수 있다. 즉 잠재적 위험이 있는 label 1을 분별하는데 비중있게 고려되는 변수를 확인할 수 있는 것이다. 앞 챕터에서 설명한 방식대로 Undersampling 10번을 시행하여 만든 10개의 모델에서 각각 Feature Importance Matrix를 추출하였고 각 모델에서 Importance가 상위 20위('상위권')안에 드는 변수들을 모았다. 이후 각 모델에서 상위권에 몇 번 들었는지 카운팅 해주었는데, 예를들어 10개의 모델에서 모두 상위권에 들었다며 10, 7번 들었다면 7이라고 표시된다. 10번 bootstrap Undersampling하여 앙상블한 결과이기 때문에 수가 많았던 label 0 집단의 특성을 효과적으로 대표하여 label 1집단과 비교할 수 있었고, 따라서 10번의 모델에서 모두, 혹은 7번 이상 주요 변수로 고려된 변수들은 label 0 과 1의 실질적인 구분자라고 판단할 수 있다. 이를 막대그래프로 시각화한 것이 위의 그림이다.

남자와 여자의 단 시점 모델에서 모두 높은 Importance를 보인 변수들은 AGE를 비롯하여 PLAT, LDL, V1_N_CHO 등이 높은 Importance를 보였다. 특히 LDL이 높은 것은 기존 의학분야에서도 관상동맥질환의 중요 원인 중 하나로 진단되고 있기 때문에 '아직 질환에 걸리지는 않았으나 미래 걸릴 위험군'을 판별하는데도 비중 있는 변수로 나타났다. 또한 CHOL(콜레스테롤)과 PLAT(혈소판 수)역시 매우 중요하게 관측되었는데, 관상동맥 벽에 콜레스테롤과 혈소판 등이 쌓여 발생할 수 있는 관상동맥 경화증과도 밀접한 연관이 있을 것이라 예상했다. 혈소판과 콜레스테롤 수치가 위험군(label 1)을 구분해주는 주요 변수로 사용된다는 것은 해당 위험군이 질환에 걸리기 전이 미리 파악하고 치료를 가능케 할 것이다. HBA1C(당화혈색소)와 TG(혈중 지질)과 같이 직접적인 변수들도 예상대로 주요변수로 고려되었다.

남자 feature에서만 발견된 중요 변수들 중에서는 SMDU(흡연기간)도 눈에 띈다. 리차드 W.제임스(Richard W. James) 박사와 연구팀이 발표한 자료에 따르면 관상동맥 질환에 대해 보호 역할을 하는 파라옥소나제(paraoxonase, PON) 라는 단백질이 흡연자의 혈액 중에 적게 존재한다는 사실이 밝혀졌다. 이를 감안할 때 남자 위험군을 판별하는 주요 변수로 흡연기간이 고려되었다는 것은 충분히 주목할 만하다. 흡연을 많이 할수록 무조건 관상동맥에 걸린다고 하는 것은 어불성설이지만 분명 위험군으로 분류될 가능성이 높아지는건 사실이고 따라서 관리의 필요성을 제시할 근거가 될 수 있다.

3-3. 다시점 모델링 (Multi Periods Modeling)

다음으로 두 시점의 데이터를 이용한 모델링 과정이다. 건강검진을 처음 받는 대상자를 위해 단시점 모델을 만들었으나 이는 코호트 데이터를 중요한 점을 모두 이용하지 못하는 것이라 판단을 했는데 그것은 바로 각종 변수 수치의 변화이다. 즉 우리는 여러 시점을 비교함으로써 각각의 사람의 건강검진 수치의 변화를 파악할 수 있고 이를 고려한 추가적인 모델링의 필요성을 느꼈다. 전체적인 모델링 과정은 위의 과정과 비슷하다. 한 가지 달라지는 점이 있다면 변화를 나타내는 변수를 추가할 수 있다는 것이고, 다음과 같은 과정으로 변수를 생성하였다.

	V1_CHOL	V2_CHOL	V2time (month)	V3_CHOL	V3time (month)
Obs1	100	120	25	140	49
Obs2	100	NA	NA	140	50

	CHOL	CHOL_before	Age_change (month)
Obs1	120	100	25
Obs1	140	120	49-25 = 24
Obs2	140	100	50

	CHOLchangerate	CHOLchange_time
Obs1	120/100	(120-100)/25
Obs2	140/120	(140-120)/24
Obs3	140/100	(140-100)/50

위와 같은 방식으로 모든 numeric 변수의 변화량과 변화율을 뽑아낼 수 있었다. 그런데 그 변화가 일어난 기간이 obs마다 상이하기 때문에 변화량을 절대적인 수치로 사용하기에는 무리가 있었고, 따라서 이를 추가적으로 기간으로 나눈 뒤 변수로 사용하였다.

10개의 모델을 앙상블 한 결과 계산된 평가지표(metric)값은 다음과 같다.

남자(Male)

```
> View(two_male_imp)
> re<-m2_CM$byClass[6]
> pr<-m2_CM$byClass[5]
> f2<-(5*pr*re)/(4*pr+re)
> names(f2)<-'F2-Measure'
> f2
F2-Measure
 0.7451335
```

```
   Recall
0.70102215

       F1
0.82279430
```

Test data	
실제 Label 1	모델예측 Label 1
16	10

여자(Female)

```
> re<-fe2_CM$byClass[6]
> pr<-fe2_CM$byClass[5]
> f2<-(5*pr*re)/(4*pr+re)
> names(f2)<-'F2-Measure'
> f2
F2-Measure
 0.7617078
```

```
   Recall
0.70102215

       F1
0.82279430
```

