딥러닝 기반 게임 흥행 예측

정보시스템공학과
성신여자대학교

Abstract

게임 산업에서는 하나의 게임을 제작하기 위해서 막대한 자원을 투자하고, 게임의 흥행 여부에 따라 게임사의 존립이 위협받기도 한다. 따라서 게임 산업에의 투자 및 마케팅 등에 활용하기 위해 게임 흥행 여부를 사전에 예측하는 것이 매우 중요하다. 지금까지는 주로 영화 분야에서 흥행 여부를 예측하는 연구들이 주류를 이루어 왔고, 최근에는 블로그, 기사, 영화평 등의 소셜 데이터를 활용하여 영화의 흥행을 예측하는 연구가 진행되고 있다. 본 연구에서는 게임에 대한 정보를 이용해 딥러닝 기반으로 게임의 흥행 여부를 조기에 예측하는 모델을 설계, 구현하여 게임 발매 이전에 게임의 흥행 여부를 예측하여 게임 제작 및 투자에 유용한 정보를 제공하고자 한다. 예측에는 대표적인 게임 소프트웨어 유통 플랫폼인 'Steam'에 등록된 게임과, 그 정보를 이용하였고, 데이터의 구성에 따라 게임 발매 전 정보만을 포함한 실험군과 게임 발매 이후의 데이터까지 포함한 대조군으로 구분하였다. 예측 모델은 DNN, Random forest, Ada boost를 이용하여, 게임의 흥행 여부를 예측하였다. 연구 결과 같은 모델이라도 게임 발매 이후에 얻을 수 있는 주요 정보가 다수 포함된 대조군의 예측 정확도가 더 높았고, 실험군의 예측 모델 중에서는 Ada boost 모델이 가장 높은 정확도를 보였다.

1 서론

게임 산업에서는 하나의 게임을 제작하기 위해서 막대한 인원, 시간과 예산을 투자하고 게임의 흥행 여부에 따라 크게 영향을 받아 게임사의 존립이 위협받기도 한다. 따라서 게임 산업에의 투자 및 마케팅 등에 활용하기 위해 게임 흥행 여부를 사전에 예측하는 것이 매우 중요하다. 지금까지는 주로 영화 분야에서 흥행 여부를 예측하는 연구들이 주류를 이루어 왔다. 예전에는 영화에 대한 정보만을 이용한 예측이 주류를 이루었으나, 최근에는 블로그, 기사, 영화평 등의 소셜 데이터를 활용하여 영화의 흥행을 예측하는 연구가 진행되고 있다. 이는 송정아 외 2인의 영화 흥행 예측에 영향을 미치는 새로운 변수 개발과 주간 박스오피스 예측 연구[1]에서 확인할 수 있다. 본 연구에서는 게임에 대한 정보를 이용해 딥러닝 기반으로 게임의 흥행 여부를 조기에 예측하는 모델을 설

계, 구현하여 게임 발매 이전에 게임의 흥행 여부를 예측하여 게임 제작 및 투자에 유용한 정보를 제공하고자 한다.

2 선행 연구

성민 외 2인의 영화 흥행 조기 예측[2] 연구에서는 영화 등급, 장르, 제작사 등의 내재적 정보와 온라인 리뷰 수, 티켓 판매량, 리뷰 평점 등의 외재적 정보를 이용해 영화 개봉 이후 주차 별 티켓 판매량을 예측하였다. 데이터는 전 세계 영화에 대한 정보가 있는 IMDb에서 영화의 기본 정보를, 웹크롤링을 이용해 기타 정보를 수집하였고, 각각의 특성에 맞게 원-핫 인코딩, 정규화, 비율 환산 등의 전처리를 수행했다. 이 연구는 여러 가지 머신 러닝 모델을 통해 영화의 흥행을 예측하였으며 랜덤 포레스트와 XGBoost가 가장 좋은 결과를 냈다. 또한 리뷰 수를 데이터에 추가하면 성능이 향상되는 경향을 보였다고 하나 이 연구는 정성적 요소를 평가에 포함하지 않았다.
송정아 외 2인의 영화 흥행 예측 변수 개발 연구에서는 기존에 사용되지 않았던 주차별 평균 매출 점유율, 주차별 순위 및 순위 변동 폭, 개봉 후 누적 네티즌 평가 등의 새로운 변수를 도입하여 주차 별 누적 관람객 수를 예측하고, 도입한 변수의 효용을 확인하였다. 새로운 변수를 추가한 데이터를 이용한 모델이 그렇지 않은 모델보다 통계적으로 유의미한 수준에서 좋은 성능을 보여 추가된 변수들이 정확도 향상에 기여했다고 결론 내렸다. 그리고 단일 알고리즘 모델보다 Random forest와 같은 앙상블 모델이 더 좋은 성능을 보인 것에 대해 이 연구에서 변화가 많은 다양한 요소를 활용했기 때문에 여러 가지 모델이 결합된 앙상블 모델이 이 연구에 더 적합했다고 판단하였다.
이도연과 장병희의 음악 흥행 예측 연구[3]에서는 음악 분야의 데이터와 딥러닝 기법을 이용하여 흥행 예측 모델 구축 가능성을 살펴보았다. 가수 영향력, 유통사 역량, 참여 가수 성별 등 17개 흥행 요인을 기반으로 음원이 차트 내에 상주하는 기간을 예측한다. 또한 이 연구는 탐색적인 수준에서 이루어졌기 때문에 변인을 아무것도 소거하지 않은 심층 신경망 모델과 선형회귀분석을 통해 변인 소거 후 구축한 심층 신경망 예측모델과의 예측률 비교를 진행하였다. 데이터 수집 과정에서 브랜드 이론, 노출효과 이론, 편승효과 이론과 선행 연구에 근거하여 수집할 흥행 요인을 선정하였고, 체계적으로 구축된 데이터베이스의 부재로 다양한 플랫폼에서 직접 데이터를 수집했다. 연구 결과 변인을 소거하지 않은 모델에 비해 선형회귀분석에 근거해 일부 변인을 소거하면 모델의 성능이 개선되었으며, 표본 데이터가 너무 적어 심층 신경망 모델이 스스로 학습하여 최적화가 가능하다는 개념에 따르지 않는 한계가 있었다. 심층 신경망 모델은 데이터가 많을 수록 좋은 성능을 낼 수 있으므로 데이터베이스의 구축을 강조하였다. 본 연구는 따로 구축된 데이터베이스가 없고 수집한 데이터도 많지 않아 모델의 최적화 능력을 기대하기 힘드므로 일정한 기준에 따라 속성을 선별함으로써 성능 향상을 시도할 수 있다.
강지훈 외 3인의 영화 흥행 실적 예측 기법 연구[4]는 설명력 있고 정확도 높은 관객 수

