Exploring The Determinants of Films Success

통계학과 1711803 김지은, 통계학과 1713239 정지훈, IT공학과 1716174 구본정

December 2020

1 Introduction

박스오피스 영화 관객수에 영향을 미치는 소셜 미디어 요인에 대해 분석한다. 사람들은 영화에 대한 기본 정보를 바탕으로 영화 관람 여부를 결정하고, 이러한 영화 정보들은 네이버 포털 사이트를 통해 쉽게 확인할 수 있다. 네이버 포털 사이트와 함께 사람들이 쉽게 접근할 수 있는 영화 정보 데이터를 바탕으로 요인 분석을 실시해 흥행 요인을 파악한다.

영화 산업에서 흥행은 매우 중요한 요소이다. 국내의 경우 흥행 지표를 박스오피스 관객수로 명시하고 있으며, 흥행할수록 영화는 미디어에 노출된다. 천만영화인 기생충의 경우 각종 매스컴에 등장하며 대중적인 영화로 자리잡기도 하였다. 이에 소셜 데이터 중 뉴스 데이터, 검색량 등과 영화의 기본적인 내용을 기반으로 영화 흥행 요인에 대해 분석하고자 한다. 이는 향후 새로운 영화가 개봉했을 때 기존에 흥행한 영화들의 요인 분석 결과를 참고해 새로운 마케팅 전략을 제시할 수 있을 것이며, 기업들이 향후 개봉할 영화에 투자를 할 때 고려할 요인에 대해 설명가능하다.

본 프로젝트에서는 영화 관객수와 관련있는 요인에 대해 연구를 실시한다. 영화 스크린 수, 배급사, 배우, 관람 가능한 나이, 장르 등의 영화 요소들에 대한 데이터와 영화 검색 횟수, 뉴스 게재 수와 같은 소셜 미디어 데이터를 이용하여 요인분석을 실시한다. 영화 요소 데이터는 네이버 영화 사이트에서 웹 크롤링을 통해 수집하였으며, 관객 수 등의 통계량 데이터는 KOBIS(영화관입장권 통합전산망)에서 API 스크래핑을 통해 수집하였다.

흥행 흐름을 고려하기 위해 시간 단위를 고려한다. 개봉 후 1주일, 2주일, 상영마감일으로 3가지 기간 범위를 구분하여 분석을 실시하며 흥행 속도를 고려해 흥행 요인에 대해 분석을 실시한다. 이에 개봉 후 일정 기간 내에 순위를 기록한 영화들을 분석 대상으로 선정한다.

특히 기업들은 영화 개봉 전 영화에 대한 기본적인 정보와 기존의 영화 투자 결과를 바탕으로 투자를 진행한다. 이때 고려해야할 요소에 대한 최대한의 정보를 이용할 경우 성공적으로 투자할 확률을 높일 수 있다.

이에 영화와 관련된 요인을 내부적, 외부적 요인으로 나누어 수집한 후 관람객 수를 종속변수로 설정한 후 선형 회귀 모형을 적합하여 결과를 해석한다. 여러 가설을 기반으로 결과를 비교한 후 공통적으로 유의미한 변수를 선별하는 것이 목표이다.

2 Background

박스오피스는 어떤 영화의 흥행 성적을 가늠할 때 사용되는 용어이다. 각국의 흥행 집계 방식은 차이가 있다. 국내의 경우 영화진흥위원회 영화관입장권통합전산망에 등록된 관객수를 기준으로 흥행성적을 측정하지만 북미지역의 경우 매출액으로 흥행성적을 집계하며 배급수입을 기준으로 집계하기도 하였다. 관객수와 흥행 수익은 정비례 관계이며, 개봉한지 오래 지난 영화의 경우 물가 상승률을 고려해 흥행 순위를 측정한다.

3 Data

본 논문에 사용되는 데이터는 네이버 영화와 KOBIS(영화관입장권 통합전산망), kofic(영화진흥위원회)를 통해 얻은 자료이다. 추가적으로는 네이버 데이터랩 검색어 트랜드에서 자료를 수집했다.

네이버 영화는 영화의 이름, 감독, 개봉일 등의 정보 뿐만 아니라 관람객과 전문가의 평점을 포함하고 있어서 본 분석을 실행하기 위한 자료로 택하였다. 네이버 영화를 통해 영화의 제목, 개봉일, 감독, 국내외 작품 여부, 주요 출연 배우, 주요 장르, 심의 등급에 대한 자료를 스크래핑하였다. 또한, 관람자 별점, 전문가 별점, 네티즌 별점을 구분하여 스크래핑하였고 이것을 이용하여 영화 흥행의 요인에는 무엇이 유효한 영향을 주는지에 대해 알아볼 예정이다.

KOBIS에서는 조사 기간에 개봉된 영화의 목록을 알 수 있고, 관객수와 매출액 등의 자료를 알 수 있기 때문에 자료로 택하였다. 또한 KOBIS에서는 년, 월, 요일 등의 기준으로 해당 영화의 점유율, 상영 수 등에 대해 알 수 있다는 점에서 본 프로젝트의 데이터로 선택하였다. KOBIS를 통해 본 논문에서 목표한 기간인 2018년 1월부터 2020년 9월까지 개봉된 영화의 목록과 매출액, 관람 수를 스크래핑하였다.

