머신 러닝 기법을 활용한 박스오피스 관람객 예측

박도균[O], 백주련(교신저자)[*]
[O]평택대학교 데이터정보학과,
[*]평택대학교 데이터정보학과
e-mail: yangju0411@gmail.com[O], jrpaik@ptu.ac.kr[*]

Prediction of Movies Box-Office Success Using Machine Learning Approaches

Do-kyoon Park[O], Juryon Paik(corresponding author)[*]
[O]Dept. of Digital Information & Statistics, Pyeongtaek University,
[*]Dept. of Digital Information & Statistics, Pyeongtaek University

● 요 약 ●

특정 영화의 스크린 독과점이 꾸준히 논란이 되고 있다. 본 논문에서는 영화 스크린 분배의 불평등성을 지적하고 이에 대한 개선을 요구할 근거로 머신러닝 기법을 활용한 영화 관람객 예측 모델을 제안한다. 이에 따라 KOBIS, 네이버 영화, 트위터, 구글 트렌드에서 수집한 3,143개의 영화 데이터를 이용하여 랜덤포레스트와 그라디언트 부스팅 기법을 활용한 영화 관람객 예측 모델을 구현하였다. 모델 평가 결과, 그라디언트 부스팅 모델의 RMSE는 600,486, 랜덤포레스트 모델의 RMSE는 518,989로 랜덤포레스트 모델의 예측력이 더 높았다. 예측력이 높았던 랜덤포레스트 모델을 활용, 상영관을 크게 확보하지 못 했던 봉준호 감독의 영화 '옥자'의 상영관 수를 조절하여 관람객 수를 예측, 6,345,011명이라는 결과를 제시한다.

키워드: 박스 오피스(box-office), 머신 러닝(machine learning), 관람객예측(ticketing prediction)

I. Introduction

특정 영화의 스크린 독과점이 국내 영화의 생태계를 해친다는 문제가 꾸준히 화두에 오르고 있다. 과도한 상영관 몰아주기가 관객들의 영화 선택 권리를 침해한다는 것이다. 이에 따라 특정 영화의 상영관 독점을 막기 위한 스크린 상한제의 입법 필요성에 대한 목소리가 커지고 있다. 그러나 수요에 의해서 상영관을 많이 배정받는 영화들에 상한제가 도입되면 이로 인한 역차별로 관객들의 영화 선택 권리가 침해된다는 주장도 있다. 하지만 현재 영화의 상영관 배정은 수요-공급의 원리만으로 결정되는 것은 아니다. 2018년 기준으로 한국 영화관 시장의 96.9%를 세 개의 거대 업체가 차지하고 있으며 배급 시장의 92.1%를 10개 배급사가 차지하고 있다. 문제는 거대 영화관 업체와 계열사 관계에 있는 배급사들이 순위에 올라있다는 것이다[1]. 따라서 중소 배급사 영화는 스크린 확보가 쉽지 않다. 소비자의 영화 선택 권리와 중소규모 영화 시장의 성장을 위해서 상영관 수 선정에 필요한 기준 설정이 시급하다고 할 수 있다.

본 논문은 관람객 수를 예측하여 합리적인 상영관 수 할당에 적용할 수 있는 근거를 수립하고자 한다. 이를 위해 포털 및 SNS에서 데이터를 수집 후 설명 변수에 상영관 수를 반영하여 전체 관람객 수를 예측하는 모델을 세운다. 예측력 높은 모델이 수립되면 상영관을 많이 배정받지 못 한 영화들을 대상으로 상영관 수 변수를 조절, 관람객 수를 예측하여 상영관 할당에 대한 의견을 제출하는 근거로 사용하고자 한다.

