# 머신러닝을 이용한 전자상거래 제품의 판매 증감률 예측에 관한 연구

이준하, 김대희

순천향대학교 사물인터넷학과

junha4304@sch.ac.kr, daeheekim@sch.ac.kr

# A Study on the Forecasting System for Sales Growth and Reduction Rate in E-Commerce using Machine Learning

Junha Lee and Daehee Kim

Soonchunhyang University

요 약

본 논문에서는 머신러닝을 활용한 제품별 판매 증감률 예측 시스템을 구현하고 그 활용 방법을 제안한다. 기계의 학습을 위해 오픈 데이터들을 전처리하여 데이터셋을 구축하고, 해당 데이터셋을 활용한 지도학습을 통해 제품별 판매 증감률의 예측이 가능한 회귀 모델을 생성한 후 성능평가를 수행한다.

## Ⅰ. 서론

4 차 산업혁명의 시대가 도래하며 제조와 유통, 물류의 경계가 허물어져 기존 산업들과는 전혀 다른 신 유통 서비스들이 생겨나고 있다. 이 과정에서 많은 유통 채널들이 생겨나고, 신생 채널들의 전자상거래 데이터가 제조사에 제공되지 않아 제조 업체들은 자사 제품의 수요를 예측할 수 없게 되었다. 제조사들은 전자상거래 업체들의 요구에 맞춰 생산량을 조절하는 실정이며, 능동적 마케팅을 펼 기회가 줄어들었다. [1]

이에 본 연구에서는 제조사들의 원활한 자사 제품 수요 예측을 위해 머신러닝과 공개 데이터들을 사용하여 검색량 기반 전자상거래 제품의 판매 증감률을 예측하는 시스템을 제안하고자 한다.

## Ⅱ. 본론

### 2-1. 데이터 전처리

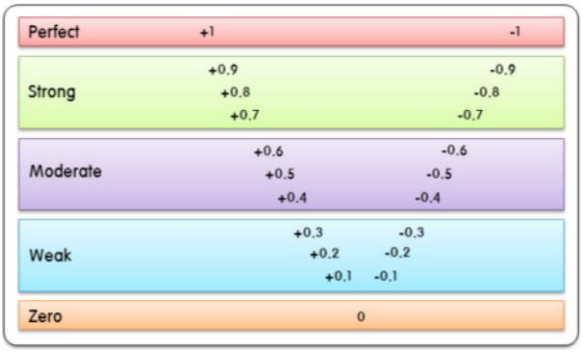
본 연구에서 사용한 데이터는 2016 년부터 2018 년까지 브라질의 전자상거래 주문 100 만개에 대한 주문, 리뷰, 지불방식, 소비자, 상품, 주문아이템, 판매자, 유통주소를 저장한 데이터[2]와 브라질 검색엔진 시장 점유율 89%를 차지하고 있는 구글의 구글 트렌드 데이터이다. [3] 전자를 가공하여 상품 군 별 일자에 따른 판매량과 판매액을 라벨로 가진 데이터셋을 생성하였고, 이를 판매데이터로 사용하였으며, 후자를 가공하여 상품 군의 일별 검색량을 판매데이터에 추가하였다. 이후 대표 카테고리들의 데이터를 통합하고, 검색량에 따른 판매량과 판매액 데이터로 구성하였으며, 그 과정에서 구글 트렌드의 검색량 데이터가 0 부터 100 까지의 상대 값이라는 특성을 고려하여 데이터를 정규화 하였고, 상대적으로 상관도가 낮은 판매량 데이터를 제거하여 [표 1]과 같은 최종 데이터셋을 생성하였다.

| Date | Num | Price | Search |
|---|---|---|---|
| 20170817 | 11 | 1268.15 | 98 |
| 20170818 | 13 | 1099.48 | 89 |
| 20170819 | 10 | 1309.42 | 91 |
| 20170820 | 20 | 1273.21 | 87 |
| 20170821 | 22 | 1782.69 | 91 |
| 20170822 | 20 | 918.71 | 93 |
| 20170823 | 20 | 1538.06 | 89 |
| 20170824 | 11 | 628.23 | 86 |
| 20170825 | 10 | 829.32 | 85 |
| 20170826 | 8 | 830.01 | 82 |
| 20170827 | 14 | 929.68 | 83 |
| 20170828 | 16 | 1377.4 | 91 |
| 20170829 | 20 | 1360.55 | 97 |
| 20170830 | 11 | 1118.38 | 89 |
| 20170831 | 12 | 1218.44 | 90 |
| 20170901 | 25 | 1556.67 | 84 |
| 20170902 | 9 | 533.27 | 80.18182 |

[표 1] 건강, 미용 관련 제품의 최종 데이터셋 (일부)

생성된 최종 데이터셋으로 검색량과 판매액, 판매량 간의 상관 분석을 진행하였다. 상관 분석은 연속 변수로 측정된 두 변수 간의 선형 관계를 분석하는 기법으로 한 변수가 증가하면 다른 한 변수도 선형적으로 증가 혹은 감소하는가를 나타내는 것이다. 상관분석에는 두 변수 사이의 선형적인 관계 정도를 나타내기 위해 상관계수를 사용한다. 일반적으로 1 에 가까울수록 양의 선형관계, -1 에 가까울수록 역의 선형관계가 있다고 판단하며, 0 인 경우 상관관계가 없음을 의미한다. 해석의 차이가 있을

