TEAM 3조

김빛나, 박종호, 박준섭, 이주현, 추성민

국내 영화 흥행 예측을 위한 텍스트 마이닝 기반 머신 러닝 프로젝트

목차

주제 및 선정 배경

- 코로나와 OTT로 인해 가속화된 영화관의 수익성 약화
- 고예산 영화에 대한 투자가 과열되어 중, 저예산 영화에 대한
  투자 감소로 인해 국내 영화 시장의 다양성 및 안정화 저해가 우려
- 국내 개봉예정 영화의 흥행 성적 예측을 통한 투자 의사결정 지원 시스템을 개발

기획 의도 및 활용 방안

- 정확한 흥행 예측으로 인한 투자 견인
- 더욱 다양한 영화에 대한 투자로 업계 고착화 현상을 해결
- AI, 머신러닝 등의 키워드를 활용한 마케팅 활용

내용

- 과거 국내 개봉작 데이터의 텍스트 마이닝과 실제 흥행 성적 간의 분석을 통해
  개봉 예정작의 주요 정보로 흥행 성적 예측

활용 장비 및 재료

- 언어 : Python
- 라이브러리 : Pandas, Numpy, tqdm, Selenium, Konlpy, Matplotlib, Seaborn, Sklearn

프로젝트 구조

- 기획 > 데이터 수집 및 분석 > 데이터 전처리 > 모델 선정 > 모델 학습 > 모델 성능 평가 및 검증

팀 구성 및 역할

훈련생	역할	담당 업무
김 빛 나	팀 원	주제 선정 및 기획
이 주 현	팀 원	데이터 수집
박 종 호	팀 원	모델 선정 및 검증
박 준 섭	팀 원	데이터 전처리
추 성 민	팀 원	PPT 제작 및 발표

수행 절차 및 방법

구분	기간	활동	비고
사전 기획	12/11(월) ~ 12/15(수)	• 프로젝트 기획 및 주제 선정 • 기획안 작성	• 아이디어 선정
데이터 수집	12/13(수) ~ 12/15(금)	• 필요 데이터 및 수집 절차 정의 • 외부 데이터 수집	• KOBIS, 다음 영화
데이터 전처리	12/14(수) ~ 12/15(금)	• 데이터 정제 및 정규화	• 전처리, 토큰화, TD-IDF
모델 선정 및 구현	12/16(토) ~ 12/17(일)	• 모델 선정 및 구현	• 선형 회귀 모델
모델 검증 및 평가	12/17(일) ~ 12/18(월)	• 모델 검증 및 상향 검토	• GridCV
총 개발 기간	• 12/11(월) ~ 12/18(월) (총 1주)		

데이터 수집

KOBIS (영화관입장권통합전산망) 에서

2017~2023 년도 국내 개봉작의

영화 기본 정보

(제목, 제작국가, 등급,

장르, 제작사, 배급사)

및 상영 정보

(개봉일, 매출액, 관객 수,

스크린 수, 상영횟수)

