석사학위논문

머신러닝 기법을 활용한 국내 영화의 흥행 요인에 관한 연구

The Determinants of Box Office Performance in Korea with Machine Learning Technique

황 인 직

한양대학교 대학원

2020년 2월

석사학위논문

머신러닝 기법을 활용한 국내 영화의 흥행 요인에 관한 연구

The Determinants of Box Office Performance in Korea with Machine Learning Technique

지도교수 임 규 건

이 논문을 경영학 석사학위논문으로 제출합니다.

2020년 2월

한양대학교 대학원

비즈니스인포매틱스학과

황 인 직

이 논문을 황인직의 석사학위 논문으로 인준함

2020년 2월

심사위원장 김 종 우 (인)

심사위원 신 민 수 (인)

심사위원 임 규 건 (인)

한양대학교 대학원

차 례

표 차례

그림 차례

수식 차례

국문 요지

국내 영화시장은 세계에서 5번째로 큰 거대 시장이며, 영화 시장 규모는 지속적인 성장 추세를 보였다. 최근에는 OTT와 IPTV 등 미디어 시장의 성장으로 수익 채널이 다양화되었다. 또 영화 산업은 일반적으로 하이 리스크 하이 리턴이라는 특성을 가지고 있어 흥행 여부에 따라 실적의 격차가 매우 커질 수밖에 없다. 따라서 영화 흥행 예측에 관한 연구는 중요성은 매우 크고 필수적이다. 최근의 영화 산업의 트렌드는 수직적 통합을 통해 일부 대형작품들이 시장을 독점하고 있다.

본 연구에서는 영화 산업의 다양성 재고와 정확한 흥행 예측 연구를 위해 가치사슬에 따라 영화 흥행 요인을 분류한 뒤 제작단계에서 새로운 예측 변수를 제시하고 흥행 예측 모델을 제시한다. 특히, LSA 잠재의미분석과 TF-IDF 개념을 활용해 시나리오 텍스트 데이터를 분석하여 기존 연구에서는 잘 다루지 않았던 제작단계에서의 새롭고 다양한 변수를 추가하였다. 이 후 영화 산업의 가치사슬 단계별로 영화 흥행 예측 변수들을 나누고 순서형 로지스틱 회귀분석을 통해 변수의 통계적 유의성을 검증하고 머신러닝 모델들을 통해 비교한다.

이를 통해 시나리오를 활용한 영화 예측 연구와 영화 제작 과정에서의 예측 및 투자 결정을 지원하는데 기여할 것으로 기대된다.

keyword : 기계 학습(Machine Learning), 순서형 로지스틱 회귀분석(Ordinal Logistic Regression), LSA(Latent Sentiment Analysis), TF-IDF, 영화 흥행 예측(Predicting box office Performace)

제1장 서론

제1절 연구 배경 및 필요성

지난해 국내 영화 누적 관객은 약 2억 1600만 명이며, 매출액은 약 1조 8100억 원으로 역대 최고치를 경신했다(2018년 한국영화 결산, 영화진흥위원회). 또 국내 영화 시장은 세계에서 5번째로 큰 영화시장이며(MPAA, 2018), 잠시 성장 정체기가 있었지만 꾸준히 성장하고 있다. 실제로 2004년 총 영화 관람객은 69,254,626명인 반면 2016년에는 217,026,182명으로 약 3.13배 증가하였다. 2003년 개봉한 '실미도' 이후 최근 개봉한 '겨울왕국'까지 25개의 천명영화가 탄생했다(2019년 12월 KOBIS 통합전산망 기준).

영화 산업은 일반적으로 하이 리스트-하이 리턴(High Risk-High Return)이라는 특징을 가지고 있기 때문에, 영화의 흥행 여부에 따라 흥행 실적의 격차는 매우 커질 수밖에 없다(De Vany, 2004). 따라서 정확한 예측을 통해 흥행에 영향을 미치는 요인과 영향력을 분석하는 것은 연구 가치가 뛰어나다고 할 수 있으며, 실무에서도 투자 비용 절감과 수익성 증대를 위해 매우 중요하다. 때문에 영화 흥행과 관련된 연구는 과거부터 꾸준히 지속되어왔고, 국내 영화 시장에 대한 연구 역시 꾸준히 지속되어 왔다.

기존의 영화 흥행 예측과 관련된 연구 사례를 살펴보면 영화 흥행에 미치는

요인들과 영향력 등을 연구한 사례가 주를 이뤘다. 몇 가지 색다른 연구들도 존재했다. 그러나 기존의 연구 사례들을 살펴보면 영화에 투입되는 예산이 영화 성공에 있어 중요 변수임을 부정할 수 없다(Basuroy, Chatterjee & Ravid, 2003). 현재 영화 제작의 트렌드도 대규모 제작비를 투입하는 블록버스터급 영화들이 시장을 독식하고 있다. 블록버스터급 영화들은 대규모 예산을 투입하여 수직적 통합을 통해 제작 단계에서부터 배급 단계, 상영단계까지 독점하여 소비자들이 다양한 영화를 선택할 수 있는 선택권을 줄이고 있다(남기연, 2017). 이러한 현상은 결국 영화의 다양성을 감소시킨다. 국내 영화시장에서도 양극화 심화와 다양성 부재 문제는 심각하며 일부 대형 작품만 살아남는 구조는 영화산업의 붕괴를 가져올 수 있다. 따라서 영화의 다양성을 재고시킬 수 있는 다양한 방안이 필요하다(배장수, 2015).

영화 산업의 양극화 트렌드는 결국 장기적으로 영화 산업에 악영향을 가져올 것으로 예상된다. (사영준 & 유승호, 2019)는 영화산업에서 소비 집중도가 산업의 규모 및 성장성에 미치는 영향을 분석한 결과, 소비의 집중이 전체적인 산업의 규모로 연결되지는 하지만 그 성장이 일시적이라면 산업의 불안정성을 증대시킬 수 있으며 국내 영화 산업 역시 소비 집중도를 해소를 통해 더욱 성장할 수 있다고 주장하였다.

따라서 본 연구에서는 영화 산업의 다양성 재고를 위해 가치 사슬 단계별 흥행과 관련된 다양한 변수들을 이용해 영화 흥행 예측을 하고자하며, 특히 기존연구에서는 많이 다루지 않았던 시나리오 텍스트를 활용하여 영화 예측 연구에 활용하고자 한다.

제2절 연구 목적

본 연구는 영화 산업의 가치사슬 단계별로 영화 흥행에 영향을 미치는 요인을 알아보고 영화 흥행 예측을 하고자 한다. 따라서 본 연구의 목적은 세 가지이다.

첫째, 영화 산업의 가치사슬 단계별로 영화 흥행에 영향을 미치는 변수를 분류하고 순서형 로지스틱 회귀분석을 통해 통계적 유의성을 검증해 각 변수들이 영화 흥행에 미치는 영향력을 알아본다.

둘째, 시나리오 텍스트 데이터에 LSA 잠재 의미 분석과 TF-IDF 등을 적용해 시나리오 기반의 영화 예측 변수를 제시한다.

셋째, 머신러닝 중 지도학습 기법들을 사용해 영화 예측 모델을 설계하고 제시한다.

이를 통해 영화 예측 모델 연구의 새로운 방법을 제시하고 제안한 모델을 통해 영화 제작사나 배급사 등 실무에 활용할 수 있는 기초자료를 제공할 수 있을 것으로 기대된다.

제2장 선행 연구

제1절 가치사슬에 관한 연구

Poter(1985)는 가치사슬 분석은 기업 활동에서 부가가치가 생성되는 과정을 의미한다고 주장한다. 또 핵심 사업에 집중하면서 보다 높은 경쟁력을 확보하고, 수익성을 증대시킬 수 있으며, 자사의 강·약점을 파악하고 원가 발생 원천 및 경쟁 기업과 차별화할 수 있는 점 등을 분석할 수 있다. 때문에 가치사슬은 사업 프로세스를 배열하는 데에 있어 가장 많이 사용되는 방법론이다. 가치사슬은 내부 프로세스를 본원적 활동(Primary Activities)와 지원적 활동(Support Activities)로 나눈다.

