The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.5.933
28(5), 933–949

A Study on Domestic Drama Rating Prediction

Suyeon Kang[a] · Heejeong Jeon[a] · Jihye Kim[a] · Jongwoo Song[a,1]

[a]Department of Statistics, Ewha Womans University

(Received July 10, 2015; Revised July 23, 2015; Accepted July 26, 2015)

Abstract

Audience rating competition in the domestic drama market has increased recently due to the introduction of commercial broadcasting and diversification of channels. There is now a need for thorough studies and analysis on audience rating. Especially, a drama rating is an important measure to estimate advertisement costs for producers and advertisers. In this paper, we study the drama rating prediction models using various data mining techniques such as linear regression, LASSO regression, random forest, and gradient boosting. The analysis results show that initial drama ratings are affected by structural elements such as broadcasting station and broadcasting time. Average drama ratings are also influenced by earlier public opinion such as the number of internet searches about the drama.

Keywords: drama rating, linear regression, LASSO regression, random forest, gradient boosting, important variables

1. 서론

최근 드라마 시장에는 다양한 주제를 다루고 여러 유명 배우들을 캐스팅한 드라마들이 등장하고 있다. 2011년 말 방송법이 개정됨에 따라 종합편성채널이 개국한 이후 지상파 방송사 중심이었던 드라마 시장이 케이블 및 종합편성채널까지 확대되었으며 스마트폰, 태블릿 등 드라마를 시청할 수 있는 방법이 다양화되었다. 이렇게 방송 시장이 빠르게 변화하고 있음에도 불구하고 시청률은 TV를 기반으로 하는 전통적인 방법으로 측정되어 프로그램 제작자와 광고주들에게 광고비 규모를 산정하는 데 매우 중요한 척도로 활용되고 있다. 특히 드라마는 대중적 인기나 그에 따른 사회적 영향력이라는 차원에서 다른 어떤 장르의 프로그램보다도 중요한 의미를 지니기 때문에 (Bae, 2005) 드라마 시청률을 예측하는 것은 제작자와 광고주 입장에서 매우 중요하다. 드라마 시청률에 관한 연구는 이미 오래전부터 진행되어왔지만 대부분의 연구들이 일반회귀모형에 기반하고 있으며 분석 대상이 특정 방송사 또는 지상파 방송사의 드라마에 한정되어 있다. 본 연구에서는 이런 한계점을 극복하고자 최근 방송시장의 변화를 고려한 다양한 통계적 예측 모형을 제시하고자 한다.

본 연구의 목적은 지상파 방송사와 케이블, 종합편성채널을 모두 포함한 드라마를 대상으로 하여 다양한 데이터마이닝 기법을 활용한 시청률 예측 모형을 제시하고 시청률 예측에 중요한 영향을 미치는 요

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the ministry of Education, Science and Technology (No. NRF-2013 R1A1A2012817).

[1]Corresponding author: Department of Statistics, Ewha Womans University, Seoul 120-750, Korea.
E-mail: josong@ewha.ac.kr

인들을 도출하는 데 있다. 분석에 사용되는 회귀모형들은 선형회귀모형, LASSO 회귀모형 (Tibshirani, 1996), 의사결정나무 (Brieman 등, 1984), 랜덤포레스트 (Breiman, 2001), 그래디언트 부스팅 (Friedman, 2002; Ridgeway, 2012), 서포트 벡터 기계 (Cortes와 Vapnik, 1995)이며 예측력 평가 지표로는 MSE를 이용하였다. 여기서 의사결정나무와 서포트 벡터 기계는 예측력 평가 시 다른 모형에 비하여 예측력이 현저히 떨어져 모형 비교에서 제외하였다. 본 연구에서의 모든 분석은 R (R Development Core Team, 2010)을 통해 이루어질 것이며, R에서 제공하는 다양한 함수들을 이용하여 유의한 변수들을 도출할 것이다. 이를 위해 최근 5년 간(2011–2015) 종영한 드라마 260개의 시청률 자료를 수집하였고 시청률에 영향을 미칠 것이라 예상되는 드라마 내·외부적 요인들을 고려하여 변수 생성을 하였다. 본 연구의 핵심은 우선 드라마의 초기 시청률에 대한 예측 모형을 제시한 뒤 최종적으로 드라마의 평균 시청률을 예측하는 것이다. 이렇게 2개의 모형을 제시하는 이유는 다음과 같다. 드라마 방영 이전에는 연출자와 주연배우와 같이 시청자들이 드라마 방영 전에 알 수 있는 기본 정보들만을 고려한 모형을 적합한다. 그리고 드라마가 방영되기 시작한 이후에는 해당 드라마에 대한 초기의 여론이 전체 시청률에 많은 영향을 미칠 것으로 생각되기 때문에 드라마의 기본 정보들과 초기 여론을 모두 고려한 모형을 적합한다. 이렇게 모형을 적합한 후 각 모형에서 시청률 예측에 중요하게 작용하는 변수들이 어떻게 달라지는지도 살펴보고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 분석 데이터의 수집 방법과 주요 변수들에 대해 설명하고 3장에서는 모형을 적합하기 위해 사용된 다양한 데이터마이닝 기법들에 대해 간단히 설명한다. 그리고 4장에서 이 기법들을 이용한 분석 결과와 드라마의 시청률을 예측하는 최종 모형을 제시한 후 5장에서 본 연구의 결과를 요약한다.

2. 분석자료 설명

2.1. 자료수집 과정

본 연구의 연구 대상은 2011년부터 2015년 6월 1일까지 종영한 260개의 드라마이다. 단, 재방송은 고려하지 않았으며 9회작 미만인 일회성 드라마는 연구 대상에서 제외하였다. 드라마와 관련된 분석 데이터를 수집하기 위하여 다양한 웹사이트를 이용할 수 있었다. 본 연구의 핵심이 되는 드라마의 시청률은 리서치 회사인 'AGB 닐슨미디어리서치(www.agbnielsen.co.kr)'에서 제공하는 시청률 자료를 사용하였다. 시청률에 영향을 미치는 여러 가지 요인으로는 연출자, 작가, 주연배우, 프로모션, 편성시간, 날씨, 경제지표 등을 고려하였는데 이에 대한 자세한 설명은 다음 절에서 하고자 한다. 드라마의 기본 정보는 구글(www.google.co.kr), 네이버(www.naver.com)와 같은 포털 사이트로부터 얻을 수 있었다. 날씨와 경제지표는 각각 기상청(www.kma.go.kr)과 통계청(www.kostat.go.kr) 웹사이트에서 수집했다.

2.2. 변수 설명

본 연구의 목적은 시청률에 영향을 미칠 것이라 예상되는 변수들을 이용하여 드라마의 초기 시청률과 평균 시청률을 예측하는 것이다. 이를 위해 반응변수와 설명변수들을 다음과 같이 정의하였다. 우선 초기 시청률이란 드라마 1, 2회의 평균 시청률을 의미하며, 평균 시청률은 일일 시청률 전체의 평균을 의미한다. 수집된 자료에서 초기 시청률의 최솟값은 0.29, 최댓값은 27.05로, 해당하는 드라마는 각각 2012년 채널A에서 방영된 '판다양과 고슴도치', 2014년 KBS2에서 방영된 '참 좋은 시절'이다. 또한 평균 시청률의 최솟값은 약 0.25, 최댓값은 33.3으로 해당 드라마는 각각 '판다양과 고슴도치', 2012년~2013년 KBS2에서 방영된 '내딸 서영이'이다. 최종 목표인 평균 시청률 예측을 위하여 시청률에 영향을 미치는 요인으로 드라마 내·외부적인 면을 모두 고려하였는데, 분석에 이용한 설명변수들은 다음과 같다.

