공학석사 학위논문

의사결정나무와 다중회귀분석을 활용한 영화 VOD 흥행 예측 및 주요 흥행 요인 분석

Analyzing the Characteristics of Movie VOD Success using Decision Tree and Multiple Linear Regression

2017년 2월

서울과학기술대학교 일반대학원

데이터사이언스학과

문 진 희

의사결정나무와 다중회귀분석을 활용한 영화 VOD 흥행 예측 및 주요 흥행 요인 분석

Analyzing the Characteristics of Movie VOD Success using Decision Tree and Multiple Linear Regression

지도교수 금영정

이 논문을 공학석사 학위논문으로 제출함
2017년 1월

서울과학기술대학교 일반대학원
데이터사이언스학과

문 진 희

문진희의 공학석사 학위논문을 인준함
2017년 1월

심사위원장 김경옥 (인)

심사위원 금영정 (인)

심사위원 홍정식 (인)

목 차

요 약

제 목 : 의사결정나무와 다중회귀분석을 활용한 영화 VOD 흥행 예측 및 주요 흥행 요인 분석

영화 배급 플랫폼의 확대와 관객들의 콘텐츠 소비 변화로 인해 최근 영화 산업에서 디지털 온라인 시장의 중요성이 매우 커지고 있다. 디지털 온라인 시장의 중요성이 높아짐에 따라 관련 연구도 증가하고 있지만, 기존 연구의 대부분이 디지털 온라인 시장 특성에 대한 정성적 연구와 소비자들의 VOD 구매 요인에 대한 연구로, 영화 VOD 콘텐츠 흥행에 관한 실증 연구는 매우 부족하다. 따라서 본 연구에서는 데이터마이닝 기법을 활용하여 영화 VOD의 수요를 예측하고, 흥행하는 VOD와 흥행하지 못한 VOD의 특징을 살펴보았다.

본 연구는 다음과 같은 단계로 구성된다. 첫째, 2013년부터 2016년 8월까지 서비스된 영화 VOD 중에서 월별 이용 건수가 상위 100위 안에 포함되는 영화 230편의 데이터를 수집한다. 둘째, 의사결정나무 기법과 다중회귀분석을 통해 training data로 예측 모델을 구축하고, test data에 대한 6개월간의 VOD 누적 이용 건수를 예측한다. 셋째, 두 가지 예측 모델에서 사용된 공통 주요 변수인 최대 스크린 수, 홀드백, 4주차 관객 수, 리뷰 수, 청소년 관람 불가 등급의 다섯 가지 변수를 VOD 흥행에 영향을 미치는 주요 요인으로 도출하였다. 넷째, 흥행 VOD와 비흥행 VOD로 영화를 분류하고, 각 그룹 별로 군집 분석을 실시한다. 흥행 VOD에서는 신선한 VOD, 천만 관객 영화, 스타 파워, 19금 범죄액션의 4개 군집이 도출되었고, 비흥행 VOD에서는 진부한 해외영화, 소규모 마니아, 무관심 19금의 3개 군집이 도출되었다. 본 연구의 결과는 영화 산업 활성화를 위한 정책 수립과 영화 관계자들의 의사결정에 효과적으로 활용 될 수 있을 것으로 기대된다.

표 목 차

그림목차

Ⅰ. 서 론

1. 연구 배경 및 필요성

영화는 예술의 한 장르를 넘어선 거대 엔터테인먼트 산업으로서 전 세계적으로 지속적인 성장세를 보이고 있으며, 최근 영화 배급 플랫폼의 확대와 관객들의 소비 성향이 변화하면서 디지털 온라인 영화시장의 중요성이 날로 높아지고 있다(전범수, 2007; 한국콘텐츠진흥원, 2015). 국내 영화 시장 또한 2015년 극장 매출 1조 7,154억 원으로 역대 최대 액수를 기록했으며, 디지털 온라인 시장의 매출은 전년에 비해 12.7% 증가한 3,349억 원을 기록했다(영화진흥위원회, 2016). 이처럼 영화 산업에서 디지털 온라인 시장의 중요성이 부각되면서 VOD의 성공을 예측하고 주요 요인을 판단하기 위한 연구의 필요성이 커지고 있다.

특히 디지털 온라인 시장 중 영화 VOD 산업의 경우 그 중요성이 더욱 커지고 있다. 영화산업의 경우 흥행여부, 즉 극장의 입장권 매출에 크게 좌우되는 고위험 고수익 구조로, 리스크 해소를 위한 수입의 다각화 방안을 마련하는 것이 필수적으로 고려되고 있다. 실제로 극장 흥행에 대한 의존도를 줄이기 위해 배급사들이 2차 유통시장으로 움직이고 있다. 이는 원하는 시간에 원하는 영화를 볼 수 있다는 강점 때문에 OTT (Over-The-Top) 및 IPTV 서비스를 통해 VOD 수요가 증가하기 때문으로도 이해할 수 있다 (진성철 and 박원준, 2013; 임성준 and 강정헌, 2006). 따라서 VOD 의 흥행에 영향을 미치는 주요 요인들을 파악하고 VOD의 수요를 예측하는 것이 매우 중요한 활동이라 하겠다.

하지만 디지털 온라인 시장에 대한 기존 연구들은 디지털 기술이 영화 배급 시장에 끼친 영향에 대한 연구와 소비자들의 VOD 구매 요인에 대한 연구가 주를 이루고 있다. 임예원(2004)은 디지털 미디어의 등장으로 인한 영화 산업의 시장 확장 현상이 글로벌 미디어 기업, 한국 영화 산업, 국내 미디어 기업에게 미친 영향을 분석하였고, 김진욱(2010)은 디지털 기술 발전에 따라 급변하는 미디어 융·복합 환경 속에서 영화 산업의 변화를 사례분석을 통해 현상을 파악하고, 영화 산업의 발전을 위한 핵심과제로 불법문제 해결, 공정경쟁 환경의 조성, 콘텐츠 제작 및 유통 경쟁력 강화, 법제도의 개선의 네 가지 대안을 제언하였다. 이상우 and 김창완(2009)은 이산적 선택모형을 이용한 회귀분석을 통해 IPTV-VOD의 채택에 영향을 주는 요인을 분석하고, IPTV-VOD 서비스와 기존 유료방송서비스는 대체적인 성격이 있다는 것을 밝혔다. 황준석 외(2009)는 홀드백 기간에 대한 민감도, 영화 VOD 장르 선택의 다양성, 인터넷 VOD 극장의 유용성, 몰입 요인이 인터넷

VOD 극장 이용 의도에 긍정적 영향을 준다는 점을 회귀분석을 통해 도출하였다. 이와 같이 기존 연구의 대부분이 전문가의 정성적 분석에 의존하는 시장 동향 연구와 설문을 기반으로 한 VOD 서비스 선택 요인에 대한 연구에 집중되어 있다.

그러나 대부분 연구에서 VOD 서비스 선택 요인 등 정성적 관점의 연구들은 다수 진행되어 왔으나, VOD 수요를 예측하거나 VOD 흥행을 위한 데이터 기반의 요인 도출 등 정량적 관점의 연구가 부족한 실정이다.

2. 연구 목적

본 연구는 데이터 마이닝 기법을 활용하여 콘텐츠 관점에서 영화 VOD의 수요를 예측하고, 디지털 온라인 시장에서의 영화 VOD 흥행 요인과 흥행 여부에 따른 VOD의 특징을 분석하고자 한다. 영화의 고유 특성과 극장 흥행 및 관객 반응 변수 통해 VOD 수요를 예측하고, 예측 모델에서 유의한 변수를 추출하여 영화 VOD에 대한 주요 흥행 요인으로 정의한다. VOD 흥행에 따라 흥행 VOD 그룹과 비흥행 VOD 그룹으로 분류하고, 그룹별로 군집분석을 수행하여 그룹의 세부 특징을 분석한다. 영화 VOD 수요 예측을 통해 디지털 온라인 시장에서 해당 콘텐츠의 수익을 미리 파악 할 수 있으며, 영화 VOD의 흥행 특징을 통해 영화 관계자들이 콘텐츠 배급 의사결정에 활용될 수 있을 것으로 기대된다.

3. 연구 구성

본 논문의 구성은 다음과 같다. 2장에서 디지털 온라인 시장에 대한 연구와 영화 산업에서의 데이터 마이닝 활용 연구에 대해 살펴보고, 3장에서는 연구 방법을 제시한다. 4장에서는 분석한 결과를 통해 디지털 온라인 시장에서 영화 VOD 수요 예측 결과와 주요 흥행 요인 및 군집별 특성을 제시한다. 마지막으로 5장에서는 연구의 결론 및 의의와 추후 연구 방향을 서술한다.

Ⅱ. 문헌 연구

1. 영화 산업에서의 디지털 온라인 시장

영화 산업에서 디지털 온라인 시장이란, 극장 종영 이후 부수적으로 발생하는 수익을 창출하는 시장으로 과거엔 관행적으로 부가시장이라 불렸으나, 부가시장의 디지털화로 인해 그 용어를 디지털 온라인 시장으로 대체하게 되었다(영화진흥위원회, 2012). 디지털 온라인 시장은 홈비디오(VHS)를 제외한 TV VOD(IPTV 및 디지털케이블 TV), 인터넷 VOD, 모바일 VOD, DVD 및 Blu-ray 시장을 모두 포함하는 의미로 쓰이지만(영화진흥위원회, 2016), 본 연구에서는 패키지상품으로 분류되는 DVD 및 Blu-ray 시장은 제외하고 논하기로 한다.

