

REPORT



기계학습 기반 영화 관객 수 예측

퓨처스리그

TEAM PANDA



목 차

I. 프로젝트 목표

II. 데이터 정제 과정

II-1. 데이터 테이블

II-2. EDA

III. 모델링 선정 및 결과

III-1. 모델링 비교

III-1-a. Random Forest

III-1-b. Model Tree

III-1-c. Earth

III-1-b. Lasso

III-2. 모델링 결과

IV. 참고 문헌

IV-1. 서적

IV-2. 인터넷



I. 프로젝트 목표

: 개봉 예정 영화에 대한 흥행 실적 예측 (9/30일 누적 관객수 예측) 다양한 모델을 세운 뒤, RMSE Value를 비교하여 RMSE가 적은 모델을 선택한다.

II. 데이터 정제 과정

II-1. 데이터 테이블

- 제공 받은 자료는 2014년 8월부터 2016년 6월 사이에 개봉한 영화 데이터
- 데이터 관리를 위해 2014년 8월 1일 ~ 2016년 6월30일 주 별 데이터(그림1)를 한 파일로 묶음

 15.07.03-07.09.xlsx Microsoft Excel 97-2003 워크시트 1.14MB	 15.07.10-07.16.xlsx Microsoft Excel 97-2003 워크시트 1.09MB	 15.07.17-07.23.xlsx Microsoft Excel 97-2003 워크시트 976KB
 15.07.24-07.30.xlsx Microsoft Excel 97-2003 워크시트 970KB	 15.07.31-08.06.xlsx Microsoft Excel 97-2003 워크시트 938KB	 15.08.07-08.13.xlsx Microsoft Excel 97-2003 워크시트 1.06MB
 15.08.14-08.20.xlsx Microsoft Excel 97-2003 워크시트 0.99MB	 15.08.21-08.27.xlsx Microsoft Excel 97-2003 워크시트 1.14MB	 15.08.28-09.03.xlsx Microsoft Excel 97-2003 워크시트 1.12MB
 15.09.04-09.10.xlsx Microsoft Excel 97-2003 워크시트 1.14MB	 15.09.11-09.17.xlsx Microsoft Excel 97-2003 워크시트 1.29MB	 15.09.18-09.24.xlsx Microsoft Excel 97-2003 워크시트 1.68MB
 15.09.25-10.01.xlsx Microsoft Excel 97-2003 워크시트 1.11MB	 15.10.02-10.08.xlsx Microsoft Excel 97-2003 워크시트 1.07MB	 15.10.09-10.15.xlsx Microsoft Excel 97-2003 워크시트 1.13MB
 15.10.16-10.22.xlsx Microsoft Excel 97-2003 워크시트 1.39MB	 15.10.23-10.29.xlsx Microsoft Excel 97-2003 워크시트 1.44MB	 15.10.30-11.05.xlsx Microsoft Excel 97-2003 워크시트 1.57MB
 15.11.06-11.12.xlsx Microsoft Excel 97-2003 워크시트 1.33MB	 15.11.13-11.19.xlsx Microsoft Excel 97-2003 워크시트 1.36MB	 15.11.20-11.26.xlsx Microsoft Excel 97-2003 워크시트 1.39MB
 15.11.27-12.03.xlsx Microsoft Excel 97-2003 워크시트 1.86MB	 15.12.04-12.10.xlsx Microsoft Excel 97-2003 워크시트 1.45MB	 15.12.11-12.17.xlsx Microsoft Excel 97-2003 워크시트 1.36MB
 15.12.18-12.24.xlsx Microsoft Excel 97-2003 워크시트 1.15MB	 15.12.19-12.31.xlsx Microsoft Excel 97-2003 워크시트 1.13MB	 16.01.01-01.07.xlsx Microsoft Excel 97-2003 워크시트 1.07MB
 16.01.08-01.14.xlsx Microsoft Excel 97-2003 워크시트 1.07MB	 16.01.15-01.21.xlsx Microsoft Excel 97-2003 워크시트 1.14MB	 16.01.22-01.28.xlsx Microsoft Excel 97-2003 워크시트 1.21MB

[그림1] 초기 데이터 파일

[표1] 초기 데이터 테이블 변동사항

순위	예측 할 영화의 순위를 알 수 없으므로 분석 데이터에서는 제외
영화명	영화명이 같으면 말미에 알파벳을 부여하여 구분 (if, 영화명도 같고 감독이 같은 것은 개봉일을 보고 구분)
개봉일	년도, 월, 일로 셀 분할
매출액	예측 할 영화의 매출액은 알 수 없으므로 분석 데이터에서는 제외
매출액 점유율	예측 할 영화의 매출액 점유율은 알 수 없으므로 분석 데이터에서는 제외
매출액 증감 (전일대비)	예측 할 영화의 매출액 증감은 알 수 없으므로 분석 데이터에서는 제외
매출액 증감률 (전일대비)	예측 할 영화의 매출액 증감율은 알 수 없으므로 분석 데이터에서는 제외
누적매출액	예측 할 영화의 누적매출액은 알 수 없으므로 분석 데이터에서는 제외
관객수	예측 할 영화의 누적관객수만 사용하므로 분석 데이터에서는 제외
관객수 증감 (전일대비)	증감은 관객 수 데이터로부터 도출되는 정보이므로 변수로는 부적절 하다는 판단, 분석 데이터에서는 제외
관객수 증감률 (전일대비)	증감은 관객 수 데이터로부터 도출되는 정보이므로 변수로는 부적절 하다는 판단, 분석 데이터에서는 제외
누적 관객수	변동 없음.
스크린 수	변동 없음.
상영횟수	변동 없음.
대표국적	한국, 미국, 그 외 (기타)로 함
국적	대표국적을 변수로 사용하므로 제외
제작사	제작사의 변수로서의 중요도가 낮다고 판단, 제외
배급사	인수-합병한 대형회사들은 현재 회사명으로 통일 (ex. CJ, 디즈니, 워너 브라더스, 폭스...)
등급	현 등급 분류 기준에 맞춰 GA, 12세, 15세, 19세(청소년관람불가)로 구분
장르	장르별 관객수 합에 장르별 영화 총 개수로 나눠 모든 장르 중에서 차지하는 점유율 계산, 장르가 여러 개인 경우 처음 1개만 분석 데이터만 사용
감독	감독이 여러 명인 경우 처음 1명만 분석 데이터에 사용
배우	배우가 여러 명인 경우 처음 3명만 분석 데이터에 사용

