영화 메타데이터 기반 관객수와 손익분기달성 예측

Predicting The Number of Audience and Breakeven Point Attainment based on Movie Metadata

저자 (Authors)

옥정우, 김희진, 조대하, 주현수, 이효철, 테레시아, 이석원

Jung-Woo Ok, Hee jin Kim, Dae-ha Cho, Hyunsu Ju, Hyo-Cheol Lee, Theresia Saputri, Seok-Won Lee

출처 (Source)

한국정보과학회 학술발표논문집 , 2017.12, 1932-1934 (3 pages)

발행처 (Publisher)

한국정보과학회

KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE07322734

APA Style

옥정우, 김희진, 조대하, 주현수, 이효철, 테레시아, 이석원 (2017). 영화 메타데이터 기반 관객수와 손익분기달성 예측. 한국정보과학회 학술발표논문집, 1932-1934.

영화 메타데이터 기반 관객수와 손익분기달성 예측

옥정우 O[1], 김희진 [1], 조대하 [1], 주현수 [1], 이효철 [2], 테레시아 [2], 이석원 [1]
아주대학교 소프트웨어학과 [1], 아주대학교 컴퓨터공학과 [2]
{okjwoo, heeeee123, dh9306, gkstnrkf, mytion7, trdsaputri, leesw} @ajou.ac.kr

Predicting The Number of Audience and Breakeven Point Attainment based on Movie Metadata

Jung-Woo Ok O[1] , Hee jin Kim[1], Dae-ha Cho[1], Hyunsu Ju[1],
Hyo-Cheol Lee[2], Theresia Saputri[2], Seok-Won Lee[1]
Dept. of Software, Ajou University[1], Dept. of Computer Engineering, Ajou University[2]

요 약

한국 영화산업에서 전략적 투자자는 영화 투자의 20~30%를 차지하고 나머지는 모태펀드로 소기업들의 자본을 이용한다. 하지만 손익분기점을 넘기지 못하는 영화가 70% 이상 있다는 점에서 영화 투자비용의 비중이 큰 전략적 투자자에게 있어서 투자 위험요소이다. 이런 위험요소를 사전에 막기위해 영화의 메타데이터를 활용한 영화의 관객수와 손익분기달성 예측을 통해 투자자가 위험요소를 미리 발견할 수 있다는 가설을 세웠다. 본문에서는 관객수 예측을 위한 regression과 손익분기점달성 예측을 위한 classification을 통해 결과를 도출하였다. 실험결과 관객수 예측에서는 낮은 coefficient를 보여주었지만, 손익분기달성 예측에서는 낮은 False Positive로 손익분기점을 넘길 영화에 대해서는 신뢰성있는 예측을 보여주었다.

1. 서 론

영화 산업의 제작 단계에 있어서 중요한 요소 중 하나는 투자이다. 여기서 투자자들은 크게 전략적 투자자, 재무적 투자자, 모태펀드가 있다. 한국 영화 산업의 투자 구조의 경우, 전략적 투자자들이 영화 투자 전체의 20~30%를 차지하며 [1] 나머지 부분은 모태펀드의 투자유치를 통해 채운다. 이 때, 전략적 투자자는 영화의 흥행 여부로 수익을 얻게 된다. 그렇기 때문에, 전략적 투자자는 영화 흥행 여부를 예측하여 적절한 금액을 투자해야 한다. 최근 한국에서 개봉된 영화의 손익분기 실적은 좋지 않다. 일반적으로 대규모 자본을 기반으로 한 영화는 흥행에 성공할 것이라고 생각한다. <군함도>, <남한산성> 또한 이러한 이유로 흥행에 성공할 것이라고 예측했었다. 하지만 막상 이 영화들은 손익분기점도 넘지 못했다. 영화 진흥위원회의 보고자료에 따르면 2015년 한국 상업 영화 10편 중 7.3편, 2016년에는 82편 중 63편이 손익분기점을 넘기지 못했다고 발표하였다. [2] [3] 이를 통해, 투자자들의 영화 흥행 예측이 70% 이상 실패하고 있음을 보이고 있다. 영화의 흥행 예측 실패는 곧 메인 투자를 담당하는 전략적 투자자들에게 심각한 문제이다. 이러한 문제를 막기 위해 투자 전에 영화의 흥행 여부를 미리 예측하여 이러한 리스크를 줄이고자 한다. 이 때 영화 수익의 80%는 영화 관객수를 통해 결정되므로 영화 관객수를 미리 예측한다면 영화 흥행 여부의 지표로서 사용할 수 있을 것이다.

