9
정보의 부재에도 의미가 있다.

빅 데이터가 아무리 계속 커져도 세상의 모든 것에 대해 전부 다 알게 되는 날은 오지 않을 것이다. 영화를 좋아하는 이들은 잘 알겠지만 영국식 건조한 유머로 가득 찬 풍자성 코미디인 'Hitchhiker's Guide to the Galaxy'라는 영화에 등장하는 'Deep Thought'라는 슈퍼 컴퓨터에 의하면, "The Ultimate Question of Life, the Universe, and Everything", 즉 "인생, 우주와 그 모든 것에 관한 궁극적인 질문"에 대한 대답은 "42"라고 한다.(우연히도 42는 필자가 내기할 때 즐겨 쓰는 숫자이기도 하다.)

물론 웃기자고 쓴 각본이겠지만 아무튼 이 세상의 모든 정보에 다 접근할 수 있는 빌딩 몇 개만한 크기의 컴퓨터가 7백 50만년 동안 반복하여 계산하고 또 검토한 대답도 그리 확신이 차 있는 것처럼 들리지 않는다. 그 "42"란 대답은 기껏해야 복잡한 수학 공식이 내놓은 추정치 같은 숫자에 불과한 것이고, 영화 속의 Deep Thought란 컴퓨터도 "그 대답은 전혀 가치가 없는 것이, 애초에 그 컴퓨터를 디자인하고 명령을 입력한 존재도 뭐가 궁극적인 질문인지 제대로 몰랐기 때문"이라고 말하는 것이다.

분석 결과가 의미 가지려면 질문부터 제대로

그런데 그것이야말로 필자가 이 책을 통해 일관적으로 주장한 바이다.

즉 어떤 분석의 결과가 의미를 가지려면 애초에 질문부터 제대로 해야 한다는 것이다. 원하는 질문도 제대로 정립되어 있지 않은 상황에서 데이터만 많이 모아 놓는다고 해결책이 저절로 나오지 않는다. 그리고 질문에 대한 대답의 형태로 나타나는 분석의 결과를 놓고 어떤 결정을 내릴 것인지도 여전히 사람들의 몫이다.

분석이란 원래 "있는 정보의 효용을 극대화하는 것"이다. 훌륭한 분석 전문가는 완벽한 데이터 셋(Data set)이 만들어지기를 기다리지 않으며, 사실 아무리 기다려 봐야 그런 데이터 셋은 나타나지도 않는다. 게다가 정책 결정자들은 대부분 무작정 오래 기다려 주는 사람들도 아니며, 그들은 80%의 신뢰도를 지닌 대답을 오늘 당장 듣는 것이 99% 확신하는 대답을 3개월 후에 듣는 것보다 훨씬 낫다고 생각하는 경우도 많다.

데이터의 크기가 계속 커지는 이유는 세상의 많은 것들이 디지털화 되어가고 있으며, 디지털화 된 정보의 조각들은 어딘가에 데이터의 형태로 저장되기 때문이다. 예를 들자면 특정 브랜드의 모바일 정보기기에 관한 데이터를 그 회사의 고객들만을 상대로 그 회사의 사업 지역에서만 모으기 시작해도 결과적으로 데이터베이스는 순식간에 커지게 된다. 그리고 모든 비정형 데이터베이스는 일어나는 모든 정보를 이벤트 별로 기록하고 저장하는 것을 주목적으로 설계되어 있다.

그런데 그런 데이터를 90도로 관점을 돌려 봐서 '모든 사람'들이 어떻게 모바일 기기를 사용하고 있는지 알아보면 놀랍게도 적은 수의 사람들이 의미 있는 양의 정보를 가지고 있다는 것을 발견하게 된다. 그것이 많은 이들이 다른 브랜드를 사용해서이건, 사용자가 정보의 추적을 차단해서이건, 아주 구형이나 신형 기기를 사용해서이건, 모든 사람에 대해 모든 정보를 가진다는 것은 참으로 어려운 일이다. 만약 모든 사람들을 1,000개의 변수로 묘사하는 시도를 한다고 가정할 때, 몇 퍼센트의 사람들이나 그 모든 변수에 의미 있는 정보를 담고 있을 것인가? 대다수는 아닐 것이며 100%는 결코 아닐 것이다. 현재 우리는 인류 역사상 가장 많은 정보를 다루고 있지만 모든 이에 관한 모든 정보는 이루기 어려운 꿈이다.