○ 등급, 감독, 개봉일자, 국적의 누락된 정보는 포털 사이트 및 영화진흥위원회의 정보를 참조함

[표2] 추가된 데이터

상영일	기존데이터에 기재된 상영일을 셀에 추가해서 입력 단, 개봉일보다 상영일이 빠른 경우 시사회로 간주하여 상영일에 포함시키지 않음
공휴일 여부	시사회를 제외한 상영시작일부터 해당 데이터의 상영일 사이에 포함된 공휴일 개수
공휴일 횟수 17일차, 24일차	영화의 총 상영일이 17이나 24에 미치지 못하는 경우 마지막 상영일 기준으로 함

스크린당 상영 횟수	(상영횟수)/(스크린 수)
배급사 파워	각 배급사가 배급한 영화가 상영된 기간 동안의 (총 스크린 수)/ (총 상영일수)를 계산한 값을 합하여 각 배급사가 차지하는 비율을 나타냄
장르 파워	(장르의 총 관객수 / 장르의 영화 개수) 값을 구한 뒤 장르가 전체 중 차지하는 비율*10000
감독파워	감독파워는 10년 전인 2006년 이후에 개봉한 영화들의 누적 관객수 합/상영 편수(감독일 경우만)로 계산하였다. *2004년에 통신 전산망 서비스가 시행 되었기 때문에 서비스 시행 전 데이터는 없거나 부정확하다. 따라서, 10년전을 기준으로 감독 영향력이 유효하다고 보았다.
배우파워	배우파워는 10년 전인 2006년 이후 개봉한 영화들의 누적 관객수 합/출연 편수(주연, 조연의 경우만)로 계산하였다. *2004년에 통신 전산망 서비스가 시행 되었기 때문에 서비스 시행 전 데이터는 없거나 부정확하다. 따라서, 10년전을 기준으로 감독 영향력이 유효하다고 보았다.
0일차 관객수	첫날 누적 관객수 - 첫날 관객수
개봉 전 평점	네이버 영화에서 검색해서 넣음
개봉 전 참여 자수	네이버 영화의 개봉 전 평점에 참가한 참여자 수

- “개봉 전 평점”과 “개봉 전 참여자수”는 국내에서 가장 많은 고객을 보유하고 있는 포털 사이트 네이버의 영화 서비스 참고
- 그 외 지운 데이터: 당일 관객수가 -3명 이었던 007스펙터의 마지막 데이터는 삭제

표 3. 사용된 데이터

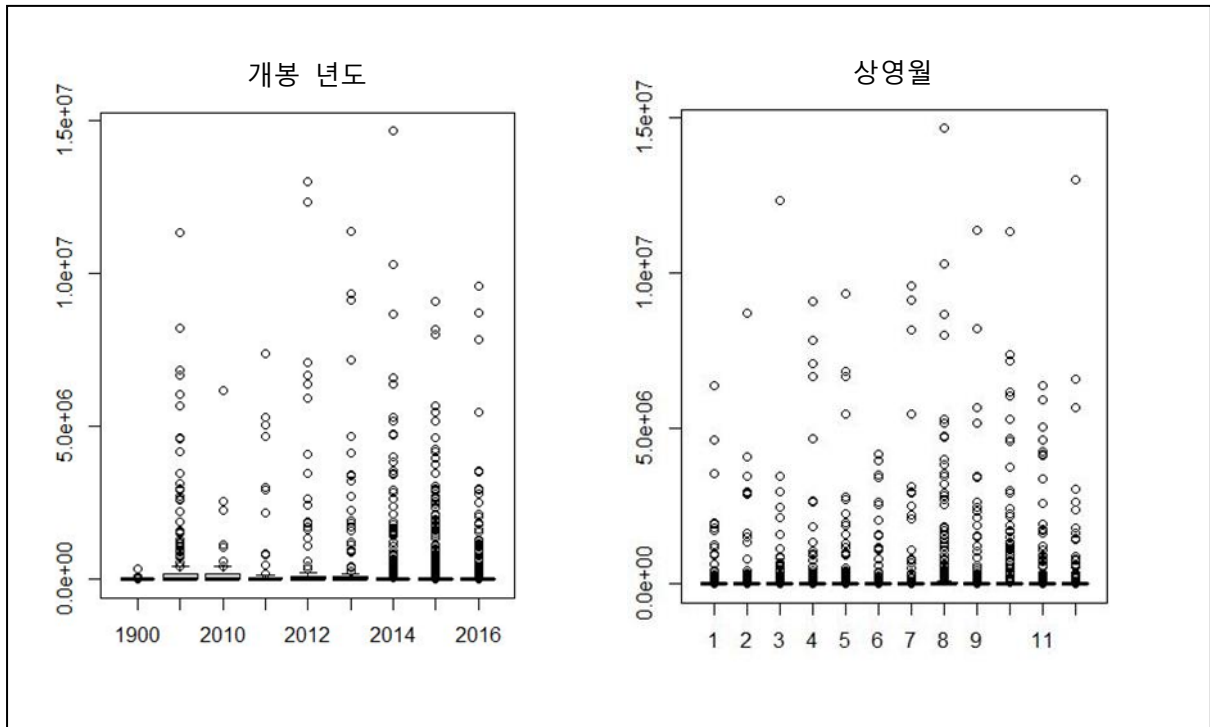
개봉연도 (year)	2000년 이전에 개봉한 영화는 개봉연도를 1900으로 넣었고 2000년부터 2009년에 개봉한 영화는 2000으로 넣고 나머지는 개봉연도를 그대로 사용함. 2006년을 기준으로 변수들을 나눠 보았으나 2000년부터 2009년에 개봉한 영화의 개수가 전체 영화 개수에서 적은 비중을 차지하고 있어서 범주 별 데이터 셋을 잘 나누거나 모으지 못하기 때문에 값을 2000으로 넣음
상영 월 (month)	상영 월
등급 (age)	전체 이용자는 0, 12세 이용자는 12, 15세 이용자는 15, 청소년 관람불가는 19로 구분
대표국적 (country)	한국은 '1', 미국은 '2', 기타는 '3'으로 치환
공휴일횟수 17일차 (holiday17)	상영 시작 후, 상영 기간 17일까지 공휴일 수 이거나 상영 기간이 17일이 되지 않는 경우 상영 종료일 사이에 포함된 공휴일 수
공휴일횟수 24일차 (holiday24)	상영 시작 후, 상영 기간 24일까지 공휴일 수 이거나 상영 기간이 24일이 되지 않는 경우 상영 종료일 사이에 포함된 공휴일 수
스크린 수 (screen)	영화 상영 첫날의 스크린 수
배급사파워 (company)	배급사의 영향력을 나타냄