2015년 PwC 보고서에 따르면, 디지털 온라인 시장은 2019년까지 연평균 10.5% 증가하여 전체 영화시장의 30.3%를 차지할 것으로 전망하였다. 이는 원하는 시간에 원하는 콘텐츠를 이용할 수 있다는 VOD의 장점으로 인해, 넷플릭스와 같은 OTT(Over-The-Top)와 IPTV 및 인터넷 VOD와 같은 서비스를 통한 영화 콘텐츠 소비가 증가하고 있기 때문이다(진성철 and 박원준, 2013; 임성준 and 강정헌, 2006). 2010년까지만 해도 전체 영화 시장의 50%를 차지하던 홈 비디오(VHS, DVD, Blu-ray) 시장이 지속적으로 감소하고 디지털 온라인 시장이 빠르게 성장함으로써, 영화 부가시장은 디지털 온라인 영화시장 중심으로 개편되는 중이라고 볼 수 있다(PwC, 2015; 영화진흥위원회, 2015).

한국은 2000년대에 영화 상영의 매체가 다양해지면서 불법 유통 및 홈비디오의 몰락 등 부가시장의 혼란과 쇠퇴의 시기도 있었으나, 2009년 이후 새 플랫폼들의 성장으로 인해 매출규모가 급속도로 증가하고 있다(영화진흥위원회, 2012; 민병현, 2013). PwC(2015)는 한국 부가시장의 매출을 2014년 2억 6,200만 달러에서 5년 후 2배 증가한 5억 2,900만 달러 규모로 전망하고 있다. 또한, 한국 영화 산업의 수익 구조가 극장 흥행에 따라 심각하게 좌우되는 형태이므로, 리스크 분산의 한 대안으로써 디지털 온라인 시장이 떠오르기도 한다. 이와 같이 산업 측면에서 중요성이 높아짐에 따라 디지털 온라인 시장에 대한 연구도 활발하게 이루어지고 있다.

기존 연구는 주로 인터넷과 정보통신 기술의 발달로 인한 영화 산업의 변화에 대한 연구(김진욱, 2010; 민병현, 2013)와 소비자들의 VOD 구매 요인에 대한 연구(황준석 외, 2009; 고정민 외, 2010)로 이루어져 있다. 민병현(2013)은 뉴미디어 기반 영화배급 플랫폼이 부가시장에 미치는 영향을 살펴보고, 해당

시장의 발전 가능성에 대한 검토와 함께 시장 확대를 위한 정책안을 제안하였다. 고정민 외(2010)는 관객들의 영화 관람 방식에 따라 응답자를 세분화 하고 세분시장별 특성 및 새로운 인터넷 VOD 서비스 이용의향을 분석하였다. 이용의향이 있는 응답자들에게서 합법적 서비스 제공, 국내외 최신작, 국내 유명포탈을 이용한 접근 용이성에 대한 기대가 높은 것으로 나타났다. 여기서, 국내외 최신작은 홀드백이 짧은 영화를 뜻하는데, 콘텐츠의 창구화 전략에 따른 홀드백에 대한 연구는 꾸준히 수행되어 왔다(김미현, 2006; 박선규 and 최성진, 2015). 박선규 and 최성진(2015)은 영화의 선행창구 성과와 장르, 홀드백이 영화 VOD 구매에 미치는 영향을 회귀분석을 통해 스크린 수가 많을수록, 홀드백이 짧을수록 VOD 구매 건수가 증가하는 것을 증명하였다.

이처럼 디지털 온라인 시장 관련 기존 연구들이 전문가 중심의 시장 동향 분석과 설문을 기반으로 소비자들의 서비스 선택요인에 관한 연구였다면, 본 연구는 기존 연구에서는 다루지 않았던 콘텐츠 관점에서의 흥행에 대한 정량적 분석을 수행하고자 한다.

2. 영화 산업에서의 데이터마이닝 활용 연구

상품이나 서비스, 콘텐츠에 대한 수요 예측 및 특징 분석 등을 위해 다양한 데이터 마이닝(data mining) 기법이 사용된다. 데이터마이닝은 데이터 속에서 의미 있는 정보나 패턴을 발견하기 위한 과정으로, 영화 산업 연구에서도 널리 활용되고 있다. 영화 VOD 콘텐츠에 대해 데이터 마이닝 방법론을 적용한 기존 연구가 거의 존재하지 않기 때문에 극장에서의 영화 흥행에 대한 연구를 살펴보기로 한다.

Ramesh and Dursun(2006)은 영화의 수익을 9개의 구간으로 나눠 인공신경망을 통해 분류모형을 만들었다. 로지스틱 회귀분석, 판별분석, 의사결정나무와 분류 성능을 비교하여 인공신경망의 분류 성능이 가장 좋음을 밝혔다. 하지만 인공신경망은 성능은 좋지만 모델에 대한 해석이 어려운 블랙박스 기법 중 하나로 흥행 요인을 분석하기 위한 방법론으로 적절하지 않다. 강지훈 외(2014)는 개봉 전 영화와 상영 중 영화에 대하여 두 가지 예측 프로세스를 제안하였으며, 영화 속성을 반영한 변수와 관객 반응을 고려한 변수, 마케팅 요소를 반영한 변수를 투입하여 의사결정나무를 통해 예측 모델을 구성함으로써 초기 스크린 수 결정에 근거를 제시하였다. 의사결정나무 기법은 데이터 분포에 대한 가정이 불필요하며 예측에 큰 영향을 미치는 소수의 설명변수를 찾을 수 있다는 장점이 있으나 분류에 자주 쓰이는 기법으로 연속형 변수에 대해서

예측 성능이 저하 될 수 있다.

또한 최근 영화 흥행 예측 연구에는 온라인 리뷰나 평점을 반영한 연구가 많이 수행되고 있다(Krushikanth et al.,2013; 허민회 외, 2013; 권선주, 2014). Krushikanth et al.(2013)은 트위터의 팔로워 수와 유투브 리뷰 데이터를 수집하여 k-means 군집분석을 통해 영화 흥행 관련 군집을 형성하고, 각 군집을 Hit, Neutral, Flop class로 할당하여 예측 모델을 설계하였다. 허민회 외(2013)는 개봉 전후 시점과 관객의 평점, 리뷰 변수 추가에 따라 예측 모델을 4개로 구성하여 각 성능을 비교하고, 평점과 평점의 개수가 들어간 모델에 온라인 리뷰 감성분석의 결과가 포함된 모델이 가장 성능이 우수함을 나타냈다. 권선주(2014)는 구전 효과를 반영하는 영화 흥행 예측 모델 구축을 위하여 관객의 평점과 기사 횟수를 변수로 활용하였다. 전체 영화, 한국 영화, 외국 영화에 따라 주차별로 회귀분석을 진행 후 유의한 변수를 도출하여, 각 시기에 맞는 마케팅 전략을 세울 수 있게 하였다.

이와 같이 영화 산업에서 데이터 마이닝 기법을 적용한 흥행 관련 연구는 크게 두 가지 관점에서 연구가 진행되고 있다. 첫 번째로 정확한 흥행 예측을 목적으로 기존 예측 모델과의 성능을 비교하여 성능의 우수함을 증명하는 연구가 있고, 두 번째로 흥행 요인을 파악하기 위해 성능보다는 설명력을 갖춘 방법론을 적용한 연구들이 있다. 본 연구에서는 흥행 예측과 더불어 흥행 요인에 대한 파악이 목적이므로 예측기법 중에서 모델의 해석이 용이한 의사결정나무와 다중회귀분석을 통해 영화 VOD 수요를 예측하며, 영화 고유 특성과 함께 관객의 평점과 리뷰 수, 미디어 노출을 나타낼 수 있는 기사 수를 변수로 포함 시켰다. 또한 영화의 특성을 묶기 위해 군집분석을 수행하였고 이때, k-means 군집분석보다 이상치의 영향을 덜 받는 k-medoids 군집 분석을 통해 VOD의 특징을 세분화하여 분석 하고자 한다.

Ⅲ. 연구방법

1. 연구 프레임워크

본 연구의 흐름은 Fig 3.1과 같이 데이터 구성, 영화 VOD 수요 예측 및 VOD 흥행 요인 도출, 영화 VOD 흥행 특성 분석으로 구성된다. 영화진흥위원회와 네이버 영화에서 수집한 데이터를 training data와 test data로 분할하여, 의사결정나무 분석과 다중회귀 분석을 통해 6개월간의 VOD 누적 이용 건수를 예측한다. 두 모델에서 공통적으로 중요 변수로 도출된 변수를 영화 VOD 흥행요인으로 선정한다. 이때, 흥행하는 VOD와 흥행에 실패한 VOD의 세부적인 특징을 살펴보기 위해 적정 VOD 사용 건수인 30만 건을 cut-off로 흥행 VOD 그룹과 비흥행 VOD 그룹으로 분류하고, 각 그룹별로 k-medoids 군집 분석을 수행하여 군집별 특징을 살펴본다.

