http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.25, No.2, May 2020, pp.49-63
https://doi.org/10.7838/jsebs.2020.25.2.049

영화 관객 수 예측을 위한 기계학습 기법의 성능 평가 연구

A Study on the Performance Evaluation of Machine Learning for Predicting the Number of Movie Audiences

정찬미(Chan-Mi Jeong)*, 민대기(Daiki Min)**

초 록

영화 제작에 막대한 비용이 투입되지만 관객수요는 매우 불확실하기 때문에 개선된 수요예측은 수익 개선을 위한 의사결정의 중요 수단으로 활용될 수 있다. 본 연구에서는 영화의 개봉 후 수요를 예측함에 있어 기계학습 기법의 적용 타당성을 예측 성능의 관점에서 검증하였다. 분석결과를 종합하면 다음과 같다. 첫째, 대안변수에 대한 통계적 검증 결과 기본 영화 특성(감독, 배우)과 함께 개봉 후 2주차까지의 스크린수, 상영횟수, 관객수, 주요 배우에 대한 관심도 등 시계열 자료가 수요예측에 유의미한 것을 확인하였다. 둘째, Random Forest Classifier와 SVM(Support Vector Machine) 등 분류 기반 기계학습 기법과 Random Forest Regressor와 k-NN Regressor와 같은 회귀모형 기반 기계학습 기법에 적용하여 예측 성능을 평가한 결과, Random Forest 기법이 우수한 결과를 보였다. 셋째, 누적관객수가 1분위보다 작은 영화에서 회귀모형 기반 기법은 낮은 예측 정확도를 보였으며, 분류기반 기법은 반대로 가장 우수한 결과를 얻었다. 즉, 영화 수요의 분포 특성에 따라서 차별화된 기계학습 기법을 적용하는 것이 필요하다.

ABSTRACT

The accurate prediction of box office in the early stage is crucial for film industry to make better managerial decision. With aims to improve the prediction performance, the purpose of this paper is to evaluate the use of machine learning methods. We tested both classification and regression based methods including k-NN, SVM and Random Forest. We first evaluate input variables, which show that reputation-related information generated during the first two-week period after release is significant. Prediction test results show that regression based methods provides lower prediction error, and Random Forest particularly outperforms other machine learning methods. Regression based method has better prediction power when films have small box office earnings. On the other hand, classification based method works better for predicting large box office earnings.

키워드 : 영화 관객 수 예측, 기계학습, 분류 모형, 회귀 모형, Random Forest, k-NN, SVM
Box Office Forecasting, Machine Learning, Classification Model, Regression Model, Random Forest, K-NN, SVM

* First Author, Graduate Student, Graduate School(Big Data Analytics), Ewha Womans University (jpraise@ewhain.net)
** Corresponding Author, Associate Professor, School of Business, Ewha Womans University(dmin@ewha.ac.kr)
Received: 2020-03-26, Review completed: 2020-04-14, Accepted: 2020-04-20

1. 서 론

영화는 제작에 막대한 비용이 투입되는 상품으로 영화 제작사 또는 배급사는 관객 수요에 따라서 향후 스크린 수와 상영 횟수를 조정하여 손실을 최소화하거나 수익을 극대화하는 의사결정이 매우 중요하다. 하지만 개봉 후 관객 수요는 불확실성이 매우 높기 때문에 관객 수에 대한 정확한 예측은 이해당사자들의 수익과 직접적으로 연결된 의사결정을 위한 중요 수단이라고 할 수 있다.

영화 관객 수 예측과 관련한 기존 연구는 선형 회귀분석 및 확률 모형을 포함한 통계적 모형[4, 8, 19, 23, 31], 확산모형(Diffusion Model) 및 벡터자동회귀(Vector Autoregression; VAR)와 같은 시계열 예측 모형[24, 31, 32], 인공신경망(Artificial Neural Network; ANN)과 같은 기계학습(Machine Learning) 모형[6, 18, 26, 34] 등 세 가지 유형의 방법론을 활용하였다. 특히, 최근 분류 기반 기계학습 모형과 회귀모형 기반 기계학습 모형을 활용한 연구가 활발하게 진행되고 있다(<Table 1> 참조).

분류 기반의 기계학습 모형을 이용한 연구의 경우 누적 관객 수를 직접 예측하는 대신에 종속변수(즉, 관객 수)를 범주화하여 예측을 수행하였다[6, 12, 18, 30, 34]. 예를 들어, Zhang et al.[34]은 과거 영화 관객 규모의 분포를 고려하여 예측 관객 수를 6개 범주로 분류하였으며, Subramaniyaswamy et al.[30]은 개봉 영화의 ROI(Return On Investment)를 네 가지 범주로 정의하여 분류모형을 적용하였다. 기존 연구에서는 분류 기반 기계학습 모형으로 ANN[6, 18, 34], Decision Tree(DT)[6, 12, 18], Bayesian Belief Network(BBN)[18], Random Forest(RF)[12], Support Vector Machine (SVM)[12, 30] 등을 활용하였다. 특히, Guo et al.[12]은 범주화된 수익 규모 예측을 위하여 RF 기법을 이용하였으며, DT, SVM, MLP(Multi-Layered Perceptron) 등의 방법론과 비교하여 예측 성능이 우수함을 제시하였다. 또한 Subramaniyaswamy et al.[30]은 범주화된 ROI의 분류 예측 과정에서 입소문(Word-of-Mouth)과 같은 새로운 변수를 고려함으로써 분류 정확도를 개선하는 결과를 얻었다.

