출석수업 과제물(평가결과물) 표지(온라인제출용)

**교과목명 : 데이터마이닝**

**학 번 : 202135-368864**

**성 명 : 홍 원 표**

**강 의 실 : 서울 (2023년 5월 14일 Zoom)**

**연 락 처 : 010-5343-4341**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1.

모수적 모형 접근방법(Parametric modeling approach)은 통계학에서 전통적으로 사용되던 방법으로 과거의 데이터를 해석해서 선형 회귀 분석 과 같은 모형식을 세우고 모형식의 모수 와 를 과거의 데이터를 활용하여 최소제곱추정법 (Least Square Estimation, LSE)이나 최대우도(가능도)추정법(maxmum likelihood estimation method) 등에 의해 모형식에 적합(fitting)하는 방법이다. 모수적 모형 접근방법의 장점으로는 결과가 복잡하지 않아 결과의 해석이 대체로 쉽지만, 단점으로는 과거의 데이터에 의존하기 때문에 데이터가 부합하지 않는 경우 예측정확도 등의 성능이 낮을 수 있다. 대표적인 모수적 모형 접근방법으로는 선형회귀분석, 로지스틱회귀모형이 있다.

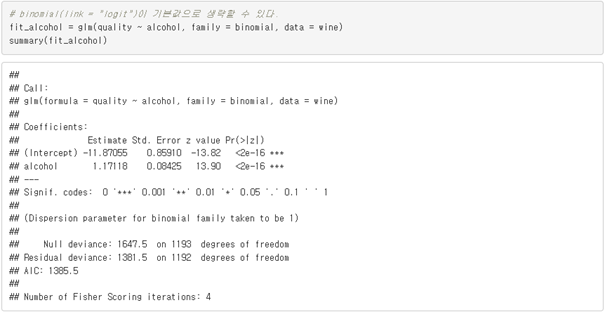
알고리즘 접근방법(algorithmic approach)은 기계학습의 발전으로 강조되고 있는 접근방법으로 정해진 알고리즘에 의해 데이터로 학습(training or learning)하여 얻은 결과를 따라 분석하는 방법이다. 장점으로는 복잡한 데이터도 적용이 가능 하고, 단점으로는 기계만이 인지할 수 있는 결과로 인해 해석이 어렵기 때문에 알고리즘의 속성이나 방법에 대한 이론적 근거를 파악하기 힘들어 모델을 지나치게 복잡하게 학습하게 되면 모델 성능은 높아지지만 새로운 데이터가 주어지면 정확한 예측/분류를 수행하지 못하게 되면 과도적합(over-fitting)이 발생하게 된다. 대표적인 알고리즘 접근방법 으로는 의사결정나무, 신경망모형, 배깅(bagging), 부스팅(boosting), 랜덤포레스트(random forest)가 있다.

2021년 말부터 ‘단계적 생활 속 거리두기’로의 전환을 시작으로 사람들은 여행을 하는데 제약이 크게 생겼습니다. 사적모임 인원 제한 및 명절에 가족들 모임까지 제한되고 실내마스크 착용 의무와 해외입국자 격리의무, 입국자 음성확인서 제출 등 많은 의무 사항도 있었습니다. 2023년 6월 사실 상 코로나19 엔데믹을 선한하고 일상회복을 하게 될 것 같습니다. 사실상 작년부터 코로나19바이러스의 치명률이 낮아 지면서 여행에 제약이 많이 풀려 많은 사람들이 여행을 시작하고 일본에 우리나라의 많은 관광객이 방문한 내용이 얼마전 뉴스에도 나왔습니다. 사람들은 여행을 하면서 그 지역의 맛집을 방문하게 되는데 대부분의 맛집들은 인플루언서의 SNS에 예쁘게 세팅하여 찍은 사진과 함께 되어 있는 게시물을 많이 볼 수 있습니다. 이런 SNS의 게시물들을 기반으로 지역별 맛집을 분석하면 여행에서의 식도락에 도움이 되지 않을까 생각됩니다.