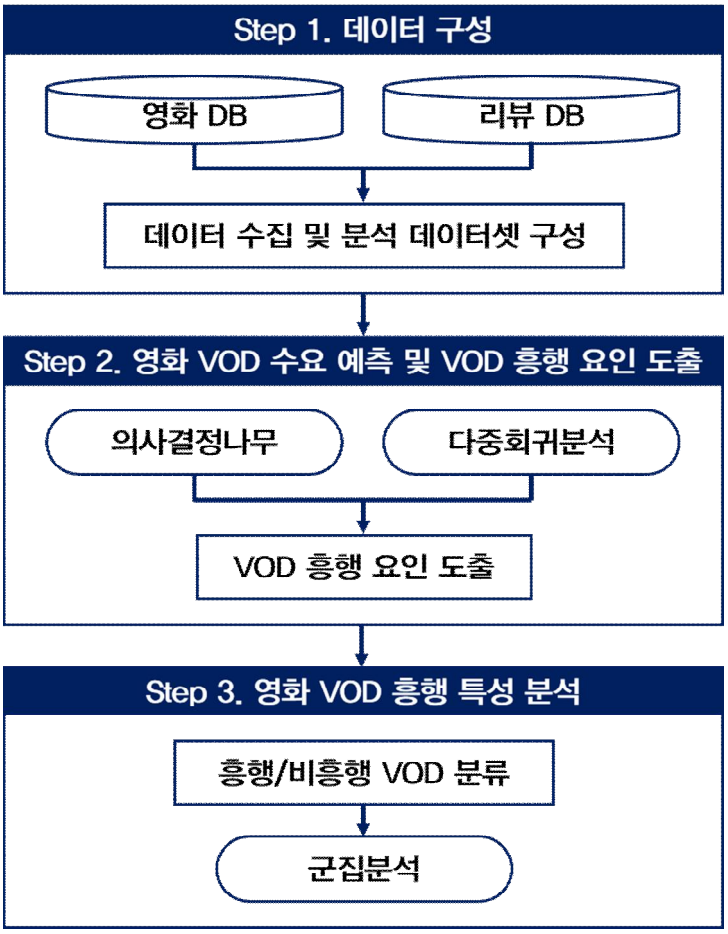

Fig 3.1 연구 프레임워크

2. 데이터 구성

우선, 영화진흥위원회(www.kofic.or.kr)로부터 2013년부터 2016년 8월까지 서비스 된 영화 VOD 중에서 월별 이용 건수가 상위 100위 안에 포함되는 영화 230편의 데이터를 수집한다. 그리고 네이버영화(http://movie.naver.com)에서 분석 대상 영화들의 시간별 평점과 리뷰 수를 python 3.4의 BeautifulSoup 패키지를 활용하여 웹 크롤링을 통해 모은다. 수집한 데이터의 출처와 설명을 Table 3.1에 정리하였다.

Table 3.1 분석용 데이터 변수 구성

변수명	유형	출처	변수 설명	최소값	평균	최대값
target	numeric	영화진흥위원회	영화 i의 vod 서비스 시작 후 6개월간 누적 이용건수	44951	275953.5	1184624
국가	categorical	영화진흥위원회	한국, 미국, 기타 국가	-	-	-
장르	categorical	영화진흥위원회	코미디, 가족/드라마, 범죄/스릴러, 액션/어드벤쳐/판타지, SF, 애니메이션, 로맨스, 공포/미스터리	-	-	-
관람 등급	categorical	영화진흥위원회	영화 관람등급(전체, 12세, 15세, 청소년관람불가)	-	-	-
감독 인지도	binary	영화진흥위원회	과거 작품의 관객이 200만 이상인 경우 1, 아니면 0	-	-	-
배우 인지도	binary	영화진흥위원회	주연 배우 과거 작품의 관객이 200만 이상인 경우 1, 아니면 0	-	-	-
배급사	binary	영화진흥위원회	연 평균 10편 이상 배급하는 배급사는 1, 아니면 0	-	-	-
원작 유무	binary	네이버, 구글 검색	원작이나 전편이 있는 경우 1, 아니면 0	-	-	-
스크린 수	numeric	영화진흥위원회	상영기간 중 최대 스크린 수	139	649.9	1602
week0	numeric	영화진흥위원회	영화 i의 개봉 전 시사회 관객 수	0	15446.6	435146
week1	numeric	영화진흥위원회	영화 i의 1주차 관객 수	40638	976151.2	6604292
week2	numeric	영화진흥위원회	영화 i의 2주차 관객 수	3007	654524	5149186
week3	numeric	영화진흥위원회	영화 i의 3주차 관객 수	0	356555	3357333
week4	numeric	영화진흥위원회	영화 i의 4주차 관객 수	0	183514.1	1914539
기사수	numeric	네이버 뉴스 검색	영화 개봉 후 2주차까지의 영화 i의 기사 수	10	769.1	5416
평균 평점	numeric	네이버 영화	영화 개봉 후 VOD 서비스 시작 전까지 영화 i의 평균 평점	3.15	7.74	9.31
평점 수	numeric	네이버 영화	영화 개봉 후 VOD 서비스 시작 전까지 영화 i의 평점 수	198	6692.9	85436
홀드백	numeric	영화진흥위원회	영화 개봉 후 VOD 서비스 시작까지 걸린 시간(단위: 월)	0	1.78	8

3. 영화 VOD 수요 예측

본 연구는 Fig 3.2와 같이 2013년 1월부터 2015년 12월 사이에 서비스를 시작한 VOD 213편을 training data, 2016년 1월부터 서비스를 개시한 17편의 VOD를 test data로 할당하여 예측 모델을 설계하였다. training data의 양이 많지 않은 관계로 training data를 leave-one-out cross-validation(LOOCV)을 통해 최적의 모델을 선택한 후, test data인 영화 17편에 대한 6개월간의 VOD 이용건수를 예측한다. 수요 예측은 의사결정나무 기법과 다중회귀분석 기법 두 가지를 수행하였고, 분석 프로그램으로는 통계분석 소프트웨어인 R을 활용하였다. 본 연구에서 의사결정나무와 다중회귀분석을 선택한 가장 큰 이유는 VOD의 수요 자체를 예측하는 것도 중요하지만 VOD의 성공여부를 결정하는 주요 변수를 도출하는 것이 연구의 핵심 목적이기 때문이다. 따라서 수요예측의 정확성에 초점을 맞추기보다는 예측의 과정 및 중요변수 등을 정확하게 파악할 수 있는 White-box 모델인 의사결정나무와 다중회귀분석을 선택하였다.

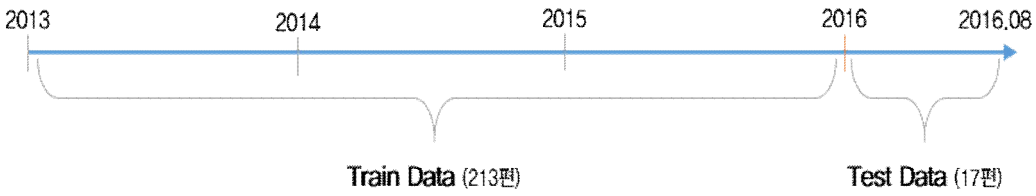

Fig 3.2 수요 예측 모델 구축을 위한 데이터 분할

1) 의사결정나무(Decision Tree) 수행

본 연구에서는 의사결정나무 알고리즘 중에서 가장 대표적인 CART (Classification and Regression Trees)를 사용하였다. CART 알고리즘은 해석의 용이성이 다른 알고리즘에 비해 뛰어난 것으로 알려져 있고(오윤학 외, 2014), 범주형과 연속형 독립변수 모두를 적용할 수 있다는 장점이 있다(전치혁, 2012). 전체 데이터를 포함하는 뿌리노드(root node)에서부터 각 독립변수를 이분화(binary split)하는 과정을 반복하여 트리 형태를 형성하며, 회귀 트리의 경우 각 노드의 분산을 감소시키는 형태로 재귀적 분기가 일어난다(Breiman et al., 1984). R의 rpart 패키지를 사용하였으며, LOOCV를 통해 최적의 파라미터 cp=0.01, min_split=25, max_depth=5를 적용하여, test data에 속한 영화 VOD들의 6개월간의 이용건수를 예측하였다.

2) 다중회귀분석(Multiple Linear Regression) 수행

다중회귀분석은 예측 모델에서 가장 많이 사용하는 방법론으로, 영화 수요 예측에서도 빈번하게 활용되고 있다(Mestyan et al., 2013; 권선주, 2014). 종속변수 Y를 설명하는 k개의 독립변수에 대하여 다중회귀모형은 다음과 같다.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

$\beta_0, \beta_1, \beta_2, ..., \beta_k$는 최소자승법(least squares method)에 의해 추정되는 회귀계수들이며, ϵ_i는 정규분포를 따르는 오차항이다. 모든 변수를 사용하는 경우와 전진 선택법(Forward Selection), 후진 소거법(Backward Elimination), 단계적 선택법(Stepwise Selection)에 의해 변수 선택을 하는 4가지 경우에 대하여 LOOCV를 통해 가장 RMSE가 낮은 단계적 선택법을 선택하였고, R의 lm 함수를 활용하였다.

3) 영화 VOD 주요 흥행 요인 도출

수요 예측 단계에서 의사결정나무 분석과 다중회귀분석을 수행하고 난 뒤, 두 예측 모델에서 공통적으로 중요 변수로 등장한 변수를 VOD 주요 흥행 요인으로 선정하고 추후 분석을 진행한다.

