Prediction of Korean Box Office using Data Crawling

양지원(통계학과), 이예진(통계학과), 한하랑(통계학과)

December 13, 2020

1 Introduction

한 달에도 수십 개의 영화가 나오는 시점에서 영화의 흥행 여부와 손익분기점을 넘을 것인가에 대한 우려가 있다. 이는 영화 누적관객수와 관련이 있는 것이다. 따라서 영화 관객수와 관련된 여러 내외부 요인들을 이용하여 여러 학습기를 통해 누적관객수의 정확한 예측을 꾀한다.

영화 관객수에 영향을 줄 것이라고 예상되는 변수들은 굉장히 많다. 하지만 실제로 이 많은 변수들이 영화 관객수에 영향을 주지는 않을 것이다. 때로는 영화 관객수와 다른 변수들과의 교란 변수로서 작용하여 영화 관객수에 영향을 주는 것처럼 보일 수도 있다. 이처럼 실제 생각과 분석 결과를 비교함으로써 우리의 사고를 확장시킬 수 있다.

더 나아가 많은 전문가들은 스크린수가 관객수에 영향을 많이 미친다고 판단한다. 한 사례로 겨울왕국2가 있다. 영화진흥위원회에 따르면 1일 기준 겨울왕국2의 누적관객수는 1천만 명 이상을 돌파했다. 이 영화의 흥행에는 스크린 독과점이 큰 몫을 했다는 의견이 있다. 전국 2천351개 스크린에 걸려 1만 3천467회 상영하여 점유율 62.5%에 달했기 때문이다. 이 때문에 한 시민단체는 겨울왕국2가 국내 상영관을 독점해 독점금지법을 위반했다며 월트디즈니컴퍼니코리아를 고발했다. 이처럼 최근 대형 배급사의 영화 혹은 특정 영화들의 스크린 독점 현상으로 인해서 스크린 상한제 도입에 대한 의견이 거론된다. 2020년 3월 문체부는 스크린 상한제를 제안했다. 이는 6개관 이상을 보유한 극장을 대상으로 관객이 집중되는 시간대(오후 1시~ 11시)에 같은 영화의 상영 횟수가 50%를 넘지 않도록 하는 것이다. 대한민국의 스크린 독점 문제는 2019년 4월에 개봉한 "어벤져스-엔드 게임"로 잘 보여진다. 이는 개봉했을 당시 상영 회차의 80.2%를 장악했다. 나머지 개봉 영화 44편이 19.8%를 나눠야 하는 상황이었다. 다양성 관점에서 영화 창작자들은 물론이고 관객들의 권리에 위협적이라고 전문가들은 말한다. 따라서 분석을 통해 실제 스크린수가 누적관객수에 영향을 미치는지 알아보고 스크린 상한제 도입이 필요한지 알아보았다.

2 Background

스크린 상한제는 관객이 몰리는 주요 시간대에 특정 영화의 상영관 스크린수를 제한하는 제도다.

현재 스크린 상한제와 관련된 정책을 시행하고 있는 나라로는 프랑스, 미국, 일본 등이 있다. 프랑스의 스크린 상한제란 한 영화가 스크린 지배를 25%를 넘지 못하게 하고 대형 영화는 15%로 묶는 것이다. 초기에 적은 수로 스크린을 릴리즈를 하고 인기가 좋아 늘려도 이 제한을 넘을 수 없다. 예를 들어 협약에 따르면 스크린 9 ∼ 14개를 갖춘 극장은 최대 3개, 스크린 15 ∼ 27개를 구비한 극장은 최대 4개 스크린에서만 같은 영화를 상영할 수 있다. 일본의 경우 상영관이 분배되는 경우가 많아 스크린 독점 논란은 없다. 일본 영화에 관련 법률이 있는 건 아니지만, 영화관 자체적으로 히트작이라 할지라도 최대 $\frac{1}{4}$의 상영 회차로 스크린 상한제가 걸려 있다. 미국의 경우 어벤져스: 엔드게임 정도 되는 영화라 할지라도 미국 영화 시장의 영화관들을 100% 차지할 수는 없다. 프라임 타임(17 ∼ 21시) 기준 최대 50%까지 상영한다. 보통은 전국 모든 영화관에서 개봉하되 영화관 내에서는 상영 회차가 최대 30 ∼ 40% 선을 지키는 편이다. 이는 상영 기간에 따라 극장과 배급사의 수익 배분율이 달라지는 '슬라이딩 시스템'이 존재해서 최대한 많은 영화를 오랫동안 장기 상영하는 게 극장 입장에서 수익적으로 더 이득이 되기 때문이다.

3 Data

영화 누적관객수 관련 데이터를 얻기 위해 영화관입장권통합전상망(KOBIS), 영화진흥위원회, 네이버로부터 데이터를 확보했다. 영화관입장권통합전상망의 기간별 박스오피스에서는 최대 조회 기간 2년에 대한 누적된 영화 정보를 제공한다. 여기서 누적된 영화 정보란 만약 2017년 1월 1일부터 2018년 12월 31일까지 총 2년을 조회했을 때, 2017년 이후의 영화를 포함하는 것이 아니라 2017년 이전 영화도 모두 포함하되, 해당 기간에서의 박스오피스 현황을 보여주는 것이라고 볼 수 있다. 기간별 박스오피스에서는 엑셀자료로 영화명, 개봉일, 매출액, 매출액점유율, 누적매출액, 관객수, 누적관객수, 스크린수, 상영횟수, 대표국적, 국적, 제작사, 배급사, 등급, 장르, 감독, 배우 정보를 제공하며, 해당 보고서에서는 앞서 언급한 정보들 중에서 영화명, 개봉일, 누적관객수, 국적, 배급사, 등급 정보를 활용하였다. 영화진흥위원회에서는 Open API를 이용하여 영화관입장권통합전상망의 자료와 비교하여 보다 정확한 정보를 얻고자 하는데 활용되었다. 그 밖에도 영화 자체의 내부적 요소 외에도 사람들의 반응과 관심과 같은 외부적 흥행 요소에 대한 정보를 얻기 위해 네이버 뉴스 및 네이버 영화 사이트를 이용하였다. 해당 사이트에서는 영화별 평점, 뉴스 검색수를 HTML 문서 스크래핑을 통해 얻었으며, 이 외에도 장르와 상영시간을 스크래핑 하였다. 기간별 박스오피스에서 얻은 원자료의 자료 수는 총 6,296개로, 여기에는 독립 · 예술 영화가 제외되었다. 이 중에서 분석의 편의성과 효과성을 위해, 분석의 타겟 변수인 누적관객수 변수가 1만 이상일 때를 첫 번째 기준으로 하였고, 이후 멀티플렉스 상영관이 상용화되었다고 생각되는 시점을 임의로 2004년으로 지정하여 2004년 이후의 영화만을 포함하였다. 따라서 위의 2가지 기준에 부합하는 영화 목록에는 총 630개의 영화가 포함되었으며 앞으로 630개 영화에 대한 분석을 진행하고자 한다. 예측변수는 개봉일, 국적, 배급사, 등급, 스크린수, 평점, 뉴스, 장르, 상영시간, 경쟁영화수이다. 파생 변수는 배급사에서 파생된 대형 배급사 지시변수, 개봉일에서 파생된 계절 지시변수, 개봉월, 개봉요일, 주말 지시변수가 있고, 장르에서 파생된 장르별 지시변수 그리고 국적에서 파생된 한국 지시변수가 있다. 다음은 위의 내용을 통해 얻은 데이터세트에 대한 자세한 설명이다.

