2021 계량경제분석과 데이터애널리틱스_프로젝트

미국수입영화의 국내흥행 예측

4조4조 최종발표

리천, 박민영, 서예은, 이준희, 정주희

목차

Literature Review

2012년 이전

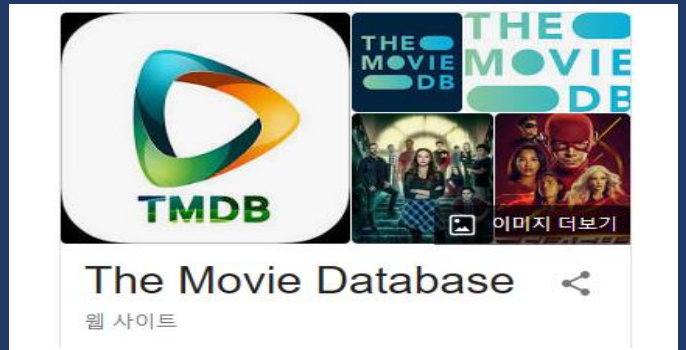
- 영화 흥행에 영향력을 미치는 **흥행 요인들의 선택**에 관한 연구
- **국가 간 영화흥행요인 비교를 위한 탐색적 연구: 한국과 미국 영화시장에서 미국 영화의 흥행요인 비교를 중심으로(이양환 외, 2007)**
  미국영화의 흥행요인이 한국에서의 **흥행요인과 유사**한지에 대해 독립변인으로 제작비, 관람등급, 상영시간, 배우, 비평가들의 평가, 관객들의 평가, 장르, 수상경력, 개봉 첫 주 상영관 수 등을 설정함.
  142개의 데이터셋 사용.
- **영화 유형별 영화 흥행 성과 예측 요인의 비교 연구: 예술 영화와 상업 영화 비교를 중심으로(김소영 외, 2010)**
  스크린수, 관객 평가, 장르는 예술 영화와 상업 영화 모두에 영향을 미침. 감독 명성, 상영등급, 전문가 평가, 배급사 영향력, 영화 제작국, 그리고 개봉 시기는 예술 영화의 흥행 성과에 대해서만 유의한 영향을 미쳤음.
  2008년 한국에 개봉된 **379편**의 영화 데이터셋 사용.

2012년 이후

- 흥행 요인들로부터 영화 **흥행 예측모델**을 연구
- **빅데이터 분석을 통한 영화 관객수, 매출액 예측 모델 (이응환, 우종필 , 2019)**
  SPSS를 통해 **천만관객 돌파유무**에 대해 회귀분석을 실시하였음.
  스크린 수, 평점 보다는 일평균뉴스 건수, 일평균TV보도 건수의 중요도가 높음을 밝혀냄.
  2015년, 2016년에 한국에서 개봉한 영화 데이터셋 사용.
- **딥러닝을 이용한 영화 흥행 예측과 주요 변수의 선택 연구 : 다변량 시계열 데이터 중심으로(변준형, 김지호, 최영진, 이홍철, 2020)**
  딥러닝(Multi-Layer Perceptron, Fully Convolutional Neural Networks, Residual Network)을 활용하여 천만관객 달성까지 걸리는 일수 예측(**누적관객수**)
  Residual Network가 가장 좋은 성능을 보였으며, 배급사 점수, 상영시간, 감독 점수, 개봉 월, 배우점수, 제작사 점수, NAVER 블로그 수, 상영 포맷을 주요 변수로써 그 유의성을 밝혀내었음.
  231개의 데이터셋 사용.

선행연구와의 차별성

- 국내에서 진행된 다수의 영화 흥행 예측 모델의 데이터가 최대 400개를 넘지 않아, 일반화시키기 힘듦.
  본 연구에서는 1,029개의 데이터를 사용함.
- 또한, 미국영화가 수입영화의 94%이상을 차지하고 있으므로 미국영화로 범위를 한정하여 더 높은 정확도의 예측모델을 구현해보고자 함.
- 기존의 연구가 국내영화와 수입영화의 구분을 짓지않고, 영화 자체의 흥행요소를 분석하고 이를 예측하기 위한 거시적인 관점에서의 접근이었다면, 본 연구를 통해 수입영화의 국내 흥행이라는 미시적으로 접근하여 보다 더 실증적이고 적용가능한 insight를 도출하고자 함.



kofic
KOBIS
영화관입장권통합전산망

Data & Methodology

- **연구목적**
  - **미국 수입영화의 국내 흥행 예측 모델 생성**
- **데이터의 수집**
  - (한국) 영화관입장권통합전산망: 2003~2020 데이터 수집 (총 3,127개의 데이터)
    - 국내 통합전산망에 관객수가 입력된 시점이 2003년부터임을 확인.
  - (미국) TMDB(The Movie DataBase): 1916~2016 Kaggle Data 수집 (총 4,803개의 데이터)
  - (중국) mtime.com -> 중국영화리뷰사이트 : 감독 및 주연배우 2명까지의 nomination 및 awards 이력수집
- **데이터의 전처리**
  - **Main Dataset**
  - 영화 영문명을 기준으로 파이썬을 이용한 merge(innerjoin) 처리
  - 결과적으로 1,029개의 데이터 사용하여 최종 데이터세트 생성
- **방법론**
  - Linear Regression
  - Random Forest

