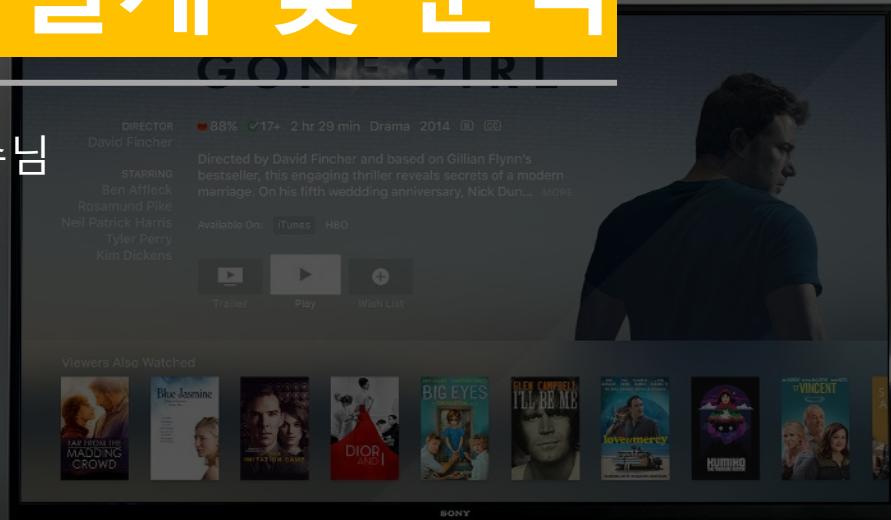


RACOI + 네이버 뉴스 감정 기반 시청률 예측 모델 설계 및 분석

미디어산업과 데이터과학입문 :: 장병희 교수님

2013313217 조상연



Introduction



- 인터넷에서 유명한 것이 과연 TV시청률로 직결될까?
 - 소위 어그로가 TV 시청률을 높혀줄까?
 - 나이대별, 성별별로 어떤 특성의 차이가 있을까?

연구 문제

1. 인터넷 **버즈량** (뉴스, 커뮤니티)과 **시청률** 간의 연관성
2. 프로그램에 대한 **감정과 시청률** 간의 연관성
3. 인터넷 **버즈량과 감정을 통한 시청률 예측**
4. **나이대 별, 성별 별 예측 특성 확인**

선행 연구

표 7. 평가지표와 시청률 및 온라인 시청과의 관계

		시청률	온라인 성과				
			VOD 시청자 수	PC 실시간	PC VOD	모바일 실시간	모바일 VOD
미디어 버즈	언론보도	.07	.51**	.47**	.64**	.52**	.53*
	동영상 수	.37**	.24**	.21**	.26**	.28**	.25**
시청자 버즈	댓글 수	-.00	.55**	.57**	.69**	.57**	.60**
	게시글 수	-.05	.49**	.54**	.60**	.53*	.52**
	동영상조회 수	.09	.44**	.47**	.72*	.70**	.70**

* $p < .05$. ** $p < .01$.

강명현 (2018). 방송 콘텐츠 가치평가 지표의 속성 및 시청률
과의 관계 연구. *한국방송학보*, 32(3), 5-30

상관계수, ANOVA 분석에 그친 기준 연구

인터넷 버즈와 가구 시청률의 관계가 뚜렷하지 않다.

연구 모형

인터넷 게시글 수

인터넷 댓글 수

영상 시청 수

인터넷 뉴스 기사 수



개인 시청률

영상 클립 수

감정 표현 수

채널 / 방송 요일

자료 수집



RACOI 주간 버즈량 및 시청률 데이터
(2018년 1월 ~ 2019년 3월)

버즈량
뉴스 기사수, 동영상 수, 게시글 수,
댓글 수, 영상시청수

시청률
10~40대별 시청률
가구 / 개인 시청률
남녀 시청률

총 프로그램 / 데이터
579개 / 9106 rows



네이버 뉴스 감정 데이터
(2018년 1월 ~ 2019년 3월)

