석사학위논문

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

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

한밭대학교 창업경영대학원

빅데이터비즈니스학과

송 정 아

2019년 2월

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

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

지도교수 김 건 우

이 논문을 경영학석사학위
청구논문으로 제출함

2018년 11월

한밭대학교 창업경영대학원

빅데이터비즈니스학과

송 정 아

송정아의 석사학위 논문을 인준함

심사위원장 _____임 재 학_____(인)

심 사 위 원 _____김 건 우_____(인)

심 사 위 원 _____최 근 호_____(인)

2018년 12월

한밭대학교 창업경영대학원

목 차

표 목 차

그 림 목 차

국 문 요 약

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

논문제출자 송 정 아
지 도 교 수 김 건 우

2013년 누적 인원 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, 매출 점유율, 흥행순위 변동 폭 등 선행연구에서 사용되지 않았던 데이터들을 새로운 요인으로 사용하고 Random Forest, Multi Layer Perception (MLP), Support Vector

Machine (SVM), Naive Bayes 등의 기계학습 기법을 이용하여 개봉 일 후, 개봉 1주 후, 개봉 2주 후 시점에 차주 누적 관객 수를 예측한다. 차주 누적 관객 수는 사분위와 십분위로 범주화 한 후 정확도를 비교하였으며 개봉 3주 후 시점에는 전체 관람객 수를 예측하였다. 정확도 비교를 위해 매 예측 시점마다 동일한 예측 요인을 사용하여 전체 관람객 수를 예측하여 비교해 보았다. 분석결과 동일한 시점에 전체 관람객 수를 예측했을 경우 보다 차주 누적 관람객 수를 예측하는 것이 더 높은 정확도를 보였으며, 모델의 신뢰도를 높이기 위해 k- fold cross validation을 사용 하였는데 10 - fold cross validation에서 가장 높은 정확도를 보였다. 새로운 요인들의 연구에서는 새로운 요인을 포함한 모델이 포함하지 않은 모델보다 대부분 높은 정확도를 나타냈다. 네가지 기계학습 기법 중에는 Random Forest가 가장 높은 정확도를 보였다. 본 연구를 통해 첫째, 새로운 요인들의 흥행요인 가능성을 확인할 수 있었으며, 둘째, 이러한 여러 요인들을 고려하여 각 시점 별 차주 누적 관람객 수 예측을 통해 영화산업 이해관계자들이 관람객들의 반응에 탄력적으로 대응할 수 있는 빠르고 정확한 의사결정을 내리는데 도움을 줄 수 있을 것이다.

주제어

영화 흥행 예측, 영화 관람객 수 예측, 박스오피스 예측, 기계학습

I. 서론

1. 연구배경

한국영화 상영 시장의 전체 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. 매해 전국 극장 수와 스크린 수는 지속적으로 증가하고 있는 반면 전체 관람객 수와 극장 매출액의 증가 폭은 미미하다. (영화진흥위원회 , 201). 이렇듯 한국 영화 시장은 2015년을 기점으로 저성장 시대에 접어들었고 앞으로도 큰 변화 없이 저성장에 머무를 것으로 전망된다(영화진흥위원회 한국영화산업결산, 2017). 그러나 인터넷을 통해 방송 프로그램, 영화, 교육 등 각종 미디어 콘텐츠를 제공하는 서비스인 OTT (Over The TOP)(방송통신위원회, 2016; 한경닷컴사전)의 등장으로 인터넷 VOD (Video On Demand) 시장과 IPTV 및 디지털 케이블 TV 등의 온라인 디지털 콘텐츠 시장은 지속적으로 성장하고 있다. OTT는 초기에서는 Over-The-Top 의 줄임말로 TV에 연결하는 셋탑박스 (Set-Top-Box)를 말하였으나 지금은 넓은 의미로 인터넷 기반의 동영상 서비스라면 모두 OTT의 한 형태라는 식으로 쓰이고 있으며 OTT 서비스의 가장 대표적인 주자로 넷플릭스를 꼽을 수 있다(서기만, 2011).

불과 몇 년 전만해도 인터넷 VOD나 IPTV에서는 영화관에서 상영 종료한 영화들이 콘텐츠로 제공되었으나 현재는 영화관 개봉과 동시에 인터넷 VOD와 IPTV에서도 콘텐츠를 제공하기 시작했기 때문에 영화소비자들은 굳이 날짜와 시간에 맞춰 영화관에 가지 않아도 가정에서 편하게 개봉 시점에 맞춰

영화를 즐길 수 있게 되었다. 이렇게 영화 소비자들의 선택의 폭이 넓어졌기 때문에 영화관들은 IMAX, 3D, 4D 영화같이 특별관에서 관람하면 좋은 영화나 옛날 명작들을 재개봉하며 관객들을 모으려고 노력 중이다.

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 초기에 매출실적을 높이고자 한다. 하지만 극장주들은 관객들에게 영화 예매를 위한 상영 시간표를 약 일주일 치 정도만 제공하고 있으며 영화 개봉 후에 관람객들의 평가와 흥행 실적을 바탕으로 다음주의 상영 여부 뿐 아니라 다음주에 개봉하게 될 영화와 상영하게 될 영화의 상영 횟수나 스크린 수 등을 결정하게 된다. 이러한 이유로 규모가 큰 멀티플렉스의 경우 교차 상영이 빈번하게 이루어지기도 한다. 영화산업을 이루고 있는 각 이해당사자인 제작사, 배급사, 극장 주 들은 개봉 영화에 대한 역동적으로 변화하는 시장의 반응을 빠르고 정확하게 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 최소화함과 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 수행되어왔다. 초기에는 영화 흥행에 영향을 미치는 직접적인 요인을 밝히고자

노력해 왔으며 (Litman B., 1983; 김휴종, 1988; 유현석, 2001; 유현석, 2002; 김병선, 2009; 강선주, 2017), 인터넷과 SNS(Social Network Service)의 발달로 온라인 데이터를 이용하여 온라인 상의 구전의 영향력을 분석하는 연구들, 특별히, 천만 관객을 돌파한 영화들의 흥행 요인 분석에 관한 연구들이 주로 수행되었다(김범수 and 서주환. 2017; 김연형 and 홍정한 2013; 장리, 최강준, 이재영. 2017). 최근의 연구들은 새로운 요인들을 규명하는 대신 과거 선행 연구에서 사용되었던 변수들에 다양한 예측 분석기법을 적용하여 흥행 예측의 정확도를 높이는데 집중하고 예측 모델에서 도출된 변수들의 영향력을 설명하고자 하는 시도들이 많이 이루어지고 있다(송종우 and 한수지, 2013; 임준엽 and 황병연, 2014; 전성현 and 손영숙, 2016; 장재영, 2017; T. G. Rhee and F. Zulkernine, 2016; Nahid 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. 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 신뢰할만한 영화 흥행 요인들을 명확히 밝히기는 쉽지 않다.

2. 연구목적

본 연구의 목적은 영화 이해 당사자들이 개봉 후 역동적으로 변화하는 영화 관람객의 수요 변화에 대해 탄력적인 대응을 할 수 있도록 영화 개봉 후 종영 시까지 전체 기간이 아닌 영화 개봉 직후 및 주차 별 누적 관람객 수를 예측하는 모델을 제안하는 것이다. 이를 위해 다양한 기계학습 기법을

활용하여 각 예측 모델을 구축하고 평가함으로 각 예측 시점에 가장 적합한 예측 모델을 찾고자 한다. 뿐만 아니라 기존 연구들에서 밝혀 지지 않은 영화 흥행 요인들을 찾고 이를 검증하고자 한다.

3. 연구범위 및 구성

(1) 연구범위

본 연구에서 분석 대상의 범위는 2015년부터 2017년까지 3년동안 개봉한 영화를 대상으로 선정하였다. 영화진흥위원회와 포털사이트 네이버 영화에서 데이터를 수집하고 데이터 전처리 과정을 통해 새로운 요인을 생성하고 예측을 위한 클래스를 분류한다. 이렇게 생성된 분석데이터 셋을 활용하여 기계학습 중 클래스 분류 모델인 Naive Bayes, Random Forest, Support Vector Machine(SVM), Multi Layer Perception(MLP) 분석기법을 이용해 다양한 예측 모델을 생성하여 새로운 요인들의 영향력을 평가하고 각 주차별로 정확도 평가와 분석기법별 정확도 평가를 통해 가장 좋은 모델을 선정하고자 한다.

(2) 연구구성

본 연구는 영화흥위원회에서 운영하는 통합전산망에서 제공하는 API를 이용하여 영화기본정보, 영화일별 흥행실적, 영화 년도별 통계, 영화인 정보등의 자료를 수집하고, 네이버 포털사이트 영화 섹션에서 전문가 평가와 네티즌 평가를 수집하여 DBMS에 저장한 후 데이터 전처리 과정을 거쳐 예측을 위한 차주 누적 관람객 수와 전체 관람객 수의 클래스를 분류하고 최종적으로 분석 데이터 셋을 구축한다. 알고리즘은 클래스 분류 기법인 Naive Bayes, Random Forest, Multi Layer Perception(MLP), Support Vector Machine(SVM) 을 이용한다. 예측 시점은 주차별 예측을 위해 개봉일 이후,

개봉 1주 후, 개봉 2주 후, 개봉 3주 후에 차주 누적 관람객 수를 예측하게 되는데 개봉 3주후에는 개봉 4주후 실적이 있는 영화의 수가 적어 전체 관람객 수만 예측하게 된다. 차주 누적 관람객 수는 사분위와 십분위로 범주화하여 클래스를 분류하고 각 예측시점마다 선행연구와의 비교를 위해 전체 관람객 수를 예측하게 된다. 또한 새로운 요인의 흥행 요인 가능성을 알아보기 위해 새로운 요인을 포함한 모델과 포함하지 않은 모델을 비교하는 실험도 구성하였다.

4. 연구방법

본 연구에서 예측을 위한 분석기법으로는 클래스 분류기법인 Naive Bayes, Random Forest, Support Vector Machine(SVM), Multi Layer Perception(MLP)을 사용하였으며 k-fold cross-validation을 통해 모델을 생성 및 평가하였다. Naive Bayes는 베이즈 정리(Bayes’ s theorem)에 근거한 조건부 확률 모델로 여러 개의 속성들 속에서 하나의 속성 값을 기준으로, 다른 속성이 독립적이라고 가정하고 해당 속성값이 클래스를 분류에 미치는 영향을 확률적 결과들로 나타내는 기법이다. 베이즈 정리는 사전 확률과 사후 확률 사이의 관계를 나타내는 정리로 두 확률변수 간의 관계를 조건부 확률로 기술한 모델이다(Heckerman D., 1997; Neapolitan, R. E., 2004).

Random Forest는 수많은 의사결정 트리가 모여서 생성된 모델로 무작위 속성 선택 기법을 엮어서 만들고 변수들을 랜덤하게 선택하여 만든 여러 개의 의사결정트리에서 나오는 빈도수를 가지고 결정하게 된다(Breiman, L.. 2001). 장점으로는 모델의 효율성이 높으며 배깅(Bagging), 부스팅(Boosting)보다 빠르며 내부 변수 추정의 중요성 파악이 가능하다는 점을 들 수 있다.

배깅(Bagging)은 여러개의 Bootstrap 즉 무작위로 여러개의 샘플을 뽑아 학습결과를 집계(Aggregating)한다는 Bootstrap Aggregating의 줄임말로 대상

데이터를 복원 랜덤 샘플링을 통해 표본을 추출하고 동일한 모델을 이용해 학습 시킨 후 집계된 학습결과로 모델을 생성한다. 배깅은 각 모델의 결과의 중간값을 맞추어주므로 데이터 과적합(Overfitting)을 피할 수 있으며 알고리즘의 안정성 뿐 아니라 정확성 역시 향상시킬 수 있다(Breiman, L.. 1996).

부스팅(Boosting)은 배깅(Bagging)처럼 복원 랜덤 샘플링을 통해 표본을 추출하지만 가중치를 부여하고 순차적으로 학습을 수행한다. 순차적 학습을 통해 이전 학습결과를 토대로 다음 학습데이터의 가중치를 조정해가면서 학습을 진행한다. 이런 반복적인 학습을 통해 정확도를 높여 나가는 방법이다(Natekin, A., & Knoll, A., 2013).

SVM은 패턴 인식, 자료 분석을 위한 지도 학습(Supervised Learning)으로, 선형, 비선형을 구분하지 않아 분류와 회귀에 주로 사용된다. 데이터를 선형으로 분류 할 수 있는 경우에는 서포트 벡터(support vector)사이의 거리인 마진(Margin)값을 이용하여 분류하고 선형으로 분류할 수 없는 경우 SVM은 입력 공간상에 커널함수(Kernel Function)를 이용하여 차원을 높여 분류한다. 커널함수는 주어진 데이터들이 선형 분류가 불가능할 경우 데이터들을 고차원의 공간으로 높이고 고차원 공간을 이용하여 선형 분류를 할 수 있는 결정 경계를 찾을 수 있게 해준다. 커널 함수의 종류에는 Polynomial Kernel, Sigmoid Kernel, 가우시안 RBF Kernel 등 많은 종류가 있으나, 그 중 가장 성능이 좋은 것은 가우시안 RBF Kernel이다. SVM은 다양한 데이터 종류에 적용하기 쉽고 노이즈 데이터에 대한 영향력이 적으나 최적의 모델을 찾기 위해 Kernel과 매개변수들 간의 여러 조합에 대한 테스트가 필요하며 입력 데이터 셋의 개수와 속성의 수가 많으면 학습이 늦어질 수 있는 단점이 있다(Gunn, S. R., 1998).