변수의 설명

[표 1]

변수	변수 영문명	변수 국문명	변수설명
종속변수	**audience_KOR**	**한국관객수**	한국에서 동원된 관객수.
독립변수	revenue_US	미국영화제작수익	미국에서의 영화 제작에 대한 수익.
	budget	제작비용	영화 제작 시 투입된 비용.
	popularity_US	선호도	TMDB 내에서 독자적으로 누적계산된 지표. 일일 투표수, 일일 시청수, 좋아요 수, 이전 점수 등을 합산하여 계산됨.
	release_diff	개봉연월 차이	미국의 개봉연월과 한국의 개봉연월의 차이. (단위: 개월)
	vote_diff	평균평점차이	미국의 평균평점과 한국의 평균평점의 차이. 미국과 한국의 상대적 감성을 표현하는 대리변수.
	vote_average	평균평점	TMDB에서 소비자에 의해 평가된 영화 평점의 평균.
	vote_count	추천수	TMDB에서 평점을 매긴 소비자의 수.
	runtime	상영시간	영화의 상영시간. 분 단위로 측정됨.
	nomi_DIR	감독노미네이션횟수	시상식에서의 감독 노미네이션 횟수.
	awards_DIR	감독수상횟수	시상식에서의 감독 수상 횟수.
	nomi_Actor1	배우1노미네이션횟수	시상식에서의 배우1 노미네이션 횟수.
	awards_Actor1	배우1수상횟수	시상식에서의 배우1 수상 횟수.
	nomi_Actor2	배우2노미네이션횟수	시상식에서의 배우2 노미네이션 횟수.
	awards_Actor2	배우2수상횟수	시상식에서의 배우2 수상 횟수.
	production_year	제작연도	영화가 제작된 연도. 더미변수.
	distributor_KOR	한국배급사	영화 배급을 맡은 한국 배급사.
	Genre	장르	영화의 장르. 여러가지가 있을 경우 대표적인 장르 하나만 선택됨.

기술통계량

[표 2]

	audience_KOR	revenue_US	budget	popularity_US	release_diff(US-KOR)	vote_diff(US-KOR)	vote_average_US	vote_count_US
count	1,029	1,029	1,029	1,029	1,029	1,029	1,029	1,029
mean	656,769	204,800,197	66,127,941	46	6	-2	6	1,660
std	1,198,509	241,753,450	59,218,786	58	17	23	1	1,880
min	0	0	0	0	0	-723	0	0
25%	38,305	45,300,000	20,000,000	20	0	-2	6	438
50%	200,038	114,281,051	50,000,000	35	1	-1	6	1,043
75%	695,566	275,293,450	99,000,000	53	4	-1	7	2,225
max	10,296,101	1,519,557,910	380,000,000	876	180	5	8	13,752

	runtime	nomi_DIR	awards_DIR	nomi_Actor1	awards_Actor1	nomi_Actor2	awards_Actor2	production_year
count	1,029	1,029	1,029	1,029	1,029	1,029	1,029	1,029
mean	111	3	1	6	2	4	1	2,010
std	19	9	4	12	5	9	4	4
min	73	0	0	0	0	0	0	2,002
25%	97	0	0	0	0	0	0	2,007
50%	108	0	0	0	0	0	0	2,010
75%	122	2	0	7	1	4	0	2,013
max	187	67	34	89	35	89	35	2,019

회귀분석 - 1

[Full Model]

$$log(audience\_KOR) = \beta0 + \beta1log(revenue\_US) + \beta2log(budget) + \beta3popularity\_US + \beta4release\_diff + \beta5vote\_diff + \beta6vote\_average\_US + \beta7vote\_count\_US + \beta8runtime + \beta9nomi\_DIR + \beta10awards\_DIR + \beta11nomi\_Actor1 + \beta12awards\_Actor1 + \beta13nomi\_Actor2 + \beta14awards\_Actor2 + dummy\_ProductionYear$$

[표 3]

종속변수와의 다중공선성 확인				
log(revenue_US)	log(budget)	popularity_US	release_diff	vote_diff
1.686639	1.440593	2.014181	1.200650	1.014426
vote_average_US	vote_count_US	runtime	nomi_DIR	awards_DIR
1.559324	3.069315	1.512951	6.666450	6.158828
nomi_Actor1	awards_Actor1	nomi_Actor2	awards_Actor2	
4.350778	3.568505	3.776937	3.223682	