수 있지만 [그림 1]과 같이 0.5 이상의 분석 결과는 양의 상관관계가 있음을 나타내며, 0.8 이상은 강한 상관관계가 있음을 의미한다. [4]




[그림 1] Correlation mean [5]

상관 분석 결과 [표 2]와 같이 검색량과 판매액, 판매량 간에 양의 상관관계가 있음을 알 수 있다.

| 상품군 | 상관 계수 | 상품군 | 상관 계수 |
|---|---|---|---|
| 스포츠 레저 | 0.650293 | 애완동물 제품 | 0.822881 |
| 파티용품 | 0.5764 | 장난감 | 0.681671 |
| 신발 | 0.587769 | 미용&건강 | 0.554486 |
| 속옷, 수영복 | 0.612894 | 아동의류 | 0.746729 |
| 남성의류 | 0.549144 | 꽃 | 0.668519 |

[표 2] 상품군별 검색량과 판매량 간 상관계수

### 2-2. 모델 학습

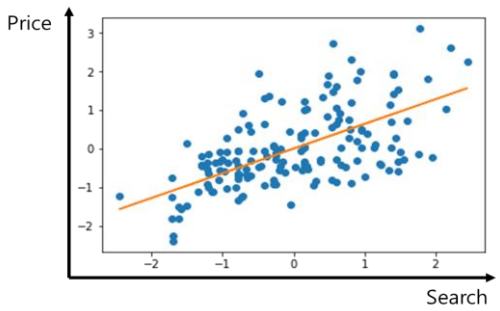
최종 데이터셋의 무작위 20%의 데이터를 검증 셋으로 분리하였고, 나머지 80%의 데이터로 모델을 학습하였다.

학습에는 Linear Regression, Random forest, Gradient Boosting Tree, Neural Network 기법을 사용하였으며, Windows 환경에서 Python 3.7 을 사용하여 분석하였다. 분석 결과 [표 3]에서 알 수 있듯이 각 기법 별 정확도는 크게 차이가 없으며, 특성이 '검색량' 하나인 영향으로 Linear Regression 을 사용한 모델이 가장 정확히 판매액을 예측했다. [6]

| Algorithm | Accuracy |
|---|---|
| Linear Regression | 82.66% |
| Random forest | 79.84% |
| Gradient Boosting Tree | 82.63% |
| Neural Network | 81.25% |

[표 3] 알고리즘 별 정확도

데이터를 통해 생성된 모델의 RMSE 는 0.8059, 검증 셋의 RMSE 는 0.8266 으로 둘의 오차율은 2.5%이므로 오버피팅이나 언더피팅이 없는 학습 모델이라는 것을 알 수 있다. 모델과 데이터를 그래프로 나타내면 [그림 2]과 같은 결과를 얻을 수 있다. 그래프를 분석하면, 기울기는 0.6167, 절편은 -0.0504 로 "판매액 = 0.6167*검색량 - 0.0504"라는 모델을 생성할 수 있었다.




[그림 2] 검색량별 판매액 모델과 데이터 산점도

## Ⅲ. 결론

본 연구에서는 정보화 사회에서 정보격차, 그 중에서도 기존에는 공유되어 왔지만 유통 채널의 다각화와 개인정보의 중요도 상승으로 인해 공유되지 못하는 판매 데이터의 격차를 해소하여 제조사가 능동적 마케팅을 할 수 있는 기반을 만들고자 머신러닝과 오픈 데이터들을 활용하여 전자상거래 제품의 판매 증감률을 예측하는 시스템을 제안하였다. 그 결과 검색량으로 제품의 판매 증감률을 예측할 수 있었고, 검색량의 시계열 데이터 또는 검색엔진 사용자의 연령, 성별 등의 검색 데이터를 활용하여 미래 검색량을 예측한다면 검색 데이터들을 통해 미래의 판매 증감률을 예측할 수 있을 것이다. 본 연구에 최적화된 데이터를 생성하기 위해 데이터 가공이 필요했으며, 최종적으로 사용되는 데이터는 raw 데이터의 50%도 채 되지 않았다. 사용된 데이터조차 타 데이터들과 모집단이 달라 둘 간의 공통 특성인 시간을 기준으로 병합하여도 유의미한 상관 관계를 갖지 않는다는 한계가 있었다. 이러한 한계를 극복하기 위해서는 각 분야의 빅데이터를 확보한 기업들의 자발적 데이터 공유가 필요하고, 정책적 동기부여를 통한 데이터 공유 문화가 정착된다면, 데이터가 핵심 자산이 되는 4 차 산업혁명 시대를 선도할 수 있을 것이다. [7]

## 참 고 문 헌


[1]박성의, "제조 vs 유통, 그 끝에 온디맨드," Mar.2018, ( http://clomag.co.kr/article/2815).

[2]"Brazilian E-Commerce Public Dataset by Olist," 2018, (https://www.kaggle.com/olistbr/brazilian-ecommerce).

[3]"google trend," (https://trends.google.com/trends).

[4] "Correlation and dependence," (https://en.wikipedia.org/wiki/Correlation_and_dependence).

[5] Dancey, C. and Reidy, J." Statistics without Maths for Psychology," 5th edition. pp. 175, Aug.2011

[6] "Machine Learning Algorithm Cheat Sheet for Azure Machine Learning designer," 2020, (https://docs.microsoft.com/azure/machine/learning/algorithm-cheat-sheet).

[7] 최희윤, "4 차 산업혁명과 연구데이터 공유 생태계," 한국일보(기고), Nov.2017, (https://www.hankookilbo.com/News/Read/201711281383267387).