Prediction of Korean Box Office using Data Crawling

Ji Won Yang, Ye Jin Lee, Ha Rang Han
Department of Statistics, Sookmyung Women's University.

Introduction

한 달에 수 십 개의 영화가 나오는 시점에서, 영화가 손익분기점을 넘을 것인가에 대한 관심은 계속 존재해왔다. 또한 지난 2월 영화진흥위원회는 영화산업 경제 민주화 제도 마련과 관련된 내용으로 스크린 상한제 도입 즉 대기업 배급사의 배급 및 상영 겸업 등으로 인한 불공정성 문제 해소 관련 내용에 대해 요청문을 발표하였다. 따라서, 이런 현황에 대하여 과연 어떤 내부적, 외부적 요인들이 영화 관객수에 영향을 끼칠지에 대하여 탐색하고 그 요인들을 통해서 영화의 누적 관객수를 예측하고자 한다. 또한 스크린 상한제가 필요한 제도인지를 확인해보고자 한다. 이러한 결과는 영화 개봉 시 배급사에서 나이브 벤치마크로써 참고 자료로 사용할 수 있을 것이고, 스크린 상한제에 대한 논의에 대한 새로운 시각을 제시 할 수 있을 것이다.

본 보고서에서는 2020년까지의 한국 누적 박스오피스에 해당하는 영화 630편을 이용하여 학습한다. 데이터 셋의 수가 적어 K-Fold를 통한 교차검증을 진행하고, 성과 평가 척도로 RMSE, MAE, F1-score, Precision을 이용하여 성능을 평가한다. Lasso, GLM , Gradient Boost, RandomForest, Decision Tree 등을 모델을 사용한다.

Data Crawling

-**Dataset** : 총 630개 영화 데이터

->누적 관객수 1만 이상, 2004년 개봉 이후의 데이터 기준

-**Crawling Site** : 뉴스검색수(네이버 뉴스), 평점&상영시간&장르(네이버 영화), 그 외 변수(영화관입장권통합전산망(KOBIS), OPEN API)

-**Target variable** : 영화 누적 관객수

-**Predictor variable** : 개봉일, 스크린 수, 국적, 배급사, 등급, 평점, 뉴스검색수, 경쟁영화수, 상영시간, 장르 계절, 개봉 요일, 개봉 월, 주말 지시변수

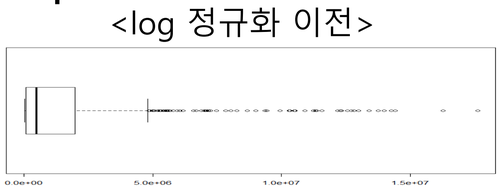
-**Correlation between factors of movie and cumulative audience**

	log(누적관객수)	평점	뉴스	경쟁영화수	상영시간	스크린수
log(누적관객수)	1					
평점	0.3909 **	1				
뉴스	0.6448**	0.1255 **	1			
경쟁영화수	-0.1342	-0.0953 **	0.1222 **	1		
상영시간	0.4886 **	0.1674 **	0.3845 **	-0.1764 **	1	
스크린수	0.7600 **	0.2073 **	0.7149 **	0.2399 **	0.3846 **	1