---

1) 송정아 · 최근호 · 김건우(2018, 12월). 영화 흥행에 영향을 미치는 새로운 변수 개발과 이를 이용한 머신러닝 기반의 주간 박스오피스 예측. *J Intell Inform Syst*, 24(4), 67-83.
2) 성민 · 김동성 · 김종우(2021, 6월). <온라인 리뷰에 대한 특성 기반 감성 분석을 활용한 영화 흥행 조기 예측>. 대한산업공학회 춘계공동학술대회 논문집. 제주: 국제컨벤션센터.
3) 이도연 · 장병희. (2020). 딥러닝을 이용한 음악흥행 예측모델 개발 연구. 한국콘텐츠학회 논문지, 20(8), 10-18.
4) 강지훈 · 박찬희 · 도형록 · 김성범(2014, 5월). 데이터마이닝 기법을 활용한 영화 흥행 실적 예측 기법. 대한산업공학회 춘계공동학술대회 논문집, 142-154.

요 예측 시스템을 구현하여 영화 산업의 투자와 관련 정책 수립을 위한 기초 자료로서의 활용이 가능하도록 하는 것을 목적으로 한다. 제작사, 등급, 감독과 주연배우의 흥행력, 개봉 전 기사 수 등 타당한 근거 하에 수집한 영화의 사전 정보와 15일간의 관객 실적 패턴으로 데이터셋을 구성하고, 모델의 강건성, 예측의 정확성, 모델의 해석 능력을 기준으로 의사결정트리를 예측 모델로써 선정했다. 상영 중인 영화는 영화의 기본 정보와 1일차 스크린 수를 입력하여 특정 일자의 누적 관객 수를 예측하고 개봉 전 영화는 중간 예측 모델을 두어 1일차 스크린 수를 예측한 후 상영 중 영화와 같은 방식으로 예측하였다. 연구 결과 다양한 영화의 누적 관객 수를 비교적 정확히 예측하였으며, 구체적으로 일정 규모 이상의 관객 확보를 위해서는 어느 정도의 스크린 확보가 필요한지 제시하였다. 본 연구는 예측의 정확성을 우선시하여 DNN과 Random forest, Ada boost로 예측 모델을 선정하였다.

이동석의 영화 관객수 예측에 관한 연구[5]에서는 영화의 흥행 요소를 개봉 전 변수와 개봉 후 변수로 나누어 수집하고 다양한 기계학습 기법을 이용하여 예측을 시도하였는데, 예측을 시행하기 전 데이터를 분석하여 장르별로 최종 관객 수의 경향이 다르다는 사실을 발견하고 장르별로 흥행 성과를 예측하였고 전체 데이터에 대한 예측 성능보다 우수함을 확인하였다.

장재영의 영화 흥행 예측 기법 연구[6]에서는 나이브베이즈 분류 모델과 신경망을 이용해 영화의 정적 데이터와 동적 데이터를 기반으로 흥행을 예측하였다. 세부적인 예측을 위해 일정 기간 내의 상위 랭크 영화만을 대상으로 분석하였고, 수집한 정적 데이터와 동적 데이터는 총 15가지 구성으로 조합하여 성능을 비교했다. 실험 결과 정적 데이터와 동적 데이터를 모두 사용한 모델의 성능이 가장 우수하였으며 실험에 사용한 두 가지 모델 중에서는 신경망이 전반적으로 더 좋은 성능을 보였다.

권신혜 외 2인의 영화 흥행 예측 방법 연구[7]에서는 영화의 제작부터 상영까지의 과정을 세 단계로 나누어 단계별로 선형회귀분석을 통해 흥행에 유의미한 변인을 파악하고, 인공신경망과 의사결정나무 모델로 흥행을 예측했다. 연구 결과 인공신경망은 선형회귀분석을 통한 변인 선정의 영향이 크지 않았지만 의사결정나무는 변인 선정으로 정확도가 향상되었고, 해당 연구에서 다루었던 세 가지 단계별 성능이 각기 달랐기 때문에 영화 산업의 단계별 모형마다 적합한 분석 기법이 다를 수 있다고 판단했다.