3만 4천 건의 뉴스 감정 데이터 크롤링
(네이버 뉴스 기준 프로그램 별 상위 5개)



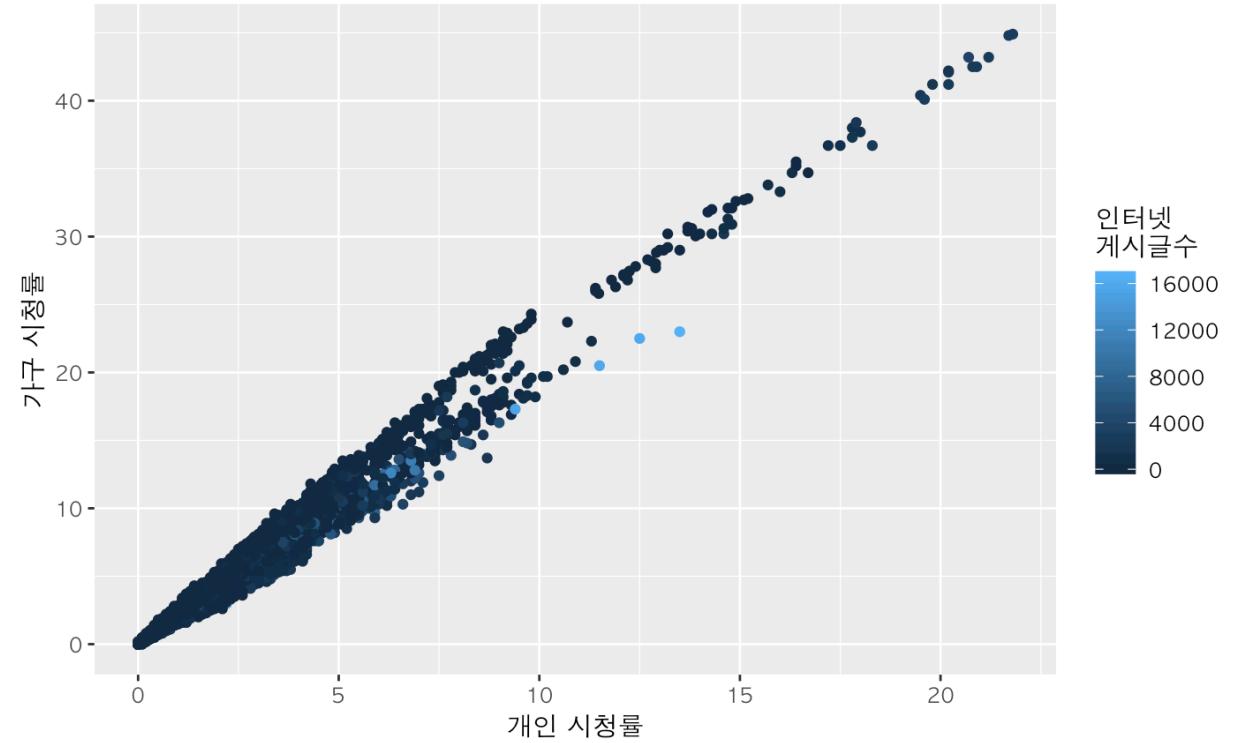
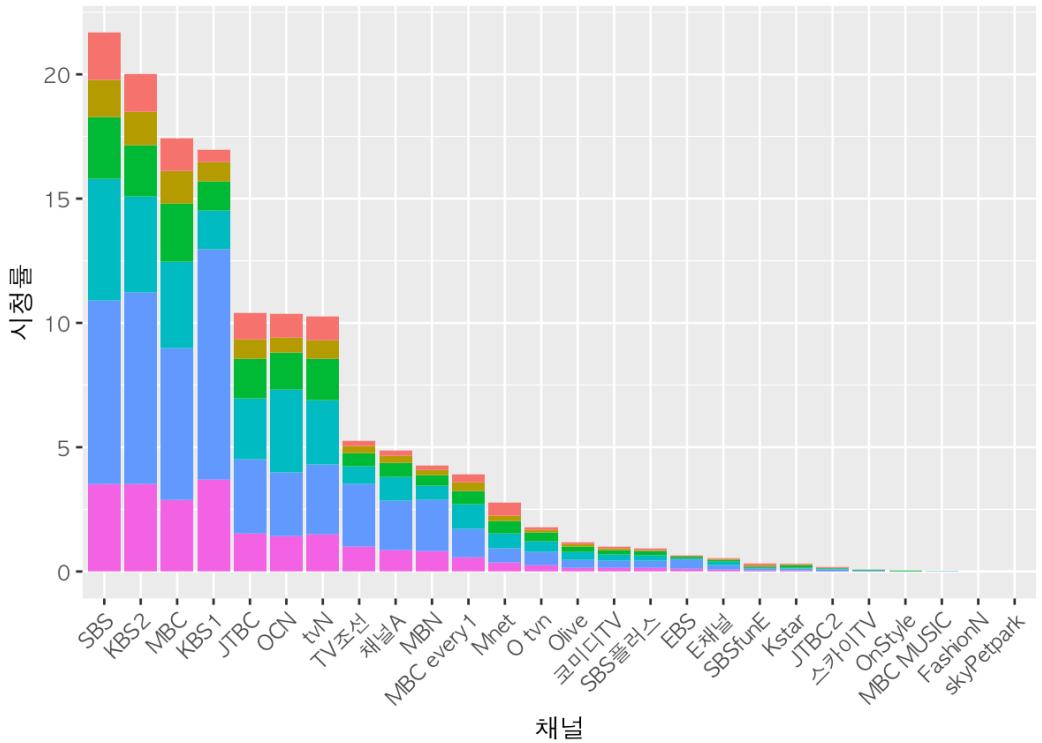
Like Warm Sad Angry Want