4. 영화 VOD 흥행 특성 분석

1) 흥행 / 비흥행 VOD 분류

영화 VOD 수요 예측 후 도출된 VOD 주요 흥행 요인에 대하여, 흥행하는 VOD와 흥행에 실패하는 VOD의 세부적인 특징을 살펴보고자 한다. 이때, 홀드백에 따라 차이가 있을 수 있지만 최소 10억 원 이상의 수익의 기준이 되는 30만 건을 기준으로 흥행 VOD 그룹과 비흥행 VOD 그룹으로 분류하고자 한다.

2) 군집분석 수행

이전 단계에서 선정된 VOD 흥행 요인들은 수요예측 모델을 통해 VOD 수요에 영향을 미치는 요인을 개별적으로 도출한 것이다. 하지만 실제로는 이러한 흥행 요인들이 복합적인 형태로 VOD 흥행에 영향을 끼칠 것이기 때문에 이를 살펴보기 위하여 흥행, 비흥행 VOD 그룹별로 군집분석을 수행한다. 영화 데이터의 특성상 소수의 영화가 천만이상의 관객을 동원하게 되면서 이상치에 가까운 값을 가질 수 있기 때문에 군집분석 방법론 중에서도 이상치에 민감하지 않은 k-medoids 군집분석을 수행한다. k-medoids 군집분석은 각 군집의 중심 좌표(centroid)를 고려하고 있는 k-means 군집분석과 다르게, 각 군집의 대표 객체(medoid)를 고려하는 기법이다. 군집의 대표 객체란 해당 군집에 속하는 객체 중 다른 객체들과의 거리가 최소가 되는 객체이다. k-medoids 군집분석에서 가장 대표적인 알고리즘인 PAM(partitioning around medoids)으로 군집분석을 수행하였고 R의 cluster 패키지 중 pam 함수를 통해 분석하였다. 적절한 군집 개수 k를 선정하기 위하여 실루엣 계수(Silhouette Width)를 활용하였다(Kaufman and Rousseeuw, 1990).

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

$a(i)$는 객체 i와 동일 군집에 속한 다른 객체들과의 평균 거리, $b(i)$는 객체 i와 가장 근접한 군집에 속한 객체들과의 평균거리를 의미한다. $s(i)$는 -1과 1 사이의 값을 갖고, 1에 가까울수록 군집화가 잘 된 것으로 평가한다.

Ⅳ. 연구 결과

1. 영화 VOD 수요 예측 결과

1) 의사결정나무 수요 예측 모델 결과

Figure 4.1와 Table 4.1는 VOD 수요예측을 위해 의사결정나무 수행한 결과이다. 각 노드별 평균이 VOD 흥행 기준인 30만 건 이상은 흥행 노드로, 30만 건 미만은 비흥행 노드로 할당하였다. 첫 번째 리프노드(4)는 개봉 첫 주 성적이 저조하며, 스크린 수가 582개 이하로 중소규모의 영화로 볼 수 있다. 데이터의 절반가량이 첫 번째 노드에 속하며, 그 특징은 배급 규모 자체가 크지 않고, 대부분의 애니메이션 영화 및 마니아층의 영화들이다. 두 번째 리프노드(20)는 대부분이 원작이나 전작이 존재하여 기대감이 높았으나, 관객의 기대를 만족시키지 못해 화제성이 떨어진 영화들이다. 세 번째 리프노드(42)는 대부분 인기 배우들의 작품이지만 흥행성적은 스타 파워에 비해 저조한 편이다. 네 번째 리프노드(43)는 화제성 있는 19세 미만 관람불가 영화로, 인터넷 기사 수가 많으며, 극장 관객 수에 비해 리뷰수가 많은 편이다. 또한 19금 영화 특성상 영화관에서 보다 VOD의 수요가 높다는 것은 세 번째 리프노드(42)번과 비교해보면 알 수 있다. 다섯 번째 리프노드(11)는 탄탄한 작품성과 관객의 좋은 평가로 입소문이 퍼져 다른 영화에 비해 뒤늦게 흥행가도에 오른 영화이다. 그러다보니 홀드백이 다른 노드에 비해 긴 편이다. 나머지 세 개의 노드는 극장에서 어느 정도 흥행성과를 거둔 영화들로 구성되어 있다. 여섯 번째 리프노드(12)는 화려한 액션과 SF 장르가 주를 이루는 극장에서 보기 좋은 할리우드 블록버스터 영화들로 구성되어 있다. 따라서 TV로 관람해야하는 VOD는 극장 흥행과 비교해 보았을 때 성과가 좋지 않은 편이며, 흥행 VOD와 비흥행 VOD가 반씩 섞여있다. 일곱 번째 리프노드(13)는 한국형 블록버스터 영화로써, 영상이 화려하고 웅장한 할리우드 블록버스터와는 다르게 우리나라 인기 배우를 중심으로 한 액션 및 범죄/스릴러 장르의 영화들로 구성되어 있다. 발빠른 VOD 진입으로 VOD에서 좋은 성과를 얻었다. 마지막으로 일곱 번째 리프노드(7)는 절반 이상이 천만 관객을 돌파한 영화이며, 나머지 영화 또한 천만에 가까울 정도로 흥행한 영화들로서, 극장에서도 VOD에서도 성공한 메가 히트작이다. 이렇게 의사결정나무 수요 예측 모델을 통하여, 새로운 영화 데이터가 들어오더라도 규칙을 따라 대략적인 VOD 사용 건수를 예측할 수 있다.

