CatBoost와 PyCaret을 기반한 영화 박스오피스 예측 모델의 성능 비교 및 SHAP 해석

김희성[1], 문지훈[2*]
[1]순천향대학교 AI·빅데이터학과 학생, [2]순천향대학교 AI·빅데이터학과 교수

Performance Comparison and SHAP Interpretation of Movie Box Office Prediction Models Based on CatBoost and PyCaret

Huiseong Kim[1], Jihoon Moon[2*]
[1]Undergraduate Student, Department of AI and Big Data, Soonchunhyang University
[2]Professor, Department of AI and Big Data, Soonchunhyang University

요 약 본 연구는 한국 영화진흥위원회에서 수집한 박스오피스 데이터를 활용하여 관람 인원수와 매출액을 예측하는 모델을 구축하고, 이를 비교 및 분석하였다. 데이터 전처리 단계에서는 불필요한 변수를 제거하고, 결측치를 범주형 및 수치형 데이터에 따라 각각 처리하여 데이터의 일관성을 유지하였다. 또한, 탐색적 자료 분석을 통해 서울 지역의 관람 인원수, 매출액, 총 상영관 수, 영화 장르, 영화 등급, 개봉 월을 주요 변수로 선정하였으며, 서울 지역의 관람 인원수와 매출액이 박스오피스 성과와 높은 상관관계를 나타냄을 확인하였다. 이러한 분석을 바탕으로 CatBoost와 PyCaret AutoML을 사용하여 예측 모델을 개발하였다. CatBoost는 감독명, 제작사명, 영화 장르와 같은 범주형 변수를 효과적으로 처리할 수 있는 특성으로 인해 적합하다고 판단되었으며, PyCaret AutoML은 비전문가도 다양한 모델을 쉽게 비교할 수 있는 도구로서 모델링 과정을 자동화하여 효율성을 극대화할 수 있다. 예측 모델의 성능은 평균절대오차, 평균제곱근오차, 결정 계수를 기준으로 평가하였으며, CatBoost가 더 높은 예측 정확도를 보였다. 또한, SHAP 기법을 적용하여 주요 변수를 해석하였으며, 서울 지역의 관람 인원수와 매출액이 가장 중요한 변수임을 확인할 수 있었다. 본 연구는 신뢰성 있는 박스오피스 예측 모델을 제시함으로써 영화 산업의 의사결정에 기여하고, 데이터 기반 전략 수립을 지원한다.

주제어 : 박스오피스 예측, 탐색적 자료 분석, 기계학습, CatBoost, AutoML, SHAP

Abstract This study uses box office data collected by the Korean Film Council (KOFIC) to develop and compare predictive models for cinema attendance and revenue. Data preprocessing removed irrelevant variables and handled missing values separately for categorical and numerical data to ensure consistency. Exploratory data analysis identified key variables, including Seoul audience size, revenue, total number of screens, film genre, rating, and month of release, which revealed a strong correlation between Seoul audience size and revenue with box office performance. Based on this analysis, predictive models were developed using CatBoost and PyCaret AutoML. CatBoost was chosen for its effectiveness in handling categorical variables such as director name, production company, and genre, while PyCaret AutoML was chosen for its ability to automate the modeling process, making it easy for non-experts to compare different models. The performance of the models was evaluated using mean absolute error (MAE), root mean squared error (RMSE), and R-squared (R^2), with CatBoost demonstrating superior accuracy. In addition, the SHAP technique was used to interpret the models, identifying Seoul's audience size and revenue as the most significant predictors. This research presents reliable box office prediction models that will improve decision-making in the film industry and support the development of data-driven strategies.

Key Words : Box Office Prediction; Exploratory Data Analysis; Machine Learning; Categorical Boosting; Automated Machine Learning; SHapley Additive exPlanations

본 연구는 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업(2021-0-01399)의 연구결과로 수행되었음.
*교신저자 : 문지훈(jmoon22@sch.ac.kr)
접수일 2024년 9월 29일 수정일 2024년 10월 03일 심사완료일 2024년 10월 14일

1. 서론

영화 산업은 문화적 가치와 경제적 이익을 동시에 창출하는 중요한 분야로, 시장의 경쟁은 날로 치열해지고 있다. 영화의 성공은 주로 매출 금액과 관람 인원수로 측정되며, 이러한 지표의 정확한 예측은 영화 제작사와 배급사가 예산 할당, 마케팅 전략 수립, 상영관 배정 등 핵심적인 의사결정을 내리는 데 필수적이다[1]. 그러나 디지털 기술의 발전과 콘텐츠 소비 채널의 다양화 그리고 COVID-19와 같은 예기치 않은 상황으로 인해 관객의 선호도와 소비 패턴이 더욱 복잡해지고 예측하기 어려워지고 있다[2]. 기존의 영화 흥행 성적 예측 연구는 주로 회귀 분석과 같은 전통적인 통계적 방법론에 의존해 왔다. 하지만 De Vany와 Walls[3]는 영화 흥행 성적이 불확실성과 장기 꼬리 분포를 갖는다는 점을 지적하며, 이러한 통계적 방법이 영화 시장의 복잡성과 비선형성을 충분히 반영하지 못함을 밝혔다. 이는 전통적인 방법론이 영화 흥행 성적 예측에 한계가 있음을 나타내며, 더욱 복잡한 모델링 접근법이 필요함을 시사한다.

인공지능(AI; artificial intelligence)은 방대한 데이터를 분석하고 복잡한 패턴을 학습하는 능력으로 인해, 영화 산업에서도 다양한 연구에 응용되고 있다. Song 등[4]은 랜덤 포레스트(RF; random forest), 서포트 벡터 머신(SVM; support vector machine), 인공 신경망(ANN; artificial neural network) 등을 이용하여 영화 흥행 성적 예측의 정확도를 비교하였으며, 랜덤 포레스트가 예측의 안정성과 정확성을 높이는 데 효과적임을 입증하였다. Quader 등[5]은 영화 투자 결정을 지원하는 방법론을 개발하여, 기계학습과 딥러닝 기술을 활용해 박스오피스 수익을 예측하였다. Subramaniyaswamy 등[6]은 예고편 조회 수, 위키백과 페이지 조회 수, 비평가 평점, 개봉 시기 등 다양한 변수가 박스오피스에 미치는 영향을 분석함으로써 예측 모델의 성능을 향상하였다.

또한, Bae와 Kim[7]은 영화 소재의 변천을 분석하고, 이러한 소재가 박스오피스 수입에 미치는 영향을 실증적으로 분석하였다. 이 연구는 2004년부터 2018년까지 한국에서 개봉한 영화를 대상으로 토픽 모델링과 회귀 모형을 적용하여, 영화 소재가 흥행에 미치는 영향을 규명하였다. 결과적으로, 특정 시기에 유행하는 소재가 박스오피스 수입에 긍정적인 영향을 미침을 발견하였다. 그러나 이 연구는 주로 영화 소재에 초점을 맞추었기 때문에, 다른 다양한 요인들이 박스오피스에 미치는 영향을 종합적으로 분석하는 데에는 한계가 있었다. Bao 등[8]은 중국 영화 플랫폼의 크라우드펀딩 성공 요인을 탐색적으로 분석하였다. 이 연구는 크라우드펀딩 프로젝트의 다양한 요인이 성공률에 미치는 영향에 관해 다중회귀분석을 통해 검증하였으며, 영화 제작 단계, 투자 자금 비율, 투자 가능 여분 금액, 수익 정산 방식 등이 성공에 유의미한 영향을 미치는 것으로 나타났다. 그러나 이 연구는 특정 국가의 플랫폼에 국한되어 있어, 글로벌 영화 시장 전반에 대한 일반화에는 한계가 있었다.

Byun 등[9]은 다변량 시계열 데이터를 중심으로 한 딥러닝 모델을 사용하여 영화 흥행을 예측하고, 주요 변수를 선택하는 연구를 수행하였다. 해당 연구에서는 랜덤 포레스트 방법으로 주요 변수를 선별하고, 다층 퍼셉트론(multilayer perceptron), 완전 합성곱 신경망(fully convolutional neural network), 잔차 네트워크(residual network) 등의 딥러닝 모델을 적용하여 예측 정확도를 향상하였다. 특히, 잔차 네트워크를 사용한 모델이 약 93%의 높은 예측 정확도를 보였다. 이러한 연구는 딥러닝의 강력한 예측 능력을 입증하였으나, 모델의 복잡성과 해석 가능성 측면에서 한계가 존재한다.

이러한 문제를 해결하고자, Leem 등[10]은 설명 가능한 인공지능(XAI; explainable AI)을 기반으로 한 박스오피스 분류 및 트렌드 분석 모델인 DRECE (Dimension REduction, clustering, and Classification for Explainable artificial intelligence)를 개발하였다. DRECE는 다양한 영화 특성을 고려하여 흥행 요인을 식별하고, 계산 효율성을 높이기 위해 특성 차원을 축소한 후 영화를 유형별로 군집화하여 군집별로 박스오피스 흥행에 기여한 요소를 분석하는 모델이다. 또한, Leem 등[11]은 다양한 영화를 군집화하고 군집 유형별 온라인 박스오피스를 예측하는 추가적인 모델을 제안하였다. 해당 모델은 다양한 알고리즘을 자동으로 구성하고 최적화함으로써 객관적인 모델 선택을 가능하게 하고자 AutoML (automated machine learning) 방법론의 하나인 PyCaret에서 제공하는 여러 기계학습 알고리즘을 비교 및 분석하여 군집된 영화에 대해 온라인 박스오피스를 예측하였다.

