졸 업 논 문

제 목 : 다중 회귀분석과 기계 학습을 통한 한국 영화 관람객 수 예측

지 도 교 수 : 이 태 욱 교수님

2 0 1 8 년 11 월 28 일

한 국 외 국 어 대 학 교 자 연 과 학 대 학 통 계 학 과

학 번 : 201302255 성 명 : 이규민

학 번 : 201302986 성 명 : 장형욱

Predicting the number of movie audience, Using Multiple Regression and Machine Learning

by

kyumin Lee

hyungook jang

Under the direction

of

Professor taewook Lee

2018. 11. 27

A thesis submitted and approved by the committee of the Department of Statistics of Hankuk University of Foreign Studies in partial fulfillment of the requirements for the degree of Bachelor

Thesis Committee : _____

DEPARTMENT OF STATISTICS

< 목 차 >

1. 서론

2014년 이후 국내영화산업 매출액은 2조원을 넘었고, 국내 영화의 수상사례가 늘어나면서 한국영화의 진흥을 보여주고 있다. 스토리 측면에서 우수한 위력을 보여주던 한국영화는 연출력이 발전함에 따라 높은 작품성을 보여주게 되었고, 이에 따라 한국영화의 높은 수요를 볼 수 있다. 완성도 높은 작품이 다수 상영됨에 따라 관객들의 잣대가 좀 더 엄격해지고 있다. 이제 관객들은 웬만한 작품에 감흥하지 못하고, 않좋은 품평을 받은 작품은 흥행에 실패한다.

본 연구의 목적은 영화의 관객 수를 개봉이전에 예측하는 것에 있다. 본 논문은 회귀분석과 기계학습을 이용한 통계적인 방법을 통해 관람객 수를 예측하고자 한다. 유의한 회귀분석 모델과 기계학습 결과를 통해 투자자들이 영화의 흥행여부를 판단 후 과도한 투자를 막고, 영화 제작자로 하여금 어떠한 변수가 추가되면 흥행이 될지 도움을 줄 것 이다. 비록 영화에는 산업적인 측면 이외에 예술적 측면이 가미되어 있기에, 완벽한 예측은 불가능하지만, 산업적 측면에서 유의미한 예측을 함으로서 관계자들의 투자결정에 도움이 되길 바라는 것이 목적이다.

2. 본론

2.1데이터 설명

본 논문의 데이터는 한국영화진흥위원회 'KOFIC' 의 '영화데이터 500순위' 이다. 1993년 (투깝스)부터 2018년 2분기 (신과함께-인과연) 까지 국내영화를 관객 수에 따라 500순위까지 지정했다. 매출액, 개봉날짜, 상영 스크린 수 의 정보를 기본적으로 얻을 수 있었고, 이외의 변수는 네이버영화에서 직접 옮겼다. 종합된 영화 데이터는 다음과 같다.

	Audience	Sales	Year	Screen	History	Action	Drama	Horror
1	17613682	136000000000	14	1587	1	1	0	0
2	14410931	116000000000	17	1912	0	1	1	0
3	14257115	111000000000	14	966	1	0	1	0
4	13414009	105000000000	15	1064	0	1	1	0
5	13019740	66716119300	6	167	0	1	0	0
	Disaster	Comic	Erotic	Crime	Actor power		Thousand	Factory
1	0	0	0	0	46		2	1
2	0	0	0	0	35		9	1
3	0	0	0	0	16		5	1
4	0	0	0	1	43		8	1
5	1	0	0	0	22		0	1

<표1 : 상위 5개 데이터의 요약된 형태> 출처: KOFIC

-변수 설명

1.종속변수(Audience): 본 연구의 종속변수는 KOFIC에서 제공해준 총관객수이다. 시계열 데이터가 아니고, 표본간의 연관성이 없기에, 자기상관의 문제는 없을 것으로 판단된다.

2.독립변수: 관객 수에 영향을 미칠 것이라 생각되는 변수를 선정하였다. 선정과정에서 분석결과 과도하게 유의하지 않은 변수는 삭제하고 다시 분석해보는 과정으로 선별하였다. 선별된 변수로는 '개봉연도', '상영관 수', '장르', '출연 배우 5년간 작품수', '1000만 배우or감독 출연여부', '3대배급사 배급여부'이다. '장르'변수는 각각의 장르를 범주형 변수로 하였고, 하나의 작품에 여러 개의 장르가 있음으로 여러 장르의 변수가 중복 될 수 있다.