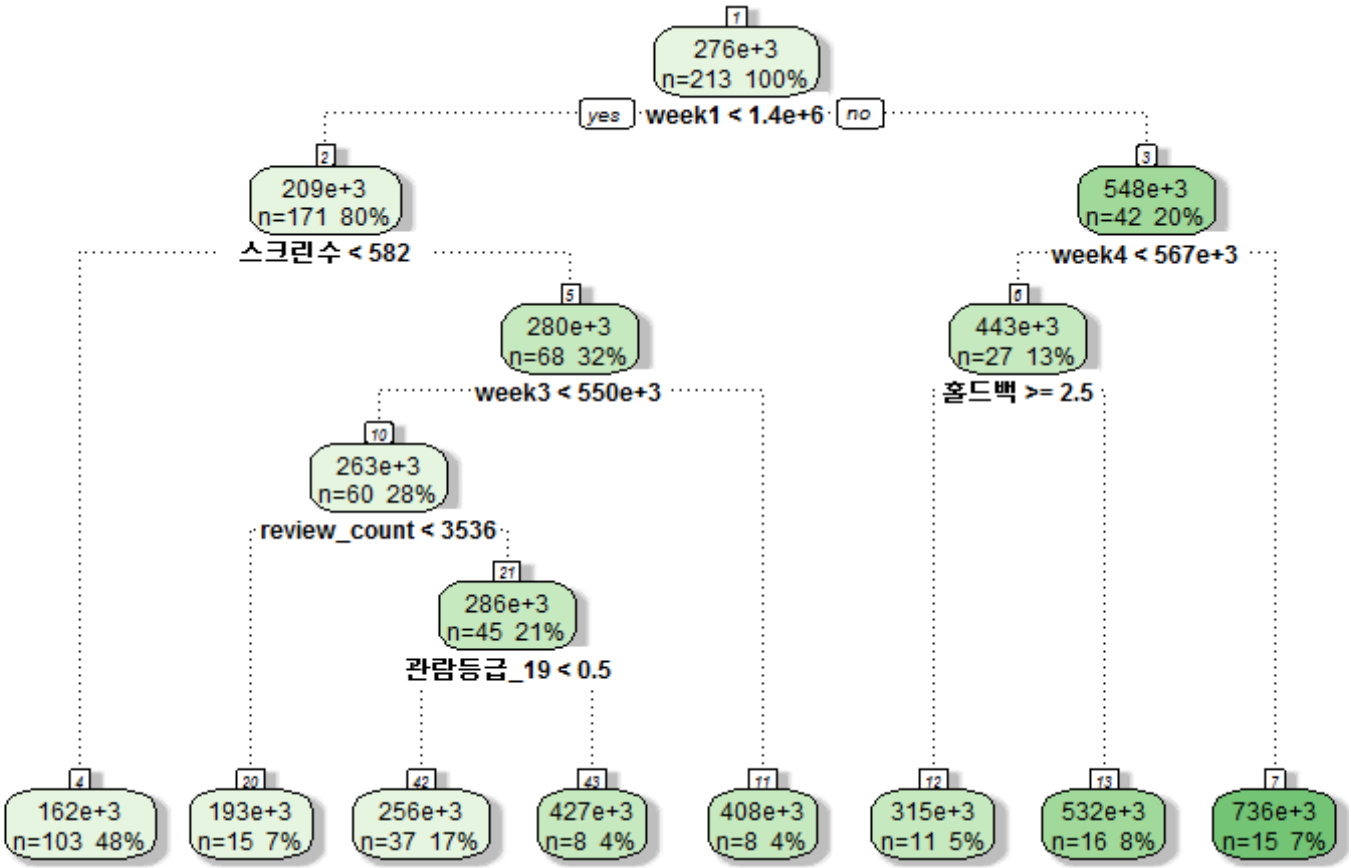

Fig 4.1 의사결정나무 VOD 수요 예측 모델

Table 4.1 의사결정나무 VOD 수요 예측 모델의 리프노드 특징

노드	규칙	대표 영화	영화 수	흥행 여부
4	1주차 관객수<1.4e+6 &스크린수<582	그레이의 50가지 그림자, 퍼시 잭슨과 괴물의 바다, 연애의 맛	103	비흥행
20	1주차 관객수<1.4e+6 &스크린수>=582 & 3주차 관객수<550e+3 &리뷰 수 < 3536	빅매치, 300: 제국의 부활, 살인의뢰	15	비흥행
42	1주차 관객수<1.4e+6 &스크린수>=582 & 3주차 관객수<550e+3 &리뷰 수 >= 3536 &not 청소년관람불가	손님, 논스톱, 스물	37	비흥행
43	1주차 관객수<1.4e+6 &스크린수>=582 & 3주차 관객수<550e+3 &리뷰 수 >= 3536 &청소년관람불가	인간중독, 간신, 내부자들: 디오리지널	8	흥행
11	1주차 관객수<1.4e+6 &스크린수>=582 &3주차 관객수>=550e+3	인턴, 분노의 질주: 더 세븐, 끝까지 간다	8	흥행
12	1주차 관객수>=1.4e+6 &4주차 관객수<567e+3 &홀드백>=2.5	트랜스포머: 사라진 시대, 혹성탈출: 반격의 서막, 월드 워 Z	11	흥행+비흥행
13	1주차 관객수>=1.4e+6 &4주차 관객수<567e+3 &홀드백<2.5	은밀하게 위대하게, 역린, 신의 한 수	16	흥행
7	1주차 관객수>=1.4e+6 &4주차 관객수>=567e+3	7번방의 선물, 암살, 명량	15	흥행

2) 다중회귀분석 수요 예측 모델 결과

단계적 변수 선택법을 적용한 다중회귀 분석의 결과가 Table 4.2에 나타나 있다. 유의한 변수 위주로 살펴보면, 미국이나 기타 국가에 비해 한국 영화일수록 VOD 이용 건수가 높고, 로맨스 장르는 VOD에서 인기가 떨어지는 것으로 나타났다. 연간 배급을 많이 하는 배급사의 영화가 VOD에서 인기가 있었지만, 의외로 감독 인지도는 영향이 없는 것으로 나타났다. 홀드백은 짧을수록 이용건수가 높아지며, 극장에서의 흥행 성적이 VOD이용 건수에 가장 큰 영향을 끼친다. 특이한 점은, 리뷰의 수가 많을수록 VOD 이용 건수가 낮아진다는 결과가 도출되었는데, 이는 재미가 없거나 실망한 영화에 대해서도 관객의 평이 많아지기 때문에 나타난 결과로 해석된다. 관람 등급에서는 15세 이상 관람가와 청소년 관람불가 등급이 유의한 변수로 나타났는데, 특히 청소년 관람불가 영화는 나이 제한으로 영화관에서 관람을 하지 못한 청소년들이나, 사적인 공간에서 관람을 원하는 소비자들에 의해 큰 인기를 얻고 있다.

Table 4.2 다중회귀분석 VOD 수요 예측 모델 결과

변수명	계수	t-값	p-값
한국***	1.27E+05	3.89	0.000137
미국	6.41E+04	1.936	0.054358
기타국가	4.55E+04	1.423	0.156187
가족/드라마	-2.49E+04	-1.616	0.107701
범죄/스릴러	-3.08E+04	-1.654	0.099697
액션/어드벤쳐/판타지	3.00E+04	1.716	0.087761
로맨스*	-6.02E+04	-2.394	0.017628
공포/미스터리	-6.52E+04	-1.579	0.115901
스크린수**	1.19E+02	2.723	0.007053
배급사*	4.66E+04	2.117	0.035536
감독 인지도*	-3.97E+04	-2.362	0.019135
홀드백**	-2.07E+04	-3.298	0.001155
week2***	1.34E-01	6.159	4.06E-09
week4***	2.02E-01	5.292	3.23E-07
review_count***	-4.57E+00	-3.329	0.001042
관람등급_15*	4.33E+04	2.43	0.016018
관람등급_19***	8.53E+04	3.704	0.000276
Multiple R-squared: 0.9158, Adjusted R-squared: 0.9085 F-statistic: 125.4 on 17 and 196 DF, p-value: < 2.2e-16			

3) 종합 결과

training data에 대하여 LOOCV을 적용하여 validation data의 RMSE(Root Mean Squared Error)와 MAPE(Mean Absolute Percentage Error)가 낮은 모델을 선택하고, 이를 test data에 적용하여 6개월간의 VOD 이용 건수를 예측한 결과가 Table 4.3이다.

Table 4.3 test data에 대한 VOD 수요 예측 결과

영화	Target	의사결정나무	다중회귀분석
히말라야	506909	735993.2	608355.3
조선마술사	88882	255823.5	149144.5
굿 다이노	215327	255823.5	210793.4
쿵푸팬더3	364904	531679.8	429476.5
몬스터 호텔 2	174749	255823.5	227304.1
극장판 꼬마버스 타요의 에이스 구출작전	101936	162375.2	94821.0
데드풀	452917	531679.8	422405.8
검사외전	874477	531679.8	758276.6
널 기다리며	119736	162375.2	309254.1
나쁜놈은 죽는다	42634	162375.2	215479.2
레버넌트: 죽음에서 돌아온 자	275263	255823.5	290152.1
스파이 브릿지	78937	162375.2	6509.4
제 5침공	182986	162375.2	150816.0
바다 탐험대 옥토넛 시즌4: 빙하탐험선S	84264	162375.2	89146.0
남과 여	94178	162375.2	230308.4
캐롤	46949	162375.2	194276.1
빅쇼트	68224	162375.2	221651.9
RMSE		132975.7	100427
MAPE		85.3	90.1%

의사결정나무의 RMSE는 132975.7, MAPE는 85.3%로 높은 MAPE를 보였으며, 다중회귀분석의 RMSE는 100427, MAPE는 90.1%로 의사결정나무 보다 RMSE의 성능은 더 좋았지만, MAPE의 측면에서 성능은 더 안 좋은 것으로 나타났다. 의사결정나무는 분류 기법에서 자주 활용되는 방법론으로, 회귀 트리에서는 노드의 평균값으로 예측이 되기 때문에 에러율이 다소 높을 수 있다는 한계점이 존재한다. 그리고 test data에 몇 개의 영화들의 실제 이용 건수가 매우 낮은 편이라 평균값이 가장 낮은 노드와도 차이가 커 오차율이 높은 것으로 확인된다. 다중회귀분석의 결과를 살펴보면, ‘굿 다이노’와 ‘레버넌트: 죽음에서 돌아온 자’와 같은 영화들은 예측이 매우 잘 되었고, ‘나쁜놈

은 죽는다’, ‘캐롤’과 같은 영화가 예측 오차가 매우 나 평균적으로 MAPE가 높게 나왔다. test data에서 발견되는 극단적인 이상치는 VOD 판권 계약 문제나 프로모션을 통한 이용건수의 변동과 같은 외부적인 요인이 있을 것으로 판단된다. 실제로 영화진흥위원회에서는 디지털 온라인 시장의 정형화 되지 않은 계약방식과 수익분배 구조, 투명하지 못한 정산 시스템에 대한 문제점을 느끼고 영화 온라인 상영관 통합 전산망 구축을 위한 정책적 논의가 오가는 중이다(영화진흥위원회, 2016). 따라서 오차가 심각하게 큰 7개의 영화는 이러한 문제가 있다고 판단하여 제외한 후 RMSE와 MAPE를 측정하면, 의사결정나무의 RMSE는 149257.2, MAPE는 38.3%, 다중회귀분석의 RMSE는 57452.7, MAPE는 12.6%로 의사결정나무의 RMSE가 조금 올라가긴 했지만, 나머지 지표의 성능은 대폭 개선되었음을 알 수 있다.

Table 4.4 test data에서 예측 이상치를 제외한 VOD 수요 예측 결과

영화	Target	의사결정나무	다중회귀분석
히말라야	506909	735993.2	608355.3
굿 다이노	215327	255823.5	210793.4
쿵푸팬더3	364904	531679.8	429476.5
몬스터 호텔 2	174749	255823.5	227304.1
극장판 꼬마버스 타요의 에이스 구출작전	101936	162375.2	94821.0
데드풀	452917	531679.8	422405.8
검사외전	874477	531679.8	758276.6
레버넌트: 죽음에서 돌아온 자	275263	255823.5	290152.1
제 5침공	182986	162375.2	150816.0
바다 탐험대 옥토넛 시즌4: 빙하탐험선S	84264	162375.2	89146.0
RMSE		149257.2	57452.7
MAPE		38.3%	12.6%

4) 영화 VOD 흥행 요인 도출

Table 4.5는 의사결정나무와 다중회귀분석을 통해 두 모델에서 중요 변수를 정리한 것이다. 이 중 공통으로 등장한 스크린 수, 홀드백, 4주차 관객 수, 리뷰 개수, 청소년 관람불가 등급의 5개 변수를 영화 VOD 흥행에 영향을 끼친 핵심 요인으로 선정하였다. VOD 흥행 요인으로 선정된 5가지 변수를 VOD 콘텐츠 관점에서 해석해보았다.