SNS의 테스트 데이터를 수집하기 위해서는 SNS 플랫폼에서 게시된 데이터를 수집해야 하는데 이는 해당 플랫폼의 API나 웹-스크래핑 기술을 활용하여 수집할 수 있습니다. 수집한 데이터는 분석을 위해서 이모티콘이나 특수문자 등을 제거하고 같은 어근 동일화 및 N-gram(유니그램, 바이그램, 트라이그램) 적용 및 품사 분석을 합니다. 수치데이터를 분석하기 위해 요약정보를 살펴보는 것과 같이 단어 빈도 분석으로 의미있는 단어들의 출현횟수로 핵심있는 단어들을 분류할 수 있는지 확인합니다. 단어의 군집분서고 해보고, 토픽모델링을 하고 감성분석 및 연관단어 분석도 진행해 보는것도 좋습니다. 동시출현기반, 통계적 가중치 기반, wordwvec, 중심성계수, 연관어 네트워크 시각화 등을 통해서 분석을 진행하여 지역별, 상호별, 호감도 별등으로 구분을 하여 데이터를 정리할 수 있을거라고 생각됩니다.

2. 

2-①



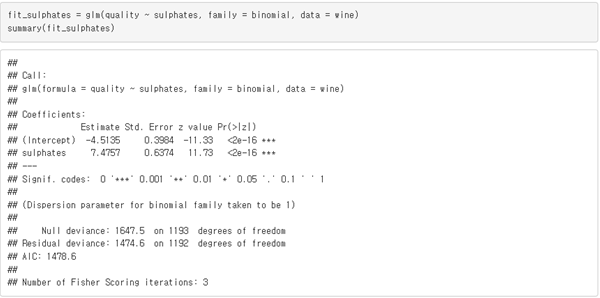
* 적합된 회귀 모형은 다음과 같다

- alcohol이 1단위 증가할 때마다 오즈비는 1.17배가 증가된다.

- AIC는 1385.5이다.



