2025-1 기계학습
팀프로젝트

영화 상영일수 예측

6조 살찐 포도는 포동포동

2021204087 이재용
2021204055 이치우
2023204078 김다솔
2023204042 홍채민
2023204083 김영우
2023204069 전서연

2025-1 기계학습
팀프로젝트

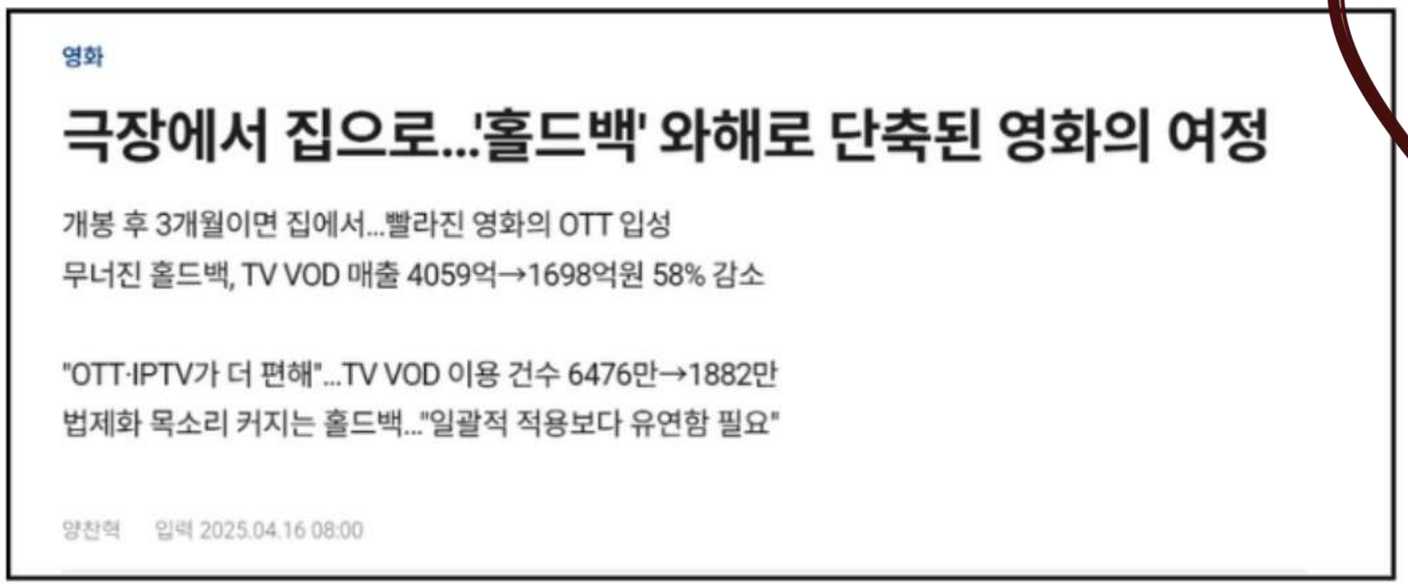
CONTENTS

Chapter
01

주제 선정 배경

Hold Back

: 극장 개봉 이후 OTT공개 전까지 갖는 유예기간



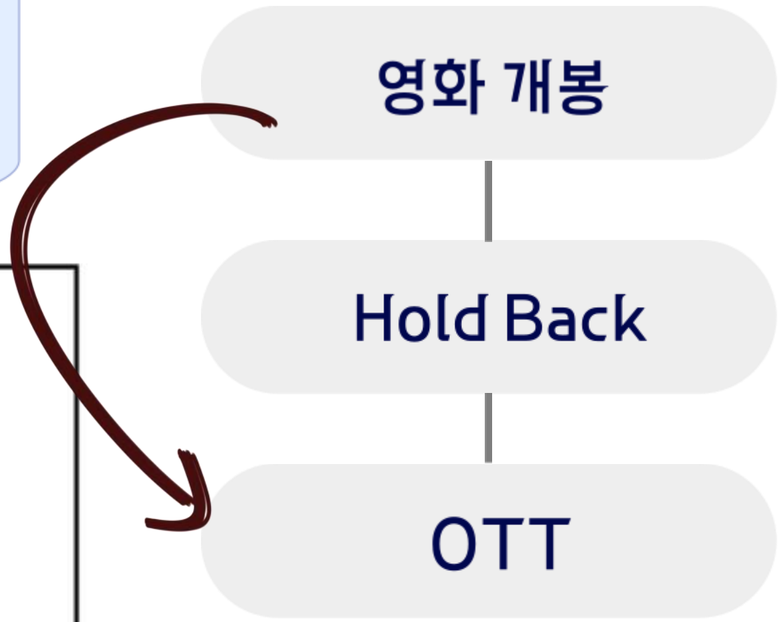
영화

극장에서 집으로...'홀드백' 와해로 단축된 영화의 여정

개봉 후 3개월이면 집에서...빨라진 영화의 OTT 입성
무너진 홀드백, TV VOD 매출 4059억→1698억원 58% 감소

"OTT·IPTV가 더 편해"...TV VOD 이용 건수 6476만→1882만
법제화 목소리 커지는 홀드백..."일괄적 적용보다 유연함 필요"

양찬혁 입력 2025.04.16 08:00



Chapter
02

데이터 수집

kofic
KOBIS
영화관입장권통합전산망

최근 5년치
일별 좌석점유율,
스크린 점유율

한 영화에 대해 조회일 기준으로 합쳐서
리스트 형태의 컬럼 생성

KCISA
한국문화정보원

문화예술 | 개인/단체 | 개인 | 여가 | CSV | JSON | 무료

문화 영화 관람 활성 지수

한국문화정보원 >

유형 CSV, JSON 가격 무료 데이터 갱신주기 Yearly

2024.07.17 업데이트 1453 261

평점 0.0 평가하기 5 평가하기

< 문화 영화 관람 활성 지수 >

2025년 06월 06일(금)

순위	영화명	개봉일	매출액	매출액 점유율	매출액증감 (전일대비)	누적매출액	관객수	관객수증감 (전일대비)	누적관객수
1	드래곤 길들이기 ↑33	2025-06-06	2,232,937,730	37.3%	2,231,659,730 (174,621.3%)	2,281,752,730	222,918	222,776 (156,884.5%)	225,989
2	하이파이브 ↓1	2025-05-30	1,446,441,160	24.2%	904,829,490 (167.1%)	8,371,616,050	156,270	95,248 (156.1%)	906,025
3	미션 임파서블: 파이널 레코닝 ↓1	2025-05-17	1,000,909,430	16.7%	569,392,090 (132.0%)	28,039,055,230	104,988	59,448 (130.5%)	2,848,355
4	신명 ↓1	2025-06-02	566,245,490	9.5%	315,877,460 (126.2%)	2,548,568,570	58,810	31,463 (115.1%)	266,862
5	릴로 & 스티치 -	2025-05-21	181,183,550	3.0%	113,142,280 (166.3%)	3,906,942,350	19,430	11,915 (158.5%)	415,289
6	브랜 뉴 백 ↑93	2025-06-06	120,300,700	2.0%	120,225,700 (160,300.9%)	142,564,300	14,950	14,945 (298,900.0%)	17,598
7	소주전쟁 ↓3	2025-05-30	129,770,600	2.2%	-8,742,260 (-6.3%)	2,183,047,720	14,074	-1,752 (-11.1%)	232,668
8	알시탈 ↑2		43,530,000	0.7%	36,970,000 (563.6%)	277,395,000	8,706	7,394 (563.6%)	55,477
9	극장판 프로젝트 세카이 부서진 세카이… ↓2	2025-05-29	41,166,030	0.7%	16,800,230 (69.0%)	445,748,400	4,486	1,740 (63.4%)	47,042
10	극장판 짱구는 못말려: 격돌! 낙서왕국과… ↑7	2021-09-15	22,533,400	0.4%	18,778,700 (500.1%)	1,963,153,090	2,544	2,117 (495.8%)	206,996

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	영화명	개봉일	장르	국적	감독	배급사	대표국적	등급	제작사	배우	누적관객수	조회일	순위	매출액	총매출액	매출액(점	매출액증	매출액증	관객수	관객수증	관객수증	스크린수	상영횟
2	너의 이름	#######	애니메이션	일본	신카이 마	메가박스중	일본	12세이상관람가		카미키 류	1233480	['2020-04-.	[36, 29, 23	[81920, 25	7.15E+08	[0.001, 0.0	[-315480,	[-0.794, 2.(	[17, 50, 40	[-60, 33, -1	[-0.779, 1.(	[3, 3, 3, 3,	[3, 4, 3, 3
3	날씨의 아	#######	애니메이션	일본	신카이 마	메가박스중	일본	15세이상관람가		다이고 코	274439	['2020-04-.	[19, 11, 11	[396780, 1	5.28E+08	[0.003, 0.0	[-574700,	[-0.592, 2.(	[77, 266, 1	[-114, 189,	[-0.597, 2.(	[9, 11, 11,	[11, 15, 1
4	더 퍼스트	#######	애니메이션	일본	이노우에	(주)넥스트	일본	12세이상관람가			1618171	['2022-12-.	[13, 11, 5,	[540000, 6	1.74E+10	[0.0, 0.006,	[540000, 6	[1.0, 1.0, 3.	[60, 378, 3	[60, 378, 2	[1.0, 1.0, 7.	[1, 1, 11, 3	[2, 1, 11,
5	라라랜드	#######	드라마,뮤지	미국	데이미언	판씨네마(	미국	12세이상관람가		엠마 스톤,	1626948	['2020-04-.	[9, 6, 6, 6,	[4947120,	4.3E+08	[0.034, 0.0	[-105520,	[-0.021, 1.(	[950, 1936	[-12, 986,	[-0.012, 1.(	[66, 67, 71	[104, 116

