3
빅데이터의 핵심은 “분석”, 분석의 중심은 “모델링”

이 책의 서두에서 언급했듯이, 빅 데이터란 단지 많은 데이터를 수집하여 쌓아 놓고 그것을 빨리 꺼내 볼 수 있도록 하는 것이 전부가 아니다. **데이터는 크건 작건 비즈니스를 더 성공적으로 수행하기 위한 도구일 뿐이다.** 의사결정자들의 질문에 대답을 해주는 형태로 정보를 제공해야하지, 많은 데이터의 조각들을 늘어놓고 그 규모에 대해서만 자랑하는 것은 예산만 낭비하고 정작 질문에 대한 대답은 주지 못하는 결과를 낳게 된다. 가공되지 않은 데이터는 광산에서 캐온 금

덩어리일 뿐이며, 그것을 가공하여 시계를 만들기 전에는 순금 자체가 시간을 말해줄 수 없는 것과 같은 이치다. **데이터를 잘 사용하기 위해서는 질문에 대한 대답의 형태를 갖추어야 하며, 그러한 데이터 가공의 중심에는 Analytics, 즉 분석 과정이 자리잡고있다.**

데이터 가공의 중심은 '분석' 과정

하지만 데이터베이스 마케팅이 수십 년간 자리를 잡아온 미국에서도 "Analytics"란 말에 대한 정의가 너무 여러 가지인지라, 일단 이에 대한 정리부터 필요할 것 같다.

그 말을 쓰는 사람에 따라 의미가 천차만별이고, 그로 인해 일을 시작하기도 전에 참여하는 팀원 간에 소통이 잘못 이루어져 첫 단추부터 잘못 꿰는 경우도 비일비재하기 때문이다. **분석가, 혹은 요즘 흔히 말하는 데이터 사이언티스트라면 비즈니스에 대한 이해도 높아야 하며 소통도 잘해야 하며 해결책에 관한 옵션을 제대로 선택하여 당면 과제에 대한 처방도 적절히 내려야 하는 것이다.**
데이터를 크건 작건 오랫동안 써온 마케팅 분야의 시각에서 볼 때 Analytics, 즉 분석이란 다음과 같이 크게 나누어 볼 수 있겠다.

Business Intelligence / BI Reporting

자동차의 계기판에 비유하여 Dashboard Reporting이라고도 하는데, 즉 현재에 어떤 일이 일어나고 있는지를 정확하고 일목요연하게 보여주는 작업이다. 빅 데이터 분석은 사실 여기서부터 출발한 것이며 대부분이 아직 이 단계에 머무르고 있다. 경영자의 입장에서는 자신들의 의사결정이 어떤 결과를 가져오고 있는지를 알 수 있게 해주는 중요한 분석과정이다. 예를 들자면 이메일 캠페인을 할 때 어떤 형태의 offer가 무슨 요일 몇 시경에 어떤 대상에게 가장 효과적인가 등의 결과를 이러한 종류의 BI Reporting을 통해 알게 되는 것이다. 그 '효과적'이란 것도 추상적인 개념이 아니라 구체적인 성공 측정기준(metrics)를 통해 반응률(response rate)과 수익률(profitability) 등을 채널, 메시지, 가격, 날짜, 시간, 고객, 상품 별로 나누어 볼 수도 있게 된다. 인터넷에서 어떠한 광고 문구(ad word)가 어떤 상품의 매출과 연관이 되어있는지 살피는 분석도 이 범주에 포함된다. 이러한 BI Reporting은 인포그래픽 분야에서의 많은 발전 덕분에 사용자가 나날이 더 쉽게 이해할 수 있도록 진화하고 있다. 하지만 이러한 류의 분석은 단지 시작일 따름이다.