장르파워 (genre)	장르의 영향력을 나타냄
감독파워 (director)	감독의 영향력을 나타냄
배우파워1 (star1)	배우1의 영향력을 나타냄
배우파워2 (star2)	배우2의 영향력을 나타냄
배우파워3 (star3)	배우3의 영향력을 나타냄
총 배우파워 (totalstar)	배우파워1, 2, 3의 합
개봉 전 평점 (pregrade)	네이버 영화에서 집계한 개봉 전 평점
개봉 전 평점 참여 자 (preparti)	네이버 영화에서 집계한 개봉 전 평점에 참여한 사람의 수
0일차관객수 (day0a)	개봉일 전의 관객수. 단, 재개봉이면 재상영일 이전까지 누적 관객수를 의미
누적 관객수 17일차 (audience17)	상영일수가 17일 일 때의 누적 관객수 관객수를 예측해야 할 영화의 개봉일로부터 9월 30일까지의 기간과 같음
누적 관객수 24일차 (audience24)	상영일수가 24일 일 때의 누적 관객수 관객수를 예측해야 할 영화의 개봉일로부터 9월 30일까지의 기간과 같음
0일차스크린수 (day0s)	개봉 일 전의 수
개봉전상영횟수 (day0r)	개봉일 전의 상영일 수