영화 데이터는 감독, 배우, 제작사, 배급사 등 다양한 범주형 변수를 포함하고 있으며, 이러한 변수들은 높은 카디널리티를 가지는 경우가 많아 전통적인 원-핫 인코딩(one-hot encoding) 방식으로 처리하기에 적합하지 않다. 높은 카디널리티의 범주형 변수는 모델의 복잡성을 증가시키고, 과적합의 위험을 높이며, 계산 자원의 효율적인 사용을 방해할 수 있다[12]. 예를 들어, 감독이나

〈Table 1〉 KOBIS (Korean Box Office Information System) box office movie information: Column definitions

No	Column Name	Column Definition	Data Type	Length
1	NO	Unique identifier for the movie	DECIMAL	30,0
2	MOVIE_NM	Name of the movie	VARCHAR	200
3	DRCTR_NM	Name of the director	VARCHAR	200
4	MAKR_NM	Name of the production company	VARCHAR	200
5	INCME_CMPNY_NM	Name of the importing company	VARCHAR	200
6	DISTB_CMPNY_NM	Name of the distributing company	VARCHAR	200
7	OPN_DE	Release date of the movie (YYYYMMDD)	VARCHAR	8
8	MOVIE_TY_NM	Type of the movie (e.g., feature film)	VARCHAR	200
9	MOVIE_STLE_NM	Format of the movie (e.g., short, feature)	VARCHAR	200
10	NLTY_NM	Nationality of the movie	VARCHAR	200
11	TOT_SCRN_CO	Total number of screens	DECIMAL	28, 5
12	SALES_PRICE	Total box office revenue	DECIMAL	28, 5
13	VIEWNG_NMPR_CO	Total number of viewers	DECIMAL	28, 5
14	SEOUL_SALES_PRICE	Box office revenue in Seoul	DECIMAL	28, 5
15	SEOUL_VIEWNG_NMPR_CO	Number of viewers in Seoul	DECIMAL	28, 5
16	GENRE_NM	Genre of the movie	VARCHAR	200
17	GRAD_NM	Rating of the movie	VARCHAR	200
18	MOVIE_SDIV_NM	Classification of the movie (e.g., general, animation)	VARCHAR	30

배우와 같은 변수는 수백에서 수천 개의 고윳값을 가질 수 있어 원-핫 인코딩 시 차원이 과도하게 증가하게 된다. 이러한 문제는 모델 학습의 효율성을 저해하고, 예측 성능에 부정적인 영향을 미칠 수 있다. 이를 해결하기 위해, 범주형 데이터 처리에 강점이 있는 CatBoost (categorical boosting) 모델이 주목받고 있다[13]. CatBoost[13]는 범주형 변수의 순서를 유지하면서 학습을 진행하여 높은 카디널리티를 가진 변수들도 효과적으로 다룰 수 있으므로, 다수의 범주형 변수가 포함된 데이터에서도 뛰어난 성능을 발휘하는 것으로 알려져 있다.

또한, 현대의 기계학습 모델은 그 예측 결과를 설명하는 것이 점점 더 중요해지고 있다[14]. 설명 가능한 인공지능 기법은 영화 산업과 같은 의사결정이 중요한 분야에서 모델의 의사결정을 이해하고, 사용자가 모델의 예측을 신뢰할 수 있도록 돕는다. 본 연구는 이러한 배경을 바탕으로, 영화 박스오피스 예측을 위해 CatBoost와 PyCaret의 성능을 비교 및 분석한다. 탐색적 자료 분석 (EDA; exploratory data analysis)을 바탕으로 개봉 월별 매출금액과 장르별 관람 인원수를 시각화하고, 매출금액과 관람 인원수 간의 상관관계를 분석한다. 이후, CatBoost 모델을 활용하여 관람 인원수와 매출 금액을 예측하고, SHAP (Shapley additive explanations) 기법[15]을 통해 예측 결과를 해석한다. 이러한 접근을 통해 본 연구는 영화 박스오피스 예측에 있어 신뢰성 있는 모델을 제시하고, 의사결정자들이 정보에 기반한 판단을 내릴 수 있도록 지원하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 영화 박스오피스 데이터를 활용한 데이터 전처리와 탐색적 자료 분석을 다루고, 3장에서는 CatBoost와 AutoML 모델, SHAP 기법에 관해 기술한다. 이어서 4장에서는 실험 방법과 결과를 제시하며, 마지막으로 5장에서는 연구의 결론을 맺는다.

2. 박스오피스 데이터 전처리 및 분석

영화 매출 및 관람 인원 예측은 영화 개봉 전 상업적 성공을 예측하고자 하는 욕구에서 출발하며, 이는 영화 산업의 다양한 이해관계자들에게 중요한 통찰력을 제공할 수 있다. 본 연구의 주요 목표는 먼저 한국 영화진흥위원회(KOFIC; Korean Film Council)[16]에서 제공하는 박스오피스 데이터를 활용하여 데이터의 결측치 처리, 중복 데이터 통합, 이상치 제거 등 여러 데이터 전처리 기법을 시행한 후, 날짜 형식 변환 및 추가 변수 생성을 통해 데이터의 일관성을 확보하는 것이다. 이러한 전처리 과정을 거친 후, 월별 평균 매출액, 장르별 평균 관람 인원수 및 매출액과 관람 인원수 간의 상관관계를 시각화하고 분석하는 것을 최종 목표로 한다.

2.1 데이터 수집 및 전처리

본 연구는 KOFIC에서 2017년부터 2023년까지 총 2,972개의 영화 박스오피스 데이터를 CSV 파일 형식으로 수집하였다. 여기서 2017년 개봉 영화는 단 한 편에 불과하여 분석의 신뢰성을 높이기 위해 제외하였으며, 최신 영화에 초점을 맞추고자 2019년부터 2023년까지 개봉한 영화들의 데이터를 사용하여 총 2,971개를 분석하였다.

데이터는 영화명, 감독명, 제작사명, 유통사명, 개봉일자, 상영 스크린 수, 매출액, 관람객 수 등 다양한 변수를 포함하고 있으며, 한국 영화뿐만 아니라 프랑스, 호주, 헝가리 등 다양한 국적의 영화들도 포함되어 있어 국제적인 분석이 가능하다. 데이터 내 모든 칼럼은 기본 키(primary key)와 NOT NULL 제약 조건이 없으며, 각 필드는 선택적 데이터임을 확인하였다. 데이터의 전체적인 구조를 파악한 후, 분석에 필요하지 않은 열들을 식별하였다. 이 중 “NO” 열은 영화의 고유번호를 나타내는 값으로, 분석에 불필요하다고 판단하여 제거하였다. 본 연구에 사용된 데이터의 칼럼 정의는 Table 1에 제시하였다.

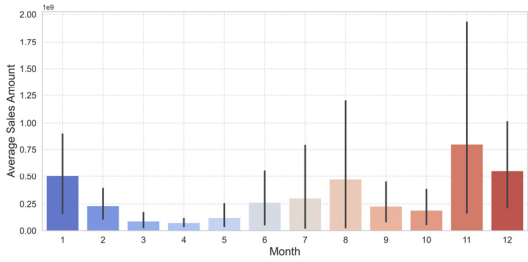
결측치는 범주형 데이터와 수치형 데이터로 나누어 처리하였다. 감독명, 제작사명, 유통사명, 장르, 국가, 등급 등과 같은 범주형 데이터에서 결측된 값들은 “Unknown”이라는 값으로 대체하였다. 이는 해당 정보가 제공되지 않은 영화들에 대해 분석에서 빠지지 않도록 하기 위함이다. 이러한 결측치 대체 작업은 분석의 일관성을 유지하면서도 중요한 정보를 누락시키지 않도록 하였다. 수치형 데이터는 상영 스크린 수, 매출, 관람객 수 등의 변수들이 포함되어 있었다. 이들 중 상영 스크린 수나 매출 정보가 결측된 영화는 분석에 적합하지 않다고 판단하여 해당 행은 제거하였다. 특히 상영 스크린 수와 매출은 영화 흥행의 주요 지표이므로 이 값들이 없는 영화는 분석에서 제외하는 것이 타당하다고 판단하였다.

개봉 일자 변수(OPN_DE)는 YYYYMMDD 형식의 문자열로 저장되어 있었으나, 일부 값은 잘못된 형식이거나 데이터 누락이 있었다. 이를 일관된 날짜 형식으로 변환하였으며, 변환할 수 없는 날짜 값은 모두 제거하였다. 이후 개봉 일자로부터 개봉 연도와 개봉 월을 각각 추출하여 새로운 변수로 생성하였다. 이 과정에서 잘못된 날짜 데이터는 대부분 1970년으로 기록되어 있었으며, 이는 잘못된 기록임을 고려하여 제거하였다.