송정아 외 2인의 연구[8]에서는 의사결정나무, MLP 신경망모형, 다항로짓모형, support vector machine을 이용해 영화의 개봉 전후 다양한 시점에 대해 총 관객 수를 예측하였다. 예측 변수로는 영화의 내재적 정보와 블로그, 뉴스, 포탈 사이트 평점 등의 외재적 정보를 활용하였는데, 외재적 정보의 경우 추가로 모델을 구성하여 추정한 값을 사용하였다. 외재적 정보를 추정하지 않고 내재적 정보만으로 구현한 모델과 다중회귀모형과 MLP 모형을 이용해 추정한 두 가지의 외재적 정보를 내재적 정보와 같이 이용한 모델의 정확도를 비교하였고, 외재적 정보를 함께 이용한 모델이 그렇지 않은 모델보다 정확도가 향상되는 경향을 확인하였다.

---

5) 이동석(2022). "양방향 국소평균 K 최근접 이웃 방법론을 이용한 영화 관객수 예측에 관한 연구." 국내석사학위논문, 한양대학교 대학원.

6) 장재영(2017). "소셜 빅데이터 분석과 기계학습을 이용한 영화흥행예측 기법의 실험적 평가", 한국인터넷방송통신학회 논문지, 17(3), 167-173.

7) 권신혜 · 박경우 · 장병희(2017). "기계학습 기반의 영화흥행예측 방법 비교: 인공신경망과 의사결정나무를 중심으로", 예술인문사회융합멀티미디어논문지, 7(4), 593-601.

8) 송정아 · 최근호 · 김건우(2018). "기계학습을 이용한 주간 박스오피스 예측", 한국지능정보시스템학회 학술대회논문집, 58-71.

임준엽과 황병연의 영화 흥행 예측 연구[9]에서는 영화가 개봉되기 전후로 영화를 관람하지 않았지만 영화에 대해 인지하고 있는 잠재 관객의 인지도를 반영하기 위해 기존 연구에서 사용하던 영화의 내/외재적 속성뿐만 아니라 포털 사이트의 데이터와 트위터의 데이터를 온라인 요소로써 이용하였다. 예측 시점은 개봉 전 1주일과 개봉 후 1주일 두 가지이고, 각 영화를 관람한 최대 관객 수를 범주화하여 목표 변수로 사용했다. 예측 모델은 나이브 베이즈 분류를 사용하였으며 해당 모델의 특성상 연속형 변수를 입력할 수 없어 가우지안 함수를 이용하였다. 실험 결과 온라인 요소와 오프라인 요소에 대해서는 온라인 요소가 대체로 예측 정확도에 미치는 영향이 컸고, 온라인 요소가 포함된 경우가 그렇지 않은 경우보다 정확도가 향상되는 경향을 보였다. 이후 실험을 통해 데이터의 개수에 따른 정확도 변화도 확인하였고, 최종적으로 온라인 요소를 포함한 데이터셋에 대해 95%의 정확도를 기록하였다.

Tiffany D. Do 외 4인의 연구[10]에서는 라이엇 게임즈의 'League of Legends(LoL)' 게임에 대하여 플레이어가 선택한 챔피언에 대한 플레이어의 경험을 기반으로 랭크 매치 결과를 예측하였다. 라이엇 게임즈가 제공하는 공개 API를 이용해 데이터를 수집하고, 게임 결과와 관련된 요소를 파악해 챔피언 마스터 포인트, 플레이어-챔피언 승률, 팀의 평균 등 44개의 속성으로 구성되고, 총 5천 개의 중복되지 않은 랭크 매치 데이터로 이루어진 데이터셋을 이용해 예측을 진행했다. 예측을 위한 모델은 Support Vector Classifiers(SVC), k-Nearest Neighbors(kNN), Random Forest(RF) trees, Gradient Boosting(GBOOST), Deep Neural Networks(DNN)로 총 네 가지를 이용했다. 각 모델의 특성에 맞게 파라미터를 조정하여 예측 모델을 구현하고 정확도를 평가한 결과 모든 모델에서 70% 이상의 정확도가 나왔으며 그중 GBOOST 모델과 DNN 모델의 정확도가 가장 높았다. 이 연구에서는 당시의 LoL 랭크 매치 시스템과 달리 플레이어-챔피언 경험만을 바탕으로 게임 플레이가 시작되기 전에 랭크 게임의 결과를 비교적 높은 정확도로 결정할 수 있음을 보여주었다.

본 연구에서는 선행 연구와 분야는 다르지만 중요 요소는 비슷하다고 판단하여 게임의 장르, 제작자, 배급사 등의 내재적 정보와 리뷰 수, 평가 등급 등의 외재적 정보를 이용해 예측을 시도하였으며, 예측 모델은 DNN과 Random forest, Ada boost 모델을 사용하였다.

3 연구 설계

3.1 데이터 수집 및 데이터 정보

본 연구는 2017년 1월 1일 이후부터 2022년 6월 1일 이전까지 ‘Steam’에 출시된 유료 게임 828개를 대상으로 하여 ‘Steam’과 ‘steamcharts.com’에서 게임의 데이터를 수집하였고, ‘steamcharts.com’은 ‘Steam’과는 별개로 ‘Steam’에서 판매되는 게임의 동시 플레이어 수에 대한 정보를 제공하는 웹사이트이다. 먼저 ‘Steam’의 web API와 ‘Steam’ 상점의 인기 제품 페이지에 노출되는 게임 목록을 이용해 게임 ID를 수집하였

9) 임준엽 · 황병연(2014). "트위터를 이용한 기계학습 기반의 영화흥행 예측", 정보처리학회논문지. 소프트웨어 및 데이터 공학 3 (7), 263-270.

