문항 1. 데이터 분석과 빅데이터 분석의 차이점에 4 가지에 대해서 설명하시오.

데이터 분석이란 문제 해결을 위해 데이터를 수집, 취합 및 변형해 원하는 정보를 찾아내는 작업이며, 이를 통해 결론을 내릴 수 있게 됩니다.

빅데이터 분석은 많은 양의 복잡한 데이터를 기존 데이터 처리 프로그램을 통해 처리하기 어렵기 때문에, 이 많은 정보 중 유용한 정보를 발견하고 의사 결정을 지원한다는 목표로 데이터를 검사, 정리, 변형 및 모델링하는 프로세스를 말합니다.

데이터 분석과 빅데이터 분석의 차이점 첫번째는 데이터의 대규모화입니다.

데이터 분석은 데이터를 수집하고 비즈니스에 대한 인사이트를 얻거나 의사결정을 위함이라면, 빅데이터는 대규모 데이터를 이용해 더욱 높은 수준의 의사결정이나 큰 수준의 시스템상의 문제를 해결할 수 있습니다.

차이점 두번째는 데이터의 다양화입니다.

데이터 분석은 정형적인 데이터 분석을 중심으로 이루어지지만, 빅데이터 분석은 사진이나 텍스트까지 모두 포함한 비정형 데이터까지 활용합니다.

차이점 세번째는 데이터 분석의 필요 기술 및 수집/처리/저장 기술입니다.

데이터 분석은 프로그래밍이나 통계, 수학 등의 지식이 필요하지만 빅데이터 분석은 프로그래밍, SQL 데이터베이스등의 시스템이나 프레임 워크에 대한 지식이 필요합니다.

차이점 네번째는 데이터의 확장입니다.

데이터 분석은 소규모 비즈니스, 과학, 건강, 에너지, 정보관리 등, 그리고 내부의 데이터 분석 위주로 이루어지지만 빅데이터는 금융 시스템, 서비스, 통신, 정보기술과 웹, SNS 등의 외부 데이터까지 모두 사용합니다.

문항 2. 데이터 전처리 기술 3 가지가 무엇인지 명칭을 작성하고, 그 개념에 대해 설명하시오.

데이터 전처리 기술은 3 가지가 있습니다.

첫째, 데이터 정제 분석하고자 하는 데이터와 친해지는 단계입니다.

데이터 정제에 대해 아래 두가지 확인 작업이 필요합니다.

1)변수 확인

독립/종속 변수의 정의, 각 변수의 유형(범주형인지 연속형인지), 변수의 데이터 타입(Date 인 지, Character 인지, Numeric 인지 등)을 확인합니다.

2)RAW 데이터 확인

(1) 단변수 분석

변수 하나에 대해 기술 통계 확인을 하는 단계, Histogram 이나 Boxplot 을 사용해서 평균, 최 빈값, 중간값 등과 함께 각 변수들의 분포를 확인합니다.

(2) 이변수 분석

변수 2 개 간의 관계를 분석하는 단계 입니다. 아래 그림과 같이 변수의 유형에 따라 적절한 시각화 및 분석 방법을 택합니다.

(3) 셋 이상의 변수

세개 이상의 변수 간의 관계를 시각화/ 분석해야 할 경우, 범주형 변수가 하나이상 포함되어 있는 경우 변수를 범주에 따라 쪼갠 후에 위 분석 방법에 따라 분석합니다.

2)데이터 스케일링

결측값이 있는 상태로 모델을 만들게 될 경우 변수간의 관계가 왜곡될수 있기 때문에 모델의 정확성이 떨어지게 됩니다. 결측값이 발생하는 유형은 결측값이 무작위로 발생하느냐, 아니면 결측값의 발생이 다른 변수 와 관계가 있는지 여부에 따라 결측값을 처리하는 방법도 조금씩 달라지게 됩니다.

