한국빅데이터학회지
제8권 제2호, 2023, pp. 201-222 https://doi.org/10.36498/kbigdt.2023.8.2.201

중소 전자상거래 판매상의 전략적 의사결정을 위한 비즈니스 인텔리전스 설계: 프로모션 전략을 중심으로

Business Intelligence Design for Strategic Decision Making for Small and Midium-size E-Commerce Sellers: Focusing on Promotion Strategy

이성주[1] · 이용현[1] · 김진현[1] · 이강현[2] · 신광섭[2†]

원제로소프트[1], 인천대학교[2]

요 약

온라인 플랫폼을 통한 전자상거래 활성화에 따라 수많은 중소 판매상들은 수익성 향상을 위해 다양한 노력을 기울이고 있다. 이를 위해서는 프로모션이나 이벤트의 범위와 할인 수준, 품목 등에 대한 전략적 의사결정이 매우 중요하다. 본 연구는 중소 전자상거래 판매상들이 효과적인 프로모션 전략을 수립하기 위한 의사결정을 지원하기 위한 도구를 개발하고자 한다. 프로모션의 시행 여부를 판단하기 위해서는 프로모션에 의한 매출 증대 수준을 예측할 수 있어야 한다. 본 연구에서는 다양한 기계학습기법 중 MLP(Multi Layer Perceptron), Gradient Boosting Regression, Random Forest, Linear Regression 모델을 통해 프로모션 시행 후의 매출변화를 예측하기 위한 모델을 개발하였다. 프로모션 데이터가 가진 복잡성과 품목의 특성이 뚜렷한 영향력을 가지는 것으로 확인되었으며, 여러 기법 중 Random Forest 모델과 MLP 모델이 가장 성능이 좋은 것으로 나타났다. 본 연구에서 개발된 방법을 통해 중소 전자상거래 판매상이 시장 변화에 능동적으로 대응하고, 데이터 기반 의사결정을 지원할 수 있을 것이다.

■ 중심어 : 전략적 의사결정, 이커머스, 비즈니스 인텔리전스, 프로모션

Abstract

As the e-Commerce gets increased based on the platform, a lot of small and medium sized sellers have tried to develop the more effective strategies to maximize the profit. In order to increase the profitability, it is quite important to make the strategic decisions based on the range of promotion, discount rate and categories of products. This research aims to develop the business intelligence application which can help sellers of e-Commerce platform make better decisions. To decide whether or not to promote, it is needed to predict the level of increase in sales after promotion. In this research, we have applied the various machine learning algorithm such as MLP(Multi Layer Perceptron), Gradient Boosting Regression, Random Forest, and Linear Regression. Because of the complexity of data structure and distinctive characteristics of product categories, Random Forest and MLP showed the best performance. It seems possible to apply the proposed approach in this research in support the small and medium sized sellers to react on the market changes and to make the reasonable decisions based on the data, not their own experience.

■ Keyword : Strategic Decision Making, e-Commerce, Business Intelligence, Promotion

2023년 11월 27일 접수; 2023년 12월 09일 수정본 접수; 2023년 12월 14일 게재 확정.

* 본 연구는 산업통상자원부의 지원을 받아 연구과제로 수행되었습니다.(연구개발 과제번호:20015463, 과제명: 2세부 온라인 유통과 물류시스템 연계를 위한 표준정보시스템 개발)

† 교신저자 (ksshin@inu.ac.kr)

Ⅰ. 서론

1.1 연구 배경 및 목적

전 세계적으로 전자상거래 시장은 2023년 매출 기준 6,310조 달러로 2022년 대비 10.4% 성장하였고, 전체 소매시장에서의 구성비도 20.8%로 전년 대비 1.1%p 성장할 것으로 예상된다. 특히, 2026년에는 매출 8.148조 달러에 구성비는 24.0%까지 성장할 것으로 전망된다[1]. 국내 시장으로는 2023년 3월 통계청 발표에 따르면 2022년 국내 전자상거래 상품거래액은 154조 6,106억 원으로 2021년 대비 8.3% 성장하였고, 전체 소매판매액 625조 5,518억 원 중 24.7%의 구성비로 국내 시장 역시 소매시장에서 전자상거래의 중요성이 크게 증가하고 있음을 확인할 수 있다[2]. 전자상거래 및 4차 산업혁명 기술의 발전에 따라 다양한 형태의 데이터가 빠른 속도로 증가하며 연결되고 있으며, 데이터 기반의 기업 거래 및 비즈니스 활동 분석은 새로운 가치 창출의 기회를 제공할 수 있다[3]. 특히, 빅데이터 관련 기술이 발전함에 따라 기업의 의사결정 방식이 기업의 경험이나 의사결정자의 직관적 판단에 의존하던 경향이 감소하고 데이터를 활용한 과학적 분석적 의사결정이 증가하고 있다. 이를 위해, 많은 기업들이 비즈니스 인텔리전스(Business Intelligence; BI) 시스템을 도입하여 의사결정을 위한 분석과 예측에 활용하고 있다[4]. 중소 규모의 영세 판매상들은 다양한 플랫폼의 유사 상품정보를 실시간으로 파악하고 합리적 의사결정을 지원하는 시스템이 필요하지만, 그들의 규모 측면의 한계로 대규모 시스템 도입이나 소프트웨어 커스터마이징이 어렵다. 데이터 활용은 기업의 존폐를 결정하는 가장 중요한 요소인데, 대기업은 데이터 확보와 활용에 많은 자원을 투입하는데 비해 중소 규모의 판매상은 상대적으로 규모가 작아 데이터 활용의 격차가 벌어지고 있어, 이를 해소하기 위한 사회적 노력이 필요하다[5]. 따라서, 국내 전자상거래 시장의 안정성을 높이고 중소 규모 판매상의 경쟁력 유지를 위해서는 실시간 수준의 데이터 확보와 이를 기반으로 한 비즈니스 인텔리전스를 활용 가능하도록 소프트웨어 및 인프라의 자원이 필요하다[6]. 중소 규모의 판매상을 파악하기 위해 산업통상자원부의 '온라인 유통과 물류시스템 연계를 위한 표준정보시스템 개발' 사업의 조사에 따르면 <그림 1>과 같이 온라인 마켓을 운영 중인 판매상의 수는 약 25만 개로 추정되며, 이들 중 약 60%는 영세한 중소 규모의 판매상으로 확인된다[7].

〈그림 1〉 물류업무 개선 필요시장 규모 추정

중소 규모의 전자상거래 판매상들은 가격 차별화 중심의 경쟁전략 외 다른 전략을 수립하거나 시행하기 어렵고, 특히 할인행사 같은 행사는 매출 증대의 요인이 될 수는 있으나, 오히려 수익성 악화의 원인이 될 수 있다. 통상적인 데이터 기반의 의사결정인 비즈니스 인텔리전스 기술을 활용하면 시장에 대한 적시성을 높일 수 있지만, 규모가 작은 중소 판매상들은 이를 도입하거나 활용하기 어렵다. 실제로 인터뷰 결과, 대다수의 중소 규모의 판매상들은 마케팅의 4가지 요소인 4P(Product, Place, Price, Promotion) 관점에서 언제, 어느 온라인 마켓에서, 어떤 상품을, 어떤 가격에, 어떻게 프로모션을 해야하는지를 의사결정권자의 감에 의존하거나, 경쟁자보다 더 저렴하게 가격을 내리는 방법을 채택하고 있다. 그 이유로는 데이터 기반의 의사결정에 대한 인

지가 없는 상태, 상품별 판매 데이터의 부족 등으로 판매가를 내리거나, 광고 마케팅의 효율분석보다는 남들이 하면 매출이 오른다는 말을 듣고 진행하는 경우가 대부분이다.

본 연구는 이러한 중소 규모의 전자상거래 판매상들의 비즈니스 인텔리전스를 활용한 경쟁력 강화방안을 제시하려 하며, 이를 통해 데이터 기반 의사결정, 다양한 프로모션 시행, 고객서비스 수준 향상 등을 통해 수익성을 높일 수 있을 것이라는 데에 의미를 두고 있다.

Ⅱ. 선행연구

2.1 수요 예측 관련 선행 연구

이재훈(2021)은 이커머스 시장의 수요 트렌드 예측을 연구하였다. 네이버 랩의 수요 트렌드 데이터를 뉴스벤더 모델의 Critical Radio를 응용하여 DNN, CNN, LSTM 모델을 비교하고 다양한 파라미터 조합을 통해 최적의 모델 도출을 다루었다. 한계점으로는 가공된 형태의 수요트렌드 예측을 수행하였기에 실제 제품 수요트렌드 예측과 비교했을 때 부정확할 수 있다. 따라서, 실제 데이터를 활용하는 것이 필요하고 추가적인 입력 변수 개발이 필요하다고 언급했다[8].

유루루(2021)는 Y의약회사의 수요 예측을 위해 특성변수 16개를 통해 훈련 데이터셋을 구축하고, 목표 변수인 의약품 판매량은 연속적인 수치형 데이터이므로 머신러닝 알고리즘 중 의사결정나무와 랜덤 포레스트, 그리고 인공신공망과 서포트 벡터 머신이 이용되었다. 구축된 모형 중 인공신경망 알고리즘으로 생성된 모형의 오차를 나타내는 MSE, RMSE에서 가장 낮은 값인 설명력을 나타내는 결정계수에서 높은 값인 0.829를 기록하여 가장 적합한 모형으로 선정하였다. 머신러닝 알고리즘을 활용해 의약품 판매량의 예측이 유의한 수준으로 가능하다는 것을 확인할 수 있었으며, 의약품 판매 예측만 아니라 다른 상품 판매 예측에도 적용할 수 있어 판매 예측의 방법론 확장에 기여하였다는 내용을 다루었다. 한계점으로는 데이터 조건이 적다는 점과 영향요인과 영향 방식 등 관련 요인을 광범위하게 발굴해야 할 필요가 있다고 언급했다[9].

박성철(2015)은 VARX 예측 모형의 분석을 위해 VAR 모형에서 그룹화된 5가지 모델을 활용해서 새로운 방법을 제시했다. 자료 분석과정에 있어 모델별 카테고리 구축을 살펴보면 상호 관련성이 높은 카테고리를 그룹화하였고, 구축된 모형을 통해 총 10가지의 회귀 방정식을 얻었다. 쇼핑몰사별 쿠폰율 정책의 영향은 장기적 효과를 나타내는 여행을 보여주고 있지만, 영구적 범주 수요 증대 현상보다 긴밀하게 연관이 되어 있는 것으로 판단이 된다고 언급했다. 한계점으로는 너무 많은 자료와 카테고리별 상세 특성 차이를 상세하게 반영되지 못한 점이라고 언급했다[10].