회귀분석 - 2

[표 4] 회귀분석 (Full Model)

	estimate	Std. Error	t-value	Pr(>| t |)
(Intercept)	-4.39	2.893	1.517	0.129508
log(revenue_US)	**0.2739**	**0.06657**	**4.115**	**4.20E-05 *** **
log(budget)	**0.5375**	**0.06917**	**7.771**	**1.93E-14 *** **
popularity_US	**0.003816**	**0.001871**	**2.039**	**0.041715 * **
release_diff	**-0.0654**	**0.004918**	**-13.297**	**< 2e-16 *** **
vote_diff	-0.002051	0.003394	-0.604	0.545769
vote_average_US	**0.4334**	**0.1221**	**3.549**	**0.000405 *** **
vote_count_US	**0.0002293**	**0.00007089**	**3.235**	**0.001258 ** **
runtime	-0.001364	0.004982	-0.274	0.784257
nomi_DIR	-0.01146	0.02304	-0.497	0.619018
awards_DIR	-0.009606	0.04509	-0.213	0.831359
nomi_Actor1	9.264E-05	0.01333	0.007	0.994457
awards_Actor1	0.00337	0.02877	0.117	0.906782
nomi_Actor2	-0.008765	0.01689	-0.519	0.603965
awards_Actor2	0.004397	0.03784	0.116	0.907502

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.434 on 997 degrees of freedom
Multiple R-squared: **0.4271**
F-statistic: 23.97 on 31 and 997 DF, p-value: < 2.2e-16

movie.2 <- movie[-c(434,435,999,993,982,371)]

[표 5] 이상치 제거 후 회귀분석 (Full Model)

	estimate	Std. Error	t-value	Pr(>| t |)
(Intercept)	-3.346	2.715	-1.232	0.218131
log(revenue_US)	**0.2495**	**0.06258**	**3.987**	**7.19E-05 *** **
log(budget)	**0.5332**	**0.0649**	**8.215**	**6.56E-16 *** **
popularity_US	**0.003118**	**0.001756**	**1.776**	**0.076089 . **
release_diff	**-0.08342**	**0.005112**	**-16.318**	**< 2e-16 *** **
vote_diff	-0.001851	0.00318	-0.582	0.560573
vote_average_US	**0.3745**	**0.1148**	**3.263**	**0.00114 ** **
vote_count_US	**0.000243**	**0.00006664**	**3.647**	**0.000279 *** **
runtime	-0.00276	0.004678	-0.59	0.555286
nomi_DIR	-0.007604	0.0216	-0.352	0.72483
awards_DIR	-0.02507	0.0423	-0.593	0.553533
nomi_Actor1	-0.000717	0.01249	-0.057	0.954271
awards_Actor1	0.005635	0.02696	0.209	0.834492
nomi_Actor2	-0.00656	0.01584	-0.414	0.678855
awards_Actor2	-0.007429	0.03551	-0.209	0.834322

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.281 on 991 degrees of freedom
Multiple R-squared: **0.4774**
F-statistic: 29.21 on 31 and 991 DF, p-value: < 2.2e-16

회귀분석 - 3

[표 6] 전진선택법 (Forward Selection)

AIC=1718.59	estimate	Std. Error	t-value	Pr(>| t |)
(Intercept)	-3.346	2.715	-1.232	0.218131
log(revenue_US)	**0.2495**	**0.06258**	**3.987**	**7.19E-05 *** **
log(budget)	**0.5332**	**0.0649**	**8.215**	**6.56E-16 *** **
popularity_US	**0.003118**	**0.001756**	**1.776**	**0.076089 . **
release_diff.US.KOR.	**-0.08342**	**0.005112**	**-16.318**	**< 2e-16 *** **
vote_diff.US.KOR.	-0.001851	0.00318	-0.582	0.560573
vote_average_US	**0.3745**	**0.1148**	**3.263**	**0.00114 ** **
vote_count_US	**0.000243**	**0.00006664**	**3.647**	**0.000279 *** **
runtime	-0.00276	0.004678	-0.59	0.555286
nomi_DIR	-0.007604	0.0216	-0.352	0.72483
awards_DIR	-0.02507	0.0423	-0.593	0.553533
nomi_Actor1	-0.000717	0.01249	-0.057	0.954271
awards_Actor1	0.005635	0.02696	0.209	0.834492
nomi_Actor2	-0.00656	0.01584	-0.414	0.678855
awards_Actor2	-0.007429	0.03551	-0.209	0.834322

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.27 on 1009 degrees of freedom
Multiple R-squared: 0.4729,
F-statistic: 69.64 on 13 and 1009 DF, p-value: < 2.2e-16

[표 7] 후진제거법 (Backward Elimination) & 단계적선택법 (Stepwise Selection)