영화명 기준으로 중복된 데이터가 존재하였다. 중복된 영화는 동일한 영화를 여러 번 기록한 경우로, 이들을 통합하여 하나의 영화 정보로 처리하였다. 범주형 변수의 경우에는 첫 번째로 기록된 값을 선택하였고, 수치형 변수의 경우에는 상영 스크린 수나 매출을 합산하여 하나의 영화 데이터로 통합하였다. 이렇게 중복된 영화 데이터를 통합함으로써 중복에 의한 분석 왜곡을 방지하였다.

2.2 탐색적 자료 분석

본 연구는 2019년부터 2023년까지의 전처리된 영화 데이터를 사용하여 탐색적 자료 분석을 수행하였으며, 다음과 같은 세부 분석을 통해 영화 산업의 경향과 특성을 심도 있게 확인하고자 하였다.

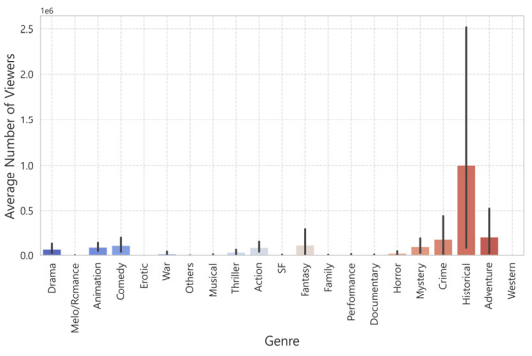


[Fig. 1] Average box office revenue by release month

Fig. 1은 개봉 월별 평균 매출 금액을 시각화한 것으로, 상세한 수치는 Table 2에서 확인할 수 있다. 11월의 매출이 799,482,182원으로 가장 높게 나타나며, 이는 연말 시즌에 대형 영화들이 몰려 개봉하기 때문으로 분석된다. 이러한 패턴은 시즌별 영화 배급 전략을 수립하는 데 중요한 지표로 활용될 수 있다.

〈Table 2〉 Average box office revenue by release month

Release Month	Average Box Office Revenue (KRW)
January	504,680,519
February	233,802,074
March	87,496,571
April	72,989,572
May	119,235,062
June	264,763,718
July	302,735,906
August	472,866,986
September	229,594,976
October	192,549,478
November	799,482,182
December	555,255,069



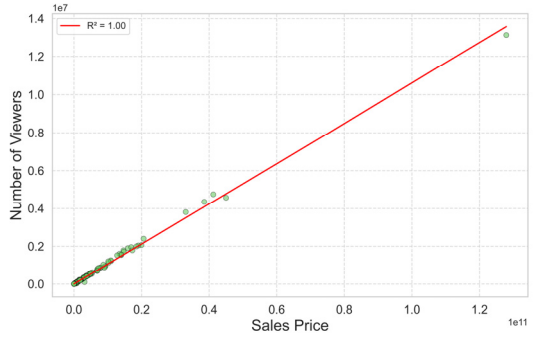
[Fig. 2] Average number of viewers by genre

Fig. 2는 장르별 관람 인원수의 평균을 시각화한 것으로, 상세한 수치는 Table 3에서 확인할 수 있다. "사극" 장르가 평균 1,004,170명으로 가장 높은 관람 인원수를 기록하였으며, "어드벤처" 장르가 그 뒤를 이어 200,842명의 관람객을 끌어모았다. 이는 대규모 투자와 높은 제작비가 소요되는 장르가 큰 인기를 끌고 있음을 시사한다. 장르 선택에 따른 관객 동향을 분석함으로써 향후 장르별 투자와 마케팅 전략을 더욱 효과적으로 수립할 수 있다.

〈Table 3〉 Average number of viewers by genre

Genre	Average Number of Viewers
SF	4,246
Family	4,241
Performance	8,385
Horror	21,972
Others	259
Documentary	5,846
Drama	65,948
Melo/Romance	1,205
Musical	7,012
Mystery	94,714
Crime	175,917
Historical	1,004,170
Western	46
Erotic	14
Thriller	33,154
Animation	88,370
Action	85,725
Adventure	200,842
War	16,700
Comedy	107,891
Fantasy	112,084

Fig. 3은 매출 금액과 관람 인원수 간의 상관관계를 분석한 결과로, 매우 높은 상관관계(R^2 = 1.00)를 확인할 수 있다. 이는 매출 증대 전략을 수립할 때 관람 인원수의 증가가 매출에 직접적으로 영향을 미친다는 강력한 근거를 제시한다. 이를 바탕으로 영화 산업에서의 관객 유치 전략 및 티켓 가격 결정에 중요한 지침을 제공할 수 있다.



[Fig. 3] Visualization of the correlation between box office revenue and number of viewers

3. 모델 구축 및 해석 방법

본 연구에서는 영화 박스오피스 예측을 위해 기계학습 알고리즘인 CatBoost와 AutoML 도구인 PyCaret을 활용하였다. 또한, 모델의 예측 결과를 해석하고 주요 변수를 파악하기 위해 SHAP 기법을 적용하였다. 이 장에서는 각 방법의 이론적 배경과 비전문가들이 고려해야 할 주요 사항을 논의한다.

3.1 CatBoost를 활용한 분류 모델 구성

CatBoost[13]는 Yandex에서 개발한 그래디언트 부스팅(gradient boosting) 알고리즘으로, 특히 범주형 변수를 효과적으로 처리할 수 있는 것이 큰 장점이다. 영화 데이터에는 감독명, 장르, 제작사명 등 고유한 범주형 변수가 많다. 최근 연구에서는 XGBoost[17], LightGBM[18]과 같은 전통적인 부스팅 알고리즘 외에도, 딥러닝 기반의 TabNet[19], FT-Transformer[20]와 같은 최신 알고리즘들이 제안되고 있으나, CatBoost는 범주형 변수 처리의 효율성과 모델 성능 면에서 여전히 우수한 선택으로 평가되고 있다. 이러한 변수들을 일반적인 기계학습 알고리즘에서 처리할 때, 범주형 변수를 수치형으로 변환하기 위해 원-핫 인코딩 또는 레이블 인코딩(label encoding)을 사용하면 이는 고차원 문제를 초래하거나 변수 간의 순서 정보(ordinal information)를 손실시킬 수

있다. 반면, CatBoost는 순열 통계(permutation statistics)를 기반으로 범주형 변수의 순서를 유지하면서 학습을 진행하여, 높은 카디널리티를 가진 변수들도 효과적으로 다룰 수 있다. 이러한 특징은 모델의 정확도를 높이고 과적합(overfitting)을 방지하는 데 도움이 된다.

영화 데이터는 장르, 배우, 개봉일 등 다수의 범주형 변수를 포함하고 있으며, 특히 배우나 감독과 같은 변수는 고유한 범주의 수가 많아 높은 카디널리티를 가진다. 이러한 데이터 특성은 일반적인 기계학습 알고리즘에서 처리하기 어려운 경우가 많다. 그러나 CatBoost는 이러한 고유한 범주형 변수를 효율적으로 처리할 수 있어, 영화 박스오피스 예측에 적합한 알고리즘이다. 이는 비전문가들도 비교적 쉽게 활용할 수 있으며, 데이터 전처리에 소요되는 시간을 줄여준다.

모델 구축 시 비전문가들이 고려해야 할 주요 옵션은 범주형 변수를 정확하게 지정하는 것과 하이퍼파라미터 설정이다. 범주형 변수는 cat_features 파라미터를 통해 명시적으로 지정해야 하며, 데이터 셋에서 범주형 변수를 정확히 식별하여 지정하는 것이 중요하다. 잘못된 변수 지정은 모델의 성능 저하를 초래할 수 있으므로 주의가 필요하다. 또한, 모델의 성능을 최적화하기 위해 학습률(learning_rate), 트리 깊이(depth), L2 정규화 계수(l2_leaf_reg) 등 주요 하이퍼파라미터를 적절히 설정해야 한다. 비전문가들은 기본값을 사용하거나, 그리드 탐색(grid search) 또는 무작위 탐색(random search)과 같은 자동화된 하이퍼파라미터 탐색 방법을 통해 최적의 값을 찾을 수 있다. 이러한 방법은 모델의 복잡도를 조절하고 일반화 능력을 향상하는 데 도움이 된다.

데이터의 분할도 중요한 요소로, 모델의 일반화 능력을 평가하기 위해 데이터 셋을 학습용과 평가용으로 분할한다. 일반적으로 70:30 또는 80:20의 비율을 사용하며, 재현성을 위해 랜덤 시드(random seed)를 고정한다. 이는 모델이 새로운 데이터에 대해서도 안정적인 예측 성능을 보이는지 확인하는 데 필수적이다. 또한, 교차검증(cross-validation)을 통해 모델의 성능을 더욱 신뢰성 있게 평가할 수 있다.