데이터 다운로드

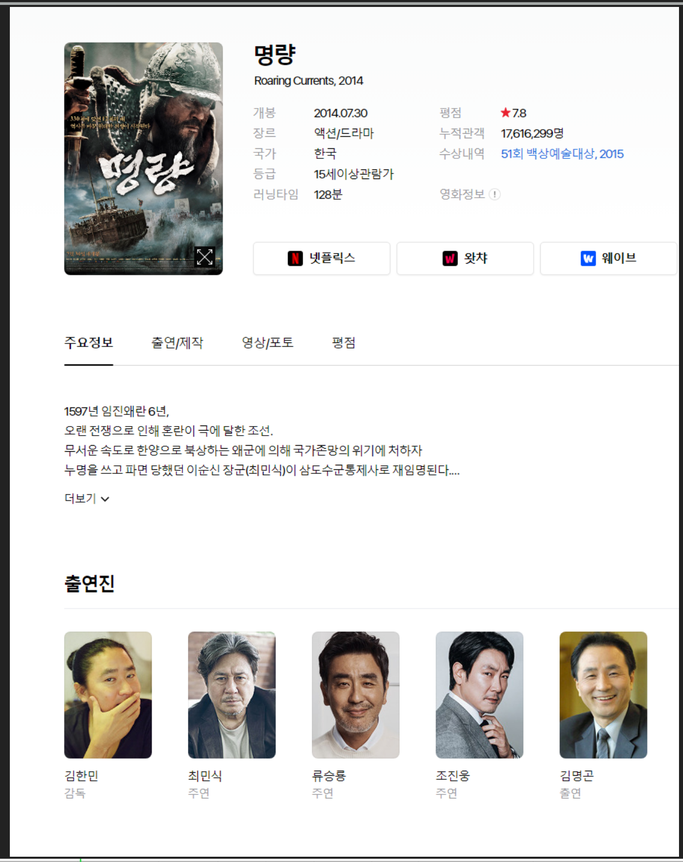
1. KOBIS 의 기간별 박스오피스 데이터 사용

순위	영화명	개봉일	매출액	매출액 점유율	누적매출액	관객수	누적관객수	스크린수	상영횟수
1	서울의 봄	2023-11-22	4,482,770,100	72.0%	86,503,273,371	445,508	8,941,109	2,259	8,887
2	뽀로로 극장…	2023-12-13	449,493,441	7.2%	1,443,533,598	48,811	159,734	850	1,594
3	3일의 휴가	2023-12-06	364,528,712	5.9%	4,027,035,896	36,426	425,868	801	1,887
4	괴물	2023-11-29	165,084,507	2.7%	2,778,576,838	15,805	288,602	401	635
5	말하고 싶은…	2023-12-13	111,830,808	1.8%	441,200,290	11,100	45,696	392	586
6	쏘우 X	2023-12-13	102,059,505	1.6%	524,346,677	9,414	50,320	483	776
7	나폴레옹	2023-12-06	90,378,181	1.5%	2,089,597,868	9,161	208,102	482	632
8	프레디의 피…	2023-11-15	67,225,743	1.1%	6,934,700,101	6,534	709,650	217	280
9	싱글 인 서울	2023-11-29	61,137,020	1.0%	3,569,200,206	6,109	381,206	365	472
10	노량: 죽음…	2023-12-20	44,706,400	0.7%	274,591,010	5,087	27,503	15	31
11	신차원! 짱…	2023-12-22	40,598,947	0.7%	80,797,825	3,763	7,564	20	24

데이터 수집

- Selenium 활용해 웹크롤링
- Daum영화 사이트에서 영화 소개, 감독, 주연 데이터 수집 (네이버는 크롤링 막힘)

```python
# pd파일 읽어서 영화명 리스트로 만들기
df = pd.read_csv('제목.csv', encoding='euc-kr')
movie_list = df['영화제목'].tolist()

# 데이터를 저장할 빈 리스트
bodies = []
URL = "https://movie.daum.net/main"
for movie in tqdm(movie_list):
    # 검색창에 영화제목 입력
    driver.get("https://movie.daum.net/search?q=" + movie)
    time.sleep(1)
    try:
        # 영화 소개 들어가기
        driver.find_element(By.XPATH, '//*[@id="mainContent"]/div/div[2]/div[2]/ul/li/div/div/strong/a').click()
        time.sleep(1)

        # 더보기 버튼이 있는지 확인
        more_button = driver.find_elements(By.XPATH, '//*[@id="mainContent"]/div/div[2]/div[2]/div[1]/div/div/a')

        # 더보기 버튼 클릭
        if more_button:
            more_button[0].click()
            time.sleep(1)

        # 영화 소개 텍스트 가져오기
        body_element = driver.find_element(By.XPATH, '//*[@id="mainContent"]/div/div[2]/div[2]/div[1]/div/div/div')
        bodies.append(body_element.text)

        # 영화 소개가 없으면 none 추가
    except Exception as e:
        bodies.append(None)
# 데이터를 DataFrame으로 변환
df = pd.DataFrame({
    '영화제목': movie_list,
    '소개': bodies
})
# CSV 파일로 저장
df.to_csv('info.csv', index=False)
```