데이터를 틀어 보면 빈 곳이 보일 수밖에 없다

이 책을 통해 필자는 의사결정이란 여러 가지 옵션들에 랭킹을 매겨 고르는 것이며,

그 순서 매김을 제대로 하려면 미래예측용 통계적 모델을 사용하여 그 랭킹의 대상을 점수로 표현해야 한다고 설명한 바 있다(제6장: 랭킹이 관건이다). 그리고 그러한 점수를 바탕으로 한 랭킹을 위한 모델을 제대로 짜려면, 데이터 자체가 그 순서를 매기는 대상, 즉 가구, 개인, 이메일 주소, 회사, 상품 등의 레벨로 먼저 집적화(summarization)되어야 한다. 그렇기

때문에 거래 별, 혹은 이벤트 별로 정리된 데이터는 그 사용의 목적이 소비자를 대상으로 한 마케팅을 위한 분석이라면 소비자 중심으로 변환되어야 한다는 것인데, 문제는 그런 식으로 데이터를 틀어서 보기 시작하면 많은 빈 곳이 생길 수밖에 없다는 점이다.

아무리 열심히 데이터를 모아도 어떤 이가 특정 상품을 구매한 적이 없다면 그 특정 상품을 중심으로 만든 변수는 비어있는 채로 남아있을 수밖에 없다. 만약에 온라인과 오프라인에 관한 성향을 분석하기 위한 변수들을 구분하여 만들어 놓았는데 어떤 특정 인물이 온라인으로 물건을 구매한 적이 없다면 온라인 쪽 변수들은 어떤 가치를 가지고 있을까. 그것 또한 비어있을 수밖에 없다.

하지만 예측적 분석(predictive analytics)에서는 일어나지 않은 일도 일어난 일만큼이나 중요한 법이다. 간단한 예로, 마케팅 캠페인에 '반응한 사람'과 '반응하지 않은 사람'들의 비교가 바로 Response Model의 기본인 것이고 그 차이를 공식으로 표현하는 것이 모델링인데, 그 결과인 모델 점수는 반응할 사람이 될 가능성을 숫자로 표현한 것뿐이다. 이런 과정에서 '반응하지 않은 사람'의 정의가 되지 않으면 이러한 '반응한 사람들'에 관한 분석 자체가 불가능하게 된다.

빈 곳에서도 의미를 찾아야

비어 있는 데이터(Missing Data)는 도처에 널려있다. 그리고 데이터가 없는 데에도 많은 이유가 있다.

먼저 예를 든 바와 같이 대상에 따라 보고할 정보가 없는 경우도 있을 것이다. 혹은 데이터를 수집하는 과정에서 오류가 생겼을 수도 있을 것이다. 그런 경우 또한 아주 흔하다. 혹은 회사나 부서에 관련된 정보제한, 법적이나 보안상의 이유, 프라이버시에 관한 규정 등으로 인해 데이터의 수집 자체가 차단 되어 있는 경우도 가정해 볼 수 있다. 아니면 여러 곳에 흩어져 있는 데이터베이스들을 합치는 과정에서 매칭이 되지 않아 빈 곳이 생길 수도 있다.