Eliashberg(2006)는 영화 산업의 가치사슬을 제작(Production) 단계, 배급(distribution) 단계, 상영(exhibition) 단계로 구분하였다. 각 단계에 대한 연구 문제는 다음과 같다.

제작 단계에서는 대본, 캐스팅(배우 및 감독), 상영등급을 이용해 정확한 영화 예측에 대한 문제이다. 영화 흥행예측에 관한 연구는 장르, 등급, 배우 및 감독 등이 영화 흥행 예측과 관련된 연구에서 중요한 변수로 주목되었다. Eliashberg(2006)과 영화 스크립트를 분석해 영화 흥행 예측 모형에 포함시켰고 유의한 결과를 얻었다.

배급 단계에서는 영화 마케팅에 관해 다룬다. 미디어에 마케팅 예산을 어떻

게 배분하는 것이 효과적이며, 영화 흥행에 미디어와 온라인 구전 등이 얼마나 영향을 미치는지에 대한 문제이다. 블로그나 SNS 등 온라인 구전은 다른 미디어보다 강력한 구전 채널로서 소비자가 영화를 선택할 때 중요한 정보 원천으로 작용하고 있다.

상영 단계에서는 주로 최적의 스크린 수 결정을 위한 문제를 다룬다. 스크린 수는 영화 흥행과 관련된 연구에서 흥행에 중요한 변수로 주목되었다.

제2절 영화 흥행 요인에 관한 연구

Litman(1983)은 창조 영역, 배급 유통 영역, 마케팅 영역 세 가지 유형으로 분류하였다. 창조 영역은 등급, 장르, 스타, 제작비를 의미하여, 배급 유통 영역은 배급사의 크기, 상영 시기를 의미한다. 마케팅 영역은 아카데미 상 등 수상 유무와 평점들을 의미한다. 회귀분석을 통해 예측한 그의 연구 결과에 따르면 제작비가 클수록, 메이저 배급사가 배급할수록, 크리스마스 연휴 기간 상영과 공포 및 SF 장르인 영화, 수상 후보나 수상작인 변수들이 영화 흥행에 긍정적 영향을 미친다고 주장했다.

유현석(2002)은 영화 흥행에 영향을 미치는 변수를 크게 제작 단계와 배급단계로 구분하고 제작단계를 중심으로 하는 변수로 영화 흥행에 미치는 영향력과 변수들 간의 중요도와 시기별 영향력 등을 분석하였다. 1988년부터 1999년까지 개봉된 한국영화 732편을 대상으로 분석했다. 배우, 감독, 제작자, 장르, 상영 등급을 예측 변수로 선정하였다. 다중 회귀 분석을 통해 분석한 결과 제

작단계 변수 중에서 스타급의 배우, 감독, 제작사와 사회풍자, 미스터리, 액션 및 코믹 장르와 15세 관람불가 등급이 영화 흥행에 긍정적 영향을 미치는 것으로 나타났다.

김은미(2003)는 Litman이 제시한 세 가지 영역에 경쟁을 추가하여 영화 흥행 성과를 예측하는 경험적 모델을 제시하였다. 그는 창조영역에서 등급, 장르, 총제작비 그리고 감독의 작품 수가 영화 흥행에 유의미한 영향을 미치고 배급 유통 영역에서는 스크린 수가 유의미하며 수상기록은 흥행에 영향을 미치는 않는 것으로 보았다.

김연형 & 홍정한(2011)은 영화 흥행에 관련 변수를 크게 영화의 내적요인과 외적요인으로 구분하고 영화의 외적요인은 다시 구전커뮤니케이션 영영과 배급유통경쟁영역으로 나누었고, 내적요인에서는 감독, 배우, 관람등급, 외적요인에서는 스크린 수, 배급사파워, 소셜미디어 변수가 영화 흥행에 유의미한 변수라고 하였다.

Sharda et al(2006)은 1998년부터 2002녀 개봉된 834편의 영화을 대상으로 MPAA 등급, 경쟁, 스타파워, 장르, 특별 효과, 속편 여부, 스크린 수를 변수로 선정한 뒤 로지스틱 회귀분석, 판별 분석, CART, MLP 모델을 통해 영화 순수익을 예측하였고, 여러 모델들의 결과를 비교한 결과 MLP 모델이 가장 우수하다고 주장했다.

Eliashberg, J et al.(2007)은 스크립트 분석을 통해 영화 제작 시점에서 수익성에 대한 예측을 시도했다. 그들은 영화 산업에 대한 지식과 자연어 처리 그리고 통계적 방법을 결합해 영화의 ROI를 예측하였다.

Zhang et al(2009)은 BP(Back Propaganda) 신경망을 이용하여 영화 수익 예측을 하였다. 그는 2005년부터 2006까지의 중국 영화시장에서 개봉한 241편의 영화를 분석하였고 국가, 스타(감독 및 배우), 광고, 장르, 개봉시기, 경쟁 영화, 스크린 수, 개봉일을 변수로 사용하여 k-fold 교차검증을 거쳐 기존 ANN(Multi Layer Peceptrom) 신경망과 BP 신경망을 비교했을 때 BP 신경망이 더 높은 정확도를 보였다고 한다.

Eliashberg et al(2014)은 1995년부터 2010년 사이 개봉된 300편의 영화 대본으로 영화 시작시점에서 영화 대본의 수익 가능성을 분석했다. 장르, 스토리라인의 내용(놀라운 결말이 있는지 여부 등), 대본의 의미 변수들(신의 수, 대화의 길이 등), Bag Of Word 변수를 활용해 Kernel 기반의 모델로 분석하였다. Kernel 기반의 방식이 기존의 회귀나 트리기반의 모델보다 더 좋은 성능을 보인 것으로 분석하였고, 대본의 내용이나 의미적 특징이 예측 성능을 향상 시킬 수 있다고 분석하였다.

Kim, T. et al.(2015)는 한국 영화시장에서 머신러닝 기반으로 SNS 데이터를 분석하여 영화 흥행예측을 하였다. 그들은 영화 예측의 초기 예측의 중요성 때문에 개봉 전, 개봉 후, 개봉 2주 후로 시점을 나누어 예측했다. SNS 언급 건수와 주간 트렌드 등을 예측 변수로 활용하였고 머신러닝 기반의 모델을 통해 예측한 결과 SNS 데이터와 머신러닝 기반 알고리즘을 결합한 모델이 예측 정확도를 향상시켰다고 주장했다.

Lash & Zhao(2016)는 2000년부터 2010년 박스오피스 데이터와 IMDB데이터를 분석하여 영화 수익성을 초기에 예측하였고 초기 예측을 제공하기 위해 영

화 투자 보증 시스템(Movie Investor Assurance System)을 제안하였다. 영화 제작단계에서 Who(배우, 감독, 네트워크) What(장르, 등급, 줄거리 개요), When(연간 평균이익, 개봉일), Hybrid(What + who, What + when) 변수를 Random Forest, Logit Boost, Logistic Regression, Navie Bayesian 모델을 사용해 교차 검증하였다. 그 결과 Random Forest가 가장 우수한 것으로 나타났고, What과 New 변수가 영화 성공에 가장 영향을 많이 미치는 것으로 나타났다.

