

# 한국 상업 영화 관객수 예측

2019년 2월 25일

딥러닝 기반 핵심산업별 빅데이터 분석 전문가 과정  
데이터 스피커  
(박희지, 임현수, 정다연)

# 목 차

1. 프로젝트 개요 .....	1
1.1 프로젝트 기획 배경 및 목표 .....	1
1.2 구성원 및 역할 .....	1
1.3 사용 언어 및 분석 도구 .....	1
1.4 프로젝트 추진 일정 .....	2
2. 프로젝트 현황 .....	3
2.1 시장 조사 .....	3
2.2 유사 분석 결과 장단점 분석 .....	3
2.3 차별화 핵심 전략 기술 .....	4
3. 프로젝트 분석 결과 .....	7
3.1 분석모형 정의 및 제시.....	7
3.2 한국 상업 영화 예측 분석서 .....	8
3.3 한국 상업 영화 예측 모델 상세 설명.....	13
3.4 활용 방안 .....	25
4. 기대 효과 .....	26
4.1 향후 개선 사항 .....	26
4.2 기대 효과 .....	27
5. 분석 후기 .....	27

# 1. 프로젝트 개요

## 1.1 프로젝트 기획 배경 및 목표

우리는 딥러닝 기반의 빅데이터 분석과정을 통해 Python 과 R 을 이용하여 한국상업영화의 관객수를 예측하였다. API 와 웹 크롤링으로 수집한 데이터를 전처리하고 다양한 분석기법 모델링인 ANN, 의사결정나무, 일반회귀예측 등을 활용하여 정확도가 95.6%까지 올라가는 예측 모형을 개발하였다. 이 프로젝트의 주된 목적은 국내에서 개봉되는 한국영화의 관객수를 예측하는 모형을 제시함으로써 배급사 홍보팀과 극장주가 보다 효과적인 마케팅과 스크린 수 배정으로 안정적인 수익을 창출할 수 있도록 하는 것이다.

## 1.2 구성원 및 역할

이름	전공	역할	구현 부분
박희지	세종대학교 (컴퓨터공학과)	팀장	데이터 수집 및 전처리, 분류 분석, SNS 텍스트 마 이닝, 사회연결망 분석
임현수	University of California, Davis (심리학과, 커뮤니케이션학과)	팀원	시장조사, 데이터 수집, 회귀분석
정다연	동서대학교 (일본어학과, 중국어학과)	팀원	데이터 수집 회귀분석

## 1.3 사용 언어 및 분석 도구

- 1) 사용 언어 : R, Python
- 2) 개발 환경 : R studio, Anaconda Jupyter Notebook, PyCharm
- 3) 사용 라이브러리

<표 1> 사용 라이브러리

R	Python
Psych / cluster / NbClust / kohonen / ggplot2 / gridExtra / scales / QunatPsyc / caret / Hmisc / PerformanceAnalytics / corrplot / gplots / RColorBrewer / d3heatmap / GPArotation	matplotlib / datetime / NumPy / pandas / seaborn / requests / selenium / nltk / konlpy /wordcloud sklearn - sklearn.preprocessing - sklearn.linear_model - sklearn.tree - sklearn.ensemble - sklearn.svm - sklearn.neighbor - sklearn.neural_network

### 1.4 프로젝트 추진 일정

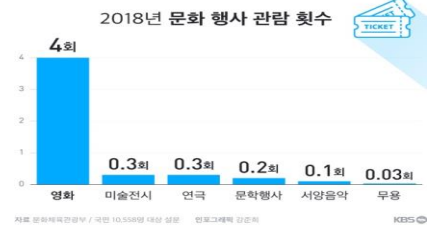
구분	기간	활동	내용
사전 기획	2018년 10월 20일 ~ 2018년 11월 20일	분석 주제 논의 및 선정	- 분석 주제 선정 및 정의
	2018년 11월 21일 ~ 2018년 11월 27일	분석 주제 정의	예측 모형 생성 방안 리서치 - 예측 모형 생성 방안 논의 후 방향 확립 - 과제 수행 일정 계획 수립
	2018년 11월 28일 ~ 2018년 12월 01일	데이터 정의	- 데이터 이해 및 가설수립 - 필요데이터 선택 및 정의
PJT 수행 및 완료	2018년 12월 02일 ~ 2018년 12월 14일	데이터 수집 데이터 전처리	- 기본 데이터 수집 - 파생변수 및 외생변수 수집 (외부데이터 API 및 크롤링) - 데이터 전처리
	2018년 12월 14일 ~ 2018년 12월 21일	데이터 탐색 1	데이터 기초통계 - 가설에 대한 검증
	2018년 12월 21일 ~ 2018년 12월 25일	데이터 탐색 2	주요 영향인자 분석 - 군집 분석을 통한 특성 분석
	2018년 12월 26일 ~ 2019년 01월 10일	모델링 생성 및 적 용과 분석 1	회귀 분석을 통한 고객 수 예측 모델링 - 기본 다중회귀 분석 - 요인분석 후 다중회귀 분석 - 딥러닝을 이용한 회귀예측
	2019년 01월 11일 ~ 2019년 01월 31일	모델링 생성 및 적 용과 분석 2	분류 분석을 통한 고객 수 예측 모델링 모형 평가 검증 및 성능 향상 - Linear Regression 분류 예측 - Decision Tree 분류 예측 - Random Forest 분류 예측 - Gradient Boosting Tree 분류 예측 - K-NN 분류 예측 - Neural Network 분류 예측
	2019년 02월 04일 ~ 2019년 02월 9일	SNS 리뷰 및 SNA 분석	- 워드 클라우드를 이용해 SNS 리뷰 분석 - 사회연결망 분석 (SNA분석)
	2019년 02월 10일 ~ 2019년 02월 14일	분석 결과 정리 및 활용방안 수립	- 분석 결과 보고서 작성 - 예측 모형 활용방안 수립 - 분석 기대효과 및 개선사항 정리
	2019년 02월 15일 ~ 2019년 02월 25일	보고서 작성 발표자료 작성 최종 발표	- 분석 및 결과 보고서 및 발표 자료 작성 - 최종 발표

## 2. 프로젝트 현황

### 2.1 영화 산업 시장조사

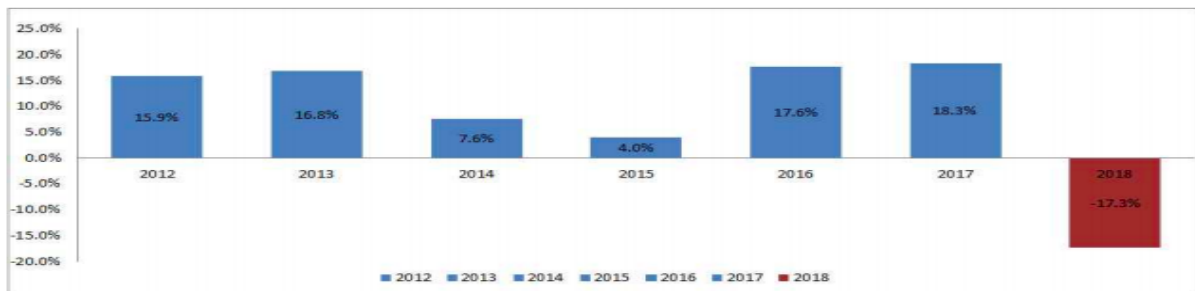
문화체육관광부에 의하면 2018년 문화 행사 관람 횟수가 가장 큰 분야는 영화였다 <그림1>. 증가하는 수요와 평균관람 요금의 상승으로 인해 전체 매출액은 증가하였지만 이상하게도 영화관계자들의 수익률은 미미하거나 떨어지고 있는 추세이다. 실제로 영화진흥위원회의 자료 <그림 2>에 따르면

<그림 1> 2018 문화 행사 관람 횟수



2018년의 추정 수익률은 과거 수익률에 비해 눈에 대폭 하락하였다 2018년 상업영화의 평균 총 제작비는 103.4억원으로 전년대비 5.7% 상승하였지만 이들의 평균수익률은 -17.3%로 추정되었다. 영화산업은 고위험-고수익 산업이라고 할 수 있다. 대작의 영화는 고수익의 확률이 높아지지만, 고위험 측면에서 보았을 때 영화산업은 소비자의 기호를 읽어 내기 어려운 관점이 많고 실패할 경우 손실비용이 높다. 이러한 산업의 불확실성은 전략적인 마케팅을 통해 소비자의 기호에 맞게 효과적인 영향을 주는 것이 중요하다.

<그림 2> 2012-2018년 한국 '상업영화' (추정)수익률 추이 (출처: 영화진흥위원회(KOFIC)) (단위 : %)



#### 2.1.1 프로젝트 주제선정 배경 및 필요성

- 1) 고객 : 극장주 & 배급사 홍보팀
- 2) 목적 : 불확실한 수익성을 데이터 기반의 딥러닝 기술을 활용하여 스크린 수 배정과 홍보를 효과적으로 할 수 있도록 고객을 돕는 것에 의의가 있다.
- 3) 핵심변수 : 관객수(종속변수), 배우 파워, 감독 파워, 트위터상의 영화 언급 수, 스크린 수

### 2.2 유사 분석 결과 장단점 분석

#### 2.2.1 기존 연구모델(변수) 및 분석방법 소개

영화 관객수 예측 모델 관련 여러 기존 연구들 중 최신성, 인용 빈도수, 우리 주제와의 관련성을 기준으로 다음의 논문들을 특히 참고하였다.

- 1) 논문: 「영화의 주별 흥행성과에 미치는 영향: 온라인 구전을 중심으로」 박승현, 송현주(2012)  
 변수: 제작비, 장르, 등급, 스타파워, 배급사 파워, 스크린 수, 전문가 평가  
 분석방법: 시계열, 다중회귀

2) 논문: 「베이지안 선택 모형을 이용한 영화 흥행 예측」 이경재, 장우진(2006)

변수: 영화 속성 (국적, 감독, 배우, 장르), 구전 효과(잡지 리뷰 수, 영화 평점, 평가자 수), 경쟁 효과(스크린 수, 계절성, 경쟁, 배급사)

분석방법: 베이지안 선택 모델 (인공지능망과 비교하여 베이지안 모델의 타당성 검증)

3) 논문: 「영화 관객 수요 예측을 위한 영향 요인 분석」 박선영, 신민정, 김태구, 신문수(2017)

변수: 연휴 여부, 감독변수(해당 작품 이전의 수상 경력, 이전 작품 중 최대 관객 수, 3년 이내 연간 50위 내에 진입한 영화의 수, 3년 이내 연간 20위 내에 진입한 영화의 수), 경쟁 변수(영화의 수, 영화의 수, 상영횟수, 상영기간, 배급력 불균형)

분석방법: 선형회귀분석, 비선형 기계학습 알고리즘(ANN, K-NN, DT, SVM)

## 2.2.2 유사 분석 결과 장단점

1) 「영화의 주별 흥행성과에 미치는 영향: 온라인 구전을 중심으로」 온라인 구전이 실제로 작동하고 있으며 흥행성과에 영향을 미친다는 것을 주별 분석을 통해 검증했다. 온라인 구전이 피상적으로 분석됐다는 한계가 있다. 온라인 구전의 총량은 네티즌 영화평의 빈도로 했으며, 리뷰 메시지를 분석해서 그 속에 담겨 있는 작품성과 대중성 등에 대한 관객들의 평가를 분석하지 못했다.

2) 「베이지안 선택 모형을 이용한 영화 흥행 예측」 감독, 배우, 장르 등의 영화 속성 변수뿐만 아니라, 입소문에 의한 영화 관람 결정 등의 구전효과와 경쟁 영화의 개봉으로 인한 효과를 반영할 수 있는 변수를 추가하여 모델의 정확성을 높였다. 하지만 각 그룹별 이질성만 고려했을 뿐 영화간 이질성은 고려하지 못했다. 영화의 마케팅에 관한 효과와 구전효과 간의 상관관계를 고려한 모델 수립 및 마케팅 변수 추가가 필요하다.

3) 「영화 관객 수요 예측을 위한 영향 요인 분석」 기존의 연구에서 다루지 못한 다양한 형태의 경쟁환경을 반영하는 방안을 도출했다. 독립 변수(배우 파워, 입소문, 속편유무, 그룹별 비교분석 추가)의 추가 및 수리 모델의 확대가 필요하다.

## 2.3 프로젝트 차별화 포인트

### 2.3.1 연구모델 측면 차별화

#### 1) 소비자 행동이론 AISAS 를 기반으로 변수를 선정

온라인 시장이 활성화되고, SNS 를 통한 커뮤니티가 활성화되면서 제품 구매정보 뿐만 아니라, 구매 후기를 공유하면서 마케팅의 채널이 온라인으로 급격하게 이동했다. AISAS 는 소비자 구매 행동 패턴으로 소비자가 제품에 대한 정보를 직접 검색하고 그 제품에 대한 자신의 경험을 공유하는 현 사회의 소비단계를 설명하는 이론이다. 이 이론에 따르면, 소비자들은 구매행동까지 Attention(주목) → Interest(흥미) → Search(검색) → Action(구매행동) → Share(공유)와 같은 5 단계의 패턴을 띄고 있다. SNS 와 인터넷 커뮤니티의 발전 및 보급에 따라, 영화 산업도 영향을 받고 있기에 해당 행동 이론을 채택하였으며, 이를 기반으로 <표 2> 와 같이 변수를 선정하였다.