10) Do, Tiffany D. et al. “Using Machine Learning to Predict Game Outcomes Based on Player-Champion Experience in League of Legends.” *The 16th International Conference on the Foundations of Digital Games (FDG) 2021* (2021): n. pag.

다. 수집한 게임 ID로 ‘Steam’ 제품 페이지 링크를 생성하여 게임의 출시일자, 가격, 개발자 및 배급사 정보, 장르, ‘앞서 해보기’ 여부, 도전과제 개수, 태그, 지원 언어 수, ‘스팀 어워드’ 수상 여부, DLC(Downlodable Contents) 유무, 최근 및 모든 평가 등급, 싱글 혹은 멀티 여부, 추천 및 비추천 리뷰 수, 플레이어가 가장 많이 달성한 도전과제의 달성률, 가장 많은 ‘유용함’ 투표를 받은 리뷰를 수집하고, 같은 방식으로 ‘steamcharts.com’의 제품 페이지 링크 또한 생성하여 각 게임 별 최근 30일 평균 동시 플레이어 수와 최고 동시 플레이어 수를 수집했다. 전처리를 수행하기 전 수집한 정보를 정리하면 <표 1>과 같다.

<표 1> 수집한 데이터

출처	이름	설명
‘Steam’ 제품 페이지	게임 ID	‘Steam’ 제품에 부여되는 고유 ID
	출시 일자	게임 출시 일자(연, 월, 일)
	가격	한국 원화 기준 정가
	개발자 정보	개발자 페이지 팔로우 수에 따른 등급
	배급사 정보	배급사 페이지 팔로우 수에 따른 등급
	장르	1개 수집, 여러 개라면 첫 번째 수집
	태그	상위 태그 1개 수집
	도전과제 개수	게임 별 도전과제 개수
	도전과제 달성률	게임 별 도전과제 중 가장 많이 클리어한 도전과제 1개의 달성률
	지원 언어 수	자막, 음성, 인터페이스 등 게임 내에서 지원하는 언어의 수
	‘앞서 해보기’ 여부	‘앞서 해보기’ 유무 수집
	스팀 어워드 수상 여부	스팀 어워드 수상 여부 수집
	DLC 유무	DLC 개수에 상관 없이 유무 수집
	최근 평가 등급	최근 30일간 게시된 리뷰의 긍정성 등급
	모든 평가 등급	모든 리뷰의 긍정성 등급
	싱글/멀티 여부	싱글 플레이, 멀티 플레이 지원 여부
	추천 리뷰 수	‘추천’으로 게시된 리뷰 수
	비추천 리뷰 수	‘비추천’으로 게시된 리뷰 수
	흥행 여부	‘Steam’ 인기 제품 페이지에 있었던 상품을 기준으로 흥행과 비흥행으로 구분
‘steamcharts.com’	최근 동시 플레이어 수	최근 30일 평균 동시 플레이어 수
	최고 동시 플레이어 수	모든 기간 최고 동시 플레이어 수

개발사 및 배급사 정보는 전부 텍스트로 수집하여 라벨링하기엔 너무 다양하여, ‘Steam’ 플랫폼 상에서 별도의 웹페이지가 존재하는지, 존재한다면 팔로워 수가 얼마나 되는지를 수집하였고, 팔로워 수에 따라 5등급, 페이지가 없을 경우까지 포함하여 총 6등급으로 나누어 등급화한 정보를 저장하였다. 장르는 일반적으로 한 게임 당 여러 개를 갖는 경우가 많은데, 이 중 첫 번째만을 수집하였다. 태그는 사용자들이 게임을 잘 나타낸다고 생각하는 것을 직접 추가하거나 이미 있는 태그에 투표하여 바꿀 수 있는 것으로, 장르와 마찬가지로 상위 1개만 수집하였다. ‘앞서 해보기’란 ‘Steam’의 기능 중 하나로 정식 발매 이전에 게임을 후원함과 동시에 개발 중인 게임을 플레이하며 피드백

을 줄 수 있는 시스템이며 제품 페이지 하단의 게임 정보 표에 표시된다. 도전과제는 'Steam'에서 '게임에서 특정 목표를 달성했거나 플레이어 상호작용에 대해 보상하고 이를 권장하는 데 사용'하는 것이라고 정의하고 있다. 게임에 따라 없거나 천 개가 넘는 등 다양하며, 플레이어가 가장 많이 달성한 도전과제의 달성률 또한 이 도전과제의 달성률을 의미한다. 지원하는 언어는 대부분 기본적으로 영어를 포함하며 이에 추가적으로 다른 언어를 지원하는 경우도 있다. '스팀 어워드'는 2016년부터 'Steam'에서 진행된 연례 이벤트로, 사용자들이 게임에 직접 투표하여 상을 수여한다. DLC는 Downlodable Contents의 줄임말로 사운드트랙, 챕터, 모드 등 게임의 추가 컨텐츠로 제공되며 무료 또는 유료로 판매된다. 'Steam'의 리뷰는 영화 리뷰와 달리 점수로 평가할 수 없고 추천 여부만 설정할 수 있으며, 플레이어가 게임에 대해 하고 싶은 말을 추가로 적을 수 있다. 'Steam'에서는 정해진 기준에 따라 전체 또는 최근 30일 간의 리뷰 수에 대한 '추천' 비율에 따라 '압도적으로 긍정적'부터 '압도적으로 부정적'까지 9가지로 평가 등급을 나누고 제품 페이지에 보여준다. 이 등급은 리뷰가 10개 이상일 때부터 집계되며 리뷰가 10개 미만인 경우 리뷰 수를 표시하거나 등급을 보여주지 않는 등 집계하지 않는다. 싱글 혹은 멀티 여부는 '앞서 해보기'와 마찬가지로 제품 페이지 하단 게임 정보 표에 표시된다. 추천 및 비추천 리뷰 수는 'Steam' 제품 페이지에서 각 게임마다 작성된 '추천/비추천' 리뷰의 수를 의미하며 두 수를 합하면 총 리뷰 수가 된다. 'Steam'에 작성된 리뷰는 다른 사용자에 의해 유용함 여부와 '재미있음' 투표를 받을 수 있는데, 'Steam'의 리뷰 페이지는 기본적으로 가장 많은 '유용함' 투표를 받은 리뷰 순으로 정렬되어 나타난다.