Descriptive Analytics

즉 묘사적인 분석이다. 마케팅에 있어서 타깃을 정밀하게 묘사하는 것이 목적으로, 흔히 쓰여지는 Profiling, Segmentation, Clustering 등이 여기에 포함된다. 예를 들자면 주 고객이 주로 30~40대 여성이며 학부형일 가능성이 많으며 그들의 주 거주지는 어느 지역이며 생활수준이 어느 정도인지 라이프 스타일은 어떤지 하는 식으로 대상을 묘사하는 것이다. 이런 분석의 결과로 마치 고객이 눈앞에 있는 것처럼 커뮤니케이션을 계획할 수도 있고, segmentation 등의 기법으로 주 대상이 어느 범주에 속하는 지도 파악할 수 있다. 시간적 개념으로 볼 때는 BI Reporting과 마찬가지로 여전히 현재형 분석이다. 그리고 segmentation이나 clustering도 분석의 끝이 아니라 단지 여러 가지 옵션 중 하나일 뿐이다.

Predictive Analytics

예측적 분석으로 대상의 미래의 행동에 관한 예측을 통계적 확률로 표현하는 것을 가능하게 흔히 언급되는 통계적 모델(statistical modeling)이 이 범주에 속하며, 사용 용도별로는 response model, clone (look-alike) model, value model, revenue model, attrition model 등이 있다. 이러한 예측적 모델 이야말로 현대 분석의 핵심이라고 할 수 있다. 그의 결과인

모델 점수, 즉 score는 다양한 데이터를 함축적으로 내포하고 있기 때문에 빅 데이터 시대에 있어서 복잡한 정보를 간결한 대답의 형태로 의사결정자들에게 전달하는 도구로서 그 중요성이 강조된다. 특히 1-to-1 마케팅에 관한 한 이러한 predictive modeling은 필수적인 도구이다. 데이터가 많아질수록 더욱 주목해야 할 분석 방법이고, 반면에 가장 어렵고 복잡하게 여겨지는 방법이기도 하다.

Optimization Model

최적화 모델이라고 할 수 있으며, 주로 광고회사나 마케팅 에이전시와 마케팅 전략을 세울 때에 필요한 분석 방법이다. 예를 들자면 마케팅 예산에서 어느 정도까지 특정 채널에 할당하여야 전체적 효율이 극대화 되느냐, 어떠한 분야에 예산을 줄여도 부정적인 결과가 최소화 되느냐 등의 "What-if Analysis"의 형태로, econometrics model을 전문으로 하는 분석가들이 주로 하는 일이다. 방법과 용도에서 predictive analytics와는 구분되어야 하지만, 많은 사람들이 이런 일도 그냥 "Analytics"라고 부르는 경향이 많아 정의를 확실하게 하는 것이 중요하다.

빅 데이터 시대 가장 유용한 것은 '예측적 분석'

이 네 가지 유형의 분석틀 중 '예측적 분석'(Predictive Analytics)이, 모든 고객과의 대화가 맞춤형이 되어가는 빅 데이터 시대에 가장 유용하고도 중요하다 할 수 있다.

그 예측적 분석의 중심에 있는 통계적 모델에 대한 고찰 또한 이 시점에서 반드시 필요할 것 같다.

여기서 일단 왜 모델이란 것을 만드는가 하는 질문부터 나올 법 하다. 철학적으로 대답하자면 인간이란 전체에 대한 진실을 알 수 없고, 바로 앞의 미래도 예측할 수 있는 능력이 없어서이다. 세상 만물의 이치를 다 안다면 그야말로 전지전능한 능력을 가진 것이니 모델이고 뭐고 다 필요 없을 것이다. 그래서 물리학자들도 양자역학적 이론 등을 가설로 세우고 또 그에 대한 실제적 실험을 가능하게 하기 위해 모델을 사용한다. 우리가 일상생활에서 늘 대하는 일기예보도 많은 정보를 압축한 모델링에 기초한 것이다.