영화의 흥행 여부가 매우 복잡하고 다양한 요인에 의해 영향을 받기 때문에 관객 수나 수익을 연속형으로 고려하는 경우 높은 예측 정확도를 기대하는 것이 어렵기 때문에 범주형 종속변수를 활용하였다. 범주형 종속변수를 이용하는 경우 관객 수를 정확하게 제시하지는

〈Table 1〉 Literature on Machine Learning Models

Dependent variable	Literature	Method
Categorical	Delen et al.[6] Lee and Chang[18] Zhang et al.[34] Guo et al.[12] Subramaniyaswamy et al.[30]	ANN(Artificial Neural Network), DT(Decision Tree) BBN(Bayesian Belief Network), ANN, DT, ANN RF(Random Forest), DT, SVM(Support Vector Machine), MLP(Multi-Layered Perceptron), SVM
Numerical	Abel et al.[1] Kim et al.[15]	Bagging REPTree(Reduced Error Pruning Tree) SVR, GPR(Gaussian Process Regression), k-NN(k-Nearest Neighbor)

못하지만 분류 정확도가 연속형 종속변수에 비하여 상대적으로 우수하고 많은 연구사례가 존재하여 광범위하게 활용되고 있다. 하지만 범주형 종속변수를 예측하는 분류 모형은 연속형 변수인 관객 수를 몇 가지 제한된 범주로 조작함으로써 정보가 손실되는 문제가 존재하며, 분류 기반의 기계학습 모형이 제한되는 단점이 존재한다.

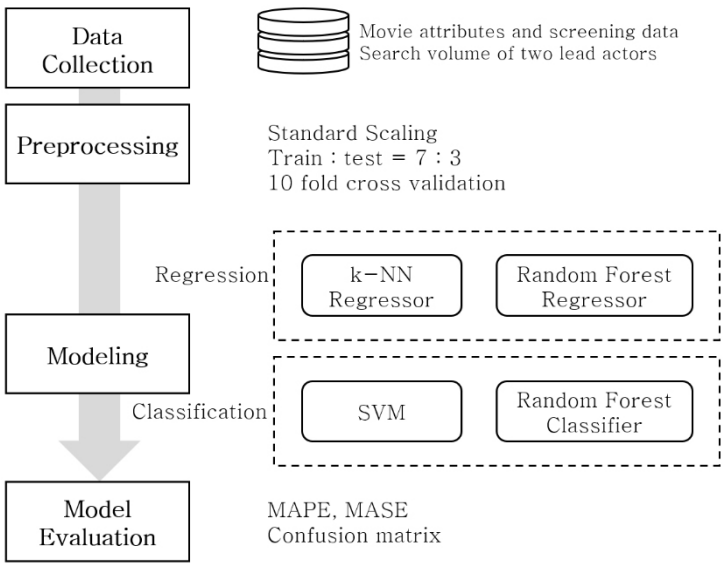
범주형 종속변수를 활용한 분류 기반의 기계학습 모형의 단점을 고려하여 최근 연속형 종속변수의 예측을 위한 회귀모형 기반의 기계학습 모형을 이용한 연구가 일부 제시되고 있다[1, 15]. Abel et al.[1]은 영화 흥행 수익과 음반 판매량 데이터를 대상으로 8개의 기계학습 기법과 단순 선형회귀모형의 예측성능을 비교하였으며, 기계학습 기법이 선형회귀모형보다 예측성능이 우수함을 확인하였다. Kim et al.[15]은 예측 정확도를 높이기 위해 회귀모형 기반의 기계학습 기법인 Support Vector Regression (SVR), Gaussian Process Regression(GPR), k-Nearest Neighbor (k-NN)을 조합한 앙상블 기법을 제안하였으며, 선형 회귀분석모형과 비교하여 예측 정확도를 개선하였다. 연속형 종속변수를 이용함으로써 관객 수를 직접 예측할 수 있는 장점이 있으나, 예측 정확도가 낮아지는 문제가 존재한다. 또한 연속형 종속변수를 이용한 연구는 영화 특성과 관련한 데이터 보다는 과거 관객 수와 선호도와 같은 시계열 자료만을 제한적으로 활용하여 예측을 수행하고 있다.

본 논문에서는 영화산업에서 개봉 영화의 수익개선을 위한 의사결정을 지원하기 위한 목적에서 개봉 영화의 관객 수 예측 문제를 고려하였다. 특히, 연속형 종속변수를 위한 회귀모형 기반 기계학습 모형과 범주형 종속변수를 대상으로 하는 분류 기반 기계학습 모형의 예측성능을 비교함으로써 의사결정에 유용한 정보의 수준에 적합한 기계학습 모형을 확인하고자 한다. 또한 기계학습 모형을 구성함에 있어 관객 수와 같은 시계열 데이터와 함께 영화 특성을 함께 입력변수로 고려함으로써 모형 특성에 따라서 예측성능 개선에 유용한 데이터의 특성을 제시한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 연구문제를 정의하고, 연구방법론을 설명한다. 연구방법론에서는 기계학습 모형과 함께 본 논문에서 고려한 주요 입력변수를 제시한다. 제 3장에서는 수치실험 결과를 통하여 기계학습 모형의 예측 성능을 비교평가하고, 제 4장에서는 본 연구의 결론과 향후 연구주제를 제시한다.

2. 연구 방법

본 논문은 개봉 영화의 관객 수 예측 문제를 대상으로 기존 연구에서 주로 사용했던 분류 기반의 기계학습 모형과 함께 연속형 종속변수를 대상으로 회귀모형 기반의 기계학습 모형의 성능을 비교 평가하는 것을 목적으로 한다. 이번 장에서는 이와 같은 연구목적을 달성하기 위한 연구 수행 절차와 단계별 주요 고려사항을 설명한다. 우선, 개봉 영화의 관객 수 예측을 위한 입력변수 선정과 관련한 문헌 연구를 살펴보고, 이를 기반으로 본 논문에서 수집한 데이터 및 전처리 과정을 설명한다. 둘째, 수집한 데이터를 활용하여 관객 수 예측에 사용한 기계학습 모형의 특성과 성능 평가 방법을 제시한다. <Figure 1>은 데이터 수집 및 전처리,



〈Figure 1〉 Research Framework

기계학습 모형의 구성, 예측 성과 평가 등의 4단계로 구성된 연구절차를 나타낸다.

