Section2 Project

발표자: 노희섭

상황: 30대 남자를 대상으로 하는 TV 마케팅

목표: TV CF를 넣을 채널/시간대 결정

방법: 시청률 회귀 분석

참조: Suyeon Kang, Heejeong Jeon, Jihye Kim and Jongwoo Song, (2015)
A Study on Domestic Drama Rating Prediction,
The Korean Journal of Applied Statistics Vol. 28, p933~949

목차

1. 데이터 전처리

Step 1

데이터 기초현황 파악
데이터 결합

>>

Step 2

분석에 용이하도록
특성 공학 진행

>>

Step 3

훈련/검증/테스트
데이터 분리

1. 데이터 전처리

Step 1: 데이터 기초 현황 파악

데이터 출처:
한국문화정보원 문화 빅데이터 플랫폼

사용한 데이터:
K-드라마 프로그램, 채널, 시청률 등 TV 콘텐츠 데이터
2021년 9월 ~ 12월
(From.TNmS)

9월	10월	11월	12월
4509 * 28	4819 * 28	4711 * 28	5315 * 28
합쳐서 훈련/검증 데이터로 사용 ☞ 14039 * 28			테스트 데이터

1. 데이터 전처리

Ste

데이
한글

사용
K-뉴
20
(Fr

9월
45
합계

순서	컬럼영문명	컬럼한글명
1	BRDCST_DE	방송일자
2	BRDCST_END_DE	방송종료일자
3	CHNNEL_NM	채널명
4	PROGRM_BEGIN_TIME	프로그램시작시간
5	PROGRM_END_TIME	프로그램종료시간
6	PROGRM_NM	프로그램명
7	PROGRM_DC	프로그램설명
8	BRDCST_TME_NM	방송회차명
9	PROGRM_BRDCST_AREA_NM	프로그램방송지역명
10	BRDCST_TIME	방송시간
11	PROGRM_GENRE_LCLAS_NM	프로그램장르대분류명
12	PROGRM_GENRE_MLSFC_NM	프로그램장르중분류명
13	PROGRM_GENRE_SCLAS_NM	프로그램장르소분류명
14	MALE_4_9YO_WTCHNG_RT	남자4_9세시청율
15	MALE_N10S_WTCHNG_RT	남자10대시청율
16	MALE_N20S_WTCHNG_RT	남자20대시청율
17	MALE_N30S_WTCHNG_RT	남자30대시청율
18	MALE_N40S_WTCHNG_RT	남자40대시청율
19	MALE_N50S_WTCHNG_RT	남자50대시청율
20	MALE_N60S_ABOVE_WTCHNG_RT	남자60대이상시청율
21	FEMALE_4_9YO_WTCHNG_RT	여자4_9세시청율
22	FEMALE_N10S_WTCHNG_RT	여자10대시청율
23	FEMALE_N20S_WTCHNG_RT	여자20대시청율
24	FEMALE_N30S_WTCHNG_RT	여자30대시청율
25	FEMALE_N40S_WTCHNG_RT	여자40대시청율
26	FEMALE_N50S_WTCHNG_RT	여자50대시청율
27	FEMALE_N60S_ABOVE_WTCHNG_RT	여자60대이상시청율
28	CST_CN	출연진내용