이렇듯 Missing Data는 늘 일상적으로 발생하는 것이며, 개인적으로 돌이켜 보아도 꽤 오래 전에 학교를 떠난 이후로 '완벽한' 데이터 셋은 본 적이 없는 듯싶다. 그야 학교에서는 교수들이 어떤 현상을 강조하여 설명하기 위해 짜깁듯 만드는 데이터를 사용하는 경우가 흔해서 그랬지만, 현실에서 볼 수 있는 데이터란 마치 스위스 치즈처럼 여기저기 구멍이 뚫려 있다고 여겨도 된다. 게다가 마케팅 전용 데이터베이스를 들여다보면 제대로 채워지지 않은 변수가 허다하며, 데이터가 흔하다고 하는 요즘에도 기존 데이터만 가지고 그 속에서 의미를 쥐어짜내야 하는 경우가 대부분이다.

여기서 근본적인 질문을 해보도록 하자. 만약 Missing Data가 피할 수 없는 것이라면 그것을 어떻게 다뤄야 할 것인가? 그러한 Missing Data를 어떻게 데이터베이스에 저장할 것인가? 그냥 있는 정보만 수집 저장하고 빈 곳은 그냥 놔둘 것인가, 아니면 그곳에 무언가를 채워 넣어야 할 것인가? 그렇다면 어떤 방법으로 빈 곳을 메울 것인가? 이러한 질문들에 대한 대답은

분명 “42”가 아닐 터이지만 분명히 강조하고 싶은 것은 정보의 부재, 즉 Missing Data에도 의미가 있다는 것이며, 게다가 모든 Missing Data가 다 같은 것이 아니라는 점이다.

더욱이 Missing Data 자체가 흥미로운 배경을 담고 있는 경우도 많다. 예를 들자면 한국에서는 개인정보라고 여겨지는 수입, 나이, 가족여부, 자녀 수 등 인구학적 데이터(demographic Data)가 비교적 자유롭게 거래되는 미국에서도, 그 대상이 아주 부자이거나 아주 가난한 가구일 때 그러한 데이터가 채워져 있지 않은 경우가 많다. 여러 이유가 있겠지만, 공통된 것은 그 두 대상이 다 ‘주소를 알아내기가 어려운 부류’라는 점이다. 그리고 그 자체가 정보인 것이다.

이처럼 어떤 정보는 특정지역이나 연령 별로 수집이 어려운 경우도 있다. 더 나아가 어떤 국가나 지역에서는 특정 변수에 관한 데이터의 수집 자체가 불법인 경우도 많다. 한국도 비교적 규제가 심한 쪽에 속한다.

또한 어떤 부류의 데이터를 볼 때 만약 온라인이나 모바일에 관련된 데이터만 빠져 있다면 그 이유야 어떻든 그런 정보가 없다는 것 자체가 정보가 된다. 그런 경우를 더 깊게 파고 들어갈 때 소비자의 행동이 아니라 회사의 사업 지역을 예측하는 오류에 빠지지만 않는다면, 데이터가 없다는 사실도 정보라고 봐야 한다는 것이다.

Missing Data를 다루는 방법

숫자로 표현되는 데이터, 즉 달러, 원, 날짜 수, 기타 셈이 가능한 변수에서의 Missing Data를 다루는 방법부터 따져보기로 하자.

어떤 변수들은 아무런 거래기록이 없을 때 자연적으로 비어있게 마련이다. 거래 수를 셈하는 변수나 거래 총액 등은 만약 해당된 카테고리 내에 거래 자체가 없었다면 영(0)으로 남아 있을 수밖에 없다. 하지만 다른 변수들은 계산 자체가 ‘불능’인 경우도 있다.

만약에 “개인당 온라인 평균 거래액수”를 계산하려는데 특정 고객에게 온라인 거래한 기록 자체가 없다면 분모가 0인 상태가 되니 그 결과는 0이 아닌 불능, 즉 ‘.’ 등으로 기록되어야 한다. 그것이 바로 요점인데, 데이터베이스에서의 영(0)은 반드시 진정한 0에 국한되어야지 Missing Data나 불능을 기록하는 수단으로 사용되어서는 안 된다는 것이다. 이 칼럼에서 꼭 기억해야만 할 요점을 하나만 고르라고 한다면, 영(0)이란 절대로 ‘정보의 부재’를 표현하는 수단으로 쓰여서는 안 된다는 점이다.