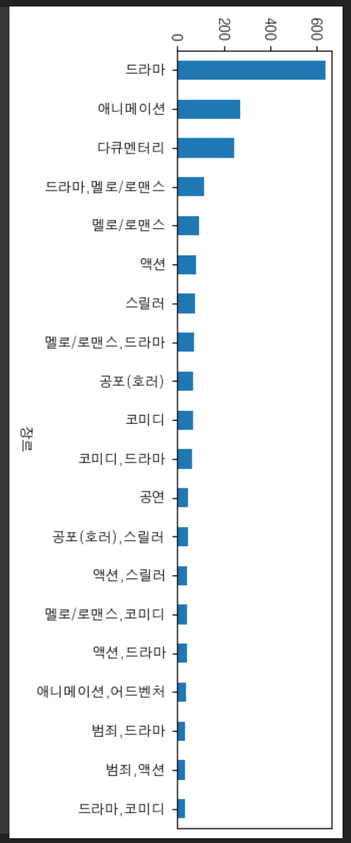


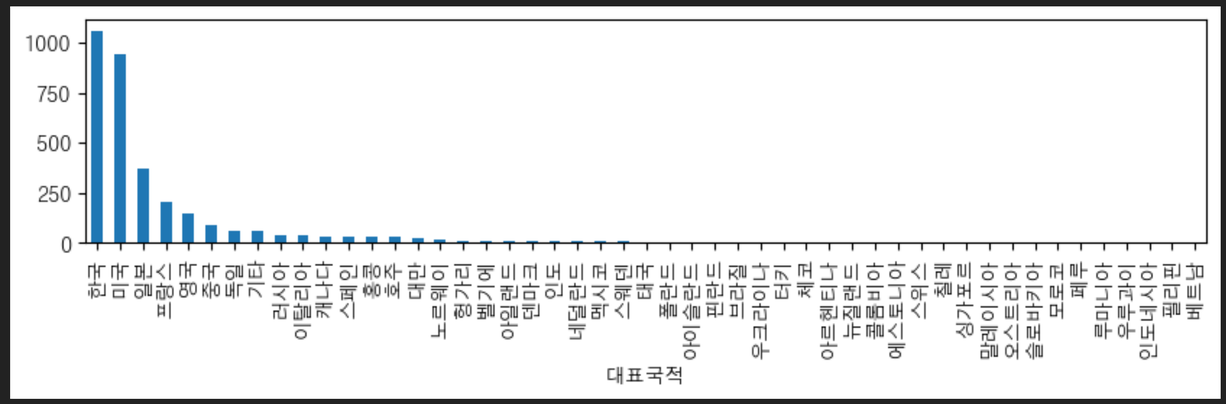
영화제목	개봉일	매출액	매출액.1	누적매출액	관객수	누적관객수	스크린수	상영횟수	대표국적	국적	제작사	배급사	등급	장르	감독	주연	영화소개
스파이더맨: 노 웨이 홈	2021-12-15	53,772,689,910	4.9%	53,772,689,910	5,369,773	5369773	2,948	137848	미국	미국		소니픽쳐스엔터테인먼트코리아주	12세이상관람	액션,어	존 왓츠	톰 홀랜드,젠데이아 콜먼,베네딕트	'미스테리오'의 계략으로 세상에
남산의 부장들	2020-01-22	41,225,216,650	3.8%	41,225,216,650	4,750,345	4750345	1,659	140051	한국	한국	(주)하이브미디어코프	(주)쇼박스	15세이상관람	드라마	우민호	이병헌,이성민,곽도원,이희준,김소	"각하, 제가 어떻게 하길 원하십니
다만 악에서 구하소서	2020-08-05	38,602,260,990	3.5%	38,602,260,990	4,357,803	4357803	1,998	193842	한국	한국	(주)하이브미디어코프	(주)씨제이이엔엠	15세이상관람	범죄,액	홍원찬	황정민,이정재,박정민,박소이,최희	2020년 여름 최고 흥행작 하드보일드 추격액션의 역사를 바
반도	2020-07-15	33,073,948,880	3.0%	33,073,948,880	3,812,455	3812455	2,575	199084	한국	한국	(주)영화사레드피터	(주)넥스트엔터테인먼트월드(NEW	15세이상관람	액션,드	연상호	강동원,이정현,권해효,김민재,구교	전대미문의 재난 그 후 4년 폐허의 땅으로 다시 들어간다!
모가디슈	2021-07-28	34,558,297,730	3.1%	34,558,297,730	3,613,984	3613984	1,688	210740	한국	한국	(주)덱스터스튜디오,(주)외	롯데컬처웍스(주)롯데엔터테인먼트	15세이상관람	액션,드	류승완	김윤석,조인성,허준호,구교환,김소	내전으로 고립된 낯선 도시, 모가디 지금부터 우리의 목표는 오로지 생
이터널스	2021-11-03	31,729,284,450	2.9%	31,729,284,450	3,050,132	3050132	2,648	162434	미국	미국		월트디즈니컴퍼니코리아 유한책임	12세이상관람	액션,어	클로이 자오	안젤리나 졸리,마동석,리처드 매든	마블 스튜디오의 <이터널스>는 수 [HOT ISSUE]
블랙 위도우	2021-07-07	29,996,075,620	2.7%	29,996,075,620	2,962,088	2962088	2,528	155016	미국	미국		월트디즈니컴퍼니코리아 유한책임	12세이상관람	액션,어	케이트 쇼트랜드	스칼렛 요한슨,플로렌스 퓨,레이철	네덜란드 마약계 큰 손이자 범죄 조 어느 날, 동료가 성폭행 당하는 장 카르멘에게 원한을 품었던 상대 조
히트맨	2020-01-22	20,614,278,000	1.9%	20,614,278,000	2,406,232	2406232	1,122	87782	한국	한국	베리굿스튜디오(주)	롯데컬처웍스(주)롯데엔터테인먼트	15세이상관람	코미디,	최원섭	권상우,정준호,황우슬혜,이이경,이	아드레날린이 폭발하는 인텔리전트 에이전트 47을 맞이하라!
분노의 질주: 더 얼티메	2021-05-19	22,059,658,060	2.0%	22,059,658,060	2,292,415	2292415	2,297	131856	미국	미국		유니버설픽쳐스인터내셔널 코리아	12세이상관람	액션	저스틴 린	빈 디젤,존 시나,성 강,샤를리즈 테	기다림은 끝났다! 전 세계가 기다려온 단 하나의 액션