Test data	
실제 Label 1	모델예측 Label 1
16	9

3-4. 다 시점 Feature Importance

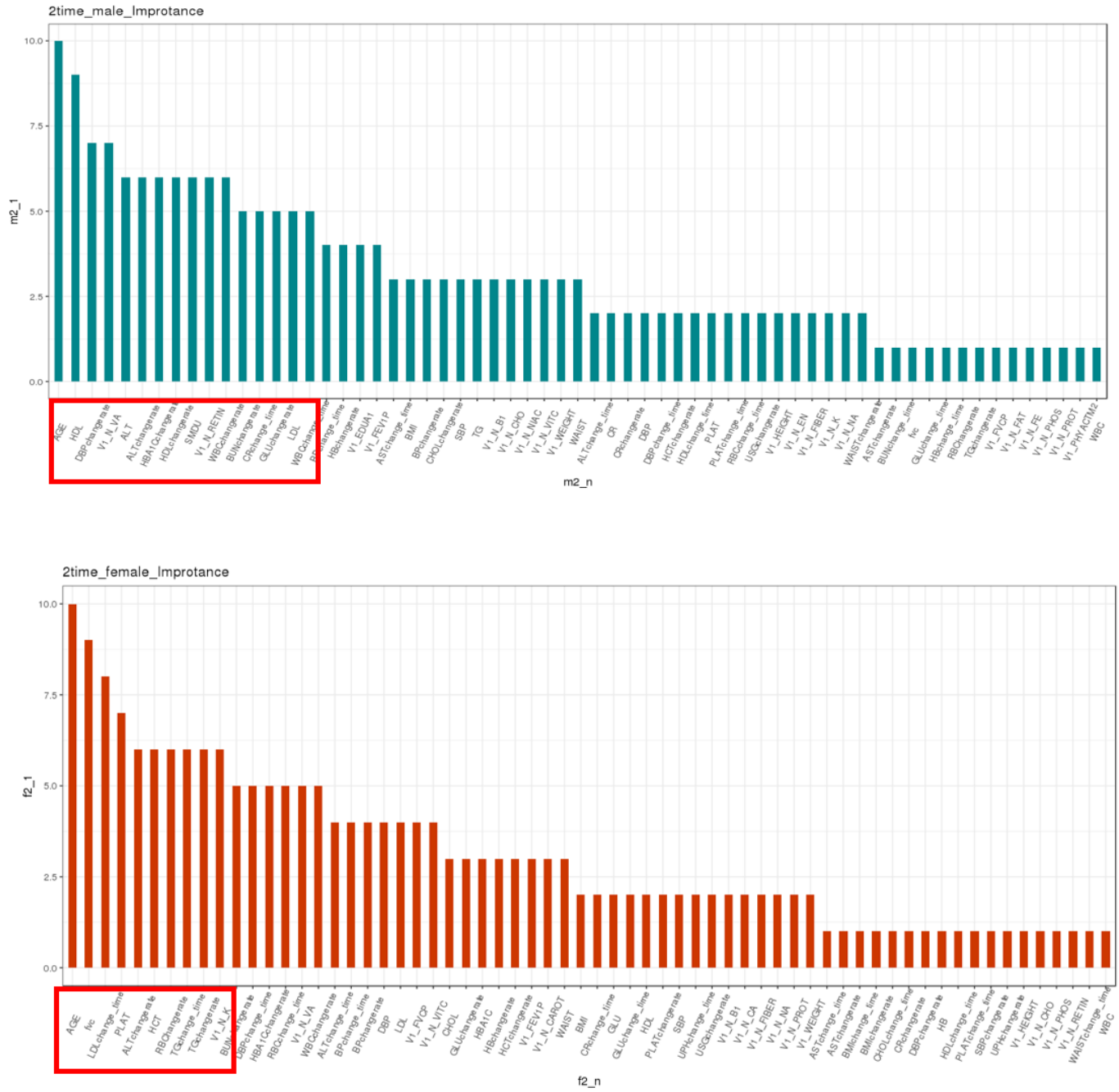

Figure 10 다 시점 주요 feature

단 시점 모델과 마찬가지로 다 시점 모델에서도 Feature Importance를 확인할 수 있었다. 계산방식과 그래프에 대한 설명은 단 시점 모델 Feature Importance 부분에서 자세히 서술하였으므로 생략하고 단 시점 Importance와 비교하여 새롭게 주요 변수로 고려된 변수들 위주로 해석을 해볼 수 있겠다.

단 시점 모델과 마찬가지로 PLAT, ALT, AGE, LDL 등은 주요 변수로 고려되었다. 또한 해당 변수들의 변화율 (Chane time, change ratio)등도 역시 높은 Importance를 보였다. 혈소판 수나 저밀도 콜레스트롤의 수치가 이전 시점 조사의 수치보다 크게 변화한 사람들은 위험군으로 분류될 가능성이 높아지는 것이다. 단 시점에서 높았

던 변수들의 변화율이 큰 Importance를 가지는 것은 위험군을 분류하는 주요 변수 수치의 변화를 모델이 비중있게 감안한다는 것이고, 따라서 실제 검사에서 수치가 크게 바뀌는 사람들을 위험군으로 의심하여 관리할 수 있다. 다 시점모델에서만 확인할 수 있었던 추가 주요 변수는 DBP Change Ratio, SBP Change Ratio 와 같은 혈압관련 수치의 변화가 있었다. 이는 간단한 신체검사 중 하나인 혈압 검사를 통해서도 관상동맥에 대한 간접적 진단을 할 수 있고, 그 수치의 차이가 크다면 충분히 위험군에 속할 확률이 높아진다는 것을 의미한다.

4. 의의 및 한계

모델링을 크게 두 방향으로 진행을 하였는데 하나는 한 시점을 이용한 것이고, 다른 하나는 두 시점을 이용한 것이다. 그런데 만약 두 시점을 이용한 모델이 한 시점을 이용한 모델보다 성능이 떨어지면 이를 사용할 이유가 없어지게 된다. 하지만, 우리의 경우 두 시점을 이용함과 더불어 그로부터 추출할 수 있는 변수를 추가해 주었고 이 변수들로 인해 F2-measure를 metric으로 했을 경우의 모델 성능이 향상됨을 알 수 있다. 비록 label1의 분류엔 큰 영향이 없고 오히려 소폭 하락하기도 하였지만 이를 극복할 만큼 label0을 구분하는데 큰 성능을 보였다.

또한, 데이터 전처리를 마친 후 label0 과 label1의 비율은 대략 98% 대 2% 일 정도로 상당히 unbalanced data였다. 일반적인 undersampling을 통해 label0을 label1의 수만큼 뽑아 모델링을 할 경우, 데이터의 손실이 상당하고, 추출한 데이터들이 전체 표본을 대표할 수 없는 문제가 생긴다. 반대로 label1을 oversampling할 경우, 과적합의 문제가 발생할 수 있다. 따라서 이 두 문제를 극복하기 위해 test set은 고정시킨 채 train test를 남녀 각각 300개씩 10번 추출함으로써 모든 데이터를 이용하고자 하였다. 또한, 10개의 모델에서 확률 값을 추출한 뒤 이를 평균 낸 값으로 분류를 진행하였는데 이 과정을 전 후로 모델이 개선되는 것을 확인할 수 있었다. 즉, 데이터의 손실은 최소화하며 모델의 성능을 개선시켰다는 점에서 의의를 지닌다고 볼 수 있다.

BIOAGE

건강지수의 정의와 이를 이용한 생체나이 측정

한대룡, 백우현

1. 배경 및 방향 제시

우리가 조사한 바에 따르면 현재 생체나이는 각종 신체검사를 통해 측정한 결과를 바탕으로 해당 나이와 비슷한 연령대 사람들의 평균값을 활용하여 그보다 좋으면 생체나이가 젊게 계산되고, 좋지 못하면 더 나이가 든 것으로 계산을 한다. 그러나 우리의 데이터에서도 볼 수 있듯이 연령 별 건강검진 데이터가 균등하지 않으며, 18세와 80세 주변의 관측치는 다른 연령보다 현저히 적음을 알 수 있다. 이 말은 즉, 이와 같은 방법으로 생체나이를 계산하게 되면 이 연령대의 사람은 다른 연령보다 부정확한 값을 얻을 가능성이 크다는 것을 의미한다. 또한, 각 나이대별로 다르게 계산이 되기에 절대적인 건강함의 정도를 나타내기 힘들다는 한계를 지닌다.

