석사학위 논문

데이터 분석을 활용한 한국 영화 흥행 예측

Prediction of Financial Success Using Data Analysis for Korean Movies

2018년 6월

숭실대학교 소프트웨어특성화대학원

소프트웨어전공

김 세 윤

석사학위 논문

데이터 분석을 활용한 한국 영화 흥행 예측

Prediction of Financial Success Using Data Analysis for Korean Movies

2018년 6월

숭실대학교 소프트웨어특성화대학원

소프트웨어전공

김 세 윤

석사학위 논문

데이터 분석을 활용한 한국 영화 흥행 예측

지도교수 신 용 태

이 논문을 석사학위 논문으로 제출함

2018년 6월

숭실대학교 소프트웨어특성화대학원

소프트웨어전공

김 세 윤

김 세 윤 의 석 사 학 위 논 문 을 인 준 함

심 사 위 원 장 최 용 락 인

심 사 위 원 박 제 원 인

심 사 위 원 신 용 태 인

2018년 6월

숭실대학교 소프트웨어특성화대학원

목 차

표 목 차

그 림 목 차

국문초록

데이터 분석을 활용한 한국 영화 흥행 예측

김세윤

소프트웨어전공

숭실대학교 소프트웨어특성화대학원

영화진흥위원회(KOFIC)의 2017년 한국 영화산업 결산 보고서에 따르면 2012년 한국영화 수익률은 15.9%로 수익률이 흑자로 돌아선 이후 현재까지 흑자를 유지하고 있다. 하지만 제작비가 100억 원 이상 들어간 영화들이 흥행에 실패하는 등 개별 영화들의 수익률은 그렇지 않다. 영화의 수익은 크게 두 가지로 볼 수 있는데, 상영 중의 영화관 입장권 매출액과 IPTV(Internet Protocol Television) 및 디지털 케이블 TV의 매출액이다. 그 중 영화관 입장권 매출의 극대화를 위하여 제작사, 배급사는 수요를 예측하고 그에 따른 전략을 세울 수 있어야한다. 하지만 영화는 대표적인 경험재이기 때문에 흥행을 예측하는 것이 쉽지 않다. 따라서 본 논문에서는 영화진흥위원회에서 제공하는 객관적인 데이터를 이용하여 흥행 예측 모델을 선정하고 나이브 베이즈 분류(Naïve Bayes Classification)를 이용하여 흥행을 예측해보고자 하였다.

ABSTRACT

Prediction of Financial Success Using Data Analysis for Korean Movies

KIM, SE YOON
Major in Software
Graduate School of Software Soongsil University

According to the Korea Film Council(KOFIC) 2017 Korea Film Industry Report, the Korean movie industry has a profit rate of 15.9% in 2012. However, the profitability of individual movies is not such that movies with production costs over 10 billion won fail to hit the market. The revenue of movies can be seen in two big ways: sales of admission tickets to movie theaters and sales of IPTV(Internet Protocol Television) and digital cable TV. In order to maximize the sales of admission tickets to the theater, producers and distributors should be able to predict demand and develop strategies accordingly. However, it is not easy to predict a movie because it is a representative experience. Therefore, in this paper, I use the Naïve Bayes Classification to predict the box office by using the objective data provided by the Korea Film Council.

제 1 장 서 론

1.1 연구 배경 및 목적

방대한 양의 데이터 발생과 데이터 분석 기술의 발달에 따라 예측이 필요한 영화 산업을 비롯한 다양한 분야에서 데이터를 이용한 연구가 활발하게 진행되고 있다.

영화진흥위원회에서 발간한 2017년 한국 영화산업 결산에 따르면 2017년 극장 시장은 관객 수 2억 1,987만 명으로 전년 대비 1.3% 증가했고, 매출액은 1조 7,566억 원으로 0.8% 증가했다. 또한 인구 1인당 연평균 관람횟수는 4.25회로 세계 최고 수준을 유지하고 있다.[1] [표 1-1]은 역대 극장 매출액과 총 관객 수를 보여준다. 그러나 순 제작비 50억 원 이상의 고 예산 영화의 흥행부진으로 수익률은 전년도에 비해 하락하였고, 개별 영화의 수익률을 따져보면 산업이 전반적으로 성장 중이라 할 수 없다.