3.2 데이터 전처리 및 대조실험 준비

게임 아이디는 데이터를 수집하기 위한 키로 사용되었을뿐 게임의 흥행 여부와는 상관이 없으므로 삭제하였고, 출시 일자는 출시 연도만 추출한 후 삭제하였다. 게임 가격은 'Steam' 상점 목록의 필터링 기준에 따라 5천 원 단위로 총 13등급으로 나누었다. 개발자 및 배급사는 데이터 수집 과정에서부터 'Steam' 내부 페이지가 없다면 6등급, 있다면 팔로워 수에 따라 다섯 등급으로 나누어 총 6가지로 등급화하였다. 장르는 결측치를 "allGames"로 대체하여 라벨링하였으며 총 11가지가 존재하였다. 태그는 전체 데이터셋에서 같은 태그가 3번 이상 등장한 경우만 남기고 나머지를 모두 "etc"로 대체한 후 라벨링하였고, 총 52가지가 존재하였다. 최근 평가 등급은 수집 결과 결측치 비율이 너무 높아 데이터셋에서 삭제하였다. 모든 평가 등급은 '압도적으로 긍정적'부터 '압도적으로 부정적' 사이에서 제대로 표기된 값만 남기고 나머지는 모두 "noData"로 대체한 후 라벨링하였다. 싱글/멀티 여부는 결측값의 경우 기본적으로 싱글 플레이가 가능하다는 전제로 "single"로 대체하였으며, 데이터 수집 과정에서 싱글 게임임이 확인된 경우 "single only"로 표기하여 결측값과 그렇지 않은 값이 구분된다. 싱글/멀티 여부는 결측치 처리 이후 'single only'(데이터 수집된), 'single'(결측치 처리됨), 'multi only', 'single and multi'로 총 4가지가 존재하였고 라벨링하였다. 최근 동시 플레이어 수와 최고 동시 플레이어 수의 결측치는 'steamcharts.com'에 별도 페이지가 존재하지 않는 것으로 간주하고 -1로 대체하였다. 도전과제 달성률은 도전과제가 없는 게임인 경우 확인할 수 없으므로 0으로 대체하였고, 도전과제가 있지만 결측치로 수집된 게임인 경우 -1로 대체하였다. 추천 리뷰 수와 비추천 리뷰 수는 두 수를 합하여 총 리뷰 수를 구하고, 추천 리뷰 수를 총 리뷰 수로 나누어 전체 리뷰에서 추천 리뷰의 비율을

구한 후 삭제하였다. 가장 많은 '유용함' 투표를 받은 리뷰는 텍스트 감성 분석 모델을 구현해 긍정성을 분석한 값을 사용하였다. 긍정적인 텍스트일수록 1에 가까운 값을 갖는다. 도전과제 개수, 지원 언어 수, '앞서 해보기' 여부, 스팀 어워드 수상 여부, DLC 유무는 그대로 사용하였다.

흥행 여부는 본 연구의 목표 변수이다. 'Steam' 인기 제품 페이지에 있었던 게임을 먼저 흥행을 나타내는 1로 표기하고, steam web API를 통해 얻은 게임 ID 중 흥행으로 표시되지 않은 제품에 모두 0으로 표기하였다. 이후 0으로 표기된 게임 중 모든 평가 긍정성 문구에 'positive'가 포함되고, 총 리뷰 수가 넘는 게임에 한하여 다시 1로 표기하였다. 이 기준은 'Steam'의 평가 등급 집계 기준을 참고했다. 총 샘플 828개 중 흥행 여부가 0인 게임은 377개, 1인 게임은 451개로 나뉘었다.

전처리를 마친 데이터셋을 게임 발매 전에 알 수 있는 정보만으로 구성된 실험군과 게임 발매 이후 알 수 있는 정보를 포함한 대조군으로 나누었다. 대조군은 흥행이 결정된 이후의 데이터가 다수 포함되고 이는 흥행 판단의 근거가 될 수 있으므로 모델을 구현했을 때 더 높은 성능을 보일 것으로 예상된다. 본 연구는 실험군으로 모델을 구현하였을 때 실제로 활용이 가능할지 확인하는 것을 목적으로 하며, 대조군 모델의 성능은 이상적인 목표치로서 사용된다. 데이터 전처리와 실험군, 대조군 분리에 대한 데이터 구성은 <표 2>에 나타난 것과 같으며, 각 속성과 목표 변수 사이의 상관 관계는 <그림 1>의 히트맵에 표현되어 있다. 색이 진할수록 절댓값이 크고, 상관 관계가 크다는 것을 의미하며 목표 변수와의 상관 관계 값은 가장 아랫줄에서 확인할 수 있다. <그림 1>에 나타난 속성 간 상관 관계 값은 실험군 데이터셋에서도 동일하므로 실험군 데이터셋에 대한 히트맵은 생략한다.