1. 데이터 전처리

Step 2: 특성공학 진행

BRDCST_D	BRDCST_E	CHNNEL_I	PROGRM_	PROGRM_	PROGRM_	PROGRM_	BRDCST_T	PROGRM_	BRDCST_T	PROGRM_	PROGRM_	PROGRM_	MALE_4_9	MALE_N1	MALE_N2	MALE_N3	MALE_N4	MALE_N5	MALE_N6	FEMALE_4	FEMALE_N	FEMALE_N	FEMALE_N	FEMALE_N	FEMALE_N	FEMALE_N	CST_CN
20210901	20210901	KBS1	203112	205919	속아도꿈결		98회	전국	2807	드라마&옽드라마	일일연속=		1.91121	3.1503	2.09781	3.07116	3.49764	5.7883	17.46663	1.01454	1.30062	1.51997	3.11875	4.5539	12.46363	18.90274	금종화역:최정우,강모란역:박준금
20210901	20210901	KBS2	92259	95100	속아도꿈결		97회	전국	2801	드라마&옽드라마	일일연속=		0.26859	0.32676	0.22904	0.50181	0.3768	1.9705	1.25558	0	0.55388	0.33854	0.13986	0.56691	2.02778	1.94509	금종화역:최정우,강모란역:박준금
20210901	20210901	KBS2	195053	202630	빨강구두		32회	전국	3537	드라마&옽드라마	일일연속=		2.43516	3.77405	2.08174	4.12124	5.31345	6.81132	18.65608	2.94364	2.28178	2.07667	4.97292	4.92582	14.83816	22.40254	민희경역:최명길,김진아역:소이현
20210901	20210901	MBC	85223	92144	두번째남편		16회	전국	2921	드라마&옽드라마	일일연속=		0.25966	1.06909	0.32441	0.38374	1.17022	1.58663	1.24675	0.20506	0.83963	0.20008	0.78139	0.97325	1.14983	1.78425	봉선화역:엄현경,윤재민역:차서원
20210901	20210901	MBC	190240	193206	두번째남편		17회	전국	2926	드라마&옽드라마	일일연속=		0.28723	0.7632	0.3151	1.9432	2.34736	1.795	5.27036	0.17365	0.7918	0.41742	1.87828	1.31816	4.54236	5.83274	봉선화역:엄현경,윤재민역:차서원
20210901	20210901	SBS	124042	135232	홍천기		2회	서울/경기,	11150	드라마&옽드라마	미니시리즈		0.94897	0.18164	0.20663	0.54833	1.43833	0.96496	0.8326	0.14105	0.27994	0.18484	0.39075	0.76853	0.95263	1.45086	하람역:안효섭,홍천기역:김유정,윤
20210901	20210901	OBS	182112	192544	이산		46회	전국	10432	드라마&옽드라마	주간연속=		0	0	0.03877	0.00157	0	0.04354	0.19661	0	0.0889	0.16056	0	0.06578	0.34101	0.36025	정조이산역:이서진,성송연역:한지
20210901	20210901	tvN	20000	24537	빈센조		3회	전국	4537	드라마&옽드라마	미니시리즈		0	0	0	0.23451	0.61305	0	0.45459	0	0	0	0.38065	0.00291	0.02888	0.33026	빈센조역:송중기,홍차영역:전여빈

컬럼 이름 수정 / 분석에 불필요한 데이터 삭제 ☞ 정확한 분석

1. 데이터 전처리

Step 2: 특성공학 진행(컬럼 수정)