[표 1-1] 2011-2017 극장 매출액, 총 관객 수

구분 \ 년도	2011	2012	2013	2014	2015	2016	2017
극장 매출 (억 원)	12,358	14,551	15,513	16,641	17,154	17,432	17,566
총 관객 수 (만 명)	15,972	19,489	21,335	21,506	21,729	21,702	21,987

1) 영화진흥위원회. 2017년 한국 영화산업 결산 p.14.

극장 개봉하는 영화의 상영 중 매출액은 극장 관람객의 수로 파악할 수 있다. 따라서 수익의 극대화를 위해서는 흥행 예측 연구를 통해 개봉 스크린 수 조정 또는 마케팅 전략 수정 등의 과정이 필요하다. 따라서 본 논문에서는 총 관객 수를 이용하여 일정 기간 동안의 영화들의 흥행 여부를 예측하여 영화산업에 기여하고자 한다.

1.2 연구 방법

영화의 흥행 예측을 위하여 2011년부터 2017년 11월까지 국내 극장 개봉한 한국 국적의 일반 상업영화의 데이터를 수집, 정제한다. 회귀분석을 통하여 흥행에 유의미한 영향을 미치는 요인에 대해 분석한 후, 그를 바탕으로 나이브 베이즈 분류(Naïve Bayes Classification)를 통해 2017년 12월부터 2018년 4월 사이에 개봉한 영화의 흥행을 예측하여 실제의 결과와 비교한다.

1.3 논문 구성

본 논문의 구성은 다음과 같다. 1장에서는 서론을 서술하고, 2장에서는 관련연구를 서술한다. 그리고 3장에서는 흥행에 유의미한 영향을 미치는 요인을 분석한 후, 흥행 예측을 연구한다. 끝으로 4장에서는 본 연구의 결론 및 향후 연구 과제와 방향을 제시한다.

제 2 장 관련 연구

2.1 회귀분석(Regression Analysis)

두 변수 사이의 관계식을 파악하여 한 변수의 값으로부터 다른 변수의 값에 대한 예측을 필요로 하는 경우에 두 변수 사이의 함수 관계에 대한 통계적 분석을 하는 방법을 회귀분석(Regression Analysis)이라고 한다.[2] 이 때 독립변수가 하나인 분석방법을 단순회귀분석이라 하고, 둘 이상의 독립변수를 고려하는 경우 다중회귀분석이라 한다. 여기에서 두 변수 사이의 관계식이 선형함수일 경우 선형회귀라고 하며 그 외의 경우 비선형회귀라고 한다. 본 논문에서는 독립변수가 여러 개인 다중회귀분석을 진행한다.

R을 이용하여 회귀분석을 하면 결과를 쉽게 확인할 수 있다. [그림 2-1]은 R을 이용하여 단순회귀분석을 실시한 결과의 한 예를 보여준다.

2) 영지문화사. 일반통계학 p.334.

```
Call:
lm(formula = total_people ~ grade)

Residuals:
    Min      1Q  Median      3Q     Max
-2093962 -164568 -164437 -163120 15519719

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1840814     185384   9.930  < 2e-16 ***
grade15        253149     218853   1.157    0.248
gradeAll     -1639396     301300  -5.441 6.69e-08 ***
grade19      -1676245     197944  -8.468  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1729000 on 978 degrees of freedom
Multiple R-squared:  0.2001,    Adjusted R-squared:  0.1976
F-statistic: 81.55 on 3 and 978 DF,  p-value: < 2.2e-16
```

[그림 2-1] R을 이용한 단순회귀분석 결과의 예

[그림 2-1]의 첫 번째 줄은 회귀분석을 실시한 변수들을 보여준다. 이 결과를 이용하여 회귀식을 추정할 수 있는데, 그것은 식 (1)과 같다.

$$\text{total\_people} = 1840814 + (253149 \times \text{grade15}) + (-1639396 \times \text{gradeAll}) + (-1676245 \times \text{grade19}) \quad (1)$$

[그림 2-1]의 하단의 Multiple R-squared과 Adjusted R-squared은 결정계수로 모형 선택의 기준으로 사용할 수 있다. 결정계수는 1에 가까울수록 회귀식이 적합하다고 할 수 있다. 결정 계수는 설명변수의 개수가 많을수록 더 커지는 경향이 있기 때문에 이를 보완한 계수가 수정된 결정 계수이다. 이때 둘 사이의 차이가 클 경우 회귀 모형을 재검토가 필요하다.