3.3 예측 모델 설계

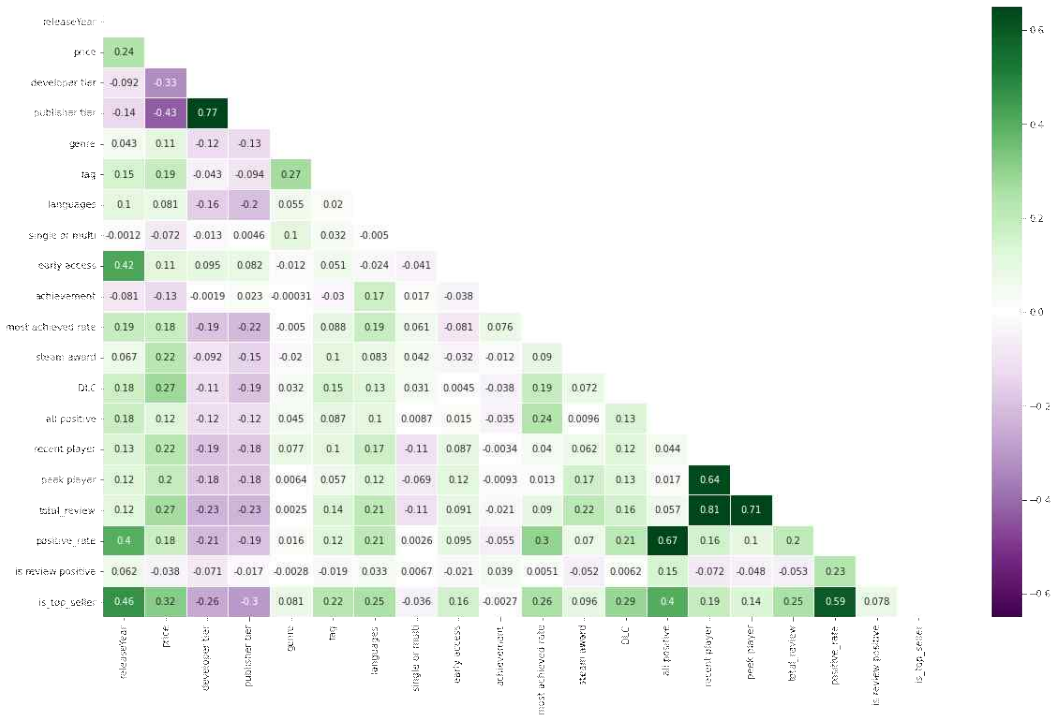
본 연구에서는 DNN, Random forest, Ada boost 총 세 가지 모델을 사용하였고 모두 10겹 교차검증을 하였다. 각각 실험군과 대조군 모델을 만들어 총 6개 모델을 학습시켰다. DNN 모델은 출력층을 포함해 총 6개의 층으로 이루어지며 이중 Dropout 층이 1개, BatchNormalization 층이 1개 포함된다. 케라스의 EarlyStopping, ModelCheckpoint 콜백을 이용하였으며 이 구조는 실험군과 대조군 모두 동일하다. Random forest 모델은 기본값으로만 생성하였고, Ada boost 모델은 최대 추정기 수(n_estimators)를 100으로 설정하였다. Random forest, Ada boost 모델 또한 실험군과 대조군에 같은 구조를 사용하였다.

4 연구 결과

실험 결과는 <표 3>과 같으며 모두 10겹 교차검증을 수행하여 10회의 평균 정확도 및 오차를 표기한 것이고, 표준편차는 10회의 정확도의 표준편차를 의미한다. 본 연구에 사용한 모든 모델에서 실험군보다 대조군의 평균 정확도가 더 높게 나온 것을 볼 수 있다. 대조군 데이터셋에 게임 발매 이후 정보가 다수 포함되었으므로 정확도가 더 높은 것은 예상한 바와 같다. 실험군 중에서는 Ada boost 모델이 가장 평균 정확도가 높고, 대조군에서는 Random forest 모델이 가장 평균 정확도가 높다. 실험군의 평균 정확도는 Ada boost 모델이 가장 높지만 표준편차는 Random forest 모델이 더 낮다.

<표 2> 전처리 후 실험군 및 대조군 구성

포함 데이터셋	속성	발매 전 정보
대조군	스팀 어워드 수상 여부(‘steam award’)	X
	모든 평가 긍정성(‘all positive’)	X
	최근 동시 플레이어 수(‘recent player’)	X
	최고 동시 플레이어 수(‘peek player’)	X
	도전과제 달성률(‘most achieved rate’)	X
	리뷰 텍스트 긍정성(‘is review positive’)	X
	총 리뷰 수(‘total_review’)	X
	추천 리뷰 비율(‘positive_rate’)	X
실험군, 대조군	가격(‘price’)	O
	개발자 등급(‘developer tier’)	O
	배급사 등급(‘publisher tier’)	O
	장르(‘genre’)	O
	‘앞서 해보기’ 여부(‘early access’)	O
	도전과제 개수(‘achievement’)	O
	태그(‘tag’)	O
	지원 언어 수(‘languages’)	O
	DLC 유무(DLC)	O
	싱글/멀티 여부(‘single or multi’)	O
	출시연도(‘releaseYear’)	O
	흥행 여부(‘is_top_seller’)	X, 목표 변수