① Year : 해당 작품의 개봉연도를 2자리 수로 나타냈다. 시간이 흐름에 따라 국내 영화의 수요가 점진적으로 증가하거나 감소할거라 예상하였다.

② Screen : 해당 작품의 상영관 수가 관객수에 직접적인 영향을 미칠 것이라 예상했다.

③ History : 해당 작품의 장르에 '사극'이 포함되면 1, 아니면 0으로 표시하였다.

④ Action : 해당 작품의 장르에 '액션'이 포함되면 1, 아니면 0으로 표시하였다.

⑤ Drama : 해당 작품의 장르에 '드라마'가 포함되면 1, 아니면 0으로 표시하였다.

⑥ Horror : 해당 작품의 장르에 '공포'가 포함되면 1, 아니면 0으로 표시하였다.

⑦ Disaster : 해당 작품의 장르에 '재난'이 포함되면 1, 아니면 0으로 표시하였다.

⑧ Comic : 해당 작품의 장르에 '코믹'이 포함되면 1, 아니면 0으로 표시하였다.

⑨ Crime : 해당 작품의 장르에 '범죄'가 포함되면 1, 아니면 0으로 표시하였다.

⑩ Erotic : 해당 작품의 장르에 '성인'이 포함되면 1, 아니면 0으로 표시하였다.

⑪ Actorpower : 주요 출연 배우 3명의 과거 5년간 작품수를 합산하였다. 배우가 과거에 많은 작품에 출연한 것은 영화계에서 해당배우에 대한 선호가 있다는 것이고, 또한 많은 작품에 출연한 배우는 관객들에게 친근하게 다가갈 수 있다고 생각된다.

⑫ Thousand : 천만배우, 천만감독의 출연여부를 기록하였다. 천만배우나 감독의 작품 표현력과 노하우는 영화의 흥행에 큰 영향을 미칠 것으로 기대된다. 이때 배우나 감독의 경우 같은 천만이더라도 어떤 사람은 몇 개의 천만 작품에 영향을 미친 반면, 어떤 배우나 감독은 단 한 개의 작품에만 영향을 미쳤을 것이다. 이런 측면을 드러내고 싶어 천만배우가 몇 명이고, 한 천만배우가 과거 몇 개의 천만작품에 출연했는지 합산하였다.

⑬ Factory : 국내 영화산업에서 가장 큰 비중을 차지하는 3개의 배급사 (CJ, 롯데, 쇼박스) 중 한 회사가 영화 배급을 맡았는지 아닌지에 따라 1또는 0으로 표기하였다.

2.2 분석방법

본 논문은 통계패키지인 R을 이용하여 분석이 이루어졌다. 데이터 전처리 작업과 회귀분석, 기계학습 모두 R을 이용하여 이루어졌다. 먼저 변수의 성질을 알기위해 양적변수는 변수간의 상관관계를 보여주었다. 질적변수는 0이거나 1일 때 반응변수인 Audience의 차이를 히스토그램으로 표현함으로서 대략적으로 변수의 유의성을 알 수 있게 만들었다. R에서 제공해주는 lm 함수를 이용하여 회귀분석 하였고, 데이터를 '2010년 이후의 데이터', '2005년 이후의 데이터', '모든 데이터'를 이용하였을 때의 결과를 비교하였다. 다중공선성의 유무를 알아보기 위해 VIF를 비교하였고, 회귀분석 결과 유의미하지 않은 변수들은 제거하였다. 각 모형에 18년도 4분기 10개 데이터에 대해 회귀 식을 이용하여 예측해 보았다. 마지막으로, 머신러닝의 렌덤 포레스트 기법을 활용하여 18년도 4분기 9개 데이터에 대한 예측을 실행하여 보았다.

3. 변수에 대한 해석

3-1. 양적변수간의 상관관계

본 논문에서 다룰 본격적인 회귀분석 전에 양적변수(Year, Screen, Actor power, Thousand) 사이의 상관관계에 대해 알아보았다.

