The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.6.809
30(6), 809–825

A study on entertainment TV show ratings and the number of episodes prediction

Milim Kim[a] · Soyeon Lim[a] · Chohee Jang[a] · Jongwoo Song[a,1]

[a]Department of Statistics, Ewha Womans University

(Received August 14, 2017; Revised September 29, 2017; Accepted October 12, 2017)

Abstract

The number of TV entertainment shows is increasing. Competition among programs in the entertainment market is intensifying since cable channels air many entertainment TV shows. There is now a need for research on program ratings and the number of episodes. This study presents predictive models for entertainment TV show ratings and number of episodes. We use various data mining techniques such as linear regression, logistic regression, LASSO, random forests, gradient boosting, and support vector machine. The analysis results show that the average program ratings before the first broadcast is affected by broadcasting company, average ratings of the previous season, starting year and number of articles. The average program ratings after the first broadcast is influenced by the rating of the first broadcast, broadcasting company and program type. We also found that the predicted average ratings, starting year, type and broadcasting company are important variables in predicting of the number of episodes.

Keywords: Entertainment TV show, ratings, number of episodes, prediction model

1. 서론

오디션, 육아, 버라이어티 등 다양한 예능 프로그램의 수가 증가하고 있다. 특히 종합편성채널이 개국한 이후에 지상파 중심이었던 예능 방송이 종합편성채널까지 크게 확대되었으며, 지상파 예능 방송의 시청률을 넘어서는 종합편성채널 예능 프로그램도 있다. 예능 프로그램의 동향을 보면, 2000년대 초반까지의 예능은 대부분 회차를 정해놓지 않고 방영하여 시청률에 따라 폐지가 결정되었지만, 최근에는 '슈퍼스타K'와 같은 서바이벌 오디션 프로그램과 '꽃보다 할배'와 같은 단기 여행프로그램이 나오면서 시청률에 관계없이 회차가 고정된 방송이 증가하였다. 이러한 변화로 시청률이 높은 프로그램의 경우 회차를 증가시키기보다 다음 시즌을 준비하는 경향을 보인다. 또한 다양한 오디션 프로그램이나 시청자들의 사연으로 구성되는 프로그램 등 시청자들의 참여를 요구하는 프로그램이 대거 등장하기도 하였다.

스마트폰과 SNS의 발달로 방송시청의 공간적인 제약이 없어진 데다 SNS 상의 방송 클립을 통해 예능 프로그램 노출도가 높아지면서 예능 시장이 더욱 활성화되었다. 이에 따라 예능 프로그램의 문화적, 산업적인 영향력이 증가하며 프로그램을 기획하는 제작자뿐만 아니라 광고 투자자들도 프로그램의 시청률

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017R1D1A1B03036078).
[1]Corresponding author: Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: josong@ewha.ac.kr

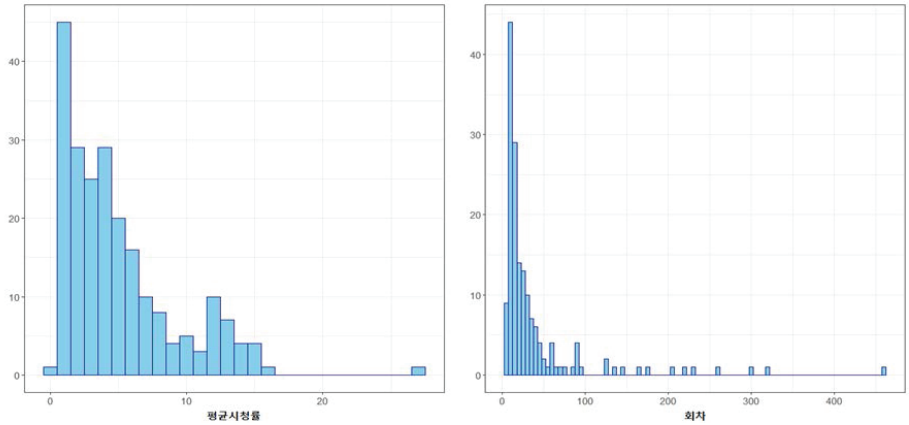


Figure 2.1. Histograms of average ratings and the number of episodes.

과 회차 예측에 대한 연구를 필요로 하고 있다. 드라마 시청률에 대한 선행 연구는 있었으나 (Kang 등 2015), 예능 프로그램을 대상으로 하는 연구는 거의 이루어지지 않고 있으며 그 중에도 예능 프로그램이 끼치는 영향력에 대한 연구 (Lee와 Choi, 2016)가 대부분이며 시청률에 대한 최근 연구는 다중회귀모형에 한정되어있다 (Han, 2016). 본 연구에서는 다중회귀모형 뿐만 아니라 다양한 비선형모형까지 이용하여 시청률을 보다 잘 예측하는 모형을 찾으려 한다.

평균 시청률 분석에 사용한 모형은 다양한 변수 선택법을 (stepwise, ridge, LASSO) (Tibshirani, 1996) 이용한 선형회귀모형, 배깅 (Breiman, 1996), 랜덤 포레스트 (Breiman, 2001), 서포트 벡터 머신 (James 등, 2013), 부분최소제곱 회귀모형(partial least squares; PLS), 주성분회귀 (Hastie 등 2001) 등이다. 모형 비교 시에는 평균제곱오차의 제곱근(root mean squared error; RMSE)을 지표로 사용하였다. 회차 분석에서는 12회 이하, 12회 초과로 범주화하여 분류 모형을 만들었다. 배깅, 랜덤 포레스트, 서포트 벡터 머신, 다항 로그 선형 모형을 사용하였으며 모형 비교 시에는 오분류율을 지표로 사용하였다. 모든 분석은 통계 분석 프로그램 R (R Development Core Team, 2010)을 사용하였다.

2장에서는 자료에 대한 설명을 한 다음, 3장에서는 평균 시청률 예측 모형, 회차 예측 모형을 적합해보고 4장에서는 결론을 내고자 한다.

2. 분석자료 설명

2.1. 자료수집 과정

본 연구의 대상은 2010년 1월 1일부터 2017년 5월 15일 중 방영한 예능 프로그램이다. 단, 파일럿과 같은 프로그램과 3회 이하의 프로그램, 특별편성은 제외하였다. 또한 6개의 방송사(KBS, SBS, MBC, tvN, Mnet, Jtbc)만 조사하였다. 총 271개의 관측치가 있으며, 시청률은 리서치 회사 'AGB 닐슨미디어리서치(www.agbnielsen.co.kr)'에서 조사된 자료를 이용하였고, 회차는 '네이버 검색(www.naver.com)'을 기준으로 하였다. 설명변수들은 네이버와 '다음 영화(http://movie.daum.net/main/new)', '구글(www.google.com)'에서 얻을 수 있었다. 자세한 변수 설명은 2.2절에서 한다.

2.2. 변수 설명

본 연구의 목적은 다양한 설명변수들을 사용하여 예능 프로그램 평균 시청률과 종영회차를 예측하는 것

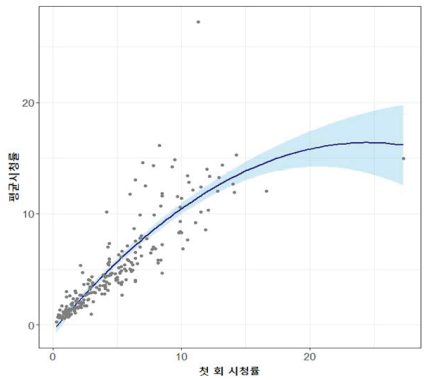


Figure 2.2. Average ratings according to the rating of first broadcasting.