2.1 관객 수 예측을 위한 입력변수

일반적으로 영화의 관객 수는 개봉 이후의 관객 수 및 스크린수와 같은 시계열 데이터와 유의미한 관계를 나타내고 있으며[5, 30], 시계열 데이터와 함께 배우, 감독, 장르 등의 영화의 기본 속성 정보가 함께 사용되고 있다[6, 15, 21]. 또한 최근에는 영화에 대한 대중의 관심도를 관객 수 예측에 활용하기 위하여 영화에 대한 검색량 데이터를 반영하고 있다[5, 9, 13, 21, 36].

관객 수 예측에 있어서 영화 속성 중 출연배우에 의한 효과는 기존 연구에서 많이 고려되고 있다. 출연배우와 관련한 속성을 예측 모형에 반영하는 방법으로는 주연배우의 유명 잡지 수록 여부[27]나 전작의 '흥행 수익 상위 10위권' 포함여부[10, 18, 19, 23] 등을 더미 변수로 반영하는 방법과 함께 구글 검색량을 활용한 인지도 척도를 이용하는 방법[34]을 고려할 수 있다. 감독의 영향력 지표 또한 배우와 유사한 방법으로 반영하고 있는데, 전작의 흥행 수익[9]을 사용하거나 전문 평론가들의 평가결과[33]를 사용하여 정량화하였다.

영화 속성 중에서 장르와 MPAA 등급(The Motion Picture Association of America film rating)은 잠재적 시장규모를 결정하는 주요 요인으로 고려되고 있고 있으며, 다수의 연구에서 이를 입력변수로 활용하고 있다. 두 변수 모두 더미(dummy) 변수의 형태로 많은 연구에서 활용되었다. 하지만 일부 연구에서는 장르와 MPAA 등급을 관객 수 예측에 활용함으로써 예측성과 개선에 긍정적 효과가 있음을 확인하였으나[3, 10, 25, 27], 다른 연구에서는 유의미한 결과를 확인하지 못하였다[19, 20, 22].

마지막으로 일부 연구에서 영화의 제작 국가

또는 상영 지역과 관객 수의 차이를 고려하고 있으나 지역별 차이에 의한 효과는 제한적으로 나타나고 있다[27, 31, 35]. 본 연구에서는 한국 영화와 해외 영화의 특성 차이를 반영하기 위하여 두 영화를 단일 모형에서 함께 다루지 않고 한국 영화와 외국 영화를 분리한 모형 또한 구축하였다.

배우, 감독과 같은 영화 속성 정보를 관객 수 예측에 활용하기 위해서는 적절하게 정량 변수로 변환하는 과정이 요구된다. 이와 관련하여 최근 수요예측에 있어서 범주요인에 의한 효과를 정량적으로 반영하기 위해서 인터넷 검색량 정보를 이용한 연구를 확인할 수 있다[5, 7, 9, 13, 15, 16, 21, 34]. 예를 들어, Demir et al.[7]은 관객 평점 예측 문제에서 영화 제목, 장르, 주연 배우에 따른 효과를 고려하기 위하여 구글 트렌드 검색량을 사용하였다. 인터넷 검색량은 검색 대상에 대한 대중의 관심도를 보여주는 유용한 자료이지만, 상대빈도를 제공하는 검색량 데이터의 특성을 고려하여 적절하게 사용하는 것이 요구된다. Chong[5]은 기간별 검색량의 평균 변화율을 고려함으로써 관심도의 지속성과 주목도를 측정하였다.

본 연구에서는 개봉 영화의 관객 수 예측과 관련한 기존 문헌에서 사용하고 있는 영화 속성 정보와 이를 정량화한 인터넷 검색량, 그리고 과거 관객 수 등의 시계열 데이터를 잠재 변수로 고려하였다. 다음 절에서는 본 논문에서 고려하고 있는 잠재 변수와 관련한 데이터의 수집 및 전처리 방안을 설명하도록 한다.

2.2 데이터 수집 및 전처리 방안

연구수행을 위한 기초 데이터로 2015년부터 2018년 7월까지 국내에서 개봉된 전체 영화들 중에서 최종 누적 관객 수 20,000명 이상이고 총 상영 일수가 21일 이상인 영화 182편을 수집하였다. 분석 대상인 영화 182편 중에서 한국영화와 외화는 각각 102편과 80편 이었다. 영화의 기본 정보 (배우, 감독, 제작사, 수입사, 배급사 등)와 영화별 상영 스크린수, 상영횟수, 관객 수 등의 기초 데이터는 영화진흥위원회 ()가 제공하는 open API를 이용하여 수집하였다.

상영 스크린수는 영화가 1회차 이상 상영된 경우를 대상으로 일별 상영 스크린수를 수집하였으며, 이를 이용하여 개봉일 스크린수, 개봉 첫 주간 누적 스크린수, 둘째 주간 누적 스크린수 등 주간 누적 스크린수를 계산하여 예측에 이용하였다. 상영횟수는 영화별로 전국에서 상영된 횟수를 의미하며, 스크린수와 동일하게 개봉일 상영횟수, 개봉 첫 주간 누적 상영횟수, 둘째 주간 누적 상영횟수를 수집하였다. 관객 수 또한 영화별로 일별 관객 수를 수집한 후에 이를 이용하여 개봉일과 주간 누적 관객수를 계산하였다.

