Demand Prediction in Retail

저자 : Maxime C. Cohen, 정리 : 김선중

December 26, 2022

이것은 **Demand Prediction in Retail**이라는 책에 대한 정리이다.

Contents

Chapter 1

Introduction

이 책은 retailer들의 수요예측 전반에 대한 과정을 다룬다. 특히, 데이터 전처리와 분석(data preprocessing and exploration), 정확도를 계산하는 모든 방법들을 포함한 모든 과정들을 수행한다. 또한, 수요예측에 공통적으로 쓰이는 여러 방법들을 제시하고, 여러 유용한 implementation detail을 제공한다. 이때, 각각의 과정들은 적절한 파이썬 코드와 함께 제시된다. 학습을 돕기 위해 이 책에 쓰이는 데이터들을 다운로드받을 수 있다 (1장 3절)

1.1 Motivation

수요예측(demand prediction/forecasting)은 retailer들에게 있어서는 가장 우선순위로서 생각하는 중요한 과제이다. 각 시간단위(일별, 주별 등) 별 수요예측을 정확하게 수행하는 것은 retailer들이 operational decision(e.g. inventory and supply chain management)를 결정하는 데 있어 중요할 수 있고, 이는 생산성과도 직결된다. 최근 IT와 컴퓨팅 능력의 발달로 인해 수요예측에 있어서 상당히 많은 가능성들이 열리고 있다. fast fashion이나 음식배달 사업 등에서 거래 데이터를 얻거나이러한 데이터를 활용해 다방면으로 활용하는 것은 흔한(ubiquitous) 일이 되었다. 2018년을 기준으로, AI와 데이터 분석을 활용하여 창출될 수 있는 retail industry에서의 잠재적인 가치는 연간 1.26T$에 이른다고 알려져있다.

수요예측에 대한 중요한 retail 활용방안으로는 inventory management decision을 개선할 수 있다는 점이 있다. 정확한 예측은 갑작스러운 수요 증가(demand surges)를 예상하고 그에 대비하는 것을 가능케 한다. 특히, 정확하게 수요를 예측하는 것은, 고객만족(customer satisfaction)과 고객유지(customer retention)에 악영향을 미칠 수도 있는 재고소진(stockout)을 예방하는 데 도움이 된다. 동시에, 정확한 수요예측은 retailer들에게 cost prohibitive할 수 있는 excessive stock levels를 완화해준다. 덧붙여서,좋은 수요예측체계를 갖추고 있으면 retailer들로 하여금, preference, substitution patterns, seasonality and elasticities to price discounts 측면에서 고객들을 더 잘 이해할 수 이해할 수 있게 해준다. 마케팅과 판매전략을 돕는데 사용할 수도 있다. 결국, 정확하게 수요를 예측하는 것은 수익을 증가시키고 비용을 감소시키는 것과 직결된다고 말할 수 있다.

최근 몇 년동안, 많은 양의 granular data를 실시간으로 수집하는 능력은 여러 산업에서의 중요한 결정들을 방해했고(?), 이것은 retail 분야도 예외는 아니었다. 거래와 고객에 대한 귀중한 정보들은, 기존의 소매(brick-and-mortar)와 온라인 소매(online retail) 모두에서 수시로 모이고 있다. 전통적인 feature들은 가격, 할인, 판매량, planogram (상품진열도) 등이다. 최근 트렌드는 유동인구 정보(foot-traffic information), 클릭 횟수(clickstream data), SNS 활동 및 체류시간(social-media activities and dwell times)등과 같은 현대적인 데이터 소스들을 수집하는 것이다. 가용한 데이터의 양은 사용할 수 있는 범위를 훨씬 초과한다. 예를 들어, 2017년의 경우, Walmart는 매 시간당 2.5 페타바이트의 데이터를 다룰 수 있다. 이렇게 엄청난 양의 데이터는 데이터 기반의 방법들을 개발할 수 있을만한 기회를 마련해준다. 이러한 방법들은 대부분 머신러닝과 인공지능에 기반을 두고 있고, 현재 많은 retailer들에 의해 다방면으로 사용되고 있다.