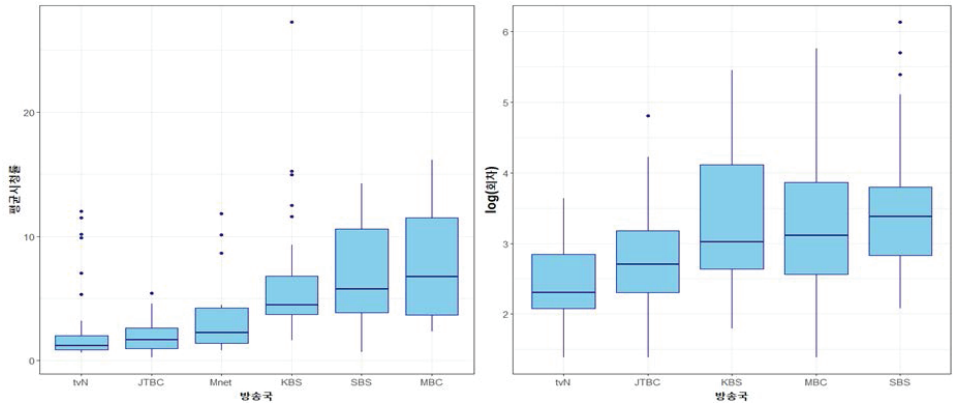


Figure 2.3. Boxplots of average ratings and log(the number of episodes) based on broadcasting company.

이다. 평균 시청률은 해당 프로그램의 전 회차 평균 시청률을 의미하며, 시청률과 회차의 분포는 Figure 2.1과 같이 나타난다. 시청률의 최대값은 27.25%로 2007년 방영된 KBS의 '1박 2일 시즌 1'이며, 최소값은 0.26%로 2012년 방영된 JTBC의 '뷰티업'이었다.

회차에서의 최대값은 529회로 2006년부터 방영중인 MBC의 '무한도전'이며, 최소값은 4회로 2012년 방영된 MBC의 '스타 다이빙 쇼 스플래시'였다. 회차 분석에서는 회차가 정해지지 않은 채로 시작하는 프로그램들만을 사용하였다. 앞으로 분석에서 사용할 설명변수들은 다음과 같으며, 설명변수와 회차의 관계를 살펴볼 때 회차가 오른쪽으로 꼬리가 긴 분포이기 때문에 log를 취하여 살펴보았다 (Figure 2.1).

2.2.1. 초반 시청률 예능에서 시청률과 회차를 예측하기 위해서 매우 중요한 요인 중 하나가 초반 시청률이다. 초반의 시청률에 따라 SNS상에서 언급되거나 입소문이 퍼지는 정도가 달라져 파급력이 달 라지기 때문에 초반 1회차 시청률을 요인으로 고려하였다. 첫 회 시청률과 평균 시청률의 관계를 Figure 2.2에서 살펴보았을 때 확실한 양의 상관관계가 보여 예측에 중요한 요인이 될 거라고 기대된다.

2.2.2. 방송사 과거의 예능 시장은 지상파 방송사들이 주를 이루고 있었지만 최근에는 케이블 채널의 예능도 많은 활약을 하고 있다. Figure 2.3은 방송국에 따른 평균 시청률과 회차를 살펴본 것이다.

Table 2.1. Frequency table of entertainment broadcasting companies

tvN	JTBC	Mnet	KBS	SBS	MBC
49	46	20	53	55	48

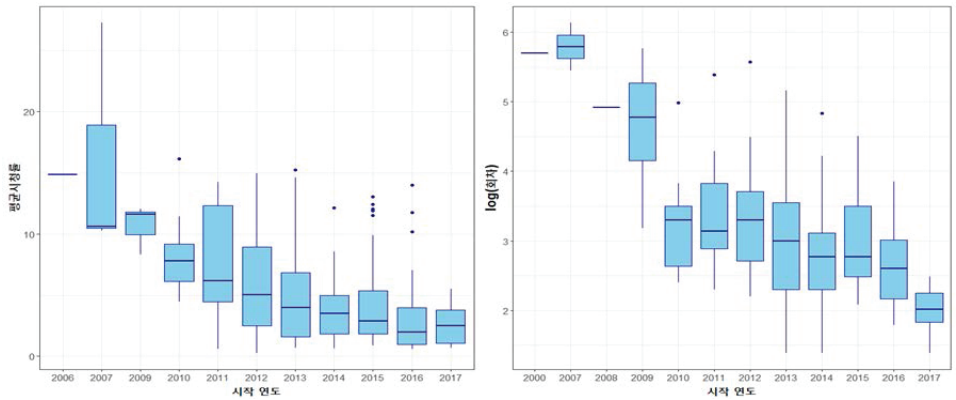


Figure 2.4. Boxplots of average ratings and log(the number of episodes) based on starting year.

지상파 방송사에서 평균 시청률과 회차가 모두 높게 나타나며 케이블 간에도 차이가 나타나는 것으로 보인다. 방송사 요인이 시청률과 회차를 예측하는 중요변수로 사용될 것으로 보인다. Table 2.1을 보면 Mnet은 음악 전문 채널이라 상대적으로 예능 프로그램의 수가 적은데다 시청률이 기록되지 않은 예능 프로그램의 수도 많았기 때문에 데이터 개수가 적다.

2.2.3. 방송 시작 연도 연도에 따른 예능 방송의 유행이 변화했을 수도 있기 때문에 요인으로 고려하였다. Figure 2.4를 보면 시작 연도가 최근일수록 평균 시청률이 감소하는 양상을 보인다. 이는 미디어의 다양화로 인해 TV 이외의 다양한 스마트 기기로 시청이 가능해지면서, TV로 시청하는 경우가 줄어들었기 때문으로 보인다.

2.2.4. 방송 편성시간 TV이용이 많은 시간대에 편성될수록 평균적으로 더 높은 시청률을 보일 것이라고 예상되어 방송 시간과 요일 요인을 설명변수에 포함하였다. 평일 중에서도 금요일이 다음날 휴무와 관련되어 시청률이 높은 것을 반영하기 위해서 '금', '주말 황금시간대(17–20시)', '그 외 시간'으로 범주화하였다. 또한 '금요일 25시'와 같은 방송시간은 실제 토요일에 방송하지만 금요일 저녁의 연장선으로 보는 것이 더 타당하다고 여겨 '금요일'로 분류하였다. Figure 2.5를 보면 주말 황금시간대가 평균 시청률과 회차 모두 높았고, '그 외 시간' 보다는 '금요일'이 평균 시청률과 회차가 더 높은 것을 볼 수 있었다.