Chapter 03

데이터 전처리

항목	결측치_개수	결측치_비율
등급	3877	58.30
대표국적	47	0.71
제작사	5219	78.48
국적	47	0.707
감독	1834	27.579
배우	3807	57.248
배급사	3828	57.564

등급·제작사 컬럼: 결측치 비중으로 인해 제거
대표국적 컬럼: 국적과 중복 정보로 제거

배급사, 감독 — 크롤링 →

범주형 변수 처리

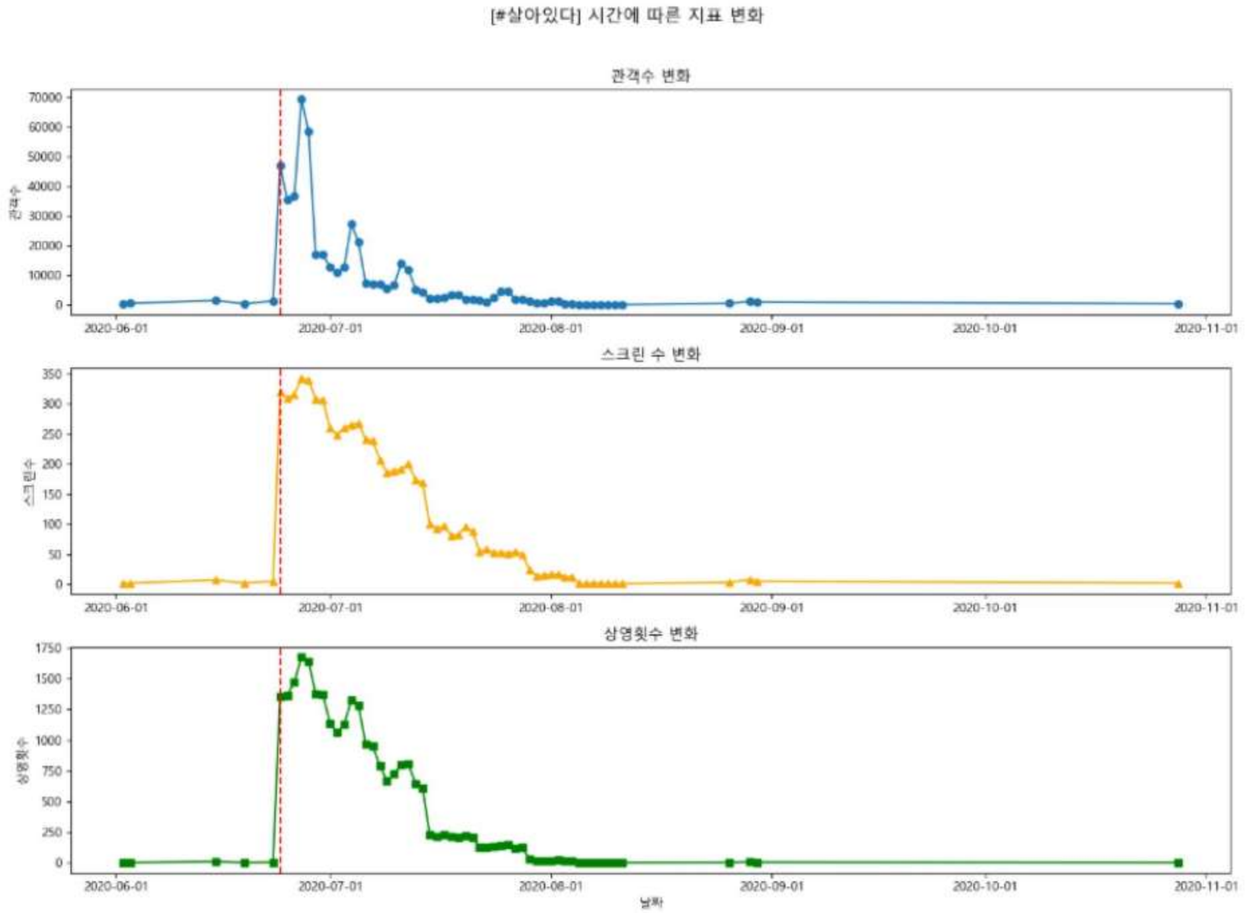
장르, 국적 → One-Hot Encoding + Target Encoding