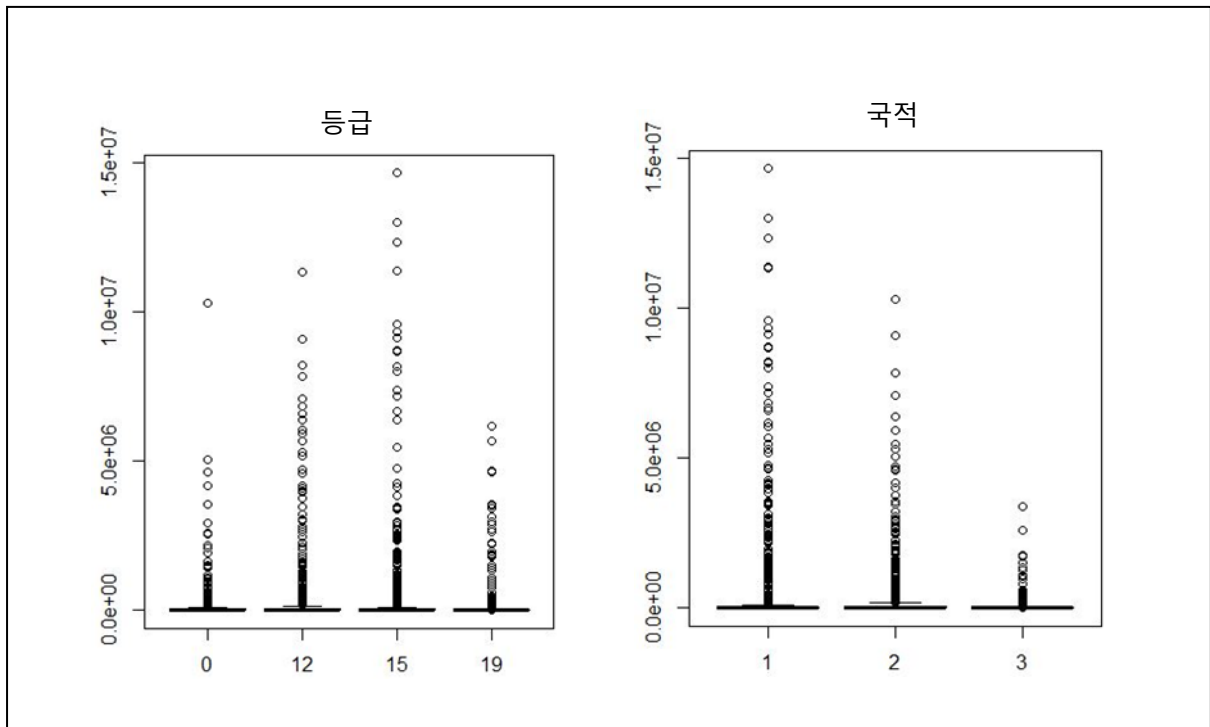
○ 영화명을 기준으로 중복되는 데이터 삭제

II-2. EDA

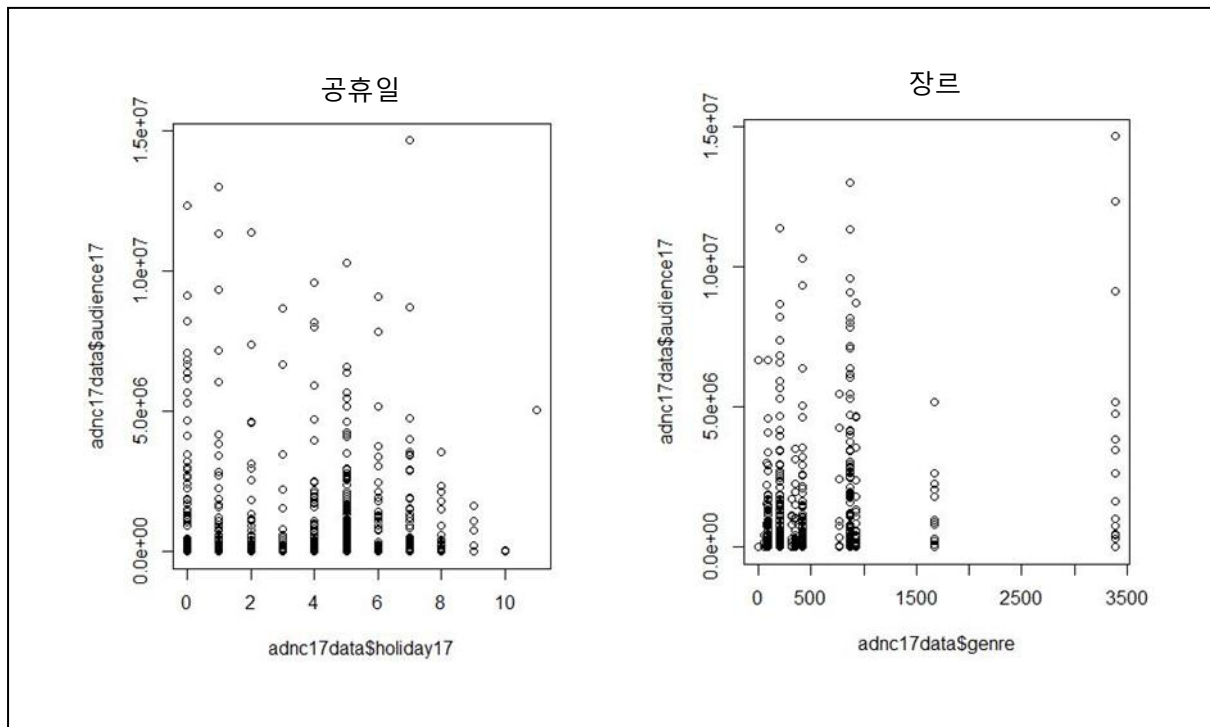
: 모델을 선택하기에 앞서 변수들의 box plot에 산점도를 나타내 보았다.



[그림2] 개봉 년도, 상영월의 box plot



[그림3] 등급, 국적의 box plot

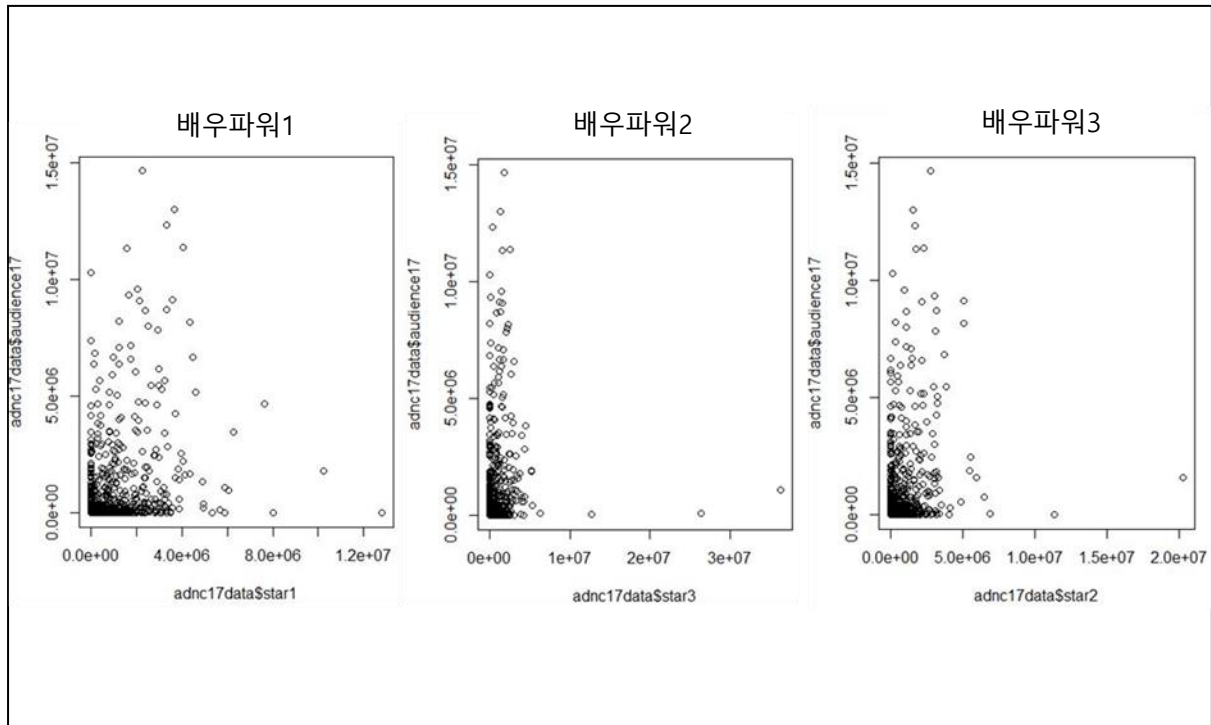


[그림4] 공휴일횟수와 장르의 산점도

- 공휴일 횟수가 10일이 넘는 경우, 영화의 개수가 현저히 작기 때문에 영화 한 두개가 미치는 영향이 지배적. 따라서, 전체 데이터의 신뢰성이 떨어지기 때문에 공휴일 횟수가 10 이상인 데이터는 제외(그림5)

	holiday17	sum(audience17)	count(audience17)	avg(audience17)
1	0	148374671	944	157176.6
2	1	89311554	338	264235.4
3	2	50701788	279	181726.8
4	3	26444342	129	204994.9
5	4	80020999	284	281764.1
6	5	180665208	550	328482.2
7	6	58166739	222	262012.3
8	7	68624066	115	596731.0
9	8	16981368	79	214954.0
10	9	3936946	16	246059.1
11	10	67253	2	33626.5
12	11	5064060	1	5064060.0