제3장 연구 설계

제1절 데이터 수집 및 전처리

1. 데이터 수집 및 전처리

본 연구에서는 2004년부터 2016년까지 13년간의 국내 영화 중 145개의 영화를 선정하였다. 우선 필름메이커스(www.filmmakers.co.kr)에서 국내 영화 시나리오 데이터 200여 개를 수집한 뒤 145개의 영화를 선정하였다. 일반 영화가 아닌 예술 영화와 독립 영화는 일반 영화와 일반적인 속성이 다를 뿐만 아니라 흥행에 관한 요인도 다르다 판단되어 제외하였다. 또 파일 형식이 PDF이거나 텍스트 파일이 완전하지 못한 것은 제외하여 최종적으로 145개의 시나리오 텍스트 데이터를 수집하였다. 이 후 특수문자와 기호, 공백은 모두 제거한 뒤 제작단계 변수 중 시나리오 변수인 '총 신(Scene)의 수', '총 대화 수', '대화 별 평균 길이'를 측정하였다. 그리고 시나리오 중 대화만 따로 추출하여 파이썬 라이브러리 konlpy의 꼬꼬마 형태소 분석기를 활용해 일반명사와 의존명사, 동사와 형용사만 따로 추출하여 LSA 잠재의미분석을 통해 차원 축소된 토픽을 추출한 뒤 해당 토픽에 대해 TF-IDF와 유사한 지표를 사용하였다.

수집한 145개의 영화에 대해서 영화진흥위원회 통합전산망 KOBIS (www.kobis.co.kr)에서 '장르', '등급', '배우파워', '감독파워', '제작사 파워', '스

크린 수', '연휴 포함 유무'와 '누적관객수'를 수집하였다. 또 배급 관련 변수인 '온라인 평점'과 '전문가 평점'은 네이버 영화(www.movie.naver.com)에서 수집하였다.

제2절 예측변수 및 타겟 변수

1. 예측 변수

예측 변수는 영화 산업의 가치사슬에 따라 크게 '제작 변수', '배급 변수', '상영 변수' 3가지로 나뉘었다.

제작 변수에는 장르, 등급, 배우 파워, 감독파워, 제작사 파워와 시나리오 구조에 해당하는 총 신(Scene)의 수, 총 대화 수, 대화별 평균길이, Bag of Word 변수들인 LSA_1, LSA_2, LSA_3이 있다. 장르는 드라마와 가족은 1, 코미디는 2, 멜로와 로맨스 그리고 에로는 3, 액션과 범죄 그리고 전쟁은 4, 공포와 스릴러 그리고 미스터리는 5으로 분류하였다. 등급은 '영화 및 비디오물의 진흥에 관한 법률 제29조(상영등급분류)'와 영상물등급위원회의 분류한 등급에 따라 '전체이용관람가', '12세 이상 관람가', '15세 이상 관람가', '청소년 관람불가 등급' 으로 구분하였고, 각각 1, 2, 3, 4로 분류하였다. 배우 파워는 해당 작품에 출연한 주연 배우들이 최근 5년간 주연으로 출연한 작품 중 최고 흥행 작품이 300만 이상이면 2, 100만 이상이면 1, 100만 이하라면 0으로 분류했다. 감독 파워는 최근 5년간 감독으로 제작한 최고 흥행 작품이 300만 이상

이면 2, 100만 이상이면 1, 100만 이하라면 0으로 분류했다. 제작사 파워는 제작사의 최근 5년간 제작 영화의 누적관람객을 모두 더한 뒤 상용로그를 취했다. 제작 변수 중 시나리오 구조와 관련된 변수 중 총 신(Scene)의 수의 수는 시나리오 데이터에서 총 신의 수를 체크하였고, 총 대화 수는 시나리오의 총 대화 수, 대화 별 평균 길이는 특수문자와 기호 및 공백을 모두 제거한 뒤 대화만 추출하여 대화 별 평균길이를 체크하였다. Bag of Word 변수인 LSA_1, LSA_2, LSA_3은 시나리오 데이터에서 대화만 추출하여 꼬꼬마 형태소분석기로 의존명사, 일반명사, 동사, 형용사를 추출하고 상위 1000개의 단어에 대해 LSA 잠재 의미 분석을 통해 10개의 토픽을 생성한 후, 그 중 3개의 토픽을 추려 TF-IDF와 유사한 값을 매겼다. 이에 대한 설명은 4장에서 자세히 설명하겠다.

배급관련 변수에는 온라인 평점과 전문가 평점이 있다. 네이버 영화 사이트에서 네티즌 평점과 전문가 평점을 사용하였다. 전문가 평점은 전문가 평점이 기록되지 않은 데이터도 있었다. 이러한 결측치는 네티즌 평점과 전문가 평점은 상관관계가 있을 것이라는 가정으로 해당 영화의 네티즌 평점에 전체 영화의 전문가 평점의 평균을 곱한 뒤 전체 영화의 네티즌 평점으로 나눈 값을 대입하였다.

상영 변수는 스크린 수와 연휴 포함 유무가 있다. 스크린 수는 해당 영화의 스크린 수를 연휴 포함 유무는 개봉일 이후 2주간 방학시즌과 크리스마스 연휴(12월 24, 25일), 그리고 설날과 추석 연휴가 포함되면 0, 포함되지 않으면 1으로 처리하였다. 가족단위 관람객이나 학생 관람객이 많을 것으로 예상되는 방학시즌(1, 2, 7, 8월)은 전체, 12세, 15세 이상 관람가에게만 적용하였다.

2. 타겟 변수

타겟 변수는 기존 선행 연구들의 기준들과 영화계에서 일반적으로 통용되는 기준을 참고하여, 해당 영화의 누적관람객을 범주형 변수로 변환하였다. 누적관람객이 700만 이상이면 '대흥행', 300만~700만이면 '흥행', 300만 이하라면 '비흥행'으로 분류하였다. 아래 [표 1]에 예측변수와 타겟 변수에 대한 정의를 정리하였다.

변수 구분	가치사슬	변수	변수 설명	변수 형태
예측변수	제작단계	장르	드라마/가족, 코미디, 멜로/로맨스/에로, 액션/범죄/전쟁, 공포/스릴러/미스터리	범주형
		등급	전체이용가, 12세 이상 관람가, 15세 이상 관람가, 청소년 관람불가	범주형
		감독파워	해당 영화의 감독이 최근 5년간 감독으로 제작한 최고 흥행 작품	수치형
		배우파워	해당 영화의 주연 배우 중 최근 5년간 주연으로 출연한 작품 중 최고 흥행 작품	수치형
		제작사파워	최근 5년간 제작사 영화의 누적관람객을 모두 더한 뒤 상용로그	수치형
		총 신의 수 (시나리오 구조 변수)	시나리오에서 총 신의 수	수치형

		총 대화 수 (시나리오 구조 변수)	시나리오에서 총 대화의 수	수치형
		대화별 평균길이 (시나리오 구조 변수)	특수문자, 기호, 공백 등을 제거한 시나리오에서 대화별 평균길이	수치형
		LSA_1 (Bag of Word 변수)	LSA로 추출된 TOPIC_1에 대한 변형된 TF-IDF을 적한 값	수치형
		LSA_2 (Bag of Word 변수)	LSA로 추출된 TOPIC_2에 대한 변형된 TF-IDF을 적한 값	수치형
		LSA_3(Bag of Word 변수)	LSA로 추출된 TOPIC_3에 대한 변형된 TF-IDF을 적한 값	수치형
	배급단계	네티즌 평점	네이버 영화의 네티즌 평점	수치형
		전문가 평점	네이버 영화의 전문가 평점	수치형
	상영단계	스크린 수	해당영화의 스크린 수	수치형
		연휴유무	개봉일 이후 2주간 방학 시즌(1,2,7,8월), 크리스마스 연휴, 명절 연휴 포함유무 (방학시즌은 청소년 관람 불가 등급 미적용)	범주형
타겟변수	-	흥행유무	해당영화의 누적관객수가 700만 이상이면 대흥행(2), 300만이상이면 흥행(1), 300만이하면 비흥행(0)	범주형

[표 1] 변수 정의

제3절 기초통계

본 연구에 앞서 탐색적 자료분석을 위해 기술통계를 실시하였다. 각 변수의 표본은 각각 145개이고 결측치는 사전에 모두 제거하였다.