감독, 배우, 배급사 → Target Encoding

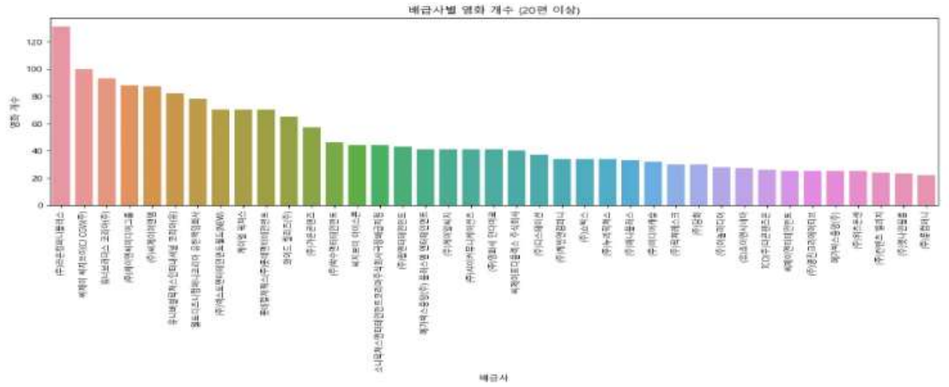
인코딩 결과 - 총 55개의 파생변수 생성

Chapter
04

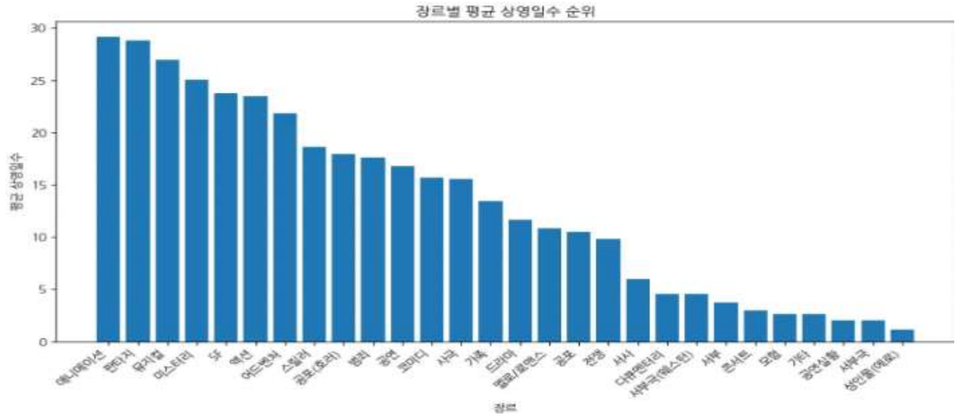
데이터 EDA



< 영화 #살아있다 시간에 따른 지표 변화 >



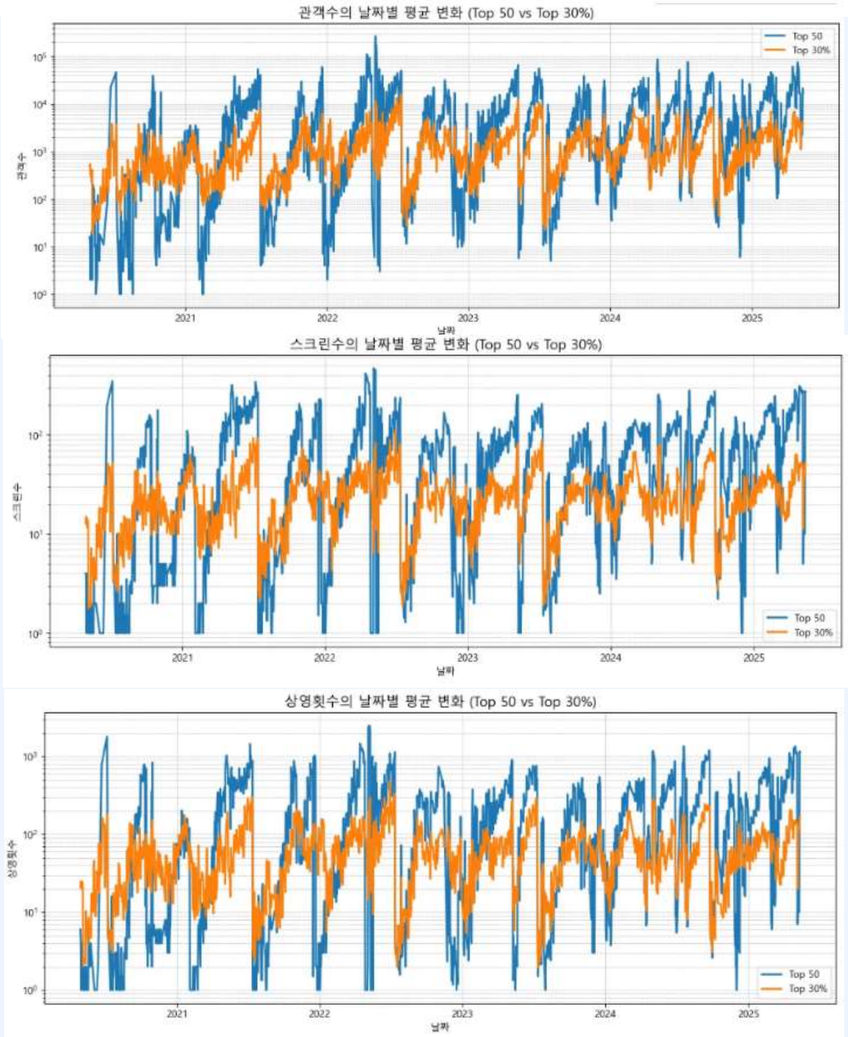
< 배급사별 영화 개수 >



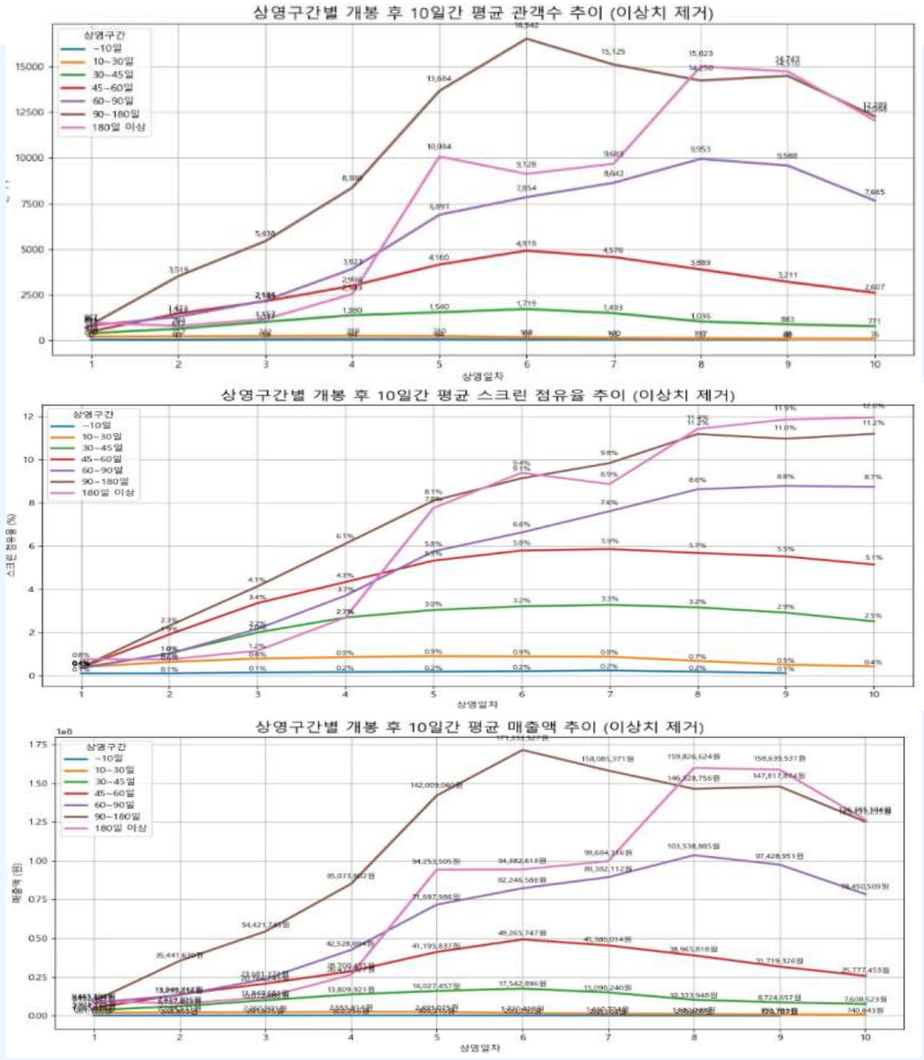
< 장르별 평균 상영일 수 >

Chapter 04

데이터 EDA



< 날짜별 평균 관객수, 스크린수, 상영횟수 변화 >

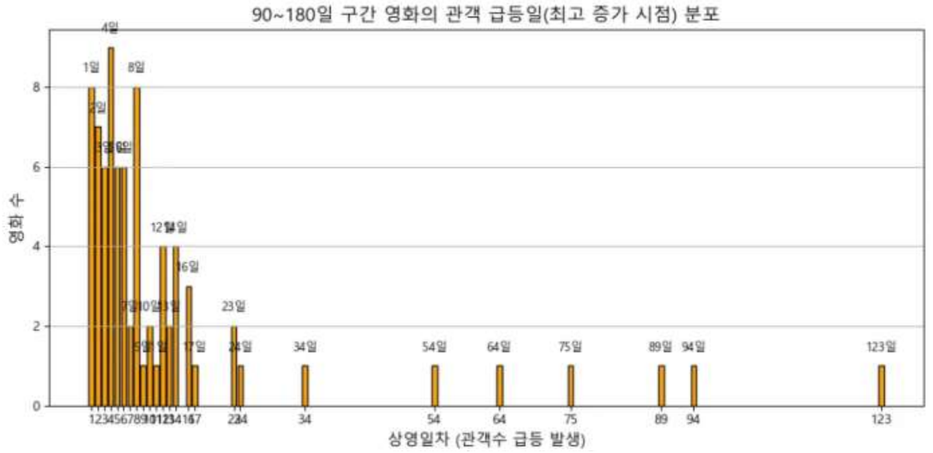


< 개봉 후 10일간 평균 관객수, 스크린 점유율, 평균 매출액 변화 >

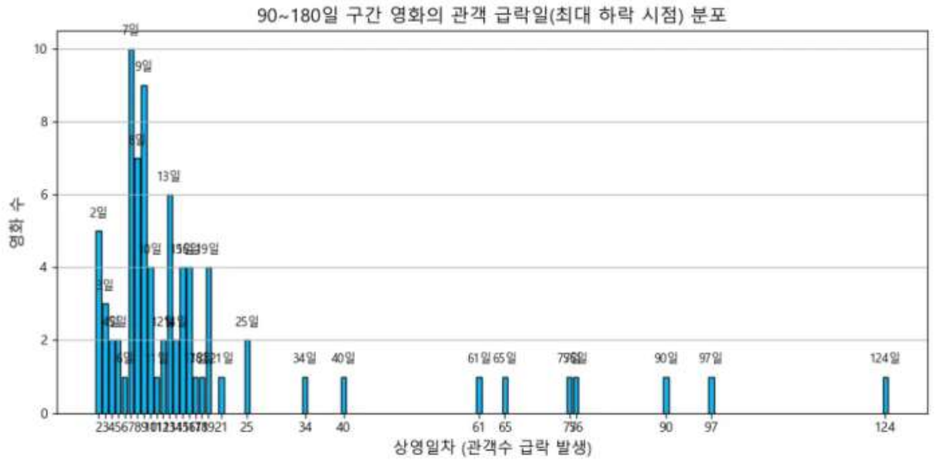
Chapter 04

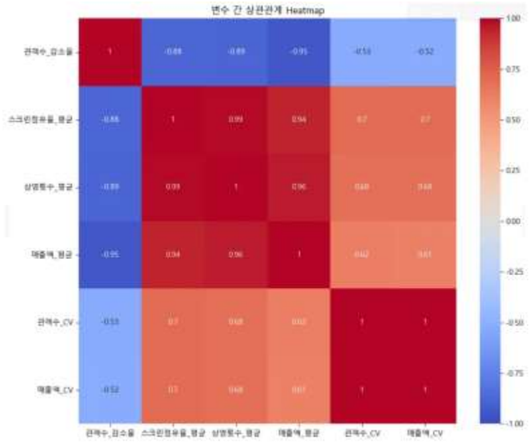
feature 추가 - 흥행 안정성 지수 3/7/10일

< 급등일 >



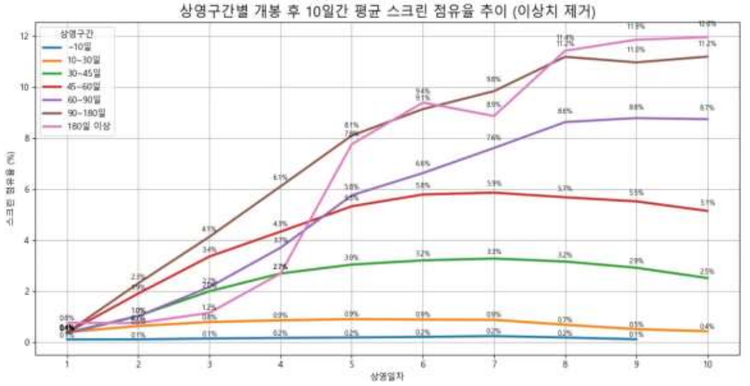
< 급락일 >





흥행 안정성 지수

