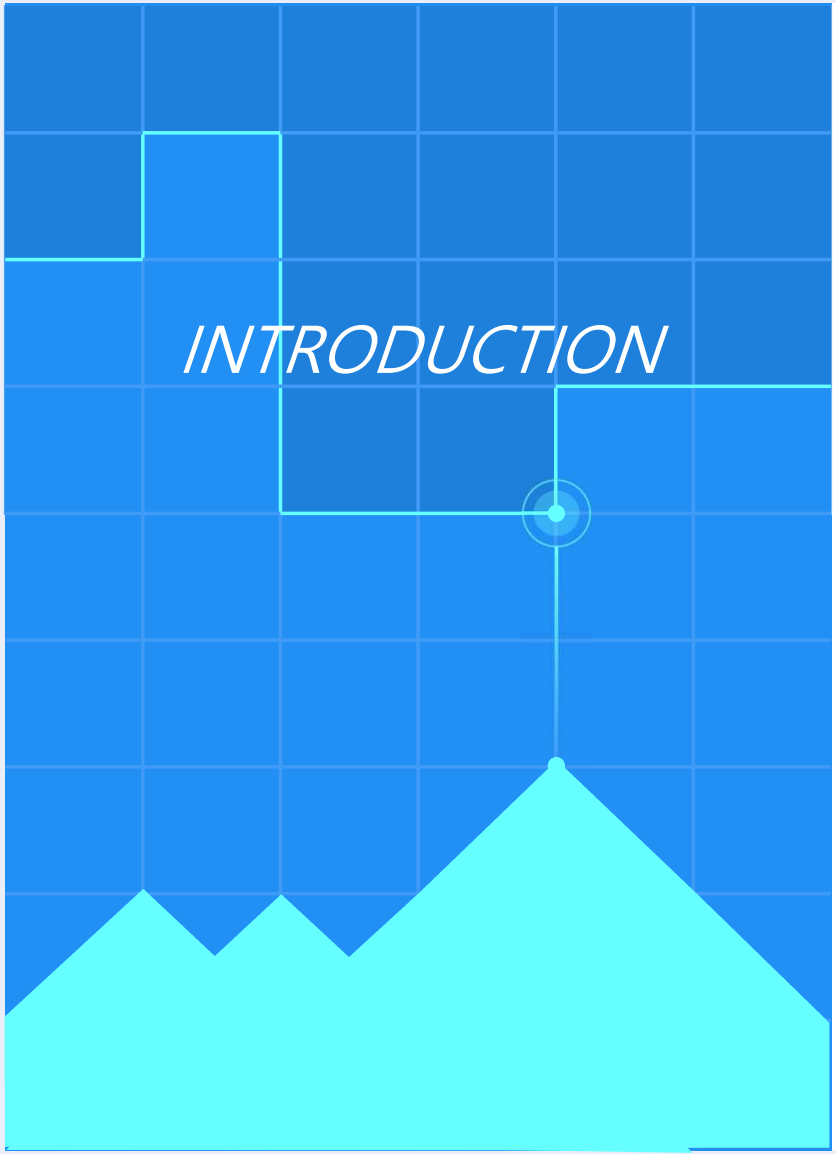
Q & M

영화 개봉이 “영화산업지수”에 미치는 영향

CONTENTS

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

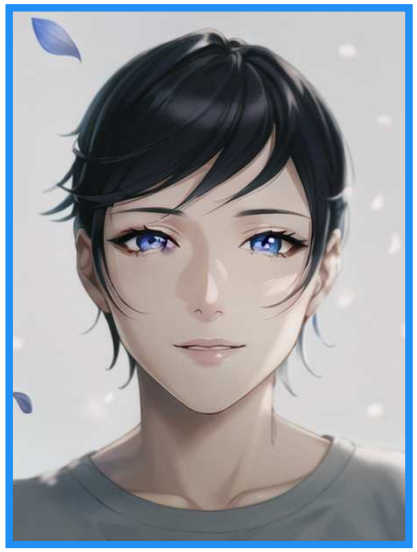


INTRODUCTION

팀원소개/ 타임라인/ WORK FLOW

R & R(팀명:Literacy)

Imacreator.zmo. ai



Kim Jeong-woo

팀장
PPT 및 일지
자료수집

Seo Su-a

메인 코더
도메인
자료수집

Ko Won-tae

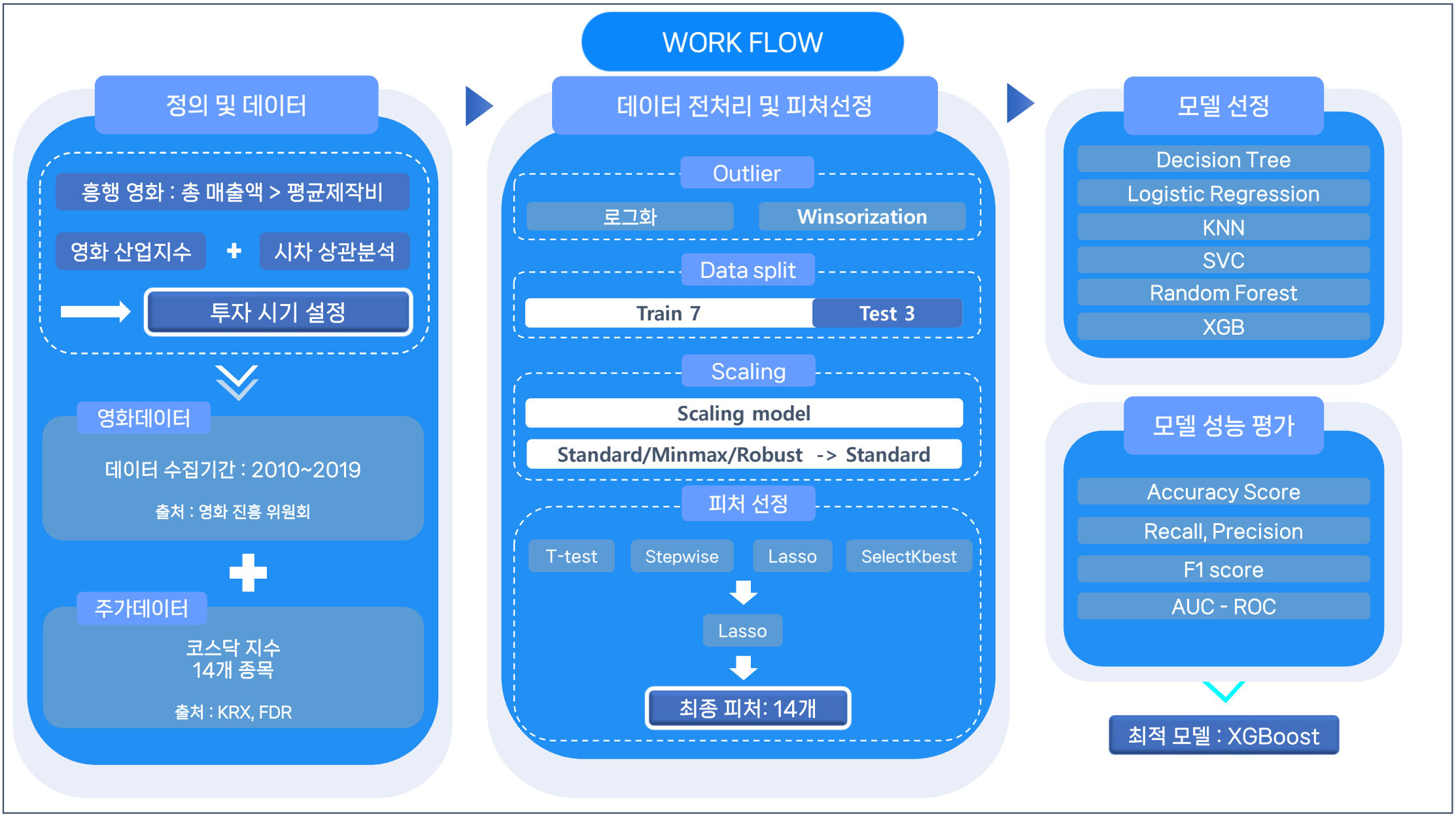
도메인
서브코더
자료수집

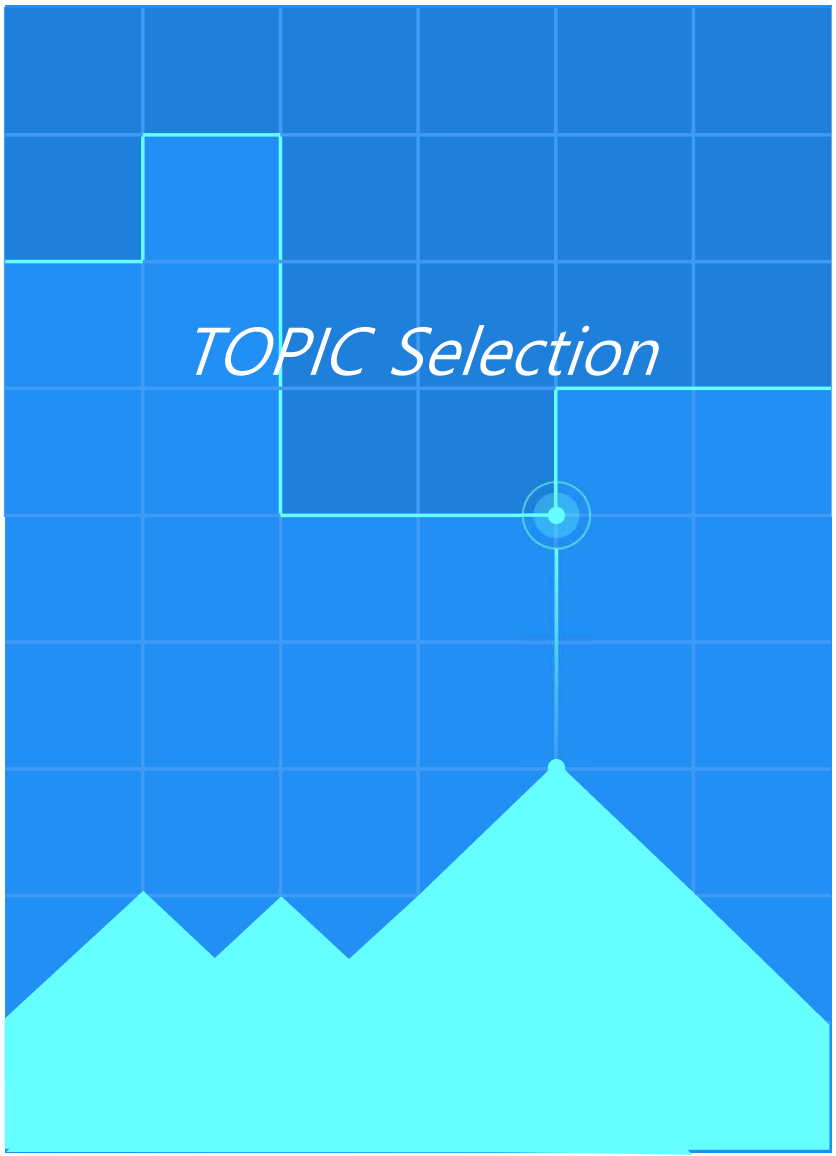
Hong Sung-il

도메인
메인코더
자료수집

TIME LINE

	2.27 ~ 3.3	3.4~3.15	3.16~3.21	3.22~3.26
주제선정 데이터 수집				
(예비)				
데이터 전처리 EDA				
(예비)				
피쳐선정 모델링				
(예비)				
인사이트 도출				
(예비)				


WORK FLOW
정의 및 데이터
흥행 영화 : 총 매출액 > 평균제작비
영화 산업지수
+
시차 상관분석
투자 시기 설정
영화데이터
데이터 수집기간 : 2010~2019
출처 : 영화 진흥 위원회
주가데이터
코스닥 지수
14개 종목
출처 : KRX, FDR
데이터 전처리 및 피쳐선정
Outlier
로그화
Winsorization
Data split
Train 7
Test 3
Scaling
Scaling model
Standard/Minmax/Robust -> Standard
피처 선정
T-test
Stepwise
Lasso
SelectKbest
Lasso
최종 피처: 14개
모델 선정
Decision Tree
Logistic Regression
KNN
SVC
Random Forest
XGB
모델 성능 평가
Accuracy Score
Recall, Precision
F1 score
AUC - ROC
최적 모델 : XGBoost



TOPIC

주제 선정 이유 및 배경

TOPIC SELECTION

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

주제 선정 배경

[특징주] 국내 음원차트 휩쓰는 신인 걸그룹 '뉴진스'…엔터테인먼트 관련주 하이브 주가 탄력받나

➢ '뉴진스(NewJeans)'의 데뷔 앨범의 호평으로
➢ 소속사 '하이브' 주가에 투자자들의 관심이 쏠리고 있다.

[증시 키워드] 엔씨소프트, 신작 '리니지W' 공개 앞두고 주가 상승세

입력 2021-11-01 08:27

조성진 기자 csjjin2002@etoday.co.kr

➢ 공개 전 거래일 '엔씨소프트'는 0.97%(6000원) 오른 62만7000원을 기록
➢ '엔씨소프트'는 오는 4일 신작 '리니지W' 출시를 앞두고 있어 주가 회복에 기대를 모으고 있다.

주제 선정 배경

증권

아바타 역대급 성적 낼까…흥행 기대감에 배급·영화사 주가 쑥

조윤희 기자 choyh@mk.co.kr
입력 : 2022-12-12 11:16:28

가

> 바른손이앤에이·CGV 등 12일 오전 강세
> 아바타2 12일 오전 기준 52만명 예약

- 13년 만에 개봉하는 영화 '아바타' 후속편에 대한 기대감
  - 바른손이앤에이와 CJ CGV등 관련주 강세

Y이슈

[Y이슈] 입소문 따라 주가도 철렁... 우영우·한산 관련주에 쏠리는 눈길

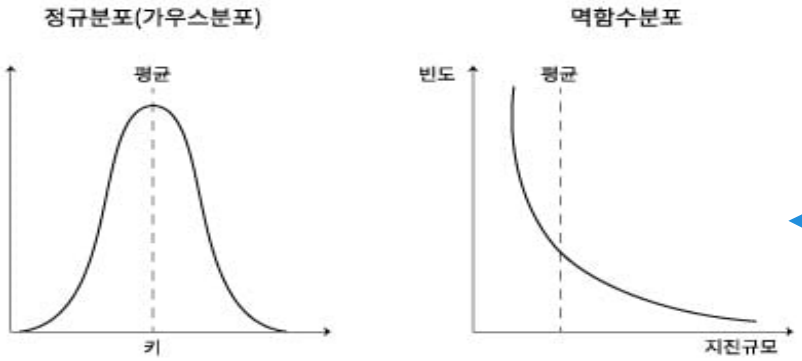
2022년 07월 26일 16시 18분 댓글

- '한산 관련주', '우영우 관련주', '롯데엔터테인먼트 관련주' 등 최근 콘텐츠관련주 투자 열기가 뜨겁다.

주제 선정 배경

영화개봉과 CJ ENM 주가 비교 (극한직업, 명량, 신과함께-죄와벌)

주제 선정 배경 이론



사람의 키 분포와 지진 규모 분포의 비교. 사람의 키는 평균을 중심으로 해서 좌우로 거의 비슷하게 분포하지만, 지진의 규모는 압도적인 다수가 평균 이하 구간에 분포한다.

정규 분포

독립

분포곡선이 평균값을 중심으로 좌우대칭인 종모양을 이루는 것

멱함수 분포

허브

평균 보다 아래에 있는 빈도수가 대다수를 차지

연구 배경

- 영화업종지수
- 기간: 2000.01.04 ~ 2020.12.31

- 2000 ~ 2010 : 횡보 구간
- 2010 ~ 2019 : 저점 상승, 상승 추세

- 2010년 이후 상승 추세로 바뀌며,
  투자전략 수립 시 참고 자료 이용 가능

연구 배경

이러한 배경 조사의 결과, 영화개봉이 주식에 영향을 유의미하게 미칠 것으로 예상되어 머신러닝을 활용하여 영화데이터 기반 투자모델을 만들고자 함.

연구 주제

주 제	영화 개봉이 "영화산업지수"에 미치는 영향

대립가설 보다는 가정이란 단어로 변경 필요

대립가설1 : 영화데이터가 배급사 주식변동 예측에 유의미 하다.

대립가설2 : 상위 순위 영화가 하위 순위 영화보다 주식변동에 미치는 영향이 크다.