BRDCST_DE, BRDCST_END_DE, PROGRM_DC
등 20가지 컬럼 삭제

MALE_N20S_WTCHNG_RT ☞ m_20
PROGRM_BEGIN_TIME ☞ time
등 8가지 컬럼 이름 수정

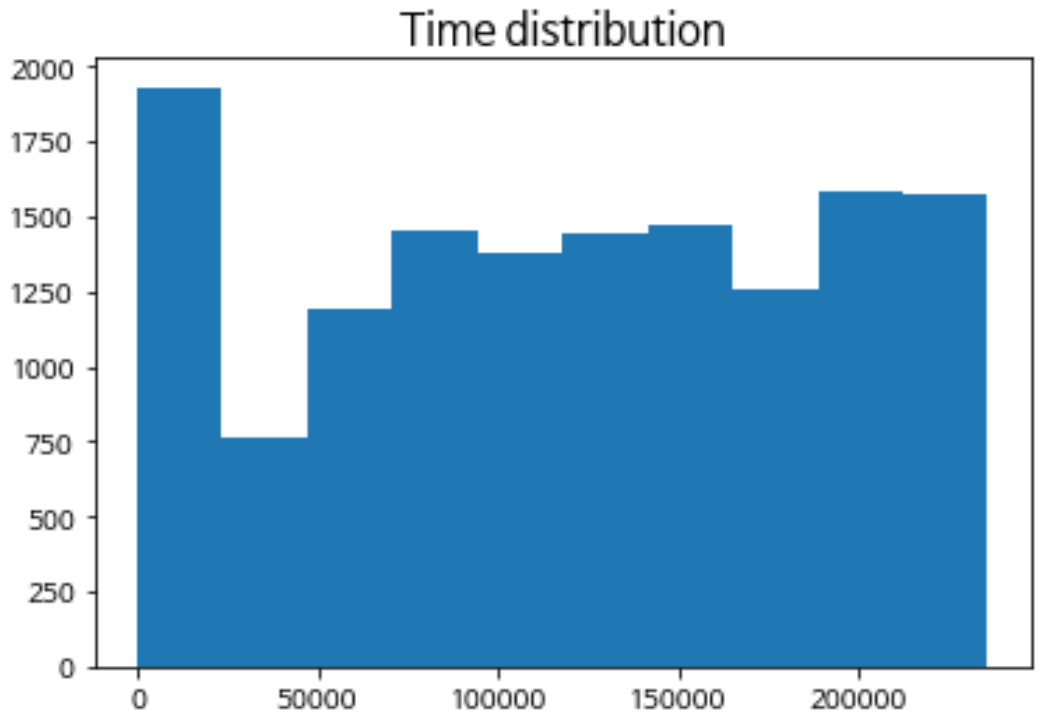
	chanel	time	name	genre	m_20	m_30	f_20	f_30
0	KBS1	G	속아도꿈결	일일연속극	2.09781	3.07116	1.51997	3.11875
1	KBS2	D	속아도꿈결	일일연속극	0.22904	0.50181	0.33854	0.13986
2	KBS2	G	빨강구두	일일연속극	2.08174	4.12124	2.07667	4.97292
3	MBC	C	두번째남편	일일연속극	0.32441	0.38374	0.20008	0.78139
4	MBC	G	두번째남편	일일연속극	0.31510	1.94320	0.41742	1.87828

1. 데이터 전처리

Step 2: 특성공학 진행(시간대 변경)

PROGRM_BEGIN_TIME	PROGRM_END_TIME
203112	205919
92259	95100
195053	202630
85223	92144
190240	193206
124042	135232
182112	192544
20000	24537
30503	43238
152833	164507
170542	182319
225430	2237
94010	110350
110510	122350
122510	134350
134510	152350
152510	164850
175010	190850
191010	202850
203010	220850
20000	21610
22921	32309