네이버 데이터랩 검색어 트랜드 api를 통해 해당 영화가 개봉된 후 7일 동안 포털사이트 네이버를 통해 언급된 수를 스크래핑했다. api를 사용하면 입력한 검색어들에 대한 네이버 통합검색에서의 검색 추이 데이터를 JSON 형식으로 반환한다. 네이버 기사에서 해당 영화가 언급된 기사의 수를 저장하였고 영화가 개봉된 후 3주일 간의 기사 수를 스크래핑하였다.

KOFIC에서는 매년 진행되는 한국 영화산업 결산, 상하반기 극장 운영 및 상영지수 소개 등 공식적으로 진행되는 영화 통계조사를 참고할 수 있기 때문에 본 논문을 위한 참고자료로 택하였다.

네이버 영화에서는 HTML 문서를 스크래핑하여 필요한 정보를 얻었고 이것을 바탕으로 제목, 개봉일, 국가, 장르, 등급, 감독, 배우에 대한 데이터베이스를 만들었다. 이때 Primary Key로는 영화 이름을 지정했고 같은 이름의 작품이 있을 경오를 대비하여 개봉일을 추가로 비교했다. 네이버 데이터랩 검색어 트랜드는 HTML 문서를 스크래핑하고 영화의 제목과 개봉일을 기준으로 검색하여 해당 영화의 언급 수를 csv파일로 다운로드했다. 스크래핑한 자료와 다운로드 받은 자료로 데이터베이스를 만들었다.

KOBIS에서는 네이버 영화에서 스크래핑한 영화 이름을 기준으로 해당 영화의 매출액과 점유율, 관객 수를 얻었다. KOBIS에서는 자료를 CSV파일로 다운로드할 수 있기 때문에 정해놓은 기간에 개봉한 영화를 검색하여 자료를 다운로드하였다.

영화 목록 중 관람가, 점문가평점이 없었던 경우는 0.0으로 처리하였다. 평점을 0.0으로 한다는 것이 극단값이라고 판단하였지만 대체적으로 평점이 0.0인 영화는 관람객이 매우 적었기 때문에 종속변수인 평점과 반응변수인 누적 관객 수 간의 상관이 있을 것이라고 판단했다. 개봉일자, 영화 상영시간, 장르, 관람 등급, 감독에 대한 결측값은 네이버 영화 검색을 통해 기입하거나 KOBIS에서 해당 영화를 검색하하여 결측치를 처리했다.

본 논문을 작성하기 위해 스크래핑한 자료 중 관람객과 전문가의 평점, 검색 수, 기사 수, 관객 수, 매출액, 영화 상영 시간은 수치형 데이터이고 영화 제목, 장르, 등급, 감독, 배우의 자료는 텍스트형 데이터이다. 또한, 각 영화의 배우는 주연을 맡은 배우들을 출력하도록 하였고 역시 각각의 배우들 사이에 ','가 출력되도록 하여 구분하였다.

스크래핑한 자료를 정리하면 다음과 같다.

텍스트 테이터 : 영화 제목, 개봉 날짜, 관람 등급, 장르, 국적,감독 이름, 주연 배우 이름
수치형 데이터 : 상영 시간, 관람가 평점, 전문가 평점, 영화 기사 수, 관람객 수, 네이버 블로그 언급 수

이 중 텍스트 형태의 자료는 제목과 개봉 날짜를 제외하고 수치형 데이터와 범주형 데이터로 대체하였다. 장르는 스크래핑한 영화의 장르를 모두 수집하여 1, 2, ... , 20의 factor로 대체하였고, 관람 등급은 전체 관람가부터 청소년 관람 불가 영화 까지 총 4개의 factor로 만들었다.영화의 국적은 국내의 영화인 것과 아닌 것으로 구분하여 0,1로 대체하였다. 감독 이름은 해당하는 감독이 수상 후보에 기재된 것과 실제 수상한 상의 횟수로 대체하였고, 데이터를 통해 얻은 모든 배우는 스크래핑한 영화의 기간동안 해당 배우가 출연한 영화들의 평균 관객 수로 대체하였다.

개봉 날짜는 YYYY-MM-DD의 형태로 표현하였고, 개봉 후 3주일 간의 네이버 기사 수를 스크래핑하고, 코로나 바이러스 발생 이전과 이후, 즉, 2020년과 2019, 2018년에 개봉된 영화를 나누기 위하여 텍스트 형태의 자료를 유지하였다.

이 과정을 거쳐 스크래핑한 자료 중 텍스트형 자료는 영화 이름과 개봉 날짜이며, 관람객 평점과 전문가 평점은 실수형 데이터로, 그 외의 데이터는 모두 정수형으로 구성하였다. 수집된 데이터를 바탕으로 전처리 전후의 E-R Diagram을 만들면 Figure1과 같다.