결측값 처리 방법의 종류

(1)삭제 결측값이 발생한 모든 관측치를 삭제하거나 (전체 삭제, Listwise Deletion), 데이터 중 모델 에 포함시킬 변수들 중 관측값이 발생한 모든 관측치를 삭제하는 방법(부분 삭제)이 있습니다. 전체 삭제는 간편한 반면 관측치가 줄어들어 모델의 유효성이 낮아질 수 있고, 부분 삭제는 모델에 따라 변수가 제각각 다르기 때문에 관리 Cost 가 늘어난다는 단점이 있습니다.

(2) 다른 값으로 대체 (평균, 최빈값, 중간값) 결측값이 발생한 경우 다른 관측치의 평균, 최빈값, 중간값 등으로 대체할 수 있습니다. 모든 관측치의 평균값 등으로 대체하는 일괄 대체 방법과, 범주형 변수를 활용해 유사한 유형의 평 균값 등으로 대체하는 유사 유형 대체 방법이 있습니다.

(3) 예측값 삽입 결측값이 없는 관측치를 트레이닝 데이터로 사용해서 결측값을 예측하는 모델을 만들고, 이 모델을 통해 결측값이 있는 관측 데이터의 결측값을 예측하는 방법입니다. Regression 이나 Logistic regression 을 주로 사용함 대체하는 방법보다 조금 덜 자의적이나, 결측 값이 다양한 변수에서 발생하는 경우 사용 가능 변수 수가 적어 적합한 모델을 만들기 어렵고, 또 이렇게 만들어 진 모델의 예측력이 낮은 경우에는 사용하기 어려운 방법입니다.

3)이상값 처리

이상값이란 데이터/샘플과 동떨어진 관측치로, 모델을 왜곡할 가능성이 있는 관측치를 말합니다.

이상값 찾아내기

이상값을 찾아 내기 위한 쉽고 간단한 방법은 변수의 분포를 시각화하는 것이며, 일반적으로 하나 의 변수에 대해서는 Boxplot 이나 Histogram 을, 두개의 변수 간 이상값을 찾기 위해서는 Scatter plot 을 사용합니다. 시각적으로 확인하는 방법은 직관적이지만 자의적이기도 하고 하나하나 확인해야 해서 번거로 운 측면이 있습니다.

(1)단순 삭제

이상값이 Human error 에 의해서 발생한 경우에는 해당 관측치를 삭제하면 됩니다. 단순 오타나, 주관식 설문 등의 비현실 적인 응답, 데이터 처리 과정에서의 오류 등의 경우에 사용합니다.

(2) 다른 값으로 대체

절대적인 관측치의 숫자가 작은 경우, 삭제의 방법으로 이상치를 제거하면 관측치의 절대량 이 작아지는 문제가 발생합니다. 이런 경우 이상 값이 Human error 에 의해 발생했더라도 관측치를 삭제하는 대신 다른 값(평 균 등)으로 대체하거나, 결측값과 유사하게 다른 변수들을 사용해서 예측 모델을 만들고, 이상 값을 예측한 후 해당 값으로 대체하는 방법도 사용할 수 있습니다.

(3)변수화

이상값이 자연 발생한 경우, 단순 삭제나 대체의 방법을 통해 수립된 모델은 설명/예측하고자 하는 현상을 잘 설명하지 못할 수도 있습니다. 자연발생적인 이상값의 경우, 바로 삭제하지 말고 좀 더 찬찬히 이상값에 대해 파악하는 것이 중요합니다.

(4) 리샘플링

자연 발생한 이상값을 처리하는 또 다른 방법으로는 해당 이상값을 분리해서 모델을 만드는 방법이 있습니다.

(5) 케이스를 분리하여 분석

위와 동일한 사례에서 실은 경력이 지나치게 길어질 경우 연봉이 낮아지는 현상이 실제로 발생할 수도 있습니다. 이 경우 이상값을 대상에서 제외시키는 것은 현상에 대한 정확한 설명이 되지 않을 수 있습니다. 보다 좋은 방법은 이상값을 포함한 모델과 제외한 모델을 모두 만들고 각각의 모델에 대한 설명을 다는 것입니다. 자연 발생한 이상값에 별다른 특이점이 발견되지 않는다면, 단순 제외 보다는 케이스를 분리 하여 분석하는 것을 추천합니다.