따라서 우리는 정상수치와의 거리를 기준으로 좀 더 정확하고 일관적인 방법으로 생체나이를 계산하고자 했으며, 동시에 기하적/시각적으로도 쉽게 표현할 수 있는 방법을 고안하였다. 또한 계층 군집분석을 통해 모든 변수를 같은 비중으로 상정하고 거리를 계산했을 때 보다 좀 더 합리적으로 접근할 수 있었으며 이를 바탕으로 그 사람의 전반적인 건강상태를 나타내는 건강 수치를 생성하였고, 이를 이용하여 최종적으로 생체나이를 계산하였다.

이 분석을 진행하면서 크게 두 가지 논점이 존재하였는데, 이는 다음과 같으며 이 논점에 대한 우리의 생각을 제시하면서 분석을 진행해보려 한다.

1. 혈액, 폐기능, 면역, 종양, 요 검사 등의 항목을 각각 분류하여 고려해야 할까?

2. 20세와 80세 두 명의 사람이 만일 모든 건강검진 항목의 수치가 동일할 경우 이 두 사람의 생체나이는 같은 것일까?

2. 데이터 전처리

2-1. 결측치 처리

생체나이 예측 분석방향은 완전정상 벡터로부터의 거리에 기초하였기에 우리의 데이터에는 결측치를 대체하거나 삭제할 필요가 있었다. 결측치 처리 방향은 크게 두 단계로 진행됐는데, 다음과 같다. 처음으로 결측치가 전체 obs 3만 개 중 1만 개 이상(30%)인 변수를 삭제해주었다. 해당 변수와 각 변수 별 결측값의 수는 아래 그림과 같은데 이와 같이 너무 많은 결측값을 가진 변수를 대체하기에는 예측의 정확성 감소는 물론 오히려 부정확한 결론이 나올 것이 우려되어 삭제를 하였다. 그런데 여기서 PSA와 CA125 항목 역시 결측치가 10,000개 이상이었으나 조사해보니 PSA(전립선 특이항원)는 남자에게만, CA125(난소암)는 여자에게만 있는 항목임을 알게 되었고, 이 변수는 삭제를 하지 않고 남자와 여자의 경우를 분리하여 생체나이를 분석하기로 하였다.

WAIST	HIP	CPK	DIRBIL	HBA1C	IRON	TIBC	UIBC	AMYLASE	WBC
22213	28788	25989	20398	17063	26807	26841	26849	25177	14993
ESR	PDW	T3	CRP						
11330	28739	21824	20880						

Figure 1 삭제 변수

위 변수들을 삭제한 이후 나머지 변수들 중 결측값이 수는 모두 7000개 이하였는데 해당 결측값들은 변수가 아닌 obs(관측치)를 삭제해 주었다. Mice같은 패키지를 이용하여 다중대체 혹은 평균대체를 할 수도 있었으나 대체가 아닌 삭제를 한 이유는 다음과 같다. 먼저 생체나이를 계산하는 새로운 방안을 고안하는 것이기에 우리가 구한 생체나이가 올바른 결과라는 것을 평가할 마땅한 척도가 존재하지 않았다. 따라서 대체값을 이용하기보단 최대한 데이터의 그 특성을 살리는 것이 맞다고 판단을 하였다.

그리고 무엇보다도 결측값이 있는 관측치를 삭제한 결과 남녀 각각 대략 8000 ~ 9000개의 관측치가 남았는데 이것만으로도 이 분석을 진행하기에 충분한 양의 데이터라고 판단을 하였다. 앞서 말했듯이 우리의 분석 방식은 각각의 연령을 구분하는 것이 아니기에 모든 데이터를 한 번에 사용할 수 있었기에 이 정도 데이터의 크기면 충분하다고 생각을 했다. 아래의 그림은 남/여 데이터의 결측값 전처리 과정후의 히스토그램을 비교해 봄으로써 전체적인 분포가 크게 변하지 않았음을 시각적으로 확인을 하였다.

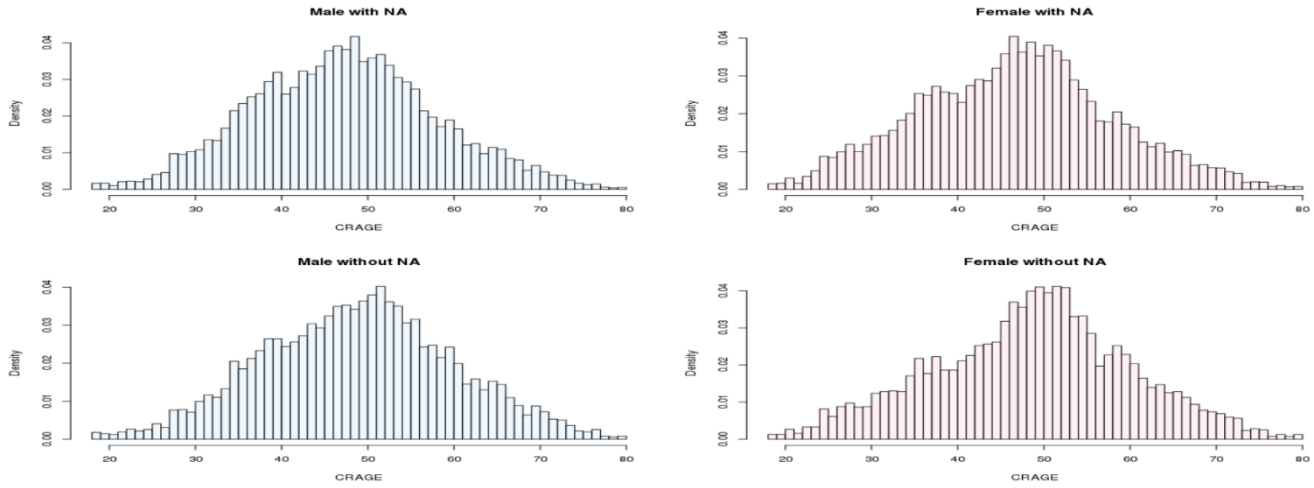

Figure 2 결측치 처리 전후 히스토그램(좌: 남자, 우: 여자)

2-2. 이상치(Outlier) 제거

일반적인 이상치 처리 방법으론 사분위수를 이용한 여러 방법이 존재하는데, 의료데이터의 특성상 일반적인 이상치 처리를 적용하면 안 된다고 판단을 하였다. 그 이유는 건강검진 데이터의 경우 정상의 경우가 대다수이고, 비정상의 값은 굉장히 극소수인 상당히 비균형적인 특성을 지닌다. 관상동맥 데이터 역시 마찬가지로 질병에 걸린 환자의 수가 3%미만이었는데 단순히 값이 작고 크다라는 기준으로 이상치를 제거할 경우 오히려 우리에게 정말 중요한 데이터를 잃는 결과를 초래할 것이라고 판단하였다.