II. The Proposed Scheme

1. 데이터

1.1 데이터 수집

본 논문에서 사용한 데이터는 KOBIS 영화관입장권통합전산망[2]에 오픈된 박스오피스 명단을 기본으로 하고 네이버 영화[3], 트위터[4], 그리고 구글 트렌드[5]에서 수집하였다. 데이터 처리와 분석의 용이함 때문에 수집된 데이터는 KOBIS에서 다운 받은 영화 명단 중 2010년 1월부터 2019년 9월까지의 수익을 목적으로 한 상업 영화로 한정하였다. VOD 판매를 목적으로 제작되어 실질적으로 극장 상영을 제대로 하지 않은 성인 영화들은 제거하였다.

KOBIS 오픈 데이터는 영화명, 감독명, 상영관 수, 관객 수, 장르, 시청 등급을 제공한다. 해당 데이터 외에 네이버 영화 플랫폼에서 주연 배우, 네티즌 평점, 평점을 준 사람의 수, 상영시간의 정보를 크롤링하여 추가하였으며, 트위터에서 개봉 전 날 영화 제목이 들어간 게시글 수와 구글 트렌드에서 개봉 전 날 구글에 검색된 영화 제목의 관심도를 크롤링하여 최종 데이터를 완성하였다. 변수의 이름은 효율적인 처리를 위해 모두 영어로 변환하였으며 Table 1은 사용된 변수들과 해당 변수들에 대한 설명을 보인다.

Table 1. 변수 설명

변수명	설명
영화명 title	영화의 한국어 제목
감독명 director	영화의 감독 이름, 범주형 변수
상영관 수 screen	영화의 상영관 수, 수치형 변수
관람객 수 popul	집계된 영화 관객 수, 수치형 변수
장르 genre	영화의 장르, 범주형 변수
시청 등급 rating	영화의 시청가능 연령, 범주형 변수
주연 배우 actor	영화의 주역 배우의 이름, 범주형 변수
평점 score	영화의 네이버 평점, 수치형 변수
평가자 수 netizen	영화의 평가한 사람 수, 수치형 변수
상영 시간 runtime	영화의 러닝 타임, 수치형 변수
트윗 수 tweet	개봉 전 날 올라온 트윗 수, 수치형 변수
구글 관심도 gtrend	개봉 전 날 구글 검색 관심도, 수치형 변수

1.2 데이터 탐색

*감독명*에 대한 3,143개의 관측값 중 중복 값을 제거한 고유 값은 총 2,053개로 정리되며, *주연 배우*에 대한 3,143개의 관측값 중 중복 값을 제거한 고유 값은 총 1,808개로 정리된다. *장르*는 총 19가지 장르로 구분되며 다음과 같다: '사극', '코미디', '판타지', '드라마', '액션', '어드벤처', '애니메이션', 'SF', '범죄', '전쟁', '미스터리', '멜로/로맨스', '스릴러', '공포(호러)', '뮤지컬', '다큐멘터리', '가족', '기타', '서부극(웨스턴)', '공연'. *시청등급*은 '전체관람가', '12세 이상 관람가', '15세 이상 관람가', 그리고 '청소년 관람 불가'의 4가지 등급으로 나뉜다. 사용된 수치형 변수의 통계량은 Table 2와 같다.

Table 2. 수치 변수의 요약 통계량

변수명	mean	stdev
상영관 수 screen	259.621	352.989
관람객 수 popul	586739.5	1562349
평점 score	6.513	2.143
평가자 수 netizen	2495.906	6136.752
상영 시간 runtime	100.781	24.111
트윗 수 tweet	118.718	349.421
구글 관심도 gtrend	32.6	39.076

