설명가능한 기계학습을 이용한 베스트셀러 예측과 영향요인 분석
Predicting Bestsellers and Key Drivers using Explainable Machine Learning

이승필(주저자) · 박은일(공저자) · 류두진(교신저자)
Seungpeel Lee(First Author) · Eunil Park(Co-Author) · Doojin Ryu(Corresponding Author)

사회평론 전무이사 Executive Director, Sahoipyoungnon Publishing Co., Inc.(*leepeel@sapyoung.com*)
성균관대학교 실감미디어공학과 부교수 Associate Professor, Department of Immersive Media Engineering, Sungkyunkwan University(*eunilpark@skku.edu*)
성균관대학교 경제학과 교수 Professor, Department of Economics, Sungkyunkwan University(*sharpjin@skku.edu*)

온라인 서점의 방대한 데이터를 기반으로 판매량이나 베스트셀러를 예측하는 연구는 판매기록과 리뷰 등 출판 이후 자료에 의존하므로, 신간도서에 대해 베스트셀러 예측을 하기 어려운 콜드 스타트 문제를 가진다. 본 연구에서는 온라인 서점의 문학 장르에서 신간도서의 메타 데이터를 사용해 베스트셀러를 예측하는 기계학습 모형을 구현하였으며, LightGBM 모형이 가장 우수한 성능을 보였다. 특성 중요도 기법과 SHAP방법으로 저자 빈도, 출판사 빈도, 카테고리 빈도, 가격, 출판월이 베스트셀러 예측에 영향을 미치는 요소임을 확인하였다. 연구 결과는 콜드 스타트 문제 해결에 기여하고, 온라인 서점이 신간도서의 성공 가능성을 예측하며 마케팅 전략을 수립하는 데 함의를 제공한다.

주제어: 기계학습, 베스트셀러, 설명가능한 인공지능, 온라인 서점, 예측

Research on predicting sales volumes or identifying bestsellers in online bookstores often relies on post-publication data, such as sales records and customer reviews, which poses a cold start problem for new books. This study addresses this issue by developing a machine learning model based solely on metadata from newly released literary books. Among the tested models, LightGBM exhibits the best predictive performance. Using feature importance analysis and the SHAP method, we identify key factors influencing bestseller prediction, including author frequency, publisher frequency, category frequency, price, and publication month. Our findings provide a solution to the cold start problem and offer actionable insights for online bookstores to anticipate a book's success potential and refine marketing strategies.

Keyword: Bestseller, Explainable AI, Machine Learning, Online Bookstore, Prediction

Ⅰ. 서 론

도서시장은 온라인 중심으로 재편되고 있다(Baye, De los Santos, and Wildenbeest, 2013; Howard, 2009; Khatun et al., 2019). 아마존(Amazon)은 미국 도서 판매량의 절반을 차지하며(bbc.com), 2023년에는 교보문고의 온라인 매출이 오프라인 매출

최초투고일: 2024. 08. 13 수정일: (1차: 2024. 11. 25) 게재확정일: 2024. 12. 26

을 15% 이상 앞서 있는 것으로 나타났다(kpa21.or.kr). 온라인 시장의 성장은 출판산업의 가치사슬 구조를 변화시키고 있다. 전통적인 가치사슬은 저자, 출판사, 유통업체, 오프라인 서점, 독자로 구성되는데(Magadán-Díaz and Rivas-García, 2019), 온라인 시장이 출현하면서 저자, 출판사, 온라인 서점, 독자로 이어지는 새로운 가치사슬이 생성된다. 출판사와 독자는 온라인 플랫폼을 통해 연결되고, 이 플랫폼에는 도서정보와 독자의 리뷰(review)를 비롯해 다양한 자료가 축적된다(Lee, Kim, and Park, 2023). 온라인 서점은 이러한 빅데이터(Big data)를 기반으로 마케팅에 유용한 여러 예측을 시도할 수 있게 되었다. 많은 신간도서와 정보가 전달되는 서점의 입장에서 판매량 예측은 중요하다. 판매가 잘 될 도서를 선별해 플랫폼에 노출하고 마케팅 예산을 효과적으로 투입하는 일과 베스트셀러(bestseller) 가능성이 큰 책의 수요를 예측하여 재고를 적절하게 확보하는 일은 온라인 서점 상품기획자의 중요한 업무이다. 온라인 서점은 판매기록, 저자의 인지도, 출판사 전략지표 등 시계열 정형데이터를 기반으로 도서의 판매량과 베스트셀러 여부를 예측해 왔다.

도서 판매량과 순위 및 베스트셀러 여부를 예측하는 연구는(김도영 외, 2023; 유지은 외, 2023; Feng et al., 2020), 판매기록, 도서 리뷰, 판매순위 등 출판 이후에 생성된 데이터를 활용한다는 한계점이 있다. 출판 이후 데이터를 활용하면 예측 정확도를 높일 수 있지만, 신간도서는 이러한 데이터가 부족하여 예측이 어려운 콜드 스타트(cold start) 문제가 존재한다. 콜드 스타트는 예측 및 추천 시스템에서 새로운 사용자나 제품에 관한 정보가 부족해 성능이 떨어지는 문제로(Bobadilla et al., 2012), 판매의 불확실성이 높은 신간도서의 판매량이나 베스트셀러 여부를 예측해야 하는 온라인 서점은 이를 고려해야 한다.

본 연구는 기계학습 알고리즘을 기반으로 출판 전 신간도서의 베스트셀러 예측모형을 제시하고 콜드 스타트를 해결하는 것을 첫 번째 목적으로 한다. 이는 온라인 서점이 신간도서의 성공 여부를 예측해 효과적으로 마케팅 자원을 배분하고 재고를 관리하는 데 기여한다. 현재 공개된 도서판매 데이터가 부재하므로 대안으로 도서 판매량을 간접적으로 보여주는 베스트셀러를 예측 대상으로 설정하였다. 두 번째 목적은 설명가능한 인공지능 모형을 이용해 예측에 영향을 미치는 요인, 즉 도서의 성공요인을 파악하는 것이다. 이는 온라인 서점이 소비자의 구매 경향성을 파악해 효과적인 마케팅 전략을 세우는 데 함의를 제공한다.

대형 온라인 서점인 예스24의 '소설/시/희곡' 분야에서 2021년부터 2023년까지 최근 3년간 출판된 도서 자료를 수집하고, 여러 기계학습 알고리즘을 활용해 출판 전 신간도서의 베스트셀러 여부를 예측하는 모형을 제시한다. 실제 데이터는 베스트셀러가 적은 클래스 불균형 자료(class imbalance data)이므로 일반적인 성능지표 외에 불균형 데이터의 예측성과를 평가하기에 적합한 지표도 추가로 사용하였다. 실험 결과 전반적으로 부스팅(boosting) 및 선형(linear) 모형의 성능이 좋아 기계학습 알고리즘으로 신간도서의 베스트셀러 예측모형을 구현하는 것이 가능함을 확인하였다. 가장 우수한 성능을 보인 LightGBM(Light Gradient Boosting Machine) 모형의 예측결과를 대상으로 베스트셀러 예측에 영향을 미친 요인을 분석하였다. 기계학습의 예측과정을 설명하는 XAI(eXplainable Artificial Intelligence) 방법론인 특성 중요도 기법과 SHAP(SHapley Additive exPlanations) 기법을 적용하였고, 저자 빈도, 출판

사 빈도, 카테고리 빈도와 가격, 출판 월 등이 예측에 큰 영향을 주는 요인으로 나타났다.

본 연구의 결과는 신간도서에 관한 베스트셀러 예측모형을 구현해 콜드 스타트 문제를 해결하고, 온라인 서점이 신간도서의 성공 가능성을 사전에 판단하여 도서 노출 전략을 최적화하고 마케팅 예산을 효율적으로 분배하는 데 도움을 줄 수 있다. 또한, 베스트셀러 예측에 영향을 주는 요인을 분석하여 온라인 서점이 소비자의 구매행동에 기반해 마케팅 전략을 세우고 재고를 관리하는 데 기여할 수 있다.

본 논문의 구성은 다음과 같다. 제2장은 도서의 판매 예측, 성공 및 구매 요인, 콜드 스타트 문제에 관한 선행연구를 살펴본다. 제3장은 데이터와 연구방법을 설명하고, 자료 수집, 변수 선정, 데이터 전처리 과정, 예측에 사용한 기계학습 모형, 성능지표를 기술한다. 제4장은 예측모형의 성능을 분석해 기계학습을 이용한 베스트셀러 예측의 가능성을 논하고, XAI 기법을 사용해 예측에 영향을 미친 요인을 분석한다. 제5장에서는 연구의 내용을 요약, 정리하고 연구의 시사점 및 한계와 향후 방향을 제시한다.

Ⅱ. 관련 연구

2.1 도서판매 예측에 관한 연구

출판 마케팅 및 유통 분야에서 도서판매의 예측은 중요하며, 선행연구는 다양한 기계학습 알고리즘을 사용하여 예측모형을 제시하였다. 유지은 외(2023)는 메타 데이터를 활용하여 도서가 베스트셀러 순위 200위 내에 3개월 또는 6개월 이상 유지될지 예측하였다. 주요 변수로 리뷰개수, 평점, 판매가, 출판 이후 개월 수, 제목, 장르, 출판사, 저자 등의 정보를 사용하고, 예측모형으로 Decision Tree(DT), Random Forest(RF), GradientBoost, XGBoost (eXtreme Gradient Boosting), LightGBM, CatBoost(Categorical Boosting), 다중 퍼셉트론을 사용하였다. 김도영 외(2023)는 국내도서가 해외에서 출판될 때 판매량을 예측하는 모형을 제시하였다. 이 연구는 한국문학번역원 데이터와 Amazon 및 Goodreads에서 크롤링한 데이터를 활용하여 578종의 도서를 분석하였다. 주요 변수로 작가의 해외 출판 횟수, 평균 평점, 출판 국가, 평점 참여자 수 등을 사용하고, 예측모형으로는 XGBoost, GradientBoost, Adaboost(Adaptive Boosting), LightGBM, RF, Support Vector Machine(SVM), Logistic Regression(LR), Deep Learning을 적용하였다. 또한 Park, Lee, & Doo(2020)는 도서의 판매와 반품에 영향을 미치는 요소로 출판부수, 반품부수, 출판연도, 카테고리, 시리즈, 주제, 베스트셀러 여부, 가격, 유명작가 여부 등을 조사하고, 판매 수요를 예측하는 회귀 모형과 RF모형을 제시하였다.

심층신경망 모형을 활용하여 도서판매 예측을 시도한 연구도 있다. Sharma et al.(2019)은 아마존 판매량 예측을 위해 회귀분석과 의사결정나무 외에 인공신경망 모형을 비교 분석하였다. 인도 아마존(Amazon.in)에서 수집된 도서의 가격, 할인 금액, 할인율, 리뷰수, 페이지 수, 평점, 긍정 및 부정 감성의 강도 등을 활용하였다. Feng, Choy, and Laik (2020)은 GAN(Generative Adversarial Network) 기반 방법론을 제시하며 Amazon 도서 판매순위를 예측하였다. 이들은 아마존에서 수집된 도서 제목, 저자, 출판사, 도서 형식, 도서 평가, 가격, 페이지 수,

언어, 대여 가능 여부, 고객 리뷰, 도서 설명, 기존 주간 판매순위 데이터를 사용하였다. GAN은 두 개의 신경망이 서로 경쟁하는 방식으로 학습되는 알고리즘으로 주로 이미지 등 다차원 데이터의 생성에 사용되었으나, 최근에는 단일 차원 데이터의 회귀 및 시계열 예측에도 적용되고 있다. Martín Sujo, Golobardes i Ribé, and Vilasís Cardona(2021)는 서점 자료에 소셜네트워크(social network)와 웹에서 수집한 정보를 추가해 베스트셀러 도서를 예측하는 모형을 제안하였다. 스페인에서 출판된 도서를 대상으로, 도서 제목, 저자, 출판일, 가격, 저자의 소셜네트워크 활동 및 웹 언급 등을 분석한다.

출판 후 데이터인 리뷰 내용 및 개수, 평점, 판매순위 등을 활용하여 도서 판매량이나 순위를 예측하는 선행연구의 접근 방식은 신간도서의 판매 예측에서 콜드 스타트 문제를 야기한다. 신간도서에는 위와 같은 데이터가 존재하지 않으므로 판매량이나 베스트셀러 여부를 예측하는 데 한계를 가질 수밖에 없다. 본 연구는 출판 후 데이터 대신 메타 데이터를 활용하여 베스트셀러를 예측하는 방법을 제안한다. 출판 전 정보를 사용함으로써 콜드 스타트를 해결하고, 신간도서의 성공 가능성을 사전에 평가하여 마케팅 전략을 수립할 수 있다.

