ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2018 December: 24(4): 67~83 http://dx.doi.org/10.13088/jiis.2018.24.4.067

영화 흥행에 영향을 미치는 새로운 변수 개발과 이를 이용한 머신러닝 기반의 주간 박스오피스 예측

송정아
한밭대학교 빅데이터비즈니스학과
(*doruddt@naver.com*)

최근호
한밭대학교 경영회계학과
(*keunho@hanbat.ac.kr*)

김건우
한밭대학교 경영회계학과
(*gkim@hanbat.ac.kr*)

2013년 누적인원 2억명을 돌파한 한국의 영화 산업은 매년 괄목할만한 성장을 거듭하여 왔다. 하지만 2015년을 기점으로 한국의 영화 산업은 저성장 시대로 접어들어, 2016년에는 마이너스 성장을 기록하였다. 영화산업을 이루고 있는 각 이해당사자(제작사, 배급사, 극장주 등)들은 개봉 영화에 대한 시장의 반응을 예측하고 탄력적으로 대응하는 전략을 수립해 시장의 이익을 극대화하려고 한다. 이에 본 연구는 개봉 후 역동적으로 변화하는 관람객 수요 변화에 대한 탄력적인 대응을 할 수 있도록 주차 별 관람객 수를 예측하는데 목적을 두고 있다. 분석을 위해 선행연구에서 사용되었던 요인 뿐 아니라 개봉 후 역동적으로 변화하는 영화의 흥행순위, 매출점유율, 흥행순위 변동 폭 등 선행연구에서 사용되지 않았던 데이터들을 새로운 요인으로 사용하고 Naive Bays, Random Forest, Support Vector Machine, Multi Layer Perception등의 기계학습 기법을 이용하여 개봉 일 후, 개봉 1주 후, 개봉 2주 후 시점에는 차주 누적 관람객 수를 예측하고 개봉 3주 후 시점에는 총 관람객 수를 예측하였다. 새롭게 제시한 변수들을 포함한 모델과 포함하지 않은 모델을 구성하여 실험하였고 비교를 위해 매 예측시점마다 동일한 예측 요인을 사용하여 총 관람객 수도 예측해보았다. 분석결과 동일한 시점에 총 관람객 수를 예측했을 경우 보다 차주 누적 관람객 수를 예측하는 것이 더 높은 정확도를 보였으며. 새롭게 제시한 변수들을 포함한 모델의 정확도가 대부분 높았으며 통계적으로 그 차이가 유의함으로써 정확도에 기여했음을 확인할 수 있었다. 기계학습 기법 중에는 Random Forest가 가장 높은 정확도를 보였다.

주제어 : 영화 흥행 예측, 영화 관람객 수 예측, 박스오피스 예측, 기계학습

논문접수일 : 2018년 8월 2일 논문수정일 : 2018년 12월 1일 게재확정일 : 2018년 12월 16일
원고유형 : 일반논문 교신저자 : 김건우

1. 서론

한국영화 상영 시장의 총 관람객 수는 2013년 2억명을 돌파한 뒤 3여 년간 지속적인 성장을 이어오다 2016년에는 소폭 감소한 후 2017년 관람객 수 2억 1,987만 명으로 역대 최다를 기록하였다. 하지만 매출액 증가 수준은 2016년에 비해 2017년 1조 7,566억 원으로 전년대비 0.8% 증가하는 수준에 그쳤다(Korean Film Council, 2017). <Table 1>을 보면 전국 극장 수와 스크린 수는 지속적으로 증가하고 있는 반면 총 관람객 수와 극장 매출액의 증가 폭은 미미하다. 이렇듯 한국 영화 시장은 2015년을 기점으로 저성장 시대에 접어들었고 앞으로도 큰 변화 없이 저성장에 머무를 것으로 전망된다(Korean Film Council, 2017). 그러나 인터넷을 통해 방송 프로그램, 영화, 교육 등 각종 미디어 콘텐츠를 제공하는 서비스인 OTT (Over The TOP) 의 등장으로 인터

〈Table 1〉 2012-2017 Key statistical indicators of Korean film industry (Korean Film Council, 2017)

Index	2012	2013	2014	2015	2016	2017
Total audience (million)	195 (22.0%)	214 (9.5%)	215 (0.8%)	217 (1.0%)	217 (-0.1%)	220 (1.3%)
Sales (Billion)	1,455 (17.8%)	1,551 (6.6%)	1,664 (7.3%)	1,715 (3.1%)	1,743 (1.6%)	1,756 (0.8%)
Number of Theaters	314	333	356	388	427	452
Number of Screens	2,081	2,184	2,281	2,424	2,575	2,766
Number of movie going times per individual per one year	3.83	4.17	4.19	4.22	4.20	4.25
Digital online market (Billion)	215 (26.3%)	267 (24.0%)	297 (11.0%)	334 (12.7%)	412 (23.2%)	436 (5.7%)

넷 VOD 시장과 IPTV 및 디지털 케이블 TV 등의 온라인 디지털 콘텐츠 시장은 지속적으로 성장하고 있다.

영화는 경험재적 성격을 가진 문화상품으로 영화 개봉 전 까지는 흥행 여부나 초기 관람객 수를 정확히 예측하는 것이 어려우며, 개봉 후에도 다양한 요인들에 의해 관람객 수는 역동적으로 변화한다. 따라서 제작사나 배급사는 영화 개봉 시에 대규모 멀티플렉스(Multiplex)를 중심으로 많은 스크린을 확보해 관람객 수를 늘려 매출실적을 높이고자 한다. 하지만 극장주들은 영화 예매를 위한 상영 시간표를 길게는 약 일주일, 짧게는 3~4일 정도만 공개하고 있으며 영화 개봉 후에 관람객들의 평가와 흥행 실적을 바탕으로 차주의 상영 횟수나 상영 여부 등을 결정한다. 따라서 영화산업을 이루고 있는 각 이해당사자(제작사, 배급사, 극장주 등)들은 개봉 영화에 대한 역동적인 시장의 반응을 예측하고 탄력적으로 대응하는 전략을 통해 시장의 이익을 극대화하려고 한다. 즉, 배급사와 제작사는 높은 예측 정확도를 바탕으로 차주 스크린 수와 상영 횟수들을 판단하여 상영 연장 또는 종료를 통해 매출을 극대화하거나 손실을 최소화 함과 동시에 디지털 온라인 판매로 판로를 바꿔 매출을 올리고자 한다. 반면 극장주들은 흥행 예측에 기반한 데이터를 참고하여 기민한 스크린 교체를 통해 손실을 최소화 하고 매출을 극대화 하고자 한다. 따라서 영화 흥행에 대한 예측은 이해당사자들에게 수익과 직접적으로 연결된, 중요한 의사결정을 내리기 위한 전략적 수단이 되어 가고 있다.