이러한 데이터들이 여러 목적을 위해 쓰이기는 하지만, 가장 많이 쓰이는(traditinoal) 활용방식은 수요예측이다. 실제로 학계와 산업계의 많은 연구들이데이터 기반 수요예측 방법을 개발하는 데 집중해오고 있다. 한편, 여러 과정들을 정리해줄

수 있는 practical guide가 부재한 것도 사실이다. 이 책은 이러한 gap을 채워줄 것이고, 코드도 마련되어 있다.

이 책에서는, 수요 예측이라는 주제에 대하여 practical한 관점에서 탐구한다. (이하 책에 대한 설명)

1.2 Dataset

우리가 사용하는 데이터는 온라인 전자제품 retailer로부터 제공되었다. 보안상의 이유로 데이터를 anonymize하였고, 일부 수정을 가하기도 했다. 다음 링크에서 다운로드 받을 수 있다. `https://demandpredictionbook.com`

이 데이터셋은 2016년 10월부터 2018년 9월까지, 100주가 넘는 기간 동안의 tech-gadget e-commenrce retailer의 주별 판매량을 나타내며, 44개 항목에 대한 주별판매량과 관련 정보들을 포함한다. 이러한 주별판매량은 SKUs(stock-keeping units)이라고도 불린다.

이 책에 나온 수요 예측 과정은 서로 다른 다양한 세팅에 적용될 수 있고, e-commerce retailer 에 한정되지도 않는다. 특히, 온라인 소매(e-commerce)와 기존의 소매(brick-and-mortal) 모두에 똑같이 적용될 수 있다. 물론, 생각하고 있는 business setting에 따라서 적용하는 방법과 feature들의 세팅은 조금씩 조정할 필요가 있을 것이다. 우리는 최대한 포괄적이고 특정 주제에 대하여 agnostic한 treatment를 제시하려고 한다.

언급한대로, retailer들은 고객과의 거래와 관련된 데이터들을 매일 수집하고 저장한다. 원본 데이터(raw data)는 때때로 isolated transaction으로 구성되어 있다. 예를 들어, 원본 데이터의 각각의 행들은 특정한 거래(transaction)를 의미하며 거래시각, 가격, 매장, 판매정보, loyalty card information, 제품정보(SKU-related features)(e.g., color, brand, size), 고객정보(과거 구매, 클릭수)와 같은 여러 개의 field와 연관되어 있다. 가장 첫번째 단계는 원본데이터(transaction level data)를 간결한 형태로 통합하는 것이다. 예를 들어, 원본 데이터를 일별 혹은 주별로 통합하는 것이다. 이런 경우, 같은 날(주)에 발생한 모든 거래들은 single observation에 합쳐지게 된다. 두번째는, 제품들(SKU)을 그대로 남길지, 아니면 다른 제품들과 합칠지(e.g. at the brand level or even at the category or subcategory level)를 결정해야 할 것이다. 마지막, 세번째로는 서로다른 여러 매장에서의 데이터를 하나로 합칠지, 아니면 각각의 매장에서의 데이터로 그대로 둘 지를 결정하는 것이다. 데이터를 통합(aggregate)하는 데에는 정답은 없고 가용한 데이터의 양이나 데이터의 variation에 따라 달라진다. 가장 granular한 방법은 데이터의 내재적인 특성들을 보존하는 것이고, 반면에 more aggregated approach는 데이터의 내재적인 특성을 유지하지 못하는 것을 감수하고 데이터의 noise를 제거하는 것이다. 예를 들어, 일별로 데이터를 통합하는 것과 주별로 데이터를 통합하는 것 사이에는 명확한 trade-off가 있다. 주별 통합(weekly aggregation)은 intra-week variation에 대한 정보(e.g. sales are often higher on weekends than weekdays)를 잃을 것이고 대신, 수많은 거래 정보들을 평균내어서 사용하므로 noise는 적어질 것이다. 불행하게도, 두루 적용되는 해답이란 존재하지 않는다. 이 책에서는 week-SKU level로 통합된 경우를 고려할 것이다. 다시 말해, 특정한 SKU에 대한 각각의 행들은 한 주(a specific week)를 나타낼 것이다.