수치형 변수					
변수	N	최솟값	최댓값	평균	표준편차
제작사파워	145	3.84	7.60	6.48	0.62
총 신의 수	145	76.00	195.00	113.86	19.32
총 대화 수	145	104.00	1373.00	702.72	238.39
대화별 평균길이	145	9.75	54.20	19.96	5.46
LSA_1	145	0	352.50	30.93	60.04
LSA_2	145	0	407.02	16.69	44.22
LSA_3	145	0	142.16	7.04	19.71
네티즌 평점	145	2.22	9.50	7.67	1.28
전문가 평점	145	2.33	8.25	5.86	1.14
스크린 수	145	8.00	1788.00	344.43	262.58

[표 2] 기술통계

범주형 변수 분포[1)]						
변수	N	Class1	Class2	Class3	Class4	Class5
장르	145(100%)	38(26%)	29(20%)	27(19%)	26(18%)	25(17%)
등급	145(100%)	65(45%)	38(26%)	33(23%)	9(6%)	-
배우파워	145(100%)	14(10%)	41(28%)	89(62%)	-	-
감독파워	145(100%)	88(61%)	30(21%)	26(18%)	-	-
연휴유무	145(100%)	96(66%)	49(34%)	-	-	-
흥행	145(100%)	118(81%)	19(13%)	8(6%)	-	-

[표 3] 범주형 변수 분포

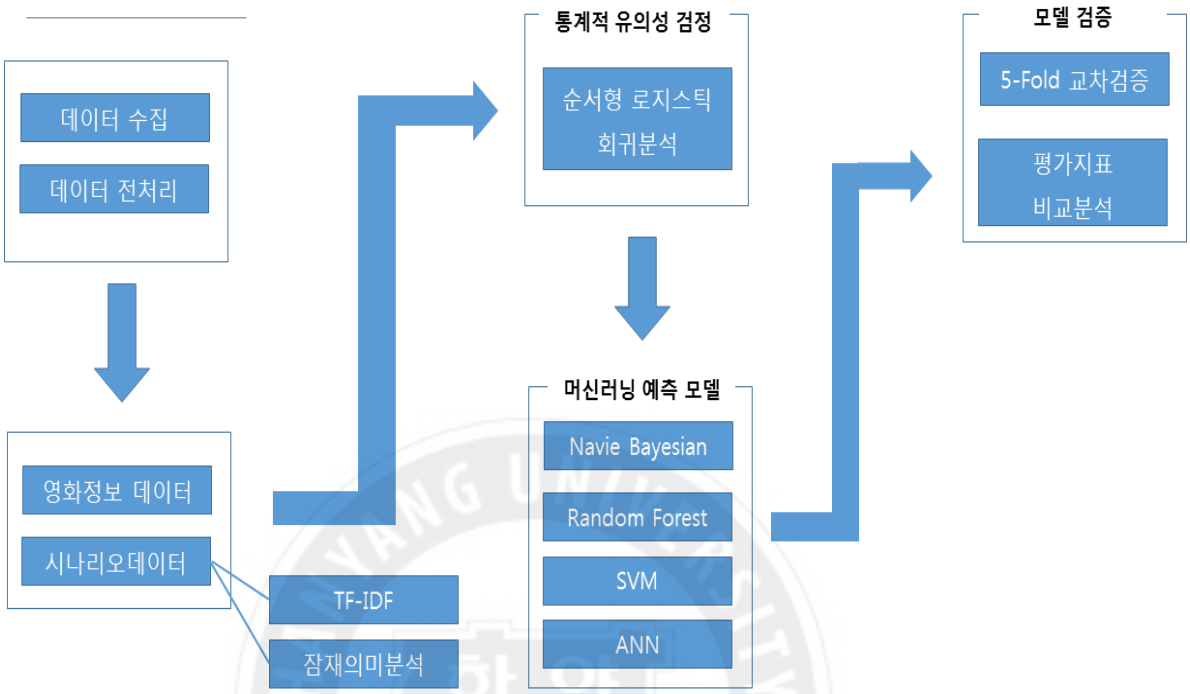
1) 장르 : 드라마 1, 멜로/로맨스/에로 2, 코미디 3, 액션/범죄/전쟁 4, 공포/스릴러/미스터리 5
등급 : 전체이용가 1, 12세 이상 관람가 2, 15세 이상 관람가 3, 청소년 관람 불가 4
배우파워 : 100만 이하 0, 100만 ~ 300만 이하 1, 300만 이상2
감독파워 : 100만 이하 0, 100만 ~ 300만 이하 1, 300만 이상2
연휴유무 : 비연휴 0, 연휴 1
행흥여부 : 비흥행 0, 흥행 1, 대흥행 2

제4장 연구 모형

제1절 연구 모형

본 연구에서는 영화 산업의 가치사슬별로 변수들을 나누고 데이터 전처리 과정을 거쳐, TF-IDF를 활용한 개념과 LSA 잠재의미분석을 사용해 한국영화에 대한 흥행 예측 연구에서는 다루지 않았던 새로운 변수를 제안한다. 이후 순서형 로지스틱 회귀분석을 통해 통계적 유의성을 살펴보고, 머신러닝 기법을 활용해 영화 흥행 예측 모델을 제안한다. 연구에 대한 개요는 [그림 1]로 도식화하였다.

앞서 설명한 연구 설계를 통해 선정한 변수들을 토대로 순서형 로지스틱 회귀분석을 통해 통계적 유의성을 검증하고, 머신러닝 모델 중 지도학습 기법을 이용하였다. 지도학습의 여러 기법 중 Naive Bayes, RandomForest(RF), Support Vector Machine(SVM), Artificial Neural Network(ANN)를 이용해 예측했다. 이후 5-Fold 교차검증 후 F1 Score와 Confusion Matrix로 머신러닝 모델들을 비교분석하였다.


데이터 수집
데이터 전처리
영화정보 데이터
시나리오데이터
TF-IDF
잠재의미분석
통계적 유의성 검정
순서형 로지스틱
회귀분석
머신러닝 예측 모델
Navie Bayesian
Random Forest
SVM
ANN
모델 검증
5-Fold 교차검증
평가지표
비교분석

[그림 1] 연구 모형

제2절 예측 기법

1. Naive Bayes

나이브 베이즈는 베이즈 정리를 기반으로 하는 통계적 기법이다. 가장 단순한 지도 학습 방법 중 하나이며, 훈련과 예측속도가 빠르고 훈련 과정을 이해하기 쉬운 장점이 있다. 또 각 특성에서 클래스별 통계를 단순 취합하기 때문에 효과적이다. 또 희소한 고차원 데이터에서 잘 작동하고 매개변수에 민감하지 않다는 점도 장점이다. 그러나 일반화 성능이 조금 떨어진다는 단점이 있다. 또 나이브 베이즈는 각 특성끼리 서로 독립적이라는 조건이 필요하다. 예를 들어 영화 흥행에 관한 예측에서 A 변수보다 B변수가 더 중요한 변수가 도리 수 있지만 나이브 베이즈는 이러한 사실을 무시하고 단순 특성끼리 서로 독립적이라는 가정 하에 단순 취합한다. 이러한 가정에도 불구하고 비교적 정확한 성능을 보이기 때문에 자주 사용되는 지도 학습 기법 중 하나이다.

베이즈 정리는 사전 확률과 추가 정보를 바탕으로 사후 확률을 추론하는 통계적 방법이다. 베이즈 정리의 기본이 되는 조건부 확률은 사건 B가 발생했을 때 사건 A가 발생할 확률을 의미하며 (식 1)과 같이 나타낼 수 있다.

$$P(A \mid B) = \frac{P(B \cap A)}{P(B)} \qquad \text{(식 1)}$$

나이브 베이즈는 (식 2)와 같이 나타낼 수 있다.

$$P(H|E) = \frac{P(E|H) \bullet P(H)}{P(E)} \quad \text{(식 2)}$$

여기서 사전확률 P(H)과 추가정보 P(E)를 바탕으로 사후확률 P(H|E)을 (식 2)와 같이 계산해 얻을 수 있다.