Table 4.5 영화 VOD 주요 흥행 요인

구분	유의한 변수
의사결정나무	1주차 관객 수, 스크린 수, 4주차 관객 수, 3주차 관객 수, 홀드백, 리뷰 수, 19세 이상 관람등급
다중회귀분석	한국영화, 2주차 관객 수, 4주차 관객 수, 19세 이상 관람등급, 스크린 수, 홀드백, 리뷰 수, 로맨스, 배급사, 감독 인지도, 15세 이상 관람가, 청소년 관람불가
공통 흥행 요인	스크린 수, 홀드백, 4주차 관객 수, 리뷰 개수, 청소년 관람불가 등급

(1) 최대 스크린 수

스크린 수는 극장 플랫폼에서의 배급사의 배급력을 나타내며, 크면 클수록 관객들에게 노출이 많이 된다. 이러한 노출은 VOD 시장에서의 홍보 효과가 되어 영화에 대한 기대감과 친숙함을 높일 수 있으므로, 스크린 수는 VOD 시장에서 친밀감이라는 성격을 지닐 수 있다.

(2) 홀드백

홀드백은 콘텐츠의 유통 과정에서 다음 창구로 중심이 이동하는 기간을 의미한다. 본 연구에서는 홀드백을 극장에서 개봉하고 VOD 서비스로 출시되는 시점까지의 기간으로 산정하였으며, 그 이유는 극장에서 해당 영화는 내리는 시점이 명확하지 않기 때문이다. 따라서 VOD 콘텐츠 관점에서 홀드백이 짧으면 짧을수록 신선한 콘텐츠로 받아들일 수 있기 때문에 홀드백은 VOD시장에서 신선함을 나타낸다.

(3) 4주차 관객 수

4주차 관객 수는 극장 흥행을 나타낸다. 흥행에 실패한 영화들이 2주 정도 극장 상영 후 종료하며, 초기 마케팅 성공으로 개봉 1주차 관람객 수는 많을 수 있으나 4주차 관객 수가 많다는 것은 흥행과 더불어 영화의 완성도나 재미를 보장할 수 있다고 볼 수 있다. 따라서 4주차 관객 수는 VOD 관점에서 만족 보장성으로 설명 될 수 있다.

(4) 리뷰 수

리뷰 수는 영화 콘텐츠의 만족도가 높은 경우, 광고나 입소문으로 부풀려진 영화에 실망한 경우, 사회적인 문제를 담고 있는 경우 등 긍정적인 경우와 부정적인 경우 모두 많을 수 있다. 이는 해당 영화의 화제성으로 간주하고 VOD에서도 화제성으로 해석하기로 한다.

(5) 청소년 관람불가 등급

마지막으로, 관람등급 중에서 특히 청소년 관람 불가 등급은 잠재 관객의 규모의 감소로 인해 극장 흥행에는 저해되는 요소일 수 있으나, 청소년 관람 불가 VOD는 선정적이거나 폭력적인 장면이 나오더라도 개인적인 공간에서 관람 할 수 있다는 장점 때문에 오히려 극장에 비해 수요가 높을 수 있다. 이러한 특성을 반영하여 청소년 관람 불가 등급 요인은 자극성이라는 성격을 부여하도록 한다.

Table 4.6 영화 VOD 주요 흥행 요인의 특징

흥행 요인	VOD 관점
최대 스크린 수	친밀감
홀드백	신선함
4주차 관객 수	만족 보장성
리뷰 수	화제성
청소년 관람불가	자극성

2. 영화 VOD 흥행 특성 분석 결과

1) 흥행 / 비흥행 VOD 그룹 분류

흥행하는 VOD와 흥행하지 못한 VOD의 특성을 분석하기 위해 VOD 이용 건수가 30만 건이 넘는 65개의 VOD를 흥행 VOD 그룹으로, 그렇지 않은 148개의 VOD를 비흥행 VOD 그룹으로 할당한다. 각 그룹별로 흥행 요인으로 선정한 5가지 변수에 대해 통계적으로 차이가 있는지 살펴본 결과가 Table 4.7에 나타나 있다. 스크린 수와 4주차 관객 수, 리뷰 수는 흥행 VOD 그룹과 비흥행 VOD 그룹 간의 큰 차이를 보이며 통계적으로 유의함을 보였다. 즉, 해당 영화의 VOD가 친숙할수록, 극장 흥행에 의해 재미가 보장될수록, 화제의 영화일수록 VOD가 흥행함을 알 수 있었다. 하지만, 홀드백과 청소년 관람불가 등급은 유의하지 않는 것으로 나타나, 이는 군집 분석을 통해 더 세부적으로 살펴보기로 한다.

Table 4.7 VOD 흥행 그룹별 흥행 요인에 대한 통계 분석

흥행 요인	흥행 VOD 그룹 평균	비흥행 VOD 그룹 평균	통계량	p-값
스크린 수***	904.97	537.88	9.6341	1.228e-15
홀드백	1.83	1.76	0.3654	0.72
4주차 관객 수***	459577.11	62270.28	6.2641	2.886e-08
리뷰 수***	13793.338	3574.405	5.7173	2.773e-07
청소년 관람불가	-	-	-	-

2) VOD 흥행 그룹별 군집 분석 결과

영화 VOD 흥행의 주요 요인으로 도출된 스크린 수, 홀드백, 4주차 관객 수, 리뷰 수, 청소년 관람불가 변수를 통해 흥행 VOD 그룹과 비흥행 VOD 그룹별로 k-medoids 군집분석을 수행하였다. 이때, 군집의 수는 3장에서 언급한 실루엣 계수를 근거로 결정한다. Fig 4.2과 Fig 4.3를 보면, 군집의 수가 2일 때 실루엣 계수가 가장 높지만, 그룹별로 다양한 특성을 살펴보기 위하여 3개 이상의 군집 중 가장 실루엣 계수가 높은 k를 선정한다. 따라서 흥행 VOD 그룹에서 Successful VOD인 S1, S2, S3, S4의 군집 4개, 비흥행 VOD 그룹에서 Failed VOD인 F1, F2, F3의 군집 3개를 도출하였고, 각 군집별 주요 흥행 요인에 대한 요약 결과를 Table 4.8, Table 4.9에 정리하였다.

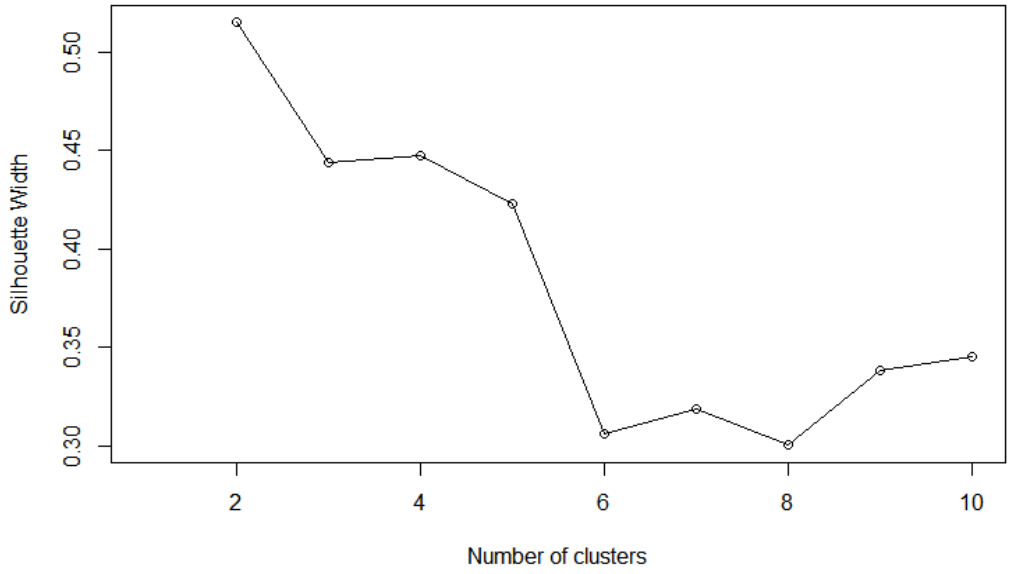

Fig 4.2 홍행 VOD 그룹의 군집 수에 따른 실루엣 계수

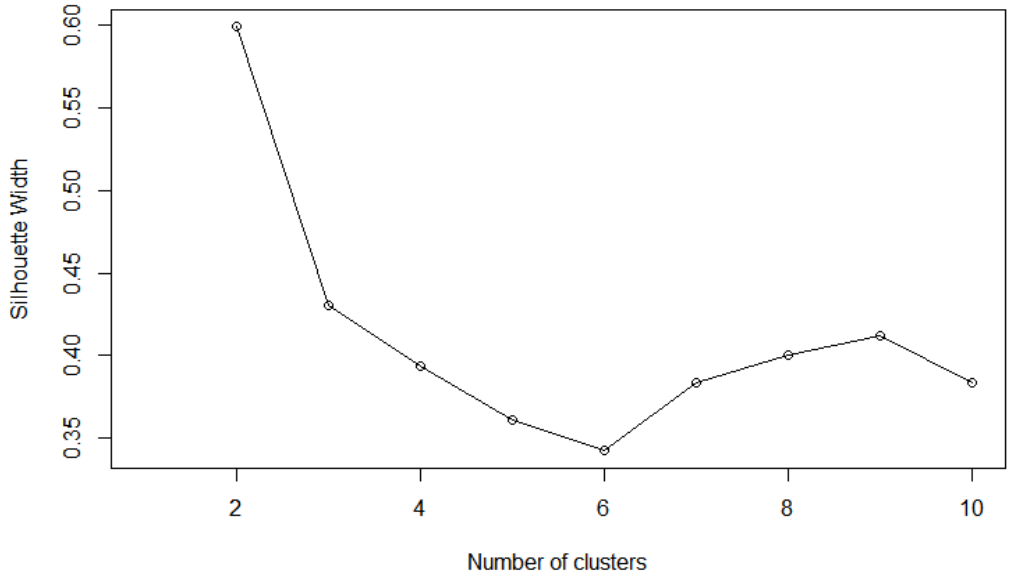

Fig 4.3 비홍행 VOD 그룹의 군집 수에 따른 실루엣 계수

Table 4.8 흥행 VOD 그룹 군집분석 결과

군집	S1	S2	S3	S4
군집 특징	신선한 VOD	천만 관객 영화	스타 파워	19금 범죄액션
영화 수	27	8	18	12
대표 영화	스파이	베테랑	관상	신의 한 수
스크린 수	705.3	1184.5	1151.4	798.3
홀드백	1.41	2.75	2.33	1.42
4주차 관객 수	156456.9	1566987.5	549683.2	268164.9
리뷰 수	5984	41579.6	16637.3	8574.2
청소년 관람불가	0	0	0	1
VOD 이용건수	388985.4	792592.8	544765	545414.3