2-②



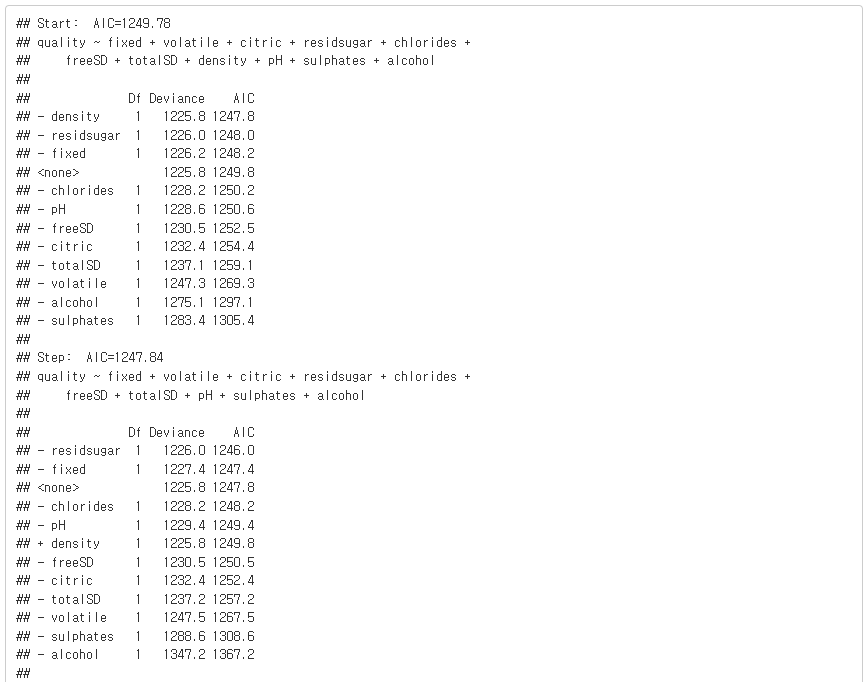
* 적합된 회귀 모형은 다음과 같다

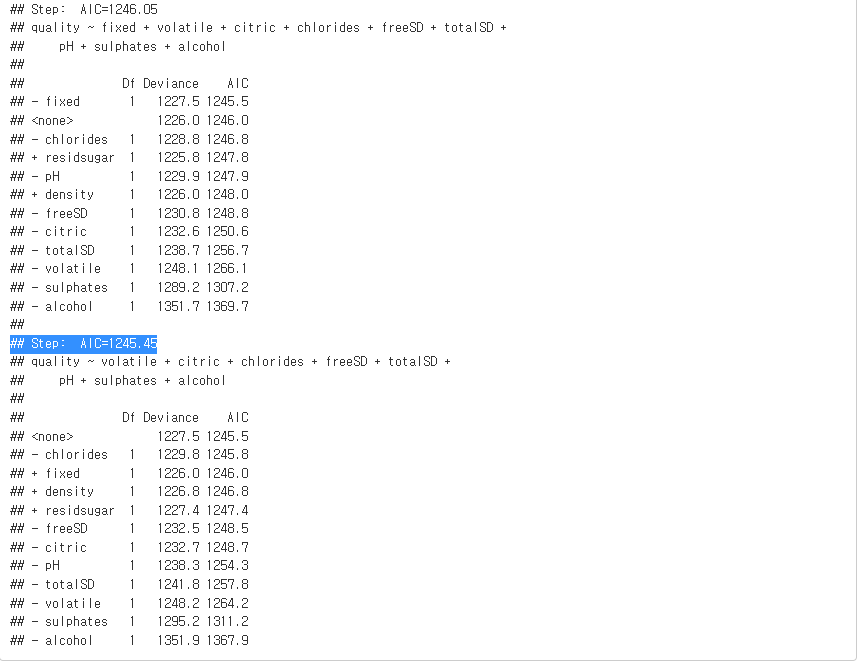
- sulphates이 1단위 증가할 때마다 오즈비는 7.48배가 증가된다.

- AIC는 1478.6이다.

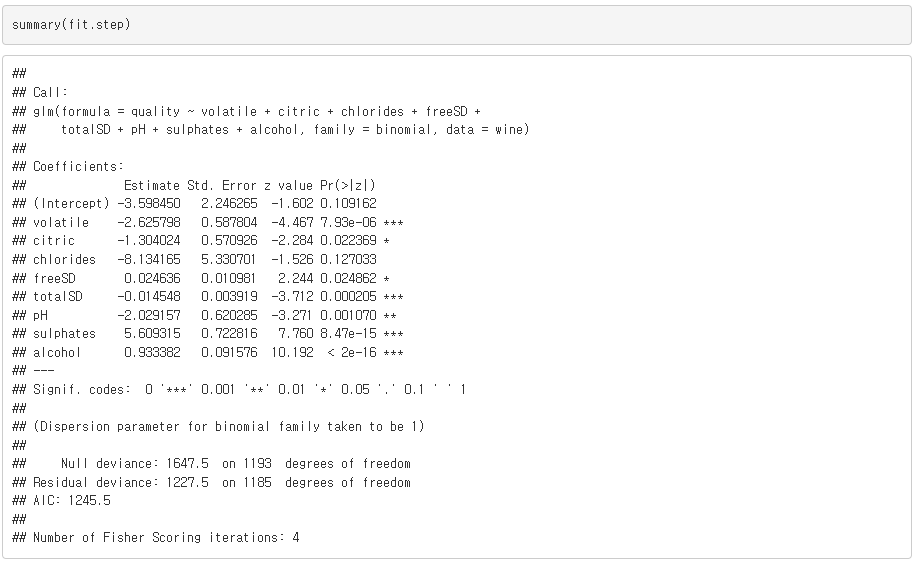


2-③





* 단계적 선택법에 의해 AIC가 제일 낮은 모형을 선택하였다.



* 단계적 선택법으로 입력변수들을 선택한 결과 density, residsugar, fixed 입력변수가 제외되고 나머지 변수들이 선택되었고 회귀식은 아래와 같다.

- AIC값은 1245.5이다.

