7

RFM Data를 넘어서

거래기록은 미래행동 예측에 가장 강력한 변수

미래예측을 위한 분석(Predictive Analytics)에서 가장 강력한 변수들은 거래기록에서 파생되는 데이터(Transaction Data)들이다.

거래기록이 아닌 인적 분포 데이터(demographic data) 등 대상을 묘사하는 데이터로 사람들의 미래의 행동을 예측하는 것도 가능하긴 하다. 간단한 예를 들자면 누가 등산을 좋아하는 사람인지 대상의 성별, 거주지, 직업, 수입, 가족사항 등을 이용하여 그런 예측을 할 수는 있다. 그렇지만, 그보다는 거래기록을 보고 예측하는 것이 훨씬 더 정확하다는 말이다. 즉 등산에 관련된 제품들의 구매기록을 볼 수 있다면 그러한 성향에 관한 예측도 단순한 성향에 관한 것을 넘어서 구매액수, 구매시기, 구매의 빈번한 정도 등으로 더 구체적으로 할 수 있게 된다.

거래기록은 미래행동 예측에 가장 강력한 변수

필자의 멘토께서 과거에 벤처 사업을 함께 할 때 만든 홍보형 표현들 중 가장 기억에 남는 것은 "Past behavior is the best predictor of the future behavior"란 말이다.

번역을 하자면 "과거의 행동은 미래의 행동을 예측하는 데에 가장 강력한 변수다"라는 것인데, 실제로 우리가 거의 15년 전에 만든 예측적 모델을 위한 데이터 플랫폼은 요즘 기준으로 봐도 분명히 빅 데이터의 범주에 쉽게 들어가고도 남을 규모였고, 결과 또한 기대 이상이었다. 그것은 데이터베이스의 규모나 모델의 성능보다도 그 비즈니스 자체가 사람들의 구매기록을 대규모로 집대성한 것이었기 때문이다.

그 데이터베이스는 마케팅하는 사람들의 입장에서 보면 그야말로 꿈의 데이터베이스라고 할 만큼 거의 2,000에 가까운 출처에서 모아놓은 구매기록(transaction data), 각종 demographic data, lifestyle data, 심지어는 promotion & response data(홍보 기록과 거기에 대한 직접적인 반응의 기록)까지 제대로 정리되어 관리되었는데, 필자는 그런 환경에서 하루에 50개에서 많게는 70개 이상의 모델들을 7년 가깝게 검토한 경험이 있다.

그렇게 다양한 데이터가 넘쳐나는 경우 어떤 종류의 데이터가 가장 예측에 유용한 것인가? 이 책의 2장에서 데이터는 크게 나눠 ① Demographic Data, ② Transaction Data, ③ Attitudinal Data가 있다고 정리한 바 있는데, 결론부터 말하자면 단연 Transaction Data, 즉 구매기록에서 파생된 데이터가 소비자의 성향을 위한 예측적 모델에 사용되는 변수들의 80% 이상을 차지하며, 나머지 demographic data나 Census 등 geo-demographic data는 조연하는 입장에서 빈 곳을 채워주고 모델의 예측 곡선을 각이 지지 않고 완만하게 해주는 역할을 한다는 것이다.

그래서 필자는 지금도 당시 늘 얘기하고 다니던 "Past Behavior is the best predictor"란 주장을 아직도 반복한다. 그러한 데이터를 수집하고 관리하는 것이 쉽지 않지만 제대로 쓰기

만 하면 그 과거의 행동의 기록인 Transaction Data가 예측의 가장 강력한 변수임에 의심이 없다는 말이다.

사람의 성향이나 행동 양식은 쉽게 변하지 않는다

그 이유는 무엇인가? 선문답같이 들릴 수도 있겠지만 그것은 사람들의 성향이나 행동양식은 쉽게 변하지 않으며, 변하더라도 빨리 변하지 않기 때문이다.