[그림 2-1]의 맨 아랫줄의 F 통계량의 p-value값이 2.2e-16으로 0.05보다 작기 때문에 위 회귀식은 모델 전체에 대하여 통계적으로 의미 있다고 볼 수 있다[3, 4].

2.2 나이브 베이즈 분류(Naïve Bayes Classification)

나이브 베이즈 분류는 베이즈 정리에 근거한 확률 분류기로 주로 스팸 메일 분류에 사용되며 영화 흥행 예측 분야에서도 많이 사용되는 방법이다. 나이브 베이즈 분류에서는 데이터의 모든 특징들이 독립적이라고 가정하는데, 분류 학습에서 매우 정확한 결과를 보여준다.

베이즈 정리에 따르면 사건 B가 사건 A에 속할 조건부 확률은 식 (2)와과 같은 수식으로 계산할 수 있다.

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)} \qquad (2)$$

이 정리를 나이브 베이즈 분류에 적용하면 사건 B는 i개의 특성을 나타내는 벡터 X로 표현할 수 있으며 이 벡터를 이용하여 n개의 가능한 확률적 결과를 식 (3)과 같이 표현할 수 있다.

$$p(A_n|x_1, x_2, x_3, .., x_n) \qquad (3)$$

i의 수가 많은 경우 베이즈 정리를 바로 적용하기 어렵기 때문에 베이즈 정리와 조건부 확률을 이용하여 식 (4)와같이 표현할 수 있다.

$$P(A_n|\chi) = P(A_n)\frac{P(\chi|A_n)}{P(\chi)} \qquad (4)$$

식 (4)는 $P(A_n|\chi)$값이 사건 χ가 A_n에 속할 확률을 의미하기 때문에 패턴 분류에 적용할 수 있다고 할 수 있다. A_n에 속한 χ는 통계적으로 독립적이라 가정하여 식(5)와 같이 표현할 수 있다. 여기서 k는 차원을 뜻한다.

$$P(\chi|A_n) = \prod_{i=1}^{k} P(x_i|c_i) \tag{5}$$

식(4)와 식(5)를 이용하면 나이브 베이즈 분류를 식(6)과 같이 표현할 수 있다. [5, 6, 7]

$$n = \arg max_n P(A_n) \prod_{i=1}^{k} p(x_i|c_i) \tag{6}$$

2.3 예측 모델에 관한 연구

이전의 연구를 살펴보면, 영화의 흥행을 예측하기 위해서 다양한 변수들을 수집하여 여러 가지 분석 방법을 이용한다. 이렇게 분석한 내용으로 어떠한 변수가 영화 흥행에 영향을 주는지 알아본다.

본 논문에서는 변수 하나하나의 영향력을 분석하기보다 여러 변수들을 다양하게 조합하여 분석하는 방법을 택한다. 이때의 변수 조합을 예측 모델이라 한다.

[8]의 연구에서는 회귀분석을 통해 예측 모델을 분석하였다. 전국 관객수를 기준으로 각 년도 별 상위 100위의 영화를 분석 대상으로 한정하였다. 독립변인으로는 영화의 장르, 영화의 등급, 영화의 국적, 감독, 배우, 속편여부, 배급사, 개봉 스크린 수, 제작비, 개봉 시기, 전문가와 관객의 평점을 설정하였다. 분석 결과 흥행에 영향을 미친 요인으로는 스크린 수, 온라인 평점, 배급사, 개봉 시점이 있었다. 또한 이 연구는 한국 영화 시장을 분석한 결과이기 때문에 외국 영화의 결과도 확인할 수 있다. 외국 영화의 경우 스크린 수, 온라인 평점, 배급사에 영향을 받는다는 결과를 얻었다.

[9]의 연구에서도 회귀분석을 통해 예측 모델을 분석하였다. 2010년 한국에서 개봉된 영화를 대상으로 흥행에 영향을 미치는 요인을 영화의 제작, 배급, 상영 단계별로 분석하였다. 영화 흥행에 관한 변수는 영화의 장르, 영화의 등급, 감독, 배우, 배급사, 스크린 수, 온라인 평점, SNS로부터 생성되는 온라인 버즈의 크기를 이용했다. 분석 결과 흥행에 영향을 미친 요인으로는 영화의 장르, 영화의 등급, 배우, 스크린 수, 배급사, SNS로부터 생성되는 온라인 버즈의 크기가 있었다.