Table 4.9 비흥행 VOD 그룹 군집분석 결과

군집	F1	F2	F3
군집 특징	진부한 해외영화	소규모 마니아	무관심 19금
영화 수	49	71	28
대표 영화	인투 더 스톰	위험한상견례2	무뢰한
스크린 수	701.1	452.7	468.1
홀드백	2.82	1.24	1.21
4주차 관객 수	149285.0	20919.2	14849.2
리뷰 수	6594.6	1943.5	2424.7
청소년 관람불가	0	0	1
VOD 이용건수	203317.6	152299.5	171720.8

흥행 VOD 그룹에 속한 영화는 65개, 비흥행 VOD 그룹에 속한 영화는 148개이며, 비흥행 VOD 그룹의 F2(소규모 마니아 영화) 군집이 71개로 가장 많은 영화를 포함하고 흥행 VOD 그룹의 S2(천만 관객 영화)가 7개로 가장 적은 영화를 포함하고 있다. 영화 VOD 흥행 요인 중심으로 군집들을 살펴보면, 최대 스크린 수는 평균과 분포에 따라 크게 세 개의 그룹으로 나눠질 수 있는데, 천개 이상의 스크린을 확보한 S2와 S3, 대략 700~800개 사이의 스크린을 동원한 S1, S4, F1, 500개 이하 스크린 수를 가진 F2와 F3이다. 두 번째 그룹에 속한 F1은 최대 스크린 수의 평균이 701.1로 흥행 VOD 그룹의 S1과 크게 차이나지 않지만 VOD 성적이 저조한 편이다. 두 군집간의 차이는 다른 요인에 비해 홀드백의 차이에서 극명하게 나타나는데, S1의 홀드백은 1.41개월, F1의 홀드백은 2.82개월로 Table 4.10에서 홀드백의 분포를 살펴 볼 수 있다.

Table 4.10 군집별 홀드백 분포

군집	1개월 미만	1개월	2개월	3개월	4개월	5개월 이상
S1	15%	48%	19%	19%		
S2	13%	13%	25%	13%	13%	25%
S3	6%	28%	28%	28%	6%	6%
S4		58%	42%			
F1		14%	24%	39%	16%	6%
F2	23%	49%	14%	10%	4%	
F3	18%	57%	14%	7%	4%	

이러한 특징에 때라 S1을 신선한 VOD, F1을 진부한 해외 영화로 군집의 특징을 서술하였다. 홀드백의 특징에 대해 좀 더 살펴보면, 각 연도별 홀드백의 평균은 2013년 2.25개월, 2014년 1.88개월, 2015년 1.16개월로 최근 들어 홀드백이 짧아지는 경향을 알 수 있다. 이는 관객들의 영화 소비 패턴 변화에 따른 영화 관계자들의 디지털 온라인 시장에 대한 의사결정 변화를 시사하고 있다. 군집 관점에서 살펴보면, S2와 S3과 같이 극장에서 흥행 대박이 난 영화들은 극장의 상영기간 자체가 길어지고, 극장의 수익을 극대화하기 위해 홀드백이 긴 편이며, 앞서 언급한 F1은 주로 판권 계약이 국내 영화와는 다른 형태인 해외 영화들이 주를 이루고 있어 홀드백이 길어진 것으로 판단된다. VOD 관점에서 만족 보장성을 나타내는 4주차 관객 수는 천만 관객 영화들로 구성된 S2가 독보적으로 큰 수치를 나타내고 있고, F2와 F3은 적은 관객 수를 보이고 있다. 실제 극장에서 흥행이 저조한 영화들은 2주 만에 상영을 중단하기도 하므로, 4주차 관객 수는 극장에서의 흥행뿐만이 아니라 상영기간을 유추할 수 있는 요인이 될 수 있다. 다음으로 화제성을 나타내는 리뷰 수에 대해 살펴보면, S2는 다른 요인들과 비슷하게 독보적인 형태를 보이고 있으며, 청소년 관람 불가 영화들이 화제성이 다소 높은 것으로 보인다. 마지막으로 자극성을 나타내는 청소년 관람불가 영화들은 각각 S4, F3에 밀집해 있으며, 다른 요인은 비슷하지만 관람 등급에 따라 VOD의 이용건수가 다름을 알 수 있다. 이를 면밀하게 살펴보기 위하여, S1과 S2, F2와 F3의 VOD 이용 건수를 t-test를 통해 살펴본 결과가 Table 4.11에 나타나 있다.

Table 4.11 관람등급에 따른 VOD 이용 건수 차이에 대한 통계 분석

비교 군집	청소년 관람불가가 아닌 영화	청소년관람불가 영화	통계량	p-값
S1 - S4	388985.4	545414.3	-3.36982	0.00229**
F2 - F3	152299.5	171720.8	-1.65897	0.052038

흥행 VOD 그룹 내의 S1과 S4의 차이가 유의하게 드러났으며, 화제성이 높으면서 자극적인 청소년 관람 불가 영화들이 VOD로써 인기가 많고, 액션 장르의 영화가 다수를 차지하고 있어, S4의 특징을 19금 범죄 액션 영화로 서술하였다.

결 론

1. 연구의 결론

본 연구에서는 최근 영화 산업에서 급격히 성장 중인 디지털 온라인 시장에서 VOD의 수요를 예측하고, VOD 흥행요인을 도출하여 흥행하는 VOD와 그렇지 않은 VOD의 특징을 분석하였다. 데이터 마이닝 기법 중 의사결정나무 기법과 다중회귀분석을 통해 수요 예측 모델을 구축하고, 두 모델에서 공통적으로 유의한 변수로 등장한 최대 스크린 수, 4주차 관객 수, 리뷰 수, 홀드백, 청소년 관람 불가 등급의 5가지 변수를 VOD 주요 흥행요인으로 도출하였다. 5가지 VOD 흥행요인을 디지털 온라인 시장 관점에서 살펴보면, 최대 스크린 수는 '친밀감', 4주차 관객 수는 '만족 보장성', 리뷰 수는 '화제성', 홀드백은 '신선함', 청소년 관람 불가 등급은 '자극성'으로 해석할 수 있다. 도출된 흥행요인에 대하여 세부적인 분석을 위해 흥행 VOD 그룹과 비흥행 VOD 그룹에 따라 군집분석을 수행하였다. 흥행 VOD 그룹에서는 신선한 VOD, 천만 관객 영화, 스타 파워, 19금 범죄액션의 4개 군집이 도출되었고, 비흥행 VOD 그룹에서는 진부한 해외영화, 소규모 마니아, 무관심 19금 영화의 3개 군집이 도출되었다. 따라서 영화 VOD 흥행분석에 대한 결과로, 영화 VOD는 극장 흥행의 결과에 영향을 받으며, 비슷한 개봉 규모와 흥행 규모의 영화라면 홀드백이 짧은 영화일수록 VOD 이용 건수가 더 높았다. 또한, 친밀하고 화제성이 큰 영화에 대해 VOD 이용 건수가 높았으며, 극장에 비해 청소년 관람 불가 영화의 흥행이 돋보였다. 본 연구로 인해 영화 관계자들이 디지털 온라인 시장에서 해당 영화 VOD의 대략적인 수요를 파악하여 수익을 미리 예측해 볼 수 있으며, VOD 배포 시기와 마케팅 의사결정을 내리는데 도움이 될 것으로 기대된다.

2. 연구의 한계점 및 추후 연구

본 연구의 한계점과 추후 연구에 대한 방향은 다음과 같다. 첫째로, 본 연구에서 사용한 데이터는 디지털 온라인 시장에서 IPTV 및 디지털케이블 TV VOD 이용 건수만이 반영된 데이터이기 때문에 추후 인터넷 VOD와 모바일 VOD에 대한 분석을 진행하여 디지털 온라인 시장 전반에 대한 분석을 할 필요성이 있다. 두 번째로, VOD 서비스를 제공하는 업체의 프로모션이나 해당

시점의 정확한 가격 정보 등을 알 수 없어 수요 예측의 정확도가 높지 않다는 한계점이 있다. VOD 가격이나 할인 혜택 등이 VOD 이용 건수에 영향을 미칠 수 있기 때문에 추후 연구에는 이러한 마케팅 정보를 반영하여 분석을 진행하도록 한다. 마지막으로, 본 연구에서는 상업 영화만을 분석하였지만, 다양성 영화가 스크린 확보가 힘든 극장 대신에 디지털 온라인 시장을 주요 시장으로 여기고 있기 때문에 다양성 영화에 대한 분석도 진행하여 상업 영화와 어떤 차이를 보이고 있는지 분석해볼 필요성이 존재한다.