MLP는 기존 입력층(Input Layer)과 출력층(Output Layer)으로만 구성된 Single Layer Perception (SLP)의 한계를 극복하기 위해, 입력층과 출력층 사

이에 은닉층(Hidden Layer)을 두어 비선형 분류가 가능하도록 고안된 인공신경망 기법이다. MLP는 은닉층이 있어 활성함수가 여러개이고 이에 따른 가중치도 여러개이다.

본 연구에서는 정확도와 신뢰도를 높이기 위해 k-fold cross validation을 사용한다. k-fold cross validation은 데이터들을 랜덤하게 k 개의 그룹으로 나누고 k-1개의 그룹을 학습에 사용하고 나머지 1개의 그룹을 모델의 정확도를 예측하는데 사용한다. 이러한 과정을 k번 반복한 후 생성된 정확도를 평균 내어 모델의 정확도로 표현한다.

예측을 위한 흥행 요인으로는 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, 개봉 1주차, 개봉 2주차 시점에는 차주 누적 관람객 수를 예측하는데 차주 누적 관람객 수를 측정한 선행 연구들이 없어 차주 누적 관람객 수를 사분위와 십분위 범주로 나누어 예측하였다. 개봉 3주차 시점에는 전체 관람객 수만을 예측하였다. 선행연구들과의 비교를 위해 개봉일 이후, 개봉 1주 이후, 개봉 2주 이후 시점에도 동일한 예측 요인을 사용하여 전체 관람객 수도 같이 예측하였다.

5. 연구결과

본 연구에서는 탄력적인 영화 흥행예측을 위해 다양한 머신러닝 기법을 이용하여 주차별 흥행예측을 실험하였다. 또한 선행연구와의 비교를 위해 전체 관람객 수도 함께 예측하였다. 실험결과 예측 시점이 뒤로 갈수록 예측 정확도가 점점 높아지고 k- fold cross validation은 대부분 10-fold cross validation에서 높은 예측정확도를 보였다. 동일한 시점에서의 결과를 보면 차주 누적 관람객 수 예측의 경우 십분위로 나눈 타겟 클래스보다 사분위로

나눈 타겟 클래스의 정확도가 높았으며 전체 관람객 수 예측 정확도 비교에서는 사분위로 나눈 차주 누적 관람객 수를 예측하는 것이 전체 관람객 수 예측보다 높은 예측 정확도를 보였다. 새로운 요인들의 연구에서는 새로운 요인을 포함한 모델이 포함하지 않은 모델보다 대부분 높은 정확도를 나타냈다. 네가지 기계학습 기법 중에서는 Random Forest가 73% ~ 88.63%로 가장 높은 예측 정확도를 보였으며 그 중 개봉 2주후 개봉 3주차 누적관람객 수를 예측한 모델이 88.63%로 가장 높게 나왔다. 마지막으로 예측정확도가 가장 높게 나온 Random Forest 모델을 이용하여 변수 선택법인 GainRatio 와 InfoGain을 적용하여 분석한 결과 초반에는 GainRatio 와 InfoGain을 이용하여 선택된 변수만 사용하는 모델이 예측 정확도가 높았으나 예측 시점이 뒤로 갈수록 GainRatio 와 InfoGain을 적용한 모델과 전체요인을 사용한 모델의 예측정확도는 차이가 없었다.

6. 연구의의

본 연구의 의의는 첫째, 기존의 흥행 요인들 뿐만 아니라 영화 개봉 후 관람객들의 평가와 흥행 실적을 바탕으로 한 네티즌 평점, 흥행 순위, 매출점유율 등 그 동안 연구들에서 다루지 않았던 흥행에 영향을 미칠만한 다른 여러 요인들을 포괄적으로 고려하였다는 점이며 이러한 요인들을 포함한 모델이 정확도가 높은 것을 통해 새로운 흥행요인의 가능성을 확인 할 수 있었으며, 둘째, 이러한 여러 요인들을 고려하여 각 시점 별 차주 누적 관람객 수를 미리 예측할 수 있는 모델을 제시함으로 영화산업 이해관계자들이 개봉 후 역동적으로 변화하는 관람객들의 반응에 탄력적으로 대응할 수 있는 빠르고 정확한 의사결정을 내리는데 도움을 줄 수 있다는 것이다.

Ⅱ. 관련연구

1. 흥행요인에 관련한 연구

일반적으로 영화 흥행과 관련된 연구는 영화 흥행에 영향력을 미치는 흥행 요인들의 선택에 관한 연구(Litman, 1983; 김휴종, 1988; 유현석, 2001; 유현석, 2002; 김병선, 2009; 강선주, 2017)와 이들 흥행 요인들로부터 영화 흥행 예측 모델을 연구하는 두 가지 주제로 분류된다(송종우 and 한수지, 2013; 임준엽 and 황병연, 2014; 전성현 and 손영숙, 2016; 장재영, 2017; T. G. Rhee and F. Zulkernine, 2016; Nahid 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 장르와 관람 등급 등이 영향력 있다는 결론을 도출 하였다.

국내의 연구에서는 김휴종(1988)이 1988년 ~ 1995년 서울에서 개봉한 영화의 흥행실적과 출연 주연배우를 50명을 대상으로 한국 영화스타들이 영화의 흥행 기여도를 알아보기 위해 경제학적 분석을 수행하였다. 그 결과 주연배우의 스타성은 추가 관객수 약 3만명을 등원하는 것으로 나타났으며 개개인의 스타성은 약 28명의 스타파워가 유의했으며 시간의 흐름에따라 스타파워가 변하는 형태도 다르다는 결론을 도출했다. 유현석(2001)은 한국영화의 흥행에 영향을 주는 변수의 영향력을 알아보기위해 제작과 관련된 요인들을 이용해 제한적 연구를 하였다. 그 결과 제작과 관련된 요인들만으로는 흥행성과를 설명하는데 어려움이 있다는 결론을 도출하였고 제작뿐 아니라 배급

단계의 요인을 포함하여 좀 더 통합적인 분석이 필요하다고 결론을 도출했다. 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 나눈 후 상호 비교를 통해 영화 유형별 흥행에 미치는 요인은 서로 다르다는 결과를 도출했고, 김소영 et al. (2010)의 연구에서는 영화 유형을 상업영화와 예술영화로 분류하고 유형에 따른 예측요인 비교를 통해 스크린 수, 관람객 평가, 장르는 상업영화와 예술영화모두 유의하나 다른 변수들은 서로 다르다는 연구결과를 발표하였다. 권선주(2014)는 전문가 평가가 영화흥행성과에 미치는 영향력을 연구를 통해 전문가 평점은 전체영화를 대상으로는 유의미하였으나 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 선행연구와 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. 권재웅과 홍병기(2012) 연구에서는 2004년부터 2011년까지 한국에서 개봉된 애니메이션 195편을 대상으로 전국 관객 동원 수를 분석대상으로 삼아서 제작, 배급, 평가 측면을 고려해 흥행에 영향을 미치는 요인을 분석한 결과 애니메이션 영화의 흥행성과에 영향을 주는 변수들은 개봉 스크린, 네티즌 빈도, 개봉 시즌에서는 여름인 것으로 나타났다. 박승현과 이푸름(2017)은 2011년

부터 2015년까지 한국에서 개봉된 애니메이션 213편을 대상으로 전국 관객수에 영향을 미친 요인을 분석하였다. 속편 형태를 인쇄물, 영상물에 근거한 형태와, 기타로 분류한 요인을 추가하여 구체적인 분석을 한 결과 인쇄물에 근거한 형태와 영상물에 근거한 형태가 유의미한 영향력을 끼치는 것으로 나타났다. 선행 연구결과처럼 온라인 평가 참여빈도와 개봉 스크린 규모가 유의미한 관계성을 보였다.

국가 간의 흥행요인을 비교한 연구들도 있는데 이양환 et al.(2007)은 미국과 한국의 흥행요인에는 어떤 차이가 있는지 알아보기 위해 미국과 한국에서 개봉한 영화 142편을 대상으로 전체 흥행실적과 개봉 첫 주의 흥행실적을 비교하였다. 그 결과 미국 영화의 흥행요인들이 한국에서 영화 흥행을 유사하게 설명할 수 없다고 하였고 개별 요인들에서는 상당한 차이가 있었으나 전체적으로는 유사성을 발견하였다. 고정민(2010)은 애국심을 중심으로 미국과 한국 영화시장의 흥행요인을 알아보았다. 애국심은 설문조사를 통해 자료를 확보하였고 제작, 배급, 온라인 구전 등의 요인을 활용하여 요인들의 상관관계와 회귀분석을 수행한 결과 미국과 한국 모두 애국심은 흥행에 영향을 주는 것으로 나타났다. 그 결과를 통해 제작에서 마케팅까지 애국심의 요소를 활용할 필요가 있다고 주장하고 있다.

<표 2-1> 흥행요인에 관련한 연구

저자	연구결과
Litman (1983)	창작, 마케팅, 배급영역으로 나누어 분석한 결과 제작비, SF/Horror, 관람등급이 영향력 있음
김휴종 (1998)	1988년 ~ 1995년 서울 개봉 영화의 주연배우 50명을 대상으로 스타성을 분석한 결과 스타성은 영향력을 가

	지며 시간의 흐름에 따라 스타파워가 변하는 형태도 다르다는 결론을 도출함
유현석 (2002)	1988년 ~ 1999년 동안 개봉한 한국영화 732편을 대상으로 분석한 결과 상식적으로 흥행 요인이라 추정되었던 출연배우, 감독, 제작사, 장르, 등급의 영향력을 실증적으로 입증함
이양환 et al. (2007)	미국과 한국의 흥행요인에는 차이를 분석한 결과 미국영화의 흥행요인들이 한국에서 영화 흥행을 유사하게 설명할 수 없으며 개별 요인들에서는 상당한 차이가 있었으나 전체적으로는 유사성을 발견함
김병선 (2009)	개봉방식과 상영방식에 따라 유형을 구분하여 분석한 결과 유형마다 흥행에 영향을 미치는 요인이 다름
김소영 et al. (2010)	상업영화와 예술영화로 유형 구분. 스크린 수, 관람객 평가, 장르는 모두 유의하나 그 외 변수들은 영화 유형에 따라 다름
고정민 (2010)	애국심을 중심으로 미국과 한국 영화시장의 흥행요인을 알아본 결과 미국과 한국 모두 애국심은 영향을 주는 것으로 나타남
권재웅 and 홍병기 (2012)	2004년부터 2011년까지 한국에서 개봉된 애니메이션 195편을 대상으로 분석한 결과 개봉 스크린, 네티즌 빈도, 개봉 시즌에서는 여름이 유의한 것으로 나타남
권선주 (2014)	전문가 평가가 영향에 미치는 영향 연구한결과 전체영화에서 유의미하나 산업/예술영화로 나누었을 때는 예술영화만 유의미 함
강선주 (2017)	2016년 개봉 상업영화 흥행요인 분석연구에서 개봉 스크린, 장르, 스타 캐스팅, 배급사 등의 영향력이 유의미하고 스타성은 필요요소이긴 하지만 필수요소는 아님
박승현 and 이푸름 (2017)	2011년부터 2015년까지 한국에서 개봉된 애니메이션 213편을 대상으로 분석한 결과 속편 형태가 인쇄물에 근거한 형태와 영상물에 근거한 형태가 유의미한 영향

	력을 끼치고, 선행 연구결과처럼 온라인 평가 참여빈도와 개봉 스크린 규모가 유의미한 관계성을 보임

2. 흥행예측에 관련한 연구

최근 들어서는 선행연구에서 도출된 변수들을 예측 변수로 활용하여 전체 관람객 수와 총 매출액의 예측 정확도를 향상시키려는 연구들이 시도되고 있는데, T. G. Rhee and F. Zulkernine (2016) 의 연구에서는 다양한 웹사이트에서 데이터를 수집해 예측 변수를 생성하고 흥행, 실패라는 두 개의 클래스로 분류한 후 역전파 인공신경망(Back-propagation neural network) 기법을 적용하여 91%의 분류 정확도를 보여줬다. Nahid Quader et al., (2017)은 서 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 Perception 가 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 5억 이상인 한국 영화 206편을 분석대상으로 하여, Linear regression analysis, Random Forest, Gradient 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 전체 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 Support Vector Machine을 사용하여 전체 관람객 수를 예측하였다. 모든 자료를 대상으로 적합 시켰을 경우 모든 시점에서 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 전체 관람객 수를 5개의 클래스로 분류하여 예측한 결과 주요한 정적 데이터와 동적 데이터를 모두 포함한 Neural network 모형이 68%로 가장 높은 예측 정확도를 보였다. 권신혜 et al.(2017)의 연구에서는 제작 및 투자, 배급, 상영단계별 모형을 구성하고 모든 변수를 사용한 모델과 회귀분석을 통해 유의한 변수들만을 사용한 모델을 의사결정트리와 인공신경망을 이용해 실험하였다. 그 결과 제작 및 투자, 배급 모형에서는 모든 변수를 사용하였을 때는 인공신경망이 더 높게 나왔으며, 회귀분석을 이용하여 유의미한 결과를 나타낸 변수를 활용하였을 경우에는 의사결정트리가 더 높은 정확도를 보였고, 상영 모형에서는 둘 다 인공신경망의 정확도가 높은 결과를 보였다. RU, Yunian, et al. (2018)은 시간 구성 요소와 정적인 영화 특성 구성 요소로 구성된 Deep-DBP라고 하는 일일 박스오피스 예측을 위한 종단간 심층 학습 모델을 제안하였고 시간 구성 요소만을 사용하였을 때는 예측 정확도가 37% 정적인 요소를 추가하였을 경우 예측정확도가 7% 높아지는 결과를 나타냈다.