CatBoost의 장점으로는 범주형 변수 처리가 효율적이며, 순열 기반의 학습으로 과적합을 줄여준다. 또한, GPU 가속을 지원하여 대용량 데이터에도 빠르게 학습할 수 있으며, 변수 중요도 및 SHAP 값 등을 통해 모델을 해석하기 쉽다. 특히, 모델 해석은 비전문가들이 모델의 동작 원리를 이해하고 결과를 신뢰하는 데 큰 도움을 준다. 따라서 CatBoost는 높은 예측 성능과 해석 가능성을 동시에 제공하는 알고리즘으로 평가받고 있다.

3.2 PyCaret 기반 AutoML 적용

AutoML[21]은 기계학습 모델의 개발 과정을 자동화하여 전문가가 아니더라도 효과적인 모델을 구축할 수 있도록 돕는 기술이다. 모델 선택, 하이퍼파라미터 튜닝, 성능 평가 등의 과정을 자동화함으로써 모델 개발에 소요되는 시간과 노력을 절감한다. PyCaret[22]은 파이썬 기반의 AutoML 라이브러리로, 간단한 코드만으로도 다양한 기계학습 알고리즘을 적용하고 비교할 수 있다. 특히, PyCaret은 최신 AutoML 기술인 자동 피처 엔지니어링, 모델 앙상블링, 메타러닝 등을 통합하여 모델 구축 과정을 더욱 효율적으로 지원한다.

모델 구축 시 고려해야 할 사항은 환경 설정, 모델 비교 및 선택, 하이퍼파라미터 튜닝, 예측 및 평가이다. 환경 설정은 *setup* 함수를 통해 데이터 셋과 종속 변수, 실험 환경을 설정하며, 이 과정에서 데이터 타입 자동 인식, 결측값 처리, 이상치 탐지, 데이터 스케일링 등의 옵션을 설정할 수 있다. 이는 모델의 성능과 신뢰성에 큰 영향을 미치며, 데이터 전처리 과정을 자동화함으로써 사용자의 실수를 줄이고 효율성을 높일 수 있다[23].

모델 비교는 *compare_models* 함수를 사용하여 다양한 알고리즘의 성능을 자동으로 비교하고 최적의 모델을 선택한다. PyCaret은 선형 회귀, 의사결정 나무, 랜덤 포레스트, XGBoost (extreme gradient boosting) 등 다양한 알고리즘을 지원하며, 각 모델의 성능을 평가지표에 따라 정렬하여 보여준다. 평가지표는 문제의 특성에 따라 RMSE (root mean square error), MAE (mean absolute error), 결정 계수(R^2) 등을 선택할 수 있으며, 복수의 지표를 함께 고려하여 모델을 선택하는 것도 가능하다.

선택된 모델에 대해 *tune_model* 함수를 사용하여 하이퍼파라미터를 자동으로 최적화할 수 있으며, 이 과정에서 교차 검증을 통해 모델의 일반화 성능을 높일 수 있다. 또한, *predict_model* 함수를 통해 평가 데이터에 대한 예측을 수행하고 성능을 평가한다. 이러한 일련의 과정은 코드 몇 줄로 구현할 수 있으며, 비전문가들도 쉽게 따라 할 수 있다. 결과적으로, PyCaret을 활용하면 모델 개발에 드는 시간과 노력을 크게 절감할 수 있으며, 고성능의 예측 모델을 구축할 수 있다.

최신 연구에서는 PyCaret이 제공하는 통합된 환경으로 인해 빠른 프로토타이핑과 반복적인 실험이 가능하여 연구 개발 속도를 크게 향상한다는 평가를 받고 있다

[24]. 더불어, PyCaret은 클라우드 기반 배포 옵션과 실시간 예측 기능을 지원하여, 데이터 과학 프로젝트의 초기 단계에서부터 최종 모델 배포에 이르기까지 전 과정에서 유연하게 활용될 수 있다. 이러한 특징들로 인해 PyCaret은 데이터 과학 프로젝트의 효율성을 극대화하고, 비전공자도 고성능 모델을 손쉽게 구축할 수 있도록 돕는다.

3.3 SHAP을 활용한 모델 해석

SHAP[13]은 게임 이론에 기반한 모델 해석 기법으로, 각 특징(feature)이 모델의 예측에 기여하는 정도를 계산하여 설명력을 제공한다. 이를 통해 모델의 투명성을 높이고 결과 해석을 용이하게 하여 의사결정에 유용한 인사이트를 제공할 수 있다. 최근 연구에서는 SHAP 외에도 LIME (local interpretable model-agnostic explanations) [25], PDP (partial dependence plot) [26], PFI (permutation feature importance) [27] 등 다양한 모델 해석 기법들이 제안되었지만, SHAP은 이들보다 일관성 있는 설명력을 제공하며 다양한 모델에 적용 가능하다는 장점을 가지고 있다. 특히, 블랙박스 모델로 여겨지는 복잡한 기계학습 알고리즘의 내부 동작을 이해하는 데 효과적이다. 또한, SHAP은 글로벌 및 로컬 해석을 모두 지원하여 모델의 전반적인 동작과 개별 예측의 세부적인 기여도를 동시에 파악할 수 있게 지원한다. 이러한 특성 덕분에 SHAP은 다른 해석 기법들보다 더욱 신뢰할 수 있는 모델 해석 도구로 평가받고 있다.

SHAP을 활용하여 기계학습 모델의 예측 결과를 해석할 때 비전문가들이 고려해야 할 사항은 모델 호환성, SHAP 값 계산, 결과 시각화이다. SHAP은 대부분의 기계학습 모델에 적용 가능하지만, 의사결정 나무 기반 모델에서는 특히 효율적으로 작동한다. 모델과 데이터 셋을 사용하여 SHAP 값을 계산하며, 데이터 셋이 클 경우 계산 시간이 오래 걸릴 수 있으므로 샘플링을 통해 데이터 크기를 조절할 수 있다. 또한, SHAP 값을 해석할 때 각 변수의 영향력이 양수인지 음수인지, 그리고 그 크기가 어느 정도인지를 파악하는 것이 중요하다.

SHAP 값은 *summary_plot*, *bar_plot*, *dependence_plot* 등의 시각화 기법을 통해 해석할 수 있다. *summary_plot*은 전체 데이터에 대한 변수 중요도를 한눈에 보여주며, 변수들이 예측에 어떻게 영향을 미치는지 시각화한다. *dependence_plot*은 특정 변수와 목표 변수 간의 관계를 보여주어 변수 간의 상호작용을 파악할 수 있다. 이러한 시각화는 복잡한 수학적 개념을 직관적으로 이해할 수 있을 뿐만 아니라 비전문가들도 결과를 해석하는 데 도움을 준다.

CatBoost 모델에 SHAP을 적용함으로써, 관람 인원수와 매출 금액 예측에 영향을 미치는 주요 변수를 식별하고 그 영향력을 정량화할 수 있다. 이는 변수 중요도 파악, 의사결정 지원, 모델 신뢰성 향상의 이점을 제공한다. 예를 들어, 특정 장르의 영화가 관람 인원수에 긍정적인 영향을 미친다면, 제작자는 해당 장르의 영화를 기획하는 데 참고할 수 있다. 또한, 모델이 어떤 변수를 중요하게 사용하는지 알 수 있어 데이터 수집 및 관리에 집중해야 할 변수를 선정할 수 있으며, 데이터 기반의 전략 수립에 활용할 수 있다.

이러한 접근을 통해 본 연구는 영화 박스오피스 예측에 있어 신뢰성 있는 모델을 제시하고, 의사결정자들이 정보에 기반한 판단을 내릴 수 있도록 지원하고자 한다. 특히, 비전문가들도 쉽게 접근할 수 있는 도구와 기법을 활용함으로써 영화 산업 내에서 데이터 기반 의사결정의 활성화를 기대할 수 있다. 이는 궁극적으로 영화 제작 및 배급 전략의 효율성을 높이고, 시장의 변화에 유연하게 대응할 수 있게 해줄 것으로 기대한다.

4. 모델 평가 및 해석

4.1 성능 평가: 모델 비교 분석

본 연구에서는 CatBoost와 PyCaret을 활용하여 영화 박스오피스 예측 모델을 구축하고, 두 모델의 성능을 비교하였다. 실험은 데이터의 70%를 학습용으로, 30%를 평가용으로 랜덤 샘플링하여 진행하였다. 모델의 일반화 능력을 평가하기 위해 CatBoost에서 최적의 초매개변수 설정과 PyCaret에서 가장 우수한 모델을 선정하고자 교차 검증을 적용하였으며, 평가지표로는 MAE, RMSE, R^2를 사용하였다. 이러한 지표들은 회귀 모델의 예측 성능을 종합적으로 평가하는 데 널리 사용된다.