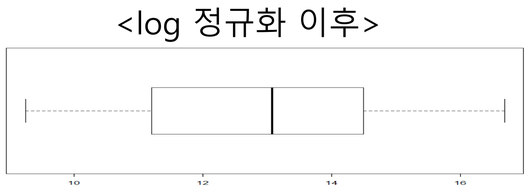
Experimental method

Box plot

<log 정규화 이전>



<log 정규화 이후>

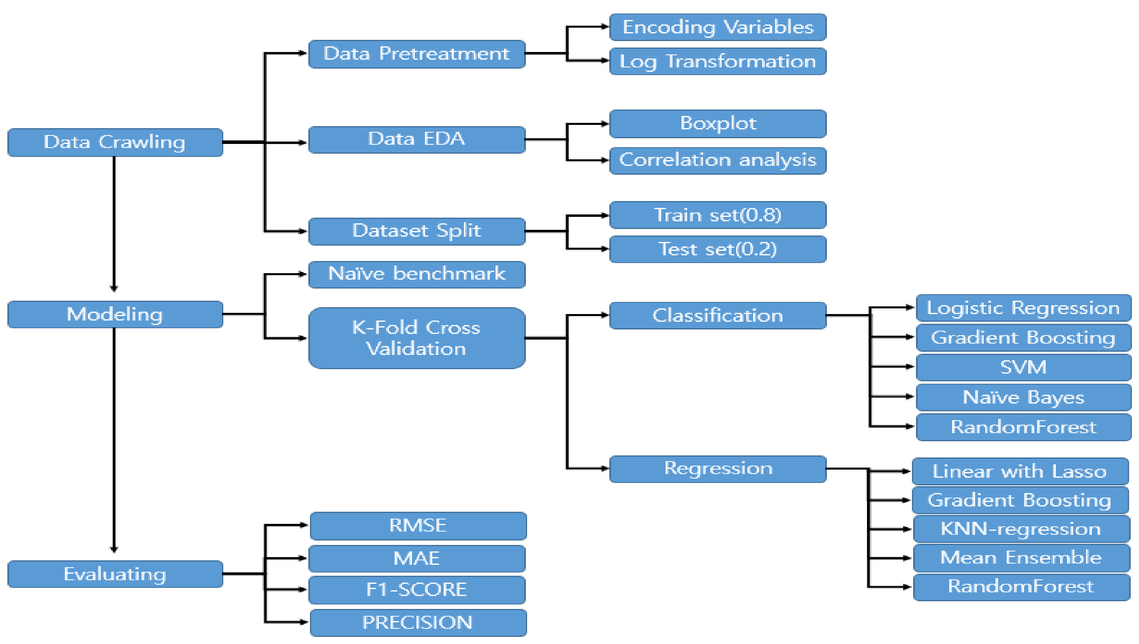


Target variable의 범위가 매우 광범위하면서도 한쪽에 쏠려 있음을 알 수 있다. 또한 1만명 이상인 영화를 대상으로 자료를 수집하여 편차가 굉장히 크다. 따라서 누적 관객수에 대하여 log 변환을 시행하여 분포를 안정화 시켜 학습에 사용한다

Naive Benchmark

Target variable인 누적 관객 수의 평균 값을 이용하여 구한 RMSE와 MAE을 나이브 벤치마크로 삼는다. 누적 관객수에 로그 변화를 하여 예측하였을 때 RMSE값은 2.0157이며 MAE값은 1.743451이다.

Flow chart



Experimental Results

Validation error for regression

	RMSE	MAE
Linear regression	0.9685	0.7683
Linear regression with lasso (lambda=0.03125)	0.9534	0.7638
KNN (k=7)	1.2337	0.9756
RandomForest	2.5829	2.0944
XGBoost	0.7759	0.6079
LightGBM	0.7762	0.6109
Mean_Ensemble	**0.7590**	**0.5942**

Validation error for classification (class=2)

		누적 관객수 300만 기준		누적 관객수 500만 기준	
		Precision	F1-score	Precision	F1-score
Logistic Regression	STD	0.83	0.84	0.90	0.92
	SMOTE	0.53	0.52	0.55	0.53
Gradient Boost	STD	0.93	0.95	0.94	0.96
	SMOTE	0.92	0.94	0.96	0.96
XGBoost	STD	0.94	0.96	0.95	0.96
	SMOTE	**0.97**	**0.96**	**0.99**	**0.97**
LightGBM	STD	0.90	0.94	0.94	0.96
	SMOTE	0.97	0.95	0.98	0.97

Validation error for classification (class=4 and class=3)

	F1-score				Precision			
	0	1	2	3	0	1	2	3
SVM	0.87	0.46	0.00	0.60	0.81	0.45	0.00	0.78
Logistic Regression	0.88	0.49	0.12	0.58	0.89	0.48	0.15	0.55
Naïve Bayes	0.84	0.38	0.16	0.55	0.87	0.35	0.15	0.56
Grandient Boost	0.87	0.46	0.12	0.51	0.81	0.49	0.23	0.63
Random Forest	0.87	0.42	0.08	0.62	0.80	0.45	0.40	0.73
LightGBM	0.86	0.54	0.31	0.43	0.87	0.50	0.30	0.50
XGBoost	0.89	0.56	0.19	0.61	0.87	0.55	0.23	0.64