2.2.5. 예능 종류 예능 종류에 따라 시청률과 회차가 다른 특징을 보일 것이라고 생각되어 '가족/동물', '경연', '버라이어티', '오디션/서바이벌', '음식', '토크/개그', '기타'의 7개의 범주로 구분하였다 (Table 2.2). 서바이벌 프로그램은 'K팝스타'처럼 프로그램이 끝날 때 1등을 정하는 오디션 종류로, 시청률이 낮아도 대부분 정해놓은 회차는 진행한다는 특징을 가지는 범주이다. 경연은 이와 다르게 매회 1등을 정하는 프로그램으로 시청률이 저조로 인한 조기종영이 가능하다. Figure 2.6을 보면 예능 종류별 평균 시청률의 분포와 log(회차) 분포를 볼 수 있는데, 버라이어티 예능이 평균 시청률도 높고 회차도 긴 편임을 알 수 있다.

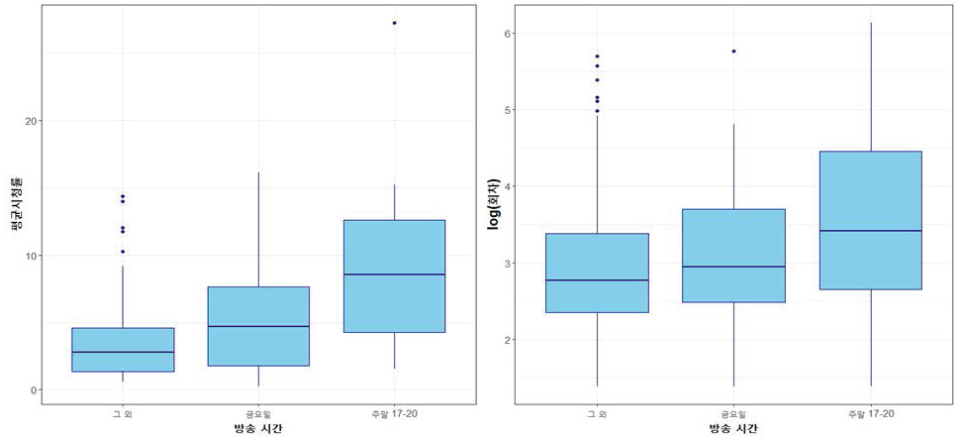


Figure 2.5. Boxplots of average ratings and log(the number of episodes) based on broadcasting time.

Table 2.2. Frequency of entertainment type

가족/동물	경연	버라이어티	오디션/서바이벌	음식	토크/개그	기타
18	26	35	46	14	39	44

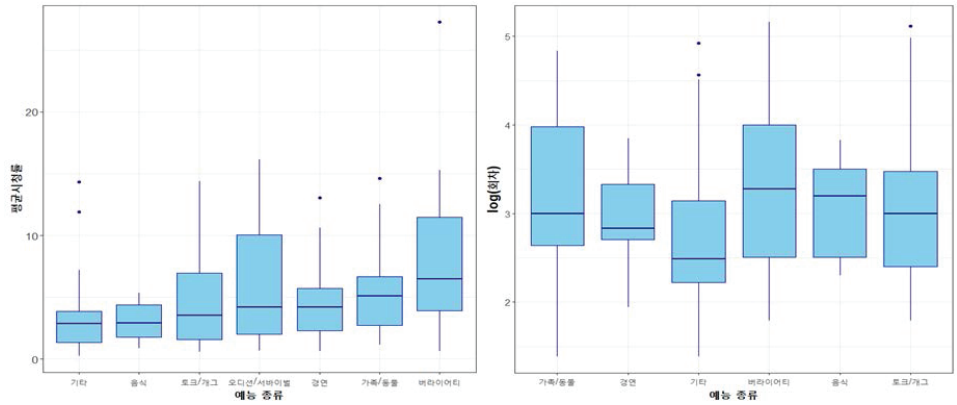


Figure 2.6. Boxplots of average ratings and log(the number of episodes) based on type.

2.2.6. 기사 개수 프로그램이 시작하기 전에 얼마나 많은 관심을 끌고 있는지 가장 알기 쉬운 방법이 기사 개수이다. 따라서 방송 시작 한 달 전부터 첫 방송 후 두 번째 방송 전까지의 기사 개수를 요인으로 고려하였다. 또한, 첫 방송 전에 시청률도 예측하려하기 때문에 article0라는 요인으로 한 달 전부터 첫 방송 전까지의 기사 개수도 변수로 만들었다. 또한 기사 개수가 시작 연도와 양의 상관관계를 보였으므로, 연도의 효과를 줄이고자 원래 값에서 프로그램 시작 해의 평균 기사개수를 빼서 사용하였다.

2.2.7. 시즌, 이전 시즌 시청률 대중들에게 인기를 얻은 예능 프로그램을 시리즈로 만들어 편성하는 경우가 많다. 또한 시즌이 2 이상이면 이전 시즌의 시청률이 프로그램의 시청률과 회차에 많은 영향을 끼치기 때문에 이전 시즌 시청률도 설명변수에 적용하였다. 다만, 해당 프로그램의 시즌이 1이면, 이전 시즌 시청률은 0으로 처리하였다.

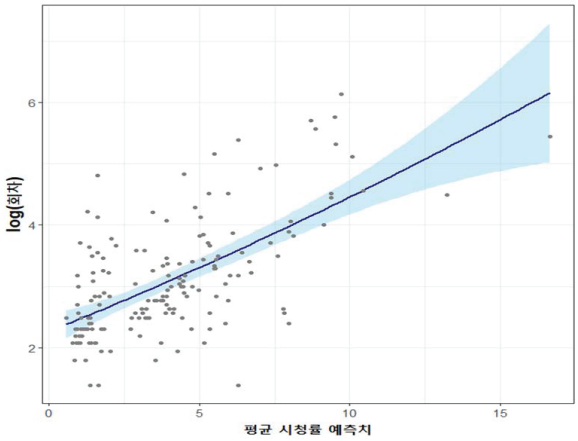


Figure 2.7. log(the number of episodes) vs. predicted values of average ratings.

2.2.8. 관람가 프로그램별로 방송 시청연령이 다르고, 시청연령에 따라 프로그램의 특징이 달라지므로 설명변수로 고려하였다. '전체 관람가 및 7세 이상', '12세 이상', '15세 이상', '19세 이상'으로 범주화하였다. 대부분의 프로그램은 15세 이상(62%)의 예능이었으며, 전체 관람가 및 7세 이상 예능(5%)이 가장 적었다.

2.2.9. 방청, 시청자 참여 여부 시청자로 하여금 방송에 직접 참여하게 함으로써 프로그램에 대한 관심과 흥미를 높일 수 있다. 따라서 방청 여부와 시청자 참여 여부가 반응변수에 영향을 미칠 것이라고 예상하였다.

2.2.10. PD 전작 예능 회차 PD는 프로그램을 기획하고 관리하는 사람으로, 이전 프로그램 회차가 차기작에 영향을 미칠 수 있다. 예를 들어 주로 회차를 정해놓고 시작하는 프로그램을 기획하는 PD는 차기작 역시 단기 프로그램인 경우가 많다. 반면 이전에 장수 프로그램을 기획한 PD의 경우 대중들이 차기작에 대한 기대가 높기 때문에 평균 시청률이 높을 거라 예상된다.