2.2.1. 연출자·작가 다양하고 창의적인 드라마가 열풍인 만큼 드라마의 내용과 완성도가 시청률에 영향을 미칠 것이고, 유명 연출자나 작가의 드라마일수록 시청자들이 큰 관심을 가질 것이다. 이에 연출자와 작가 변수를 고려하였다. 두 변수 모두 해당 드라마가 방영되기 이전까지의 이력을 이용하여 과거에 연출했던 드라마 중 최고 시청률이 20% 이상(지상파 외 채널 2% 이상)인 드라마의 수를 기록하는 방식으로 수치화하였다.

2.2.2. 주연배우 드라마에 있어서 연출자와 작가도 시청률에 영향을 미치지만, 가장 영향력이 클 것이라 생각되는 변수는 단연 배우이다. 주연배우가 유명한 배우일수록 드라마가 흥행할 가능성이 높다. 이 때 주연배우는 포털 사이트 네이버에 해당 드라마의 주연으로 명시되어있는 배우를 의미한다. 주연배우는 드라마를 대표하는 배우로서 큰 비중을 차지하고 있으므로 주연 배우를 일종의 등급과 같이 수치화하는 데 있어 5가지의 기준을 두었다. 첫 번째와 두 번째 기준은 각각 해당 드라마 이전에 출연한 방송(드라마 및 예능과 같은 TV 브라운관에서의 방송활동)과 영화의 수이다. 세 번째 기준은 수상 경력으로, 각종 시상식을 비롯하여 각 방송사에서 매년 주최하는 연말 시상식에서 수상한 상의 개수를 점수화하였다. 이 때 대상, 최우수·우수 연기상, 그 외의 상에 대하여 각각 3, 2, 1로 가중치를 주었다. 네 번째 기준은 역대 한국 드라마 시청률 100위 안에 드는 드라마에 주연으로 출연한 횟수이다. 마지막으로 다섯 번째 기준은 해당 드라마의 주연배우 수이며 앞선 4가지 변수는 모두 주연배우의 수로 나눈 평균 점수를 사용하였다.

2.2.3. 프로모션 인터넷과 스마트폰의 발달로 대부분의 사람들은 언제 어디서든 드라마 관련 기사를 접할 수 있고 관심 있는 드라마를 실시간으로 검색해볼 수 있다. 이러한 사실에 기반하여 드라마에 대한 기사의 개수와 드라마 검색 횟수를 프로모션의 척도로 사용하고자 한다. 기사의 개수는 포털 사이트 네이버에서 제공하는 인터넷 기사를 기준으로 드라마 방영 전 날부터 3개월 이전까지, 그리고 방영한 날부터 1주일 이후까지의 기사 개수를 기록하였다. 검색어 자료는 네이버 트렌드 검색(trend.naver.com) 서비스를 통하여 얻을 수 있었으며 기간 설정은 기사 변수와 동일한 방법을 이용하였다. 단, 네이버 트렌드 검색 서비스는 정확한 검색 횟수가 아닌 상대적인 값만을 제공하므로 모든 드라마의 검색어 자료를 동일한 기준 하에서 추출하여 변수로 이용하였다. 예를 들어, 5개의 드라마 이름을 검색하면 5개 드라마 중에서 가장 많이 검색된 드라마의 점수를 100으로 하고 나머지 드라마들의 검색 횟수가 상대적인 값으로 나온다. 따라서 가장 많이 검색되는 키워드 하나를 기준점으로 잡고 나머지 드라마의 상대적인 검색수를 본 분석에 사용하였다.

2.2.4. 편성시간 방송사마다 드라마가 편성되는 방송시간대는 다양하다. 일반적으로 방송 시간이 시청률에 많은 영향을 미치므로 이 변수를 분석에 포함하였다. 편성시간대 변수로 방송요일과 방송시작 시간을 기록하였다.

2.2.5. 방송사 과거의 드라마 시장은 지상파 방송사들이 지배하고 있었지만, 최근에는 채널의 다양화로 드라마를 제작하는 방송사들이 많아지고 있다. 각 방송사들은 그들만의 이미지를 구축함으로써 시청자들의 프로그램 선택에 영향을 미치며, 기존 연구들에서도 채널을 시청 행위를 예측하는 데 중요하게 작용하는 변인으로 다뤄왔다 (Cohen, 2002; Bae, 2005). 이에 따라 각 방송사명을 변수로 사용하였다.

2.2.6. 방영년도 과거에는 드라마를 접할 수 있는 매체가 TV에 국한되어 있었지만, 최근 스마트폰의 발달로 다양한 방식으로 드라마를 시청할 수 있게 되었다. 이에 따라 상대적으로 TV를 시청하는 인

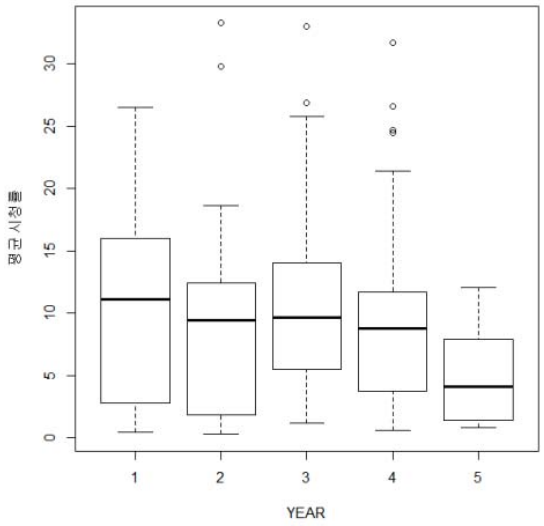

Figure 2.1. Box plots of drama ratings based on the year.

구의 수가 줄게 되므로 TV를 기반으로 측정되는 지표인 시청률 또한 그 영향을 받을 수밖에 없다. 따라서 방영년도를 시청률에 영향을 미치는 요인으로 고려하였고, 방영년도를 나타내는 변수로서 2011년부터 2015년까지 1–5의 값을 갖는 연속변수를 사용하였다. 방영년도 변수에 따른 시청률 분포는 Figure 2.1과 같다.

2.2.7. 경쟁작·이전작 방송 프로그램의 시청률은 비슷한 시간대에 다른 채널에서 하는 프로그램(경쟁작)이나 해당 프로그램 방영 이전에 같은 방송사에서 동시간대에 방영되던 프로그램(이전작)의 영향을 받는다. 경쟁 드라마를 시청하고 있는 사람이 많을수록 해당 시간대에 방영을 시작하는 드라마의 진입장벽이 높다고 할 수 있다. 그리고 연속적인 이야기라는 드라마의 특성상, 어떤 작품이 종영하면 다른 채널의 작품을 중간부터 시청하기보다는 해당 채널에서 방영하는 후속작을 시청하게 될 가능성이 높다. 또 일반적으로 작품의 후반부에 후속작의 예고편을 반복해서 방영하므로 인기드라마의 후속작의 경우 방영 전부터 좋은 광고효과를 볼 수 있다. 이런 이유로 경쟁작과 이전작의 평균 시청률을 변수로 사용하였다.

2.2.8. 원작 유무 일반적으로 만화나 소설을 원작으로 하는 드라마의 경우 원작의 높은 인기를 바탕으로 만들어진 작품이기 때문에 어느 정도의 시청률이 보장될 가능성이 높다. 이에 원작이 있는 경우 1, 없는 경우 0을 나타내는 원작 유무 변수를 고려하였다.