[그림4] 개봉일 17일차 기준, 공휴일 횟수에 따른 데이터

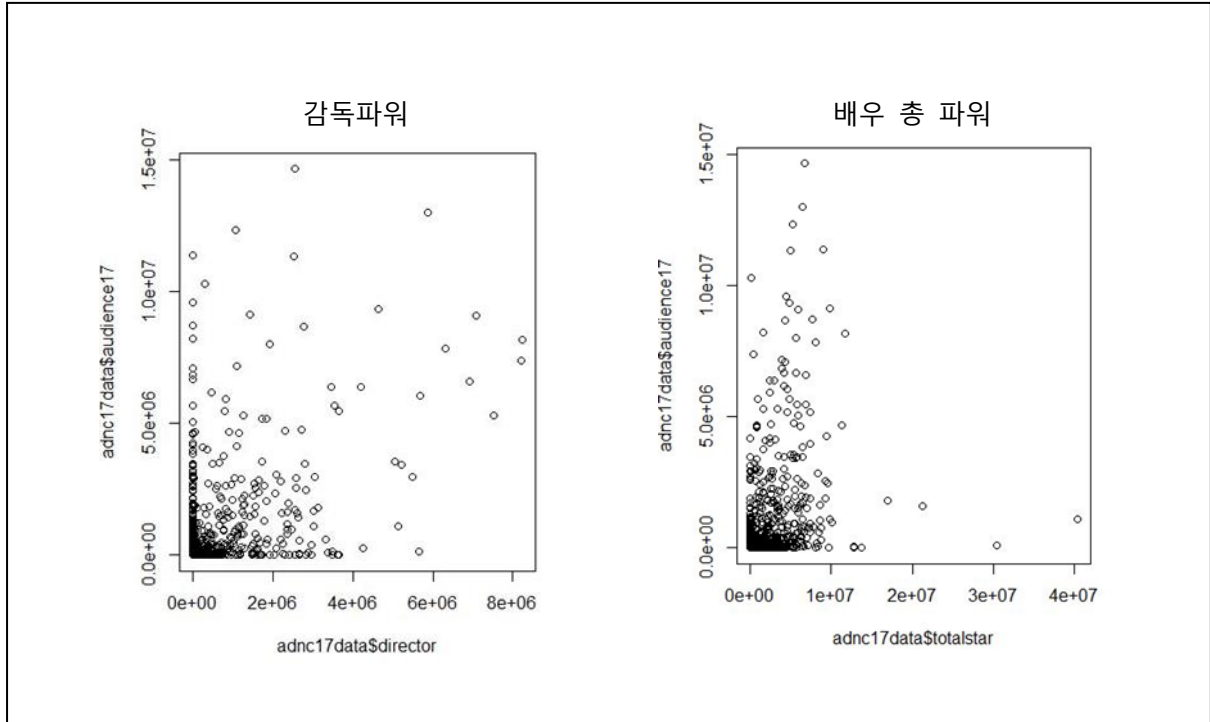


[그림5] 배우파워 1,2,3의 산점도

- 배우파워가 1에서 3으로 갈수록 상관성이 떨어지는 것으로 보임(그림7 참조)
- 배우파워를 모두 합한 총 배우파워는 상관성이 0.4를 넘는 양의 상관관계인 것으로 보임.

```
> cor(adnc17data$star1, adnc17data$audience17)
[1] 0.4232671
> cor(adnc17data$star2, adnc17data$audience17)
[1] 0.3716228
> cor(adnc17data$star3, adnc17data$audience17)
[1] 0.2046198
> cor(adnc17data$totalstar, adnc17data$audience17)
[1] 0.4349051
> |
```

[그림6] 배우파워 1,2,3 과 배우파워 합의 상관성



[그림7] 감독파워와 총 배우파워의 산점도

- 감독파워는 0을 제외하고는 선형성을 보이는 듯 하는데 그 이유로는 감독파워 산정 시 기간 설정에 의한 것으로 보임.