2.2.11. 사회자 유무, 대표자 및 사회자 이전 방송 프로그램 수 예능에서는 사회자가 매우 중요한 역할을 한다. 어떤 사회자가 출연하는지는 그 프로그램의 평균 시청률과 회차는 물론 제작비에도 크게 영향을 미치기 때문이다. 사회자가 있는지 없는지에 따라 범주형 변수를 만들었고, 그에 따라 사회자 또는 대표자 1명을 선택하였다. 대표자의 경우는 프로그램 시작 이전 최근 6개월 간 기사 개수가 가장 많은 사람으로 선택하였다. 이전 방송 프로그램 수는 네이버 인물검색을 사용하여 '방송' 카테고리의 프로그램 수로 하였다.

2.2.12. 파생 변수 예측력을 좀 더 높이기 위해 설명변수들을 이용하여 새로운 파생 변수들을 생성하였다. 다음은 파생 변수에 대한 설명이다.

- 사회자 및 대표자 유형

  사회자와 대표자를 각각 1명씩 선택하여서 그 인물의 주요 직업을 분류하였다. '가수', '배우', '예능', '기타' 4가지로 범주화 하였다. 가수로 데뷔했어도 가수활동보다 예능활동에 중점을 둔 인물이면

Table 2.3. Description of variables

Variable	Description	Type
Input varibles		
rate1	첫 회 시청률	Numerical
year	시작 연도	
PD1	PD 전작 예능 회차	
MC1	대표자 및 사회자 이전 방송 프로그램 수	
newarticle	프로그램 시작 한 달 전부터 2회 방송 전까지의 기사 수 (시작 연도 기준으로 centering)	
newarticle0	프로그램 시작 한 달 전부터 첫 방송 전까지의 기사 수 (시작 연도 기준으로 centering)	
season	시즌	
prev	이전 시즌 시청률	
rate.hat	평균 시청률 예측치 (회차 분석에서만 사용)	
com	방송사(KBS, SBS, MBC, JTBC, Mnet, tvN)	Categorical
time	방송시간(금요일, 주말 17시 20시, 그 외)	
type	예능 종류(가족/동물, 경연, 버라이어티, 오디션/서바이벌, 음식, 토크/개그, 기타)	
age	관람가(전체 관람가 및 7세 이상, 12세 이상, 15세 이상, 19세 이상)	
par1	방청 여부	
par2	시청자 참여 여부	
starPD	유명PD 여부	
Response variables		
$y1$	회차(∼12회, 13회∼)	Categorical
$y2$	평균 시청률	Numerical

예능으로 분류하였다.

- 유명PD

  대중들이 알만큼 유명한 PD가 제작한 프로그램의 경우, 해당 예능 프로그램에 대한 관심이 높아지게 된다. 이를 반영하기 위해 PD의 유명 여부를 변수로 생성하였다. 2010년부터 2017년 5월 15일까지의 PD 기사 개수를 기준으로 10,000개 이상인 경우 유명PD라고 정의하였다.

- 평균 시청률 예측치

  회차 예측을 하고자 할 때, 평균 시청률이 많은 영향을 줄 것이라고 예상되어 평균 시청률 모형을 이용하여 예측치를 만들어 설명변수로 고려하였다. 평균 시청률이 높을수록 회차가 길어질 것이라고 예상한 바와 같이 Figure 2.7을 보면 회차가 클수록 시청률이 높은 경향이 있다는 것을 확인할 수 있다. 본 논문에 사용된 모든 변수들을 Table 2.3에 정리해 놓았다.

3. 분석 결과

3.1. 평균 시청률 분석 결과

이번 장에서는 통계적 기법들을 이용하여 예능 프로그램의 평균 시청률을 예측하는 모형을 구축하고 평균 시청률에 영향을 미치는 변수가 무엇인지 파악하고자 한다. 평균 시청률의 분포는 Figure 2.1에서 보듯 오른쪽으로 긴 분포이므로 모형 구축 시 반응변수에 로그를 취하였다. 3.1.1절에서는 예능 프로그

Table 3.1. The result of cross validation error of each model (Model1)

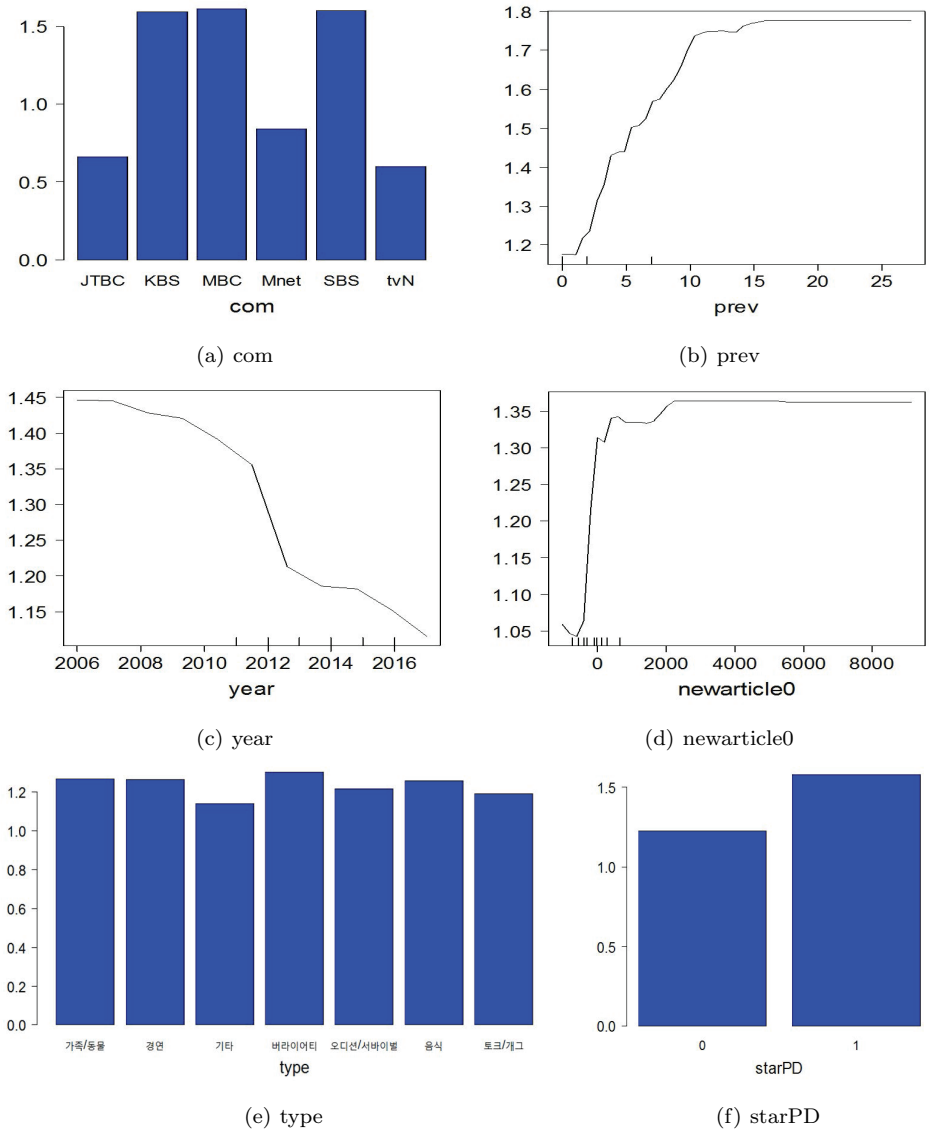
	CV error
Linear regression	3.0485 (0.1825)
Linear regression with stepwise	2.9847 (0.1755)
Ridge regression	2.9830 (0.1801)
LASSO regression	2.9941 (0.1809)
Partial least squares	3.0293 (0.1821)
Principal component regression	3.0442 (0.1908)
Bagging	2.5584 (0.1505)
Random forest	**2.5189 (0.1511)**
SVM - Linear	3.1345 (0.2089)
SVM - Radial	2.7194 (0.1604)
SVM - Polynomial	3.1681 (0.2191)

램이 시작하기 전에 평균 시청률을 예측하는 모형(Model1)을 적합해보고 3.1.2절에서는 프로그램 1회 방영 후 평균 시청률을 예측해본다(Model2). 적합된 모형들의 예측력을 비교하기 위해 10-fold 교차 평가 방법을 1,000번 반복하여 평균 교차오차를 계산하였다. 앞으로의 분석은 전체 데이터를 이용하여 구한 10-fold 교차오차의 평균값이 최소인 모형을 최종 모형으로 선정하고, 그 모형에서 반응변수에 영향을 미치는 중요 변수를 도출하도록 하겠다.