2.2.9. 경제 지표 대표적으로 사용되는 경제 지표로는 국내총생산(GDP)과 국민총소득(GNI)이 있다. 국내총생산과 국민총소득의 값이 증가하는 추세라면 전반적으로 나라의 경제가 활성화되고 있으며 사람들이 여유로운 생활을 하고 있다는 의미이다. 그리고 실업률 또한 전반적인 경제 상황을 나타내는 변수로 활용 가능하다. 이런 경제 지표들이 사람들의 여가시간이나 일상생활에 영향을 미칠 가능성이 있기 때문에 드라마의 시청률에 간접적인 영향을 줄 것이라 판단하여 변수로 사용하였다. 단, 세 변수 모두 해당 드라마의 첫 방영일을 기준으로 전 분기와 해당 분기의 값을 사용하였다.

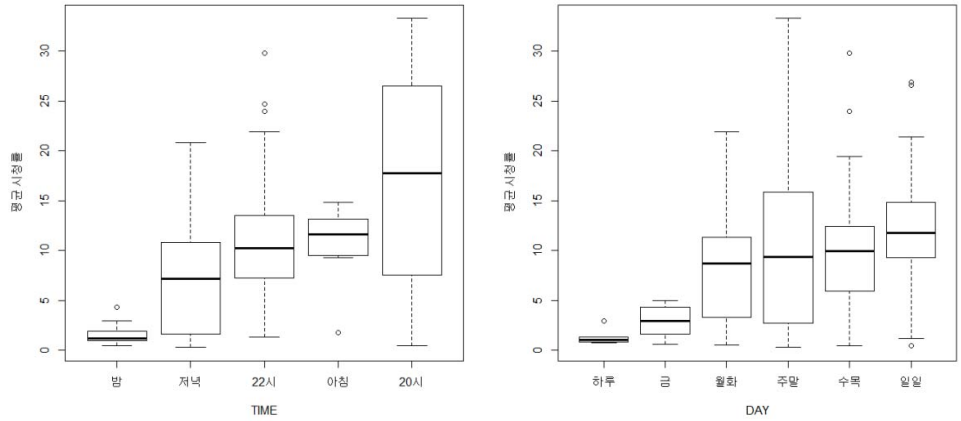

Figure 2.2. Box plots of drama ratings based on the time and day.

2.2.10. 날씨 날씨에 따라 사람들의 행동 패턴이 많이 좌우된다. 맑고 화창한 날씨라면 외출을 많이 하는 경향이 있는 반면 흐리거나 비가 내리는 등 날씨가 좋지 않을 때는 되도록 외출을 삼가는 경우가 많다. 외출을 하지 않는다면 가정에서 다른 일을 할 수도 있지만 TV 드라마를 볼 가능성이 있고 이런 이유로 시청률 예측에 영향을 미치는 요인으로 날씨를 고려하였다. 날씨에 대한 정보로는 평균기온, 강수량, 습도를 사용하였고 경제지표와 유사하게 해당 드라마의 첫 방영일을 기준으로 전 달과 해당 달의 값을 사용하였다.

2.2.11. 2차적 변수 검증된 웹사이트를 통해 모든 변수가 정의되었고, 분석을 위한 정형화된 형태의 자료가 구축되었다. 그런데 이 변수들을 그대로 사용하여 예측 모형을 도출한다면 높은 예측력을 보장할 수가 없다. 따라서 구축된 자료로부터 새로운 변수를 생성하거나 보정을 함으로써 예측력을 강화시키고자 한다. 이런 과정을 통해 연출자, 작가, 방송요일, 방송시작시간, 방송사 변수를 보정하거나 재범주화하였다. 이에 대한 설명은 아래와 같다.

연출자·작가

흥행 드라마를 제작했던 횟수를 수치화하여 만든 연출자, 작가 변수 사이에서는 명확한 선형관계가 나타나지 않아 GMM(Gaussian Mixture Model)을 이용하여 군집분석을 수행하였다. 그 결과 BIC를 최대화하는 최적 군집의 개수가 두 경우 모두 4개로 도출되었고, 이를 통해 새로운 변수를 생성한 뒤 실제 분석에 사용하였다.

방송요일·방송시간

드라마의 편성시간대를 나타내는 방송요일과 방송시간 모두 평균 시청률을 기준으로 하여 비슷한 값을 나타내는 범주들은 동일 범주로 묶었다. 이런 방식으로 다시 범주화한 결과는 다음과 같다. Figure 2.2을 보면 범주에 따른 시청률의 분포 차이를 확인할 수 있다. TIME 변수의 범주 중 '밤'은 23시 이후 방영되는 드라마를 의미하며 '저녁'은 18, 19, 21시에 방영하는 드라마를, '아침'은 아침드라마를 나타낸다. '저녁'과 '20시' 범주를 나눈 것은 두 그룹의 시청률 차이가 크기 때문이다. 그리고 DAY 변수는 드라마가 방영되는 요일을 나타내며 DAY 변수의 범주 중 '하루'는 일주일에 1번 방영하는 드라마를 의미한다. Figure 2.2에서 범주에 따른 시청률의 분포 차이에 비해 최솟값이 비슷한 이유는 모든 범주에서 전반적으로 낮은 시청률을 보이는 기타 채널의 영향 때문인 것으로 보인다.

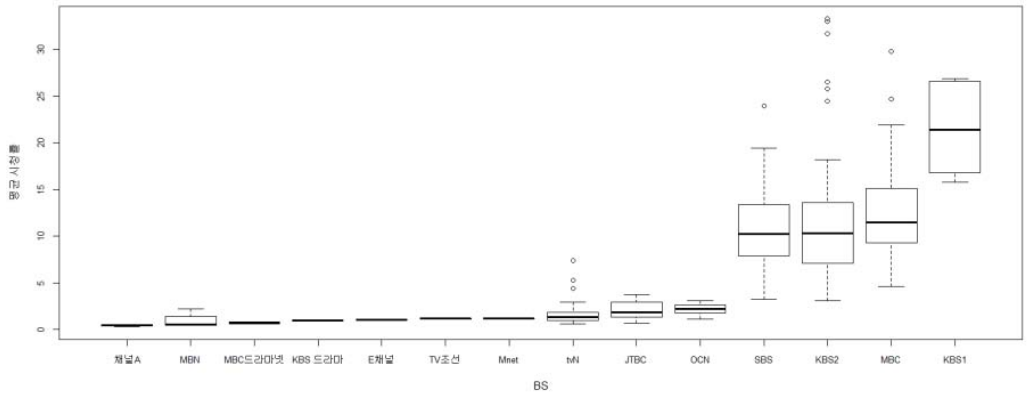

Figure 2.3. Box plots of drama ratings based on the broadcasting stations.

방송사

앞서 방송사 변수로 방송사의 이름을 이용하기로 하였다. 그런데 지상파 방송이 아닌 방송사의 경우 1개 또는 2개의 드라마만을 제작한 경우가 많아 분석에 어려움이 따르므로 각 방송사 이름 대신 OTHERS로 동일하게 범주화하였다. 실제로 기타 채널에 속하는 방송사들의 드라마 평균 시청률이 서로 매우 비슷하며 지상파 방송사의 드라마 평균 시청률과는 큰 차이를 보이므로 이러한 범주화 방식에 무리가 없을 것으로 판단된다. 각 방송사들에 따른 시청률 분포는 Figure 2.3과 같다. Figure 2.3에서 왼쪽 9개의 범주는 기타 채널, 오른쪽의 4개의 범주는 지상파 방송사에 해당한다. 따라서 분석 시 사용할 방송사 변수는 OTHERS, SBS, KBS2, MBC, KBS1의 범주를 갖는 변수이다.

이렇게 생성된 변수들을 정리한 결과는 다음의 Table 2.1과 같다. 모형 적합 시 주의할 점은 이 표의 설명변수들 중 AT1과 SC1은 드라마 방영 후에 알 수 있는 변수이므로 Model 1에는 들어가지 않고, Model 2에서 Model 1의 반응변수였던 my2가 설명변수로 사용된다는 점이다.