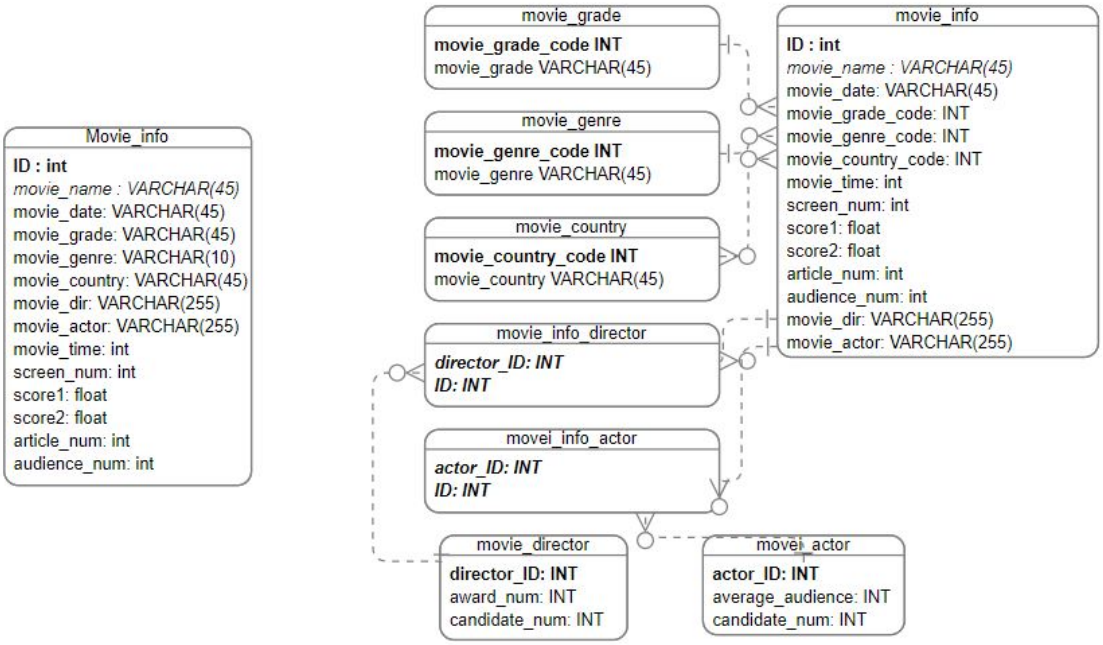

Figure 1: Picture 1. E-R Diagram

4 Methods and hypotheses

4.1 영화의 외부적, 내부적 요인을 고려한 영화 흥행 요인 분석

4.1.1 연구 방법

영화 흥행 요인 분석을 위해 다중회귀모형을 수행하기 이전에, 설명변수와 종속변수에 대한 탐색적 자료 분석 및 가공을 수행하였다. 2018년부터 2020년 10월까지 개봉한 영화 586편에 대해 분석을 진행하였으며 필요 변수에 대해서는 정규화 스케일링 등의 전처리를 추가로 진행하였다. ***Table*1**은 본 가설에서 사용한 주요 변수들의 기초통계량 값을 나타낸 것이다.

Table 1: Descriptive statistics of variables

	director	score1	score2	time	actor	article	screen	audience
count	586.00	586.00	586.00	586.00	586.00	586.00	586.00	586.00
mean	28.57	8.02	5.97	108.25	7.76e+05	103.17	445.60	800000
std	50.42	1.76	1.34	19.40	1.10e+06	110.80	458.24	1.77e+06
min	0.00	0.00	012	67.00	0.00	0.00	1.00	0.00
max	441.00	10.00	9.54	209.00	5.62e+06	486.00	2235.00	1.63e+07

586편의 영화의 평균 관객수는 80만명인 가운데 편차가 1.77e+06으로 매우 크다는 것을 알 수 있다. 실제 관객수 분포를 살펴보면 600만명의 관객수를 동원한 영화는 전체 영화 중 매우 일부분이며 그 수가 아주 적다. 통계적 모형은 주어진 자료에 대해 정규분포 가정에 기반한 추론이 이루어지기 때문에 종속변수에 대해 로그 변환을 시도하여 정규분포와 가까운 모형이 나오도록 하였다.

Table 2: Correlation between variables

	score1	score2	time	director	actor	article	screen	audience
score1	**1.00000**	0.327021	0.181776	0.084179	0.123239	0.009472	0.082785	0.138029
score2	0.327021	**1.00000**	0.147900	0.345565	0.047191	-0.01895	-0.03978	0.100082
time	0.181776	0.147900	**1.00000**	0.322578	0.395093	0.355094	0.268811	0.295039
director	0.084179	0.345565	0.322578	**1.00000**	0.112353	0.170465	0.017140	0.095998
actor	0.123239	0.047191	0.395093	0.112353	**1.00000**	0.610942	0.595514	0.814005
article	0.009472	-0.01895	0.355094	0.170465	0.610942	**1.00000**	0.472993	0.505084
screen	0.082785	-0.03978	0.268811	0.017140	0.595514	0.472993	**1.00000**	0.593457
audience	0.138029	0.100082	0.295039	0.095998	0.814005	0.505084	0.593457	**1.00000**

Table2는 더미변수를 제외한 독립변수와 종속변수에 대해서 스피어만 상관분석을 실시한 결과이다. 관객수와 스크린 수, 배우 점수, 그리고 기사 수가 유의미한 상관관계를 보인 반면 관람객 평점, 전문가 평점, 영화 상영 시간, 그리고 영화감독 점수와는 상관관계가 없는 것으로 나타났다. 또한 스크린 수는 배우 점수, 기사 수와 상관 관계를 보였으며 배우 점수는 기사 수와 상관관계를 보였다. 이러한 상관관계 분석 결과를 기반으로 다음 3개의의 모형을 제시하였다.