오지연(2019)은 빅데이터 분석은 재고 관리, 수요 예측, 생산, 가격 결정, 시장 흐름 파악, 소비자 인식 및 트렌드 분석 등에 있어서 기존의 전형적인 데이터 수집 방식으로는 얻을 수 없던 새로운 정보를 창출할 수 있다. 지난 5년간 온라인 쇼핑몰 'A' 업체에서 실시간으로 누적된 데이터를 바탕으로 상품별 판매량을 분석하고 각 변수의 변화에 따른 판매량을 예측하여 판매량 증대와 효율적인 재고 관리를 위한 빅데이터를 활용한 온라인 판매 수요 예측 모델 BAPP(Big data Analysis for sales Prediction Platform)를 제안하였다. 비교 분석한 결과, 반팔 티셔츠는 더울 때뿐만 아니라 평소에도 인 웨어(In Wear)로 많은 사람이 활용하기 때문에 꾸준한 판매량을 보이며 예측과 실제 판매량이 ±1.5%의 오차만 발생했지만, 아우터 웨어(Outer Ware)의 경우에는 예측과 실제 판매량에서 ±8%의 오차가 발생하였다. 그 원인으로는 2016년 10월 29일부터 2017년 3월 10일까지 촛불 집회가 진행되어 아우터 웨어 판

매량이 순간적으로 급증했기 때문이라고 언급했다[11].

이강현, 방선호, 장지영, 신광섭(2022)은 기존의 전통적인 수요 예측 기법의 한계를 극복하기 위해 상품의 분류 체계 혹은 상품이 가진 속성에 기반하여 유사한 상품을 그룹핑하고 각 그룹별 수요 예측 모형을 개발하는 방법들이 제시되었다. 그러나, 같은 상품군에 속하더라도 시장의 판매 수요에는 큰 변동성이 존재하기 때문에 근본적인 한계점을 극복하는 데는 한계가 존재한다. 이러한 한계를 극복하기 위해 기존 군집기반의 수요예측 모형 개발 방법을 개선하여 상품별 판매 패턴과 관련된 새로운 입력 변수를 생성하고, 판매 패턴 기반의 군집화를 수행하였다. 군집별 판매 데이터를 통합하고, 과거 데이터로부터 미래 수요를 예측하기 위한 딥러닝 기반 시계열 수요 예측 모델을 설계하였다고 언급했다[12].

정윤재(2021)는 전통 패션기업들이 D2C 공급망으로 전환함에 따라 기존에 온라인채널을 통해 상품을 판매하던 쇼핑몰 기업들은 경쟁우위를 가져가기 위해 더욱 다양한 신상품을 출시하고 초도 발주량을 높여 고객 서비스율을 높이는 전략을 취하였다. 그 결과, 계절성이 강한 패션상품의 특성상 판매시즌을 놓친 상품들이 불용재고로 전환되면서 재고비용 절감에 대한 방안으로 수요예측이 제시되었다. 기존 문헌연구에서 패션상품 수요예측을 위한 다양한 기법이 제시되었으나, 짧은 주기로 많은 SKU 단위의 신상품을 출시하여 경쟁력을 확보하는 온라인 패션기업 특성상 상품 수요예측에 대한 불확실성이 빠르게 증가하고 있어 통계적 기법의 적용이 어렵다는 것을 확인할 수 있었다. 이에 현업에서는 통계적 기법 대신 판매량이 높은 상품의 특성을 분석 및 추출하여 상품특성에 따라 주문 수량을 조정하는 직관에 의존하는 수요예측 기법을 통해 초도 발주량을 산정하고 있다. 수요예측 정확도를 측정하는 A/F ratio 방법론을 사용하여 온라인 쇼핑몰기업 S사의 수요예측 정확도를 측정하고 초도 발주량을 설정하는 의사결정 프로세스에서 발생하는 수요왜곡요인들을 분석하여 개선된 수요예측 프로세스를 제시한다고 언급했다[13].

김광호, 장병훈, 최황규(2019)는 전력수요예측 데이터의 시계열 특성을 고려하여 딥러닝 기법 중 LSTM 알고리즘을 사용하였고, 전력수요량 등의 입출력 값에 원-핫 인코딩 기법을 적용하는 새로운 시도를 하였다. 성능평가에서 일반 DNN과 본 논문에서 구현된 LSTM 예측모델은 각각 평균 제곱근 오차 4.50, 1.89를 나타내어 LSTM 모델이 예측정확도가 높게 나타났다고 언급했다[14].

김영남, 모혜란, 이지홍, 류상천, 김현(2022)은 Amazon은 비즈니스에 최적화된 수요예측 모델을 구축하였고 이를 AWS에 탑재하고 판매를 하는 또 다른 비즈니스를 창출하고 있다. 이와 달리 C사는 N사, 11사와 같은 다양한 e커머스 주문 플랫폼과 연계하는 협업 비즈니스를 하고 있어 Amazon의 수요예측(매입 상품 기반 주문 예측)과는 다른 수요예측 기술(물류센터 운영에 필요한 수요예측)이 필요하다. e커머스 플랫폼을 통해 주문이 들어오고 해당제품을 6~8시간 내 배송하기 위해서는 자체적으로 해당 제품이 언제 팔릴지를 예측해서 해당 물건의 재고를 준비하고 피킹(picking), 패킹(packing), 출고, 배송 프로세스를 수행해야 하는데, 이를 위한 수요예측 기술은 반드시 필요하다. 그리고 Amazon 비즈니스와 달리 주문에 대한 예측과 주문 플랫폼 할인행사, 제조사 수행 이벤트들의 예측을 통해 풀필먼트센터에서 적시에 포장되어 배송이 이루어질 수 있도록 주문 예측과 이벤트 예측이 가능한 딥러닝&머신러닝 기반의 멀티 조합 수요예측 모델을 만들었다고 언급했다[15].

주종문, 황승국(2004)은 구매자의 구매수요예측을 위한 첫 단계인 클러스터링을 통해 과거 거래내역을 통하여 구매실적이 우수한 구매자와

그렇지 않은 구매자를 구분하여 적절한 대응이나 생산계획을 수립하고, 구매 패턴 분석을 위해 분석방법으로 데이터마이닝 방법을 이용할 수 있다. 또한, 이러한 패턴은 월간의 패턴은 물론 주간, 분기간, 반기간, 연간 등으로 다양하게 분석이 가능하다. 분석의 대상으로 시간에 따른 주문수량뿐만 아니라 기술변화에 대한 요구 등으로 다양하게 분석할 수 있다. 패턴에 대한 분석이 완료되면 구매 물량에 대한 수요예측이 필요하다. 이것은 구매자의 구매수요가 공급자의 능력에 맞춰 평균적으로 이루어지는 것이 아니고 시장의 상황이나 구매자의 전략변화에 따라 한 곳에 집중될 수 있기 때문에 구매자의 요구 변화에 능동적으로 대응하기 위해서는 구매자의 수요를 예측할 필요가 있는 것이라고 언급했다[16].

안세희, 정재윤(2023)은 경진대회 M5 Competition 데이터를 대상으로 Temporal Fusion Transformer(TFT) 모형을 적용하였고, 이 대회에서 우승한 DRFAM 기법과 정확도를 비교하였다. M5 Competition의 Walmart 데이터셋 중 CA_1 매장의 판매량 데이터를 대상으로 성능을 평가하였으며, 매장(store) 수준과 카테고리(category) 수준의 데이터풀(data pool)로 각각 TFT 모형을 학습한 후 예측값을 산술평균하는 방식을 사용하였다. 그 결과, 세 가지 수준의 데이터풀에 대해 직접적 예측모형(direct forecasting)과 재귀적 예측모형(recursive forecasting)으로 총 6개의 LightGBM 모형을 학습하여 산술평균으로 예측하는 DRFAM 기법보다 평균적으로 개선된 예측 정확도를 달성하였다. 이를 통해 TFT 모형이 자기-어텐션 구조를 사용하여 시계열에서 변수와 판매량 간의 관계를 충분히 학습하였음을 알 수 있었다. DRFAM 기법의 직접적 예측모형과 재귀적 예측모형이 28일 간의 예측을 위하여 28회 반복호출을 해야 하지만, TFT 모형은 다중 출력 구조이기 때문에 한 번 모형 호출로 28개의 시계열 예측이 가능하다. 따라서, TFT 기반의 예측모형은 보다 빠르고 정확한 시계열 예측을 제공하여 다양한 분야에 확대 적용할 수 있을 것으로 기대한다고 언급했다[17].

유지현(2019)은 관중 수에 대한 예측을 위한 여러 가지 기존 모델을 검토하고, 그 중에서 효율적인 머신러닝 모델을 제안하였다. 또한 딥러닝과 랜덤포레스트 모델을 혼용하여 일별 관중 수 예측과 비정상적 관중 수 예측에 대한 연구를 진행하였다고 언급했다[18].

정원희, 정다운, 강아영, 구영현, 유성준(2020)은 수요예측 성능 향상을 위해 품목의 수요 패턴을 분석해 4가지 유형으로 구분하고, 각 유형에 적합한 모델을 제안한다. 성능 비교를 위해 사용한 데이터는 대한민국 공군 T-50 단일 기종의 수리 부속 품목의 분기별 수요 데이터이다. 품목의 수요 패턴은 수요발생구간(average demand interval, ADI)과 변동 계수(coefficient of variation, CV)를 사용해 네 가지 smooth, lumpy, intermittent, erratic으로 구분하며 다양한 알고리즘으로 구현한 수요예측 모델의 성능을 비교하기 위해 5가지 기계학습 알고리즘과 2가지 딥러닝 알고리즘을 사용해 수요예측 모델을 구현한다. 기계학습 알고리즘 중에는 앙상블 알고리즘인 random forest regression, adaboost, extra trees regression, bagging, gradient boosting regression과 딥러닝 알고리즘인 long-short term memory (LSTM), deep neural network(DNN)을 사용한다. 수요 패턴에 따른 네 가지 유형에 적합한 모델을 선정해 수요예측 결과를 도출한 경우가 일관된 모델을 사용한 경우에 비해 품목 정확도가 0.61%, 수량 정확도가 0.09 우수한 것을 확인할 수 있다. 제안하는 모델을 적용한다면 전문가의 효율적인 수요 관리가 이루어질 수 있을 것으로 기대한다고 언급했다[19].

정세훈(2021)은 이전 판매 실적을 기반으로 미래 판매량과 재고를 예측하기 위해 머신러닝 알고리즘을 적용하고 각 알고리즘의 성능을 비교

하였다. 테스트 결과, 서포트 벡터 머신이 MSE 수치가 거의 0에 근접해있어 관련 연구 항목에서 예상했던 바와 같이 가장 압도적이고 뛰어난 성능을 보이고 있으나 나머지 회귀분석 모델인 랜덤포레스트와 K-NN은 성능 순위가 경우에 따라 변동되는 추세를 보인다. 그러나 데이터의 양이 적어 전체적으로 성능을 판단하기 어려운 것으로 보인다. 테스트용 샘플을 10개로 변화를 주었을 때 역시 마찬가지로 서포트 벡터 머신의 성능이 월등히 뛰어났고 나머지 두 모델도 이전과 비슷한 양상을 보였다. 최고 MSE 수치의 경우 각각 0.25, 0.55, 0.35, 0.75에 달했으며 큰 격차는 없는 것으로 보인다고 언급했다[20].