사람들이 구매하는 것, 보는 것, 클릭하는 것, 탐내는 것은 다음의 행동으로 이어지게 마련이고, 수학적으로 말하자면 어떠한 예측적 곡선 상에 있다는 말이다. 물론 구매는 하지 않고 그냥 예쁜 구두를 보는 것을 좋아하는 경우가 있듯이 항상 그렇다는 것은 아니지만, 데이터만 제대로 모으고 분석하면 일반적인 행태, 즉 누가 아웃도어 스포츠를 좋아할 것이며, 누가 바캉스를 크루즈 선박에서 보낼 것인지, 누가 위험부담이 적은 쪽의 투자를 선호할지, 누가 와인을 즐기는 사람인지, 누가 화초를 기르는 타입인지, 누가 패션에 관심이 많을 것인지 등의 예측이 어느 정도의 정확성을 가지고 가능하다는 말이다. 그리고 여기서 언급하는 Transaction Data가 충분히 마련된다면 이러한 고객 성향에 관한 예측을 넘어, 특정고객의 구매 빈도, 소비량, 심지어는 어떤 시기에 그가 규칙적인 구매를 그만 둘 것인지도 예측할 수 있다.

분명 그런 예측 활동은 데이터가 제대로 모아져야 가능한 것이지만, 요즘처럼 사람들의 모든 활동이 디지털화되어 데이터가 사방에 널려 있는 시대에 제대로 된 분석이 안되고 있다면 그것은 정보가 모자라서 예측이 안 되는 것이 아니라 있는 정보를 제대로 소화하지 못해서 그러한 고등적 분석이 이루어지지 않는다라는 게 더 맞는 말일 것이다. 근접한 미래에 냉장고가 알아서 물품을 주문하는 시절, 즉 IoT(Internet of Things) 시대가 도래한다고 하지만, 그러한 변화는 데이터를 모으는 과정만이 바뀌는 것이고, 축적된 데이터를 제대로 가공하고 분석하는 것은 여전히 반드시 필요한 작업이다.

이 책의 1장에서 언급했듯이 데이터를 제대로 다루려면 ① Collection, ② Refinement, 그리고 ③ Delivery까지 잘해야 하는 것이며, 그 1단계인 수집단계에만 머물러 단지 데이터의 양과 데이터베이스의 크기만 내세워서는 빅 데이터란 표현이 무색해진다. 그리고 그 가공(Refinement)에 관한 것은 Transaction Data를 다룰 때 가장 필요한 일이기도 하다. 왜냐하면 가공이 되지 않은 Transaction Data란 '구매'를 묘사하는 데이터이지 '사람'을 묘사하는 데이터가 아니기 때문이다.

RFM Data는 시작일 뿐이다

그런데 이러한 Transaction Data를 언급하면 많은 마케터들은 마케팅에서 흔히 언급되는 'RFM Data', 즉 Recency, Frequen

cy, Monetary Data를 뜻하는 것이냐고 묻는다.

그것은 분명 거래기록의 중요한 부분을 차지하는 변수들이긴 하지만 RFM Data는 Transaction Data의 전부는 결코 아니다. 이런 데이터를 다룰 때 RFM이란 단지 체크 리스트로 인식되어야지, 그것만 가지고 구매성향을 결정하려고 들면 더 좋은 양질의 분석자료와 예측변수를 간과하게 되는 수가 있다.

누가 얼마나 최근에, 어떠한 빈도로, 얼마나 썼느냐는 분명 중요한 질문이고 그 답은 반드시 나와야 하는 것이지만, 그런 식으로만 데이터를 보면 "지난 12개월 동안 무슨 물건을 샀던 건 당 십만 원 이상의 거래가 있었던 고객의 명단을 뽑아라" 하는 식의 단순한 필터링(filtering)으로 이어지는 수가 많다. 이런 query가 복잡하다고 생각하는 사용자들도 있겠지만 이런 식의 탐색은 단지 일차원적인 것이고, 더 고등적인 분석의 걸림돌이 되는 경우도 많다.