배우, 감독, 제작사, 수입사 등 영화의 기본 정보를 정량화하기 위하여 관객 수를 이용한 파생 변수를 생성하였다. 예를 들어, 주연배우의 전작 관객 수를 해당 배우의 영향력 지표로 반영한 선행연구들[10, 19]을 참고하여 영화별로 주연배우 2인의 과거 3년간 전작 관객 수 평균을 배우의 영향력 점수로 사용하였다. 동일한 방법으로 감독, 배급사, 제작사, 수입사의 영행력 점수 또한 과거 3년간 전작 관객 수 평균을 이용하여 계산하였다. 애니메이션 장르의 경우 배우의 영향력 점수는 제외하였으며, 외화의 제작사 점수와 국내 영화의 수입사 영향력 점수 또한 0점으로 처리하였다.

관객 수를 이용하여 도출한 주연배우의 영향력 점수와 함께 주연배우에 대한 대중의 관심도를 반영하기 위해 개봉일 전후 3주, 즉 총 6주 동안 구글 트렌드(https:// trends.google.co.kr)에서 수집한 주연배우 2인의 일별 검색량을 함께 고려하였다. 구글 트렌드의 검색량 데이터는 검색 기간 중 검색 빈도가 가장 높은 값을 100으로 정의하고, 이를 기준으로 일별 검색 빈도를 상대도수로 제시한다. 따라서 구글 트렌드의 검색량 데이터는 주연배우에 대한 관심도를 절대적으로 단순 비교하는데 적절하지 않다. 본 연구에서는 일별 상대도수 값을 직접 적용하는 대신에 기간별 누적값을 사용함으로써 검색 기간 내에서 대중들의 관심이 얼마나 꾸준하게 지속되었는지 정량적으로 나타내는 지표로 활용하였다.

일반적으로 개봉 영화의 흥행여부는 개봉 후 첫 3주 기간 내에 결정되는 것으로 알려져 있으며[17], 다수의 국내연구에서 개봉 후 3주 차 시점에서의 관객 수 예측 문제를 고려하고 있다[13, 17]. 본 논문 또한 개봉 3주차까지의 누적 관객 수 예측을 위하여 앞서 제시한 데이터를 개봉 후 3주차까지 수집하였다. 또한 본 연구에 사용된 데이터는 모두 수치형 자료이며, 원자료의 편차가 매우 큰 문제를 해결하고 학습의 안정적 수렴을 확보하기 위하여 분포가 평균이 0이고 표준편차가 1이 되도록 표준화 스케일링(Standard Scaling) 전처리 과정을 수행하였다.

2.3 관객 수 예측을 위한 기계학습 모형

개봉 후 3주차 시점에서의 누적 관객 수 예측을 위한 기계학습 모형은 앞서 언급한 바와 같이 종속변수의 유형에 따라서 회귀모형 기반의 k-NN Regressor, Random Forest Regressor와 분류모형 기반의 SVM, Random Forest Classifier를 사용하였다. Random Forest 모형은 영화 관객 수 예측문제에서 성능이 우수한 것으로 알려져 있어 분류기반 모형과 회귀모형 기반 모형에서 모두 사용하였다[33]. K-NN과 SVM은 다양한 유형의 데이터에 적용하기 쉽고 노이즈(noise)에 대한 영향이 적은 기법으로 노이즈가 존재하는 본 연구 데이터를 대상으로 Random Forest Classifier와 우수한 성능을 갖는지 비교하고자 한다.

Breiman[2]이 제안한 Random Forest 기법은 부트스트랩(Bootstrap) 표본을 다수 생성하고 이를 Decision Tree에 적용하여 도출한 표본결과를 종합하여 예측을 수행하는 앙상블 기법(ensemble methods)이다. 일반적으로 Random Forest 기법은 표본수가 증가할수록 예측 오차가 작아지며, 과적합이 발생하지 않는 장점을 갖는 것으로 알려져 있다. 또한 입력변수가 매우 많은 경우 예측력이 높으며 안정적인 성능을 제공한다[28, 35].

기계학습 모형의 성능을 최적화하기 위하여 복수의 하이퍼 파라미터 조합에 대한 그리드 탐색(Grid Search) 기법을 적용하여 성능이 가장 우수한 결과를 제공하는 하이퍼 파라미터 조합을 설정하였다. Decision Tree의 개수와 깊이는 각각 100과 5로 결정하였다. 분류기반 모형인 Random Forest Classifier의 경우 Decision Tree의 개수는 30으로 설정하였으며, 깊이는 'pure leaf'까지 확장하는 것을 설정하였다. 또한 내부 노드를 분할하는 데 필요한 최소 샘플 수는 5를 활용하였다.

k-NN Regressor는 사례 기반 학습기법으로 속성값 (즉, 입력변수)이 유사한 다른 영화들의

누적 관객 수의 가중평균을 활용하여 예측 대상 영화의 누적 관객 수를 예측한다. 본 논문에서는 Euclidian 거리함수를 이용하여 속성간 유사도를 측정하였으며, 근접 이웃의 개수 k는 사전 실험을 통하여 성능이 가장 우수한 값인 5를 선정하였다.

SVM은 패턴 인식 및 자료 분석을 위한 지도학습(Supervised Learning) 기법으로, k-NN과 같이 다양한 데이터 유형에 적용하기 쉽고 노이즈에 대한 영향이 적은 장점이 있다. 하지만 최적의 모델을 찾기 위해 Kernel과 매개변수들 사이의 조합에 대한 성능 실험이 필요하며 입력변수와 데이터가 많은 경우 학습이 오래 걸리는 단점이 있다[8, 9]. SVM의 주요 파라미터인 오분류 비용 C (cost of misclassification)와 gamma 값은 모두 0.0001을 적용하였다.