<표 2-2> 흥행예측에 관련한 연구

저자	연구결과
T. G. Rhee and F. Zulkernine (2016)	인공신경망의 역전파 기법을 적용한 결과 91%의 분류 정확도를 보임
Nahid Quader et al, (2017)	개봉 전 요소와 개봉 후 요소로 구분 후 SVM과 인공신경망 기법을 적용하여 분석한 결과 개봉 후 요소를 포함한 모델이 정확도가 높게 나옴 인공신경망이 89.27% 가장 높은 예측률을 나타냄
송종우 and 한수지 (2013)	2008년~2011년 매출규모 5억이상인 영화 206편을 대상으로 Linear regression analysis, Random Forest, Gradient Boosting 기법을 적용하여 분석한 결과 Gradient Boosting이 예측률이 가장 좋음
임준엽 and 황병연 (2014)	2013년 무작위 60편을 대상으로 오프라인 요소와 온라인 요소로 구분하여 나이브 베이지안 기법 사용 개봉일 78.4%, 개봉 1주일 후 95% 온라인 요소를 포함한 예측모델이 전체적으로 정확도 높음
전성현 and 손영숙 (2016)	2012년 ~2015년 관람객수 50만 이상 : 276편을 대상으로 개봉 전, 개봉 1주일, 개봉 2주일 시점에 의사결정나무, SVM MLP, 인공신경망, 다항로짓모형 예측모델을 이용하여 분석한 결과 신경망 모형이 약 100%의 정확도를 보임
장재영 (2017)	예측변수를 정적/동적데이터를 구분하여 나이브 베이

	지안과 신경망기법 적용하여 분석한 결과 정적/동적데이터를 모두 포함한 신경망이 68% 가장 높음
권신혜 et al. (2017)	제작 및 투자, 배급, 상영단계별 모형을 구성하여 모든 변수를 사용한 모델과 회귀분석을 통해 유의한 변수들만을 사용한 모델로 구성한 후 의사결정트리와 인공신경망을 이용해 실험한 결과 제작 및 투자, 배급 모형에서는 모든 변수를 사용하였을 때는 인공신경망이 더 높게 나왔으며, 회귀분석을 이용하여 유의미한 결과를 나타낸 변수를 활용하였을 경우에는 의사결정트리가 더 높은 정확도를 보였고, 상영 모형에서는 둘 다 인공신경망의 정확도가 높은 결과를 보임
RU, Yunian, et al. (2018)	시간 구성 요소와 정적인 영화 특성 구성 요소를 사용하여 일일 박스오피스 예측한 결과 시간 구성 요소만을 사용하였을 때는 예측 정확도가 37% 정적 인 요소를 추가하였을 경우 예측정확도가 7% 높아지는 결과를 나타냄

3. 온라인구전에 관련한 연구

모바일 기술의 발달과 소셜네트워크서비스(SNS)의 활성화로 인해 온라인구전(eWOM)의 양은 폭발적으로 증가하고 있고 영향력도 점차 커지고 있다. 이시내와 이경렬(2013)은 SNS 이용자들은 SNS 의견 전달을 가장 많이 하고 그 다음으로 의견 추구, 의견 제공 순으로 이용한다고 밝혔다. SNS 이용자들의 증가로 인해 영화 홍행요인 분석대상으로 SNS 비정형 데이터 분석이 최근 들어 많은 연구가 진행되어 오고 있는데 이오준 et al. (2014) 에서는 관객들의 영화 선택 근거를 분석하기 위해 영화<변호인>의 트위터 데이터를 이용하여 빈번하게 언급된 명사들을 추출하여 토픽으로 정의하고 토픽분석을 수행하였다. 그 결과 관객들은 영화의 스토리, 배우, 감독의 연출력이 영

화 선택의 동기가 됨을 밝히고 이러한 속성이 흥행에 영향을 미치는 중요한 요인임을 제시하였다.

전성현과 손영숙(2016)은 2012년부터 2015년까지 국내개봉 영화 중 총 관객수가 50만 이상인 276편의 영화를 대상으로 온라인 구전의 효과를 알아보기 위해 통계분석을 수행하였다. 그 결과 개봉 후 평가자 수, 개봉 후 뉴스 수 등의 온라인 구전의 크기를 나타내는 변수들이 영화의 제작 및 배급정보들 보다 더욱 연관성 있다는 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.

<표 2-3> 온라인구전에 관련한 연구

저자	연구결과
이오준 et al. (2014)	영화<변호인>의 트위터 데이터를 이용하여 토픽분석을 수행한 결과 관객들은 영화의 스토리, 배우, 감독의 연출력이 영화 선택의 동기가 되고 흥행에 영향을 미치는 중요한 요인임을 밝힘
전성현 and 손영숙 (2016)	2012년부터 2015년까지 국내개봉 영화 276편을 대상으로 통계분석을 수행한 결과 개봉 후 평가자수, 개봉 후 뉴스 수 등의 온라인 구전 변수들이 영화의 제작 및 배급정보들 보다 더욱 연관성 있다는 결론을 도출함

박승현 and 송현주 (2012)	2010년 개봉한 영화 68편을 대상으로 온라인 구전이 흥행에 미치는 영향을 개봉 7주차까지의 주별 흥행성과와 전체 흥행성과로 나누어 연구한 결과 온라인 구전의 빈도는 상영이 끝나는 시점까지 지속적인 영향을 미쳤으나 평점은 개봉 초기에만 유의미한 영향을 미친다는 결론을 도출함

Ⅲ. 데이터 설명

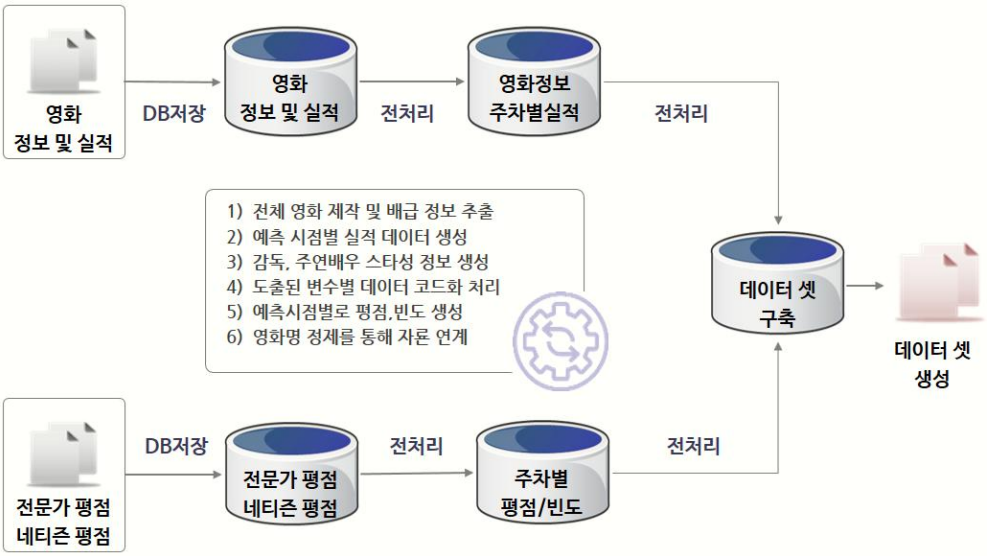
1. 데이터 수집

본 연구에서는 영화 흥행 예측을 위해 영화의 제작 및 배급 정보와 영화의 일별 실적은 영화진흥위원회에서 운영하는 영화 입장권 통합전산망 웹사이트에서 제공하는 API를 이용하여 수집하였다. 영화진흥위원회를 통해 수집된 데이터는 영화 일별 실적, 영화 월별 실적, 영화 년도별 통계, 영화인 정보, 영화 상세정보 등이다. 온라인 구전인 전문가 평점과 네티즌 개봉 전/후의 평점 정보는 국내 최대 포털 사이트인 네이버의 영화 섹션에서 일자별로 수집하였다. 분석대상 건 2015년부터 2017년까지 3년 동안 개봉한 영화를 대상으로하였다.

2. 데이터 전처리

본 연구에서는 2015년부터 2017년까지 3년 동안 개봉한 영화를 분석 대상으로 하였는데 영화진흥위원회에서 수집된 데이터는 총 4,670건 이었다. 데이터 전처리 과정은〔그림 3-1〕과 같다.

영화진흥위원회에서 수집한 데이터와 네이버에서 수집된 데이터를 MySQL DBMS에 저장하고 전처리 과정을 진행 한 후 데이터 분석을 통해 최종 분석 데이터 셋을 구축하였다. 전처리 과정을 살펴보면 영화진흥위원에 수집된 데이터를 기본으로 영화의 기본 정보인 배급 정보와 제작 정보를 활용해 데이터를 생성하였고 일자별 실적으로 통해 개봉일 실적을 추출하였으며 월별



〔그림 3-1〕데이터 전처리

실적을 기준으로 주차별 실적을 추출하고 주차별 집계실적을 생성하였다. 선행연구에서 사용되었던 흥행요인들은 참고문헌들을 중심으로 데이터를 범주화하는 전처리 과정을 거쳤다. 네이버 영화 섹션에서 수집된 전문가 및 네티즌 평가 데이터는 전문가 평점과 네티즌 개봉 전 평점은 각 평가자 수 즉 빈도와 평점의 평균값을 생성하였으며. 네티즌 개봉 후 평점은 일자별로 수집된 데이터를 영화정보와 동일하게 개봉일 평가 빈도와 평점의 평균 및 주차별 평가 빈도와 평균 평점을 계산하여 생성하였다. 영화진흥위원회에서 수집된 데이터와 네이버 영화 섹션에서 수집된 데이터 연계를 위해 분석한 결과 각각 사용되는 영화 코드 체계가 다르고 영화제목 표기 방식도 달라 영화진흥위원회의 영화 제목을 기준으로 네이버의 영화 제목을 정제하였다. 정제된 영화 제목을 통해 각각의 데이터를 연계하여 하나의 데이터로 생성하였다. 분석 대상 선정은 영화진흥위원회의 핵심상업영화군 기준이 ‘최대 개봉관 수 300개관 이상이거나 순 제작비 30억 원 이상’ 임을 고려하여 개봉관 수 300개관 이상인 데이터를 추출하였고 주차별 예측을 위해 영화들의

상영 기간이 3주 이상인 영화만을 추출하였다. 영화진흥위원회에서 제작비는 공개하지 않기 때문에 분석대상 조건에서 제외하였다.

<표 3-1>은 분석대상으로 선정된 영화를 년도별 상영기간을 나타낸 표로 2015년은 66편 2016년은 72년 2017년은 73편으로 총 211편이다. 2015년 ~ 2016년 까지 국내에서 개봉한 영화들이 4,670여편 정도였으나 전처리과정 및 분석을 위한 데이터를 추출한 결과 활용할 수 있는 분석데이터는 많지 않은 것으로 나타났다.

<표 3-1> 년도별 영화 수

년도	2015	2016	2017	합계
3 주 상영	25	28	31	84
4 주 이상 상영	41	44	42	127
합계	66	72	73	211

<표 3-2>는 주차별로 개봉스크린이 300개 이상인 영화를 대상으로 전체 관람객 수 기준 영화 상영 기간을 주단위로 나타낸 표다. 주차별 상영 영화들을 보면 전체 관람객 수 백만명 이상을 동원한 영화들이 4주 이상 상영한 것을 확인 할 수 있다.

<표 3-2> 전체 관람객 수 기준 영화 수

전체 관람객 수 (백만)	~ 0.5	0.5 ~ 1	1 ~ 3	3 ~	합계
3주 상영	19	29	35	1	84
4주 이상 상영	2	11	54	60	127
합 계	21	40	89	61	211

분석 대상 211편을 대상으로 구축된 최종 데이터 셋은 <표 3-3>에 나타난 것처럼 개봉시점을 기준으로 개봉 후에도 변하지 않는 영화의 속성정보 즉 제작 및 배급 단계의 정보와 전문가 평점, 개봉 전 네티즌 평점 정보와 개봉 후 변화하는 다양한 흥행실적 데이터 등을 주차별로 생성하여 데이터 세트를 구성하였다.