왜 일차원적인가 하면 그것은 애초에 '거래'를 중심으로 한 질문이지 '사람'이나 '대상'을 중심으로 한 질문이 아니기 때문이다. 더 깊은 질문을 하자면 질문 자체가 '고객중심'(Buyer-centric)이어야지 상품, 채널, 구매기록, 회사 및 부서가 중심이어서는 곤란하다. 고객 중심의 사고방식이 결여되어 있으면 대상을 채널이나 상품별로 가둬 두게 되며, 그것은 적절하지 않은 캠페인을 하게 되는 지름길이 된다.

지난 호에는 의사결정이란 종국에는 여러 옵션(혹은 대상)을 '랭킹'별로 정렬하여 보는 것이고, 그래서 데이터베이스가 모델링이나 고객 중심의 분석을 위해 잠시나마 형태가 "고객을 묘사하는 모습"으로 바뀌어야 하는 것이며, 그러한 분석을 위해 최적화된 데이터베이스나 데이터마트를 Analytical Sandbox란 개념으로 소개한 바 있다. 그리고 그러한 데이터의 변환 과정의 중심에는 데이터의 집적과정(Summarization)이 있는 것인데, 그것은 많은 고객의 정보는 바로 그 고객의 거래기록에서 비롯되기 때문이다. 그리고 거래기록이란 정의 자체가 거래를 중심으로 한 것이고, 고객의 행동을 예측하자면 질문과 그 질문에 대한 대답을 줘야 하는 데이터 자체가 고객 중심이 되어야 한다.

즉 모든 변수는 "고객을 묘사하는 형태"를 갖춰야 한다는 것인데, 그것은 같은 내용을 다른 각도에서(정확히 말하자면 90도로 틀어서) 보는 작업이다. 같은 집안을 대문을 통해 보느냐 옆으로 트인 창문으로 보느냐의 차이지만 그 효과는 엄청나다.

단순하게 단 하나만의 거래기록을 보더라도 그 차이는 분명하다. 예를 들어, 거래기록을 보자면 "2016년 7월 15일에 가격이 1만 8천 원인 Wireless 마우스를 1개 구매"로 나타나겠지만, 구매자 중심으로 같은 기록을 보자면 "지난 3개월 이내에 컴퓨터 액세서리 카테고리 안에서의 총 지출액이 2만 원 이하인 구매자"로 나타나게 된다.

요점은 분석을 위한 데이터는 구매자를 묘사해야지 구매 상품이나 기록 자체를 묘사하는 게 아니라는데 있다. 만약에 그 고객이 여러 번의 구매기록이 있다면 그 사람에 대한 묘사는 더욱 구체적이고 다양하게 될 것이다. 예를 들자면 "지난 24개월간 총 지출액이 10만원에서 25

만원 사이이며 총 거래횟수는 13번이며 구매평균액수는 1만원에서 2만원 사이인, 컴퓨터 부품 및 각종 전자제품과 게임기기에 관심이 있는 고객"이라는 구체적 묘사가 개개인 별로 정리되어야 한다는 것인데, 물론 데이터베이스 안에서는 이런 정보가 총 지출액, 총 거래횟수, 평균 지출액 등이 날짜 별로, 또 상품 카테고리 별로 나누어진 여러 변수로 기록되겠지만, 보는 관점을 바꾸면 기록 자체가 다르게 보인다는 것이다.

고객중심의 묘사적 기록 (Buyer-centric Portrait)

구매기록(Transaction Data)을 다룰 때 구매자 중심의 기록이란 데이터를 구매자 별로 집적(Summarization or De-normalization)하는 과정으로부터 시작된다.

구매 기록이란 원래 구매나 거래가 발생할 때마다 새로운 데이터를 새로운 엔트리(entry)로 만들게 되기 마련이며, 데이터베이스 관리자들은 그것을 정상적인 상태(normal state)라고 부른다고 언급한 바 있다. 하지만 의사결정을 위해 가구, 개인, 이메일, 회사, 접촉상대를 중요한 순서대로 '랭킹'을 하는 것이 목적이라면 그 데이터 자체가 그런 레벨로 재정리되어야 한다는 것이 지난 챕터 내용의 핵심이었다.