2.2 도서의 성공 및 구매 요인에 관한 연구

도서의 성공 및 구매 요인은 문학, 출판, 마케팅 등 다양한 분야에서 연구된 중요한 주제이다. 도서의 성공 및 구매 요인은 도서의 내용과 관련된 내적 요인과 도서의 외부 환경과 관련된 외적 요인으로 구분된다. 내적 요인에 관한 연구로는 작가의 글쓰기 스타일과 도서 성공의 관계를 분석한 연구가 있다. Ashok, Feng, and Choi(2013)는 소설의 성공 예측에서 작문 스타일의 영향을 분석하였다. Project Gutenberg에서 수집한 다양한 장르의 소설 데이터를 기반으로 LibLinear(Library for Large Linear Classification) SVM을 사용하여 예측성능을 평가하였다. 통계적 스타일 분석을 통해 소설의 성공을 최대 84%의 정확도로 예측하였으며, 성공적인 글쓰기 스타일의 특징요소를 밝혀냈다. 외적 요인에 관한 연구로는 리뷰와 도서 성공의 관계를 분석한 연구가 있다. Chevalier and Mayzlin(2006)은 온라인 서점의 리뷰가 도서판매에 미치는 영향을 분석하였다. Amazon과 Barnes & Noble에서 수집한 데이터를 분석하였는데, 리뷰와 평점이 판매에 영향을 미치며, 특히 부정적 리뷰가 큰 영향을 미친다는 사실을 발견하였다. Nakamura(2013)는 디지털 시대의 독서와 소셜 네트워킹의 관계를 탐구하기 위해 Goodreads 플랫폼에서 독자가 상호작용하고 리뷰를 공유하는 방식을 분석하였다. 독서 플랫폼은 독서 경험을 사회적인 활동으로 변화시키며, 이는 독서의 상업화와 관련된 도전과 기회를 제공한다고 주장하였다. 비평, 수상, 광고와 도서 성공의 관계를 분석한 연구도 있다. Clement et al.(2007)은 독일의 문학 프로그램에서 소개된 책과 베스트셀러 목록을 비교하고 혼합회귀모형(mixed regression model)을 사용하여, 비평가의 극단적인 평가가 책의 판매를 증가시키며, 비평가의 의견 불일치가 인지도를 높이는 것을 발견하였다. Kovács and Sharkey(2014)는 문학상 수상이 작품에 미치는 영향을 탐구하기 위해 수상작 및 후보작에 관한 Goodreads 리뷰 자료를 분석하였다. 수상작이 더 많은 주목을 받으면서 평가 점수가 낮아지는 경향이 있다는 사실을 발견하였는데, 이는 기대치와 평가자의 다양성이 증가하면서 나타난 변화라고 설명한다. Shehu et al.(2014)은 독일 도서시장에서 광고가 판매에 미

치는 영향을 조사하기 위해 약 600종의 데이터를 사용하여 잠재적 선택효과와 광고의 영향을 분석한다. 광고의 영향력을 과대평가하는 선택효과가 존재하며, 인지도가 낮은 저자의 책이 광고를 통해 판매가 증가함을 발견하였다.

독자의 동기와 필요가 도서의 성공이나 구매에 영향을 미친다는 연구도 있다. Lee, Yi, and Kim(2023)은 독자의 리뷰에서 성장, 관계, 인지적 필요가 도서의 장기적 성공에 영향을 미치는 중요한 요인임을 밝혀냈다. Leitão et al.(2018)은 독자는 제목, 줄거리, 주제, 지인의 추천, 할인 여부 등 다양한 요인을 고려해 책을 구매하며, 여성 독자는 자신을 위한 책을 구매할 때 더 충동적인 성향을 보인다고 분석한다. Ozturk et al.(2006)은 독서 및 도서구매 행동을 분석하여 독서가 개인적 발전, 교육적 목적, 즐거움과 도피의 수단으로 활용됨을 밝혔다. 시각적 요소와 도서구매의 관계를 분석한 연구도 있다. Visentin and Tuan(2021)은 책 표지의 '벨리 밴드'가 구매 결정에 미치는 영향을 조사하여 시각적 단서가 주의를 끌고 감정적 반응을 유도하여 구매행동에 긍정적인 영향을 줄 수 있음을 발견하였다. Park et al.(2023)은 책 표지의 시각적 매력도가 구매 결정에 미치는 영향을 분석하였다. 표지의 사진과 색상이 소비자에게 감정적인 즐거움을 주며, 이러한 감정적 반응이 책의 인식 가치를 높여 결과적으로 구매로 이어질 수 있음을 보여주었다.

도서의 성공 및 구매 요인을 분석한 기존 연구 중 다양한 요인을 기반으로 도서의 성공을 예측하는 연구는 부족하다. 도서의 성공에는 여러 요인이 복합적으로 작용하므로 광범위한 요인과 도서 성공의 관계를 분석하는 연구가 필요하다. Schmidt-Stölting et al.(2011)은 독일 시장에서 소설 판매 성공요인을 분석하기 위해 대규모 데이터를 사용하여 인기 작가, 장르, 출판사 역량, 책 표지 디자인 등의 요인을 분석하였으나, 단순한 선형모형 사용으로 성능이 제한적이었다. 본 연구는 기계학습 모형을 사용하여 도서의 성공을 예측하고 예측에 영향을 미친 다양한 요인을 분석한다는 점에서 차별점을 갖는다. 구체적으로 본 연구는 확률모형, 선형모형, 트리기반 모형, 부스팅 모형 등 다양한 기계학습 모형을 사용해 신간도서의 베스트셀러 여부를 예측하고, 예측과정을 설명하는 XAI 기법을 활용해 예측에 영향을 준 요인을 탐구한다. 이를 통해 신간도서의 성공요인을 밝혀내고 온라인 서점이 도서 마케팅 전략을 수립하는 데 유용한 정보를 제공한다.

2.3 콜드 스타트에 관한 연구

콜드 스타트는 예측 및 추천 시스템에서 발생하는 대표적 문제로, 새로운 사용자나 제품에 관한 데이터가 부족하여 알고리즘이 효과적으로 작동하지 못하는 상황을 의미한다(Lika et al., 2014; Son, 2016). 이 문제는 예측 및 추천의 정확도를 떨어뜨려 사용자 만족도를 감소시키고, 시스템의 성능을 저해한다. 마케팅, 인공지능 분야에서 콜드 스타트를 해결하기 위한 선행연구는 비정형 데이터가 포함된 소셜네트워크 데이터, 신뢰 데이터, 위치기반 데이터, 크로스 도메인 데이터 등 여러 유형의 데이터를 활용한다. 소셜네트워크 데이터는 주로 소셜미디어나 음악 추천 시스템과 같은 사용자 활동 정보가 풍부한 도메인에서 연구되었다. Abel et al.(2013)은 소셜미디어 플랫폼에서 사용자가 생성한 태그를 통해 사용자 프로필을 보강하고, 소셜네트워크 데이터가 사용자 관심사를 효과적으로 반영하며, 초기 사용자에게 적합한 추천을 제공할 수 있음을 보여주었다. Nie et al.(2014)은 소셜네트워크 사용자의 관심

사와 사회적 영향력을 반영한 편향랜덤워크 알고리즘을 사용하여 초기 데이터가 부족한 사용자에 관한 추천의 정확도를 높이는 데 기여한다. Givon and Lavrenko(2009)는 소셜 태그 기반 예측모형을 활용해 사용자 태그를 바탕으로 도서 간 유사도를 계산하여, 새로운 책에 관한 평가 데이터가 부족한 상황에서도 유의미한 추천을 제공할 수 있음을 보여주었다. 신뢰 데이터는 사용자 간 신뢰관계를 활용하여 부족한 평가 데이터를 보완하는 접근 방식에서 사용되고 주로 전자상거래나 리뷰 플랫폼과 같은 도메인에서 연구되었다. 이러한 도메인에서는 사용자 간의 평판이나 신뢰도가 구매결정에 중요한 역할을 한다. Ghavipour and Meybodi(2019)는 전자상거래 플랫폼에서 신뢰 네트워크를 통해 사용자 간 신뢰도를 반영하여 추천 품질을 향상하는 방안을 연구하였다. 신뢰도가 높은 사용자의 평가를 반영하여 새로운 사용자에게 추천을 제공함으로써 콜드 스타트를 완화하였다. Zou et al.(2015)은 사용자 간 신뢰관계를 기반으로 새로운 사용자나 데이터가 부족한 상황에서 추천 정확도를 향상시키는 TrustRank 알고리즘을 제안하였다. 이 알고리즘은 친구 또는 친구의 친구 관계를 확장하여 사용자가 신뢰할 수 있는 추천을 생성하였다. 위치기반 데이터는 지역 비즈니스 추천 시스템이나 위치기반 서비스에서 사용된다. Liu and Meng(2015)은 이동 경로와 위치 정보를 기반으로 자주 방문하는 지역 맞춤 추천을 통해 새로운 사용자에게도 유의미한 추천을 제공한다. Rosli et al.(2015)은 Facebook의 위치 관련 정보를 기반으로 사용자 유사도를 계산하여 새로운 사용자에게 맞춤형 추천을 제공하였다. 지역 비즈니스의 관심을 반영하여 보다 개인화된 추천을 가능하게 하였다. 크로스 도메인 데이터는 한 도메인에서 얻은 데이터를 다른 도메인에 적용하여 추천의 정확도를 높이는 데 사용된다. Gao et al.(2019)은 영화 추천 데이터를 기반으로 보조 도메인에서 대상 도메인으로 아이템만 전송하여 사용자 프라이버시를 보호하면서 데이터 희소성 문제를 해결하는 방법을 제안하였다. Zhang et al.(2017)은 영화, 책, 음악 데이터를 기반으로 도메인 간 일관성을 유지하며 정보 전이를 가능하게 하여, 새로운 제품에 관한 추천 정확도를 높였다.

한편, 초기 구매 데이터나 구매행동 데이터, 광고 데이터를 활용해 콜드 스타트 해결을 시도하는 마케팅 연구도 있다. Padilla and Ascarza(2021)는 고객관계관리(CRM; Customer Relationship Management)에서 콜드 스타트를 해결하기 위한 접근법으로, 고객의 첫 구매 데이터를 활용하여 고객의 선호도를 예측하는 기계학습 방법을 제안하였다. Padilla et al.(2025)은 고객의 구매행동을 분석하여 데이터가 부족한 상황에서도 개인화된 추천이 가능한 방법론을 제공하였다. Yang et al.(2024)은 광고 데이터를 활용하여 새로운 제품의 품질을 평가하고, 이를 검색 순위에 반영하여 콜드 스타트를 해결하는 접근법을 제안하였다. 이처럼 다양한 연구가 존재하지만, 도서 분야에서 이 문제를 해결하려는 연구는 드물다. 본 연구는 출판계에서 실용적으로 중요한 과제인 신간도서의 베스트셀러 예측과 콜드 스타트를 주제로 삼아 신간도서 메타 데이터에 기반한 예측모형을 제안하였다는 점에서 음악, 영화, 전자상거래 등 다른 도메인의 콜드 스타트를 다룬 연구와 차별된다.

Ⅲ. 데이터 및 방법론

3.1 데이터 수집과 변수 선정

본 연구는 예측모형에 사용할 자료를 수집하기 위해 예스24(company.yes24.com)의 분야별 매출에서 수험서, 학습서, 만화 다음으로 높은 비중을 차지하는 '소설/시/희곡' 분야에서 2021년 1월부터 2023년 12월까지 출판된 22,447개 도서의 정보를 Python 라이브러리인 BeautifulSoup과 Selenium을 활용하여 웹 크롤링(web crawling)을 통해 수집하였다. Park, Lee, and Doo(2020)와 Sharma, Chakraborti, and Jha(2019)를 참고해 수집할 정보와 변수를 선정하였다. 기존 연구에서 사용된 독립변수로는 제목, 저자, 출판사, 가격, 리뷰, 평점, 순위, 페이지 수, 출판일, 할인율 등이 있는데 출판 후 생성된 정보인 리뷰, 평점, 순위는 제외하였고 국내 온라인 서점은 할인율이 10%로 동일하므로 할인율도 수집 대상에서 제외하였다. 여기에 온라인 서점이 출판사로부터 제공받는 도서정보인 카테고리, 책 소개, 저자 소개를 추가하였고, ISBN(International Standard Book Number)도 수집 대상에 포함하였다. ISBN은 국제표준도서번호로 구매에 영향을 미친다고 보기 어렵지만, 실험과정에서 개별도서를 식별하기 위해 수집하였다. 카테고리는 '소설/시/희곡' 분야의 하위 카테고리, 예를 들어 한국소설, 영미소설, 장르소설, 시 등을 의미한다. 저자 정보에는 동명이인이 존재할 수 있으므로 동명이인 리스트를 수집하고, 해당 저자가 동명이인 리스트에 있으면 저자명을 구분하였다. 문학 장르에서는 동일한 원전이 여러 번역본으로 출판되기도 하는데 일반적으로 출판사에 따라 다른 번역본이 출판되므로 각 번역본은 출판사로 구별하였다.