예를 들어 ‘가구당 자녀 수’라는 간단한 정보를 수집, 저장한다고 할 때, 그 0이란 숫자는 정말로 해당 가정에 자녀가 없다는 것이 확인된 경우에만 써야 한다. 자녀가 없는 가정이라는 것이 확인이 되지 않았으면 ‘.’등으로 표시하거나 아예 빈 곳으로 놔두는 것이 0을 사용하는 것보다 훨씬 바람직하다. SAS등 분석 전용 소프트웨어는 없거나 계산이 불능인 숫자를 ‘.’으로

처리하는 경우가 많다.

‘정보의 부재’를 영(0)으로 표현해서는 안 된다

만약에 그런 정보를 외부 데이터베이스에서 들여오는 경우에는

그 변수가 ① 원래 진정한 0인지, ② 데이터 수집 전문 회사도 그 정보를 얻을 수가 없었는지, ③ 수집은 가능했지만 그 변수를 내부 데이터베이스로 들여오는 과정에서 매칭이 되지 않았는지 구분을 해야 한다. 타 데이터베이스와 개인이나 가구 별로 매칭이 되지 않았다는 것 자체도 정보이며, 그런 식으로 생긴 Missing Data는 모델 안에서 다른 유효한 데이터의 값과 예측곡선 상에서 방향을 같이 하는 경우도 많다.

숫자가 아닌 데이터에도 비슷한 룰이 적용될 수 있겠다. 어떤 경우는 빈 곳으로 남아있는 것이 유효한 것이고, ‘정보부재’ 등으로 기록을 분명히 할 수도 있다. 연습 삼아 코드(code), 텍스트(text)나 다른 카테고리 데이터(categorical Data)에서의 Missing Data를 구분해 보자면:

- ‘ ’ – blank or ‘null’ (빈 곳으로 남겨둠)
- ‘N/A’, ‘Not Available’, or ‘Not Applicable’ (정보부재, 적용불가 등)
- ‘Unknown’ (모름)
- ‘Other’ – 기타 (설문조사나 웹 상에서의 메뉴에서 파생된 데이터)
- ‘Not Answered’ or ‘Not Provided’ – ‘거부’ (설문 등에 대답 자체를 거부한 경우로 ‘모름’과 구분되어야 한다.)
- ‘0’ (숫자로 대답이 가능한 경우 확인된 경우에 한해 ‘0’을 사용할 수 있다)
- ‘Non-match’ (타 데이터베이스나 외부 데이터와의 매칭이 안되거나 오류가 있는 경우)
- 기타 등등

이러한 변수 값들이 서로 높은 상관관계(correlation)를 가지고 있어 예측곡선에서 같은 방향으로 움직일 수도 있겠지만, 그렇지 않은 경우도 많기 때문에 반드시 따로 관리가 되어야 하는 것이다. 구분되어 있는 값들은 나중에 합쳐질 수 있지만, 모든 Missing Data가 한 가지 값으로 표현되어 있으면 차후 분석단계에서 나누어 보는 것이 불가능해진다.

실제로 필자는 이런 여러 가지 Missing Data의 표시 값들이 확실하고 유효한 데이터와 모델에서 합쳐져 사용되는 사례를 많이 보아왔다. 예를 들자면 ‘가구 당 수입’이라는 변수 내에서 높은 값과 빈값이 모델 내에서 같은 방향으로 움직이고 또 통합해 사용되어 버리는 경우가 발생하게 되는 것이다. 다시 강조하지만 정보의 부재, 즉 Missing Data도 나름대로의 의미를 가지고 있다.

빈 곳을 메우는 방법

하지만 정보가 없다고 빈 곳이 늘 빈 곳으로 남아있어야 하는 것은 아니다.