네 가지 기계학습 모형을 이용하여 학습과 검증 실험을 수행하고 예측 성능을 비교 평가하였다. 수집한 데이터는 학습용 데이터 70%과 검증용 데이터 30%로 분류하고 10-fold 교차검증(cross validation)을 수행하였다. 회귀모형 기반의 기계학습 모형의 예측 성능은 MAPE (Mean Absolute Percentage Error)와 MASE (Mean Absolute Scaled Error)를 사용하였으며, 분류기반 모형의 예측성능 평가를 위하여 Confusion matrix를 제시하였다.

3. 분석결과

3.1 종속변수: 누적 관객 수

영화진흥위원회에서 수집한 영화 182편의 개봉 후 3주차 누적 관객 수는 최소 14만 명, 최대 1,150만 명, 평균 약 268만 명이었다. <Table 2>에 제시한 바와 같이 누적 관객 수의 분포를 고려하여 분류기반 기계학습 모형의 종속변수(즉, 3주차 누적 관객 수)를 4분위수를 이용하여 네 구간(A, B, C, D)으로 범주화하였다.

〈Table 2〉 Categorized Target Variable

Category	A	B	C	D
Cum. Tickets Sold(million)	Less than 2	2~4	4~6	More than 6
Frequency	87	55	25	15

3.2 입력변수의 선정

2장에서 설명한 바와 같이 영화 속성과 시계열 데이터로 구성된 28개 잠재 입력변수를 대상으로 예측에 유의미한 입력변수를 선정하기 위한 분석을 수행하였다. 종속변수(즉, 개봉 후 3주차의 누적 관객 수)와 잠재 입력변수 사이의 상관성에 대한 통계적 유의성 검정을 수행하여 최종 입력변수를 선정하였다. 모형에서 사용한 입력변수는 표로 정리하여 Appendix에 제시하였다.

<Table 3>은 종속변수와 개별 잠재 입력변수 사이의 상관분석 결과로 도출된 검정통계량 및 유의성 결과를 나타낸다. 분석결과 영화 속성 및 시계열 데이터와 관련 변수의 경우 배우 1의 영향력 점수(Actor 1), 감독/배급사/제작사의 영향력 점수(Director, Distributor, Producer), 개봉 후 1주차와 2주차의 누적 관객수, 스크린수, 상영횟수(scrnCnt, audiAcc, showCnt) 등이 종속변수와 유의미한 상관성을 갖는 것으로 확인되었다. 반면에 배우 2의 영향력 점수(Actor 2)와 개봉일(release)의 관객수, 스크린수, 상영횟수는 모두 상관관계가 존재하지 않았다.

〈Table 3〉 Correlation Analysis between Input Variables and Target Variable

Movie attributes and Num. of screens		Search volume of two lead actors	
Actor1	0.6014764**	act1_b_days7	-0.06282008
Actor2	0.6861844	act1_b_days14	-0.07325798
Director	0.6672405**	act1_b_days21	-0.06759939
Distributor	0.3938774**	act2_b_days7	0.05834574
Producer	0.5976366**	act2_b_days14	0.04775171
Importer	-0.004866111	act2_b_days21	0.04714715
days14_scrnCnt	0.7585896**	act1_day7	0.1526174*
days14_showCnt	0.7688824**	act2_day7	0.2111242**
days14_audiAcc	0.845056**	act1_day14	0.2553678**
days7_audiAcc	0.4722535**	act2_day14	0.3325477**
days7_scrnCnt	0.2871365**	act1_day1	0.005992385
days7_showCnt	0.3151093**	act2_day1	0.08952526
release_audiCnt	0.06506025		
release_scrnCnt	-0.01922876		
release_showCnt	0.0005978781		

significant code: **1%, *5%

주연배우에 대한 관심도를 나타내는 배우 검색량의 경우 개봉 후 1주 및 2주까지의 누적 검색량(sum_day7, sum_day14)이 종속변수와 유의미한 상관성을 보였다. 이외에 개봉 전 기간(b_days7, b_days14, b_days21)과 개봉 당일(day1)의 검색량은 상관성이 통계적으로 유의미하지 않았다. 이와 같은 분석결과를 요약하면 개봉 3주 차 누적 관객 수는 영화의 기본 속성 정보와 함께 개봉 후 2주 기간 동안의 상영 이력과 정보에 영향을 받고 있음을 알 수 있다.

3.3 예측 성능 분석

네 가지 기계학습 모형을 이용하여 10-fold cross validation을 수행한 결과를 <Table 4>에 요약하여 제시하였다. 실험결과 Random Forest Regressor, Random Forest Classifier, SVM, k-NN Regressor 순으로 예측 정확도가 우수한 것으로 확인되었다. 분류기반 모형과 회귀기반 모형에 대한 전반적인 성능을 비교해 볼 때 모두 Random Forest의 성능이 우수한 것으로 나타났다.

