Exploring The Determinants of Films Success

김지은(kje5076@gmail.com), 구본정(hellowd97@gmail.com), 정지훈(tnraud1713239@gmail.com)

1. Introduction

영화 산업에서 흥행은 매우 중요한 요소이며 국내에서는 흥행 지표를 박스오피스 관객수로 명시하고 있다. 이에 영화 관객수와 관련있는 요인에 대해 분석하고자 한다. 이는 향후 새로운 영화가 개봉했을 때 기존에 흥행한 영화들의 요인 분석 결과를 분석해 새로운 마케팅 전략을 세울 수 있을 것이며, 기업들이 향후 개봉할 영화에 투자를 할 때 고려할 요인에 대해 설명 가능하다.

본 프로젝트를 통해 기업과 개인 투자자가 투자 전략을 세울 때 고려할 요소들에 대한 제안을 한다. 크라우드 펀딩은 개인이 다수를 대상으로 자금을 모으는 투자 방식이다. 현재 여러 영화에 대해서 크라우드 펀딩이 이뤄지고 있고, 수집 가능한 정보가 한정된 개인 투자자들이 적절한 투자를 할 수 있도록 한다.

본 프로젝트에서는 감독, 주연 배우, 장르, 관람 등급, 장르 등의 영화의 내부적 요소들과 더불어 평점, 스크린 수, 관련 기사 수, 등의 외부적 데이터를 이용해 요인분석을 실시한다. 이때, COVID-19로 인한 사회적 거리두기 정책이 영화 관람에도 영향을 미쳤을 것이라는 가정하에 분석을 실시한다. 2020년 초부터 발생한 COVID-19로 인해 상황의 차이가 있을 것이라 판단해 2020년 1월 전후의 자료를 따로 분석한다.

2. Data

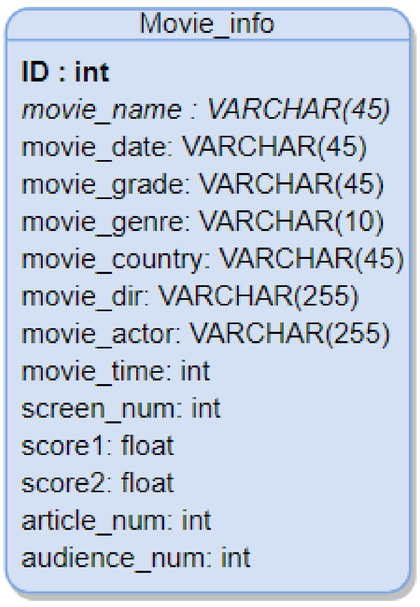
- **데이터 수집**

- 웹 스크래핑(네이버), API 스크래핑(영화진흥위원회)

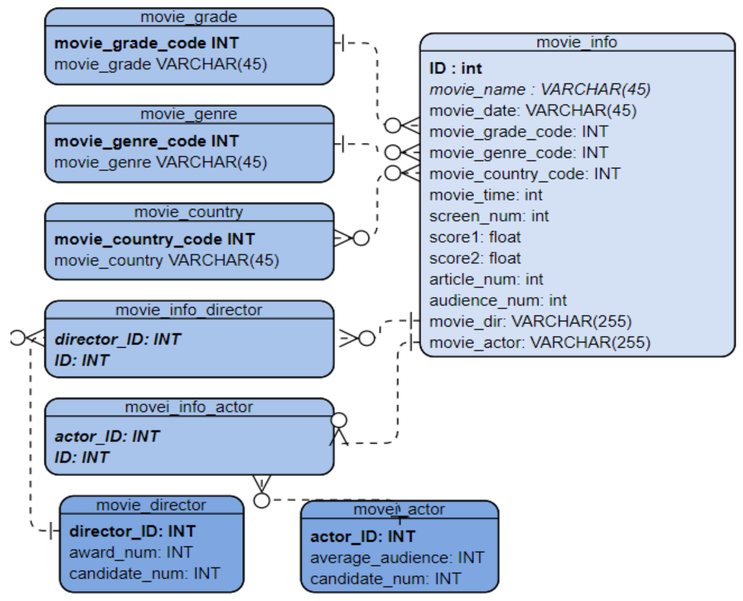
수집 방식	웹 스크래핑	범주형	장르	movie_genre
			나라	movie_country
			관람등급	movie_level
		연속형	관람객평점	movie_score1
			전문가평점	movie_score2
			상영시간	movie_time
			감독	movie_dir
			주연배우	movie_actor
	API 스크래핑	범주형		
		연속형	네이버 데이터랩	naver_datalab
			스크린수	movie_screen
			누적관객수	movie_audience