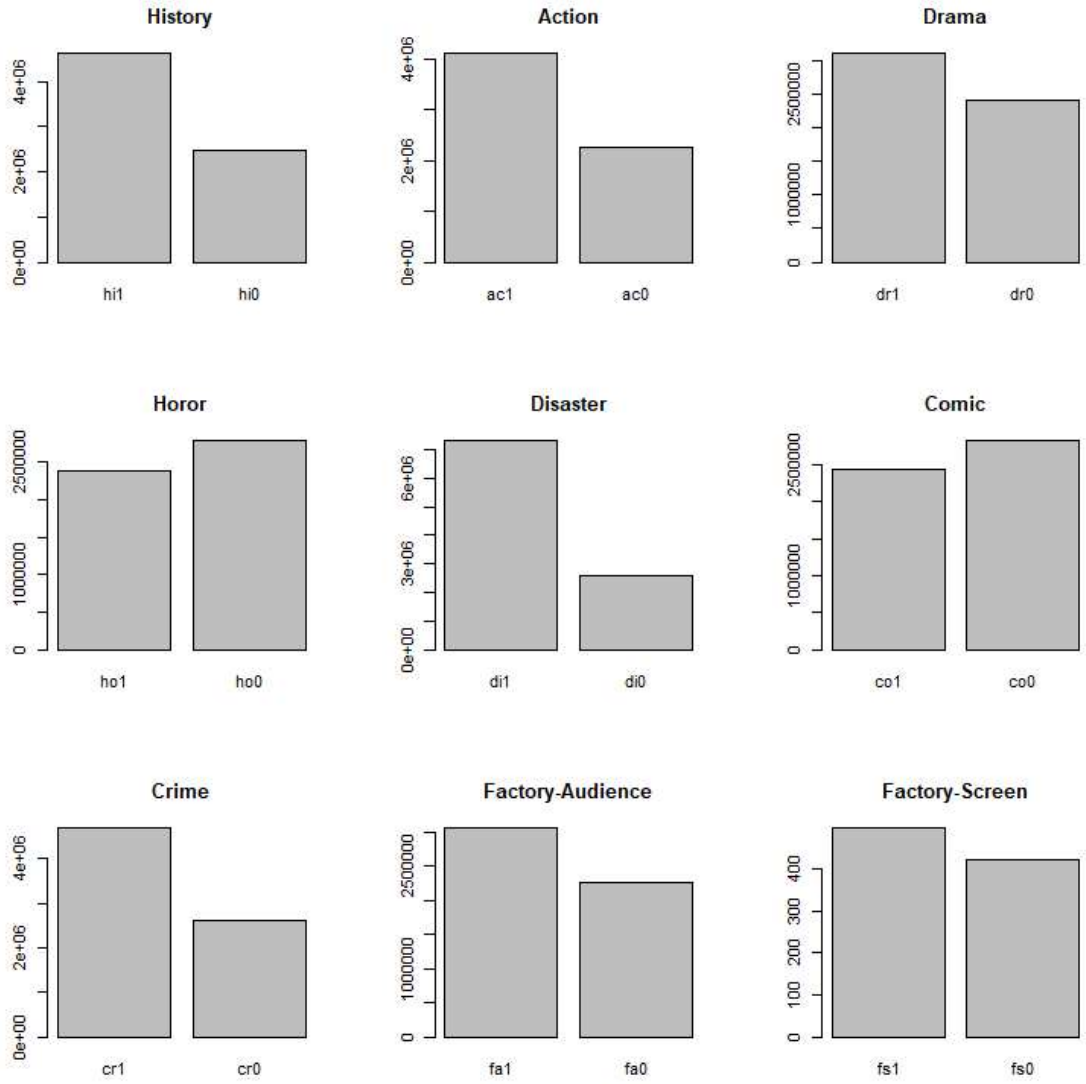
	Year	Screen	Actorpower	Thousand
Year	1	0.85377837	0.01780342	00.3871747
Screen	0.85377837	1	0.09972819	0.5497567
Actorpower	0.01780342	0.09972819	1	0.1250046
Thousand	0.3871747	0.5497567	0.1250046	1

<표2 : 상관관계 도표>

<표2> 로 보았을 때 큰 상관관계를 보이는 변수는 Year-Screen 사이밖에 보이지 않는다. 이는 한국의 영화관이 해가 거듭할수록 증가하고 있다는 해석이 가능하다. 또한 두 변수 사이의 상관관계가 존재하기 때문에 회귀분석 시 다중공선성에 대한 분석을 유의해 봐야함을 의미한다.