>>

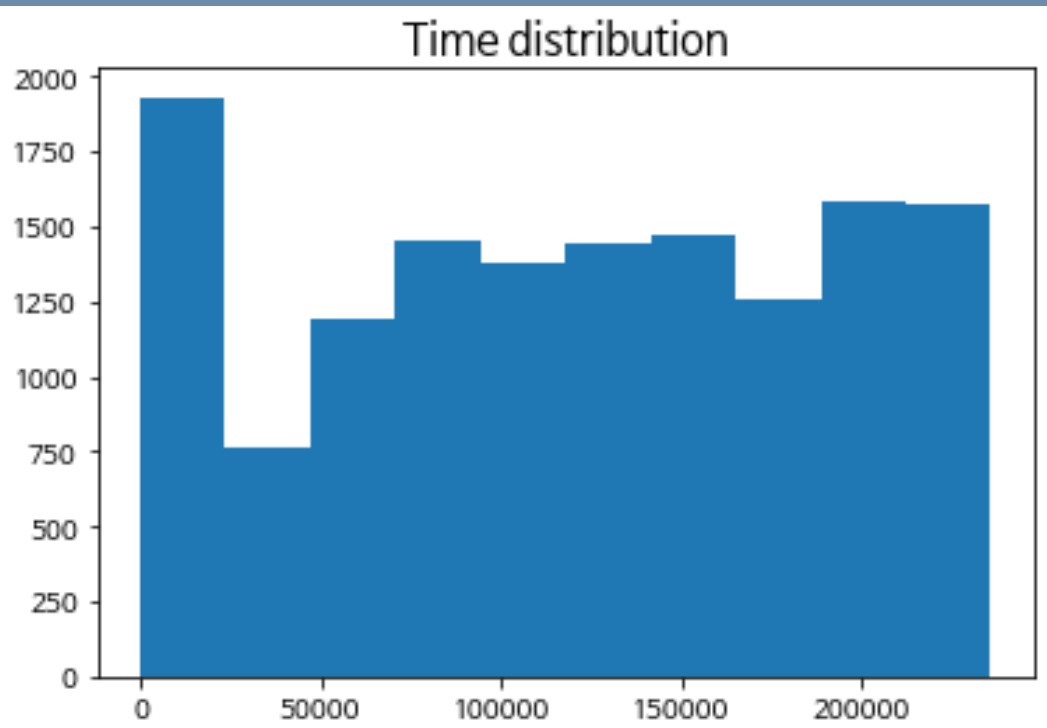


1. 데이터 전처리

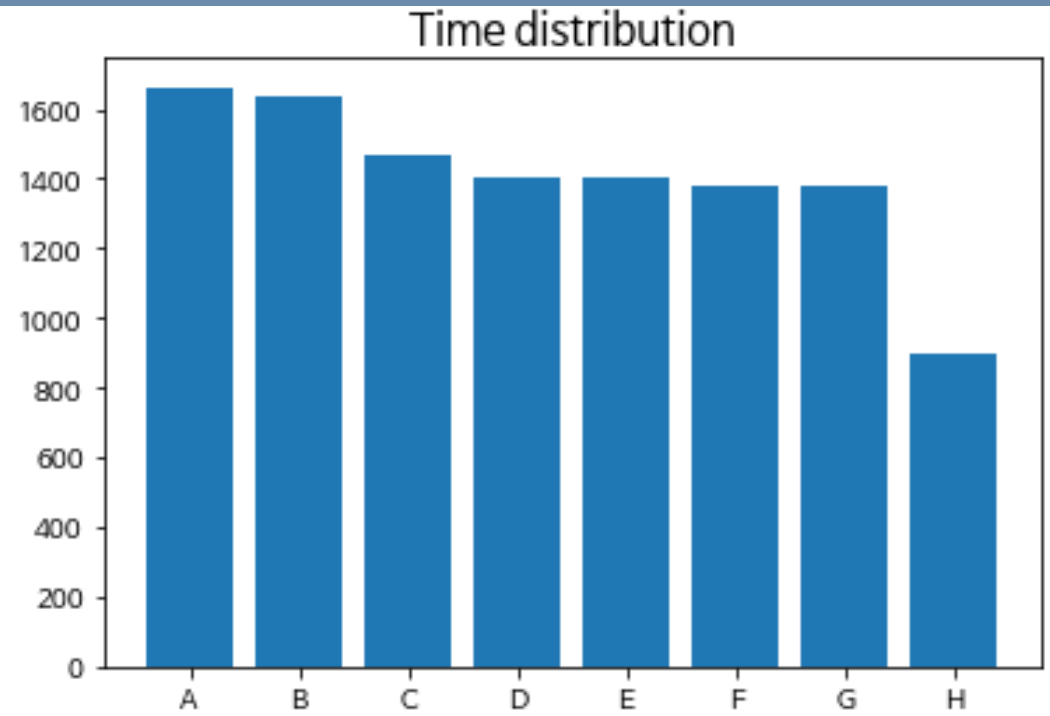
Step 2: 특성공학 진행(시간대 변경)