그렇다면 물리학, 수학, 혹은 통계학과 인연이 별로 없는 경영인들과 마케팅 전문가들에게 이러한 모델링은 무슨 상관이 있을까? 돌이켜 보자면 마케팅 분야에서 모델링과 그 사용의 역사는 의외로 길다. 이미 60년대 말, 70년 대 초에 Reader's Digest등 당시 미국 굴지의 출판 회사들은 통계를 이용하여, 비용이 많이 드는 우편물을 보내기 전에 누가 특정 상품을 구매할 확률이 높은 지 미리 알아내어 어떻게 하면 비용을 절감할 것인가에 대한 해법을 가지고 있었다. 당시 사용했던 컴퓨터들은 그야말로 골동품 수준이라 그 모든 데이터 처리와 모델링은 장시간이 걸리는 작업이었지만, 통계적 이론 자체는 아직도 그대로 적용할 만한 수준급이었다. 그 회사들은 컴퓨터 화면이라는 것도 없어서 펀치카드를 쓰던 시절에 이미 regression model이나 logistics model의 기술을 사용하여 실제로 엄청난 효과를 거두었던 것이다.

간단한 예로 데이터베이스에 수백만 명의 기록이 있는데 그들에게 모든 상품을 다 권한다면 우편물 하나의 발송에 1불만씩만 든다 하여도 그 비용 총액은 수백만 달러에 이르게 된다. 당시의 기초적인 모델링 기술만 가지고도 특정 상품의 offer에 반응할 확률을 모델로 짜서 메일을 보낼 대상을 10분의 1로 줄일 수 있다면 그 비용절감은 엄청나게 된다. 한마디로 마케팅 캠페인을 하기 전에 누가 물건을 살 지 미리 예측하고 골라서 상대했다는 말이다.

21세기로 넘어와 고객을 상대하는 비용이 많이 줄어든 시점에서도 누가 무슨 물건에 대한 관심이 있는지를 미리 안다는 것은 대단한 정보이다. 그리고 그것은 단지 과거에 무슨 상품을 얼마를 주고 샀으며 그 사람이 어느 동네에 살며 얼마나 자주 고객센터에 전화를 해 어떠한 불평을 해왔는지를 기록해두고 꺼내 보는 것보다 훨씬 고차원적인 정보이기도 하다. 사실 단순한 정보의 나열은 의사결정자나 실무자들에게 그리 도움이 되는 것이 아니다. 마케팅의 예를 들자면 마케팅 전문가가 알고 싶은 것은 크게 나누어 ① 과연 어떤 특정 상품을 판매할 때 누구에게 우선적으로 접근해야 하는가, 그리고 ② 어떤 대상에게 접근하기로 정했다면 과연 어떠한 통로로 무슨 offer를 가지고 접근할 것인가 이다. 이러한 질문에 대한 대답은 바로 예측적 분석, 즉 모델링을 통해 이루어질 수 있다.

모델링으로 얻을 수 있는 이점

데이터의 최적화와 자동화가 가능해지며 복잡한 데이터를 사용하기 쉽게 만들어준다.

모델링으로 얻을 수 있는 이점에 대해 말해보겠다

① 이미 예를 들었듯이 호의적으로 반응할 대상을 미리 알아서 마케팅과 영업 비용을 절감한다.
② 직관만 가지고 접근하는 것보다 타깃을 훨씬 더 정확히 파악할 수 있다.
③ 마케팅 노력에 대한 일관성이 있는 결과를 가지게 된다.
④ 되풀이 될 수 있는 과정을 설립하여 커뮤니케이션과 상품 최적화의 자동화가 가능하게 되며, 현존하는 CRM 프로그램의 효과도 증진시킨다.
⑤ 타깃으로 할 대상을 확장하기가 수월 해지며, 그러한 확장에 따르는 위험도까지 감안할 수 있게 된다.
⑥ 직관만으로는 알 수 없는 데이터의 패턴을 발견하게 된다.
⑦ **복잡한 데이터를 사용하기 쉬운 대답의 형태로 만들어주며, 빈 공간을 효과적으로 메워주는 역할을 한다. 이것이 빅 데이터 시대에 가장 중요한 이점이다.**
⑧ 항상 고객에게 적절한 상품과 서비스를 제공할 수 있게 한다.

모델링에 관한 상담을 하다 보면 이미 그러한 기법이 오래 자리를 잡아온 미국에서도 그 효용에 대해 의문을 갖는 사람들을 많이 만나보게 된다. 재미있는 것은 마케팅에 오래 종사해 온 사람일수록 그런 경우가 많다는 것이다. 자신들의 직관으로 잘 하고 있는데 무슨 복잡한 수학 얘기냐는 식의 반응이 많다. 미국인들의 수학이나 통계에 대한 이해가 높지 않은 경우가 많아서 그렇기도 하지만, 일반적으로 사람들은 빅 데이터나 분석 등 자신이 잘 모르는 분야에 일단 의심을 가지기 마련이다.