조성일(2019)은 2015년 1월부터 2019년 12월까지 60개월 동안의 월 단위 두유제품 판매자료를 시계열 분석과 다항식에 의한 추세분석을 통해 두유제품에 최적인 수요예측 방법을 찾고자 하였다. 분석단위는 두유제품의 용기 구분에 따라 팩, 병, 파우치로 나누었으며, 매출의 70%가 넘는 팩은 다시 블랙팩과 플레인팩, 기타팩 그리고 소용량팩으로 세분화해서 실무에서 즉시 활용할 수 있도록 하였다. 분석단위별로 단순 이동평균법, 가중 이동평균법, 단순 지수평활법, 선형 지수평활법, 윈터스 가법, 윈터스 승법, 다항식에 의한 추세분석(2차, 3차 다항식) 등 8가지 방법을 통해 예측치를 구하고, 예측치와 실제값과의 정확도를 MAPE(Mean Absolute Percentage Error) 값을 이용해서 평가하는 방법으로 가장 정확한 수요예측 방법을 찾도록 하였다. 그 결과 두유제품에는 가중 이동평균법이 거의 모든 품목에서 MAPE값 10% 이하를 보여, 정확한 수요예측방법으로 나타났다고 언급했다[21].

임설아(2021)는 머신러닝 추정모델인 CHAID, 랜덤트리, GenLin(일반화선형), 인공신경망, 선형, 선형회귀분석 등 총 6가지 기법으로 매출을 예측하는 알고리즘을 개발하였다. 분석 결과, 인공신경망 기법이 가장 높은 예측력을 나타내었으며, CHAID 기법이 가장 낮은 예측력을 나타내었다. 각 알고리즘에 따른 주요변수의 영향력 순위가 다르게 나타났는데, 인공신경망, GenLin과 선형회귀는 생방송 여부를 가장 높은 변수로 분석하였고, 랜덤트리, CHAID와 선형은 홈쇼핑 사업자의 명칭(브랜드)을 가장 영향력이 높은 변수로 분석하였다. 결과를 요약하면, 홈쇼핑매출을 예측하는 알고리즘을 통해 도출된 주요변수는 홈쇼핑 브랜드명, 방송시간, 방송속성, 기후・날씨 그리고 상품구성 등 총 5가지이다. 알고리즘에 투입된 주요변수들의 영향력 순위를 분석하여 어떤 투입(독립)변수가 매출증대에 큰 효력을 발휘하는가를 밝혀내었다. 이러한 연구결과는 향후 사업자의 효율적인 운영 방안 고안 및 상품재고관리를 위한 기초자료와 전략수립에 근간이 될 수 있다. 또한, 본 연구의 결과를 통해 종합적이고 거시적인 차원의 논의가 가능할 뿐만 아니라, 학술 및 실무차원에서 실제 적용될 수 있다는 점에 시사점이 있다고 언급했다[22].

2.2 연구의 차별성

본 연구의 선행연구와 차별성은 기존 수요 예측에서 제시되지 않은 중소 전자상거래 판매상의 프로모션을 통한 전략적 의사결정의 지표를 찾는 것에 초점이 맞춰져 있다. 그러나, 안타깝게도 중소 전자상거래 판매상들이 보유한 판매자료에 프로모션에 의한 판매라는 구분이 되어있지 않다. 본 연구에서는 기존 연구들에서 나타난 공통적인 한계점이었던 데이터의 부족을 해결하고 실질적으로 사용 가능한 수준의 입력 변수를 채택하여 중소 전자상거래 판매상의 프로모션 수요 예측에 적합한 머신러닝 모델을 제시하고자 한다. 실험 및 검증은 두 단계로 진행된다. 첫 번째, 단일 기업의 데이터를 바탕으로 설계 가능성과 입력 변수에 대한 검증을 진행하기 위한 사전 설계 모형(Preliminary Model)을 제시하고자

한다. 두 번째, 제시된 사전 설계 모형을 기반으로 중소 전자상거래 판매상의 원천적 문제인 데이터의 양, 질을 확보하기 위해 여러 판매상의 통합 데이터를 통해 중소 판매상 단일 데이터가 아닌 동일 품목 구분을 취급하는 타 판매상의 데이터를 융합하여 활용할 수 있는 통합 데이터 모형을 제시하고자 한다.

Ⅲ. 연구모형 설계

3.1 연구 문제

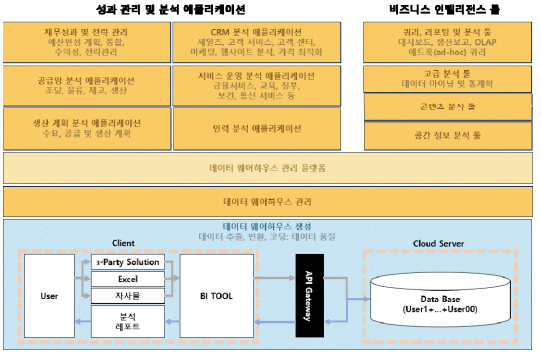
본 연구에서는 중소 전자상거래 판매상이 프로모션 진행할 때 활용할 수 있는 수준의 데이터를 분석하여 효과적인 수요 예측 모델을 제안하려 한다. 분석을 위한 데이터는 전자상거래 판매상 특성을 볼 때, 직접 제조를 하거나 자사 브랜드의 제품을 유통하는 중소 판매상은 매우 적다. 따라서, 데이터 선정의 조건은 첫 번째, 인지도가 없는 비 브랜드 제품군은 프로모션의 효율과 연관이 없을 수 있기에 인지도가 있는 제품군이어야 하며, 계절 지수의 영향을 받지 않는 제품군을 선정하는 것이 중요하다. 두 번째, 평소 판매량이 많은 기업의 데이터는 중소 판매상과 규모적으로 맞지 않기에 디지털 트윈에 적합한 프로모션 이전 상시 판매 수량이 낮은 판매상 데이터가 필요하다. 세 번째, 데이터에 프로모션 진행 유무가 중요하다. 네 번째 품목의 분류가 필요하다. 상기 조건을 모두 만족하는 제품군과 판매량을 보유한 중소 판매상의 데이터가 필요하였다. 그러나, 가장 중요한 조건인 '세 번째 조건'인 프로모션의 진행 여부가 포함된 데이터를 확보하는 것은 상당히 어려운 상황이다. 이는 실제 프로모션 진행 여부를 별도로 관리하지 않기 때문이며, 단기간동안 일회성으로 진행하는 이벤트에 대한 기록과 추적을 진행하지 않기 때문이다. 이러한 한계점을 극복하기 위해 프로모션 수행 결과라고 판단될 정도로 매출액이 급증하는 구간을 기준으로 분석을 위한 데이터를 확보하였다. F사의 최근 1년간 판매 데이터를 바탕으로 데이터를 정제한 결과는 <표 1>과 같다.

〈표 1〉 F사 원시 데이터

명칭	정의
Date	날짜(YYYY-MM-DD)
Style Code	F사의 품목 코드
D-AVG(30)	전일 기준 30일 평균 매출액
D-1	전일 매출액
D-Day	당일 매출액
Sex	남성, 여성, 공용 분류
Category	속옷, 신발, 의류, 용품 분류

확보한 데이터는 중소 전자상거래 판매상으로 F사의 국내 전자상거래 판매 중계 플랫폼(G사)에 지난 1년(2021년)간 판매 데이터이다. 판매 일자별 품목 코드의 당일 매출액, 전일 매출액, 품목코드의 성별(남성용, 여성용, 남녀공용), 품목의 카테고리 정보(속옷, 신발, 의류, 용품)가 수록되어 있다. 확보한 데이터를 기반으로 선정한 연구 모델은 복잡한 데이터 구조를 다룰 때 강점이 있는 '다층 퍼셉트론(MLP: Multi Layer Perceptron' 그리고, 전통적인 통계 및 머신러닝 작업에 자주 사용되는 '랜덤 포레스트(Random Forest)', '선형 회귀 분석(Linear Regression)', 마지막으로 약한 예측 모델을 순차적으로 결합하여 강한 예측 모델을 만드는 앙상블 기법인 '그래디언트 부스팅(Gradient Boosting)'이다. 총 4가지 모델을 비교 분석하여 예측 결과에 대한 신뢰성이 높은 사전 설계 모형을 제시하고자 한다. 또한, 본 연구에서 제시된 사전 설계 모형을 기반으로 중소 전자상거래 판매상이 타 사의 데이터를 활용할 수 있는 방안을 모색하고, 데이터 부족으로 인한 수요 예측의 어려움을 해소하고자 한다. 중소 판매상의 데이터를 통합하고 활용할 수 있는 예상

아키텍처는 <그림 2>와 같다.



〈그림 2〉 통합 데이터 활용 수요 예측(아키텍처)

중소 판매상이 보유한 프로모션 판매 기록과 유사 규모의 타사 데이터를 결합하여 본 연구에서 선정된 사전 설계 모형으로 학습시키고, 학습된 데이터를 Data Base에 수록되며, 수록된 Data Base는 다른 중소 판매상의 수요 예측 자료에 사용되는 순환 구조의 통합 데이터 활용을 통한 수요 예측을 할 수 있는 통합 데이터 모형을 제시하고자 한다.

3.2. 연구 모델

3.2.1 MLP 모델

본 연구에서는 수요 예측 분석을 위해 MLP 모델을 채택하였다. MLP는 인공신경망의 일종으로, 복수 레이어와 각 레이어에 존재하는 다수의 뉴런으로 구성된다. 각 레이어는 이전 레이어의 출력을 입력으로 사용하는 심층 신경망이며, 비선형 문제 및 복잡한 데이터 패턴학습에 적합하다. MLP의 구조는 <표 2>와 같이 정의된다.

$h^{(1)}$는 레이어l의 출력 벡터이며, $W^{(l)}$은 각 레이어 l의 가중치 행렬, $b^{(l)}$은 바이어스 벡터이다. f는 활성화 함수(ReLU, 시그모이드, 하이퍼볼릭 탄젠트)이며, 이전 레이어 l-1의 출력은 $h^{(l-1)}$, $h^{(0)}$은 입력 레이어일 때, 각 층 l에 대해 출력 $h^{(1)}$은 $h^{(1)} = f(W^{(l)}h^{(l-1)} + b^{(l)})$로 계산된다. MLP에서

〈표 2〉 MLP 모델의 구조

항목	정의
입력 레이어 (Input Layer)	원시 데이터의 특징의 수에 해당하는 노드 수를 가진다.
은닉 레이어 (Hidden Layers)	여러 층의 신경망을 통해 특징과 복잡한 패턴을 추출하고 학습한다. 각 뉴런은 가중치(W)와 편향(b)를 통해 입력 신호를 처리하고 활성화 함수를 사용하여 출력 신호를 생성한다.
출력 레이어 (Output Layer)	예측 또는 분류를 위한 결과를 출력한다.