<표 2> 소비자 구매 행동 패턴 AISAS 기반 변수 선정

단 계	설 명	변 수
Attention(주목)	제품이나 서비스에 대해 알게 되고 주목하는 단계	기사 수, 개봉 전 트위터 수
Interest(흥미)	제품이나 서비스에 관심과 흥미를 갖는 단계	배우 수상횟수, 감독 수상횟수, 배우파워, 감독파워
Search(검색)	흥미를 가진 제품이나 서비스에 대해 검색하는 단계	개봉 전 트위터 수
Action(구매행동)	제품이나 서비스를 구매하는 단계	개봉이전 스크린 수, 개봉이전 상영횟수
Share(공유)	제품에 대한 정보와 경험을 공유하는 단계	개봉 전 트위터 수

## 2) 새로운 파생변수를 추가

기존 연구에서는 다루지 않았던 '개봉이전 스크린 수 평균, 상영횟수 평균' 등의 파생변수를 새로이 추가했다. 그 외에도 사계절을 반영한 'season', 방학 여부를 반영한 'holiday', '개봉 전후 트위터 언급 수', '기사 수' 등의 변수를 추가했다.

## 3) 변수에 대한 새로운 정의

감독 파워, 배우 파워, 배급사 등 변수에 대한 정의를 기존 연구와는 다르게 정의했다.

<표 3> 기존연구와 변수에 대한 정의 비교

변수	기존 연구	데이터 스피어버그
배우 파워	박승현(2012)은 분석 대상에 포함된 출연작의 개봉 시점으로부터 4년 전까지 주연으로 출연했던 영화들 중 1백만 명 이상의 관객을 동원한 영화가 있는지에 따라 결정	배우 필모그래피 수상횟수와 지금까지의 참여 작품 중 관객 수가 500만 이상인 영화 개수의 합계 (손익분기점을 고려하지 않았을 때, 전 국민의 약 10%가 보았다면 어느 정도 흥행한 영화로 간주할 수 있다는 영화관계자의 자문을 참고하여 반영)
	안용대(2017)는 오프라인 마켓 리서치 데이터를 수집하여 인지도 및 선호도에 따라 배우 파워를 설정	
	김유진(2017)는 영화 개봉시점 전 3년 간 연출작 수의 합으로 집계	
감독 파워	안용대(2017)는 오프라인 마켓 리서치 데이터를 수집하여 인지도 및 선호도에 따라 감독 파워를 설정	감독 필모그래피 수상횟수와 지금까지의 참여 작품 중 관객 수가 500만 이상인 영화 개수의 합계 (손익분기점을 고려하지 않았을 때, 전 국민의 약 10%가 보았다면 어느 정도 흥행한 영화로 간주할 수 있다는 영화관계자의 자문을 참고하여 반영)
	김유진(2017)의 경우 영화 개봉시점 전 3년 간 연출작 수의 합으로 집계	
배급사	박승현(2012)은 메이저, 중간 급, 마이너로 분류함 메이저: 극장 체인을 직접 소유한 배급사 중간 급: 2010년 한 해 기간 동안 6편씩 배급한 배급사 마이너: 그 이외의 소규모 배급사	Big: 5년(2014-2018)동안 배급한 영화가 30편 이상인 배급사 Middle: 5년(2014-2018)동안 배급한 영화가 10편 이상 30편 미만인 배급사 Small: 5년(2014-2018)동안 배급한 영화가 10편 이하인 배급사
	박선영(2017)은 첫 주 좌석수를 기준으로 배급력을 Big, Small로 분류함	

## 2.3.2 분석 기법 측면 차별화

### 1) 다양한 기법으로 분석 시행

로지스틱 회귀, 의사 결정 트리, 랜덤 포레스트, K-NN, 인공신경망(ANN) 분석 기법 등 다양한 방법을 시도 후, 정확도 및 예측력이 가장 높은 모델을 선택했다. 또한, 기존 연구에서는 시도하지 않았던 워드 클라우드와 사회연결망 분석(SNA)기법을 시도하여 인사이트를 도출해 내려 했다는 점에서 의의가 있다.

### 2) 분류 예측 시 분류 범위를 세분화

기존 연구에서는 관객수를 100 만 단위로 분류 기준을 정했다. 400 만 이하의 영화의 개수 밀집도가 높은 것을 확인하고, 관객수가 400 만 이하인 영화는 50 만 단위로 세부적으로 범위를 나눴다는 점에서 기존 연구와의 차이가 있다.

### 3) 온라인 리뷰 텍스트 분석

기존 연구와는 달리, 네이버 영화 온라인 리뷰 및 영화 명 연관 트윗 텍스트를 크롤링했으며, 텍스트 마이닝 기법을 시도했다. 빈출 어휘 및 감정 단어를 분석해서 워드 클라우드를 도출하고 특징을 파악하였다는 점에서 기존 연구와 차이가 있다.

## 2.3.3 데이터 측면 차별화

### 1) 웹 크롤러 제작을 통한 새로운 변수 수집

<표 4> 크롤링을 이용한 데이터 수집 및 생성 변수

사이트명	크롤링 대상	생성한 파생변수
온라인비즈니스센터(KoBiz)	영화인 출품내역 리스트 중 수상작	감독수상, 배우수상
영화진흥위원회(KOFIC)	영화인 필모그래피의 각 영화들 중 관객수가 500 만 이상인 영화	감독파워, 배우파워
네이버 뉴스	개봉일 전 '영화+영화명'이 포함된 뉴스 수	개봉 전 뉴스 수
네이버 영화	영화 정보	상영 등급
트위터(SNS)	개봉일 기준 전, 후 일주일의 기간 동안의 '영화' + '영화 명' 언급 수	개봉 전/후의 SNS 트위터 영화 언급 수

### 2) 비정형 데이터 활용 차별화

기존연구에서는 온라인 구전을 영화평의 빈도와 평점에만 국한되었기에, 관객의 반응을 확인하기에는 한계점이 있었다. 또한 기존 연구에서는 대부분 영화진흥위원회 API 에서 제공하는 정형데이터를 기반으로 연구를 진행한 것에 비해 우리는 네이버 영화 온라인 리뷰 데이터, SNS 트위터 영화 언급 수와 같은 비정형 데이터까지 포함시켜 분석 및 활용방안에 반영하였다는 점에서 차이가 있다.

## 2.3.4 활용시사점 측면

### 1) 영화 개봉 전 관객수 예측 모델 활용

기존 연구와는 달리, 영화 개봉 전 관객수 예측 모델로 배급사의 배급 팀, 홍보 팀 그리고 극장주가 스크린 수 편성 및 마케팅 전개 시에 활용 가능성이 엿보인다. 더 나아가 뮤지컬, 연극, 오페라 등 다른 문화 산업의 수요 예측 모델 생성 및 응용을 시도해볼 수 있다.

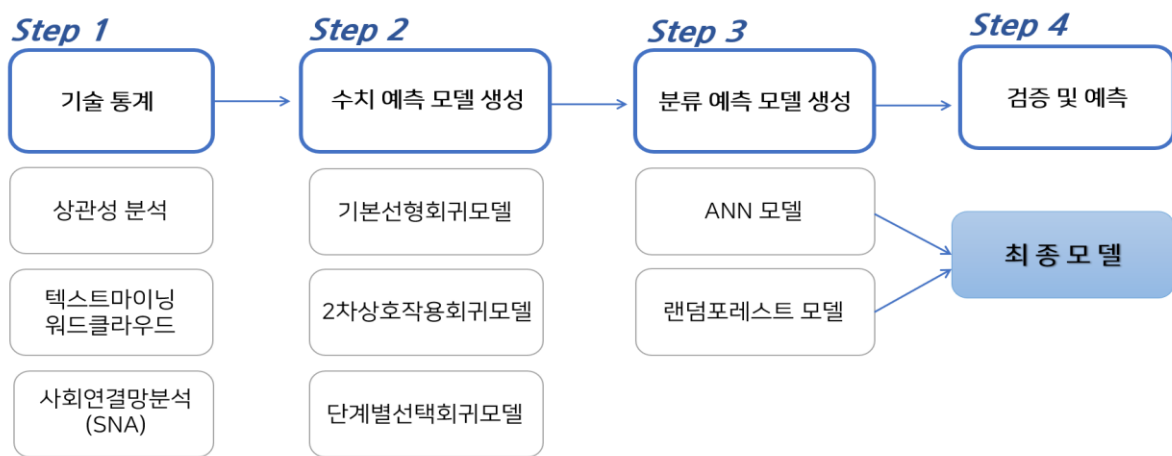


2) 배급사의 영화 마케팅에 활용

상관성 분석을 통해 중요한 변수를 체크할 수 있으며, 이를 통해 시사회 현장 SNS 이벤트를 다양화하거나 인플루언서(Influencer) 마케팅을 전개하는 등의 방안을 제안했다는 점에서 기존연구와의 차이가 있다.

### 3. 프로젝트 분석 결과

#### 3.1 분석모형 정의 및 제시



#### 3.1.2 코딩 북

<표 5> 코딩 북 정의 및 출처

컬럼 명	데이터	출처	방식	정의
영화명	Index	영화진흥위원회	DB	
관객수	연속	영화진흥위원회	DB	최종 예측 값
배급사 크기	범주	영화진흥위원회	DB	배급사의 크기 파악 (big/mid/small)
Holiday	범주	개봉일 응용	전처리	방학 시즌 여부 (0: 일반 /1: 방학)
Season	범주	개봉일 응용	전처리	계절 (1: 봄/2: 여름/3: 가을/4: 겨울)
감독수상	연속	영화진흥위원회	크롤링	감독의 수상횟수
감독파워	연속	영화진흥위원회	크롤링	감독의 이전 영화 흥행여부
개봉 전 스크린 수	연속	영화진흥위원회	크롤링	시사회 스크린 수
개봉 전 상영횟수	연속	영화진흥위원회	크롤링	시사회 상영횟수
개봉 전 뉴스 수	연속	네이버 뉴스	크롤링	개봉 7 일전 ~ 개봉일 전까지 기사 수
배우수상	연속	영화진흥위원회	크롤링	배우의 수상횟수
배우파워	연속	영화진흥위원회	크롤링	배우의 이전 영화 흥행여부
상영등급	범주	네이버 영화	크롤링	청소년관람불가/전체/12 세/15 세관람가
개봉 전 트윗 수	연속	트위터	크롤링	개봉 7 일전 ~ 개봉일 전까지 트윗 수

- \* 처음 데이터 셋에는 '장르'라는 변수를 넣었으나, 현재 영화 산업에서의 장르는 경계선이 모호하고 여러 장르가 합쳐진 종합적인 장르로 나타나기 때문에 최종 모델의 독립변수로는 사용하지 않았다.
- \* 또한 입소문 효과를 위해 네이버에서 영화 상영 전 조사하는 '기대지수', '비기대지수' 지표도 사용했으나, 영화마다의 모집단의 크기가 다르고 영화 특성에 대한 기대감과 비기대감이 아닌 외부적인 요소가 많이 영향을 주어 최종 모델의 독립변수로 사용하지 않았다.

### 3.2 한국 상업 영화 예측 분석서

#### 3.2.1 탐색적 데이터분석

##### 1) 수집 경로

영화진흥위원회API	네이버영화 API	웹 크롤링	트위터 크롤링
<ul style="list-style-type: none"> <li>• 영화명</li> <li>• 관객수</li> <li>• 배급사</li> <li>• 개봉일</li> <li>• Holiday</li> <li>• Season</li> <li>• 개봉 전의 스크린 수</li> <li>• 개봉 전의 상영횟수</li> </ul>	<ul style="list-style-type: none"> <li>• 감독명</li> <li>• 상영등급</li> </ul>	<ul style="list-style-type: none"> <li>• 감독 수상</li> <li>• 감독 파워</li> <li>• 배우 수상</li> <li>• 배우 파워</li> <li>• 개봉 전 기사 수</li> <li>• 개봉 영화 리뷰</li> </ul>	<ul style="list-style-type: none"> <li>• 개봉 전의 트위터 영화 언급 수</li> <li>• 개봉 전 트위터 리뷰</li> <li>• 개봉 후 트위터 리뷰</li> </ul>

##### 2) 데이터 전처리

###### 2.1) 5년치 한국 상업 영화 데이터 (스크린 수 100이상의 영화만 추출)

영화명	개봉일	관객수	스크린 수	상영횟수	배급사
명량	2014-07-30	17613682	1587	188611	씨제이이엔엠(주)
해적: 바다로 간 산적	2014-08-06	8666046	910	133350	롯데쇼핑(주)롯데엔터테인먼트
수상한 그녀	2014-01-22	8657982	1027	136975	씨제이이엔엠(주)
국제시장	2014-12-17	5345677	966	63603	씨제이이엔엠(주)

###### 2.2) 영화진흥위원회의 '영화정보 API'를 이용하여 '영화명', '개봉일'을 이용하여 해당 영화의 감독명, 장르, 영화코드 컬럼 추가

영화명	개봉일	관객수	스크린 수	상영횟수	배급사	영화코드	장르	감독
명량	2014-07-30	17613682	1587	188611	씨제이이엔엠(주)	20129370	사극	김한민
해적: 바다로 간 산적	2014-08-06	8666046	910	133350	롯데쇼핑(주)롯데엔터테인먼트	20136064	어드벤처	이석훈
수상한 그녀	2014-01-22	8657982	1027	136975	씨제이이엔엠(주)	20131102	드라마	황동혁
국제시장	2014-12-17	5345677	966	63603	씨제이이엔엠(주)	20137048	드라마	유제균

###### 2.3) 감독 및 배우 파워 컬럼 추출: '감독수상', '감독파워', '배우수상', '배우파워'

- '감독수상', '배우수상' 정의: 감독 및 배우가 해당 영화 상영 이전 수상을 얼마나 하였는지 파악하였다.
- '감독파워', '배우파워' 정의: 감독 및 배우가 해당 영화 상영 이전 개봉한 영화의 최종 관객수가 500만명 이상인 영화의 수 파악하였다. (감독을 보고 관람하는 경우를 생각하였다.)