하지만 곰곰이 따져보자. 필자는 다년간 정말 많은 똑똑한 사람들과 일을 해왔지만 불행히도 두세 가지 이상의 변수를 머리 속으로 따져가며 사용할 수 있는 능력을 가진 사람은 만나본 적이 없다. 설사 그럴 수가 있다 치더라도, 그 변수들의 상관관계까지 암산으로 따진다는 것은 불가능에 가까운 일일 것이다. 반면, 통계적 모델에는 일반적으로 대여섯 개에서 많게는 스무 가지 이상의 변수가 어우러져 있다.

모델링이란 '차이점'에 대한 수학적 표현

그렇다면 모델링이란 과연 어떤 작업인가? 한마디로 통계적 모델이란 어느 상이한 두 집단(dichotomous groups)의 '차이점'에 대한 수학적 표현이다.

일단 비교할 수 있는 두 집단의 예를 들어보자면, 마케팅의 경우 그것은 어떤 물품이나 서비스의 사용자와 비사용자, 캠페인에 반응한 사람과 무관심한 사람, 크레딧을 줄 만한 사람과 그렇지 않은 사람, 고정 고객과 단발성 고객, 높은 수익률을 주는 고객과 그렇지 않은 고객, 자주 찾는 고객과 가끔씩만 들르는 고객 등이 있겠다. 모델링에서의 첫 번째 관문은 이러한 추상적인 개념을 수학적으로 표현하여 '타깃'을 만드는 일이다.

그러려면 일단 그에 대한 데이터를 확보하는 것이 우선이다. 예를 들어 "돈을 많이 쓰는 고객"이라는 인간적 표현을 수학적으로 표현하는 것도 분석과정의 시작이다. 돈을 많이 쓰다니? 수백만 원을 쓰는 고객을 원하는 것인가? 어느 정도의 시간을 두고서? 어느 채널로? 단번에 그런 돈을 써도 타깃에 넣어줘야 하나? 자주 들리면서 소액을 쓰는 고객은 어떻게 하고? 외상으로 그 액수를 긋고 간 고객도 포함해야 하나? 등등, 간단하다고도 할 수 있는 이런 질문에 대해 모든 경우를 다 고려하고 그 대상의 크기도 살펴보고 (너무 크거나 작은 타깃은 모델링에서 효용이 없다) 수학적 표현을 도출해 내야 비로서 타깃을 정했다고 할 수 있다. 타깃을 잘못 걸어 놓으면 아무리 총이 좋아도 목적을 이룰 수 없으니 이것이야 말로 중요한 첫 단추라고 할 수 있으며, 사용자와 분석가가 반드시 머리를 맞대야 하는 작업이다.

'차이'를 변별하는 '변수'의 중요성

일단 타깃이 정해졌으면 타깃과 타깃이 아닌 두 집단의 차이를 변별하는 변수를 찾는 작업에 들어가게 된다.

과거에 수백 가지의 변수로도 시간이 꽤 걸리는 작업이었으니, 흔히 수천 개의 변수가 포함된 현대의 데이터베이스를 다루려면 이 과정 이야말로 가장 어려운 것일 수도 있다. 여기서 중요한 것은 이 모든 과정이 수학적으로 이루어진다는 것이다. 즉, 직관적으로 오랫동안 사용해온 변수들도 만약 두 집단의 차이를 설명하는 데 유용하지 않다면 가차없이 버려지게 된다. 또 이 과정에서 많은 변수들은 합쳐 지기도 하고 숫자의 경우 그룹으로 나누어지던가

공식을 통한 변환을 겪기도 한다. 통계전문가에 따라 공식에서의 모델 내의 이상적인 변수의 수는 천차만별이지만 통계적 이론을 전공하지 않은 사용자의 입장에서는 모델이 유용하기만 하면 별로 신경을 써야 할 부분은 아니다. 고양이가 무슨 색이던 쥐만 잘 잡으면 고양이의 털 색깔은 별 의미가 없는 것과 비슷하다고 하겠다.