이러한 중요성에 기인하여 영화 흥행을 예측하기 위한 많은 연구들이 수행 되어왔다. 초기에는 영화 흥행에 영향을 미치는 여러 요인들을 밝히고자 노력해 왔으며(Litman, 1983; Yoo, 2002; Kim, 2009; Kang, 2017) 최근의 연구들은 새로운 요인들을 규명하는 대신 과거 선행 연구에서 사용되었던 변수들에 다양한 예측 분석기법을 적

용하여 흥행 예측의 정확도를 높이는데 집중하고 예측 모델에서 도출된 변수들의 영향력을 설명하고자 하는 시도들이 많이 이루어지고 있다 (Song and Han, 2013; Lim and Hwang, 2014; Jeon and Son, 2016; Chang, 2017; Rhee , T. G., and F. Zulkernine, 2016; Quader et al., 2017).

그러나, 대부분의 기존 연구들은 영화 흥행을 예측하기 위해 설정한 목표 변수로 영화 개봉시점에서 종영시점까지 전체 기간 동안 발생한 총 누적 관람객 수 또는 총 누적 매출액을 사용하고 있는데 이는 영화 개봉 시부터 종영 시까지 역동적으로 변화하는 시장 수요를 선제적으로 예측하고 탄력적으로 대응하기에는 한계점이 존재한다. 이는 영화 흥행 예측의 정확도를 떨어뜨리고 나아가 실제 영화 산업의 현실을 제대로 반영하지 못해 영화 흥행 예측 모델을 사용하려는 사용자들이 느끼는 모델의 효용성을 떨어뜨리게 된다. 또한 흥행 요인 연구들에서는 동일한 요인이 연구마다 다른 결과를 보여 주는 사례가 많아 변수와 영화 흥행 사이의 요인 규명의 복잡도를 증가시키고 있다. 이런 혼재된 결과로 인해 신뢰할 만한 영화 흥행 요인들을 명확히 밝히기는 쉽지 않다.

따라서, 본 연구에서는 기존 연구들의 한계점을 극복하고 올바른 시점에 보다 정확한 영화 흥행 예측 결과를 이해당사자에게 제공하기 위해 영화 개봉 후 종영까지의 전체 기간이 아닌 주차별 누적 관람객 수를 예측하고자 한다. 이를 위해 선행연구에서 사용되었던 요인 뿐 아니라 선행연구에서 사용되지 않았던 영화의 흥행순위, 매출 점유율, 순위 변동 폭 등 개봉 후 역동적으로 변화하는 여러 변수들을 포괄적으로 사용하여 주차별 누적 관람객 수를 예측하였다. 본 연구를 위한 예측 방법으로 Naive Bayes, Random Forest, Support Vector Machine(SVM)은 10-fold cross-validation을 사용하였고, Multi Layer Perception(MLP)는 4-fold cross-validation을 사용하였다. 또한 새로운 변수들의 예측요인의 가능성을 알아보기 위해 새롭게 제시한 변수들을 포함한 모델과 포함하지 않은 모델로 구성하였다. 영화 개봉 후에도 변하지 않는 스타성, 장르, 등급, 배급사, 국가 등 제작과 배급 단계의 요인들과 더불어 네티즌 평점, 흥행 순위, 매출 점유율 등 개봉 후 변화하는 요인들을 이용하여 개봉일 후, 개봉 1주 후, 개봉 2주 후 시점에는 차주 누적 관객수를 예측하고 개봉 3주 후 시점에는 총 관람객 수를 예측하였으며 비교를 위해 각 시점마다 동일한 예측 요인을 사용하여 총 관람객 수도 같이 예측하였다.

실험결과 예측 시점이 뒤로 갈수록 예측 정확도가 점점 높아지며 동일한 시점에 총 관람객 수를 예측했을 경우 보다 차주 누적 관람객 수를 예측하는 것이 더 높은 예측 정확도를 보였고 기계학습 기법 중에서는 Random Forest가 73.9% ~ 88.6%로 가장 높은 예측 정확도를 보였으며 새롭게 제시한 변수를 포함한 모델이 그렇지 않은 모델보다 높은 정확도를 보였다.

마지막으로, 본 연구의 기여점은 다음과 같다. 첫째, 기존의 흥행 요인들뿐만 아니라 영화 개봉 후 관람객들의 평가와 흥행 실적을 바탕으로 한 네티즌 평점, 흥행 순위, 매출 점유율 등 그 동안 연구들에서 다루지 않았던 흥행에 영향을 미칠 만한 다른 여러 요인들을 포괄적으로 고려하여 보다 정확한 영화 흥행 예측 모델을 제안하였고 새로운 변수들의 흥행요인 가능성을 확인하였다. 둘째, 이러한 여러 요인들을 고려하여 각 시점 별 다음주 누적 관객수를 미리 예측할 수 있는 모델을 제시함으로 영화산업 이해관계자들이

실제 현장에서 개봉 후 역동적으로 변화는 관람객들의 반응에 탄력적으로 대응할 수 있는 빠르고 정확한 의사결정을 내리는데 도움을 줄 수 있다. 이는 모델이 보다 현실적인 상황에서 많은 사람들에게 널리 사용될 수 있는 가능성을 높여 모델이 가진 효용성을 극대화해 줄 수 있다.