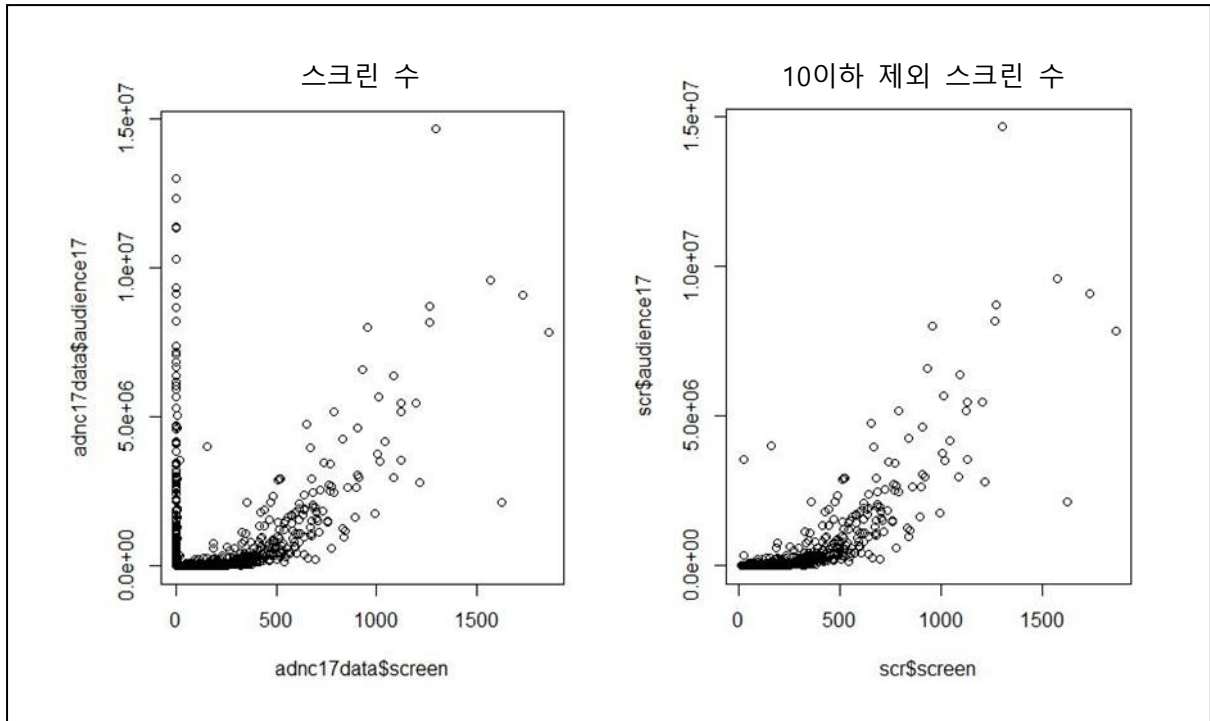
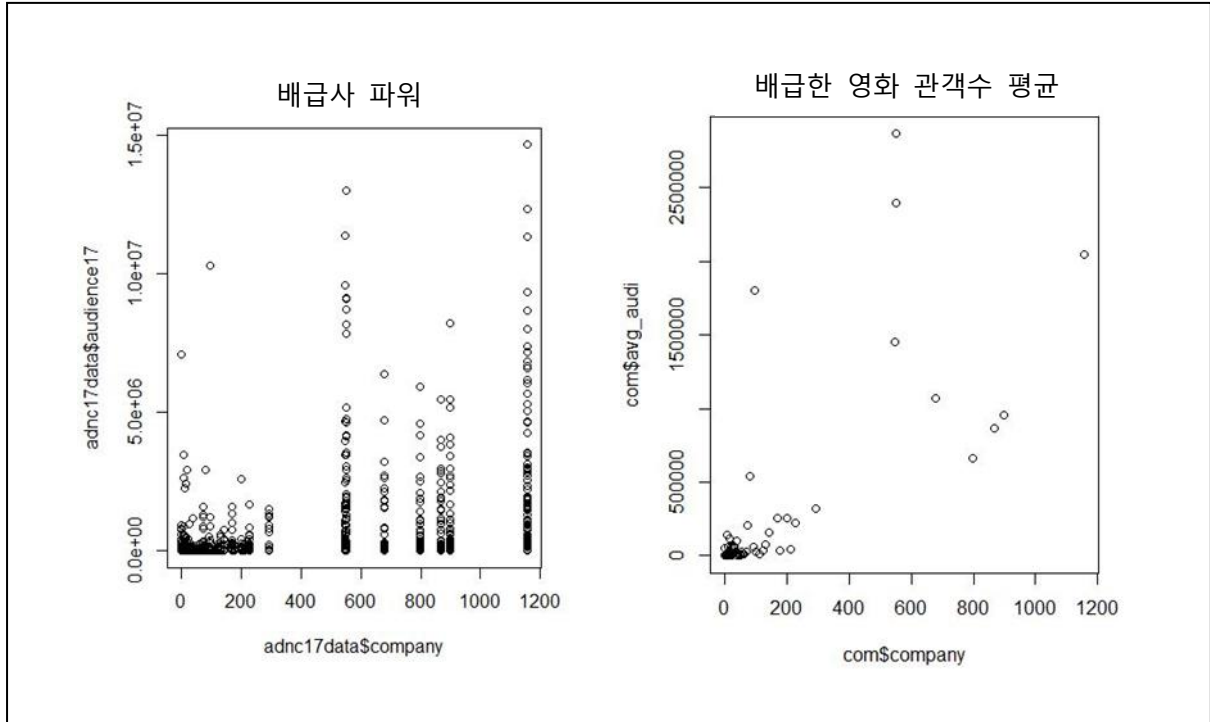
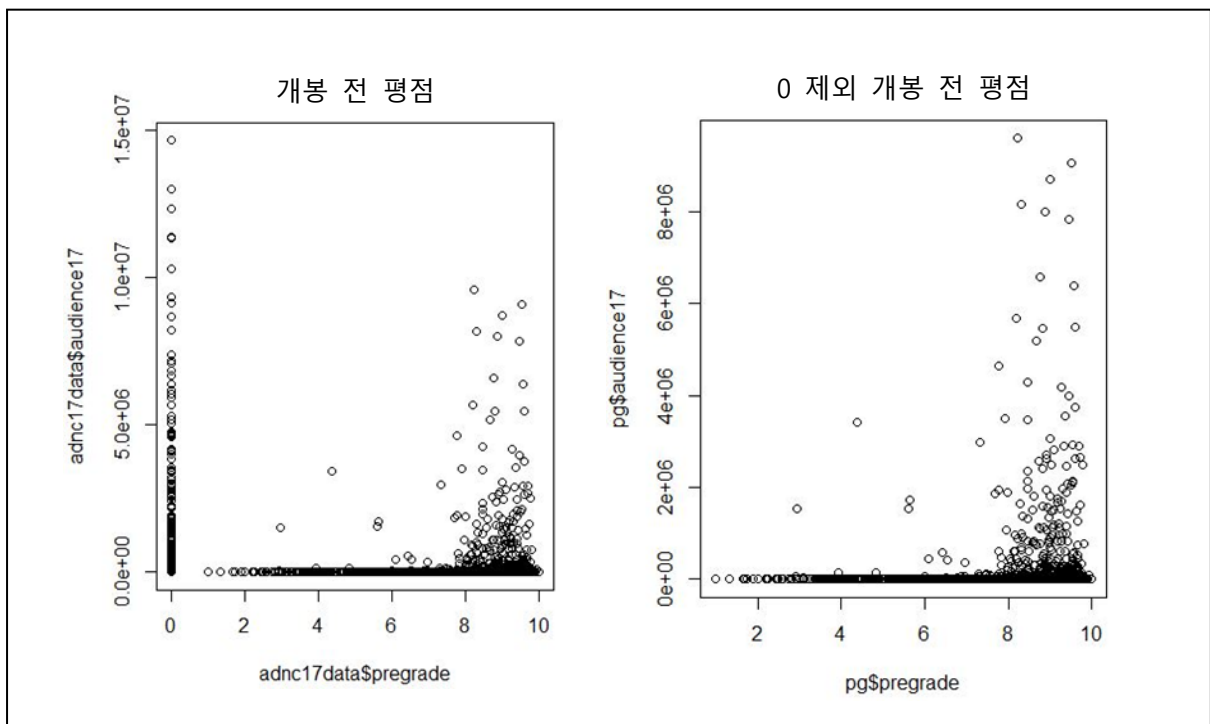