2.3. 결측치 처리 방법

데이터를 수집하는 과정에서 일부 변수에 대한 결측치가 발생하였다. 경제지표인 GDP, GNI, 그리고 실업률의 경우 해당 분기가 지나야 통계청에서 그 평균값을 제공하기 때문에 조사 시점인 2015년 6월에는 2015년도 2분기에 방영된 드라마에 대한 경제지표 값을 얻을 수 없었다. 월평균 강수량, 습도, 온도 등 날씨 관련 변수도 마찬가지다. 그리고 동시간대 이전 드라마의 평균 시청률과 경쟁 드라마의 평균 시청률에 대한 정보 역시 드라마의 편성시간대 변경으로 인해 경쟁작이나 이전작이 존재하지 않는 경우 결측값이 발생하였다. 본 연구에서는 이렇게 발생한 결측치를 처리하기 위해 k-nearest neighbor 대치법을 사용하였다. 이 방법은 결측치와 가장 가까운 거리에 있는 k개의 이웃 개체들을 이용하여 결측된 값을 대체하는 방법으로, R 패키지 “cluster”를 이용할 수 있으며 연속형, 범주형 등 변수의 유형에 제한 없이 거리 계산이 가능하다는 장점이 있다. 본 연구에서는 유클리드 거리를 사용하였으며 결측치가 발생한 변수가 모두 연속형이므로 해당 변수와 상관관계가 높은 10개의 변수를 이용하여 이들의 중앙값으로 결측치를 대체하였다.

3. 회귀모형

본격적인 드라마 시청률 예측 분석에 앞서 본 연구에서 사용할 여러 가지 통계적 기법을 간략히 소개하

Table 2.1. Description of variables

Variable	Description	Type
Input variables		
NUM	드라마의 총 방영 횟수	
DIR.score	연출자 등급(1–4)	
WRI.score	작가 등급(1–4)	
ACT_NUM	주연배우의 수	
ACT1	주연배우의 평균 방송활동 개수	
ACT2	주연배우의 평균 영화 출연 횟수	
ACT3	주연배우의 평균 수상 경력	
ACT4	주연배우의 시청률 순위 TOP100 드라마 평균 출연 횟수	
AT0	방영 전 3개월간 해당 드라마 관련 기사 개수	
AT1	방영 후 1주일간 해당 드라마 관련 기사 개수	
SC0	방영 전 3개월간 해당 드라마 검색량	
SC1	방영 후 1주일간 해당 드라마 검색량	
UR0	첫 방영일 기준 이전 달 실업률	
UR1	첫 방영일 기준 해당 달 실업률	numerical
GDP0	첫 방영일 기준 이전 분기 국내 총생산	
GDP1	첫 방영일 기준 해당 분기 국내 총생산	
GNI0	첫 방영일 기준 이전 분기 국민 총소득	
GNI1	첫 방영일 기준 해당 분기 국민 총소득	
TEMP0	첫 방영일 기준 이전 달 평균 기온	
TEMP1	첫 방영일 기준 해당 달 평균 기온	
PT0	첫 방영일 기준 이전 달 평균 강수량	
PT1	첫 방영일 기준 해당 달 평균 강수량	
HM0	첫 방영일 기준 이전 달 평균 습도	
HM1	첫 방영일 기준 해당 달 평균 습도	
COMP	경쟁 작품들의 초반(1–5회) 평균 시청률	
PRE	이전작의 평균 시청률	
YEAR	드라마의 방영년도(1–5)	
TIME	방송시간(아침, 저녁, 밤, 20시, 22시)	
DAY	방송요일(금, 일일, 주말, 하루, 월화, 수목)	
BS	방송사 종류(KBS1, KBS2, MBC, SBS, OTHERS)	categorical
AGE	연령제한 유무(유, 무)	
ORIG	원작 유무(유, 무)	
Response variables		
my2	해당 드라마의 초반 2회 평균 시청률	numerical
y	해당 드라마의 전체 평균 시청률	

고자 한다. 이 때 선형회귀모형을 제외한 모든 방법론에서 모수 최적화 단계를 거친 후 얻어진 최적 모수를 이용하여 모형을 적합하였다. 다시 말하면, train data에서 예측오차(MSE)를 최소화 하는 모수를 이용하여 모형을 적합하였다.

3.1. 선형회귀모형(Linear Regression)

회귀분석에 있어서 가장 기본적인 방법으로서 반응변수와 설명변수간의 관계를 선형모형으로 적합하는 방법이다. 모형 적합이 쉽고 모형에 대한 해석도 쉽기 때문에 많이 쓰이는 방법이다. 특히 설명변수의

수가 많을 때 stepwise method (Venables과 Repley, 2003)를 사용하여 AIC를 기준으로 한 최적 모형을 손쉽게 찾을 수 있고 통계적으로 유의하지 않은 변수를 모형에서 제거할 수 있다.

3.2. LASSO 회귀모형(LASSO Regression)

LASSO 회귀모형은 회귀계수의 절댓값의 합이 어떤 양의 상수보다 작다는 제약조건 하에서 회귀모형의 계수를 추정한다. 따라서 LASSO 회귀계수는 다음과 같이 정의된다.

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

위 식의 t는 shrinkage factor로서 LASSO의 튜닝 모수에 해당하며, R 패키지 "lars"를 이용하여 LASSO 회귀모형을 적합하였다. 앞서 살펴본 stepwise method는 특정 변수를 모형에 넣거나 빼는 불연속적인(discrete) 변수 선택법이므로 분산이 커질 가능성이 있어 예측오차를 큰 폭으로 감소시키지는 못한다. 이에 비해 LASSO는 회귀계수의 크기를 점차 줄여나가는 연속적인(continuous) 변수 선택법으로, 약간의 편의를 허용하는 대신 분산을 크게 줄여 전체 예측오차를 감소시킬 수 있다. LASSO의 회귀계수는 특정 값보다 작아지는 순간 0이 되어 모형에 유의한 변수들이 자동적으로 선택된다.

다음으로 살펴볼 방법론들은 나무모형을 기반으로 한 방법론이다. 의사결정나무 (Brieman 등, 1984)는 전체 데이터를 몇 개의 소집단으로 구분하는 분류 및 예측 기법으로, 구축과정을 나무 형태로 도식화 한다. 나무모형은 매우 많은 장점을 가지고 있다. 빠르게 적합되며, 설명력이 좋고 설명변수의 유형에 구애받지 않으며 결측치 또한 잘 다룬다. 하지만 이런 장점에도 불구하고 single tree는 예측력이 낮으며 계층적인 구조 때문에 불안정하다는 결정적인 단점이 존재하는데, 이를 보완할 수 있는 방법론이 바로 앙상블이다 (Park 등, 2011). 앙상블이란 single tree로 적합한 여러 개의 예측모형들을 결합하여 하나의 최종 예측모형을 만드는 방법으로 배깅 (Breiman, 1996)과 부스팅 (Friedman, 2002)이 이에 해당한다. 배깅의 대표적인 방법으로는 랜덤 포레스트가 있으며 부스팅 방법으로는 그래디언트 부스팅이 있다. 이런 앙상블 기법은 여러 개의 모형을 결합해야 하므로 적합 속도와 설명력은 조금 떨어지지만 예측 정확도의 측면에서 다른 어떤 방법론과 비교해도 경쟁력 있는 결과를 얻을 수 있다. 이들에 대한 설명은 다음과 같다.