데이터 수집 및 전처리

E.D.A

타겟 설정

데이터 수집 및 전처리

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

FEATURES 후보군

영화 데이터		주가 데이터	
KOFIC	KOFIC(OPEN API)	KRX	FDR
순번(순위) 영화명 제작사 수입사 배급사 개봉일 영화유형 영화형태 국적 등급 영화구분 장르 전국스크린수 전국매출액 전국관객수 서울매출액 서울관객수 평균제작비	개봉일 관객수 개봉일 매출액 개봉일 매출비율 배우(Actor) 감독(Director)	‹**코스피** (5종목)› CJ CGV 롯데쇼핑 아센디오 IHQ 콘텐트리중앙 ‹**코스닥** (9종목)› CJ ENM 쇼박스 NEW 애니플러스 위지윅 스튜디오 덱스터 바른손이앤에이 판타지오 스튜디오 산타클로스	코스닥 지수

데이터 수집 및 전처리

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

FEATURES 후보군

영화 데이터		주가 데이터	
KOFIC	KOFIC(OPEN API)	KRX	FDR
순번(순위) 영화명 제작사 수입사 배급사 개봉일 영화유형 영화형태 국적 등급 영화구분 장르 전국스크린수 전국매출액 전국관객수 서울매출액 서울관객수 평균제작비	개봉일 관객수 개봉일 매출액 개봉일 매출비율 배우(Actor) 감독(Director)	‹**코스피** (5종목)› CJ CGV 롯데쇼핑 아센디오 IHQ 콘텐트리중앙 ‹**코스닥** (9종목)› CJ ENM 쇼박스 NEW 애니플러스 위지윅 스튜디오 덱스터 바른손이앤에이 판타지오 스튜디오 산타클로스	코스닥 지수

데이터 수집 및 전처리

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

FEATURES 후보군

영화 데이터	주가 데이터
KOFIC	

소수의 배급사가 시장 장악 → 영화 제작 시 배급사와 계약을 맺고 제작

국내 영화배급사와 주식시장의 영향력을 판단

순번(순위)
영화명
제작사
수입사
배급사
개봉일
영화유형
영화형태
국적
등급
영화구분
장르
전국스크린수
전국매출액
전국관객수
서울매출액
서울관객수
평균제작비

개봉일 매출액
개봉일 매출비율
배우(Actor)
감독(Director)

CJ CGV
롯데쇼핑
아센디오
CJ ENM
쇼박스
NEW
애니플러스
위지윅 스튜디오
덱스터
바른손이앤에이
판타지오
스튜디오 산타클로스

데이터 수집 및 전처리

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

RAW DATA 생성

흥행 영화의 선정 기준

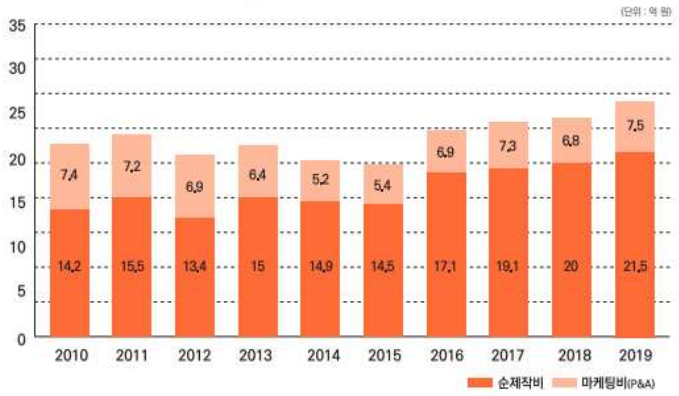
총매출액 – 제작비(연평균 제작비)>0

	영화명	개봉일	국적	전국스크린수	전국매출액	전국관객수	서울매출액	서울관객수	장르	등급	...	평균 제작비	국내배급사	상장배급사	주요배급사	개봉일관객수	개봉일매출액	개봉일매출비율	타겟	배우가중치	감독가중치
0	명량	2014-07-30	1.0	1587	1357483989 10	17613682	3.312123e+10	4163666	사극	1	...	20.1	0.0	1.0	1	682701.0	4.708879e+09	61.6	0	39.000000	2.000000
1	극한직업	2019-01-23	1.0	1978	139647979516	16264944	3.185866e+10	3638287	코미디	1	...	29.0	0.0	1.0	1	368582.0	3.004763e+09	73.2	1	21.000000	3.000000
2	신과함께-죄와 벌	2017-12-20	1.0	1912	115698654137	14410754	2.753083e+10	3346172	판타지	2	...	26.3	0.0	1.0	1	406365.0	2.984151e+09	63.2	0	27.000000	3.000000
3	국제시장	2014-12-17	1.0	966	110828014630	14245998	2.584252e+10	3233946	드라마	2	...	20.1	0.0	1.0	1	184756.0	1.356870e+09	31.0	0	29.000000	1.000000
4	어벤져스: 엔드게임	2019-04-24	0.0	2835	122182694160	13934592	3.357714e+10	3597963	액션	2	...	29.0	1.0	0.0	1	1338729.0	9.678990e+09	97.1	0	22.000000	3.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
909	스파이 브릿지	2015-11-05	0.0	425	2015360478	260721	7.222464e+08	90018	스릴러	2	...	19.9	1.0	0.0	1	18404.0	1.352783e+08	5.6	1	0.333333	0.100000
910	함정	2015-09-10	1.0	445	2088685200	257716	4.114176e+08	50477	스릴러	0	...	19.9	0.0	0.0	0	32277.0	2.502558e+08	13.1	0	0.633333	0.033333
911	파이널 데스티네이션5	2011-09-08	0.0	168	2370348500	257193	8.800060e+08	87499	공포(호러)	0	...	22.7	1.0	0.0	1	12613.0	1.180325e+08	7.5	1	0.033333	0.066667
912	언브로큰	2015-01-07	0.0	401	1997947292	256475	6.057764e+08	75489	액션	1	...	19.9	1.0	0.0	1	26707.0	1.993983e+08	5.8	1	0.233333	0.033333
913	하늘을 걷는 남자	2015-10-28	0.0	415	2267273653	238257	9.107222e+08	92402	드라마	2	...	19.9	1.0	0.0	1	19675.0	1.432000e+08	6.4	0	0.200000	0.066667

연 평균 제작비

순제작비 + 마케팅비

〈그림 31〉 2010-2019년 한국영화 실질개봉작 평균 제작비 구성



데이터 수집 및 전처리

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

RAW DATA 생성

흥행 영화의 선정 기준

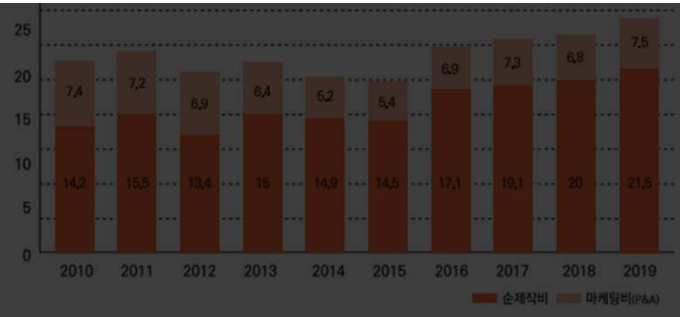
총매출액 − 제작비(연평균 제작비) > 0

연 평균 제작비

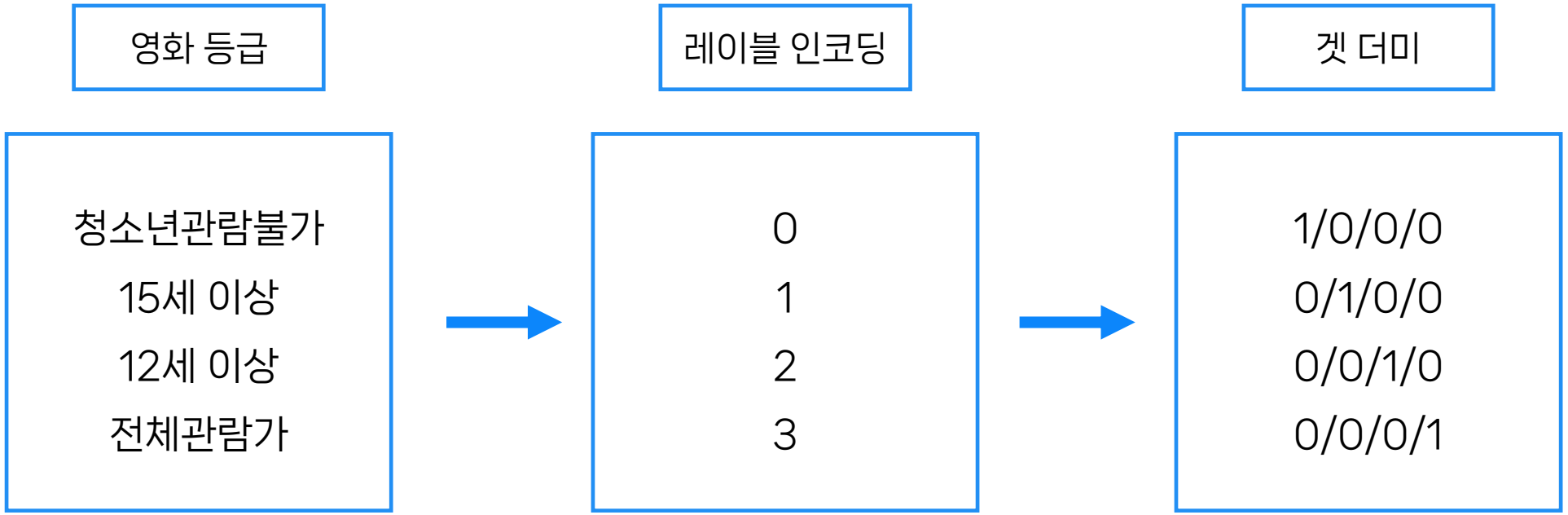
순제작비 + 마케팅비

	영화명	개봉일	국적	전국스크린수	전국매출액	전국관객수	서울매출액	서울관객수	장르	등급	...	평균제작비	국내배급사	상장배급사	주요배급사	개봉일관객수	개봉일매출액	개봉일매출비율	타겟	배우가중치	감독가중치
3	국제시장	2014-12-17	1.0	966	110828014630	14245998	2.584252e+10	3233946	드라마	2	...	20.1	0.0	1.0	1	184756.0	1.356870e+09	31.0	0	29.000000	1.000000
4	어벤져스: 엔드게임	2019-04-24	0.0	2835	122182694160	13934592	3.357714e+10	3597963	액션	2	...	29.0	1.0	0.0	1	1338729.0	9.678990e+09	97.1	0	22.000000	3.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
909	스파이 브릿지	2015-11-05	0.0	425	2015360478	260721	7.222464e+08	90018	스릴러	2	...	19.9	1.0	0.0	1	18404.0	1.352783e+08	5.6	1	0.333333	0.100000
910	황정	2015-09-10	1.0	445	2088685200	257716	4.114176e+08	50477	스릴러	0	...	19.9	0.0	0.0	0	32277.0	2.502558e+08	13.1	0	0.633333	0.033333
911	파이널 데스티네이션5	2011-09-08	0.0	168	2370348500	257193	8.800060e+08	87499	공포(호러)	0	...	22.7	1.0	0.0	1	12613.0	1.180325e+08	7.5	1	0.033333	0.066667
912	언브로큰	2015-01-07	0.0	401	1997947292	256475	6.057764e+08	75489	액션	1	...	19.9	1.0	0.0	1	26707.0	1.993983e+08	5.8	1	0.233333	0.033333
913	하늘을 걷는 남자	2015-10-28	0.0	415	2267273653	238257	9.107222e+08	92402	드라마	2	...	19.9	1.0	0.0	1	19675.0	1.432000e+08	6.4	0	0.200000	0.066667

11,819개(전체 영화) → 914개(흥행 영화)



Get dummie


영화 등급
레이블 인코딩
겟 더미
청소년관람불가
15세 이상
12세 이상
전체관람가
0
1
2
3
1/0/0/0
0/1/0/0
0/0/1/0
0/0/0/1

파생 변수

감독가중치	계산식	$\frac{\text{감독 연출 횟수}}{\text{영화 관객 가중치}}$	
배우가중치	계산식	$\frac{\text{배우 출연 횟수}}{\text{영화 관객 가중치}}$	
상장배급사	분류	상장배급사 : 1	비상장배급사 : 0
국내배급사	분류	국내배급사 : 1	해외배급사 : 0
주요배급사	분류	메이저배급사 : 1	마이너배급사 : 0

파생 변수

감독가중치	계산식	1000만 초과 : 1 (2.3%) 500만 초과 : 10 (7.2%) 100만 초과 : 20 (44.5%) 100만 이하 : 30 (46%)	감독 연출 횟수 영화 관객 가중치
배우가중치	계산식	영화별 상위 배우 3명 선정	배우 출연 횟수 영화 관객 가중치
상장배급사	분류	상장배급사 : 1	비상장배급사 : 0
국내배급사	분류	국내배급사 : 1	해외배급사 : 0
주요배급사	분류	메이저배급사 : 1	마이너배급사 : 0

파생 변수

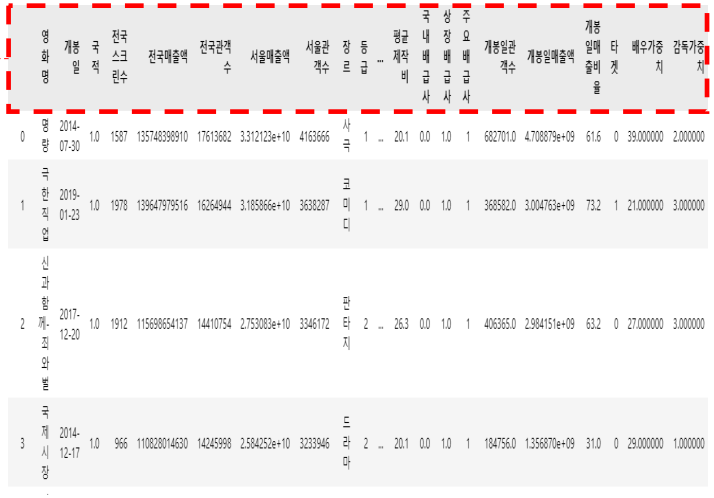
감독가중치	계산식	$\frac{\text{감독 연출 횟수}}{\text{영화 관객 가중치}}$	
배우가중치	계산식	$\frac{\text{배우 출연 횟수}}{\text{영화 관객 가중치}}$	
상장배급사	분류	상장배급사 : 1	비상장배급사 : 0
국내배급사	분류	국내배급사 : 1	해외배급사 : 0
주요배급사	분류	메이저배급사 : 1	마이너배급사 : 0