활성화 함수는 비선형 패턴을 학습하기 위해 사용된다. 또한, 훈련 과정에서 예측 값과 실제 값의 차이를 나타내는 손실 함수를 최소화하기 위해 가중치와 바이어스를 조정하고, 최적화를 위해 역전파 알고리즘과 최적화 방법(Adam, SGD)을 사용하여 가중치를 갱신시킨다. 그러나, 하이퍼파라미터의 선택이 중요하며, 과적합(Overfitting)의 위험이 있다. 그러므로, 교차 검증과 정규화 기법이 중요하다.

본 연구에서는 특징 추출과 복잡한 비선형 관계 및 고차원 데이터를 효과적으로 학습할 수 있는 MLP 모델을 사용하여 성능 평가를 수행하였다.

3.2.2 그래디언트 부스팅 모델(Gradient Boosting)

본 연구에서는 수요 예측 분석을 위해 머신러닝에서 널리 사용되는 앙상블 기법인 그래디언트 부스팅 모델을 채택하였다. 그래디언트 부스팅은 약한 결정 트리(Decision Tree)를 순차적으로 결합하여 강한 예측 모델을 만드는 방식으로 주어진 학습 데이터셋에 연속적인 최적화를 수행하여 모델의 성능을 향상시킨다. 이전 학습의 실제 값과 예측 값 사이의 차이에 대해 새롭게 학습 훈련을 시키는 방식이다. 즉, 함수 공간에서의 최적화 문제를 해결하는 방법으로, 주어진 손실 함수 $L(y,F(x))$에 대해 반복적인 최적화 수행을 통해 손실 함수 L을 최소화하는 예측 모델 $F(x)$을 찾는 것이다. 그래디언트 부스팅의 각 단

계에서 현재 모델 $F_t(x)$에 대한 손실 기울기를 계산하고, 이 기울기를 최소화하는 방향으로 약한 결정 트리 $h_{t+1}(x)$를 훈련시켜 새로운 모델 $F_{t+1}(x)$는 $F_t(x)+\nu \cdot h_{t+1}(x)$로 갱신된다. ν는 Learning rate로 각 단계에서 약한 결정 트리의 기여도를 조정한다. 이 과정은 미리 정의된 반복 횟수에 도달하거나, 추가 학습이 손실 함수를 감소시키지 못할 때까지 반복된다.

본 연구에서는 데이터셋에서 수요 예측을 하기 위해 가장 중요한 조건이라고 가정했던 '프로모션의 진행 유, 무'의 부재로 '품목의 분류별' 상관성을 앙상블 기법을 활용하여 예측하고자 한다.

3.2.3 랜덤 포레스트 모델(Random Forest)

본 연구에서는 수요 예측 분석을 위해 랜덤 포레스트 모델을 채택하였다. 랜덤 포레스트의 장점으로는 결정 트리(Decision Tree)의 집합으로 구성되기에 분석하고자 하는 상품군의 특성을 평가할 수 있기에 각기 다른 트리의 예측을 결합하여 단일 결정 트리보다 높은 정확성 및 안전성을 가질 수 있다. 또한, 오버 피팅에 강하고 다양한 데이터셋에 유용하게 사용할 수 있다. 랜덤 포레스트는 3가지 단계로 구성된다. 첫 번째 단계는 부트스트랩 샘플링으로 원본 데이터를 반복적으로 샘플링하여 여러개의 부트스트랩 데이터셋을 생성하고 B번의 반복을 수행한다. 두 번째 단계는 결정 트리 구축이다. 데이터의 다양한 특성을 기반으로 회귀와 분류 규칙을 형성하며, 노드 분할 시 무작위로 선택된 특성 부분집합을 사용하여 오버피팅을 감소시킨다. 세 번째 단계는 결정 트리의 앙상블은 개별 결정 트리의 예측을 집계하여 최종 예측을 결정한다. 분류 문제의 경우 다수결 투표 방식을 채택하고 회귀 문제의 경우 평균을 채택한다. b번째 결정트리의 예측값은 $T_b(x)$이고, B는 결정 트리의 수인 경우, 입력 x에 대한 예측값은 $\frac{1}{B}\sum_{b=1}^{B} T_b(x)$로 계산된다. 본 연구에서는 수요 예측의 정확도를 높이기 위해 랜덤 포레스트 모델을 방법론으로 채택하였다.

3.2.4 선형 회귀 분석 모델(Linear Regression)

본 연구에서는 수요 예측 분석을 위해 선형 회귀 분석 모델을 채택하였다. 선형 회귀 분석은 통계학에서 가장 기본적이고 많이 사용되는 예측 기법으로, 하나 또는 다수 독립 변수 $(X_1, X_2, ..., X_n)$와 종속 변수(Y) 간의 선형 관계를 나타낸다. 모델의 파라미터$(\beta_0, \beta_1, ..., \beta_n)$ 추정 방법은 관측된 데이터와 예측된 데이터 간의 차이의 제곱합을 최소화한다. 즉, 선형 회귀 방법론은 주어진 데이터에 가장 잘 맞는 선을 찾기 위해 실제 관측값과 예측값 간의 차이의 제곱을 최소화하여 예측한다.

또한, 선형 회귀 분석 모델이 유효하기 위해 다음 <표 3>의 가정사항에 크게 의존한다.

〈표 3〉 선형 회귀 분석 모델의 가정사항

항목	정의
선형성	독립 변수와 종속 변수 간에는 선형 관계가 존재해야 한다.
독립성	오차 항은 서로 독립적이어야 한다.
등분산성	모든 독립 변수 값에 대해 오차 항의 분산이 일정해야 한다.
정규 분포	오차 항은 정규 분포를 따라야 한다.

선형 회귀 분석 모델의 성능 평가는 결정 계수(R^2) 및 조정된 R^2, 평균 제곱 오차(MSE), 평균 절대 오차(MAE)의 지표를 사용하여 수행되며, 회귀 계수의 통계적 유의성은 t-통계량과 p-값을 통해 평가된다.

본 연구에서 채택한 선형 회귀 분석 모델은 수요 예측의 정확도를 높이기 위해 데이터의 특성과 가정을 고려하여 진행되었다.

Ⅳ. 실험 및 결과분석

4.1 사전 설계 모형

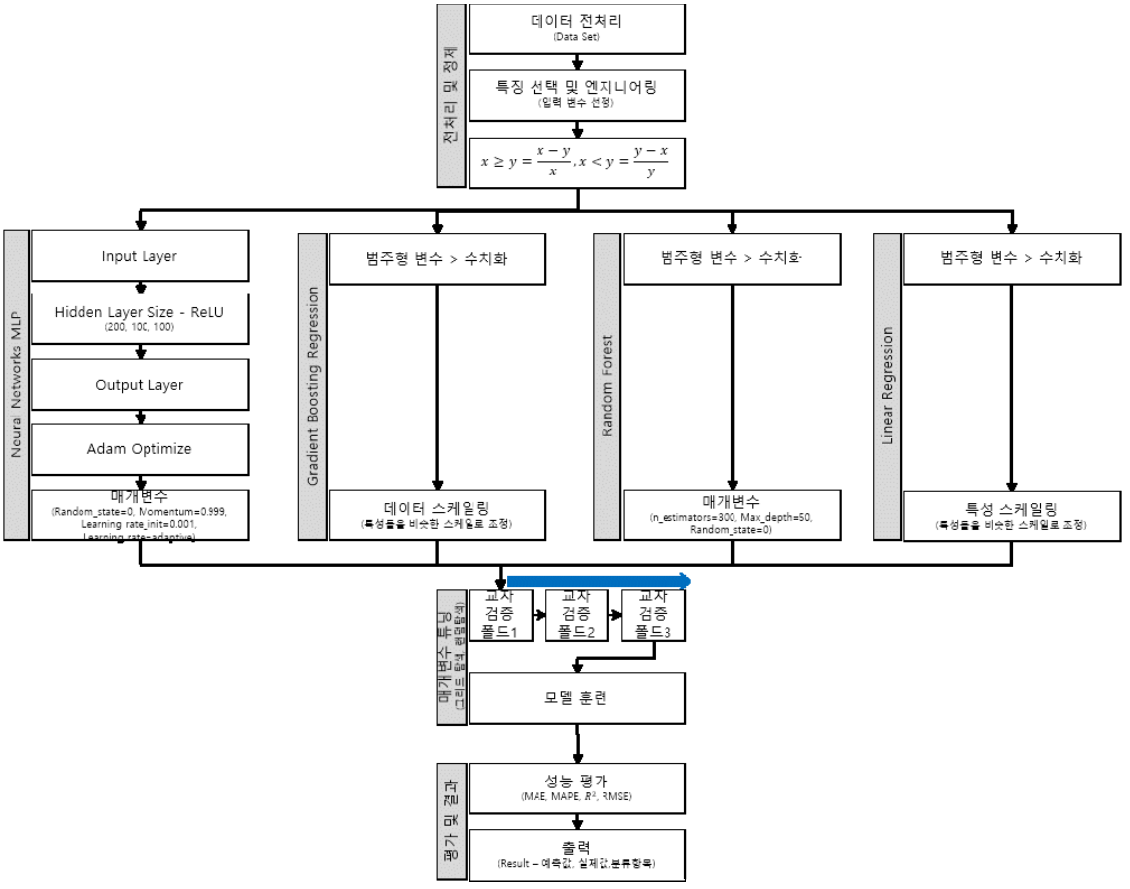
4.1.1 실험 프레임워크

본 연구에서는 중소 전자상거래 판매상인 'F사'의 데이터(단일 기업)를 바탕으로 프로모션 진행할 때 활용할 수 있는 수준의 데이터를 분석하고, 효과적인 수요 예측 모델을 제안하기 위해 분석 모델로 'MLP', '그래디언트 부스팅', '랜덤 포레스트', '선형 회귀 분석'을 채택하였다. 실험 모형의 프레임워크는 <그림 3>과 같다.

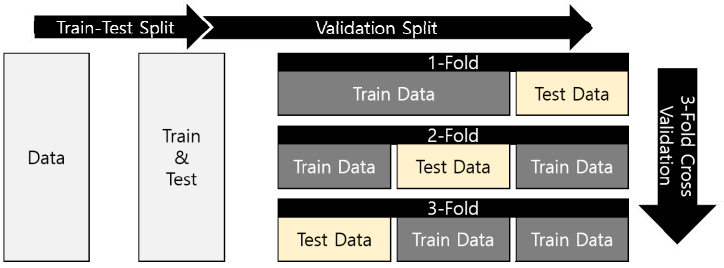
데이터셋의 '품목 분류'는 입력 변수로 '그래디언트 부스팅', '랜덤 포레스트', '선형 회귀 분석'을 위해 범주형 변수를 수치형 변수로 변환하여 분석을 진행하였고, 선형 회귀 분석에서는 True, False 값으로 분석을 진행하였다. 또한, 모델 평가에 객관성을 부여하기 위해 <그림 4>와 같이 교차 검증(Cross Validation)을 공통적으로 적용하였고, 전체 데이터를 3조각(Folds), 3회(Splits) 분리하고 각 분리 당 1조각을 나머지 2조각의 훈련 데이터로 사용해 학습을 총 3회 진행하였다.