3.3. 랜덤 포레스트(Random Forest)

자료로부터 복원추출을 허용하여 붓스트랩 표본을 여러 번 추출하고 각 표본에 대해 나무모형을 적합한다. 이렇게 추출된 표본들은 서로 독립일 수가 없고 이들의 상관관계가 높을수록 분산을 감소시키는 배깅의 효과는 줄어들게 된다. 이 때 적합된 나무모형들 간의 상관관계를 줄여 배깅의 효과를 극대화하고자 제시된 방법이 랜덤 포레스트 (Hastie 등, 2009)로, 나무모형을 적합할 때 전체 설명변수 p개를 사용하지 않고 $m\,(< p)$개의 설명변수를 사용하는 것이 핵심이다. 각 노드에서 선택되는 설명변수의 개수인 m이 바로 랜덤 포레스트의 튜닝 모수에 해당하며, 이런 방식으로 나무모형의 불안정성을 줄일 수 있다. 본 연구에서는 R 패키지인 "randomForest"를 이용하여 랜덤 포레스트 모형을 적합하였다.

3.4. 그래디언트 부스팅(Gradient Boosting)

그래디언트 부스팅은 손실 함수의 기울기를 바탕으로 여러 개의 약한 예측 모형을 단계적으로 생성한 뒤 앙상블 방법으로 결합하여 강한 예측력을 가지도록 하는 부스팅 기법을 의미한다. 특히 그래디언트 부스팅 알고리즘은 약한 예측력을 가지는 나무모형들을 결합하는 데 적용되어 그 예측력을 높일 수 있다.

Table 4.1. MSE of each model (Model 1)

	Linear	LASSO	Random Forest	Gradient Boosting
MSE	7.0640 (1.2640)	6.3165 (1.2419)	5.6835 (1.3880)	6.7742 (1.5110)

나무모형을 이용한 그래디언트 부스팅 알고리즘은 참고 문헌 (Friedman, 2002; Ridgeway, 2012)에 자세히 나와 있으며, R 패키지 "gbm"을 이용하여 모형을 적합할 수 있다.

4. 분석 결과

이번 장에서는 앞서 설명한 통계적 방법들을 이용하여 드라마 초기 시청률 예측 모형(Model 1)과 평균 시청률 예측 모형(Model 2)을 적합하고 시청률에 영향을 미치는 변수가 무엇인지 파악하고자 한다. 적합된 모형의 예측력을 공정하게 비교하기 위해 전체 자료의 70%을 training data, 나머지 30%을 test data로 임의로 나누어 test data에서의 MSE(Mean Squared Error)를 계산하는 과정을 100번 반복하였다. 이 때 MSE의 평균이 가장 작은 모형이 예측력이 가장 좋은 모형이 된다. 앞으로의 분석은 우선 각 모형에 대한 MSE를 계산하여 예측력을 비교한 후, 최적 모형에서의 중요변수를 도출하여 이 변수들이 반응변수인 시청률에 어떤 영향을 미치는지 알아보는 것을 중심으로 진행하겠다.

4.1. 드라마 초반 시청률 예측 모형(Model 1)

초반 시청률을 예측하기 위하여 AT1(방영 후 1주일간 드라마 관련 기사 개수), SC1(방영 후 1주일간 드라마 검색량)을 제외한 모든 변수를 사용하여 선형회귀분석을 하였고 stepwise 과정을 통해 결정계수 R^2의 값이 0.8271인 모형이 적합되었다. 설명력이 80% 이상이므로 모형이 데이터를 잘 설명하고 있다고 볼 수 있다. 3장에서 설명한 4가지 모형에 대한 MSE를 계산한 결과는 Table 4.1과 같다. 랜덤 포레스트 모형이 가장 좋은 예측력을 보이지만, LASSO 회귀모형 또한 비슷한 예측력을 보이며 선형모형이기 때문에 설명력이 매우 좋으므로 최적 모형인 랜덤 포레스트 모형을 살펴보기 전에 선형모형에 대한 해석을 먼저 하고자 한다.

선형모형인 Linear 회귀모형과 LASSO 회귀모형에서 도출된 회귀계수 및 유의성은 Table 4.2에서 살펴볼 수 있다. 단, Linear 회귀모형의 stepwise 과정과 최적 LASSO 회귀모형에서 모두 유의하지 않은 것으로 나타난 변수는 표에서 제외되었다. Linear 회귀모형은 각 회귀계수의 p-value를 제시해주므로, 매우 유의한 것으로 도출된 변수들을 중심으로 살펴보자. 유의수준 0.001 하에서 유의한 변수는 TIME(방송시간), BS(방송사), PRE(이전작의 평균 시청률)이었다. 선형모형의 주요변수들을 통해 알 수 있는 사실은 다음과 같다. 초반 시청률은 방송시간과 방송사의 영향을 받는데, 저녁 8시와 KBS1에서 방영하는 드라마의 시청률이 타 시간대, 타 방송사의 드라마에 비해 초반 시청률이 높은 것으로 드러났다. 또한 해당 시간대에 같은 방송사에서 방영했던 이전 드라마의 평균 시청률 역시 드라마의 초기 시청률에 중요한 영향을 미치는 변수로 도출되어 인기 드라마의 후속작일수록 높은 초기 시청률을 나타낸다고 할 수 있다.

이제 최적 모형인 랜덤 포레스트 모형에 대해서 살펴보자. 랜덤 포레스트와 같은 tree-based 모형은 선형모형과 달리 특정 변수에 대한 회귀계수 값이 주어지지 않기 때문에 어떤 설명변수에 따른 반응변수의 변화를 쉽게 알 수 없다. 따라서 특정 변수가 반응변수인 초반 시청률에 어떠한 영향을 미치는 지를 알아보기 위해 해당 변수를 제외한 다른 모든 변수들을 일정 수준으로 두고 해당 변수의 값을 변화시키면서 초반 시청률의 변화를 살펴보았다.

Table 4.2. The result of linear regression using stepwise procedure and LASSO regression (Model 1)

	Model 1	
	Linear Regression	LASSO Regression
(Intercept)	20.174915***	7.799098
NUM	0.030053*	0.028904
DIR.score	0.310189′	0.359693
WRI.score		0.041906
ACT_NUM	−0.106337	−0.091331
ACT1		−0.009654
ACT2		−0.000605
ACT4	0.541048*	0.543534
SC0	0.009173**	0.007218
TIME22시	−3.718814***	−2.640786
TIME밤	−1.839456*	−0.997020
TIME아침	−2.718911*	−2.717199
TIME저녁	−3.750803***	−3.051015
DAY수목	0.656965	
DAY월화	0.772547	
DAY일일	−2.810880′	−2.169798
DAY주말	0.949978	0.468834
DAY하루	0.650545	
BSKBS2	−3.357367**	−0.160527
BSMBC	−3.053715**	−0.017656
BSSBS	−3.734622**	−0.685998
BSOTHERS	−8.731225***	−5.303632
AGE유		0.405114
UR0	−0.770494*	−0.523464
UR1	0.250557	0.179529
GDP1	−0.000022′	−0.000014
TEMP1		−0.022036
PT1		0.001170
HM1		0.005129
COMP	−0.088394′	−0.107215
PRE	0.363533***	0.387700
YEAR		−0.045772
ORIG유	0.863924′	0.838630

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

랜덤 포레스트 모형에서 주요 변수로 도출된 BS(방송사), PRE(이전작 평균 시청률), TIME(방송시간), SC0(방영 전 3개월간 드라마 검색량)에 대하여 그래프를 그려본 결과는 Figure 4.1과 같다. 지상파 채널에서의 초기 평균 시청률이 기타 채널에 비해 상대적으로 높으며 지상파 방송사 중에서는 KBS1이 가장 높은 시청률을 보였다. 그리고 이전작의 평균 시청률이 증가함에 따라 초기 평균 시청률이 증가하는 것을 볼 수 있으며, 저녁 8시에 방영되는 드라마의 초기 시청률이 다른 시간대에 방영되는 드라마보다 높음을 알 수 있다. 마지막으로 SC0 변수의 경우 검색량이 매우 작은 부분을 제외하고는 해당 드라마 방영 전 드라마에 대한 검색량이 많을수록 초기 시청률도 높아지는 것을 볼 수 있다. 이 결과는 앞서 살펴본 회귀모형에서의 결과와 일치한다.