00:00 ~ 03:00 ~ 06:00 ~ 09:00 ~ 12:00 ~ 15:00 ~ 18:00 ~ 21:00 ~24:00

A B C D E F G H



>>

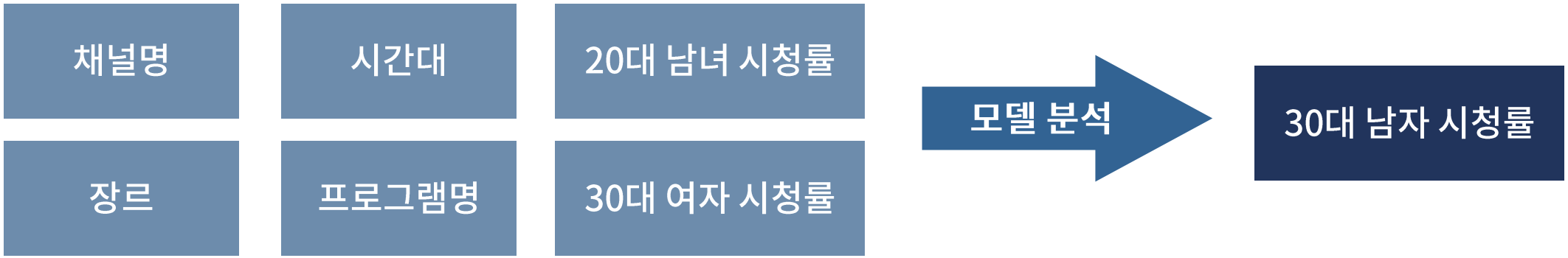


1. 데이터 전처리

Step 3: 훈련/검증/테스트 데이터 분리

회귀 분석의 목표(타겟): ‘m_30’
Why? 30대 남자가 많이 보는 채널, 시간대를 알아내기 위해

훈련 데이터		검증 데이터		테스트 데이터	
독립 변수	타겟	독립 변수	타겟	독립 변수	타겟
11231 * 7	11231	2808 * 7	2808	5315 * 7	5315



2. 모델 설정 및 평가

Model 1

BaseLine: 기준모델

평가

>>

Model 2

RandomForestRegression:
랜덤포레스트회귀모델

RandomizedSearchCV:
최고의 성능을 내는 모델 찾기

비교 및 평가

>>

Model 3

MultipleLinearRegression:
다중선형회귀모델

비교 및 평가

>>

Model 4

XGBoostRegression:
XG부스트회귀모델

RandomizedSearchCV:
최고의 성능을 내는 모델 찾기

비교 및 평가

2.모델설정및평가

Model 1:Base Line 기준모델

기준 모델이란?
: 모델의 성능을 평가하는 기준이 되는 모델

평가지표:

r2: 모델의 설명력, 1에 가까울수록 설명력 높음

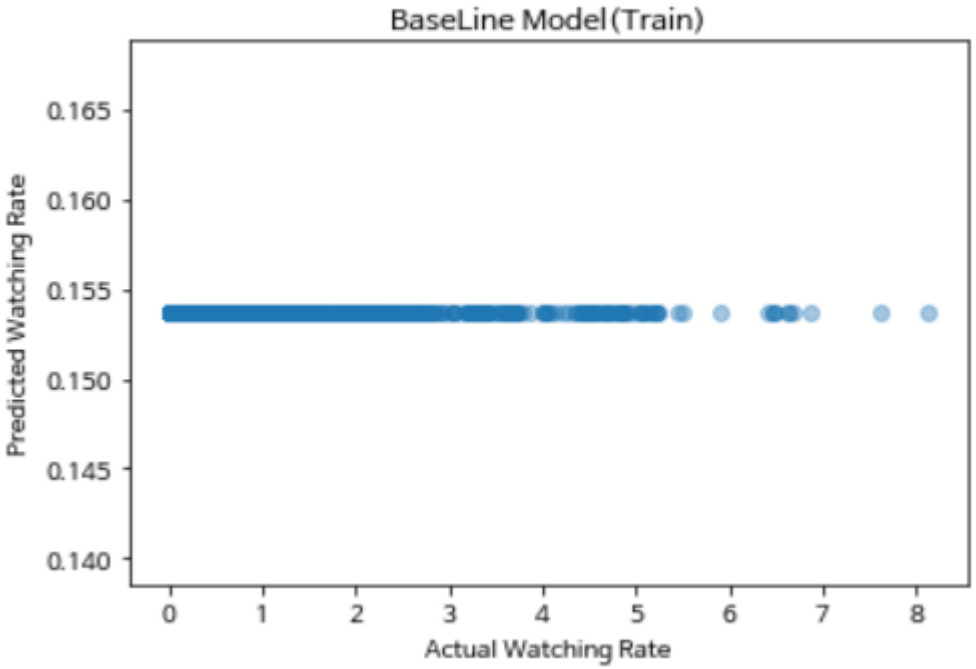
MSE: 모델의 오차, 0에 가까울수록 정확도 높음

훈련데이터 타겟 값의 **평균**
Baseline = 0.1537

2.모델설정및평가

Model 1:Base Line 기준모델

```
BaseLine_훈련 세트 r2: 0.0000000000
BaseLine_훈련 세트 MSE: 0.2696117100
```