참고문헌

강지훈, 박찬희, 도형록, 김성범 (2014). 데이터마이닝 기법을 활용한 영화 흥행 실적 예측 기법. 대한 산업공학회 공동학술대회 논문집, 142-154.

권선주 (2014). 영화 흥행성과의 분석과 예측: 뉴스와 웹사이트 데이터 이용. 한국문화경제학회, 17(1), 35-56.

고정민, 안성아, 이백헌 (2010). 영화관람방식에 따른 인터넷VOD의 소비의향 분석. 한국문화산업학회 학술대회, 120-142.

김미현 (2006). 영화의 창구화와 유통전략에 대한 연구-DVD의 홀드백과 차별화 정책을 중심으로. 영화연구, 28, 79-103.

김진욱 (2010). 디지털 환경-미디어 융복합-내에서 영화산업의 패러다임 변화-영화산업 발전을 위한 대안적 모형 제시. 한국콘텐츠학회논문지, 10(4), 133-140.

민병현 (2013). 뉴미디어 시대 한국영화 디지털 온라인 시장 활성화 방안 연구. 중앙대학교 첨단영상대학원 석사 논문

박선규, 최성진 (2015). 디지털케이블TV에서 영화의 선행창구 성과, 장르, 홀드백 기간이 영화 VOD 구매에 미치는 영향. 방송공학회논문지, 20(6), 950-962.

박찬호, 권혜정, 곽현 (2014). IPTV 이용자의 VOD 구매에 영향을 미치는 요인에 대한 실증적 연구. 한국정보기술학회논문지, 12(11), 153-163.

영화진흥위원회(2012). 영화산업 부가시장 통계는 왜 믿을 수 없는가?. KOFIC Issue Paper 2012, 2.

영화진흥위원회(2015). 세계 영화시장 현황 및 전망. KOFIC Issue Paper 2015, 13.

영화진흥위원회(2016). 2015년 한국 영화산업 결산.

영화진흥위원회(2016). 영화 온라인 시장 구조분석. KOFIC Report 2016, 4.

오윤학, 김한, 윤재섭, 이종석 (2014). 데이터마이닝을 활용한 한국프로야구 승패예측모형 수립에 관한 연구. 대한산업공학회지, 40(4), 8-17.

이상우, 김창완 (2009). IPTV-VOD 서비스 선택의 결정요인 분석. 한국언론정보학보, 9-36.

임예원 (2004). I디지털 미디어의 등장과 영화산업의 시장 확대 연구. 중앙대학교 예술대학원 석사 논문

전범수 (2007). 영화 소비 창구의 구조와 특성. 한국언론정보학보, 221-248.

전치혁 (2012). 『데이터마이닝 기법과 응용』, 서울: 한나래출판사

진성철, 박원준 (2013). 인터넷 VOD 서비스 이용자의 영화 콘텐츠 이용에 관한 연구. 한국전자통신학회 논문지, 8(2), 255-261.

한국콘텐츠진흥원 (2015). 디지털 시대, 영화 VOD 시장 동향. 통계로 보는 콘텐츠 산업, 15(4).

허민회, 강필성, 조성준 (2013). Predicting Box-office with Opinion mining reviews. 대한산업공학회 춘계공동학술대회 논문집, 487-500.

황준석, 이준기, 이재경 (2009). 인터넷 VOD 이용의도에 영향을 미치는 요인에 관한 연구. 한국컴퓨터정보학회논문지, 14(2), 221-229.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees.* Wadsworth, CA, USA

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis..* Wiley, New York, USA

Kruchikanth R., Merin J., Supreme M., C.-C., Kathy J. and Federico G.(2013). Prediction of Movies Box Office Performance Using Social Media, 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

Mestyan M., Yasseri T. and Kertesz J. (2013). Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLoS ONE, 8(8)

Pricewaterhousecoopers(PWC) (2015). PwC Global Entertainment and Media Outlook 2015-2019

Ramesh S. and Dursun D. (2006). Predicting box-office success of motion pictures with neural networks. Expert Systems with Application, 30, 243-254.

Abstract

Analyzing the Characteristics of Movie VOD Success using Decision Tree and Multiple Linear Regression

Moon, Jin Hee
(Supervisor Geum, Youngjung)
Dept. of Data Science
Graduate School
Seoul National University of Science and Technology

Due to the drastic change of the movie distribution platform, the digital market has become very important in the recent movie industry. This is especially true for the movie VOD market where customer demand is significant growing due to its convenience. However, previous studies on digital movie market has focused on qualitative approach to investigating the characteristics of the movie VOD market and customer behavior for VOD consumption. Also, there is a lack of quantitative and analytic research to characterize and analyze the success of movie VOD market. Addressing the limitation of previous studies, this study try to predict the demands for movie VOD and to analyze the success factor for the movie VOD in a quantitative and analytic way. This study consists of the following steps. First, data are collected 230 movie VODs between 2013 and 2016, from Korean Film Council(www.kofic.or.kr). Second, two data mining techniques – decision tree and multiple linear regression – are used to forecast the demands of movie VOD. Third, five success factors are identified based on the results for the decision tree and multiple linear regression; the maximum number of screens, holdbacks, number of audiences in fourth week, number of reviews and NC-17 rated. Fourth, characteristics for successful VODs and unsuccessful VODs are analyzed based on the cluster analysis. In successful VOD group, four clusters are identified: the latest VOD, megahit movie, star power, and NC-17 rated criminal action. In unsuccessful VOD group, three clusters are

identified: outdated foreign movie, low budget movie, and indifferent NC-17 rated movie. The results of this study are expected to be effectively applied in decision making of movie VOD market for revitalizing the movie industry.

감사의 글

석사 학위논문 마지막 페이지를 쓰기 위해 7년이라는 시간이 걸렸습니다. 서울과학기술대학교 산업정보시스템공학과에 입학하여, 데이터 사이언스 석사과정 졸업에 이르기까지 우연 같았던 매 순간의 선택과 좋은 인연들이 지금의 저로 성장하게 한 것 같습니다. 아직은 많이 부족하고 미숙하지만, 마음속에 품고 있는 조그마한 희망과 소신을 가지고 사회에 나가려고 합니다. 학교에 너무 오래 얽매어 있었던 것은 아닌가 하는 두려움도 있지만 어엿한 한 사회인이 되기 위해 노력할 것입니다.

석사과정 내내 진심어린 충고와 날카로운 지성으로 지도해주신 금영정 교수님께 진심으로 감사의 말씀을 드립니다. 개인 공부와 연구에만 몰두할 수 있도록 좋은 연구 환경을 조성해 주시고, 출산휴가 와중에도 논문 지도에 힘써 주셔서 석사과정을 잘 마칠 수 있었습니다. 앞으로도 교수님을 롤 모델로 삼아 매사에 열정적이며, 공정하면서도 온화한 성품을 갖춰나갈 수 있도록 노력하겠습니다. 그리고 학부시절 지도 교수님이셨던 안재경 교수님께도 정말 감사드립니다. 석사 진학하고 나서도 항상 신경써주시고, 아버지처럼 믿고 격려해주셔서 많은 힘이 되었습니다. 기대에 부응할 수 있도록 언제나 열심히 살겠습니다. 두 분의 지도 교수님 외에도 정말 훌륭하고 좋은 학과 교수님들을 만나 행복한 학교생활을 했습니다. 즐겁고 편안한 학교생활을 만들어주신 학과 교수님들과 송미자 조교님께 진심으로 감사하다는 말씀드리고 싶습니다.

마지막으로 가장 고생하신 부모님, 부모님의 원조가 없었다면 석사 진학을 꿈도 꾸지 못했을 것입니다. 딸이 서울에서 편하게 공부할 수 있게 언제나 신경써주시고 챙겨주셔서 무사히 졸업을 하게 되었습니다. 이젠 부모님 품을 벗어나 한 사회인으로서 제 몫을 해나가는 사람이 되겠습니다. 이 외에도 정말 가족처럼 서로 챙기고 같이 공부했던 데이터사이언스 연구실 선후배 및 동기들과 안랩 식구들 모두 감사합니다. 힘들 때 곁에서 아낌없는 조언과 격려로 학교생활을 무사히 마칠 수 있었습니다. 옆에서 항상 잘 할 수 있다고 격려해 주고 고민 상담을 들어줬던 고향 친구들과 대학 선배와 친구들, 모두 믿어주고 지지해줘서 정말 고맙습니다.

힘들었지만 즐거웠던 석사 생활을 무사히 마치고, 마지막 석사 학위 논문을 완성하게 된 것은 제 주변에 있는 모든 분들의 도움과 지지덕분입니다. 날카로운 비판, 따뜻한 응원, 친절한 도움들 모두 마음에 새기며 감사하며 살겠습니다. 제가 받은 것들을 다 베풀며 살 수 있도록 노력하겠습니다. 끝까지 곁에서 지켜봐 주십시오.