2. 관련연구

일반적으로 영화 흥행과 관련된 연구는 영화 흥행에 영향력을 미치는 흥행 요인들의 선택에 관한 연구(Litman, 1983; Yoo, 2002; Kim, 2009; Kang, 2017)와 이들 흥행 요인들로부터 영화 흥행 예측 모델을 연구하는 두 가지 주제로 분류된다(Song and Han, 2013; Lim and Hwang, 2014; Jeon and Son, 2016; Chang, 2017; Rhee , T. G., and F. Zulkernine, 2016; Quader et al., 2017). 영화산업의 규모가 커져가면서 영화 흥행을 예측하는 다양한 연구들이 국내외로 진행되어왔다. 연구 초기에는 영화 흥행 요인을 규명하는데 집중되어 있었다. Litman(1983)의 연구는 창작 영역의 장르, 관람 등급, 스타 캐스팅 유무, 제작비와 마케팅 영역의 아카데미상 수상 여부, 평론 등, 배급 영역의 배급사의 유형, 개봉시기들을 변수로 활용하여 제작비, SF / Horror, 관람등급 등이 영향력 있다는 결론을 도출 하였다. 유현석(2002)은 그동안 상식적으로 흥행에 영향을 준다고 추정되었던 출연배우, 감독, 제작사, 장르, 등급 등 제작관련 변수들의 영화 흥행 요인들의 영향력을 연구를 통해 실증적으로 입증하였다. 김병선(2009)은 영화 개봉방식과 상영기간에 따라 유형을 나누고 상호 비교를 통해 영화 유형별 흥행에 미치는 요인은 서로 다르다는 결과를 도출했고, 김소영 등(2010)의 연구에서는 영화 유형을 상업영화와 예술영화로 분류하고 유형에 따른 예측 요인 비교를 통해 스크린 수, 관람객 평가, 장르는 상업영화와 예술영화모두 유의하나 다른 변수들은 서로 다르다는 연구결과를 발표하였다. 권선주(2014)는 전문가 평가가 영화흥행 성과에 미치는 영향력을 연구를 통해 전문가 평점은 전체영화를 대상으로는 유의미했으나 상업영화와 예술영화로 나누었을 때에는 예술영화에서만 유의미하였다. 네티즌 평가의 빈도는 모두 유의미하나 평점은 예술 영화만 유의미하며 내생성 제거 후 네티즌 평가의 빈도는 상업영화에서 개봉 2주까지만 유의미하고 평점은 상업영화에서만 유의미하다는 결과를 발표하였다. 강선주(2017)는 선행 연구들의 흥행 요인들을 2016년 개봉한 상업영화를 중심으로 분석해 개봉 스크린 수, 장르, 스타 캐스팅, 배급사의 영향력이 유의미하나 제작비 규모는 비례한다고 볼 수 없으며 스타성도 필요요소이긴 하지만 필수요소는 아니며 영화산업은 경제적, 사회적인 것들이 반영되기 때문에 흥행에 영향을 줄 수 있는 요소들을 한정 짓고 영향력을 판단하기는 어렵다는 결론을 도출하였다. 박승현과 송현주(2012)는 2010년 개봉한 영화 중 68편을 대상으로 온라인 구전이 흥행에 미치는 영향을 개봉 7주차까지의 주별 흥행성과와 전체 흥행성과로 나누어 연구하였다. 그 결과 온라인 구전의 빈도는 상영이 끝나는 시점까지 지속적인 영향을 미쳤으나 평점은 개봉 초기에만 유의미한 영향을 미친다는 결론을 도출하였다. 이 연구는 구체적이고 역동적인 설명 모형을 통해 주별 흥행 성과에 영향을 미치는 흥행 요인을 찾아내고 특히 온라인 구전의 영향력을 규명하는데 큰 의의를 두고 있다.

최근 들어서는 선행연구에서 도출된 변수들을

예측 변수로 활용하여 총 관람객 수와 총 매출액의 예측 정확도를 향상시키려는 연구들이 시도되고 있는데, Rhee, T. G., & Zulkernine, F. (2016)의 연구에서는 다양한 웹사이트에서 데이터를 수집해 예측 변수를 생성하고 흥행, 실패라는 두 개의 클래스로 분류한 후 Back-propagation neural network 기법을 적용하여 91%의 분류 정확도를 보여줬다. Quader, N., et al., (2017)은 Support Vector Machine 과 Multi Layer Perception 을 이용하여 개봉 전 요소와 개봉 후 요소를 기반으로 영화관 총 매출액을 5개의 클래스로 분류 예측하였는데 전체적으로 개봉 후 요소가 포함된 경우에 예측 정확도가 향상되었으며 Support Vector Machine 보다는 Multi Layer Perception가 89.27%로 가장 높은 예측력을 보이며, 제작비와, IMDb 평가자 수, 스크린 수가 중요하다는 결론을 도출하였다. 국내에서는 송종우와 한수지(2013)가 2008년부터 2011년까지 개봉된 영화 중 매출 규모가 5억이상인 한국 영화 206편을 분석대상으로 하여, Linear regression analysis, Random Forest, Grdient Boosting 기법을 적용하여 예측한 결과 Gradient Boosting이 예측률이 가장 좋으나 차이가 작기 때문에 해석이 쉬운 Linear regression analysis가 적절하며 장르, 감독과 배우의 스타성이 흥행 요인으로 영향력이 있다는 결론을 도출하였다. 임준엽과 황병연(2014)은 2013년 4월부터 10월까지 개봉된 영화 중 무작위 60편을 분석대상으로 선택 후 총 관람객 수를 5개의 범주로 분류하고 오프라인 요소와 온라인 요소로 변수를 구분 한 후 Naive bayes 기법을 사용하여 분석하였다. 그 결과, 개봉 일에는 78.4%, 개봉 1주일 후에는 95%의 적합도를 보였으며 온라인 요소(포털 평점, 트위터 언급 수)를 포함하여 예측한 결과가 전체적으로 더 높다는 결론을 도출하였다. 전성현과 손영숙(2016)은 2012년부터 2015년까지 관람객 수 50만 이상인 국내 영화 276편을 대상으로 개봉 전, 개봉 일, 개봉 1주일, 개봉 2주일 각 시점에서 Decision Tree, Multi Layer Perception, Multinominal Logistic Regression, 그리고 SVM을 사용하여 총 관람객 수를 예측하였다. 모든 자료를 대상으로 적합 시켰을 때 모든 시점에 Neural network 모형의 적합도는 거의 100%의 정확도를 보였다. 장재영(2017)은 예측 변수들을 정적 데이터와 동적 데이터로 구분하고 Naive bayes와 Neural network 기법을 적용하여 총 관람객 수를 5개의 클래스로 분류하여 예측한 결과 주요한 정적 데이터와 동적 데이터를 모두 포함한 Neural network 모형이 68%로 가장 높은 예측 정확도를 보였다.

본 연구에서는 선행 연구에서 사용된 기계학습 기법 중 Naive Bays, Random Forest, Support Vector Machine, Multi Layer Perception을 이용하여 주차별 누적 관람객 수 예측을 수행하였다.