유지력 * 총관객수 * 선형회귀 기울기



Chapter
04

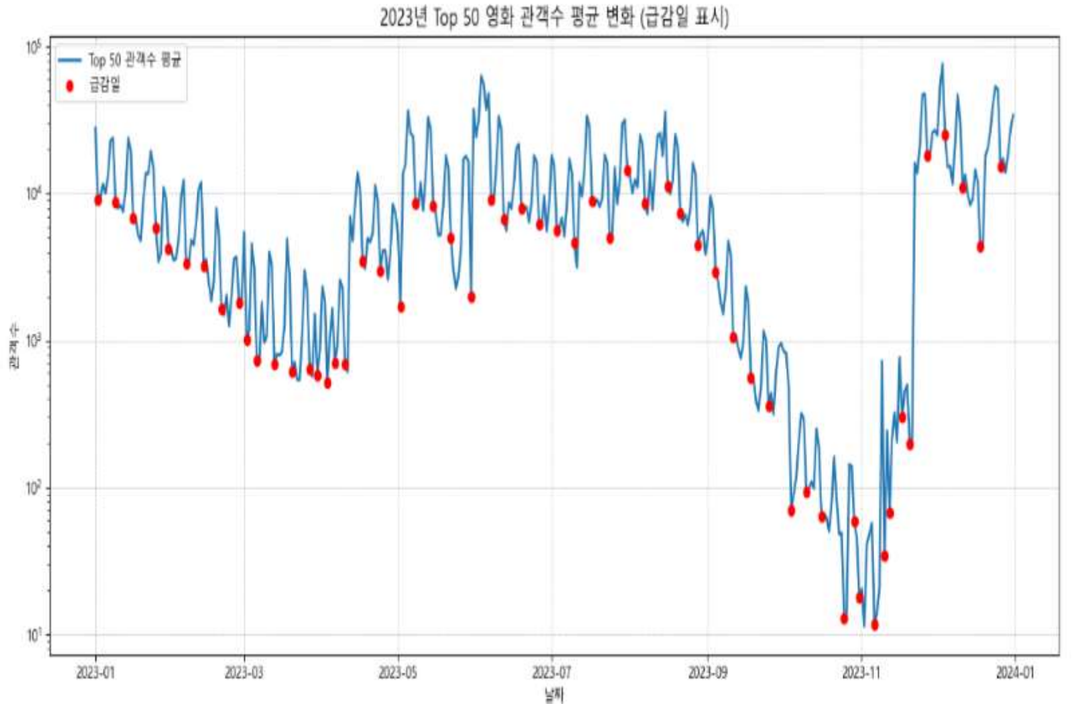
feature 추가 - 반감기

반감기
첫 관객 수의 절반 이하로 떨어지는 데 걸린 일 수

1. 관객수와 날짜 컬럼을
리스트 형태로 변환하여 일별 데이터 확보

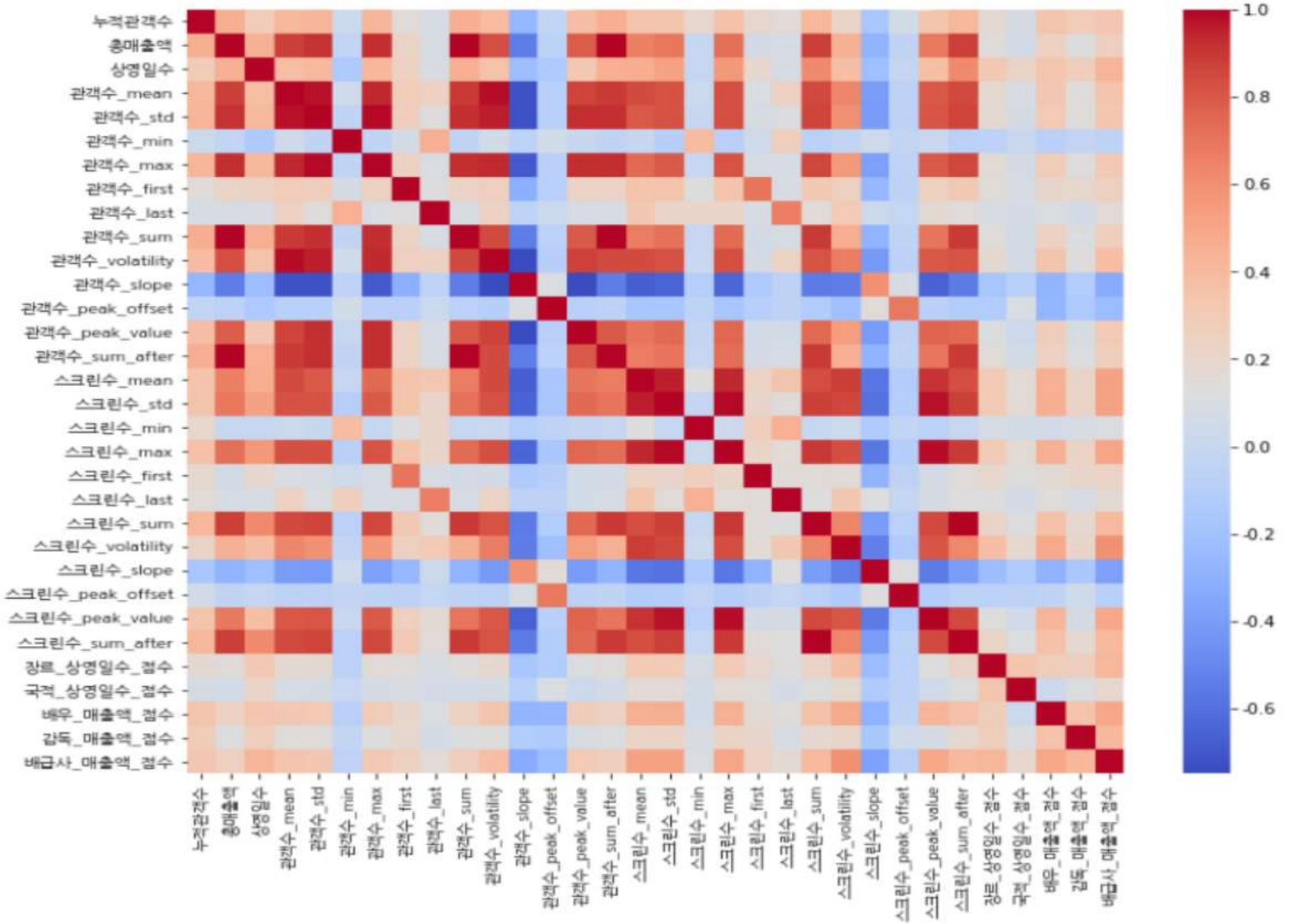
2. 첫날 관객수를 기준으로
관객 수가 절반 이하로 떨어지는 첫 번째 시점을 탐색

3. 해당 시점까지의 일 수를 반감기로 정의



Chapter
04

feature 간 상관관계 히트맵



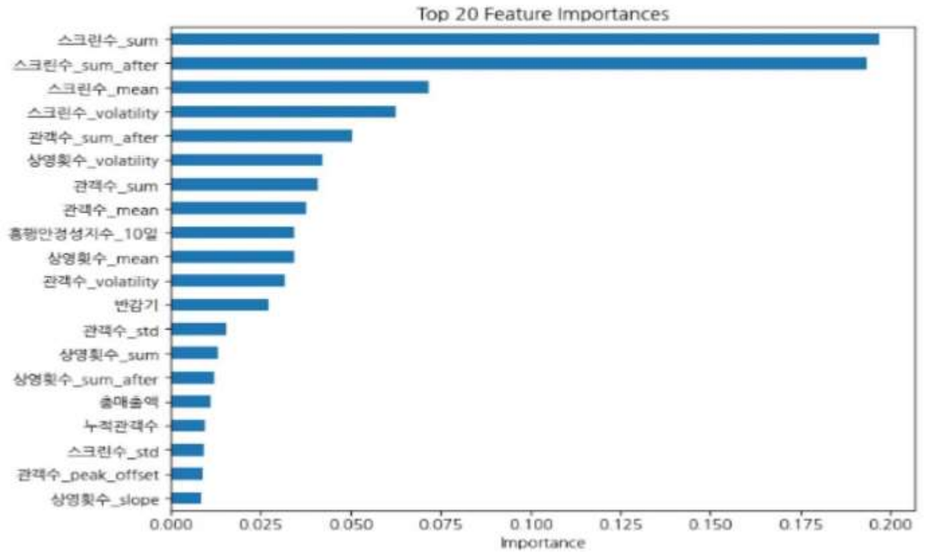
Chapter 05

데이터 분석 - Random Forest

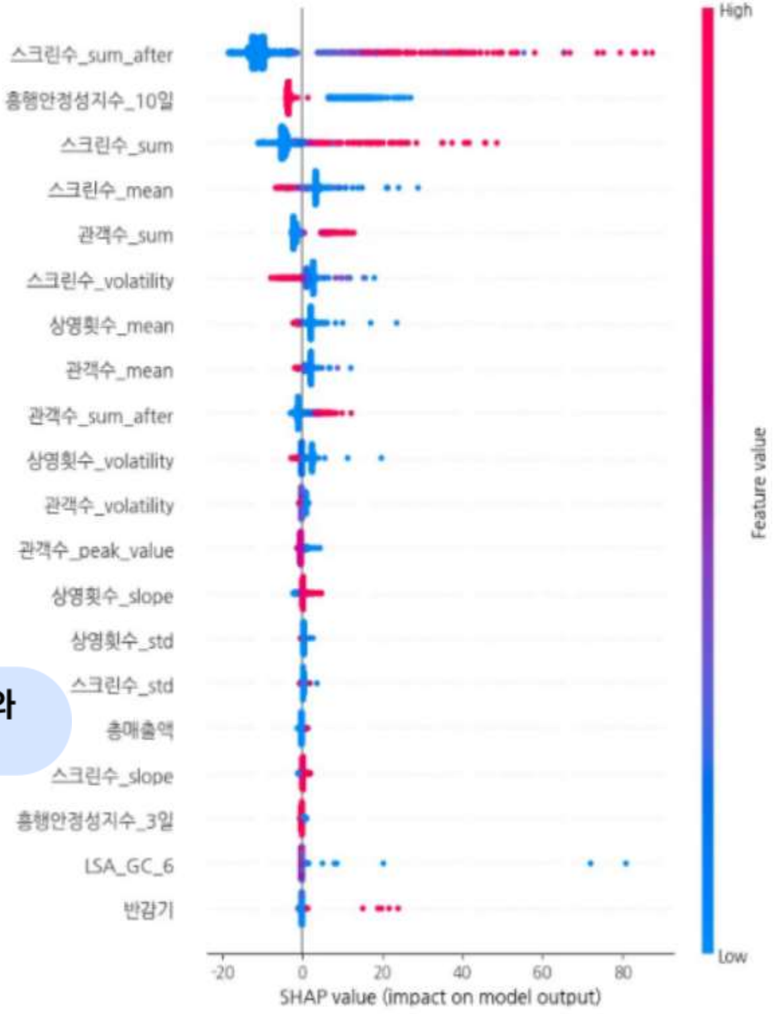
< 하이퍼파라미터 튜닝 및 모델 학습 >

```
best_params = {
    'n_estimators': 100,
    'min_samples_split': 2,
    'min_samples_leaf': 1,
    'max_features': 0.5,
    'max_depth': None,
}
cv_mse = 227.59954039432697
```