되는 모델

록 설명력 높음

록 정확도 높음

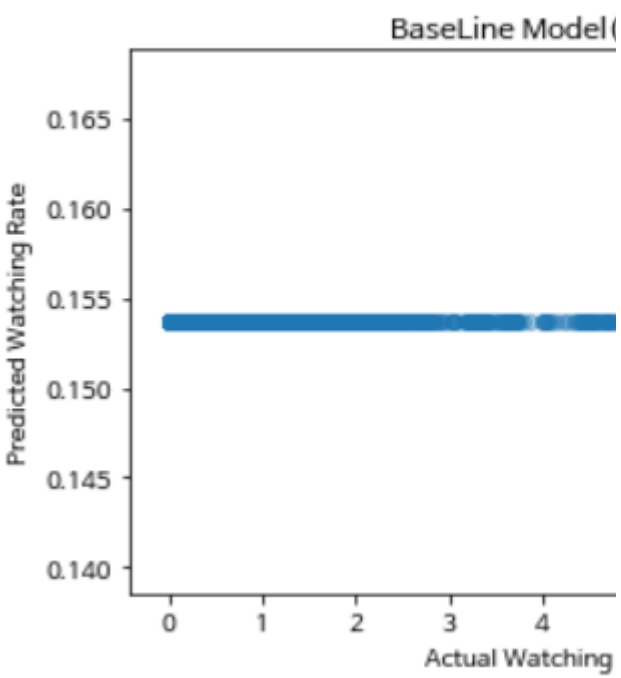
훈련데이터 타겟 값의 **평균**
Baseline = 0.1537

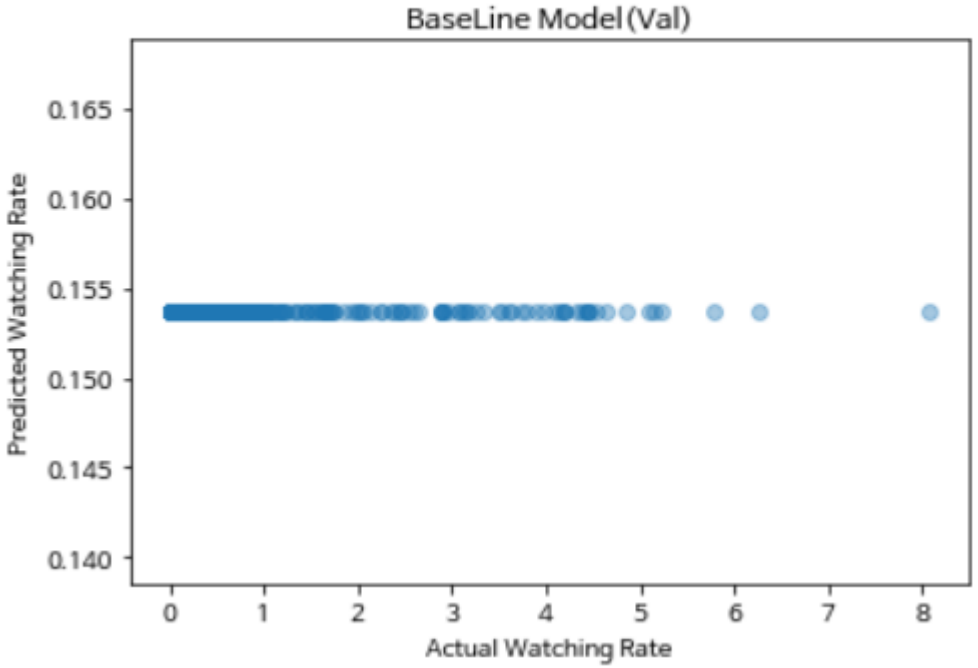
2.모델설정및평가

Model 1:Base Line 기준모델

```
BaseLine_훈련 세트 r2: 0.0000000000
BaseLine_훈련 세트 MSE: 0.2696117100
```

```
BaseLine_검증 세트 r2: -0.0002229936
BaseLine_검증 세트 MSE: 0.3085190294
```