<표 3-3> 분석 데이터

구분	변수 명	변수ID	개봉일	주차(t=1,2,3)
변하지 않는 요인 (정적 변수)	배급사	DISTCD	O	O
	개봉월	OPENMM	O	O
	성수기	PEAKYN	O	O
	제작국가	NATION	O	O

	장르	GENRECD	O	O
	관람등급	WATCHGROUP	O	O
	감독 스타성	D_STAR	O	O
	배우 스타성	A_STAR	O	O
	전문가 평가 수	SPECIAL_CNT	O	O
	전문가 평점	SPECIAL_GRADE	O	O
	개봉 전 네티즌 평가 수	NET_BF_CNT	O	O
	개봉 전 네티즌 평점	NET_BF_GRADE	O	O
변하는 요인 (동적 변수)	매출 점유율	SALESSHARE	O	O
	관람객 수	AUDICNT	O	O
	스크린 수	SCRNCNT	O	O
	상영횟수	SHHOWCNT	O	O

	순위	RANK	O	O
	순위 증감여부	RANKID		O
	순위 증감 폭	RANKINTEN		O
	개봉 후 네티즌 평가 수	NET_AF_CNT	O	O
	개봉 후 네티즌 평점	NET_AF_GRADE	O	O
예측 클래스	차주 누적 관람객 수 (사분위 수)	WEEK_4_T	O	O
	차주 누적 관람객 수 (십분위 수)	WEEK_10_T	O	O
	총 관람객 수	TOT_CNT	O	O

3. 데이터 설명

영화진흥위원회의 핵심상업영화군 기준인 ‘최대 개봉관 수 300개관 이상 이거나 순 제작비 30억 원 이상’으로 대상을 선정하려고 했으나 제작사들이 제작비를 공개를 하지 않기 때문에 정확한 대상을 선정할 수가 없어 개

봉스크린이 300개 이상인 영화와 주차별 예측을 위해 3주 이상 상영한 영화를 분석대상으로 선정하였다.

본 연구에서는 개봉 일, 개봉 1주, 개봉 2주에는 사분위와 십분위로 범주화한 차주 누적 관람객 수와 전체 관람객 수를 예측하고 개봉 3주에는 전체 관람객 수를 예측하였다. 모델에 사용한 요인들은 영화 개봉을 기점으로 변하는 요인과 변하지 않는 요인으로 나눌 수 있다. 변하지 않는 요인으로는 영화의 제작 단계와 배급 단계에서 알 수 있는 요인들과 영화 개봉 전 알 수 있는 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 전체 관람객 수를 범주화 하였다. 배우 스타성도 감독과 마찬가지로 2010년도 이후부터 분석대상 작품 전까지 출연한 작품의 평균 전체 관람객 수를 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 수직결합 배급사인 씨제이이앤엠(주), 롯데쇼핑㈜롯데엔터테인먼트와 자사영화관이 없는 비수직결합 배급사인 (주)쇼박스, (주)넥스트엔터테인먼트월드(NEW), 해외영화를 직접 배급하는 유니버셜픽쳐스 인터내셔널, 월트 디즈니, 워너 브라더스로 재분류 하였다(최성희, 2017).

개봉 월은 개봉일자에서 개봉 월을 추출하였다. 성수기의 경우 우리나라는 일반적으로 크리스마스와 명절연휴, 그리고 여름방학이 포함된 12월, 1월, 7

월, 8월을 성수기로 정의한다(강선주, 2017). 본 연구에서는 2015년 ~ 2017년도의 설날과 추석을 확인하여 12월, 1월, 7월, 8월외 명절이 있는 달을 성수기로 정의하였다.

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(강선주, 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, 순위변경폭등을 예측요인으로 사용하였다. 또 실시간으로 변화하는 네티즌의 평점 정보 역시 예측 요인으로 사용하였다. 네티즌 평점 정보의 평균 평점과 네티즌 평가 수를 예측 요인으로 도출하였다.

예측 시점은 개봉 일, 개봉 1주, 개봉 2주, 개봉 3주로 총 네 번의 시점에서 실험을 하기 때문에 각 예측 시점마다 스크린 수, 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 해당 주차 기간 점유율을 평균으로 표현하였다. 관람객 수와 스크린 수는 예측시점까지의 누적값으로 표현하였으며 순위

역시 해당 주의 평균 순위로 표현하였다. 개봉일의 경우 순위증감여부와 순위변경값은 확인할 수 없기 때문에 사용하지 않고 주차별 예측에서는 지난 주 실적과 비교하여 증가, 동일, 감소의 세가지로 순위 증감을 표현하였으며 순위 변경 값은 지난 주와의 차이값을 이용하였다. 개봉 후의 네티즌 평가자 수는 개봉일부터 예측시점까지의 누적 평가자 수이고 평점은 평균평점으로 표현하였다. 최종적으로 예측 변수를 정리한 내용은 <표 3-4>와 같다.

<표 3-4> 변수 내용

구분	예측요인	내용
	감독 스타성	2010년 이후 감독한 영화의 평균 총 관람객 수
	배우 스타성	2010년 이후 출연한 영화의 평균 총 관람객 수
	배급사	메이저 배급사 (수직결합, 비수직결합, 직수입) 기타 배급사
	개봉월	개봉 월
	성수기	크리스마스, 명절,여름방학(12월,1월,7월,8월) 명절이 있는 월
	국가	한국, 미국, 기타
	장르	사극, 액션/범죄/스릴러, 드라마, SF/어드벤쳐/판타지, 전쟁, 기타
	등급	전체 관람가, 12세 이상 관람가, 15세 이상 관람가, 청소년관람불가

	전문가 평점	평균 평점
	전문가 평가 수	빈도
	개봉 전 네티즌 평점	평균 평점
	개봉 전 네티즌 평가 수	빈도
	관람객 수	주간 동안의 누적 관람객 수
	스크린 수	주간 동안의 누적 스크린 수
	상영횟수	주간 동안의 누적 상영횟수
	네티즌 평점	평균 평점
	네티즌 평가 수	빈도
	매출 점유율	주간 동안의 평균 매출 점유율
	흥행순위(관람객)	주간동안의 평균 순위
	순위 증감 구분	지난 주 실적과 비교하여 상승/동일/하락
	순위 증감분	지난 주 실적과 비교하여 증감폭

예측 대상인 주차별 누적 관람객 수는 수치형으로 제공되기 때문에 수집된 영화의 주차별 누적 관람객 수의 데이터를 사분위와 십분위로 범주화하여 등급을 나누었다. 주차별 관람객 수는 누적 값을 이용하였으며 예측 시점마다 예측 데이터가 변하기 때문에 주차별 등급의 데이터 범주는 서로 다르다. 주차별 관람객 수는 <표 3-5>와 <표 3-6> 처럼 구성하여 실험에 사용하였다.

<표 3-5> 타겟 클래스 정의1 (주차별 관람객 수_사분위)

타겟 클래스	개봉 1주 후 누적 관람객 수	개봉 2주 후 누적 관람객 수	개봉 3주 후 누적 관람객 수
A	~ 747,921.8	~ 1,322,505	~ 1,365,492
B	~ 1,375,140	~ 2,537,592	~ 2,901,012
C	~ 2,471,757	~ 4,381,621	~ 5,266,025
D	2,471,757 ~	4,381,621 ~	5,266,025 ~

<표 3-6> 타겟 클래스 정의2 (주차별 관람객 수_십분위)

타겟 클래스	개봉 1주 후 누적 관람객 수	개봉 2주 후 누적 관람객 수	개봉 3주 후 누적 관람객 수
A	~ 173,997	~ 609,170	~ 686,516
B	~ 257,530	~ 934,387	~ 1,022,447

C	~ 349,632	~ 1,349,035	~ 1,508,374
D	~ 456,177	~ 1,741,250	~ 1,965,940
E	~ 583,340	~ 2,246,916	~ 2,752,144
F	~ 708,407	~ 2,763,246	~ 3,332,560
G	~ 840,405	~ 3,465,038	~ 4,067,262
H	~ 1,128,396	~ 4,304,720	~ 5,276,493
I	~ 1,417,061	~ 6,044,317	~ 7,804,388
J	~ 3,216,109	~ 12,545,377	~ 15,910,462

주차별 관람객 수 예측률과 비교하기 위한 또 다른 예측 대상인 전체 관람객 수 역시 수치형으로 제공되기 때문에 여러 등급으로 나누었다. 주차별 관람객 수와 달리 전체 관람객 수 범주의 기준은 임준엽 and 황병언(2014)의 연구를 참고하였다. 이들의 연구에서는 5개의 등급으로 나누었으나 영화 편수의 편차가 커서 본 연구에서는 <표 3-7>과 같이 4개의 등급으로 나누어 실험에 사용하였다.

<표 3-7> 타겟 클래스 정의3 (전체 관람객 수)

타겟 클래스	A	B	C	D
전체 관람객수 (백만)	~ 0.5	0.5 ~ 1	1 ~ 3	3 ~
영화 편 수	21	40	89	61

Ⅳ. 예측모델 생성

1. 모델 생성 방법

본 연구에서는 예측 모델 생성을 위해 기계학습의 지도학습 분류기법을 활용하였다. 지도학습의 다양한 분류기법 중 Naive Bayes, Random Forest, Support Vector Machine(SVM), Multi Layer Perceptron(MLP)을 이용하여 평

가하였다. 예측 요인은 <표 3-1>에 표현된 것처럼 예측시점이 뒤로 갈수록 예측 요인의 개수는 늘어난다. 예를 들어 예측 시점이 개봉 2주 후 일 경우 개봉일, 개봉 1주일 실적을 모두 활용하여 예측을 수행하였다. 예측모델은 <표 4-1>에 표현된 것처럼 예측시점마다 사분위와 십분위로 범주화 된 다음 주 누적 관람객 수도 예측하지만 전체 관람객 수도 함께 예측해 정확도를 비교하였다. 전체 관람객 수는 분석대상이 개봉 후 3주이상 상영한 영화이기 때문에 개봉 4주차 실적이 없는 데이터들이 있어 개봉 3주 후에는 전체 관람객 수를 예측한다. 예측을 위한 분석 도구로는 WEKA를 사용하였다. 예측모델의 신뢰성을 높이기 위해 Naive Bayes, Random Forest, Support Vector Machine(SVM), Multi Layer Perception(MLP)의 분류기법에 k-fold cross-validation을 사용하였다. k-fold cross-validation은 2 ~ 10 회를 수행하였다.

<표 4-1> 예측모델

<table>
<tr><th>예측시점</th><th>예측요인</th><th>타겟 클래스</th></tr>
<tr><td rowspan="3">개봉일 이후</td><td rowspan="3">영화기본속성
(변하지 않는요인)
개봉일 실적</td><td>개봉 1주일 후
누적관람객수
(사분위)</td></tr>
<tr><td>개봉 1주일 후
누적관람객수
(십분위)</td></tr>
<tr><td>전체 관람객 수</td></tr>
<tr><td rowspan="2">개봉 1주일 후</td><td rowspan="2">영화기본속성
(변하지 않는요인)
개봉일 실적
개봉 1주일 후 실적</td><td>개봉 2주일 후
누적관람객수
(사분위)</td></tr>
<tr><td>개봉 2주일 후
누적관람객수</td></tr>
</table>

		(십분위)
		전체 관람객 수
개봉 2주일 후	영화기본속성 (변하지 않는요인) 개봉일 실적 개봉 1주일 후 실적 개봉 2주일 후 실적	개봉 3주일 후 누적관람객수 (사분위)
		개봉 3주일 후 누적관람객수 (십분위)
		전체 관람객 수
개봉 3주일 후	개봉 3주일 후 실적	전체 관람객 수

2. 모델 비교 방법

본 연구에서는 네가지 분석기법 Naive Bayes, Random Forest, Support Vector Machine(SVM), Multi Layer Perception(MLP) 을 이용하여 개봉 일 이후, 개봉 1주 후, 개봉 2주 후, 개봉 3주 후 네번의 예측 시점에 사분위와 십분위 범주화로 구성된 차주 누적 관람객 수와 최종 관람객 수를 예측한다. 신뢰도를 높이기 위해 k-fold cross validation을 2 ~ 10회를 수행하여 모든 실험 결과를 비교한다. 각 예측 시점 별로 가장 예측 정확도가 높은 모델을 찾고자 한다.

<표 4-2> 실험 순서 및 모델 비교 방법

No	내용	비교 방법
1	차주 누적 관람객수 예측	사분위와 십분위로 구성된 분류 클래스 중 정확도가 높은 모델 선택
2	전체 관람객 수 예측	차주 누적관람객 수와 전체 관람객 수와의 정확도 비교
3	신규 요인 효과	신규 요인을 포함한 모델과 포함하지 않은 모델의 정확도 비교
4	모델 선정	정확도가 가장 높게 나온 예측모델 선정
5	모델 고도화	변수선택법을 활용하여 모델의 정확도를 높임

3. 모델 성과 검증지표

본 연구에서는 네가지 분석기법 Naive Bayes, Random Forest, Support Vector Machine(SVM), Multi Layer Perception(MLP) 을 이용하여 개봉 일 이후, 개봉 1주 후, 개봉 2주 후, 개봉 3주후 네 번의 예측 시점에 사분위와 십분위 범주화로 구성된 차주 누적 관람객 수와 최종 관람객 수를 k-fold cross validation을 2 ~ 10회를 수행한다. 각각의 분석 기법과 예측 시점 중 차주 누적 관람객 수, 최종 관람객 수 예측 중 가장 높은 정확도를 나타낸 k-fold 횟수를 이용해 최종 모델로 선정하고 각 분석기법별로 예측 정확도를 이용해 모델의 성능을 평가하고자 한다. 본 연구의 모델별 성과 검증지표로

서 구현된 분석기법들의 가장 높은 예측 정확도를 이용하고 가장 높게 나온 분석기법을 선택 한 후 변수 선택법들을 이용해 새로 도출된 변수들의 성능을 검증할 수 있도록 한다. 모델들의 성과 지표로 정확도(Precision)를 사용하였으며 혼동행렬(Confusion matrix)을 활용하여 정확도(Precision), 재현율(Recall)도 같이 확인 하였다. Confusion Matrix는 분류 모델을 평가하는 지표로 사용되며 <표 4-3>과 같다. True와 False의 값은 실제 값을 맞춘 것을 나타내는데 True는 실제 값과 예측 값이 일치한 경우를 나타내고 False는 실제 값과 예측 값이 일치한 경우를 나타낸다. Positive와 Negative는 예측된 값을 나타낸다. <표 4-3>처럼 True Positive(TP) 와 True Negative (TN)은 예측 값과 실제 값이 일치한 경우를 나타내며 False Positive(FP)와 False Negative(FN)은 예측 값과 실제 값이 불일치하는 경우를 나타낸다.