- 영화 관계자의 자문에 의하면, 전국민의 10% 이상이 보았으면 흥행했다고 본다고 한다.
- 그림 코드의 예시는 감독 코드 추출 내용만 명시한다.

<그림 3> 감독 영화인코드 추출과정

1. 감독명과 영화명이 포함된 별도의 감독데이터를 생성
2. 감독 수상내역과 필모그래피를 파악하기 위해 영화진흥위원회의 영화인API를 통해 각 감독의 코드번호를 추출

감독	영화명	감독	영화명	peopleCd
김한민	명량	김한민	명량	10006204
이석훈	해적: 바다로 간 산적	이석훈	해적: 바다로 간 산적	10055965
황동혁	수상한 그녀	황동혁	수상한 그녀	10090050
윤제균	국제시장	윤제균	국제시장	10054492
윤종빈	군도: 민란의 시대	윤종빈	군도: 민란의 시대	10054495

- '감독수상' 추출과정: 추출한 개개인별의 peopleCD(영화인코드)를 이용하여 수상내역 크롤링 후 개수 컬럼화 하였다.
- '감독파워' 추출과정: 추출한 개개인별의 peopleCD 를 이용하여 감독의 필모그래피 안의 영화 전체와 각각의 관객수 및 참여역할 크롤링 후 이 중 '감독으로 참여한 영화들 중 관객수('공식통계' 기준)가 5 백만명이 넘는 영화 개수 컬럼화 하였다.

2.4) 네이버 영화 API를 통해[영화명+개봉일+국가]로 영화 검색 후 해당 영화의 배우, 장르, 관객 평점 수, 영화 별 링크 컬럼 생성

영화명	개봉일	관객수	...	장르	감독	기사 수	관객평점	링크
명량	2014-07-30	17613682	...	사극	김한민	56134	8.46	https://movie.naver.com/movie/bi/mi/basic.nhn?..
해적: 바다로 간 산적	2014-08-06	8666046	...	어드벤처	이석훈	22656	8.27	https://movie.naver.com/movie/bi/mi/basic.nhn?..
수상한 그녀	2014-01-22	8657982	...	드라마	황동혁	61025	9.00	https://movie.naver.com/movie/bi/mi/basic.nhn?..
국제시장	2014-12-17	5345677	...	드라마	유제균	53038	9.01	https://movie.naver.com/movie/bi/mi/basic.nhn?..

2.5) 네이버 뉴스에서 영화 개봉 일주일 전부터 개봉 하루 전까지의 '영화' + '영화명'이 포함된 기사 횟수 크롤링

2.6) 영화진흥위원회 크롤링

- 기존 데이터셋의 영화코드를 사용하여 개봉 전 스크린 수/상영횟수/관객수와 개봉 1일차~ 10일차 스크린 수/상영횟수/관객수 추출하였다.

영화명	날짜	스크린 수	상영횟수	관객수	누적 관객수
명량	개봉이전	37	157	22,500	22,500
명량	개봉1일	1,159	6,147	682,701	705,201
...	...	...	...	...	...
명량	개봉10일	1,278	7,026	690,123	8,672,578

2.7) 트위터 트윗 수 크롤링

- 영화명 전처리를 위해 영화명 중에서 '해적: 바다로 간 산적'이나 '신과 함께-죄와 벌'과 같이 특수문자가 들어 있는 경우 특수문자(/ - )는 제거 후 검색 특정 영화를 지칭하는 것이

아닌 다른 의미로도 사용할 수 있으니 '영화' 를 앞에 붙여준 후 검색 ex) ['영화' + '해적']이 들어간 트윗 수를 추출하였다.

- 기존의 데이터셋에서 개봉일을 분리하여 개봉일 기준으로 개봉 전 7일 ~ 개봉 후 7일 간의 트위터 트윗 횟수 추출하였다.

- 정제과정 예: 명량, 개봉일: 2014-07-30

개봉 전 [영화 명량]이 포함된 트윗 수		
Date	Frequency	Movie
2014-07-23	37	명량
2014-07-24	39	명량
2014-07-25	38	명량
2014-07-26	39	명량
2014-07-27	36	명량
2014-07-28	37	명량
2014-07-29	37	명량

개봉 후 [영화 명량]이 포함된 트윗 수		
Date	Frequency	Movie
2014-07-30	41	명량
2014-07-31	39	명량
2014-08-01	44	명량
2014-08-02	59	명량
2014-08-03	49	명량
2014-08-04	41	명량
2014-08-05	47	명량

각각의 트윗 개수를 '영화명'으로 합쳐서 '개봉 전 트윗 수' '개봉 후 트윗 수' 컬럼 생성

영화명	개봉전트윗	영화명	개봉후트윗
명량	263	명량	320
해적: 바다로 간 산적	230	해적: 바다로 간 산적	289
수상한 그녀	223	수상한 그녀	264
국제시장	150	국제시장	251



영화명	개봉일	관객수	배급사	장르	진사수	감독수상	감독국적	개봉이전 스크린수	개봉이후 스크린수	상영횟수	거대지수	비거대지수	holiday	season	배우수상	배우국적	개봉전트윗	개봉후트윗	
명량	20140730	17613682	씨제이이엔엠(주)	사극	56134	28	2	37.0	...	867437	15세, 큰형가	13639	1126	1	2	32	11	263	320
해적: 바다로 간 산적	20140806	8666046	롯데쇼핑(주)롯데엔터테인먼트	어드벤처	22656	9	4	27.0	...	297432	12세, 큰형가	4941	1537	1	2	19	4	230	289
수상한 그녀	20140122	8657982	씨제이이엔엠(주)	드라마	61925	27	2	19.0	...	175833	15세, 큰형가	0	0	1	4	21	11	223	264
국제시장	20141217	5345677	씨제이이엔엠(주)	드라마	53038	37	2	68.0	...	241578	12세, 큰형가	12784	2052	0	4	14	19	150	251
관두만관 의 시대	20140723	4774931	(주)노바스	사극	21502	24	0	46.0	...	468320	15세, 큰형가	14319	1262	1	2	23	16	199	278

### 3) 파생변수 생성

3.1) 'Holiday' 컬럼 생성: '개봉일' 데이터를 기반으로 방학의 여부를 파악하였다.

- 3월/4월/5월/6월/9월/10월/11월/12월: 일반 (0)
- 1월/2월/7월/8월: 방학 (1)

3.2) 'Season' 컬럼 생성: '개봉일' 데이터를 기반으로 계절 파악하였다.

- 12월/1월/2월: 겨울
- 3월/4월/5월: 봄
- 6월/7월/8월: 여름
- 9월/10월/11월: 가을

3.3) '배급사크기' 컬럼 추출: 5년 동안 배급한 영화 수에 따라 big/middle/small로 분리하였다.

big: 5년 동안 배급한 영화의 수가 30개 이상인 배급사

middle: 5년 동안 배급한 영화의 수가 10개 이상 30개 미만인 배급사

small: 5년 동안 배급한 영화의 수가 10개 미만인 배급사

<표 6> 배급사 크기 그룹별 구분

배급사	배급사 수	배급사 크기
씨제이이엔엠(주)	51	big
(주)넥스트엔터테인먼트월드(NEW)	34	big
롯데쇼핑(주)롯데엔터테인먼트	34	big

(주)쇼박스	32	big
메가박스중앙(주)플러스엠	19	middle
(주)리틀빅픽처스	15	middle
...	...	middle
오피스픽처스	7	small
...	...	small

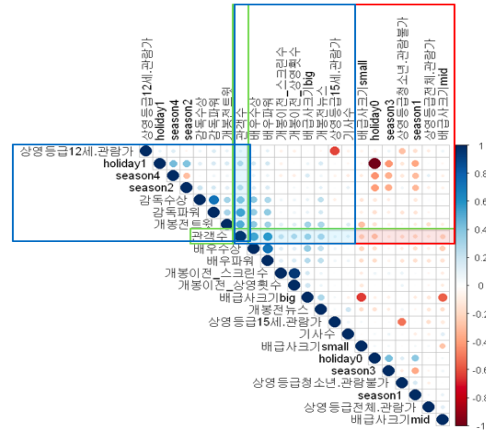
#### 4) 기술통계분석

##### 4.1) 관객수(종속변수)와 독립변수의 상관성 분석

종속변수와 독립변수들의 상관성을 그래프로 나타났을 때, 양의 상관성과 음의 상관성이 뚜렷이 나누어지는 것을 쉽게 볼 수 있었다.

- **양의 상관성:** '12 세관람가'/'방학기간개봉(holiday 1)', '여름(season2)', '감독수상'/'감독파워', '개봉전트윗', '배우수상'/'배우파워'/'개봉이전 스크린 수'/'개봉이전 상영횟수'/'배급사크기 big'/'15 세관람가'/'개봉전뉴스(기사 수)'
- **음의 상관성:** '배급사크기 small'/'방학 아닐 때'/'가을(season3)'/ '청소년관람불가'/'봄(season1)'/ '전체 관람가'/'배급사크기 mid'

<그림 4> 전체 변수들의 상관성



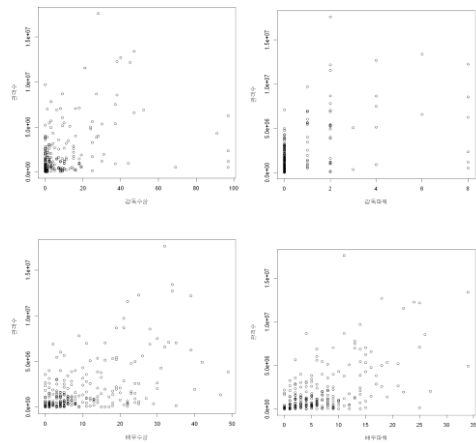
##### 4.2) 관객수 ~ 감독수상 / 관객수 ~ 감독파워 / 관객수 ~ 배우수상 / 관객수 ~ 배우파워

감독과 배우의 수상횟수가 많을수록 수상횟수가 단 한번도 없는 감독의 영화보다 흥행가능성이 더 높다.

감독파워와 배우파워가 높을수록 감독과 배우가 500 만명 이상의 관객수를 가진 영화 수가 많을수록 흥행 가능성이 있다.

- \* **감독수상:** 한 영화당 감독(들)의 누적 수상 횟수의 합
- \* **배우수상:** 한 영화당 출연한 주연배우(들)의 누적 수상 횟수의 합
- \* **감독파워:** 한 영화당 감독(들)의 과거 500만 이상의 관객수를 기록한 횟수의 합
- \* **배우파워:** 한 영화당 출연한 주연배우(들)의 과거 500만 이상의 관객수를 기록한 횟수의 합

<그림 5> 관객수와 감독 수상횟수, 감독파워, 배우 수상횟수, 감독파워의 상관관계 그래프

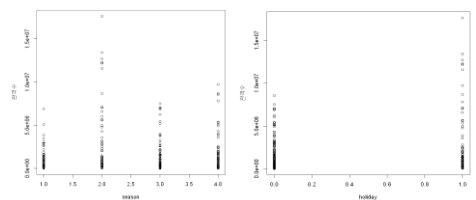


##### 4.3) 관객수 ~ 개봉시기

계절로는 여름(2)이 가장 높은 흥행가능성을 가지고 있다. 또한 방학에 개봉한 영화(1)가 그렇지 않은 영화(0)보다 더 높은 흥행가능성을 가지고 있다.