데이터 분석 알고리즘

- Correlation
- ANOVA
- Multiple Linear Regression – 간단한 예측 모형을 위한 선형 회귀
- Decision Tree – 예측 모델 발전 및 트리 시각화
- Random Forest – 높은 정확도의 예측 모델 ($n\text{tree} = 300$, $m\text{try} = 6$)

분석 결과

1) 탐색적 데이터 분석



분석 결과

1) 탐색적 데이터 분석

2018년 1월 이후 연령대 별 평균 시청률 Top 10

pj_seq	pjname	10대	a10
<int>	<fcctr>		<dbl>
1267	황금빛 내인생		10.44
10454	SKY 캐슬		6.44
82	무한도전		6.07
10050	슬기로운 감빵생활		5.60
10123	같이 살래요		5.54
10347	하나뿐인 내편		5.37
825	다시쓰는 육아일기–미운 우리 새끼		4.84
10087	윤식당 시즌2		4.83
93	* 나혼자산다		4.59
10276	미스터 션샤인		4.48

pj_seq	pjname	20대	a20
<int>	<fcctr>		<dbl>
1267	황금빛 내인생		7.83
10347	하나뿐인 내편		5.88
10123	같이 살래요		4.87
82	무한도전		4.65
10454	SKY 캐슬		4.57
10580	세상에서 제일 예쁜 내 딸		4.25
825	다시쓰는 육아일기–미운 우리 새끼		4.10
10022	돈꽃		3.90
93	* 나혼자산다		3.89
10087	윤식당 시즌2		3.86