기획, 제작, 투자, 배급, 수입 전부 한번에 담당 가능한 배급사

데이터 수집 및 전처리

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

최종 데이터 컬럼

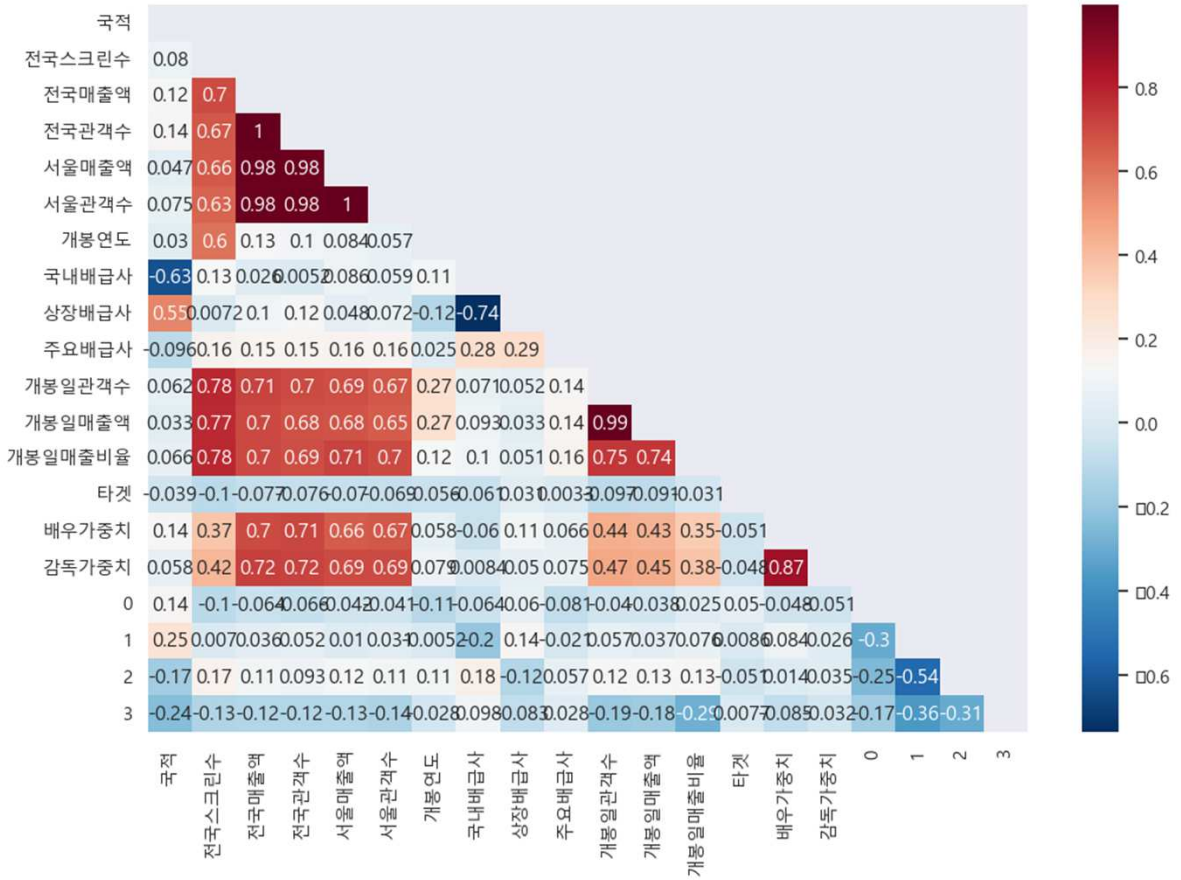


	영화명	개봉일	국적	전국스크린수	전국매출액	전국관객수	서울매출액	서울관객수	장르	등급	...	평균제작비	국내배급사	상장배급사	주요배급사	개봉일관객수	개봉일매출액	개봉일매출비율	타겟	배우가중치	감독가중치
0	영랑	2014-07-30	1.0	1587	135748398910	17613682	3.312123e+10	4163666	사극	1	...	20.1	0.0	1.0	1	682701.0	4.708879e+09	61.6	0	39.000000	2.000000
1	극한직업	2019-01-23	1.0	1978	139647979516	16264944	3.165866e+10	3638287	코미디	1	...	29.0	0.0	1.0	1	368582.0	3.004763e+09	73.2	1	21.000000	3.000000
2	신과함께-죄와벌	2017-12-20	1.0	1912	115698654137	14410754	2.753083e+10	3346172	판타지	2	...	26.3	0.0	1.0	1	406365.0	2.984151e+09	63.2	0	27.000000	3.000000
3	국제시장	2014-12-17	1.0	966	110828014630	14245998	2.584252e+10	3233946	드라마	2	...	20.1	0.0	1.0	1	184756.0	1.356870e+09	31.0	0	29.000000	1.000000

타겟을 제외한 19개 컬럼 생성

기본정보	영화명, 개봉일, 국적, 장르, 등급, 전국스크린수, 개봉연도
관객수	전국관객수, 서울관객수, 개봉일관객수
매출액	전국매출액, 서울매출액, 개봉일매출액, 개봉일매출비율
배급사	국내배급사, 상장배급사, 주요배급사
티켓파워	감독가중치, 배우가중치

Heatmap, VIF



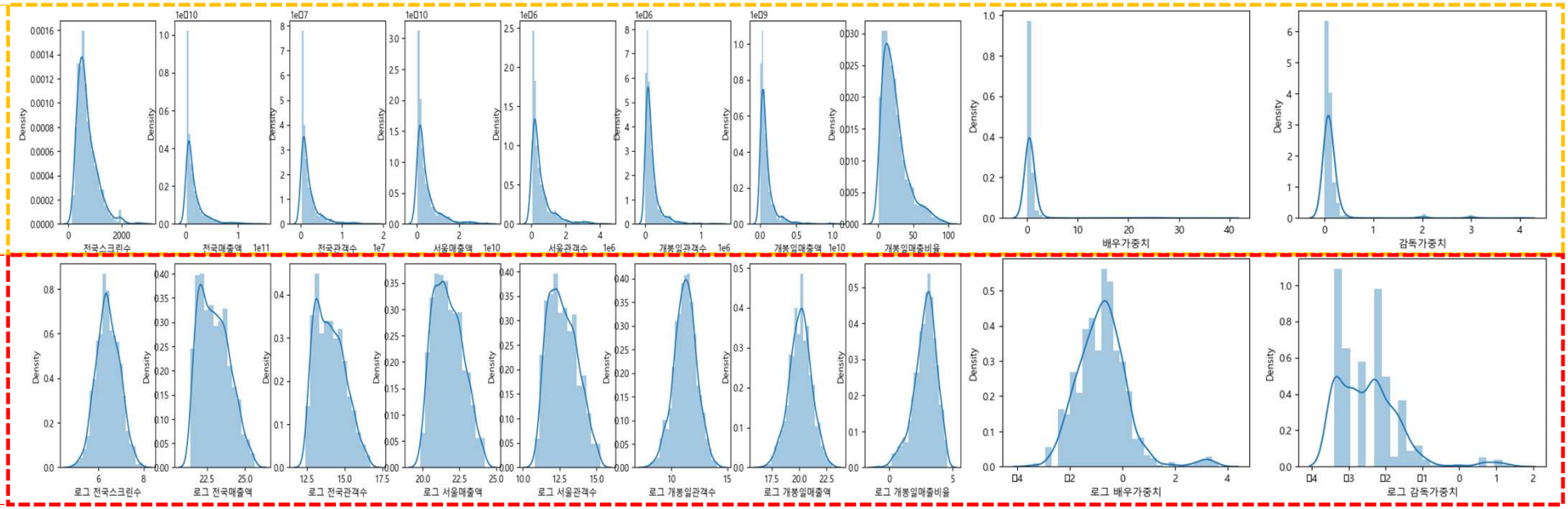
	VIF Factor	features
0	2299.679555	전국관객수
1	2206.205315	전국매출액
2	1882.506061	서울관객수
3	1783.670279	서울매출액
4	70.783115	개봉일관객수
5	68.886048	개봉일매출액
6	12.170841	전국스크린수
7	6.874839	국내배급사
8	6.718222	상장배급사
9	5.946032	배우가중치
10	5.854939	개봉일매출비율
11	5.729070	감독가중치
12	4.977840	1
13	4.084213	2
14	3.750760	개봉연도
15	3.197159	주요배급사
16	2.724068	3
17	2.644202	0
18	2.268158	국적

정규성확인 - 히스토그램

로그화의 이유(상용로그(log_{10})시 값들이 0,1,2로 줄어드는 효과를 발생
(minmaxscaling과 비슷한 효과(0~1))