[10]의 연구에서는 변량분석을 통하여 예측 모델을 분석하였다. 이 연구에서는 배우, 감독, 제작사, 영화의 등급, 영화의 개봉시기, 영화의 장

르는 영화 흥행에 영향을 끼칠 것이라 가정한 후 개별 영화의 관객 동원을 종속변수로 하여 변량분석 하였다. 분석 결과 흥행에 영향을 미친 요인으로는 배우, 영화의 등급, 영화의 장르가 있었다.

[11]의 연구에서는 회귀분석을 통하여 예측 모델을 분석하였다. 이 연구에서는 분석을 위해 배우, 감독, 제작비, 영화의 등급, 영화의 개봉 시기, 영화의 장르를 변수로 정하였다. 분석 결과 흥행에 영향을 미친 요인은 제작비, 개봉스크린 규모, 전문가 평점, 영화의 개봉시기가 있었다.

2.4 흥행 예측에 관한 연구

앞 절에서 살펴본 예측 모델에 대한 연구 결과를 응용하여 흥행을 예측한 연구들이 있다. 과거의 연구는 여러 가지 예측 방법의 예측률 비교 연구, 같은 날 개봉하는 영화의 흥행을 비교 예측하는 연구가 있다.

2.4.1 예측률 비교 연구

[8]의 연구에서는 회귀분석을 통해 얻은 예측 모델을 이용하여 연구를 진행하였다. 회귀분석을 통해 흥행에 영향을 미치는 요인을 선정하여 다항로짓모형, 인공신경망모형, 판별분석으로 예측률을 비교한 결과 인공신경망(94.1%)과 다항로짓모형(95.1%) 모두 전반적으로 우수하다는 결과를 얻었다.

[12]의 연구에서는 선형회귀분석을 통해 얻은 예측 모델을 이용하여 연구를 진행하였다. 선형회귀분석을 통해 얻은 결과를 인공신경망과 의사결정나무에 반영하여 예측률을 비교한 연구결과는 크게 두 가지로 볼 수 있었다. 먼저 선형회귀분석결과를 반영하지 않은 결과는 인공신경망이 73.53%로 68.40%의 의사결정나무보다 성능이 더 좋은 것으로 나타났다. 선형회귀분석결과를 반영한 결과 또한 인공신경망이 71.31%로

70.11%의 의사결정나무보다 성능이 더 좋은 것으로 나타났다.

2.4.2 같은 개봉일 영화의 흥행 예측 연구

[13]의 연구는 Stylometry라는 텍스트 마이닝 기법으로 먼저 개봉한 <명량>의 온라인 댓글 데이터를 기반으로 같은 날에 개봉할 두 영화를 비교하였다. <명량>의 개봉일 이전의 댓글과 개봉 후 5일간의 댓글을 수집하여 분석한 후 같은 날 개봉하는 <제보자>와 <슬로우 비디오>의 댓글과 비교 분석한다. <명량>의 댓글과 일치 단어 빈도수는 제보자가 66회로 46회의 <슬로우 비디오>보다 높게 나타났다. 또한 가장 빈번하게 사용된 주요 단어들의 분포를 분석한 결과 또한 <제보자>의 댓글이 <명량>과 유사함을 알 수 있었다. 따라서 <제보자>가 <슬로우 비디오> 보다 더 많은 관객 수를 얻을 것이라고 예측하였다. 실제 최종 관객 수는 <제보자>가 1,755,181명으로 <슬로우 비디오>보다 더 많은 관객 수를 얻었고 분석 결과와 일치함을 알 수 있었다.

제 3 장 데이터 분석을 이용한 영화 흥행 예측

3.1 흥행의 기준

한국 영화 산업에서의 흥행 여부에 대한 판단은 관객 수를 기준으로 한다. 미국을 비롯한 중국, 일본 등의 국가에서는 흥행 여부를 가리기 위해 극장 매출액을 기준으로 하지만 우리나라에서는 그렇지 않다[14]. 현재는 영화진흥위원회에서 영화관 입장권 통합전산망으로 실시간으로 집계하기 때문에 개별 영화의 매출액을 쉽게 파악할 수 있지만 과거에는 서울의 관객 수 만을 집계하였고, 극장별 영화관 입장권의 가격도 제각각이었기 때문에 파악할 수 없었다. 또한 시스템을 통해 수집된 데이터가 아닌 개별 극장에서 발표한 자료로 만들어진 매출액이기 때문에 객관적일 수 없었다. 따라서 한국 영화 산업에서는 관객 수를 기준으로 흥행 여부를 판단하고 있다. 또한 매출액을 알아냈다고 해서 개별 영화에 대한 손익분기점을 파악할 수도 없다. 영화를 만드는데 쓰인 제작비 정보를 알아낼 수 없기 때문이다. 제작비를 따로 공개하지 않고, 공개하더라도 그 수는 전체 영화에 비하면 극히 일부이다. 과거의 한 연구에서는 잡지나 기사를 통해 먼저 수집하고, 그 외의 경우에는 제작사나 배급사, 투자자들에게 개별 연락하여 제작비 정보를 수집하였으나 전체 영화의 제작비 정보는 얻을 수 없었다[15, 16]. 따라서 본 논문에서는 통합적인 흥행의 기준으로 총 관객 수 정보만 이용한다.