3-1. 질적 설명변수에 대한 해석

본격적인 회귀분석 진행에 앞서 질적 변수들이 데이터에 얼마나 영향을 미치는지 알아볼 필요가 있다.



위의 그래프는 X축에 장르와 3대배급사 여부를 나태내고, Y축에 관객 수를 나타낸 간단한 막대그래프이다. 이때 X축 변수의 1은 해당 장르를 포함하는 것이고 0은 포함하지 않음을 의미한다. 이 표를 통해 우리는 몇 가지 사실을 도출해 낼 수 있다.

* 장르부문에서 공포영화와 코미디 영화는 많은 관객을 이끌어 내기에 좋지 않다.
* 재난요소를 포함하는 영화가 아닌 영화에 비해 관객 수에 많은 영향을 끼친다.
* 3대 배급사인 영화의 경우 아닌 영화에 비해 약 1/4더 많은 스크린을 차지할 수 있다.

4. 회귀분석모형

본 논문에서는 너무 과거의 데이터는 앞으로 개봉할 영화예측에는 타당하지 않을 것이라는 의문을 품고 3가지 모형(All data, Over 2005, Over 2010)을 만들어 본 후에 어떠한 모델이 가장 좋은 모델일지 검정해보는 순서이다. 이때 각 모델에서 유의하지 않은 변수(Erotic)는 제거한 후 유의한 변수만을 이용하여 더욱 좋은 모델을 만들어 보려 시도하였다. 이때 종속변수는 Audience인 관람객 수 이다.

4-1. 모든 데이터를 이용한 모델

Variable	Estimate	Std. Error	t value	Pr(>[t])
Intercept	-566603.7	404141.1	-1.402	0.161554
Year	-228581.1	31622.7	-7.228	1.90e-12 ***
Screen	3678.5	436.2	8.433	3.84e-16 ***
History	2407092.7	263462.7	9.136	< 2e-16 ***
Action	2314112.1	233536.8	9.909	< 2e-16 ***
Drama	2389354.6	217764.4	10.972	< 2e-16 ***
Horror	2019955.6	308867.0	6.540	1.56e-10 ***
Disaster	3578527.2	473755.1	7.554	2.11e-13 ***
Comic	1826528.5	233073.6	7.837	2.93e-14 ***
Crime	1057063.7	328422.2	3.219	0.001374 **
Actorpower	25265.7	9713.1	2.601	0.009573 **
Thousand	214607.8	60262.6	3.561	0.000405 ***
Factory	325625.7	158175.4	2.059	0.040060 *
Residual standard error: 1694000 on 487 degrees of freedom				
Multiple R-squared: 0.5902, Adjusted R-squared: 0.5801				
F-statistic: 58.45 on 12 and 487 DF, p-value: < 2.2e-16				

<표3 : 회귀분석 결과 (All data) >

* <표3>은 역대 모든 한국 영화 데이터를 이용해서 다중회귀분석을 실시한 결과 이다. 통계적으로 유의한 모든 변수들을 이용하였고, 이를 통해 추정된 회귀식은 다음과 같다.

Audience=-580052.5 - 227303.3Year + 3679.7Screen + 2405245.0History + 2313426.6Action + 2385120.5Drama + 1976499.3Horror + 3585302.2Disaster + 1813742.8Comic + 981308.8Crime + 25756.4Actorpower + 216153.9Thousand + 307547.6Factory

이때 회귀모형에서 독립변수 사이의 영향력을 확인하기 위해 VIF (분산팽창인자)를 이용하여 공선성에 의해 커지는 추정량의 분산의 정도의 크기를 알아보려 한다. VIF가 10이상이면 보통 다중공선성에 심각한 문제가 있는 것으로 판별한다.

	Year	Screen	History	Action	Drama	Horror
VIF	4.305375	5.575208	1.349280	1.854495	2.058197	1.574431
	Disaster	Comic	Crime	Actorpower	Thousand	Factory
VIF	1.063823	1.547430	1.092438	1.070676	1.531712	1.040361

<표4 : 독립변수의 VIF>

<표4>를 통해 어떠한 독립변수도 VIF>10을 만족하지 않으므로 다중공선성의 문제는 존재하지 않는다.