통계적 모델을 이용해 추정된 값으로 비어 있는 곳을 메우는 방법도 있다. 실제로 미국에서 전문적으로 데이터를 수집하여 판매하는 회사들은 그러한 방법으로 Missing Data를 통계적으로 추론된 추정치로 대체하여 사용하곤 한다. 그러한 추정치들 또한 도처에 널려 있으며, 그러한 변수를 사용하는 것도 제대로 알고만 사용한다면 별 문제가 되지 않는다. 어차피 세상의 모든 이들에 관한 모든 정보를 가진다는 것은 불가능한 일이며, 통계적 추정치를 사용하는 것이 어림짐작으로 의사결정을 하는 것보다 백 번 낫기 때문이다.

다만 그런 추정치에 관한 한계에 대해서는 사용자들도 어느 정도 알고 있어야 하는데, 예를 들자면 가구당 연 수입을 다른 변수들을 이용하여 추정한다고 할 때에 아주 높은 액수(예: 5억원 이상의 단위)나 특정 액수(예: 8천 6백 5십만 원 등)를 정확히 추정하는 것은 가능하지도 않고 바람직하지도 않다. 흔히 그 결과는 '8천만 원~9천만 원' 등의 '범위(range)'로 주어지게 된다.

그러한 범위를 추정하는 모델을 만드는 것도 여러 단계의 데이터 손질이 필요한 복잡한 일이지만 그것은 전문가들이 걱정할 일이고, 사용자들은 어떤 것이 실제 데이터(real data)이고 어떤 것이 추정치(inferred data)인가에 유념하여 의사결정을 하면 되는 것이다.

그런 추정치를 만드는 과정을 대치법(Imputation)이라고 하는데, 그런 과정에서 반드시 통계적 모델을 사용해야 하는 것은 아니다. 통계전문가들은 많은 Imputation 방법들을 사용하고 있고, 또 그런 방법에 대해 쉽게 의견을 같이 하지도 않는다. 사실 그래서 변수 별로 그 Imputation method에 대해 룰을 만들어 조직 내의 모든 사용자가 공유하는 것이 중요한 것이다.

다수의 분석가들이 다른 방법을 이용함으로 인해 분석이나 모델의 적용단계에서 불규칙하고 일관성이 없는 결과가 나오는 경우가 많다. 가장 바람직한 방법은 데이터베이스를 구축하거나 업데이트하는 과정에서 일관된 룰을 적용하여 미리 빈 곳들을 채워 놓고 모든 분석가나 사용자들이 공통된 추정치를 바탕으로 분석을 시작하는 것이다.

추정치를 만드는 과정인 대치법(Imputation)

구체적인 Imputation 방법에 대해서는 많은 토론이 필요할 것이다.

간단히 존재하는 데이터의 평균치를 사용할 수도 있는데, 그럴 경우 과연 어느 정도의 데이터 존재비율이 적정선일 것인가? 아는 5%의 정보로 95%를 추정하는 것은 누가 봐도 바람직하지 않을 것이다. 아니면 여기서 예를 든 대로 모델을 짜서 빈 곳을 채울 것인가? 그럴 경우 어디에서 타깃 데이터를 가져올 것인가? 만약에 타깃 데이터 자체에 수집의 오류나 정보의 제

한으로 편향성(bias)이 있다면 그런 문제를 어떻게 대처할 것인가? 타깃은 어떻게 정의할 것이며 어떠한 범위(range)로 정할 것인가? 아니면 타깃을 범위가 아닌 연속적 변수(continuous variable)로 놔둘 것인가? 나중에 추정치가 데이터베이스에 기록될 때 그것을 어떻게 실제로 수집된 데이터와 구분할 것인가? 아니면 사용자의 편리, 처리속도, 자동화 등을 위해 그런 표시 자체를 하지 않을 것인가?

중요한 점은 이러한 질문에 대한 구체적 대답보다 조직 내 규칙의 일관성이다. 일관성을 가지면 모든 사용자와 분석가들이 분석의 목적에 상관없이 같은 시작점에서 프로젝트를 추진하게 된다. Imputation Method, 즉 추정 방법에 대해서는 치열한 토론이 있겠지만, 일단 결정이 내려지면 데이터베이스 업데이트 과정에서 정해진 룰에 따라 빈 곳을 채우는 것이 추후 분석과 그 적용 과정에서 발생할 수 있는 많은 에러를 피하는 길이다. 많은 경우 일관성이 없는 추정방법이 일관성이 없는 결과를 가져온다.