또한 도서의 저자, 출판사, 카테고리의 인기도가 구매에 영향을 미칠 수 있다고 가정하고 이를 수치화하였다. 선행연구(Martín Sujo, Golobardes i Ribé, and Vilasís Cardona, 2021)에서는 저자의 소셜네트워크 플랫폼 활동 정보(팔로워 수, 게시물 수, 상호작용 수 등), 웹페이지 및 뉴스 등에서의 언급 빈도를 저자 인기도의 지표로 삼았다. 그러나 모든 저자가 소셜네트워크 활동을 하는 것이 아니고, 출판사나 카테고리는 이와 같은 지표로 인기도를 측정하기에는 적합하지 않다. 본 연구는 도서의 저자, 출판사, 카테고리가 출판일을 기준으로 최근 3년 동안의 주별 베스트셀러 순위에 몇 회 들어 있는지 그 빈도를 계산해 인기도의 지표로 삼았다. 주별 베스트셀러 순위 범위는 1위부터 1,000위로 설정하였다. 예스24를 비롯해 알라딘, 인터넷교보문고 등 모든 온라인 서점은 주별 베스트셀러 순위를 상위 1,000개까지 공개하고 있어 이러한 관례에 따라 베스트셀러 기준을 정하였다.

저자, 출판사, 카테고리의 빈도를 계산하기 위해 2018년 1월 첫째 주부터 2023년 12월 마지막 주까지 예스24 주별 베스트셀러 1등부터 1,000등까지에 속하는 도서의 저자, 출판사, 카테고리를 수집하였다. 이 데이터는 2021년부터 2023년까지 출판된 도서의 저자, 출판사, 카테고리가 최근 3년간 베스트셀러 순위에 포함된 빈도를 계산하기 위한 것이다. 예를 들어 2021년 1월 1일에 출판된 도서의 경우 3년 전인 2018년 1월 1일부터 출판 직전인 2020년 12월 31일까지의 기간 동안 해당 도서의 저자, 출판사, 카테고리가 주별 베스트셀러 순위에 몇 회 포함되어 있는지 계산하였다. 계산 결과는 '저자 빈도', '출판사 빈도', '카테고리 빈도'라는 이름의 변수로 저장하였다. '저자 빈도', '출판사 빈도', '카테고리 빈도'

는 도서의 출판일을 기준으로 3년 전부터 출판일 직전까지 베스트셀러 순위에 포함된 빈도이므로 동일한 저자, 출판사, 카테고리라 하더라도 도서의 출판일에 따라 값이 달라진다.

다음으로 2021년부터 2023년까지 출판된 도서의 출판 후 베스트셀러 여부를 분류하였다. 예스24에서 '소설/시/희곡' 분야의 2021년 1월 첫째 주부터 2024년 4월 첫째 주까지 주별 베스트셀러 순위 도서의 ISBN 리스트를 수집하였다. 앞에서 수집한 22,447개 도서의 ISBN이 이 리스트에 1회 이상 포함되어 있으면 클래스(class) 1, 포함되어 있지 않으면 클래스 0으로 분류하였다. 베스트셀러인 클래스 1에 속하는 데이터 개수는 4,671개, 베스트셀러가 아닌 클래스 0에 속하는 데이터 개수는 17,776개로 나타났다. 두 클래스의 비율이 20 대 80으로 클래스 0 데이터의 비중이 월등히 높으므로 예측모형이 두 클래스를 균형 있게 학습하지 못할 가능성이 있다. 이에 본 연구는 모형 학습 시 불균형 문제를 해소하기 위한 SMOTE(Synthetic Minority Over-sampling Technique) 기법을 적용하였고, 불균형한 데이터에서 성능을 잘 나타내는 지표인 매튜 상관계수(MCC; Matthews Correlation Coefficient)를 추가로 사용하였다.

수집 데이터 중 독립변수는 제목, 저자, 출판사, 가격, 페이지 수, 출판 월(특정 출판 월이 베스트셀러 예측에 영향을 주는지 파악하기 위해 출판일을 변환하여 사용), 카테고리, 책 소개, 저자 소개, 저자 빈도, 출판사 빈도, 카테고리 빈도이고, 종속변수는 예측하고자 하는 출판 후 베스트셀러 여부이다. 독립변수 중 가격, 페이지 수, 출판 월, 저자 빈도, 출판사 빈도, 카테고리 빈도는 수치형 변수이다. 수치형 변수는 숫자로 표현되며, 산술적 계산이 가능한 변수로 데이터의 양적 속성을 나타낸다. 저자, 출판사, 카테고리는 범주형 변수이다. 범주형 변수는 고유한 값이나 범주로 표현되며, 각기 다른 그룹이나 분야를 나타내기 위해 사용된다. 제목, 책 소개, 저자 소개는 텍스트 변수이다. 텍스트 변수는 문자 데이터로 이루어져 있으며, 주로 자연어 처리를 통해 분석

〈Table 1〉 독립변수의 정의와 자료 유형

변수	정의	자료 유형
제목	책의 제목	텍스트
저자	책의 저자(복수일 경우 1저자만 사용)	범주형
출판사	도서를 출판한 출판사	범주형
가격	책의 가격	수치형
페이지 수	책의 페이지 수	수치형
출판 월	책이 출판된 월(1~12월)	수치형
카테고리	'소설/시/희곡' 내 하위 카테고리	범주형
책 소개	책의 주요 내용을 설명하는 글	텍스트
저자 소개	저자의 이력을 설명하는 글	텍스트
저자 빈도	도서의 저자가 최근 3년간 주별 베스트셀러 순위에 포함된 빈도	수치형
출판사 빈도	도서의 출판사가 최근 3년간 주별 베스트셀러 순위에 포함된 빈도	수치형
카테고리 빈도	도서의 카테고리가 최근 3년간 주별 베스트셀러 순위에 포함된 빈도	수치형

된다. 〈Table 1〉은 베스트셀러 여부 예측을 위해 사용한 독립변수의 정의와 자료 유형을 나타낸다. '자료 유형' 항목에서 텍스트는 텍스트 변수, 범주형은 범주형 변수, 수치형은 수치형 변수를 의미한다.

변수 선정 후 예측모형에 입력하기 위해 각 변수 데이터를 전처리하였다. 전처리는 데이터 분석이나 모델링을 수행하기 전에 데이터를 정제하고 변환하는 과정을 의미한다. 데이터에서 누락된 값인 결측치는 모형의 성능에 부정적인 영향을 미칠 수 있으므로 적절한 처리가 필요하다. 본 연구에서는 텍스트 변수의 결측치를 빈 문자열로, 수치형 변수의 결측치를 0으로 대체하였다. 수치형 변수에 속하는 가격, 페이지 수, 출판 월, 저자 빈도, 출판사 빈도, 카테고리 빈도는 스케일이 다른데, 이 차이로 인해 기계학습 알고리즘에서 특정 변수가 다른 변수보다 더 큰 영향을 미칠 수 있다. 이러한 문제를 방지하기 위해 Scikit-learn 패키지의 StandardScaler 라이브러리를 사용해 수치형 변수의 값을 표준화하였다. 표준화는 데이터의 평균을 0, 표준편차를 1로 변환하는 과정이다. 이는 데이터가 정규 분포를 따르지 않을 때도 사용할 수 있으며, 기계학습 알고리즘에서 널리 사용된다.

범주형 변수와 텍스트 변수는 기계학습 모형에 입력하기 위해 숫자 형태의 벡터로 변환하는 임베딩(embedding) 기법을 적용해야 한다(Bengio, Ducharme, and Vincent, 2000; Johnson, Murty, and Navakanth, 2024). 범주형 변수에 속하는 저자, 출판사, 카테고리는 Scikit-learn 패키지의 OneHotEncoder 라이브러리를 사용하여 단어 집합을 만들고, 표현하고 싶은 단어의 인덱스에는 1을, 다른 단어의 인덱스에는 0을 부여하는 방법인 원-핫 인코딩(One-Hot Encoding) 처리를 하였다(Karthiga, Usha, Raju, and Narasimhan, 2021; Okada, Ohzeki, and Taguchi, 2019).

텍스트 변수인 제목, 책 소개, 저자 소개는 TF-IDF (Term Frequency-Inverse Document Frequency) 기법을 적용하여 벡터로 변환하였다. TF-IDF는 기본적으로 단어 집합에서 각 단어에 빈도수를 부여하는 방식인데, 해당 문서(본 연구에서는 제목, 책 소개, 저자 소개) 내에 있는 각 단어의 중요도를 가중치로 부여한다(Havrlant & Kreinovich, 2017; Ramos, 2003). 본 연구는 TF-IDF 기법을 사용하기 위해 Scikit-learn 패키지의 TfidfVectorizer 라이브러리를 사용하였다. TF-IDF는 파라미터로 max_features 값을 설정할 수 있는데, 이는 벡터화할 때 사용할 최대 단어 수를 의미한다. 벡터화 과정에서 중요한 단어만 분석에 활용하기 위한 요소인데, 본 연구는 max_features 값을 10,000으로 설정하였다. 이는 단어 집합에서 빈도순으로 상위 10,000개의 단어만을 고려하여 벡터를 생성하는 것을 의미한다.

한편, 본 연구의 사용 데이터는 클래스 불균형 데이터이기 때문에 이 문제를 해결하기 위해 SMOTE 기법을 적용하였다. SMOTE는 소수 클래스의 데이터를 합성하여 데이터의 균형을 맞추는 기법으로, 기존의 소수 클래스 데이터 포인트 사이의 벡터를 따라 새로운 데이터를 생성한다(Chawla, Bowyer, Hall, and Kegelmeyer, 2002; Han, Wang, and Mao, 2005). 이를 통해 훈련 데이터에서 베스트셀러 클래스의 데이터양을 증가시켜 모형이 각 클래스에 대해 균형 잡힌 학습을 할 수 있도록 한다. SMOTE 기법은 데이터의 불균형을 완화하여 예측 모형의 성능을 향상하는 데 기여한다.

3.2 기계학습 예측모형

데이터 전처리 후 기계학습 알고리즘을 활용하여 예측

모형을 구현하였다. 예측 모델링에는 회귀(regression)와 분류(classification) 기법이 있는데 회귀 기법은 연속적인 숫자 값을 예측하는 데 사용되고, 분류기법은 두 개 이상의 범주 중 하나를 예측하는 데 사용된다. 본 연구는 베스트셀러 여부를 예측하는 실험을 진행하므로 분류기법을 선택하였다(Loh, 2011). 선행연구(문동지 외, 2020; 이중원&박철, 2021; 장동률&박민재, 2021; Lee, Ji, Kim, and Park, 2021; Lee, Kim, and Park, 2023)를 참고해 본 연구에서 사용한 모형은 확률기반 모형인 Naive Bayes(NB), 선형모형인 Logistic Regression, SVM, 트리기반 모형인 DT, RF, 부스팅 모형인 AdaBoost, CatBoost, GradientBoost, XGBoost, LightGBM이다.[1)]

확률기반 모형은 주어진 데이터가 특정 클래스에 속할 확률을 계산하여 분류를 수행한다. 대표적인 모형인 NB는 베이즈 정리를 활용한 분류 알고리즘이다. 베이즈 정리는 주어진 데이터에 관한 사전 확률(prior probability)과 데이터가 특정 클래스에 속할 조건부 확률(likelihood)을 기반으로 사후 확률(posterior probability)을 계산한다(Glickman and Van Dyk, 2007). 선형모형은 데이터 특성과 출력 사이의 선형 관계를 가정하여 예측을 수행한다. 대표적인 모형인 LR은 이진분류 문제를 해결하는 선형모형이다. 입력 특징의 선형조합을 통해 로짓(logit) 함수를 계산하고, 이 값을 시그모이드(sigmoid) 함수를 통해 0과 1 사이의 확률로 변환해 이 확률을 기반으로 분류한다(Peng, Lee, and Ingersoll, 2002).[2)] SVM은 두 클래스 간의 데이터를 선형적으로 구분하는 초평면(hyperplane: 데이터 공간에서 두 클래스를 나누는 경계)을 찾아 분류하는 알고리즘이다(Cortes and Vapnik, 1995).

트리기반 모형은 데이터 특징(feature)을 사용하여 의사결정 트리를 생성하고 예측을 수행한다. 대표적인 예로 DT와 RF가 있다. DT는 트리 구조를 사용하여 데이터를 분할하고 예측을 수행한다. 각 노드는 하나의 특징에 기반한 조건 분할을 수행하며, 리프 노드(leaf node: 트리의 마지막 노드)는 최종 예측을 나타낸다(Quinlan, 1986). RF는 여러 개의 결정 트리를 앙상블(여러 개의 모형을 결합하여 하나의 모형보다 더 좋은 성능을 도출하는 기법)하여 예측을 수행한다. 각각의 트리가 독립적으로 학습하고, 최종 예측은 트리의 예측을 평균 내거나 투표하는 방식으로 결정된다(Breiman, 2001).