MAE는 실젯값과 예측값의 절대 차이의 평균으로, 예측값이 실젯값에서 평균적으로 얼마나 벗어나는지를 나타낸다. 이는 오차의 크기를 직관적으로 이해할 수 있게 해주며, 단위가 원 데이터와 동일하여 해석이 용이하다. RMSE는 오차의 제곱을 평균하여 제곱근을 취한 값으로, 큰 오차에 대해 민감하게 반응한다. 이는 모델의 예측 정확도를 평가하는 데 유용하며, 값이 낮을수록 모델의 성능이 우수함을 의미한다. R^2는 모델이 데이터의 변동성을 얼마나 설명하는지를 나타내며, 0에서 1 사이의 값을

가지는데, 1에 가까울수록 모델의 설명력이 높다.

CatBoost 모델은 하이퍼파라미터 최적화를 위해 그리드 탐색을 활용하였으며, 주요 하이퍼파라미터로는 *depth*, *l2_leaf_reg*, *learning_rate*을 사용하였다. 관람 인원수 예측에서는 *depth* = 4, *l2_leaf_reg* = 9, *learning_rate* = 0.03, 매출 금액 예측에서는 *depth* = 4, *l2_leaf_reg* = 9, *learning_rate* = 0.1이 최적의 조합으로 나타났다. 이러한 하이퍼파라미터 설정은 모델의 복잡도와 일반화 능력을 조절하여 과적합을 방지하고 예측 성능을 향상하는 데 기여하였으며, CatBoost 모델의 성능은 Table 4와 같다.

〈Table 4〉 CatBoost model performance evaluation results

Metric	Number of Viewers	Box Office Revenue
MAE	2,637.42	21,480,344.63
RMSE	14,666.76	166,802,524.21
R^2	0.97	0.95

학습 데이터 셋에서 PyCaret을 활용하여 다양한 회귀 알고리즘의 성능을 비교하였으며, 그 결과는 Table 5와 Table 6과 같다. 각 알고리즘의 성능 지표를 통해 모델의 예측 능력을 종합적으로 평가할 수 있으며, 이는 알고리즘 선택에 중요한 인사이트를 제공한다.

〈Table 5〉 Performance comparison of PyCaret algorithms (prediction of number of viewers)

Algorithm	MAE	RMSE	R^2
Huber Regressor	3,369.40	22,329.53	0.97
K-Neighbors Regressor	5,177.85	54,116.26	0.91
Linear Regression	12,202.29	63,905.53	0.82
Lasso Least Angle Regression	12,199.44	63,912.45	0.82
Ridge Regression	12,195.18	63,931.19	0.82
Extra Trees Regressor	6,678.73	64,262.72	0.87
Lasso Regression	12,267.70	64,468.35	0.82
Bayesian Ridge	12,122.07	64,623.63	0.82
Elastic Net	12,204.53	65,080.97	0.82
Orthogonal Matching Pursuit	12,463.42	71,868.73	0.78
Random Forest Regressor	7,912.06	78,198.57	0.79
Extreme Gradient Boosting	9,606.25	84,708.50	0.79
AdaBoost Regressor	12,401.04	88,010.19	0.77
Gradient Boosting Regressor	9,679.17	88,394.25	0.72
Decision Tree Regressor	9,578.96	92,077.37	0.51
Passive Aggressive Regressor	11,456.82	109,394.13	0.50
Light Gradient Boosting Machine	20,475.92	123,882.55	0.40
Least Angle Regression	85,702.53	162,878.36	-14.91
Dummy Regressor	33,835.57	171,840.87	-0.01

Table 5는 관람 인원수 예측에 대한 PyCaret 알고리즘의 성능을 나타낸다. Huber Regressor는 이상치에 강인한 회귀 알고리즘으로, 본 데이터 셋에서 가장 우수한 성능을 보였다. K-Neighbors Regressor와 Extra Trees Regressor도 비교적 좋은 성능을 나타내었으나, 선형 회귀 계열의 알고리즘은 상대적으로 낮은 성능을 보였다.

매출 금액 예측에 대한 PyCaret 알고리즘의 성능은 Table 6과 같다. 매출 금액 예측에서도 Huber Regressor가 가장 우수한 성능을 보였다. 그러나 일부 알고리즘은 R^2가 음수로 나타나거나 매우 낮은 값을 보여 예측 성능이 좋지 않음을 알 수 있다. 이는 해당 알고리즘이 데이터의 특성을 충분히 반영하지 못했거나, 과적합 또는 과소적합 문제가 발생했을 가능성을 시사한다.

두 모델의 성능을 비교하면, CatBoost 모델이 관람 인원수 예측에서 MAE와 RMSE 측면에서 PyCaret의 Huber Regressor보다 낮은 값을 보여 더욱 정확한 예측을 수행하였다. 매출 금액 예측에서는 두 모델의 성능이 비슷한 수준이지만, CatBoost 모델이 약간 더 낮은 MAE를 보여주었다. 이는 CatBoost가 범주형 변수가 많은 영화 데이터의 특성을 효과적으로 반영하였기 때문이라고 판단된다.

실험 결과는 데이터의 특성에 따라 모델 선택의 중요함을 보여준다. CatBoost는 범주형 변수가 많고 고차원인 데이터 셋에서 강점을 보이며, 하이퍼파라미터 튜닝을 통해 모델의 성능을 최적화할 수 있다. 반면에 PyCaret은 AutoML 도구로서 다양한 알고리즘을 손쉽게 비교하고 최적의 모델을 선택할 수 있게 해주며, 비전문가들도 쉽게 활용할 수 있다는 장점이 있다. 따라서 두 접근 방식은 상호 보완적으로 활용될 수 있으며, 데이터 분석 목적과 상황에 따라 적절한 방법을 선택하는 것이 중요하다.

또한, PyCaret을 통해 다양한 알고리즘의 성능을 한눈에 비교함으로써, 모델 선택 과정에서의 객관성을 높일 수 있다. 이는 비전문가들이 인공지능 기술을 활용하여 데이터 기반 의사결정을 내리는 데 큰 도움이 될 수 있다. 특히, 영화 산업 분야에서 이러한 접근은 제작 및 배급 전략 수립, 마케팅 계획 등에 유용한 인사이트를 제공하여 경쟁력을 향상할 수 있을 것으로 기대한다.

4.2 XAI: 모델 해석과 통찰

SHAP bar plot과 SHAP summary plot을 종합적으로 분석하면, 각 변수가 모델의 예측에 미치는 영향을 더

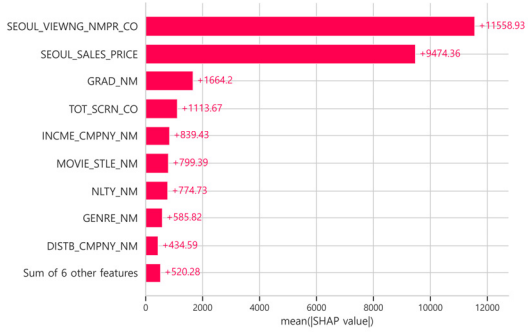
〈Table 6〉 Performance comparison of PyCaret algorithms (prediction of box office revenue)

Algorithm	MAE	RMSE	R^2
Huber Regressor	25,380,496.91	168,493,814.13	0.98
Passive Aggressive Regressor	59,445,143.20	367,288,276.04	0.95
Decision Tree Regressor	51,858,847.05	516,030,723.21	0.62
Linear Regression	100,093,721.79	519,927,084.95	0.85
Lasso Least Angle Regression	100,093,719.52	519,927,090.75	0.85
Ridge Regression	100,021,613.49	520,131,811.69	0.85
Lasso Regression	100,093,725.65	520,373,067.22	0.85
Elastic Net	99,383,171.38	525,310,677.99	0.84
Bayesian Ridge	94,090,900.75	535,812,778.28	0.84
Orthogonal Matching Pursuit	94,581,719.24	536,416,382.85	0.84
Extra Trees Regressor	54,007,726.18	546,018,515.40	0.88
Random Forest Regressor	61,542,037.25	623,334,938.39	0.82
Gradient Boosting Regressor	71,735,333.96	650,469,300.67	0.79
Extreme Gradient Boosting	71,045,466.40	665,989,884.80	0.82
K-Neighbors Regressor	75,842,223.20	716,213,561.60	0.80
AdaBoost Regressor	95,790,024.89	769,584,692.62	0.77
Light Gradient Boosting Machine	168,666,774.84	1,071,906,784.02	0.40
Dummy Regressor	295,427,363.20	1,496,258,240.00	-0.01
Least Angle Regression	583,131,082,775.22	1,310,818,230,414.92	-1,257,146.66

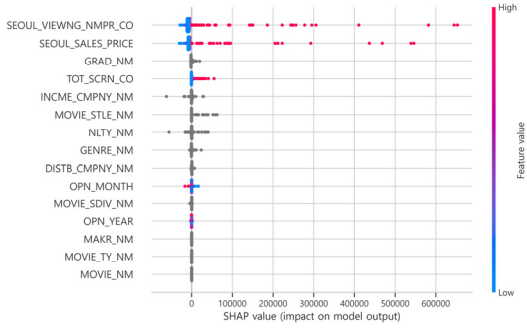
욱 구체적으로 파악할 수 있다. 두 플롯 모두 변수의 중요도를 시각적으로 제시하지만, 서로 다른 관점에서 변수가 모델에 기여하는 방식을 보여준다. 이를 바탕으로 관람 인원수 예측과 매출 예측에서 중요한 변수를 분석하였다.