본 연구에서는 영화의 메타데이터를 통한 영화의 관객수와 손익 분기 달성 여부를 예측하는 모델을 제안한다. 먼저, 영화 제작 전 단계에서 영화 메타데이터를 분석하여 영화의 관객수와 손익 분기 달성 여부에 영향을 주는 요소들을 찾고, 이를 이용하여 regression과 classification의 결과로 각각 관객수와 손익 분기점 달성 여부를 예측하였다.

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음(20150009080031001)

2. Data

2.1. Data collection

학습모델을 위해 2004년 1월 1일부터 2017년 8월 31일까지 개봉한 영화 총 16020개의 영화 메타데이터를 영화진흥위원회에서 수집하였다.

2.2. 변수 목록

Data collection 때 얻은 변수는 순위, 영화명, 개봉일, 매출액, 매출액 점유율, 누적매출액, 누적관객수, 스크린수, 상영횟수, 대표국적, 국적, 제작사, 배급사, 등급, 장르, 감독, 배우가 있다. 이외에 관련 연구 조사로 추가한 변수는 감독의 관객동원력, 제작사의 관객동원력, 배우의 관객동원력 [4], 개봉 월, 원작 유무, 원작 종류, 원작의 시각화 [5], 제목 3글자 여부 [6], 순제작비, 영화 메인 장르, 영화 인기 장르 여부 [7], 구간별 관객수, 손익분기 달성 여부, 성수기 여부 [8], 영화 평점, 평점 참여자 수가 있다.

2.3. Data preprocessing

그 중 국내 투자자가 국내 영화에 투자한다는 이유 때문에 대표국적이 한국인 영화만을 선정하였고 제작단계에서만 사용할 수 없는 스크린 수, 상영횟수, 좌석 점유율, 매출액 등 변수들은 제거하였다. 또한 2006년부터 영화 산업의 크기가 성장하여 그 이전의 데이터는 우리의 분석과 알맞지 않다고 판단하여 제거하였다. [9] 결과적으로 분석을 위해 총 689개의 instance와 12개의 독립변수를 사용하였다. 이렇게

존재하는 데이터를 분석한 결과 관객수의 경우 100만 이하의 관객수를 동원한 영화가 많았고 손익분기점의 경우 손익분기점을 넘기지 않은 영화가 더 많은 skewed 데이터였다. 정규화를 해도 결과의 차이가 없어 그대로 사용하였다.

3. Solution/Method

3.1. Attribute selection

총 독립변수의 수가 12개로 종속 변수를 설명하기에 충분하지 않다고 판단되어 Attribute selection을 수행하여 만든 모델과 Attribute selection 없이 만든 모델의 성능을 비교 하였다.

3.2. 관객수 예측

관객수 예측을 위해 관객수 변수를 연속형 변수와 범주형 변수로 변환하여 사용하였다. 연속형 변수인 관객수 변수를 범주형 변수로 변환할 때, 영화 분야에서 heuristic하게 사용되는 기준을 사용하였다. 이는 100만 이하는 1, 100만이상 300만 이하는 2, 300만 이상 500만 이하는 3, 500만 이상 1000만 이하는 4, 1000만 이상은 5로 지정하여 범주화 시켰다.[2] Class1은 class3보다 class2와 더 유사한 경향을 가진다.

3.2.1. 다중 선형 회귀분석

관객수 변수가 연속형 변수인 경우에는 회귀분석을 사용하여 결과 값을 예측한다. 위 회귀 분석을 통해 어떤 변수가 관객수를 예측하는데 영향력을 끼치는 지 알 수 있고, 영화의 예상되는 관객수를 예측할 수 있다.

3.2.2. Classification

관객수 변수가 범주형 변수인 경우에는 classification 기법을 사용하여 어떤 class로 분류되는지 알아본다. 이때 사용하는 classification 알고리즘에는 Logistic regression, Decision tree, Naïve Bayes가 있다. 그리고 각 모델의 성능을 평가하여 가장 좋은 성능의 모델을 classification모델로 채택한다. 그리고 이를 모델의 성능을 비교하는데 고려한다.