3.1.1. 예능 프로그램의 시작 전 평균 시청률 예측 모형 (Model1) 프로그램 시작 전 평균 시청률을 예측하기 위하여 newarticle(프로그램 시작 한 달 전부터 2회 방송일 전날까지 기사 개수를 시작 연도에 따라 중심화한 변수)과 1회 시청률을 제외하고 모든 변수를 사용하였다. 시청률을 예측하기 위해 선형회귀분석 방법으로는 선형회귀모형, Ridge, LASSO (Tibshirani, 1996), 부분최소제곱, 주성분회귀(principal component regression; PCR) 방법을 사용하였고 비선형회귀분석 방법으로는 배깅, 랜덤 포레스트, 서포트 벡터 머신(support vector machine; SVM)을 사용하였다. Table 3.1은 각 모형에서 교차오차 평균의 결과를 나타내며, 랜덤 포레스트 모형이 오차가 2.5189%로 가장 좋은 예측력을 보인다.

최적 모형으로 선택된 랜덤 포레스트 모형에서 가장 중요도가 높은 변수들을 선택하여 각 변수가 변화함에 따라 평균 시청률 예측치에 어떻게 영향을 미치는지 살펴보고자 한다. Figure 3.1(a)–(f) plot은 특정 변수를 제외한 다른 변수들의 값은 일정 수준으로 고정하고 특정 변수의 값을 변화시켜 평균 시청률 예측값의 변화를 관측한 것이다.

랜덤 포레스트 모형에서 com(방송사)이 가장 중요한 변수로 선택되었으며 prev(이전 시즌의 시청률), year(시작 연도), newarticle0, type, starPD(유명 PD 여부) 또한 중요한 변수로 나타났다. Figrue 3.1(a) com을 보면 공중파 프로그램(SBS, MBC, KBS)의 평균 시청률이 케이블 채널 프로그램보다 높은 것으로 나타났다. 공중파 프로그램 중에서는 MBC가 평균 시청률이 높고 케이블 채널 프로그램 중에서는 Mnet이 높음을 알 수 있다. Mnet의 경우 대중들에게 인기를 끈 오디션 프로그램이 많이 편성하였기 때문으로 보인다. Figure 3.1(b) prev의 경우 이전 시즌의 시청률이 높을수록 평균 시청률이 증가하다가 10% 이상이 되면 그 차이가 미미하다. Figure 3.1(c)의 year의 경우 최근에 시작된 프로그램일수록 평균 시청률이 전반적으로 감소하는 추세를 보이며 특히 2011년과 2012년 사이에 평균 시청률이 가파르게 감소하는 것을 볼 수 있다. 이는 2011년부터 스마트폰 시장의 활성화로 TV 시청이 줄어들었기 때문으로 보인다. newarticle0 변수의 그래프를 보면 프로그램 시작 연도의 평균 기사 개수보다 적을 경우 기사 개수가 증가할수록 시청률이 급격히 증가하지만 평균 기사 개수보다 2,000개 이상 많은 프



(a) com (b) prev

(c) year (d) newarticle0

(e) type (f) starPD

Figure 3.1. log(Average ratings) according to com, prev, year, newarticle0, type and starPD.

로그램들의 경우 기사 개수에 따른 평균 시청률 차이가 거의 없는 것으로 나타났다. 따라서 프로그램 기획 시 그 해 평균 보다 2,000건 넘게 기사를 낸 경우 홍보비용에 비해 시청률 증가 효과가 적게 나타날 수 있다. 또한 프로그램의 유형(type)이 버라이어티일 때, 유명 PD가 기획한 프로그램일 때 그렇지 않은 프로그램보다 평균 시청률이 높다.

선형회귀모형 중 Stepwise Regression, Ridge 회귀모형이 랜덤 포레스트 모형과 예측력이 비슷한 것으로 나타났다. Table 3.2는 선형회귀모형 중 예측력이 가장 좋은 Ridge 회귀모형을 통해 추정한 변수들의 회귀 계수와 표준오차, p-value를 정리한 표이다.

Table 3.2. The table of coefficients of ridge regression model

변수	회귀계수	표준오차	p-value	
comMnet	0.044674	0.054661	0.41376	
comJTBC	0.027342	0.044653	0.54032	
comKBS	0.330353	0.046764	1.62E−12	***
comSBS	0.431372	0.045665	<2.00E−16	***
comMBC	0.397419	0.045390	<2.00E−16	***
year	−0.182920	0.042699	1.84E−05	***
time2	0.051061	0.040716	0.20982	
time3	0.070243	0.046236	0.12870	
type경연	−0.040540	0.047317	0.39153	
type기타	−0.109820	0.045871	0.01666	*
type버라이어티	−0.017060	0.046500	0.71372	
type오디션/서바이벌	−0.120110	0.054655	0.02798	*
type음식	0.046841	0.039643	0.23737	
type토크/개그	−0.076900	0.047497	0.10543	
season	0.098107	0.050290	0.05108	·
prev	0.135836	0.043666	0.00187	**
age2	0.092391	0.056691	0.10316	
age3	0.083687	0.058377	0.15170	
age4	0.050706	0.060573	0.40254	
par11	0.011992	0.049281	0.80775	
par21	0.135338	0.046862	0.00388	**
PD1	0.042995	0.036160	0.23444	
MC1	−0.074990	0.047627	0.11538	
isMC1	−0.032700	0.042642	0.44315	
REPtype기타	0.004772	0.043156	0.91195	
REPtype배우	0.046145	0.048565	0.34202	
REPtype예능	0.092720	0.058527	0.11314	
starPD1	0.204607	0.038948	1.49E−07	***
newarticle0	0.047521	0.037050	0.19963	

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

Table 3.2의 변수 중 유의수준 0.05 하에서 유의한 변수들에 대해서만 설명하겠다. 공중파 채널 프로그램의 평균 시청률이 tvN 채널에 비해 높은 것으로 나타났다. 시즌이 거듭될수록, 이전 시즌의 시청률이 클수록 평균 시청률이 증가한다. 추가적으로 시청자가 참여하는 프로그램일수록, 유명 PD일수록 평균 시청률이 증가하는 것으로 나타났다.