데이터 전처리

- 설명변수 (X) : 문자열 자료 인덱싱
  - 각 영화의 특성을 잘 반영하는 feature 로 장르, 등급, 제작사, 배급사, 감독, 주연배우 를 선정해 one-hot encoding 처리
    - 한 컬럼에 여러 값들이 포함된 경우 분리하여 인덱싱
  - 텍스트 마이닝 주요 대상인 영화 소개(시놉시스)은 토큰화, 불용어 제거, TF-IDF 처리 후 one-hot encoding 하여 Max Feature로 최빈 단어 5,000개만 남김
  - 수치형 설명변수로 사용하려던 평점, 개봉일은 배제
    - 평점은 사후적 지표이고, 흥행 성적과는 무관하게 평점을 작성한 개인 간 다양한 변수가 작용하고 있어 제거
    - 개봉일은 datetime 으로 변환해 연도, 월 분리하여 분석하려 했으나, 2020~2022년 코로나 이슈로 제거

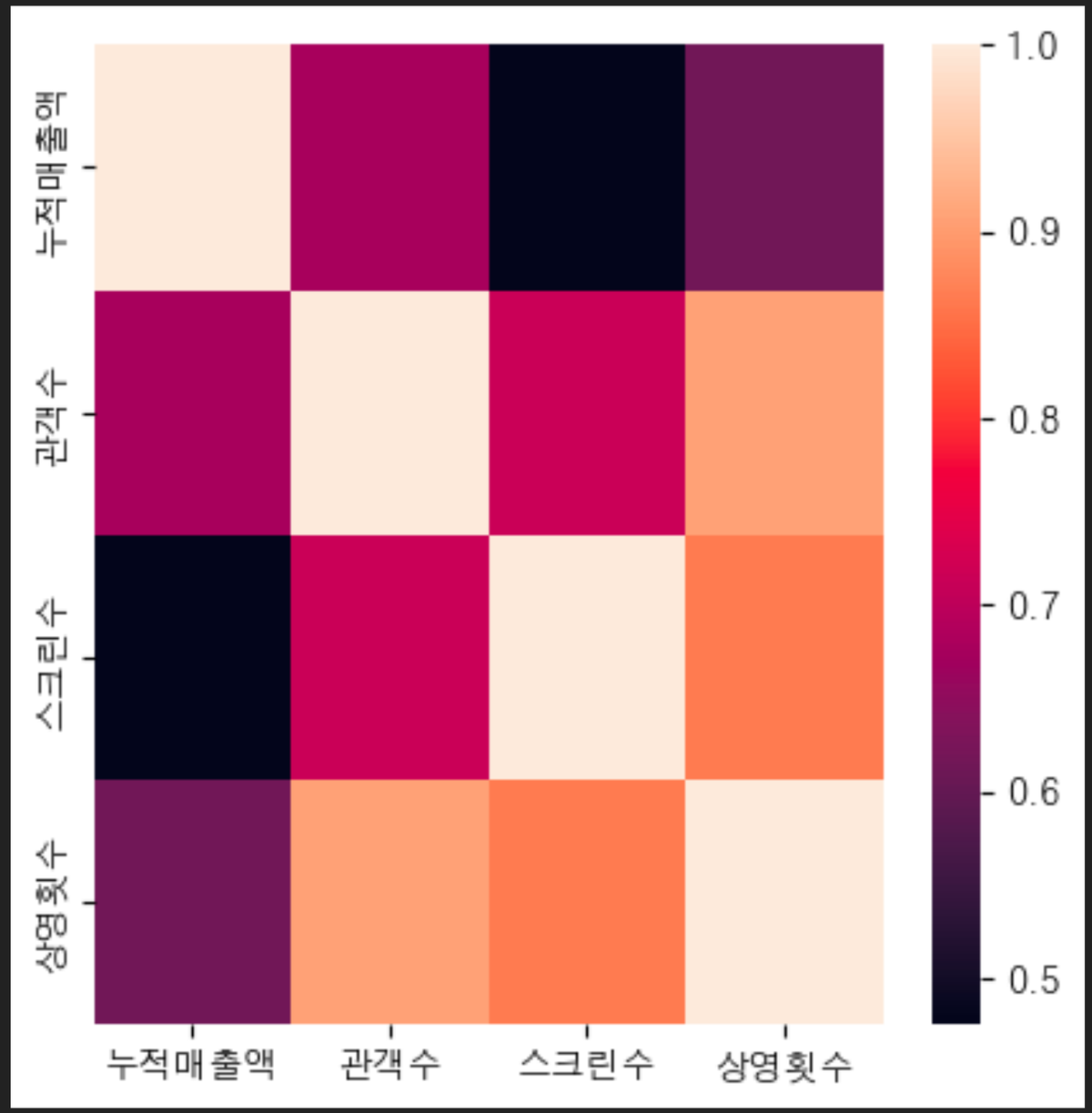
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3299 entries, 0 to 3298
Data columns (total 18 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   영화제목    3299 non-null   object
 1   개봉일     3272 non-null   object
 2   매출액     3299 non-null   object
 3   매출액.1   3299 non-null   object
 4   누적매출액   3299 non-null   object
 5   관객수     3299 non-null   object
 6   누적관객수   3299 non-null   int64
 7   스크린수    3299 non-null   object
 8   상영횟수    3299 non-null   int64
 9   대표국적    3298 non-null   object
 10  국적      3298 non-null   object
 11  제작사     1149 non-null   object
 12  배급사     3263 non-null   object
 13  등급      3282 non-null   object
 14  장르      3293 non-null   object