부스팅 모형은 여러 약한 학습기(weak learner)를 순차적으로 학습하여 강한 학습기(strong learner)로 만드는 앙상블 학습 방법이다. 대표적으로 AdaBoost, CatBoost, GradientBoost, XGBoost, LightGBM 등이 있다. AdaBoost의 초기 모형은 데이터를 학습하고, 후속 모형은 이전 모형이 잘못 분류한 데이터에 더 많은 가중치를 부여하면서 학습한다. 이 과정을 반복하여 각 모형의 성능을 점진적으로 개선한다(Freund and Schapire, 1997). CatBoost는 카테고리형 데이터를 효율적으로 처리할 수 있는 알고리즘이다. 고유의 데이터 전처리 기법을 사용하여 카테고리형 특징을 자동으로 인코딩하고, 부스팅 과정에서 과적합을 방지하는 최적화 기법을 적용한다(Prokhorenkova et al., 2018). GradientBoost는 모형의 예측 오류를 점진적으로 줄이기 위해 순차적으로 학습기를 추가하는 알고리즘이다. 각 단계

1) 계산자원을 많이 필요로 하고, 추정의 안정성이 높지 않은 심층신경망 모형은 제외하였다.
2) 로지스틱 회귀에서 로짓 함수가 사용되며, 로짓은 확률 p를 로그 오즈(log odds)로 변환한 값이다. 시그모이드 함수는 로지스틱 회귀의 출력을 확률값으로 변환하는 데 사용된다(Yin et al., 2003).

에서 새로운 모형은 이전 모형의 잔차[3]를 예측하도록 학습된다. XGBoost는 성능과 효율성을 극대화한 GradientBoost 알고리즘이다. 병렬 학습을 통해 빠른 학습 속도를 제공하며, 과적합을 방지하기 위한 다양한 정규화 기법을 적용한다(Chen and Guestrin, 2016). LightGBM은 대용량 데이터 처리에 최적화되고 학습 속도를 향상시킨 GradientBoost 알고리즘이다. 의사결정 트리 알고리즘의 한 종류로, 트리의 리프 노드를 중심으로 나누는 방법인 리프 중심 트리 분할 알고리즘을 사용하여 효율성을 높인다(Ke et al., 2017).

예측모형의 구현을 위해 훈련자료(Training Data)와 모형 평가를 위한 시험자료(Test Data)를 각각 8:2의 비율로 나누어 훈련과 테스트를 진행하였다. 또한 기계학습에서 훈련자료를 과하게 학습하는 과대적합(overfitting)을 개선하기 위해 교차검증(Cross Validation)을 이용하였다(Bates, Hastie, and Tibshirani, 2023). 본 연구에서 제시하는 성능은 5겹 교차검증을 수행한 실험 결과의 평균값이다.

3.3 성능평가 지표

예측모형의 성능을 평가하기 위한 지표로는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수(F1-Score), 매튜 상관계수, ROC AUC(Receiver Operating Characteristic Area Under the Curve), PR AUC(Precision-Recall Area Under the Curve) 등 7개를 사용하였다(Lee et al., 2022; Lee et al., 2023; Lee and Park, 2024; Naidu, Zuva, and Sibanda, 2023; Oh, Kim, Lee, and Park, 2021; Varoquaux and Colliot, 2023).

정확도는 전체 예측 중에서 올바르게 예측된 샘플의 비율을 의미한다. 식(1)에서 TP는 True Positives로 실제 Positive인데 예측도 Positive인 경우, TN은 True Negatives로 실제 Negative인데 예측도 Negative인 경우, FP는 False Positives로 실제 Negative인데 예측은 Positive인 경우, FN은 False Negatives로 실제 Positive인데 예측은 Negative인 경우를 의미한다. 본 연구에서 Positive는 베스트셀러 클래스 1, Negative는 베스트셀러가 아닌 클래스 0이다.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

식 (2)에서 정밀도는 Positive로 예측한 샘플 중 실제로 Positive인 샘플의 비율을 나타낸다. 본 연구에서는 베스트셀러로 예측한 경우 중 실제로 베스트셀러인 경우의 비율이다.

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \quad (2)$$

식 (3)에서 재현율은 실제 Positive 샘플 중 올바르게 Positive로 예측한 샘플의 비율을 나타낸다. 본 연구에서는 실제 베스트셀러 중 베스트셀러로 예측한 경우의 비율이다.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

3) 예측모형에서 실제 값과 예측값의 차이를 나타내는 값이다(Friedman, 2001).

식 (4)에서 F1 점수는 정밀도와 재현율의 조화 평균이다.

$$F1-Score=2\times\frac{\mathrm{Pr}ecision\times Recall}{\mathrm{Pr}ecision+Recall} \quad (4)$$

식 (5)에서 매튜 상관계수는 이진분류 문제에서 예측의 정확도를 측정하는 상관계수이다. 아래 식에서 나타난 바와 같이 TP, TN, FP, FN 등 혼동행렬(모형이 예측한 결과와 실제 값을 비교하여 각 예측의 정확성을 시각화한 것)의 모든 구성요소를 반영해 계산한다. -1에서 1 사이의 값을 가지며, 1은 완벽한 예측, 0은 무작위 예측, -1은 완전히 잘못된 예측을 의미한다. 클래스 불균형 문제에 강건하다는 특징이 있다.

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

ROC AUC에서 ROC 곡선은 다양한 임곗값에서의 재현율(TPR; True Positive Rate)과 거짓 양성 비율(FPR; False Positive Rate)을 나타낸 곡선이다. ROC 곡선은 TPR을 Y축에, FPR을 X축에 두고 여러 임곗값에서의 값을 그래프로 나타낸다. AUC는 이 ROC 곡선 아래의 면적을 의미하며, 모형의 전체적인 판별력을 평가하는 데 유용하다. ROC AUC는 0에서 1 사이의 값을 가지는데, 1에 가까울수록 좋은 성능을 나타낸다.

PR AUC에서 PR 곡선은 다양한 임곗값에서의 정밀도(Precision)와 재현율(Recall)을 나타낸 곡선이다. PR 곡선은 정밀도를 Y축에, 재현율을 X축에 두고 여러 임곗값에서의 값을 그래프로 나타낸다. PR AUC는 이 PR 곡선 아래의 면적을 의미하며, 특히 클래스 불균형이 심한 경우에 유용하다. PR AUC는 Positive 클래스의 예측성능을 더 잘 평가하므로 클래스 불균형 데이터에서 모형의 성능을 더 정확하게 반영할 수 있다. PR AUC도 0에서 1 사이의 값을 가지는데, 1에 가까울수록 좋은 성능을 나타낸다.

3.4 XAI 기법

베스트셀러 예측에 영향을 미친 요인을 분석하기 위해, 기계학습 모형의 예측이나 결정 과정을 설명하는 XAI 기법을 사용하였다(Guidotti et al., 2018). XAI는 기계학습 모형의 투명성을 높이고, 사용자가 모형의 작동 방식을 명확히 설명하는 데 도움을 준다. XAI의 주요 특징으로는 AI 모형이 예측이나 결정을 어떻게 내렸는지 그 과정을 설명할 수 있는 설명 가능성, 모형의 작동을 해석하고 그 과정이 어떻게 이루어졌는지 이해할 수 있는 이해 가능성, AI 모형이 완전히 공개되어 작동 과정을 이해할 수 있는 투명성, 결과와 관련된 의미 있는 설명을 제공할 수 있는 해석 가능성이 있다(Arrieta et al., 2020). 본 연구는 여러 XAI 기법 중 특성 중요도 기법과 SHAP 기법을 사용하였다.

특성 중요도 기법은 모형 예측 시 각 특성이 얼마나 중요한 역할을 하는지를 평가하는 방법이다(Zien et al., 2009). 특성 중요도 분석을 통해 모형의 성능을 최적화하고, 어떤 특성이 예측에 가장 큰 영향을 미치는지 파악할 수 있다. DT와 RF같은 트리 기반 모형은 특성 중요도를 평가하기에 적합한 구조이다. 각 분할 단계에서 Gini Impurity(결정트리에서 각 노드가 특정 클래스에 순수하게 분류되어 있는지를 나타내는 지표)나 정보 이득(결정트리에서 노드 분할 시 특정 특성이 데이터를 잘 분류하는지

측정하는 지표)을 기준으로 특성을 선택하므로 분할에 얼마나 기여했는지 누적하여 중요도를 계산한다(Wang et al., 2024).

SHAP 기법은 게임이론에서 유래한 샤플리 값(shapley value)[4]을 활용한 기법으로 각 특성이 기계학습 모형의 예측에 얼마나 기여하는지를 평가한다. SHAP 값은 각 특징이 예측값에 기여하는 정도를 나타내며, 모든 가능한 특성 조합에 관한 평균적 기여도를 계산하여 공정하게 기여도를 분배한다(Lundberg and Lee, 2017). SHAP값을 시각화하면 특정 예측에 어떤 특징이 긍정적 또는 부정적으로 작용했는지 명확히 알 수 있다. 이는 모형이 예측을 내린 이유를 직관적으로 이해할 수 있게 하여 모형의 신뢰성을 높이고 모형 예측과정을 설명할 때 유용하다(Wang, Liang, Hancock, and Khoshgoftaar, 2024). 또한 SHAP 기법은 모형의 전체적인 중요도와 특정 예측값에 관한 개별적 기여도를 동시에 계산할 수 있다.

평가 방식과 상호작용 반영 여부에 차이가 있는데, 특성 중요도 기법은 주로 전체 모형의 전역적 기여도(global contribution)를 평가하므로 예측마다 특성이 미치는 영향은 반영하지 못하지만, SHAP 기법은 각 예측에 대해 지역적 기여도(local contribution)를 평가하여 모든 특성 조합에서 개별 특성의 기여도를 계산한다. 특성 중요도 기법은 주로 개별 특성의 단독 중요도를 평가하고 특성 간 상호작용은 충분히 고려하지 않지만, SHAP 기법은 특성 간 상호작용을 모든 가능한 조합에서 평가해 기여도를 계산한다(Guidotti et al., 2018).

Ⅳ. 결과 분석

4.1 베스트셀러 예측성능 비교

10개의 기계학습 모형에 대한 성능평가 결과를 제시한다. 〈Table 2〉는 10개 기계학습 모형의 성능지표 중 정밀도, 재현율, F1 점수, 정확도, 매튜 상관계수를 비교한 표이다. 표에서 class 항목의 'non-best', 'best'는 각각 베스트셀러가 아닌 클래스, 베스트셀러인 클래스를 의미한다. 〈Figure 1〉은 ROC AUC, PR AUC 값을 비교한 그림이다. X축은 10개 모형을 나타내고, Y축은 성능 값을 나타낸다. 파란색 막대(bar)는 ROC AUC 값, 주황색은 PR AUC 값을 나타낸다. X축에서 AB는 AdaBoost, CB는 CatBoost, GB는 GradientBoost, XGB는 XGBoost, LGBM는 LightGBM을 의미한다.

일반적으로 성능을 평가할 때는 정확도를 기준으로 하지만, 클래스가 불균형한 경우에는 정확도만으로는 성능을 제대로 평가하기 어려우므로 재현율, F1 점수, 매튜 상관계수와 같은 지표를 종합적으로 평가해야 한다. 특히, 매튜 상관계수는 클래스 불균형이 있는 이진 분류 문제에서 다른 성능지표와 달리 모든 성능 요소를 고려하므로 클래스에 치우치지 않고 모형의 전반적인 성능을 평가할 수 있다(Rainio et al., 2024). 성능평가 결과, 부스팅 모형과 선형모형의 성능이 좋으며, 그중에서도 LightGBM이 MCC 0.6603, 정확도 0.8889, ROC AUC 0.9243, PR AUC 0.7926을 보여 전체적으로 가장 좋은 성능을 보인다. ROC AUC는 이진 분류 모형의 성능을 평가하는 지표 중 하나로 모형이 양성클래스와

4) 협동 게임이론에서 참여자가 협력하여 얻은 성과를 공정하게 분배하는 방법(Fatima et al., 2008).