3.3. 손익분기점 달성 여부

손익분기점 달성 여부를 나타낸 범주를 1과 0을 가진 범주형 변수를 종속형 변수로 사용한다. 위 달성 여부를 알기 위해 classification 기법을 사용한다. 이때 사용하는 classification 알고리즘에는 Decision tree, Naïve Bayes, SVM을 사용한다. 각 모델의 성능을 평가하여 가장 좋은 성능의 모델을 채택한다.

표 1 관객수 Classification Evaluation

		With Attribute selection					Without Attribute selection				
		Class1	Class2	Class3	Class4	Class5	Class1	Class2	Class3	Class4	Class5
Naive Bayes	F-measure	0.771	0.344	0.00	0.250	0.10	0.771	0.317	0.036	0.209	0.083
	AUC	0.756	0.659	0.636	0.756	0.942	0.743	0.647	0.637	0.762	0.946
J48	F-measure	0.758	0.323	0.148	0.059	0.00	0.783	0.297	0.047	0.090	0.316
	AUC	0.636	0.527	0.652	0.545	0.566	0.635	0.514	0.485	0.547	0.558
Logistic	F-measure	0.774	0.323	0.00	0.133	0.250	0.784	0.352	0.115	0.235	0.320
	AUC	0.759	0.667	0.714	0.762	0.763	0.742	0.642	0.639	0.754	0.794

표 2 관객수 Regression과 Classification Evaluation

	With attribute selection		Without attribute selection	
	accuracy	F-measure	accuracy	F-measure
Regression	0.595	-	0.584	-
Logistic	0.61	0.56	0.62	0.587

표 3 손익분기 달성 Classification Evaluation

	TP Rate	FP Rate	Accuracy	Precision	Recall	AUC
Ranker Selected & Naive Bayes	0.308	0.143	0.637	0.590	0.308	0.649
Ranker Selected & SMO	0.236	0.131	0.615	0.546	0.236	0.552
Ranker Selected & Decision tree & post pruning (confidence factor = 0.5)	0.402	0.259	0.605	0.509	0.402	0.593
Ranker Selected & Decision tree (confidence factor = 0.5)	0.471	0.291	0.613	0.520	0.471	0.601
Without Attribute Selection & SMO	0.279	0.143	0.625	0.566	0.279	0.568
Without Attribute Selection & Naïve Bayes	0.337	0.150	0.644	0.600	0.337	0.632
Without Attribute Selection & J48	0.475	0.320	0.598	0.498	0.475	0.587

[2] 서울신문, 멈춰선 대박 행진… 사라진 중박 영화… 불안한 쏠박 행렬

4. Evaluation & Result

4.1. 관객수 예측 모델 비교

4.1.1 Classification

Classification을 평가하는 척도로 F-measure를 사용하였다. [표 1]에 따르면 F-measure의 값이 0임에도 AUC는 높은 것을 알 수 있다. 그러므로 AUC가 아닌 F-measure를 채택하였다.

Attribute selection에 의해 모델을 만들 경우 class 3 과 class 5를 하나도 예측 하지 못하는 경우가 생긴다. 또한 Attribute selection을 진행하지 않은 모델보다 class 3,4,5의 F-measure가 낮다. 이는 관객수가 많은 영화들에 대해 정확도가 떨어짐을 나타낸다. 이번 연구에 사용된 데이터는 100만 이하의 관객수를 동원한 영화가 많았기 때문에 관객수가 많은 영화에 대한 학습이 부족하다. 그러므로 관객수가 많은 영화들을 예측할 때, 상대적으로 error가 적은 모델을 선택해야한다. 이는 class 3, 4, 5에 대해 F-measure가 높은 모델이다. 결과적으로 변수선택을 하지 않은 Logistic model을 classification model로 사용하였다.

4.1.2. Regression과 Classification 비교

[표 2]의 결과로 보아 최종적으로 관객수를 예측하기 위해 accuracy가 조금 더 높은 attribute selection을 하지 않은 logistic regression model을 사용하였다.

4.2. 손익분기점 달성 예측 모델 비교

4.2.1. 비교 기준

비즈니스 환경에 적용시킬 때 TP는 예측한 영화가 정말 손익분기점을 넘길 경우를 말하는 것이다. FP는 손익분기점을 넘기지못하지만 넘긴다고 예측한 경우이다. 이 경우에는 영화 투자비를 소모하고 이익을 챙길 수 없는 경우이므로 최악의 경우이다. 이번 classification 모델에서는 높은 TP rate와 낮은 FP rate를 갖는 모델을 최적의 모델이라 판단한다. 즉 높은 AUC를 보일수록 좋은 모델임을 나타낸다.