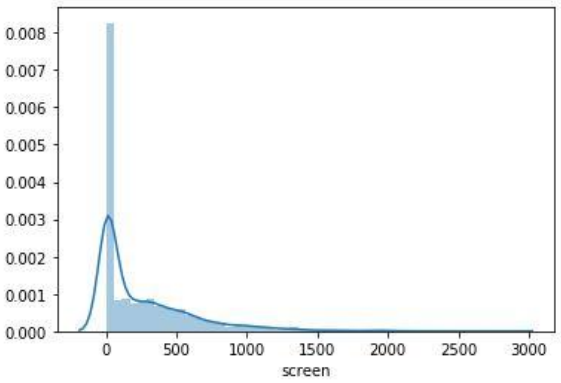

Fig. 1. 상영관 수 히스토그램

상영관의 분포는 Fig. 1에 제시된 것처럼 왼쪽으로 심하게 편중되어 있는 경향을 보였다. 상영관을 100개 이하로 할당받은 영화가 1587개로 전체의 50.5%를 차지하는 상영관 불균형 현상이 있었다. 반대로 오른쪽에는 크기가 큰 이상값을 가져 길게 꼬리를 늘어뜨린 것을 볼 수 있다. 이는 상영관 독과점 논란이 있었던 영화들이다.

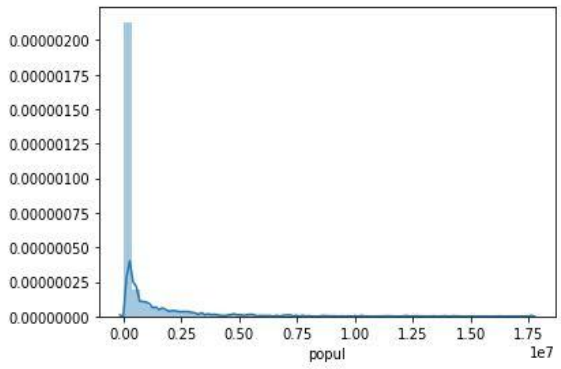

Fig. 2. 관람객 수 히스토그램

관람객의 분포는 왼쪽으로 심하게 솟고 오른쪽으로 긴 꼬리가 있는 모양을 보인다. 전체에 비하면 소수의 영화만이 흥행에 성공한

것이다. 상위 5%(157개)의 영화가 전체 관람객 수의 52.6%(18억 4천만 명 중 13억 4천만 명)를 차지했다.

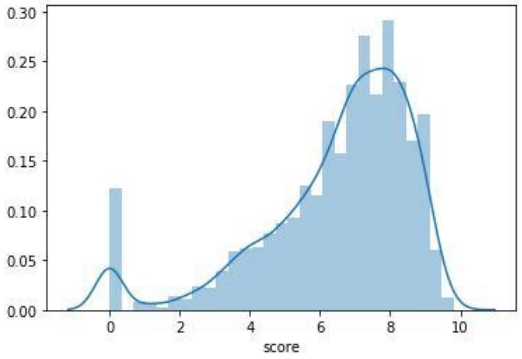

Fig. 3. 평점 히스토그램

평점은 1점부터 10점까지 네티즌들이 영화에 대한 평가를 메긴 것의 평균으로 평가가 된 적 없는 영화는 0점으로 나타내어져 있다. 차후 의사결정나무 기반의 모델로 예측을 할 것으로 결측값도 의미가 있기 때문에 별도의 결측값 대체는 하지 않는다.

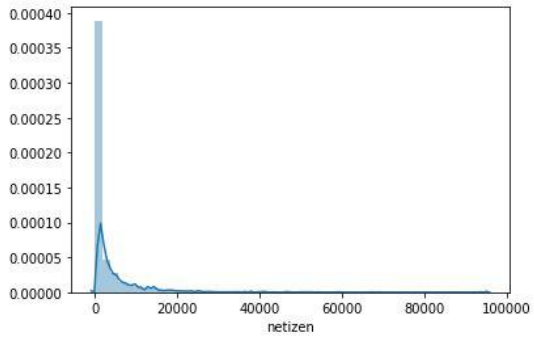

Fig. 4. 평가자 수 히스토그램

영화의 평점을 평가한 사람 수는 관람객 수와 비슷한 분포를 보였다. 관람객과의 상관계수도 0.89로 매우 높은 양의 상관관계를 보였다.