- **데이터 전처리 - 독립변수**

[전처리 전]

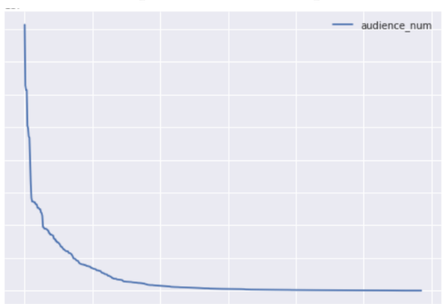


[전처리 후]

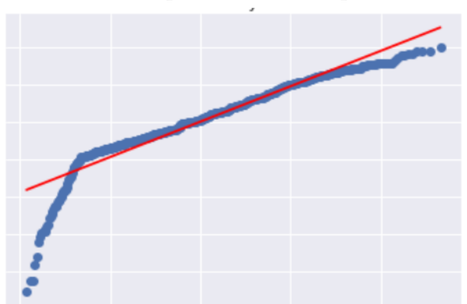


- **데이터 전처리 - 종속변수**

[로그 변환 전]



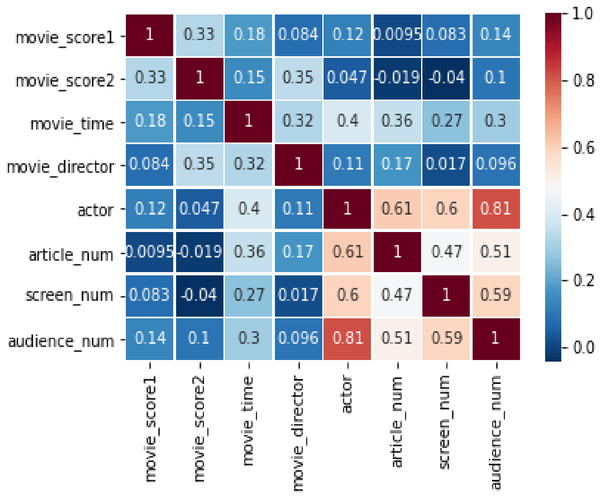
[로그 변환 후]