〈Table 2〉 모형 성능 비교(Precision, Recall, F1-score, Accuracy, MCC)

models	class	Precision	Recall	F1-score	Accuracy	MCC
NB	non-best	0.9448	0.8299	0.8836	0.8269	0.5695
	best	0.5577	0.8155	0.6623		
LR	non-best	0.9318	0.9126	0.9221	0.8779	0.6407
	best	0.6917	0.7459	0.7177		
SVM	non-best	0.9151	0.9264	0.9207	0.8737	0.6102
	best	0.7063	0.6729	0.6891		
DT	non-best	0.9101	0.8987	0.9044	0.8495	0.5513
	best	0.6321	0.662	0.6465		
RF	non-best	0.8882	0.9343	0.9107	0.8549	0.5299
	best	0.6888	0.5523	0.6129		
AdaBoost	non-best	0.9218	0.8913	0.9063	0.8541	0.5786
	best	0.6329	0.7123	0.6701		
CatBoost	non-best	0.93	0.8782	0.9033	0.8541	0.5786
	best	0.6176	0.7482	0.6701		
GradientBoost	non-best	0.931	0.9031	0.9168	0.8703	0.6238
	best	0.669	0.7454	0.7051		
XGBoost	non-best	0.9213	0.9327	0.927	0.8836	0.641
	best	0.6329	0.6969	0.7136		
LightGBM	**non-best**	**0.9275**	**0.9326**	**0.93**	**0.8889**	**0.6603**
	best	**0.7381**	**0.7225**	**0.7301**		

〈Figure 1〉 성능 비교(ROC AUC, PR AUC)

음성클래스를 얼마나 잘 구별하는지를 측정한다. 연구에 따르면(Fawcett, 2006; Liu et al., 2022), ROC AUC 값이 0.9 이상인 이진분류 모형은 매우 우수한 성능으로 간주한다. LightGBM의 ROC AUC 값은 0.9243으로 선행연구의 기준에 비추어 볼 때 예측성능이 우수하다고 볼 수 있다. 또한 불균

형 클래스의 이진 분류를 연구한 선행연구(Chicco and Jurman, 2020; Vihinen, 2012)는 MCC 값이 0.6 이상이면 데이터 불균형 상황에서 상당히 우수한 성능을 나타내는 지표로 보고하고 있다. 이에 따르면 MCC 값이 0.6603인 LightGBM은 데이터 불균형 상황에서도 예측을 잘 수행하는 우수한 모형이라고 평가할 수 있다. LightGBM 다음으로 동일한 부스팅 모형인 XGBoost가 MCC 0.641로 우수한 성능을 보였으며, 전통적인 방법인 LR 또한 MCC 0.6407로 균형 잡힌 성능을 제공함을 확인할 수 있다. 예측모형의 강건성을 살펴보기 위해 베스트셀러의 기준을 500위, 300위, 100위로 설정하고 LightGBM을 실행하였고, 그 결과 이 기준에서도 유의미한 예측성능을 확인할 수 있었다.[5)]

본 연구의 첫 번째 목적은 기계학습을 이용해 출판 전 신간도서의 베스트셀러 여부를 예측하는 모형을 구현하는 것이다. 여러 성능지표를 분석한 결과 LightGBM 등의 기계학습 알고리즘으로 베스트셀러 예측모형을 구현할 수 있음을 확인하였다.

4.2 베스트셀러 예측 영향요인 분석

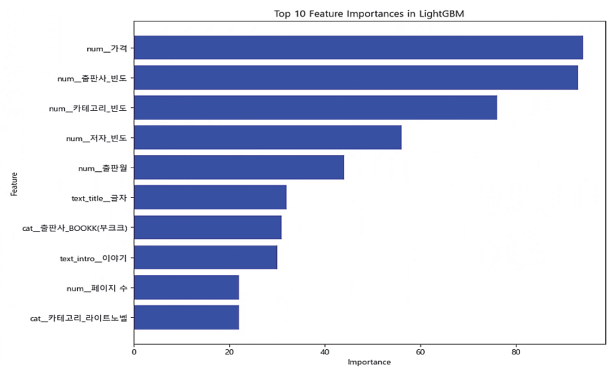
본 연구의 두 번째 목적은 베스트셀러 예측에 영향을 미치는 요인을 분석하는 것이다. 이 분석을 통해 온라인 서점은 소비자의 구매행동을 이해하고, 이를 바탕으로 마케팅 전략을 수립할 수 있다. 본 연구는 베스트셀러 예측에 영향을 미치는 요인을 분석하기 위해 특성 중요도 기법과 SHAP 기법을 사용하였다.

4.2.1 특성 중요도 기법

특성 중요도 기법은 해당 특성이 모형의 예측에 얼마나 기여하는지를 나타낸다. 이는 곧 해당 특성이 베스트셀러 여부를 예측하는 데 있어서 중요한 정보를 제공한다는 것을 의미한다. 특성 중요도는 모형 유형에 따라 다양한 방법으로 계산될 수 있는데 LightGBM과 같이 트리를 활용한 모형에서는 분기에서의 정보 이득을 통해 중요도를 계산한다. 〈Figure 2〉는 베스트 성능을 보인 LightGBM 모형의 상위 10개 특성 중요도를 나타낸다. 그래프에서 X축의 수치는 특성 중요도 값이다. 수치가 클수록 해당 특성이 예측성능에 더 기여함을 의미한다. Y축에 나타나 있는 것은 예측모형에서 사용한 특성인데, 특성 이름 앞에 붙어 있는 num, text, cat는 각각 수치형 변수, 텍스트 변수, 범주형 변수를 의미한다. 모형에 값을 입력할 때 변수 유형 구분을 위해 표시한 것이다. 예를 들어 'num_가격'은 수치형 변수인 가격, 'text_title_글자'는 텍스트 변수인 제목에 쓰인 '글자'라는 단어, 'cat_출판사_BOOKK(부크크)'는 범주형 변수인 출판사 중 부크크 출판사를 의미한다. 제목, 책 소개, 저자 소개 등 텍스트 변수는 벡터화를 위해 단어별로 분리되었으므로 단어가 특성으로 입력되었다.

〈Figure 2〉에서 책의 가격이 가장 중요한 특성으로 나타났다. 가격이 예측에 상당한 영향을 미치며, 특정 가격대의 책이 베스트셀러가 될 가능성이 크고, 소비자의 도서구매 결정에 큰 영향을 미침을 보여준다. 출판사 빈도는 두 번째 중요한 특성으로 나타났

5) 클래스가 불균형한 데이터에 관한 이진 분류 모형에서는 ROC AUC 값과 MCC 값이 중요하다. 500위, 300위, 100위 기준 ROC AUC 값은 각각 0.9119, 0.9180, 0.9510으로 1,000위 기준 ROC AUC 값(0.9243)과 유사하거나 약간 더 높다. 500위, 300위, 100위 기준 MCC 값은 각각 0.5478, 0.5171, 0.5272이다. MCC 값은 -1에서 1 사이의 값을 가지고 있고, 선행연구(Chicco and Jurman, 2020; Vihinen, 2012)에서는 불균형한 상황에서 MCC 값이 0.5 이상이면 의미 있는 분류 성능을 보인다고 해석한다. 이처럼 ROC AUC 값과 MCC 값을 비교하였을 때 모형이 500위, 300위, 100위 기준에서도 유의미한 예측성능을 보인다고 판단할 수 있다.



〈Figure 2〉 LightGBM의 특성 중요도 그래프

다. 이는 베스트셀러 순위에 자주 등장한 출판사의 책이 베스트셀러가 될 가능성이 크다는 것을 의미한다. 특정 출판사의 책이 더 많이 베스트셀러에 오를 가능성이 있다는 것을 보여준다. 카테고리 빈도는 세 번째 중요한 특성으로 나타났다. 이는 베스트셀러 순위에 자주 등장한 카테고리의 책이 베스트셀러가 될 가능성이 크다는 것을 의미하며, 특정 카테고리의 책이 더 자주 베스트셀러에 오를 가능성이 있음을 보여준다. 저자 빈도는 네 번째 중요한 특성으로 나타나는데, 베스트셀러 순위에 자주 등장한 저자가 쓴 책이 베스트셀러가 될 가능성이 크고, 특정 저자의 책이 자주 베스트셀러에 오르는 경향이 있다는 것을 나타낸다.

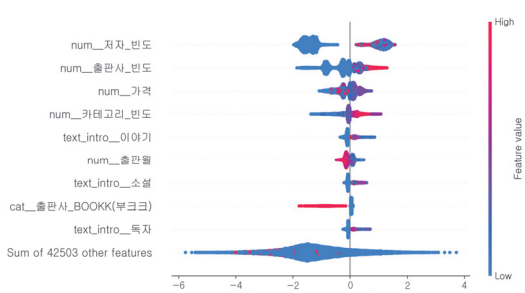
이외에 출판 월, 제목에 포함된 '글자'라는 단어, 출판사 '부크크', 책 소개에 포함된 '이야기'라는 단어, 책의 페이지 수, '라이트노벨' 카테고리도 중요한 특성으로 나타났다. 출판 월이 중요한 특성으로 나타난 것은 특정 시기에 출판된 책이 베스트셀러가 될 가능성이 있다는 것을 의미하며, 계절적 트렌드나 마케팅 전략과 관련이 있을 수 있다. 제목에 포함된 '글자'라는 단어, 책 소개에 포함된 '이야기'라는 단어의 중요도가 높게 나타난 것은 책 제목과 책 소개의 특정 단어가 베스트셀러 여부와 상관관계가 있음을 의미한다. '부크크'라는 출판사가 중요한 특성으로 나타난 것은 이 출판사에서 출판된 책이 베스트셀러 예측에 영향을 미친다는 것을 시사한다. 책 페이지 수가 중요한 특성으로 나타난 것은 책의 분량이 독자의 구매 결정에 영향을 미칠 수 있음을 의미하고, '라이트노벨' 카테고리가 중요한 특성으로 나타난 것은 이 장르의 책과 베스트셀러 여부 간에 상관관계가 높다는 것을 시사한다.

특성 중요도 그래프를 통해 가격, 출판사 빈도, 카테고리 빈도, 저자 빈도와 같은 특성이 베스트셀러 예측에서 중요한 역할을 한다는 것을 알 수 있다. 가격이 가장 중요한 특성이며, 소비자의 구매결정에 큰 영향을 미친다. 출판 시기, 제목과 책 소개의 특정 단어, 특정 출판사나 카테고리, 페이지 수도 예측에 중요한 요소로 작용한다.

4.2.2 SHAP 기법

특성 중요도 그래프는 모형의 전반적인 특성 중요도를 이해하는 데 유용하지만, 각 특성이 예측값에 대해 미치는 영향을 구체적으로 보여주는 데는 SHAP

그래프가 유용하다. SHAP는 트리 기반 모형, 선형 모형, 딥러닝 모형까지 다양한 모형에서 사용될 수 있다. 〈Figure 3〉은 가장 좋은 성능을 보인 LightGBM 모형의 상위 10개 주요 특성에 관한 SHAP 값을 시각화한 Summary Plot으로, 신간도서 베스트셀러 예측에 영향을 미친 각 특성의 중요도를 파악할 수 있다. X축은 SHAP 값으로 각 특성이 예측에 미치는 영향을 나타낸다. SHAP 값이 0 이상이면 긍정적인 영향, 0 이하이면 부정적인 영향을 의미한다. Y축은 특성 이름을 나타낸다. 색상은 특성값의 크기를 나타내는데, 파란색은 낮은 값, 빨간색은 높은 값을 의미한다.



〈Figure 3〉 LightGBM의 Summary Plot

〈Figure 3〉을 보면 저자 빈도가 베스트셀러 예측에 가장 중요한 특성으로 나타났다. 저자 빈도가 높은 경우(붉은색 점)는 SHAP 값이 양의 영역(+)에 있어 예측에 긍정적인 영향을 미치고, 반대로 저자 빈도가 낮은 경우(파란색 점)는 SHAP 값이 양의 영역, 음의 영역(-) 둘 다에 있지만, 음의 영역에 좀 더 많아 예측에 부정적인 영향을 미친다고 볼 수 있다. 이는 베스트셀러 순위에 자주 등장한 저자의 책이 베스트셀러가 될 가능성이 크고, 반대로 자주 등장하지 않은 저자의 책은 베스트셀러가 될 가능성이 작다는 것을 의미한다. 이를 통해 베스트셀러의 가장 중요한 요인은 저자 빈도이고, 독자는 베스트셀러를 많이 배출한 저자의 책을 구매하려는 경향이 있다고 해석할 수 있다. Wang et al.(2019)도 문학 장르에서 저자의 이전 판매기록과 인지도가 도서판매에 큰 영향을 미치며, 이는 독자가 유명 작가나 과거에 성공한 작가의 책을 선호하는 경향이 반영된 결과라고 분석하였다. 또한 캐나다 도서 산업을 지원하는 비영리 기관인 BookNet Canada가 조사한 보고서에 의하면 도서구매 결정에 영향을 미치는 요소 중 '작가와의 친숙함'이 1순위로 꼽혔다.