따라서 데이터 전처리 과정에서는 이상치 처리는 각각 변수의 Boxplot을 일일이 살펴본 후 아래의 그림처럼 BMI가 45이상(이는 키 150cm, 몸무게 100kg을 의미)와 같이 직관적으로 결측 오류라고 판단되는 관측치만 삭제를 해주었으며 남자의 경우 14개, 여자의 경우 17개의 관측치를 삭제를 하였다. 설령 이 값이 실제 값이라 하더라도 이 값으로 인해 우리가 만든 건강지수가 왼쪽으로 지나치게 쏠리는 현상이 발생하기에 이를 해결하기 위해서는 삭제를 하는 것이 맞다고 판단하였다. 대신 건강지수를 만든 뒤 그 분포를 확인 후 그 수치가 극단적으로 높은 값을 삭제하는 방식으로 이상치의 악영향을 방지할 수 있을 것이라 생각했다.

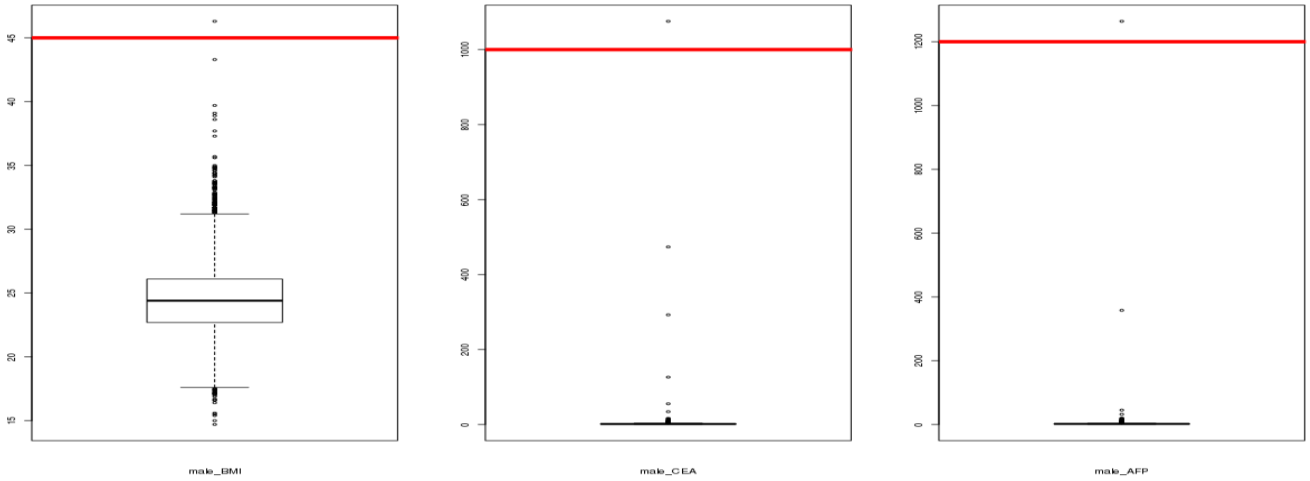

Figure 3 BMI, CEA, AFP Boxplot

3. 분석 과정

3.1. 건강지수: 정상범위와의 격차

총 78개의 건강검진 변수가 존재하였는데 이 변수를 원래의 상태로 사용하기에는 무리가 있었다. 그 이유는 예를 들어 BMI의 정상 범위가 20 ~ 25라고 가정할 경우 BMI가 19인 사람과 26인 사람 모두 정상이 아닌 사람으로 고려가 되어야 하기 때문이다. 따라서 우리는 각각의 변수를 그대로 사용한 것이 아니라 변수 설명에 있는 각 변수 별 정상수치 범위를 참고하여 정상수치와의 거리로 값을 수정해주었다. (참고로 정상수치가 표시 되어있지 않은 변수들은 일반적인 경우 측정되는 임상참고치를 정상수치로 대체하여 참가했다.) 이는 사람이 얼마나 건강한지, 다시 말해 얼마나 완전 정상수치로부터 떨어져있는지를 수치상으로 표현한 것이다. 즉, 0을 기준으로 값을 재배치해주었기 때문에 숫자가 클수록 건강이 안 좋은 것을 의미하며 이와 같은 과정을 진행함으로써 후에 선형회귀에 데이터를 적합할 수 있는 기반을 마련하였다.

추가적으로 조사한 바에 따르면 FVCPP는 그 자체 변수만으로 건강함의 정도를 판단하는 것이 아니라 FEV1_FVC의 변수와 같이 고려함으로써 건강의 상태를 나타내기에 이 두 변수를 통합하여 fvc라는 새로운 변수를 생성하였다.

각 obs별로 변수의 값이 정상수치 범위로부터 떨어져 있는 거리를 계산하여 삽입해준 후 변수 별로 scale이 상당히 다양하기 때문에 정상수치와의 거리를 일정하게 통일시켜줄 필요가 있었다. 따라서 계산된 거리를 각 변수별로 최대/최솟값 기준 0~1로 Scaling을 해주었다. (MinMax Scaling) 아래는 계산과정의 단순한 예시이다. (*'변수이름 [정상수치범위]'*로 표시) 해당 과정을 남성데이터와 여성데이터에 각각 적용시켜주었다.

기존데이터

BAID	SBP [0~120]	DBP[0~80]	BMI[18.5~25]	FEV1_FVC[0.7이상]
BA00102	124	83	21.1	0.683
BA00109	131	85	25.6	0.865
BA00110	133	90	24.0	0.784

정상범위와의 거리

BAID	SBP [0~120]	DBP[0~80]	BMI[18.5~25]	FEV1_FVC[0.7이상]
BA00102	4	3	0 (정상)	0.017
BA00109	11	5	0.6	0 (정상)
BA00110	13	10	0 (정상)	0 (정상)

변수별로 MinMax Scaling (0 ~ 1)

BAID	SBP [0~120]	DBP[0~80]	BMI[18.5~25]	FEV1_FVC[0.7이상]
BA00102	0.0444	0.0667	0 (정상)	0.0651
BA00109	0.1222	0.1111	0.024	0 (정상)
BA00110	0.1444	0.2222	0 (정상)	0 (정상)