MSE: 25.9230
R^2 Score: 0.9487



feature importance 그래프와 SHAP 적용 결과 유사



Chapter 05

데이터 분석 - Random Forest

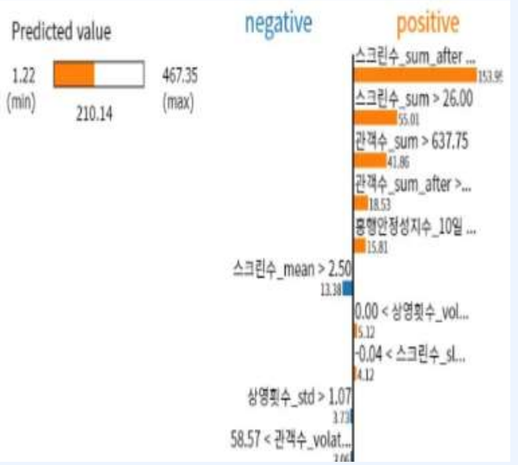
```
=== 샘플 분석 결과 ===
영화명       : 극장판 아이돌리쉬 세븐; 라이브 4비트 비욘드 더 피리어드 - 데이 1
예측 상영일수: 207.43
실제 상영일수: 276
절대 오차    : 68.57
상대 오차    : 24.84%
```

Feature	Value
스크린수_sum_after	835.00
스크린수_sum	841.00
관객수_sum	22117.00
관객수_sum_after	21502.00
장르_서사	0.00
스크린수_mean	3.05
장르_서부	0.00
관객수_volatility	86.12

sparse feature로 인한 편향을 완화하기 위해
장르,국적 컬럼에 Truncated SVD 적용

< Truncated SVD 적용 후 예측 안정화 >

```
=== 샘플 분석 결과 ===
영화명       : 극장판 아이돌리쉬 세븐; 라이브 4비트 비욘드 더 피리어드 - 데이
예측 상영일수: 210.14
실제 상영일수: 276
절대 오차    : 65.86
상대 오차    : 23.86%
```



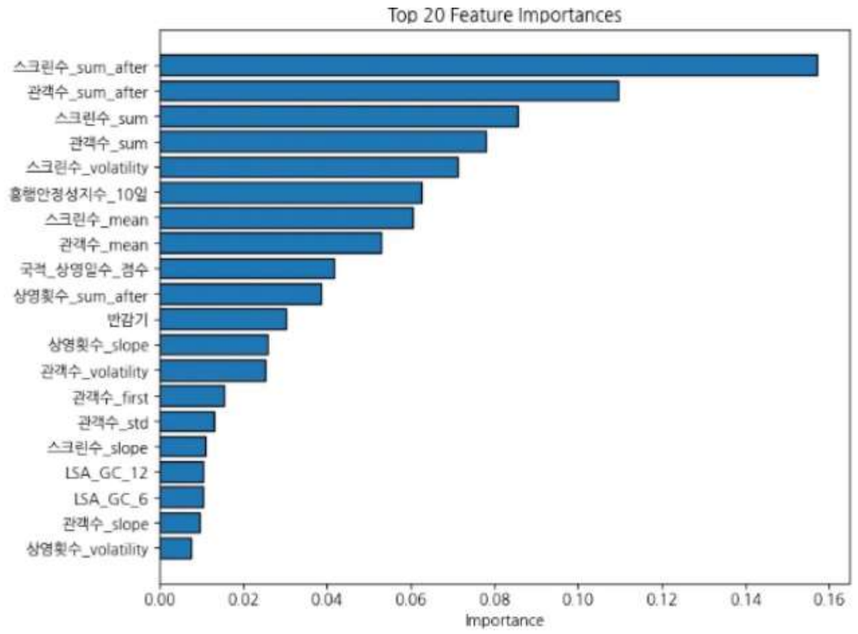
스크린수_sum	841.00
관객수_sum	22117.00
관객수_sum_after	21502.00
흥행안정성지수_10일	1006.49
스크린수_mean	3.05
상영횟수_volatility	1.06
스크린수_slope	-0.02
상영횟수_std	3.44
관객수_volatility	86.12

Chapter 05

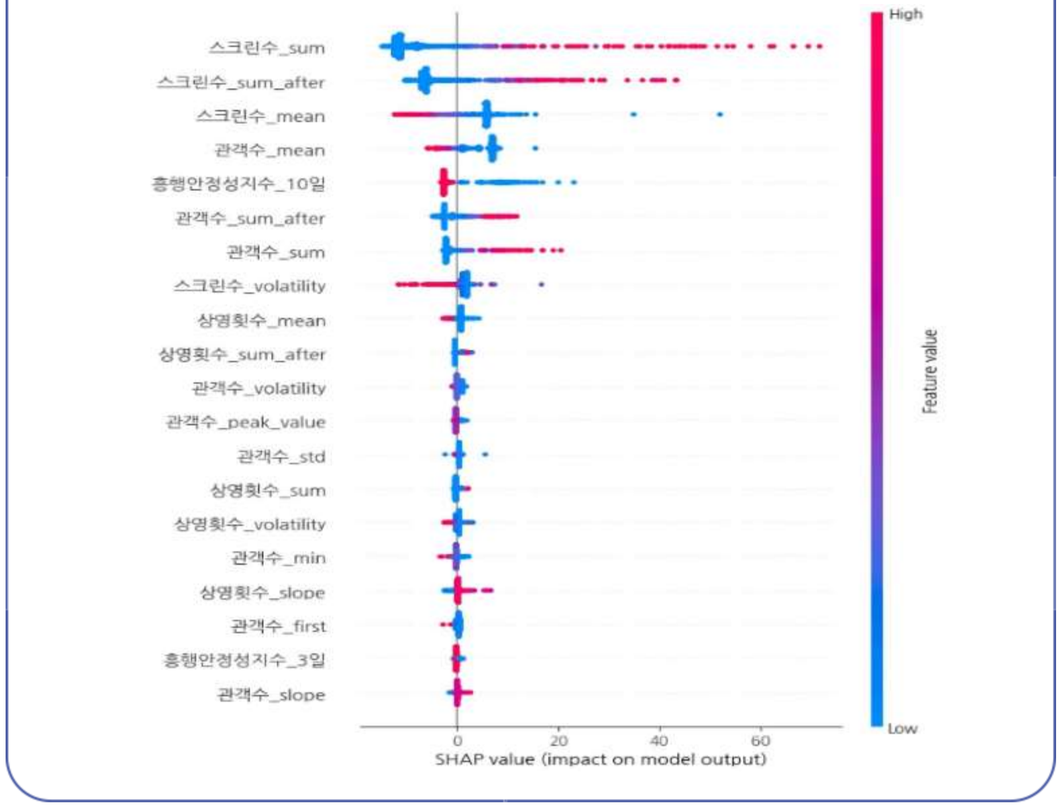
데이터 분석 - XGBoost

< 모델 학습 결과 >

Test | RMSE: 3.531 | MAE: 1.047 | R^2: 0.973



< SHAP 적용 >



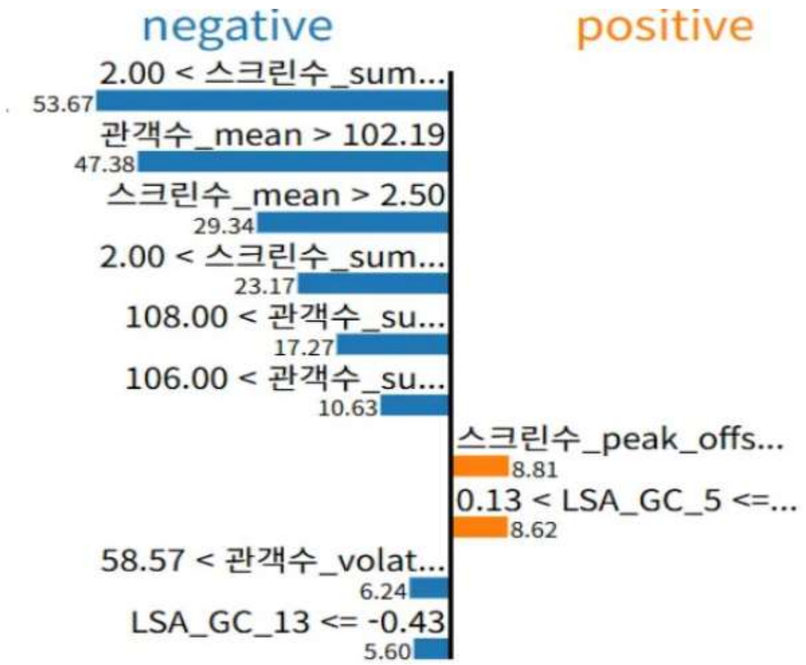
개봉 이후 추가 확보한 스크린수·관객 수를 특히 강하게 반영
안정성, 변동성 지표는 보조적으로 활용