4-2. 2005년 이후 데이터를 이용한 모델

Variable	Estimate	Std. Error	t value	Pr(>[t])
Intercept	977495.4	562238.4	1.739	0.082955 .
Year	-368523.8	45220.4	-8.150	5.94e-15 ***
Screen	4856.8	505.9	9.600	< 2e-16 ***
History	1994784.7	308528.0	6.465	3.25e-10 ***
Action	1840588.6	280037.2	6.573	1.71e-10 ***
Drama	2202683.3	255833.8	8.610	2.24e-16 ***
Horror	1645656.6	365437.0	365437.0	9.02e-06 ***
Disaster	3385053.7	482747.7	482747.7	1.15e-11 ***
Comic	1943666.7	271160.7	271160.7	4.27e-12 ***
Crime	1010341.6	358690.1	358690.1	0.005116 **
Actorpower	24610.0	11245.7	11245.7	0.029275 *
Thousand	220535.2	61362.9	61362.9	0.000371 ***
Residual standard error: 1715000 on 364 degrees of freedom				
Multiple R-squared: 0.6297, Adjusted R-squared: 0.6186				
F-statistic: 56.28 on 11 and 364 DF, p-value: < 2.2e-16				

<표5 : 회귀분석 결과 (2005년 이후 데이터) >

<표5>는 2005년 이후의 데이터만을 이용하여 관람객 수를 예측한 다중회귀직선이다. 유의하지 않은 Erotic변수와 Factory변수를 제거한 후 추정한 다중 회귀 직선 식은 다음과 같다.

Audience=960200.4 - 368971.9Year + 4861.3Screen + 1997030.1History + 1826374.0Action + 2193360.7Drama + 1642712.4Horror + 3392074.7Disaster + 1928623.9Comic + 1016527.2Crime + 25454.7Actorpower + 220962.7Thousand

	Year	Screen	History	Action	Drama	Horror
VIF	3.527040	4.872182	1.450835	2.041964	2.091777	1.623697
	Disaster	Comic	Crime	Actorpower	Thousand	
VIF	1.068458	1.311653	1.096752	1.070979	1.467460	

<표6 : 2005년 이후 회귀 식 독립변수의 VIF>

표6은 2005년 이후 다중 회귀직선 식의 VIF를 나타낸 표이다. 어떠한 독립변수도 VIF가 10이 넘지 않으므로 다중공선성의 문제는 없다.

4-3. 2010년 이후 데이터를 이용한 모델

Variable	Estimate	Std. Error	t value	Pr(>[t])
Intercept	2936732.5	987303.0	2.974	0.003255 **
Year	-521518.5	70171.4	-7.432	2.21e-12 ***
Screen	5108.1	604.3	8.454	3.57e-15 ***
History	2272846.4	400645.0	5.673	4.29e-08 ***
Action	1863608.6	386302.0	4.824	2.59e-06 ***
Drama	2299455.9	341497.8	6.733	1.36e-10 ***
Horror	1974918.9	473925.5	4.167	4.40e-05 ***
Disaster	1706040.4	576284.5	2.960	0.003401 **
Comic	1889523.1	371185.5	5.091	7.54e-07 ***
Crime	1008170.9	438652.4	2.298	0.022461 *
Actorpower	26932.5	14711.6	1.831	0.068468 .
Thousand	247160.9	67632.9	3.654	0.000321 ***
Residual standard error: 1801000 on 225 degrees of freedom				
Multiple R-squared: 0.6737, Adjusted R-squared: 0.6577				
F-statistic: 42.22 on 11 and 225 DF, p-value: < 2.2e-16				

<표7 : 회귀분석 결과 (2010년 이후 데이터) >

<표7>은 2010년 이후의 데이터만을 이용하여 관람객 수를 예측한 다중회귀직선이다. 유의하지 않은 Erotic변수와 Factory변수를 제거한 후 추정한 다중 회귀 직선 식은 다음과 같다.