더욱 구체적으로 Transaction Data의 집적(summarization) 과정을 설명해 보자면, 일단 그것은 반복된 엔트리를 제거하는 작업이 아니라는 것이다. 구매기록에서 버릴 것은 하나도 없는 것이고, 이것은 그런 기록들의 재정리를 하는 과정이다. 한 고객이 여러 번의 구매를 했다면 그 고객의 총 구매건수는 얼마이며 총 지출액은 얼마인가? 단순한 연산과정으로 더하거나 뺄 수 없는 구매 날짜는 한 고객에게 데이터 한 줄만이 주어진다면 어떻게 표현되어야 할 것인가?

일단 시작은 그 고객의 가장 첫 구매일과 가장 마지막 구매일을 정리하는 것이 되겠다. 그런 날짜를 안다면 몇 년 동안이나 고객이었는지, 얼마나 최근에 거래가 있었는지, 더 나아가 구매 간의 평균 날짜수는 얼마나 되는지도 알 수 있게 된다. 물론 이런 변수들은 흔히 말하는 RFM Data의 범주에 포함 되겠지만 훨씬 더 구체적이고 더 나아가 더욱 고등적인 분석에 쓰일 수 있는 '준비된' 자료가 된다.

Data Summarization Example(구매기록이 고객 중심으로 집적화 된 예)

여기에 첨부된 도표는 이러한 집적과정의 전과 후를 보여주는 아주 간단한 사례이다. 도표의 왼쪽은 전형적인 거래 별 Order Table로 Customer ID와 거래의 일련번호인 Order ID와 함께 거래 날짜(Order or Transaction date)와 액수(Dollar Amount)가 기록되어 있다. 만약에 어떤 고객이 여러 번의 거래를 하였다면 거기에 상응하는 기록이 독립된 엔트리로 남게 된다. 실제로는 이러한 테이블에 지불방식(Payment Method), 기타 세금이나 할인 및 쿠폰 액수, 또한 발송이나 배달비용 등도 같이 기록될 것이다.

Order Table

Cust ID	Order #	Order Date	$Amount
000123	100011	2011-05-06	$199.99
000123	100128	2012-08-30	$50.49
000123	103082	2013-12-21	$128.60
003859	100036	2012-06-06	$43.99
003859	101658	2013-01-20	$43.99
003859	102189	2013-04-15	$119.45
003859	106458	2014-02-18	$43.99
004593	104535	2014-07-30	$354.72
016899	107296	2013-07-14	$199.99
019872	102982	2012-09-07	$128.60
019872	103826	2013-04-30	$499.99
019872	109056	2014-03-12	$59.99

Order Summary Table

Cust ID	Order #	$ Total	First Order Data	Last Order Data
001223	3	$379.08	2011-05-06	2013-12-21
003859	4	$251.42	2012-06-06	2014-02-18
004393	1	$354.72	2014-07-30	2014-07-30
016899	1	$199.99	2013-07-14	2013-07-14
019872	3	$688.58	2014-09-07	2014-03-12

오른쪽에는 고객별로 데이터가 압축된 결과가 있는데, 거기에는 고객 한 명 당 여러 산술적 합산의 결과가 나열된 단 한 줄만이 있게 된다. 이런 압축과정만을 위해서라도 데이터베이스에는 일관되고 정확한 개인의 일련번호가 반드시 필요하다. 수집 과정에서의 오류나 고객의 이사, 매장이나 채널 별 정보 수집 방법의 차이 등 다른 여러 가지 이유로 한 개인에게 여러 개의 Customer ID가 부여된다면 개인별 통계나 정보는 정확할 수가 없다. 이메일 주소나 전화번호를 이용한 정보의 통합도 위험한 것이, 많은 사람들은 용도별로 여러 개의 이메일 주소와 번호를 가지고 있는 경우가 흔하기 때문이다. 일단 그런 통합 시스템이 확립이 되었다면 개인별 정보의 통합과 압축은 비교적 쉬운 연산 기능에 불과하며, 그 결과는 더 고등적인 분석과 모델에 크게 도움이 된다.