Chapter 05

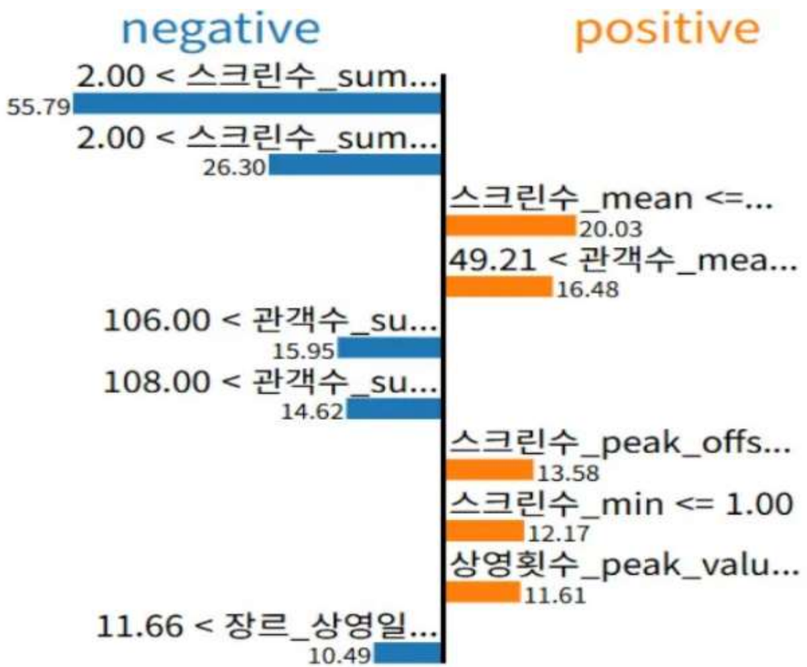
데이터 분석 - XGBoost

< LIME 적용 >

영화명 : 노바디즈 히어로
실제 상영일수 : 4일
예측 상영일수 : 4일



영화명 : 크리미널 트위스트
실제 상영일수 : 1일
예측 상영일수 : 1.1일



관객수 확보와 스크린 peak 지표 - 긍정적 작용

Chapter
05

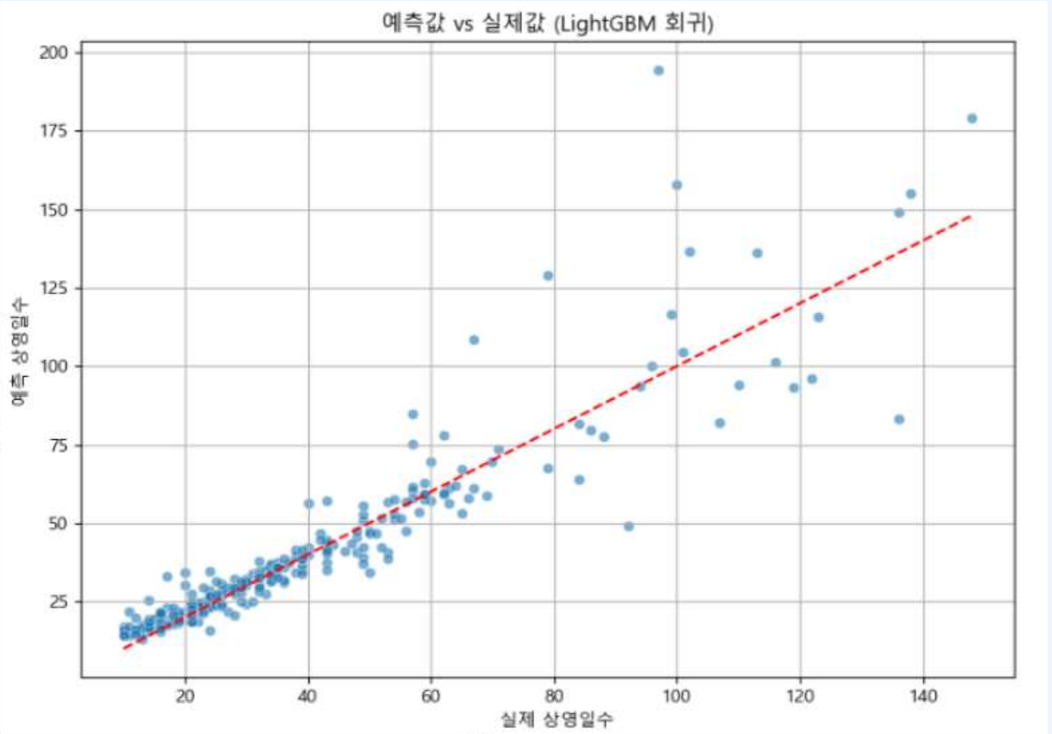
데이터 분석 - Light GBM

< 모델 학습 결과 >

MAE: 5.71
RMSE: 11.30
R^2: 0.826

< 상위 N개 피처로 재학습 결과 >

N = 5	N = 10
MAE: 5.91 RMSE: 10.56 R^2: 0.848	MAE: 5.80 RMSE: 10.29 R^2: 0.855
N = 20	**N = 30**
MAE: 6.58 RMSE: 10.56 R^2: 0.848	MAE: 6.54 RMSE: 10.66 R^2: 0.845