3.1 데이터 가공 및 변수 획득 과정

개봉일, 국적, 배급사 변수는 영화관입장권통합전산망에 나와 있는 자료를 그대로 활용하였다. 개봉일은 2004년부터 2020년 6월까지 포함하며 대부분의 영화가 2010년 이후의 영화임을 확인할 수 있다. 개봉일 변수로부터 계절 지시변수, 개봉월, 개봉요일, 주말 지시변수가 파생되었는데 각각을 설명하면 다음과 같다. 계절 지시변수는 봄, 여름, 가을, 겨울 계절에 대해 해당 계절에 개봉되었을 경우 1, 아닐 경우 0을 넣는 one-hot encoding을 하였다. 대체로 4개의 계절이 비슷한 영화 개봉수를 가지지만 그중 계절이 188개로 가장 많았으며 이후 여름, 가을, 봄 순이었다. 개봉월의 경우, 개봉일에서 개봉월만 추출한 변수이며 3월이 전체 630개 중 31개(약 5%)를 포함하여 가장 적은 개봉 수를 보였다. 반대로 2월이 67개(약 10.6%)로 가장 많은 개봉 수를 가진 달이었다. 개봉요일은 수요일이 342개(약 54%), 목요일이 242개(약 38%)로 대부분의 영화가 수요일과 목요일에 개봉하였으며, 의외로 금요일, 토요일, 일요일의 비율이 전체의 약 3%로 매우 낮았다. 조사결과, 개봉 첫 주 관객동원 성적이 스크린수 유지에 영향을 미치기 때문에 첫 주 관객수를 늘리기 위한 목적과 주말 관객을 이끌기 위한 입소문을 퍼뜨리는 전략으로 수요일이나 목요일 영화 개봉을 선호하는 것으로 나타났다. 또한 주5일제의 영향 또한 포함된다고 한다. 주말 지시변수는 토요일과 일요일 개봉일 때 1, 아닐 경우 0을 배정하였는데 앞의 결과들을 참고하였을 때, 주말 개봉은 현저히 낮을 것이라는 점을 예상할 수 있었다.

파생변수	설명
개봉 요일	0:월, 1:화, 2:수, 3:목, 4:금, 5:토, 6:일

국적 변수에는 한국, 미국, 영국, 일본, 프랑스, 2개국 합작 등이 포함되었으며, 국적변수에서 한국이 포함될 경우를 1, 한국이 포함되지 않을 경우를 0으로 두어 한국 지시변수를 생성하였다. 배급사 변수는 128개의 값이 있었는데 여기에는 단일 배급사 혹은 2개 이상의 배급사가 포함된 경우를 포함한다. 배급사 변수를 이용하여 흔히 메이저 배급사라고 불리는 CJ엔터테인먼트, 롯데엔터테인먼트, 쇼박스, 넥스트엔터테인먼트월드, 월트디즈니컴퍼니, 소니픽쳐스엔터테인먼트 그리고 워너브러더스를 배급사로 하고 있다면 대형배급사 지시변수를 1, 아닐 경우 0으로 하였다. 총 630개 영화 중에서 403개(약 64%)의 영화가 대형배급사 지시변수가 1이라는 점을 확인할 수 있는데, 이러한 점에서 대부분의 영화가 대형배급사에 의해 대중들에게 공개된다는 사실과 영화 시장에 있어서 대형배급사의 역할이 크다는 점을 확인할 수 있다. 등급 변수는 다음과 같이 정리되었다.

변수	설명
등급	0:전체관람가, 1:12세이상, 2:15세이상, 3:18세이상(청소년관람불가)

총 4개의 등급 중 15세이상관람가가 227개(약 36%)로 가장 많았고 이후 12세이상관람가, 전체관람가, 18세이상(청소년관람불가) 순이었다. 스크린수의 경우 영화관입장권통합전산망의 기간별 박스오피스와 스크린점유율 자료 등에서 얻을 수 있지만, 보다 정확한 자료를 얻기 위해 영화진흥위원회 Open API를 통해 확보하였다. 여기서 말하는 스크린수란 영화 개봉 후 일주일간 최대 스크린수를 말한다. API를 통해 630개 영화 중 503개 영화에 대한 스크린수 데이터를 얻을 수 있었다. 남은 127개의 영화에 대해서는 먼저 영화관입장권통합전산망의 자료를 이용하여 대체하였으며, 영화관입장권통합전산망과 Open API에서 모두 NULL값을 가지는 영화에 대해서는 knn-imputation을 이용하여 값을 모두 채워 넣었다. 스크린수 변수를 자세히 살펴본 결과, 제2사분위수가 537.5, 제3사분위수가 930.25, 최대값이 2835라는 점에서 대부분의 영화가 개봉 첫 주 1000개 이하의 스크린수를 가지며, 몇몇 영화에 대해서만 큰 값을 가진다는 것을 확인할 수 있었다. 평점 변수의 경우, 포털사이트와 영화관사이트 중에서 가장 이용인원이 많고, 보편적으로 사용된다고 예상되는 곳을 선정하였으며, 선정결과 네이버 포털을 이용하기로 하였다. 따라서 네이버 검색창에 “영화 ◯◯” (◯◯에는 영화제목)으로 입력하였을 때 나타나는 네이버 영화 정보창에서 평점 정보를 이용하였다. 평점은 10

점 만점이며, 히스토그램과 사분위수를 확인해 보았을 때, 중간값이 8.525이고 오른쪽으로 치우쳐진 그래프가 나온다는 점에서 평점이 대체로 높다는 것을 확인할 수 있었다. 같은 방식으로 가장 보편적으로 사용된다고 생각되는 네이버 포털을 이용하였다. 네이버 검색창에 "영화 ○○"을 검색한 뒤, 영화 개봉일 한 달 전후의 총 기사 수를 스크래핑 하였다. 스크래핑 후 뉴스 변수에 대해 자세히 살펴본 결과, 영화 "#살아있다" 의 뉴스 기사 수가 326,204,560으로 두 번째로 큰 값을 가지는 영화 "기생충"의 기사 수 657,676과 비교했을 때 상당히 큰 값을 가진다는 것을 확인할 수 있었다. 해당 영화의 기사 수가 다른 영화보다 비이상적으로 큰 값이 나온다는 점에서 영화 "#살아있다"의 뉴스 기사 수를 NULL값으로 처리한 후, knn-imputation을 통해 값을 대체하였다. 뉴스 기사 수는 중간값이 34,380이었지만 boxplot, 사분위수를 통해 불균형한 분포를 가지고 있음을 확인할 수 있었다.

참고사항으로 자료 수집 단계에서는 총 기사 수가 뉴스 탭 상단에 기재되어 있었으나, 네이버 뉴스의 업데이트로 인해 현재는 최대 4,000건만 출력되어 총 기사 수는 확인 불가능하다. 평점과 기사 수를 얻는 과정에서 10개 미만의 영화에 대해서는 결측치가 발견되어 평점의 경우에는 평균을, 뉴스 기사 수에 대해서는 knn-imputation을 통해 값을 대입하여 처리하였다. 장르 변수는 영화관입장권통합전산망에서도 자료를 구할 수 있었지만, 대표적인 장르를 파악하기 위해 네이버영화 정보를 활용하였다. 영화 평점을 구한 방식과 동일한 방식을 이용하였으며 총 19개의 장르가 있었다. 19개의 장르 중 액션이 123개로 가장 많은 비중을 차지하였고 그 다음이 멜로/로맨스 였다. 장르 변수에 대해 이후 해당 장르를 포함하면 1, 아닐 경우 0으로 하는 one-hot encoding을 하여 장르별 지시변수를 생성하여 분석을 시작하였다. 상영시간 역시 영화 평점과 장르변수를 구한 방식과 동일하다. 상영시간에 대한 히스토그램을 그린 결과, 정규분포와 유사한 형태를 가진다는 점을 확인할 수 있었으며, 평균이 약 110분, 중간값이 111분이었다. 가장 상영시간이 긴 영화는 181분으로 2019년에 개봉한 "어벤져스:엔드게임" 과 2012년에 개봉한 "어벤져스" 였다. 마지막으로 경쟁영화수는 해당 영화가 개봉한 달에 함께 개봉한 영화의 수로 지정하였고, 영화관입장권통합전산망에서 자료를 얻었다. 경쟁영화수의 범위는 [29,951]로 2019년 11월이 951개의 영화가 개봉하여 조사기간 중 가장 많은 영화가 개봉한 달이었다. 본격적인 분석에서는 예측변수와 파생변수에서 중복된 변수는 배제하였다. 배제된 변수들은 개봉일, 국적, 배급사, 장르이다. 이 변수들은 파생변수가 그 내용을 대체하거나 더 세부적으로 다루기 때문에 배제하였다. 따라서 분석에 사용되는 변수들은 스크린수, 등급, 평점, 뉴스, 경쟁영화수, 상영시간, 대형배급사 지시변수, 계절별 지시변수, 개봉 요일, 개봉월, 주말 지시변수, 장르별 지시변수 그리고 한국 지시변수이다. 다음은 파생변수를 제외한 특정 영화에 대한 수집 자료를 보여주는 샘플 데이터이다.