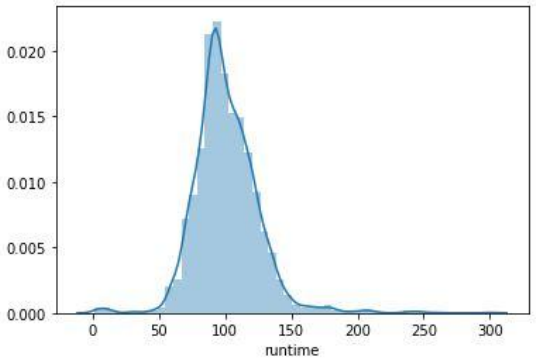

Fig. 5. 상영시간 히스토그램

상영시간의 분포는 비교적 대칭적이나 첨도가 높아 높게 솟은 모습을 보여준다.

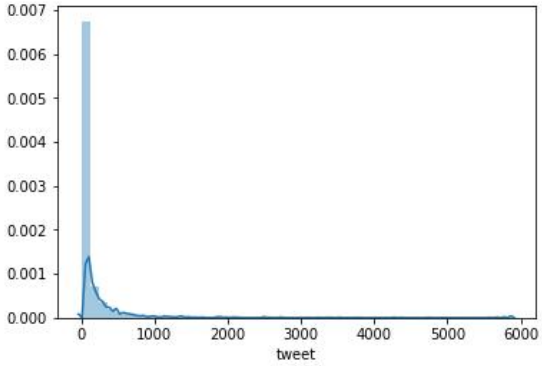

Fig. 6. 트윗 수 히스토그램

트윗 수의 분포 모양도 관람객 수와 비슷하나 관람객 수와 선형적인 상관관계가 보이지는 않았다.

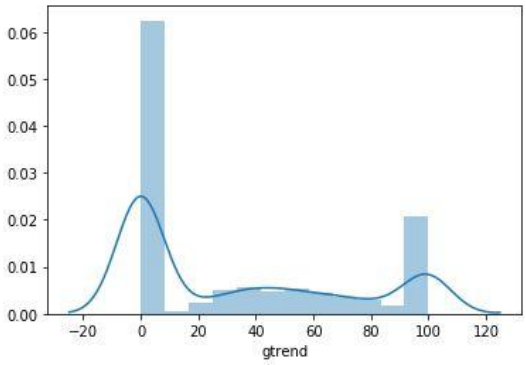

Fig. 7. 구글 관심도 히스토그램

구글 관심도는 양 극단에 자료의 분포가 집중되어 있었다.

1.3 데이터 처리

데이터 탐색 결과, 감독명과 주연의 범주는 고유한 값이 2,053명과 1,808명으로 매우 다양한 것을 알 수 있다. 따라서 이를 바탕으로 해당 감독과 주연이 영화 흥행 실적이 있는지 여부를 나타내는 변수를 새로 만들어 모델에 활용하였다. 영화 흥행실적 여부는 박스오피스 200위 안에 있는 영화를 연출하거나 출연한 적이 있는지를 기준으로 삼았다. 장르와 시청등급은 N개의 범주를 각 범주에 해당하는 지의 여부를 0과 1로 나타내는 N차원의 벡터로 표현하는 One-Hot-Encoding을 통해 각각 4개와 19개의 더미변수로 변경하여 주었고, 수치형 변수들은 그 분포가 매우 치우쳐져 있어서 데이터에 로그 변환을 해주었다.

모델을 학습하고 이를 평가하기 위해서 훈련용 데이터와 테스트용 데이터로 분리했으며 이 때 각 데이터 그룹에서의 데이터 분포 유지를 위해 관람객 수 기준 내림차순으로 정렬 후 계통추출법을 이용하여 데이터를 분리했다. 그 결과 훈련용(75%, 2357개)과 테스트용(25%, 786개)으로 나뉘어졌다.

