데이터마이닝과 텍스트마이닝을 활용한 영화 흥행 예측

조효정
성균관대학교 소프트웨어학과
jhjung07188@gmail.com

Box Office Hit Prediction Using Data mining and Text mining

Hyo-jung Jo
Dept. of Software, Sung-Kyun-Kwan University

요 약

영화 수익에 있어 영화의 흥행 여부는 중요한 영향을 끼친다. 영화 흥행 요인은 영화 산업의 규모가 커지면서 많은 제작사들 및 투자자들이 고려해야 하는 사항이 되었다. 따라서 영화의 흥행을 예측하기 위한 많은 모델이 연구되었다. 본 연구의 목적은 선행연구에서 흥행에 유의미한 영향을 끼친다고 밝혀진 스크린 수, 감독명, 제작사명 등의 내재적인 속성과 더불어 온라인 구전 변수를 사용하여 영화 흥행 예측 모델을 만드는 것이다. 이때 기사 수, 블로그 수와 같이 온라인 구전의 크기를 나타내는 변수들을 사용하는 대신 개봉 후 첫 주간의 관람객 리뷰를 텍스트마이닝을 이용하여 전체 리뷰 중 긍정 리뷰의 비율에 따라 점수를 매긴 후 독립변수로 사용한다. 그 후, 데이터 마이닝 기법을 활용하여 만든 모델에 앞서 언급한 독립변수를 입력 값으로 사용하여 영화의 흥행을 예측한다. 최종적으로 의사결정트리와 로지스틱회귀를 수행한 결과 영화 흥행에 영향을 주는 독립변수를 찾고 모델의 성능을 평가하였다. 로지스틱회귀의 결과 관객 수, 평점이 영화의 흥행에 특히 유의한 영향을 끼치는 변수로 선정되었고 리뷰 역시 유의한 변수로 선정되었다. 이때 만들어진 모델은 약 90%의 높은 수준의 정확도를 보여주었다. 의사결정트리의 결과 관객 수가 가장 중요한 변수로 선정되었다.

키워드: 영화 흥행 예측, 의사결정트리, 로지스틱회귀, 텍스트마이닝, 데이터마이닝

1. 서론

현재 영화 산업은 과거와 비교했을 때 그 규모가 상당히 크며 꾸준히 성장하는 추세이다. 실제로 한 기사의 영화진흥위원회 '2018 년 한국영화 결산 자료를 보면 지난해 한국영화 누적 관객은 2 억 1639 만 명으로 6 년 연속 2 억명을 넘겼으며 매출액은 1 조 8140 억원으로 역대 최고치를 경신하였다. 또한 미국 영화협회(MPAA)에 따르면 지난해 한국영화 시장 규모는 세계 5 위이며 세계 영화시장 전체 규모인 411 억 달러 중 16 억달러로 북미, 중국, 일본, 영국의 뒤를 잇고 있다.[1] 라고 밝히고 있다. 이렇게 영화시장이 커짐에 따라 영화를 흥행시키는 요인을 분석하고 흥행을 예측하는 것에 대한 중요성이 대두되었다. 영화의 흥행은 곧 영화의 수익성과 직결되고 수익을 내지 못하는 영화는 제작자에게 큰 금전적 부담을 안길 것이다. 제작비가 많이 투자된 영화로 개봉전에 입소문을 날렸지만 결국 흥행에는 실패한 영화가 있는 반면 제작비가 다른 영화에 비해 적게 들었어도 흥행에 성공하는 영화가 있으며 유명한 배우나 감독이 제작에 참여하지 않아도 흥행에 성공하는 영화가 있다. 이렇게 영화의 흥행은 제작비를 얼마나 들였는가, 혹은 누가 제작에 참여하였는 가와 같은 단순한 요인으로만 이루어지는 것이 아니다.

따라서 영화 흥행에 영향을 끼치는 요인을 분석하고 이를 통해 영화 흥행을 예측하는 연구가 많이 이루어졌다. 기존의 연구들에서는 영화의 내재적인 속성을 독립변수로 사용해서 예측모델을 만들거나 온라인 구전 변수들을 추가하여 예측모델을 만들었다. 이때 온라인 구전 변수로는 기사 수, 포탈의 평가자 수, 블로그 수, 평점 등의 수치를 나타내는 정보들을 주로 활용하였다. 혹은 관람객들의 리뷰, 즉 텍스트로 된 정보를 이용하여 영화의 흥행을 예측하는 데에 사용하였다. 이러한 연구들은 관람객의 리뷰 역시 영화 흥행의 요인에 해당한다는 사실을 밝혔다. 정회윤, 양형정 (2013)의 연구[2] 에서는 다중회귀 분석을 통해서 유의하다고 선정된 흥행 요소들만 고려하였고, 이