- 모형#1: 관객수 = 감독점수 + 배우점수 + 국적 + 장르 + 관람등급 + 상영시간
- 모형#2: 관객수 = 감독점수 + 배우점수 + 국적 + 장르 + 관람등급 + 상영시간 + 평점 + 기사수
- 모형#3: 관객수 = 감독점수 + 배우점수 + 국적 + 장르 + 관람등급 + 상영시간 + 평점 + 기사수 + 스크린수

모형1은 영화의 내부적 속성인 감독 점수, 배우 점수, 국적, 장르, 관람 등급, 그리고 영화 상영 시간만을 포함시킨 모형이다. 모형2는 모형1에서 영화 흥행의 외부적 요인이라고 할 수 있는 관람객평점, 전문가평점, 그리고 영화 개봉 후 3주간 총 기사 수 변수를 포함시켰다. 마지막으로 모형3은 관객수와 비교적 상관관계가 높았던 스크린수 변수를 포함시켰다.

4.1.2 분석 방법

종속변수가 존재하고 연속형 변수이므로 위 세 모형에 대해 회귀분석을 진행하였다. 잔차제곱합을 최소화하는 가중치 벡터를 구하는 방법인 최소자승법을 이용해 선형회귀분석을 진행하였다. 또한 후진소거법을 이용하여 변수를 하나씩 제거해가며 최적의 모형을 만들고자 하였다. ***Table3***은 세 모형에 대한 회귀 분석 결과이다.

Table 3: OLS Regression Result

	모형#1	모형#2	모형#3
adj. R-Squared	0.663	0.667	0.687
F-statistic	45.26	41.42	43.86
Prob(F-statistic)	1.73e-119	4.28e-119	7.18e-126
No. Observation	586	586	586
min(eigenvalue)	4.61e-19	1.52e-18	1.47e-18

모형#1로 분석을 진행한 결과 66.3%의 설명력을 보였으며 배우 점수가 주요 요인임을 확인할 수 있었다. 기사수와 관람객평점, 전문가 평점 변수를 추가한 모형#2로 분석을 진행한 결과 66.7%로 모형#1보다 상승된 설명력을 보였으며 배우 점수 변수 이외에도 genre20(전쟁 장르)가 주요 요인으로 확인되었다. 마지막으로 모형#3은 68.7%의 설명력을 보였으며 배우 점수, genre20(전쟁 장르), 그리고 새롭게 추가한 스크린수 변수가 영화 흥행에 유의미한 영향을 주는 것으로 나타났다.

위 결과는 다중공선성을 고려하지 않은 모형이며, 분석의 다음 단계로는 VIF 값이 10 이상인 변수를 하나씩 제거해 나가는 방식으로 다중공선성을 해결하고자 하였다.

우선 모형#1에서 각 변수에 대해 독립 변수간 상관 관계가 있는지 측정하는 척도인 VIF 값을 확인해보니 다른 변수는 모두 1점대의 값이 나온 반면 15세 관람가를 나타내는 level3 변수가 전체관람가를 나타내는 level1 변수와 12세관람가를 나타내는 level2 변수와 함께 20 이상의 높은 값을 보였다. 영화등급 변수간 상관관계가 존재함을 확인하고 level3 변수 제거를 진행한 후 다시 VIF 값을 확인하는 작업을 진행하였으며 비로소 VIF 값이 모두 10 이하가 되었다.

기사 수와 관람객평점, 전문가 평점 모형을 포함시킨 모형#2에 대해서도 VIF 값을 출력해았고, 그 결과 모든 변수가 3 이하의 값이 나온 반면 level1부터 level4 변수는 10 이상의 높은 값이 나왔다. 따라서 가장 높은 VIF 값을 가지고 있는 level3 변수를 제거한 후 다시 VIF 값을 출력해주었다. 모형#2에서는 모형#1에서와는 달리 level3 변수 제거 후 관람객평점1, 관람객평점2, 그리고 영화상영시간 변수가 20 이상의 높은 VIF 값이 나왔다. 따라서 가장 높은 VIF 값을 가지고 있는 영화상영시간 변수, 그 다음으로는 전문가평점 변수를 제거해주었다.

모형#3에 대해서도 같은 작업을 반복했으며 모형#2와 같이 level3, 영화상영시간, 그리고 전문가평점, 그리고 관람객 평점 변수를 순서대로 제거하는 작업을 진행하였다. ***Table*4**는 각 모형에 대해 다중공선성을 고려한 변수 제거 작업을 완료한 후 확인한 결과이다.

Table 4: OLS Regression Result after VIF check

	모형#1	모형#2	모형#3
adj. R-Squared	0.663	0.666	0.684
F-statistic	45.26	44.15	47.94
Prob(F-statistic)	1.73e-119	4.28e-119	1.37e-126
No. Observation	586	586	586
min(eigenvalue)	7.05e-18	1.52e-18	7.05e-18

4.2 연도별 영화 흥행 요인 분석 및 사회적 거리두기 영향 분석

4.2.1 고려한 모형

먼저 2018-2019년도 영화 데이터와 2020 영화 데이터에서 유의미한 변수의 차이를 확인하기 위한 것이므로 날짜를 기준으로 두 개의 데이터 프레임을 만들었다.