본 논문에서 흥행의 기준으로 이용하기 위하여 사용하는 데이터는 영화진흥위원회에서 제공하는 역대 박스오피스 내역으로 1971년부터 현재까지 국내 개봉한 한국 상업영화 전체 관객 수 순위를 기준으로 상위 500편을 제시한다. [표 3-1]은 역대 박스오피스 내역의 통계를 보여준다.

2014년 개봉한 <명량>이 총 관객 수 17,613,682명으로 전체 1위이고, 2011년 개봉한 <고양이: 죽음을 보는 두 개의 눈>이 총 관객 수 672,071명으로 전체 500위이다. 이 500편의 관객 수의 평균은 약 2,650,000명으로 본 논문에서는 총 관객 수가 2,650,000명 이상일 경우 흥행한다고 판단한다.

[표 3-1] 역대 박스오피스 내역 통계

(단위: 명)

	평균	최소 관객 수	최대 관객 수
관객 수	2,657,711	672,071	17,613,682

3.2 분석 대상의 선정

분석 대상은 2011년부터 2017년 11월 사이에 개봉한 한국영화 1,118편이다. 영화진흥위원회의 영화관입장권통합전산망에서 대상 영화의 배우, 감독, 관객 수, 스크린 수, 배급사 등의 데이터를 수집, 정제한다. 2011년 이전의 데이터는 전국 극장의 데이터를 가지고 있지 않거나 배급사를 통해 확인한 데이터이기 때문에 객관적이지 않다고 판단하여 제외하였다. 또한 제작국가가 대한민국이지만 출연배우가 외국인인 경우 분석 대상에서 제외하였고, 출연배우의 정보를 정확히 알아낼 수 없는 경우도 제외하였다. 하지만 다수의 배우가 대한민국 국적이고, 감독 또한 대한민국 국적인 경우 외국인 배우 데이터만을 제외하였다. 이 데이터를 이용하여 흥행에 영향을 미치는 요인들을 분석한 후 2017년 12월부터 2018년 4월까지 국내 개봉하는 한국영화의 흥행을 예측한다.

3.3 데이터 수집 및 정제

분석에 필요한 데이터는 영화관입장권통합전산망에서 수집하였다. 영화에 관련된 배우를 제외한 데이터는 엑셀 파일로 제공되었고, 배우 데이터는 파이썬을 이용하여 웹 크롤링(Web Crawling)하였다. 웹 크롤링을 통해 얻은 데이터는 분석에 활용하기 위하여 엑셀 파일로 저장하였다. 데이터에는 영화의 제목, 영화의 감독, 제작사, 수입사, 배급사, 개봉일, 전체 스크린 수, 총 매출액, 전체 관객 수, 영화의 장르, 영화의 등급, 주연 배우, 조연 배우가 포함되어있다. 영화 정보에 주연 배우만 입력되어있는 경우 조연 배우는 제외하고 주연 배우만 포함하였다. 감독과 제작, 배급사가 둘 이상이어도 하나로 취급하였고, 배우는 모두 한 사람으로 취급하여 최대 28명까지 포함되어있다. 또한 분석을 위해 흥행여부 항목을 만들어 총 관객 수가 2,650,000명 이상일 경우 Y, 아닐 경우 N으로 입력하였다.

3.4 예측 모델 분석

앞서 선정한 데이터를 바탕으로 회귀분석을 통하여 모델을 선정한다. 회귀분석은 R을 이용하여 진행하였다. 예측 모델 분석에는 흥행 여부 항목을 사용하지 않고, 총 관객 수 데이터를 이용하였다. 독립변수는 총 관객 수이고, 감독, 배우, 장르, 등급, 제작사, 배급사 변수를 종속변수로 [표 3-2] 와 같이 조합하여 예측 모델을 선정하였다.