Audience=2963769.3 -523496.1Year + 5133.5Screen + 2298562.8History + 1861830.1Action + 2284030.0Drama + 1981836.9Horor + 1706455.0Disaster + 1927674.7Comic + 1017732.5Crime + 26038.6Actorpower + 247258.6Thousand

	Year	Screen	History	Action	Drama	Horror
VIF	1.795950	3.008569	1.511093	2.288199	2.130180	1.656824
	Disaster	Comic	Crime	Actorpower	Thousand	
VIF	1.074109	1.175870	1.086436	1.114654	1.531712	

<표8 : 2010년 이후 회귀 식 독립변수의 VIF>

표8은 20010년 이후 다중 회귀직선 식의 VIF를 나타낸 표이다. 어떠한 독립변수도 VIF가 10이 넘지 않으므로 다중공선성의 문제는 없다.

5. 세 가지 회귀직선 모델의 비교

모든 데이터, 2005년 이후의 데이터, 2010년 이후의 데이터를 이용한 세 가지 회귀직선 모델 중 어떤 회귀직선 모델이 가장 예측력이 좋은지를 알아보고자 한다. 이때 오차의 정도를 나타내는 MSE , 오차의 비율을 나타내는 MAPE 통계량을 이용하여 알아보겠다. 회귀직선모델이 2018년 08월까지의 데이터를 이용하여 추정되었으므로 2018년 09월부터 개봉한 영화들의 시험용 데이터는 다음과 같다.

안시성	관객 5,436,779 연도18 스크린 수1538 사극/액션 배우파워3+3+19=25 천만 0+0+1 배급사X
협상	관객 1,743,437 연도18 스크린 수918 범죄 배급사CJ 배우파워6+4+16=26 천만3
명당	관객 1,968,493 연도 18 스크린 수1,114 역사 배급사X 배우파워5+4+13=22 천만1
독전	관객 5,063,620 연도18 스크린 수1532 범죄, 액션 배우파워 20+10+16=46 천만3 배급사x
원더풀 고스트	관객449,906 연도18 스크린 수351 범죄 코미디 배우파워21+5+6=32 천만3 배급사X

암수살인	관객3,739,085 연도18 스크린 수1177 범죄 드라마 배우파워9+9+9=27 천만3 배급O
창권	관객1,596,687 연도18 스크린 수1351 액션 배우파워25 천만0 배급사x
미쓰백	관객721,204 연도18 스크린 수638 드라마 배우파워7+0+20 천만x 배급사x
완벽한 타인	관객4,427,723 연도18 스크린 수1313 코미디 배우파워28+22+2=52 천만5 배급사o

<표9: test data>

<표9>의 시험용 데이터들을 각각 회귀직선 모델에 대입한 후 오차에 대한 검정은 다음과 같다.

5-1. Using MSE

test 데이터를 대입 후 오차정도를 보기위해 평균제곱오차인 MSE를 사용하였다. 이때 관객 수 단위가 너무 커서 MSE가 크므로 단위를 만 단위로 끊어서 구해주었으며, $MSE = \frac{1}{n}(A_{관측값} - A_{실제값})^2$ 를 이용하여 구한 MSE는 다음과 같다.

	Total model	Over 2005 model	Over 2010 model
MSE	86.8133374289663	94.0183433638824	88.0574307546564

<표10 : 각 모델의 MSE>

<표10>을 통해 MSE 가 가장 작은 Total model이 가장 우수한 것을 알 수 있다.

5-1. Using MAPE

또 다른 통계량으로 MAPE(Mean Absolute Percentage Error)를 이용하였다. 이때 영화 데이터는 이상치가 많은 것을 감안해 Mean이 아닌 Median을 이용한 MAPE를 구해보았다.

$MAPE = median(\frac{A_{관측값} - A_{실제값}}{A_{실제값}})$ 을 이용한 MAPE는 다음과 같다.

	Total model	Over 2005 model	Over 2010 model
MAPE	0.246639253382829	0.267367811960613	0.224111801438497

<표11 : 각 모델의 MAPE>

<표 11>을 통해 MAPE가 가장 작은 2010년 이후 모델이 가장 우수한 것을 알 수 있으며, 오차가 약 22%임을 알 수 있다.