4.1.2 데이터 전처리 및 가정사항

본 연구에서 분석할 데이터 중 입력 변수인 품목 분류 값 중 성별에 해당하는 '남성', '여성', '남녀공용'과 카테고리에 해당하는 '속옷', '신발', '의류', '용품'에 대해 수치형 변수로 전처리를 진행하였다. 실험에 영향을 미치지 않는 'Date'



〈그림 3〉 실험 모형 프레임워크



〈그림 4〉 교차검증(Cross Validation) 구조

와 'Style Code'는 제거하였다. 가장 큰 요인으로 작용할 '프로모션 유, 무'가 원시 데이터상에 없기에 'D-1'의 값이 'D-Day' 값보다 큰 데이터와 'D-Day'의 값이 'D-1' 값보다 3배수 미만인 데이터는 제거하고, 3배수 이상의 자료를 프로모션 자료로 채택하였다. 가정사항을 반영한 데이터 전처리 및 정제 후 실험 데이터셋은 <표 4>와 같다.

4.1.3 모델별 실험

본 연구에서는 프로모션의 수요 예측을 목표로 사전 설계 모형을 선정하기 위해 실험을 진행하였다. 사전 설계 모형은 단일 기업의 데이터를 바탕으로 설계의 가능성과 입력 변수에 대한 검증을 위해 진행되었다. MLP, 그래디언트 부스팅, 랜덤 포레스트, 선형 회귀 분석 총 4개의 모델을 통해 MAPE, R^2(결정 계수), 성능 정확도 분포(실제 값의 선과 예측 값의 분포)를 구하고 편차가 적으며 높은 상관성을 나타내는 모델을 선정하고자 한다.

4.1.3.1. MLP 모델

MLP 모델의 결과는 MAPE 0.14, R^2 0.93로 나타났다. 또한, 표본 중 가장 큰 오차는 <표 5>의 색상 표기와 같이 0.61로 나타났다.

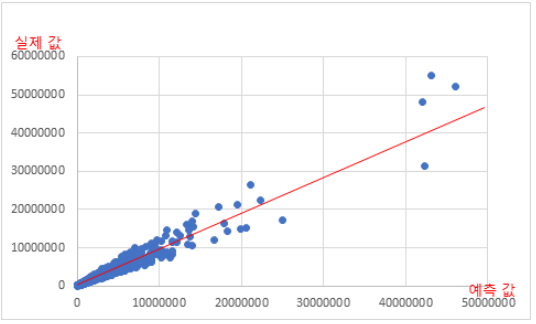
〈표 5〉 MLP 모델 결과

No.	예측 값	실제 값	오차
1	244,976.3	245,000	0.00009
2	2,108,752.7	2,108,000	0.00035
3	41,974.0	42,000	0.00061
⋮			
1765	803,203.2	472,000	0.41235
1766	801,859.8	354,000	0.55852
1767	900,928.0	351,000	0.61040

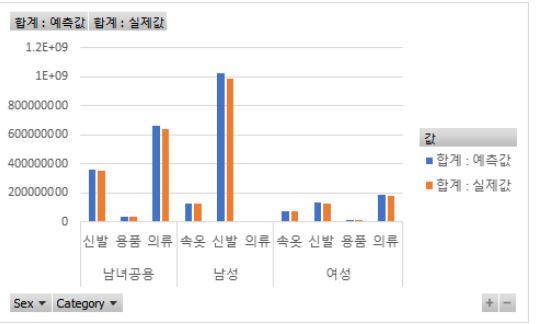
'예측 값'과 '실제 값'을 통해 얻은 성능의 정확도는 <그림 5>와 같이 표본이 적은 구간일수록 정확도가 낮아지는 것을 확인할 수 있었다.

〈표 4〉 실험 데이터셋

No.	D-AVG (30)	D-1	D-Day	Sex 남성	Sex 여성	Sex 남녀 공용	Cate 속옷	Cate 신발	Cate 의류	Cate 용품
1	119,966	472,000	3,009,000	0	1	0	0	1	0	0
2	11,700	39,000	312,000	0	1	0	0	1	0	0
3	52,266	196,000	1,372,000	0	1	0	0	1	0	0
⋮										
1765	35,676	194,600	1,167,600	0	1	0	0	0	1	0
1766	13,533	377,000	2,494,000	0	1	0	0	0	0	1
1767	5,933	8,900	53,400	0	0	1	1	0	0	0



〈그림 5〉 성능 정확도(MLP)분포



〈그림 6〉 MLP 모델 결과(분류별) 그래프

가장 오차가 큰 구간을 확인하기 위해 분류별 (*Sex, -Category) 그룹화하여 '예측 값'과 '실제 값'의 오차를 확인했다. 가장 오차가 큰 분류명은 <표 6>과 같이 *남성-의류(12%)으로 나타났다.

〈표 6〉 MLP 모델 결과(분류별)

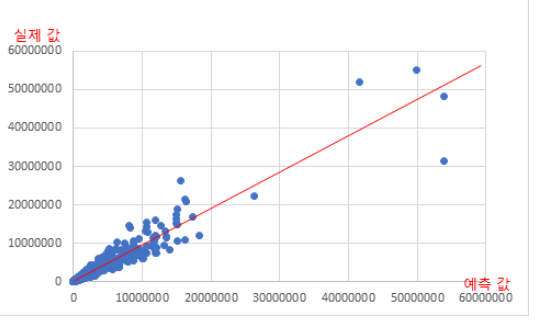
분류명	예측 값	실제 값	오차
* 남녀 공용	1,057,707,875	1,019,718,900	4%
-신발	361,664,232	349,625,200	3%
-용품	35,308,091	34,294,000	3%
-의류	660,735,552	635,799,700	4%
* 남성	1,154,127,785	1,110,381,500	4%
-속옷	128,991,773	122,855,600	5%
-신발	1,022,444,380	985,126,600	4%
-의류	2,691,632	2,399,300	12%
* 여성	409,674,466	387,885,300	6%
-속옷	76,418,135	70,868,700	8%
-신발	136,459,359	128,905,500	6%
-용품	11,067,988	10,310,900	7%
-의류	185,728,984	177,800,200	4%
총 합계	2,621,510,126	2,517,958,700	4%

4.1.3.2 그래디언트 부스팅 모델

그래디언트 부스팅 모델의 결과는 MAPE 0.20, R^2 0.90로 나타났다. 표본 중 가장 큰 오차는 <표 7>의 색상 표기와 같이 0.50으로 나타났다.

〈표 7〉 그래디언트 부스팅 모델 결과

No.	예측 값	실제 값	오차
1	234,000	234,000	0.00000
2	414,000	414,000	0.00000
3	354,000	354,000	0.00000
⋮			
1765	2,451,000	1,309,800	0.46561
1766	631,994.7	331000	0.47626
1767	3,081,000	1,519,000	0.50698



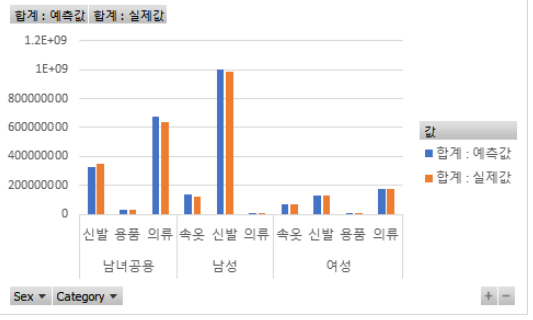
〈그림 7〉 성능 정확도(Gradient Boosting)분포

〈표 8〉 그래디언트 부스팅 모델 결과(분류별)

분류명	예측 값	실제 값	오차
* 남녀 공용	1,038,436,645	1,019,718,900	2%
-신발	330,059,216	349,625,200	-6%
-용품	33,521,557	34,294,000	-2%
-의류	674,855,872	635,799,700	6%
* 남성	1,141,421,311	1,110,381,500	3%
-속옷	134,451,293	122,855,600	9%
-신발	1,004,151,202	985,126,600	2%
-의류	2,818,816	2,399,300	17%
* 여성	391,709,142	387,885,300	1%
-속옷	71,700,585	70,868,700	1%
-신발	129,996,692	128,905,500	1%
-용품	11,762,275	10,310,900	14%
-의류	178,249,589	177,800,200	0%
총 합계	2,571,567,098	2,517,958,700	2%

‘예측 값’과 ‘실제 값’을 통해 얻은 성능의 정확도는 <그림 7>과 같이 실제 값과 예측 값이 10,000,000인 구간 이후부터 MLP 모델에 비해 성능이 좋지 않았다.

<표 8>과 같이 분류별(*Sex, -Category) ‘예측 값’과 ‘실제 값’의 오차가 가장 큰 분류명은 3가지 분류로 *남성-의류(17%), *여성-용품(14%), *남성-속옷(9%)순으로 나타났다.



〈그림 8〉 그래디언트 부스팅 모델 결과(분류별) 그래프

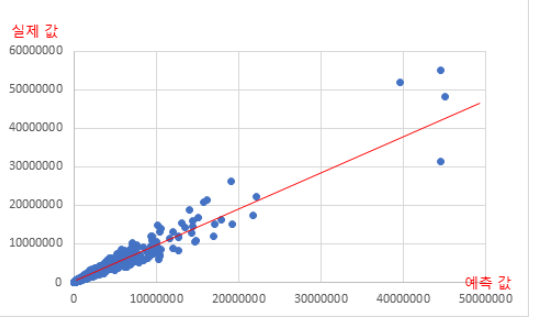
4.1.3.3. 랜덤 포레스트 모델

랜덤 포레스트 모델의 경우 결과는 MAPE 0.16, R^2 0.94로 나타났다. 표본 중 가장 큰 오차는 <표 9>의 색상 표기와 같이 0.42로 나타났다.

〈표 9〉 랜덤 포레스트 모델 결과

No.	예측 값	실제 값	오차
1	342,964	343,000	0.00010
2	693,105.3	693,000	0.00015
3	2,204,160.6	2,205,000	0.00038
⋮			
1765	252,733	423,000	0.40252
1766	10,345,997.6	6,084,000	0.41194
1767	31,509.3	18,000	0.42874

‘예측 값’과 ‘실제 값’을 통해 얻은 성능의 정확도는 <그림 9>와 같이 MLP 모델과 성능이 비슷한 결과가 나타났다.