다음으로 출판사 빈도가 중요한 특성으로 나타났다. 출판사 빈도가 높은 경우(붉은색 점)는 SHAP 값이 양의 영역에 있어 예측에 긍정적인 영향을 미치고, 반대로 출판사 빈도가 낮은 경우(파란색 점)는 대부분 SHAP 값이 음의 영역에 있어 예측에 부정적인 영향을 미친다고 볼 수 있다. 이는 베스트셀러를 많이 출판한 출판사의 책이 베스트셀러가 될 가능성이 크고, 그렇지 못한 출판사의 책은 베스트셀러가 될 가능성이 작다는 것을 의미한다. 이를 통해 출판사 빈도도 베스트셀러의 매우 중요한 요인이고, 독자는 베스트셀러를 많이 출판한 출판사의 책을 구매하려는 경향이 있다고 해석할 수 있다. 이는 출판사의 명성, 자원, 마케팅 범위가 도서판매에 영향을 미칠 수 있음을 함의하는데, Wang et al.(2019)은 저자와 출판사는 모두 도서판매에 큰 영향을 미치며, 문학 장르에서는 저자의 인지도와 과거 판매 이력이 더 중요하고 비문학 장르에서는 출판사의 평판이 판매에 더 영향을 미친다고 분석하였다. 문학 장르 도서를 중심으로 분석한 본 연구에서도 저자 빈도가 출판사 빈도보다 베스트셀러 예측에 더 영향을 미치는 특성으로 나타나 선행연구의 결과와 유사함을 보인다.

책의 가격이 높은 경우(붉은색 점)는 SHAP 값이

음의 영역에 있어 예측에 부정적인 영향을 미치는 것으로 나타난다. 가격이 낮은 경우(파란색 점)도 SHAP 값이 대체로 음의 영역에 분포하여 예측에 부정적인 영향을 미치는 것으로 나타난다. 가격이 중간인 경우(보라색 점)는 양의 영역에 위치해 가격이 중간일 때 예측에 긍정적인 영향을 미친다고 볼 수 있다. 이는 책의 가격이 너무 높거나 낮은 경우 베스트셀러가 될 가능성이 작고, 중간 가격대일 경우 베스트셀러가 될 가능성이 크다는 것을 의미한다. 이러한 특징을 통해 독자는 너무 높거나 낮은 가격의 책보다는 중간 가격대의 책을 구매하려는 경향이 있다고 해석할 수 있다. 가격 변수의 전체 분포를 분석하면 보라색에 해당하는 중간 가격대의 실제 가격은 10,000~16,000원이다. 즉, 10,000원과 16,000원 사이에 있는 가격을 가진 도서가 베스트셀러가 될 가능성이 크다고 해석할 수 있다. 가격과 소비자 구매 의도를 분석한 선행연구(Chang and Wildt, 1994; Chernev, 2003; Rao and Monroe, 1989)의 결과에 기반하면 소비자는 낮은 가격대의 책에 관해서는 품질이 낮을 것이라고 인식해 구매하지 않는 경향이 있고, 높은 가격대의 책에 관해서는 가격이 부담되어 구매하지 않는 경향이 있으므로 중간 가격대의 책을 구매하는 경향이 높다고 해석된다.

카테고리 빈도도 중요한 특성으로 나타났다. 카테고리 빈도가 높은 경우(붉은색 점) SHAP 값이 양의 영역에 있어 예측에 긍정적인 영향을 미치고, 반면 카테고리 빈도가 낮은 경우(파란색 점) 대부분 음의 영역에 있어 예측에 부정적인 영향을 미친다고 볼 수 있다. 이는 베스트셀러 순위에 자주 등장한 카테고리의 책이 베스트셀러가 될 가능성이 크고, 그렇지 않은 카테고리의 책은 베스트셀러가 될 가능성이 작다는 것을 의미한다. 이를 통해 독자는 다른 독자가 많이 찾는 카테고리의 책을 선호하고 구매하려는 경향이 있다고 해석할 수 있다. 카테고리 빈도가 높은 장르는 판타지, 무협, 추리, 미스터리 등 '장르 소설'로 나타났다. Wang et al.(2019)도 카테고리를 도서 성공에 큰 영향을 미치는 요소로 언급하였는데, 특히 추리, 미스터리 장르에서는 작가의 과거 판매 기록이 영향을 미친다고 분석하였다.

책 소개 글에 나온 '이야기', '소설', '독자'는 그 값이 큰 경우(붉은색 점) SHAP 값이 크지는 않지만 대체로 양의 영역에 있어 예측에 긍정적인 영향을 미친다고 볼 수 있다. 이는 책 소개 글에 '이야기', '소설', '독자' 단어가 많이 들어간 책이 베스트셀러가 될 가능성이 있다는 것을 의미한다. '이야기', '소설', '독자' 단어가 많이 들어간 책은 '소설/시/희곡' 카테고리의 책 중 '소설' 카테고리에 해당하는 책일 가능성이 크다. 독자가 문학 분야의 책 중에서 시나 희곡보다는 소설을 선호하고 구매하려는 경향이 있다고 해석할 수 있다.

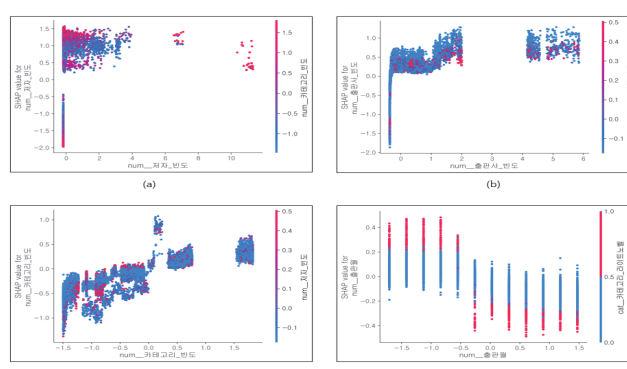
출판 월도 중요한 특성으로 나타났다. 출판 월은 값이 큰 경우(붉은색 점) SHAP 값이 주로 음의 영역에 있어 예측에 부정적인 영향을 미치고, 반대로 작은 경우(파란색 점)는 양의 영역에 있어 예측에 긍정적인 영향을 미친다고 볼 수 있다. 이는 값이 작은 월, 즉 연초나 상반기에 출판된 책이 베스트셀러가 될 가능성이 크고, 값이 큰 월, 즉 연말이나 하반기에 출판된 책은 베스트셀러가 될 가능성이 작다는 것을 의미한다. 이를 통해 독자가 연말이나 하반기보다는 연초나 상반기에 책을 구매하려는 경향이 있다고 해석할 수 있다. 소비자행동을 분석한 선행연구에서는 새해가 시작될 때 많은 사람이 새로운 목표를 설정하는 경향이 있고, 이는 특히 건강, 자기계발, 재정계획 등의 영역에서 나타나며 관련 상품이나 서비스에 관한 소비행동에도 영향을 미친다고 분석하였다(Gollwitzer and Sheeran, 2009; Scott and Williams, 2023).

이러한 분석에 의하면 새해에 독서나 자기계발을 목표로 설정한 소비자가 책을 구매해 연초나 상반기의 책 구매 경향이 높아진 것으로 해석할 수 있다.

출판사 부크크(BOOKK)는 값이 큰 경우(붉은색 점) SHAP 값이 음의 영역에 있어 예측에 부정적인 영향을 미친다고 볼 수 있다. 이는 부크크 출판사에서 출판된 책이 베스트셀러가 될 가능성이 작다는 것을 의미한다. 부크크 출판사는 개인이 쉽게 책을 출판할 수 있도록 돕는 자가출판(Self-Publishing) 플랫폼 서비스를 운영하고 이 서비스를 통해 만들어진 책을 출판 및 유통한다. 부크크 출판사는 자가출판 플랫폼 서비스를 통해 신속히 많은 책을 출판할 수 있지만, 대형 출판사에 비해 유명 저자가 적고, 마케팅에 투자하기 어려우므로 출판된 책이 베스트셀러가 될 가능성이 작다고 해석할 수 있다. 앞에서 살펴본 출판사 빈도의 결과 해석과 결합해 분석하면 베스트셀러를 많이 출판한 대형 출판사의 책은 베스트셀러가 될 가능성이 크지만, 자가출판한 책은 베스트셀러가 될 가능성이 작다고 볼 수 있다.

〈Figure 4〉는 앞에서 살펴본 특성 중 저자 빈도, 출판사 빈도, 카테고리 빈도, 출판 월의 Dependency Plot을 나타낸다. Dependency Plot은 단일 특성과 모형 예측 간의 관계를 좀 더 상세하게 나타내고, 단일 특성과 연관이 높다고 판단되는 다른 특성과의 관계를 보여준다. Dependency Plot의 X축은 단일 특성의 값, Y축은 SHAP 값을 의미하고, 색상은 상호연관이 높다고 판단되는 다른 특성값의 크기를 나타낸다.

〈Figure 4〉의 Panel (a)는 저자 빈도에 관한 Dependency Plot이다. 저자 빈도가 0에 가까우면 SHAP 값이 양의 영역과 음의 영역에 다 분포하지만, 저자 빈도가 높아지면 SHAP 값은 모두 양의 영역에 있다. 이는 모형이 베스트셀러를 자주 출판한 저자의 책에 대해 베스트셀러 가능성을 크게 예측한다는 것을 의미한다. 상호연관이 높은 특성은 카테고리 빈도로 나타났는데, 저자 빈도가 양의 영역에 위치하면서 카테고리 빈도가 높은 경우(빨간색 점) SHAP 값도 모두 양의 영역에 있다. 이는 베스트셀러를 많이 집필한 저자가 베스트셀러가 많이 나온 카테고리의 책을 출판하면 베스트셀러가 될 가능성이 크다는 것을 의미한다. Panel (b)는 출판사 빈도에 관한 Dependency Plot이다. 출판사 빈도가 0에 가까우면 SHAP 값이 음의 영역에 위치하고, 출판사 빈도가 높아질수록 SHAP 값이 커지고 양의 영역에 있다.



〈Figure 4〉 LightGBM의 Dependency Plot

이는 모형이 베스트셀러를 자주 낸 출판사의 책에 대해 베스트셀러 가능성을 크게 예측한다는 것을 의미한다. 상호연관이 높은 특성으로 저자 빈도가 나타났는데, 저자 빈도가 높고(빨간색 점) 출판사 빈도가 높은 경우 출판사 빈도가 낮은 경우보다 SHAP 값이 더 높아지는 경향을 보인다. 이는 베스트셀러를 많이 출판한 출판사에서 베스트셀러를 많이 집필한 저자가 책을 출판할 때 베스트셀러 가능성이 더욱 커진다는 것을 의미한다. Panel (c)는 카테고리 빈도에 관한 Dependency Plot이다. 카테고리 빈도가 낮으면 SHAP 값이 음의 영역에 위치하지만, 카테고리 빈도가 높아질수록 SHAP 값은 양의 영역에 있다. 이는 모형이 베스트셀러가 자주 나온 카테고리의 책에 대해 베스트셀러 가능성을 크게 예측한다는 것을 의미한다. 상호연관이 높은 특성은 저자 빈도로 나타났는데, 카테고리 빈도가 높으면서 저자 빈도도 동시에 높은 경우(빨간색 점) SHAP 값이 더 높아지는 경향을 보인다. 이는 베스트셀러가 많이 나온 카테고리에서 베스트셀러를 많이 집필한 저자가 책을 출판할 때 베스트셀러 가능성이 더욱 커진다는 것을 의미한다. Panel (d)는 출판 월에 관한 Dependency Plot이다. 출판 월값이 작은 경우(1월부터 5월) 상대적으로 SHAP 값이 양수에 있어 모형이 이 시기에 출판된 도서의 베스트셀러 가능성을 크게 예측하고, 반대로 출판 월값이 큰 경우(6월부터 12월) 상대적으로 SHAP 값이 음수에 있어 모형이 이 시기에 출판된 도서의 베스트셀러 가능성을 작게 예측한다는 것을 알 수 있다. 출판 시기와 연관이 높은 특성으로 라이트노벨 카테고리가 나타났는데, 라이트노벨 카테고리 도서(빨간색 점)는 출판 월값이 작을 때 베스트셀러 가능성이 커지고(SHAP 값이 큼), 반대로 출판 월값이 클 때 베스트셀러 가능성이 작아진다(SHAP 값이 낮음). 이러한 특징을 통해 라이트노벨 카테고리 도서는 출판 월이 1~5월일 때 상대적으로 베스트셀러 가능성이 크다고 해석할 수 있다.