위의 결과들을 통해 두 통계량에서 모두 채택되지 못한 2005년 모델은 가장 성능이 낮아 보인다. MSE 통계량의 특성을 생각해 보았을 때, 실제 관객 수가 큰 데이터가 오차도 크게 나올 가능성이 높다. 따라서 영화의 흥행정도를 알아보고 싶을 때에는

제작예정 영화의 목표가 아주 높다면 Total model을 사용한 회귀직선 식을 사용하는 것이 좋아 보인다. 반면 MAPE는 모든 데이터가 비율로써 계산되므로 관객 수에 대한 오차의 가중치는 존재하지 않는다. 따라서 목표가 아주 높지 않은 영화의 관객 수 예측을 위해서는 MAPE가 가장 작은 2010년 모델을 이용하는 것이 합리적이다.

6. 기계학습을 통한 예측

본 논문은 관객 수를 예측할 수 있도록 도움을 주기 위함이다. 따라서 다양한 방법을 이용하고자 회귀분석 이외에도 머신러닝을 이용한 예측을 시도해 보았다. 랜덤포레스트 방법을 이용한 500개 데이터의 기계학습 후 <표9>의 Test data를 시험했다. 이때 각 영화를 3등급으로 나누어 관객 수를 상, 중, 하 (1, 2, 3) 로 범주화 시켜주었다.

pridicted observed	1.17309679619897	1.39525708279977	1.39560693099976	1.44252196507196	1.48765887349366
1	1	1	0	1	1
2	0	0	1	0	0
3	0	0	0	0	0
observed	1.81525567642836	1.85943701298701	2.09178113830221	2.75485058970936	
1	0	0	0	0	
2	1	1	0	0	
3	0	0	1	1	

<표12 : 기계학습을 통한 관람객 수 예측>

<표12>는 데이터를 통한 예측 값과 실제 값을 나타내준 표이다. 첫 번째 데이터의 경우 독립변수들을 통한 관람객 수 의 예측 값은 약 1.173 이고, 실제 값은 1 (상) 임을 의미한다. 따라서 이 결과는 잘 맞춘 것이라고 볼 수 있다. 각 예측 값들을 반올림하여 실제 값들과 비교하면 9개의 Test data 중 7개를 맞추어 약 77%의 적중률을 보인다.

7. 결론

영화산업 분야는 상업성과 예술성을 함께 가지고 있다. 예술적인 특성으로 인해 영화 데이터는 숫자만으로는 모든 것을 나타내기는 힘든 사실이다. 실제로 소비자 입장에서 영화 예고편 광고를 보았을 때, 자신들도 모르게 끌리는 면이 있다. 그것은 배우, 감독, 장르, 스토리가 아닌 영화특유의 분위기, 재밌을 것이라는 예감 때문이다. 그러기에 본 연구는 이러한 예술적 측면은 제외하고, 상업적인 측면에서만 영화를 분석하였다. 영화를 소개하는 TV프로그램에서 영화스토리 이외에 강조하는 부분은 감독과 배우의 이력, 영화의 장르, 등급, 원작여부, 실화여부 이다. 하지만 등급과 원작, 실화 유무는 회귀분석 결과 유의미하지 않기에 제외시켰다. 본 연구가 영화의 모든 측면을 고려했다고 말할 수 없지만 가장 대표적인 영화의 특징들을 변수로 대입하였기에 유의미한 결과를 도출 할 수 있었다.

질적 변수들을 히스토그램으로 비교해 보았을 때 (본 논문 3-1 참고) 재난영화의 여부가 관객 수에 큰 영향을 미친다는 것을 알 수 있다. 이는 재난 영화의 경우 대부분 스케일이 크게 제작되기 때문으로 보인다. 또, 범죄영화는 현실에 밀접한 부분이고, 이를 바탕으로 탄탄한 스토리를 이끌어 낼 수 있으며, 배우들의 연기력이 돋보인다. 또한 범죄영화의 대부분은 액션으로 이루어져 많은 볼거리를 제공하기에 선호된다. 사극은 국민들에게 친숙한 스토리로 구성되어있고, 상위 작품들은 대부분 대한민국의 과거 외세에 의해 침략 받은 역사 (명량, 암살, 밀정 등)나 정부의 부패 (왕의남자, 택시운전사), 분단의 아픔 (국제시장, 태극기 휘날리며)을 다뤘다. 이는 범국민적으로 공감할 수 있고, 때로는 통쾌함을 줌으로서 스토리상의 카타르시스를 제공한다. 때문에 사극이라는 장르는 취향과 별로 연관이 없고, 대부분 상위 순위를 차지한다. 이에 반대로 공포나 코믹과 같은 요소는 취향에 따라 호불호가 명백히 나뉘기에 평균적으로 높은 성적을 달성하지 못하는 것으로 보인다. 3대 배급사 유무는 관객 수에 미약하게 영향을 미치는 것으로 보아 적절한 배급전략도 흥행에 무시하지 못할 요소라는 것을 알 수 있다.