Numeric 변수는 위와 같은 과정으로 정상으로부터의 거리를 쉽게 계산할 수 있었다. 문제는 HBSAG, RF와 같은 factor형 변수였는데 결론부터 말하자면 위와 같은 방식으로 처리를 하였다. 즉, 변수별로 정상인 level을 기준으로 정상 level 이면 0, 정상보다 한 단계 위면 1, 두 단위 위면 2로 데이터를 바꾸고, 이것 역시 MinMax Scaling을 진행하였다. 예를 들어, UBLD(요잠혈)은 Level이 Negative, Trace, Positive(1,2,3)의 5개로, 정상기준은 Negative(음성)이다. 따라서 Negative는 0, Trace는 1, Positive 3 (양성 3)은 4로 설정을 하였고 이를 Scaling 함으로써 0 ~ 1사이의 값으로 맞춰주었다. (0, 0.25, 0.5, 0.75, 1)

분석의 초기에는 factor 변수로 위와 같은 방식으로 처리하지 않고 factor 상태에서 "Gower" method를 사용하여 뒤의 군집분석에 사용될 거리를 구하려 했다. 그러나 곧 이 방식은 우리 데이터 적합하지 않다고 판단했는데, 우리 factor 변수의 경우 levels에 분명히 '순위'가 매겨진 형태가 존재했기 때문이다. 다시 말해, 우리는 factor 변수 내 각 level들 간의 거리를 동일하게 보는 것이 아니라 정상수치를 기준으로 더욱 부정적인 level은 거리가 더 멀어져야 하는 것이다. 위의 UBLD(요잠혈) 변수를 다시 보자면 Negative(음성) 정상으로부터 Positive 3

은 Positive 2보다, Positive 2는 Positive 1보다 더욱 좋지 않은 상태임을 알 수 있다. 따라서 이를 고려하기 위해선 factor 변수 역시 numerice 변수와 마찬가지로 정상과의 거리로 바꾼 후 정상수치(원점)과의 거리를 구할 필요가 있었다.

위의 모든 전처리 과정이 끝나면 각 obs는 전체 변수 개수만큼의 차원으로 벡터화 시킬 수 있다. 정상수치 범위 내에 있다면 해당 변수의 값은 0이므로 모든 변수에서 정상이라면 그 점은 완전정상(완전건강)인 포인트고 이는 기하적으로 표현하면 좌표의 원점에 해당할 것이다. 따라서 정상범위와의 거리로 값을 표시한 후에는 해당 obs가 원점으로부터 얼마나 떨어져 있는지, 즉 distance를 구하여 완전 정상수치와의 거리를 구하는 것이 가능해졌다.

. 아래는 지금까지의 과정과 Distance 아이디어를 보다 쉽게 이해할 수 있도록 시각화한 자료이다. X, Y, Z 축에는 각각 GLU, LDL, MCH가 들어가 있다. 각 값들은 위의 Reshape 과정을 거쳐 정상범위로부터 거리를 구한다음 0~1로 스케일링 되었다. 원점에는 모든 수치가 정상인 완전정상 Point가 있으며 원점으로부터 멀어질수록 색이 붉게 변하도록 설정했다. 하얀색 원점(0,0,0)으로부터의 거리가 멀수록 건강이 나빠지는 것이다. 시각화를 위한 샘플은 전체 데이터 중 300개를 랜덤 추출하여 사용했다.

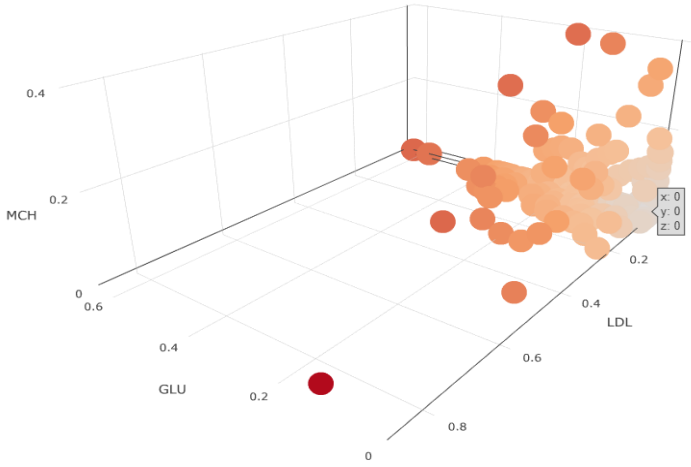

Figure 4 distance 시각화

3.2. 군집분석을 통한 그룹 생성

위의 과정에서 데이터 전처리를 모두 마무리하였고, 이제 그 데이터를 활용하는 일만 남았다. 즉, 각각의 obs가 원점으로부터 거리가 얼마나 떨어져 있는지를 계산하여 그 값을 더하면 그것이 바로 그 사람의 건강상태를 나타내는 지표이다. 그런데 한 가지 더 고려할 점이 존재했는데 그것은 앞서 말한 논점1과 관련된 부분이다. 각각의 변수를 독립적으로 보는 것은 인체의 메커니즘을 전혀 고려하지 못한 방법이라고 생각을 하였는데 우리가 신체 내부에 대한 모든 것을 알 순 없지만 상당히 밀접하고, 연관이 있는 하나의 유기체라는 것은 누구나 아는 사실이기 때문이다. 그렇다면 혈액 검사는 혈액 검사끼리, 폐기능은 폐기능 검사끼리 구분한 후 정상수치로부터 거리를 구하면 될까? 이에 대한 대답 역시 "No"이다. 앞서 말했듯이 인체 내부의 메커니즘을 우리가 다 알 수는 없고 실제로 혈액검사 A가 면역 검사 B와 상당 수 연관이 있을 수도 있다. 이를 고려하지 않고 정상수치를

구하게 되면 A와 B를 각각 독립적으로 구한 거리는 A와 B를 함께 묶어 구한 거리보다 클 수 밖에 없고 이는 생체나이가 그만큼 중첩되어 증가된다는 것을 의미한다.

A	B	Distance
c(0,0), c(1,1)	c(0,0), c(1,1)	$2\sqrt{2}$

A, B	Distance
c(0,0,0,0), c(1,1,1,1)	2

따라서 우리는 변수를 각각의 항목으로 볼 것이 아니라 연관이 있는 변수끼리 묶어서 볼 필요성을 느끼게 되었고, 계층 군집 분석을 통해 이를 구현하기로 하였다. 일반적인 클러스터링은 관측치끼리의 거리를 바탕으로 가까운 거리의 관측치는 같은 클러스터로 먼 거리의 관측치는 다른 클러스터로 구분을 해주는 기법이다. 그러나, 우리는 이 분석에서 관측치를 구분하는 것이 아니라 연관이 있는 변수를 구분하고 싶은 것이었기에 다음과 같은 방법을 이용하였다. 먼저 일반적인 클러스터링처럼 dist 함수의 Euclidean 메소드를 이용하여 관측치끼리의 거리를 구하였다. 우리가 앞선 전처리 과정에서 정상인 0을 기준으로 건강이 악화되면 숫자가 커지도록 전처리를 해주었기에 Euclidean 거리르 사용하여도 무방하다고 판단하였다. 그리고 다음의 과정이 중요한데 hclust의 경우 row 단위로 클러스터링을 진행하게 된다. 즉 dist를 그냥 적합시킬 경우 이는 변수간의 클러스터가 아닌 관측치간의 클러스터가 되는 것이다. 따라서 우리는 그냥 dist값이 아닌 이에 transpose를 취함으로써 row에는 변수가, column에는 각각의 관측치 값이 들어가도록 설정을 해주었고, 이를 통해 계층 군집을 나눔으로써 변수간의 클러스터를 설정할 수 있었다. 아래의 그림은 남자를 대상으로 계층 군집을 진행한 후 볼 수 있는 덴드로그램을 시각화한 것이다.