만약에 분석가나 통계 전문가에게 그들 마음대로 빈 곳을 메우는 자유가 주어진다면, 그들이 만들어내는 모든 모델 공식(model algorithm)에는 반드시 그 추정치를 계산하는 방법도 포함되어야 한다. 그런 방식은 모델 적용 과정을 길게 만들고 에러가 생길 확률도 높이게 마련이지만, 어떤 단계에서든 변수들의 추정치가 제대로 계산되지 않으면 모델도 따라서 망가지게 되어있다. 그래서 그런 추정치는 데이터베이스를 다루는 부서를 중심으로 관리되어야 한다는 것이고, 데이터베이스의 데이터 사전(Data dictionary)에는 사용된 Imputation Method도 변수마다 꼼꼼히 기록되어야 한다.

Missing Rate에 관심을 둘 것

흔히 모델이 제대로 성능을 발휘하지 못하고 캠페인의 결과가 좋지 않을 때

그 배후를 들여다보면 모델에 사용된 변수 중 데이터의 수집이 제대로 되지 않아 Missing Rate, 즉 빈 곳의 비율이 높아져 있는 것을 발견하게 된다. 뒤집어 말하면 Missing Data의 비율이 모델의 예측능력과 분석의 결과에 지대한 영향을 미친다는 것이다.

두말할 나위 없이 새로운 데이터의 일관적인 공급은 모델이나 공식의 질보다 훨씬 중요한 것이다. 미국식 표현으로 'Garbage-in-garbage-out'(GIGO)의 전형적인 케이스인 것인데, 그래서 요즘 화두가 되고 있는 데이터 거버넌스(Data Governance)가 중요한 것이다. 주지하다시피 데이터베이스는 원래 구축보다 관리가 더 어렵고 중요하다.

그리고 데이터 관리에는 주요 변수들의 Missing Rate이 시간대(time-series)로 정리된 리포트가 작성되고 검토되는 절차가 반드시 포함되어야 한다. 어느 중요한 변수의 데이터 수집이 제대로 안되고 Missing Rate이 허용범위를 벗어나게 되면 아무리 날고기는 통계전문가라도 다른 변수를 사용한 새로운 모델을 짜는 것 이외에 구제방법이 없게 된다. 그러한 일이 피할 수 없는 것이라 하여도 미리 알고 계획하여 대처하는 것이 현명한 것이다. 어차피 통계적 모

델이라는 것도 한정된 유효기간이 있는 것인데, 불규칙하고 일관적이지 않은 데이터는 그 모델의 수명을 더 빨리 단축시키는 역할을 하고, 오르내리는 Missing Rate은 그런 현상의 좋은 지표가 된다.

또한 모델 점수의 분포(model score distribution)가 개발 당시의 그것과 큰 차이를 보인다면, 그 모델에 사용된 모든 변수의 Missing Rate을 검토하는 것이 바람직하다. 앞으로 더 자세히 다루겠지만, 모델 점수 분포의 급작스러운 변화는 데이터베이스 내에서의 바람직하지 않은 현상을 나타내는 경우가 많다.

이번 챕터에서 열거한 Missing Data에 대한 가이드라인을 따르면 분석과 모델에 사용되는 변수들이 훨씬 다양해질 것이고 모델들의 예측능력과 사용기간도 더 연장될 것이다. '정보의 부재'도 의미를 함축하고 있는 것이다. 그런 숨은 의미는 Missing Data를 제대로 다룰 때에만 모습을 드러내는 법이다. 그리고 데이터를 다루는 사람이라면 우리가 모든 것에 대한 모든 것을 알게 되는 날까지 Missing Data의 처리와 관리에 관심을 두어야 한다. 단지 그 "42"라는 대답에 만족할 것이 아니라면 말이다.