그림 8. 영화 별 개봉 첫날 스크린 수와 수정한 개봉 첫날 스크린 수의 산점도

- 개봉당일 스크린 수가 10 미만인 경우(대체로 재개봉영화)를 제외하면 선형성을 보이는 듯 함



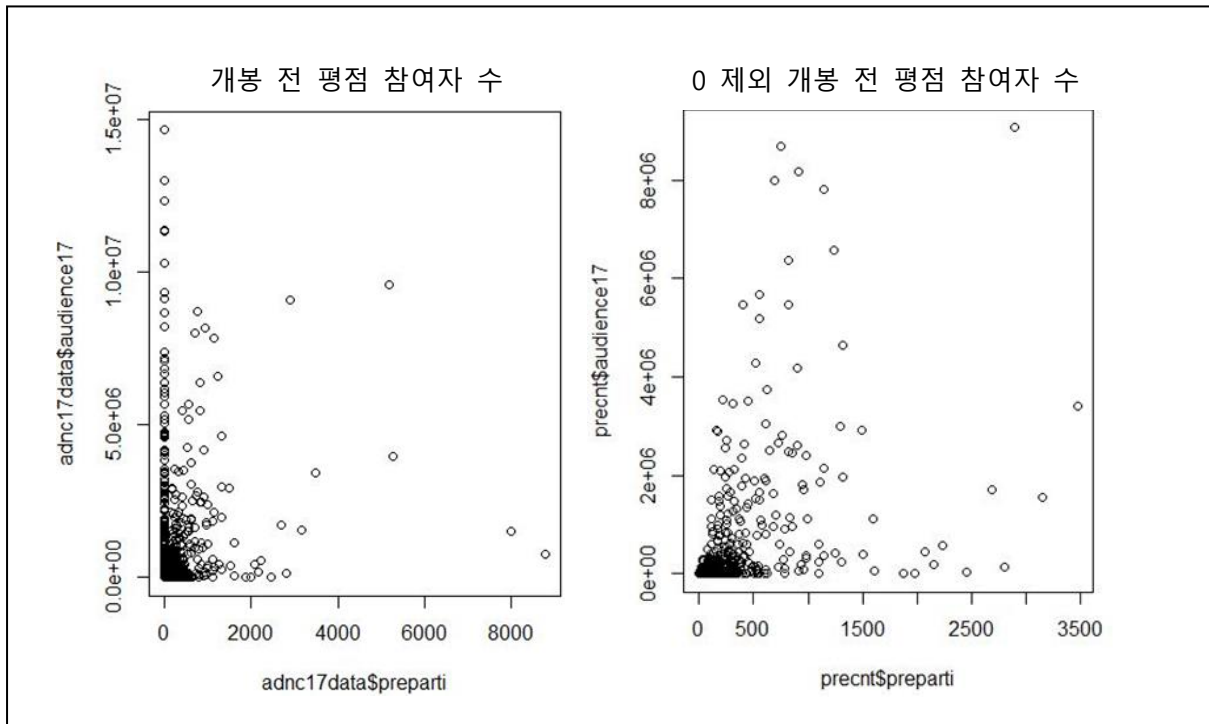
[그림9] 배급사 파워와 배급한 영화 관객수 평균의 산점도



[그림10] 개봉 전 평점과 수정한 개봉 전 평점의 산점도

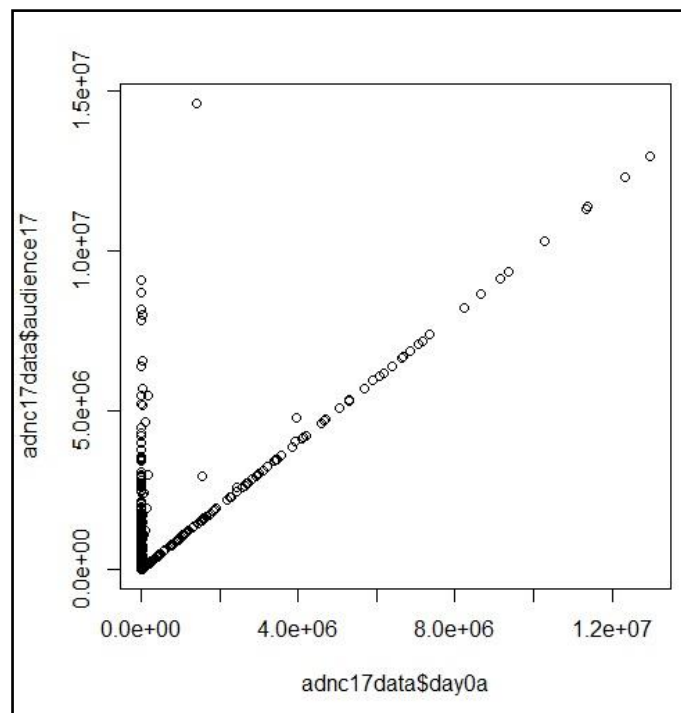
- "개봉 전 평점"의 값이 0인 경우는 평점에 참여한 사람이 없는 경우에 해당
- 네이버 영화 평점 서비스가 2012년 10월 27일 이후에 "개봉 전"과 "개봉 후"로 개편되었기 때문에 이전에 개봉한 영화의 경우, "개봉 전 평점"과 "개봉 전 평점 참여자 수"의 값을 0으로 줌

- 개봉 전 평점의 값이 0인 경우를 제외하니 선형성을 보이는데 함



[그림11] 개봉 전 평점 참여자 수와 수정한 개봉 전 평점 참여자 수의 산점도

- "개봉 전 평점 참여자 수"의 범위를 0이상 5000미만으로 제한했을 때, 상관성이 향상될 것으로 보임
- 개봉 전 평점 참여자 수가 5000 이상인 경우는 이상치로 보임



[그림12] 0일차 관객수 산점도

- <그림 13>을 보면 0일차 관객수가 0보다 클 경우, 0일차 관객수에 따른 17일차 관객수가 $y=x$ 의 정비례 함수처럼 보인다. 따라서, 재개봉 영화는 재개봉 후에 관객수 증가가 거의 없었다는 것으로 볼 수 있다. 이때, 상관관계는 약 0.8이었다.

따라서, 위 산점도를 보아 데이터가 선형으로 나타나 보인다. 따라서 Lasso regression, Model Tree, Random Forest 또는 Earth를 적용 할 수 있음을 확인하였다. 알맞은 모델을 선정하기 위한 모델링 결과 비교나 코드는 “Ⅲ. 모델링 선정 및 결과”에서 자세히 다루도록 하겠다.