[표 3-2] 예측 모델 분석을 위한 모델

모델
감독, 배우, 장르
감독, 배우, 제작사
감독, 배우, 배급사
감독, 장르, 등급
감독, 장르, 제작사
감독, 장르, 배급사
감독, 등급, 제작사
감독, 등급, 배급사
배우, 장르, 제작사
배우, 장르, 배급사
감독, 배우, 장르, 제작사
감독, 배우, 장르, 배급사

[표 3-3]은 다양한 모델의 회귀 분석 결과를 나타낸다. 모든 데이터의 p-value가 0.05보다 작기 때문에 통계적으로 의미 있다고 볼 수 있다. 또한 데이터의 양이 많기 때문에 설명력이 높게 나타났다. 분석 결과 모델에서 배우가 제외된 경우는 회귀계수의 설명력이 비교적 떨어지는 것을 확인할 수 있었다. 이 모델 중에서 회귀계수의 설명력이 가장 높은 배우, 장르, 배급사 모델로 흥행을 예측한다.

[표 3-3] 회귀 분석 결과

				상수	R^2	$Adjusted R^2$	$p-value$
감독	배우	장르	-	-5.85E+02	0.999	0.9863	<2.2e-16
감독	배우	제작사	-	8.35E+01	0.999	0.9857	<2.2e-16
감독	배우	배급사	-	-4.94E+02	0.999	0.9863	<2.2e-16
감독	장르	등급	-	8.12E+05	0.8429	0.6391	<2.2e-16
감독	장르	제작사	-	-5.04E+04	0.9463	0.8113	<2.2e-16
감독	장르	배급사	-	206451.6	0.8616	0.5796	<2.2e-16
감독	등급	제작사	-	4.27E+06	0.9526	0.8389	<2.2e-16
감독	등급	배급사	-	-1.88E+05	0.8545	0.574	<2.2e-16
배우	장르	제작사	-	-2.01E+02	0.999	0.9871	<2.2e-16
배우	장르	배급사	-	7.59E+01	0.999	0.9879	<2.2e-16
감독	배우	장르	제작사	-3.93E+01	0.999	0.9848	<2.2e-16
감독	배우	장르	배급사	-3.37E+02	0.999	0.9855	<2.2e-16

3.5 흥행 예측

예측 모델을 이용하여 실제 흥행 여부를 예측해보기 위한 데이터는 2017년 12월 1일부터 2018년 4월 30일까지의 국내에서 개봉한 한국 상업영화 167편중 배우와 감독 등의 정보를 정확하게 파악할 수 있는 영화 가운데 전국 관객 수가 2명 이상이고 총 스크린 수가 5개 이상인 영화 56편으로 선정한다. 데이터의 정제는 예측 모델 분석을 위해 수집했던

데이터 정제 방법과 동일하게 감독이 외국인인 경우 제외하고, 출연 배우 중 외국인이 포함되어있으면 해당 배우를 제외시켰다. 또한 출연 배우 모두가 외국인일 경우 영화를 제외하는 방식으로 진행하였다. 이 데이터의 배우는 최대 13명까지 입력하였다.

[표 3-4]은 나이브 베이즈 분류를 통해 분석한 흥행 예측 결과를 보여준다.

[표 3-4] 흥행 예측 결과

제목	실제 흥행 여부	예측된 흥행 여부
신과함께-죄와 벌	Y	Y
1987	Y	N
그것만이 내 세상	Y	N
곤지암	Y	N
지금 만나러 갑니다	N	N
조선명탐정: 흡혈괴마의 비밀	N	N
리틀 포레스트	N	N
골든슬럼버	N	N
궁합	N	Y
사라진 밤	N	N
염력	N	N
7년의 밤	N	N
강철비	N	N
흥부: 글로 세상을 바꾼 자	N	N
치즈인더트랩	N	N

1급기밀	N	N
게이트	N	N
미니특공대X	N	N
뽀로로 극장판 공룡섬 대모험	N	N
젝스키스 에이틴	N	N
비밥바룰라	N	N
터닝메카드W: 반다인의 비밀 특별판	N	N
돌아와요 부산항애(愛)	N	N
라라	N	N
광인옥한흠	N	N
예스 평창!	N	N
풍뎅이뎅이	N	Y
섹스인더게임	N	N
처녀 사냥	N	N
엄마 애인	N	N
어린 엄마	N	N
첫경험	N	N
발칙한 동거	N	N
쾌락 도우미 무삭제	N	N
아들의 아내 무삭제	N	N
거짓말 애인-감독판	N	N
공즉시색 2	N	N
가래떡의 맛: 감독판	N	N
엄마친구 4	N	N
친구누나2	N	N
미용실 : 특별한 서비스 2	N	N
밀애 : 친구엄마	N	N