<표 4-3> 혼동 행렬 (Confusion Matrix)

		예측 값	
		Positive	Negative
실제 값	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive	True Negative

		(FP)	(TN)

<표 4-4>는 각 지표들의 설명이다. 정확도(Accuracy)는 전체 데이터 중에서, 정확하게 예측한 데이터의 비율이고, 정밀도(Precision)는 Positive로 예측한 값 중, 실제 Positive인 값의 비율을 나타내며, Precision = TP / (TP + FP)로 구한다. 재현율(Recall)은 실제 Positive인 값들 중 Positive로 예측된 값들의 비율로, Recall = TP / (TP + FN) 이다. 정확도와 정밀도는 반비례한다. 마지막으로 F1-score는 Precision과 Recall의 조화 평균을 이용하여 2개를 모두 고려해서 평가하게 된다.

<표 4-4> 혼동 행렬 분석지표

	설명	수식
정확도 (Accuracy)	전체 데이터 중 정확하게 예측한 데이터의 비율	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$
정밀도 (Precision)	Positive로 예측한 값 중, 실제 Positive인 값의 비율	$\frac{(TP)}{(TP+FP)}$
재현율 (Recall)	실제 Positive인 값들 중 Positive로 예측된 값들의 비율	$\frac{(TP)}{(TP+FN)}$
F1-Score	Precision과 Recall의 조화 평균을 이용하여 2개를 모두 고려해서 평가	$\frac{\mathrm{Precison} \times Recall}{\mathrm{Precison} + Recall} \times 2$

Ⅴ. 실험결과

1. 모델별 예측결과

(1) 차주 누적 관람객 수 예측 결과

1) Naïve Bayes 예측 정확도 결과

본 실험을 통해 Naïve Bayes의 결과는 <표 5-1>과 같다. 전체적으로 10-fold 에서 높은 정확도를 보이며 주차별 누적 관람객 수는 사분위 범주가 십분위 보다 높은 정확도를 보였다. 예측 시점 별로는 예측 시점이 뒤로 갈수록 정확도가 높아지는 것으로 나타낸다.

<표 5-1> Naïve Bayes의 k-fold 결과

예측 시점	타겟 변수	k - fold								
		2	3	4	5	6	7	8	9	10
개봉 일	1주차 누적 관람객 수 사분위	63.51	63.51	62.56	66.82	66.82	67.30	63.98	66.82	67.77
	1주차 누적 관람객 수 십분위	35.07	35.55	34.60	36.97	36.49	37.91	32.23	36.02	35.55
개봉 1주 후	2주차 누적 관람객 수 사분위	68.72	70.14	70.62	69.67	69.67	71.56	71.09	71.56	70.14
	2주차 누적 관람객 수 십분위	40.28	40.76	42.65	43.13	40.28	42.18	41.71	40.76	40.28
개봉 2주 후	3주차 누적 관람객 수 사분위	77.25	76.30	79.15	75.83	76.78	76.78	76.78	77.25	76.78
	3주차 누적 관람객 수 십분위	53.08	51.18	49.76	52.13	54.98	54.50	50.71	52.61	53.08
개봉 3주	전체 관람객	77.73	74.41	76.78	75.36	77.25	76.30	76.78	75.83	75.83

후	수									

2) Random Forest 예측 정확도 결과

Random Forest의 결과는 <표 5-2>와 같다. 전체적으로 10-fold 에서 높은 정확도를 보이며 주차별 누적 관람객 수는 사분위 범주가 십분위 보다 높은 정확도를 보였다. 예측 시점 별로는 예측 시점이 뒤로 갈수록 정확도가 높아지는 것으로 나타낸다.

<표 5-2> Random Forest의 k-fold 결과

예측 시점	타겟 변수	k - fold								
		2	3	4	5	6	7	8	9	10
개봉일	1주차 누적 관람객 수 사분위	73.93	70.62	71.56	73.93	72.99	75.83	74.41	75.36	73.93
	1주차 누적 관람객 수 십분위	27.49	35.55	34.12	35.55	31.75	36.97	32.70	37.91	37.91
개봉 1주 후	2주차 누적 관람객 수 사분위	80.57	79.15	80.57	78.20	79.62	82.94	79.62	80.09	83.89
	2주차 누적 관람객 수 십분위	13.17	15.84	17.79	20.24	16.99	17.39	17.39	17.19	16.80
개봉 2주 후	3주차 누적 관람객 수	83.89	87.20	86.26	88.63	87.68	88.63	88.15	88.15	88.63

	사분위									
	3주차 누적 관람객 수 십분위	52.61	58.29	56.40	63.03	61.14	59.24	61.61	57.82	60.66
개봉 3주 후	전체 관람객 수	81.99	82.46	85.31	84.36	84.36	87.20	87.20	87.20	87.20

3) Support Vector Machine (SVM) 예측 정확도 결과

Support Vector Machine의 결과는 <표 5-3>와 같다. 전체적으로 10-fold 에서 높은 정확도를 보이며 주차별 누적 관람객 수는 사분위 범주가 십분위 보다 높은 정확도를 보였다. 예측 시점 별로는 예측 시점이 뒤로 갈수록 정확도가 높아지는 것으로 나타낸다.

<표 5-3> Support Vector Machine (SVM) 의 k-fold 결과

예측 시점	타겟 변수	k - fold								
		2	3	4	5	6	7	8	9	10
개봉 일	1주차 누적 관람객 수 사분위	43.13	56.40	55.92	57.35	56.40	58.29	54.50	54.03	61.14
	1주차 누적 관람객 수 십분위	20.38	18.48	20.85	19.91	18.01	19.43	18.48	19.91	20.38
개봉 1주 후	2주차 누적 관람객 수 사분위	53.55	57.35	52.61	54.03	57.82	61.61	55.45	53.08	54.50
	2주차 누적 관	22.75	23.70	19.91	24.17	21.80	25.12	26.07	26.07	27.01

	람객 수 십분위									
개봉 2주 후	3주차 누적 관람객 수 사분위	59.72	63.51	62.09	68.25	64.93	61.14	63.03	63.98	66.82
	3주차 누적 관람객 수 십분위	23.70	30.33	26.54	29.86	31.75	28.44	26.07	29.86	28.91
개봉 3주 후	전체 관람객 수	67.77	70.14	70.14	72.99	73.46	68.72	72.04	69.19	72.51

4) Multi Layer Perception (MLP) 예측 정확도 결과

Multi Layer Perception의 결과는 <표 5-4>와 같다. 전체적으로 10-fold 에서 높은 정확도를 보이며 주차별 누적 관람객 수는 사분위 범주가 십분위 보다 높은 정확도를 보였다. 예측 시점 별로는 예측 시점이 뒤로 갈수록 정확도가 높아지는 것으로 나타낸다.

<표 5-4> Multi Layer Perception (MLP) 의 k-fold 결과

예측 시점	타겟 변수	k - fold								
		2	3	4	5	6	7	8	9	10
개봉 일	1주차 누적 관람객 수 사분위	48.34	63.03	56.40	55.45	55.45	55.92	52.61	54.98	58.77
	1주차 누적	20.85	19.43	18.96	19.91	23.22	22.27	19.43	24.17	21.80

	관람객 수 십분위									
개봉 1주 후	2주차 누적 관람객 수 사분위	54.03	59.24	50.24	52.61	52.61	58.29	55.92	56.87	52.61
	2주차 누적 관람객 수 십분위	20.38	25.59	24.17	27.49	28.91	28.91	26.07	24.17	24.17
개봉 2주 후	3주차 누적 관람객 수 사분위	58.29	66.35	62.09	65.40	64.46	63.51	64.46	63.51	63.03
	3주차 누적 관람객 수 십분위	24.64	30.81	25.12	25.12	29.38	27.96	27.01	30.33	27.96
개봉 3주 후	전체 관람객 수	63.98	68.72	70.62	71.09	69.67	71.56	70.62	66.82	71.09

(2) 전체 관람객수 예측 결과

1) Naïve Bayes 예측 정확도 결과

본 실험을 통해 Naïve Bayes의 결과는 <표 5-5>와 같다. 전체적으로 10-fold 에서 높은 정확도를 보이며 예측 시점 별로는 예측 시점이 뒤로 갈수록 정확도가 높아지는 것으로 나타낸다.

<표 5-5> Naïve Bayes의 k-fold 결과

예측 시점	타겟 변수	k - fold								
		2	3	4	5	6	7	8	9	10
개봉일	전체 관람객수	61.14	60.66	61.14	66.82	62.56	62.09	61.14	59.24	61.61
개봉 1주 후	전체 관람객수	69.19	67.30	65.88	67.77	66.82	67.77	67.77	66.82	67.77
개봉 2주 후	전체 관람객수	73.93	76.30	81.04	76.20	80.57	78.20	78.20	80.09	78.20
개봉 3주 후	전체 관람객수	81.99	82.46	85.31	84.36	84.36	87.20	87.20	87.20	87.20

2) Random Forest 예측 정확도 결과

Random Forest의 결과는 <표 5-6>과 같다. 전체적으로 10-fold 에서 높은 정확도를 보이며 예측 시점 별로는 예측 시점이 뒤로 갈수록 정확도가 높아지는 것으로 나타낸다.

<표 5-6> Random Forest의 k-fold 결과

예측 시점	타겟 변수	k - fold								
		2	3	4	5	6	7	8	9	10
개봉일	전체 관람객수	61.14	60.66	61.14	66.82	62.56	62.09	61.14	59.24	61.61
개봉	전체	69.19	67.30	65.88	67.77	66.82	67.77	67.77	66.82	67.77

1주 후	관람객 수									
개봉 2주 후	전체 관람객 수	73.93	76.30	81.04	78.20	80.57	78.20	78.20	80.09	78.20
개봉 3주 후	전체 관람객 수	81.99	82.46	85.31	84.36	84.36	87.20	87.20	87.20	87.20

3) Support Vector Machine (SVM) 예측 정확도 결과

Support Vector Machine의 결과는 <표 5-7>과 같다. 전체적으로 10-fold 에서 높은 정확도를 보이며 예측 시점 별로는 예측 시점이 뒤로 갈수록 정확도가 높아지는 것으로 나타낸다.

<표 5-7> Support Vector Machine (SVM) 의 k-fold 결과

예측 시점	타겟 변수	k - fold								
		2	3	4	5	6	7	8	9	10
개봉 일	전체 관람객 수	52.13	57.35	54.50	51.18	56.87	51.66	55.92	54.50	53.08
개봉 1주 후	전체 관람객 수	53.55	57.82	57.82	60.19	62.09	58.29	60.66	61.61	61.61
개봉 2주 후	전체 관람객 수	60.66	65.40	63.03	69.19	70.62	66.35	67.30	67.30	66.82
개봉 3주 후	전체 관람객 수	67.77	70.14	70.14	72.99	73.46	68.72	72.04	69.19	72.51

4) Multi Layer Perception (MLP) 예측 정확도 결과

Multi Layer Perception의 결과는 <표 5-8>과 같다. 전체적으로 10-fold 에서 높은 정확도를 보이며 예측 시점 별로는 예측 시점이 뒤로 갈수록 정확도가 높아지는 것으로 나타낸다

<표 5-8> Multi Layer Perception (MLP) 의 k-fold 결과

예측 시점	타겟 변수	k - fold								
		2	3	4	5	6	7	8	9	10
개봉일	전체 관람객 수	47.39	51.66	52.13	50.71	54.50	56.87	52.13	51.66	55.45
개봉 1주 후	전체 관람객 수	57.82	53.08	55.92	54.50	56.87	54.98	58.29	52.13	57.82
개봉 2주 후	전체 관람객 수	57.35	67.77	61.14	65.88	64.46	62.56	67.30	60.66	63.98
개봉 3주 후	전체 관람객 수	63.98	68.72	70.62	71.09	69.67	71.56	70.62	66.82	71.09

(3) 신규 요인 비교 실험결과

본 연구에서 선행연구에서 사용되었던 요인들 외에도 새로운 흥행요인을 추가하였는데 추가한 변수는 매출 점유율, 흥행 순위, 순위 증감 구분, 순위 증감 분으로 전체 요인은 <표5-9>와 같다. 신규 요인들은 예측 시점별로 각각 생성하여 적용하였다.