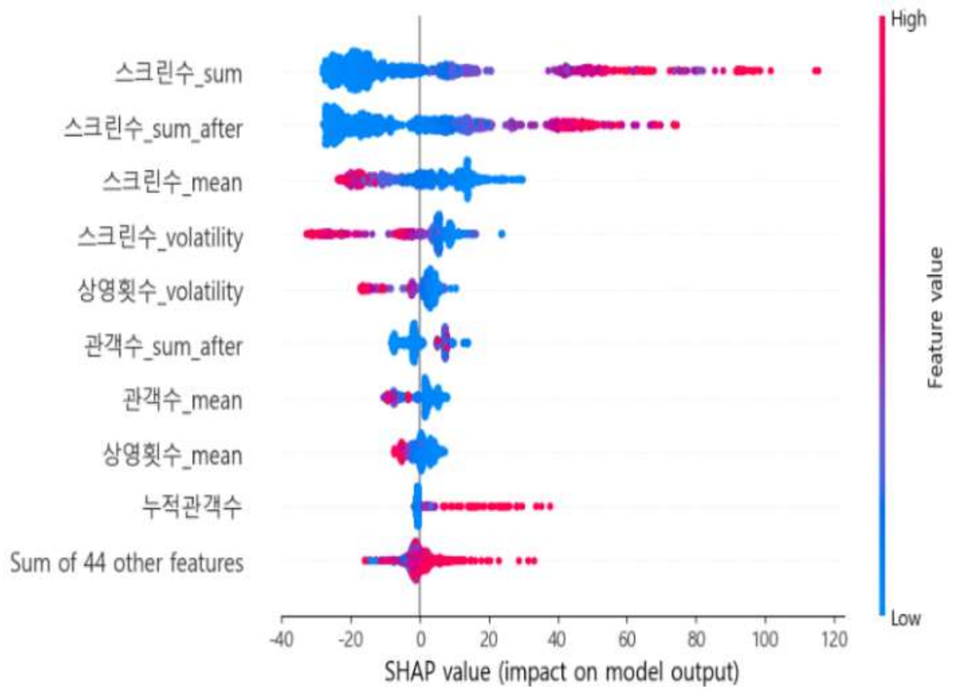


전반적으로 실제값과 유사
일부 고상영일 구간에서 예측 편차 증가

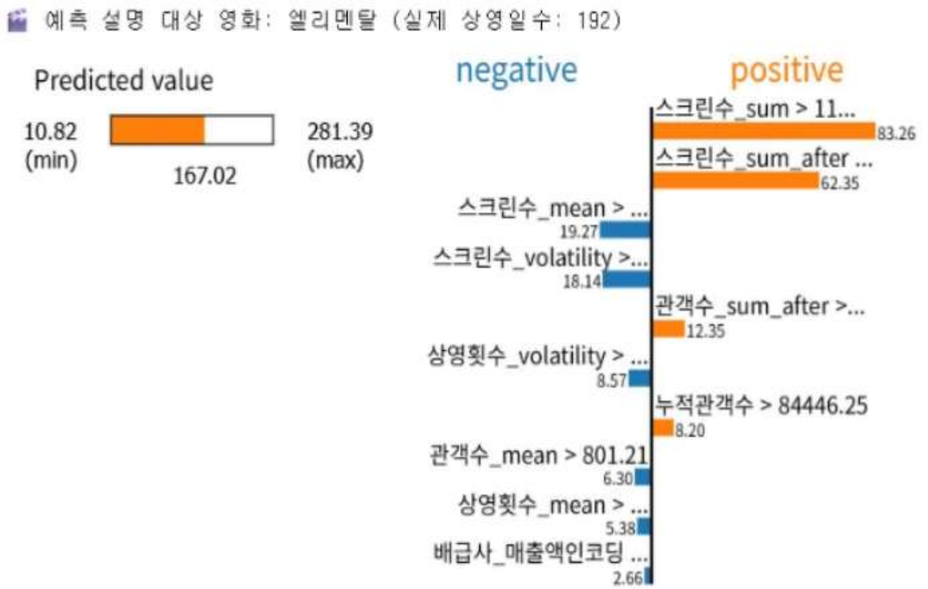
Chapter 05

데이터 분석 - Light GBM

< Feature 영향 분석 >



SHAP
스크린 수와 관객 수가 예측 결과에 주요하게 작용



LIME
해당 영화에서 스크린 수가 예측 상승에 주요하게 작용

Chapter 05

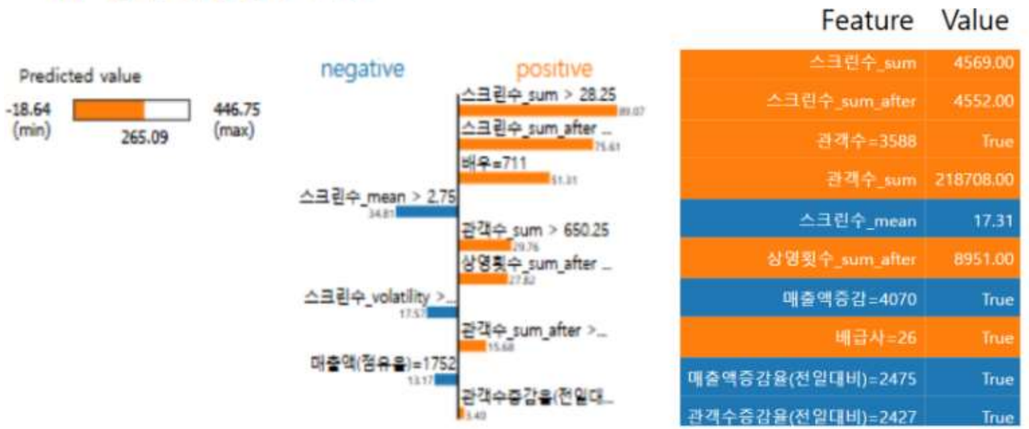
데이터 분석 - CatBoost

< 모델 학습 결과 >

테스트 MAE: 1.412, 테스트 R^2: 0.953

< LIME 적용 >

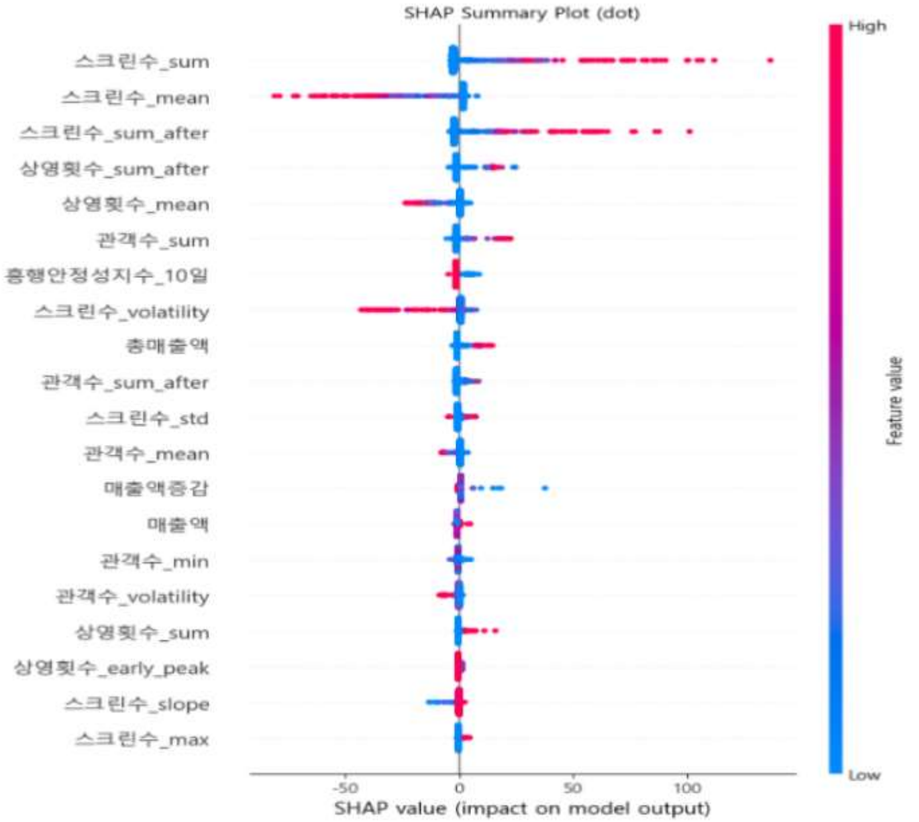
영화명 : 기적
예측 상영일수 : 91.94
실제 상영일수 : 88



Feature	Value
스크린수_sum	4569.00
스크린수_sum_after	4552.00
관객수=3588	True
관객수_sum	218708.00
스크린수_mean	17.31
상영횟수_sum_after	8951.00
매출액증감=4070	True
배급사=26	True
매출액증감율(전일대비)=2475	True
관객수증감율(전일대비)=2427	True

전일대비 매출 증감률이 부정적 영향 → 이례적 급등 판단
매출 변화의 지속성 부족 → 인기 하락으로 예측

< SHAP 적용 >



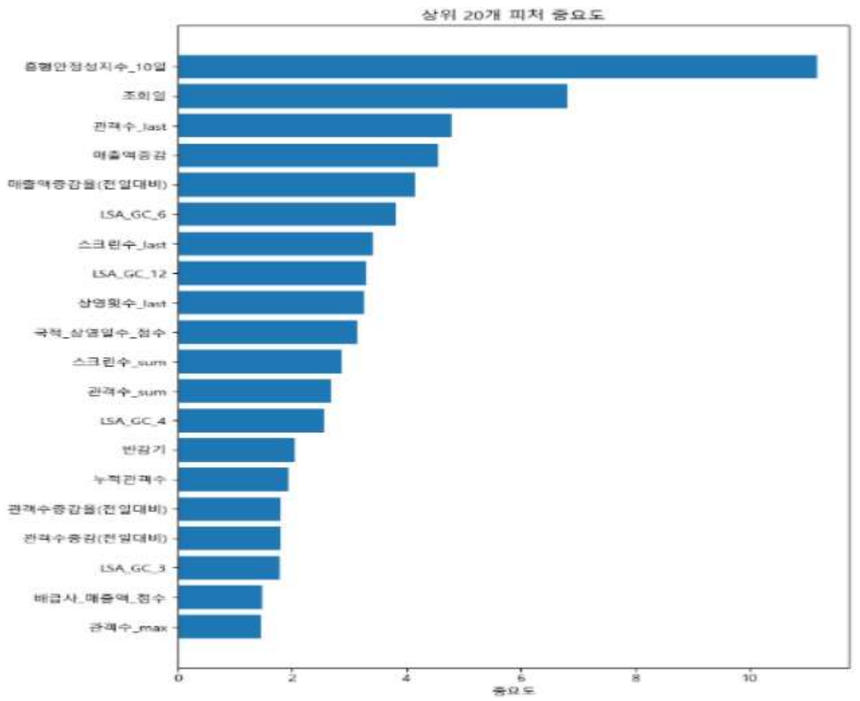
스크린수, 상영횟수의 지속성 → 예측에 긍정적 영향
변동성 높은 변수, 불안정한 지표 → 예측값 감소 영향