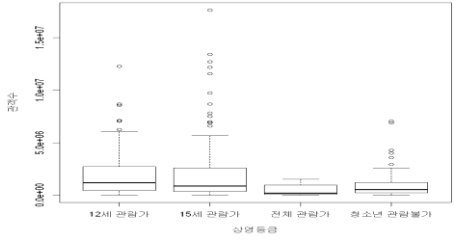
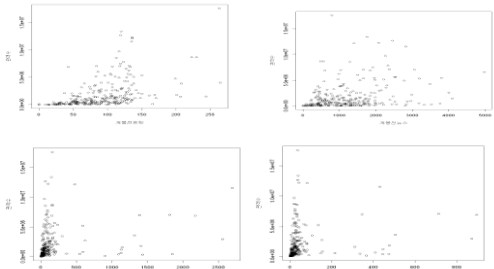
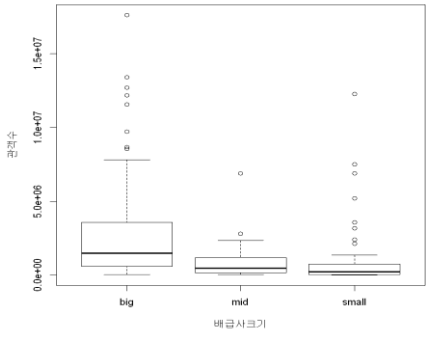
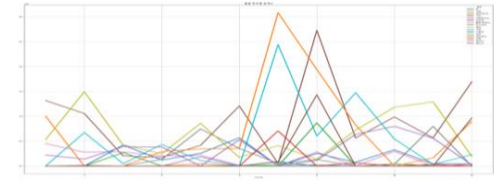
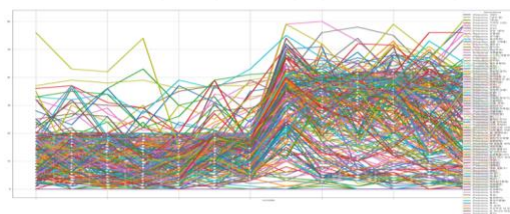
- \* **season:** 1-봄, 2-여름, 3-가을, 4-겨울
- \* **holiday:** 0 - 방학이 아닐 때 개봉, 1 - 방학에 개봉

<그림 6> 관객수와 계절, 방학의 유무의 상관관계



##### 4.4) 관객수 ~ 상영등급

<그림 7> 상영 등급별 평균 관객수와 이상치 boxplot

<p>'15 세관람가'의 박스 플롯을 보았을 때, 다른 상영 등급보다 좋은 흥행성적을 가지고 있는 아웃라이어들이 더 많다. 이를 바탕으로 '15 세 관람가'의 상영등급을 가진 영화가 가장 큰 흥행 가능성을 가지고 있다는 것을 알 수 있다. 두번째로는 '12 세 관람가'의 영화가 큰 흥행 가능성을 가지고 있다.</p> <p>'청소년 관람불가'와 '전체 관람가'의 영화는 흥행 가능성이 다른 두개의 등급보다 훨씬 낮은 것을 볼 수 있다. 특히, '전체 관람가'의 영화가 가장 낮은 흥행 가능성을 가지고 있는 것으로 해석된다.</p>	
<p><b>4.5) 관객수 ~ 마케팅요소</b></p> <p>개봉 전 트윗수가 3 개의 마케팅변수 중 관객수에 가장 큰 영향을 미친다.</p> <p>개봉 전 기사수(뉴스)는 개봉 전 트윗수와 관객수의 상관성보다는 일정하지 않지만 흥행가능성을 높이는 요소로 보인다.</p> <p>개봉 전 상영횟수를 시사회 횟수로 보았을 때, 시사회 횟수는 흥행결과가 큰 연관이 없는 것으로 나타난다.</p>	<p>&lt;그림 8&gt; 관객수와 개봉전 트윗, 개봉전 뉴스, 개봉전 상영횟수, 개봉 전 스크린 수의 상관관계</p> 
<p><b>4.6) 관객수 ~ 배급사크기</b></p> <p>큰 배급사일수록 평균적으로 중간 또는 작은 배급사보다 더 많은 관객수를 갖는다. 하지만, 작은 배급사도 오히려 중간크기의 배급사보다 이례적으로 500 만 관객수를 넘을 가능성이 낮지 않은 것으로 나타난다. 이것을 통해 큰 배급사와 작은 배급사의 콜라보로 더 큰 시너지효과를 낼 수 있을 것이라 생각할 수 있다.</p> <p>* big: 5년간100개 이상의 영화 배급 * mid: 5년간 30개 이상의 영화 배급 * small: 5년간 30미만의 영화 배급</p>	<p>&lt;그림 9&gt; 관객수와 배급사 크기의 상관관계</p> 
<p><b>4.7) 월별 기준 - 장르별 개봉 영화 누적 관객 수</b></p> <ul style="list-style-type: none"> <li>- 액션물과 사극물이 7 월에 가장 인기가 많다.</li> <li>- 8 월에는 판타지가 많은 관객을 이끌어냈다.</li> <li>- 범죄물은 2 월에 가장 많은 사람들이 보았고,</li> <li>- 공포물은 8 월에 많은 사람들이 보았다.</li> </ul>	<p>&lt;그림 10&gt; 월별 장르별 개봉 영화 누적 관객 수</p> 
<p><b>4.8) 개봉 전/후 기준 - 영화별 트위터 수</b></p> <ul style="list-style-type: none"> <li>- 영화의 언급 수가 트위터에서 확실히 개봉전보다 개봉후의 트윗 수가 더 많은 것을 명확히 보여준다.</li> </ul>	<p>&lt;그림 11&gt; 영화별 개봉 전 후의 트윗 언급 수</p> 

### 3.3 '한국 상업 영화 예측 모델' 상세 설명

#### 3.3.1 회귀모델(Linear regression 분석)

##### 1) 데이터 준비 및 변수 선택

데이터는 2014년 1월 1일~ 2018년 12월 31일의 기간동안 개봉한 한국 상업 영화 260편의 기존 데이터와 2019년 1월 1일 이후에 개봉한 새 영화 7편 데이터를 준비했다. 모델 생성시, 변수는 두 가지의 경우로 선택했다. 첫번째 방법으로 '개봉 이전'과 관련된 22개 변수를 모두 넣은 회귀모델을 생성하고, 두번째 방법으로는 변수를 요인분석으로 축소한 후 회귀모델을 생성을 시도한다.

##### 2) 스케일링 및 더미변수화

데이터셋에서 독립변수 중 범주형 데이터(상영등급, 배급사크기, 계절, 방학)를 제외한 나머지 변수 컬럼을 연속형으로 지정했다. 그리고 독립변수 중 범주형 변수 컬럼을 일괄 팩터화했다. 그리고 독립변수 중 연속형 변수를 일괄 스케일링 했다. 더미변수 생성 알고리즘 적용시 제외되는 종속변수를 나중에 재결합해주기 위해서 종속변수를 별도로 준비한 후, 더미변수생성 원핫인코딩(One-Hot Encoding)함수인 `dummyVars` 에 전체데이터를 투입했다. 더미변수화를 진행한 후, 더미변수화 과정에서 제외되었던 종속변수를 재결합했다.

##### 3) 회귀모델 생성

최종 관객 수에 영향을 미치는 요인들의 영향력을 파악하기 위해 관객 수를 종속변수로 한 회귀분석을 실시했다. 기본선형회귀모델, 2차상호작용 선형회귀모델, 그리고 단계별 변수선택 회귀모델로 3가지 분석 모델을 설계했다. 첫번째는 독립변수를 단독으로 선형결합한 모델이며, 두번째는 독립변수 뿐만 아니라 상호작용으로 결합된 통합 버전의 모델이다. 그리고 마지막 세번째는 두번째 모델에서 필요 없는 것을 제외한 모델로 예측력에 무의미한 영향력을 갖고 있는 변수를 제거한 모델이다.

<표 7> 관객 수에 영향을 미치는 요인

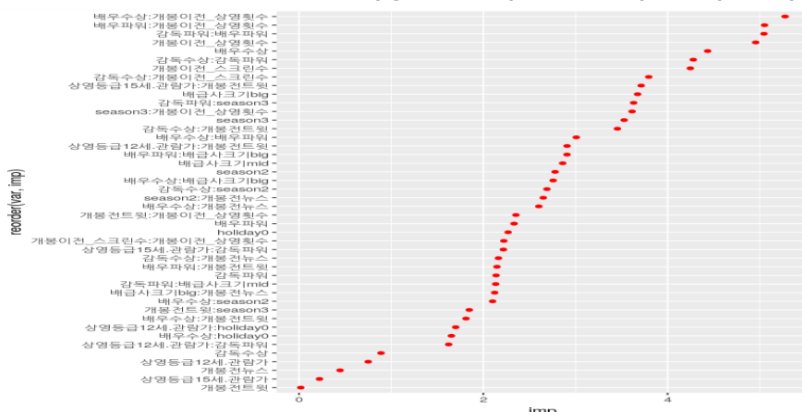
독립변수	비표준화 계수		유의확률
	B	표준오차	
(상수)	1436375	517591	0.00610
12 세관람가	48721	419332	0.90763
15 세관람가	339542	357202	0.34309
전체 관람가	871584	903667	0.33608
감독수상	-191374	225966	0.39816
감독파워	1135635	205711	1.16e-07
배우수상	217077	201989	0.28394

배우파워	556383	191447	0.00412
개봉트릿	644632	156479	5.77e-05
개봉이전 스크린 수	-556638	409899	0.17616
개봉이전 상영횟수	936244	405471	0.02207
배급사 크기(대)	500026	347478	0.15187
배급사 크기(중)	-8533	401426	0.98306
개봉 전 뉴스	9285	150454	0.95086

Residual standard error: 1806000, R-squared: 0.6069, Adjusted R-squared: 0.5699, F-statistic: 16.43, p-value: < 2.2e-16

분석결과, 몇 개의 요인을 제외하고 대부분의 요인들이 개별 영화의 총 관객 수에 영향을 미치고 있는 것으로 나타났다. 분석에 투입된 독립변수들 중에서 감독파워가 상대적으로 가장 큰 영향력( $\beta=5.521$ ,  $p=1.16e-07$ )을 미치고 있는 것으로 나타났다. 이외에도 개봉트릿( $\beta=4.120$ ,  $p=5.77e-05$ ), 배우파워( $\beta=2.906$ ,  $p=0.0412$ ), 개봉이전 상영횟수( $\beta=2.309$ ,  $p=0.02207$ ) 등이 최종 관객 수에 유의미한 영향을 미치고 있었다. 흥미로운 결과는 제작 시 결정되는 감독과 배우 관련 변수인 수상횟수와 파워가 높을수록 총 관객 수가 많아진다는 것이다. 또한, 개봉이전의 트릿과 상영횟수를 고려했을 때, 관객 수가 개봉 초반에 관객의 입소문, 개봉 전 시사회의 횟수 등을 중심으로 빠르게 결정된다는 것을 의미한다. 기본선형회귀모델은 낮은 R-squared 값을 가지므로, 상대적으로 설명력이 높은 2 차상호작용 회귀모델(R-squared: 0.81)을 살펴보기로 한다.

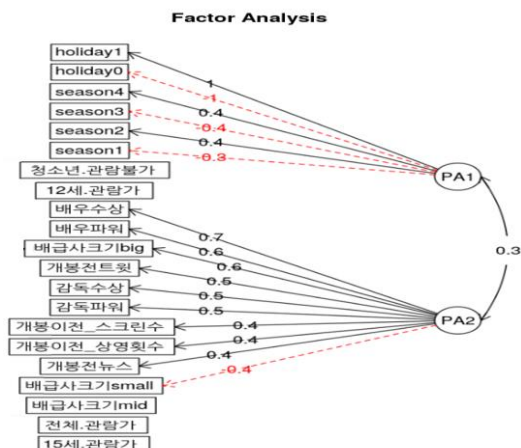
<그림 12> 2 차상호작용 회귀모델에 사용한 투입변수 중요도 그래프



<변수 중요도 Top 7순위>

1. 배우수상:개봉이전\_상영횟수
2. 배우파워:개봉이전\_상영횟수
3. 감독파워:배우파워
4. 개봉이전 상영횟수
5. 배우수상
6. 개봉이전 스크린 수
7. 감독수상:개봉이전 스크린 수

<그림 12>은 변수를 두 개씩 상호작용하는 효과를 살펴볼 수 있는 회귀모델에 사용한 투입변수의 중요도를 살펴본 그림이다. 중요도가 상대적으로 높았던 조합은 배우수상과 개봉이전 상영횟수, 배우파워와 개봉이전 상영횟수, 감독파워와 배우파워, 개봉이전 상영횟수, 배우수상, 개봉이전 스크린 수, 감독수상과 개봉이전 스크린 수, 상영등급 15세 관람가와 개봉 트릿, 배급사크기(대) 순서였다. 변수 간의 상호작용을 봄으로써 이를 추후 활용방안에도 접목시킬 수 있으리라 기대한다.



#### 4) 요인분석 실시

<그림 13> 사각회전방식 요인분석 다이어그램

요인분석은 전형적인 데이터 축소기법으로 여러 개의 변수데이터를 활용해서 공통적인 새로운 변수를 만들 수 있다. 서로 상관관계가 있는 여러 개의



변수들을 선형결합으로 만들어 변수가 가진 자료의 변동을 최대한 보존하는 축소된 개수의 변수로 변환시키는 방법이다. 현재 데이터로는 변수의 개수가 많기에 요인분석으로 차원을 축소한 후 회귀분석을 시도해보기로 한다.

R 의 Psych 패키지를 활용하여 잠재적 요인 개수가 2 개인 것을 확인했다. 잠재적 요인 개수를 기반으로 무회전방식, 직교(직각)회전방식, 사각회전방식을 시도하였고, 요인 간에 상관관계가 있는 경우에 실시하는 경우에 하는 방식인 사각회전방식을 채택했다.