관람 인원수 예측 모델 분석에서 Fig. 4와 Fig. 5에서 확인할 수 있듯이, SEOUL_VIEWNG_NMPR_CO (서울 관람 인원수)는 관람 인원수 예측에서 CatBoost 모델의 SHAP 분석 결과 가장 중요한 변수로 나타났다. SHAP bar plot에서는 해당 변수가 다른 변수들에 비해 높은 기여도를 보이며, SHAP summary plot에서는 이 변수의 높은 값이 오른쪽으로 몰려 있는 것을 확인할 수 있다. 이는 서울에서 관람 인원이 많을수록 전국적인 관람 인원 예측이 증가하는 경향을 시사한다. 따라서, 서울에서의 관람 인원수가 전국적인 영화 흥행을 결정짓는 핵심 변수로 작용한다는 것을 알 수 있다.

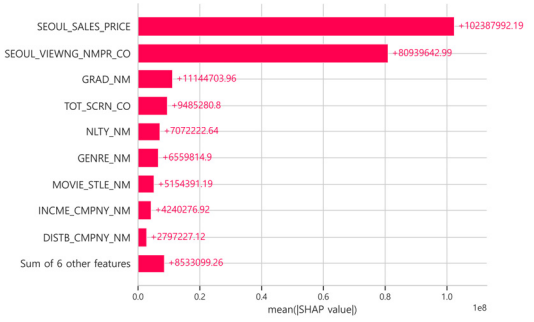
또한, SEOUL_SALES_PRICE (서울 매출 금액)도 관람 인원수 예측에 중요한 변수로 나타난다. Fig. 5에서 확인할 수 있듯이, 서울에서의 높은 매출은 전국적으로 더 많은 관람객을 유치하는 데 기여하며, 매출이 높을수록 관람 인원수 예측도 증가하는 경향을 보인다. 이는 서울에서 매출이 영화의 인지도와 인기에 긍정적인 영향을 미친다는 점을 시사한다. 이 외에도 GRAD_NM (영화 등급), TOT_SCRN_CO (총 상영관 수) 역시 관람 인원수에 큰 영향을 미친다. SHAP 분석 결과, 영화 등급이 낮을수록 더 많은 연령대가 영화를 관람할 수 있어 관람 인원수가 증가하며, 상영관 수가 많을수록 관람 인원수가 더 효과적으로 증가하는 경향이 있다.



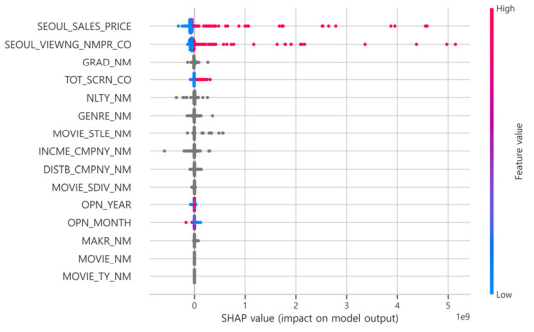
[Fig. 4] SHAP bar plot for key predictors of number of viewers in CatBoost model



[Fig. 5] SHAP summary plot for key predictors of number of viewers in CatBoost model



[Fig. 6] SHAP bar plot for key predictors of box office revenue in CatBoost model



[Fig. 7] SHAP summary plot for key predictors of box office revenue in CatBoost model

매출 예측 모델에 대한 Fig. 6과 Fig. 7에서 확인할 수 있듯이, CatBoost 모델의 SHAP 분석 결과, SEOUL_SALES_PRICE (서울 매출 금액)은 매출 예측에서 가장 중요한 변수로 나타났다. SHAP bar plot에서는 이 변수의 중요도가 두드러지며, SHAP summary plot에서도 서울 매출이 높을수록 전국적인 매출에 큰 기여를 한다는 것을 확인할 수 있다. 이는 서울이 대한민국 영화 시장의 경제적 중심지로서, 영화 매출에 결정적인 역할을 한다는 점을 강조한다.

SEOUL_VIEWNG_NMPR_CO (서울 관람 인원수)도 매출 예측에서 중요한 변수로 작용한다. 서울에서 더 많은 관객이 영화를 관람할수록 전국적인 매출이 증가하는 경향이 있으며, 관람 인원수가 많을수록 매출 예측값도 더 증가하는 모습을 확인할 수 있다. 이는 영화의 인기가 매출에 직결된다는 점을 보여준다. 또한, GRAD_NM (영화 등급)과 TOT_SCRN_CO (총 상영관 수) 역시 매출 예측에 큰 영향을 미친다. 등급이 낮을수록 더 많은 관객을 유치할 수 있고, 상영관 수가 많을수록 매출이 증가하는 경향을 보인다.

CatBoost 모델의 SHAP 분석에서 공통적으로 SEOUL_VIEWNG_NMPR_CO와 SEOUL_SALES_PRICE가 가장 중요한 변수로 나타난다. 관람 인원수와 매출 예측 모두에서 서울 지역의 성과가 전국적인 영화 흥행과 매출에 매우 큰 영향을 미친다. SHAP bar plot은 각 변수의 상대적인 중요도를 명확히 보여주며, SHAP summary plot은 각 변수값이 모델의 예측에 어떤 영향을 미치는지를 직관적으로 시각화한다. 이를 통해 영화 흥행 성과를 예측하고 마케팅 전략을 수립하는 데 중요한 인사이트를 제공할 수 있다.

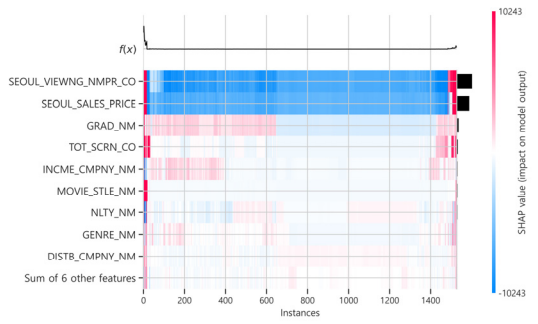
Fig. 8과 Fig. 9는 각각 관람 인원수 예측 모델과 매출 예측 모델에 대한 SHAP heatmap을 시각화한 것이다. 해당 heatmap은 각 변수의 SHAP 값이 여러 샘플에 걸쳐 어떻게 분포하는지 시각적으로 보여주며, 특정 변수들이 모델 예측에 미치는 영향을 색상으로 표현한다. 빨간색은 해당 변수가 예측값을 증가시키는 긍정적인 영향을 나타내며, 파란색은 예측값을 감소시키는 부정적인 영향을 나타낸다.

관람 인원수 예측 모델 분석에서 Fig. 8에서 나타난 바와 같이, SEOUL_VIEWNG_NMPR_CO (서울 관람 인원수)는 관람 인원수 예측에서 중요한 변수로 작용한다. 대부분 샘플에서 파란색으로 나타나, 서울 관람 인원수가 적을 때 전국 관람 인원수 예측이 감소하는 경향을 보여준다. 반면, 일부 샘플에서는 빨간색이 강하게 나타나, 서울 관람 인원수가 많을 때 전국적인 관람 인원 예측이 증가하는 것을 확인할 수 있다.

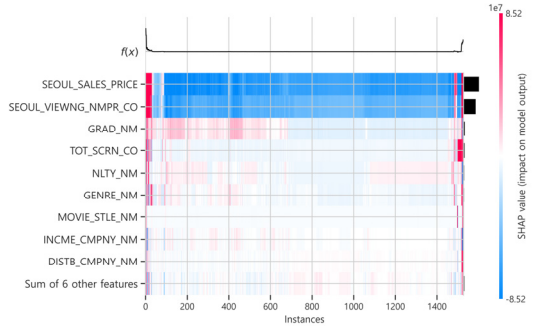
SEOUL_SALES_PRICE (서울 매출 금액)도 관람 인원수 예측에 중요한 변수이다. 서울에서의 매출이 높을수록(빨간색) 관람 인원 예측이 증가하고, 매출이 낮을 경우(파란색) 관람 인원 예측이 감소하는 경향을 보인다. 이는 서울 매출이 관람 인원수에 큰 영향을 미친다는 점을 시사한다.

GRAD_NM (영화 등급)은 영화 등급이 관람 인원수에 미치는 영향을 보여준다. 낮은 등급(모든 연령대가 관람 가능)은 관람 인원수를 증가시키는 경향을 나타내며, 높은 등급(예: 청소년 관람 불가)은 관람 인원수를 줄이는 영향을 미친다. TOT_SCRN_CO (총 상영관 수) 역시 상영관 수가 많을수록 관람 인원수가 증가하는 중요한 변수로 나타났다.

매출 예측 모델 분석(Fig. 9)에서는 SEOUL_SALES_PRICE (서울 매출 금액)가 매출 예측에서 가장 중요한 변수로 나타난다. 이 변수는 샘플에 따라 파란색과 빨간색이 혼재되어 있으며, 빨간색은 서울 매출이 높을 때 전국 매출이 크게 증가하는 경향을 보여준다. 반대로 파란색은 매출이 낮을 때 전체 매출 예측이 감소하는 것을 시사한다.