현재 데이터 셋에선 종속변수가 존재하기에 지도 학습만을 고려하였다. 선형 회귀 모형을 이용하여 p-value를 기준으로 변수를 선정하였다. 영화 관객 수와 범주형 및 수치형 자료의 연관성 파악을 위해 상관관계를 확인하였다. 영화의 관객 수의 영향을 미치는 요소를 범주형 변수와 수치형 변수로 분류하여 각각의 경우에 대해 선형회귀분석을 실시하였다. 영화의 장르와 관람 등급 등 범주로 표현할 수 있는 변수들은 범주형으로 범주화하여 분석을 실시하였다. 또한, 영화 관객 수에 영향을 주는 변수를 찾기 위하여 단계별 변수 선택법(stepwise)을 사용하여 변수를 택하였다. 각각에 해당하는 변수들이 정규성을 띄는지에 대해서는 shapiro wilk 검정을 통하여 정규성 검사를 하였다.

결과를 비교하기 위해 다양한 언어 환경에서 모델링이 가능한 h2o를 사용하여 결과를 비교하였다. h2o의 경우 비교지표로 mse를 사용한다. 이를 위해서 h2o로 데이터를 불러온 후, 데이터 간의 상관관계를 파악하여 변수를 선택하였다. 이후 feature importance를 확인하여 선형 회귀 모형의 결과와 비교하였다. 또한 반응변수가 정규성을 띄는 지 확인하기 위해 qq-plot으로 값의 분포를 확인하여 min-max 정규화를 진행하였다. 이후, 영화 점수, 영화 상영시간 변수에 대해서도 동일하게 min-max 정규화를 실시하여 최종적인 모형 적합에 사용하였다.

1-2. 분석 방법

반응변수인 영화의 누적 관객 수를 설명하는 모델을 만들기 위하여 다음과 같은 방법을 분석하였다.

1.1) 반응변수를 설명하는 유의미한 수치형 데이터를 선택하기 위하여 회귀분석을 실시하였고, 단계별 변수 선택 방법인 stepwise를 실시
1.2) 반응변수를 설명하는 유의미한 범주형 데이터를 선택하기 위하여 회귀분석을 실시하였고, 단계별 변수 선택 방법인 stepwise를 실시
1.3) 반응변수를 설명하는 유의미한 수치형 데이터, 범주형 데이터를 선택하기 위하여 회귀분석을 실시하였고, 단계별 변수 선택 방법인 stepwise를 실시

2.1) 종속변수 중 정규성을 띄는 변수를 확인하기 위하여 shapirowilk test를 실시
2.2) 변수의 정규화 실시

3) 정규화된 변수를 바탕으로 1.1, 1.2, 1.3을 동일하게 실시

4) 동일한 데이터에 대해 h2o를 적용해 결과 비교

h2o 방법에서는 train 데이터 셋 0.6, validation, test set을 각각 0.2로 비율을 맞춰 분할하여 모형 진단에 사용하였다. 정규화를 진행한 최종 데이터 셋에 gradientboosting 모형을 사용하여 mse를 구하고, 구한 mse를 바탕으로 변수 중요도를 파악하였다. 분석에 포함된 변수는 장르에 대해 one-hot 인코딩을 진행한 변수, 영화 관람가 나이, 영화 배포한 국가 정보, 반응변수가 포함되어 있다. 최종적으로 특정 장르에 대해 중요하다는 결과를 파악할 수 있었으며, 영화를 배포한 국가정보, 영화 관람가 나이 등 영화의 내부적 요인이 중요하다는 것을 확인할 수 있었다.

5 Experimental Results

5.1 영화의 외부적, 내부적 요인을 고려한 영화 흥행 요인 분석 결과

영화의 내부적, 그리고 외부적 요인을 고려해서 2018년부터 2020년까지의 영화를 대상으로 영화 흥행 요인을 분석을 진행해보았다. 분석은 세 개의 모형으로 나눠서 각각에 대해 진행을 하였으며 각 모형에 대해 상이한 결과가 나옴을 확인할 수 있었다. 다중공선성 고려 전후 모형#1의 최종 모형은 다음과 같다.

$$\mathbf{y} = (1.35e + 05) + (1.37) * \boldsymbol{actor}$$

$$\mathbf{y} = (1.54e + 05) + (-4.30e + 05) * \boldsymbol{genre20} + 1.37\boldsymbol{actor}$$

다중공선성 고려하기 전에는 배우점수인 actor만이 유의한 변수로 나왔다. 반면 VIF 값이 가장 높았던 12세관람가를 나타내는 level3 변수 제거 후 다시 회귀분석을 진행한 결과 actor 변수와 함께 범죄 장르를 나타내는 genre20 변수도 유의한 변수로 확인이 되었다. 회귀계수를 살펴보면 genre20은 음수, actor 양수로 배우점수는 영화 흥행과 양의 상관관계를 가지고 있지만 점죄 장르는 그 반대임을 확인할 수 있다. 다음으로 다중공선성 고려 전후 모형#2의 최종 모형은 다음과 같다.