수천 개의 변수 중 두 집단의 차이를 설명할 수 있는 변수로 뽑혔다 하여도 그 중요성까지 같은 것은 아니다. 그 차이를 나타내는 중요성에 따라 값이 달리 매겨지며, 그것을 합산한 것이 우리가 말하는 모델 점수, 즉 score이다. 요점은 이 점수에는 그야말로 수천 개의 변수가 고려되고 그 과정에서 걸러진 많은 정보가 사용자가 쓰기 편하게 함축되어 있다는 것이다. 사용자가 알아야 할 것은 단지 모델의 목적과 거기에 따르는 점수이다. 모델이 만약 영업대상에 대한 가치를 고려한 것이라면, 사용자인 영업사원은 일단 점수가 높은 대상부터 상대하는 것이 영업 실적을 높이는 지름길이 된다. 그리고 그 영업사원은 굳이 그 모델 안에 얼마나 많은 양의 정보가 집적되어 있는지 그의 지적 호기심을 충족시키는 이유 외에는 별 상관을 할 필요가 없게 되는 것이다. 정보가 넘쳐나는 시대에 데이터의 효과적인 집적화는 정말 필요한 일이 아닐 수 없다.

직관이 정확할 때도 있지만 매번 그럴 수는 없다

여기서 다시 강조하자면 이러한 과정을 거친 통계적 모델은 일반인의 직관에 비해 훨씬 정확하고 효과적이며, 위의 예 중 3번에 해당하는 일관적인 효과도 기대할 수 있게 된다.

사람의 직관은 정확할 때도 있지만 매번 그렇게 하기가 쉬운 일은 아니다. 현대 경영에서는 예측이 가능한 성공 방식이 중요한 것이며, CRM 등을 대하며 무수한 자동화를 꾀하는 시점에서 모델의 이러한 정보의 집적과 일관성은 효과적으로 정보의 가공과 사용 시 많은 시간을 절약해주는 이점도 제공한다.

게다가 직관을 통한 데이터의 사용이 아무리 성공적이었다 하더라도, 그러한 성공 사례를 자꾸 반복하여 사용하다 보면 효용성이 점차 떨어지거나 그 대상이 아예 고갈되어 버리는 경우도 비일비재하게 생긴다. 그런 경우 목적에 부합하는 새로운 데이터 사용법칙을 만들어야 하는데, 이게 쉬운 일이 아닌 것이다.

미국에서의 마케팅 사례를 하나 들어보겠다. 어떤 마케터가 많은 고소득자를 팔고자 하는 상품에 따라 분류해야 하는데, 수영장과 벽난로를 가진 사람들을 우선으로 타깃으로 삼아 재미를 본 경우가 실제로 있었다. 그야말로 소가 뒷걸음질 치다가 쥐 잡은 격인 것인데, 문제는 그런 성공을 되풀이할 만한 변수를 또 우연히 찾는다는 것은 기적에 가까운 일이라는 것이다. 그런데 애초부터 통계적 이론에 기초하여 그 상품이나 서비스에 밀접히 관련된 예상 고객들을 찾았더라면 얘기가 달라진다. 일단 그는 점수가 높은 사람부터 상대를 했을 것이며, 그 마켓이 포화상태가 되면 조금 낮은 점수로 이동해 가면 될 일인 것이다. 심지어는 그런 낮은 점수로 이동해 가는 과정에서 효용성과 위험도까지 봐 가면서 타깃을 공략할 수도 있게 된다.

직관으로 볼 수 없던 데이터의 상관관계, 모델링으로 찾는다

다음에 설명할 이점은 필자가 가장 선호하는 것 중 하나인데, 즉 직관으로는 볼 수 없었던 데이터의 상관관계를 찾게 된다는 점이다.