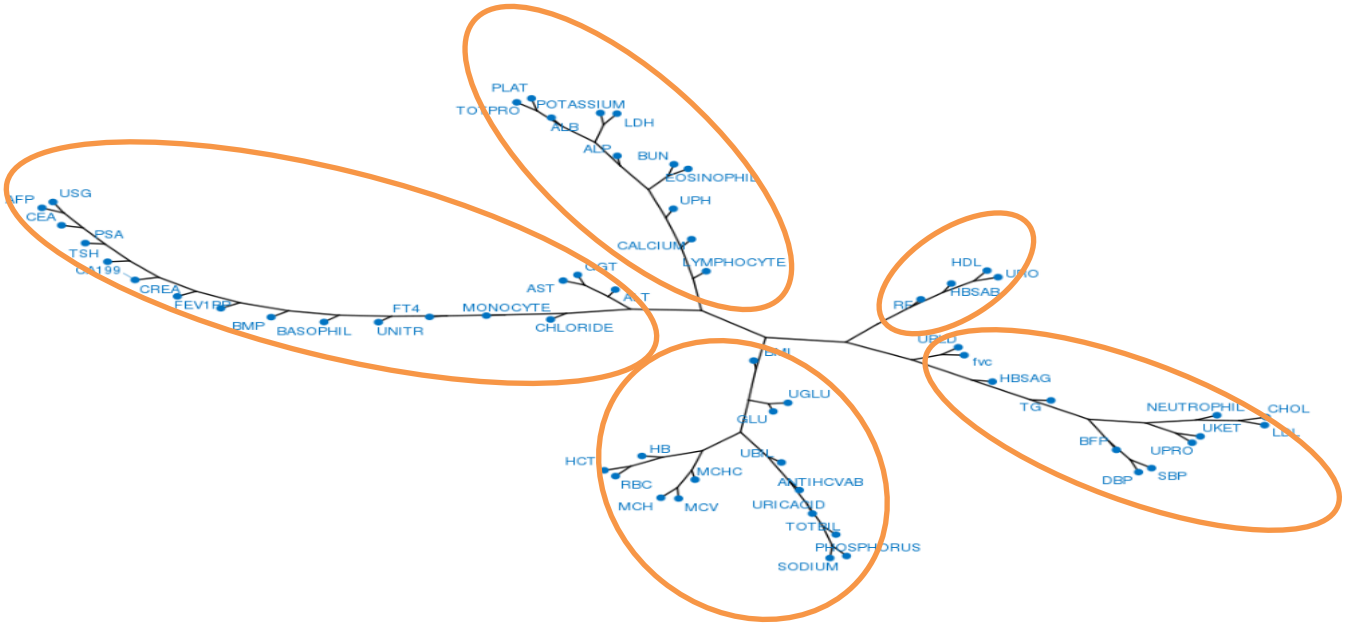

Figure 5 남자 계층 군집 덴드로그램

여기서 흥미로운 점은 우리가 혈액 검사, 종양 검사, 면역 검사 등의 변수들을 따로 구분하지 않아도 대체로 이 변수들이 모여 있다는 점이었다. 이는 대체로 검사 유형에 따라 검출되는 노화의 정도가 비슷함을 의미한다. 덴드로그램에 따라 그룹을 5개로 분류를 하였는데 그룹1과 혈액과 갑상선 기능, 그룹2는 혈액, 그룹3은 BMI, GLU(공복혈당), UGLU(요당)과 같은 체내 당 수치, 그룹 4은 요검사의 항목들이 대체로 모여있었다. 허나 우리는 이보다 더욱 나아 다른 검사일지라도 연관이 있는 변수들을 함께 묶어줌으로써 우리가 알지 못한 신체의 메커니즘을 조금이라도 반영하기 위해 노력하였다. 여자의 경우 역시 아래 그림과 같이 대체로 비슷한 양상을 띄웠으며, 마찬가지로 그룹을 5개로 나누어 분석을 진행하였다.

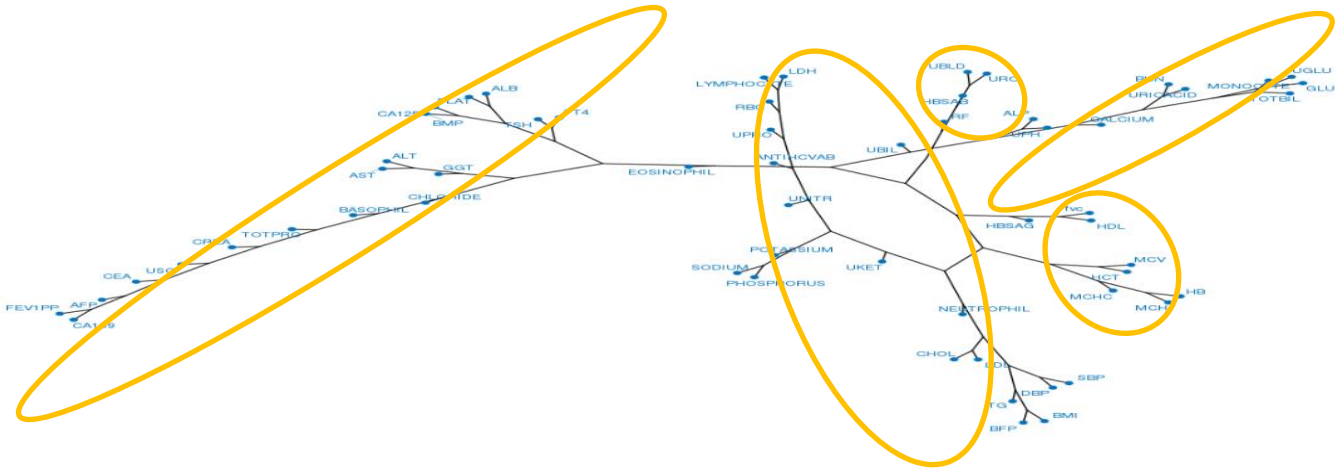

Figure 6 여자 계층 군집 덴드로그램

3.3. 건강지수 생성

이제 각각의 그룹으로 나누어 원점으로부터의 거리를 각각 구한 뒤 그 합을 계산하면 된다. 그리고 이것이 바로 그 사람의 전반적인 건강상태를 나타내는 건강지표인 것이며 그 값이 클수록 건강 상태가 좋지 않음을 의미한다. 우리의 데이터 상으론 모든 평가에서 0점, 즉 모든 변수가 정상 범위 내에 속한 사람은 없었으며 남자의 경우 0.244~6.473, 여자의 경우는 0.08~5.884의 분포를 따른다. 그러나 우리가 정상을 0을 기준으로 비정상의 거리를 구하였기에 그 분포가 상당히 왼쪽으로 쏠린 비균형적인 분포였다. 따라서 이를 일반화하기엔 무리가 있다고 판단을 하였고, 상위 1%의 값은 이상치라고 판단을 하여 이 값들을 삭제해주었고 이를 -10 ~ 10의 값으로 스케일링을 진행하였다.