이와는 반대로 최근 시작된 프로그램일수록 평균 시청률이 낮아진다. 또한 프로그램 유형이 오디션/서바이벌이거나 기타일 때, 가족/동물일 때보다 평균 시청률이 낮아진다. 이는 앞서 Figure 3.1에서 확인한 결과와 일치한다.

3.1.2. 예능 프로그램의 1회 방영 후 평균 시청률 예측 모형 (Model2) 이번 절에서는 프로그램의 1회 방영 후 평균 시청률을 예측하고자 한다. Model2에서는 Model1에서 포함된 변수에서 newarticle0를 제외하고 1회 시청률, newarticle이 포함되었다. 3.1.1절에서 사용된 방법론 11가지를 그대로 적용하여 모형을 구축하였다.

Table 3.3. The result of cross validation error of each model (Model2)

	CV error
Linear regression	2.5724 (0.2275)
Linear regression with stepwise	2.5218 (0.2307)
Ridge regression	2.4018 (0.1808)
LASSO regression	2.4593 (0.2054)
Partial least squares	2.4963 (0.2092)
Principal component regression	2.6984 (0.1877)
Bagging	1.7590 (0.1174)
Random forest	**1.7332 (0.1176)**
SVM - Linear	2.5813 (0.2348)
SVM - Radial	2.1311 (0.1615)
SVM - Polynomial	2.5812 (0.2351)

Table 3.3을 보면 1회 시청률을 알면 모든 방법론에서 평균 시청률 예측이 개선된다는 사실을 알 수 있다. 특히 배깅 모형과 랜덤 포레스트 모형의 교차오차가 1%대로 확인되어 매우 정확하게 추정되었다.

Model1의 Figure 3.1과 같이 Model2의 랜덤 포레스트 모형에서 가장 중요도가 높은 세 변수(rate1, com, type)에 대한 그래프를 그려보았다. rate1의 경우, 다른 변수에 비해 월등히 중요한 변수로 나타났다. 1회 시청률이 클수록 평균 시청률이 증가하다가 1회 시청률이 10% 이상일 때는 증가량에 큰 차이가 없는 것으로 나타났다. 그리고 공영 방송사일수록, 프로그램 유형이 음식일수록 평균 시청률이 증가한다.

다음으로 모델 설명이 쉬운 선형회귀모형을 살펴보겠다. Table 3.4는 선형회귀모형 중 예측력이 가장 좋은 Ridge 회귀모형의 회귀 계수와 표준오차, p-value를 정리한 표이다.

Table 3.4의 변수 중 유의수준 0.05 하에서 유의한 변수들에 대해서만 설명하겠다. 공중파 또는 JTBC 채널 프로그램일 때 tvN에 비해 평균 시청률이 높은 것으로 나타났다. 또한 Model1의 결과와 마찬가지로 시즌이 거듭될수록, 시청자가 참여하는 프로그램일수록, 유명 PD일수록 평균 시청률이 증가하는 것으로 나타났다. 그리고 시작 연도가 최근일수록, type이 기타이거나 토크/개그일 때 평균 시청률이 낮아지며, 이전 시즌의 시청률이 증가할 때, 평균 시청률이 감소하는 것으로 나타났다. 이는 Model1과 다른 결과로 Model2에 이전 시즌의 시청률과 상관이 높으면서(55%) 반응변수에 영향을 크게 미치는 1회 시청률이 포함됨으로써 반대의 결과가 나온 것으로 추측된다.

3.2. 회차 분석 결과

이번 장에서는 앞서 적합한 평균 시청률 예측 모형을 이용해 새로운 설명변수를 만든 후 다양한 통계적 방법들을 이용하여 예능 프로그램이 3개월 이상 지속될지를 예측하는 분류 모형을 적합하고자 한다. 또한 어떤 변수가 예능 프로그램의 수명을 결정하는 데에 영향을 미치는지 알아볼 것이다. $y1$(회차) 변수를 12회 이하/12회 초과로 나누어 0과 1의 값을 매긴 다음 이를 group이라는 변수로 지정하여 반응변수로 사용하였다. 회차 변수의 경우 12회의 빈도수가 가장 높았는데, 이를 통해 방송사 입장에서 프로그램의 인기 등을 척도로 방영 지속 여부를 결정하는 시기가 12회일 것이라고 예상하였다. 따라서 12회 초과 방영되는 프로그램은 장기 프로그램으로 발전할 가능성이 있다고 판단하였다. 오디션과 같이 회차가 정해져있는 포맷의 프로그램들은 데이터에서 제외한 후, 12회 초과 방영된 프로그램은 112개이고 12회 이하 방영된 프로그램은 53개이다. 분석에 사용한 방법론은 로지스틱 회귀모형, 부분최소제곱, 주성분회귀, Ridge, LASSO, 랜덤 포레스트, 서포트 벡터 머신이다. 이 중 회귀모형의 경우 수치형 반응

Table 3.4. The table of coefficients of ridge regression model

변수	회귀계수	표준오차	p-value	
rate1	0.554616	0.053109	<2E−16	***
comMnet	0.080938	0.055755	0.14660	
comJTBC	0.112006	0.043610	0.01022	*
comKBS	0.276867	0.049549	2.30E−8	***
comSBS	0.346090	0.049109	1.82E−12	***
comMBC	0.335228	0.047133	1.14E−12	***
year	−0.067580	0.040279	0.09338	.
time2	0.015834	0.035847	0.65869	
time3	0.019442	0.041662	0.64075	
type경연	−0.027530	0.051162	0.59053	
type기타	−0.104890	0.047999	0.02887	*
type버라이어티	−0.047420	0.048021	0.32339	
type오디션/서바이벌	−0.058760	0.062624	0.34808	
type음식	0.053484	0.037956	0.15880	
type토크/개그	−0.102790	0.051133	0.04440	*
season	0.111667	0.047665	0.01914	*
prev	−0.118500	0.045092	0.00859	**
age2	0.040149	0.063244	0.52554	
age3	0.025898	0.067334	0.70052	
age4	0.052422	0.068330	0.44297	
par11	0.005483	0.046020	0.90516	
par21	0.095748	0.043988	0.02950	*
PD1	0.020493	0.030530	0.50207	
MC1	0.007399	0.045510	0.87085	
isMC1	−0.052070	0.038763	0.17915	
REPtype기타	0.045780	0.040814	0.26200	
REPtype배우	0.073105	0.049413	0.13902	
REPtype예능	0.056597	0.063285	0.37115	
starPD1	0.141475	0.034642	4.43E−5	***
newarticle	−0.000850	0.032896	0.97935	