〈Table 4〉 Cross Validation Results

	Regression		Classification	
	k-NN	Random Forest	SVM	Random Forest
Average Accuracy	0.7727	0.944	0.8249	0.8576

3.3.1 회귀모형 기반 기계학습의 예측성능

<Table 5>는 회귀모형 기반의 두 가지 기계학습 모형인 Random Forest Regressor와 k-NN Regressor의 예측성능을 비교한 결과로 Random Forest Regressor의 예측 오차가 k-NN과 비교하여 매우 낮은 수준임을 알 수 있다. MAPE와 MASE를 대상으로 paired t-test를

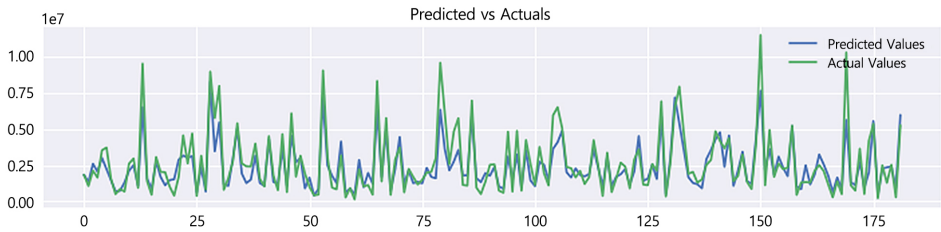
수행한 결과 Random Forest 모형의 예측 오차가 통계적으로 유의미하게 더 낮은 것을 확인하였다. 이와 같은 성능 차이는 상대적으로 입력변수의 수가 많고 데이터양이 적은 본 연구 데이터의 특성을 고려할 때 다수 샘플을 이용한 앙상블 기반의 Random Forest 모형이 단일 모형을 사용하는 k-NN과 비교하여 우수한 성능을 보인 것으로 판단된다.

참고로 두 모형의 예측 성능을 비교하기 위하여 영화 182편의 영화별 실제 총 관객 수와 예측 결과를 비교한 결과를 <Figure 2>에 제시하였다. 결과를 살펴보면 k-NN Regressor는 실제 값이 급격하게 증가하거나 감소하는 경우 이를 제대로 반영하지 못하고 과소 또는 과대 예측하는 경우가 많은 것으로 나타나고 있다. 이와 같은 결과는 유사한 조건의 과거 관객 수를 참조하는 사례 기반 학습 알고리즘인 k-NN의 특성에 따라서 비슷한 조건의 영화이지만 예상치 못하게 관객 수가 변화하는 경우 이를 적절하게 예측하지 못하게 된다.

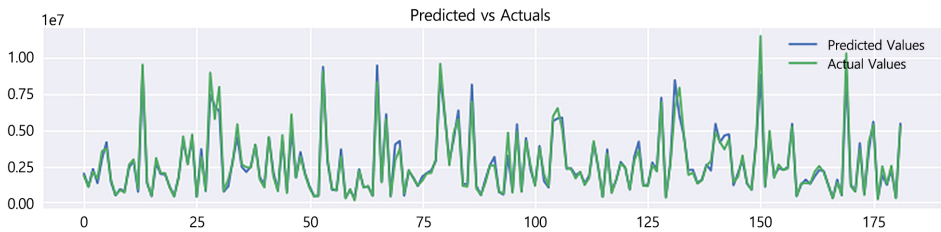
3.3.2 분류기반 기계학습의 예측성능

<Table 6>은 검증용 데이터를 대상으로 SVM과 Random Forest Classifier를 적용하여 분류를 실행한 결과를 정리한 confusion matrix이다. SVM과 Random Forest Classifier의 예측 성능을 비교했을 때, 전반적으로 Random Forest Classifier가 모든 범주에서 높은 분류 정확도를 보였다. 또한 적용 기계학습 모형과 관계없이 누적 관객 수가 적은 범주에 대하여 전반적으로 분류 성능이 우수하게 나타났다. 누적 관객 수가 400만 명~600만 명 구간을 의미하는 범주 C의 precision이 가장 낮은 수준을 보이고 있는데, 이는 해당 구간에 속하는 데이터가 다른 범주와 비교하여 상대적으로 매우 적어 학습이 적절하게 이루어지지 이유로 판단된다.

분류 성능을 한국영화와 외화로 구분하여 살펴보면 성능 차이를 확인할 수 있었다. 한국영화의 경우 누적 관객 수가 적은 범주 A의 분류 성능이 우수한 반면 외화의 경우 누적 관객 수가 많은 범주 C 또는 D에서 분류 성능이 가장



(a) k-NN Regressor



(b) Random Forest Regressor

〈Figure 2〉 Predicted and Actual Values of Regression Models

〈Table 6〉 Confusion Matrix of Classification Models

(a) SVM

	A	B	C	D	recall
A	22	1	0	0	95.65
B	5	14	2	0	66.67
C	0	3	5	0	62.5
D	0	0	1	2	66.67
precision	81.48	77.78	62.5	100	

(b) Random Forest Classifier

	A	B	C	D	recall
A	23	0	0	0	100
B	4	15	2	0	71.43
C	0	0	1	2	87.5
D	0	0	1	2	66.67
precision	85.19	93.75	70	100	

우수하였다. 외화의 경우 검증용 데이터가 많지 않아 실험 결과의 신뢰도가 다소 떨어진다는 한계가 있으나, 한국영화와 비교하여 정보가 제한되어 입력변수를 제한적으로 사용했음에도 불구하고 분류기반 기계학습 모형이 적절하게 적용될 수 있음을 확인하였다.

4. 결 론

본 연구는 영화 수익 개선을 위한 의사결정에서 중요한 정보이지만 높은 불확실성으로 정확한 예측이 어려운 개봉 영화의 누적 관객 수 예측 문제를 대상으로 기계학습 모형의 성능을 평가하였다. 특히, 기존 연구에서 주로 사용하던 분류기반의 기계학습 예측 모형(k-NN, Random Forest Regressor)과 함께 회귀모형 기반의 기계학습 모형(SVM, Random Forest Classifier)의 적용 타당성을 분석하였다. 또한 영화의 기본 속성 정보와 시계열 데이터로부터 관객 수 예측에 유의미한 입력변수를 선정하기 위한 데이터 처리 방법과 통계적 검정 분석 결과를 제시하였다.