<표 5-9> 예측요인 구분

구분	예측요인	기존 요인	신규 요인
	감독 스타성	0	
	배우 스타성	0	
	배급사	0	
	개봉월	0	
	성수기	0	
	국가	0	
	장르	0	
	등급	0	
	전문가 평점	0	
	전문가 평가 수	0	
	개봉 전 네티즌 평점	0	
	개봉 전 네티즌 평가 수	0	
	관람객 수	0	
	스크린 수	0	
	상영횟수	0	
	네티즌 평점	0	
	네티즌 평가 수	0	
	매출 점유율		0
	흥행순위(관람객)		0
	순위 증감 구분		0
	순위 증감분		0

새롭게 추가한 요인의 흥행요인 가능성을 알아보기 위해 신규 요인을 포함한 전체 요인을 적용한 모델과 신규 요인를 제외한 모델을 생성하여 실험하였다. 차주 예측은 앞의 차주 누적 관람객 수 예측 결과에서 알 수 있듯이 사분위로 클래스를 적용했을때와 십분위로 클래스를 적용한 실험을 비교했을 때 사분위로 클래스를 나눈 실험의 예측정확도가 훨씬 높게 나왔다. 신규 요인을 포함한 모델과 신규 요인을 제외한 모델을 구축하여 사분위로 클래스를 나눈 데이터를 이용해 차주 관람객 수 와 전체 관람객 수를 예측하여 비교 하였다.

실험결과는 <표 5-10>과 같으며 전체적으로 신규 요인을 포함한 모델이 전체적으로 예측정확도가 높은 결과를 나타냈다. 개봉 시점이 뒤로 갈수록 예측 시점이 뒤로 갈수록 신규 요인을 포함한 모델과 제외한 모델 모두 예측정확도가 높아졌으며 개봉 3주차에 전체 관람객 수를 예측 했을 경우 가 가장 높은 예측정확도를 보였다. 이 실험을 통해 신규 요인의 흥행 가능성을 확인 할 수 있었다.

<표 5-10> 신규요인 비교 결과

		Naive-Bayes		MLP		SVM		Random Forest	
		포함	**제외**	**포함**	**제외**	**포함**	**제외**	**포함**	**제외**
개봉일	전체 관람객 수	62.08	58.77	50.23	43.6	53.55	49.29	61.61	62.09
	1주차 누적 관람객 수	67.77	60.19	56.39	51.19	61.13	55.92	73.93	72.99

개봉 1주 후	전체 관람객 수	67.77	65.4	55.92	49.29	61.61	55.92	67.77	67.3
	2주차 누적 관람객 수	70.14	68.72	50.23	52.6	54.5	54.98	83.88	79.62
개봉 2주 후	전체 관람객 수	72.03	69.19	64.92	59.24	65.4	62.09	74.4	79.15
	3주차 누적 관람객 수	76.77	75.36	62.08	65.88	66.82	63.98	88.62	89.1
개봉 3주 후	전체 관람객 수	75.82	73.93	70.61	65.88	72.51	67.3	87.2	89.1

2. 모델별 성능비교

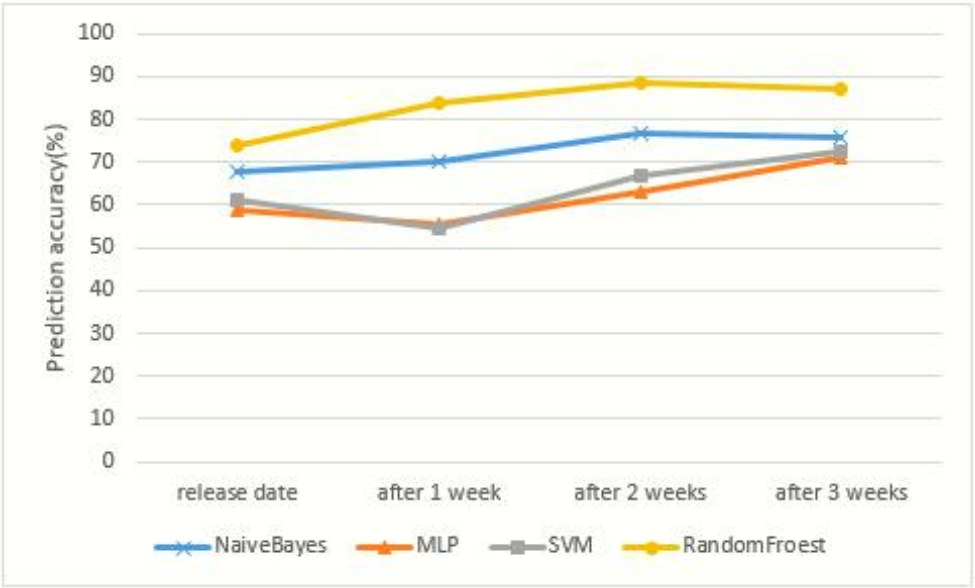
(1) 모델별 성능비교

예측시점별 결과는 전체적으로 k-fold cross validation의 결과는 10-fold 전체적으로 높게 나왔으며 신규요인을 포함한 전체요인을 이용하여 실험한 결과 MLP, SVM, Naive Bayes, Random Forest의 순서로 정확도가 높게 측정되었으며 예측시점마다 전체 관람객 수를 예측하는 것 보다 다음 주 누적 관람객 수를 예측한 것이 정확도가 높게 측정되었다. 각 분석기법별 10-fold의 결과와 사분위로 범주화한 차주 누적 관람객 수의 결과는 <표 5-10>과 같다.

<표 5-11> 예측시점별 예측 정확도 비교

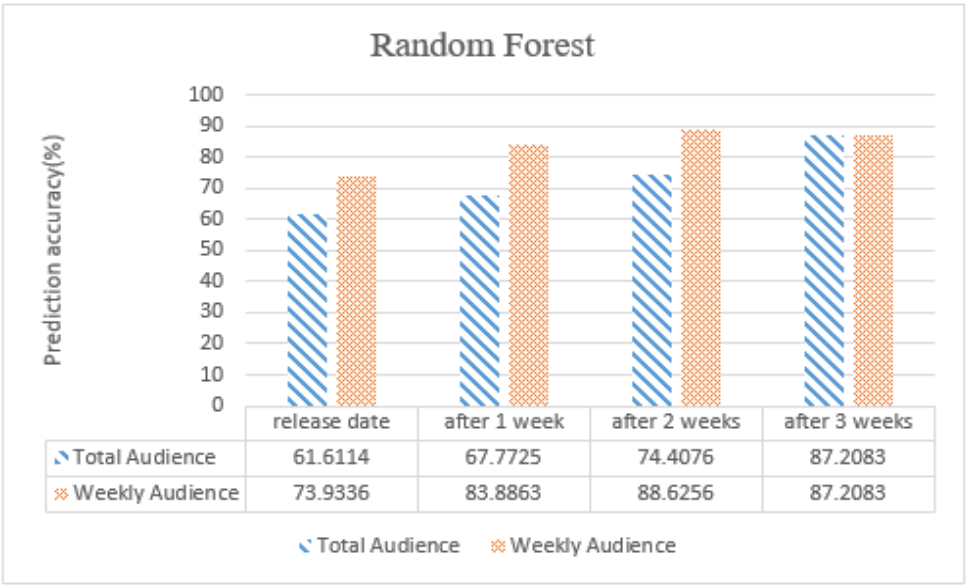
예측 시점	타겟 변수	Naive-Bayes	MLP	SVM	Random Forest
개봉일	전체 관람객 수	62.08	55.45	53.55	61.61
	개봉 1주 후 누적 관람객 수	67.77	58.77	61.13	73.93
개봉 1주 후	전체 관람객 수	67.77	57.82	61.61	67.77
	개봉 2주 후 누적 관람객 수	70.14	52.61	54.50	83.88
개봉 2주 후	전체 관람객 수	72.03	63.98	65.40	74.40
	개봉 3주 후 누적 관람객 수	76.77	63.03	66.82	88.62
개봉 3주 후	전체 관람객 수	75.82	71.39	72.51	87.20

예측 시점에 따라 주차별 누적 관람객 수를 예측한 결과를 나타낸 [그림 5-1]을 보면 전반적으로 시간이 지날수록 예측 정확도가 높아지는 것으로 나타났다. Random Forest가 73% ~ 87%로 가장 높게 측정되었고 Naive Bayes가 67% ~ 75%의 정확도를 보였다. MLP와 SVM의 경우 개봉일에 예측한 정확도 보다 개봉 1주 후에 예측 정확도가 떨어졌으며 이후에는 예측 정확도가 높아지는 유사한 패턴을 보여주고 있다.



〔그림 5-1〕 시점별 예측정확도 비교

본 실험에서 가장 높은 정확도를 나타낸 모델은 Random Forest로 대부분의 모든 예측시점에서 높은 예측 정확도를 나타냈다. Random Forest의 예측률을 〔그림 5-2]를 통해 구체적으로 살펴보면 시간이 흐를수록 전체 관람객 수를 예측했을 경우와 차주 누적 관람객 수를 예측하는 것 모두 점점 높아지는 것으로 측정되었다. 또한 대체적으로 차주 누적 관람객 수를 예측하는 것이 더 높은 정확도를 보였다. 비교 결과를 보면 각 시점마다 약 12% ~ 16%의 예측 정확도 차이를 보이고 있다.



〔그림 5-2〕 Random Forest의 예측 정확도

〔그림 5-3〕은 Random Forest의 confusion matrix 이다. 4*4의 matrix이며 4가지의 클래스 별로 정밀도(precision)와 재현율(recall)을 확인할 수 있다. 가장 높은 예측 정확도를 보인 개봉 2주후 3주차 누적 관람객 수 예측 결과를 보면 A클래스는 precision이 100이며 recall이 96.49 이다. 전체 클래스에 대한 평균 precision은 88이고 recall은 88.6로 측정되었다.

본 연구의 결과를 보면 역동적으로 변화하는 다양한 경쟁 요소를 활용해 Random Forest를 기법을 적용하여 차주 누적 관람객 수를 예측하는 것이 가장 높은 정확도를 보인다고 결론을 내릴 수 있다.

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	

after 1 week

	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	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	

	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	

〔그림 5-3〕 Random Forest의 Confusion Matrix

(2) 모델 고도화

본 연구에서 가장 높은 예측 정확도를 보인 Random Forest 분석기법을 적용하여 변수 선택법 중 GainRatio 와 InfoGain의 Ranker기법을 이용하

여 각 예측시점별로 사분위로 범주화한 차주 누적 관람객 수와 전체 관람객 수를 이용해 선택된 변수만을 사용하여 모델을 생성하였다. Random Forest는 데이터에 따라 학습 결과의 성능과 변동이 큰 의사결정 트리의 단점을 보완하고자 다수의 의사결정 트리를 하나의 모형으로 결합하여 생성하는 방법이다. 의사결정 트리의 학습 알고리즘은 <표 5-12>처럼 ID3 , C4.5, C5.0, CART, CHAID, QUEST, CRUISE 등등이 있으며 ID3을 보완하여 C4.5가 개발되었고, C4.5를 보완하여 C5.0이 개발되었다.

<표 5-12> 의사결정 트리 알고리즘

알고리즘	평가지수(선택방법)	비고
ID3	Entropy	다지분리(범주)
C4.5	Information gain	다지분리(범주) 및 이진분리(수치)
C5.0	Information gain	C4.5와 거의 유사(차이점)
CHAID	카이제곱(범주), F검정(수치)	통계적 접근 방식
CART	Gini index(범주), 분산의 차이(수치)	통계적 접근 방식, 항상 2진 분리

정보 이득(Information Gain)은 어떤 분류를 통해 정보(Information)에 대한 이득(Gain)이 있는지를 알아보는 것으로 엔트로피(Entropy)를 통해 계산된다. 엔트로피(Entropy)는 무질서의 정도를 뜻하며 엔트로피가 높을수록 다양한

데이터가 혼잡하다는 뜻으로 엔트로피는 작을수록 좋다. Information Gain(IG)은 상위 엔트로피에서 하위노드의 엔트로피를 뺀 값으로 클수록 좋다. 정보 이득 비율(Information Gain Ratio)은 정보 이득(Information Gain)값을 구한 다음 가지가 많은 것에 대한 패널티를 부여하여 정보 이득 비율(Information Gain Ratio)를 계산한 후 가장 높은 값을 갖는 속성을 선택 한다(Quinlan, J. R., 1987; Quinlan, J. R., 1993; Quinlan, J. R., 1997).

GainRatio 와 InfoGain 변수선택법을 이용하여 개봉일 이후, 개봉 1주 후, 개봉 2주 후, 개봉 3주 후 시점마다 사분위로 범주화된 차주 누적 관람객 수와 전체 관람객 수에 사용된 변수들을 선택한 결과 개봉일 이후 선택된 변수들은 〔그림 5-4〕와 같다. 선택되지 않은 변수들을 보면 개봉월, 전문가 평점, 개봉전 네티즌 평점들이 제외 되었다.

개봉 1주 후 변수들은 〔그림 5-5〕에 따르면 전문가 평점, 네티즌 평점, 개봉전 네티즌 평점, 개봉 월 등이 선택되지 않았다.