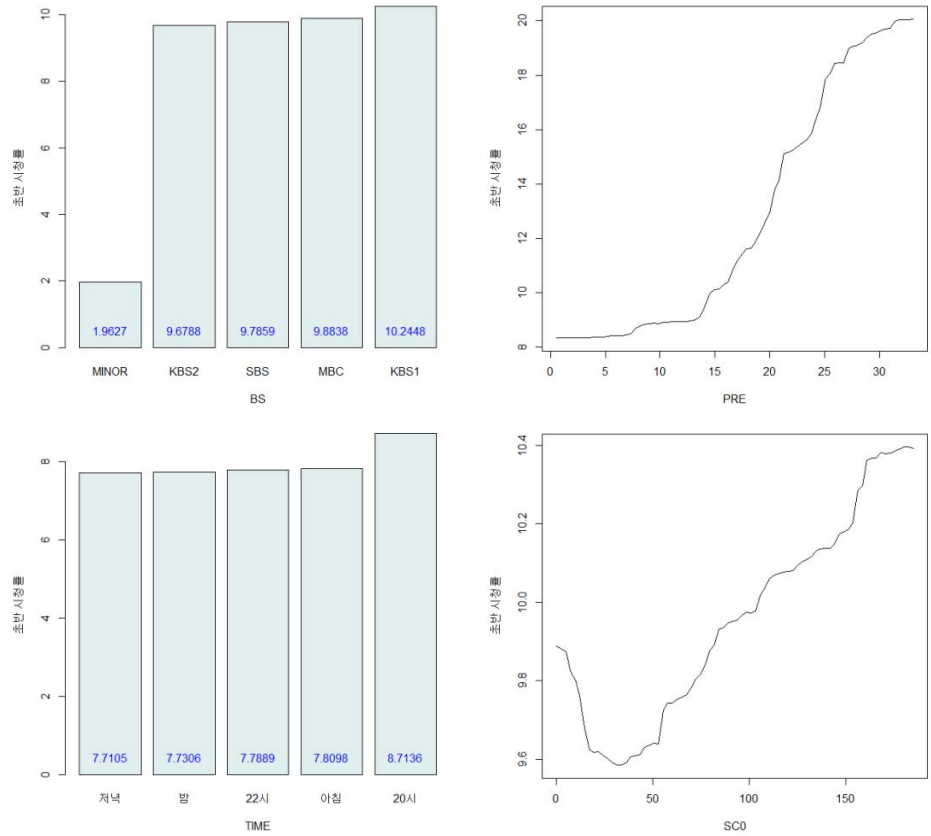

Figure 4.1. Drama ratings according to BS, PRE, TIME, and SC0.

Table 4.3. MSE of each model (Model 2)

	Linear	LASSO	Random Forest	Gradient Boosting
MSE	8.4862 (1.4802)	7.6725 (1.2866)	7.9382 (1.3468)	8.7630 (1.9491)

4.2. 드라마 평균 시청률 예측 모형(Model 2)

이번 절에서는 드라마의 평균 시청률 예측을 하고자 한다. Model 2에서는 Model 1에서 사용한 변수를 포함하여 방영 후 트렌드를 반영한 변수 AT1(방영 후 1주일간 드라마 기사 개수), SC1(방영 후 1주일간 드라마 검색량)이 사용되었다. 또 Model 1의 반응변수였던 my2(1,2회 평균 시청률)가 설명변수로 사용되었다. 이 변수들을 사용하여 선형회귀분석에서 stepwise 과정을 거친 결과 R^2는 0.877로 설명력이 큰 모형이 적합되었고 Table 4.3에서 각 모형의 MSE를 비교해본 결과 LASSO 회귀모형의 예측력이 가장 좋았다.

우선 선형모형에서의 결과에 대해 살펴보자. 유의수준 0.001 하에서 SC1, my2 변수가 유의한 것으로 도출되었다. Table 4.2와 Table 4.4를 비교해보면 Model 1, 2에서 도출된 주요변수들 중 공통된 변수는 NUM, TIME이었다. 또 주요변수들의 차이점도 볼 수 있는데, Model 1에서는 방송사 및 방영 전의 트렌드가 고려된 변수들이 선택되었지만, Model 2에서는 초반 시청률과 방영 후의 트렌드가 고려된 변

Table 4.4. The result of linear regression using stepwise procedure and LASSO regression (Model 2)

	Model 2	
	Linear Regression	LASSO Regression
(Intercept)	13.046520*	9.341045
NUM	0.036321**	0.014131
DIR.score		0.116317
WRI.score	0.554775*	0.529055
ACT_NUM	−0.165247*	−0.107918
ACT1	0.054361*	0.033918
ACT4		−0.001468
AT1	−0.001017	−0.000746
SC1	0.044151	0.025354
TIME22시	−1.316839	
TIME밤	−2.724068	−0.815428
TIME아침	−0.573485	
TIME저녁	−1.925977	−0.222722
DAY수목	1.449138	0.333476
DAY월화	−0.191194	−0.308830
DAY일일	−2.087597	
DAY주말	1.813535	1.126670
DAY하루	1.182518	
BSOTHERS		−0.553676
AGE유	−1.553771	−1.108126
UR0	−1.084475	−0.282578
UR1	0.000208	0.177523
GNI0		−0.000017
TEMP1	0.053468	0.004355
PT0	−0.003661	
HM1		0.016801
COMP	−0.120098	−0.105496
ORIG유		0.076797
my2	0.982413	1.019800

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

수가 선택되었다. Table 4.4의 회귀계수들을 살펴보면, Model 1에서와 마찬가지로 드라마 총 방영횟수가 많을수록 드라마 평균 시청률이 높아지며 저녁 8시 드라마 평균 시청률이 타 시간대에 비해 높은 것을 알 수 있다. 또한 드라마 방영 후 1주일간 드라마 검색량이 많을수록 평균 시청률이 높아지는 것으로 나타났다. 가장 유의한 변수는 my2로, 드라마의 초반 2회 평균 시청률이 전체 시청률에 매우 큰 영향을 미치므로 방영 초기의 시청률이 드라마의 흥행여부에 매우 중요하게 작용한다고 할 수 있다.