특성 중요도 기법과 SHAP 기법의 결과를 보면 중요한 특성으로 도출되는 변수나 중요도 순위에 차이가 있다. 두 기법이 중요도를 계산하는 방식이 다르기 때문이다. 앞에서도 기술했듯이 특성 중요도 기법은 각 특성이 모형의 분할 노드에서 얼마나 자주 사용되었는지, 또는 해당 특성이 모형의 성능에 얼마나 기여했는지를 계산하지만, SHAP 기법에서 SHAP 값은 게임이론에 기반해 각 특성이 예측결과에 기여하는 정도를 공정하게 분배해 계산된다. 그러나 이러한 계산 방식의 차이에도 불구하고 저자 빈도, 출판사 빈도, 카테고리 빈도, 가격, 출판 월은 두 기법의 결과에서 모두 상위 중요도로 나타난다. 이는 이 특성이 출판 전 신간도서의 베스트셀러 여부를 예측하는 데 결정적인 영향을 미치는 요인임을 보여준다.

Ⅴ. 결론 및 함의

본 연구는 출판 전 신간도서의 베스트셀러 여부를 예측하는 모형을 제시하고 예측에 영향을 미치는 요인을 분석하는 데 목적을 두었다. 기계학습 기법을 활용하여 다양한 예측모형을 구현하고 정확도, 정밀도, 재현율, F1 점수, 매튜 상관계수, ROC AUC, PR AUC 등 여러 성능지표를 비교하였다. 부스팅 모형과 선형모형이 좋은 성능을 보였고, 그중에서도 LightGBM이 베스트셀러 예측에서 가장 높은 성능을 보였다. 특성 중요도 기법, SHAP 기법을 이용해 베스트셀러 예측에 영향을 미친 요인을 분석한 결과, 저자 빈도, 출판사 빈도, 카테고리 빈도, 가격, 출판

월 등이 예측에 영향을 많이 미치는 것으로 나타났다. 이전 베스트셀러 순위에 자주 등장한 저자일수록, 이전 베스트셀러 순위에 자주 등장한 출판사일수록, 이전 베스트셀러 순위에 자주 등장한 카테고리일수록 베스트셀러 가능성이 크다고 볼 수 있다. 또한 책 가격의 경우 중간 가격대의 책이 베스트셀러 가능성이 커지는 경향을 보였고, 출판 월의 경우 연초나 상반기에 출판된 책이 베스트셀러 가능성이 커지는 경향을 보였다.

기존에 인공지능 기술을 이용해 도서 판매량이나 순위, 베스트셀러 여부 예측을 수행한 선행연구는 주로 출판 후 데이터(리뷰, 평점, 판매순위 등)를 활용하여 도서 판매량이나 순위 예측모형을 제시한다. 이러한 접근은 신간도서의 판매를 예측할 때 콜드 스타트를 발생시킨다. 신간도서에는 리뷰, 평점, 판매순위 등의 데이터가 존재하지 않으므로 이 데이터를 사용하는 예측모형은 신간도서의 판매량이나 베스트셀러 여부를 예측하는 데 한계가 있다. 본 연구는 출판 후 데이터를 사용하지 않고 신간도서의 메타 데이터를 기반으로 베스트셀러 여부를 예측하는 모형을 제안하여 콜드 스타트를 해결한다는 점에서 선행연구와 차별되는 학문적 의의가 있다. 또한 본 연구는 특성 중요도 기법, SHAP 기법 등 국내 경영학 및 마케팅 연구에서 많이 활용되지 않은 '설명가능한 인공지능' 기법을 활용해 베스트셀러 예측에 영향을 주는 요인을 분석하였다는 점에서 학문적 시사점을 가진다.

콜드 스타트 해결을 위한 본 연구의 제안은 실용적 의의가 있다. 본 연구는 온라인 서점이 신간도서의 성공 가능성을 사전에 판단하여 마케팅 자원을 효율적으로 배분하고, 소비자의 구매행동을 기반으로 효과적인 마케팅 전략을 수립하는 데 기여한다. 구체적으로 살펴보면 다음과 같다. 첫째, 온라인 서점은 본 연구의 예측모형을 활용해 베스트셀러 가능성이 큰 책을 선별하고 해당 도서에 마케팅 예산을 투입함으로써 수익률을 개선할 수 있다. 둘째, 베스트셀러 가능성이 큰 책에 대해 사전 예약, 할인 쿠폰 등 프로모션을 기획하거나 해당 책을 주요 화면이나 추천 목록 상단에 배치하여 더 많은 노출을 유도하고 초기 판매량을 높일 수 있다. 셋째, 베스트셀러 가능성이 큰 책에 관한 수요를 사전에 예측하여 재고를 적절하게 확보함으로써 품절을 방지하고 판매기회를 확대할 수 있다. 넷째, 고객에게 베스트셀러를 추천해 고객 만족도를 높이고 충성 고객을 확보할 수 있다. 다섯째, 베스트셀러 예측 영향요인 분석을 통해 소비자의 구매행동에 영향을 미치는 요소를 고려한 마케팅 전략을 수립할 수 있다. 예를 들어 저자 빈도, 출판사 빈도, 카테고리 빈도가 도서구매에 영향을 미친다는 분석을 기반으로 블로그 포스트, 이메일 뉴스레터, 소셜미디어 콘텐츠 등에서 도서를 홍보하거나 광고 메시지를 전달할 때 저자, 출판사, 카테고리를 강조해 책에 관한 관심을 높일 수 있다. 또한 출판 시기가 판매에 영향을 미친다는 분석을 고려해 특정 시기에 맞춰 신간도서를 집중적으로 홍보할 수도 있다. 출판사에서도 영향요인 분석 결과를 기반으로 광고 메시지를 작성하거나 출판 시기를 조정하는 마케팅 전략을 수립할 수 있다.

베스트셀러를 많이 낸 작가가 베스트셀러를 집필할 가능성이 크고, 베스트셀러를 많이 낸 출판사가 베스트셀러를 출판할 가능성이 큰 것은 인공지능 모형을 이용하지 않고도 경험적으로 알 수 있는 사실이다. 그러나 베스트셀러 시장은 매우 경쟁적인 시장이어서 베스트셀러 작가나 출판사가 다음에 반드시 베스트셀러를 출판한다는 보장이 없는 것도 사실이다, 이러한 베스트셀러 예측의 불확실성은 서점 업계의 고민이기도 하다. 본 연구의 예측모형은 작가, 출판

사뿐만이 아니라, 제목, 줄거리, 가격, 카테고리, 출판 월 등 다양한 요인을 고려해 예측의 정확도를 높였다는 점에서 서점 업계의 고민을 해결하는 데 실질적인 도움을 줄 수 있다. 서점이 전통적인 판단기법과 본 연구의 예측모형을 함께 사용하면 좀 더 정확하게 베스트셀러를 예측할 수 있을 것이다. 본 연구의 예측모형에 사용된 데이터는 모두 서점이 기본적으로 보유하고 있는 정보이므로 서점에서 쉽고 편리하게 모형을 실행할 수 있다.

본 연구에는 한계점이 존재한다. 첫째, 실험대상이 '소설/시/희곡' 분야에 한정되어 있어 본 연구의 결과를 일반화시키기 어렵다는 것이다. 본 연구의 예측모형은 '소설/시/희곡' 분야의 상품기획자에게 실용적인 도움을 줄 수 있지만, 다른 분야 도서의 예측과 영향요인을 파악하려면, 추가 검증이 필요하다. 둘째, 공개된 판매량 데이터가 없어 베스트셀러를 예측 대상으로 삼았지만, 도서판매를 직접적으로 보여주는 데이터를 사용한다면, 결과가 달라질 수도 있다. 셋째, 예스24, 알라딘, 인터넷교보문고의 베스트셀러 공개 순위를 기반으로 베스트셀러 순위 범위를 상위 1,000개로 설정하였으나, 1위부터 1,000위까지 도서의 판매량 차이를 반영하지 못했다는 한계가 있다. 향후 연구에서는 이러한 한계를 보완하기 위해 다양한 분야 도서에 대한 실제 판매량 자료를 수집하여 예측모형을 구현하거나, 순위 간의 판매량 차이를 반영해 순위를 범주화한 예측실험을 진행할 수 있다.

참고문헌

김도영, 김나연, 김현희(2023), "앙상블 학습 기반 국내 도서의 해외 판매 굿셀러 예측 및 굿셀러 리뷰 키워드 분석," **정보처리학회논문지**, 제12권 4호, pp.173-178.

(Kim, D., Kim, N., and Kim, H.(2023), "Ensemble learning-based prediction of good sellers in overseas sales of domestic books and keyword analysis of reviews of the good sellers," Transactions of the Korea Information Processing Society, 12(4), pp.173-178.)

문동지, 윤상혁, 최수빈, 김희웅(2020), "머신러닝 기반의 보상형 크라우드펀딩 성공 예측 모델링," **Korea Business Review**, 제24권 3호, pp.125-143.

(Moon, D., Yoon, S., Choi, S., and Kim, H.(2020), "A machine learning approach for the success prediction of reward crowdfunding project," *Korea Business Review*, 24(3), pp.125-143.)

유지은, 조솔비, 유석종(2023), "베스트셀러 도서 예측을 위한 머신러닝 알고리즘 성능평가," **한국정보기술학회논문지**, 제21권 7호, pp.1-6.

(Yu, J., Cho, S., and Yu, S.(2023), "Performance evaluation of machine-learning algorithms for bestseller book prediction," *Journal of Korean Institute of Information Technology*, 21(7), pp.1-6.)

이중원, 박철(2021), "소셜미디어 콘텐츠 주제와 고객 인게이지먼트 간의 관계분석: 머신러닝 방법론을 중심으로," **경영학연구**, 제50권 1호, pp.115-142.

(Lee, J., and Park, C.(2021), "An analysis on the relationship between content topics of social media and customer engagement using machine learning methodology," *Korean Management Review*, 50(1), pp.115-142.)

장동률, 박민재(2021), "결정 트리 기반 학습 모형을 이용한 미술품 경매 가격 예측," **경영학연구**, 제50권 2호, pp.357-381.

(Jang, D., and Park, M.(2021), "Art price prediction using decision tree-based machine learning methods," *Korean Management Review*, 50

(2), pp.357-381.)

Abel, F., Herder, E., Houben, G. J., Henze, N., and Krause, D.(2013), "Cross-system user modeling and personalization on the social web," *User Modeling and User-Adapted Interaction*, 23, pp.169-209.

Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... and Herrera, F.(2020), "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, 58, 82-115.

Ashok, V. G., Feng, S., and Choi, Y.(2013), "Success with style: using writing style to predict the success of novels," In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp.1753-1764.

Bates, S., Hastie, T., and Tibshirani, R.(2023), "Cross-validation: what does it estimate and how well does it do it?," *Journal of the American Statistical Association*, 119, pp. 1-12.

Baye, M. R., De los Santos, B., and Wildenbeest, M. R.(2013), "Searching for physical and digital media: The evolution of platforms for finding books," *Economic Analysis of the Digital Economy*, University of Chicago Press, pp.137-168.

Bengio, Y., Ducharme, R., and Vincent, P.(2000), "A neural probabilistic language model," *Advances in Neural Information Processing Systems*, 13, pp.1-7.

Bobadilla, J., Ortega, F., Hernando, A., and Bernal, J.(2012), "A collaborative filtering approach to mitigate the new user cold start problem," *Knowledge-Based Systems*, 26, pp.225-238.

Breiman, L.(2001), "Random forests," *Machine Learning*, 45, pp.5-32.

Chang, T. Z., and Wildt, A. R.(1994), "Price, product information, and purchase intention: An empirical study," *Journal of the Academy of Marketing Science*, 22, pp.16-27.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W.P.(2002), "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, 16, pp. 321-357.

Chen, T., and Guestrin, C.(2016), "Xgboost: A scalable tree boosting system," In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785-794.

Chernev, A.(2003), "Reverse pricing and online price elicitation strategies in consumer choice," *Journal of Consumer Psychology*, 13(1-2), pp.51-62.

Chevalier, J. A., and Mayzlin, D.(2006), "The effect of word of mouth on sales: online book reviews," *Journal of Marketing Research*, 43(3), pp.345-354.

Chicco, D., and Jurman, G.(2020), "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, 21, pp.1-13.

Clement, M., Proppe, D., and Rott, A.(2007), "Do critics make bestsellers? opinion leaders and the success of books," *Journal of Media Economics*, 20(2), pp.77-105.

Cortes, C., and Vapnik, V.(1995), "Support-vector networks," *Machine Learning*, 20, pp.273-297.

Fatima, S. S., Wooldridge, M., and Jennings, N. R.(2008), "A linear approximation method for

the Shapley value," *Artificial Intelligence*, 172(14), pp.1673-1699.

Fawcett, T.(2006), "An introduction to ROC analysis," *Pattern Recognition Letters*, 27(8), pp.861-874.

Feng, T. Q., Choy, M., and Laik, M. N.(2020), "Predicting book sales trend using deep learning framework," *International Journal of Advanced Computer Science and Applications*, 11(2), pp.28-39.

Freund, Y., and Schapire, R. E.(1997), "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, 55(1), pp. 119-139.