Table 1. Data after processing

항목명	샘플데이터	항목설명
영화명	극한직업	영화 제목
개봉일	2019-04-24	영화 개봉 날짜
누적관객수	16266338	개봉부터 마감까지 누적된 관객 수
국적	한국	영화 제작한 나라
배급사	씨제이이앤엠(주)	영화 배급한 회사
등급	15세이상관람가	영화의 상영 등급으로 4개의 등급이 있다. (전체 관람가, 12세 이상 관람가, 15세 이상 관람가, 청소년 관람 불가)
평점	9.2	네이버 영화 기준 평점 (10점 만점)
뉴스	358109	개봉 한 달 전후 뉴스 검색 수
경쟁영화수	511	동시 상영된 영화 수
상영시간	111	영화 상영 시간(분 단위)
장르	코미디	영화의 대표 장르
전국스크린수	1978	개봉 첫 주 상영된 최대 스크린 수

4 Methods and hypotheses

본 연구에서 사용된 방법론들은 크게 분류와 회귀로 나누어질 수 있다. 분류는 세부적으로 300만 누적관객수, 500만 누적관객수를 기준으로 흥행성공과 흥행실패로 분류하는 이진 분류와, 누적관객수의 class를 3개, 4개로 나누어 분류하는 다중분류로 분석을 진행하였다.

Table 2. Classification

class=2		
흥행기준(y)	흥행성공(>y)	흥행실패(<y)
300만	1	0
500만	1	0

class=4		class=3	
0(흥행실패)	100만 미만	0(흥행실패)	100만 미만
1(중박)	100만 이상 300만 미만	1(중박)	100만 이상 500만 미만
2(대박)	300만 이상 500만 미만	2(대박)	500만 이상
3(초대박)	500만 이상		

Logistic Regression, Gradient Boost Classifier, XGBoost Classifier 그리고 LightGBM Classifier 이 이진분류를 하는데 사용되었고, Support vector machine(SVM), Logistic Regression, Gradient boosted model(GBM), XGBoost Classifier, Naive Bayes, LightGBM Classifier 그리고 Random Forest Classifier 이 다중분류를 하는데 사용되었다. 회귀에서는 Linear Regression, Linear Regression with lasso penalty, KNN Regression 그리고 앙상블을 기반으로 하는 Random Forest Regression, XGBoost Regression, LightGBM Regression 이 사용되었다.

각각의 방법론들은 훈련데이터를 이용하여 모델링을 하였는데, 훈련데이터는 전체 630개의 데이터를 훈련데이터, 테스트데이터로 8:2 비율로 무작위로 선별한 다음에 구해진 데이터로 테스트데이터와는 독립적인 특징을 가지고 있다. 회귀를 이용한 분석과 다중 분류의 경우, 자료의 개체수(N)가 작은 것을 감안하여 K-fold Cross Validation을 이용하여 검증오류를 측정하며, 이때 K는 일반적으로 많이 사용되는 5를 사용한다. 5-fold CV를 이용하여 검증오류를 계산한 뒤 가장 검증오류가 작은, 즉, 가장 성능이 좋은 모형을 하나 선택할 것이다. 이 때 분류의 검증오류로는 F1-score와 정밀도(precision)이 사용되었으며, 회귀의 검증오류로 RMSE와 MAE가 사용되었다.

추가적으로 LightGBM과 XGBoost 모델의 경우 OOF(Out-of-Fold)예측을 사용하여 성능을 측정 하였다. 즉, 5-fold를 통해서 나온 5세트의 예측을 합쳐서 전체 훈련 데이터에 대한 예측을 만들어 예측을 평가하는 것이다. K-fold Cross Validation에 경우 각 폴드에서 나온 에러값(RMSE)들은 서로 비슷해야 한다. 만약, 값들이 서로 비슷하지 않다면, 이때의 모델로는 테스트세트에서 안정적인 예측치를 얻기 어려울 것이다. 따라서 더 좋은 성능 향상을 위하여 사용되었다.

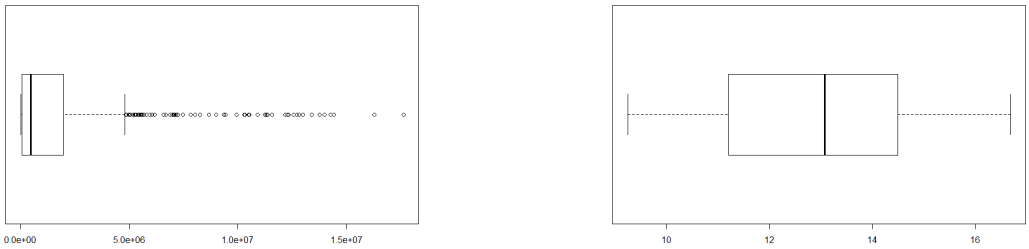
분류의 검증오류로 F1-score와 정밀도를 선택한 이유는 다음과 같다. 우선 분석에 사용되는 데이터세트이 누적관객수의 흥행 기준에 따라 불균형한 class를 가지고 있다는 것을 사전에 확인할 수 있었다. F1-score는 이러한 자료의 불균형성을 가진 자료의 성능지표로 주로 사용되어 왔다. 따라서 최종 모형을 선택할 때, F1-score를 첫 번째 기준으로 삼는다. 이후에는 정밀도를 두 번째 기준으로 삼는다. 정확도, 재현율, 민감도와 같은 성능지표들 중에서 정밀도를 두 번째 기준으로 삼은 이유는 정밀도는 흥행할 것이라고 예측된 영화들 중에서 실제로 흥행을 한 영화의 비율을 뜻하기 것으로, 분류문제에서 흥행 성공 영화를 흥행 실패 영화로 예측하는 것을 피하기기 위해 두 번째 기준으로 삼게 되었다.

회귀의 검증오류로 RMSE와 MAE를 선택한 이유는 RMSE는 대략적인 성능을 보기 위함으로 선택하였으며, 실제로는 이상점에 로버스트한 MAE를 기준으로 최종 모형을 선택할 것이다. 모든 방법론을 사용하기 이전에 먼저 훈련데이터에서 록로그 변환한 누적관객수(log(Y))의 평균을 이용하여 검증오류를 계산해보는 작업을 하여 이후에 사용될 회귀 방법론들의 기준을 만들어 놓는다. 회귀를 이용한 방법론들은 누적관객수의 평균을 이용하여 구한 검증오류보다는 적어도 작은 검증오류를 가져야 한다. 만약 이러한 나이브한 방식으로 구한 검증오류보다 큰 검증오류를 가진다면 해당 방법론은 과감하게 최종 모형에서 배제시킨다.

본 연구의 목적은 스크린 상한제의 실효성과 반응변수 예측력 최대화이다. 첫 번째 목적을 달성하기 위해서 전체 데이터세트을 선형회귀로 적합한 후, 스크린 변수의 유의성을 알아보는 방법과 우도비 검정을 통한 변수의 유의성 탐색, 그리고 마지막으로 변수 중요도(Feature Importance, FIMP)를 이용하여 스크린 변수가 누적관객수를 예측하는데 상대적으로 중요한 변수인지 확인해보는 방법을 이용하여 다양한 방식으로 스크린 상한제의 효용성을 논해볼 것이다. 두 번째 목적을 달성하기 위해 검증오류가 가장 작은 최종 방법론을 선택하고, 테스트 데이터로 남겨둔 126개의 영화 중 6개 영화를 선택하여 이들에 대해서는 실제 누적관객수를 얼마만큼 예측하였는지를 확인해볼 것이다. 따라서 테스트 데이터세트의 오류는 120개 영화에 대한 오류를 말한다는 것을 미리 언급한다.