본 논문의 분석결과 다음과 같은 몇 가지 흥미로운 결과를 도출하였다. 첫째, 개봉 후 3주차의 누적 관객 수 예측에서는 개봉 후 2주 동안의 정보(예: 관객수, 스크린수, 배우에 대한 관심도 등)가 유의미한 관계를 보였으며, 개봉 이전 및 개봉 당일의 정보에서는 유의미한 관계를 확인할 수 없었다. 둘째, 분류기반 모형과 비교하여 회귀모형 기반의 기계학습 모형이 보다 안정적인 예측 성능을 보였다. 또한 분류기반과 회귀모형 기반 기계학습 모형에서 모두 Random Forest 기법이 가장 우수한 예측 정확도를 보였다. 마지막으로, 분류기반 기계학습 모형의 경우 누적 관객 수가 적은 범주에 대하여 전반적으로 분류 성능이 우수하게 나타났다.

Random Forest 기법이 전반적으로 예측 성능이 우수한 결과는 입력변수가 많고 데이터가 적은 본 연구 데이터 특성에 따라서 다수 표본을 활용한 앙상블 기법 기반의 Random Forest 기법의 예측 정확도가 높게 나온 것으로 판단된다. 따라서 가용 정보가 제한적이고 입력변수가 적은 상황에서 동일한 연구결과를 기대할 수 있는지 추가 실험이 요구된다. 또한, 약 3년간 국내에서 개봉한 총 182편의 영화를 대상으로 분석을 수행하였으나 학습모형을 적용함에 있어 데이터가 충분하지 않은 한계가 존재하며, 향후 학습과 검증을 위한 데이터를 추가로 확보하여 실험을 진행하는 것이 필요하다.

References

[1] Abel, F., Diaz-Aviles, E., Henze, N., Krause, D., and Siehndel, P., "Analyzing the Blogosphere for predicting the success of music and movie products," International Conference on Advances in Social Networks Analysis and Mining, pp. 276-280, 2010.

[2] Breiman, L., Machine Learning 45:5. Kluwer Academic Publishers, 2001.

[3] Brewer, S. M., Kelley, J. M., and Jozefowicz, J. J., "A blueprint for success in the US film industry," Applied Economics, Vol. 41, No. 5, pp. 589-606, 2009.

[4] Chintagunta, P. K., Gopinath, S., and Venkataraman, S., "The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets," Marketing Science, Vol. 29, No. 5, pp. 944-957, 2010.

[5] Chong, M., "Evaluating real-time search query variation for intelligent information retrieval service," Journal of Digital Convergence, Vol. 16, No. 12, pp. 335-342, 2018.

[6] Delen, D., Sharda, R., and Kumar, P., "Movie forecast Guru: A Web-based DSS for Hollywood managers," Decision Support Systems, Vol. 43, No. 4, pp. 1151-1170, 2007.

[7] Demir, D., Kapralova, O., and Lai, H., "Predicting IMDB movie ratings using Google Trends," 2012.

[8] Eliashberg, J. and Shugan, S. M., "Film Critics: Influencers or Predictors?," Journal of Marketing, Vol. 61, No. 2, pp. 68-78, 1997.

[9] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M., and Brilliant, L., "Detecting influenza epidemics using search engine query data," Nature, Vol. 457, No. 19, pp. 1012-1015, 2009.

[10] Gong, J. J., Young, S. M., and der Stede, W. A. V.. "Real options in the motion picture industry: Evidence from film marketing and sequels," Contemporary Accounting Research, Vol. 28, No. 5, pp. 1438-1466, 2011.

[11] Gunn, S. R., "Support vector machines for classification and regression," ISIS technical report, Vol. 14, No. 1, pp. 5-16, 1998.

[12] Guo, Z., Zhang, X., and Hou, Y., "Predicting box office receipts of movies with pruned Random Forest," International Conference on Neural Information Processing ICONIP 2015: Neural Information Processing, pp. 55-62, 2015.

[13] Jung, J., Hwang, S., and Kwon, C., "Forecasting Korean Unemployment Rate with Web Queries," Korean Institute of Industrial Engineers, pp. 3373-3377, 2015.

[14] Kim, J. and Kim., J., "Relationship between Internet Buzz Share and Market Share : Movie Ticket Case", The Journal of Society for e-Business Studies, Vol. 18, No. 2, pp. 241-255, 2013.

[15] Kim, T., Hong, J., and Kang, P., "Box office

forecasting using machine learning algorithms based on SNS data," International Journal of Forecasting, Vol. 31, pp. 364-390, 2015.

[16] Koo, P. and Kim, M., "A Study on the Relationship between Internet Search Trends and Company's Stock Price and Trading Volume", The Journal of Society for e-Business Studies, Vol. 20, No. 2, pp. 1-14, 2015.

[17] Kwon, S. J., "Factors influencing Cinema Success: using News and Online Rates," Review of Culture & Economy, Vol. 17, No. 1, pp. 35-55, 2014.

[18] Lee, K. J. and Chang, W., "Bayesian belief network for box-office performance: A case study on Korean movies," Expert Systems with Applications, Vol. 36, pp. 280-291, 2009.

[19] Litman, B. R., "Predicting Success of Theatrical Movies: An Empirical Study," The Journal of Popular Culture, Vol. 16, No. 4, pp. 159-175, 1983.

[20] Lovallo, D., Clarke, C., and Camerer, C., "Robust analogizing and the outside view: two empirical tests of case-based decision making," Strategic Management Journal, Vol. 33, No. 5, pp. 496-512, 2012.

[21] Preis, T., Moat, H., and Stanley, H., "Quantifying trading behavior in financial markets using Google trends," Science Report, Vol. 3, p. 1684, 2013.