GainRatio (Ranker)	
최종관객	1주차
Ranked attributes:	Ranked attributes:
0.4933 21 AUDICNT_1W	0.6783 21 AUDICNT_1W
0.3812 14 AUDICNT_D	0.5254 14 AUDICNT_D
0.3545 22 SCRNCNT_1W	0.465 22 SCRNCNT_1W
0.3416 20 SALESSHARE_1W	0.4277 23 SHOWCNT_1W
0.3306 16 SHOWCNT_D	0.4258 27 NET_AF_CNT_1W
0.3284 13 SALESSHARE_D	0.3982 20 SALESSHARE_1W
0.3264 15 SCRNCNT_D	0.3918 13 SALESSHARE_D
0.3072 27 NET_AF_CNT_1W	0.372 16 SHOWCNT_D
0.305 23 SHOWCNT_1W	0.3343 15 SCRNCNT_D
0.2908 18 NET_AF_CNT_D	0.326 24 RANK_1W
0.2796 24 RANK_1W	0.3098 18 NET_AF_CNT_D
0.2604 17 RANK_D	0.2682 17 RANK_D
0.2075 9 SPECIAL_CNT	0.1598 9 SPECIAL_CNT
0.1315 11 NET_BF_CNT	0.156 4 NATION_G2
0.0766 5 GENRECD1	0.1415 11 NET_BF_CNT
0.0628 7 D_STAR_1	0.0834 5 GENRECD1
0.0466 6 WATCHGRADECD	0.0585 7 D_STAR_1
0.0444 8 A_STAR_1	0.0544 8 A_STAR_1
0.0197 25 RANKID_1W	0.0339 6 WATCHGRADECD
0.0121 1 DISTCD3	0.0234 1 DISTCD3
0 10 SPECIAL_GRADE	0.0149 25 RANKID_1W
0 12 NET_BF_GRADE	0 3 PEAKYN_2
0 2 OPENMM	0 10 SPECIAL_GRADE
0 3 PEAKYN_2	0 2 OPENMM
0 4 NATION_G2	0 26 RANKINTEN_1W
0 26 RANKINTEN_1W	0 12 NET_BF_GRADE
0 19 NET_AF_GRADE_D	0 19 NET_AF_GRADE_D
0 28 NET_AF_GRADE_1W	0 28 NET_AF_GRADE_1W

InfoGain (Ranker)	
최종관객	1주차
Ranked attributes:	Ranked attributes:
0.9533 21 AUDICNT_1W	1.4906 21 AUDICNT_1W
0.6858 14 AUDICNT_D	0.8959 20 SALESSHARE_1W
0.6786 20 SALESSHARE_1W	0.8384 23 SHOWCNT_1W
0.632 13 SALESSHARE_D	0.8262 14 AUDICNT_D
0.6014 23 SHOWCNT_1W	0.7308 13 SALESSHARE_D
0.6009 27 NET_AF_CNT_1W	0.7238 22 SCRNCNT_1W
0.6006 16 SHOWCNT_D	0.6729 27 NET_AF_CNT_1W
0.5605 22 SCRNCNT_1W	0.6652 16 SHOWCNT_D
0.5141 15 SCRNCNT_D	0.6445 15 SCRNCNT_D
0.439 24 RANK_1W	0.5117 24 RANK_1W
0.414 18 NET_AF_CNT_D	0.4744 18 NET_AF_CNT_D
0.4024 17 RANK_D	0.4051 17 RANK_D
0.2547 9 SPECIAL_CNT	0.2174 9 SPECIAL_CNT
0.1655 5 GENRECD1	0.1802 5 GENRECD1
0.1313 11 NET_BF_CNT	0.1412 11 NET_BF_CNT
0.1048 7 D_STAR_1	0.1096 8 A_STAR_1
0.0893 8 A_STAR_1	0.0977 7 D_STAR_1
0.0855 6 WATCHGRADECD	0.0622 6 WATCHGRADECD
0.0218 1 DISTCD3	0.0549 4 NATION_G2
0.0195 25 RANKID_1W	0.042 1 DISTCD3
0 10 SPECIAL_GRADE	0.0147 25 RANKID_1W
0 2 OPENMM	0 3 PEAKYN_2
0 4 NATION_G2	0 10 SPECIAL_GRADE
0 3 PEAKYN_2	0 2 OPENMM
0 26 RANKINTEN_1W	0 26 RANKINTEN_1W
0 12 NET_BF_GRADE	0 12 NET_BF_GRADE
0 19 NET_AF_GRADE_D	0 19 NET_AF_GRADE_D
0 28 NET_AF_GRADE_1W	0 28 NET_AF_GRADE_1W

〔그림 5-5〕 개봉 1주 후 선택된 변수

개봉 2주 후 변수들은 〔그림 5-6〕에 따르면 전문가 평점, 네티즌 평점, 개봉전 네티즌 평점, 개봉 월, 성수기 여부, 국가등이 선택되지 않았다. 개봉 3주 후 변수들 역시 〔그림 5-7〕에 따르면 전문가 평점, 네티즌 평점, 개봉전 네티즌 평점, 개봉 월, 성수기 여부, 국가 등이 선택되지 않았다.

변수선택법에 의해 선택된 변수들을 이용하여 Random Forest 모델을 생성하여 실험한 결과는 <표 5-11>과 같다. 전체 요인을 이용한 결과와 선택된 변수만을 이용한 결과는 예측시점마다 혼재된 결과를 보여주고

있다. 개봉 일과 개봉1 주후의 예측 시점에는 선택된 변수들을 이용한 모델들이 예측정확도가 높았으나 예측 시점이 뒤로 갈수록 요인 전체를 사용한 모델과 큰 차이가 나지 않았다.

GainRatio (Ranker)	
최종관객	1주차
Ranked attributes:	Ranked attributes:
0.5501 14 AUDICNT_D	0.3812 14 AUDICNT_D
0.4934 13 SALESSHARE_D	0.3306 16 SHOWCNT_D
0.4315 16 SHOWCNT_D	0.3284 13 SALESSHARE_D
0.3868 15 SCRNCNT_D	0.3264 15 SCRNCNT_D
0.3767 18 NET_AF_CNT_D	0.2908 18 NET_AF_CNT_D
0.3024 17 RANK_D	0.2604 17 RANK_D
0.2259 4 NATION_G2	0.2075 9 SPECIAL_CNT
0.1633 12 NET_BF_GRADE	0.1315 11 NET_BF_CNT
0.158 9 SPECIAL_CNT	0.0766 5 GENRECD1
0.1542 11 NET_BF_CNT	0.0628 7 D_STAR_1
0.0774 5 GENRECD1	0.0466 6 WATCHGRADECD
0.047 6 WATCHGRADECD	0.0444 8 A_STAR_1
0.0463 7 D_STAR_1	0.0121 1 DISTCD3
0.0384 8 A_STAR_1	0 3 PEAKYN_2
0.0251 1 DISTCD3	0 2 OPENMM
0 2 OPENMM	0 19 NET_AF_GRADE_D
0 19 NET_AF_GRADE_D	0 4 NATION_G2
0 3 PEAKYN_2	0 12 NET_BF_GRADE
0 10 SPECIAL_GRADE	0 10 SPECIAL_GRADE

InfoGain (Ranker)	
최종관객	1주차
Ranked attributes:	Ranked attributes:
0.6858 14 AUDICNT_D	1.0684 14 AUDICNT_D
0.632 13 SALESSHARE_D	0.9203 13 SALESSHARE_D
0.6006 16 SHOWCNT_D	0.7511 15 SCRNCNT_D
0.5141 15 SCRNCNT_D	0.6702 16 SHOWCNT_D
0.414 18 NET_AF_CNT_D	0.5692 18 NET_AF_CNT_D
0.4024 17 RANK_D	0.4568 17 RANK_D
0.2547 9 SPECIAL_CNT	0.2149 9 SPECIAL_CNT
0.1655 5 GENRECD1	0.1672 5 GENRECD1
0.1313 11 NET_BF_CNT	0.1539 11 NET_BF_CNT
0.1048 7 D_STAR_1	0.0943 12 NET_BF_GRADE
0.0893 8 A_STAR_1	0.0862 6 WATCHGRADECD
0.0855 6 WATCHGRADECD	0.0796 4 NATION_G2
0.0218 1 DISTCD3	0.0773 7 D_STAR_1
0 3 PEAKYN_2	0.0773 8 A_STAR_1
0 2 OPENMM	0.0451 1 DISTCD3
0 19 NET_AF_GRADE_D	0 2 OPENMM
0 4 NATION_G2	0 19 NET_AF_GRADE_D
0 12 NET_BF_GRADE	0 3 PEAKYN_2
0 10 SPECIAL_GRADE	0 10 SPECIAL_GRADE

〔그림 5-4〕 개봉일 선택된 변수

GainRatio (Ranker)	
최종관객	1주차
Ranked attributes:	Ranked attributes:
0.6944 30 AUDICNT_2W	0.78568 30 AUDICNT_2W
0.4933 21 AUDICNT_1W	0.58523 21 AUDICNT_1W
0.4492 31 SCRNCNT_2W	0.53295 29 SALESSHARE_2W
0.4087 32 SHOWCNT_2W	0.51119 32 SHOWCNT_2W
0.3947 36 NET_AF_CNT_2W	0.48753 31 SCRNCNT_2W
0.3816 29 SALESSHARE_2W	0.45629 14 AUDICNT_D
0.3812 14 AUDICNT_D	0.44248 36 NET_AF_CNT_2W
0.3545 22 SCRNCNT_1W	0.43768 23 SHOWCNT_1W
0.3416 20 SALESSHARE_1W	0.40285 22 SCRNCNT_1W
0.3407 33 RANK_2W	0.39112 20 SALESSHARE_1W
0.3306 16 SHOWCNT_D	0.38917 27 NET_AF_CNT_1W
0.3284 13 SALESSHARE_D	0.36531 13 SALESSHARE_D
0.3264 15 SCRNCNT_D	0.35534 33 RANK_2W
0.3072 27 NET_AF_CNT_1W	0.34604 16 SHOWCNT_D
0.305 23 SHOWCNT_1W	0.31077 24 RANK_1W
0.2908 18 NET_AF_CNT_D	0.2895 15 SCRNCNT_D
0.2796 24 RANK_1W	0.2669 18 NET_AF_CNT_D
0.2604 17 RANK_D	0.2521 17 RANK_D
0.2075 9 SPECIAL_CNT	0.17264 9 SPECIAL_CNT
0.1706 35 RANKINTEN_2W	0.16222 4 NATION_G2
0.1315 11 NET_BF_CNT	0.13184 35 RANKINTEN_2W
0.0988 34 RANKID_2W	0.11982 11 NET_BF_CNT
0.0766 5 GENRECD1	0.10295 34 RANKID_2W
0.0628 7 D_STAR_1	0.07487 5 GENRECD1
0.0466 6 WATCHGRADECD	0.05699 8 A_STAR_1
0.0444 8 A_STAR_1	0.05473 7 D_STAR_1
0.0197 25 RANKID_1W	0.03378 6 WATCHGRADECD
0.0121 1 DISTCD3	0.01669 1 DISTCD3
0 2 OPENMM	0.00914 25 RANKID_1W
0 4 NATION_G2	0 2 OPENMM
0 3 PEAKYN_2	0 3 PEAKYN_2
0 37 NET_AF_GRADE_2W	0 37 NET_AF_GRADE_2W
0 10 SPECIAL_GRADE	0 10 SPECIAL_GRADE
0 28 NET_AF_GRADE_1W	0 28 NET_AF_GRADE_1W
0 26 RANKINTEN_1W	0 26 RANKINTEN_1W
0 12 NET_BF_GRADE	0 12 NET_BF_GRADE
0 19 NET_AF_GRADE_D	0 19 NET_AF_GRADE_D

InfoGain (Ranker)	
최종관객	1주차
Ranked attributes:	Ranked attributes:
1.2996 30 AUDICNT_2W	1.79161 30 AUDICNT_2W
0.9533 21 AUDICNT_1W	1.29693 21 AUDICNT_1W
0.8672 29 SALESSHARE_2W	1.04495 29 SALESSHARE_2W
0.8137 32 SHOWCNT_2W	0.99598 32 SHOWCNT_2W
0.7364 36 NET_AF_CNT_2W	0.96738 31 SCRNCNT_2W
0.7118 31 SCRNCNT_2W	0.88935 20 SALESSHARE_1W
0.6858 14 AUDICNT_D	0.80493 36 NET_AF_CNT_2W
0.6786 20 SALESSHARE_1W	0.71757 14 AUDICNT_D
0.6331 33 RANK_2W	0.68873 23 SHOWCNT_1W
0.632 13 SALESSHARE_D	0.68643 13 SALESSHARE_D
0.6014 23 SHOWCNT_1W	0.65593 16 SHOWCNT_D
0.6009 27 NET_AF_CNT_1W	0.61971 22 SCRNCNT_1W
0.6006 16 SHOWCNT_D	0.61469 27 NET_AF_CNT_1W
0.5605 22 SCRNCNT_1W	0.57711 15 SCRNCNT_D
0.5141 15 SCRNCNT_D	0.54512 33 RANK_2W
0.439 24 RANK_1W	0.4888 18 NET_AF_CNT_D
0.414 18 NET_AF_CNT_D	0.48782 24 RANK_1W
0.4024 17 RANK_D	0.38962 17 RANK_D
0.2547 9 SPECIAL_CNT	0.23484 9 SPECIAL_CNT
0.1655 5 GENRECD1	0.16176 5 GENRECD1
0.1608 35 RANKINTEN_2W	0.14523 34 RANKID_2W
0.1394 34 RANKID_2W	0.1243 35 RANKINTEN_2W
0.1313 11 NET_BF_CNT	0.11958 11 NET_BF_CNT
0.1048 7 D_STAR_1	0.11473 8 A_STAR_1
0.0893 8 A_STAR_1	0.09141 7 D_STAR_1
0.0855 6 WATCHGRADECD	0.06196 6 WATCHGRADECD
0.0218 1 DISTCD3	0.05713 4 NATION_G2
0.0195 25 RANKID_1W	0.03002 1 DISTCD3
0 2 OPENMM	0.00905 25 RANKID_1W
0 4 NATION_G2	0 2 OPENMM
0 3 PEAKYN_2	0 3 PEAKYN_2
0 37 NET_AF_GRADE_2W	0 37 NET_AF_GRADE_2W
0 10 SPECIAL_GRADE	0 10 SPECIAL_GRADE
0 28 NET_AF_GRADE_1W	0 28 NET_AF_GRADE_1W
0 26 RANKINTEN_1W	0 26 RANKINTEN_1W
0 12 NET_BF_GRADE	0 12 NET_BF_GRADE
0 19 NET_AF_GRADE_D	0 19 NET_AF_GRADE_D