위에서 언급한 이진 분류에서의 class imbalance을 해결하기 위해 F1-score를 성능지표를 삼는 것과 다른 방식으로 전처리를 진행해보았다. 먼저 누적관객수 300만을 기준으로 흥행 여부를 정했을 때, 총 504개의 훈련 데이터세트에서 흥행에 성공한 영화는 88개, 흥행에 실패한 영화는 416개로 나누어졌다. 흥행 성공 영화가 전체 훈련 데이터세트에 약 1/5에 해당하는 것으로, class imbalance가 있는 것이 확인할 수 있었고, 이를 해결하기 위해 SMOTE Sampling 기법을 이용하여 두 class의 비율을 맞추었다. 그 결과로, 흥행 성공한 영화와 실패한 영화 각각 416개로 전체 훈련 데이터세트이 832개로 샘플링되었다. Class imbalance가 예측 결과에 많은 영향을 주는지 확인하기 위해 각각의 이진 분류 모델에 대하여 SMOTE Sampling 이전과 이후의 훈련 데이터세트을 사용하여 분류 분석을 시행하였다. 위와 같이 누적관객수 500만을 기준으로 흥행 여부를 나누었을 때, 총 504개의 훈련 데이터세트에서 흥행에 성공한 영화는 52개, 흥행에 실패한 영화는 452개로 흥행 성공 영화가 전체 훈련 데이터세트에 차지하는 비율이 누적관객수 300만을 기준으로 하였을 때보다 더 작다는것을 확인할 수 있었다. 따라서 같은 방식으로 SMOTE Sampling기법을 이용하여 두 class의 비율을 맞추어 주었다. 그 결과로는 흥행 성공한 영화와 실패한 영화 각각 452개로 전체 훈련 데이터세트이 904개로 샘플링되었고, class imbalance가 예측 결과에 많은 영향을 주는지 확인하기 위해 각각의 모델에 대하여 SMOTE Sampling이전 이후 훈련 데이터 세트을 사용하여 분류 분석을 시행하였다. 구체적인 성능지표는 5장 Experimental results에서 확인할 수 있다.

5 Experimental results



(a) Boxplot before log transformation

(b) Boxplot after log transformation

Figure 1: Boxplot of Y

왼쪽 Boxplot을 통해, 반응 변수인 누적관객수의 분포가 매우 한쪽으로 쏠려있음을 발견할 수 있다. 이러한 점은 누적관객수의 범위가 [10315, 14411782] 으로 매우 광범위하고, 자료 수집 시 누적관객수가 1만 이상인 영화를 대상으로 자료들을 수집하였기 때문이다. 따라서 누적 관객수에대한 변수변환이 필요하다고 결론지을 수 있다. 오른쪽 Boxplot은 반응변수 누적관객수에 대해 로그 변환을 한 것이다. 분포가 안정되었음을 확인할 수 있다. 본격적인 모형 적합 및 비교에 앞서 2가지 방식으로 벤치마킹을 하고자 한다. 첫 번째는 R 프로그램에 소개되어 있는 AutoML(Automated Machine Learning)를 이용하여 데이터세트가 주어졌을 때 어떤 모형이 가장 좋은 성능을 보여주는지 간단하게 확인하는 것이다. 두 번째는 타켓변수인 누적관객수의 평균을 이용하여 성능 비교에 사용되는 RMSE와 MAE를 알아보는 것이다. 첫 번째로 AutoML을 통해 박스오피스 데이터세트에 대해 누적관객수 예측을 가장 잘하는 모형이 무엇인지에 대해 확인해보면, 아래의 표와 같은 결과를 얻을 수 있다.

Table 2. result of AutoML

model	RMSE	MAE
StackedEnsemble_BestOfFamily	0.7933445	0.6056057
StackedEnsemble_AllModels	0.7955141	0.6062511
GBM_3	0.8014172	0.6115859
GBM_grid_1_model_3	0.8074838	0.6291050
GBM_2	0.8092251	0.6177355
GBM_grid_1_model_4	0.8114113	0.6240635
GBM_4	0.8212646	0.6277964
GBM_grid_1_model_5	0.8660153	0.6679476
GBM_grid_1_model_2	0.8876976	0.6728762
GBM_1	0.9028782	0.6945839
GLM_1	0.9424287	0.7375050

실제로는 Mean Residual Deviance, RMSE, MSE, MAE 그리고 RMSLE 다섯 개의 평가 지표와 20개의 모델이 나오지만 보고서에서 주로 다루는 RMSE와 MAE만 요약하여 위와 같이 나타내었다. RMSE과 MAE를 기준으로 StackedEnsemble_BestOfFamily 모델이 0.7933445 와 0.6056057로 가장 작은 값을 가지는 것을 확인할 수 있다. 표를 확인해보면 전반적으로 앙상블을 기반으로 하는 모형들이 예측 성능이 좋다는 것을 볼 수 있다. 중간에 GLM이 있지만 StackedEnsemble에 비하면 낮은 성능을 보여준다. 여기서 StackedEnsemble이란, 기본 학습기에 대해 최적의 조합을 찾기 위해서 2단계로 meta learner를 훈련시키는 것을 포함하는 알고리즘을 말한다. StackedEnsemble은 특히 기본 학습기가 개별적으로 강하고, 상관관계가 없는 오류들을 가질 때 잘 수행되는 특징을 가진다. Best of Family와 All Model의 차이점은 말 그래도 모든 모델을 포함하는지, 아니면 각각의 알고리즘에서 가장 성능이 좋은 모형만을 포함하는지에 따라 달라진다. 위의 표를 통해 앙상블 모형이 최종 모형으로 선택될 것이라는 예상을 할 수 있다.

두 번째로 누적관객수의 평균으로 예측한 모형의 RMSE와 MAE를 나타낸 것이다.

Table 3. RMSE & MAE from Naïve Benchmark

반응변수	RMSE	MAE
log(누적관객수)	2.0157	1.743451

분류 문제에서 성능지표의 기준을 0.5로 삼는 것처럼, 여기서 구한 RMSE와 MAE를 기준으로 이들 보다 높은 값을 가지는 모형은 최종 모형에서 배제할 것이다. 즉 적어도 이후 최종 모형으로 선택된 모형들은 적어도 누적관객수의 평균으로 예측한 것보다 더 우수한 성능을 낸다.

5.1 상관관계 분석

Table 4. Correlation between factors of Box Office (p-value *: <0.05, **: <0.01)

	log(누적관객수)	평점	뉴스	경쟁영화수	상영시간	스크린수
log(누적관객수)	1					
평점	0.3909 **	1				
뉴스	0.6448**	0.1255 **	1			
경쟁영화수	-0.1342	-0.0953 **	0.1222 **	1		
상영시간	0.4886 **	0.1674 **	0.3845 **	-0.1764 **	1	
스크린수	0.7600 **	0.2073 **	0.7149 **	0.2399 **	0.3846 **	1

log(누적관객수), 뉴스, 경쟁영화수, 상영시간, 스크린수 등 연속형 변수들의 관계를 파악하기 위해 상관관계 분석을 실시하였다. 분석 결과, Table 4에서 log(누적관객수)와 경쟁영화수를 제외한 모든 변수들은 유의수준 0.01에서 통계적으로 유의미한 양의 상관관계가 있음이 관찰되었다. 특히 log(누적관객수)와 스크린수, 뉴스와 스크린수는 0.76, 0.71로 강한 양의 상관관계를 보여주고 있다. 각 변수들의 산점도는 부록에 첨부하였다.

5.2 회귀 분석

Table 5. Result of Regression

	RMSE	MAE
Linear regression	0.9685365	0.7683119
Linear regression with lasso ($\lambda = 0.03125$)	0.9534337	0.7638288
knn (k=7)	1.2336853	0.9755847
RandomForest	2.5828698	2.0944038
XGBoost	0.7759194	0.6078935
LightGBM	0.7761568	0.6108657
Mean_Ensemble	0.7590318	0.5942206

lasso는 λ=0.03125 일 때, RMSE와 MAE가 모두 최솟값을 가지며, RMSE값이 0.9534이고 MAE값이 0.7838이다. knn 회귀 역시 k=7 일 때, RMSE와 MAE가 모두 최솟값을 가지며, RMSE값이 1.2337이고 MAE값이 0.9756이다. 이상점에 민감한 RMSE와 이상점에 덜 민감한 MAE 값을 구해보았을 때, 두 개의 값에 대해서 lasso 벌점항을 가진 선형회귀가 0.9534, 0.7638로 전반적으로 좋은 성능을 보이고 이후는 기본적인 선형회귀, knn regression 순이었다. RMSE와 MAE에 대해 세 모형 모두 나이브한 방식으로 구한 값인 2.0157, 1.7435보다 낮은 값을 가진다.

RandomForest Regression의 경우 앙상블 배깅 테크닉의 모델이며 훈련시간에 다수의 Decision Tree를 구성하여 평행하게 작동되고 평균 예측을 출력하여 작동한다. 즉, 여러개의 예측을 결합한 것이다. 훈련 결과 예측값에 대한 RMSE 값은2.5829 MAE 값은 2.0944이다.