RFM + P&C

더욱 유용한 정보는 이러한 숫자의 합산 결과를 상품의 종류나 채널, 그리고 다른 중요한 범주적 변수(즉 숫자가 아닌 변수)들과 연관되어 통합될 때 발생하게 된다.

그중에서도 특히 상품(혹은 서비스)과 채널 별로 고객들의 행동양식이 크게 달라지기 때문에, 그 Product와 Channel의 앞 글자를 이미 유명한 RFM에 더하여 "R, F, M, P & C"라고 부르며 체크 리스트로 사용하는 것이 어떨까 하는 것이 필자의 제안이다.

상품, 혹은 상품의 카테고리는 사람들의 행동을 구분하는 중요한 변수이다. 예를 들자면 아무

리 스포츠 용품이나 음악에 관한 제품을 많이 구매하는 사람들도 패션전문 카탈로그에는 전혀 관심을 보이지 않을 수 있다는 것이다. 그러니 특정 개인이 아무리 자주 거래를 한다 하여도 그 모든 종류의 상품에 같이 반응한다고 가정해서는 바람직하지 않고, 구매 기록을 개인을 묘사하는 변수로 압축, 전환하는 과정에서 그러한 상품의 카테고리로 나누어 보는 것이 올바른 방법이다.

상품을 구매하는 채널, 즉 매장, 인터넷, 모바일 기기 및 스마트폰, 전화 등도 사람들의 구매 패턴에 지대한 영향을 미치는데, 예를 들어 인터넷으로 자주 상품을 구매하는 사람도 특정한 종류의 의류나 가구는 반드시 매장에 가서 눈으로 보고 손으로 만져보고 사는 경우가 많다. 같은 카테고리 안의 상품이라도 골프 공은 온라인으로 주문해도 골프클럽은 반드시 만져보고 사는 사람들이 있는 것이다.

그래서 구매채널도 중요한 구분 변수가 되는 것인데, 여기서 말하는 채널이란 "구매자가 상품을 구매할 때 이용한 채널"을 말하는 것이고 "마케터가 그 고객을 상대로 캠페인을 할 때 이용한 채널"과는 분명히 구분되어야 한다. 채널이란 원래 쌍방향으로 인식되어야 하는 법이고, 그 어떤 마케터도 고객을 한 채널에 가두어 놓을 권리란 없다. 즉 어떤 개인이 이메일을 보내도 된다고 허락했다 해도, 그 고객은 반드시 온라인으로만 거래하는 고객이라고 봐서는 곤란하다는 것이다. 이메일을 받아보고 온라인으로 상품을 주문하지 않고 매장에 직접 찾아와 구매했다 해서 그를 오프라인 고객이라고 부르기 곤란한 것과 마찬가지이다.

하지만 많은 회사들은 아예 마케팅 부서를 광고 채널 별로 관리하며 캠페인도 마치 고객들이 그런 채널 안에 갇혀 독립적으로 존재한다는 가정 아래 운영하는 경우가 많은데, 그것은 여기서 말하는 '고객 중심'의 마케팅이나 데이터의 관리의 개념과 크게 동떨어진 것이다. 그런 사고방식은 일관되지 않고 고객의 성향과 관련성이 없는 메시지를 반복적으로 보내게 되는 근본적인 이유이기도 하다.

RFM Data에 기초한 변수 작성 연습

여기서 연습 삼아 실제로 RFM Data에 기초한 변수들을 만들어 보기로 하자.