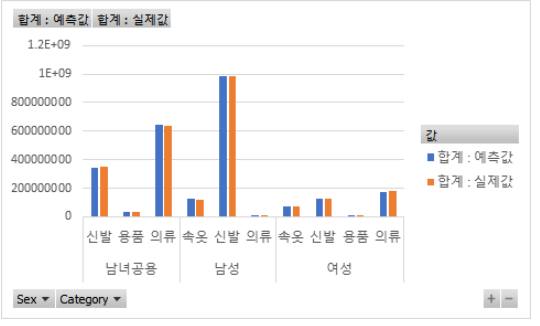


〈그림 9〉 성능 정확도(Random Forest)분포

<표 10>과 같이 분류별(*Sex, -Category) ‘예측 값 ‘과’실제 값 ‘의 오차가 가장 큰 분류명은 *남성-의류(14%)로 가장 낮게 나타났다.

〈표 10〉 랜덤 포레스트 모델 결과(분류별)

분류명	예측 값	실제 값	오차
* 남녀 공용	1,024,778,020	1,019,718,900	0%
-신발	342,437,667	349,625,200	-2%
-용품	33,386,006	34,294,000	-3%
-의류	648,954,345	635,799,700	2%
* 남성	1,119,937,025	1,110,381,500	1%
-속옷	129,803,290	122,855,600	6%
-신발	987,396,340	985,126,600	0%
-의류	2,737,393	2,399,300	14%
* 여성	390,093,604	387,885,300	1%
-속옷	73,015,385	70,868,700	3%
-신발	130,040,450	128,905,500	1%
-용품	10,866,968	10,310,900	5%
-의류	176,170,799	177,800,200	-1%
총 합계	2,534,808,649	2,517,958,700	1%



〈그림 10〉 랜덤 포레스트 모델 결과(분류별) 그래프

4.1.3.4 선형 회귀 분석 모델(Linear Regression)

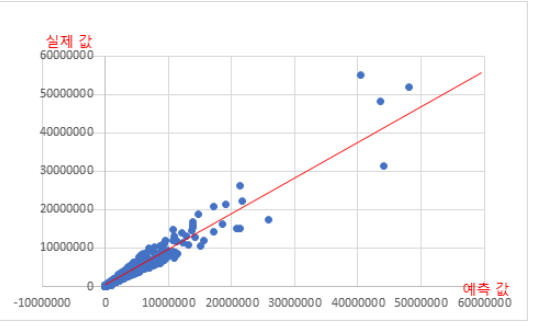
선형 회귀 분석 모델의 경우 결과는 MAPE 0.24, R^2 0.94로 나타났다.

그러나, 다음 <표 11>의 색상 표기와 같이 표본 중 가장 큰 오차는 2.62로 과적합(Overfitting)이 되었기에 수치상으로 높은 R^2 값을 나타냈지만, 실제로 새로운 데이터에 대한 예측이 어렵다는 의미로 해석되었다.

〈표 11〉 선형 회귀 분석 모델 결과

No.	예측 값	실제 값	오차
1	693,492.2	693,000	0.00071
2	767,666.3	767,000	0.00086
3	441,639.3	441,000	0.00144
⋮			
1765	-58622.7	48,000	2.22130
1766	-58636.3	48,000	2.22159
1767	-71496.4	44,100	2.62123

'예측 값'과 '실제 값'을 통해 얻은 성능의 정확도는 <그림 11>과 같이 나타났지만, 수치 및 데이터를 확인한 결과 음수 값(-)의'예측 값 '은 <그림 11>에 표현이 되지 않았다.

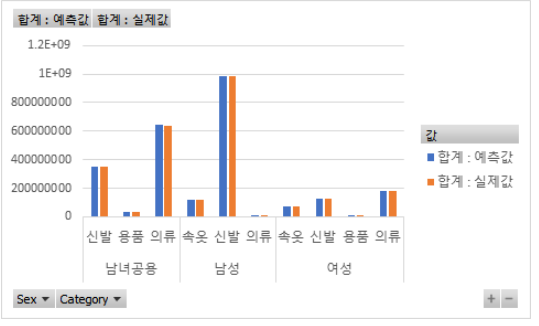


〈그림 11〉 성능 정확도(Linear Regression)분포

<표 12>와 같이 분류별(*Sex, -Category) '예측 값 '과'실제 값 '의 오차가 가장 큰 분류명은 *여성-용품(11%)로 가장 낮게 나타났다.

〈표 12〉 선형 회귀 분석 모델 결과(분류별)

분류명	예측 값	실제 값	오차
* 남녀 공용	1,026,179,978	1,019,718,900	1%
-신발	351,305,652	349,625,200	0%
-용품	33,089,620	34,294,000	-4%
-의류	641,784,706	635,799,700	1%
* 남성	1,105,527,652	1,110,381,500	0%
-속옷	121,580,855	122,855,600	-1%
-신발	981,722,633	985,126,600	0%
-의류	2,224,163	2,399,300	-7%
* 여성	391,146,768	387,885,300	1%
-속옷	70,207,736	70,868,700	-1%
-신발	130,273,589	128,905,500	1%
-용품	11,442,626	10,310,900	11%
-의류	179,222,817	177,800,200	1%
총 합계	2,522,854,398	2,517,958,700	0%



〈그림 12〉 선형 회귀 분석 모델(분류별) 그래프

4.1.4 모델별 비교 결과

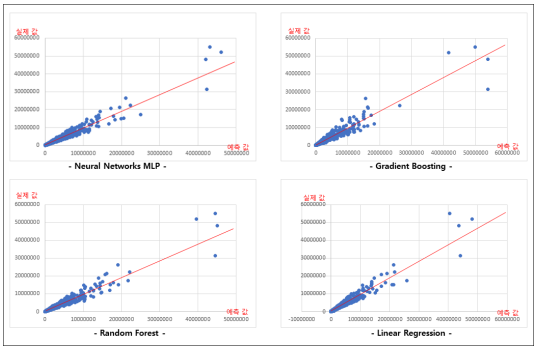
본 연구에서는 중소 전자상거래 판매상의 프로모션 수요 예측 진행할 때 활용할 수 있는 수준의 항목인 'D-AVG(30)' 평균 매출액, 'D-1'매출액, 'D-DAY'매출액, 'Sex'품목 분류1, 'Category' 품목 분류2로 분석한 연구 모델(MLP 모델, 그래디언트 부스팅 모델, 랜덤 포레스트 모델, 선형 회귀 분석 모델)의 결과를 다음 <표 13>과 같이 도출했다.

〈표 13〉 연구 모델 결과

연구 모델	MAPE	R^2
MLP	0.16	0.94
그래디언트 부스팅	0.20	0.90
랜덤 포레스트	0.16	0.94
선형 회귀 분석	0.24	0.94

MAPE 값은 MLP 모델과 랜덤 포레스트 모델 0.16, 그래디언트 부스팅 모델 0.20, 선형 회귀 분석 모델 0.24로 MAPE 값이 0.16인 MLP 모델과 랜덤 포레스트 모델이 가장 우수했다.

R^2 값의 경우 그레디언 부스팅 0.90, MLP 모델, 랜덤 포레스트 모델, 선형 회귀 분석 모델이 0.94로 높게 나왔지만, 선형 회귀 분석은 과적합이 발생하였기에 제외하면 R^2 값이 0.94인 MLP 모델과 랜덤 포레스트 모델이 가장 우수했다. 준비된 데이터셋 기준으로 실험을 진행한 결과, MLP 모델, 그래디언트 부스팅 모델, 랜덤 포레스트 모델, 선형 회귀 분석 모델 중 <그림 13>과 같이 MAPE 및 성능 정확도 분포가 기준선에 가장 가깝고, 상관성이 높게 나타낸 모델인 MLP 회귀 분석을 중소 전자상거래 판매상의 수요 예측을 위한 사전 설계 모형에 적합한 모델로 선정하였다.

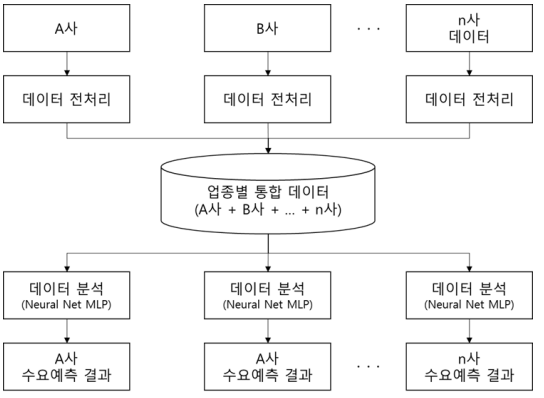


〈그림 13〉 모델별 성능 정확도 분포 비교

4.2 통합 데이터 모형

4.2.1 실험 프레임워크

본 연구에서는 사전 설계 모형을 확장하여 통합 데이터 모형을 설계하고 검증하여 중소 전자상거래 판매상이 프로모션 수요 예측 시, 데이터 부족 문제를 해소할 수 있는 방안을 제시하고자 한다. 통합 데이터 모형의 프레임워크는 <그림 14>와 같다.



〈그림 14〉 통합 데이터 모형 프레임워크

4.2.2 데이터 전처리 및 가정사항

본 연구에서는 중소 전자상거래 판매상이 프로모션 수요 예측을 하기 힘든 근본적인 문제인 데이터의 부족 문제를 해결하기 위해 통합 데이터를 적용한 실험을 진행하였다. '4장 2절 모델별 실험 결과'에서 성능이 가장 좋았던 MLP모델을 다른 중소 전자상거래 판매상의 프로모션 판매 데이터에 적용시키기 위해 채택하였다. 실험에 사용된 데이터셋은 F사의 데이터 및 F사 제품과 같은 분류를 취급하는 A사, I사, P사, T사의 데이터 총 5개사 데이터를 융합하여 적용하였다. 통합 데이터를 적용하는 실험에 사용된 데이터셋의 항목은 <표 14>와 같다.

〈표 14〉 통합 데이터 활용 실험 데이터셋

항목	내용
D-1	프로모션 전일 매출액
D-Day	프로모션 당일 매출액
Company	회사명
D-AVG(30)	프로모션 전일기준 -30일 평균 매출액
Sex 남성	분류 남성 원-핫 인코딩
Sex 여성	분류 여성 원-핫 인코딩
Sex 남녀공용	분류 남녀공용 원-핫 인코딩
Cate 속옷	분류 속옷 원-핫 인코딩
Cate 신발	분류 신발 원-핫 인코딩
Cate 의류	분류 의류 원-핫 인코딩
Cate 용품	분류 용품 원-핫 인코딩

통합 데이터 모형 실험에 사용된 데이터는 'A사' 데이터 4,496행, 'F사' 데이터 1,727행, 'I사' 데이터 3,242행, 'P사' 데이터 1,015행, 'T사' 데이터 1,814행으로 총 12,294행을 사용하였다. 또한, 입력 변수는 'A_Co.', 'F_Co.', 'I_Co.', 'P_Co.', 'T_Co.', 'Sex 남성', 'Sex 여성', 'Sex 남녀공용', 'Cate 속옷', 'Cate 신발', 'Cate 의류', 'Cate 용품' 총 12열로 구분하였다. 연구에 활용된 데이터셋의 전처리 후 기초 통계량은 <표 15>와 같다.