이것이 왜 중요하냐 하면 수천 가지의 데이터가 떠돌아 다니는 시대에 흔히 쓰던 몇몇 가지의 변수에만 의존하는 것은 너무나도 전근대적인 방식이기 때문이다. 실제로 미국에서 가장 흔히 쓰이는 데이터는 수입, 나이, 지역, 자택 소유여부, 자녀여부 등과 같은 것들인데, 그보다 훨씬 다양하고 예측에 효과적인 변수를 흔하게 찾을 수 있는 요즘과 같은 시대에 과거의 관행에 이끌려 다니는 것은 마치 어린 아이가 수십 가지 색깔이 있는 크레파스 통을 앞에 놓고도 두세 가지 색깔만 반복해서 사용하는 것과 비슷한 행태이다.

앞에서 설명했듯이 모델링의 첫 과정은 변수의 선택, 즉 variable selection인데, 거기에서 우리는 뜻하지 않은 많은 상관관계를 마주치게 된다. 예를 들자면 한이 없지만 흥미로운 한 가지만 소개하자면, 높은 가격대의 가구를 카탈로그로 판매하는 회사를 위한 모델을 살펴보다가 "정화조를 설치한 주택의 비율"이라는 미 통계국이 제공하는 지역적 변수가 포함되어 있는 것을 본 적이 있었다. 고급 가구를 파는데 정화조라니? 언뜻 보기에는 전혀 상관이 없어 보였지만 더 생각해보니 그 변수는 대상가구가 큰 도시의 중심에서 얼마나 멀리 떨어져 있나를 효과적으로 예측하고 있는 것이었다. 즉 도시의 중심에서 멀리 살면서 비교적 큰 단독 주택에 사는 사람들일수록 고급 가구를 카탈로그로 주문할 확률이 더 높다는 것이다. 물론 그 모델에는 다른 많은 변수도 포함되어 있었지만 이것은 정말 무릎을 칠만한 발견이었다. 문제는 그냥 사람더러 변수를 고르라고 하면 백 년을 고르고 있어도 가구를 팔면서 정화조란 변수를 사용할 사람은 필자를 포함해서 없을 것이란 점이다. 수학은 그러한 고정관념에 얽매어 있지 않다. 어떤 상관관계가 존재한다면 그것은 그냥 수학적인 발견일 뿐인 것이다.

지금까지의 사례는 주로 1-to-1 마케팅에서 얻은 것들이지만, 그 다음 요점인 "복잡한 데이터를 사용하기 쉬운 대답의 형태로 만들어주며 빈 공간을 효과적으로 메워주는 역할을 한다"는 7번과 같은 효과는, 요즘과 같은 빅 데이터 시대에 있어서 모델링의 중요성을 가장 부각시켜 주는 것이라 하겠다. 많은 데이터를 효과적으로 집약한다는 점에 대해서는 많은 설명이 있었지만, "빈 공간을 메워준다"라는 말에는 고개를 갸우뚱하시는 독자들이 계실 것 같다. 결론부터 말하자면 빅 데이터는 자세히 들여다 보면 구멍투성이이다. 서론에서 언급했듯이 모델은 우리가 전체적인 진실을 알 수 없기 때문에 만드는 것이다. 빅 데이터 운동이란 ① 수많은 데이터에서 잡음을 제거하고 요점을 찾는 것, ② 빈 곳을 채우는 것, 이 두 가지가 우선되어야 한다고 필자는 믿는다. 데이터가 쌓이고 쌓이다 보면 그것이 태산만큼 커 보일 수도 있고, 실제로 그 사이즈가 엄청나기 때문에 빅 데이터란 말도 나온 것이다. 하지만 아무리 데이터가 커져도 우리가 알고 싶은 것에 대한 대답이 거기에 다 있는가, 우리는 과연 모든 사람에 대한 모든 것을 알 수가 있는가 하면 천만의 말씀이다.

"알고 있는 확실한 정보"로 "지금은 알 수 없는 대상"을 예측

데이터는 쌓아 놓기만 한다고 능사가 아니다. 대답을 찾으려면 다시 정리되어야 하는데, 그 정리란 구하는 대답에 대한 대상의 순으로 이루어져야 한다.