XGBoost Regression의 경우 여러 개의 Decision Tree들을 조합해서 사용하는 앙상블 알고리즘이며, 그래디언트 부스팅(가중치를 경사하강법으로 진행) 알고리즘을 분산 환경에서도 실행할 수 있도록 구현한 것이다. 또한 과적합 방지가 가능한 규제가 포함되어 있으며 Early Stopping을 제공한다는 특징이 있다. 각 Fold마다 Early Stopping을 지정하여 성능이 좋아지지 않으면 다음 Fold로 넘어가도록 하였다. 훈련 결과 예측값에 대한 RMSE 값은 0.7759이며 MAE 값은 0.6079이다.

LightGBM Regression은 XGBoost Regression의 느린 학습시간과 많은 하이퍼 파라미터의 단점을 보완하기 위해 나온 모델이다. 이는 대용량 데이터 처리가 가능하고, 다른 모델들 보다 더 적은 메모리를 사용하며 빠르다는 장점이 있다. 하지만 너무 적은 수의 데이터를 사용하면 과적합의 문제가 발생할 수 있다. 균형트리분할(level-wise)를 사용하는 기존 GBM과 달리 리프중심 트리분할(leaf-wise)을 사용하였다. 이로 인해, 손실을 더 줄일 수 있다는 장점이 있다. 훈련 결과 예측값에 대한 RMSE 값은 0.7761이며 MAE 값은 0.6108이다.

추가적으로 파이썬 사이킷런의 Voting 라이브러리를 사용하지 않고 직접 각각의 예측 값들에 대해 산술평균을 이용하여 Ensemble 해주었다.(prediction_Ensemble=(predictions_ XGBoost+predictions_LightGBM)/2) 이때의 경우 성능이 비슷한 모델의 예측값을 이용해야 더 좋은 성능을 얻을 수 있다. 따라서 우리는 성능이 비슷한 LightGBM Regression과 XGBoost Regression의 예측값을 선택했으며 이 두 예측값의 평균을 가지고 다시 성능을 측정해본 결과 RMSE 값은 0.7590이며 MAE 값은 0.5942이다.

최종적으로 alpha=0.08, gamma=0.06, eta=0.04 로 XGBoost와 LightGBM의 초모수를 지정하였을때의 Mean_Ensemble을 회귀모형중에서 가장 우수한 성능의 모델로 선택하였다. 테스트 데이터를 이용하여 최종 모형을 검증한 결과 RMSE 값은 0.7898이며 MAE 값은 0.5811값을 확인할 수 있다. 이는 AutoML을 통해 구한 가장 성능이 좋았던 StackedEnsemble_BestOfFamily 모델의 RMSE 값 0.7933과 MAE 값 0.6056에 비하여 더 좋은 성능임을 확인할 수 있다.

5.3 분류 분석

영화 누적관객수에 대하여 300만 명과 500만 명을 기준으로 하여 흥행 성공은 "0"으로, 흥행 실패는 "1"로 분류하는 두 가지의 이진 분류를 하였다.

먼저 누적관객수 300만 명을 기준으로 한 경우이다.

Table 6. Result of Classification, 누적관객수 300만 기준

	Class imbalance		SMOTE Sampling	
	Precision	F1-score	Precision	F1-score
Logistic Regression	0.83	0.84	0.53	0.52
Gradient Boost	0.93	0.95	0.92	0.94
XGBoost	0.94	0.96	0.97	0.96
LightGBM	0.90	0.94	0.97	0.95

SMOTE Sampling 이전의 경우, Logistic Regression의 정밀도는 0.83, F1-score는 0.84의 결과가 나왔으며, Gradient Boost Classifier 모델은 정밀도는 0.93, F1-score는 0.95, XGBoost Classifier의 경우 정밀도는 0.94, F1-score는 0.96이다. 마지막으로 LightGBM Classifier는 정밀도는 0.90, F1-score는 0.94이다.

SMOTE Sampling 이후의 경우, Logistic Regression의 정밀도는 0.53, F1-score는 0.52의 결과가 나왔으며, Gradient Boost Classifier은 정밀도는 0.92, F1-score는 0.94, XGBoost Classifier의 경우 정밀도는 0.97, F1-score는 0.96이다. 마지막으로 LightGBM Classifier는 정밀도는 0.97, F1-score는 0.95이다.

SMOTE Sampling 데이터세트를 이용한 XGBoost Classifier이 정밀도는 0.97, F1-score는 0.96으로 가장 좋은 성능을 보여 누적 관객수 300만을 흥행 기준으로 하는 모델에 대하여 최종 모델로 선정하였다. 이에 테스트 데이터세트를 이용하여 최종 모형을 검증한 결과 정밀도는 0.93, F1-score는 0.94임을 확인할 수 있다.

다음으로 흥행 기준을 누적관객수 500만으로 한 경우이다.

Table 7. Result of Classification, 누적관객수 500만 기준

	Class imbalance		SMOTE Sampling	
	Precision	F1-score	Precision	F1-score
Logistic Regression	0.90	0.92	0.55	0.53
Gradient Boost	0.94	0.96	0.96	0.96
XGBoost	0.95	0.96	0.99	0.97
LightGBM	0.94	0.96	0.98	0.97

SMOTE Sampling 이전의 경우, Logistic Regression의 정밀도는 0.90, F1-score는 0.92의 결과가 나왔으며, Gradient Boost Classifier의 정밀도는 0.94, F1-score는 0.96, XGBoost Classifier의 경우 정밀도는 0.95, F1-score는 0.96이다. 마지막으로 LightGBM Classifier는 정밀도는 0.94, F1-score는 0.96이다.

SMOTE Sampling 이후의 경우, Logistic regression의 정밀도는 0.55, F1-score는 0.53의 결과가 나왔으며, Gradient Boost Classifier의 정밀도는 0.96, F1-score는 0.96, XGBoost Classifier의 경우 정밀도는 0.99, F1-score는 0.97이다. 마지막으로 LightGBM Classifier의 정밀도는 0.98, F1-score는 0.97이다.

Sampling 이전 이후의 두 경우를 비교하였을때 정밀도와 F1-score 성능 모두 어느정도의 차이로 SMOTE Sampling 한 경우 더 향상됨을 확인 할 수 있다. class imbalance를 고려한 성능척도를 이용하였음에도 성능 향상이 있는것으로 보아 class imbalance가 예측하는 것에 있어서 많은 영향을 끼친다고 할 수 있다.

SMOTE Sampling 데이터세트를 이용한 XGBClassifier이 정밀도는 0.99, F1-score는 0.97로 가장 좋은 성능을 보여 누적관객수 500만을 흥행 기준으로 하는 모델에 대하여 이진 분류에서의 최종 모델로 선정하였다. 이에 테스트 데이터세트를 이용하여 최종 모형을 검증한 결과 정밀도는 0.96, F1-score는 0.96임을 확인할 수 있다. 또한 누적관객수 300만 명과 500만 명을 기준으로 흥행 성공/실패 척도를 나뉘어 모델링 하였을 때, 성능적으로는 500만 명을 기준으로 두었을 경우 더 좋았음을 확인할 수 있다.

훈련데이터세트에서의 성능에 비하여 테스트 데이터세트에 대한 성능이 낮은 것을 보아 전체 데이터 자체의 개수가 적어 과적합(Over fitting) 현상이 있음을 확인하였다.

5.4 다중 분류 분석

영화의 누적관객수를 기준으로 0: 흥행 실패(100만 미만), 1: 중박(100만 이상 300만 미만), 2: 대박(300만 이상 500만 미만), 3: 초대박(500만 이상)으로 세분화하여 다중 분류를 하였다. 전과 동일하게 8:2 비율로 훈련 데이터세트과 테스트 데이터세트을 나누어 진행했다.

위 기준으로 504개의 훈련 데이터세트를 분류했을 때, 316, 97, 35, 56 순으로 분류되었다. 약 60% 이상의 영화가 흥행 실패로 분류되어 이진 분류 처럼 class imbalance가 확인되었다. 따라서 성능척도를 F1-score와 정밀도로 기준을 정하여 모델의 성능을 비교하였다. 대부분의 영화는 최종 관객수 100만 이하로 상영을 마무리한다. 100만 이하를 예측하는 것보다는 100만 이상을 예측하는 것에 더 초점을 두었으며 최종 모델을 선정할 때, 흥행 실패(100만 미만)을 제외한 class에 더 초점을 두었다. RandomForest에서 tree의 개수는 100으로 지정했으며, 다른 초모수들은 default 값으로 고정하였다. RandomForest에 기반한 다른 모델링들의 tree의 개수는 200으로 지정했으며, 다른 초모수들은 default값으로 고정하였다.