2. Random Forest

랜덤 포레스트는 의사결정나무(Decision Tree) 기반의 앙상블(Ensemble) 모형으로 Breiman(2001)에 의해 제시되었다. 의사결정 나무는 반응변수(Response Variable)를 가장 잘 설명하는 설명 변수(Explanatory Variable)로 가지를 뻗어나가도록 하는 알고리즘이다(유진은, 2015). 하나의 트리로 구성된 의사결정 모형은 단순하고 시각화가 가능하고 해석이 용이하다는 장점이 있다. 그러나 분산이 높고 정확성이 떨어져 예측 결과의 안정성과 예측 정확도가 낮은 점이 단점이다. 이러한 단점을 보완하는 방법이 앙상블 모형이며, 여러 개의 모형을 만들어 각 모형의 예측을 다수결이나 평균 등으로 결정하는 방식이다. 배깅(Bagging), 부스팅(Boosting), 스태킹(stacking)이 있다. 배깅은 부트스트랩 후 각각의 샘플에 모형을 학습시킨다. 이후 이들의 예측을 합쳐 최종 예측하는 방식이다. 부스팅은 오답에 대한 가중치를 부여하여 오답 문제를 잘 맞힌 모델을 최종 모델로 선정하는 방식이다. 배깅은 병렬적으로 학습한다면 부스팅은 순차적으로 학습하는 방식이다. 학습이 끝나면 나온 결과에 따라 가중치를 재분배한다. 오답에 높은 가중치를 부여하고 정답에 낮은 가중

치를 부여하기 때문에 정확도는 높으나 이상치(Outlier)에 취약한 특징이 있다. 스태킹은 서로 다른 모델들을 조합하여 최고의 성능을 보이는 모델을 생성하는 방식이다.

랜덤 포레스트는 배깅 기법을 이용하며 이름에서 알 수 있는 의사결정 나무들이 숲처럼 많이 모인 모형이다. 의사결정나무 모형을 다수 생성하여 더 정확한 예측을 하는 것이 목적이며, 무작위성을 최대로 부여하여 예측오차를 줄이는 방식이다. 랜덤 포레스트의 수행 과정은 다음과 같다.

1. ntree와 mtry 값을 설정한다. ntree는 랜덤 포레스트를 구성할 전체 의사결정나무의 개수이고, mtry는 랜덤 포레스트를 구성할 의사결정나무에 사용될 입력 변수의 개수를 의미한다.
2. 무작위 방식으로 부트스트랩 방식을 이용하여 표본을 다수 생성하고, 의사결정나무 모형을 적용하여 학습시킨다.
3. 예측결과를 산출하고 예측결과와 실제결과가 얼마나 일치하는지 확인하여 OOB(Out of Bag) 데이터에 대한 오류율을 산출한다.
4. 2~3의 결과를 ntree회 반복하여 랜덤 포레스트를 최종적으로 구성한다(김성진 외, 2016).

랜덤 포레스트의 특징으로는 잡음이나 이상치로부터 받는 영향이 적고, 학습 시간이 빠르며 효율적인 장점이 있다. 또 굳이 훈련자료와 시험자료로 나누어 모형 평가를 시도할 필요 없이 OOB 분석을 할 수 있다. 또 분포에 대한 통계적 가정을 필요로 하지 않는 비모수적 방법이며, 설명 변수가 많고 설명변수 간 상호작용이 복잡한 고차원 자료에서 예측력이 높다(Cutler et al., 2007; Strobl et al., 2009).

3. Support Vector Machine

Vladimir Naumovich Vapnik에 의해 제안된 SVM은 최대 마진 초평면(Maximum Margin Hyperplane)을 기본 아이디어로 한다. 즉, SVM은 마진을 최대화하는 분류 경계면을 찾는 기법이다. 데이터 상의 n차원의 벡터 공간이 있을 때 n차원의 벡터 공간을 두 개로 구분하는 초평면에서 마진을 최대화하는 선을 구하는 방식이다. 여기서 마진은 초평면 집합과 가장 가까운 벡터와의 수직 거리를 의미하고, 구분하는 초평면과 가장 가까운 벡터를 Support Vector라고 한다. SVM은 고차원 공간의 데이터를 선형회귀 함수로 재구성하고 커널 함수(kernel Function)를 사용하여 비선형 문제를 선형 문제로 해결할 수 있다.

4. Artificial Neural Network

인공신경망은 인간의 뇌의 신경망 구조를 모방하여 모델링한 이론이다. 입력값을 받아서 Neural Network을 복잡하고 비선형적이 구조를 가진 자료에서 예측 문제를 해결하기 위해 사용된다. 인공신경망은 입력층(Input Layer), 은닉층(Hidden Layer), 출력층(Output Layer)으로 총 3개의 층으로 구성된다. 한 층은 여러 개의 노드로 이루어질 수 있으며 노드는 일정 크기 이상의 자극을 받으면 반응하는데 일정 수준이 넘어서면 활성화되어 출력값으로 내보낸다.

인공신경망의 은닉층과 출력층은 결합함수(Combination Function)와 활성화 함수(Activation Function)로 구성되어 있는데, 결합함수의 경우 입력층 또는 은닉층의 마디들을 결합하는 형태를 의미하고 선형함수가 사용이 된다. 활성화 함수는 입력 또는 은닉 마디의 결합을 변환하는 함수이며, 대표적으로 시그모이드 함수(Sigmoid Function)이 많이 사용된다. 시그모이드 함수는 모든 실수 값을 0과 1사이의 값으로 변환시키고 입력값에 대한 가중치로 계산된 결과를 시그모이드 합수에 입력하여 0과 1사이의 값으로 변환한 뒤, 반응 변수가 이분형일 경우 보통 0.5를 기준으로 0.5 미만은 0, 0.5이상은 1에 대응되도록 한다. 인공신경망의 활성화 함수에는 시그모이드 함수 외에도 RELU, Tanh, Identity, Max Out, Logistic, Softmax 등 다양한 함수가 존재한다.

인공신경망은 기존의 통계적 기법과는 다르게 특별한 통계적 가정이 필요하지 않고 대량의 데이터가 존재할 경우 내재된 정보를 데이터에 내재된 정보를 잡아내고 복잡한 모델을 만들 수 있다. 또 충분한 연산 시간과 데이터가 있고 매개변수를 잘 조정한다면 다른 알고리즘과 비교해 우수한 예측력을 보이이며, 예측 변수와 반응 변수간의 관계가 복잡할 때 유용하고 잡음이 많은 데이터에도 안정적이라는 장점이 있다. 반면 모형의 구조를 설명하지 못하는 블랙박스라는 점과 복잡한 학습과정을 거치기 때문에 분석과정에서 많은 시간과 자원이 소요되며 최적의 모형을 도출하는 것이 상대적으로 어렵다는 단점이 있다.

제3절 Bag of Word 변수

1. LSA(Latent Semantic Analysis) 잠재 의미 분석

LSA는 문서 내의 숨겨진 의미 관계를 파악할 수 있는 방법으로 토픽모델링이나 정보검색, 문서분류 등 다양한 분야에서 활용된다. 단어의 의미는 함께 사용되는 단어를 통해 알 수 있듯, 단어의 의미는 문서의 맥락과 매우 밀접한 관계가 있다. 이러한 관점에서 LSA는 연구·발전되어왔고, 단어의 의미를 맥락으로부터 계산해낼 수 있다. LSA는 특이값 분해 SVD(Singular Value Decomposition)을 통해 행렬을 적절한 수의 요인으로 차원 축소한 뒤 행렬을 분해하는 절차를 거친다. 여기서 단어 행렬을 기준으로 상관분석을 실시하면 단어들 사이의 거리 측정할 수 있고, 코사인 유사도(Cosine Similarity) 등을 이용해 단어들 사이의 거리를 계산해낸다.

본 연구에서 LSA 잠재 의미 분석을 통해 토픽을 추출한 결과는 [표]와 같다. 이 중 다른 토픽과 중복이 적고 비교적 문서의 정보를 잘 담아내었다고 판단되는 Topic 2, 4, 6을 선정하였다.