$$\mathbf{y} = (-2.12e + 05) + (-3.94e + 05) * \boldsymbol{genre20} + (-2.37e + 05) * \boldsymbol{country\_kor} + (1.32) * \boldsymbol{actor}$$

$$\mathbf{y} = (-4.93e + 05) + (-2.38e + 05) * \boldsymbol{genre1} + (-4.22e + 05)\boldsymbol{genre20} + 1.31 * \boldsymbol{actor}$$

다중공선성 고려 전 모형#2의 독립변수 중 범죄 장르를 나타내는 genre20, 국내영화 여부를 나타내는 country_kor, 그리고 배우 점수를 나타내는 actor 변수가 유의한 변수로 확인된 반면 15세 관람등급, 영화상영시간, 그리고 전문가평점을 차례로 제거한 후 회귀분석을 진행한 결과 드라마 장르를 나타내는 genre1 변수와 범죄 장르를 나타내는genre20 변수, 그리고 배우 점수를 나타내는 actor 변수가 유의한 변수로 확인이 되었다. 마지막으로 모형#3의 다중공선성 고려 전후 최종 모형은 다음과 같다.

$$\mathbf{y} = (-2.552e+05) + (7.93e+04) * \boldsymbol{score2} + (-2.79e+05) * \boldsymbol{kor} + (1.1897) * \boldsymbol{actor} + (743.4385) * \boldsymbol{screen}$$

$$\mathbf{y} = (-2.44e + 05) + (-2.48e + 05) * \boldsymbol{kor} + (1.19)\boldsymbol{actor} + (736.8) * \boldsymbol{screen}$$

다중공선성 고려 전 모형#3의 유의미한 변수는 전문가평점인 score2, 국내영화 여부, 배우 점수, 그리고 스크린수로 나타났다. 반면, 변수 제거 후 유의미한 변수는 국내영화 여부, 배우 점수, 그리고 스크린 수가 남았다. 총 여섯가지 모형을 비교해본 결과 영화 흥행에 큰 영향을 미치는 요인은 배우 점수임을 알 수 있다. 배우 변수는 위 여섯 모형에 모두 등장함으로써 영화 흥행의 주요 요인으로 자리하고 있다. 여기서 배우 점수란 영화의 주연 배우가 최근 3년 출연한 영화의 관객수 평균을 나타낸다. 즉, 주연배우의 과거 출연작의 관객수가 다음 출연작의 흥행에 영향을 미친다는 것이라고 판단할 수 있다.
다음으로는 위 모형들에서 국내영화를 나타내는 country_kor 변수와 범죄 장르를 나타내는 genre20 변수가 있음을 볼 수 있다. 하지만 두 변수위 회귀변수를 보면 모두 음수로 최근 3년간 국내영화가 아닌 해외영화가, 그리고 범죄 장르가 아닌 다른 장르의 영화가 흥행에 영향 더 미친다는 것을 알 수 있다.
마지막으로는 스크린 수를 나타내는 screen 변수가 두 모형에 존재한다. 이 때 screen 변수의 회귀계수가 다른 계수에 비해 매우 큰 값을 가짐으로써 영화 흥행에 가장 큰 영향을 미치는 변수임을 알 수 있다.

5.2 연도별 영화 흥행 요인 분석 및 사회적 거리두기 영향 분석 결과

가설2를 검정하기 위하여 코로나 바이러스가 없었던 2018, 2019년에 개봉한 영화와 2020년에 개봉한 영화를 분리하여 모델을 만들어보고 적합성을 비교 분석했다. 또한, 수집을 통해 저장된 데이터를 범주형 데이터와 수치형 데이터로 분류하여 분석하고 추가적으로는 두 분류의 데이터를 통합하여 분석하였다.

1) 영화의 누적관객 수를 설명하는데 있어서 유의미한 설명력이 있는 수치형 데이터는 코로나 바이러스 전과 후에 차이가 있을 것이다.

수치형 데이터에는 관람객과 전문가 평점, 영화 상영 시간, 영화 감독의 수상 횟수, 영화 주연 배우의 평균 관람객 수가 속해있다.
2020년의 영화를 바탕으로 유의미한 설명력을 갖는 변수를 찾기 위해 회귀분석을 실시하였고 그 결과 영화 주연 배우의 평균 관객수를 의미하는 movie actor의 p-value가 0.1보다 작기 때문에 유의미한 변수라는 것을 알 수 있었다. 유의미한 변수인 movie actor를 종속변수로 하는 회귀식은 다음과 같다.

$$\mathbf{y} = (3.735e + 04) + (9.321e - 01) * \boldsymbol{movie\_actor} + \boldsymbol{\epsilon}$$

2019년과 2018년의 영화를 바탕으로 동일한 분석을 실시한 결과, 2020년과 동일한 결과가 나왔고 따라서 코로나 바이러스의 발생 전과 후와 관계없이 영화의 주연 배우가 다른 수치형 데이터보다 유의미한 설명력을 갖는다는 것을 알 수 있었다.

$$\mathbf{y} = (1.408e + 05) + (8.483e - 01) * \boldsymbol{movie\_actor} + \boldsymbol{\epsilon}$$

2) 영화의 누적 관객 수를 설명하는데 있어서 유의미한 설명력이 있는 범주형 데이터는 코로나 바이러스 발생의 전과 후에 차이가 있을 것이다.