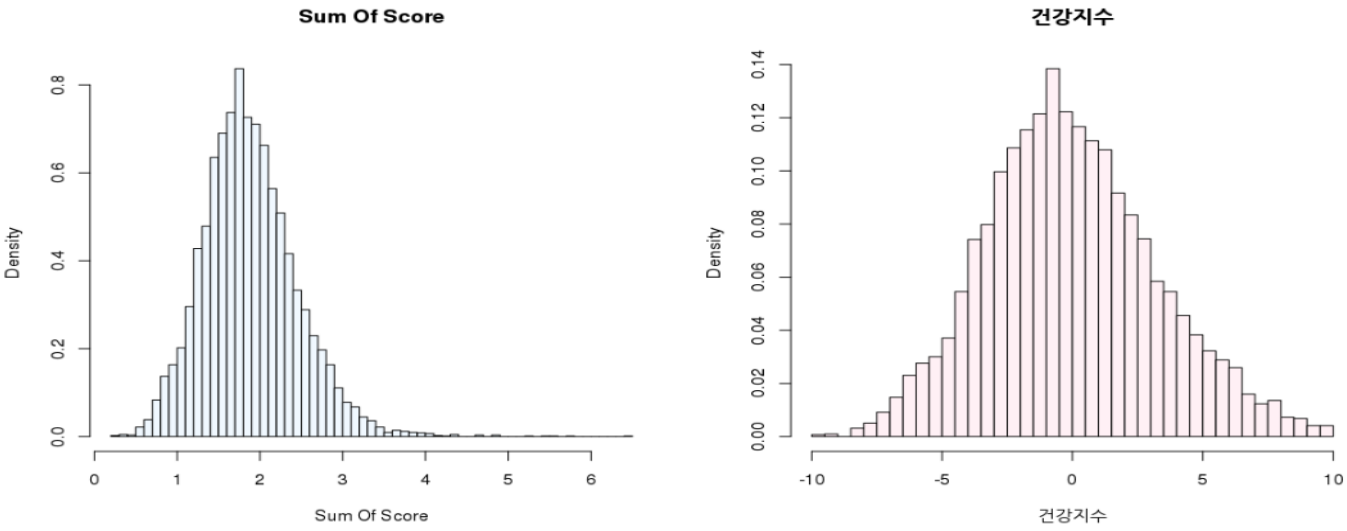

Figure 7 건강지수 스케일링(-10~10)

-10~10에 스케일링을 한 이유는 약간의 우리의 주관이 반영된 것이며 앞서 말한 두 번째 논점과 관련이 깊다. 현재 건강검진 데이터에는 대략 70개의 변수가 존재하지만 우리가 생각하기에 이 변수들이 인간의 노화의 일부분을 나타낼 뿐이지 모든 노화를 나타낼 순 없다고 판단을 하였다. 즉, 18세와 80세 두 명의 건강검진 수치가 모두 같더라도 이는 건강검진 내에서 파악된 변수 상에서만 수치가 같은 것뿐이지 이것만으로 두 사람의 생체나이를 같다고 단정지을 수 없다고 생각을 하였다. 따라서, 생체나이를 구할 때 절대적인 나이를 생성하는 것이 아니라 본인의 나이에서 건강이 좋으면 나이를 감소시키고, 건강이 좋지 않으면 나이를 증가시키는 방식으로 생체나이를 계산하는 것이 합리적이라고 판단을 하였고 우리가 생각하기에 생체 나이는 일반적으로 본인 나이의 +10에서 -10 사이라고 생각을 하여 스케일링을 진행한 것이다.

3.4. 선형회귀모델

이제 우리는 -10~10의 값을 갖는 건강지수라는 새로운 변수를 생성하였다. 그리고 이를 원래의 데이터에 추가를 하여 건강지수를 반응변수로, 그 외 건강검진 변수를 종속변수로 하는 선형회귀 모델을 만들 수 있다. 마지막에 건강 지수를 다시 회귀식에 적합한 이유는 다음과 같다.

1. 우리가 만든 생체 지수를 나이로 사용할 경우 생체나이에서 도출될 수 있는 값이 -10 ~ 10으로 한정되어 버린다. 이 역시 생체나이를 구하는 하나의 방안이 될 수는 있지만 실제로 본인의 나이보다 지나치게 건강하거나, 그 반대의 경우 생체 나이가 그 이상으로 차이가 날 수 있기 때문에 이는 문제가 있다고 판단을 하였다. 즉, 회귀모형을 적합함으로써 변수의 값이 지나치게 높거나 낮다면 -10 이하 혹은 10 이상의 값을 가질 수 있게 된다.

2. 대략 70개의 변수 중 인체의 노화와는 무관한 변수가 존재할 것이라 판단을 하였다. 즉 회귀 모델에 적합을 함으로써 그 안에 유의미한 변수의 영향력은 높이고, 무의미한 변수의 영향력은 낮게 해줌으로써 좀 더 명확한 생체나이의 기준을 제시하고자 하였다.

4. 결과 해석 및 시뮬레이션

Figure 8 estimate 상위 5개 변수

순위	변수	estimate
1	BMP	7.48
2	UNITR	7.22
3	FEV1PP	7.06
4	FT4	6.56
5	BASOPHIL	6.50

남자와 여자 각각 선형 회귀 모형에 적합을 시켰는데 처음에 적합 시 estimate 값이 NA가 뜨는 값이 있었다. 이는 다중공선성을 의미한다고 생각을 하여 선형 회귀의 단계적 방법(stepwise selection)을 변수를 제거해주었고 이렇게 하여 남자의 경우 58개, 여자의 경우 59개 변수를 사용한 선형회귀식 모델이 만들어졌다. 건강지표의 악화정도를 설명변수로, 건강지수를 반응변수로 설정을 하였기에 estimate 값이 큰 변수는 전반적인 건강상태와 관련이 있다고 해석이 가능한데 상위 5개를 추출한 결과는 옆의 그림과 같다. 반대로 estimate 값이 작은 변수는 전반적인 건강상태에 영향을 끼치기보단 특정 한 부위에 국한된 질병일 가능성이 높음을 의미한다.