Topic 1	형사	검사	반장	회장	언니
Topic 2	**검사**	**형사**	**범인**	**반장**	**사건**
Topic 3	지요	느냐	그림	옵니다	어찌
Topic 4	**감독**	**선수**	**회장**	**경기**	**야구**
Topic 5	누나	검사	부대	아부지	경석
Topic 6	**부대**	**전쟁**	**공격**	**군인**	**작전**
Topic 7	감독	현우	언니	선수	수연
Topic 8	기봉	신부님	검사	마라톤	이장
Topic 9	기봉	형사	언니	실장	할매
Topic 10	누나	승희	기봉	형사	경석

[표 4] LSA 토픽 모델링 결과

2. TF-IDF

TF-IDF(Term Frequency Inverse Document Frequency)는 대표적인 단어 가중치 기법 중 하나이며, 어떤 단어가 특정 문서 내에서 얼마나 중요한 지를 측정할 수 있는 지표이다. TF-IDF는 전체 문서 중에서 해당 단어를 포함한 문서가 적을수록 TF-IDF값이 높아지는 특징이 있다. 이러한 특징으로 TF-IDF는 모든 문서 내에 나타나는 흔한 단어는 걸러내고, 특정 단어가 가지는 중요도를 측정할 수 있다. 단어 빈도를 나타내는 TF는 특정 단어가 문서 내에서 얼마나 자주 나타나는 지를 나타내는 값이다. 그러나 TF만으로는 문서 내의 단어의 중요도를 측정할 수 없다. 문서의 길이가 길어지면 해당 단어가 출현할 확률도 높아지기 때문이다. 역문서 빈도를 나타내는 IDF는 해당 단어의 일반적인 중요도를 나타내는 값으로, 전체 문서 수에 해당 단어가 포함된 문서들의 수로 나눈 값에 로그를 취해 얻는다. TF와 IDF는 다음과 같은 식을 가진다.

$$TF_{i,j} = \frac{N_{i,j}}{\sum_k n_{k,j}} \quad \text{(식 3)}$$

$$IDF = \log\left(\frac{|D|}{d_j : t_j \in d_j}\right) \quad \text{(식 4)}$$

$$TF-IDF_{i,j} = TF_{i,j} \times IDF_i \quad \text{(식 5)}$$

여기서 $N_{i,,j}$은 문서 d_j에서 단어 t_i가 나타는 빈도수이고, $\Sigma_k N_{k,j}$는 문서 d_j내의 단어 빈도수이다. |D|는 전체 문서 수를 나타내며, $|\{d_j:t_j \in dj\}|$는 단어 t_i가 출현하는 문서 수이다. TF-IDF는 (식 5)과 같이 TF와 IDF의 곱으로 계산되면 다음과 같은 식을 가진다(송민, 2012).

3. Bag of Word 변수

시나리오 속 단어들은 영화의 정서나 테마를 담고 있다(Eliashberg, 2014). 시나리오 속 단어들은 단순히 장르로 구분하는 방식에서는 담지 못했던 정보를 담고 있을 것이라 생각된다. 때문에 제작 단계에서 보다 정확한 영화 예측을 위해서는 시나리오 내용에 대한 분석이 필요하다고 판단되었다. 그래서 본 연구에서는 기존의 선행 연구들을 참고하여 시나리오 텍스트 데이터를 분석하여 Bag of Word 변수(LSA_1, LSA_2, LSA_3)들을 추출하여 영화 흥행 예측에 활용하고자 한다.

먼저 시나리오 텍스트 데이터에서 특수문자, 기호 등을 제거하는 등 전처리 과정을 거친 후 파이썬 라이브러리의 꼬꼬마 형태소 분석기를 사용하여 형태소 분석을 하였다. 이후 의존명사, 일반명사, 동사, 형용사를 추출하였다. 의존명사, 일반명사, 동사, 형용사만을 추출한 이유는 시나리오 텍스트 데이터 내에 캐릭터 이름과 같은 고유명사가 너무 많아 토픽모델링을 통해 토픽을 추출하면 추출된 토픽이 대부분 캐릭터 이름으로 이루어지는 문제점이 발생했기

때문이다. 때문에 명사 중에서 의존명사, 일반명사, 고유명사 등을 구별하여 형태소 분석기인 꼬꼬마 형태소 분석기를 사용하였다.

형태소 분석 후 빈도수 기반 상위 1000개의 단어에 대해 LSA 잠재 의미 분석을 통해 10개의 토픽을 생성한 후, 그 중 3개의 토픽을 추려 추출된 토픽에 대해 TF-IDF와 비슷한 값을 매겼다. 최초 추출된 토픽 [표 4]에서 토픽2, 4, 6,을 추출하여 토픽1, 2, 3으로 정했다.

이후 선정된 토픽들에 TF-IDF와 비슷한 지표를 개발하여 사용했다. TF-IDF의 수식은 (식 5)과 같이 $TF\text{-}IDF_{i,j} = TF_{i,j} \cdot IDF_i$으로 나타낼 수 있다. 이 개념을 활용해 각각의 토픽 내의 단어들을 하나의 벡터로 보고, $VF_{i,j}$는 해당 벡터 내 단어들이 해당 문서에서 출현한 횟수이며, IDF는 전체 문서 수를 해당 단어 벡터가 등장하는 단어 수에 1을 더한 값으로 나눠준 뒤 로그를 취한 값으로 계산하여 TF-IDF에서 변형된 VF와 IDF를 곱해준 값으로 계산하였고, (식 6)로 나타내었다. 여기서 $|\{d_j : v_j \in dj\}|$는 단어 벡터 v_j가 등장하는 문서 수이다.

$$VF_{i,j} \times \log\left(\frac{|D|}{|d_j : t_i \in d_j|}\right) \qquad \text{(식 6)}$$

제4절 데이터 불균형(Imbalanced Data)

데이터 불균형(Imbalanced Data) 문제는 데이터 셋에서 한 범주에 속하는 패턴의 수가 다른 범주에 속하는 패턴의 수보다 매우 적거나 많은 경우를 말한다. 대부분의 기계학습 알고리즘은 범주들의 비율이 동일하다는 가정 하에 적용되는데 데이터 불균형 문제가 있을 경우 여러 가지 문제가 나타날 수 있다. 우리의 바람과는 다르게 실제 문제에서는 데이터가 불균형한 문제가 자주 존재하는데, 사기 탐지와 연체 예측 그리고 본 연구와 같은 영화 흥행 예측도 그러한 경우다. 앞서 살펴본 탐색적 자료 분석 결과 종속 변수인 흥행의 비율이 비흥행, 흥행, 대흥행이 각각 118(81%), 19(13%), 8(6%)로 나타났다.

데이터 불균형 해결을 위한 방법은 여러 가지가 있지만 본 연구에서는 이러한 문제를 극복하기 위해 오버샘플링(Over Sampling), K-fold 교차검증(k = 5)을 실시하였고 평가지표로는 Confusion Matrix와 F-1 Score을 사용하였다.

제 5장 연구결과

제1절 순서형 로지스틱 회귀분석을 통한 통계적 유의성 검증 결과

본 연구에서 종속변수인 흥행에 변수들이 미치는 영향에 대한 통계적 유의성을 살펴보기 위해, 순서형 로지스틱 회귀분석(Ordinal Logistic Regression)을 사용했다. 순서형 로지스틱 회귀분석은 종속변수가 이산형 변수일 경우 사용되는 로지스틱 회귀 분석의 확장 개념이며, 종속변수가 서열척도일 경우 사용된다. 종속변수가 서열척도이기 때문에 순서형 로지스틱 회귀분석은 종속변수가 1단위씩 변화 할 때마다 각각의 독립변수가 종속변수에 동일하게 영향을 준다는 기본 가정을 가지고 있다. 이를 검정하기 위해 평행성 검정을 실시했으며, 평행성 검정 결과 유의수준이 .000으로 나타나 순서형 독립변수의 기본 가정을 만족하는 것으로 나타났다.