Before log

After log

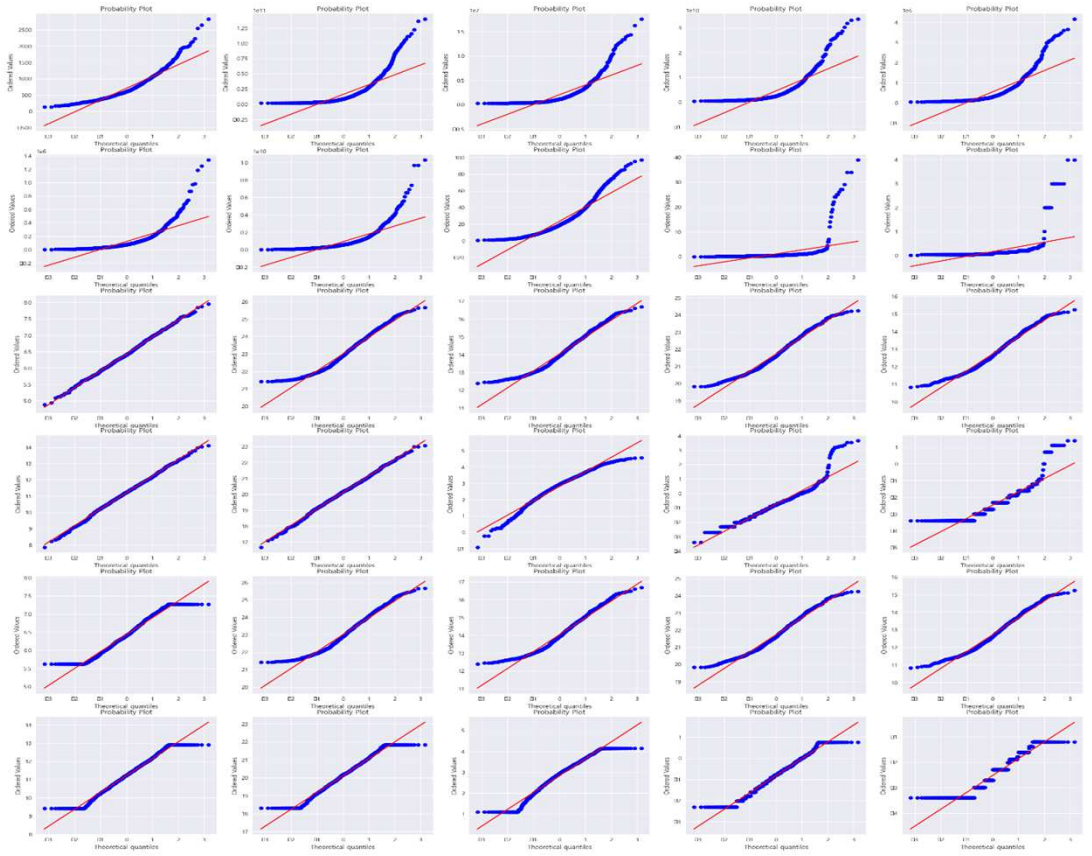


정규성 확인 - QQ plot

로그 전 데이터

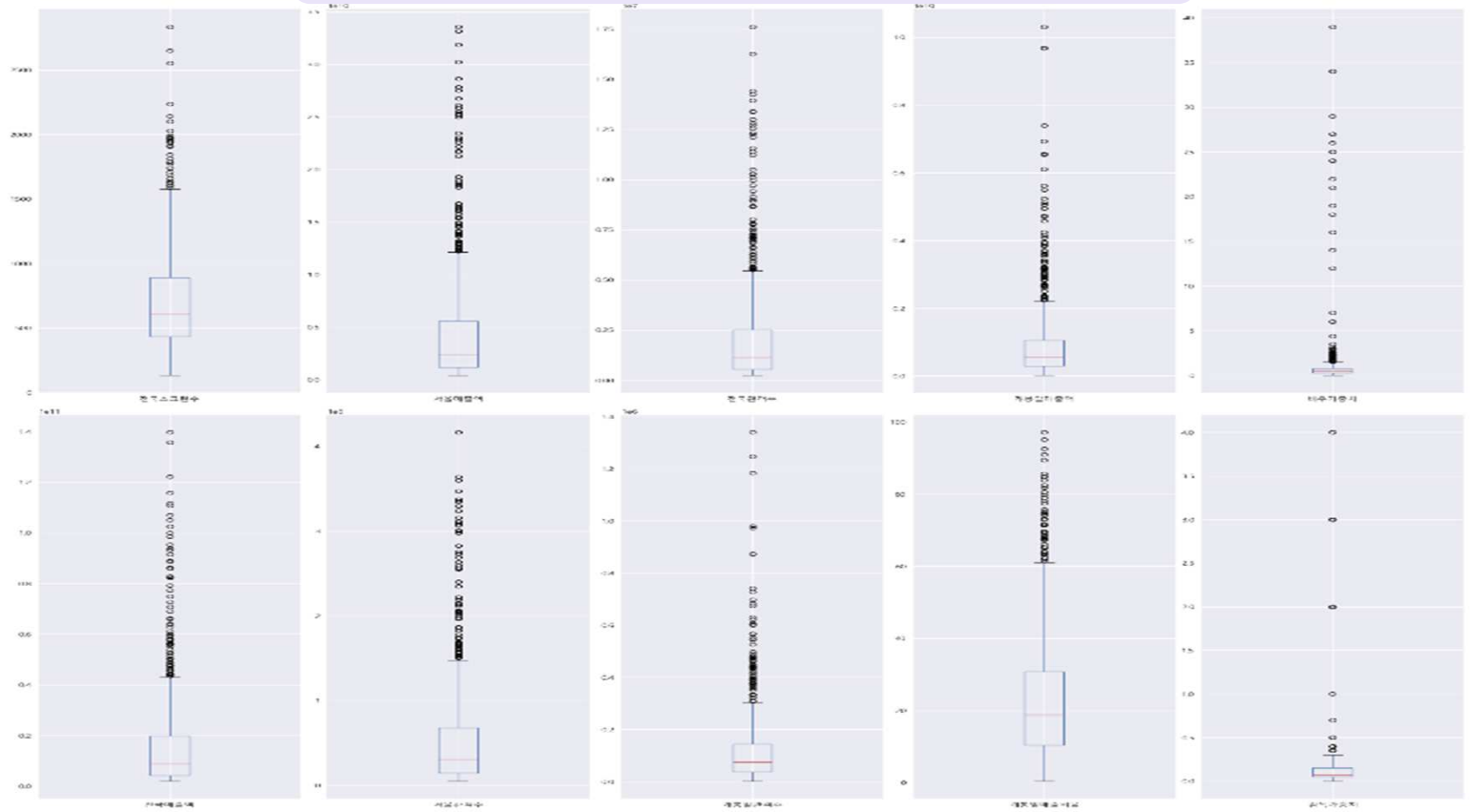
로그 후 데이터

Winsorization 후 데이터



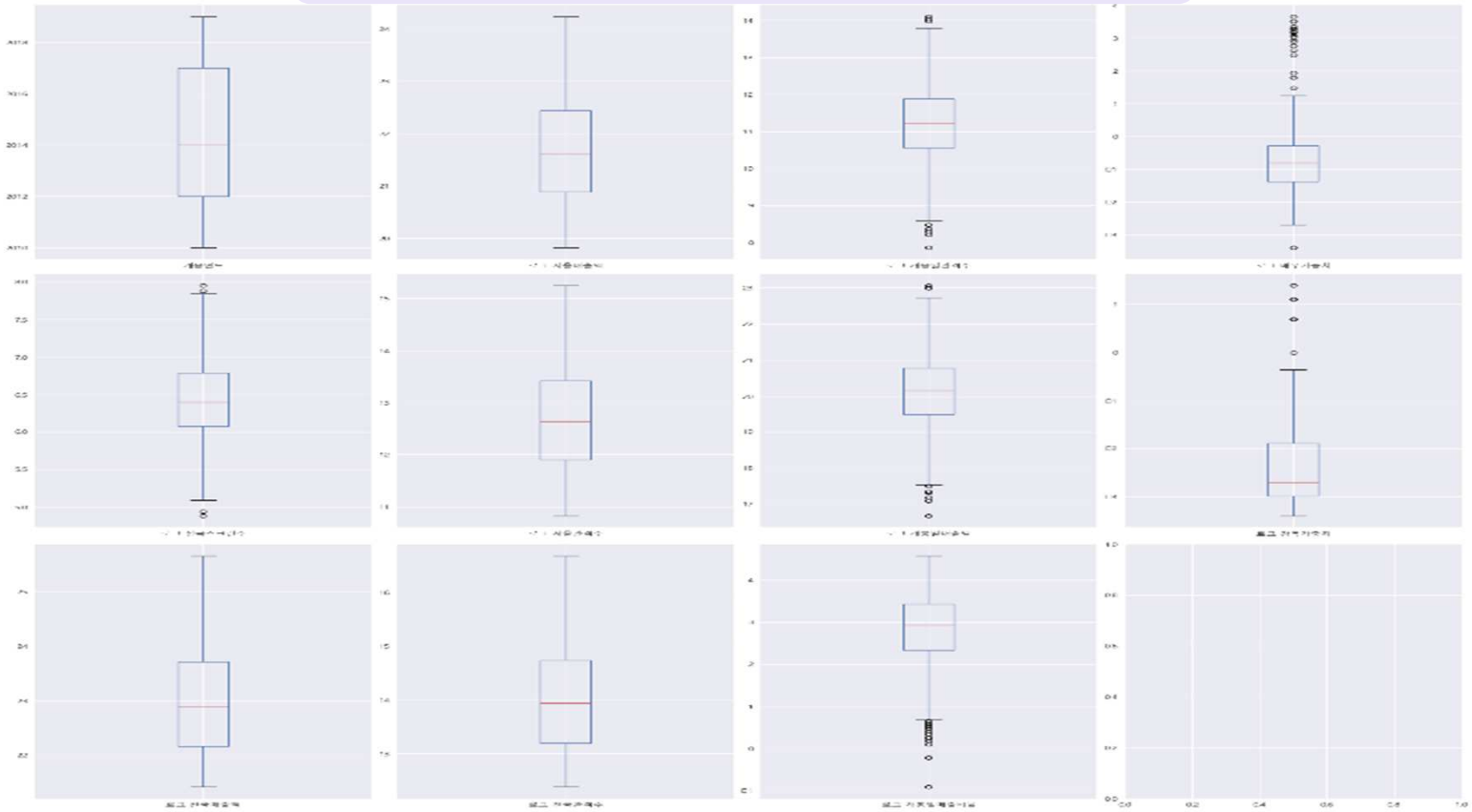
이상치 처리

기본 데이터



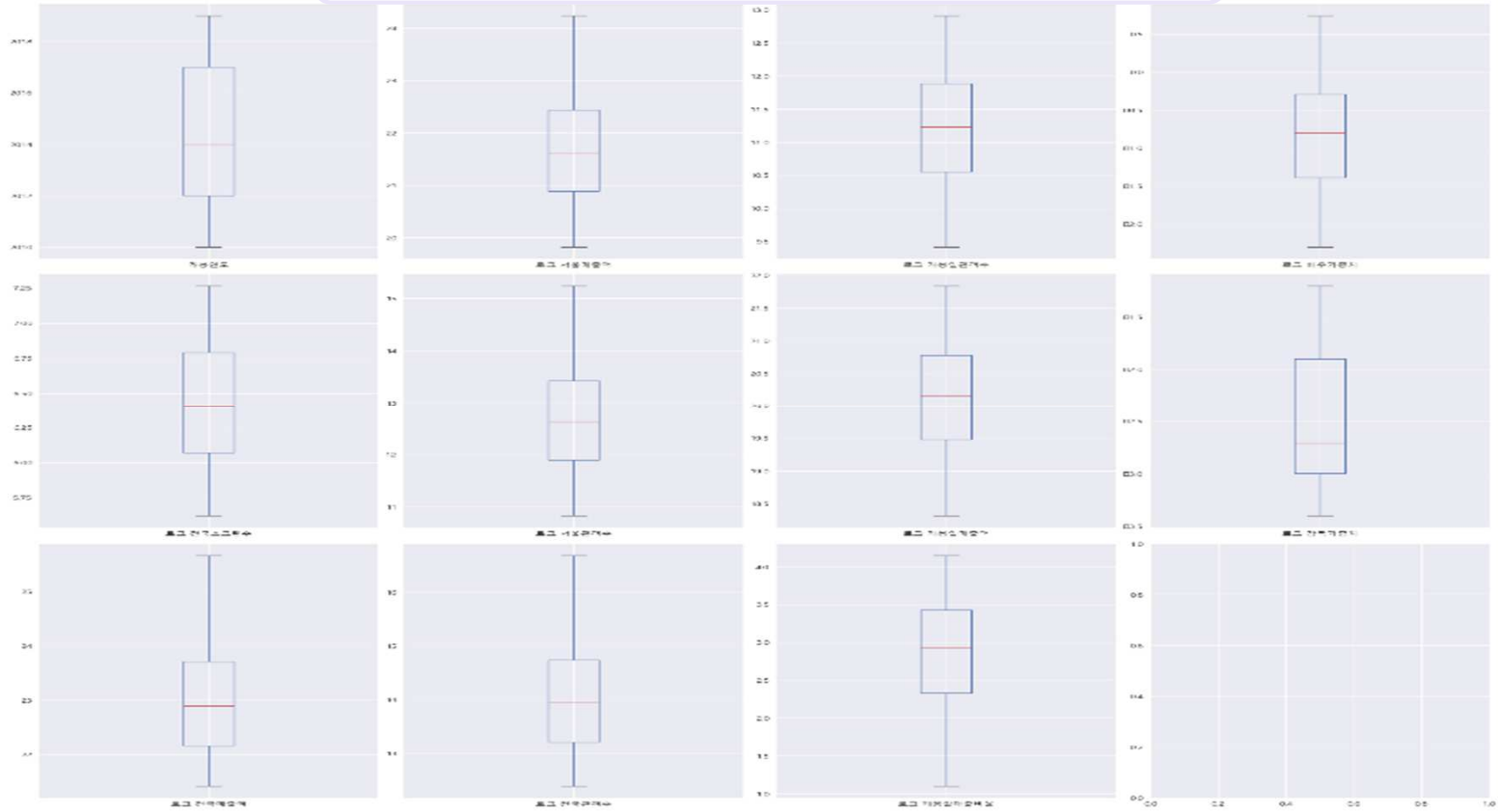
이상치 처리

로그화



이상치 처리

윈저라이징(범위 설정 설명)



타겟 설정

업종 지수화

KRX

업종	방송서비스
종목	CJ ENM

업종	오락 / 문화
종목	스튜디오 드래곤

네이버 금융

업종	방송과 엔터
종목	CJ ENM 스튜디오 드래곤

TS2000

업종	방송업
종목	CJ ENM

업종	영상 및 기록물 배급업
종목	스튜디오 드래곤

타겟 설정

업종 지수화

코스피(5종목)

CJ CGV
롯데쇼핑
아센디오
IHQ
콘텐트리중앙

코스닥(9종목)

CJ ENM
쇼박스
NEW
애니플러스
위지윅 스튜디오
덱스터
바른손이앤에이
판타지오
스튜디오 산타클로스

종목 선정 : 14개

타겟 설정

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

업종 지수화

<선정 이유>

KOSPI, KOSDAQ이 시가 총액식으로 구성

1. 주가지수 산정 방식

시가총액방식

$$\text{시가총액식 주가지수} = \frac{\text{비교시점의 시가총액}}{\text{기준시점의 시가총액}} \times 100$$

ex) KOSPI(Korea Composite Stock Price Index), S&P , Nasdaq

- KOSPI : 파셰식, 상장주식수 가중 방식
- KOSPI200 : 파셰식, 유동주식수 가중 방식

주가평균방식

$$\text{주가평균식 주가지수} = \frac{\text{비교시점의 수정주가 평균}}{\text{기준시점의 수정주가 평균}} \times 100$$

ex) 다우존스방식, 니케이225

타겟 설정

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

업종 지수화

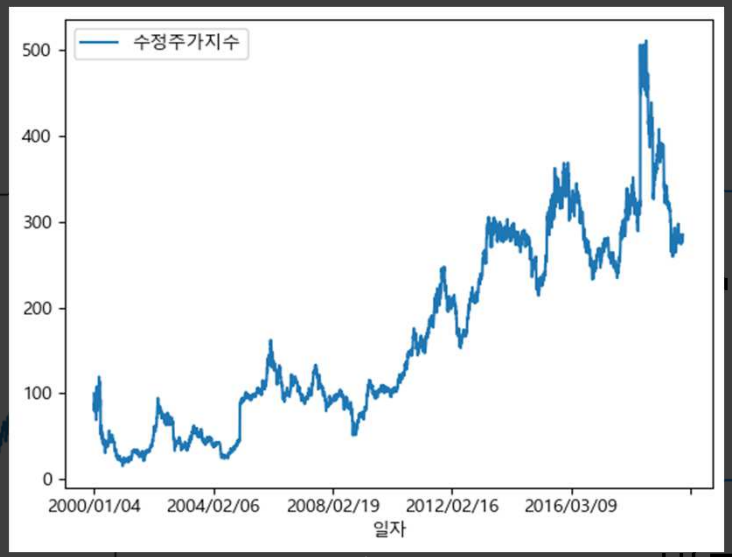
- 공식의 단순 대입을 통한 새로운 주가 지수 그래프 (2000~2019)

$$\frac{\text{비교시점의 시가총액}}{\text{기준시점의 시가총액}} \times 100$$

타겟 설정

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

업종 지수화



공식의 단순 대입을 통한
새로운 주가 지수 그래프
(2000~2019)

$$\frac{\text{비교시점의 시가총액}}{\text{기준시점의 시가총액}} \times 100$$

해당 일자에 상장된 기업수로 나눈 수정주가지수

$$\frac{\dfrac{\text{해당일자의 시가총액}}{\text{기준시점의 시가총액}} \times 100}{\text{해당일자의 기업개수} \times 4}$$

타겟 설정

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

결측치 처리

결측치 처리 시: 1) 오늘의 종가 다음의 시가의 변화율 계산하여 채움 2)결측치 삭제

선형 보간법

- 2개의 인접한 관측값에 대하여 그 사이에 위치한 값을 추정하기 위해 임의적인 직선을 가정하고 선형적으로 계산하는 방법
- 사용METHOD: interpolote()

선형 보간법 이유

- 영화데이터와 주가지수의 괴리를 줄이기 위해 선정

	날짜	관객수	수정주가지수
0	2014-07-30	682701	289.105978
1	2014-07-31	705070	286.584181
2	2014-08-01	867437	286.026735
3	2014-08-02	1232529	NaN
4	2014-08-03	1257380	NaN
5	2014-08-04	990022	289.639448
6	2014-08-05	869153	283.977458
7	2014-08-06	702887	284.349510
8	2014-08-07	652776	286.725349
9	2014-08-08	690123	285.683865

	날짜	관객수	수정주가지수
0	2014-07-30	682701	289.105978
1	2014-07-31	705070	286.584181
2	2014-08-01	867437	286.026735
3	2014-08-02	1232529	287.230973
4	2014-08-03	1257380	288.435211
5	2014-08-04	990022	289.639448
6	2014-08-05	869153	283.977458
7	2014-08-06	702887	284.349510
8	2014-08-07	652776	286.725349
9	2014-08-08	690123	285.683865

타겟 설정

시차 상관 분석

시차상관분석이란?

시차를 갖고 있는 두 시계열의 상관을 분석

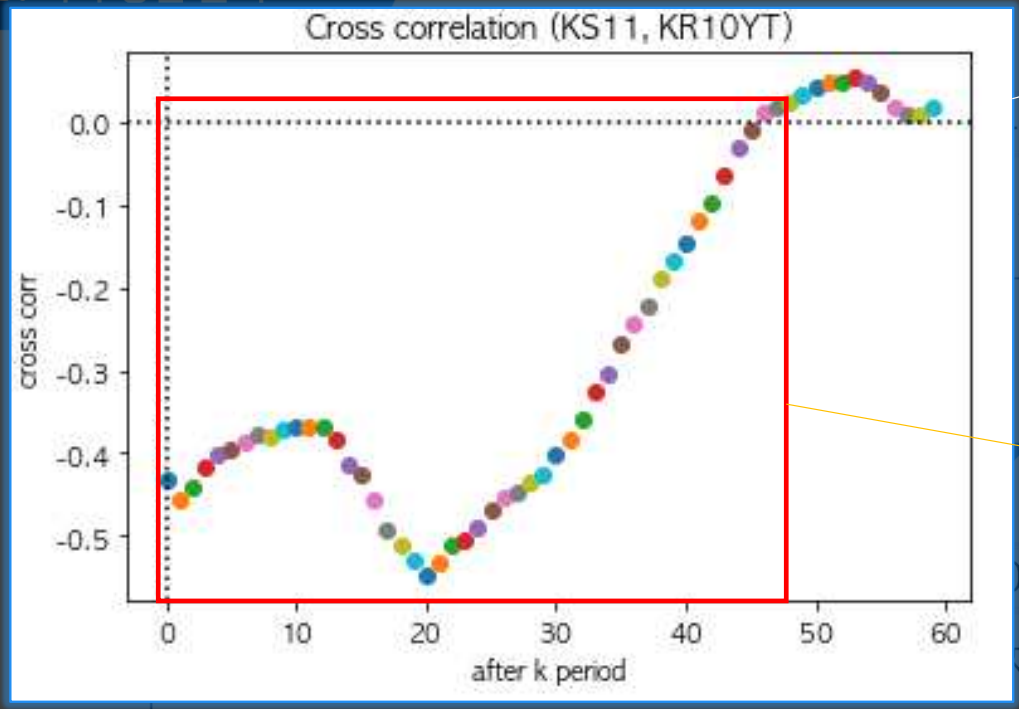
선정 이유

시차상관분석을 통해 '개별 영화'와 '업종 지수'의 상관계수를 통한 투자기간 설정

타겟 설정

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

C. 시차 상관 분석



사용 라이브러리

```
# 시차상관계수(비교상관계수) 분석
cc = sm.tsa.stattools.ccf(a1, a2, adjusted=False, fft=False)
```

Statsmodels의 api에서 추출한 함수, ccf

그래프의 해석

X축 : 시간(일)
Y축 : 상관계수

시간의 흐름에 따른 상관계수의 변화

석이란 :

계열의 상관을 분석

이유

지수'의 상관계수를 통한 투자기간 설정

타겟 설정

Enjoy your stylish business and campus life with BIZCAM

시차 상관 분석

산술평균

데이터 값들의 대표값을 산출할 때 주로 사용하는 기법입니다. 이 기법은 데이터 분포가 종 모양으로 중앙 근처에서 많이 분포하고 양끝단에서 작아지는 유형에 적합(실생활에서의 평균)

기하평균

인구증가률, 물가상승률, 경제성장률 등과 같이 연속적인 변화율 데이터를 기반으로 어느 구간에서의 평균 변화율을 구할 때 사용하는 것(변화율들에 대한 평균 변화)

조화평균

구간별 데이터 값의 역수를 취하여 산술 평균을 구한 후에 다시 역수를 취하여 구한다. 평균속력을 구할때 사용(구간별 속력 값으로부터 전체의 평균 속력)

시차 상관 분석

- 상관계수들의 평균값 : 12(기하평균)
- 사용 라이브러리 : *from scipy.stats import gmean*

기하평균

인구증가률, 물가상승률, 경제성장률 등과 같이 연속적인 변화율 데이터를 기반으로 어느 구간에서의 평균 변화율을 구할 때 사용하는 것(연속적 데이터들에 대한 평균 변화)

조화평균

구간별 데이터 값의 역수를 취하여 산술 평균을 구한 후에 다시 역수를 취하여 구한다.
평균속력을 구할때 사용(구간별 속력 값으로부터 전체의 평균 속력)

타겟 설정

Enjoy your stylish business and campus life with BIZCAM

시차 상관 분석

산술평균

데이터 값들의 대표값을 산출할 때 주로 사용하는 기법입니다. 이 기법은 데이터 분포가 종 모양으로 중앙 근처에서 많이 분포하고 양끝단에서 작아지는 유형에 적합(실생활에서의 평균)

12일 영화업종 투자수익률 > 12일 코스닥 수익률 → 1

12일 영화업종 투자수익률 < 12일 코스닥 수익률 → 0

조화평균

구간별 데이터 값의 역수를 취하여 산술 평균을 구한 후에 다시 역수를 취하여 구한다.
평균속력을 구할때 사용(구간별 속력 값으로부터 전체의 평균 속력)

타겟 설정

Enjoy your stylish business and campus life with BIZCAM

최종 타겟 설정

시차상관분석

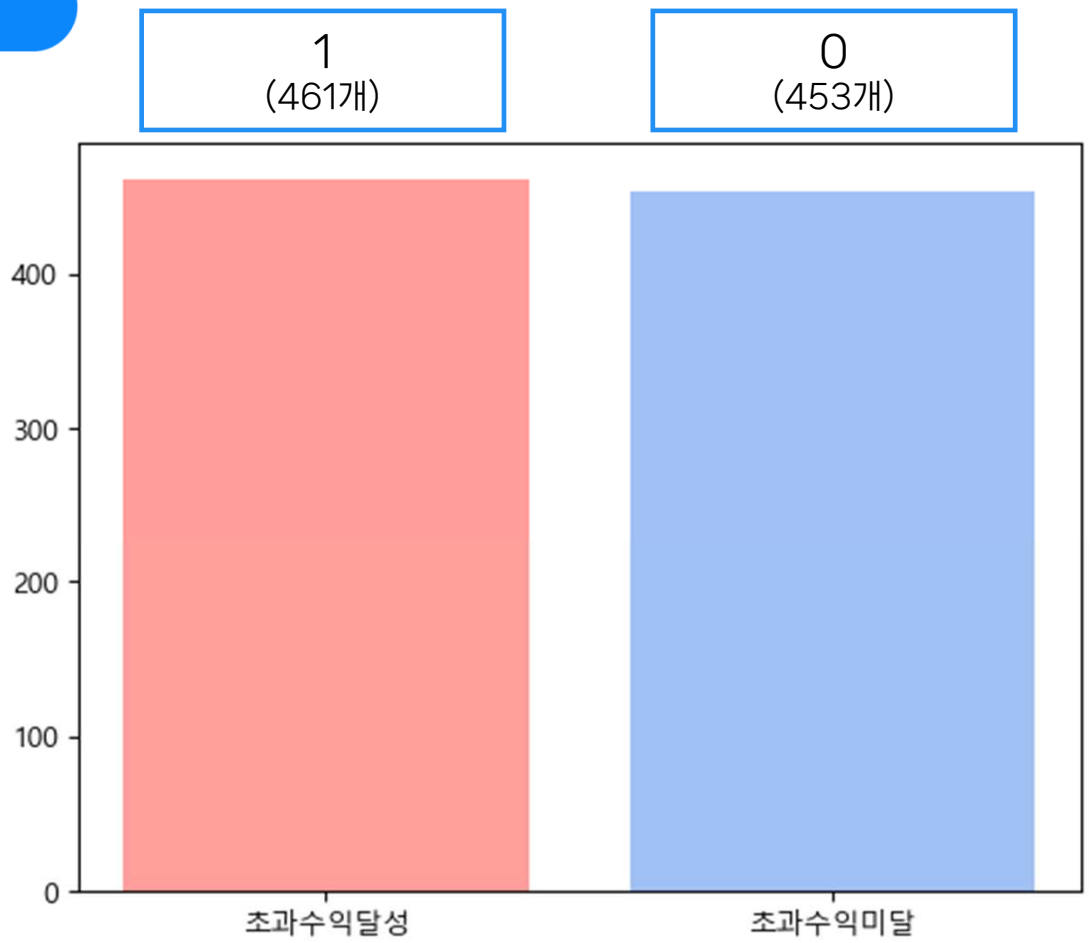
영화산업지수 15일 수익률 > 코스닥 15일 수익률 : 1

영화산업지수 15일 수익률 < 코스닥 15일 수익률 : 0

타겟 설정

Enjoy your stylish business and campus life with BIZCAM

최종 타겟 설정



정규성 검정

Feature Selection

피쳐 선정

정규성 검정

Shapiro	Anderson Test	KS Test	Jarque-Bera	Normal Test
-	-	로그 전국스크린수 로그 개봉일관객수 로그 개봉일매출액 로그 배우가중치	-	-