선정 결과를 바탕으로 묶인 두개의 요인은 <그림 13>과 같이 방학, 계절 변수가 묶인 요인 1(PA1)과 배우수상, 배우파워, 배급사크기(대), 개봉전트윗, 감독수상, 감독파워, 개봉이전 스크린 수, 개봉이전 상영횟수가 묶인 요인 2(PA2)이다. 아이겐벨류(eigenvalues; 고유 값)<sup>1</sup>이 요인 1 의 경우 0.16, 요인 2 의 경우 0.097 로 1 보다 작다. 즉, 결합된 변수들의 데이터 변화량을 대표할 수 있는 1.0 이상의 충분한 표준화된 분산량을 가지고 있지 않다. 이는 결합되는 변수들의 공통적 변화량 보다도 결합변수들의 고유성이 강해 하나의 대표성을 갖는 요인이 되지 못하고 있음을 의미한다.

### 5) 분석 결과 비교

두가지 요인을 바탕으로 회귀모델을 생성했을 때, 표 3 과 같이 모델의 설명력이 기존 전체 변수를 넣은 회귀모델보다 떨어지는 것을 알 수 있다. 스케일링 및 더미변수화 등과 같은 전처리 후, 기본패키지의 sample 함수를 이용해 데이터를 학습용 데이터(75%)와 검증용 데이터(25%)로 분리하였다. 훈련 모델 자체 및 검증했을 때의 설명력은 요인분석 하지 않고 전체변수를 회귀모델에 넣었을 때가 더 높았다.

<표 8> 회귀모델 분석결과표

	'개봉이전' 관련 22개 변수		요인분석으로 축소된 변수	
	기본선형회귀모델	2 차상호작용 회귀모델	기본선형회귀모델	2 차상호작용 회귀모델
RMSE	1753051	1194000	1744829	1797000
R squared (훈련 모델)	0.56	0.81	0.49	0.57

이처럼, 회귀 수치예측을 실시했으나 모델의 설명력 및 테스트 예측력이 예상만큼 높게 나오지 않았기에, 분류예측 등과 같은 다른 분석기법을 시도할 필요가 있다. 로지스틱 회귀, 다중회귀, 다항회귀(2 차, 3 차)를 시도하여 회귀모델을 강화해 나가는 것을 추후 개선 사항으로 고려하고 있다. 그리고 기본회귀모델 이외의 다양한 분석 기법을 시도하여 모델을 보완해 나가야 할 것이다.

<표 9> 2 차상호작용 회귀모델을 통한 새 영화 관객수 예측

영화명	2 차상호작용회귀모델 관객수 예측	실제 관객수 (2019.02.21 기준)	오차
말모이	3,593,213	2,858,130	-735,083
스윙키즈	2,718,602	1,471,248	-1,247,360
마약왕	3,255,880	1,864,077	-1,391,803
트와이스랜드	-8,214	18,254	26,468
리벤저	275,707	3,232	-272,475

1

극한직업	3,889,498	14,856,794	10,967,296
뽕만	3,300,221	1,825,750	-1,474,471

### 3.3.2 다양한 분류 예측 분석과 검증

기존의 회귀분석의 아쉬웠던 정확도 측면을 높이면서 실무에서 배급사들의 기준에 맞추어 관객수를 분류 후 그 기준으로 예측을 시도해보았다.

2014년부터 2018년까지 5년간 한국영화진흥위원회에서 제공하는 한국상업영화와 최종 관객수와 위에서 찾아낸 독립변수들을 사용하여 관객수 예측을 다양한 분류 분석기법을 통해 검증하였다.

<표 10> 관객 수에 따른 그룹화와 각 그룹의 빈도 수

분류 기준	설명	빈도수	오버샘플링	콤비네이션 샘플링
ABOVE10m	관객수 1000 만 이상의 영화	6	6	102
ABOVE9m	관객수 900 만 이상의 영화	1	6	102
ABOVE8m	관객수 800 만 이상의 영화	3	6	102
ABOVE7m	관객수 700 만 이상의 영화	10	10	101
ABOVE6m	관객수 600 만 이상의 영화	6	6	102
ABOVE5m	관객수 500 만 이상의 영화	9	9	102
ABOVE4m	관객수 400 만 이상의 영화	8	8	102
ABOVE3.5m	관객수 350 만 이상의 영화	6	6	101
ABOVE3m	관객수 300 만 이상의 영화	5	10	102
ABOVE2.5m	관객수 250 만 이상의 영화	13	13	102
ABOVE2m	관객수 200 만 이상의 영화	11	11	101
ABOVE1.5m	관객수 150 만 이상의 영화	17	17	102
ABOVE1m	관객수 100 만 이상의 영화	31	31	101
ABOVE500K	관객수 50 만 이상의 영화	37	37	101
UNDER500K	관객수 50 만 이하의 영화	102	102	99

<표 10>을 살펴보면, 기본적으로 100 만단위로 분류를 하였고, 그 중에서 빈도가 많은 백만 명에서 300 만명 사이의 관객수는 50 만 단위로 한번 더 세부적으로 분류하였다. 100 만 미만의 영화의 빈도수 역시 많지만, 실무자의 이야기를 들어본 결과 평균 200 만 이하로 예측되는 영화에는 주의 깊게 살펴보지 않는다고 한다.

#### 1) 범주형 독립변수 더미변수로 변환

피쳐 셋 중 텍스트 형식으로 되어 있던 '배급사크기'와 '상영등급'을 더미변수로 변환하였다.

- 배급사크기: '배급사크기\_mid'/'배급사크기\_small'/'배급사크기\_big'
- 상영등급: '상영등급\_12 세관람가'/'상영등급\_15 세관람가'/'상영등급\_전체 관람가'/'상영등급\_청소년관람불가'

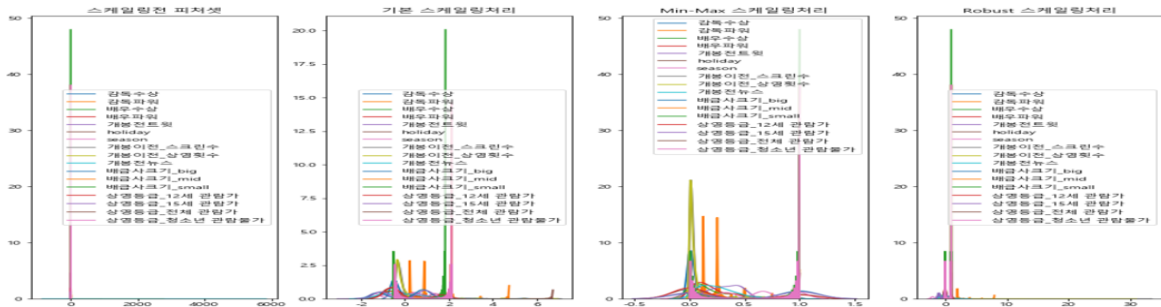
#### 2) 독립변수 스케일링

스케일링은 자료 집합에 적용되는 전처리 과정으로 모든 자료에 선형 변환을 적용하여 전체 자료의 분포를 평균 0, 분산 1이 되도록 만드는 과정이다. 스케일링은 자료의 오버플로우나 언더플로우를 방지하고 독립 변수의 공분산 행렬의 조건수를 감소시켜 최적화 과정에서의

안정성 및 수렴 속도를 향상시킨다. 여기서 우리는 Scikit-Learn에서 제공하는 스케일링 기법 중 3가지 기법(기본/Min-Max/Robust 스케일링)을 사용해 스케일링을 진행하였다.

- **scale(X)**: 기본 스케일. 평균과 표준편차 사용
- **robust\_scale(X)**: 중앙값(median)과 IQR(interquartile range) 사용. 아웃라이어의 영향을 최소화
- **minmax\_scale(X)**: 최대/최소값이 각각 1, 0이 되도록 스케일링

<그림 14> 독립변수 Feature set 스케일링 결과



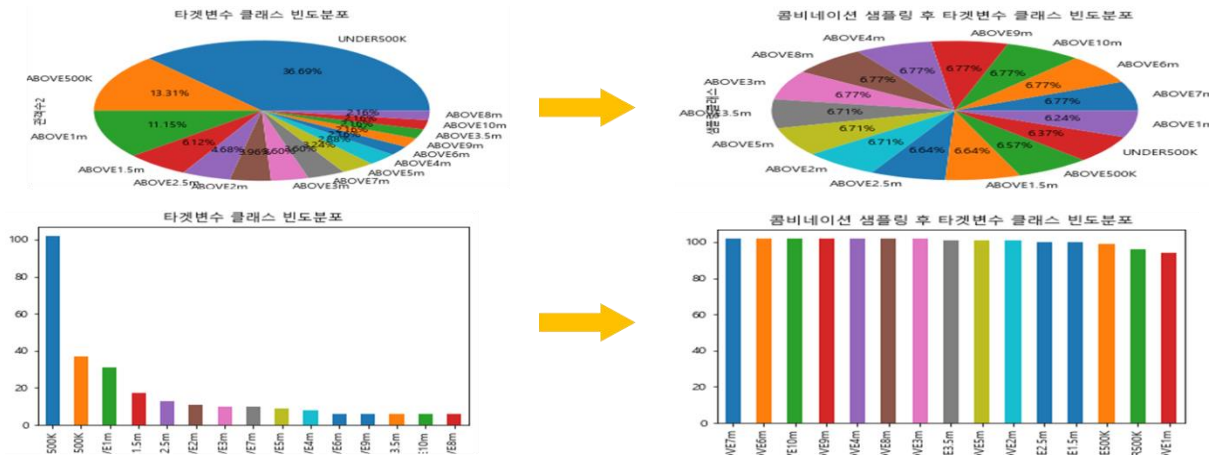
<그림 14> 을 살펴보면, Robust 스케일링 처리 결과가 가장 피쳐셋의 분포를 전체적으로 통일시켜주었기에 Robust 스케일링으로 처리한 피쳐셋을 모델링에 사용할 것이다.

### 3) 클래스 불균형(Imbalanced datasets) 전처리

분류 예측을 하기 위해서는 사전 전처리 작업이 필요하다. <표 10>에서 보는 바와 같이, 분류 단위 별 빈도수가 제각기 다를 수 있다. 타겟 변수의 클래스가 불균형인 상태에서는 특정 클래스에 대한 학습을 더 강하게 하게 되어 각 클래스 별 정확도, 즉 민감도와 특이도가 균형을 이루지 못하게 된다. 즉, 클래스 비율이 너무 차이가 나면 우세한 클래스를 택하는 모형의 정확도가 높아져 성능판별이 어려워진다. 또한 정확도가 높아도 데이터 개수가 적은 클래스의 재현율(recall-rate)이 급격히 작아지는 현상이 발생한다. 이를 방지하기 위해, 우선적으로 적은 클래스를 오버샘플링(Over-sampling) 과정이 필요하다. 오버샘플링이란, 다수(majority) 클래스의 크기를 기준으로 소수(minority) 클래스 규모를 확대하는 방법이다. <표 10>에서 살펴보면, 분류 클래스 중 'ABOVE9m', 'ABOVE8m', 'ABOVE3m'이 오버샘플링이 필요하다. 이들의 클래스의 최소한의 요구 개수인 6개 이상으로 맞춰주었다.

그 후 아직 완벽하게 해소되지 않은 클래스의 불균형을 해결하기 위해 '콤비네이션 샘플링(Combining Over-and Under-Sampling)'을 진행하였다. 이 샘플링 방법은 언더샘플링과 오버샘플링의 방법을 적절히 결합해 소수 클래스 데이터는 증폭시키고 다수 클래스의 데이터는 제거해 클래스 비율을 조정하는 샘플링이다. 해당 샘플링들의 결과로 260개의 레이블이 1522개의 레이블로 증가하였다. 샘플링을 통해 균형이 맞게 변화한 클래스의 비율 분포는 결과는 아래의 그림을 통해 알 수 있다.

<그림 15> SMOTETomek 콤비네이션 샘플링 실행 후 클래스 비율 변화



#### 4) 분류 모델

##### 4-1) 사용 분류 모형 설명

Multiclass classification 은 두 개 이상의 클래스를 가진 분류 작업을 의미한다. 우리가 모델링에 사용한 알고리즘 패키지인 scikit-learn 속의 분류 알고리즘은 다중 클래스 분류를 지원하고 있다. 해당 scikit-learn 패키지 속의 분류 알고리즘 중 사용할 알고리즘은 Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Tree, K-NN, Neural-Network 이다.

##### 4-2) 분류 모형 평가

위와 같은 전처리 후, K-Fold 로 Cross validation Test 를 진행하였고, 새 영화 관객수 예측을 위해 Train\_test\_Split 명령을 통해 데이터를 학습용 데이터 (75%)와 검증용 데이터(25%)로 분리하였다. 하단의 <표 11>는 다양한 분류 기법 알고리즘을 사용하여 학습용 데이터를 학습시키고, 학습시킨 모델을 사용해 검증용 데이터로 테스트한 결과이다.