총 15개의 예측변수 중에서 통계적으로 유의한 변수(유의수준 0.1)는 제작단계에서는 제작사 파워, 배우 파워, LSA_3 배급단계에서는 네티즌 평점, 상영단계에서는 스크린 수, 연휴 유무로 확인되었다. 나머지 9개 변수는 유의하지 않았다. 실험 결과 제작사 파워가 클수록, 배우 파워가 클수록 LSA_3 값이 높을수록, 네티즌 평점이 높을수록, 스크린 수가 많을수록, 연휴일수록 흥행 확률이 증가하는 것으로 나타났다.

		B 추정값	표준오차	Wald	자유도	유의확률
임계값	[흥행여부 = 1]	26.080	8.774	8.836	1	.003
	[흥행여부 = 2]	29.224	8.964	10.628	1	.001
위치	[장르=1]	.918	1.992	.212	1	.645
	[장르=2]	-1.614	2.388	.457	1	.499
	[장르=3]	1.091	2.087	.273	1	.601
	[장르=4]	1.260	2.009	.394	1	.530
	[장르=5]	0	.	.	0	.
	[등급=1]	-.041	1.722	.001	1	.981
	[등급=2]	1.806	1.136	2.530	1	.112
	[등급=3]	1.192	1.017	1.374	1	.241
	[등급=4]	0	.	.	0	.
	총대화수	-.002	.002	.840	1	.359
	대화별평균길이	-.104	.104	.998	1	.318
	총신의수	.016	.019	.710	1	.399
	[배우파워=0]	2.128	1.350	2.484	1	.115
	[배우파워=1]	1.795	.928	3.744	1	.053

	[배우파워=2]	0	.	.	0	.
	[감독파워=0]	-.749	.917	.667	1	.414
	[감독파워=1]	-.260	.959	.073	1	.787
	[감독파워=2]	0a	.	.	0	.
	제작사파워	9.403E-8	5.542E-8	2.879	1	.090
	LSA_1	.005	.007	.597	1	.440
	LSA_2	.011	.007	2.556	1	.110
	LSA_3	0.17	.014	3.800	1	.051
	네티즌평점	2.308	.874	6.979	1	.008
	전문가평점	.356	.454	.617	1	.432
	스크린수	.006	.002	9.255	1	.002
	[연휴유무=0]	-1.317	.808	3.103	1	.078
	[연휴유무=1]	0	.	.	0	.

[표 5] 모수 추정값

모형	-2 로그 우도	카이제곱	자유도	TPL 유의확률
절편만	17.798			
최종	77.688	94.110	55	.000

[표 6] 모형 적합 정보

Cox 및 Snell	.480
Nagelkerke	.689
McFadden	.548

[표 7] 유사 R^2

모형	-2 LOG 우도	카이제곱	자유도	TPL 유의확률
영가설	77.668			
일반	.000	77.688	22	.000

[표 8]평행성 검정 결과

제2절 흥행 예측 결과

본 연구에서는 영화 산업의 가치 사슬 단계에 따라 변수들을 분류하였다. 로지스틱 회귀분석을 실시하고 통계적 유의성을 검증을 통해 각 변수들이 흥행에 영향을 미치는지 알아보았다. 분석 결과 통계적으로 유의한 변수는 제작사 파워, 배우 파워, LSA_3 배급단계에서는 네티즌 평점, 상영단계에서는 스크린 수, 연휴 유무로 확인되었다. 장르, 등급, 감독파워, 시나리오 구조, 전문가평점 등 나머지 9개 변수는 유의하지 않았다. 실험 결과 제작사 파워가 높을수록, 배우 파워가 높을수록 LSA_3 값이 높을수록, 네티즌 평점이 높을수록, 스크린 수가 많을수록, 연휴일수록 흥행이 증가하는 것으로 나타났다.

이후 머신러닝 모델들을 이용하여 영화 흥행 예측 모델을 만들고, 각 모델별로 예측력을 비교하였다. 예측 결과 SVM이 F1 Score가 85.34%로 가장 좋은 성능을 보였다.

흥행 예측 결과			
naive bayesian	Radom Forest	**SVM**	ANN
f1_score			
83.65	84.37	**85.34**	80.59
accuracy			
83.45	86.21	**86.90**	82.07
precision			
84.72	83.97	**85.23**	79.48
recall			
83.45	86.21	**86.90**	82.07

[표 9] 흥행 예측 결과

	비흥행	**흥행**	**대흥행**	**recall**
비흥행	108	10	0	92%
흥행	8	10	1	53%
대흥행	1	4	3	38%
precision	92%	42%	75%	

[표 10] Naive Bayesian Confusion Matrix

	비흥행	흥행	대흥행	recall
비흥행	115	3	0	97%
흥행	9	9	1	47%
대흥행	2	5	1	23%
precision	91%	53%	50%	

[표 11] Radom Forest Confusion Matrix

	비흥행	흥행	대흥행	recall
비흥행	115	3	0	97%
흥행	9	9	1	47%
대흥행	3	3	2	25%
precision	91%	60%	67%	

[표 12] SVM Confusion Matrix

	비흥행	흥행	대흥행	recall
비흥행	111	4	3	94%
흥행	11	7	1	97%
대흥행	3	4	1	12%
precision	89%	47%	20%	

[표 13] ANN Confusion Matrix

제6장 결론 및 한계점

영화의 가치사슬 단계별로 제작 단계, 배급 단계, 상영 단계로 나누어 '장르', '등급', '감독 파워', '배우 파워', '제작사 파워', '총 신의 수', '총 대화 수', '대화별 평균 길이', 'LSA_1', 'LSA_2', 'LSA_3', '네티즌 평점', '전문가 평점', '스크린 수', '연휴유무'를 통해 영화의 흥행을 예측해보았다.

순서형 로지스틱 회귀분석을 통해 예측변수들에 대해 통계적 유의성을 검증한 결과 제작 단계에서는 '배우파워', '제작사 파워', 'LSA_3'이 통계적으로 유의한 것으로 나타났다. 시나리오 구조나 장르 및 등급과 감독 파워는 유의하지 않은 것으로 나타났다. LSA_3이 Topic_3 ['부대', '전쟁', '공격', '군인', '작전']에 대한 변형된 TF-IDF인 것으로 보아 전쟁관련 작품들이 흥행에 유의한 것으로 판단된다. 배급 단계에서는 네티즌 평점이 유의한 것으로 나타났다. 사영 단계에서는 스크린 수와 연휴 유무가 유의한 것으로 나타났다. 이후 머신러닝 모델들을 통해 예측한 결과 SVM이 F1 Score 85.34%로 가능 좋은 성능을 보였다.

본 연구의 기여점은 영화의 가치 사슬 단계별로 영화 흥행에 영향을 미치는 변수들을 재정의하고, 최근 대규모의 예산이 투입되는 대작들이 영화 시장을 잠식해나가는 상황에서 영화 제작의 다양성 재고를 위해 기존 국내 영화 예측 연구에서는 잘 다루지 않았던 시나리오 텍스트 데이터를 영화 예측 연구에 활용하여 유의한 의미를 찾았다는 데 있다. 이는 다양한 예측 변수를 개발 및 제안함으로써 차후 영화 예측 연구나 실무에서 영화 제작 및 기획단계에서 활용할 수 있는 정보를 제공할 수 있다고 생각된다.

연구의 한계점은 클래스가 편향된 데이터와 데이터 수의 부족으로 흥행과 대흥행에 대한 낮은 예측률을 극복하는데 한계를 보였다. 이는 차후 연구에서 더 많은 데이터 확보를 통해 극복할 수 있다고 기대된다.