활용된 중소 판매상별 데이터는 'A사' 4,496행으로 가장 많은 양의 데이터를 보유하였고, 'I사'는 3,242행으로, 'F사'와 'T'사는 각 1,727행, 1,814행으로 비슷했으며, 'P사'는 1,015행으로 가장 적었다. 입력 변수인 성별 기준에서 'Sex 남녀공용'이 공통적으로 높게 분포되어 있다. 또 다른 입력 변수인 카테고리 기준에서 'Cate 용품'이 다른 카테고리에 비해 공통적으로 가장 낮게 분포되어 있었다. 각 중소 판매상이 취급하는 브랜드가 다르다는 점에서 미루어볼 때, 판매상별 인기 있는 성별, 카테고리가 다를 수 있다. 따라서, 본 연구에서는 다양한 분포가 존재하는 데이터를 기반으로 연구를 진행하였다.

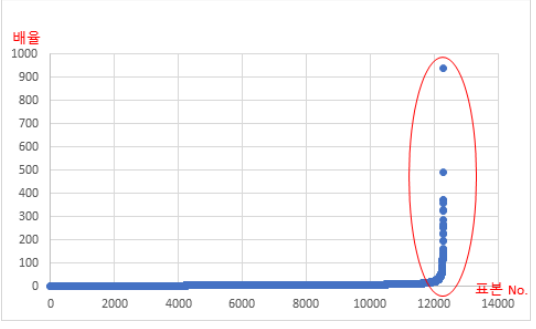
4.2.3 모형 실험

본 연구에서는 도출된 수요 예측 모델인 MLP 모델을 활용한 통합 데이터 적용 실험에 앞서 <그림 15>와 같이 'D-1'(프로모션 전일) 대비

〈표 15〉 전처리 후 활용된 기초 통계량

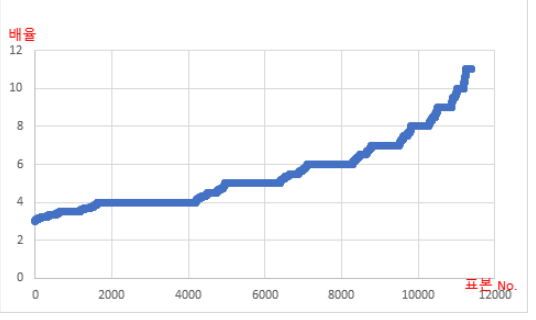
중소 판매상	분류 구분	Cate 속옷	Cate 신발	Cate 의류	Cate 용품	합계
A사	Sex 남성	447	565	41	-	1,053
	Sex 여성	1,179	6	113	21	1,319
	Sex 남녀공용	9	499	1,183	433	2,124
	합계	1,635	1,070	1,337	454	4,496
F사	Sex 남성	151	414	5	-	570
	Sex 여성	203	115	245	23	586
	Sex 남녀공용	-	160	352	59	571
	합계	354	689	602	82	1,727
I사	Sex 남성	268	433	40	-	741
	Sex 여성	560	31	184	10	785
	Sex 남녀공용	4	649	833	230	1,716
	합계	832	1,113	1,057	240	3,242
P사	Sex 남성	61	223	5	-	289
	Sex 여성	90	2	28	5	125
	Sex 남녀공용	-	203	307	91	601
	합계	151	428	340	96	1,015
T사	Sex 남성	140	198	70	2	410
	Sex 여성	233	11	127	11	382
	Sex 남녀공용	1	427	408	186	1,022
	합계	374	636	605	199	1,814

'D-Day'(프로모션 당일)의 매출액이 900배 이상 증가한 이상 구간을 발견하여 Outlier를 진행하였다. 데이터 Outlier 기준은 표본이 적으며 특정 구간의 데이터가 급격한 상승을 이루는 구간을 진행하였다.



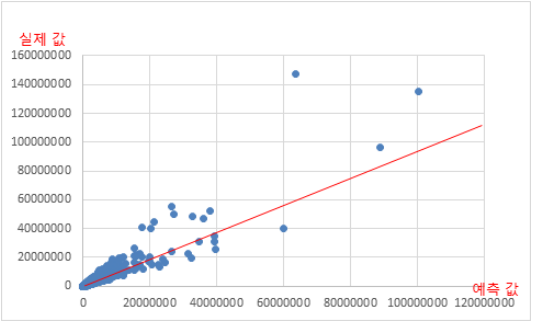
〈그림 15〉 프로모션 배율 기준 Outlier 구간

그 결과, 5개사의 데이터 표본 총 12,294줄 중 <그림 15>에 표시된 영역의 표본 918줄 제거하여 실험에 사용할 표본은 11,376줄을 채택하여 최종 데이터의 분포는 다음 <그림 16>과 같다.



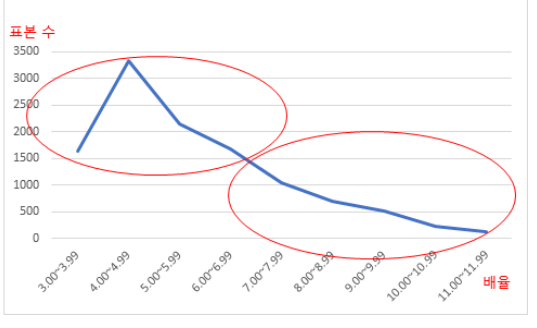
〈그림 16〉 Outlier 결과

그러나, Outlier 이후 산출된 데이터셋을 기준으로 통합 데이터의 MLP모델 결과는 <그림 17>과 같이 MAPE 0.44, R^2 0.83으로 성능이 좋지 않게 나타났다.



〈그림 17〉 성능 정확도 분포(통합 데이터)

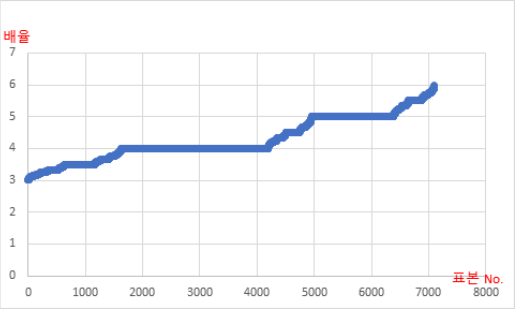
F사 단일 데이터셋의 실험 결과와는 다르게 5개사 데이터를 합친 후 성능이 좋지 않은 이유를 찾기 위해 채택한 데이터를 <그림 18>과 같이 배율을 역산하여 '매출액 After/Before(배율의 표본 수)'를 확인하였다. 분석 결과, 성능이 좋지 않았던 이유는 제공한 중소 판매상의 매출액 배율 Scale에 있었다. '비슷한 Scale을 찾아 그룹화를 진행한다면 성능이 다시 의미 있는 수치가 나올 것이다'라는 가설을 세우고 계산한 결과, 매출액('D-Day'/'D-1') 배율별 표본의 분포는 <그림 18>과 같이 4.00~4.99배율 바로 전, 후와 이후로 Scale의 차이가 크게 나타났으며,



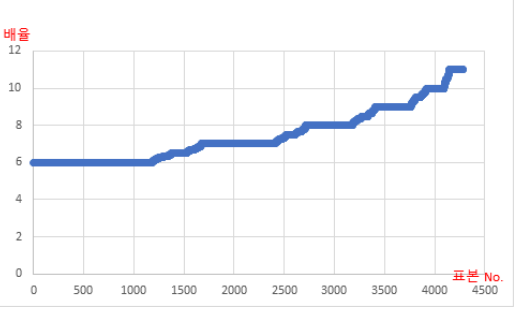
〈그림 18〉 매출액 'D-Day'/ 'D-1' 배율별 표본 수

이는 여러 규모의 중소 전자상거래 판매상의 프로모션 판매 규모 또는, 취급 품목군의 가격에 따라 편향될 수 있다는 점을 확인했다. 이 문제를 해결하기 위해 Scale에 따른 분류를 시행하였다.

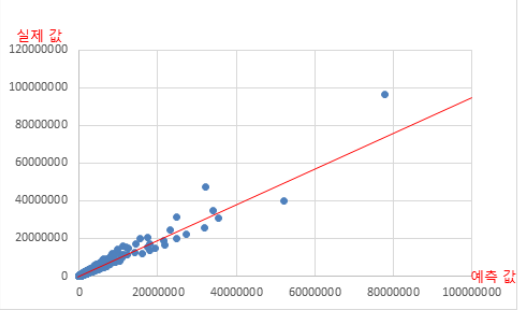
Scale에 따른 분류의 기준점은 매출액('D-Day'/'D-1') 배율별 평균 표본 수 1,264으로 설정하였다. 표본 수 1,264를 기준으로 이상과 이하 두 그룹으로 분할하였다. 분할한 두 그룹의 표본 분포는 다음 <그림 19>, <그림 20>과 같다.



〈그림 19〉 기준(평균1,264)이상 표본



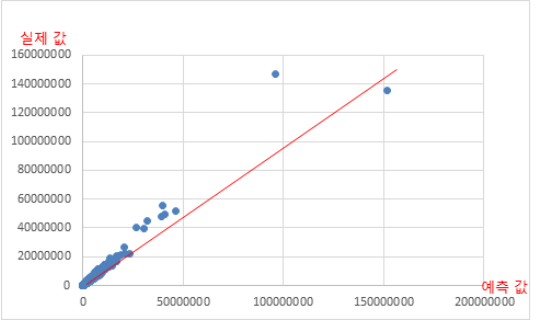
〈그림 20〉 기준(평균1,264)이하 표본



〈그림 21〉 성능 정확도 분포(기준이상 표본)

Scale에 따른 분할한 그룹별 MLP 모델의 결과는 다음과 같다. 첫 번째 그룹인 '기준(평균 1,264) 이상 표본'의 실험 결과는 MAPE 0.14, R^2 0.95로 <그림 21>과 같이 '실제값'과 '예측값'이 상관성이 좋으며 및 정확도가 향상된 의미 있는 수치가 나타났다.

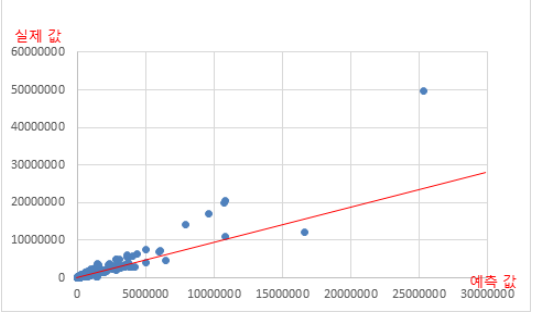
두 번째 그룹인 '기준(평균 1,264) 이하 표본'의 실험 결과는 MAPE 0.16, R^2 0.93으로 <그림 22>와 같이 '실제값'과 '예측값'이 상관성이 좋으며 및 정확도가 향상된 의미 있는 수치가 나타났다.