[Fig. 8] SHAP heatmap for key predictors of number of viewers in CatBoost model



[Fig. 9] SHAP heatmap for key predictors of box office revenue in CatBoost model

SEOUL_VIEWNG_NMPR_CO (서울 관람 인원수) 역시 매출 예측에 큰 영향을 미친다. 관람 인원수가 많을 경우(빨간색) 전체 매출이 증가하고, 관람 인원수가 적을 경우(파란색) 매출이 감소하는 경향을 보여준다. 또한, GRAD_NM (영화 등급)은 등급이 낮을 때 매출이 증가하는 경향이 있으며, 높은 등급일수록 매출이 감소하는 것을 확인할 수 있다. TOT_SCRN_CO (총 상영관 수)도 상영관 수가 많을수록 매출이 증가하는 중요한 변수로 나타난다.

두 heatmap plot은 SEOUL_VIEWNG_NMPR_CO와 SEOUL_SALES_PRICE가 관람 인원수와 매출 예측 모두에서 중요한 변수임을 명확히 보여준다. 서울 지역의 성과가 영화의 전국적인 흥행 성과와 매출에 큰 영향을 미친다는 것을 확인할 수 있다. 또한, 각 변수의 SHAP 값이 샘플별로 어떻게 변동하는지 시각화하여, 변수 간의 상호작용과 모델 예측에 미치는 영향을 구체적으로 이해할 수 있게 한다.

4.3 추가 분석 및 연구 한계

개봉 시점은 계절적 요인, 관객의 선호도 변화, 경쟁 작품의 개봉 여부 등과 밀접한 관련이 있어 영화의 흥행 성과에 큰 영향을 미친다. 특히 개봉 월을 예측함으로써 영화 산업에서 마케팅 전략 수립, 상영 일정 계획, 관객 타겟팅 등 다양한 측면에서 실질적인 도움을 줄 수 있다. 본 연구에서는 개봉 월(OPN_MONTH)을 종속 변수로 설정하여 영화의 개봉 시점을 예측하는 분류 모델을 구축하였다.

〈Table 7〉 Performance comparison of classification models for movie release month prediction

Algorithm	Accuracy	F1 Score
Categorical Boosting	0.20	0.18
K Neighbors Classifier	0.15	0.13

앞서와 마찬가지로 데이터는 전체 데이터의 70%를 훈련 집합으로, 30%를 평가 집합으로 분할하여 모델의 일반화 능력을 평가하였다. CatBoost 알고리즘과 PyCaret에서 가장 성능이 우수한 모델인 K-최근접 이웃(K Neighbors Classifier)을 비교한 결과, Table 7과 같이 CatBoost가 K-최근접 이웃 알고리즘보다 우수한 성능을 도출하였다. 그러나 전체적으로 모델의 성능이 높지 않았으며, 이는 개봉 월 예측에 있어 추가적인 연구와 모델 개선의 필요성을 시사한다.

본 연구에서는 관객 수와 매출액을 예측하는 모델을 구축하여 우수한 성능을 달성하였다. 특히, CatBoost 알고리즘을 활용한 모델은 관객 수 예측에서 결정 계수(R^2) 0.97, 매출액 예측에서 R^2 0.95를 기록하며 높은 정확도를 보였다. 이러한 결과는 영화 산업에서 데이터 기반 의사결정에 크게 기여할 수 있을 것으로 기대된다. 반면에, 개봉 월 예측 모델의 성능은 상대적으로 낮게 나타났다. 개봉 월은 영화 제작사의 전략, 마케팅 일정, 경쟁 작품의 개봉 계획 등 복합적인 요인에 의해 결정되므로, 단순한 변수만으로는 정확한 예측에 한계가 있었다. 따라서 개봉 월 예측은 향후 추가적인 변수 도입과 모델 개선을 통해 성능을 향상할 여지가 있다.

향후 연구 방향은 다음과 같다:

- 추가적인 변수 도입: 날씨 정보, 사회적 이벤트 일정, 경쟁 작품의 개봉 일정, 마케팅 예산, 온라인 관심도(검색 트렌드, 소셜 미디어 언급량) 등 개봉

월 결정에 영향을 미치는 다양한 요인을 데이터에 포함하여 모델의 설명력을 높이고자 한다.

- 데이터 규모 확대 및 다양화: 데이터 수집 기간을 연장하고, 다양한 국가와 장르의 영화를 포함하여 데이터의 양과 다양성을 확보함으로써 모델의 일반화 능력을 향상하고자 한다. 이를 통해 계절적 패턴이나 문화적 차이도 고려하고자 한다.
- 고급 모델의 적용 및 최적화: 딥러닝 기반의 모델이나 시계열 분석 기법을 활용하여 비선형적이고 복잡한 패턴을 포착함으로써 예측 성능을 개선할 수 있다. 예를 들어, 그래프 신경망이나 전이 학습을 사용하여 시간적 흐름을 고려한 예측을 시도하고자 한다.
- 실무 적용을 위한 모델 개선: 영화 산업의 실제 의사결정 과정에 예측 모델을 적용할 수 있는 방안을 모색하고자 한다. 예를 들어, 배급사의 개봉 일정 계획, 마케팅 전략 수립 등에 모델의 예측 결과를 활용할 수 있는 프로세스 구축 및 모델의 실효성을 검증한다.

이를 통해 개봉 월 예측 모델의 정확도를 높이고, 실시간 데이터 수집 및 분석을 통해 영화 산업에서의 데이터 기반 의사결정에 이바지할 수 있을 것으로 기대한다. 향후 연구에서는 더 풍부한 데이터를 활용하고, 연결된 디바이스 및 복합적인 요인을 고려한 모델링을 통해 예측 성능을 향상하며, 실질적인 비즈니스 인사이트를 제공하고자 한다.

5. 결론

본 연구에서는 KOFIC에서 제공하는 영화 박스오피스 데이터를 활용해, 영화의 관람 인원수와 매출액의 예측 모델을 개발하였다. 데이터 전처리 과정에서는 불필요한 변수를 제거하고, 결측치는 범주형 및 수치형 데이터에 따라 각각 처리하여 데이터의 일관성을 유지하였다. 범주형 데이터는 “Unknown”으로 대체하고, 수치형 결측치는 분석에서 제외하였다. 또한, 개봉일자 형식을 통일하고 중복된 데이터를 통합함으로써 분석 왜곡을 방지하였다. EDA을 통해 개봉 월별 매출과 장르별 관람 인원수의 특성을 시각화하였으며, 매출 금액과 관람 인원수 간의 높은 상관관계를 확인하였다.

예측 모델링 단계에서는 CatBoost와 PyCaret AutoML을 사용하여 성능을 비교하였다. CatBoost는 고차원 범주형 변수를 효과적으로 처리하며, 그리드 탐색을 통해 최적의 하이퍼파라미터를 찾아 높은 예측 성능을 보였다. PyCaret AutoML은 다양한 회귀 알고리즘을 자동으로 비교하고 최적화하여 예측 모델을 구성할 수 있으며, 사물인터넷(IoT; Internet of Things) 기반 환경에서도 실시간 데이터를 처리하기에 적합하다. 특히 비전문가도 쉽게 사용할 수 있는 점에서 강점을 보였다. 결과적으로, CatBoost는 관람 인원수와 매출액 예측 모두에서 PyCaret보다 약간 더 높은 정확도를 보였으나, PyCaret의 자동화된 파이프라인은 사용 효율성에서 강점을 보였다.

또한, SHAP 기법을 통해 모델의 예측 결과를 해석하였다. 분석 결과, 서울 지역의 관람 인원수와 매출액이 두 모델 모두에서 가장 중요한 변수로 나타났으며, 이는 서울에서의 성과가 영화 흥행에 큰 영향을 미친다는 점을 시사한다. SHAP 분석을 통해 각 변수의 기여도를 시각적으로 확인할 수 있었으며, 이는 IoT 기술이 접목된 스마트 영화관의 데이터 기반 전략적 의사결정에 유용한 인사이트를 제공할 수 있다. 따라서 본 연구에서 도출된 예측 모델은 영화 산업의 마케팅 전략, 상영관 배정, 예산 할당 등 실무적 의사결정에 중요한 데이터를 제공하며, 영화 흥행 예측의 정확성을 높이는 데 기여할 수 있다.