3. 데이터 설명

3.1 데이터 수집

흥행 예측을 위해 2015년부터 2017년까지 3년 동안 개봉한 영화를 분석 데이터로 사용하였다. 상세 데이터는 영화진흥위원회에서 운영하는 영화입장권 통합전산망 웹사이트에서 제공하는 API를 이용하여 영화의 상세정보와 흥행 실적, 영화인 정보를 일자별로 수집하였다. 전문가 평점과 네티즌 개봉 전/후의 평점 정보는 국내 최대 포털 사이트인 네이버의 영화 섹션에서 일자별로 수집하였다.

3.2 데이터 전처리

본 논문은 2015년부터 2017년까지 3년동안 개봉한 영화를 분석 대상으로 하였다. 영화진흥위원회의 핵심상업영화군 기준인 '최대 개봉관 수 300개관 이상이거나 순 제작비 30억 원 이상'으로 대상을 선정하려고 했으나 제작사들이 제작비를 공개를 하지 않기 때문에 정확한 대상을 선정할 수가 없어 개봉스크린이 300개 이상인 영화를 대상으로 하였다. <Table 2>는 개봉스크린이 300개 이상인 영화를 대상으로 전체 관람객 수 기준으로 영화상영기간을 주단위로 나타낸 표이며 주차별 예측을 위해 최소 3주차까지 실적이 있는 영화를 대상으로 하였다. 최종적으로 본 연구는 2015년부터 2017년까지 3년동안 개봉한 영화 중 개봉 스크린이 300개이상이고 3주이상 상영한 영화 211편을 분석 대상으로 선정하였다.

네이버에서 211편을 대상으로 일자별로 수집된 데이터를 전문가 평점과 평가자 수, 네티즌의 개봉 전 평점과 평가자 수, 개봉 후 주차별 평점과 평가자 수로 구성하였다.

개봉일을 기준으로 변하지 않는 영화의 속성정보, 전문가 평점, 개봉 전/후 네티즌 평점정보, 개봉 후 변화하는 다양한 흥행실적 데이터 등을 주차별로 생성하여 데이터 세트를 구성하였다.

3.3 데이터 설명

본 연구에서는 개봉 일 후, 개봉 1주 후, 개봉 2주 후에는 다음주 누적 관람객 수를 예측하고 개봉 3주 후에는 총 관람객 수를 예측하였다. 요인들은 영화 개봉을 기점으로 변하는 요인과 변하지않는 요인으로 나뉜다. 변하지 않는 요인으로는 영화의 제작 단계와 배급 단계에서 알 수 있는 요인들, 전문가 평점, 그리고 개봉 전 네티즌 평점이 있다.

제작 단계와 배급 단계의 요인들을 살펴보면 감독 스타성은 2010년도 이후부터 분석대상 작품 전까지 감독을 맡은 작품의 평균 총관람객수를 범주화 하였다. 배우스타성도 감독과 마찬가지로 2010년도 이후부터 분석대상 작품 전까지 출연한 작품의 평균 총관람객수를 범주화 하였다. 배우의 경우 기준이 모호하여 영화진흥위원회에서 수집된 데이터 중 처음에 나오는 한 명으로 제한하였다.

배급사는 메이저 배급사와 기타 배급사로 분류하였으며 메이저 배급사의 경우 국내 4개 외에 해외 배급사 3개로 총 7개로 정의하였다. 해외 배급사들이 직접 배급하는 영화들이 늘어나고 있으며 해외 배급사들이 국내 영화를 배급하는 경우도 늘어나고 있다. 그 중 메이저 배급사는 배급사가 자사 영화관을 운영하고 있는 수직

〈Table 2〉 Number of movies by class (Total Audience)

CLASS Total Audience (million)	A (~ 0.5)	B (0.5 ~ 1)	C (1 ~ 3)	D (3 ~)
3 WEEKS	19	29	35	1
OVER 4 WEEKS	2	11	54	60
TOTAL	21	40	89	61

결합 배급사인 씨제이이앤엠(주), 롯데쇼핑㈜롯데엔터테인먼트와 자사영화관이 없는 비수직결합 배급사인 (주)쇼박스, (주)넥스트엔터테인먼트월드(NEW), 해외영화를 직접배급하는 유니버설픽쳐스인터내셔널, 월트디즈니, 워너브러더스로 재분류 하였다(Choi, 2017).

개봉 월은 개봉일자에서 개봉 월을 추출하였다. 우리나라는 일반적으로 크리스마스와 명절 연휴, 그리고 여름방학이 포함된 12월, 1월, 7월, 8월을 성수기로 정의한다(Kang, 2017). 본 연구에서는 2015년도~ 2017년도의 설날과 추석을 확인하여 12월, 1월, 7월, 8월외 명절이 있는 달을 성수기로 정의하였다.

제작 국가는 대부분의 흥행 영화들이 한국, 미국 영화이고 그 외의 국가들은 거의 없기 때문에 한국, 미국, 그 외 국가로 분류하였다. 상영 등급의 경우 한국영화연감에 기초하여 전체 관람가, 12세 이상 관람가, 15세 이상 관람가, 청소년관람불가 네 가지로 분류하였다. 장르의 경우 영화진흥위원에서 제공되는 세분화된 데이터를 사극, 액션/범죄/스릴러, 드라마, SF/어드벤처/판타지, 전쟁, 기타로 총 7개로 분류하였다(Kang, 2017). 그 외 전문가 평점과 개봉 전 네티즌 평점은 전문가 평가 수와 평점들의 평균으로 표현하였다.

영화가 개봉한 후에는 다양한 흥행 지표들이 역동적으로 변하기 시작하는데 대부분의 선행연구들에서는 예측 시점의 스크린 수와 관람객 수만을 활용하였다. 본 연구에서는 예측 시점의 스크린 수와 관람객 수 이외에 영화진흥위원회에서 제공하는 데이터 중 다른 영화와의 경쟁요소라고 볼 수 있는 매출 점유율, 순위, 순위변경폭 등을 예측요인으로 사용하였다. 또 실시간으로 변화하는 네티즌의 평점 정보 역시 예측 요인으로 사용하였다.

예측 시점은 개봉일 후, 개봉 1주 후, 개봉 2주 후, 개봉 3주 후로 총 네 번의 예측 실험을 하기 때문에 각 예측 시점마다 매출 점유율, 순위 등의 요인들을 새로 생성하였으며 전 주 예측에서 사용되었던 요인들을 모두 포함하여 실험하였다. 각 주차는 월요일부터 일요일까지를 기준으로 하였다.