상훈, 조장식, 강창완, 최승배 (2015)의 연구[3] 에서는 관람객들의 리뷰를 사용해서 영화 흥행을 예측하는 모델을 제시하였다. 하지만 각각의 선행연구는 영화의 객관적인 정보와 관람객의 리뷰 같은 주관적인 정보를 개별적으로 생각하여 모델을 만들었다.

따라서 본 연구에서는 영화 흥행을 예측하는 모델을 만들기 위해서 스크린 수, 관객 수, 감독명, 배급사명과 같은 영화의 내재적 요인과 함께 개봉 후 첫 주 관람객이 남긴 리뷰를 텍스트마이닝을 이용하여 점수화 시킨 후 독립변수로 만들고 이를 데이터마이닝 기법을 통해 모델을 만들 때 사용하고자 한다. 이와 같은 데이터마이닝과 텍스트마이닝을 통합적인 접근은 영화 흥행 예측에 있어서 조금 더 유의미한 모델을 제시하고 영화 흥행에 어떤 요인이 주요한 영향을 미치는지 파악하는 데에 도움이 될 것이라고 기대하는 바이다.

2. 데이터 수집 및 전처리

본 연구에서 사용하는 관객 수, 스크린 수, 감독명, 제작사명, 매출액 데이터는 한국영화진흥위원회(KOFIC)에서 운영하는 통합전산망 사이트(KOBIS)에서 다운로드 받았다. 이때 한국 액션 영화 중에서 독립영화 및 예술영화를 제외하고 2010 년 1 월 1 일 이 후 2021 년 1 월 1 일 이전에 개봉한 영화를 대상으로 하였다. 이 결과 총 154 개의 영화 목록을 얻을 수 있었고 수집한 데이터의 통계기준일은 2021 년 1 월이다.

종속변수는 영화의 흥행여부로 매출액이 100 억 이상인 영화를 흥행한 영화, 그 이하인 영화를 흥행에 실패한 영화로 분류하였다. 감독명과 제작사명 변수는 각각 해당 영화 개봉 이전 필모그래피와 제작했던 영화 중 흥행한 영화가 몇 개인지를 기준으로 범주화하였다. 기존 흥행작이 5 개 이상인 경우는 10, 3~4 개인 경우는 7, 1~2 개인 경우는 5, 0 개인 경우는 0 의 4 그룹으로 나누었다.

평점, 리뷰 데이터는 네이버 영화에서 제공하는 것을 기준으로 하였고 이때 검색이 되지 않는 영화는 0 을 기입하였다. R 을 이용한 웹 크롤링을 통해서 리뷰 데이터를 수집하였으며 개봉 후 1 주일간의 리뷰만을 대상으로 하였다. 만약 리뷰의 수가 1000 개 이하인 경우는 감성분석을 할 때 충분한 양의 데이터가 제공되지 않아 특이값을 도출할 가능성이 높다고 판단하여 이 경우는 0 으로 설정했다. 이렇게 크롤링을 통해 수집한 리뷰들은 텍스트마이닝 기법 중 감성분석을 통해 다시 가공하였다. 감성분석 역시 R 을 이용하여 수행하였다. 긍정 및 부정어를 판단하기 위해 KNU 한국어 감성사전을 다운로드 받았고 리뷰의 특성상 인터넷 용어 및 은어, 줄임말 등의 기존 감성사전에 없는 단어가 많기에 이를 새로 추가하여 감성사전을 재구축하였다. 감성분석 결과 리뷰는 긍정, 부정, 중립으로 분류되었으며 단순히 개수로 비교하기에는 영화마다 리뷰 수가 천차만별이기 때문에 긍정 리뷰와 부정 리뷰의 비율을 비교하여 어떤 것이 더 높은 지를 판단하였다. 이렇게 얻은 데이터들은 모두 엑셀 시트에 정리하고 모델 학습 시 데이터셋으로 사용할 수 있도록 csv 파일로 저장하였다.