앞서 첨부된 도표와 같이 기본적인 개인 별 RFM 측정 변수들로 시작해 보자면,

Number of Transactions/Orders (거래건수)
Total Dollar Amount (총 거래액수)
Number of Days (or Weeks) since the Last Transaction (마지막 거래일로부터의 기간)
Number of Days (or Weeks) since the First Transaction (처음 거래일로부터의 기간)

여기서 주목할 점은 날짜가 달력 상의 날짜로 표시되지 않고 어느 시점으로부터(현실적으로 데이터베이스가 업데이트된 날로부터) 날짜의 수나 주간의 수로 표현되어야 한다는 점이다.

그것은 날짜 자체의 의미가 시간이 지나가면서 변하기 때문인데, 예를 들자면 2월의 한 날짜가 4월 기준으로 보는 것과 11월 기준으로 보는 것이 두 달 전이냐 9개월 전 일이냐의 정도로 큰 차이가 있다는 것을 제대로 나타내야 한다는 것이다.
그러한 'Recency' 라는 것이 원래 상대적 개념이고, 그래서 시간에 관한 변수를 만들 때에는 상대적으로 표현해야 옳다. 그렇게 하지 않으면 모델에 날짜가 포함된 경우 사용할 때마다 날짜의 기준을 업데이트해야 하는 번거로움이 생기며, 그런 작업을 간과하면 시간이 가면서 그 모델 자체가 제대로 돌아가지 않게 된다. 반면에 처음부터 시간적 변수들을 상대적인 개념으로 표현하면 그런 업데이트 자체가 필요 없는 것이다.

위에 열거된 기본적인 변수들을 통해 다음과 같은 유용한 개인별 변수들도 쉽게 창출해 낼 수 있다.

Average Dollar Amount per Customer (개인별 평균 거래액수)
Average Dollar Amount per Transaction (거래당 평균 거래액수)
Average Dollar Amount per Year (연 평균 거래액수)
Lifetime Highest Amount per Item (개인별 상품 당 최고 거래액수)
Lifetime Lowest Amount per Transaction (개인별 거래 건 당 최저 거래액수)
Average Number of Days Between Transactions (거래 간 평균 날짜수)
Etc., etc...

이러한 평균치 형태의 변수들은 단순 총액에 비해 모델에 아주 유용한데, 그것들을 모델을 짤 때 만들어 쓰면 된다는 의견도 있지만 정형화되고 일관된 형태의 데이터는 응용과 자동화를 훨씬 효율적으로 만들기 때문에 미리 계획하여 개발하기를 권한다. 그리고 이런 모든 변수들은 고객 하나하나의 개념으로만 볼 것이 아니라, 매 개인의 기록을 매장, 인터넷, 카탈로그, 모빌, 메일 등 온갖 채널 별로 쪼개어 볼 수 있고, 또 상품 카테고리 별로도 반복할 수 있다.

아주 커다란 Excel Spreadsheet을 상상해 보자면 이런 식으로 압축된 테이블은 고객 수만큼의 줄(row)이 있겠고, 어찌 보면 한없이 오른쪽으로 퍼져 나가는 난(column)들이 파생되는 것이다. 나중에 숫자가 아닌 범주적 데이터에 대해 더 깊이 다루겠지만, 여기서는 일단 채널과 상품별로 여기 나열된 총액과 평균치들이 계속 반복된다고 생각하면 이해가 쉬울 것이다.

그 결과로 우리가 흔히 말하는 Recency의 표현도 불특정 채널이 아닌 "온라인에서의 최종 거래일"이란 식이 된다. Frequency에 관한 변수도 "다이어트 관련 상품의 거래 수"라고 말하는 것이지 무작정 총 거래 숫자를 말하는데 그치는 것이 아니다. Monetary의 표현도 "아웃도어 스포츠 관련 상품을 온라인으로 구매한 경우의 평균 액수"라는 식으로 보다 구체적이 될 것이다.