<표 11> 다양한 분류 기법을 사용한 관객수 예측 결과 비교

분류 기법	Logistic Regression	Decision Tree	Random Forest		Gradient Boosting Tree		K-NN		Neural Network
			n_estimators	n_estimators	n_estimators	n_estimators	n_neighbors	n_neighbors	
손실함수					Logistic regression	AdaBoost			
파라미터	penalty: l1, C: 1		n_estimators : 100, n_jobs : 100	n_estimators : 100, n_jobs : 1	n_estimators: 100, learning_rate: 0.1	n_estimators: 100, learning_rate: 0.3	n_neighbors = 1	n_neighbors = 3	adam hidden_layer_sizes=(100,) max_iter=2000
10-Fold Accuracy	0.632	0.851	0.942	0.93	0.879	0.877	0.940	0.899	0.956
아래는 Train(75%)-Test(25%) Split을 통해 도출한 검증 결과									
Accuracy	0.654	0.853	0.924	0.921	0.879	0.879	0.924	0.866	0.929
Precision	0.636	0.846	0.925	0.922	0.88	0.878	0.92	0.861	0.926
Recall	0.646	0.848	0.926	0.923	0.882	0.879	0.921	0.864	0.93

<b>F1</b>	0.621	0.844	0.923	0.922	0.878	0.877	0.915	0.851	0.927
-----------	-------	-------	-------	-------	-------	-------	-------	-------	-------

<표 11>에서의 Logistic Regression, Random Forest, Gradient Boosting Tree, K-NN, ANN의 분류 모형들은 각각 모형마다의 파라미터들을 조정하여 정확도가 가장 높게 나왔었을 때의 값들만 작성한 것이다. 또한 cross-validation 을 통한 모델 검증과 예측을 위한 train-test 셋의 정확도를 각각 별도로 표시하였다. 그 결과, 모든 독립변수들을 투입하여 대체적으로 모든 모형에서 높은 정확도를 나타내었다. Random Forest 의 정확도는 94.2%, K-NN(n-neighbors=1)의 정확도는 94.0%, ANN(인공신경망)의 정확도는 95.6%로 ANN 모형의 성과가 가장 좋았다.

<표 12> 정확도가 높은 분류 모형 Top3 결과

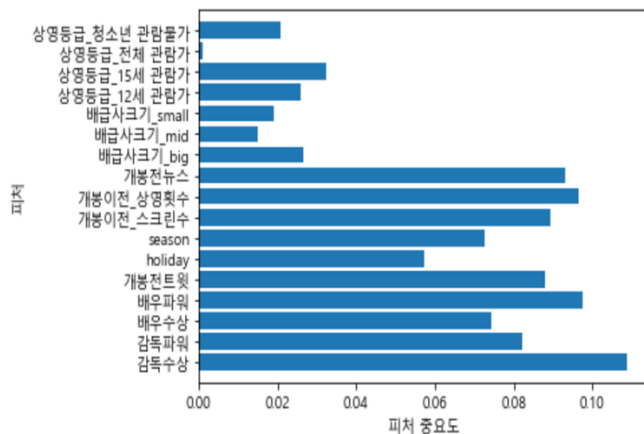
ANN	Random Forest	K-NN
<pre> confusion matrix [[31 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  [0 20 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  [0 0 29 0 0 0 0 0 0 0 0 0 0 0 0 0]  [0 0 0 15 0 0 0 0 0 0 0 0 0 0 0 0]  [0 0 0 0 32 0 0 0 0 0 0 0 0 0 0 0]  [0 0 0 0 0 29 0 0 0 0 0 0 0 0 0 0]  [0 0 0 0 0 0 27 0 0 0 0 0 0 0 0 0]  [0 0 0 0 0 0 0 22 0 0 0 0 0 0 0 0]  [0 0 0 0 0 0 0 0 15 0 0 0 0 0 0 0]  [0 0 0 0 0 0 0 0 0 15 0 0 0 0 0 0]  [0 0 0 0 0 0 0 0 0 0 27 2 0 0 0 1]  [0 0 0 0 1 0 0 2 0 2 0 26 0 1 0]  [0 0 0 0 0 0 2 0 0 0 0 19 0 3]  [0 0 0 0 0 0 0 0 1 1 1 0 2 21 0]  [0 0 0 0 0 0 0 0 0 0 0 1 4 3 14]]  accuracy : 0.929 precision : 0.926 precision : [1. 1. 1. 1. 0.97 1. 1. 0.846  0.938 0.9 0.964 0.897  0.76 0.84 0.778] recall : 0.93 recall : [1. 1. 1. 1. 1. 1. 1. 1. 1.  1. 0.9 0.812  0.792 0.808 0.636] F1 : 0.927 F1 : [1. 1. 1. 1. 0.985 1. 1. 0.917 0.968  0.947 0.931 0.852  0.776 0.824 0.7 ] </pre>	<pre> ----- n_estimators : 100, n_jobs : 100 ----- confusion matrix [[31 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  [0 20 0 0 0 0 0 0 0 0 0 0 0 0 0]  [0 0 29 0 0 0 0 0 0 0 0 0 0 0 0]  [0 0 0 15 0 0 0 0 0 0 0 0 0 0 0]  [1 0 0 0 30 1 0 0 0 0 0 0 0 0 0]  [0 0 0 0 29 0 0 0 0 0 0 0 0 0 0]  [0 0 0 0 0 27 0 0 0 0 0 0 0 0 0]  [0 0 0 0 0 0 22 0 0 0 0 0 0 0 0]  [0 0 0 0 0 0 0 15 0 0 0 0 0 0 0]  [0 0 0 0 0 0 0 0 27 2 0 0 0 0 0]  [0 0 0 0 1 0 0 0 0 2 0 26 0 1 0]  [0 0 0 0 0 0 1 1 0 2 0 27 0 0 1]  [0 0 0 0 0 0 2 0 0 2 0 19 0 1]  [0 0 0 0 0 0 0 0 1 0 1 0 1 18 2]  [0 0 0 0 0 0 0 0 0 1 0 2 1 2 16]]  accuracy : 0.924 precision : 0.925 precision : [0.969 1. 1. 0.938 0.968 0.906 0.9 0.917  1. 0.844 0.964 0.871  0.905 0.9 0.8 ] recall : 0.926 recall : [1. 1. 1. 1. 0.938 1. 1. 1. 1.  1. 0.9 0.844  0.792 0.892 0.727] F1 : 0.923 F1 : [0.984 1. 1. 0.968 0.952 0.951 0.947 0.957 1.  0.915 0.931 0.857  0.844 0.783 0.762] </pre>	<pre> ----- knn : 1 ----- confusion matrix [[31 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  [0 20 0 0 0 0 0 0 0 0 0 0 0 0 0]  [0 0 29 0 0 0 0 0 0 0 0 0 0 0 0]  [0 0 0 15 0 0 0 0 0 0 0 0 0 0 0]  [0 0 0 0 32 0 0 0 0 0 0 0 0 0 0]  [0 0 0 0 0 29 0 0 0 0 0 0 0 0 0]  [0 0 0 0 0 0 26 0 0 0 0 0 0 0 0]  [0 0 0 0 0 0 0 21 0 0 0 1 0 0 0]  [0 0 0 0 0 0 0 0 15 0 0 0 0 0 0]  [0 0 0 0 0 0 0 0 1 26 0 0 0 0 0]  [0 0 0 0 0 0 0 0 0 1 0 26 2 1 0]  [0 0 0 0 0 1 1 0 0 0 0 29 0 1 0]  [0 0 0 0 0 0 0 0 1 0 2 0 21 0]  [0 0 0 0 0 0 0 0 0 1 0 0 23 2]  [0 0 0 0 0 0 1 0 2 0 0 0 2 5 3 9]]  accuracy : 0.924 precision : 0.92 precision : [1. 1. 1. 0.938 0.97 0.935 1. 0.875  0.892 0.897 1. 0.853  0.778 0.852 0.818] recall : 0.921 recall : [1. 1. 1. 1. 1. 1. 0.963 0.955 1.  0.983 0.867 0.906  0.875 0.885 0.409] F1 : 0.915 F1 : [1. 1. 1. 0.968 0.985 0.967 0.961 0.913 0.938  0.929 0.929 0.879  0.824 0.868 0.545] </pre>

<표 12>에서와 같이 정확도가 높은 모델들을 토대로 기존의 데이터 셋에 포함되지 않은 새 영화 7 가지의 관객수를 예측하였다. 그 결과는 2.5 절 '새 영화 데이터 예측'에서 설명한다.

<참고> 중요도 변수

<표 11>에서 보는 바와 같이 10-fold 로 살펴본 정확도는 ANN > Random Forest > K-NN 순으로 모델링의 정확도가 높게 나왔음을 알 수 있다. 모델을 구성하다 보니 어떠한 변수가 중요도를 갖는지에 대한 궁금증이 생겼다. 구성한 모델 중 가장 높은 정확도를 나타내고 Black-Box 모델이 아닌 Random Forest 의 모델에서의 중요 변수를 살펴보았다

<그림 16> Random Forest 의 Feature 중요도



<Feature 중요도 Top 7 순위>

1. 감독수상
2. 배우파워
3. 개봉이전 상영횟수
4. 개봉전뉴스
5. 개봉이전 스크린 수
6. 개봉전트윗
7. 감독파워

'Random Forest' 모델링 기법에서 중요 변수로 사용한 변수들로, 1,2 번은 영화 속성 변수이고 3~6 번까지는 마케팅 변수이다.

### 5) 새 영화 관객수 예측

새 영화 5 편을 추가하여 최종 관객수 예측을 시도하였다. 정확도 값이 높았던 ANN 모델과 Random Forest 모델에 새 영화 데이터를 투입하여 예측한 결과는 <표 13>과 같다.

<표 13> 정확도가 높은 분류 모형 2 가지를 사용하여 새 영화 관객수 예측

영화명	ANN 모델 관객수 예측	Random Forest 모델 관객수 예측	실제 관객수 (2019.02.21 기준)	실제 관객수
말모이	ABOVE1.5M	ABOVE2M	2,858,130	ABOVE2.5M
스윙키즈	ABOVE1.5M	ABOVE1.5M	1,471,248	ABOVE1M
마약왕	ABOVE1M	ABOVE2.5M	1,864,077	ABOVE1.5M
트와이스랜드	UNDER500K	UNDER500K	18,254	UNDER500K
리벤저	UNDER500K	UNDER500K	3,232	UNDER500K
극한직업	ABOVE3.5M	ABOVE3.5M	14,856,794	ABOVE10M
뽕반	ABOVE1.5M	ABOVE1.5M	1,825,750	ABOVE1.5M

단순히 관객 수 예측으로 끝낼 수 있었지만, 실제 최근 영화 관객수 예측을 잘 맞는지에 대해 궁금증이 생겼고, 2019 년 02 월 21 일을 기준으로 그날까지의 실제 관객수를 대입해 비교해보았다. <표 13>는 모델의 예측 값과 실제 관객수 값을 명시하고 있다.

전체적으로 살펴보면, 인공지능망 모델의 새 영화 예측 값과 Random Forest 모델의 예측 값은 서로 유사한 결과가 나왔다. 하나하나 자세히 살펴보면, 오차 범위가 특정 하나의 영화(극한직업)를 제외하고 모두 유사하거나 동일한 결과 값이 나왔다.

영화 '뽕반', '트와이스랜드', '리벤저' 이 세가지 영화는 모두 실제 관객 수와 동일한 예측 값이 도출되었다. '영화 '스윙키즈'의 경우에는 분류 단위로는 50 만명이라는 차이가 있지만, 실제 관객 수 147 만명과 비교했을 때 매우 유사한 수치임을 확인할 수 있다. 영화 '마약왕'과 '말모이'의 경우도 오차 범위가 크지 않음을 알 수 있다.

그렇다면, '극한직업'의 예측이 실제 관객수에 비해 수치가 낮게 나온 이유는 무엇일까? 우리는 우리가 제시하지 않은 외부 환경 즉 다른 외생변수의 작용이 클 것으로 보았다. 영화 평론가나 전문가들 역시 외부적 환경 요인, 예를 들면 연초라는 시기적 특성과 경쟁 작품의 부재 등 다양한 외생요인들이 '극한직업'을 예상치 못한 흥행 성공의 길로 이끌었다고 보았다.

우리는 위와 언급한 외생 요인 이외의 다른 외부적 환경요인들이 존재하는지 파악하기 위해 이어지는 3 절에서 '네이버 영화 리뷰'와 '트위터 영화 리뷰'를 분석해보았다.



### 3.3.3 SNS 리뷰 분석

위에서 제시한 모델의 피쳐 셋 이외의 외부적 환경 요소를 찾기 위함과 흥행 요소를 파악하기 위해 '네이버 영화 리뷰'와 '트위터 영화 리뷰'를 크롤링하여 관객들의 반응을 워드 클라우드 기법으로 살펴보았다.

- 트위터 크롤링 : 개봉일 ~ 개봉 후 7일까지의 '영화명'을 포함한 메시지를 트윗 한 리뷰만을 크롤링
- 네이버 영화 리뷰 크롤링 : 개봉일 이후 네이버 영화 리뷰 전체 크롤링.

트위터와 네이버의 리뷰의 기한을 다르게 한 이유는 트위터 리뷰를 통해 개봉 1 주차의 실시간 반응들을 통해 입소문이 어떤 방식과 내용으로 퍼지는지 파악하기 위함이고, 네이버 영화의 전체 리뷰를 통해 영화 자체를 파악하기 위함이다.