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

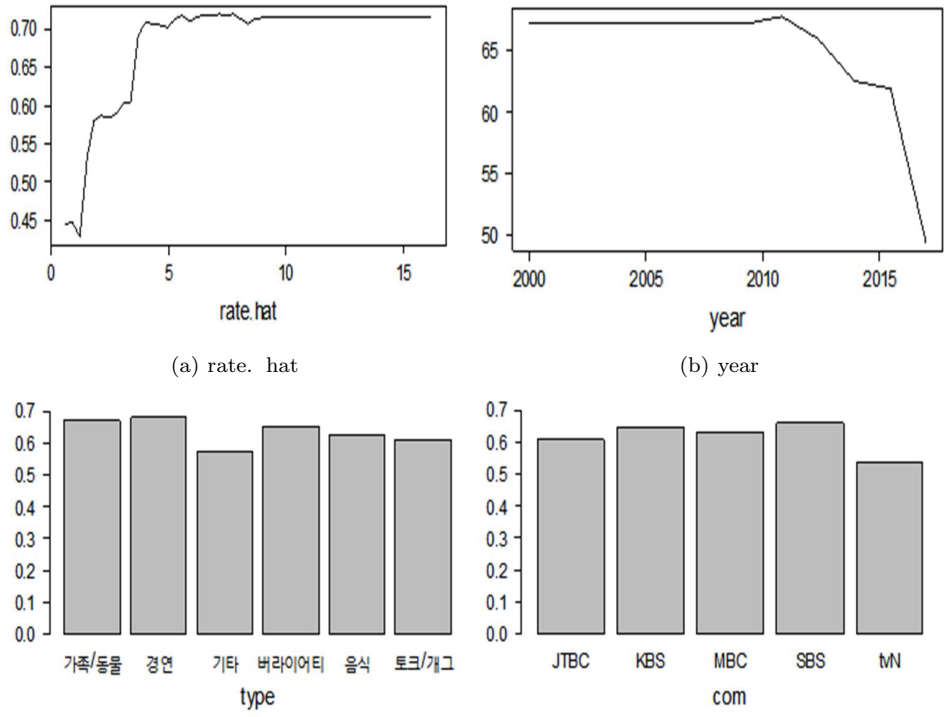
변수로 수정한 후(1 = 12회 이하, 2 = 12회 초과)모형을 적합하고 그 결과를 이용해 각 그룹(0 = 12회 이하, 1 = 12회 초과)으로 분류하였다. 각 모형 간 성능을 비교하기 위하여 10-fold 교차 평가 방법을 사용하여 모형 별로 평균 오분류율을 계산하는 과정을 1,000번 반복하였다. 모형들 중 오분류율이 가장 작은 모형을 최적 모형으로 선택한 다음 최적 모형에서 중요 변수를 살펴보려 한다.

3.2.1. 예능 프로그램의 회차 예측 모형 앞서 3.1.2절에서 적합한 최적 평균 시청률 예측 모형을 이용하여 새로운 파생변수 rate.hat(평균 시청률 예측치)를 생성한 다음, 이를 설명변수로 포함시켜 각 모형에 적합하였다. 최적 평균 시청률 예측 모형은 프로그램이 1회 방영 후 데이터를 이용한 모형이기 때문에 회차 예측 모형에서도 역시 newarticle(방영 한 달 전부터 2회 방영 전날까지의 기사 개수를 연도별 평균으로 중심화한 값) 변수가 설명변수에 포함된다.

Table 3.5를 살펴보면 랜덤 포레스트 모형의 CV error가 가장 낮게 나타났다. 따라서 랜덤 포레스트 모

Table 3.5. Misclassification rate of each model

	CV error
Logistic regression	0.2712 (0.0172)
Logistic regression with stepwise	0.2724 (0.0183)
Ridge	0.2281 (0.0129)
LASSO	0.2418 (0.0129)
Partial least squares	0.3223 (0.0049)
Principal component regression	0.2362 (0.0158)
Bagging	0.2383 (0.0137)
Random forest	0.2198 (0.0103)
SVM - Linear	0.2443 (0.0220)
SVM - Radial	0.2384 (0.0180)
SVM - Polynomial	0.2401 (0.0140)



(a) rate. hat (b) year

(c) type (d) com

Figure 3.2. $P(Y = 1|X)$ according to the value of important variables.

형을 최적 모형으로 선정하였다.

이번에는 중요 변수와 그들이 반응변수에 미치는 영향을 알아보자. 랜덤 포레스트 모형에서 중요 변수로 선택된 변수는 rate.hat(평균 시청률 예측치)와 com(방송사), year(시작연도), type(프로그램 유형)이다. 그 중에서도 rate.hat 변수의 중요도가 월등히 높았다. 네 변수의 값에 따라 반응변수가 어떻게 변하는지 알아보고자 한다.

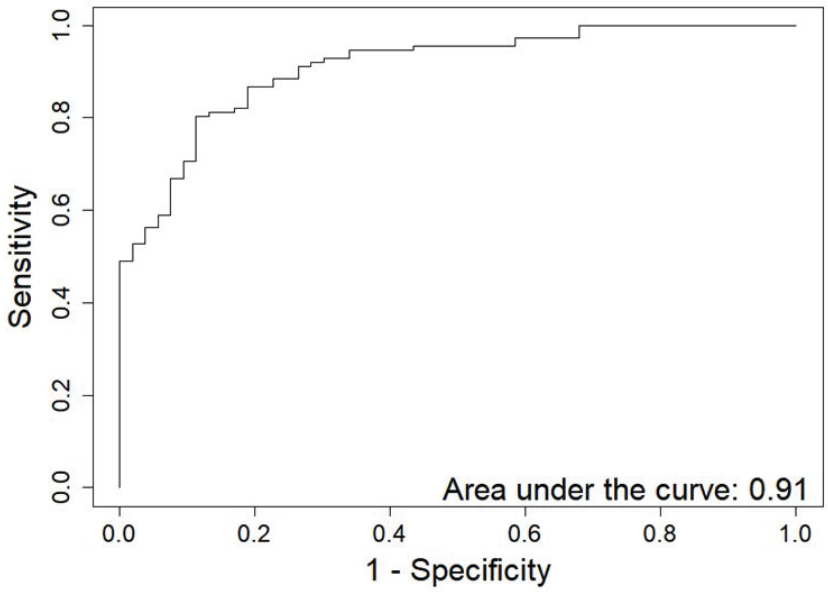


Figure 3.3. ROC curve of random forest model.

Figure 3.2는 중요변수들의 값에 따라 그룹 1(12회 초과)로 분류될 확률을 보여준다. 평균 시청률 예측치가 증가할수록 12회 초과 그룹으로 갈 확률이 증가하는데, 약 1.8%–3.4%까지는 그 값의 변화가 없다가, 다시 약 4%까지는 증가한다. 이후 4% 이상인 경우에는 증가량이 미미한 것으로 나타났다. 이는 평균 시청률이 높을수록 프로그램이 더 오래 방영된다는 의미이다. 또한 2010년 이후로는 해가 갈수록 12회 초과 그룹으로 분류될 확률이 낮아지는데, 이 시기에 회차가 30회 이하인 프로그램이 전체 중 70% 이상을 차지하기 때문으로 보인다. 프로그램 유형의 경우에는 경연, 가족/동물의 경우가 12회 초과 그룹으로 분류될 확률이 상대적으로 크다. 방송사의 경우 공영사의 프로그램이 케이블채널의 프로그램보다 더 오래 방영될 것이라는 것을 알 수 있다. 이는 앞서 2장에서 살펴본 내용과 일치한다.

Figure 3.3은 데이터를 이용하여 랜덤 포레스트 모형의 (OOB estimates를 이용한) ROC 곡선을 그린 것이다. AUC 값이 0.91로 매우 높은 것을 보아 랜덤 포레스트 모형의 성능이 매우 좋은 것을 알 수 있다.