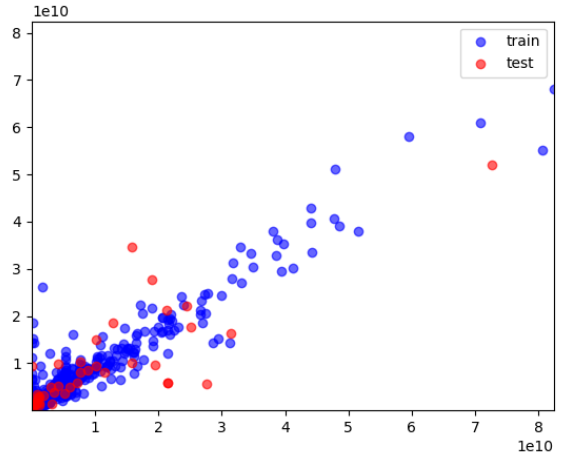
AIC=1691.39	estimate	Std. Error	t-value	Pr(>| t |)
(Intercept)	**-4.443**	**1.504**	**-2.955**	**0.003203 ** **
log(revenue_US)	**0.2511**	**0.06182**	**4.061**	**5.26E-05 *** **
log(budget)	**0.5215**	**0.06332**	**8.236**	**5.49E-16 *** **
popularity_US	**0.0031**	**0.001738**	**1.784**	**0.074744 . **
release_diff	**-0.08242**	**0.005002**	**-16.478**	**< 2e-16 *** **
vote_average_US	**0.3696**	**0.1084**	**3.409**	**0.000677 *** **
vote_count_US	**0.0002293**	**0.00006399**	**3.583**	**0.000355 *** **
awards_DIR	**-0.04477**	**0.01745**	**-2.565**	**0.01045 * **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

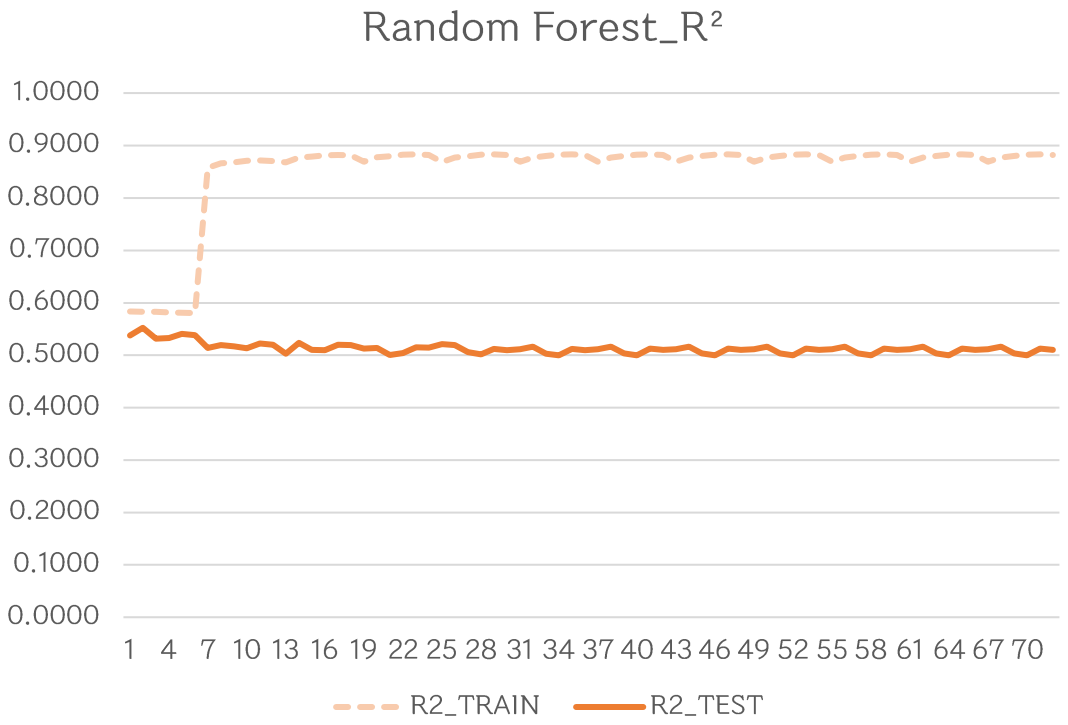
Residual standard error: 2.27 on 1009 degrees of freedom
Multiple R-squared: 0.4729,
F-statistic: 69.64 on 13 and 1009 DF, p-value: < 2.2e-16