그리고 상품과 채널의 구분에서 멈출 이유가 없으니 Offer Type(캠페인이나 판매 시의 인센티브나 할인, 쿠폰 등), Customer Status(사업체가 분류한 VIP등의 고객분류방식), Payment Method(지불방식), Time Intervals(예: 평생, 12개월, 24개월, 48개월...) 등 다른 여러 변수들과도 조합을 할 수 있겠다. 모든 RFM변수들이 무한정 쪼개질 필요는 물론 없겠지만, 예를 들자면 "지불방식 별 거래 수"는 사람들이 때와 장소, 혹은 구매하는 상품에 따라 꺼내는 신용카드나 지불방식 자체가 바뀌는 경우가 많기 때문에 예측적 모델에서 훌륭한 변수가

될 수 있는 것이다. 미국의 경우 American Express 카드 사용자들의 프로파일은 이미 고급 상품이나 사치품의 구매와 관련이 높다고 밝혀진 바 있다. 다른 지불방식도 특정상품의 구매 성향과 연관관계가 있다는 것을 알아내려면 일단 그러한 변수부터 만드는 것이 올바른 순서다. 이런 모든 작은 변수들은 예측적 모델을 만드는 데 있어서 한 장의 벽돌과 같은 존재이며, 그것들이 제대로 만들어지고 쌓여야 좋은 결과가 나오는 법이다.

물론 너무나 많은 변수들 또한 바람직하지 않기 때문에 어디서 멈추어야 할지도 정해야 하는데, 그 적정선을 제대로 긋는 것도 경험과 지식, 그리고 사전분석이 반드시 필요한 것이니 전문가의 조언을 구하는 것이 바람직하다. 여기서의 요점은 RFM Data란 단지 단순한 세 변수가 결코 아니며, 거기에서 파생되어 나올 수 있는 개인별 변수들이 무궁무진할 수 있다는 것이다. 그리고 이것은 직접적인 거래에 관련되지 않은 행동적 데이터(Behavioral Data), 즉 Click, Page-view, 혹은 minutes의 단위로 표현되는 것들은 아직 손도 대지 않은 단계의 얘기이다.

이와 같이 빅 데이터의 정리는 많은 생각과 준비가 필요한 작업이며, 그래서 모아놓은 데이터의 크기만 자랑할 게 아니라는 것이다. 게다가 성공적인 모델들을 들여다보면 단지 수학적인 요소만의 우수함이 아니라 창조적이고 다양한 데이터 변수들을 먼저 만들고 시작한 것이 성공의 주 이유인 경우가 대부분이다.

The Time Factor (시간적 요소)

그렇게 압축된 데이터가 고등적 분석과 예측적 모델에 반드시 필요하다면, 사용자는 될 수 있는 대로 많은 데이터를 얻기 위해 데이터가 창출되기 시작한 시점으로 매번 되돌아가야만 하는가?

거기에 대한 대답은 불행히도 명확하지가 않다. 그것은 팔고자 하는 상품, 소비자가 그 상품을 보는 관점, 마케터의 목표 등에 따라 답이 달라지기 때문이다.

무한정 과거로 돌아가서 있는 데이터를 전부 압축하는 것('Life-to-date' Summary)도 바람직하지 않은 결과를 가져올 수 있다. 왜냐하면 그런 방식은 "오랫동안 거래해 온 고객"이 "정보가 많이 쌓이지 않은 새로운 고객"보다 항상 더 좋게 보이게 할 수 있기 때문이다. 하지만 현실적으로 볼 때 새로운 고객의 잠재력은 아주 예전에 자주 거래를 했지만 현재에는 거래가 뜸해진 오래된 고객들의 가치보다 훨씬 더 클 수 있다. 그래서 그러한 시간적 선입관이나 편견을 없애기 위해 미국식 표현으로 'Level playing field', 즉 공평한 경기장을 먼저 만드는 것이 바람직하다.