훈련데이터 타겟 값의 **평균**
Baseline = 0.1537

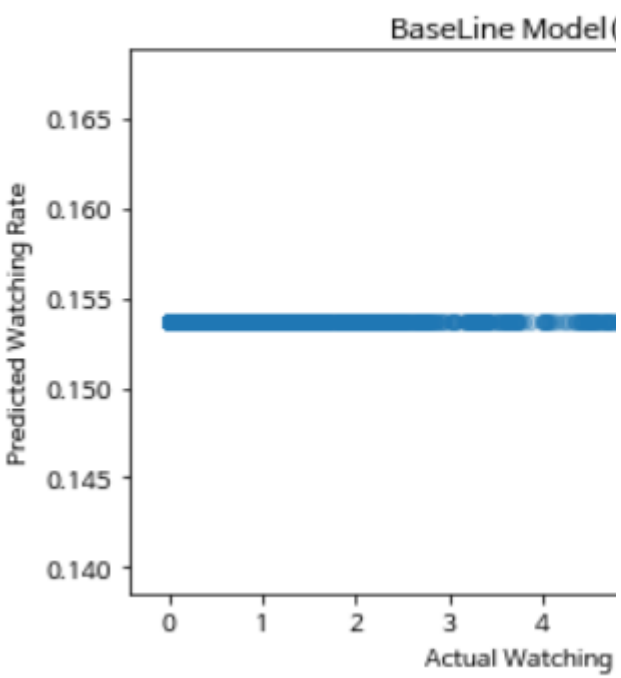
2.모델설정및평가

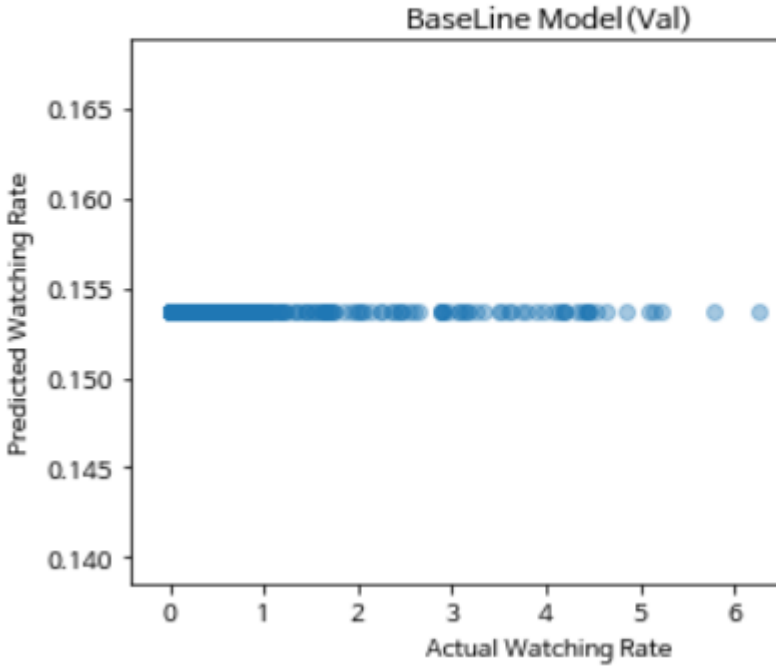
Model 1:Base Line 기준모델

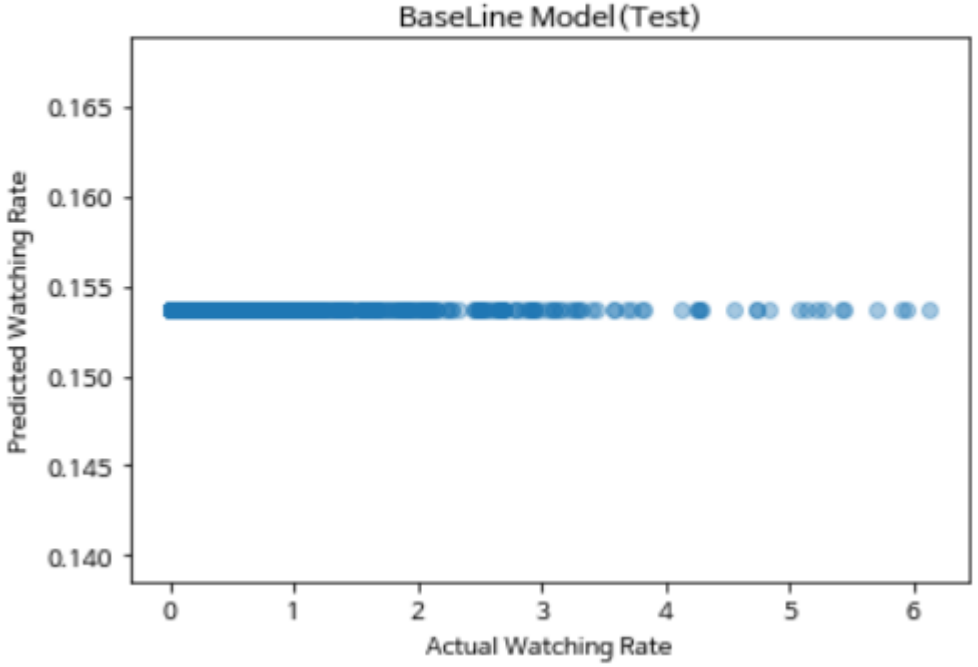
```
BaseLine_훈련 세트 r2: 0.0000000000
BaseLine_훈련 세트 MSE: 0.2696117100
```

```
BaseLine_검증 세트 r2: -0.0002229936
BaseLine_검증 세트 MSE: 0.3085190294
```

```
BaseLine_테스트 세트 r2: -0.0000344486
BaseLine_테스트 세트 MSE: 0.3085190294
```







훈련데이터 타겟 값의 **평균**