범주형 데이터에는 영화의 장르, 관람등급, 국내 영화의 여부가 속해있다.
2020년의 영화를 바탕으로 유의미한 설명력을 갖는 범주형 변수를 찾기 위해 다중회귀분석을 실시하였고 그 결과 영화의 복합 장르 여부와 국내 영화 여부의 p-value가 0.1보다 작기 때문에 영화의 누적 관객 수를 설명하기에 유의미한 변수라는 것을 알 수 있었다.

$$\mathbf{y} = (-350004) + (350808) * \boldsymbol{multi} + (687899) * \boldsymbol{country\_kor} + \boldsymbol{\epsilon}$$

2019년과 2018년의 영화를 바탕으로 동일한 분석을 실시한 결과, 영화의 국내 영화 여부의 p-value가 0.1보다 작았다. 따라서 영화의 누적 관객 수를 설명하기에 유의미한 범주형 변수는 국내 영화 여부를 의미하는 country kor 라는 것을 알 수 있었다.

$$\mathbf{y} = (618497) + (349056) * \boldsymbol{country\_kor} + \boldsymbol{\epsilon}$$

따라서, 2)에서 언급한 가설과 같이 영화의 누적 관객 수를 설명하는데 있어서 유의미한 설명력이 있는 범주형 데이터는 코로나 바이러스의 발생 이전과 이후에 차이가 있다는 것을 알 수 있었다.
추가적으로 h2o 패키지의 gradientboosting 모형을 이용해 변수 중요도를 출력한 결과 2020년 자료의 경우 관람가 나이가 중요한 것을 확인할 수 있었고, 회귀 분석의 결과와 마찬가지로 country kor 변수가 중요한 변수임을 확인할 수 있었다.

3) 영화의 누적 관객 수를 설명하는데 있어서 유의미한 설명력이 있는 수치형과 범주형 데이터는 코로나 바이러스의 발생의 전과 후에 차이가 있을 것이다.

수치형과 범주형 데이터는 가설 1과 2를 분석하기 위해 사용되었던 모든 변수이다.
2020년의 영화를 바탕으로 유의미한 설명력을 갖는 변수를 찾기 위해 분석을 실시한 결과 영화 주연 배우의 평균 관객수를 의미하는 movie actor의 p-value가 0.1보다 작기 때문에 유의미한 변수라는 것을 알 수 있었다.

2019년과 2018년의 영화를 바탕으로 동일한 분석을 실시하였고, 영화의 상영시간을 의미하는 변수인 movie time과 movie actor의 p-value가 0.1보다 작기 때문에 유의확률 0.1에서 영화의 누적 관객 수를 설명하기에 유의미한 변수라는 것을 알 수 있었다.

따라서, 3)의 가설과 같이 영화의 누적 관객 수를 설명할 때 유의미한 설명력이 있는 변수는 코로나 바이러스의 발생 전과 후에 차이가 있다는 것을 알 수 있었다.

각 변수의 정규성을 확인하기 위하여 shapirowilk test를 실시하였고, QQ-plot을 사용하여 시각적으로도 정규성을 알 수 있는지 확인하였다. shapirowilk test를 실시한 결과, 2019년과 2018년 영화의 전문가 평점을 의미하는 movie score2만이 정규성을 띈다는 것을 알 수 있었다. 해당하는 변수의 정규성을 시각적으로 확인하기 위하여 QQ-plot을 사용하였고 결과는 다음의 그림과 같다.

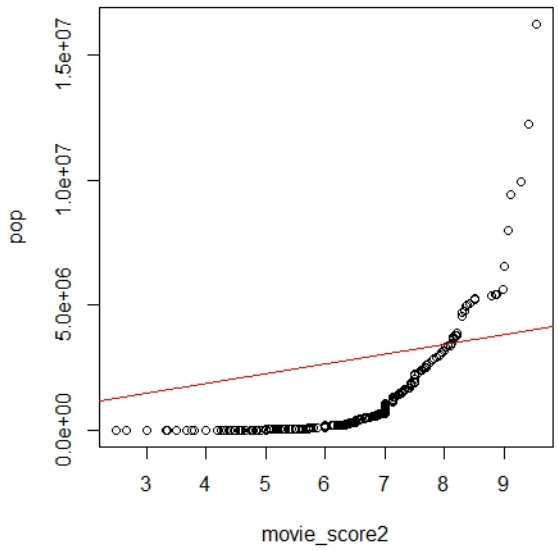

Figure 2: QQ plot with score2(x) and pop(y)

변수를 정규화 한 후 수치형 데이터와 반응변수의 분석을 실시하였다. 그 결과, 위에서의 결과와 동일하게 movie actor가 유의미한 설명력이 있다는 결과가 나왔고 2020년과 2019, 2018년의 영화 모두 동일한 결과를 알 수 있었다.

마찬가지로, 수치형과 범주형 데이터를 모두 사용하여 다중회귀분석을 실시하였다. 그 결과, 위에서의 결과와 동일하게 2020년에 개봉한 영화는 movie actor가 가장 유의미한 설명력을 갖는다는 결과가 나왔고, 2019년과 2018년에 개봉한 영화에 대해서는 movie time과 movie actor가 영화의 누적 관객 수인 pop을 설명하는데 가장 유효하다는 결과를 알 수 있었다.