Table 8. Result of Multi-Class Classification (Class=4)

	F1-score					Precision				
	0	1	2	3	mean	0	1	2	3	mean
SVM	0.87	0.46	0.00	0.60	0.48	0.81	0.45	0.00	0.78	0.51
Logistic Regression	0.88	0.49	0.12	0.58	0.52	0.89	0.48	0.15	0.55	0.52
Naïve Bayes	0.84	0.38	0.15	0.55	0.48	0.87	0.35	0.15	0.56	0.48
Gradient Boost	0.87	0.46	0.12	0.51	0.49	0.81	0.49	0.23	0.63	0.54
RandomForest	0.87	0.42	0.08	0.62	0.50	0.80	0.45	0.40	0.73	0.59
LightGBM	0.86	0.53	0.31	0.43	0.53	0.87	0.50	0.30	0.50	0.54
XGBoost	0.89	0.56	0.19	0.61	0.56	0.87	0.56	0.16	0.58	0.54

Support vector machine(SVM)은 분류 규칙을 찾아내는 기법 중 하나로, n 차원 공간에서 최적의 분할선을 찾아내어 분류를 해주는 알고리즘이다. 훈련 결과, 각 class 별 F1- score은 0.87, 0.46, 0, 0.60이며, 정밀도는 0.81, 0.45, 0, 0.78이다. 이진 분류와는 다르게 성능이 낮은 이유는 중박과 대박에서 분류를 잘 못하는 것을 확인했다. 특히 대박 영화 중, 단 한 영화도 대박으로 예측하지 못했으며 대박 영화의 약 80%를 흥행 실패와 중박으로 예측하였다.

Multi-Class Logistic Regression 모델은 다중 분류에 이용되는 Logistic Regression이다. 훈련 결과, F1-score은 0.88, 0.49, 0.12, 0.58이며, 정밀도는 0.89, 0.48, 0.15, 0.78로 나왔다. 여전히 성능이 좋다고 말할 수 없는 수치이지만 초대박 영화를 제외한 나머지 class에서 SVM보다는 성능이 나아짐을 확인했다. Naïve Bayes 분류 모델은 통계 기반 분류 기계학습법이다. 훈련 결과, class 별 F1-score은 0.86, 0.48, 0.16, 0.55이며, 정밀도는 0.87, 0.35, 0.15, 0.56으로 나왔다.

GBM Classifier 모델의 F1-score은 0.87, 0.46, 0.12, 0.51이며, 정밀도는 0.81, 0.49, 0.23, 0.63이다. RandomForest 모델의 F1-score은 0.87, 0.42, 0.08, 0.62이며, 정밀도는 0.80, 0.45, 0.40, 0.73이다. RandomForest의 대박으로 예측할 F1-score은 다른 class에 비해 0.1 이하로 매우 낮게 나왔다. 그 이유는 재현율에 있다. 실제 대박으로 분류된 영화 중에 대박으로 예측된 영화는 38개의 영화 중에 하나이다. LightGBM의 F1-score은 0.86, 0.54, 0.31, 0.43이며, 정밀도는 0.87, 0.5, 0.3, 0.5이다. XGBoost의 F1-score은 0.89, 0.56, 0.19, 0.61이며, 정밀도는 0.87, 0.55, 0.23, 0.62로 나온다.

4-class 다중 분류 결과, 중박을 기준으로 성능이 제일 좋은 모델은 XGBoost이며, 대박은 LightGBM이 제일 좋았다. 대박을 기준으로 LightGBM의 F1-score은 다른 모델링에 비해 두 배 가까이 높다. 대박을 기준으로 RandomForest가 제일 성능이 좋았다. XGBoost의 F1-score의 평균이 제일 좋았으며, 정밀도 기준으로는 RandomForest가 제일 좋았다. 따라서 최종 모형은 XGBoost로 선정되었다. 하지만 이진 분류에 비해 성능이 낮은 이유는 Confusion matrix를 통해 찾을 수 있었다. SVM과 비슷하게 흥행 실패와 초대박에 비해 중박과 대박을 분류하는 과정에서 성능이 낮은 것을 확인했다.

성능을 향상하기 위해 흥행 실패(100만 미만), 중박(100만 이상 500만 미만), 대박(500만 이상)으로 세 범주로 다시 분류하여 다중 분류를 실시하였다. 위 기준으로 504개의 훈련 데이터 세트을 분류했을 때, 흥행 실패, 중박, 대박은 각각 396, 168, 66으로 분류되었다. 세 범주로 나누었을 때 역시 절반 이상의 영화가 흥행 실패로 분류되어 이전과 같이 class imbalance가 확인되었다. 따라서 세 범주의 성능 척도 또한 F1-score과 정밀도를 기준으로 모델의 성능을 비교하였다.

Table 9. Result of Multi-Class Classification (Class=3)

	F1-score				Precision			
	0	1	2	mean	0	1	2	mean
SVM	0.89	0.64	0.49	0.67	0.85	0.65	0.79	0.76
Logistic Regression	0.88	0.61	0.58	0.69	0.89	0.60	0.63	0.71
Naïve Bayes	0.83	0.54	0.54	0.64	0.87	0.50	0.57	0.65
Gradient Boost	0.88	0.59	0.45	0.64	0.83	0.62	0.68	0.72
RandomForest	0.87	0.61	0.54	0.67	0.83	0.63	0.81	0.77
LightGBM	0.88	0.60	0.45	0.64	0.90	0.56	0.49	0.65
XGBoost	0.89	0.64	0.57	0.70	0.87	0.64	0.63	0.71

세 범주로 나누어 성능 척도를 비교해 본 결과, 모든 모델링에서 흥행 실패를 제외한 모든 class에서 성능이 좋아졌음을 확인했다. SVM의 F1-score은 0.89, 0.64, 0.49이며, 정밀도는 0.85, 0.65, 0.79이다. 4-class 분류에서는 단 하나의 영화도 대박 영화(300만 이상 500만 미만)로 분류로 예측하지 않은 문제는 3-class에서는 일어나지 않았다. 중박을 기준으로 성능이 제일 좋은 모델은 SVM이며, 대박을 기준으로 성능이 제일 모델은 Logistic Regression이다. F1-score의 평균으로 최종 모델을 선정한다면 XGBoost의 성능이 제일 좋아 최종 모델로 선정하였다. 4-class 분류와 비교하면, 평균 F1-score와 정밀도가 최대 0.2 이상 높아졌으며 평균적으로 0.1 이상 높아진 것을 확인할 수 있었다. 분류 분석 결과, 이진 분류의 성능이 더 좋게 나오므로 최종 모델로는 SMOTE Sampling한 데이터 세트를 이용하는 이진 분류의 XGBoost로 결정하였다.

5.5 테스트세트 예시

Table 10. Predition rate

	Mean_Ensemble regression			XGBoost for binary classification		
	$log\hat{Y}$	$logY$	$log\hat{Y}/logY$	$\hat{Y}$	Y	예측 성공 여부
#살아있다	14.3069	13.9936	1.02	1	1	성공
광해, 왕이 된 남자	15.2790	16.3271	0.94	1	0	실패
배테랑	15.8727	16.4119	0.97	0	0	성공
국가부도의날	14.7711	15.1387	0.98	1	1	성공
데드풀2	15.5702	15.1465	1.03	0	1	실패
악질경찰	13.5332	12.4770	1.08	1	1	성공

다음은 회귀와 분류분석에서 최종모형으로 뽑힌 2개의 모형에 대해, 테스트세트에서 분리시킨 6개 영화에 대한 예측정도를 살펴보았다. 대상 영화는 "#살아있다", "광해, 왕이 된 남자", "배테랑", "국가부도의날", "데드풀2" 그리고 "악질경찰" 이다. 회귀에서는 해당 영화에 대해 실제 누적관객수에 log를 취하고, 회귀에서 선택된 최종모형인 XGBoost와 LightGBM의 평균으로 구한 앙상블(Mean_Ensemble)에서 얻은 예측값과 비교해보았다.