데이터를 살펴보았을 때, 최근의 데이터가 많다는 것을 알 수 있다. 1993년 서편제, 투캅스 두개의 영화만 데이터에 있으며, 1994-1998년 데이터는 없고, 1998년에 쉬리를 포함한 3개 영화만 있다. 영화산업의 변화를 예측하는 목적에 오래된 데이터는 맞지 않다는 결론을 내렸고, 데이터를 '2005년 이후', '2010년 이후', '모든 데이터'로 나눠 회귀분석을 진행하였다. 회귀분석 결과 각 모델별로 유의미한 변수가 조금씩 다름을 보였다. 하지만 3가지 모델 모두 다중공선성은 존재하지 않았다.

영화산업의 흥행은 시간의 흐름에 직접적인 영향을 받는다. 대표적인 원인으로 요즈음 에는 Netflix, 헬로TV등 집에서도 충분히 영화를 즐길 수 있어서 영화가 최근으로 올수록 영화 관객 수는 조금씩 줄어든다. 이는 회귀식의 Year변수의 계수가 공통적으로 음의 값을 가짐을 보면 알 수 있다. 장르는 모두 관객 수에 긍정적인 영향을 보였지만, 계수를 비교해 적절히 선택해야 할 것이다. 모든 장르를 가지는 작품은 없고, 많아봐야 3개의 장르가 최대이다. 배우 이력이나 천만배우 유무 또한 관객 수에

긍정적인 영향을 보였다. 이것은 소비자들이 영화 자체의 작품성 보다는 직관적으로 보이는 출현 배우들에 더 큰 반응을 한다는 것을 알 수 있다.

또한 3가지 모델을 비교를 위해 18년 4분기 영화를 직접 예측해 보았다. 9개 영화 데이터를 각 모델에 대입하여 예측된 결과와 실제 결과를 비교하여 MSE와 MAPE를 통해 가장 적절한 모델을 찾았다. 이로써 상위권 성적을 원하는 영화제작에는 Total model을 이용하는 것이, 중위권 성적을 원하는 영화제작에는 Over 2010 model을 이용하는 것이 가장 적합하다.

마지막으로 랜덤포레스트 기법을 이용한 머신러닝을 통해서도 영화 관람객 수를 예측해 보았다. 영화 관람객 수의 범주화를 위해 영화 등급을 관람객 수로 3등급으로 나누어 예측했다. 그 결과 test data의 적중률은 77%를 보였다. 77%라는 수치는 예측이 어려운 영화산업에서 충분히 활용 가능한 모델이라 판단된다. 앞서 말했다시피, 영화는 작품성이 관객을 끌어들이는 중요한 요소이다. 예술성은 객관적으로 평가할 수 없기에, 본 연구는 산업적인 요소만으로 분석을 진행하였다. 본 연구가 이후의 영화산업 관련 연구에 좋은 근간이 되어 본 논문 이후에는 예술성까지 포함하여 영화의 흥행정도를 사전에 미리 예측할 수 있는 모델로 발전하기를 바란다.

<참고문헌>

-영화 투자자를 위한 흥행성과 예측지표 발굴
(Journal of the Korean Data Analysis Society (August 2017) Vol. 19, No. 4 (B), pp. 1963-1975, 김유진, 권오경)

-영화 흥행성과 예측을 위한 온라인 리뷰 마이닝 연구: 개봉 첫 주 온라인 리뷰를 활용하여
(조승연, 김현구, 김범수, 김의웅 - 연세대학교)

-개봉 전 영화의 수요예측모형
(김병도-서울대학교 경영대학 교수, 표태형-서울대학교 경영대학 석사과정)

-Applied regression analysis (Terry E. Dielman)