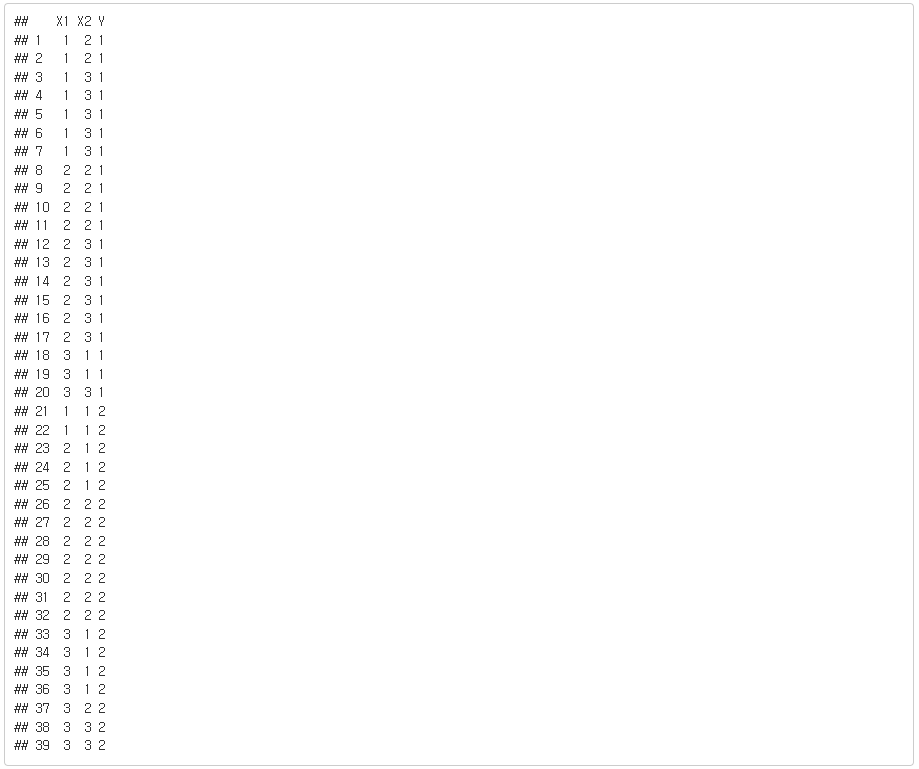


* 적합한 모형에서 얻은 예측값과 관측값으로 만든 정오분류표를 확인해보면 예측정확도는 75.1%, 오분류율은 24.9%이고 민감도는 76.7%, 특이도는 73.2%이다.

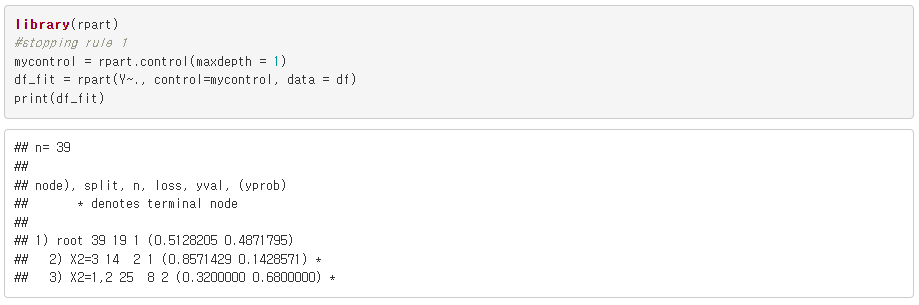
2-①②③

* ①②③ 3가지 모형의 적합도를 AIC값으로 비교해 보면 ③번 모형이 1245.5로 가장 낮은 값을 갖는것으로 보아 가장 잘 적합된 모형이다.
* 예측정확도도 확인해 보면 ①번 모형의 예측정확도는 71.2%이고 ②번 모형은 64.5% ③번 모형은 75.1%로 ③번 모형의 예측정확도가 가장 높게 나타나는 것을 알 수 있다.

3-(1)



3-(2)



최초분할은 X2=3과 X2=1,2인 경우로 분할이 되고 최초 분할 후 지니지수는 아래와 같다.

X2=3일 경우는

X2=1,2일 경우는

가중평균은

3-(3)

3-(2)에서 구한 의사결정 나무에서 X=3인 경우에는 Y=1인 경우는 85.71%를 차지하고 Y=2인 경우는 14.29%를 차지하고 있다. 그리고 X2=1,2인 경우에는 Y=1인 경우는 32%에 해당하고 Y=2인 경우에는 68%를 차지하고 있다.

3-(4)

root인 경우이 지니지수는 였기 때문에 한번분할 후 가중평균은 0.3669이기 때문에 불순도가 0.1328 정도 감소하는 효과가 발생했다.

<참고자료>

장영재, 김현중,조형준 / 데이터마이닝 / 한국방송통신대학교출판문화원 / 2023.1.25

서대호 / 잡아라!텍스트마이닝with파이썬 / 비제이퍼블릭 / 2019년 4월

* 2023년 5월 14일 데이터마이닝 출석수업 과제물 끝 -