KU-BIG
개론스터디

2019-1 학기

PART.Ⅱ 전처리

데이터 전처리 과정은 모델링을 하기 전에 모델에 맞는 형태로 변환하거나 필요한 데이터만을 뽑는 등의 작업으로 필수적이면서 매우 중요한 부분이다. 그러나 데이터 전처리는 정말 많은 과정을 요구하며, 자동화되기도 힘들다. 또한, 현실의 데이터는 매우 다양하기에 분석 목적에 따라 전처리 하는 방법도 달라 전처리를 개념적, 이론적으로 딱 잘라서 설명하기는 힘들다. 그렇지만 여기서는 기본적이면서 큰 범주로 나누어 전처리에 대해서 설명하였다.

1. EDA(탐색적 자료 분석)란?

데이터의 특징과 내재하는 구조적인 관계를 알아내기 위한 분석 기법으로 이러한 자료의 탐색 과정을 통하여 얻은 정보를 기초로 통계모형을 세울 수 있음, 미지의 특성을 파악하고 자료구조를 파악할 수 있는 증거 수집의 과정.

다시 말해, 가설 검정 등의 이론이 아니라 데이터를 있는 그대로 보여주는 과정에 중점을 둠. 다음의 4가지 Issue가 있음

(a) 저항성(Resistance to outliers)

이상치, 결측치, 이상오류의 영향을 받지 않는 tool을 사용

예를 들어, 평균값은 Outlier에 의해 크게 영향을 받지만, 중앙치는 변화가 없음.

(b) 잔차의 해석

각 값들이 주된 흐름에서 얼마나 벗어나 있는지를 탐색하고, 그 이유가 무엇인지 생각해보는 작업

(c) 자료의 재표현

자료의 여러 가지 성질을 나타낼 수 있는 다양한 형태로 표현하는 것을 의미.

예를 들어서 같은 값이라도 log를 씌워 본다든지 하는 것은 비선형적 데이터를 선형적으로 나타내어볼 수 있음.

(d) 자료의 현시성

시각화(visualization)라고도 불림.

시각화를 통해 데이터 의미를 직관적으로 전달할 수 있기 때문

2. 데이터 확인

(a) 변수 확인

- 독립/종속 변수의 정의, 각 변수의 유형, 변수의 타입을 확인

- 기본 적인 부분이지만, 변수의 Type에 따라 모델에서 다른 결과를 도출

(b) RAW data 확인

1)단변량 분석(연속형)

-군집이 존재

-집중도 높은 구간

-대칭성 여부

-자료의 범위 및 산포

-Y와의 관계

2) 단변량 분석(범주형)

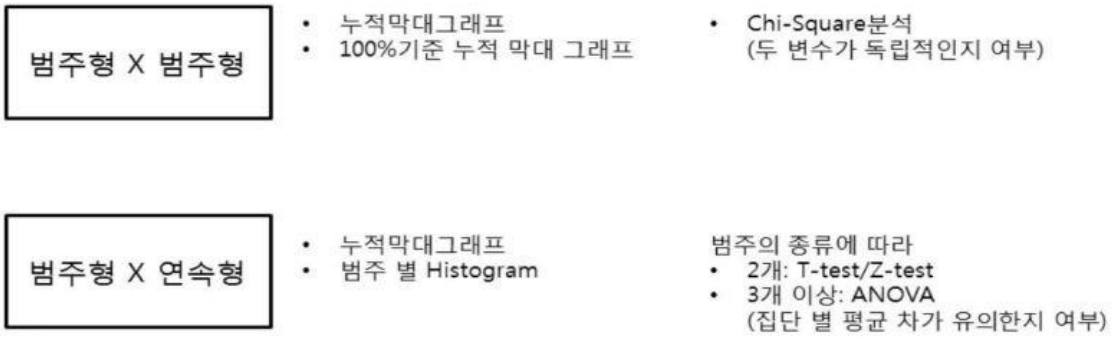
-자료의 분포

3) 이변량 분석(연속형)

-Scatter plot

-Correlation분석

4) 이변량(범주형)