Ⅲ. 모델링 선정 및 결과

Ⅲ-1. 모델링 비교

Ⅲ-1-a. Random Forest

```
> p.rf<-predict(m.rf, adnc17data.te[, -19])
> error_rf <-sqrt(mean((adnc17data.te$audience17)-(p.rf))^2)
> error_rf
[1] 44068.09
```

그림 13. Random Forest의 RMSE

Ⅲ-1-b. Model Tree

```
> p.m5p <-predict(m.m5p, adnc17data.te[, -19])
> error_mt<-sqrt(mean((adnc17data.te$audience17)-(p.m5p))^2)
> error_mt
[1] 5633.891
```

그림 14. Model Tree의 RMSE

Ⅲ-1-c. MARS

```
> error=sqrt(mean(adnc17data.te$audience17-p.be)^2)
> error
[1] 3434.395
```

그림 15. MARS의 RMSE

III-1-b. Lasso

```
> error_lars <-sqrt(mean(pre-adnc17data.te$audience17)^2)
> error_lars
[1] 20768.8
```

그림 16. Lasso의 RMSE

- 모델들의 RMSE를 비교한 결과 MARS, Model Tree, Lasso, Random Forest 순으로 낮았다. 따라서, MARS와 Model Tree의 모델을 쓰는 것이 타당해 보였음.

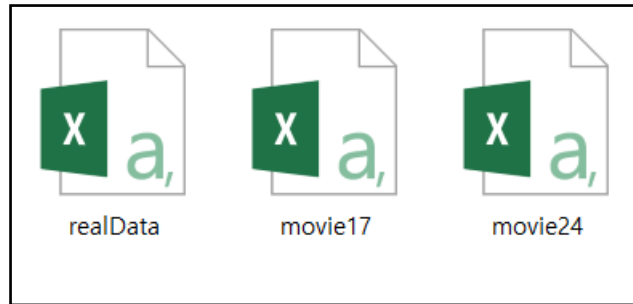
[1]	39187.558	9308.137	91922.241	56210.868	8949.294	8371.646	6882.040	31922.782	6561303.006
[10]	38731.091	1898970.753	35272.133	3276933.850	42207.612	126431.583	96611.546	10887.501	1006875.128
[19]	10976.370	6572.074	922356.633	1006277.861	49051.120	138333.011	449749.914	65746.614	628457.534
[28]	10023.883	543651.108	9957.423	15329.860	345725.208	307725.100	33342.225	57076.576	6506.075
[37]	12502.003	33435.366	145601.856	92917.752	10330.350	51080.619	53734.491	33674.558	8586.430
[46]	9663.774	437386.533	5612.881	8288.199	11855.824	13214.791	9370.785	9481.875	8158.086
[55]	12213.974	45344.913	11383.395	6828.107	36938.280	-201121.352	7859.659	121450.378	2396343.883
[64]	35092.785	12115.239	9944.050	75373.600	-450238.001	33777.934	8348.479	7212.498	2668531.859
[73]	8650.145	8674.469	-223859.261	783001.721	-483693.088	37702.076	239819.229	33535.257	-351970.316
[82]	118985.200	34289.820	5354077.804	14039.960	10635.732	8771.332	308222.821	3010434.790	75872.642
[91]	726936.801	7325.656	674855.843	14337.211	1203276.416	10393.100	35816.122	11550.291	7426.433
[100]	52008.021	14327.080	6951.582	15462.501	6487.487	9378.470	489200.012	59947.876	8003.595
[109]	1850017.609	9208.932	1854626.308	8988.439	1817920.312	10369.206	88534.627	33679.853	-303463.022
[118]	27218.496	117483.846	11044.014	10620.297	141194.160	8570.382	9829.529	6552.951	8798.073
[127]	-339319.975	2071524.900	193374.698	5514.106	94590.952	11138.194	9724.330	14934.992	99124.454
[136]	794018.010	11879.299	2210992.946						

그림 17. Model Tree의 예측 값

- 그렇지만 <그림 18>에서 보이 듯이 Model Tree의 예측 값이 음수가 많이 나오는 것으로 보여 로그변환을 시켰음. 그러자, RMSE가 매우 커져서 Lasso를 사용하기로 하였음.
- MARS와 Lasso를 이용한 실 예측은 “III-2. 모델링 결과”에서 자세히 다루도록 함

III-2. 모델링 결과

: “III-1. 모델링 비교”를 통해 MARS와 Lasso가 적합한 것을 확인



[그림18] 실 예측에 사용하는 데이터

- <그림 19>는 실 예측에 사용하는 데이터 파일. realData는 예측해야 하는 영화의 정보가 담겨있으며 movie17은 영화 정보를 17일 기준으로 담고 있고 movie24는 영화 정보를 24일 기준으로 담고 있음

[표4] realData의 스크린 값의 기준

영화명	스크린(screen)
고산자	9월 7일 개봉 첫 날 스크린 수
매그니피센트7	유니버설 픽처스 배급사의 스크린 수가 많은 영화 5개의 첫 날 스크린 수의 평균
카페 소사이어티	CGV 아트하우스 배급사의 최근 예술영화 3개의 첫 날 스크린 수의 평균

- 17일차 누적 관객수(audience17)와 24일차 누적 관객수(audience24)의 값이 3000보다 큰 경우의 데이터만 이용할 경우 RMSE가 좋아진 것으로 보임
- 따라서, 누적 관객수가 3000보다 작을 경우 실 예측에 이용하는 데이터에서 제외함

결과(답안)

[고산자 24일차 예상관객수]

```
> final_output2
[1] 2608574
```

[매그니피센트7, 카페 소사이어티 17일차 예상관객수]

```
> p.be<-predict(fit.be, adnc17data.re[-19])
> p.be
[1] 1288692.38 63037.44
```

IV. 참고 문헌

IV-1. 서적

- 1) 브레트 란츠, 〈R을 활용한 기계 학습〉, 에이콘출판
- 2) 장병희, 〈영화 흥행 요인〉, 커뮤니케이션북스

IV-2. 인터넷

- 1) http://pop.heraldcorp.com/view.php?ud=201607251648296472169_1
(H헤럴드POP, 부산행 흥행진단)
- 2) http://topepo.github.io/caret/train-models-by-tag.html#Ensemble_Model
(Regression Model about ensemble)
- 3) <http://rpubs.com/cardiomoon>
(RPubs, brought to you RStudio)