비뇨기과 여의사들	N	N
사촌 누나	N	N
정사 : 착한 며느리들	N	N
내 아내의 언니 2	N	N
두처제2-무삭제	N	N
정사2: 친구 새엄마-무삭제판	N	N
정사 : 바람난 유부녀들 무삭제판	N	N
연변 아가씨의 맛섹사 - 감독판	N	N
바람 바람 바람	N	N
덕구	N	N
나를 기억해	N	N
머니백	N	N
파파 오랑후탄	N	N
정사 : 친구의 엄마 2	N	N

이 분석의 성능을 평가하기 위하여 오차행렬(Confusion Matrix)을 이용한다. 오차행렬은 기계학습 분야에서 알고리즘의 성능을 시각화할 수 있는 행렬이다. [표 3-5]는 이 분석의 성능을 평가한 오차행렬의 결과값을 보여준다.

[표 3-5] 오차행렬의 결과값

	Accuracy	Sensitivity	Specificity
결과값	0.9	0.9565	0.2500

오차행렬로 평가한 결과 나이브 베이즈 분류의 정확도(Accuracy)는 91%로 높게 나타났다. 이 평가의 'Positive' Class는 다수의 데이터인 영화가 흥행하지 않을 것이라는 'N'이다. 따라서 올바르게 식별된 흥행하지 않는 영화의 비율(Sensitivity)은 96%이고 올바르게 식별된 흥행하는 영화의 비율(Specificity)은 25%이다. 이는 흥행한 영화의 수가 상대적으로 너무 적었기 때문에 비율이 낮게 나타났다고 볼 수 있다.

분석 결과 총 56편 중 51편의 영화가 실제 흥행 여부와 동일하게 나타났다. 다소 적합하지 않은 모습을 보인 영화의 대부분은 과거에 흥행 기준을 넘는 영화에 출연한 유명 배우가 주연으로 있지만 한두 명의 유명 배우를 제외하고는 과거의 출연작이 흥행의 기준을 넘지 못했던 영화인 배우가 속해있는 경우이다. 또한 영화 <곤지암>의 경우는 [표 3-6]에서 볼 수 있듯이 <곤지암>이 첫 작품인 배우들이 출연하였기 때문에 유명 배급사의 작품이지만 참고할만한 과거의 데이터가 충분하지 않았기 때문에 흥행을 정확하게 예측할 수 없었다.

[표 3-6] 영화 <곤지암>에 출연한 배우의 출연작

구분	배우	영화 제목	
주연	위하준	곤지암	-
주연	박지현	곤지암	반드시 잡는다
주연	오아연	곤지암	-
주연	문예원	곤지암	-
주연	박성훈	곤지암	-
주연	이승욱	곤지암	-
주연	유제윤	곤지암	-

영화 <풍뎅이뎅이>는 국산 애니메이션으로 이 연구에서는 성우 중심이 아닌 배우를 중심으로 분석하였기 때문에 정확한 예측 결과를 볼 수 없었다.

제 4 장 결론 및 향후 연구 과제

예측 모델 분석 결과 다양한 모델 중 배우, 배급사, 장르를 포함하는 모델이 회귀계수의 설명력이 높게 나타났다. 데이터의 양이 많아질수록 설명력은 높게 나올 수밖에 없었지만 그 중에 가장 높게 나온 모델을 선택하였다. 설명력이 높은 변수는 대기업의 배급사들이 포함되어 있었다. 또한 주, 조연 배우들은 영화나 드라마 등의 매체를 통해서 잘 알려진 배우들이 포함되어 있었다. 장르는 드라마, 멜로/로맨스, 코미디가 흥행에 영향을 미치는 것을 알 수 있었다.