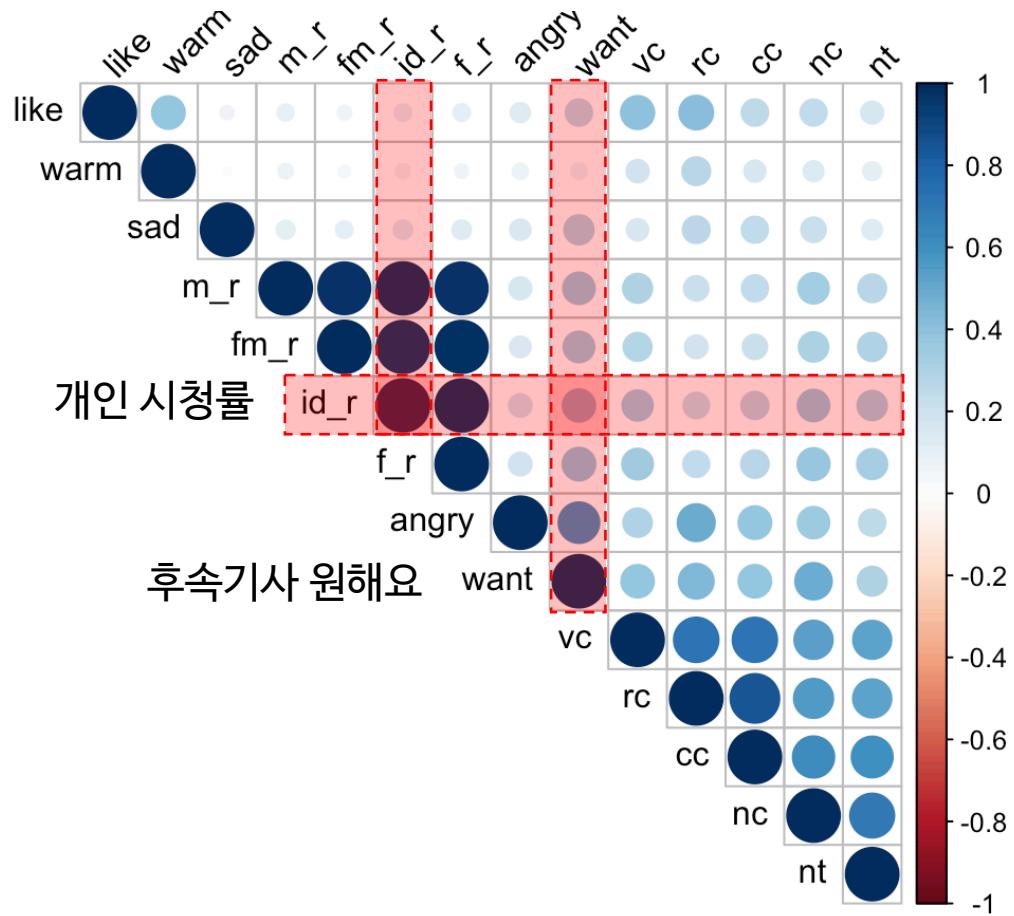
pj_seq	pjname	30대	a30
<int>	<fcctr>		<dbl>
1267	황금빛 내인생		10.31
10347	하나뿐인 내편		8.24
82	무한도전		8.07
10087	윤식당 시즌2		6.81
825	다시쓰는 육아일기–미운 우리 새끼		6.61
93	* 나혼자산다		6.54
10123	같이 살래요		6.53
10050	슬기로운 감빵생활		6.33
10450	신서유기6		6.20
10276	미스터 션샤인		6.14

pj_seq	pjname	40대	a40
<int>	<fcctr>		<dbl>
1267	황금빛 내인생		22.85
10123	같이 살래요		13.80
10347	하나뿐인 내편		12.91
10276	미스터 션샤인		12.08
10087	윤식당 시즌2		11.69
10454	SKY 캐슬		11.49
825	다시쓰는 육아일기–미운 우리 새끼		11.36
10541	열혈사제		11.36
10050	슬기로운 감빵생활		10.03
10458	알함브라 궁전의 추억		9.86

