빅데이터 애널리틱스

< 데이터 전처리 계획 수립 >

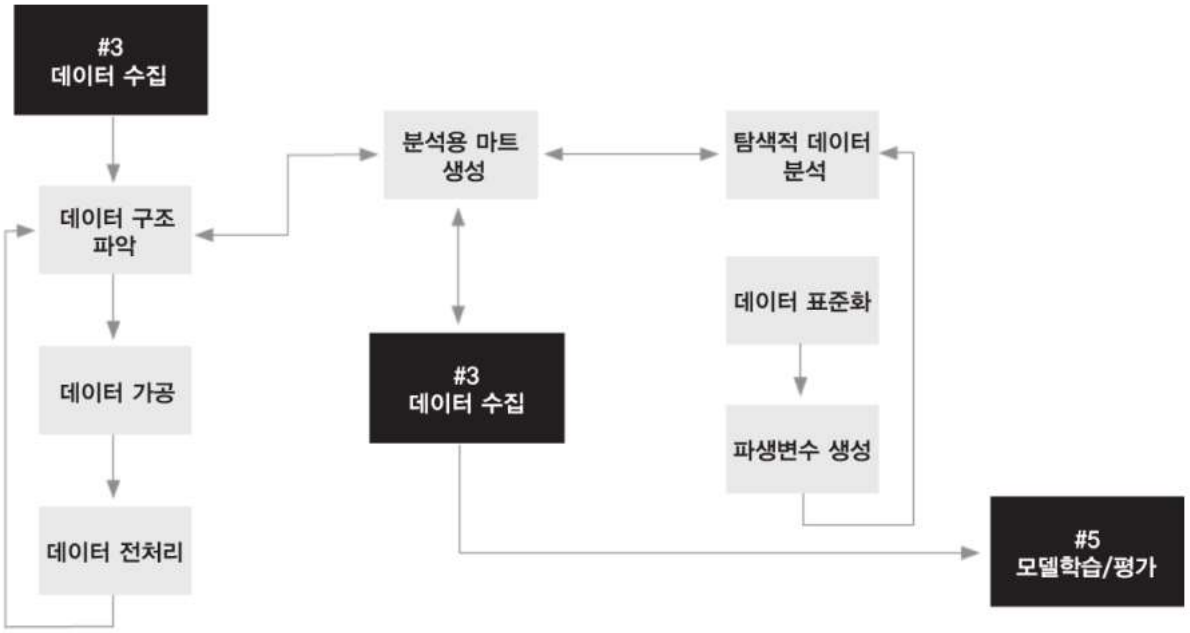
빅데이터 애널리틱스 8조

2018045214 최준희

데이터 탐색과 전처리 단계 - 요약

데이터 탐색과 데이터 전처리 단계

데이터 모델링은 아래와 같은 방식으로 진행되고, 여기서는 데이터 수집 과정과 모델학습 과정은 배제한다.



데이터 구조 파악 > 변수 유형 확인 > 표준화 필요성 > 데이터 양 확인 > 탐색 완료

데이터 결합 > 목표 변수 생성 > 이산화 변수 생성 > 이상치 처리 > 이상치 정제 > 결측치 처리 -> 결측치 정제 > 필요시 추가 처리 > 데이터 전처리 완료

데이터 전 처리의 각 단계별로 어떠한 이론이나 알고리즘을 사용한 근거가 필요하다. (질문 무조건 들어옴)

아래는 데이터 탐색과 전처리에 대한 간략한 가이드이다.

1. 데이터 구조 파악

   데이터 구조에 대해 정량/정성적 파악이 필요하다. 이는 변수 유형 확인, 표준화 필요성, 데이터 양의 균형 확인 등을 거친다. (다행히 SECOM 데이터에는 약간의 표준화, 데이터 불균형 해결을 중점으로 진행하면 된다.)

2. 데이터 가공 과정 - 변수(Features)

   데이터의 구조를 파악했다면, 이제 목표했던 모델의 효율적인 학습을 위하여 데이터를 가공해야 한다. 분석하고자 하는 정보들을 하나의 데이터 셋으로 새로 만들어 분석하거나, 목표 변수 (우리의 경우 Pass/Fail)를 사용할 수도 있다. 또한 연속형 변수는 변환없이 사용하지만 범주형 변수로 변환하였을 때 더 많은 정보를 얻을 가능성이 있다.

3. 데이터 가공 과정 - 데이터 처리

   데이터에 있는 극단적인 값이나, 비어 있는 특성들을 다시 채워줄 필요가 있다. 이 때, Boxplot을 통해 분포를 확인하여 이상치를 쉽게 확인할 수 있다.

   이상치는 상한/하한을 정해 제거하는 편이 매우 건강에 이롭다. 결측치는 지난 발표 2에서 발표했떤 내용들을 바탕으로 채워 넣는다. (제거, 평균, 단순확률, KNN, 다중대치 등등등)

4. 데이터 처리 마무리

   1~3과정에서 얻은 인사이트를 기반으로 데이터 셋을 새롭게 조합하거나, 테스트 데이터 셋과 트레인 데이터 셋의 분리 및 모델링에 있어 필요한 데이터를 최대한 tidy 하게 마무리한다.

References

논문 자료만 APA 스타일로 기록함.

[데이터 전처리와 데이터 탐색]

https://dataonair.or.kr/4-%EB%8D%B0%EC%9D%B4%ED%84%B0-%EC%A0%84%EC%B2%98%EB%A6%AC-%EB%8D%B0%EC%9D%B4%ED%84%B0-%ED%83%90%EC%83%89/

[데이터 전처리에 대한 가이드 예제]

https://wikidocs.net/16582

http://www.dodomira.com/2016/10/20/how_to_eda/

[깔끔한 데이터]

https://partrita.github.io/posts/tidy-data/

[데이터 전처리 이론 및 실습]

https://velog.io/@kimdukbae/%EB%8D%B0%EC%9D%B4%ED%84%B0-%EC%A0%84%EC%B2%98%EB%A6%AC-%EC%9D%B4%EB%A1%A0-%EB%B0%8F-%EC%8B%A4%EC%8A%B51

[데이터 전처리와 머신러닝]

https://skyil.tistory.com/100