#### 1) 예측보다 높은 관객수 동원 영화 '극한직업' 리뷰 분석

 <p>&lt;그림 17&gt; 영화 '극한직업'의 트위터 리뷰 분석</p>	<ul style="list-style-type: none"> <li>* 긍정적인 단어가 많이 나타남. Ex) 웃기, 웃음, 기대, 최고, 추천, 코믹, 풀잼 등의 단어</li> </ul>
 <p>&lt;그림 18&gt; 영화 '극한직업'의 네이버 리뷰 분석&gt;</p>	<ul style="list-style-type: none"> <li>* 캐릭터명, 영화 키워드와 같은 영화 자체 속성 단어가 많이 나타남. Ex) 형사, 테드, 무배, 치킨, 통닭, 수원, 갈비 등에 대한 단어.</li> <li>* 경쟁 작 "뺑반" 1 가지</li> </ul>

#### 2) 예측보다 낮은 관객수 동원 영화 '마약왕' 리뷰 분석

 <p>&lt;그림 19&gt; 영화 '마약왕'의 트위터 리뷰 분석</p>	<ul style="list-style-type: none"> <li>* 부정적인 단어가 많이 나타남. Ex) 무슨 소리, 굳이, 노잼, 최악, 쓰레기, 별로</li> <li>* 영화 자체 속성 단어보다는 영화 이외의 단어가 비교적 많다. Ex) 연출, 편집, 무대인사, 롯데시네마, 월드타워, 박스오피스</li> </ul>
------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------





\* 경쟁 영화명이 많이 나타남.  
 Ex) 보헤미안, 스파이더맨,  
 스윙키즈, 아쿠아 맨

### 3) 워드 클라우드 분석 결과

위와 같은 영화 리뷰를 분석하면서 경쟁 작품이 많은 경우 예측보다 관객 수가 적을 수 있으며 이와 대조적으로 경쟁작품이 적을 경우 비교적 흥행할 가능성이 크다는 것을 도출할 수 있었다. 또한 흥행작품에 대한 리뷰에는 영화 자체를 표현하는 단어와 긍정적인 단어가 많았다면, 흥행에 실패한 영화의 경우는 영화 이외의 속성의 단어가 많고 부정적인 단어가 많은 것을 볼 수 있다.

두 영화 리뷰를 통해 공통적으로 얻어낸 인사이트로는 '마케팅 요소의 효과'이다. 두개의 영화 리뷰 모두에서 빈번하게 사용된 단어는 '평점'과 '무대인사'라는 요소이다. 이를 통해 '평점'의 경우에는 관객들의 입소문의 효과의 중요성을 파악할 수 있었고 (긍정적인 단어 혹은 부정적인 단어를 통한 입소문 효과), '무대인사'와 '광고', ' 예고편'과 같은 단어를 통해 마케팅 적 요소 역시 관객들에게 어떠한 영향력을 행사하고 있을 것으로 보인다.

## 3.3.4 사회연결망 분석 (SNA 분석)

### 1) 사회연결망 정의 및 목표

우리는 앞서 모델링 과정에서 배우의 파워가 관객수 흥행에 있어 중요한 핵심 요인으로 도출하였다. 만약 어떤 배우가 흥행에 영향력을 갖는지를 사전에 파악해 우리의 타겟층인 배급사에게 이 분석결과를 제공한다면 한국 영화에 있어서 긍정적 효과를 줄 것으로 기대한다.

사회연결망 분석이란 연결망 구조의 특징을 도출하고 인간관계와 사회현상을 설명하는 것으로 참여자들의 관계유형을 중점으로 하여 그 참여자를 체계적으로 표현하는데 효과적이다.

우리는 사회연결망 분석에서 많이 사용되는 척도인 중심성(centrality)에 초점을 맞추고 5백만명 이상의 관객수를 동원한 한국 영화에 출연한 배우들이 영화에서 공통출연한 관계로 형성된 사회연결망을 분석하였다. 관객수를 한정함으로써 배우의 연결망 중심성이 높을수록 작품 흥행할 수 있는 가능성이 높다고 보았고 이를 통해 중심성이 높은 배우들의 특징을 알아보고자 한다.

### 2) 사회연결망에서의 중심성(centrality)

한 연결망에서 '중심'이란 많은 구성원과 관계를 맺고 있거나, 혹은 단순히 연결망 내의 구조적인 위치가 중개자 역할인 경우를 의미한다. '중심'에 있다는 것은 결국 그 구성원이 연결망 상에서 힘이 있다는 것으로 해석할 수 있다. 사회연결망 분석에서 말하는 중심성은 주로



연결정도(Degree) 중심성, 매개(Betweenness) 중심성과 근접(closeness) 중심성의 세 가지 개념으로 구분된다

- **연결정도 중심성:** 한 명의 구성원에 직접 연결된 인접 구성원의 개수에 초점을 맞춤
- **근접 중심성:** 연결된 인접 구성원 수와 근접 정도를 측정하는 지표
- **매개 중심성:** 전체 연결망 내에서 다른 두 구성원 사이의 최단경로 (Geodesic)에 위치하는 정도를 측정하는 지표

### 3) 중심성(centrality) 분석

우리는 분석 조건을 지난 5 년동안의 한국상업영화 데이터 중 5 백만명의 관객수를 유지한 영화를 흥행영화라 보고 이를 기준으로 데이터를 제한하였고, 해당 영화들에 출연하는 주연을 기준으로 배우들 간의 전체적인 연결 구조를 파악하고, 어떤 배우가 연결망에서 중심적 위치에 있는지 파악하였다.

<표 14> 각 중심성 상위 10 명

순위	연결정도 중심성			근접 중심성			매개 중심성		
	배우	값	작품 수	배우	값	작품 수	배우	값	작품수
1	유해진	0.226667	5	유해진	0.377344	5	유해진	0.290300	5
2	황정민	0.200000	6	황정민	0.362604	6	황정민	0.209520	6
3	조진웅	0.160000	3	오달수	0.348972	3	오달수	0.149850	3
4	마동석	0.160000	4	유아인	0.320092	2	하정우	0.147417	4
5	류준열	0.146667	3	송강호	0.311499	3	류준열	0.135255	3
6	*김수안	0.133333	2	류준열	0.307373	3	마동석	0.068018	4
7	하정우	0.120000	4	*김수안	0.301385	2	송강호	0.066186	3
8	공유	0.093333	2	공유	0.297521	2	이정재	0.060541	2
9	오달수	0.093333	3	하정우	0.295626	4	*김수안	0.059790	2
10	송강호	0.080000	3	조진웅	0.284744	3	조진웅	0.058979	3

분석결과는 5 년간 '5 백만 이상의 관객수를 모집한 영화'에 출연한 총 76 명의 배우의 순위로 나타나지만, 여기에서는 상위 10 명의 결과만 작성하였다. <표 5> 를 살펴보면 배우 유해진, 황정민, 오달수가 상위권을 차지하고 있는 것을 확인할 수 있으며, 상위 10 위 안에 유일한 아역배우 아역 '김수안'의 영향력을 살펴볼 수 있었고, 여자 아역배우 '김수안' 배우를 제외한 성인 여성배우가 Top10 에 부재함에 따라 성인여성배우보다 남성배우가 배우 연결망 중심성 분석에서 월등히 우세한 것을 파악하였다. 또한 상위 10 위 안에 포함된 배우들의 연령대를 살펴보면, 20 대 배우들이 등장하지 않는 것을 알 수 있다.

### 4) 연결정도 중심성 수치 분석

연결정도 중심성 분석결과에는 유해진, 황정민, 조진웅의 순서로 1, 2, 3 위에 올랐다. 연결정도 중심성은 연결망 내에서 다른 참여자들과 연결이 많을수록 높은 수치가 나오는데 이는 배우들의 영향력에 대한 지표로 볼 수 있다. 연결정도 중심성의 분석 결과 중 2 위에 있는 황정민보다 출연 편수가 더 적은 유해진이 1 위에 위치하였는데 이는 유해진이 5 년동안 황정민보다 더 많은 다른 배우들과 공동출연 하였음을 의미한다. 연결정도 중심성은 배우들 간의 연결에 해당하는 공동 출연 배우 수에 중점을 두었다.

### 5) 근접 중심성 수치 분석

연결정도의 경우 전체 사회연결망에서 중앙에 위치하지 않아도 높은 값을 가질 수 있으나, 근접 중앙성의 경우에는 중앙에 위치해야만 높은 값을 보이는 특성이 있다. 근접 중앙성 척도가 해당 구성원이 다른 모든 구성원과 정보를 주고받을 때 거쳐야 하는 단계의 수에 반비례하기 때문이다. 그 결과 유해진이 가장 높은 값인 0.377 을 보이며 연결망 내의 중앙에 위치하고 있음을 알 수 있다. 근접 중앙성은 역수로 계산되기 때문에, 이 값은 유해진이 다른 모든 배우들과 평균 2.65 (= 1/0.377) 단계 만에 연결될 수 있음을 의미한다. 유해진과 같이 근접 중앙성이 높은 배우들은 사회연결망에서 다른 모든 배우들과의 정보 교류가 더욱 빠르고 효율적일 것으로 예상할 수 있다.

### 6) 매개 중심성 수치 분석

매개 중심성은 연결망 내에서 중재자 역할을 할 수 있는가에 해당하는 척도이며, 전체 연결망 내에서 다른 구성원들 간의 최단 경로상에 위치하는 정도로 계산한다. 매개 중심성이 높다는 것은 어떠한 특정 배우가 다른 두 배우 간의 연결 사이에 위치하고 있는 경우가 많다는 것을 의미한다. 즉, 많은 작품에 출연하였거나, 비교적 소수 작품에 출연하였다면 출연작품의 성격이나 출연 배우 구성이 이질적인 작품인 것들에 출연하였음을 파악할 수 있다. 이처럼 매개 중심성의 수치가 높은 배우들은 이질적인 복수의 집단들의 구성원과 쉽게 상호작용이 가능한 배우들로 보인다.

### 7) 위세 중심성 수치 분석

<표 15> 위세 중심성 상위 10 명

순위	위세 중심성		
	배우	값	작품수
1	유해진	4.288063e-01	5
2	조진웅	3.439085e-01	3
3	송하윤	2.463571e-01	1
4	김지수	2.463571e-01	1
5	염정아	2.463571e-01	1
6	이서진	2.463571e-01	1
7	윤경호	2.463571e-01	1
8	류준열	2.401724e-01	3
9	황정민	1.901086e-01	6
10	김주혁	1.893904e-01	2

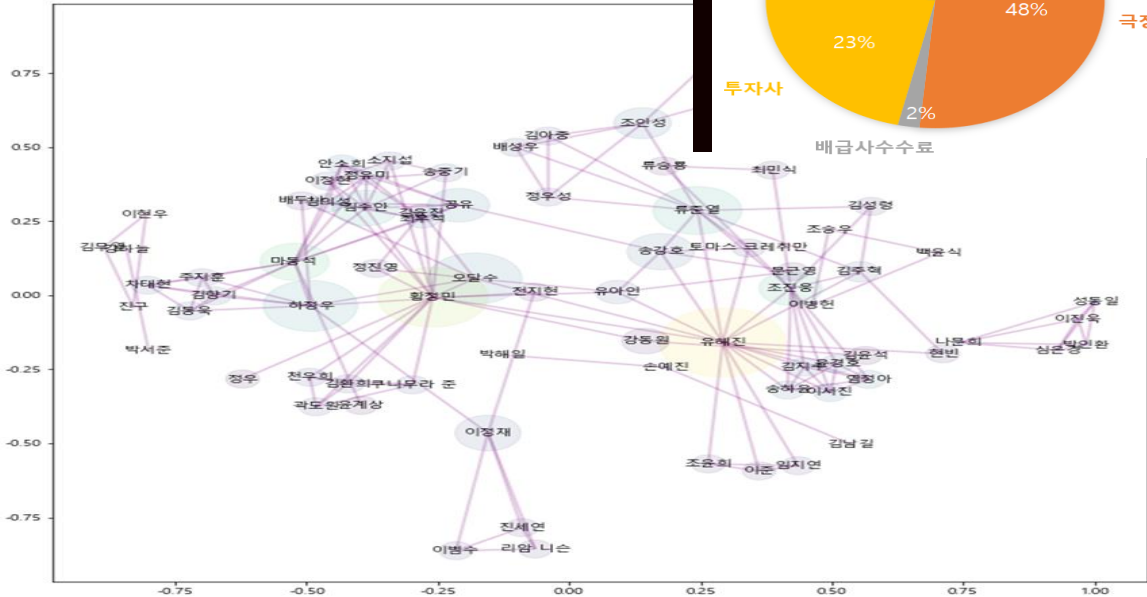
위세 중심성은 배우들 간의 연결에 해당하는 공동 출연 배우 수와 더불어 함께 출연한 배우의 연결정도까지 함께 고려한 개념이다. 위세 중심성의 1, 2, 3 위 순서는 유해진, 조진웅, 송하윤으로 이루어졌다. 분석 결과 중 8,9 위에 위치하는 배우 류준열, 황정민 보다 더 적은 작품 출연 편수임에도 불구하고 송하윤, 김지수, 염정아, 이서진, 윤경호 등이 높은 순위에 위치했다. 이것은 이들이 그동안 많은 다른 배우들과 함께 합작했으며, 연결망 내에서 상대적으로 중요한 역할을 하는 배우들과 자주 합작하였음을 의미해 주조연으로서의 역할로 잘 어울릴 것으로 보인다. 반면에, 연결정도, 매개, 근접 중심성이 높으나 위계중심성에서는 비교적 수치가 낮게 나온 류준열이나 황정민의 경우에는 그들 자체로 상대적 중요한 역할을 하기 때문에 수치가 비교적 낮게 나온 것으로 보인다.