〔그림 5-6〕 개봉 2주 후 선택된 변수

GainRatio (Ranker) 최종관객	
Ranked attributes:	
0.778	39 AUDICNT_3W
0.6944	30 AUDICNT_2W
0.4935	38 SALESSHARE_3W
0.4933	21 AUDICNT_1W
0.4492	31 SCRNCNT_2W
0.4304	45 NET_AF_CNT_3W
0.425	40 SCRNCNT_3W
0.4183	41 SHOWCNT_3W
0.4087	32 SHOWCNT_2W
0.3947	36 NET_AF_CNT_2W
0.3816	29 SALESSHARE_2W
0.3812	14 AUDICNT_D
0.3545	22 SCRNCNT_1W
0.3416	20 SALESSHARE_1W
0.3407	33 RANK_2W
0.3306	16 SHOWCNT_D
0.3284	13 SALESSHARE_D
0.3264	15 SCRNCNT_D
0.3223	42 RANK_3W
0.3072	27 NET_AF_CNT_1W
0.305	23 SHOWCNT_1W
0.2908	18 NET_AF_CNT_D
0.2796	24 RANK_1W
0.2604	17 RANK_D
0.2075	9 SPECIAL_CNT
0.1718	44 RANKINTEN_3W
0.1706	35 RANKINTEN_2W
0.1639	43 RANKID_3W
0.1315	11 NET_BF_CNT
0.0988	34 RANKID_2W
0.0766	5 GENRECD1
0.0628	7 D_STAR_1
0.0466	6 WATCHGRADECD
0.0444	8 A_STAR_1
0.0197	25 RANKID_1W
0.0121	1 DISTCD3
0	10 SPECIAL_GRADE
0	3 PEAKYN_2
0	2 OPENMM
0	12 NET_BF_GRADE
0	4 NATION_G2
0	37 NET_AF_GRADE_2W
0	28 NET_AF_GRADE_1W
0	26 RANKINTEN_1W
0	19 NET_AF_GRADE_D
0	46 NET_AF_GRADE_3W

InfoGain (Ranker) 최종관객	
Ranked attributes:	
1.4983	39 AUDICNT_3W
1.2996	30 AUDICNT_2W
0.9723	40 SCRNCNT_3W
0.9705	38 SALESSHARE_3W
0.9612	41 SHOWCNT_3W
0.9533	21 AUDICNT_1W
0.9237	45 NET_AF_CNT_3W
0.8672	29 SALESSHARE_2W
0.8137	32 SHOWCNT_2W
0.7364	36 NET_AF_CNT_2W
0.7118	31 SCRNCNT_2W
0.6984	42 RANK_3W
0.6858	14 AUDICNT_D
0.6786	20 SALESSHARE_1W
0.6331	33 RANK_2W
0.632	13 SALESSHARE_D
0.6014	23 SHOWCNT_1W
0.6009	27 NET_AF_CNT_1W
0.6006	16 SHOWCNT_D
0.5605	22 SCRNCNT_1W
0.5141	15 SCRNCNT_D
0.439	24 RANK_1W
0.414	18 NET_AF_CNT_D
0.4024	17 RANK_D
0.2558	44 RANKINTEN_3W
0.2547	9 SPECIAL_CNT
0.1655	5 GENRECD1
0.1608	35 RANKINTEN_2W
0.1394	34 RANKID_2W
0.1317	43 RANKID_3W
0.1313	11 NET_BF_CNT
0.1048	7 D_STAR_1
0.0893	8 A_STAR_1
0.0855	6 WATCHGRADECD
0.0218	1 DISTCD3
0.0195	25 RANKID_1W
0	37 NET_AF_GRADE_2W
0	3 PEAKYN_2
0	12 NET_BF_GRADE
0	2 OPENMM
0	4 NATION_G2
0	28 NET_AF_GRADE_1W
0	10 SPECIAL_GRADE
0	26 RANKINTEN_1W
0	19 NET_AF_GRADE_D
0	46 NET_AF_GRADE_3W

〔그림 5-7〕 개봉 3주 후 선택된 변수

<표 5-13> 변수선택법 별 예측 정확도

예측 시점	타겟 변수	GainRatio (Ranker)	InfoGain (Ranker)	요인 전체
개봉일	최종 관람객 수	67.29	62.09	61.61
	개봉 1주 후 누적 관람객 수	78.19	74.88	73.93
개봉 1주차	최종 관람객 수	66.82	69.67	67.77
	개봉 2주 후 누적 관람객 수	81.51	81.04	83.88
개봉 2주차	최종 관람객 수	77.25	69.19	74.40
	개봉 3주 후 누적 관람객 수	88.26	77.25	88.62
개봉 3주차	최종 관람객 수	87.20	80.57	87.20

Ⅵ. 결론

1. 결론

본 연구에서는 기계학습의 분류기법을 이용하여 주차별 누적 관람객 수를 예측하는 기법을 제안하였고 그 결과를 확인하였다. 전체 관람객 수와 총 매출액 예측만 시도되었던 기존 연구와는 달리 주차별 누적 관람객 수 예측을 통해 개봉 후 역동적으로 변하는 관람객들의 반응과 영화 흥행 실적에 따라 빠르게 대응하고 입체적으로 분석할 수 있는 방법을 제안하였으며, 총 실험 결과 Random Forest가 73% ~ 87% 의 정확도를 보였다. 신규 요인을 포함한 모델과 제외한 모델을 비교 실험한 결과 전체적으로 신규 요인을 포함한 모델이 전체적으로 높은 예측정확도를 보였다. 가장 높게 나온 Random Forest를 변수 선택법을 이용하여 추가 분석한 결과 전체적으로 GainRatio의 Ranker기법을 이용하여 선택된 요인들을 적용한 실험결과가 전체적으로 높은 예측정확도를 나타냈다.

향후 연구로는 본 연구 결과를 바탕으로 일 단위 예측을 통해 좀 더 실용적 적용 가능성이 높은 연구를 진행해야 할 것으로 판단된다. 또한 일 단위 예측을 위해 좀 더 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.

Ⅶ. 참고문헌

강선주, 2017, “흥행영화 요소 분석”, 한국엔터테인먼트산업학회논문지, 제11권 제5호, pp. 1-15

고정민, 2010, “미국영화와 한국영화의 흥행요인에 관한 비교연구”, 문화산업연구, 제10권 제2호, pp. 71-96

김범수, 서주환, 2017, “소비자 오피니언이 영화흥행에 미치는 영향에 관한 연구”, 유통연구, 제22권 제2호, pp. 65-91

김병선, 2009, “영화 유형에 따른 흥행 예측 요인 비교 연구”, 한국언론학보, 제53권 제1호, pp. 257-287

김소영, 임승희, 정예슬, 2010, “영화 유형별 영화 흥행 성과 예측 요인의 비교 연구”, 한국콘텐츠학회논문지, 제10권 제2호, pp. 381-393

김연형, 홍정한, 2013, “영화흥행 영향요인 선택에 관한 연구”. 응용통계연구, 제26권 제3호, pp. 441-452

김휴종, 1988, “한국 영화스타의 스타파워 분석”, 문화경제연구, 제1권 제1호, pp. 165-200

권선주, 2014, “전문가 평가가 영화흥행성과에 미치는 영향력”, 문화경제연구, 제17권 제3호, pp. 3-21

권신혜, 박경우, 장병희, 2017, “기계학습 기반의 영화흥행예측 방법 비교: 인공신경망과 의사결정나무를 중심으로”. 예술인문사회융합멀티미디어논문지, 제7권 제4호, pp. 593-601

권재웅, 홍병기, 2012, “애니메이션영화의 흥행성과 결정 요인에 관한 연

구”, 문화산업연구, 제12권 제1호, pp. 93-106

방송통신위원회, 2016, 미디어의 새로운 방향, OTT 서비스 알아보기, http://blog.daum.net/kcc1335/6279 (05 January, 2019)

박승현, 송현주, 2012, “영화의 주별 흥행성과에 미치는 영향”, 한국언론학보, 제56권 제4호, pp. 210-235

박승현, 이푸름, 2017, “한국 영화시장 내 애니메이션영화의 흥행결정 요인 분석”, 문화산업연구, 제17권 제3호, pp. 1-7

서기만, 2011, “OTT 서비스의 이해와 전망”, 방송과 미디어, 재16권, 제1호, *pp.* 91-101

이양환, 장병희, 박경우, 2007, “국가 간 영화흥행요인 비교를 위한 탐색적 연구”, 언론과학연구, 제7권 제1호, pp. 185-222

유현석, 2001, “영화 흥행 변수에 관한 연구”, 문화정책논총, 제13권, pp. 231-254

유현석, 2002, “한국영화의 흥행 요인에 관한 연구”, 한국언론학보, 제46권, 제3호, pp. 183-213

이경렬, 이시내, 2013, “SNS 이용자들의 온라인 구전(eWOM) 행동에 영향을 미치는 요인에 관한연구 개인적 특성, SNS 특성, 대인적 영향, 사회적 자본을 중심으로”, 한국광고홍보학보(구 한국광고학보), 제15권 제4호, pp. 273-315

임준엽, 황병연, 2014, “트위터를 이용한 기계학습 기반의 영화흥행 예측”, 정보처리학회지, 제3권 제7호, pp. 263-270

장재영, 2017, “소셜 빅데이터 분석과 기계학습을 이용한 영화흥행예측 기

법의 실험적 평가”, 한국인터넷방송통신학회 논문지, 제17권 제3호, pp. 167-173

전성현, 손영숙, 2016, “데이터마이닝을 이용한 박스오피스 예측”, 한국통계학회, 제29권 제7호, pp. 1257-1270

전성현, 손영숙, 2016, “영화 흥행 예측변수로서 온라인 구전 변수의 효과”. 응용통계연구, 제29권 제4호, pp. 657-678

최성희, 2017, “한국 영화 상영시장에서 배급사의 영향”, 문화경제연구, 제20권 제1호, pp. 105-128

한경닷컴사전 http://dic.hankyung.com/apps/economy.view?seq=12495 (05 January, 2019)

한국 영화 연감 (2017). 2017 영화진흥위원회 한국영화산업결산,

http://www.kofic.or.kr/ (Downloaded 13 February, 2018)

Breiman, L., 1996, “Bagging predictors”, *Machine learning,* Vol.24. No.2, pp. 123-140

Breiman, L., 2001. “Random Forests”, *Machine Learning*, Vol.45, pp. 5-32

Fawcett, T., 2006, “An introduction to ROC analysis”, *Pattern recognition letters*, Vol.27, No.8, pp. 861-874

Gunn, S. R., 1998, “Support vector machines for classification and regression”, *ISIS technical report*, Vol.14, No.1, pp. 5-16

Heckerman, D. 1997, “Bayesian networks for data mining”, *Data mining*

and knowledge discovery, Vol.1, No.1, pp. 79-119

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

Natekin, Alexey, Alois Knoll., 2013, "Gradient boosting machines, a tutorial", *Frontiers in neurorobotics,* Vol.7, pp. 21

Neapolitan, R. E., 2004, *Learning bayesian networks*, New Jersey, Pearson Prentice Hall

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

Quinlan, J. R., 1987, "Simplifying decision trees", *International journal of man-machine studies*, Vol.27, No.3, pp. 221-234

Quinlan, J. R., 1993, *C4.5: Programs for Machine Learning.* Morgan Kaufmann,

Quinlan, J. R., 1998, "Induction of decision trees", *Machine Learning*, Vol.1, No.1 pp. 81-106

Rhee, T. G., F. Zulkernine, 2016. "Predicting Movie Box Office Profitability: A Neural Network Approach," *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*

Ru, Y., Li, B., Liu, J., & Chai, J. 2018. "An effective daily box office prediction model based on deep neural networks." *Cognitive Systems*

Research, Vol.52, pp. 182-191

Sharda, R., Delen, D, 2006, "Predicting Box-Office Success of Motion Pictures with Neural Networks", *Expert Systems with Applications*, Vol.30, pp. 243-254

Simonoff, J. S., Sparrow, I. R., 2000, "Predicting movie grosses: Winners and losers, blockbusters and sleepers." *Chance*, Vol.13, No.3, pp. 15-24

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

Zhang, W., Skiena, S. 2009, "Improving movie gross prediction through news analysis. In Web Intelligence and Intelligent Agent Technologies," *WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on. IEEE* Vol.1, pp. 301-304

ABSTRACT

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

Song Jung A

Department of Big Data Business

Master of Business Administration

Hanbat National University

Advisor : Kim Gun Woo

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 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 Bayes, 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