랜덤포레스트

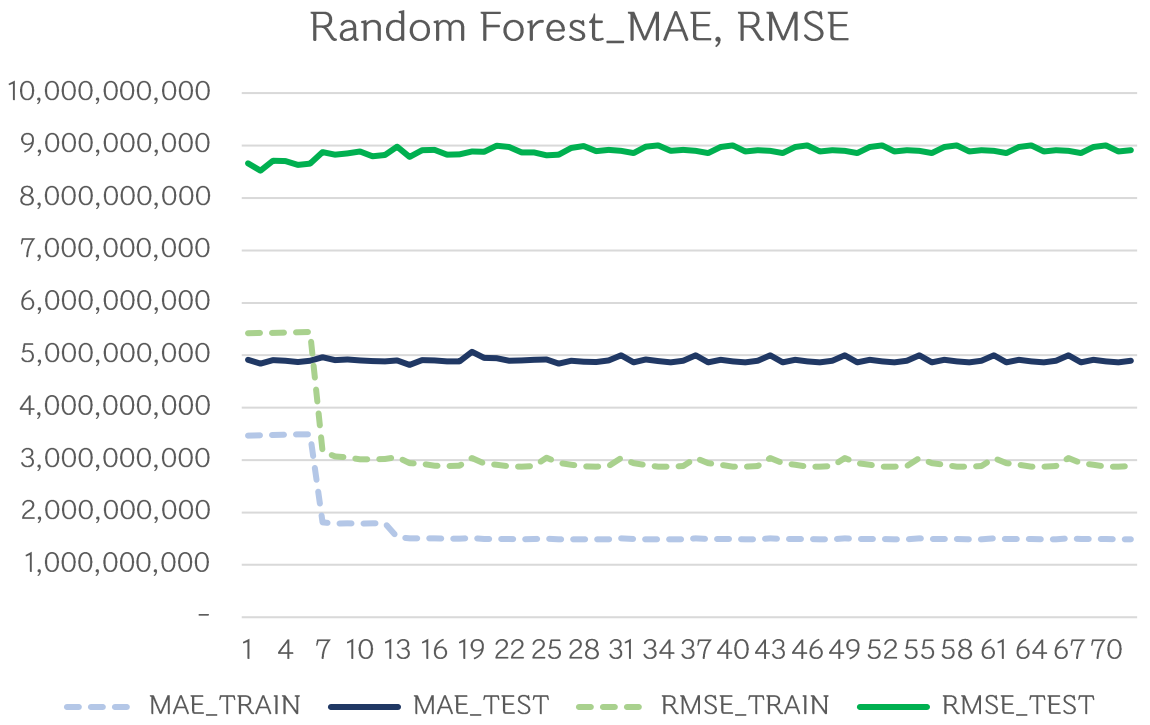
[그림 3]



[그림 1]



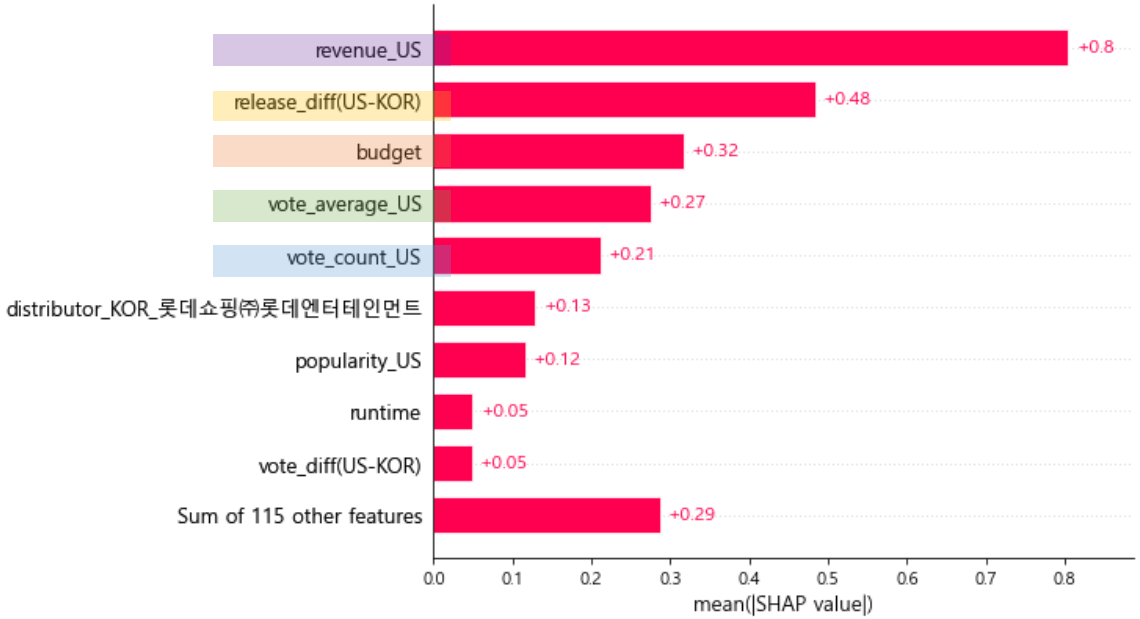
[그림 2]