실제로 12회 초과 그룹에 속하지만 12회 이하 그룹으로 예측된 데이터들의 특징을 살펴보자면, 우선 방송국의 경우 대부분이 케이블 채널 프로그램이었다. 또한, 평균 시청률 예측치가 매우 낮아 대부분이 2%가 채 안 되었다. 반대로 12회 이하 그룹이지만 12회 초과 그룹으로 예측된 데이터들의 특징을 살펴보면 방송국의 경우 공영방송사가 많았고, 예능 유형의 경우 기타와 토크/개그가 많았다. 이러한 결과는 앞서 Figure 3.2에서 확인한 결과와 일치한다.

12회 이하 그룹으로 오분류된 데이터들 중에는 출연진이 많거나 유명한 MC가 출연하는 등 제작비가 많이 들었을 것이라고 예상되는 데이터가 매우 많았다. 방송사 입장에서는 투자 대비 성과가 좋지 않은 프로그램을 계속 이어가기엔 부담이 클 것이다. 이렇듯 프로그램의 회차를 예측하는 데 있어서 제작비는 매우 큰 영향력을 가지는데, 그에 대한 데이터는 구할 수 없어 모형에 포함시키지 못하였다. 제작비에 대한 정보가 있다면 모형의 예측력은 더 좋아질 것으로 예상된다.

4. 결론

본 연구에서는 선형 모형과 비선형 모형, 다양한 데이터마이닝 기법을 활용하여 평균 시청률과 대략적인 프로그램 회차를 예측하는 모형을 구축하고 그들의 성능을 비교해보았다.

시청률 예측 모형 중에서는 배깅, 랜덤 포레스트 모형의 성능이 좋았는데, 그 중에서도 첫 방송 후 적합한 랜덤 포레스트 모형의 RMSE 값이 1.733%로 가장 낮았다. 랜덤 포레스트 모형들에서 공통적으로 중요 변수로 선택된 변수들은 방송국과 프로그램 유형이었다. 첫 회 시청률과 이전 시즌 시청률, 연도, 프로그램 기사 개수, 유명 PD 여부도 중요도가 높은 것으로 나타났다. 이러한 변수들은 2장에서 살펴보았을 때 시청률과 뚜렷한 관계를 보였던 변수들이다.

회차 예측 모형 중에서는 랜덤 포레스트 모형의 교차평가 오분류율이 0.2198로 가장 낮게 나타났다. 회차를 잘 예측하기 위해서는 평균 시청률 예측치, 시작 연도, 예능 유형과 방송국 변수의 역할이 중요하다는 사실을 알 수 있었다.

예능 시장이 시시각각 변하면서 생기는 데이터의 변화를 추적하기 위해서는 새로운 경향이 나타난 뒤 일정 시간이 지난 후 경향성을 포함하는 데이터가 생성되었을 때 모형을 적합하는 것이 합리적일 것이다. 또한 프로그램의 수명을 결정하는 데 중요한 역할을 하는 제작비 정보도 포함된다면 모형의 예측력을 향상시킬 수 있을 것이다.

본 연구를 수행하면서 데이터 테이블을 완성시키는 데 시간이 오래 걸리는 어려움이 있었다. 방송사 별로 종영된 프로그램 정보를 제공하긴 했지만, 각 프로그램 페이지에서 제공하지 않는 정보들이 있어 이를 포털 사이트 등에서 알아내야 했던 번거로움이 있었다. 또한 본 연구는 프로그램 회차를 정확히 예측하지 못했다는 한계점이 있다. 단지 3개월 이상 방영 여부를 분류하는 데 그치지 않고 프로그램의 정확한 수명을 예측할 수 있었다면 상황에 따라 다양한 방안을 제시할 수 있었을 것이다. 이를 보완하기 위해서는 예능 프로그램 회차와 관련이 높은 제작비 정보가 필요할 것이다. 또한 본 논문에서는 2017년 5월 기준으로 종영한 프로그램만을 다루었지만 생존분석모형을 이용한다면 현재 방영중인 프로그램도 포함해서 분석이 가능할 것이다.

본 논문에서 프로그램 평균 시청률의 오차가 1%대로 매우 낮았으며 프로그램 회차에서도 오분류율이 20% 전후로 매우 좋은 예측력을 보였다. 때문에 프로그램을 제작하는 입장에서 프로그램의 평균 시청률과 회차 예측이 시청자들을 사로잡을 수 있는 프로그램 기획에 도움이 될 것이라는 데에 본 연구의 의의가 있을 것이다. 뿐만 아니라 광고주나 방송사로 하여금 예산을 효율적으로 분배할 수 있도록 도울 것이다.

References

Breiman, L. (1996). Bagging predictors, *Machine Learning*, **24**, 123–140.

Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5–32.

Han, B. K. (2016). An explorative analysis of the factors affecting entertainment TV show ratings: focused on channel, program schedule and production factors.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer, New York.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to statistical Learning*, Springer, New York, USA.

Kang, S., Jeon, H., Kim, J., and Song, J. (2015). A study on domestic drama rating prediction, *The Korean Journal of Applied Statistics*, **28**, 933–949.

Lee, H. E. and Choi, H. S. (2016). A study on the effect of entertainment show on the tourism. In *Korea Contents Association 2016 Spring Conference*, 215–216.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society B*, **58**, 267–288.

국내 예능 시청률과 회차 예측 및 영향요인 분석

김미림[a] · 임소연[a] · 장초희[a] · 송종우[a,1]

[a]이화여자대학교 통계학과

(2017년 8월 14일 접수, 2017년 9월 29일 수정, 2017년 10월 12일 채택)

요 약

오디션, 육아, 버라이어티 등 다양한 예능 프로그램들의 수가 점점 증가하고 있다. 특히 종합편성채널이 개국한 이후에 예능 시장 경쟁이 심화되고 있다. 그에 따라 시청률과 회차에 대한 연구의 필요성이 대두되고 있다. 본 연구의 목적은 예능 프로그램 시청률과 회차의 예측모형을 제시하고 주요요인을 살펴보는 데 있다. 모형 적합 시 선형회귀모형, 로지스틱 회귀모형, LASSO 회귀모형, 랜덤 포레스트, 그래디언트 부스팅, 서포트 벡터 머신 등과 같은 다양한 분석 방법을 고려하였다. 예능 시청률 예측 모형에서는 첫 회가 방영되기 전과 방영된 후 두 가지 모형을 적합하였고, 회차 예측 모형에서는 예능 시청률 예측 모형의 예측치를 추가 변수로 생성하여 모형을 적합하였다. 그 결과 첫 회 방영 전 예능 시청률 예측에서는 방송사, 이전 시즌 시청률, 시작 연도, 기사 수가 큰 영향을 주는 것으로 나타났다. 첫 회 방영 후 예능 시청률 예측에서는 첫 회 시청률, 방송사, 예능 유형이 중요한 변수로 나타났으며, 두 모형 모두 랜덤 포레스트 모형에서 가장 좋은 결과를 보였다. 예능 회차 예측에서는 평균 시청률 예측치, 시작 연도, 예능 유형, 방송국 등이 중요한 변수로 나타났다.

주요용어: 예능 프로그램, 시청률, 회차, 선형회귀모형, 로지스틱 회귀모형, Ridge, LASSO, 서포트 벡터 머신, 랜덤 포레스트, 그래디언트 부스팅, 중요 변수

이 논문 또는 저서는 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017R1D1A1B03036078).

[1]교신저자: (03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과.
E-mail: josong@ewha.ac.kr