서로 다른 예측 방법을 실행하고 성능을 평가할 수 있도록 되어 있는 코드들이 자세한 설명과 함께 제공된다. jupyter notebook 파일도 함께 제공된다.

이 절(section)과 관련되어 있는 파일은 다음 웹사이트에서 볼 수 있다.

- website : `https://demandpredictionbook.com`
- github : `https://github.com/demand-prediction-guide/public`
- 1/Introduction.ipynb
- data_raw.csv

가상환경 없이도 이 코드는 돌아간다. 혹시 처음 하는 사람이라면 Google Colab을 사용할 수도 있을 것이다. companion website에다가 기본적인 파이썬 설정을 제공해놓았다. 이 책에 담긴 learning experience들을 모두 활용할 수 있으려면, 기본적인 파이썬 코딩 능력을 갖추는 것이 추천된다.

가장 먼저는 데이터셋을 살펴보고 high-level의 탐색 및 분석을 해본다. 먼저 panndas library를 불러오자.

```
import pandas as pd
sales = pd.read_csv('data_raw.csv',parse_dates=['week'])
sale
```

결과는 그림 1.1에 표시되어 있다. 이 데이터셋은 4400개의 행들과 8개의 열들로 이루어져 있다. 각각의 행들은 SKU-week pair에 대응된다. (44 개의 SKU(제품), 100개의 주)

이 데이터셋의 feature들에 대한 설명은 다음과 같다

- Week : 이 데이터셋은 2016년 10월 31일부터 2018년 9월 24일까지의 데이터를 담고 있다. 총 100개 주에 대한 정보를 담고 있고, 각 주들은 그 주 월요일과 동일시(identify)된다.
- SKU : 44개의 제품(SKU)들이 있고, 각각은 1부터 44까지 index되어 있다. 총 행의 개수는 $44 \times 100 = 40,000$개이다.
- Features on the main page : 특정 제품들의 판매를 촉진하기 위해서, 마케팅 팀에서는 홈페이지에 이 제품들이 부각되도록 표시하기도 한다.
- Color : 총 9개의 색깔들이 있다. 그것들은 검정, 금색, 핑크, 파랑, 빨강, 갈색, 초록, 흰색, 보라이다. 두 개의 제품들은 "none"으로 되어 있다. 색깔이 특정되지 않았다는 뜻이다. 다시 말해서, 제품들은 여러 색깔 중 하나이든지, 아니면 특정한 색상을 하나만 가지고 있지 않거나 (예를 들어, 컴퓨터 부품은 색깔을 특정하기 힘들다.) 그것도 아니면 결측치이다.
- Price : 각 주마다 아이템들은 같은 가격으로 고정되어 있다. pricing team은 매 주마다 아이템들의 가격을 변동할 수 있다.
- Vendor : 이 회사는 전자제품 브랜드의 retailer 역할을 한다. vendor는 해당 제품의 브랜드를 나타낸다. 이 데이터셋의 제품들은 열 개의 서로 다른 vendor들을 가지고 있다. 보안 상의 이유로, vender name들은 1부터 10까지의 숫자로 index되어 있다.
- Functionality : 해당 제품의 주된 쓰임새 혹은 해당 제품에 대한 묘사를 나타낸다. 총 12가지의 functionality 값이 존재한다. illustration purpose를 위해 product category의 이름들을 각각의 functionality에 대응시켜 나타냈다. 다시 말해, 12개의 서로 다른 functionality들은 그에 맞는 카테고리를 가지고 있다 : streaming sticks, portable smartphone charger, bluetooth speakers, selfie sticks, bluetooth tracker, mobile phone accesaries, headphones, digital pencils, smartphone stands, virtual reality headset, fitness trackers, flash drives.

마지막으로, weekly_salies 에 대한 설명이다. 이것은 해당 제품에 대한 focal week 동안의 판매량이며, 예측해야 하는 양이다. target variable 혹은 outcome variable이라고 불러도 될 것이다.