3. 모델 학습 및 결과 분석

우선 모델 학습을 위해서 데이터셋을 8:2 의 비율로 훈련 데이터, 테스트 데이터로 랜덤하게 나누었으며 로지스틱회귀와 의사결정트리를 수행하기 위해서 glm, rpart 패키지 및 필요한 패키지들을 설치해 주었다. 그 후 훈련 데이터를 사용해서 모델을 학습하였다. 이때 로지스틱회귀 방법으로 모든 변수를 사용한 모델과 backward selection 을 사용해서 유의하지 않은 변수를 제거하고 만든 모델, 총 두 모델을 만들었다. 먼저 모든 변수를 사용한 로지스틱 회귀 결과는 다음과 같다.

```
> summary(result)

Call:
glm(formula = 매출액 ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4447  -0.0426  -0.0019   0.0010   3.1809

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.890e+00  1.885e+00  -1.003  0.31602
감독         1.610e-01  2.227e-01   0.723  0.46965
제작사       2.322e-01  1.841e-01   1.261  0.20723
스크린수    -2.281e-03  2.330e-03  -0.979  0.32746
관객수       7.763e-06  2.572e-06   3.019  0.00254 **
평점        -1.464e+00  6.417e-01  -2.281  0.02256 *
리뷰         1.262e-01  9.163e-02   1.377  0.16861
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 170.116  on 122  degrees of freedom
Residual deviance:  23.347  on 116  degrees of freedom
AIC: 37.347

Number of Fisher Scoring iterations: 10
```

(그림 1) 로지스틱회귀 결과

coefficient 부분을 살펴보면 관객수와 평점의 p-value 가 0.00254, 0.02256 으로 0.05 보다 작아서 종속변수, 즉 영화의 흥행에 유의한 영향을 끼치고 있음을 확인할 수 있다. 특히 관객 수가 가장 영향을 많이 주는 것으로 보인다. 다른 독립변수 중 리뷰 역시 상대적으로 낮은 p-value 를 보이고 있는데 이는 리뷰 변수 역시 영화 흥행에 있어서 어느 정도의 영향력을 미치고 있는 것을 의미한다. 이때 중요하지 않은 변수를 제거한 모델을 만들기 위해서 backward selection 를 수행했다. 그 결과 관객 수, 평점, 리뷰의 세 가지 변수

만이 유의한 변수로 선정되었고 이를 사용해서 다시 로지스틱회귀 모델을 만들었다. 이렇게 만든 모델의 성능을 평가하기 위해 predict() 함수를 사용하여 만든 모델에 테스트 데이터를 넣고 도출된 예측값과 실제 종속변수의 값을 비교하였을 때 약 0.9 의 높은 정확도를 보였다.

두번째로 의사결정트리를 통해서 모델을 학습한 결과는 다음과 같다.

```
> summary(tree)
Call:
rpart(formula = 매출액 ~ ., data = train)
  n= 123

         CP nsplit  rel error     xerror       xstd
1 0.9655172      0 1.00000000 1.00000000 0.09545312
2 0.0100000      1 0.03448276 0.03448276 0.02418394

Variable importance
  관객수 스크린수   제작사     평점     감독     리뷰
      33       22       13       13       10        9

Node number 1: 123 observations,    complexity param=0.9655172
  predicted class=0  expected loss=0.4715447  P(node) =1
    class counts:    65    58
   probabilities: 0.528 0.472
  left son=2 (67 obs) right son=3 (56 obs)
  Primary splits:
      관객수    < 1377218 to the left,  improve=57.420220, (0 missing)
      스크린수 < 508.5   to the left,  improve=32.069010, (0 missing)
      제작사    < 2.5     to the left,  improve=15.374020, (0 missing)
      평점      < 6.17    to the left,  improve=11.186190, (0 missing)
      감독      < 2.5     to the left,  improve= 9.002029, (0 missing)
  Surrogate splits:
      스크린수 < 508.5   to the left,  agree=0.854, adj=0.679, (0 split)
      제작사    < 2.5     to the left,  agree=0.732, adj=0.411, (0 split)
      평점      < 6.64    to the left,  agree=0.724, adj=0.393, (0 split)
      감독      < 2.5     to the left,  agree=0.691, adj=0.321, (0 split)
      리뷰      < 16.665  to the left,  agree=0.667, adj=0.268, (0 split)

Node number 2: 67 observations
  predicted class=0  expected loss=0.02985075  P(node) =0.5447154
    class counts:    65     2
   probabilities: 0.970 0.030

Node number 3: 56 observations
  predicted class=1  expected loss=0  P(node) =0.4552846
    class counts:     0    56
   probabilities: 0.000 1.000
```