따라서, 변수의 정규화와 상관없이 수치형 데이터 중에서는 영화 주연 배우가 영화의 관객수에 가장 큰 영향을 미친다는 것을 알 수 있었다. 범주형 데이터 중에서는 코로나 바이러스의 발생 이전에는 영화의 국적이 가장 유효한 설명력을 지니고, 코로나 바이러스의 발생 이후에는 영화의 장르의 갯수와 국적이 영화의 관객수를 설명하는데 가장 유효한 설명력을 갖는다는 것을 알 수 있었다. 수치형과 범주형 데이터를 모두 사용한다면 코로나 바이러스의 발생 이전에는 영화 배우가 가장 유효한 변수가 될 수 있고, 코로나 바이러스의 발생 이후에는 영화 상영 시간과 영화 주연 배우가 영화 관객 수를 설명하는 가장

유효한 변수가 될 것이다.

6 Concluding remarks

1) 본 연구의 가치

해당 연구를 통해 영화 흥행에 유의미한 설명력을 갖는 변수를 영화 내적인 요인과 외적인 요인으로 분류하여 알아볼 수 있었다.

또한, 코로나 바이러스로 인한 사회적 거리두기가 진행되기 전과 후로 나누어 분석을 실시하였다는 점에서 유사 주제의 다른 연구와 차이가 있다고 생각한다. 앞으로의 우리 생활은 코로나 바이러스 등과 같은 여러 질병적 재난들이 빈번할 것이다. 그렇기 때문에, 본 연구의 결과를 바탕으로 논의해보거나 추가적인 분석을 하기에 가치있는 연구가 된다고 생각한다.

2) 본 연구의 한계

본 연구에서는 총 586편의 영화 데이터를 이용해 분석을 진행하였다. 586편의 영화 데이터는 빅데이터라고 보기 힘들며 586편의 영화를 대상으로 분석을 진행해 사실상 유의미한 결과를 얻기에는 한계가 존재한다. 많지 않은 데이터 수이기 때문에 새로운 결과가 아닌 다소 뻔한 결과가 나올 수 밖에 없었다. 538편의 영화 중 2018-2019년의 영화는 총 438편, 2020년의 영화는 총 160편으로 데이터 불균형이 존재한다. 본 연구에서는 데이터 불균형에 대한 별다른 처리 없이 그대로 진행을 하였다.

해당 연구를 진행하면서 해당 영화의 네이버 블로그 언급 수, 그리고 네이버 기사 수에 대한 데이터를 크롤링하였다. 하지만 수집 도중 해당 사이트 구조가 바뀌면서 더이상 수집을 진행하기 힘들게 되어 결국 분석에서는 수집된 데이터를 사용하지 못했다. 추가적인 진행을 할 수 있다면 네이버 블로그 언급 수와 네이버 기사 수 중에서 영화의 관객 수에 유의미한 영향을 미치는 변수는 무엇인지에 대한 분석을 실시해 보고 싶다.

3) 그 외 추가적으로 진행하고 싶은 분석

현재는 영화의 개봉 일시 중 년도로 영화를 분류하였지만, 주어진 시간이 더 있었더라면 개봉 월을 기준으로 영화를 분류하여 추가적인 분석을 할 수 있었을 것이다. 만약 그렇게 분류된 영화를 기준으로 분석을 실시한다면 영화의 개봉 계절이 영화의 관객 수에 영향을 미치는지, 그렇다면 어느 계절에 개봉한 영화가 가장 많은 관객 수를 갖는 경향을 보이는지 등을 가설로 세울 것이다.

또한, 방학 기간과 학기 중의 기간을 나눠 방학 여부가 영화의 관객 수에 영향을 미치는지, 그렇다면 방학 기간과 학기 중의 기간 중에서 언제 개봉한 영화가 더 많은 관객 수를 갖는지를 box-plot 등을 통해 실시해 보고 싶다.

페이스북과 트위터, 인스타그램 등의 SNS에서도 해당 영화의 언급 수를 스크래핑하고 싶었지만 시간적인 부족함과 각 사이트의 보안 정책으로 인하여 스크래핑하지 못했다. 추가된 시간이 주어진다면 대표 SNS를 하나 선택하여 해당 영화의 언급 수에 대한 자료를 스크래핑해보고 그것이 네이버 블로그 언급수, 네이버 기사 언급 수와 비교했을 때 더 유의미한 설명력을 갖는 변수가 될 수 있을지에 대한 분석을 진행해보고 싶다.

7 References

[1] 소셜 빅데이터를 이용한 영화 흥행 요인 분석(Movie Box-office Analysis using Social Big Data), 이오준, 박승보, 정다울, 유은순, 2014

[2] 빅데이터 분석을 통한 영화 관객수, 매출액 예측 모델(Movie attendance and scales forcast model through big data analysis), 이응환, 우종필, 2019

[3] 통계 분석으로 본 천만 영화, KOFIC, 2014

[4] 딥러닝을 이용한 영화 흥행 예측과 주요 변수의 선택 연구: 다변량 시계열 데이터 중심으로(Movie Box-office Prediction using Deep Learning and Feature Selection: Focusing on Multivariate Time Series), 변준형, 김지호, 최영진, 이홍철, 2020

[5] 한국 영화시장의 흥행결정 요인에 관한 연구,(The Determinants of Motion Picture Box Office Performance: Evidence from Movies Released in Korea, 2006-2008, 박승현, 정완규, 2009