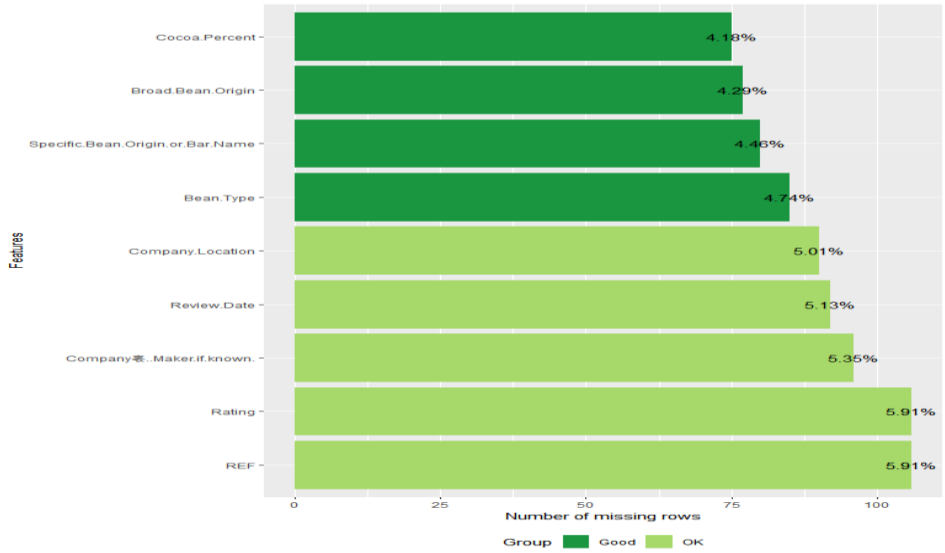
5) 다변량

3. 데이터 정제

결측치를 제거하거나 채우고, 이상치를 다루고 모순된 데이터는 정합성이 맞는 데이터로 교정하는 작업이다.

(a) 결측값

각 변수마다 결측치의 비율을 확인하여 결측치가 10% 이상인 변수는 제거하는 것이 바람직하다. 10% 이상이 아니라면 그 결측치를 대체할 수 있는 방법으로 평균값, 중간값 활용 혹은 해당 변수의 분포에 따른 랜덤추출이 있다.

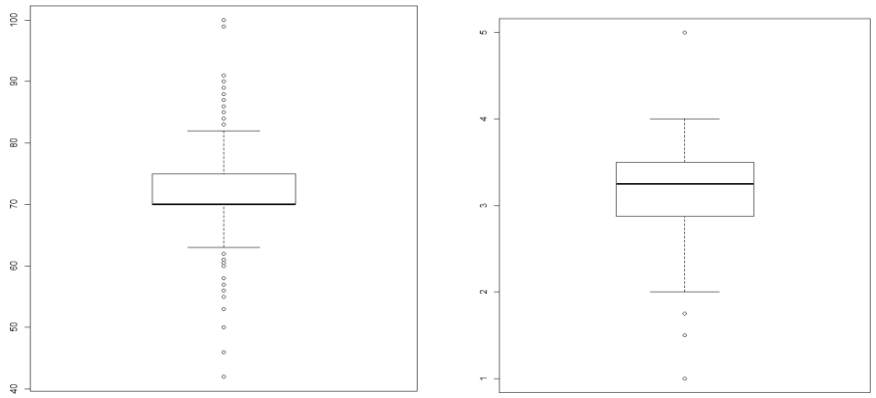


- 삭제
- 다른 값으로 대체
- 예측 값 삽입

(b) 이상치

이상치를 검출하는 방법으로는 가장 기본적으로 Box plot을 활용하는 방법이 있다. 하지만 변수 하나만을 가지고 이상치라고 속단할 수는 없다. 키도 이상치고, 몸무게도 이상치라면 이 사람은 그럴 수 있는 사람이다. 키는 1m 60cm인데, 몸무게가 110kg이라면 이상치라고 할 수 있다. 따라서 다양한 방법을 통한 이상치 검출이 필요하다. 일반적으로 이상치는 제거하는 방법을 많이 사용하지만 로그변환 같은 방법으로 이상치를 안정시키는 방법을 사용할 수도 있다.

- 검출 : 내면 스튜던트화 잔차 확인, Leverage, Cooks'D
- 처리 : 삭제, 상'하한선 제한, 케이스 분리 분석



4. 데이터 축소

샘플링 등을 통해 데이터의 양을 줄이거나 차원(변수,속성)을 줄이는 작업이다. 분석하려는 데이터에 너무 많은 변수가 존재한다면, 데이터 분석 작업의 시간 효율이 떨어질 수 밖에 없다. 또한 데이터 분석에 영향을 미치지 않거나 타 변수와 중복적 성격을 띠는 것도 많이 존재할 수 있다. 이러한 변수들을 충분히 제거하지 않고 분석작업을 시행하면, 알고리즘에 혼동을 줄 수 있다. 따라서 연관성이 낮은 변수를 제거하고, 중복된 데이터 차원(변수)를 제거하거나, 통합하여 데이터 집합의 크기를 줄이는 노력이 필요하다. 가장 쉬운 방법은 변수들 간의 상관계수를 확인하는 것이다. 상관계수가 클수록 굳이 두 변수 모두 넣을 필요 없이 하나만 선택하여 넣어도 충분하다. 또 하나의 방법으로는 주성분분석(PCA) 있다. 이는 변수들의 선형결합을 통해 새로운 변수를 만드는 것으로서 차원을 축소시키는 대표적인 방법이다.

5. 데이터 변환

데이터 변환은 주로 변수의 대칭성 혹은, 종속변수와 독립변수간의 선형성 확보를 위해 시행되곤 한다. 그리고 변수간의 척도(scale)를 맞춰주는데 사용한다. 대표적인 방법으로는 정규화가 있다.

(X – mean(x))/sd(x) (표준정규화)

(X- min(x))/(max(x) – min(x)) (MIN-MAX 정규화)

- Scaling : max-min, Log, 표준화

- Binning : 연속형 변수를 묶어 범주화

- 파생변수 : 변수에서 또 다른 변수를 만들어 내는 것