Friedman, J. H.(2001), "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, 29(5), pp.1189-1232.

Gao, C., Chen, X., Feng, F., Zhao, K., He, X., Li, Y., and Jin, D.(2019), "Cross-domain recommendation without sharing user-relevant data," In *The World Wide Web Conference*, pp.491-502.

Ghavipour, M., and Meybodi, M.R.(2019), "Stochastic trust network enriched by similarity relations to enhance trust-aware recommendations," *Applied Intelligence*, 49, pp.435-448.

Givon, S., and Lavrenko, V.(2009), "Predicting social-tags for cold start book recommendations," In *Proceedings of the third ACM conference on Recommender Systems*, pp.333-336.

Glickman, M. E., and Van Dyk, D. A.(2007), "Basic bayesian methods," *Topics in Biostatistics*, Humana Press, pp.319-338.

Gollwitzer, P. M., and Sheeran, P.(2009), "Self-regulation of consumer decision making and behavior: The role of implementation intentions," *Journal of Consumer Psychology*, 19(4), pp. 593-607.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D.(2018), "A survey of methods for explaining black box models," *ACM Computing Surveys*, 51(5), pp.1-42.

Han, H., Wang, W. Y., and Mao, B. H.(2005), "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," In *International Conference on Intelligent Computing*, pp.878-887.

Havrlant, L., and Kreinovich, V.(2017), "A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)," *International Journal of General Systems*, 46(1), pp.27-36.

Howard, N.(2009), *The book: The life story of a technology*, Johns Hopkins University Press.

Johnson, S. J., Murty, M. R., and Navakanth, I.(2024), "A detailed review on word embedding techniques with emphasis on word2vec," *Multimedia Tools and Applications*, 83(13), pp.37979-38007.

Karthiga, R., Usha, G., Raju, N., and Narasimhan, K.(2021), "Transfer learning based breast cancer classification using one-hot encoding technique," In *International Conference on Artificial Intelligence and Smart Systems*, pp.115-120.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y.(2017), "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, 30, pp.3149-3157.

Khatun, F., Chowdhury, S. M. H., Tumpa, Z. N.,

Rabby, S. F., Hossain, S. A., and Abujar, S.(2019), "Sentiment analysis of amazon book review data using lexicon based analysis," In *Proceedings of the International Conference On Computational Vision and Bio Inspired Computing*, pp.1303-1309.

Kovács, B., and Sharkey, A.J.(2014), "The paradox of publicity: How awards can negatively affect the evaluation of quality," *Administrative Science Quarterly*, 59(1), pp.1-33.

Lee, I., Yi, J., and Kim, S.H.(2023), "Standing the test of time: What makes a book survive on the bestseller list?," *Journal of Business Research*, 164, 114013.

Lee, S., Ji, H., Kim, J., and Park, E.(2021), "What books will be your bestseller? A machine learning approach with Amazon kindle," *Electronic Library*, 39(1), pp.137-151.

Lee, S., Kim, J., Choi, E. B., Shin, S., Kim, D., Yu, H., Kim, S., Na, W.S., and Park, E.(2022), "Computational analysis of a collaboration network on human-computer interaction in Korea," *Mathematical Biosciences and Engineering*, 19(12), pp.13911-13927.

Lee, S., Kim, J., Kim, D., Kim, K. J., and Park, E.(2023), "Computational approaches to developing the implicit media bias dataset: Assessing political orientations of nonpolitical news articles," *Applied Mathematics and Computation*, 458, 128219.

Lee, S., Kim, J., and Park, E.(2023), "Can book covers help predict bestsellers using machine learning approaches?," *Telematics and Informatics*, 78, 101948.

Lee, S., and Park, E.(2024), "AutoCaCoNet: Automatic Cartoon Colorization Network using self-attention GAN, segmentation, and color correction," In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.403-411.

Leitão, L., Amaro, S., Henriques, C., and Fonseca, P.(2018), "Do consumers judge a book by its cover? A study of the factors that influence the purchasing of books," *Journal of Retailing and Consumer Services*, 42, pp.88-97.

Lika, B., Kolomvatsos, K., and Hadjiefthymiades, S.(2014), "Facing the cold start problem in recommender systems," *Expert Systems with Applications*, 41(4), pp.2065-2073.

Liu, S., and Meng, X.(2015), "A Location-Based Business Information Recommendation Algorithm," *Mathematical Problems in Engineering*, 2015(1), 345480.

Liu, S., Sun, X., Roemer, F. W., Guermazi, A., and Kwoh, C.K.(2022), "Comparison of various metrics for evaluating the performance of deep learning binary classification, particularly when underlying imaging data are imbalanced," *Osteoarthritis Imaging*, 2, 100017.

Loh, W. Y.(2011), "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), pp. 14-23.

Lundberg, S. M., and Lee, S. I.(2017), "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, 30, pp.4768-4777.

Magadán-Díaz, M., and Rivas-García, J.(2019), "Crowdfunding in the Spanish publishing industry," *Publishing Research Quarterly*, 35, pp.187-200.

Martín Sujo, J. C., Golobardes i Ribé, E., and Vilasís Cardona, X.(2021), "Cait: a predictive tool for supporting the book market operation

using social networks," *Applied Sciences*, 12(1), 366.

Naidu, G., Zuva, T., and Sibanda, E. M.(2023), "A review of evaluation metrics in machine learning algorithms," In *Computer Science On-line Conference*, pp.15-25.

Nakamura, L.,(2013), "Words with friends: socially networked reading on Goodreads," *PMLA*, 128(1), pp.238-243.

Nie, D. C., Zhang, Z. K., Dong, Q., Sun, C., and Fu, Y.(2014), "Information filtering via biased random walk on coupled social network," *Scientific World Journal*, 2014(1), 829137.

Oh, S., Kim, J., Lee, S., and Park, E.(2021), "Jujeop: Korean puns for k-pop stars on social media," In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pp.170-177.

Okada, S., Ohzeki, M., and Taguchi, S.(2019), "Efficient partition of integer optimization problems with one-hot encoding," *Scientific Reports*, 9(1), 13036.

Ozturk, S. A., Sevim, N., and Eroglu, E.(2006), "Leisure book reading and purchasing: an analysis of Turkish consumers," *International Journal of Consumer Studies*, 30(4), pp. 378-388.

Padilla, N., and Ascarza, E.(2021), "Overcoming the cold start problem of customer relationship management using a probabilistic machine learning approach," *Journal of Marketing Research*, 58(5), pp.981-1006.

Padilla, N., Ascarza, E., and Netzer, O.(2025), "The customer journey as a source of information," *Quantitative Marketing and Economics*, In-press.

Park, J. Y., Kim, C., Park, S., & Dio, K.(2023), "Do you judge a book by its cover? Online book purchases between Japan and France," *Asia Pacific Journal of Marketing and Logistics*, 35(10), pp.2345-2360.

Park, M. H., Lee, J. S., and Doo, I.C.(2020), "A study of the demand forecasting model for publishing business using business analysis," *International Journal of Computing and Digital Systems*, 90(5), pp.801-812.

Peng, C. Y. J., Lee, K. L., and Ingersoll, G. M.(2002), "An introduction to logistic regression analysis and reporting," *Journal of Educational Research*, 96(1), pp.3-14.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A.(2018), "CatBoost: unbiased boosting with categorical features," *Advances in Neural Information Processing Systems*, 31, pp.6639-6649.

Quinlan, J. R.(1986), "Induction of decision trees," *Machine Learning*, 1, pp.81-106.

Rainio, O., Teuho, J., and Klén, R.(2024), "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, 14(1), 6086.

Ramos, J.(2003), "Using tf-idf to determine word relevance in document queries," In *Proceedings of the first Instructional Conference on Machine Learning*, 242(1), pp.29-48.

Rao, A. R., and Monroe, K. B.(1989), "The effect of price, brand name, and store name on buyers' perceptions of product quality: An integrative review," *Journal of Marketing Research*, 26 (3), pp.351-357.

Rosli, A. N., You, T., Ha, I., Chung, K. Y., and Jo, G. S.(2015), "Alleviating the cold-start problem by incorporating movies facebook pages," *Cluster Computing*, 18, pp.187-197.

Schmidt-Stölting, C., Blömeke, E., and Clement,

M.(2011), "Success drivers of fiction books: an empirical analysis of hardcover and paperback editions in Germany," *Journal of Media Economics*, 24(1), pp.24-47.

Scott, S. E., and Williams, E. F.(2023), "In goal pursuit, I think flexibility is the best choice for me but not for you," *Journal of Marketing Research*, 60(5), pp.1008-1026.

Shehu, E., Prostka, T., Schmidt-Stölting, C., Clement, M., and Blömeke, E.(2014), "The influence of book advertising on sales in the German fiction book market," *Journal of Cultural Economics*, 38, pp.109-130.

Sharma, S. K., Chakraborti, S., and Jha, T.(2019), "Analysis of book sales prediction at amazon marketplace in India: a machine learning approach," *Information Systems and e-Business Management*, 17(2), pp.261-284.

Son, L. H.(2016), "Dealing with the new user cold-start problem in recommender systems: A comparative review," *Information Systems*, 58, pp.87-104.

Varoquaux, G., and Colliot, O.(2023), "Evaluating machine learning models and their diagnostic value," *Machine Learning for Brain Disorders*, Humana Press, pp.601-630.

Vihinen, M.(2012), "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis," *BMC Genomics*, 13, pp.1-10.

Visentin, M., and Tuan, A.(2021), "Book belly band as a visual cue: Assessing its impact on consumers' in-store responses," *Journal of Retailing and Consumer Services*, 59, 102359.

Wang, H., Liang, Q., Hancock, J. T., and Khoshgoftaar, T. M.(2024), "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods," *Journal of Big Data*, 11(1), 44.

Wang, X., Yucesoy, B., Varol, O., Eliassi-Rad, T., and Barabási, A. L.(2019), "Success in books: predicting book sales before publication," *EPJ Data Science*, 8(1), pp.1-20.

Yang, J., Sahni, N. S., Nair, H. S., and Xiong, X.(2024), "Advertising as information for ranking e-commerce search listings," *Marketing Science*, 43(2), pp.360-377.

Yin, X., Goudriaan, J., Lantinga, E. A., Vos, J., and Spiertz, H. J.(2003), "A flexible sigmoid function of determinate growth," *Annals of Botany*, 91(3), pp.361-371.

Zhang, Q., Wu, D., Lu, J., Liu, F., and Zhang, G.(2017), "A cross-domain recommender system with consistent information transfer," *Decision Support Systems*, 104, pp.49-63.

Zien, A., Krämer, N., Sonnenburg, S., and Rätsch, G.(2009), "The feature importance ranking measure," In *Machine Learning and Knowledge Discovery in Databases*, pp.694-709.

Zou, H., Gong, Z., Zhang, N., Zhao, W., and Guo, J.(2015), "Trustrank: A cold-start tolerant recommender system," *Enterprise Information Systems*, 9(2), pp.117-138.

• 저자 이승필은 서울대 국사학과를 졸업하였으며, 성균관대 인공지능융합학과 대학원에서 공학 석사학위를 취득한 후 박사과정에 재학 중이다. 현재 (주)사회평론 출판사 전무이사로 재직 중이며, 한국도서출판정보센터 기술총괄위원, 성균관대 융합소프트웨어전공 겸임교수를 겸직하고 있다.

• 저자 박은일은 성균관대에서 전자전기컴퓨터공학, 인터랙션사이언스학으로 학사 및 석사학위를 수여받았으며, KAIST에서 사용자 혁신으로 박사학위를 수여받았다. 한국건설기술연구원과 한양대학교 ICT융합학부 교수를 거쳐 현재 성균관대학교 인공지능융합학과 부교수로 재직 중이다. 현재 Sustainable Development와 IEEE Transactions on Automation Science and Engineering과 같은 최우수 국제 학술지의 Associate Editor를 맡고 있으며, ICT혁신인재4.0사업단과 딥페이크 연구센터의 단장으로 활동하고 있다.

• 저자 류두진은 서울대 전기공학부를 졸업하였으며, KAIST 테크노경영대학원에서 경영공학 박사학위를 취득하였다. 국민연금공단 연구위원으로 근무했으며, 한국외대 국제경영학과 학과장과 중앙대 경제학부 교수를 거쳐 현재 성균관대 경제학과 교수이다. 한국경영학회 상임이사, 한국재무관리학회 부회장, 제23대 한국금융공학회 학회장을 지냈으며, 현재 한국경제학회 이사, 재무관리논총 편집위원장, 성균관대 경제연구소 소장, Global Finance Research Center 센터장을 담당하고 있다. Investment Analysts Journal (SSCI)의 Editor이며, Emerging Markets Review (SSCI)와 Journal of Multinational Financial Management (SSCI)의 Subject Editor이다.