### 8) 사회 연결망 분석 시각화

- **노드의 색상:** 노드의 색상은 배우 각각의 degree 로 나타냈다. Degree 가 깊을수록 더 많은 다른 배우들과 공동출연 하였음을 의미한다. 배우들의 데이터를 5 백만 이상의 영화 즉 흥행영화로 한정 지었기에, 노드의 색상이 밝을수록 배우들의 활동력이 높고 여러 배우들과 잘 어울릴 수 있는, 즉 흥행 보증 배우들과의 케미를 잘 낼 수 있는 배우라고 보인다. [유해진 > 황정민 > 조진웅]

- **노드의 크기:** 노드의 크기는 매개중심성 수치로 나타내었고, 크기가 클수록 많은 작품에 출연하여 5년간의 활동력이 강했던 경우나 출연 작품수가 적다면 이질적인 복수의 집단들의 구성원과 쉽게 상호작용이 가능한 배우로 볼 수 있다. [유해진> 황정민> 오달수> 하정우> 류준열]

<그림 21> 배우들의 사회 연결망 분석 시각화



### 3.4 활용 방안

#### 3.4.1 극장주의 스크린 수 편성에 활용

영화 관련 실무자들은 영화 개봉 전 그 영화가 얼마나 흥행에 성공할 수 있는가에 관심이 많다. 그 중에서도 영화관은 개봉 예정 영화의 예측 관람객 수요에 따라 적합한 좌석수의 영화관에서 상영하고 몇개의 상영관에서 상영할 것인가를 결정한다. 또한 상영관은 예측수요에 따라 영화사와 라이선스 계약을 어떤 조건으로 체결할 것인가를 결정할 수 있다.

관객수에 따라 수익을 가장 많이 가져가는 것은 극장주라는 것을 고려해 보았을 때, 이들을 타깃으로 우리의 정확도 95.6%의 개봉이전 관객수 예측 모델을 활용할 수 있을 것이다.

#### 3.4.2 배급사의 영화 마케팅에 활용

##### 1) 시사회 현장 SNS 이벤트의 혜택 다양화

관객수와 개봉 전 트윗의 상관관계 및 회귀계수를 고려했을 때, 개봉이전 SNS 를 통한 입소문 효과가 중요하다고 볼 수 있다. 이벤트를 더 활성화하고 참여를 유도하기 위해서 <표 16>와 같이 기존보다 혜택을 다양화할 것을 제안한다.

<표 16> 시사회 현장 SNS 혜택 다양화

	방법	혜택
기존	상영 전 스크린에 이벤트 공지, 당첨자 추후 발표	커피 기프트콘 등 보편적 이벤트의 혜택 제공
제안	당첨자를 영화 상영 종료 후 즉시 발표하여 현장에서의 참여를 유도	영화 티켓, 배우 사인이 들어간 영화 포스터, 영화 제작 과정 비하인드 포토 등 영화 및 배우의 특성을 살린 혜택 제공
관련	개봉 전 트윗	배우파워, 배우수상 등

변수	
----	--

## 2) 인플루언서(Influencer) 마케팅 전개

관객수와 개봉 전 트윗의 상관관계를 고려했을 때, 인플루언서 마케팅을 전개할 것을 제안한다. 얼마전까지만 해도 인플루언서 마케팅이라고 하면 블로거들을 통한 체험단 마케팅이 대부분이었고, 최근 들어 페이스북이나, 인스타그램 등의 빅 유저를 통한 인플루언서 마케팅이 굉장히 활발해졌다. 그러다가 최근 1년 사이에 인플루언서 마케팅의 판도를 바꿀 만큼 큰 시장이 열렸는데, 바로 유튜브 크리에이터와 콜라보이다. SNS의 트렌드가 기존 페이스북과 트위터에서 인스타그램과 유튜브로 흘러가는 시대적 흐름에 따라 개봉 전 트윗은 새로운 소셜미디어의 대리변수라고 할 수 있다. 유튜브는 구독자 기반의 플랫폼으로 크리에이터의 신뢰도를 기반으로 추천 영상을 통한 유입을 유도할 수 있다. 우리가 제안하는 바는 <표 17>와 같이 유튜브를 통한 마케팅을 전개하는 것이다.

<표 17> 인플루언서(Influencer) 마케팅 확대

	방법	관련변수
기존	<ul style="list-style-type: none"> <li>- 다음, 네이버 등 포털 및 커뮤니티 사이트의 배너광고</li> <li>- 모든 매체를 통해 배포되는 기사 관리</li> <li>- 파워 블로거 및 블로거모임 컨택을 통한 포스팅</li> <li>- 페이스북, 공식 홈페이지 등을 통한 시사회 초대 및 시사회 운영</li> </ul>	개봉 전 트윗 (대리변수)
제안	<ul style="list-style-type: none"> <li>- 최소 10만명 이상의 구독자를 보유한 유튜버를 시사회에 초대</li> <li>- 영화 관련 영상 및 흥미롭게 소개하는 영상을 게재하는 협약 체결</li> <li>- 소비자들이 원하는 니즈를 가장 잘 이해하는 유튜버들과 커뮤니케이션 방식을 모색 가능</li> <li>- 영화를 흥미롭게 소개하는 등 원하는 메시지를 다양하게 표현해서 전달할 수 있음</li> </ul>	

### 3.4.3 문화 공연 산업의 수요 예측 모델로 활용

한국 상업 영화 관객수 예측 모델의 변수 선택 및 정의, 분석 기법 등을 활용하여 뮤지컬, 연극, 오페라 등 문화공연의 수요 예측 및 영향요인 분석에 활용할 수 있으리라 기대한다.

### 3.4.4 웹드라마, 웹툰 등의 댓글 분석에 활용

지금까지의 활용방안 외에도 온라인 웹툰 및 웹드라마의 조회수 및 수요 예측 등 영화 이외의 콘텐츠 산업의 수요 예측 모델 개발에도 활용할 수 있을 것이다. 더 나아가, 텍스트 마이닝과 워드 클라우드 기법을 활용하여 해당 콘텐츠의 리뷰 및 댓글을 분석하는 방안도 고려된다.

## 4. 기대 효과

### 4.1 향후 개선 사항

첫째, 분석 데이터 측면에서 봤을 때, 제작비 및 손익분기점 데이터를 구하기가 어려웠다는 점이 대표적인 제약사항 중 하나다. 향후 손익분기점을 고려하여 예측 모델을 생성한다면 수익률까지 도출해낼 수 있으리라 생각된다. 또한 경쟁 작품 등 외부 환경을 고려한 외생변수, 영화진흥위원회에 API에서는 제공하지 않는 시나리오 작가의 수상횟수 및 파워 변수를 수집 및 추가하여 상관성을 살펴보고, 모델을 보완해 나가야 할 것이다.

둘째, **분석 기법 측면**에서는 모델 알고리즘에 대한 수학적 개념까지 이해하는 데 시간이 많이 소요되었다는 점이 제약사항이다. 향후 프로젝트를 업그레이드할 때에 모델링에 대한 더 깊이 있는 이해를 바탕으로 분석을 실시해보고자 한다. 또한 사회네트워크분석(SNA)을 현재 '배우'에서 더 나아가, '감독'만으로, '감독+배우'로 넓혀 분석을 시도할 필요성이 있다. 이를 통해, 시너지 효과를 파악할 수 있으리라 기대한다. 그 외에도, 크롤링한 리뷰 텍스트 데이터를 활용한 텍스트 마이닝 분석, 감정분석 등을 진행하여 관련 모델을 생성하는 것과 워드 클라우드를 더 많은 표본으로 분석해 볼 필요가 있다.

셋째, **분석 모델 측면**에 우리의 최종 모델과는 별개로 우리는 개봉 후 마케팅 변수(개봉 1일차~3일차 스크린 수, 상영횟수, 개봉 후 뉴스, 개봉 후 트윗 수)를 추가하여 개봉이 시작된 후에도 최종 관객수 예측을 할 수 있는 모델도 만들었다. 개봉 후의 마케팅 변수들은 최종 예측 값과 유사한 그래프를 띄고 있고 상관성이 높은 변수들이 들어가 확실히 최종 관객수 예측에 더 높은 정확도(98%)를 보였다. 하지만 우리가 설정한 타겟 고객층인 극장 주 및 배급사의 홍보팀은 개봉 전의 한정된 정보만으로 판단을 한다는 점에서 이 모델은 본 보고서에서 제외하였다. 지금 프로젝트의 연장선으로 추후에 이 모델에 대한 분석 및 실무적 활용성을 살펴보기로 한다.

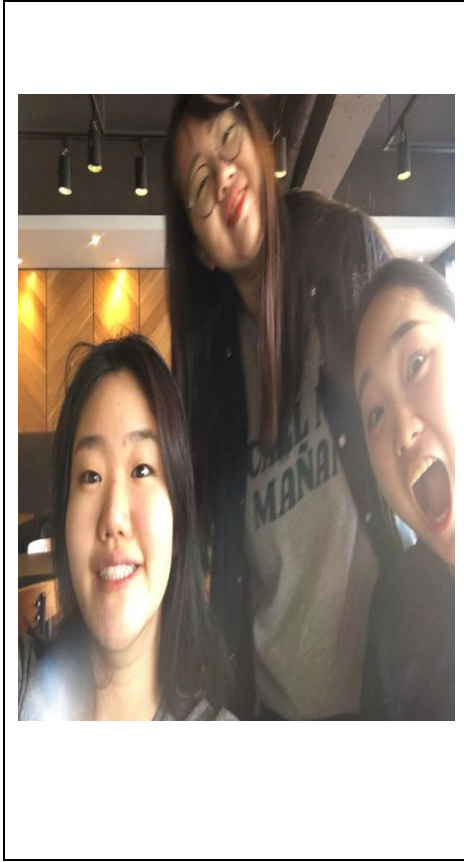
넷째, **향후 프로젝트 대상 및 주제의 측면**에서 '해외영화 관객수 및 수요 예측' 등을 시도한다면 보다 업그레이드할 수 있을 것이라 생각된다. 또한, 영화산업 전체를 고려해 보았을 때, 공급채널을 현재 분석 초점인 '극장'보다 더 넓혀서 VOD 시장, 인터넷 온라인 콘텐츠 시장 등에서의 수요 예측 모델을 개발함으로써 향후 프로젝트 주제를 발전시킬 수 있을 것이다.

## 4.2 기대 효과

API 와 웹 크롤링으로 수집한 데이터를 전처리하고 다양한 분석기법 모델링인 ANN, 의사결정나무, 일반회귀예측 등을 활용하여 정확도가 95.6%까지 올라가는 '개봉예정 한국 상업영화의 관객수 예측모형'을 개발하였다. 이 예측 모형을 활용함으로써 극장주가 보다 효과적으로 스크린 수를 편성하는 데 기여할 수 있을 것이다. 그리고 배급사의 영화 마케팅에 활용하는 방안으로는 시사회 현장 SNS 이벤트의 다양화, 인플루언서(Influencer)마케팅 전개 등을 제시했다. 추후 더 나아가 이 모델을 다른 문화 공연 산업의 수요 예측 모델에도 응용 가능할 것으로 기대하는 바이다.

성명	후기
박희지	학부 수업에서는 듣지 못했던 전문적인 통계 분석과 다양한 분석 방법론을 배우며 한걸음 더 성장할 수 있었다. 다양한 전공의 사람들, 멘토님들과 함께한 프로젝트 역시 시야를 넓게 볼 수 있게 해준 좋은 경험이었다. 아직 실력이 충분하진 않지만 이 과정을 기반으로 좋은 데이터 사이언티스트가 되고 싶다.

## 5. 분석 후기



<p>임현수</p>	<p>처음 시도해보는 분야가 많아서 어려웠지만 배우는 과정 하나하나가 뜻 깊었다. 웹 크롤링과 전처리를 통해 모델링을 만들고 이로써 새로운 인사이트를 도출하는 과정이 매우 흥미로웠다. 좋은 분들을 많이 만나 프로젝트를 잘 마무리 할 수 있었던 것 같다.</p>
<p>정다연</p>	<p>웹 크롤링 코드를 처음 짜면서 어려움도 많았지만 원하는 데이터를 직접 수집해 볼 수 있었다는 점이 가장 기억에 남는다. 또한, 모델링을 하는 일련의 단위 작업은 여러 경우의 수를 시도해 보는 것이 중요하며, 해당 도메인 업계 실무자의 관점에서 문제를 바라보는 시각을 길러야 한다는 점을 배울 수 있었다.</p>