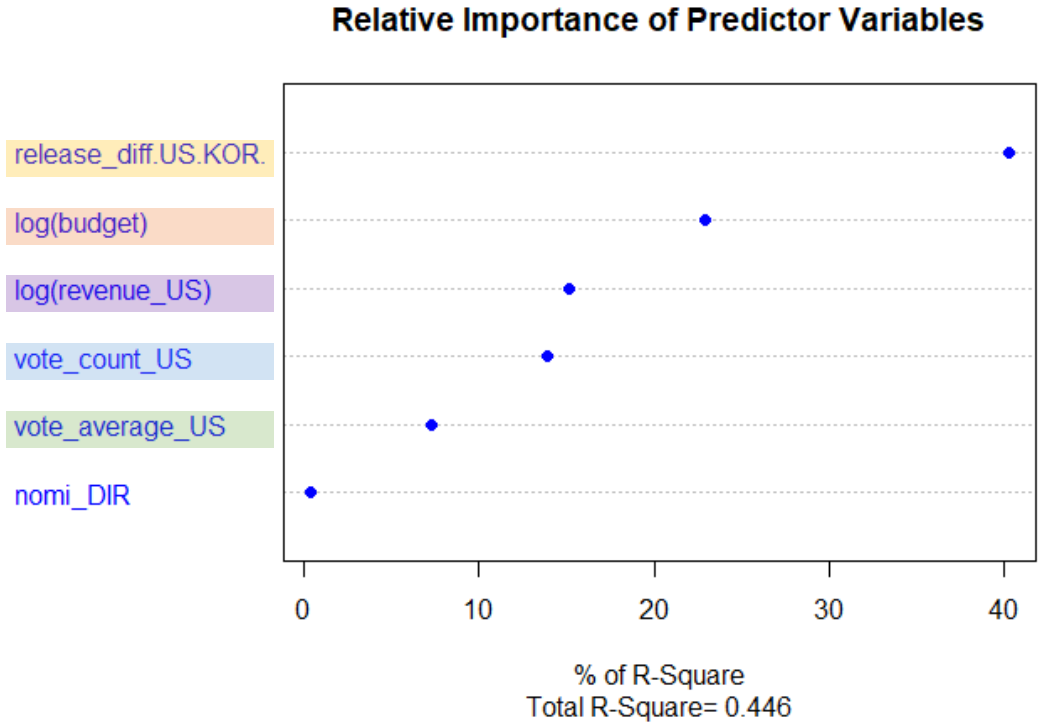
SHAP value - 1

SHapley Additive exPlanations

[그림 4] SHAP value 절대값의 평균



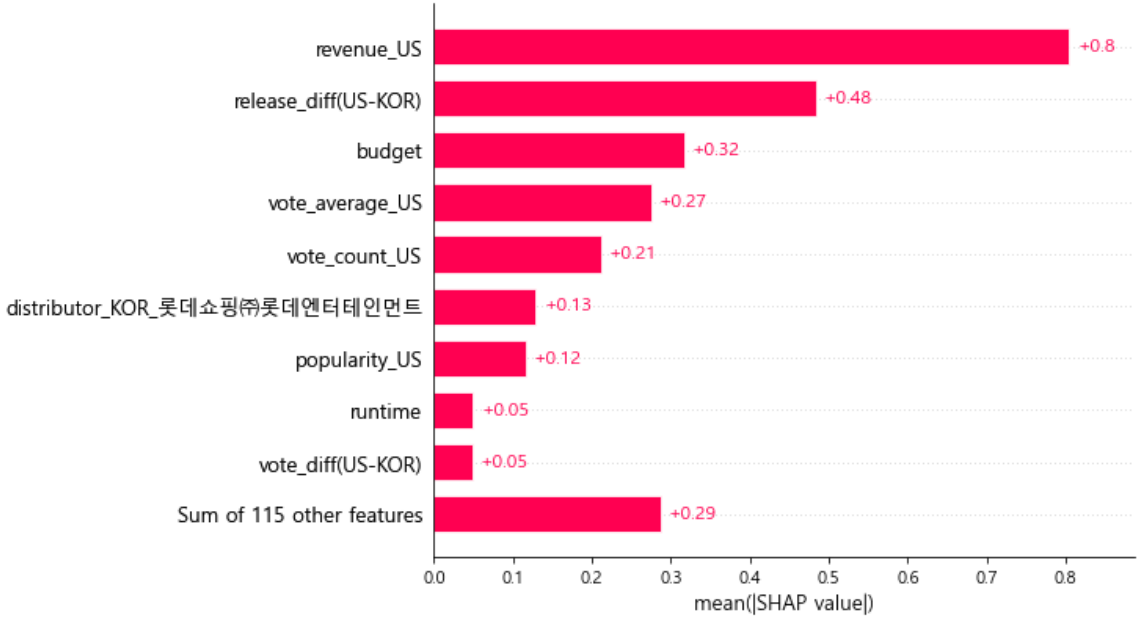
[그림 5] R을 이용한 주요인 판별



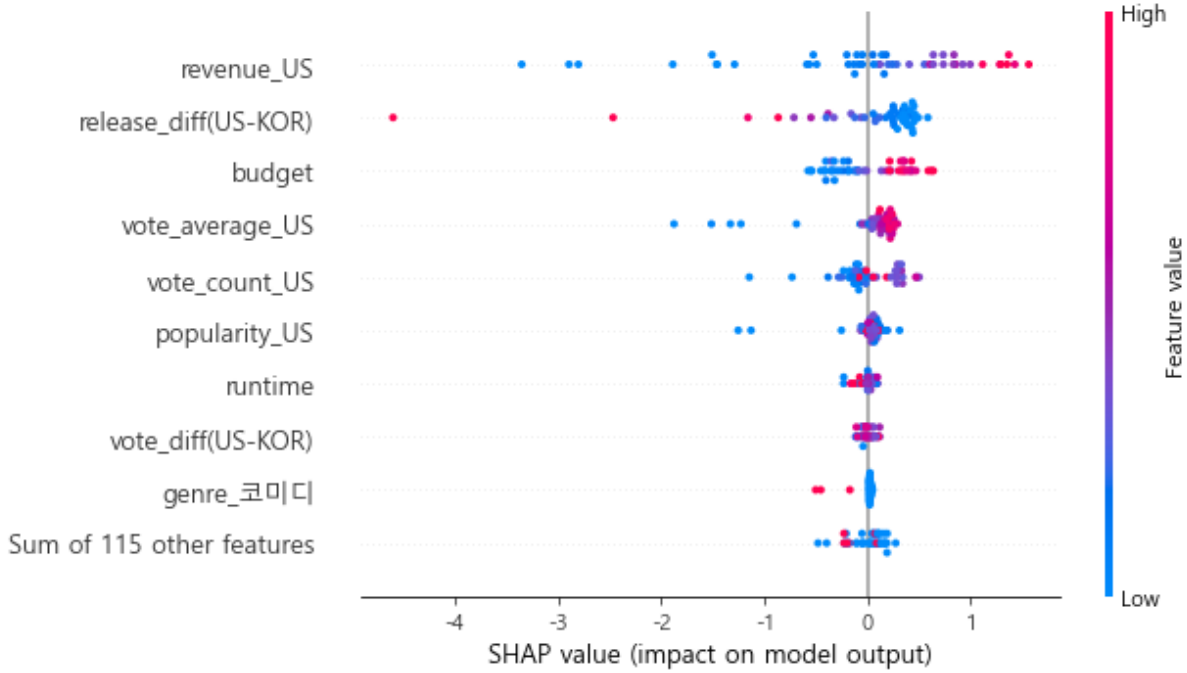
SHAP value - 2

SHapley Additive exPlanations

[그림 4] SHAP value 절대값의 평균

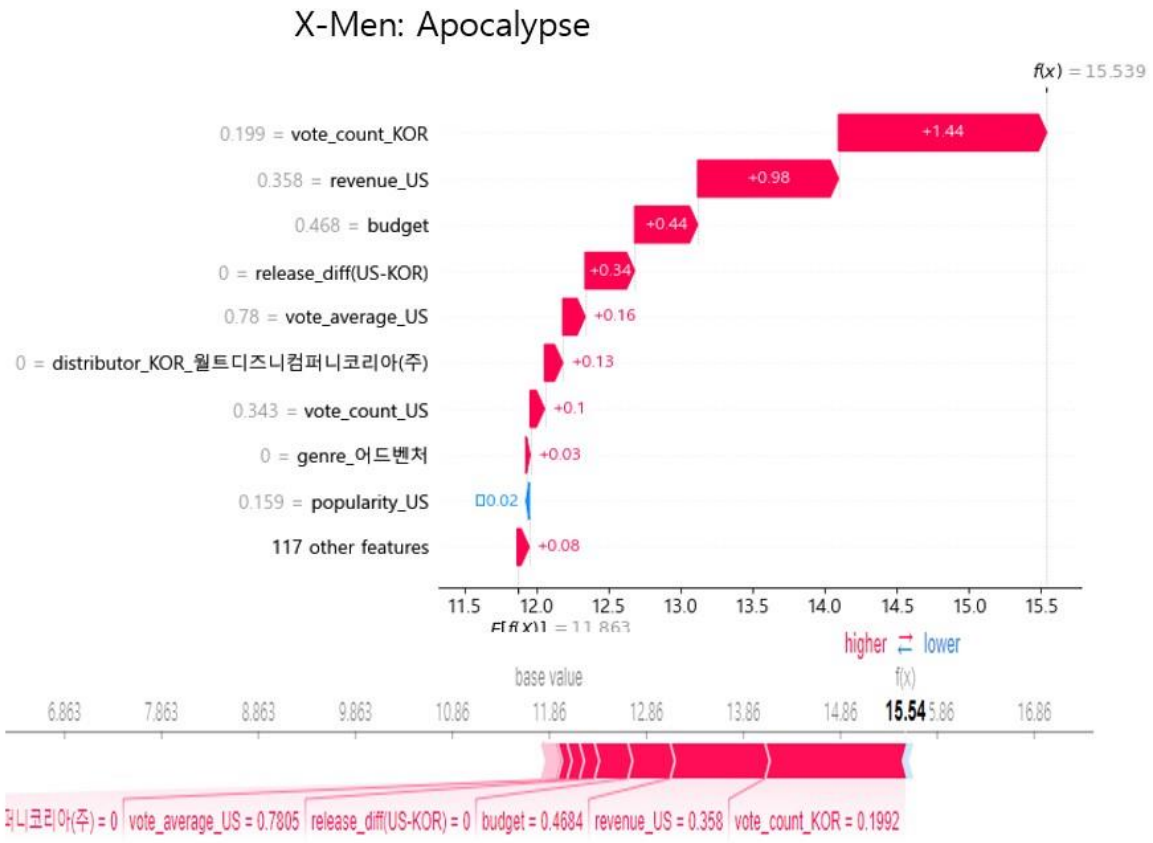