그렇다면 지금까지의 알고리즘이 실제 상황에서는 어떻게 작동될 수 있고 사람들에게 어떻게 보여지는지 시뮬레이션 해볼 수 있겠다. 우선 해당 알고리즘이 작동되는 '생체나이 분석기계'를 가정하고 사람들이 검진받은 결과를 기계에 입력하는 것에서 시작된다. 다음 페이지의 그림에서 보이는 것처럼 A,B,C,D,E 다섯 명의 사람들이 자신들의 건강 검진결과를 기계에 입력했다고 생각해보자. 최초 데이터에 제공된 상태처럼, 다섯 명의 사람들은 건강검진 70여개의 변수들을 입력하게 된다. 물론 기계는 모델링될 때 제거된 변수들 (위의 결측치 처리 부분 참조)을 제거한 상태로 수치 Reshape 및 생체 Score 계산을 하게 된다.

BAID	CRAGE	SBP	DBP	BMP	BFP	BMI	FEV1PP
A	62	127	75	74.9	19.3	21.2	117
B	53	128	66	67.2	27.0	25.4	88
C	38	134	84	67.1	27.1	27.5	110
D	61	132	78	73.0	21.2	22.0	83
E	46	115	74	68.2	26.2	25.5	89

Data Reshape (정상수치로부터의 거리)

BAID	CRAGE	SBP	DBP	BMP	BFP	BMI	FEV1PP
A	62	0.08641975	0.00000000	0	0.00000000	0.00000000	0
B	53	0.09876543	0.00000000	0	0.19830028	0.02185792	0
C	38	0.17283951	0.05970149	0	0.20113314	0.13661202	0
D	61	0.14814815	0.00000000	0	0.03399433	0.00000000	0
E	46	0.00000000	0.00000000	0	0.17563739	0.02732240	0

군집별 계산 (원점으로부터의 거리)

```
      Group1    Group2    Group3    Group4    Group5
A 0.00000000 0.24038462 0.1087568 0.08641975 0.5270463
B 0.03870968 0.00000000 0.3811701 0.34766045 0.6167668
C 0.05585375 0.00000000 0.2682863 1.15060519 1.1180340
D 0.00000000 0.07692308 0.4135547 0.52259304 0.8112823
E 0.00678196 0.02711586 0.1596139 0.26617380 0.5121969
> mean(group1(test))> mean(group2(test)) > mean(group3(test)) > mean(group4(test))> mean(group5(test))
[1] 0.02090346     [1] 0.06699471     [1] 0.1534961     [1] 0.497129     [1] 1.125894
```

Bioage 산출 회귀식에 Fitting

```
        A         B         C         D         E
-4.863927 -1.776311  4.564618 -0.620937 -5.494829
```

NAME	A	B	C	D	E
AGE	62	53	38	61	46
Bioage	**57.136**	**51.224**	**42.565**	**60.379**	**40.505**

여기서 검사자들은 도출되는 자신의 생체나이를 알 수 있을 뿐 아니라 자신이 정상수준을 벗어나게 측정된 부분을 세부적으로 파악할 수 있다. 예를 들어 B와 C는 Group1에서 일반적인 평균치(0.02)보다 높은 수치를 보였으므로 ALT, GGT 등의 수치에서 이상을 보였을 수 있다. 따라서 간 검사를 추가적으로 받아보는 것이 추천된다. 또한 5명중 4명이 (B,C,D,E)가 Group 3에서 평균치(0.1535)를 초과했는데 이를 통해 BMI, GLU, UGLU 등에서 이상을 보였을 수 있다. 따라서 당뇨, 고혈당 검사 등을 추가로 받는 것이 추천된다. 이처럼 세부적인 건강도 파악을 통해 추가적인 검사가 추천되며 최종적으로는 자신들의 실제 나이를 감안한 생체나이를 확인하게 된다.

5. 의의 및 한계

생체나이의 범위를 -10에서 10 라고 정한 것은 우리의 어느 정도의 주관이 반영된 부분이라 볼 수 있다. 이 부분에 있어 고민을 많이 해봤으나 합리적인 기준을 구할 마땅한 방법은 존재하지 않았고, 우리의 주관대로 분석을 진행할 수 밖에 없었다. 그러나 앞서 말했듯이 이 값을 직접 사용한 것이 아닌 선형회귀의 반응변수로 설정하였다는 점에서 생체나이가 반드시 -10에서 10값을 가지는 것은 아니며, 특정 변수가 크게 증거하거나 혹은 감소한다면 생체 나이 역시 그에 맞게 10이상 혹은 10이하의 값을 가질 수 있기에 그런 한계점을 조금은 극복한 것이라 판단한다.

우리의 생체 나이 모델은 혈관, 면역, 갑상선과 같은 각각의 분야를 각각의 나이로 계산하기보단 신체의 메커니즘을 반영하여 노화의 정도가 중첩되지 않도록 설계를 하였다. 이로써 더욱 정확하고 체계적인 생체 나이를 계산할 수 있을 것이라 예상된다. 또한 단순히 생체나이가 얼마인지만 알려주는 것이 아니고, 각 건강 분야 중 어느 부분에서 내 건강지수가 나쁘고 추가검사 및 건강관리를 집중적으로 해야 하는지 구체적으로 제시한다. 따라서 보다 정확하게 자신의 건강함의 정도를 알 수 있다. 추가적으로 연령을 구분하지 않았다는 것 역시 큰 장점으로 작용하는데 앞서 말했듯이 우리의 데이터 역시 마찬가지로 18세와 80세 주변 연령대 사람의 표본은 굉장히 적은 것을 알 수 있다. 이 말은 이 연령을 구분하여 생체 나이를 구할 경우 이 연령대는 표본의 수가 적기 때문에 과소적합의 문제가 발생할 수 있고 이는 곧 옳지 않은 생체 나이 계산으로 이어질 수 있다. 반면 우리의 모델은 각 변수의 정상 수준에서의 거리를 사용하였기에 연령이 다를지라도 각각의 정상 기준으로부터 거리를 각각 구해주면 되며, 심지어 PSA와 CA125와 같이 남성에만 있는 변수 혹은 여성에게만 있는 변수를 이용하지 않는다면 성별이 다르더라도 모두 통합하여 사용이 가능하다. 즉, 같은 관측치라도 이전보다 우리는 더 많은 관측치를 이용하는 효과를 얻을 수 있는 것이며 기본정보와 직결된 건강 검진 데이터의 특성 상 데이터를 생성하는 것부터 유지 보수하는 것이 상당히 까다로운데 우리의 생체 나이 계산 방안은 이런 어려움에 큰 도움이 될 것이라 생각한다.