표본 데이터 수는 914개로 위의 정규성 검정이 만족하지 않더라도

중심극한정리에 따라 정규분포를 따른다고 가정

피쳐 선정

Feature Selection

T-test (5개)

2
로그 감독가중치
로그 전국관객수
로그 전국매출액
로그 전국스크린수

Stepwise (2개)

2
로그 감독가중치

Lasso (14개)

개봉연도
로그 전국스크린수
로그 서울매출액
로그 개봉일관객수
로그 개봉일매출비율
로그 배우가중치
로그 감독가중치
1
2
3
국적
국내배급사
상장배급사
주요배급사

SelectKbest (13개)

로그 전국스크린수
로그 전국매출액
로그 서울매출액
로그 전국관객수
로그 서울관객수
로그 개봉일관객수
로그 개봉일매출액
로그 배우가중치
로그 감독가중치
0
2
국적
국내배급사

피쳐 선정

최종 피쳐 선정

피쳐 선정 (4가지 방법)

	Feature	ttest	stepwise	lasso	kbest	total
0	국적	0	0	1	1	2
2	국내배급사	0	0	1	1	2
6	로그 전국스크린수	1	0	1	1	3
7	로그 전국매출액	1	0	0	1	2
8	로그 서울매출액	0	0	1	1	2
9	로그 전국관객수	1	0	0	1	2
11	로그 개봉일관객수	0	0	1	1	2
14	로그 배우가중치	0	0	1	1	2
15	로그 감독가중치	1	1	1	1	4
18	2	1	1	1	1	4

라쏘 선정

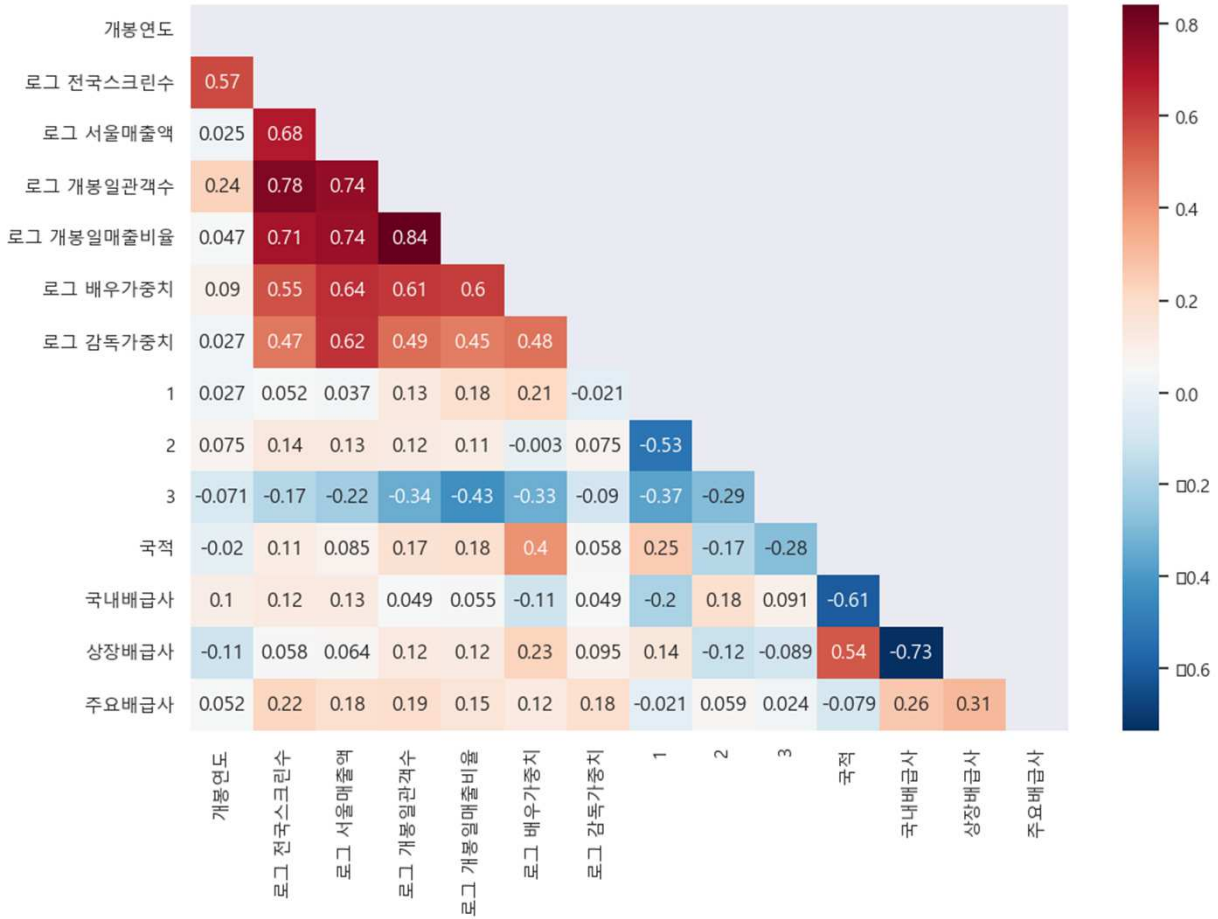
	feature	coef
0	개봉연도	1.229336
1	로그 전국스크린수	-2.512595
3	로그 서울매출액	0.579889
6	로그 개봉일관객수	-0.870024
8	로그 개봉일매출비율	2.568619
9	로그 배우가중치	0.341002
10	로그 감독가중치	-0.500918
12	1	-0.281264
13	2	-0.452264
14	3	0.185219
15	국적	-0.518649
16	국내배급사	-1.134638
17	상장배급사	-0.596875
18	주요배급사	0.911597

최종 14개 선정

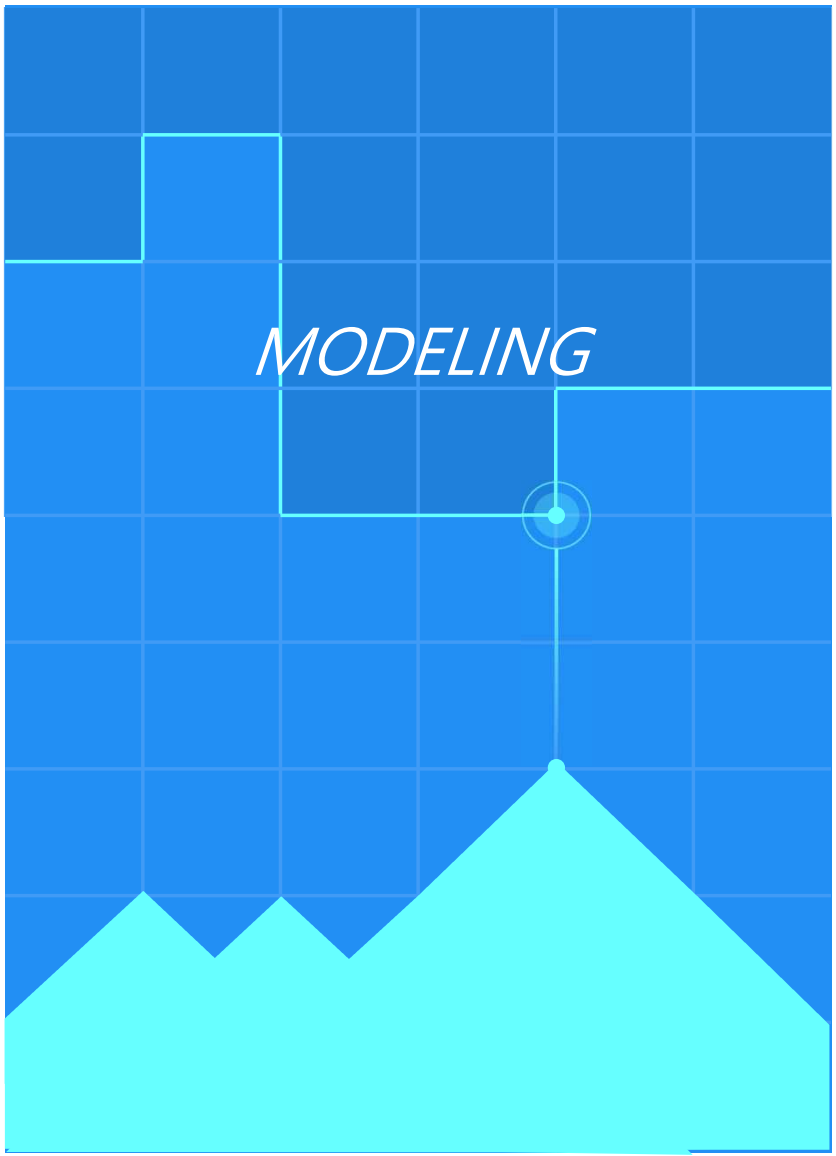
피쳐 선정

Literacy - The Impact of Movie Releases on the "Movie Industry Index"

Heatmap, VIF



	VIF Factor	features
0	23.024088	로그 개봉일매출비율
1	21.605904	주요배급사
2	21.110937	로그 개봉일관객수
3	19.543428	로그 전국스크린수
4	17.036575	로그 서울매출액
5	11.958641	상장배급사
6	10.142176	로그 배우가중치
7	8.758550	국내배급사
8	7.096156	개봉연도
9	4.418050	로그 감독가중치
10	3.591652	국적
11	3.295489	1
12	2.645349	2
13	1.841192	3



데이터 분리 및 Resampling

모델링

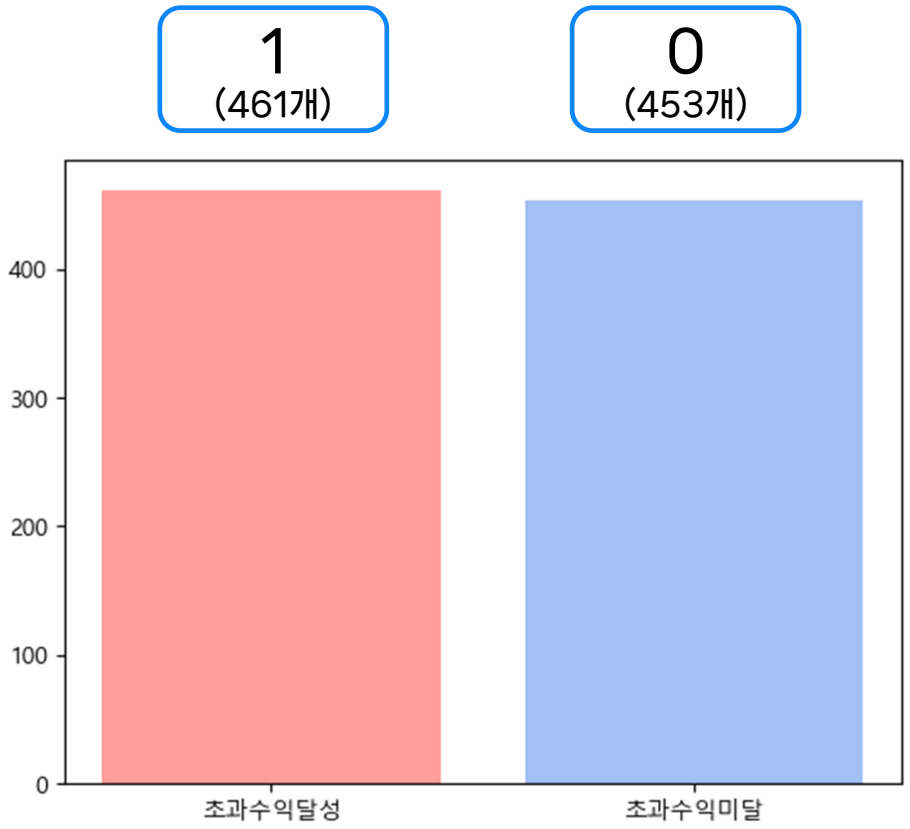
최적 모델링 선정

데이터 분리

Train Set	Test Set
7 (639개)	3 (275개)

영화데이터의 경우 시계열데이터가 아니기 때문에 임의분리
Test Set 활용 백테스팅 진행

Resampling



불균형이 심하지 않아
Resampling 실시 X

분류 모델

단일분류

Logistic
KNN
Decision Tree
SVC

앙상블

Random Forest
XGB

GridsearchCV(조정 전)

	Train set		Test set				
	ACC	AUC	ACC	Precision	Recall	F1-score	AUC
Logistic	0.58	0.57	0.49	0.43	0.55	0.48	0.50
KNN	0.70	0.70	0.53	0.47	0.62	0.53	0.54
Decision Tree	1.0	1.0	0.49	0.41	0.39	0.40	0.48
SVC	0.60	0.60	0.45	0.39	0.49	0.44	0.46
Random Forest	1.0	1.0	0.51	0.44	0.48	0.46	0.51
XGB	1.0	1.0	0.49	0.41	0.39	0.40	0.48

모델별 최적 파라미터

Logistic	C : 5, penalty : L2, solver : lbfgs, max_iter : 200, multi_class : auto
DecisionTree	criterion : gini, max_depth : 9, max_features : sqrt, min_samples_leaf : 1, min_samples_split : 16, splitter : best
RandomForest	criterion : log_loss, max_depth : 12, max_features : log2, max_leaf_nodes = 9, n_estimators : 25
XGB	min_child_weight : 1, gamma : 3, subsample = 0.5, colsample_bytree : 0.5, max_depth : 9, learning_rate : 0.2

파라미터 조정 후

	Train set		Test set				
	ACC	AUC	ACC	Precision	Recall	F1-score	AUC
Logistic	0.55	0.55	0.61	0.61	0.60	0.61	0.61
Decision Tree	0.65	0.65	0.50	0.50	0.73	0.60	0.50
Random Forest	0.71	0.71	0.58	0.58	0.60	0.59	0.55
XGB	0.89	0.89	0.64	0.64	0.64	0.64	0.64

모델 최종평가

최적 하이퍼 파라미터 대입전에 비해 성능이 향상됨

6개의 모델 중 ACC, F1-score 모두 **XGB classifier**가 가장 뛰어남

XGB 결과값 해석 : 영화 데이터는 여러 특성이 상호작용을 갖는 경우가 많고,
XGB의 모델은 특성 상호작용을 고려하여 모델링 수행하기 때문에 높은 예측 값이 나온 것으로 판단.