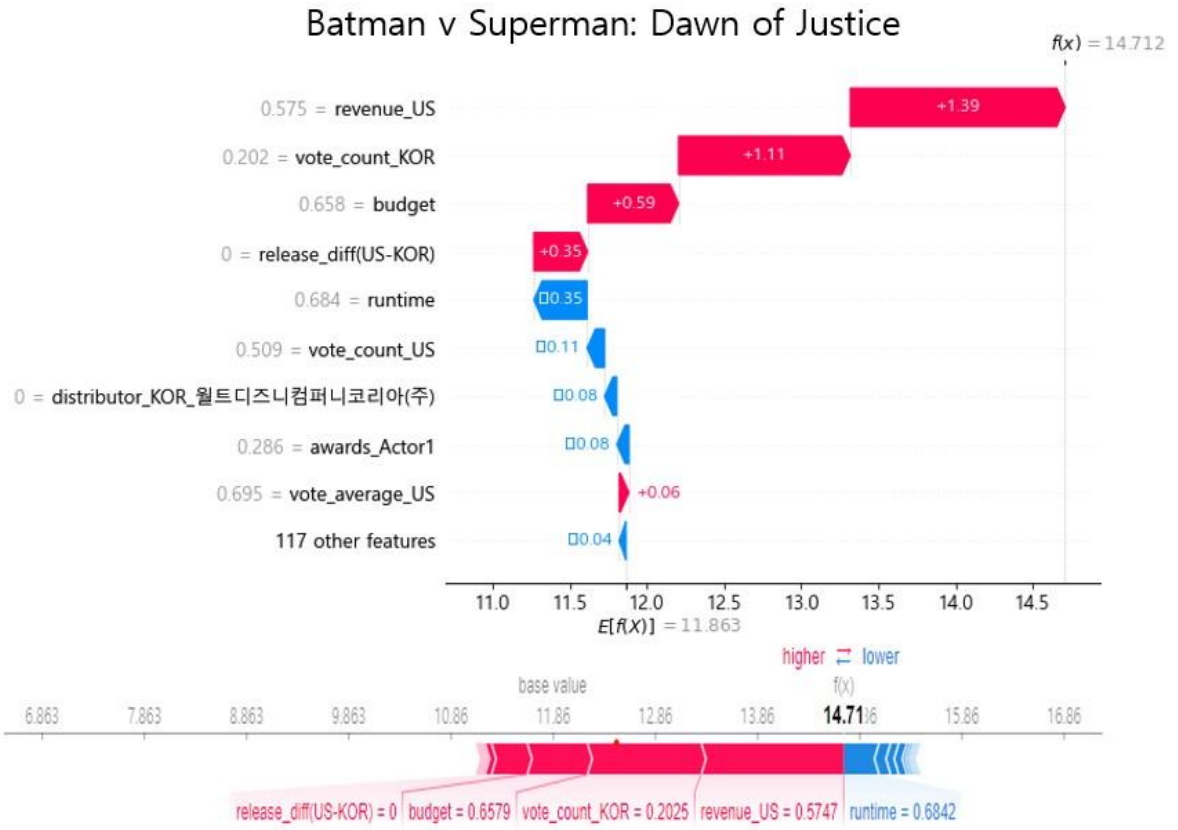
[그림 6] SHAP value (Impact on model output)



SHAP value - 3 SHapley Additive exPlanations

[그림 7] 각 영화에 대한 SHAP value





Conclusion

Finding

- 미국수입영화의 흥행수익, 투여된 제작비용, 영화의 선호도, 평점은 국내 관객수(영화흥행)에 긍정적인 영향을 미친다.
- 미국수입영화의 국내 개봉일은 늦춰질 수록 국내관객수(영화흥행)에 부정적인 영향을 미친다.

Implication

- 수입영화가 국내에 흥행할 수 있는 요인들을 분석해 내어 의사결정 시 합리적인 판단이 가능한 가이드를 제시함.
- SHAP 방법론을 사용한 각 영화에 대한 개별적 흥행 요인 분석과 이를 통한 insight 도출이 가능함.

Future Study

- 미국 수입영화 데이터만 제한적으로 사용됨. 데이터의 다양화 필요.
- 프로젝트에 활용된 변수의 다양성 확장 필요. (배우파워, 감독파워 등)
- 머신러닝 및 딥러닝 예측 방법에 대한 다양한 시도 및 예측 정확도 개선 필요.

Q & A

감사합니다.

본 프로젝트를 위해 고생해주신 4조4조 조원분들 감사합니다 ☺