4.2.2 모델 비교

[표 3]은 Classification model들의 결과 성능을 분석한 표이다. 각 성능 지표의 최적의 값을 비교하여 최적의 모델을 선택한다. TP rate는 대부분의 decision tree에서 높은 결과를 보여주지만 높은 FP rate와 그로 인한 상대적으로 낮은 AUC를 보인다. SMO는 상대적으로 가장 낮은 FP rate를 보이지만 TP rate도 낮아 낮은 AUC를 보인다. Naive Bayes는 상대적으로 높은 TP rate, 낮은 FP rate, 높은 AUC를 보인다. Naive Bayes 모델 중에서도 attribute selection을 진행하지 않은 모델을 선택했다. 그 이유는 TP와 FP에서 TP의 비율을 나타내는 Precision이 더 높기 때문이다.

5. Conclusion

우리 연구의 목적은 제작사가 영화에 투자를 할 때 수반되는 투자의 위험성을 줄여주는 것이다. 이를 위해서 영화 메타데이터를 이용하여 관객수와 손익분기점 달성 여부를 예측하였다. 결과적으로 관객수를 예측 할 때는 Logistic regression을 사용한 classification기법을 채택하였고, 손익분기점 달성 여부에서는 Naïve Bayes 기법을 채택하였다. 두 예측 모두 종속 변수가 고르게 분포된 변수가 아닌 한 쪽으로 치우친 변수였다. 이를 기반으로 만든 모델은 상대적으로 적은 쪽에 위치한 데이터를 예측하는 데 낮은 성능을 보였다. 그러므로 우리의 모델은 성공할 영화를 예측하는 것 보다는 실패할 영화를 피하는 것에 더 유용하게 이용될 수 있다. 이를 통해 영화 투자자에게 있어 영화 흥망에 따른 리스크를 사전에 방지할 수 있다.

결과적으로 영화의 흥행을 예측함에 있어 이번 프로젝트는 시나리오를 완전히 배제하였다. 앞으로의 발전된 모델을 위해 시나리오 변수를 추가해야 한다. 또한 이번 모델에서는 혼합된 장르 [10]의 경우의 수가 적어 장르를 한 개만 고려하였다. 그러나 영화에는 다양한 장르의 조합이 있다. 조금 더 많은 데이터를 수집하여 혼합 장르를 고려해야한다.

참고 문헌

[1] 한국수출입은행 해외경제연구소, “한국 영화산업 투자구조와 수익성 현황 및 개선방향,” 한국수출입은행 해외경제연구소, 서울특별시, 2014.

[2] 영화진흥위원회 산업정책연구팀, “2015년 한국영화산업 결산 보고서,” 영화진흥위원회 산업정책연구팀, 부산, 2016.

[3] 영화진흥위원회 산업정책연구팀, “2016년 한국영화산업 결산 보고서,” 영화진흥위원회 산업정책연구팀, 부산, 2017.

[4] 김연형 그리고 홍정한, “영화 흥행 결정 요인과 흥행 성과 예측 연구,” %1 *한국통계학회*, 2011.

[5] 이윤정 그리고 신형덕, “원작의 유무와 형태가 영화 흥행에 미치는 영향,” %1 *한국콘텐츠학회*, 2013.

[6] 이정미, 신형덕 그리고 조성호, “제목의 패턴이 영화 흥행에 미치는 영향,” %1 *한국상품학회*, 2014.

[7] 영화진흥위원회, “2016 극장 영화 소비자 조사,” 영화진흥위원회, 2017.

[8] 김소영, 임승희 그리고 정예슬, “영화 유형별 영화 흥행 성과 예측 요인의 비교 연구,” %1 *한국콘텐츠학회*, 2010.

[9] 영화진흥위원회 영화정책센터, “2010년 한국 영화산업 결산,” 영화진흥위원회 영화정책센터.

[10] 김혜원, “영화마케팅 현장에서 영화 장르구분이 미치는 영향,” %1 *제12회 전주국제영화제 컨퍼런스*, 2011.