<그림 1> 모든 속성 상관 관계 히트맵

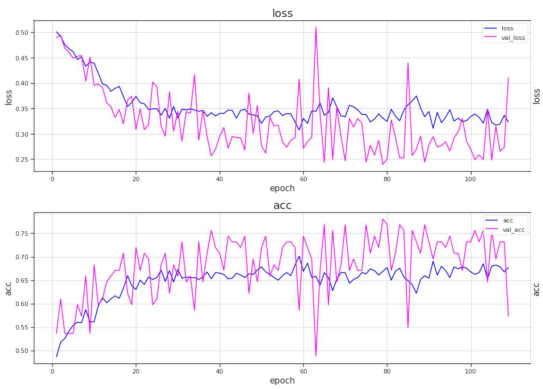
<표 3> 실험 결과

모델	DNN		Random forest		Ada boost	
데이터셋	정확도	오차	정확도	표준편차	정확도	표준편차
실험군	63.19 %	0.3719	75.54 %	0.0364	76.12 %	0.0707
대조군	92.59 %	0.1867	96.38 %	0.0263	96.07 %	0.0194

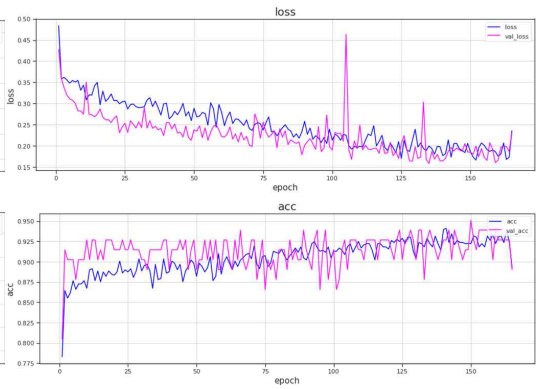
<그림 2>와 <그림 3>은 DNN 모델의 10겹 교차검증 중 마지막 10회차 모델의 실험군과 대조군 학습 과정에서의 오차와 정확도를 기록한 것이다. 실험군은 epoch 109에서 종료되었고, 대조군은 epoch 165에서 종료되었다. 두 모델 모두 오차는 하향 곡선을, 정확도는 상향 곡선을 그리고 있으나 불안정한 학습 경향을 보이며, 대조군이 더 많이 학습하였음을 알 수 있다.

<표 4>는 연구에 사용한 모델 중 Random forest 모델과 Ada boost 모델의 네 가지 분류 성능 평가 지표를 나타낸 것이다. Accuracy는 정확도로, 전체 데이터에서 모델이 바르게 분류한 데이터의 비율을 나타낸다. Precision은 정밀도이고 모델이 True라고 분류한 것 중 실제로 True인 것의 비율을 나타낸다. recall은 재현도, 실제 True인 것 중 모델이 True라고 분류한 것의 비율이다. F1 score는 정밀도와 재현도의 조화평균이다. 두 모델 모두 대조군에 있어서는 유사한 성능을 보였으나 실험군에서는 Random forest 모델이 더 우수한 성능을 보이고 있다. 10겹 교차검증의 평균 성능은 Ada boost 모델이 근소하게 우수하였으나 표준 편차가 비교적 높았던 점으로 보아, Ada boost 모델은 여러 번의 테스트에서 성능이 고르게 나오지 못한 것으로 판단된다.

<그림 4>는 성능이 가장 안정적이고 우수했던 Random forest 모델의 변수 중요도 그래프이다. 실험군에서는 'tag'의 중요도가 0.1860으로 가장 높았고, 그 뒤를 'price'와 'languages'가 각각 0.1435와 0.1418로 두 번째와 세 번째 중요 속성으로 자리했으며 'genre'가 중요도 0.0119로 가장 낮은 중요도를 보였다. 대조군에서는 'single or multi'가 중요도 0.3374로 첫 번째, 'total_review'가 0.1535, 'publisher tier'가 0.1306으로 세 번째에 위치했다. 가장 중요도가 낮은 속성은 'most achieved rate'이며 중요도는 0.001로 사실상 관계가 없다고 볼 수 있다.



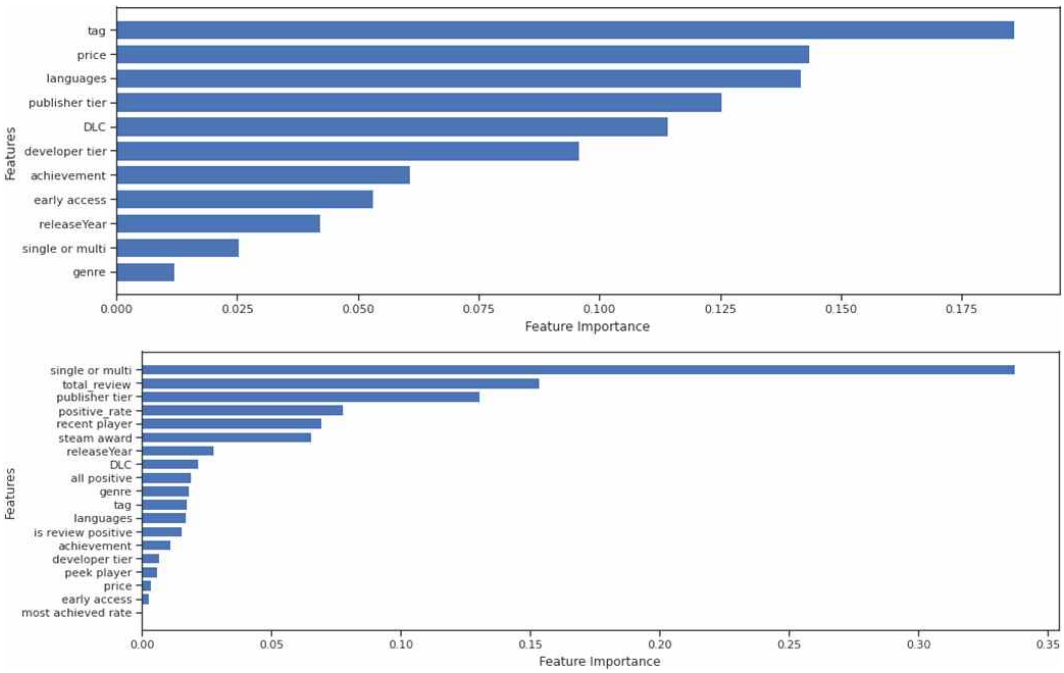
<그림 2> 10회차 DNN 실험군 모델 오차와 정확도 그래프(상: 오차, 하: 정확도)