평균매출점유율은 개봉일의 경우 당일 매출 점유율을 사용하였고 주 단위 예측에서는 일요일을 마감으로 해당데이터의 평균으로 표현하였다. 관람객 수와 스크린 수는 예측시점까지의 누적값으로 표현하였으며 순위 역시 해당 주의 평균 순위로 표현하였다. 개봉일의 경우 순위증감 여부와 순위변경값은 확인할 수 없기 때문에 사용하지 않고 주차별 예측에서는 지난 주 실적과 비교하여 증가, 동일, 감소의 세가지로 순위 증감을 표현하였고 순위변경값은 지난 주와의 차이값을 이용하였다.

개봉 후의 네티즌 평가자 수는 개봉일부터 예측시점까지의 누적 평가자 수이고 평점은 평균 평점으로 표현하였다. 최종적으로 예측 변수를 정리한 내용은 <Table 3>과 같다.

예측 대상인 주차별 누적 관람객 수는 수치형으로 제공되기 때문에 수집된 영화의 주차별 누적 관람객 수의 데이터의 사분위수를 구해 범주화하여 등급을 나누었다. 주차별 관람객 수는 누적값을 이용하였으며 예측 시점마다 예측 데이터가 변하기 때문에 주차별 등급의 데이터 범주는 서로 다르다. 주차별 관람객 수는 <Table 4 > 처럼 구성하여 실험에 사용하였다.

주차별 관람객 수 예측률과 비교하기 위한 또 다른 예측 대상인 총 관람객 수 역시 수치형으로 제공되기 때문에 여러등급으로 나누었다. 주차

〈Table 3〉 Definition of variable

Variable Type	Variable Name	Variable Description	Release	After release t weeks (t=1,2,3)
Static Variable	DISTCD	Distributor	O	O
	OPENMM	Release month	O	O
	PEAKYN	Peak season	O	O
	NATION	Nation(Korea, USA, ETC)	O	O
	GENRECD	Genre	O	O
	WATCHGROUP	Film rating	O	O
	D_STAR	Director star power	O	O
	A_STAR	Actor Star Power	O	O
	SPECIAL_CNT	Expert raters	O	O
	SPECIAL_GRADE	Expert rating	O	O
	NET_BF_CNT	Netizen Raters before release	O	O
	NET_BF_GRADE	Netizen Rating before release	O	O
Dynamic Variable (Predict point)	SALESSHARE	Average revenue share	O	O
	AUDICNT	Number of audiences after release	O	O
	SCRNCNT	Number of screens after release	O	O
	SHHOWCNT	Number of shows after release	O	O
	RANK	Rank	O	O
	RANKID	Increase or decrease in ranking		O
	RANKINTEN	Rank change value		O
	NET_AF_CNT	Netizen Raters after release	O	O
	NET_AF_GRADE	Netizen Rating after release	O	O

〈Table 4〉 Definition of target variable (Weekly Audience)

CLASS	A	B	C	D
Number of audiences after release 1 week	~ 747,921.8	747,921.8 ~ 1,375,140	1,375,140 ~ 2,471,757	2,471,757 ~
Number of audiences after release 2 weeks	~ 1,322,505	1,322,505 ~ 2,537,592	2,537,592 ~ 4,381,621	4,381,621 ~
Number of audiences after release 3 weeks	~ 1,365,492	1,365,492 ~ 2,901,012	2,901,012 ~ 5,266,025	5,266,025 ~
Total audience	~ 500,000	500,000 ~ 1,000,000	1,000,000 ~ 3,000,000	3,000,000 ~

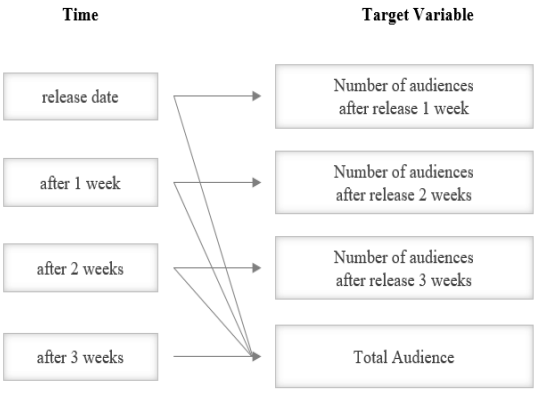
별 관람객 수와 달리 총 관람객 수 범주의 기준은 임준엽 and 황병언(2014)의 연구를 참고하였다. 이들의 연구에서는 5개의 등급으로 나누었으나 영화 편수의 편차가 커서 본 연구에서는 <Table 5>와 같이 4개의 등급으로 나누어 실험에 사용하였다.

〈Table 5〉 Definition of target variable (Total Audience)

CLASS	A	B	C	D
Total audience(million)	~ 0.5	0.5 ~ 1	1 ~ 3	3 ~
Number of Movies	21	40	89	61

4. 예측 모델 생성 및 실험 결과

예측을 위해 기계학습의 지도학습 분류기법을 활용하였다. 지도학습의 다양한 분류기법 중 Naive Bayes, Random Forest, Multi Layer Perceptron(MLP), Support Vector Machine(SVM)을 이용하여 평가하였다. 예측 요인은 <Table 3>에 표현된 것처럼 사용하였으며 예측시점이 뒤로 갈수록 예측 요인의 개수는 늘어난다. 예를 들어 예측시점이 개봉 2주 후 일 경우 개봉일, 개봉 1주일 실적을 모두 활용하여 예측을 수행하였다. 예측 모델은 <Figure 1>에 표현된 것처럼 예측시점마다 차주 누적 관람객 수도 예측하지만 총 관람객 수도 함께 예측해 정확도를 비교하였다. 분석대상이 개봉 후 3주이상 상영한 영화이기 때문에 개봉 4주차 실적이 없는 데이터들이 있어 개봉 3주 후에는 총 관람객 수를 예측한다. 예측을 위한 도구로는 WEKA를 사용하였다. 예측 모형의 신뢰성을 높이기 위해 Naive Bayes, Random Forest, Support Vector Machine(SVM)은 10-fold cross-validation, Multi Layer Perception (MLP)는 4-fold cross-validation을 사용하였다. 새롭게 제시한 요인들의 흥행요인 가능성을 알아보기 위해 우선적으로 새롭게 제시한 변수들을 포함한 모델(prop.)과 포함하지 않은 모델(conv.)을 구성하여 새로운 변수들의 흥행요인 가능성을 알아보았다.