본 연구는 한국 영화 데이터를 기반으로 진행되었으나, CatBoost와 PyCaret의 유연성과 범주형 변수를 처리하는 특성은 국제적인 영화 시장에도 적용될 수 있는 잠재력을 가진다. 현재 데이터 셋에는 38개국의 박스오피스 데이터가 포함되어 있으나, 한국 영화의 비중이 상대적으로 높아 데이터 불균형 문제가 존재하였다. 이러한 불균형을 해소하기 위해 향후 연구에서는 데이터 증강 기법을 도입하거나, 샘플링 방법을 활용하여 다양한 국가의 영화를 균형 있게 분석할 계획이다. 특히, 글로벌 영화 시장에서 중요한 할리우드 영화를 중심으로 분석을 확장함으로써 모델의 일반화 가능성을 높이고자 한다. 이를 통해 글로벌 영화 데이터를 분석하고, 다양한 국가의 영화 산업에서의 의사결정 지원에 기여할 수 있을 것으로 기대된다. 추가로, 다양한 문화적 배경과 시장 특성을 반영한 맞춤형 예측 모델을 개발하여 국제적 활용성을 더욱 강화할 예정이다. 이를 통해 기업들은 시장 수요를 더욱 정확하게 반영하고, 전략적 결정을 내리는 데 신뢰성 있는 예측 도구를 활용할 수 있을 것으로 기대한다.

끝으로, 본 연구는 순천향대학교 SW중심대학사업단의 ‘SW 명문중학교 만들기’ 프로그램[28]의 일환으로 진행되었으며, 순천향대 AI·빅데이터학과 학생인 김희성과 아산중학교 학생인 김승혁이 멘토-멘티로 참여하였다.

연구 과정에서 김승혁 학생은 PyCaret AutoML을 활용하여 영화 박스오피스의 관람 인원수와 매출액을 예측하는 모델을 구축하였으며, 이를 통해 기계학습 모델의 자동화 및 활용 방법을 학습하였다. 다른 연구인 데이터 전처리, CatBoost 모델 개발 및 SHAP 해석 등은 문지훈 교수의 지도하에 김희성 학생이 수행하였다. 이러한 협력은 대학생과 중학생 간의 지식 공유와 교육적 연계를 강화하여 중학생들의 SW 및 데이터 분석 분야에 관한 관심과 이해를 높이는 데 이바지하고자 하였다. 본 연구를 통해 지역 사회를 넘어 국가 차원의 SW 교육 발전과 미래 인재 양성에 이바지할 수 있는 시작점이 될 수 있기를 기대한다.

REFERENCES

[1] R.Behrens, N.Z.Foutz, M.Franklin, J.Funk, F.Gutierrez-Navratil, J.Hofmann and U.Leibfried, "Leveraging analytics to produce compelling and profitable film content," *Journal of Cultural Economics*, Vol.45, pp.171-211, 2021.

[2] A.Barrios-Rubio, "The Colombian media industry on the digital social consumption agenda in times of Covid-19," *Information*, Vol.13, No.1, p.11, 2021.

[3] A.De Vany and W.D.Walls, "Uncertainty In The Movie Industry: Does Star Power Reduce The Terror Of The Box Office?" *Journal of Cultural Economics*, Vol.23, No.4, pp.285-318, 1999.

[4] J.Song, H.Cho and K.Kim, "Development Of New Variables Affecting Movie Success And Prediction Of Weekly Box Office Using Them Based On Machine Learning," *Journal of Intelligence and Information Systems*, Vol.24, No.4, pp.67-83, 2018.

[5] N.Quader, M.O.Gani, D.Chaki and M.H.Ali, "A machine learning approach to predict movie box-office success," in Proceedings of the *2017 20th International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2017, pp.1-7.

[6] V.Subramaniyaswamy, M.V.Vaibhav, R.V.Prasad and R.Logesh, "Predicting movie box office success using multiple regression and SVM," in Proceedings of the *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, Palladam, India 2017, pp.182-186.

[7] G.Bae, H.-J.Kim, "Using Topic Modeling to Analyze Transition of Movie Themes in South Korea," *Korean Journal of Marketing*, Vol.36, No.3, pp.1-24, 2021.

[8] T.Bao, H.Kim and B.-H.Chang, "An Expoloratory Study on Influencing Factors of Film Equity Crowdfunding Success: Based on Chinese Movie Crowdfunding," *The Journal of the Korea Contents Association*, Vol.21, No.2, pp.1-14, 2021.

[9] J.H.Byun, J.H.Kim, Y.J.Choi, and H.C.Lee, "Movie Box-office Prediction using Deep Learning and Feature Selection: Focusing on Multivariate Time Series," *Journal of The Korea Society of Computer and Information*, Vol.25, No.6, pp.35-47, 2020.

[10] S.Leem, J.Oh, D.So and J.Moon, "Towards Data-Driven Decision-Making in the Korean Film Industry: An XAI Model for Box Office Analysis Using Dimension Reduction, Clustering, and Classification," *Entropy*, Vol.25, No.4, p.571, 2023.

[11] S.Leem, J.Moon and S.Rho, "A Box Office Type Classification and Prediction Model Based on Automated Machine Learning for Maximizing the Commercial Success of the Korean Film Industry," *Journal of Platform Technology*, Vol.11, No.3, pp.45-55, 2023.

[12] B.Avanzi, G.Taylor, M.Wang and B.Wong, "Machine learning with high-cardinality categorical features in actuarial applications," *ASTIN Bulletin: The Journal of the IAA*, Vol.54, No.2, pp.213-238, 2024.

[13] L.Prokhorenkova, G.Gusev, A.Vorobev, A.V.Dorogush, and A.Gulin, "CatBoost: unbiased boosting with categorical features," *Advances in Neural Information Processing Systems*, Vol.31, 2018.

[14] J.Moon, S.Rho and S.W.Baik, "Toward explainable electrical load forecasting of buildings: A comparative study of tree-based ensemble methods with Shapley values," *Sustainable Energy Technologies and Assessments*, Vol.54, p.102888, 2022.

[15] S.M.Lundberg, G.Erion, H.Chen, A.DeGrave, J.M.Prutkin, B.Nair, R.Katz, J.Himmelfarb, N.Bansel and S.-I.Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, Vol.2, No.1, pp.56-67, 2020.

[16] Korean Film Council, Korean Film Council official website [Internet], https://www.kofic.or.kr/kofic/business/main/main.do

[17] T.Chen and C.Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp.785-794.

[18] G.Ke, Q.Meng, T.Finley, T.Wang, W.Chen, W.Ma, Q.Ye and T.Y.Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Proceedings of the *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017.

[19] S.O.Arik and T.Pfister, "TabNet: Attentive Interpretable Tabular Learning," in Proceedings of *the AAAI Conference on Artificial Intelligence*, New York, NY, USA, 2020, pp.6679-6686.

[20] Y.Gorishniy, I.Rubachev, V.Khrulkov and I.Oseledets, "Revisiting Deep Learning Models for Tabular Data," *arXiv preprint* arXiv:2106.11959, 2021.

[21] V.K.Singh and K.Joshi, "Automated machine learning (AutoML): an overview of opportunities for application and research," *Journal of Information Technology Case and Application Research*, Vol.24, No.2, pp.75-85, 2022.

[22] J.O.Q.Quispe, A.C.F.Quispe, N.C.L.Calvo and O.C.Toledo, "Analysis and Selection of Multiple Machine Learning Methodologies in PyCaret for Monthly Electricity Consumption Demand Forecasting," *Materials Proceedings*, Vol.18, No.1, p.5, 2024.

[23] P.Whig, K.Gupta, N.Jiwani, H.Jupalle, S.Kouser and N.Alam, "A novel method for diabetes classification and prediction with Pycaret," *Microsystem Technologies*, Vol.29, No.10, pp.1479-1487, 2023.

[24] Y.Kim, Y.C.Byun, and S.J.Lee, "A Study on Sugar Content Improvement and Distribution Flow Response through Citrus Sugar Content Prediction Based on the PyCaret Library," *Horticulturae*, Vol.10, No.6, p.630, 2024.

[25] M.R.Zafar and N.Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," *Machine Learning and Knowledge Extraction*, Vol.3, No.3, pp.525-541, 2021.

[26] J.Moon, S.Park, S.Rho and E.Hwang, "Robust building energy consumption forecasting using an online learning approach with R ranger," *Journal of Building Engineering*, Vol.47, p.103851, 2022.

[27] J.Chung, J.Li, A.I.Saimon, P.Hong and Z.Kong, "Predicting the stereoselectivity of chemical reactions by composite machine learning method," *Scientific Reports*, Vol.14, No.1, p.12131, 2024.

[28] Campus News, The Chosun Daily [Internet], https://lifenlearning.chosun.com/pan/site/data/html_dir/2024/05/31/2024053101931.html

김 희 성(Huiseong Kim) [준회원]

■ 2021년 3월 ~ 현재 : 순천향대학교 AI·빅데이터학과 학사과정 재학 중

〈관심분야〉

데이터 과학, 알고리즘 개발, 사물인터넷, 인공지능 응용

문 지 훈(Jihoon Moon) [정회원]

■ 2015년 2월 : 한성대학교 정보통신공학과 (공학사)
■ 2021년 2월 : 고려대학교 전기전자공학과 (공학박사)
■ 2021년 6월 ~ 2022년 8월 : 중앙대학교 박사후연구원
■ 2022년 9월 ~ 현재 : 순천향대학교 AI·빅데이터학과 조교수

〈관심분야〉

지속 가능한 솔루션, 설명 가능한 인공지능, 딥러닝 응용, 데이터 마이닝, 시계열 분석 등