 15  감독      3064 non-null   object
 16  주연      2953 non-null   object
 17  영화소개    3299 non-null   object
dtypes: int64(2), object(16)
```





데이터 분석

- 종속변수(y) 를 선정하기 위한 분석
  - 매출액 : 물가 상승, 상영관/좌석 등급 가격 차이 등 외부요인의 영향을 받기 때문에 적합하지 않음
  - 스크린 수, 상영횟수 : 영화의 흥행과 인기는 반영하지만, 적은 스크린으로 오래 지속되는 영화 등 다양한 흥행 패턴을 반영하지 못함
  - (누적)관객 수 : 가장 직접적인 성과를 반영하며 다양한 흥행 패턴이나 다른 외부요인에 대한 영향이 가장 적음
    - 후보 간 상관관계 분포에서도 다른 변수에 비해 더 높은 상관관계를 보임



모델 선정

- 종속변수인 누적 관객수가 연속성을 가지고 있기 때문에
  이를 예측하는 목적이므로 분류가 아닌 회귀에 해당
  -> 수업에서 배운 여러 회귀 모델들을 비교하기로 함

- 설명변수가 텍스트 데이터의 TF-IDF 와 인코딩된 범주형 데이터로, 인덱싱 후
  Feature 개수가 약 3만 개에 달할 정도로 증가, 수업 내용에 기반하여
  L1 규제를 사용한 모델이 적합하다는 가설 설정함

  - 데이터의 특성 수가 샘플 수 보다 많거나, 특성 간의 상관관계가 높은 경우 :
    L1 규제가 불필요한 특성을 0으로 만드는
    특성 선택을 수행하여 모델의 복잡도를 줄여줌

  - 모델이 대부분의 값이 0이고 일부 값 만이 0이 아닌
    희소한(sparse) 특성을 가지는 경우 :
    L1 규제가 희소한 특성에 대해 더 높은 가중치를 부여하고,
    불필요한 특성을 제거하여 모델의 성능을 향상함

- 실제 여러 모델 비교 결과, L1 규제 가중치가 높은 (L1 : 0.8)
  ElasticNet 성능이 압도적으로 높았음 (R2 : 0.68)

모델 성능 평가

- 성능 평가 지표 R2 기준, 선형회귀 모델 중 ElasticNet(L1 & L2) 및 Ridge(L2) 성능이 우수
- SVM

3가지 커널 poly, rbf, linear 모델 모두 결과가 좋지 않음

- Decision Tree

낮은 성능을 보임

선형회귀 모델