6개 영화 중 3개의 영화에 대해서는 실제 누적관객수에 log를 취한 값보다 작게 나타났으며, 나머지 3개에 대해서는 더 큰 값을 보여주었다. $log\hat{Y}/logY$을 보니 6개 영화가 대체적으로 1근방의 값을 가진다는 것을 확인할 수 있고, 따라서 실제 누적관객수에 log를 취한 값과 예측값이 비슷하다는 것을 알 수 있다. 분류에 대해서도 예측 성공 여부를 확인해보았다. 분류에서는 누적관객수 500만을 기준으로, 누적관객수가 500만보다 크다면 0, 500만보다 작다면 1로 분류하였을 때, XGBoost가 가장 좋은 모형으로 나타났다. 따라서 XGBoost 분류기를 이용하여 분류 결과를 구했을 때, 6개 중 4개의 영화에 대해 예측 성공을 하였다.

5.6 스크린 상한제 관련 분석

Table 11. Result of Regression(p-value *: <0.01, **: <0.001, ***: <0)

	coef	t	p-value			coef	t	p-value	
(Intercept)	9.542e+00	9.199	<2e-16	***	스릴러	7.098e-02	0.075	0.94052	
평점	1.507e-01	5.553	4.23e-08	***	공포	4.999e-01	0.525	0.59986	
뉴스	1.054e-06	1.612	0.10758		판타지	3.949e-01	0.411	0.68155	
경쟁영화수	-3.020e-03	-12.186	<2e-16	***	미스터리	2.247e-01	0.234	0.81538	
상영시간	6.029e-03	2.315	0.02095	*	모험	5.322e-01	0.551	0.58209	
국적	5.385e-01	5.061	5.55e-07	***	SF	4.682e-01	0.479	0.63182	
배급사	8.366e-01	8.363	4.32e-16	***	가족	-3.984e-01	-0.393	0.69465	
등급	1.139e-01	1.969	0.04936	*	뮤지컬	3.883e-01	0.385	0.70043	
개봉월	9.255e-02	6.246	8.00e-10	***	공연	-1.437e+00	-1.389	0.16540	
요일	-1.053e-01	-1.593	0.11165		전쟁	-3.053e-02	-0.028	0.97740	
주말	1.397e+00	2.732	0.00648	**	다큐멘터리	-2.680e-01	-0.248	0.80436	
애니메이션	2.782e-01	0.294	0.76859		사극	4.766e-01	0.443	0.65807	
액션	3.938e-01	0.418	0.67626		느와르	NA	NA	NA	
드라마	3.145e-01	0.335	0.73805		스크린수	2.462e-03	17.163	<2e-16	***
코미디	3.313e-01	0.350	0.72640		봄	-7.240e-02	-0.659	0.51001	
범죄	1.845e-01	0.195	0.84522		여름	2.743e-02	0.257	0.79711	
멜로/로맨스	4.581e-01	0.481	0.63086		가을	-1.486e-01	-1.179	0.23885	

분석의 목적 중 하나인 스크린 상한제의 실효성을 알아보기 위해 전체 데이터세트를 회귀분석으로 적합시켰다. 느와르 변수가 NA인 이유는 느와르 변수가 1인 것이 하나밖에 없기 때문이다. 회귀분석으로 적합시킨 결과, 유의수준 0.001에서 스크린수는 다른 변수들을 모두 일정하게 유지하였을 때, 누적관객수를 예측하는데 유의미한 변수라고 말할 수 있다. 또 다른 방식으로 스크린수 변수가 유의미한지 알아보기 위해, 우도비 검정을 시행하였다. 우도비 검정은 두 모형의 우도비를 구한 후, 우도비의 차이가 크다면 모형을 축소하지 않고, 차이가 작다면 두 모형 중 모수의 수가 작은 축소 모형을 고려하는 검정방법이다. 첫 번째 모형으로는 전체 변수를 고려한 full model을, 두 번째 모형으로는 스크린수 변수를 제외한 reduced model을 생각한다.

Table 12. Deviance

	deviance	df
full model	509.15	597
reduced model	760.36	598

Full model과 reduced model의 우도비는 각각 509.15와 760.36이다. 우도비의 차이가 유의한지에 대해서 귀무가설 " β_{screen}=0 " , 유의수준이 0.001일 때 검정한 결과, p-value가 0으로 유의수준에서 우도비의 차이가 유의미한 것으로 나타나 귀무가설을 기각할 수 있다. 따라서 스크린수는 유의미한 변수이며, reduced model보다 full model로 적합시키는 것이 좋다고 결론지을 수 있다.

두 가지의 결과로 스크린수 변수가 누적관객수 예측에 유의미하다는 것을 밝혔다. 다른 변수들을 통제했을 때 스크린수가 증가할수록 log(누적관객수) 역시 증가하므로, 특정 영화의 스크린수 독점으로 인해 다른 영화가 피해를 볼 수 있다는 점이 일정 부분 인정된다고 볼 수 있다.

다음으로는 5.2 회귀 분석과 5.3 분류 분석, 5.4 다중 분류 분석에서 구한 최종 모형들의 변수 중요도이다.

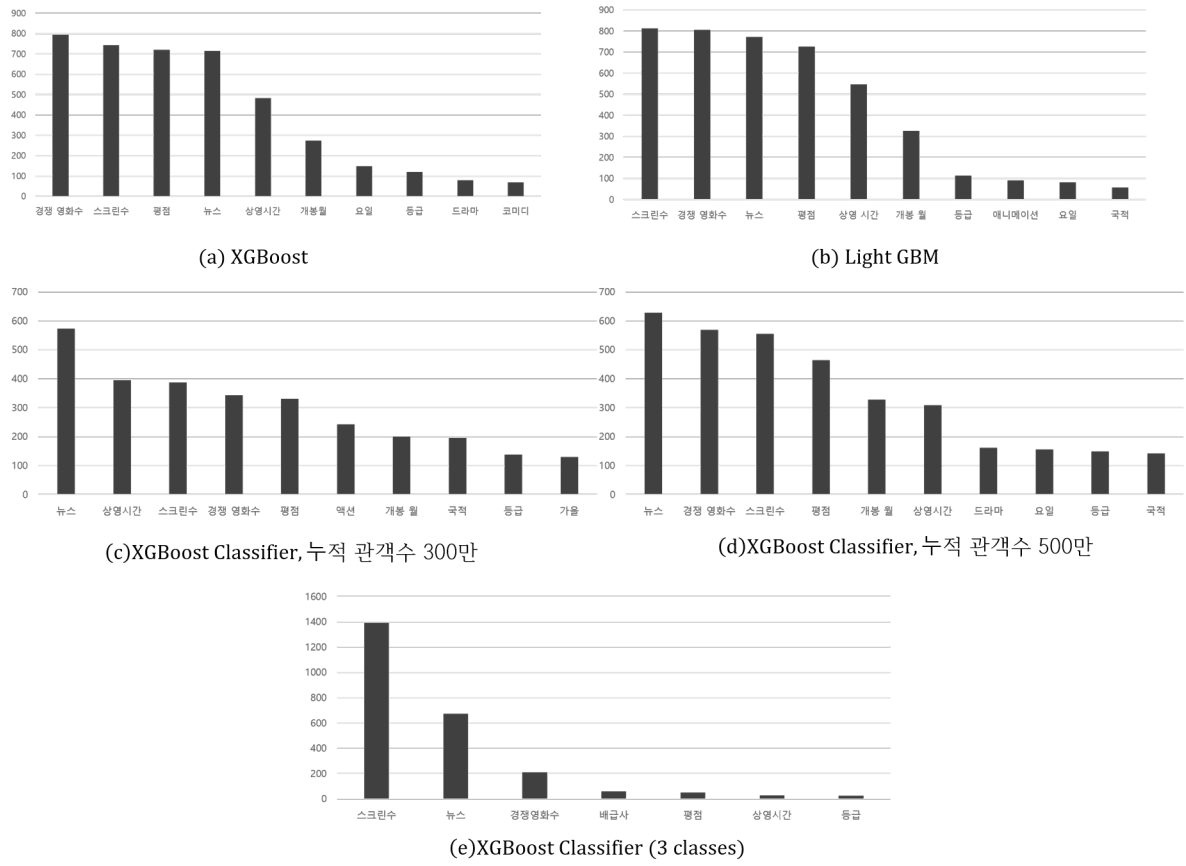


Figure 2: Feature Importance of final model

변수 중요도의 경우 각 모델 라이브러리에 내장되어 있는 plot_importance를 사용하였다. 여기서 얻어진 변수 중요도 척도의 경우 절대적인 값이 아닌 변수들 내에서의 상대적인 중요 척도이다. 교차 검증을 통해서 얻어진 각 fold의 변수 중요도를 변수별로 평균값을 계산하여 수치화하였다. 위 그래프는 34개의 변수 중요도 중에서 상위 10개에 해당하는 변수들을 시각화하였다. 회귀 분석의 경우 XGBoost와 LightGBM 예측값의 평균을 사용하였기 때문에 두 모델의 변수 중요도를 확인하였으며, 분류의 경우 흥행 기준 누적관객수 모두 XGBoost모델이 최종 모델로 선정되어 이 모델들의 변수 중요도를 확인하였다.