〈그림 22〉 성능 정확도 분포(기준 이하 표본)

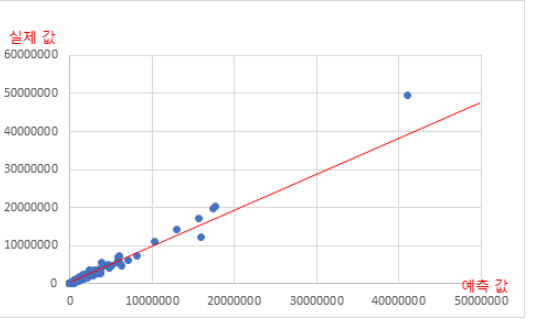
4.2.4 실험 결과

본 연구에서는'중소 전자상거래 판매상이 데이터 부족으로 프로모션 수요를 예측하기 어려운 현실이다'라는 문제를 풀기 위해 통합 데이터를 통해 부족한 데이터를 확보하고, Scale에 따른 그룹화를 통해 수요 예측의 정확도를 높일 수 있었다. 통합 데이터를 활용한 수요 예측의 수치를 확인하기 위해 'P사'를 기준으로 비교 분석하였다. 단일 데이터 실험을 위해 <표 15>와 같이 P사의 데이터 1,015 표본 중 Outlier 이후 920 표본을 기준으로 실험을 진행하였다. 결과는, MAPE 0.31, R^2 0.78로 나타났다. 단일 데이터(920표본 활용)의 성능 정확도는 <그림 23>과 같이 분포가 나타났다.



〈그림 23〉 P사 성능 정확도(단일 데이터) 분포

통합 데이터 및 Scale 그룹 적용을 통해 산출한 P사의 학습 결과는 두 그룹으로 분할하여 나타난 각각의 결과 예측 값과 실제 값으로 정확도를 평가하였기에 MAPE 및 R^2 값을 나타낼 수 없었다. 통합 데이터(11,376줄 활용 후 P사 데이터 920표본 추출)의 성능 정확도는 <그림 24>와 같이 분포가 나타났다.



〈그림 24〉 P사 성능 정확도
(통합 데이터-Scale 적용)분포

실험을 진행한 P사의 결과는 데이터 표본이 적은 단일 데이터를 활용한 성능 정확도 분포보다 통합 데이터를 활용한 성능 정확도 분포가 향상된 결과를 나타냈다.

Ⅴ. 결론

5.1 연구 결과 요약

본 연구에서는 첫 번째, 사전 설계 모형을 선정하기 위해 중소 전자상거래 판매상들의 수요

예측을 위한 다양한 모델을 탐구하고 분석하였다. MLP, 그래디언트 부스팅, 랜덤 포레스트, 그리고 선형 회귀 분석 모델의 각각의 성능을 평가한 결과 각 모델은 복잡한 데이터 구조와 다양한 품목 조건을 처리하는 데 특유의 장점을 나타냈다. 각 모델의 실험 결과, 중소 전자상거래 판매상이 보유한 수준의 데이터를 기반으로 우수한 예측 성능을 보인 MLP 모델을 사전 설계 모형으로 선정하였다. 두 번째, 본 연구에서 목적으로 하는 판매상들이 데이터 부족으로 인해 결정권자의 감에 의존하는 수요 예측이 아닌 데이터 기반의 수요 예측을 할 수 있도록 통합 데이터 모형을 제시했다. 보유한 데이터(A사, F사, I사, P사, T사)를 결합하고 적용하여 실험한 결과 의미있는 결과를 나타내기도 하였다. 더 나아가 중소규모의 전자상거래 판매상들이 데이터 기반의 전략적 의사결정의 중요성을 인지하고, 시장 동향을 반영시킬 수 있는 지표 및 취급하는 품목의 특성(입력 변수)을 잘 이해한다면 프로모션의 수요 예측을 통해 효과적인 판매 전략을 수립하고 경쟁력을 강화하는데 중요한 역할을 할 것이다. 또한, 본 연구는 데이터 기반 의사결정의 중요성을 강조하며, 중소 규모의 판매상들이 효과적인 마케팅 전략 수립에 의미가 있다. 마지막으로, 이 논문은 중소 전자상거래 판매상들이 시장 변화에 능동적으로 대응하고, 의사결정권자의 감에 의한 의사결정이 아닌, 데이터 기반의 의사결정을 내리는데 효과적인 방법론을 제시한다.

5.2 연구의 한계점

본 연구에서의 한계점은 가장 중요한 입력 변수였던 프로모션의 시행 여부와 같은 중요한 변수가 원시 데이터에 명시되지 않아, ‘D-Day’가 ‘D-1’보다 3배 이상인 자료들을 기반으로 제약을 둔 상태로 연구를 진행했던 부분 매우 아쉬운 부분이다. 또한, 이를 대체할 수 있는 데이터를 가진 중소 판매상의 데이터 또한 찾을 수 없었다. 품목의 특성, 시장의 특성을 나타내는 중소 판매상의 자료를 확보하는데 추가적인 어려움이 있었다. 예를 들어, 시장 트렌드, 소비자 행동의 급격한 변화, 경제적 요인 등은 판매 자료에 기록이 되어있지 않아, 연구에서 고려되지 않았다. 이러한 한계점들을 인식하고, 향후 연구에서는 데이터의 다양성과 범위를 확장하고, 모델의 범용성과 실용성을 높이는 방향으로 중소 판매상을 위한 프로모션 전략 수요 예측 연구가 필요하다.

참 고 문 헌

[1] Ethan Cramer-Flood, “Worldwide e-commerce growth drops to single digits, while overall retail muddles through”, Insider intelligence, Aug2. 2022,

[2] 통계청, “2023년 3월 온라인쇼핑동향”, 2023.

[3] 하나금융경영연구소, “데이터 연결 분석이 중요한 시대”, 하나Knowledge, 2023.

[4] 권영옥. (2014). 비즈니스 인텔리전스 시스템의 활용 방안에 관한 연구. 지능정보연구, 20(4), 155-169.

[5] 이종주, “‘데이터 격차’, 다가올 중소벤처기업의 위험”, 『소프트웨어 정책연구소 산업동향』, 2021. pp4-11.

[6] 김근환, 권태훈, 전승표, “공공 정보지원 인프라 활용한 제조 중소기업의 특징과 성과에 관한 연구”, J Intell Inform Syst, Vol.25 NO.4, December 2019. pp.1-33.

[7] ㈜원제로소프트, “(2세부) 온라인 유통과 물류 시스템 연계를 위한 표준정보시스템 개발”. 지식서비스산업기술개발. 산업통상자원부. 제 2021-51호.

[8] 이재훈. “이커머스 시장의 수요 트렌드 예측을 위한 딥러닝 모델에 대한 연구.” 국내석사학위

논문 인천대학교 동북아물류대학원, 2021. 인천 2021.

[9] 유루루. “머신러닝을 활용한 온라인 약국 의약품 판매량 예측 모형 개발.” 국내석사학위논문 전남대학교, 2021. 광주.

[10] 박성철. “시계열 분석을 통한 중개 사이트에서의 쇼핑몰 매출액 예측 모형 구축.” 국내석사학위논문 연세대학교 공학대학원, 2015. 서울.

[11] 오지연. “빅데이터 분석을 활용한 온라인 판매 수요 예측.” 국내석사학위논문 한신대학교 대학원, 2019. 경기도.

[12] 이강현, 방선호, 장지영, 신광섭. (2022). 기계학습 기법을 활용한 수요 예측 모형 개발 -몽골 유통 기업 사례-. 물류학회지, 32(6), 111-120.

[13] 정윤재. “소비자 직접판매 D2C 패션산업의 수요예측 및 초도발주량 최적화.” 국내석사학위논문 인천대학교 동북아물류대학원, 2021. 인천.

[14] 김광호, 장병훈, 최황규. (2019). 원-핫 인코딩을 이용한 딥러닝 단기 전력수요 예측모델. 전기전자학회논문지, 23(3), 852-857.

[15] 김영남, 모혜란, 이지홍, 류상천, 김현. (2022). e커머스 풀필먼트 비즈니스를 위한 멀티 조합 수요예측 모델 연구. 대한전자공학회 학술대회.

[16] 주종문(Jong-Moon Ju),and황승국(Seung-Gook Hwang). “전자상거래 효율화를 위한 Web Mining기반 수요예측방법.” 한국지능시스템학회 학술발표 논문집 14.1 (2004): 7-12.

[17] 안세희, 정재윤. (2023). Temporal Fusion Transformer를 이용한 대형마트 판매량의 다단계 시계열 수요예측. 한국전자거래학회지, 28(3), 43-53, 10.7838/jsebs.2023.28.3.043

[18] 유지현. (2019). 머신러닝을 이용한 관중 수요예측에 관한 연구. 전기전자학회논문지, 23(4), 128-134.

[19] 정원희, 정다운, 강아영, 구영현, and 유성준. “수요 패턴 별 최적 머신러닝 수요예측 모델 성능 비교.” 한국차세대컴퓨팅학회 논문지 16.6 (2020): 76-89.

[20] 정세훈. “머신러닝 회귀 모델 알고리즘을 적용한 기업의 판매량 예측 연구 및 분석.” 국내석사학위논문 부경대학교, 2021. 부산.

[21] 조성일. “두유제품의 수요예측 정확도 향상에 관한 연구.” 국내석사학위논문 인하대학교 물류전문대학원, 2019. 인천.

[22] 임설아. “홈쇼핑 매출 알고리즘 개발에 관한 연구.” 국내박사학위논문 세종대학교 대학원, 2021. 서울.

저 자 소 개

이 성 주(Seung-Joo Lee)

・2012년~현재: ㈜원제로소프트 차장/팀장

・2022년~현재: 인천대학교 동북아물류대학원 석사과정

<관심분야> 컨설팅, 이커머스, BL분석 및 기업진단

이 용 현(Young-Hyun Lee)

・2000년~현재: ㈜원제로소프트 이사/연구소장

・2002~2004년: 경희대학교 대학원 컴퓨터공학 (석사)

<관심분야> 웹프로그래밍/개발, SCM, 데이터 가공, 데이터 정제, 데이터 분석

김 진 현(Jin-Hyun Lee)

・2017년~현재: ㈜원제로소프트 대리/선임
・2013~2017년: 장안대학교 게임컨텐츠과 (학사)
<관심분야> 웹프로그래밍/개발, 데이터 정제, 데이터 분석

이 강 현(Kang-Hyun Lee)

・2022년 2월: 청주대학교 전자공학과 (학사)
・2022년 3월~현재: 인천대학교 동북아물류대학원 물류시스템학과 석사과정
<관심분야> 빅데이터, 머신러닝

신 광 섭(Kwang-Sup Shin)

・2003년 2월: 서울대학교 산업공학과 (공학사)
・2006년 2월: 서울대학교 산업공학과 (공학석사)
・2012년 2월: 서울대학교 산업공학과 (공학박사)
・2012년 2월~현재: 인천대학교 동북아물류대학원 교수
<관심분야> 빅데이터 활용, 솔루션