```
LinearRegression
-1.3362477751646 28e+22
1.6795592910201862e+17
8.088309433647163e+16

Ridge
0.6472657002021429
862930.2398963502
421440.10255462996
Lasso
0.21075146307697135
1290799.3406537543
592174.4460937895
ElasticNet
0.33669135263081695
1183340.3450775836
494630.08630791714
```

Decision Tree

SVM

param_C	param_gamma	params	split0_test_score	split1_test_score	mean_test_score	std_test_score	rank_test_score
0.1	0.1	{'C': 0.1, 'gamma': 0.1}	0.000758	0.000758	0.000758	2.871781e-07	1
1	0.1	{'C': 1, 'gamma': 0.1}	0.000758	0.000758	0.000758	2.871781e-07	1
0.1	1	{'C': 0.1, 'gamma': 1}	0.000000	0.000000	0.000000	0.000000e+00	3

param_C	param_gamma	params	split0_test_score	split1_test_score	mean_test_score	std_test_score	rank_test_score
0.1	0.1	{'C': 0.1, 'gamma': 0.1}	0.000758	0.000758	0.000758	2.871781e-07	1
0.1	1	{'C': 0.1, 'gamma': 1}	0.000758	0.000758	0.000758	2.871781e-07	1
0.1	10	{'C': 0.1, 'gamma': 10}	0.000758	0.000758	0.000758	2.871781e-07	1

param_C	param_gamma	params	split0_test_score	split1_test_score	mean_test_score	std_test_score	rank_test_score
0.1	0.1	{'C': 0.1, 'gamma': 0.1}	0.0	0.0	0.0	0.0	1
0.1	1	{'C': 0.1, 'gamma': 1}	0.0	0.0	0.0	0.0	1
0.1	10	{'C': 0.1, 'gamma': 10}	0.0	0.0	0.0	0.0	1

