

2. 모델링이란?

모델이란 단어는 현실에서 굉장히 자주 그리고 광범위하게 사용하는 말입니다. 패션 모델, 모델 하우스, 프라모델 등등 다양한 곳에서 모델이라는 단어를 사용하고 있죠. ~~모델 (model)~~은 우리 말로 번역하면 '모형'인데 이 말의 사전적 의미는 '어떤 대상을 본떠 만든'입니다.

이를 테면, 모델 하우스는 실제 살게될 집을 본떠 만든 것이고, 프라모델은 실제(혹은 가상의) 로봇이나 기계를 본떠 만든 것이죠. 패션 모델 역시 어떤 대상(수많은 옷을 입을 사람들)을 대표하는 사람이라는 측면에서 비슷한 의미를 갖습니다. 즉, 이들은 모두 공통적으로 1) 실체가 아니라는 점과 2) 실체와 비슷한 대체물이라는 특징을 갖고 있습니다. 그리고 이런 모델을 만드는 작업을 우리는 모델링이라고 부릅니다.

그럼 이제 우리같은 데이터 분석가가 만드는 데이터 모델에 대해서도 비슷한 정의를 내릴 수 있습니다. 데이터 모델은 실제 데이터를 대신하여 본떠 만든 대체물인 것이죠. 그리고 이렇게 어떤 데이터를 이용해서 그 데이터를 대체하는 모형을 만드는 작업이 바로 데이터 모델링입니다. 때론 기계 학습 모델이라는 말을 쓰기도 하는데 이것은 기계 학습 알고리즘을 이용해서 만든 모델이라는 뜻으로 해석할 수 있습니다.

그러면 왜 우리는 모델을 만들까요? 그 이유는 모델을 만들면 실제 대상보다 더 편하게 (혹은 적은 비용으로) 활용할 수 있기 때문입니다. 모델 하우스의 목적은 사람들에게 실제 집을 지어서 보여주지 않아도 최대한 실제 집을 경험할 수 있게 해줍니다. 패션 모델 역시 모든 사람들이 일일이 직접 옷을 입어 보지 않더라도 그 옷을 입었을 때의 모습을 간접적으로 보여주는 역할을 하며, 프라모델은 내가 실제 건담을 만들지 못하더라도 그 비슷한 만족감을 주는 것이죠.

데이터 모델 역시 마찬가지입니다. 데이터 모델은 내가 수많은 데이터를 매번 확인하지 않아도 비슷한 작업을 할 수 있게 해줍니다. 가령, 고양이를 인식하는 모델을 잘 만들면 나는 매번 어떤 이미지 데이터가 고양이를 나타낸 이미지인지를 기존에 내가 알고 있는 모든 고양이 및 다른 대상의 이미지 데이터와 비교해 보지 않더라도 모델만을 이용해서 판단할 수 있겠죠.

결국 이렇게 모델링의 목적을 생각한다면 우리는 어떻게 좋은 모델인지 얘기할 수 있습니다. 좋은 모델은
① 실체가 갖고 있는 속성을 최대한 많이 포함하되 ② 실체보다 최대한 단순해야 합니다. 이런 측면에서 보자면 모델링은 일종의 데이터를 압축하는 작업이라고 볼 수 있습니다. 데이터를 효과적으로 압축하려면 중복되거나 정보로써의 가치가 없는 자료는 최대한 버리고 꼭 필요한 정보만 남기는 것이 필요합니다.