3. Methods

◆ **Analysis 1: 영화의 외부적, 내부적 요인을 고려한 영화 흥행 요인 분석**

- **1단계:** 주요 변수간 상관관계 분석



- **2단계:** 다항회귀분석

<모형1>

누적 관객수(y) = 감독 점수 + 배우 점수 + 국적 + 장르 + 관람 등급 + 상영 시간

<모형2>

누적 관객수(y) = 감독 점수 + 배우 점수 + 국적 + 장르 + 관람 등급 + 상영시간 **+ 평점1 + 평점2 + 기사 수**

<모형3>

누적 관객수(y) = 감독 점수 + 배우 점수 + 국적 + 장르 + 관람등급 + 상영시간 + 평점 + 기사수 **+ 스크린수**

- **3단계:** 다중공선성 문제 해결: VIF 이용한 변수 제거

◆ **Analysis 2: 연도별 영화 흥행 요인 분석 및 사회적 거리두기 영향 분석**

- **1단계: 연도별 변수 Split**

2018-2019	2020
438 x 28	160 x 28

- **2단계: 다항회귀분석**

<모형1>

누적관객수(y) = 평점1 + 평점2 + 상영 시간 + 감독 점수 + 배우 점수

<모형2>

누적관객수(y) = 장르 + 관람 등급 + 나라

<모형3>

누적관객수(y) = 평점1 + 평점2 + 상영 시간 + 감독 점수 + 배우 점수 + 장르 + 관람 등급 + 나라

- **3단계: 정규화 스케일링 후 다항회귀분석**

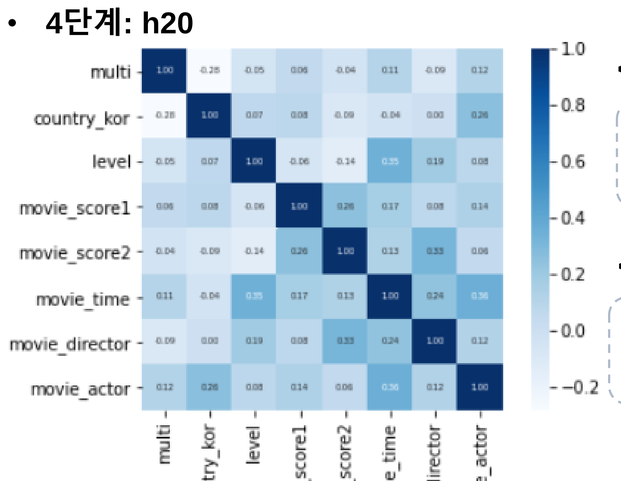
<모형4>

누적관객수(y) = 평점1 + 평점2 + 상영 시간 + 감독 점수 + 배우 점수

<모형5>

누적관객수(y) = 평점1 + 평점2 + 상영 시간 + 감독 점수 + 배우 점수 + 장르 + 관람 등급 + 나라

- **4단계: h20**



<모형1> GradientBoosting

Mse: 0.003
Train : Valid : Test = 0.6 : 0.2 : 0.2로 분할

<모형2> GradientBoosting

Mse: 0.016
Train : Valid : Test = 0.6 : 0.2 : 0.2로 분할

4. Experimental Result

◆ **Analysis 1: 영화의 외부적, 내부적 요인을 고려한 영화 흥행 요인 분석 결과**

- 다중공선성 제거한 최종 모형:

$$y = 1.54e^{05} - 4.30e^{+05} * genre20 + 1.37 * actor$$
$$y = -4.93e^{05} - 2.38e^{+05} * genre1 - 4.22e^{+05} * genre20 + 1.31 * actor$$
$$y = -2.44e^{+05} - 2.48e^{+05} * kor + 1.19 * actor + 736.8 * screen$$

◆ **Analysis 2: 연도별 영화 흥행 요인 분석 및 사회적 거리두기 영향 분석 결과**

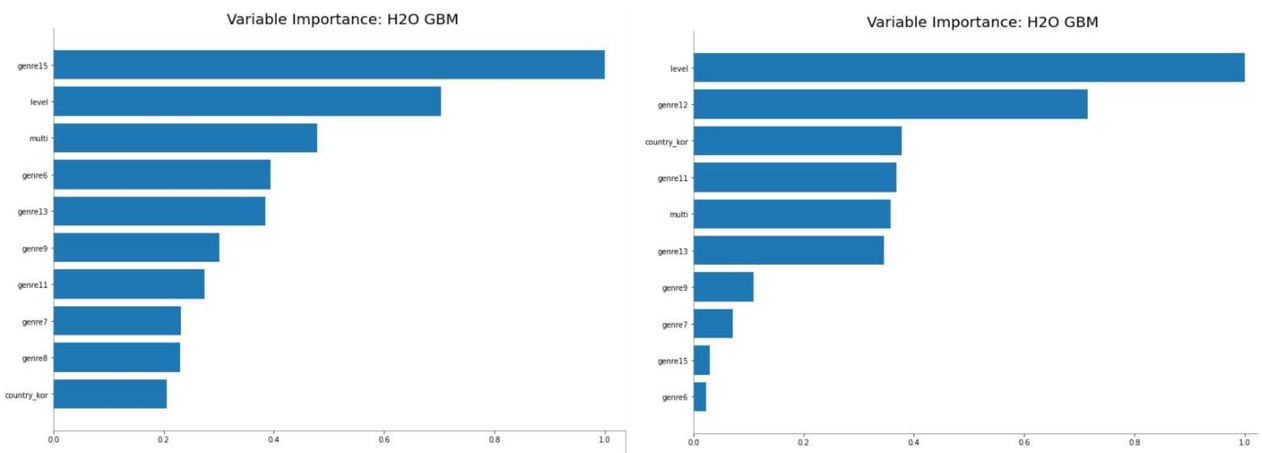
수치형 데이터와 반응변수

$$\mathbf{y} = (3.735e+04) + (9.321e-01) * \boldsymbol{movie\_actor} + \boldsymbol{\epsilon}$$
$$\mathbf{y} = (1.408e+05) + (8.483e-01) * \boldsymbol{movie\_actor} + \boldsymbol{\epsilon}$$

범주형 데이터와 반응변수

$$\mathbf{y} = (-350004) + (350808) * \boldsymbol{multi} + (687899) * \boldsymbol{country\_kor} + \boldsymbol{\epsilon}$$
$$\mathbf{y} = (618497) + (349056) * \boldsymbol{country\_kor} + \boldsymbol{\epsilon}$$

◆ **H2o의 gradientboosting 모형을 통해 파악한 변수 중요도**



다른 모형과 공통 적으로 특정 장르에 대한 변수가 중요하다는 것을 확인할 수 있었으며, 영화 내부 요인에 대해선 중요한 변수에 대한 결과가 분석1 방법과 차이가 있었다.

5. Reference

[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