나이브 베이즈 분류를 이용하여 2017년 12월부터 2018년 4월까지의 영화 흥행을 예측한 결과 56편 중45편의 영화가 실제와 일치하는 결과를 얻을 수 있었다. 예측 결과 흥행하지 않을 것이라 예측된 영화의 수가 흥행할 것이라 예측된 영화의 수 보다 높게 나타났다. 이는 2011년부터 2018년 4월까지의 영화 총 관객 수 기록에서 확인할 수 있듯 전체 1,309편의 영화 중 흥행 기준인 2,650,000명 이상의 관객 수를 얻은 영화는 86편으로 흥행 기준을 넘는 영화는 약 7%로 소수인 것처럼 흥행하지 않을 것이라고 예측되는 영화의 비율이 높을 수밖에 없다. 분석을 위해 정한 기간 동안 개봉한 영화 중 흔히 성인 영화로 분류하는 영화는 총 24편으로 전체의 약 43%를 차지하고 있었다. 대다수의 성인 영화는 극장 수익보다 IPTV로 얻는 수익이 더 크기 때문에 성인 영화에 대한 흥행 예측 연구는 극장 관객 수를 대상으로 하는 것이 아니라 IPTV 및 디지털 케이블 TV의 이용 건수를 대상으로 분석하는 것이 더 적합하다고 판단할 수 있다. 따라서 예측 모델 분석에서부터 이러한 성인영화가 모두 포함되었기 때문에 분석 결과에 영향을 미쳤을 가능성도 생각해볼 수 있다.

또한 애니메이션의 경우 더빙하는 성우가 기존에 상업 영화에서 활동하던 배우일 수도 있고, 애니메이션 더빙을 전문으로 하는 배우일 수도 있기 때문에 분석에 영향을 미칠 수 있을 것이다. 이전 연구에서는 배우와 감독을 배우 파워, 감독 파워 등과 같은 이름으로 점수화하여 분석을 하였는데 <곤지암>과 같은 영화는 이 수치가 전혀 영향을 줄 수 없는 영화였다. 따라서 향후 연구에서는 신인배우나 신인 감독에 대한 분석도 필요할 것으로 보인다.

향후 연구에서는 TV 드라마나 예능 등에서 인지도를 쌓고 영화에 출연하는 배우에 대한 영향력 분석도 필요하다. 또한 SNS상에서의 배우, 감독, 배급사 등의 평판 및 영향력 분석도 필요하다.

참고문헌

[1] 영화진흥위원회 (2018). 2017년 한국영화산업 결산. 영화진흥위원회.

[2] 김우철 (2006). 일반통계학. 영지문화사.

[3] 강근석, 유현조 (2016). R을 활용한 선형회귀분석. 교우사.

[4] 김종웅 (2014). 음성 향상을 위한 다중 선형회귀 분석 기반의 음성 존재 불확실성 추정 기법. 한양대학교 대학원 석사학위논문.

[5] 장미소 (2016). 나이브 베이즈 분류기를 이용한 분류기 통합 모델. 명지대학교 대학원 석사학위논문.

[6] David D. Lewis (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. European Conference on Machine Learning, 4-15.

[7] 김제욱, 김한준, 이상구 (2001). 한국데이타베이스학회 국제 학술대회, 2001.6, 331-341.

[8] 박승현, 정완규 (2009). 한국 영화시장의 흥행결정 요인에 관한 연구. 언론과학연구, 9(4), 243-276.

[9] 김연형, 홍정한 (2011). 영화 흥행 결정 요인과 흥행 성과 예측 연구. 한국통계학회 논문집, 18(6), 859-869.

[10] 유현석 (2001). 영화 흥행 변수에 관한 연구. 문화정책논총, 13, 231-254.

[11] Byeng-Hee Chang, Eyun-Jung Ki (2005). Devising a Practical Model for Predicting Theatrical Movie Success: Focusing on the Experience Good Property. JOURNAL OF MEDIA ECONOMICS, 18(4), 247-269.

[12] 권신혜, 박경우, 장병희 (2017). 기계학습 기반의 영화흥행예측 방법

비교: 인공신경망과 의사결정나무를 중심으로. 예술인문사회융합멀티미디어논문지, 7(4), 593-601.

[13] 백광일, 김구곤, 최승배, 강창완 (2015). Stylometry를 이용한 영화 흥행 예측. 한국자료분석학회, 17(2), 719-728.

[14] 씨네21, http://www.cine21.com (2018.05.30).

[15] 박승현, 송현주 (2012). 영화의 흥행성과와 제작비 규모와의 관계. 사회과학연구, 51(1), 45-79.

[16] 김상호, 한진만 (2014). 한국 영화의 흥행성과 결정요인 분석. 사회과학연구, 53(1), 191-214.