[22] Qin, L., "Word-of-Blog for movies: A predictor and an outcome of box office revenue?," Journal of Electronic Commerce Research, Vol. 12, No. 3, pp. 187-198, 2011.

[23] Ravid, S. A., "Information, blockbusters, and stars: A study of the film industry," The Journal of Business, Vol. 72, No. 4, pp. 463-492, 1999.

[24] Rogers, E. M., "New product adoption and diffusion," Journal of Consumer Research, Vol. 2, No. 4, pp. 290-301, 1976.

[25] Sawhney, M. S. and Eliashberg, J., "A parsimonious model for forecasting gross box-office revenues of motion pictures," Marketing Science, Vol. 15, No. 2, pp. 113-131, 1996.

[26] Sharda, R. and Delen, D., "Predicting box-office success of motion pictures with neural networks," Expert Systems with Applications, Vol. 30, pp. 243-254, 2006.

[27] Simonoff, J. S. and Sparraw, I. R., "Predicting movie grosses: winners and losers, blockbusters and sleepers," Chance, Vol. 13, No. 3, pp. 15-24, 2000.

[28] Siroky, D. S., "Navigating Random Forests and related advances in algorithmic modeling," Statistics Survey, Vol. 3, pp. 147-163, 2009.

[29] Song, J., Choi., K., and Kim. G., "Development of New Variables Affecting Movie Success and Prediction of Weekly Box Office Using Them Based on Machine Learning," Journal of Intelligent Information System, Vol. 24, No. 4, pp. 67-83, 2018.

[30] Subramaniyaswamy, V., Viginesh, V. M., Vishnu, P. R., and Logesh, R., "Predicting

movie box office success using multiple regression and SVM," 2017 International Conference on Intelligent Sustainable Systems(ICISS), pp. 182-186, 2017.

[31] Wang, F., Zhang, Y., Li, X., and Zhu, H., "Why do moviegoers go to the theater? The role of prerelease media publicity and online word of mouth in driving moviegoing behavior," Journal of Interactive Advertising, Vol. 11, No. 1, pp. 50-62, 2010.

[32] Wen, K. and Yang, C., "Determinants of the box office performance of motion picture in China-indication for Chinese motion picture market by adapting determinants of the box office(part II)," Journal of Science and Innovation, Vol. 1, No. 4, pp. 17-26, 2011.

[33] Yu, L., Zhao, Y., Tang, L., and Yang, Z., "Online big data-driven oil consumption forecasting with Google trends," International Journal of Forecasting, Vol. 35, pp. 213-223, 2019.

[34] Zhang, L., Luo, J., and Yang, S., "Forecasting box office revenue of movies with BP neural network," Expert Systems with Applications, Vol. 36, pp. 6580-6587, 2009.

[35] Zhang, W. and Skiena, S., "Improving movie gross prediction through news analysis," 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Workshops, pp. 301-304, 2009.

[36] Zhang, Z., Li, B., Deng, Z., Chai, J., Wang, Y., and An, M., "Research on movie box office forecasting based on internet data," 2015 8th International Symposium on Computational Intelligence and Design, 2015.

〈Appendix〉

〈Appendix Table〉 Description of Variables

	Variables	Description
Movie attributes and screening	actor1	Average number of previous movie audiences of actor1
	actor2	Average number of previous movie audiences of actor2
	director	Average number of previous movie audiences of director
	distributor	Average number of previous movie audiences of distributor
	producer	Average number of previous movie audiences of producer
	importer	Average number of previous movie audiences of importer
	release_audiCnt	Number of audiences on the day of release
	release_scrnCnt	Number of screens on the day of release
	release_showCnt	Number of screenings on the day of release
	days7_scrnCnt	Number of accumulated screens in 7 days
	days7_showCnt	Number of accumulated screenings in 7 days
	days7_audiAcc	Number of accumulated audiences in 7 days
	days14_scrnCnt	Number of accumulated screens in 14 days
	days14_showCnt	Number of accumulated screenings in 14 days
	days14_audiAcc	Number of accumulated audiences in 14 days
	days21_ audiAcc	Number of accumulated audiences in 21 days
Search volume of two lead actors	act1_sum_b_days7	Cumulated search volume of actor1 for 7 days before release
	act1_sum_b_days14	Cumulated search volume of actor1 for 14 days before release
	act1_sum_b_days21	Cumulated search volume of actor1 for 21 days before release
	act1_sum_day7	Cumulated search volume of actor1 for 7 days after release
	act1_sum_day14	Cumulated search volume of actor1 for 14 days after release
	act2_sum_b_days7	Cumulated search volume of actor2 for 7 days before release
	act2_sum_b_days14	Cumulated search volume of actor2 for 14 days before release
	act2_sum_b_days21	Cumulated search volume of actor2 for 21 days before release
	act2_sum_day7	Cumulated search volume of actor2 for 7 days after release
	act2_sum_day14	Cumulated search volume of actor2 for 14 days after release
	act1_day1	Search volume of actor1 on the day of release
	act2_day1	Search volume of actor2 on the day of release

저 자 소 개

정찬미 (E-mail: jpraise@ewhain.net)
2018년 울산대학교 글로벌경영학 (학사)
2018년~2020년 이화여자대학교 빅데이터분석학 (석사)
관심분야 수요예측, 빅데이터 분석

민대기 (E-mail: dmin@ewha.ac.kr)
1999년 서울대학교 산업공학과 (학사)
2001년 서울대학교 산업공학과 (석사)
2010년 퍼듀대학교 산업공학과 (박사)
2001년~2006년 LG CNS
2010년~현재 이화여자대학교 경영대학 부교수
관심분야 빅데이터, 강화학습, Stochastic programming, 전력 시스템