다음으로 tree-based 모형의 결과에 대해 살펴보자. 랜덤 포레스트의 예측력 또한 최적모형인 LASSO 회귀모형과 비슷하게 좋았다. 랜덤 포레스트 모형에서 중요한 변수로 my2(드라마 초기 시청률), NUM(총 방영횟수), PRE(이전작의 평균 시청률), SC1(방영 후 1주일간 드라마 검색량), AT1(방영 후 1주일간 드라마 기사 개수)가 도출되었다. 선형모형을 포함해 공통적으로 선택된 중요변수는 my2, NUM, SC1, AT1으로 이들이 평균 시청률에 어떠한 영향을 미치는지를 그래프를 통해 살펴보았다. 그 결과는 Figure 4.2와 같다. 초반 시청률인 my2가 증가할수록 드라마 평균 시청률 역시 증가하였으며, NUM 변수의 그래프를 보면 약 50부작 이하인 드라마에 대해서는 총 방영횟수가 많아질수록 드라마 평

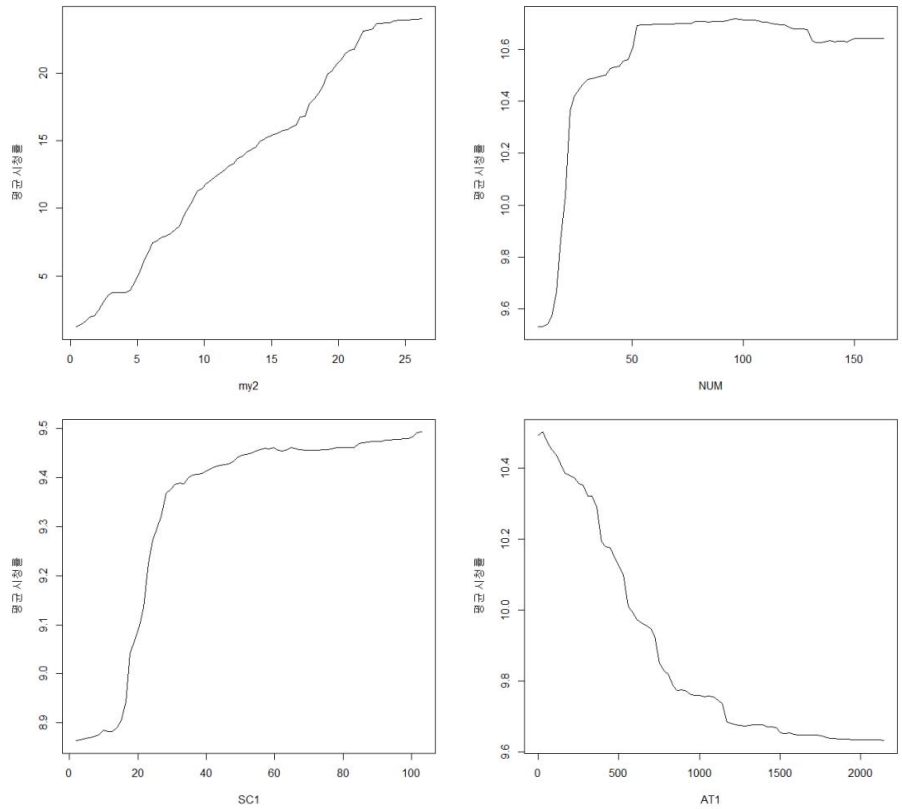

Figure 4.2. Drama ratings according to my2, NUM, SC1, AT1.

Table 4.5. R^2 of each model depending on air channel broadcasting

	지상파		기타 채널	
	Model 1	Model 2	Model 1	Model 2
R^2	0.6815	0.8089	0.5617	0.8357

균 시청률이 증가하는 추세를 보이지만 50부작 이상인 드라마의 경우 총 방영횟수에 따른 시청률 차이가 크지 않은 것으로 나타났다. 또 SC1이 증가할수록 드라마 평균 시청률이 증가하는 추세를 보이므로 인기 있는 드라마의 경우 방영 후 입소문 효과로 인해 많은 시청자들의 관심을 끌 수 있어 평균 시청률이 증가하는 것으로 보인다. 마지막으로 AT1 변수의 그래프를 보면 드라마 방영 후 1주일간 드라마 관련 기사 개수가 증가할수록 시청률이 감소하는 재미있는 결과를 볼 수 있는데, 이는 초반 시청률이 부진한 드라마의 경우 기사를 통한 프로모션을 더 적극적으로 하게 되기 때문인 것으로 추측된다.

4.3. 지상파 여부를 고려한 예측 모형

다음은 추가적으로 지상파 드라마 여부에 따른 시청률을 예측하고자 한다. Figure 4.3을 보면 지상파 드라마의 평균 시청률이 기타 채널에 비하여 월등히 높음을 볼 수 있다. 이에 지상파 드라마 여부가 시청률에 상당한 영향을 미칠 것이라고 생각되어 260개의 드라마를 지상파 여부에 따라 지상파 드라마

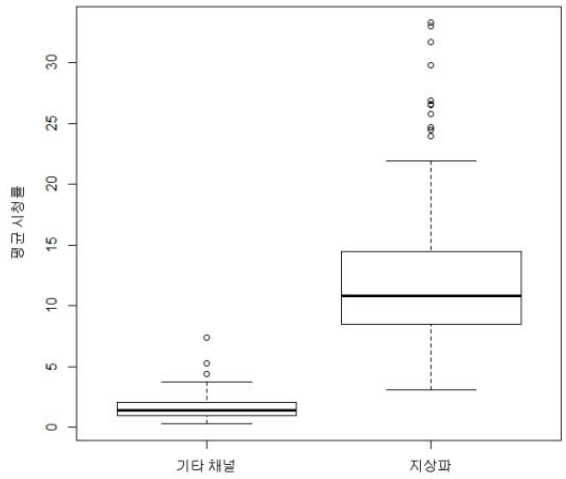

Figure 4.3. Box plots of drama ratings depending on air channel broadcasting.

Table 4.6. The important variables of each model

	지상파		기타 채널	
	Model 1	Model 2	Model 1	Model 2
Linear Regression	NUM, TIME, PRE	my2, SC1, TIME	ACT4, TEMP0, PT0	my2, SC1, COMP
Random Forest	PRE, TIME, DAY	my2, NUM, AT1	ACT4, GDP1, GNI1	my2, SC0, AT0
Gradient Boosting	PRE, TIME, SC0	my2, NUM, PRE	ACT4, PT0, ACT2	my2, AT1, SC1

194개와 기타 채널 드라마 66개로 나누어 각각 다른 모형을 적합하였다. 이 때 방송사 변수인 BS는 앞서 살펴본 모형에서 매우 유의한 변수로 도출되었고, 방송사를 나타내는 변수라는 점에서 지상파 여부와 동일한 의미를 내포하고 있어 모형 적합 시 제외하였다. 이렇게 가장 유의한 변수를 제외했을 때 각 모형에서 주요변수가 어떻게 변하는지도 살펴보고자 한다.

그 결과, 각 모형에 대한 R-square 및 중요변수는 각각 Table 4.5, Table 4.6과 같이 도출되었다. 지상파 여부를 고려한 모형을 적합하였을 때 R-square는 고려하지 않았던 경우보다 감소하였다. 이는 가장 유의하였던 방송사 변수를 제외한 후 모형을 적합한 결과라 볼 수 있다. 이어서 각 모형에서 도출된 주요변수를 살펴보자. 우선 Linear 회귀모형을 기준으로 선택된 중요변수에 대한 설명은 다음과 같다. 초기 시청률에 대하여 지상파와 기타 채널에서 선택된 중요변수는 각각 NUM(총 방영횟수), TIME(방송시간), PRE(이전작 평균 시청률) 그리고 ACT4(주연배우의 시청률 TOP100 드라마 평균 출연 횟수), TEMP0(드라마 방영 이전 달 평균 온도), PT0(드라마 방영 이전 달 평균 강수량)이었다. 이 중 음의 회귀계수를 갖는 변수는 TEMP0와 PT0로 드라마 방영 이전 달의 평균 온도가 낮고 강수량이 적을수록 초반 시청률이 증가하는 경향을 보였다. 범주형 변수인 TIME의 경우 저녁 8시에 방영하는 드라마의 초반 시청률이 타시간대의 드라마보다 유의하게 높았다. 다음으로 평균 시청률에 대하여 살펴보면, 지상파와 기타 채널에서 공통적으로 선택된 중요변수는 my2(드라마 초기 시청률)와 SC1(드라마 방영 후 1주일간 드라마 검색량)이며 지상파에서만 선택된 변수로는 TIME, 기타 채널에서는 COMP(경쟁작