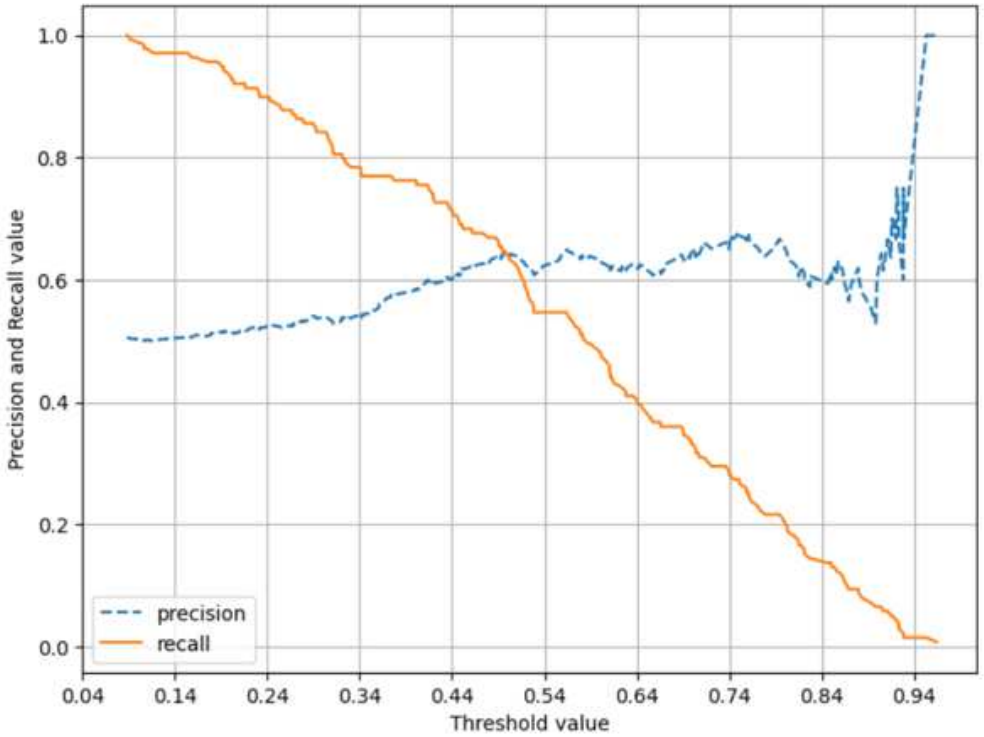
DecisionTree 결과값 해석 : 데이터 수가 적을 수록 불안정하며, 새로운 자료에 대한 예측력이 떨어진다.

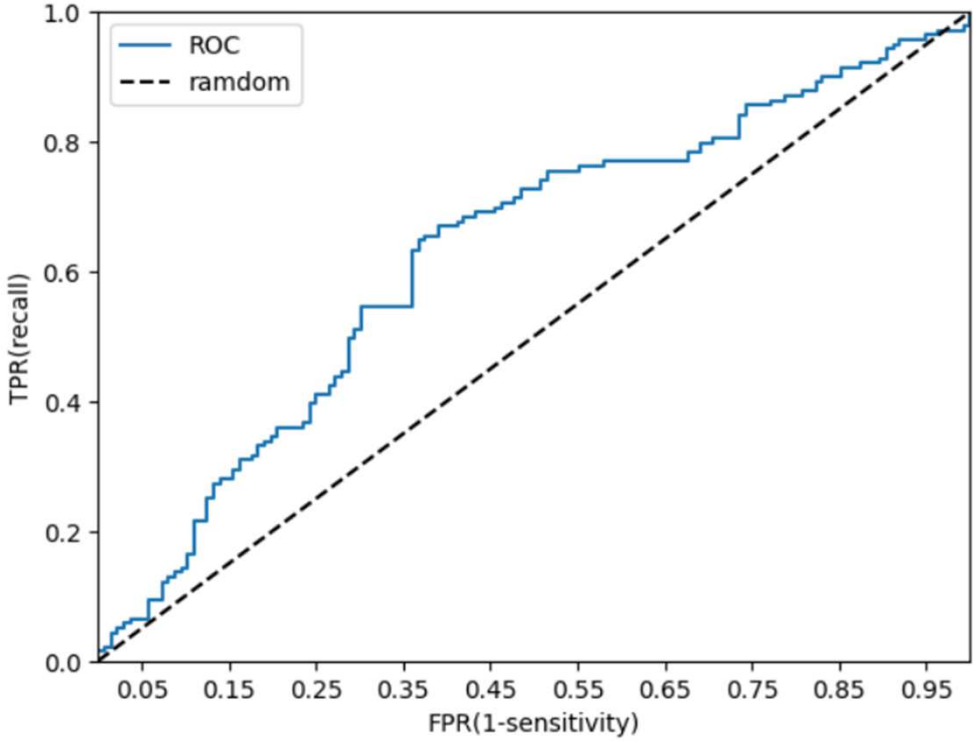
파라미터 조정 후

	Train set		Test set				
	ACC	AUC	ACC	Precision	Recall	F1-score	AUC
Logistic	0.64					0.61	0.60
Decision Tree	0.70	0.70	0.56	0.57	0.53	0.55	0.55
Random Forest	0.70	0.70	0.58	0.58	0.60	0.59	0.59
XGB	0.89	0.89	0.64	0.64	0.64	0.64	0.64

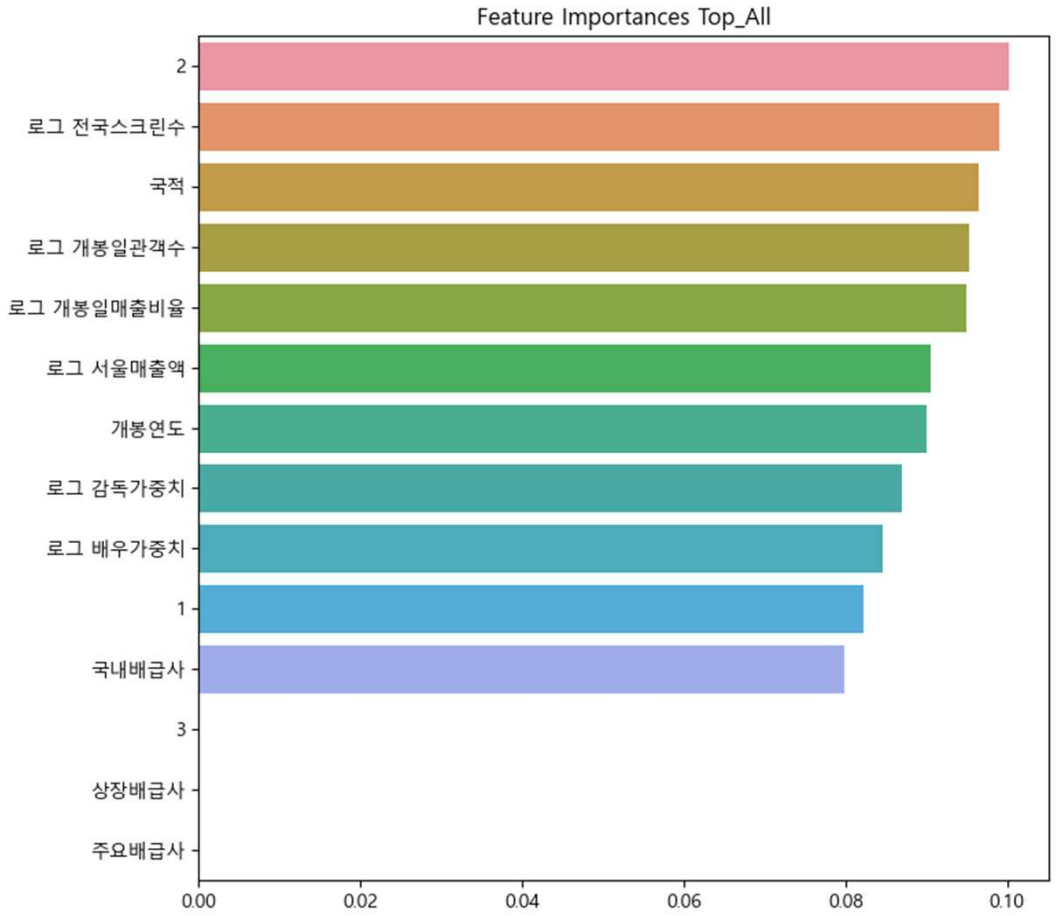
모델선정 : XGB Classifier

XGB


Precision and Recall value
Threshold value
precision
recall
1.0
0.8
0.6
0.4
0.2
0.0
0.04
0.14
0.24
0.34
0.44
0.54
0.64
0.74
0.84
0.94


TPR(recall)
FPR(1-sensitivity)
ROC
ramdom
1.0
0.8
0.6
0.4
0.2
0.0
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95

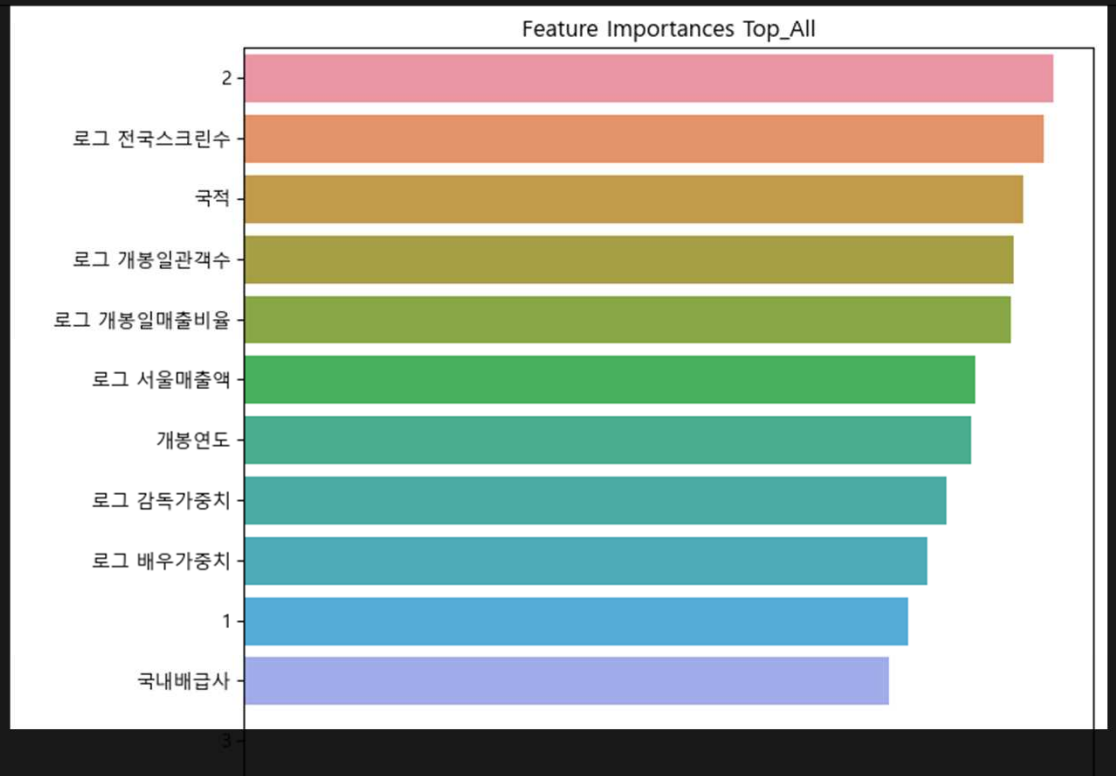
피쳐 임포턴스


Feature Importances Top_All
2
로그 전국스크린수
국적
로그 개봉일관객수
로그 개봉일매출비율
로그 서울매출액
개봉연도
로그 감독가중치
로그 배우가중치
1
국내배급사
3
상장배급사
주요배급사
0.00
0.02
0.04
0.06
0.08
0.10

MODELING

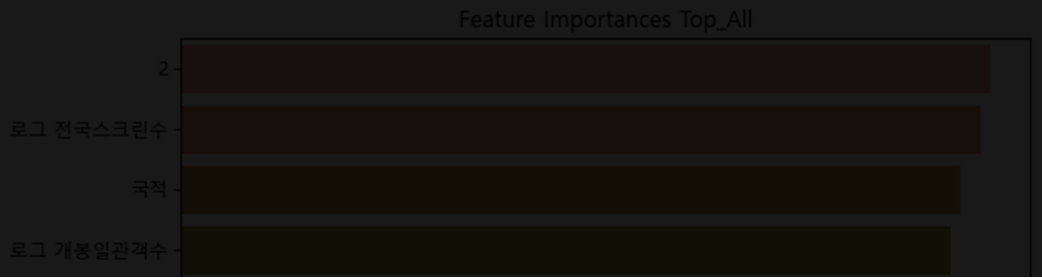
Literacy - The Impact of Movie Releases on the "Movie Industry Index"

