마케팅 사례로 배우는 빅테이터 기초와 비지니스 활용 Q&A

Q1. 데이터 분석과 빅데이터 분석의 차이점에 4가지에 대해서 설명하시오. (48점)

A. 데이터 분석과 빅데이터 분석의 차이점은 **첫번째, 데이터의 확장입니다.** 데이터 분석은 조직 내부의 데이터 분석을 중심으로 이루어졌다면, 빅데이터 분석은 웹 상, SNS상의 외부 데이터까지 활용합니다. 예를 들어, 과거 기업들은 자사의 매출 분석만 시행했다면 요즘은 웹과 SNS상의 외부 데이터까지 분석해 자사 제품에 대한 소비자의 인식을 확인한다고 볼 수 있습니다. **두번째, 데이터의 다양화입니다.** 데이터 분석은 정형 데이터 분석 중심으로 이루어졌지만, 빅데이터 분석은 사진, 동영상, 텍스트 모두 포함하여 비정형 데이터까지 활용한다는 점에서 차이가 있습니다. **세번째, 데이터의 대규모화입니다.** 데이터 분석에 비해서 빅데이터 분석은 분석 대상 데이터의 규모에 큰 차이가 있습니다. **네 번째, 데이터 수집, 처리, 저장 기술입니다.** 데이터 분석에 비해 NoSQL, Hadoop 등의 다양한 프레임워크가 활용됩니다.

Q2. 데이터 전처리 기술 3가지가 무엇인지 명칭을 작성하고, 그 개념에 대해 설명하시오.(필요시, 사용되는 세부적인 기술을 설명하시오.) (52점)

A. 데이터 전처리 과정은 가장 많은 노력과 고생이 필요한 단계이다. 전처리 과정없이 수집한 데이터를 바로 분석에 적용하는 경우는 거의 없다. 전처리가 필요한 경우가 대부분이다. 왜냐하면 데이터를 생성할 때 분석을 전제로 데이터를 생성하지 않았기 때문이다.

데이터 전처리 기술은 다음 3가지가 있다.

1. 데이터 정제
분석하고자 하는 데이터와 친해지는 단계입니다. 데이터 정제에 대해 아래 두가지 확인 작업

A. 변수 확인
독립/종속 변수의 정의, 각 변수의 유형(범주형인지 연속형인지), 변수의 데이터 타입(Date인지, Character인지, Numeric 인지 등)을 확인

B. RAW 데이터 확인
B-1. 단변수 분석
변수 하나에 대해 기술 통계 확인을 하는 단계, Histogram이나 Boxplot을 사용해서 평균, 최빈값, 중간값 등과 함께 각 변수들의 분포를 확인
B-2. 이변수 분석
변수 2개 간의 관계를 분석하는 단계 입니다. 아래 그림과 같이 변수의 유형에 따라 적절한 시각화 및 분석 방법을 택
B-2. 셋 이상의 변수
번거롭지만 세개 이상의 변수 간의 관계를 시각화, 분석해야 할 경우, 범주형 변수가 하나이

상 포함되어 있는 경우 변수를 범주에 따라 쪼갠 후에 위 분석 방법에 따라 분석

2. 데이터 스케일링

결측값이 있는 상태로 모델을 만들게 될 경우 변수간의 관계가 왜곡될수 있기 때문에 모델의 정확성이 떨어지게 됨.

결측값이 발생하는 유형은 결측값이 무작위로 발생하느냐, 아니면 결측값의 발생이 다른 변수와 관계가 있는지 여부에 따라 결측값을 처리하는 방법도 조금씩 달라짐.

결측값 처리 방법의 종류

A. 삭제

결측값이 발생한 모든 관측치를 삭제하거나 (전체 삭제, Listwise Deletion), 데이터 중 모델에 포함시킬 변수들 중 관측값이 발생한 모든 관측치를 삭제하는 방법(부분 삭제)이 있음

전체 삭제는 간편한 반면 관측치가 줄어들어 모델의 유효성이 낮아질 수 있고, 부분 삭제는 모델에 따라 변수가 제각각 다르기 때문에 관리 Cost가 늘어난다는 단점이 있음

B. 다른 값으로 대체 (평균, 최빈값, 중간값)

결측값이 발생한 경우 다른 관측치의 평균, 최빈값, 중간값 등으로 대체할 수 있는데요, 모든 관측치의 평균값 등으로 대체하는 일괄 대체 방법과, 범주형 변수를 활용해 유사한 유형의 평균값 등으로 대체하는 유사 유형 대체 방법이 있음

C. 예측값 삽입

결측값이 없는 관측치를 트레이닝 데이터로 사용해서 결측값을 예측하는 모델을 만들고, 이 모델을 통해 결측값이 있는 관측 데이터의 결측값을 예측하는 방법입니다. Regression이나 Logistic regression을 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 방법임

3. 이상값 처리

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

이상값 찾아내기

이상값을 찾아 내기 위한 쉽고 간단한 방법은 변수의 분포를 시각화하는 것, 일반적으로 하나의 변수에 대해서는 Boxplot이나 Histogram을, 두개의 변수 간 이상값을 찾기 위해서는 Scatter plot을 사용함

시각적으로 확인하는 방법은 직관적이지만 자의적이기도 하고 하나하나 확인해야 해서 번거로운 측면이 있음

A. 단순 삭제

이상값이 Human error에 의해서 발생한 경우에는 해당 관측치를 삭제하면 됨, 단순 오타나, 주관식 설문 등의 비현실 적인 응답, 데이터 처리 과정에서의 오류 등의 경우에 사용함

B. 다른 값으로 대체

절대적인 관측치의 숫자가 작은 경우, 삭제의 방법으로 이상치를 제거하면 관측치의 절대량이 작아지는 문제가 발생함

이런 경우 이상 값이 Human error에 의해 발생했더라도 관측치를 삭제하는 대신 다른 값(평균 등)으로 대체하거나, 결측값과 유사하게 다른 변수들을 사용해서 예측 모델을 만들고, 이상

값을 예측한 후 해당 값으로 대체하는 방법도 사용할 수 있음

C. 변수화

이상값이 자연 발생한 경우, 단순 삭제나 대체의 방법을 통해 수립된 모델은 설명/예측하고자 하는 현상을 잘 설명하지 못할 수도 있음

자연발생적인 이상값의 경우, 바로 삭제하지 말고 좀 더 찬찬히 이상값에 대해 파악하는 것이 중요함

D. 리샘플링

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

E. 케이스를 분리하여 분석

위와 동일한 사례에서 실은 경력이 지나치게 길어질 경우 연봉이 낮아지는 현상이 실제로 발생할 수도 있음 이 경우 이상값을 대상에서 제외시키는 것은 현상에 대한 정확한 설명이 되지 않을 수 있음, 보다 좋은 방법은 이상값을 포함한 모델과 제외한 모델을 모두 만들고 각각의 모델에 대한 설명을 다는 것임

자연 발생한 이상값에 별다른 특이점이 발견되지 않는다면, 단순 제외 보다는 케이스를 분리하여 분석하는 것을 추천