그렇다면 어디에서 그 시간적 선을 긋는 것이 옳은 것인가? 거기에 관한 가이드라인은 파는 물건에 따라 다르다. 자동차나 가구를 파는 경우라면 최소한 4~5년의 기간을 봐야 할 것이다. 그 상품이 회전이 빠른 품목이라면 1~2년만 봐도 무방하겠지만 계절적 차이도 감안하자면 2년 이상의 데이터를 수집하여 압축하고 관리하는 것이 좋겠다. 왜냐하면 상품에 따라 계절

을 탈 수 있는 것이고, 그런 패턴을 월별이나 사분기 별로 검토해 발견하려면 최소한 2~3년의 데이터가 필요하다. 만약에 계절에 따른 성향이 크게 다르고, 또 같은 범주의 상품이라도 골프공과 골프 드라이버처럼 구매간격의 차이가 큰 상품을 같은 매장이나 채널을 통해 판매하고 있다면 12개월, 24개월, 48개월 등으로 같은 데이터를 다르게 쪼개 놓아야 할 것이다.

더 나아가 신용카드나 통신 서비스 등 장기적 관계가 요구되는 상품을 위해 Lifetime Value나 Time-series 등의 모델을 계획하고 있다면 그 시간대는 월별이나 그 이하로 더 잘게 나누어져야 한다. 그리고 그런 고등적 분석으로 넘어가는 시점에서 전문가의 조언이 반드시 필요하며, 그와 더불어 여태까지 누차 강조한 바와 같이 그런 결정에는 사용자와 정책 결정자, 그리고 마케터의 목표와 의견이 충분히 반영되어야 한다. 이 글을 쓰는 가장 중요한 목적들 중 하나가 IT나 통계의 전문가가 아닌 사람들도 데이터와 분석에 관한 기본적인 옵션에 대한 지식을 갖추게 하기 위한 것이다.

Analytical Sandbox

마지막으로, 이러한 데이터의 압축과 변수의 창출의 실제적인 계획과 개발은 누가 해야 하는 일인가?

지난 챕터에서 소개한 'Analytical Sandbox'의 개념은 모든 데이터의 변환과 수정, 카테고리화, 그리고 여기서 다룬 압축 과정이 원활하게 이루어지고, 고등적 분석과 모델의 개발이 그러한 작업에 최적화된 도구와 언어로 신속하고 정확하게 진행되는 분석전문가들을 위한 곳이다. 그리고 그것이 데이터마트나 다른 어떤 이름으로 불리더라도 그 개발은 분석이나 통계전문가에게 맞길 일은 아니다. 그런 분석가들은 그 Analytical Sandbox에서 주로 활동하게 되는 사람들이며 원하는 것을 주문하는 클라이언트이지 그 장소의 설계자나 개발자가 아니라는 뜻이다. 빅 데이터이건 스몰 데이터이건 데이터를 수집, 저장, 관리하고 또 분석을 통해 비즈니스에 관한 대답을 얻는 것이 목적이라면, 분석 전문가들의 팀을 구성함과 동시에 그런 인력들이 제대로 활동할 공간도 따로 계획하여 마련해야 한다.

데이터와 분석 관련 분야에 오랫동안 종사해온 사람으로서 필자가 데이터베이스를 디자인을 할 때 가장 중요한 목적은 분석이나 통계 전문가들에게 "전혀 더 이상의 가공이나 수정이 필요 없는" 양질의 데이터를 그야말로 은쟁반에 얹어서 제공하는 것이다. 그런 환경에서는 분석가들이 타깃의 정의와 비즈니스의 목적을 이루기 위한 방법을 생각하는 데에 대부분의 시간을 쓰게 될 것이다. 질문에 대한 대답은 그러한 노력의 결과인 모델 공식(model algorithm)에서 비롯되는 것이며, 그러한 공식들은 유용한 변수들(variables)로 이루어져 있다. 그래서 건물을 제대로 지으려면 벽돌부터 잘 만들어야 하듯이, 대답을 제대로 얻자면 데이터의 변수들부터 잘 갖추고 시작해야 한다. 유념해야 할 것은, 지금처럼 데이터가 넘쳐나는 시대에는 그 변수들은 간단한 RFM 세가지로 끝날 것은 결코 아니라는 점이다.