피쳐 임포턴스

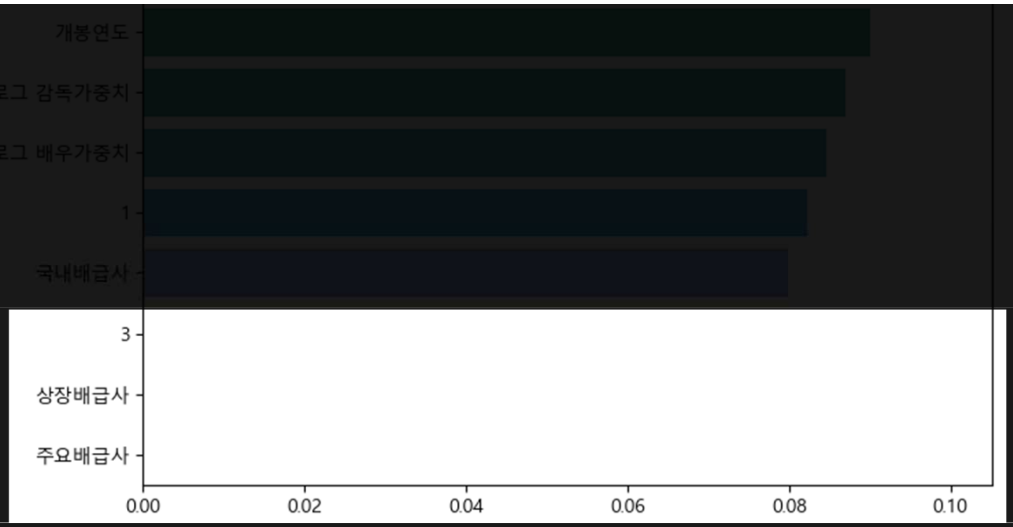


피쳐들 간의 중요도 차이가 크지 않아 특정 피쳐의 의존도가 높지 않음

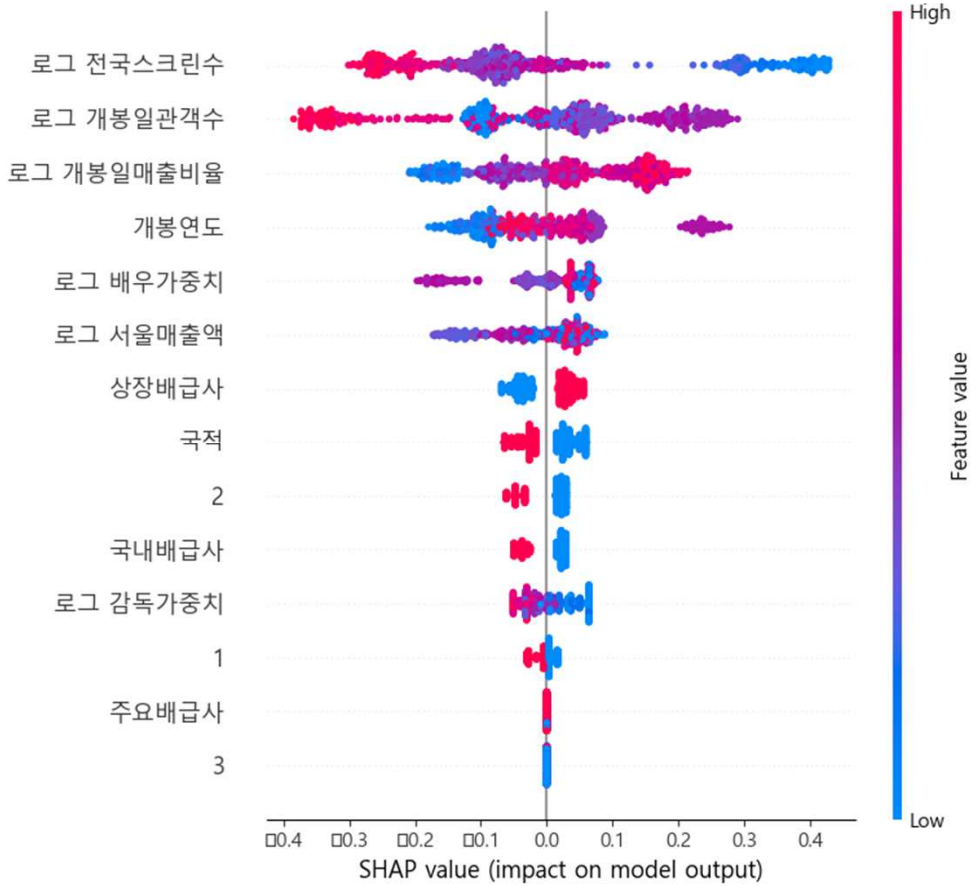
피쳐 임포턴스



라쏘 피쳐선정의 경우 회귀기반, XGB의 경우 분류기반으로 인한 차이로 판단



Shapley value

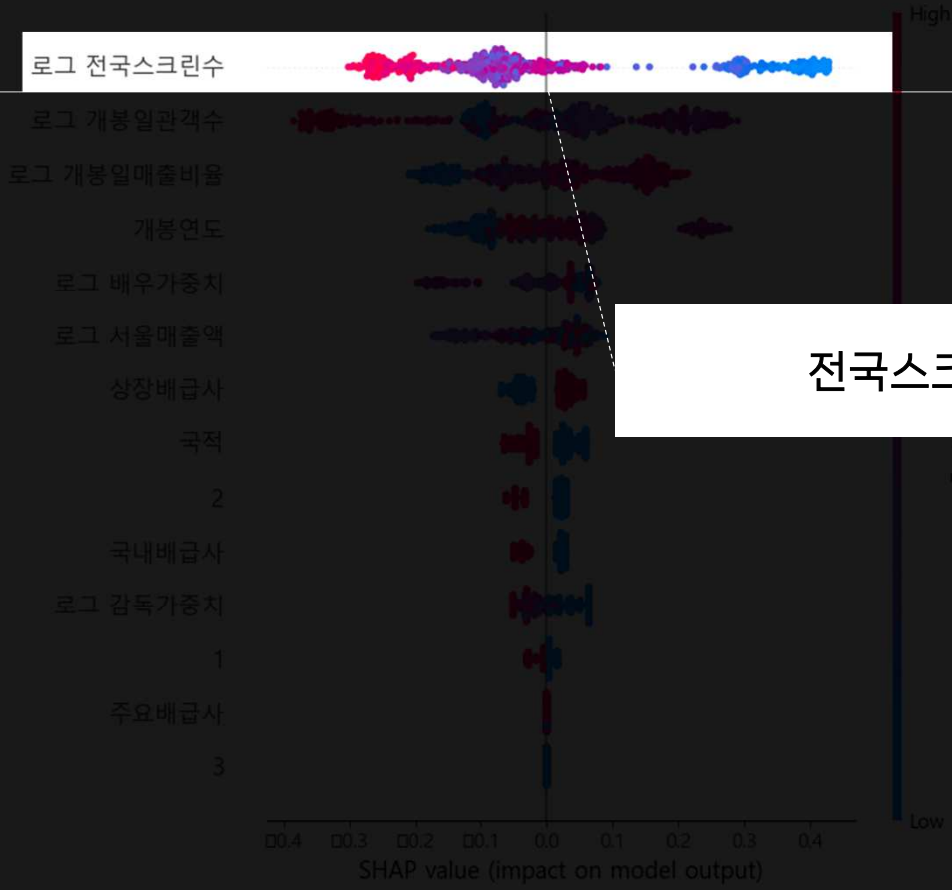


예측에 영향을 준 순서대로 나열

X축은 예측 영향력을 의미(0기준)
좌측일 수록 예측 결과 감소에 기여
우측일 수록 예측 결과 향상에 기여

빨강색은 높은값을 의미
파란색은 낮은값을 의미

Shapley value



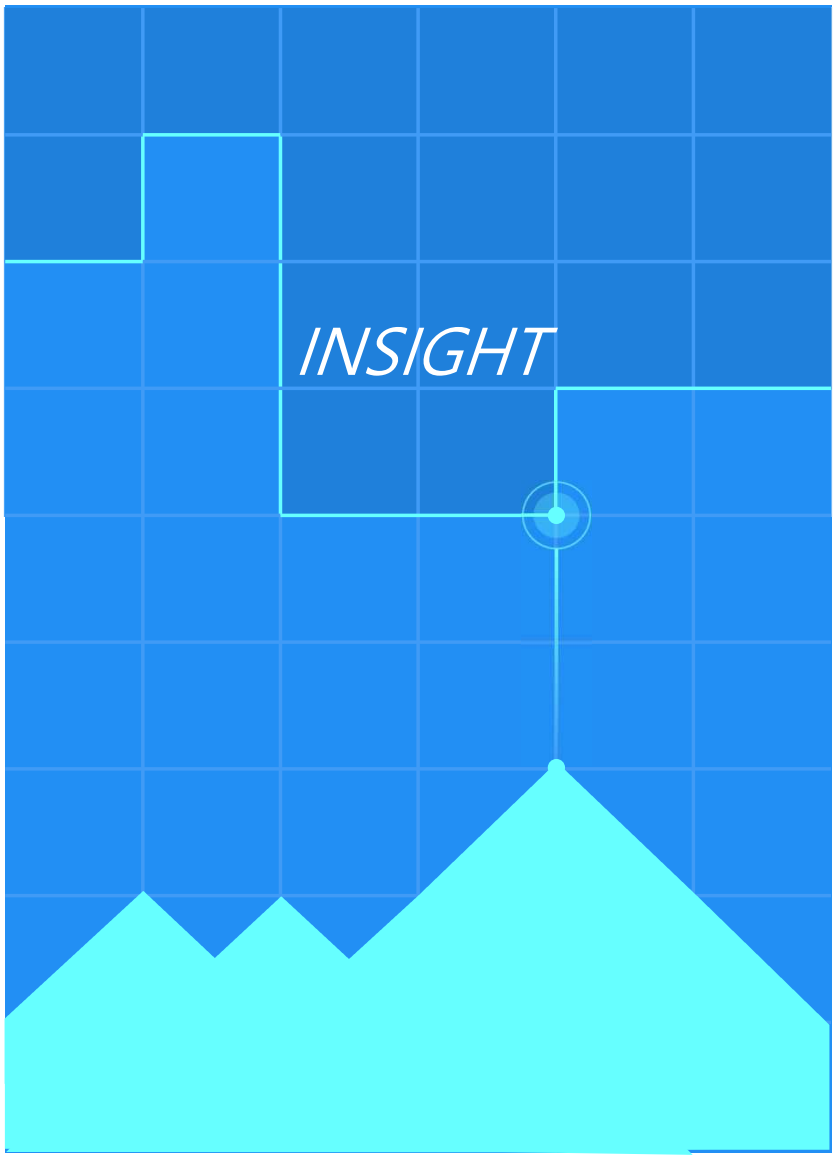
전국스크린수가 낮을 수록 예측결과 향상에 기여

예측에 영향을 준 순서대로 나열

우측일 수록 예측 결과 향상에 기여

빨강색은 높은값을 의미

파란색은 낮은값을 의미



백테스팅

EDA 인사이트

백테스팅

	1000만 초과	900만 초과	800만 초과	700만 초과	600만 초과	500만 초과	500만 이하
시행 횟수	10	10	10	10	10	10	10
평균 정확도	0.645	0.658	0.652	0.618	0.622	0.613	0.531

	1000만 초과	900만 초과	800만 초과	700만 초과	600만 초과	500만 초과	500만 이하
시행 횟수	100	100	100	100	100	100	100
평균 정확도	0.663	0.664	0.647	0.594	0.588	0.598	0.526

백테스팅

	1000만 초과	900만 초과	800만 초과	700만 초과	600만 초과	500만 초과	500만 이하
시행 횟수	1000	1000	1000	1000	1000	1000	1000
평균 정확도	0.655	0.649	0.635	0.584	0.588	0.593	0.526

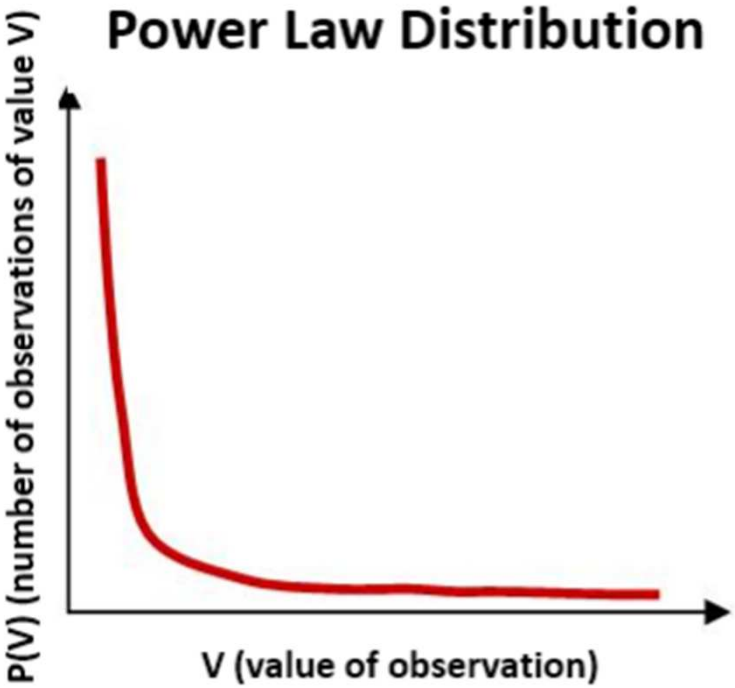
	1000만 초과	900만 초과	800만 초과	700만 초과	600만 초과	500만 초과	500만 이하
시행 횟수	5000	5000	5000	5000	5000	5000	5000
평균 정확도	0.653	0.649	0.635	0.584	0.588	0.591	0.526

백테스팅

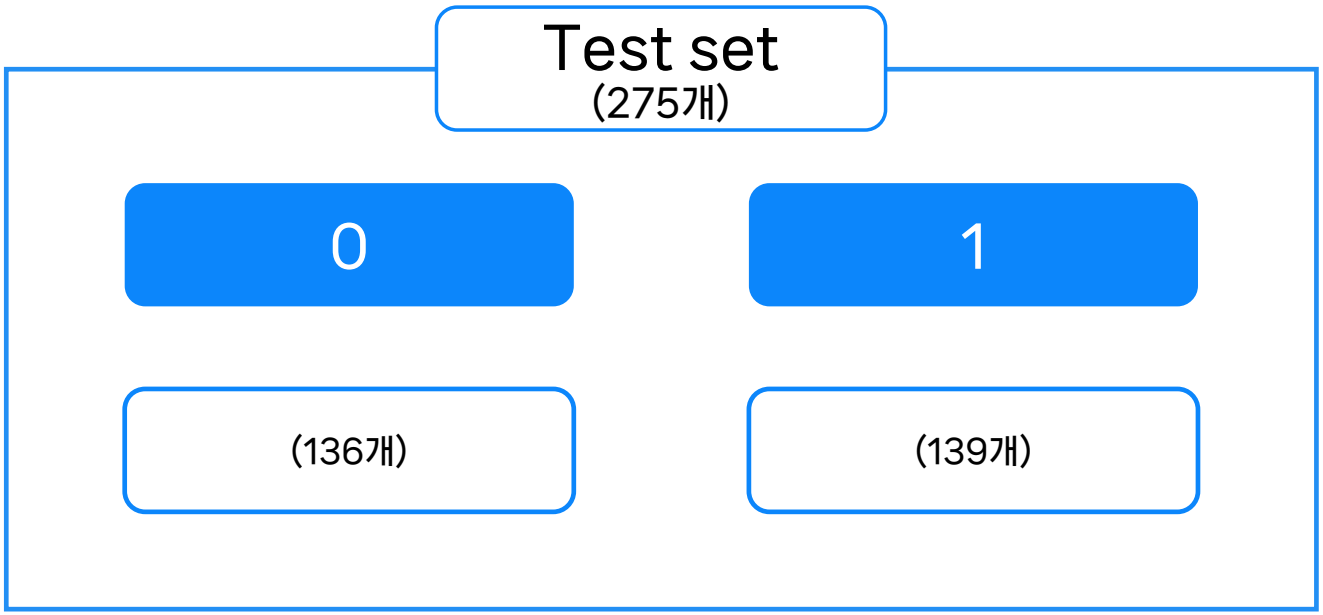
	1000만 초과	900만 초과	800만 초과	700만 초과	600만 초과	500만 초과	500만 이하
시행 횟수	1000	1000	1000	1000	1000	1000	1000
평균 정확도	0.655	0.649	0.635	0.584	0.588	0.593	0.526

	1000만 초과	900만 초과	800만 초과	700만 초과	600만 초과	500만 초과	500만 이하
시행 횟수	5000	5000	5000	5000	5000	5000	5000
평균 정확도	0.653	0.649	0.635	0.584	0.588	0.591	0.526

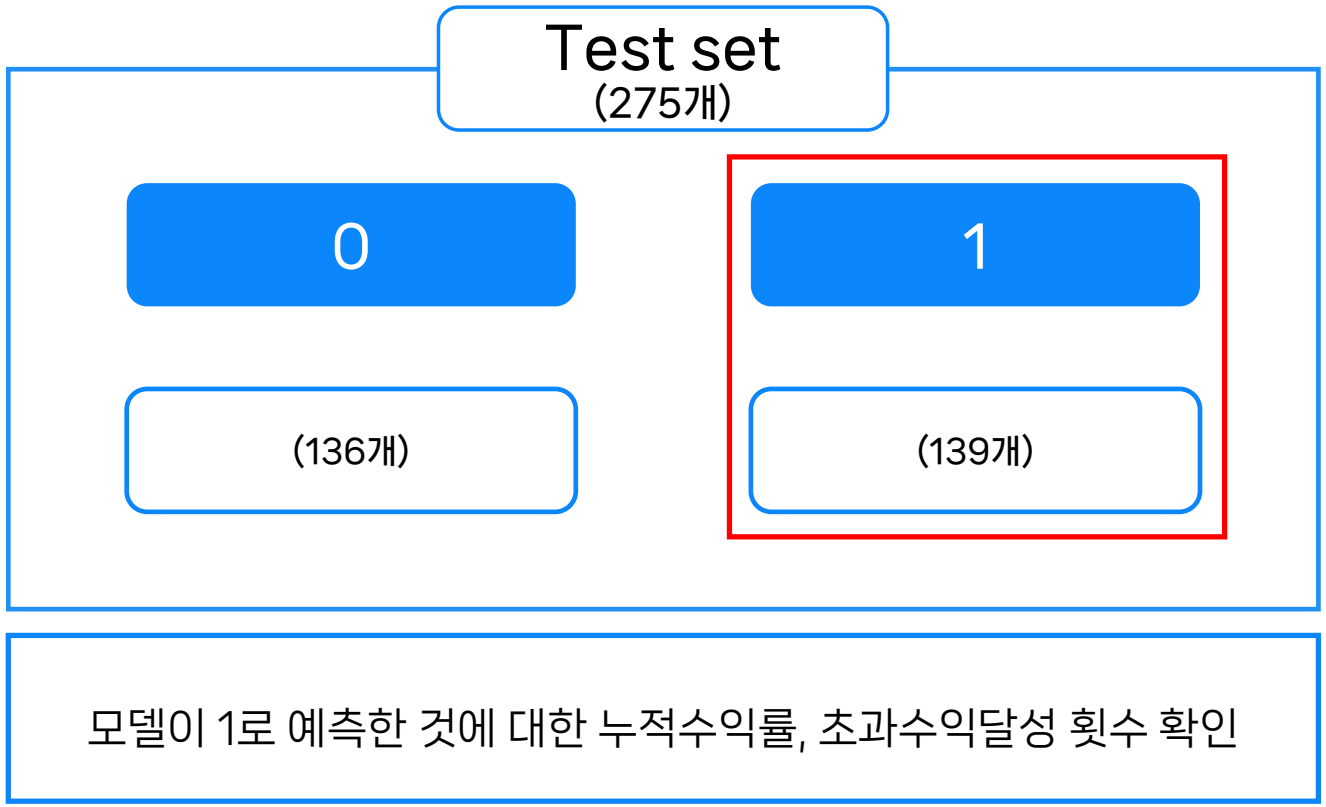
백테스팅 결과해석


Power Law Distribution
P(V) (number of observations of value V)
V (value of observation)

백테스팅



백테스팅



백테스팅 - 누적수익률

	타겟	예측값	영화누적수익률	코스닥누적수익률	초과수익여부
0	1	1	-0.010482	-0.034579	True
1	1	1	0.013565	-0.062699	True
2	1	1	0.038626	-0.044095	True
3	0	1	0.029705	-0.049098	False
4	1	1	0.073785	-0.059655	True
...	...	...	...	...	...
134	1	1	1.127189	-0.589678	True
135	1	1	1.084413	-0.629122	True
136	0	1	1.092223	-0.646178	True
137	1	1	1.155667	-0.640715	True
138	0	1	1.069188	-0.652093	False

139 rows × 5 columns

<누적수익률>

영화산업지수 : +1.06%

코스닥 : -0.65%

백테스팅 - 누적수익률

	타겟	예측값	영화누적수익률	코스닥누적수익률	초과수익여부
0	1	1	-0.010482	-0.034579	True
1	1	1	0.013565	-0.062699	True
2	1	1	0.038626	-0.044095	True
3	0	1	0.029705	-0.049098	False
4	1	1	0.073785	-0.059655	True
...	...	...	...	...	...
134	1	1	1.127189	-0.589678	True
135	1	1	1.084413	-0.629122	True
136	0	1	1.092223	-0.646178	True
137	1	1	1.155667	-0.640715	True
138	0	1	1.069188	-0.652093	False