분석 결과

2) 상관 분석

```
res <- cor(df[,c('fm_r','id_r','f_r','m_r','nc','nt','rc',
               'cc','vc','like','angry','warm','sad','want')])
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



ANOVA Result

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
nc	1	4992	4992	1310.865	< 2e-16	***
nt	1	267	267	70.210	< 2e-16	***
rc	1	20	20	5.347	0.0208	*
cc	1	20	20	5.121	0.0237	*
vc	1	1014	1014	266.245	< 2e-16	***
rp_count	1	109	109	28.736	8.53e-08	***
reactions	1	59	59	15.618	7.82e-05	***
like	1	319	319	83.849	< 2e-16	***
angry	1	7	7	1.748	0.1861	
warm	1	20	20	5.214	0.0224	*
sad	1	1837	1837	482.436	< 2e-16	***
Residuals	7829	29813	4			

분석 결과

3) 다중 선형 회귀 분석

Train Data
2018년 1월~12월
총 6000 Rows

Test Data
2019년 1월~3월
총 1841 Rows

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.324 on 5951 degrees of freedom
Multiple R-squared:  0.6196,    Adjusted R-squared:  0.6165
F-statistic: 201.9 on 48 and 5951 DF,  p-value: < 2.2e-16

[1] 0.5854488
```

R-Squared: 0.6196
Adjusted R-squared: 0.6165

Test Result

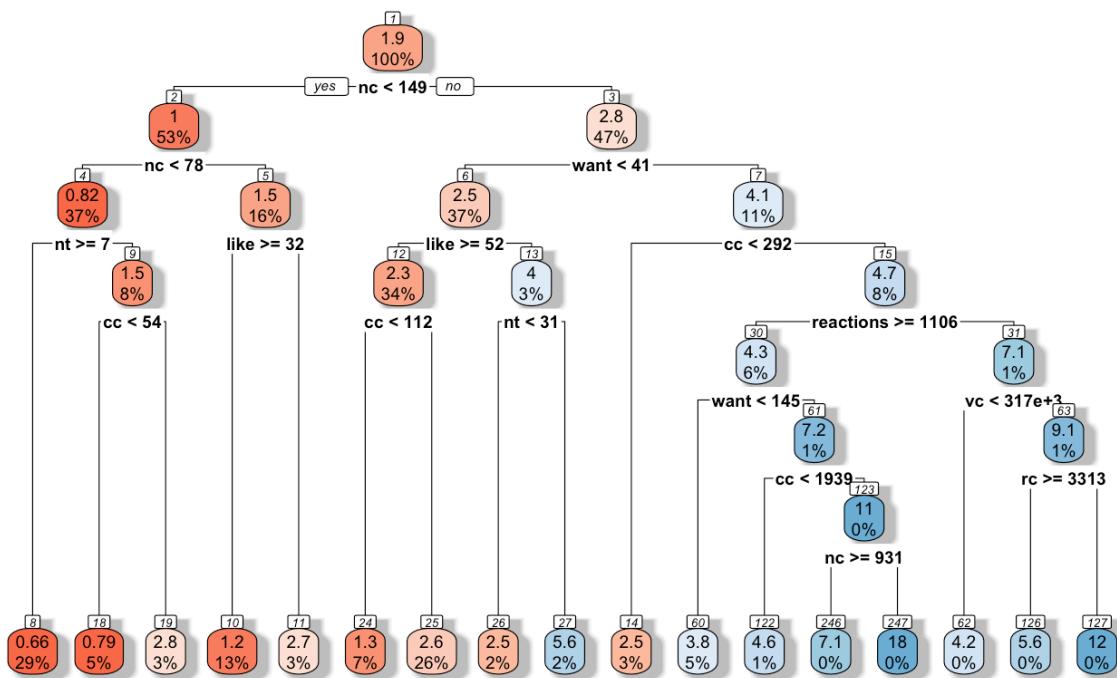
Predict R-squared
0.585

RMSE
1.574

Accuracy
40.0%

분석 결과

4) Decision Tree



Test Result

Predict R-squared

0.6942

RMSE

1.352

Accuracy

44.0%

분석 결과

5) Random Forest_발전 및 비교

파라미터 조정을 통한 최적의 값 도출

```
Call:  
randomForest(formula = id_r ~ nc + nt + rc + cc + vc + rp_count +  
service_day, data = df_train, ntree = 300, mtry = 6)  
Type of random forest: regression  
Number of trees: 300  
No. of variables tried at each split: 6  
  
Mean of squared residuals: 0.3956104  
% Var explained: 91.34
```