(그림 2) 의사결정트리 결과

Variable importance 부분을 살펴보면 관객수와 스크린 수가 가장 높고 리뷰가 가장 낮게 나타난다. 앞서 수행한 로지스틱회귀와 같이 관객수가 가장 유의한 변수라는 것을 확인할 수 있었다. 하지만 평점과 리뷰가 영화의 흥행에 유의미한 변수로 나타난 로지스틱회귀와는 달리 의사결정트리에서는 관객 수와 스크린 수가 유의한 변수로 나타났다. 또한 primary splits 와 surrogate splits 를 통해 어떤 기준으로 모델이 데이터를 분류하였는지 확인할 수 있는데 관객수가 1.4e+6 보다 작으면 영화 흥행 실패 (54%), 그보다 크면 영화 흥행 성공 (46%)으로 구분되었다. 마지막으로 이 모델의 성능을 평가하기 위해서 오분류표를 만들었다. 이때 모델에 사용한 데이터가 적어 앞서 언급한 기준만으로 모든 데이터가 정확히 분류되었다. 따라서 정확도는 높지만 이것이 신뢰성 있는 모델이 만들어졌다는 것을 보장한다고 말하기는 어렵다고 볼 수 있다.

4. 결론 및 소감

본 연구를 진행하면서 얻은 결론은 다음과 같다. 로지스틱회귀와 의사결정트리를 통해 공통적으로 관객 수가 영화 흥행에 유의한 영향을 끼치는 요인이라는 것을 확인할 수 있었다. 특히 로지스틱회귀는 관객수와 더불어 평점 역시 영화 흥행에 유의한 영향을 끼친다는 것을 밝혔다. 또한 backward selection 을 사용하여 의미 없는 변수를 제거하였을 때 관객 수, 평점, 리뷰가 유의한 변수로 남게 되었고 이는 리뷰 역시 영화 흥행에 있어서 영향을 미치는 요소라는 것을 알 수 있게 해주었다. 의사결정트리는 관객수의 영향력이 너무 커서 모델을 만들었을 때 그 정확도가 지나치게 높게 나타났다. 이는 모델 학습에 사용한 데이터 양의 부족으로 인해서 발생한 결과라고 생각되며 더 많은 데이터를 사용하면 관객 수 뿐만 아니라 다른 독립변수들도 데이터 분류 기준으로 사용될 것이라고 보인다.

본 연구를 진행하면서 아쉬웠고 부족하다고 느꼈던 점은 웹 크롤링 기법을 사용해서 관람객의 리뷰를 수집하는 것은 성공하였지만 오타, 비속어, 줄임말, 신조어 등 기존 감성사전에 없었던 어휘가 많아 많은 리뷰가 중립으로 분류되었다는 것이다. 또한 소위 댓글알바라고 불리는 리뷰 조작으로 인해서 평점이 낮아도 긍정적인 리뷰가 많아 감성분석을 했을 때 실제와 다른 결과가 도출되는 경우도 있었다. 감성사전에 좀 더 많은 어휘를 추가하고 평점과 감성분석 결과의 괴리가 지나치게 큰 영화에는 주의를 기울여야겠다는 것을 느꼈다. 또한 시간 부족과 개인 능력의 한계로 인해서 완벽하게 기대한 결과를 얻지는 못했지만 인터넷 상에서 다운로드 받을 수 있는 데이터베이스와 같이 쉽게 얻을 수 있는 데이터가 아닌 직접 영화 흥행에 어떠한 요인들이 영향을 미칠 것인가를 판단하고 관련 데이터를 수집 및 가공하면서 기를 수 있었던 데이터 처리 능력은 추후에 연구 및 과제를 수행할 때 도움이 많이 될 것이라고 생각한다.

참고문헌

[1] 김시균, “할리우드에도 K 무비…한국영화 관객 2억 세계 5 위 시장으로”, 매일경제, 2019 년 4 월 28 일, https://www.mk.co.kr/news/culture/view/2019/04/270213/” (2020 년 09 월 14 일)

[2] 정회윤, 양형정, “다중회귀 분석을 이용한 영화 흥행 예측”, 한국컴퓨터정보학회 학술발표논문집, 21 권, 2 호, 275-278 쪽, 2013 년

[3] 이상훈, 조장식, 강창완, 최승배, “텍스트 마이닝을 활용한 영화흥행 예측 연구”, 한국데이터정보과학 회지, 26 권, 6 호, 1259-1269 쪽, 2015 년