```
R2-train 데이터: 0.9592678171934174
R2-test 데이터: 0.4612058379361438
RMSE-train 데이터: 316779.0621010667
RMSE-test 데이터: 1066505.9503668454
```

모델 검증

GridSearchCV 사용하여
최적의 하이퍼 파라미터 탐색

결과

alpha : 0.01

l1_ratio : 0.8

성능

R2 : 0.676

RMSE : 826909 - 4472333104

MAE : 397414

```python
import numpy as np
import warnings
warnings.filterwarnings("ignore")

from sklearn.model_selection import GridSearchCV

model = ElasticNet()

parametersGrid = {
                "alpha": [0.0001, 0.01, 1, 10, 100],
                "l1_ratio": np.arange(0.0, 1.2, 0.2)}

grid_els = GridSearchCV(model, param_grid = parametersGrid, cv=3, refit = True ,  scoring='r2')

# refit : True가 디폴트, True이면 가장 좋은 파라미터 설정으로 학습시켜서 모델 반환

grid_els.fit(X_tr, y_tr)

print(grid_els.best_estimator_)
print(grid_els.best_params_)
```

```
ElasticNet(alpha=0.01, l1_ratio=0.8)
{'alpha': 0.01, 'l1_ratio': 0.8}
```

```
ElasticNet
0.6760990490793224
826909.4472333104
397414.6527120774
```

모델 상향 검토

As-Is

1. [영화 소개] 문장을 TF-IDF 처리, 빈도수 기준으로 최빈 단어 5천개 추출해 feature로 사용
2. [장르, 등급] 유사하지만 이름만 다른 값들을 모두 개별 feature들로 분리
3. [기타 설명변수] 감독, 주연, 제작사 등 대부분 희소한 feature들이 많음
4. [관객수] 수치형 종속변수의 연속성만을 고려하여 회귀 모델을 선정

To-Be

1. 다른 텍스트 마이닝 기법들(분류, 클러스터링등) 사용하여 보다 유의미한 정보를 추출하거나, 흥미 유발 단어에 가중치를 차등적으로 부여
2. 유사한 장르/등급 등 그룹화(예: 공포=호러=스릴러) 하여 feature와 모델 간소화
3. 해당 문자열 데이터를 수치화(예: 평균 출연작 수, 동원 관객수, 흥행작 수 등) 하여 보다 유의미한 feature로 변환
4. 종속변수를 범주화하여(예: 관객수 별 구간 설정) 분류 모델을 사용하거나, 여러 종속변수 후보들을 앙상블 하면 전반적으로 더 높은 성능이 기대됨

자체 평가 및 추후 방향

- 주제 선정에 시행착오를 겪어 여러 번의 주제 변경이 있어 최종 주제 선택 후 절대적 시간 부족
  - 데이터 수집 전처리 및 가공 과정에서 더 다양한 방법을 시도해보지 못함
  - 더 적합한 모델 선정 및 파라미터 튜닝을 통한 성능 향상 가능성이 있었으나, 시간 부족으로 다 검토 못함

- 보다 체계적인 역할 분담의 필요성을 느낌
  - 프로젝트 flow 상 앞 단계 task가 지체되며 프로젝트 진행 bottleneck으로 작용
  - 전처리 할 사람이 준비를 해두면 주제가 정해지고 원본 데이터가 들어왔을 때 바로바로 다음 단계에 들어갈 수 있었을 텐데, 전처리 준비가 미흡했음

- 다양한 모델에 대한 기본 지식 부족
  - 목적이 정해지면 어떤 모델을 사용할지 정해지는게 맞으나, 정확하게 모델들의 장단점, 차이점을 인지하고 있지 못하여서 여러 모델을 적용해보고 학습하다 보니 시간이 많이 지체
    => 여러 모델들의 장단점과 어떤 상황에서 사용하는지 정확하게 인지해서 시간을 단축 시켜야 할것 같다