Predict R-squared

0.84

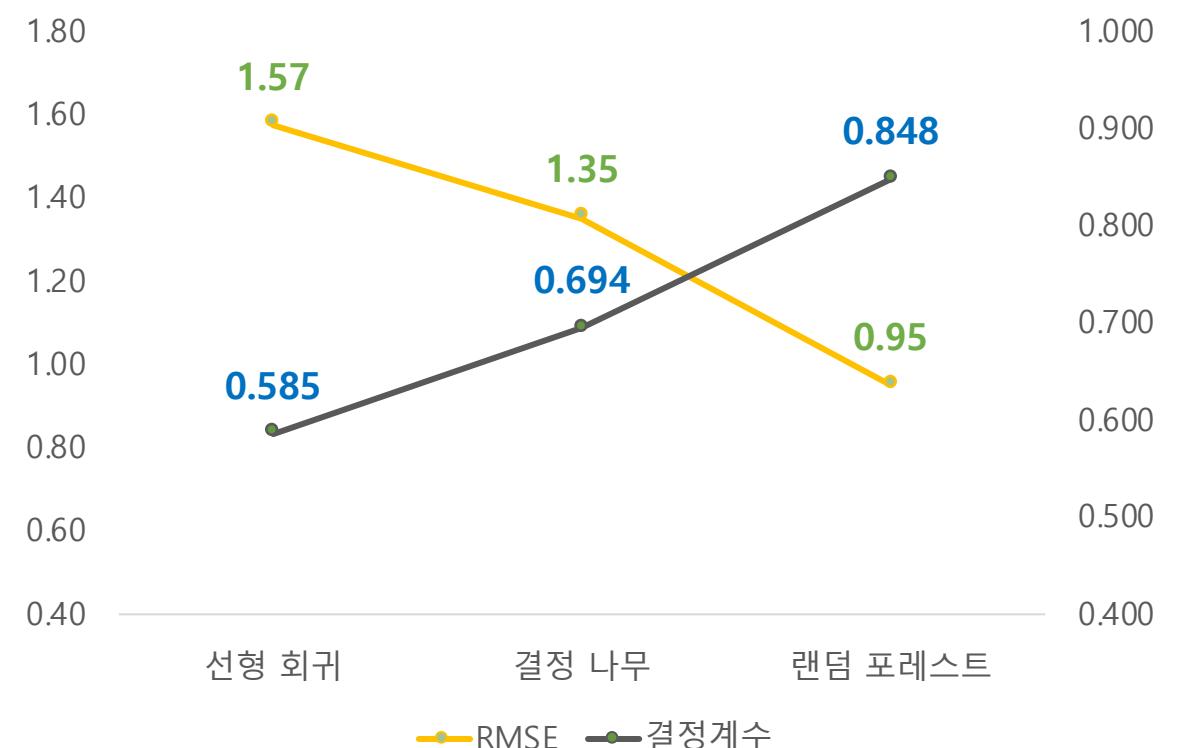
RMSE

0.95

Accuracy

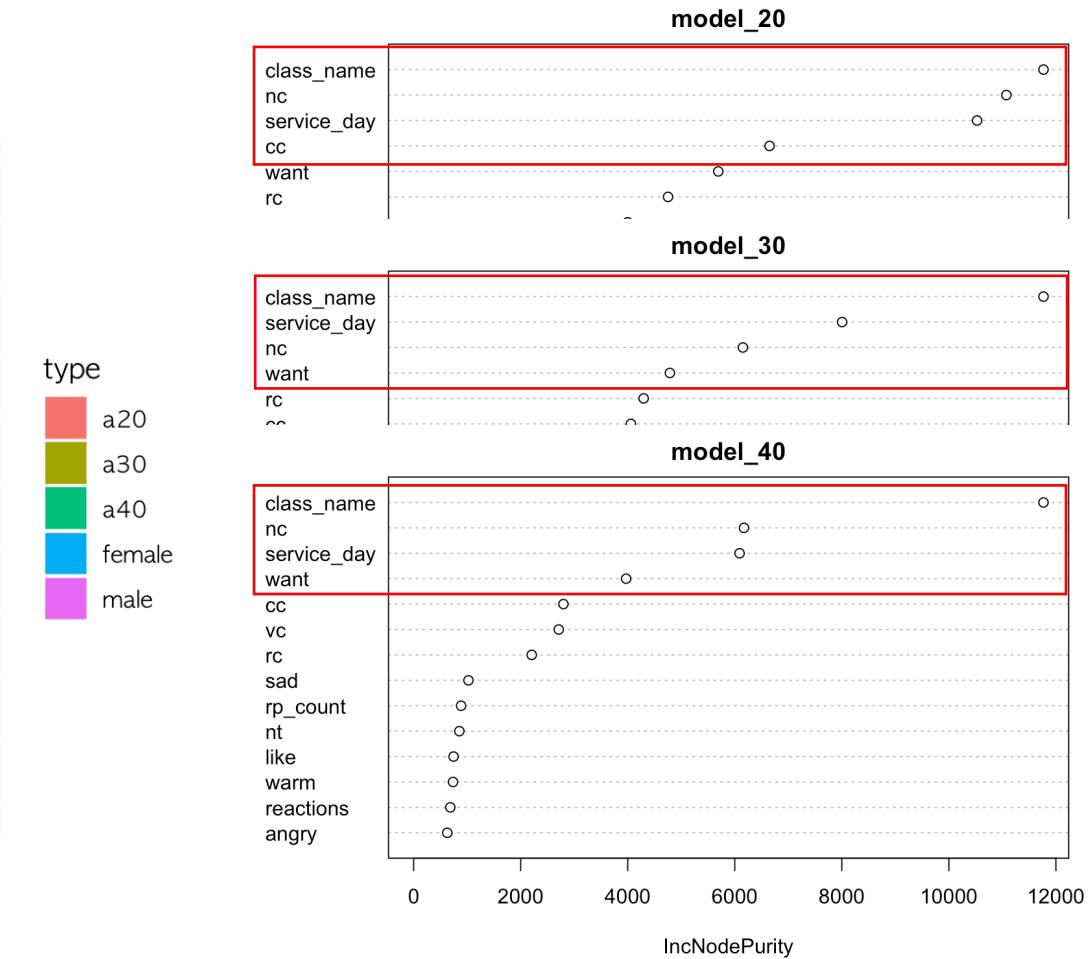
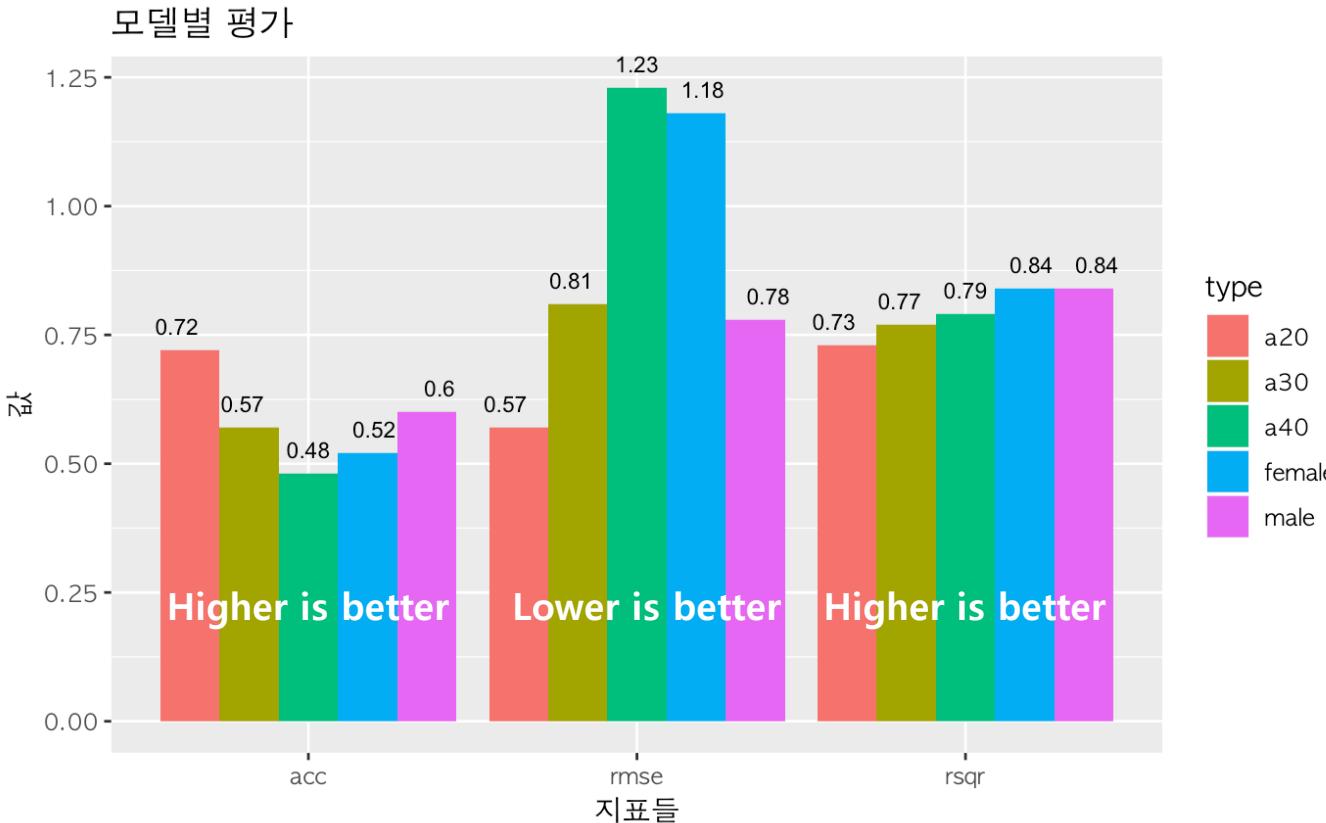
57.3%

모델별 결정계수 (R-Squared)



분석 결과

5) Random Forest_인구특성별 모델 생성



시사점

1. 인터넷 버즈량 (뉴스, 커뮤니티)과 시청률 간의 연관성

- ✓ 뉴스 기사 수가 상관성을 보인다.
- ✓ 그외 인터넷 버즈량과 시청률 간의 큰 상관성을 보이진 않는다.

2. 프로그램에 대한 감정과 시청률 간의 연관성

- ✓ 가장 연관성을 보이는 지표는 “후속 기사 원해요” 이다.
- ✓ 좋아요나 싫어요는 시청률과의 상관성이 거의 없다.

3. 인터넷 버즈량과 감정을 통한 시청률 예측

- ✓ 랜덤 포레스트를 이용한 모델이 가장 좋은 성능을 보인다.
- ✓ 채널과 요일이 미치는 영향이 너무 크다.

4. 나이대별, 성별별 예측 특성 확인

- ✓ 20대와 40대는 30대 보다 요일에 구애 받지 않는다.
- ✓ 20대가 다른 나이대에 비해 인터넷 버즈량을 통한 시청률 예측이 훨씬 더 잘 맞는다.

2013313217 조상연