또 영화 산업에서 수익은 티켓 판매뿐만 아니라 광고, 굳즈, DVD 등 다양한 채널이 존재한다. 최근에는 디지털 기술의 발달로 IPTV, OTT(Over The Top) 등 수익 창출 채널이 더욱 다양해졌다. 그러나 본 연구에서는 티켓 판매 이외의 다른 매출 창구에 대한 데이터를 구할 수 없어 흥행에 대한 기준을 누적관람객을 바탕으로 만들었다. 따라서 본 연구에서는 티켓 판매 외의 다른 매출에 대한 고려를 하지 못했다. 차후 연구에서 영화의 다양한 수익 구조까지 연구범위를 확장한다면 더 좋은 연구결과가 있으리라 예상된다.

또 본 연구에서는 가수, 아이돌이나 브라운관에서 주로 출연하는 등 비영화인 스타들의 흥행에 대한 영향력은 고려하지 못하였고, 연구 범위를 국내 영화를 대상으로 했지만 연구 범위를 외국 영화 시장까지 확대하여 국내 시장과 외국 시장의 차이를 분석하고 외국 시장의 흥행 예측 연구를 한다면 보다 좋은 연구 모델을 제안할 수 있을 것으로 예상된다.

참고문헌

[국내문헌]

김은미 (2003). "한국 영화의 흥행 결정 요인에 관한 연구." 한국언론학보 47(2): 190-220.

김성진, and 안현철. "기업신용등급 예측을 위한 랜덤 포레스트의 응용." 산업혁신연구 32.1 (2016): 187-211.

김종국 "미디어생태학 관점의 다양성영화에 관한 고찰."

남기연 (2017). "영화산업의 독과점 실태와 해소 방안." 스포츠엔터테인먼트와 법 (JSEL) 20(4): 125-146.

배장수 (2015). "영화의 다양성과 한국영화산업의 현주소." 영산법률논총 12(1): 125-154.

변재란 (2006). "문화다양성, 영화다양성 그리고 다양성영화-영화문화다양성을 둘러싼 다양한 논의들에 대한 담론 분석." 영상예술연구 9: 9-49.

사영준 and 유승호 (2019). "영화산업에서의 소비 집중도와 산업의 성장성의 관계: 49 개국의 영화 흥행 데이터를 중심으로." 문화산업연구 19(2): 1-8.

송민. "텍스트 마이닝." (2012).

유진은. "랜덤 포레스트." 교육평가연구 28 (2015): 427-448.

유현석 (2002). "한국영화의 흥행 요인에 관한 연구: 제작 관련 변수를 중심으로." 한국언론학보 46(3): 183-213.

한국 and 통계 (2011). "영화 흥행 결정 요인과 흥행 성과 예측 연구." 한국통계학회논문집 18(6): 859-869.

[해외 문헌]

Ahmad, J., et al. (2017). Movie success prediction using data mining. 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE.

Bae, G. and H.-j. Kim (2019). "The impact of movie titles on box office success." Journal of Business Research 103: 100-109.

Basuroy, S., et al. (2003). "How critical are critical reviews? The box office effects of film critics, star power, and budgets." Journal of marketing 67(4): 103-117.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T.,

Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88, 2783-2792.

De Vany, A. (2003). Hollywood economics: How extreme uncertainty shapes the film industry, Routledge.

Eliashberg, J., et al. (2006). "The motion picture industry: Critical issues in practice, current research, and new research directions." Marketing science 25(6): 638-661.

Eliashberg, J., et al. (2014). "Assessing Box Office Performance Using Movie Scripts: A Kernel-Based Approach." IEEE Transactions on Knowledge and Data Engineering 26(11): 2639-2648.

Eliashberg, J., et al. (2007). "From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts." Management Science 53(6): 881-893.

Hunter, I., et al. (2016). "Predicting box office from the screenplay: A text analytical approach." Journal of Screenwriting 7(2): 135-154.

Kim, T., et al. (2015). "Box office forecasting using machine learning algorithms based on SNS data." International Journal of Forecasting 31(2): 364-390.

Kim, Y.-H. and J.-H. Hong (2011). "A Study for the Development of Motion Picture Box-office Prediction Model." Communications for Statistical Applications and Methods 18(6): 859-869.

Lash, M. T. and K. Zhao (2016). "Early Predictions of Movie Success: The Who, What, and When of Profitability." Journal of Management Information Systems 33(3): 874-903.

Lee, K., et al. (2018). "Predicting movie success with machine learning techniques: ways to improve accuracy." Information Systems Frontiers 20(3): 577-588.

Litman, B. R. (1998). The motion picture mega-industry, Allyn & Bacon.

Poter, M. E. (1985). Competitive advantage: Creating and sustaining superior performance, New York: The Free Press.

Sharda, R. and D. Delen (2006). "Predicting box-office success of motion pictures with neural networks." Expert Systems with Applications 30(2): 243-254.

Zhang, L., et al. (2009). "Forecasting box office revenue of movies with BP neural network." Expert Systems with Applications 36(3): 6580-6587.

Zhou, Y., et al. (2019). "Predicting movie box-office revenues using deep neural networks." Neural Computing and Applications 31(6): 1855-1865.

Abstract

The Determinants of Box Office Performance in Korea with Machine Learning Technique

Hwang Injik
Department of Business Informatics
Graduate School of
Hanyang University

Korea film market is the fifth largest market in the world and the scale of the industry has steadily grown up. Recently, revenue streams have diversified due to the growth of the media market such as OTT and IPTV. In addition, as the characteristics of film industry are generally high-risk and high-return, a box office record of the movies has a major impact on difference in performance. Therefore, research on predicting box office success of movies is important and essential. Recent trends in film market are that a few of large films dominate the market through vertical integration.

This paper categorizes the determinants of box office performance according to the value chain and proposes new prediction variable in

production section and forecasting model for box office success in order to reconsider the diversity of film industry and improve the quality of research on forecasting box office success.

In particular, after scenario text data were analysed by using latent sentiment analysis and the concepts of TF-IDF, new and diverse variables in the production phase that were not well addressed in previous studies were added. Then, the predictors of box office success are divided by value chain stage of the movie industry and are compared through machines learning models after validating the statistical significance of the variables through ordinal logistic regression.

This is expected to contribute to supporting research on film prediction using scenarios and decision making for investment in the process of film production.

keyword : Machine Learning, Ordinal Logistic Regression, LSA(Latent Sentiment Analysis), TF-IDF, Predicting box office Performace

연구 윤리 서약서

본인은 한양대학교 대학원생으로서 이 학위논문 작성 과정에서 다음과 같이 연구 윤리의 기본 원칙을 준수하였음을 서약합니다.

첫째, 지도교수의 지도를 받아 정직하고 엄정한 연구를 수행하여 학위논문을 작성한다.

둘째, 논문 작성시 위조, 변조, 표절 등 학문적 진실성을 훼손하는 어떤 연구 부정행위도 하지 않는다.

셋째, 논문 작성시 논문유사도 검증시스템 "카피킬러"등을 거쳐야 한다.

2019년12월13일

학위명 : 석사

학과 : 비즈니스인포매틱스학과

지도교수 : 임규건

성명 : 황인직 (서명)

한 양 대 학 교 대 학 원 장 귀 하

Declaration of Ethical Conduct in Research

I, as a graduate student of Hanyang University, hereby declare that I have abided by the following Code of Research Ethics while writing this dissertation thesis, during my degree program.

"First, I have strived to be honest in my conduct, to produce valid and reliable research conforming with the guidance of my thesis supervisor, and I affirm that my thesis contains honest, fair and reasonable conclusions based on my own careful research under the guidance of my thesis supervisor.

Second, I have not committed any acts that may discredit or damage the credibility of my research. These include, but are not limited to : falsification, distortion of research findings or plagiarism.

Third, I need to go through with Copykiller Program(Internet-based Plagiarism-prevention service) before submitting a thesis."

DECEMBER 13, 2019

Degree :	Master
Department :	DEPARTMENT OF BUSINESS INFORMATICS
Thesis Supervisor :	Gyoo Gun Lim
Name :	HWANG INJIK (Signature)