2. 모델 수립, 예측 및 평가

2.1 모델링

예측에는 의사결정나무의 앙상블 기법인 랜덤포레스트[6]와 그라디언트 부스팅 기법[7]을 사용하였다. 랜덤포레스트는 데이터에서 중복을 허용하여 여러 개의 샘플을 추출하여 독립적인 의사결정나무를 학습시켜 예측한 후 평균을 내어 예측값으로 하는 기법이다. 그라디언트 부스팅은 부스팅 기법의 하나로 예측력이 약한 모델을 단계적으로 연결시켜 이전 모델에서 잘못 예측한 데이터에 가중치를 주어 학습, 오류를 최소화 시키는 방향으로 나아가며 예측력이 강한 모델을 만드는 기법이다.

2.2 모델 평가

학습용 데이터를 통해 모델을 학습 후 테스트용 데이터에 대해 예측 수행 후 테스트용 데이터의 실제 관람객 수와 비교하였다. 평가는 RMSE (Root Mean Squared Error : 평균 제곱근 오차)가 작을수록 예측이 잘 된 모델이라고 보았다. 그 결과 그라디언트 부스팅의 RMSE는 600,486이고 랜덤포레스트의 RMSE는 518,989으로 랜덤포레스트의 RMSE가 더 낮기 때문에 랜덤포레스트 모델의 예측력이 더 높았다.

2.3 상영관 수 조절을 통한 예측

대형 멀티플렉스 3사의 보이콧으로 인해 중소 상영관 111개만을 확보하여 상영한 봉준호 감독의 영화 '옥자'는 관람객 321,550명을 동원하였다. 만약 스크린 수를 증가하여 상영했다면 관람객 수는 어떻게 되었을까? 제안한 모델링 방법을 적용하여 예측해보았다. 영화 '옥자'의 상영관 수를 제외한 봉준호 감독의 평균 영화 상영관 수인 1,538로 조절하여 랜덤포레스트 모델을 통한 예측을 실시하였다. 그 결과 6,345,011명의 관객을 동원했을 것이라는 예측이 나왔다. 이 수치가 의미하는 것은 만약 대형 영화관에서 제대로 상영되었다면 많은 관람객을 동원하여 큰 수익으로 연결되었을 가능성이 높다는 것이다.

III. Conclusions

학습시킨 모델을 통해 상영관을 적게 할당 받은 옥자가 상영관을 제대로 받았을 경우의 흥행을 예측해 볼 수 있었으며 이를 근거로 상영관 배정에 대한 제안을 할 수 있을 것이다. 이런 사용 방법 이외에도 관람객 예측 모델은 영화 흥행 전략을 세우는 등 사용 주체와 목적에 따라 다양한 활용이 가능하기 때문에 한국 영화 시장 발전에 여러 방면으로 기여할 수 있을 것으로 기대된다. 본 연구에서는 네이버 영화, 트위터, 구글 트렌드를 크롤링했지만 차후 연구에서는 다른 플랫폼의 데이터를 크롤링하여 관람객 수에 영향을 미치는 데이터를 더 추가하고 랜덤포레스트 이외의 모델들을 비교하여 모델의 예측력을 더 높이는 방향으로 나아가고자 한다.

ACKNOWLEDGEMENT

이 논문은 2019년도 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2017R1A2B1007015).

REFERENCES

[1] KOFIC, "2018 Closing Report for Korean Movie Industry", pp. 37-39, Feb. 2019.
[2] KOBIS, http://www.kobis.or.kr/
[3] Naver Movies, https://movie.naver.com/
[4] Twitter, https://twitter.com/
[5] Google Trend, https://trends.google.co.kr/
[6] Breiman, L., "Random Forests, Machine Learning", Journal Machine Learning Vol. 45, No. 1, pp. 5 - 32, Oct. 2001.
[7] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", The Annals of Statistics Vol. 29, No. 5, pp. 1189-1232, Oct. 2001.