Chapter 06

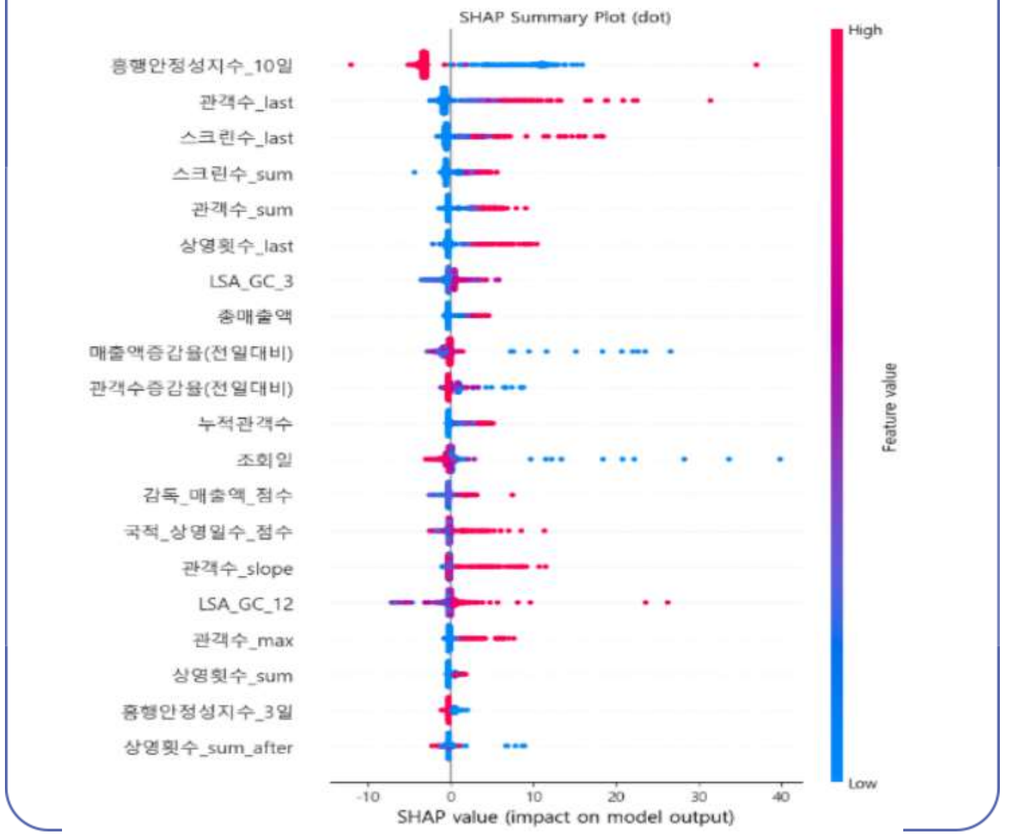
최종 모델 - CatBoost

< 10일치 모델 학습 결과 >

테스트 MAE: 4.766, 테스트 R²: 0.596



< SHAP 적용 >



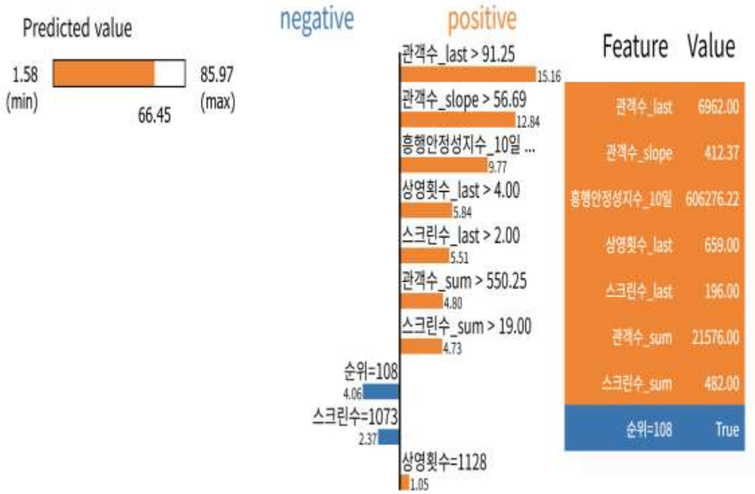
흥행안정성지수의 중요도 ↑
→ 관객의 안정적 유입이 결정적인 요인

Chapter 06

최종 모델 - CatBoost

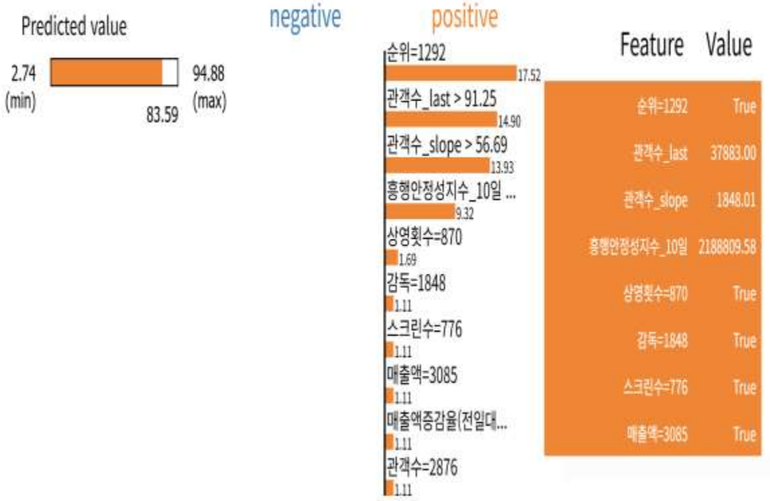
< LIME 적용 >

영화명 : 기적
실제 상영일수 : 88일
예측 상영일수 : 66.45일



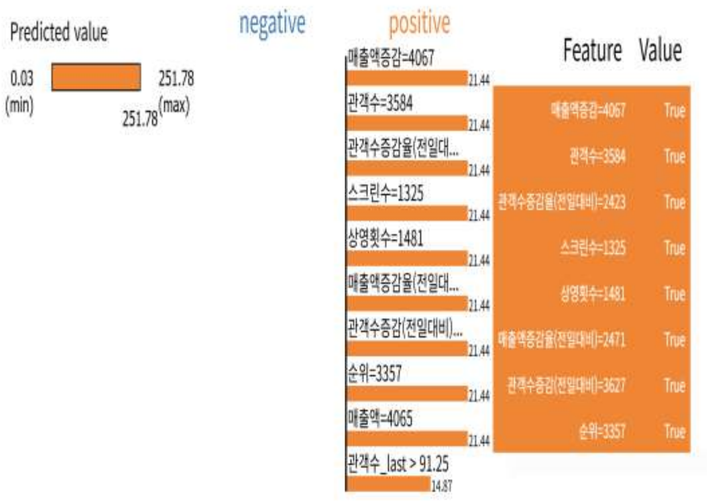
과거 흥행 저조 영화들과 유사한 패턴으로 모델이 보수적으로 판단

영화명 : 탈주
실제 상영일수 : 88일
예측 상영일수 : 83.59일



특히 흥행안정성지수가 긍정적인 영향으로 작용

영화명 : 극장판 하이큐!! 쓰레기장의 결전
실제 상영일수 : 264일
예측 상영일수 : 251.78일



흥행 흐름은 매우 뜨겁지만 흥행안정성지수가 낮아 단기 집중형 흥행으로 판단
→ 장기 상영일수 예측을 낮게 산출

Chapter
06

결론 및 기대효과

상영 스케줄 및 극장 운영 최적화

장기 상영 가능성 높은 작품 선별

스크린 확보 및 상영 기간 조정에 활용

OTT 진출 및 수익 모델 개선

Hold Back 기간 및 OTT 전환 시점 조정

극장 수익과 OTT 수익 간의 균형 조정 가능

2025-1 기계학습
팀프로젝트

감사합니다

Chapter
06

BUSINESS
PLAN

PART 2. 이곳에 이어질 내용의 주제를 입력해 보세요.

Chapter 1 키워드

Chapter 2 키워드

Chapter 3 키워드

ANY

Chapter
06

결론 및 기대효과

상영 스케줄 및 극장 운영 최적화

- 장기 상영 가능성 높은 작품 선별
- 스크린 확보 및 상영 기간 조정에 활용

OTT 진출 및 수익 모델 개선

- Hold Back 기간 및 OTT 전환 시점 조정
- 극장 수익과 OTT 수익 간의 균형 조정 가능

Chapter
02

4단 키워드 레이아웃

키워드에 대해 입력해주세요.

- 주제에 대한 세부내용을 간략하게 입력해주세요.

키워드에 대해 입력해주세요.

- 폰트는 프리텐다드 레귤러, 크기는 18pt 입니다.

키워드에 대해 입력해주세요.

- 주제에 대한 세부내용을 간략하게 입력해주세요.

키워드에 대해 입력해주세요.

- 폰트는 프리텐다드 레귤러, 크기는 18pt 입니다.

Goal

목표 키워드를
입력해주세요.

BUSINESS
PLAN

PART 2. 이곳에 이어질 내용의 주제를 입력해 보세요.

Chapter 1 키워드

Chapter 2 키워드

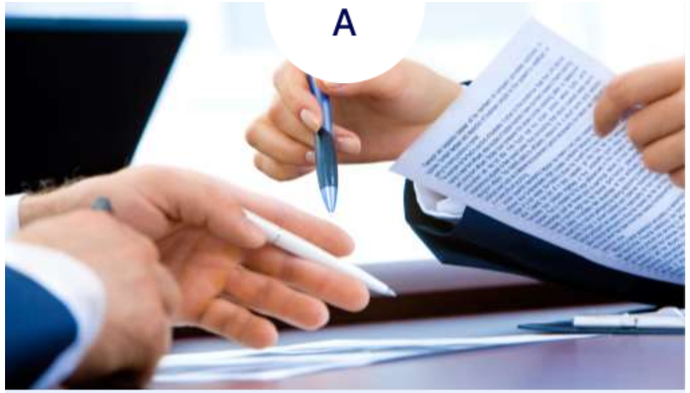
Chapter 3 키워드

ANY

Chapter
03

사진이 있는 3단 레이아웃

A



키워드A에 대해 입력해주세요.

주제에 대한 세부내용을 간략하게 입력해주세요.
폰트는 프리텐다드 레귤러, 크기는 18pt 입니다.
이곳에 위의 사진과 관련된 내용을 입력해주세요.

B

키워드B에 대해 입력해주세요.

주제에 대한 세부내용을 간략하게 입력해주세요.
폰트는 프리텐다드 레귤러, 크기는 18pt 입니다.
이곳에 위의 사진과 관련된 내용을 입력해주세요.

C

키워드C에 대해 입력해주세요.

주제에 대한 세부내용을 간략하게 입력해주세요.
폰트는 프리텐다드 레귤러, 크기는 18pt 입니다.
이곳에 위의 사진과 관련된 내용을 입력해주세요.

Chapter
04

수치로 설명하는 레이아웃

주요 키워드에 대한 수치를 함께 나타내야할때 사용하면 좋은 페이지입니다. 가독성의 핵심은 얼마나 내용을 간결하게 전달하는가 입니다. 쉽게 읽을 수 있고, 눈에 잘띄는 것이 중요해요.

45% 키워드 A

키워드 A에 대해 입력해주세요.

주제에 대한 세부내용을 간략하게 입력해주세요.
폰트는 프리텐다드 레귤러, 크기는 18pt 입니다.

30% 키워드 B

키워드B에 대해 입력해주세요.

이곳에 숫자에 대한 내용을 설명해주세요.
주제에 대한 내용을 간략하게 입력해주세요.

25% 키워드 C

키워드C에 대해 입력해주세요.

주제에 대한 세부내용을 간략하게 입력해주세요. 텍스트를 입력해주세요.

Chapter
05

2가지 비교 레이아웃

To Be		As - Is
키워드1에 대해 입력해주세요.	Keyword 1	키워드1에 대해 입력해주세요.
키워드2에 대해 입력해주세요.	Keyword 2	키워드2에 대해 입력해주세요.
키워드3에 대해 입력해주세요.	Keyword 3	키워드3에 대해 입력해주세요.

Insight 현재 상황과 이상적인 지향점에 대한 내용을 간략하게 기재해주세요. 폰트는 프리텐다드 입니다.

Chapter
06

일수 예측 모델 활용 가능성

상영 스케줄 및 극장 운영 최적화

장기 상영 가능성이 높은 작품을 조기에 선별

스크린 확보 및 상영 기간 조정에 활용

OTT 진출 및 수익 모델 개선

Hold Back 기간 및 OTT 전환 시점 조정

극장 수익과 OTT 수익 간의 균형 조정 가능