〈Figure 1〉 Prediction model

실험결과는 <Table 6>과 같다. 우선, 4가지 예측 알고리즘 모두에서 본 연구에서 새롭게 제시한 변수들인 매출액 점유율, 흥행 순위, 순위 증감구분, 순위 변화폭, 포함한 모델(prop.)이 그렇지 않은 모델(conv.)에 비해 통계적으로 유의한 수준($p<0.05$)에서 정확도가 높게 나타난 경우가 많았으며(<Table 6>에 bold로 표시), 그 반대의 경우는 발견되지 않았다. 이는 본 연구에서 제시한 새로운 변수들이 예측 모델의 정확도를 향상시키는데 기여를 한 것으로 볼 수 있다.

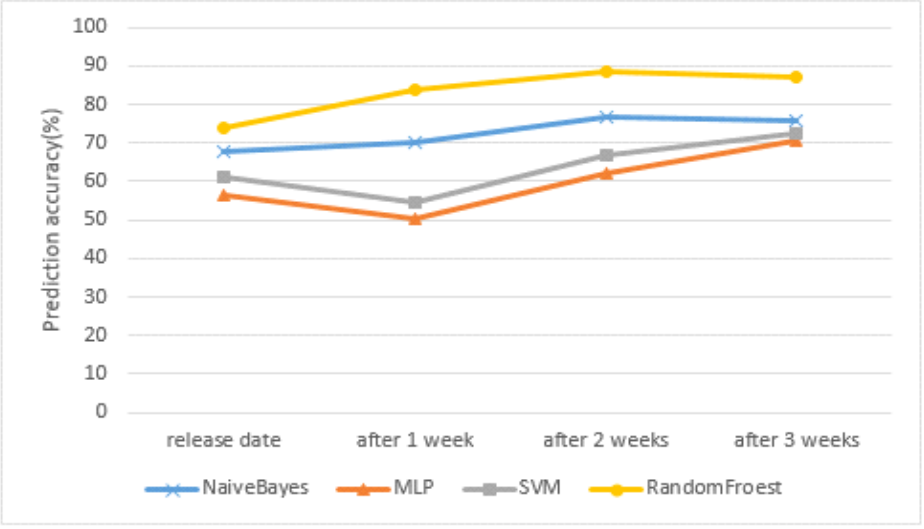
또한 전체적으로 MLP, SVM, Naive Bayes, Random Forest의 순서로 정확도가 높게 측정되었으며 예측시점마다 총 관람객 수를 예측하는 것 보다 차주 누적 관람객 수를 예측한 것이 정확도가 높게 측정되었다.

〈Table 6〉 Comparison of prediction accuracy

Time	Target Variable	Naïve-Bayes		MLP		SVM		Random Forest	
		Prop.	Conv.	Prop.	Conv.	Prop.	Conv.	Prop.	Conv.
release date	Total audience	**63.51***	58.77	52.13	49.76	53.08	49.29	61.61	62.09
	Number of audiences after release 1 week	**67.77***	60.19	**56.40***	50.24	**61.14***	55.92	73.93	72.99
after 1 week	Total audience	67.77	65.40	**55.92***	48.82	**61.61***	55.92	67.77	67.30
	Number of audiences after release 2 weeks	70.14	68.72	50.24	56.87	54.50	54.98	**83.89***	79.62
after 2 weeks	Total audience	72.04	69.19	61.14	55.92	66.82	62.09	74.41	79.15
	Number of audiences after release 3 weeks	76.78	75.36	62.09	63.03	66.82	63.98	88.63	89.10
after 3 weeks	Total audience	75.83	72.99	**70.62***	60.19	72.51	67.30	87.21	88.15

* $p<0.05$

예측 시점에 따라 주차별 누적 관람객 수를 예측한 결과를 나타낸 <Figure 2>를 보면 전반적으로 시간이 지날수록 예측 정확도가 높아지는 것으로 나타났다. Random Forest가 약 73.9% ~ 88.6%로 가장 높게 측정되었고 Naive Bayes가 약 67.7% ~ 75.8%의 정확도를 보였다. MLP와



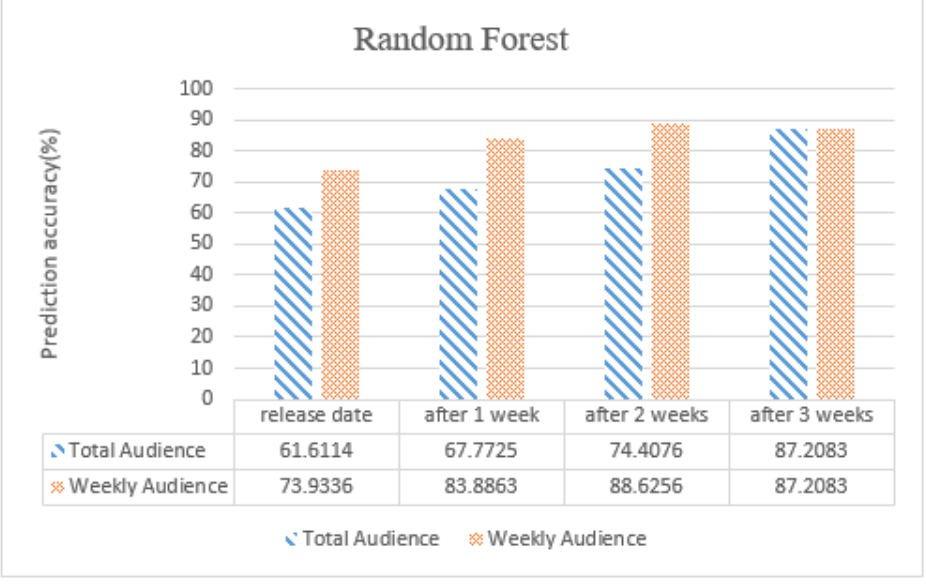
〈Figure 2〉 Comparison of prediction accuracy over time

SVM의 경우 개봉일에 예측한 정확도 보다 개봉 1주후에 예측 정확도가 떨어졌으며 이후에는 예측 정확도가 높아지는 유사한 패턴을 보여주고 있다.

가장 높은 정확도를 나타낸 Random Forest의 예측률을 <Figure 3> 통해 구체적으로 살펴보면 시간이 흐를수록 총 관람객 수를 예측했을 경우와 다음주 누적 관람객 수를 예측하는 것 모두 점점 높아지는 것으로 측정되었다. 또한 대체적으로 다음주 누적 관람객 수를 예측하는 것이 더 높은 정확도를 보였다. 비교 결과를 보면 각 시점마다 약 12% ~ 16%의 예측 정확도 차이를 보이고 있다.