최종 모델의 변수 중요도를 확인해 본 결과 평점과 뉴스같이 영화 내부적인 요인이 아닌 외부적으로 대중의 영향을 많이 받는 변수들도 누적관객수를 예측하는 것에 꽤 많은 영향을 주고 있다. 하지만 상대적으로 경쟁영화수와 스크린수 변수가 각 모델에서 다른 변수들과 비교 하였을 때 예측 성능에 더 많은 영향을 주고 있음을 확인할 수 있다. 또한 상위 10개의 변수들이 회귀와 분류 모두 비슷한 변수들을 가지고 있다는 특징을 확인 할 수 있다.

따라서, 종합하여 보았을 때 관객이 몰리는 주요 시간대에 특정 영화의 상영관 스크린수를 제한하는 제도인 스크린 상한제를 실시할 근거를 마련해 준다.

6 Concluding remarks

본 연구의 한계점으로는 분석에 사용 되어지는 전체 데이터세트의 크기가 작아, 과적합 현상이 나타난다는 것이다. 추가적으로 손익분기점에 대한 분석을 시도 하려고 하였으나, 손익분기점을 계산하는 것에 있어 객관적인 자료 부족과 영화사의 제작비 비공개로 인하여 추가적인 흥행 관련 분석을 진행하지 못했다. 또한, 다중 분류 모델에서 범주별 영화의 수가 충분하지 않아 양 끝 범주를 제외한 중간 범주에 대해 예측 성능이 좋지 않았다. 따라서 이후 분석에서는 이러한 점들을 고려하여 더 많고 다양한 범주의 영화를 가지고 연구할 필요성이 있다.

끝으로 본 보고서는 영화 누적관객수 예측력 최대화와 스크린 상한제의 효용성에 대한 분석이다. 누적관객수 예측력 최대화를 위해서 회귀와 분류모형을 사용하였으며, 각 분석방법에서 1개의 최종모형을 선택했다. 그 결과, Xgboost와 lightGBM 의 평균을 이용한 앙상블 모형(Mean_Ensemble)과, XGBoost Classifier를 사용한 모형이 최종모형으로 선택되었다. 스크린 상한제 분석에서는 스크린수 변수가 우도비 검정과 변수중요도를 통해 유의하고 중요한 변수임을 확인할 수 있었다. 따라서 오랜 시간 논의가 되어온 스크린 상한제는 비록 관객의 볼 권리를 침해하는 경향이 있지만 필요한 제도임을 확인해볼 수 있다.

부록

연속형 변수들의 산점도

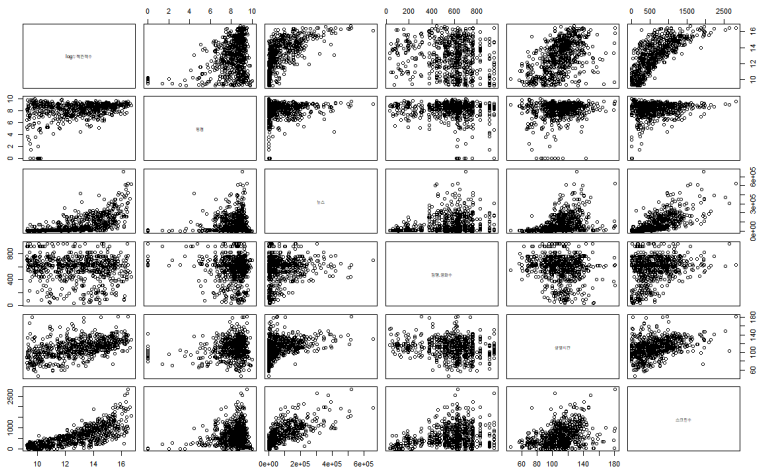
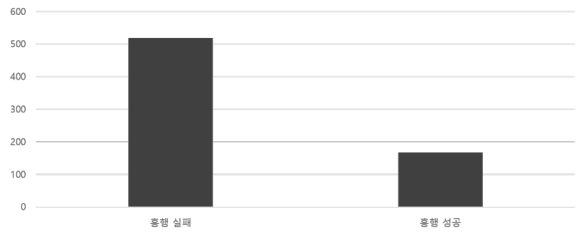
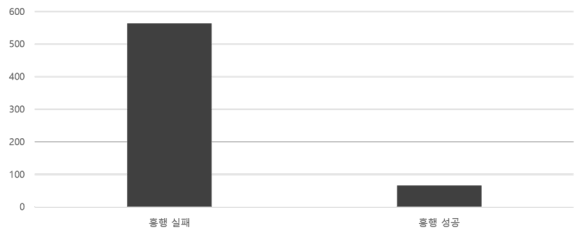


Figure 3: Scatter plot for box office

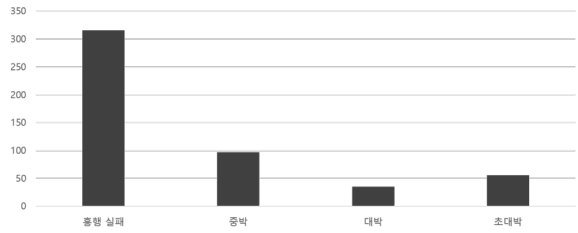
분류 분석에서의 종속변수 분포 히스토그램



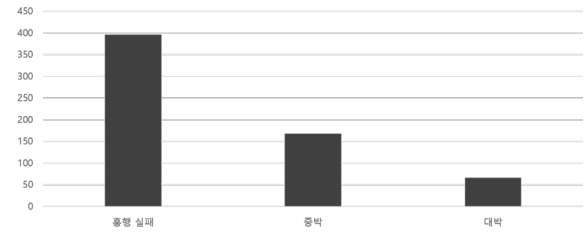
(a) 300만 기준 classification



(b) 500만 기준 classification



(c) 4 classes classification



(d) 3 classes classification

Figure 4: Class imbalance for classification

Table result of AutoML (full ver)

model	RMSE	MAE
StackedEnsemble_BestOfFamily	0.7933445	0.6056057
StackedEnsemble_AllModels	0.7955141	0.6062511
GBM_3	0.8014172	0.6115859
GBM_grid_1_model_3	0.8074838	0.6291050
GBM_2	0.8092251	0.6177355
GBM_grid_1_model_4	0.8114113	0.6240635
GBM_4	0.8212646	0.6277964
GBM_grid_1_model_5	0.8660153	0.6679476
GBM_grid_1_model_2	0.8876976	0.6728762
GBM_1	0.9028782	0.6945839
GLM_1	0.9424287	0.7375050
XRT_1	0.9514860	0.7265629
DRF_1	0.9578091	0.7200839
DeepLearning_grid_1_model_1	0.9923735	0.7648340
GBM_grid_1_model_1	1.0556940	0.8403271
GBM_5	1.0633147	0.8466254
DeepLearning_grid_2_model_1	1.0714181	0.8098127
DeepLearning_grid_1_model_2	1.2053098	0.9605742
DeepLearning_1	1.3355457	1.0477706
GBM_grid_1_model_6	1.3371830	1.1209705

참고문헌

김보경(2019). 머신러닝 기법을 활용한 정형비정형 데이터에 대한 예측 연구와 응용.중앙대학교 대학원 석사학위논문.

Jeon, S. H, Son, Y. S(2016). Prediction of box office using data mining. The Korean Journal of Applied Statistics, 29(7), 1257-1270