139 rows × 5 columns

<초과수익달성 횟수>

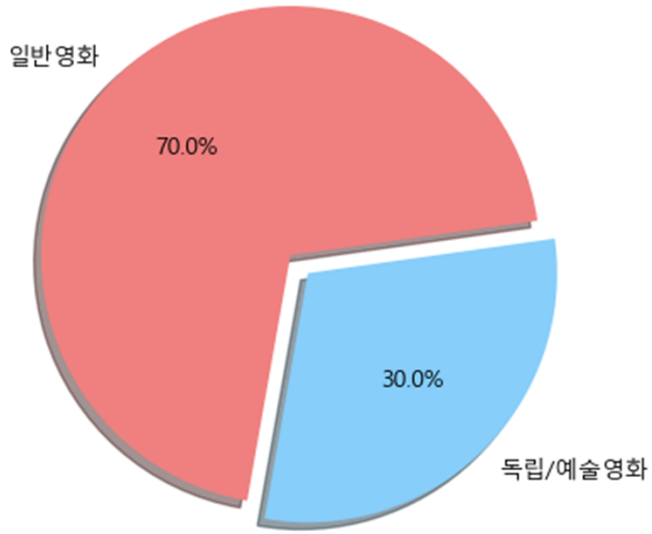
영화산업지수 : 91회

코스닥 : 48회

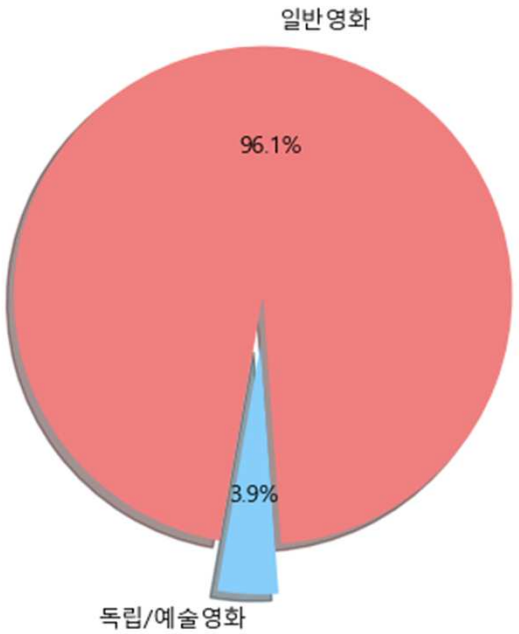
승률 : 65.4%

인사이트(EDA) - 구분

원본(11,819)

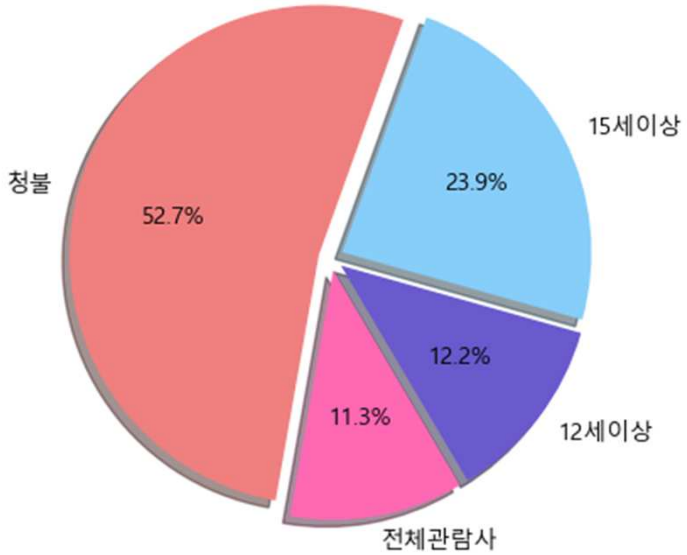


추출(914)

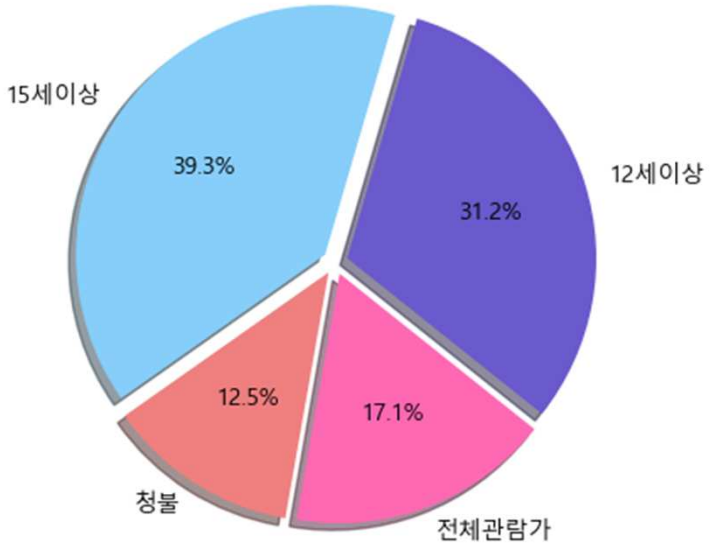


인사이트(EDA) - 등급

원본(11,819)

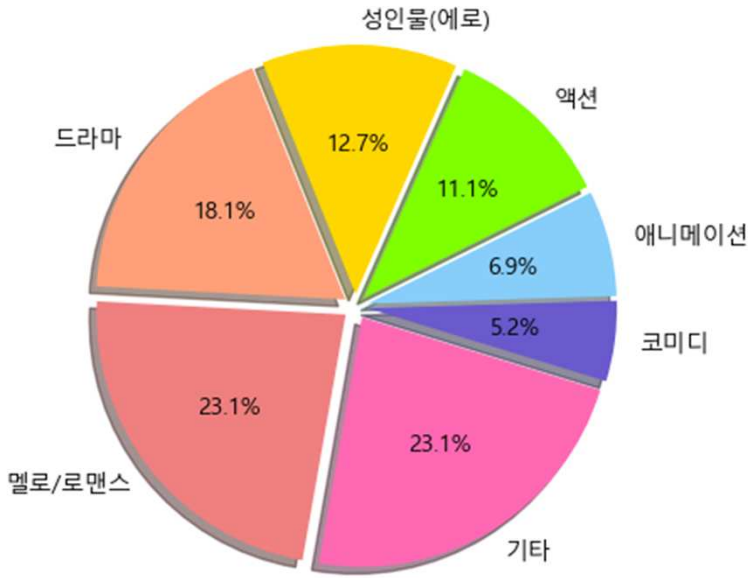


추출(914)

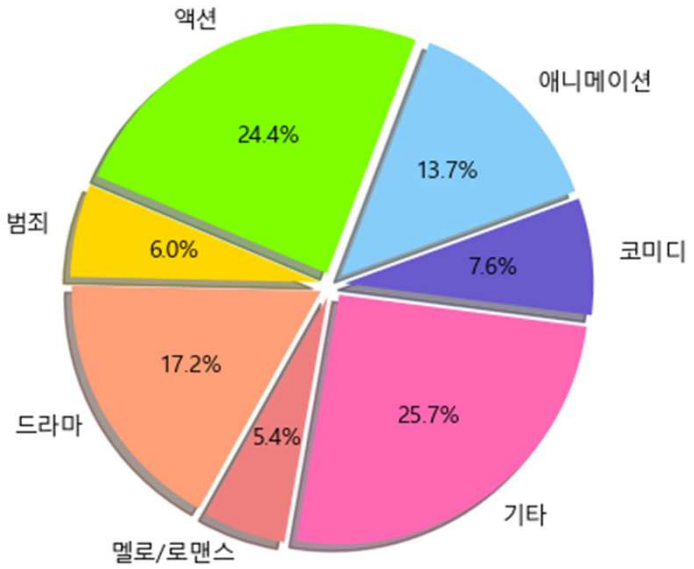


인사이트(EDA) - 세부장르

원본(11,819)

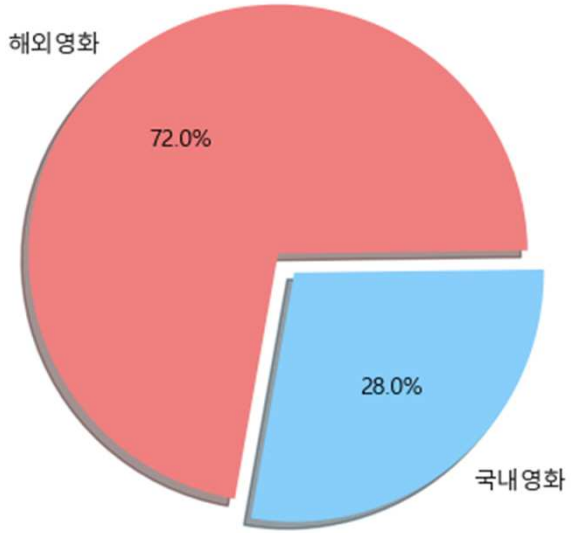


추출(914)

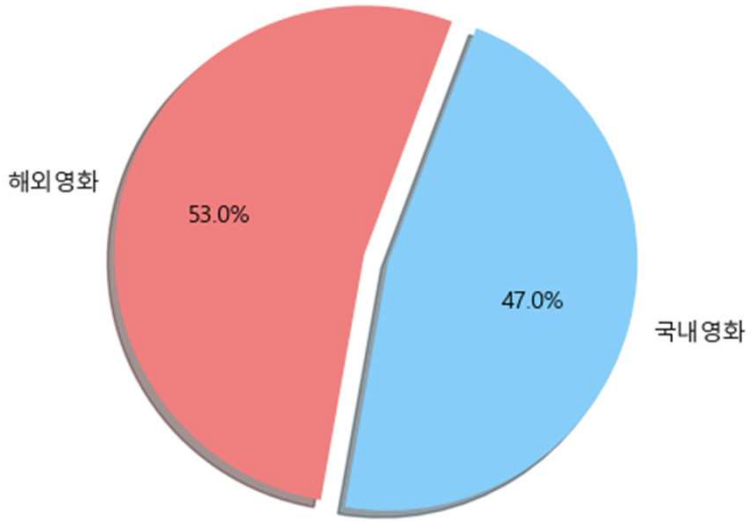


인사이트(EDA) - 국적

원본(11,819)



추출(914)



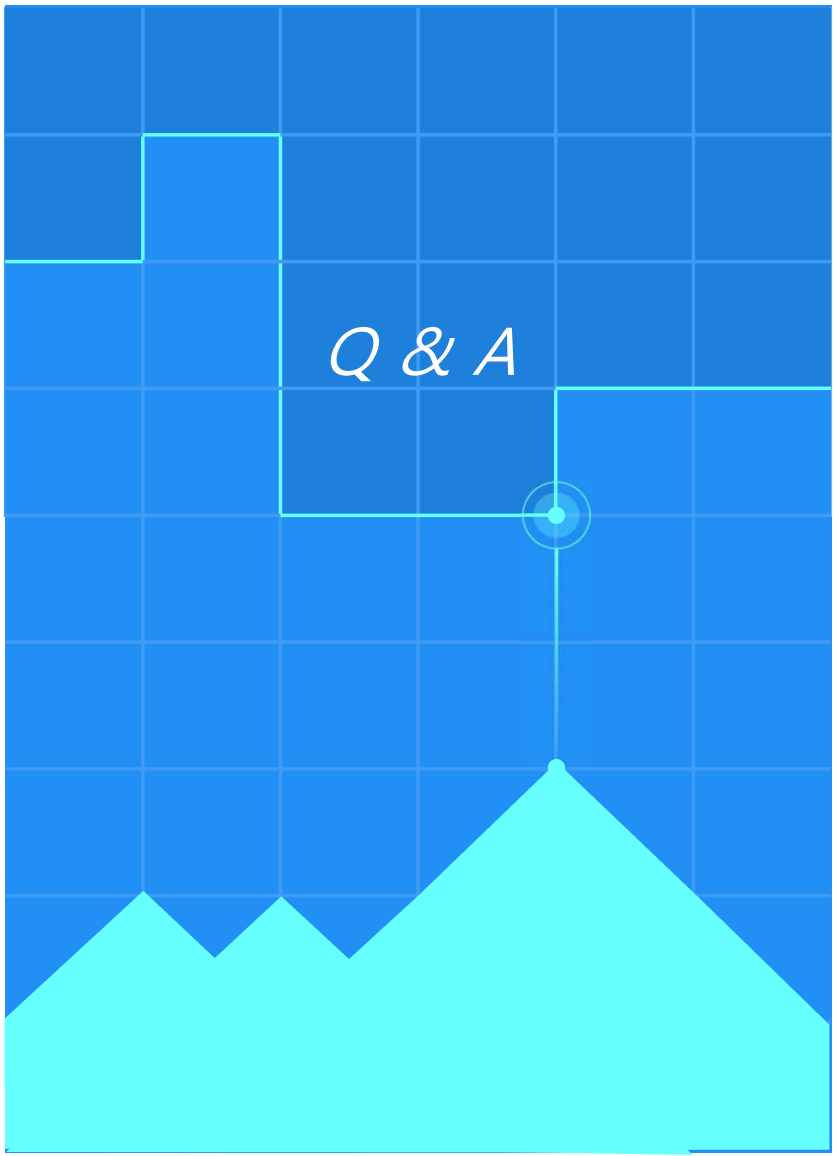
의의 및 한계점

의의

- 영화데이터가 배급사 주식간의 관계성을 파악해 보았다.
- 영화데이터 분석을 통해 흥행영화의 트렌드와 패턴을 확인했다
- 기존의 영화 데이터 외에도 새로운 파생변수 등을 활용하여 피쳐 후보군을 넓혔다.
- 선행연구가 없어 새로운 분야의 연구를 시도해 보았다
- 새로운 업종 지수를 통해 영화 업종 시장, 투자 전략 성과를 측정하기 위한 벤치마크로 활용이 가능하다

한계점

- 개별 영화의 손익분기점은 미공개 데이터로 흥행영화 분류의 기준으로 적용할 수 없었다
- 파생변수의 경우 참고할 선행연구가 없어 자체 기준을 설정하였다
- 연구의 범위가 개봉영화로 한정되어 영화시장을 반영하기 위해 OTT까지 범위를 넓힐 필요가 있다



출처 및 Q & A

출처/ 질문 과 답변

데이터 출처

데이터 \ 내용 및 출처	내용	출처
데이터	➢ 흥행 영화 선정을 위한 10년치 평균 제작비	➢ 한국 영화산업 결산 보고서(KOFIC)
	➢ 타겟 변수 설정을 위한 업종 지수화 ➢ 코스피, 코스닥 종목 선정 - 영화와 직접 관련된 종목 선별 ➢ 기준 : 2000년 1월 1일 ~ 2019년 12월 31일	➢ -코스피(KRX) ▪ 아센디오, IHQ, 콘텐트리중앙, CJ CGV,롯데쇼핑 ➢ -코스닥(KRX) ▪ CJ ENM, 위지윅스튜디오, 덱스터, 쇼박스, 애니플러스, NEW, 바른손이앤에이, 판타지오, 스튜디오산타클로스
	➢ 영화 데이터	➢ 영화진흥위원회 통합전산망
	➢ 영화 일일데이터	➢ https://www.kobis.or.kr/kobisopenapi/homepg/main/main.do(OPEN API)
	➢ 타겟 설정을 위한 벤치마크 선정	➢ 코스닥지수(Finance Data Reader)

reference

논문

✓ 김유신, 김남규, 정승렬(2012) 뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자의사결정모형
Stock-Index Invest Model Using News Big Data Opinion Mining

✓ 데이터 분석을 활용한 한국 영화 흥행 예측
Prediction of Financial Success Using Data Analysis for Korea Movies

✓ 황예나, 남윤재(2017) 한국 흥행영화의 배우관계망 분석: 2012 ~ 2016년도 한국 흥행 영화 출연 배우들을 중심으로

reference

뉴스기사

- ✓ [특징주] 국내 음원차트 휩쓰는 신인 걸그룹 '뉴진스'...엔터테인먼트 관련주 하이브 주가 탄력받나
  https://www.nbntv.co.kr/news/articleView.html
- ✓ 아바타 역대급 성적 낼까...흥행 기대감에 배급·영화사 주가 쑥
  https://www.mk.co.kr/news/stock/10564769
- ✓ [뉴욕 e종목] 디즈니, '아바타2' 글로벌 신드롬에 주가 급등
  https://news.genews.com/kokr/news/article/news_all/20230131085403114l6b49b9d1da_1/article.html?md=20230131091659_U
- ✓ [Y이슈] 입소문 따라 주가도 철렁... 우영우·한산 관련주에 쏠리는 눈길
  https://www.ytn.co.kr/_ln/0117_202207261618026636
- ✓ [증시 키워드] 엔씨소프트, 신작 '리니지W' 공개 앞두고 주가 상승세
  https://www.etoday.co.kr/news/view/2073830

질문과 답변

감사합니다