<Figure 4>는 Random Forest의 confusion matrix 이다. 4*4의 matrix이며 4가지의 클래스별로 precision과 recall을 확인할 수 있다. 가장 높은 예측 정확도를 보인 개봉 2주후 3주차 누적 관람객수 예측 결과를 보면 A클래스는 precision이 100%이며 recall이 96.49% 이다. 전체 클래스에 대한 평균 precision은 88%이고 recall은 88.6%로 측정되었다.

본 연구의 결과를 보면 역동적으로 변화하는 다양한 경쟁 요소를 활용해 Random Forest 기법을 적용하여 차주 누적 관람객 수를 예측하는 것이 가장 높은 정확도를 보인다고 결론을 내릴 수 있다. 상대적으로 독립변수의 수가 많고 레코드의 수가 적은 본 연구의 분석 데이터 특성 상, 여러 모델을 결합한 앙상블 기법의 Random Forest가 단일 알고리즘을 사용한 모델에 비해 정확도가 높게 나온 것으로 판단된다.



〈Figure 3〉 Comparison accuracy by the target variable (Random Forest)

release date

Weekly Audience

	A	B	C	D	recall
A	48	8	3	1	80
B	9	33	10	0	63.46
C	0	13	32	5	64
D	0	0	6	43	87.76
precision	84.21	61.11	62.75	87.76	

Total Audience

	A	B	C	D	recall
A	8	11	2	0	38.1
B	9	18	13	0	45
C	1	9	66	13	74.16
D	0	1	22	38	62.3
precision	44.44	46.15	64.08	74.51	

after 1 week

Weekly Audience

	A	B	C	D	recall
A	55	4	0	0	93.22
B	5	43	6	0	79.63
C	0	8	44	5	77.19
D	0	0	6	35	85.37
precision	91.67	78.18	78.57	87.5	

Total Audience

	A	B	C	D	recall
A	10	10	1	0	47.62
B	10	19	11	0	47.5
C	1	7	71	10	79.78
D	0	2	16	43	70.49
precision	47.62	50	71.72	81.13	

after 2 weeks

Weekly Audience

	A	B	C	D	recall
A	55	2	0	0	96.49
B	0	51	2	0	96.23
C	0	7	43	4	79.63
D	0	0	6	38	86.36
precision	100	85	84.31	90.48	

Total Audience

	A	B	C	D	recall
A	10	10	1	0	47.62
B	10	19	11	0	47.5
C	1	7	71	10	79.78
D	0	2	16	43	70.49
precision	47.62	50	71.72	81.13	

after 3 weeks
Total Audience

	A	B	C	D	recall
A	18	3	0	0	85.71
B	3	34	3	0	85
C	1	3	80	5	89.89
D	0	0	9	52	85.25
precision	81.82	85	86.96	91.23	

〈Figure 4〉 Confusion Matrix for Random Forest

5. 결론 및 향후 연구계획

본 논문에서는 기계학습의 분류기법을 이용하여 주차별 누적 관람객 수를 예측하는 기법을 제안하였고 그 결과를 확인하였다.

본 연구의 이론적 시사점은 영화 흥행에 대한 예측을 위해 기존 문헌들에서 많이 다루었던 변수들뿐만 아니라, 기존 문헌들에서 다루지 않았던 새로운 변수들을 추가하여 모델을 개발하였고, 이 변수들이 모델의 성능 향상에 미치는 영향을 밝혀 영화 흥행 예측 모델의 정확도 향상을 위한 새로운 변수를 발견하였다는 점이다. 더불어, 본 연구의 실무적 시사점은 총 관람객 수와 총 매출액에 대한 예측만 시도되었던 기존 연구와는 달리 주차별 누적 관람객 수 예측을 통해 보다 현실적인 환경에서 개봉 후 역동적으로 변하는 관객들의 반응과 영화 흥행 실적에 따라 빠르게 대응하고 입체적으로 분석할 수 있는 방법을 제안하였다는 점이다.

향후 연구로는 본 연구 결과를 바탕으로 일 단

위 예측을 통해 좀 더 실용적 적용 가능성이 높은 연구를 진행해야 할 것으로 판단된다. 또한 일 단위 예측을 위해 좀더 다양하게 변화하는 경쟁요인들을 발굴하고 각 분석 기법에 맞는 변수들을 활용하여 보다 높은 정확도 높은 예측모델을 구현할 수 있을 것으로 기대된다. 마지막으로 본 연구에서 사용한 분석 대상 영화의 수가 적은 점은 본 연구의 한계점으로 볼 수 있으나, 이는 흥행 실적의 편차가 상당히 큰 영화 자체의 특수성으로 인해 기존의 영화 흥행 예측 연구들도 공통적으로 갖고 있는 문제로서 향후 여러 기간에 걸친 데이터 수집을 통해 완화할 수 있을 것으로 기대된다.

참고문헌(References)

Korean Film Council (2017). 2017 Korean film industry settlement, Korean Film, Available at http://www.kofic.or.kr/ (Downloaded 13 February, 2018)

Chang, J. Y., “An Experimental Evaluation of Box office Revenue Prediction through Social Bigdata Analysis and Machine Learning.” *The journal of the institute of internet, broadcasting and communication*, Vol.17, No.3(2017), 167-173.

Choi, S. H., "The Impact of Distributors in the Movie Exhibition Market : Focusing on Distributor Types," *Review of Culture & Economy*, Vol.20, No.1(2017), 105~128.

Jeon, S. H., and Y, S, Son, “Prediction of box office using data mining,” *The Korea Journal of Applied Statistics*, Vol.29, No.7(2016), 1257-1270.

Kang, S. J., "Analysis Box Office Success of A Movie - Focused on Commercial Film Released in 2016," *Journal of the Korea Entertainment Industry Association*, Vol.11, No.5(2017), 1~15.

Kim, B. S., "Comparison of Factors Predicting Theatrical Movie Success : Focusing on the Classification by the Release Type and the Length of Run," *Korean Journal of Journalism & Communication Studies*, Vol.53, No.1(2009), 257~287.