만약에 고객을 효과적으로 상대하는 것이 목적이라면, 그 데이터베이스도 "고객 중심"으로 잠깐이나마 재구성이 되어야 그 대상의 우선순위를 정하는 모델을 짤 수 있게 된다. 대상이 상품이라면 상품별로 다시 구성되어야 하고, 지역을 찾는 게 목적이라면 지역별로 뒤집어 봐야 한다. 문제는 그런 식으로 데이터를 보면 없는 부분, 즉 missing data가 엄청나게 발생한다는 것이다.

예를 들어 모바일 기기의 사용에 대한 데이터가 엄청나게 모였다고 치자. 그 기기를 통한 정보검색, 데이터 사용량, 사용 시기와 시간, 구매한 상품, 이동구간 등 정말로 많은 데이터가 모일 수 있다. 하지만 그것을 "사용자 중심"으로 뒤집어 보고, 모든 데이터를 사용자를 묘사하는 방식으로 바꾸어 보면 (모델링에 있어서 필수적인 준비요건이다) 의외로 많은 사용자들의 변수가 빈 곳으로 남아 있는 것을 보게 될 것이다. 데이터 수집의 문제이건, 프라이버시의 문제이건, 사용약관의 차이이건, 수집과정에서의 에러이건, 혹은 더 근본적으로 다른 회사의 서비스를 사용 해서이건, missing data의 발생은 데이터를 다루면서 피할 수 없는 것이다. 극단적인 예로는 어떤 구체적인 질문에 대한 대답을 전체의 1%도 못 미치게 줄 수밖에 없는 경우도 많다.

모델은 그 빈 곳을 효과적으로 채워 넣는 역할도 할 수 있다. 그것이 꼭 정답이 아니고 "아는" 데이터에 기초한 것이 아닐지라도, 질문에 대한 대답을 빈 곳이 없이 확률이나 점수로 표현할 수 있다는 것은 대단한 일이다. 우리가 늘 대하는 일기예보도 그 점수에 대한 사람들의 해석인 것이다. 어떠한 사람이 모바일 기기로 어떤 특정 서비스를 사용할 확률도 바로 그런 것이다. 모델은 "알고 있는 확실한 정보"를 이용해 "지금은 알 수 없는 대상"에 관한 예측을 하는 것이다. 예를 들어, 모바일 기기에 어떤 새로운 기능이 생겼을 때 그것을 이미 사용한 사람들에 대한 데이터를 이용해 아직 사용하지 않는 사람들 중 누가 사용자들의 패턴을 가지고 있는지를 알아보는 것이 통계적 분석이라는 뜻이다. 그리고 그 과정은 우리가 질문을 논리적으로 할 수 있게 하고, 그 질문에 대한 대답을 하기 위해 필요한 데이터를 가공하게 하며, 거기에서 전에 알 수 없었던 새로운 패턴들을 찾게 해주고, 질문에 대한 대답을 수많은 데이터를 함축한 숫자, 즉 점수로 표현해 주며, 그 점수를 가지고 모르는 부분에 대한 빈 곳을 채우고 또 미래에 대한 예측을 가능하게 하는 것이다.

그리고 그 결과는 (마케팅의 예로) 마케터들로 하여금 고객에게 적절한 상품과 서비스를 맞춤형으로 제공할 수 있게 한다. 그 고객에 대해 지금은 별로 아는 것이 없어도 그가 호의적으로 반응할 상품과 서비스를 사람의 직관에만 의존하지 않고 수학적으로 추측하여 모든 이들이 효과적인 구매경험을 할 수 있도록 하는 것이다. 마케팅 이외에도 이러한 데이터의 분석과정은 질문에 대한 대답을 효과적으로 주기 위해선 반드시 필요한 일이다.

쌓아만 놓았다고 데이터가 저절로 대답을 주는 경우는 없으며, 아무리 데이터가 크다 하여도 자세히 들여다 보면 그것은 스위스 치즈만큼이나 구멍이 많이 뚫려있는 것이 현실이다. 그것이 바로 빅 데이터의 핵심은 분석이며, 그 분석의 중심에는 모델링이 있는 이유이다.