의 평균 시청률)가 있었다. 이 중 음의 회귀계수를 갖는 변수는 COMP로 경쟁작의 평균 시청률이 높을수록 해당 드라마의 평균 시청률이 감소하였다. 그 외에 랜덤포레스트와 그래디언트 부스팅에서 선택된 중요변수는 Table 4.6에서 확인할 수 있다. 지상파 드라마와 기타 채널 드라마 모형의 주요변수를 비교해보면 지상파의 경우 방송시간, 총 방송횟수, 이전작 시청률 등 드라마 자체의 속성과 구조적 요인들이 유의한 변수로 도출된 반면 기타 채널 드라마의 경우 주연배우의 인기드라마 출연 횟수와 같은 배우 관련 요소와 평균 온도, 평균 강수량, GDP, GNI 등 날씨와 경제지표를 나타내는 변수들이 시청률에 유의한 영향을 미치는 것으로 나타났다. R-square를 기준으로 모형의 설명력을 비교한다면 지상파 여부를 고려하여 2가지 모형을 따로 적합했을 때의 설명력이 그렇지 않은 경우보다 다소 낮았지만, 지상파와 기타 채널 드라마에 대해 서로 다른 모형을 적합하고 각 경우의 주요변수를 도출하는 과정도 지상파 여부에 따른 차이점을 살펴볼 수 있다는 점에서 의미가 있다고 본다.

5. 결론

본 연구에서는 드라마와 직접적 관련이 있는 변수와 경제지표, 날씨 등 드라마 시청률에 간접적인 영향을 미칠 것으로 보이는 외부적 변수를 포함한 다양한 설명변수를 생성하여 국내 드라마 시청률을 예측하는 예측모형을 제시하고 시청률에 유의한 영향을 미치는 변수들을 도출하였다. 모형의 적합에는 선형회귀모형, LASSO 회귀모형, 랜덤 포레스트, 그래디언트 부스팅을 이용하였으며 모형 평가 지표로는 MSE를 이용하였다.

분석 결과는 다음과 같다. 우선 드라마의 초반 시청률을 예측한 첫 번째 모형에서 가장 중요한 요인들로 도출된 설명변수는 방송사, 이전작 평균 시청률, 방송시간, 드라마 방영 3개월 전 드라마 검색량 변수였다. 따라서 시청자들은 해당 방송 시간대의 이전 작품의 영향을 받는 경향이 있으므로 인기 드라마의 후속작일수록 초반 시청률이 높아질 수 있으며 어느 방송사에서 방송하는 드라마인지 또한 시청률에 크게 영향을 미친다는 사실을 알 수 있었다. 분석 결과 KBS1에서 방영하는 드라마의 시청률이 타 방송사의 드라마 시청률에 비해 유의하게 높았으며 지상파 방송사일수록 시청률이 높았다. 그리고 드라마 방영 전의 드라마 검색량도 초반 시청률에 중요한 영향을 미치는 변수로 도출되었다. 다음으로 드라마의 전체 평균 시청률을 예측하는 두 번째 모형에서는 드라마의 초반(1, 2회) 평균 시청률이 가장 중요한 요인으로 도출되었으며, 다음으로는 드라마의 총 방영횟수와 드라마 방영 후 1주일간 검색량과 기사 개수가 중요한 변수로 도출되었다. 두 모형을 비교해보면, 시청자들은 어떤 드라마를 시청하기 시작할 때 이전작, 방송시간, 방송사, 방영 전 드라마 검색량 등 드라마의 구조적 요소나 입소문 효과에 큰 영향을 받으며 드라마가 방영되기 시작한 이후에는 드라마 초반 시청률과 방영 이후의 홍보효과가 많은 시청자들을 끌어들이는 데 큰 영향을 미치는 것으로 보인다. 추가적으로 두 모형에서 공통적으로 시청률에 큰 영향을 미쳤던 방송사 변수를 제외한 후 지상파 여부에 따라 드라마를 나누어 지상파 드라마와 기타 채널 드라마에 대해 다른 모형을 적합해보았다. 그 결과, 지상파 드라마의 시청률은 드라마의 구조적 요인들에 큰 영향을 받는 반면 기타 채널 드라마의 경우에는 주연배우와 날씨, 경제지표의 영향을 많이 받는 것으로 나타났다.

드라마 시청률 예측을 위해 추가적인 설명변수를 고려할 때 드라마 제작비와 같은 금전적 요인이 적지 않은 영향을 미칠 것이라 생각되었지만 국내 제작사들의 보안으로 정보를 수집하는 데 제약이 있었다. 하지만 제작비는 유명 작가의 원고료나 유명 배우의 출연료 등과 높은 상관관계를 보일 것이므로 본 연구에서 사용한 설명변수들로 충분히 보완이 가능한 부분인 것으로 생각된다. 본 연구에서 제시한 시청률 예측 모형의 R-square는 약 8–90%로 높은 값을 보였다. 이 모형들이 앞으로 방영될 드라마의 시청률을 예측하는 데 어느 정도의 가이드라인을 제시해 줄 수 있으리라 생각된다.

References

Bae, J. (2005). An analysis on the factors in drama ratings - focusing on the drama attributes and audience factors, *Korean Journal of Broadcasting and Telecommunication Studies*, **19**, 270–309.

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

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

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Chapman and Hall, New York.

Cohen, J. (2002). Television viewing preferences: Programs, schedules, and the structure of viewing choices made by Israeli adults, *Journal of Broadcasting & Electronic Media*, **46**, 204–221.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning*, **20**, 273–297.

Friedman, J. (2002). Stochastic gradient boosting, *Computational Statistics & Data Analysis*, **38**, 367–378.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, Springer, New York. USA.

Park, C., Kim, Y., Kim, J., Song, J. and Choi, H. (2011). *Datamining using R*, Kyowoo, Seoul.

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.

Ridgeway, G. (2012). Generalized Boosted Models: A guide to the gbm package.

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

Venables, W. N. and Ripley, B. D. (2003). *Modern Applied Statistics with S*, Springer, New York.

국내 드라마 시청률 예측 및 영향요인 분석

강수연[a] · 전희정[a] · 김지혜[a] · 송종우[a,1]

[a]이화여자대학교 통계학과

(2015년 7월 10일 접수, 2015년 7월 23일 수정, 2015년 7월 26일 채택)

요 약

최근 상업방송의 도입과 채널의 다양화로 국내 드라마 시장의 시청률 경쟁이 심화되었다. 이에 시청률에 대한 실증적인 연구의 필요성이 대두되고 있다. 본 연구의 목적은 다양한 데이터마이닝 기법을 이용하여 최근 방송시장의 변화를 고려한 국내 드라마 시청률 예측 모형을 제시하고 시청률에 유의한 영향을 미치는 변수들을 도출하는 데 있다. 모형 적합 시 선형회귀모형, LASSO 회귀모형, 랜덤 포레스트, 그래디언트 부스팅 등과 같은 다양한 분석 방법을 고려하였다. 이 때 드라마 방영 전 알 수 있는 기본 정보들만을 고려하여 드라마의 초반 시청률을 예측하는 모형을 적합한 후 방영 초기의 여론을 고려한 평균 시청률 예측 모형을 적합하였다. 그 결과 드라마 초반 시청률은 방송사, 방송시간, 드라마 방영 이전 드라마 관련 검색량 등 드라마의 구조적 요인과 임소문 효과의 영향을 크게 받으며, 평균 시청률은 드라마 초반 시청률과 드라마 방영 이후 드라마 관련 검색량 등 방영 초기의 여론에 큰 영향을 받는 것으로 나타났다.

주요용어: 드라마 시청률, 선형회귀모형, LASSO 회귀모형, 랜덤 포레스트, 그래디언트 부스팅, 주요변수

이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2013R1A1A2012817).

[1]교신저자: (120-750) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과.
E-mail: josong@ewha.ac.kr