<그림 3> 10회차 DNN 대조군 모델 오차와 정확도 그래프(상: 오차, 하: 정확도)

<표 4> Random forest와 Ada boost 실험 결과

모델	Random forest		Ada boost	
데이터셋	실험군	대조군	실험군	대조군
Accuracy	0.7590	0.9819	0.7108	0.9940
Precision	0.7500	0.9759	0.7125	0.9880
recall	0.7683	0.9878	0.6951	1.0000
F1 score	0.7590	0.9818	0.7037	0.9939



<그림 4> Random forest 모델 변수 중요도 그래프(상: 실험군, 하: 대조군)

5 결론 및 논의

앞서 확인한 연구 결과에서 실험군의 정확도가 아주 높지는 않았지만 Random forest와 Ada boost 모델에서 약 75% 이상의 정확도가 나오므로 게임 제작사의 입장에서, 기획 이후 개발을 시작하기 전 흥행을 선제적으로 예측하여 투자 결정을 내리는 데에 도움을 줄 수 있을 것으로 보인다. 이 중 특히 Random forest 모델이 여러 번의 테스트에서도 안정적인 성능을 보이고 있으므로 본 연구에 사용한 세 가지 모델 중 실제 활용에 가장 적절할 것으로 판단된다. 모델의 예측 정확도는 향후 연구에서 더 많은 데이터를 확보하고, 예측에 유효한 새로운 변수를 밝혀낸다면 개선할 수 있을 것이다. 데이터를 일자별로 수집할 수 있다면 발매 초기 게임에 대한 흥행 예측도 수행할 수 있을 것으로 기대된다.

본 연구는 지금까지 영화를 주류로 연구되어 왔던 흥행 예측 분야에서 게임에 대한 흥행 예측을 시도하였고, 게임에 대한 기본적인 정보를 기반으로 예측을 수행함으로써 현실에 적용할 수 있는 가능성을 보았다. 본 연구에서는 정량적 데이터를 중심으로 사용하여 정성적 데이터를 반영하지 못하였고, 2022년 7월 steam web API 기준 'Steam'

에 게임 ID를 부여받은 게임이 6만 개 이상 존재하는 데 비해 샘플은 800여 개밖에 사용하지 못했다는 점에서 한계가 있고, 시대의 흐름에 따른 유행과 트렌드 변화도 반영하기 어렵다는 문제가 있으나 추가 연구를 통해 정성적 요소를 반영하도록 한다면 극복할 수 있을 것으로 보인다.

참고문헌

성민 · 김동성 · 김종우(2021, 6월). 온라인 리뷰에 대한 특성 기반 감성 분석을 활용한 영화 흥행 조기 예측. 대한산업공학회 춘계공동학술대회 논문집. 제주: 국제컨벤션센터.

송정아 · 최근호 · 김건우(2018, 12월). 영화 흥행에 영향을 미치는 새로운 변수 개발과 이를 이용한 머신러닝 기반의 주간 박스오피스 예측. *J Intell Inform Syst*, 24(4), 67-83.

이도연 · 장병희(2020). 딥러닝을 이용한 음악흥행 예측모델 개발 연구. 한국콘텐츠학회 논문지, 20(8), 10-18.

강지훈 · 박찬희 · 도형록 · 김성범(2014, 5월). 데이터마이닝 기법을 활용한 영화 흥행 실적 예측 기법. 대한산업공학회 춘계공동학술대회 논문집, 142-154.

이동석(2022). "양방향 국소평균 K 최근접 이웃 방법론을 이용한 영화 관객수 예측에 관한 연구." 국내석사학위논문, 한양대학교 대학원.

장재영(2017). "소셜 빅데이터 분석과 기계학습을 이용한 영화흥행예측 기법의 실험적 평가", 한국인터넷방송통신학회 논문지, 17(3), 167-173.

권신혜 · 박경우 · 장병희(2017). "기계학습 기반의 영화흥행예측 방법 비교: 인공신경망과 의사결정나무를 중심으로", 예술인문사회융합멀티미디어논문지, 7(4), 593-601.

송정아 · 최근호 · 김건우(2018). "기계학습을 이용한 주간 박스오피스 예측", 한국지능정보시스템학회 학술대회논문집, 58-71.

임준엽 · 황병연(2014). "트위터를 이용한 기계학습 기반의 영화흥행 예측", 정보처리학회논문지. 소프트웨어 및 데이터 공학 3 (7), 263-270.

Do, Tiffany D., Seong Ioi Wang, Dylan S. Yu, Matthew G. McMillian and Ryan P. McMahan. “Using Machine Learning to Predict Game Outcomes Based on Player-Champion Experience in League of Legends.” *The 16th International Conference on the Foundations of Digital Games (FDG) 2021* (2021): n. pag.