0: 100만 미만, 1: 100만 이상 300만 미만, 2: 300만 이상 500만 미만, 3: 500만 이상

	F1-score			Precision		
	0	1	2	0	1	2
SVM	0.89	0.64	0.49	0.85	0.65	0.79
Logistic Regression	0.88	0.61	0.58	0.89	0.60	0.63
Naïve Bayes	0.83	0.54	0.54	0.87	0.50	0.57
Grandient Boost	0.88	0.59	0.45	0.83	0.62	0.68
Random Forest	0.87	0.61	0.54	0.83	0.63	0.81
LightGBM	0.88	0.60	0.45	0.90	0.56	0.49
XGBoost	0.89	0.64	0.57	0.87	0.64	0.63

0: 100만 미만, 1: 100만 이상 500만 미만, 2: 500만 이상

Test error for final model

	Mean_Ensemble
RMSE	0.7898
MAE	0.5811

	XGBoost
F1-score	0.96
Precision	0.96

Prediction rate

	Mean_Ensemble regression			XGBoost for binary classification (500만)		
	$log\hat{Y}$	$logY$	$log\hat{Y}/logY$	$\hat{Y}$	Y	예측 성공 여부
#살아있다	14.3069	13.9936	1.02	1	1	성공
광해, 왕이 된 남자	15.2790	16.3271	0.94	1	0	실패
배테랑	15.8727	16.4119	0.97	0	0	성공
국가부도의날	14.7711	15.1387	0.98	1	1	성공
데드풀2	15.5702	15.1465	1.03	0	1	실패
악질경찰	13.5332	12.4770	1.08	1	1	성공

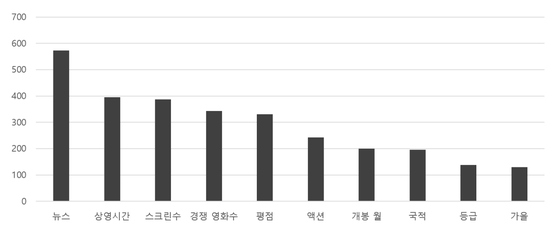
1 : 누적관객수 500만 미만, 0 : 누적관객수 500만 이상

Likelihood ratio test

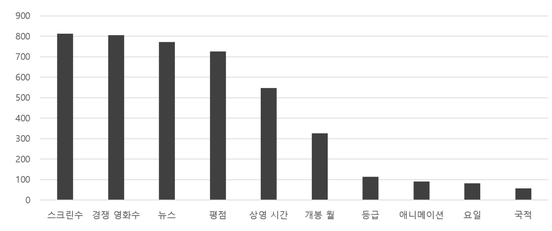
	deviance	df
Full model	509.15	597
Reduced model	760.36	598

Variable Importance (VIMP)

<XGBoost Classifier(500만 기준)>



<LightGBM Regressor>



Conclusion

영화 누적관객수 예측력 최대화와 스크린 상한제 효용성에 대한 분석을 진행하였다. 그 결과, 누적관객수 예측을 최대화하는 모형은 회귀에서는 XGBoost와 LightGBM의 평균을 이용한 모형(Mean_Ensemble) 이었으며, 분류에서는 누적관객수를 500만 기준으로 나누었을 때 XGBoost 분류기가 가장 좋은 예측력을 보여주었다. 스크린 상한제의 효용성에 대한 분석에서는 선형회귀모형을 이용한 우도비 검정과, 회귀분석과 분류분석에서의 변수중요도를 살펴보았을 때, 스크린수 변수가 유의미한 변수이며 중요한 변수라는 결론을 얻을 수 있었다. 따라서 스크린 상한제를 시행하는 것을 생각해볼 필요가 있다고 생각된다.

References

- 김보경(2019). 머신러닝 기법을 활용한 정형·비정형 데이터에 대한 예측 연구와 응용. 중앙대학교 대학원 석사학위논문.
- Jeon, S. H, Son, Y. S(2016). *Prediction of box office using data mining. The Korean Journal of Applied Statistics*, 29(7), 1257-1270.