Kim, S.Y, S. H. Im, and Y. S. Jung, "A Comparison Study of the Determinants of Performance of Motion Pictures: Art Film vs. Commercial Film," *Journal of The Korea Contents Association*, Vol.10, No.2(2010), 381~393

Kwon. S. J., "The Impact of Critics on Movie Market Performances: Art versus Commercial," *Review of Culture & Economy*, Vol.17, No.3(2014), 3~21.

Litman B., “Predicting success of theatrical movies: An empirical study”, *The Journal of Popular Culture*, Vol.16, No.4(1983), 159~175.

Park, S. H. and H. J. Song, "Word of Mouth and Box Office Performance: WOM’s Impact on Weekly Box Office Revenues," *Korean Journal of Journalism & Communication Studies*, Vol.56, No.4(2012), 210~235.

Quader, N., Gani, M. O., Chaki, D., & Ali, M. H., "A machine learning approach to predict movie box-office success", *Computer and Information Technology (ICCIT),* Vol.20 (2017), 1-7.

Rhee , T. G., and F. Zulkernine, “Predicting Movie Box Office Profitability: A Neural Network

Approach," *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA),* 2016.

Song, J. W. and S. J. Han, "Predicting gross box office revenue for domestic films," *Communications for Statistical Applications and Methods*, Vol.20(2013), 301-309.

Yim, J. Y. and B. Y. Hwang, "Predicting Movie Success based on Machine Learning Using Twitter," *KIPS transactions on Software and Data Engineering*, Vol.3, No.7(2014), 263-270.

Yoo, H. J. "The Determinants of Motion Pictures Box Office Performances - For Movies Produced in Korea Between 1988 and 1999," *Korean Journal of Journalism & Communication Studies*, Vol. 46, No. 3 (2002), 183~213.

Abstract

Development of New Variables Affecting Movie Success and Prediction of Weekly Box Office Using Them Based on Machine Learning

Junga Song* · Keunho Choi** · Gunwoo Kim***

The Korean film industry with significant increase every year exceeded the number of cumulative audiences of 200 million people in 2013 finally. However, starting from 2015 the Korean film industry entered a period of low growth and experienced a negative growth after all in 2016. To overcome such difficulty, stakeholders like production company, distribution company, multiplex have attempted to maximize the market returns using strategies of predicting change of market and of responding to such market change immediately. Since a film is classified as one of experiential products, it is not easy to predict a box office record and the initial number of audiences before the film is released. And also, the number of audiences fluctuates with a variety of factors after the film is released. So, the production company and distribution company try to be guaranteed the number of screens at the opining time of a newly released by multiplex chains. However, the multiplex chains tend to open the screening schedule during only a week and then determine the number of screening of the forthcoming week based on the box office record and the evaluation of audiences. Many previous researches have conducted to deal with the prediction of box office records of films. In the early stage, the researches attempted to identify factors affecting the box office record. And nowadays, many studies have tried to apply various analytic techniques to the factors identified previously in order to improve the accuracy of prediction and to explain the effect of each factor instead of identifying new factors affecting the box office record. However, most of previous researches have limitations in that they used the total number of audiences from the opening to the end as a target variable, and this makes it difficult to predict and respond to the demand of market which changes dynamically. Therefore, the purpose of this study is to predict the weekly number of audiences

* Department of BigData Business, Hanbat National University
** Department of Business & Accounting, Hanbat National University
*** Corresponding Author: Gunwoo Kim
Department of Business & Accounting, Hanbat National University
125, Dongseo-daero, Yuseong-gu, Daejeon, Korea, 34158
Tel: +82-42-821-1290, Fax: +82-42-821-1597, E-mail : gkim@hanbat.ac.kr

of a newly released film so that the stakeholder can flexibly and elastically respond to the change of the number of audiences in the film. To that end, we considered the factors used in the previous studies affecting box office and developed new factors not used in previous studies such as the order of opening of movies, dynamics of sales. Along with the comprehensive factors, we used the machine learning method such as Random Forest, Multi Layer Perception, Support Vector Machine, and Naive Bays, to predict the number of cumulative visitors from the first week after a film release to the third week. At the point of the first and the second week, we predicted the cumulative number of visitors of the forthcoming week for a released film. And at the point of the third week, we predict the total number of visitors of the film. In addition, we predicted the total number of cumulative visitors also at the point of the both first week and second week using the same factors. As a result, we found the accuracy of predicting the number of visitors at the forthcoming week was higher than that of predicting the total number of them in all of three weeks, and also the accuracy of the Random Forest was the highest among the machine learning methods we used. This study has implications in that this study 1) considered various factors comprehensively which affect the box office record and merely addressed by other previous researches such as the weekly rating of audiences after release, the weekly rank of the film after release, and the weekly sales share after release, and 2) tried to predict and respond to the demand of market which changes dynamically by suggesting models which predicts the weekly number of audiences of newly released films so that the stakeholders can flexibly and elastically respond to the change of the number of audiences in the film.

Key Words : Movie, Box Office, Box Office Revenue, Box Office Factors, Prediction of Box Office, Predicting Number of Audience, Machine Learning

Received : August 2, 2018 Revised : December 1, 2018 Accepted : December 16, 2018
Publication Type : Regular Paper Corresponding Author : Gunwoo Kim

저 자 소 개

송정아
현재 대전에 소재한 국립한밭대학교 창업경영대학원 빅데이터 비즈니스학과 석사과정에 재학중이다. 관심분야는 머신러닝, 데이터마이닝 등이다.

최근호
현재 대전에 소재한 국립한밭대학교에서 경영회계학과 조교수로 재직하고 있다. 고려대학교 경영학과에서 박사 학위(MIS 전공)를 받았으며, 근로복지공단 근로복지연구원에서 데이터 분석 업무를 총괄하는 책임연구원으로 근무하였다. 주요 관심분야는 추천 시스템, 의료 빅데이터 분석, 딥러닝, 머신러닝, 데이터마이닝 등이다.

김건우
현재 대전에 소재한 국립한밭대학교에서 경영회계학과 부교수로 재직하고 있다. 연세대학교 공과대학에서 컴퓨터 사이언스를 전공하였으며 고려대 경영학과에서 석사를 졸업하고 동대학에서 박사 학위를 받았다. 현재 한국창업학회 부회장을 맡고 있으며 ICT플랫폼학회 빅데이터과 위원장을 맡고 있다 그 외 다수의 학회에서 편집위원 및 이사로써 활동하고 있다. 주요 관심분야는 비즈니스 애널리틱스, 온톨로지 모델 설계 및 적용, 빅데이터 분석 및 블락체인 기술 및 적용 등이다.