Baseline = 0.1537

2.모델설정및평가

Model 2:Random Forest 모델

랜덤포레스트 모델
: 의사결정나무를 기반으로 만든 모델

평가지표:

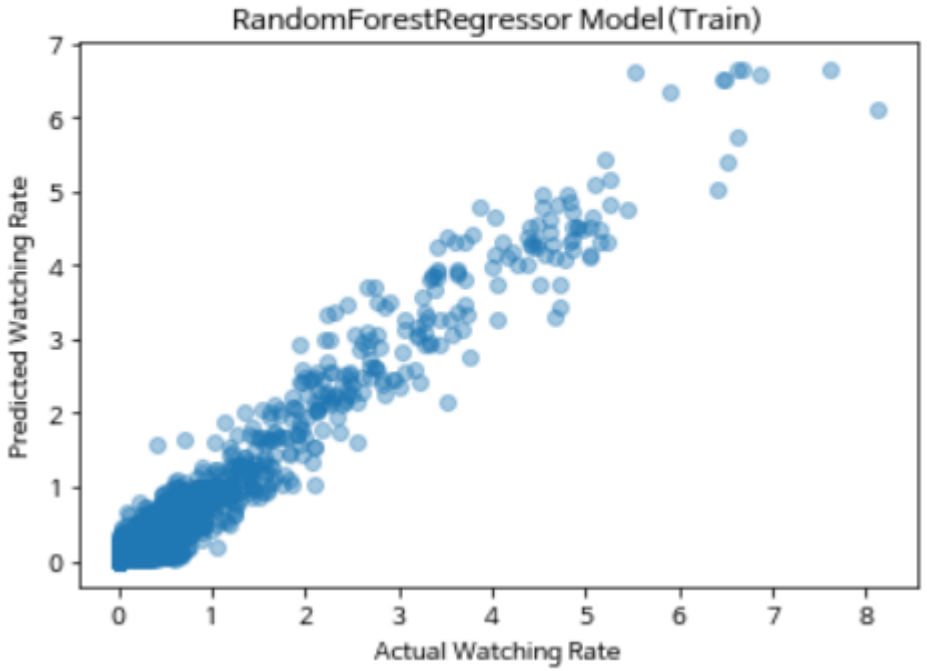
r2: 모델의 설명력, 1에 가까울수록 설명력 높음

MSE: 모델의 오차, 0에 가까울수록 정확도 높음

기준모델 평가지표
r2= -0.00003 MSE= 0.309

2.모델설정및평가

```
RandomForestRegressor_훈련 세트 r2: 0.9481815248
RandomForestRegressor_훈련 세트 조정된 r2: 0.9481492046
RandomForestRegressor_훈련 세트 MSE: 0.0139708677
----------------------------------------------------------
```



모델

록 설명력 높음

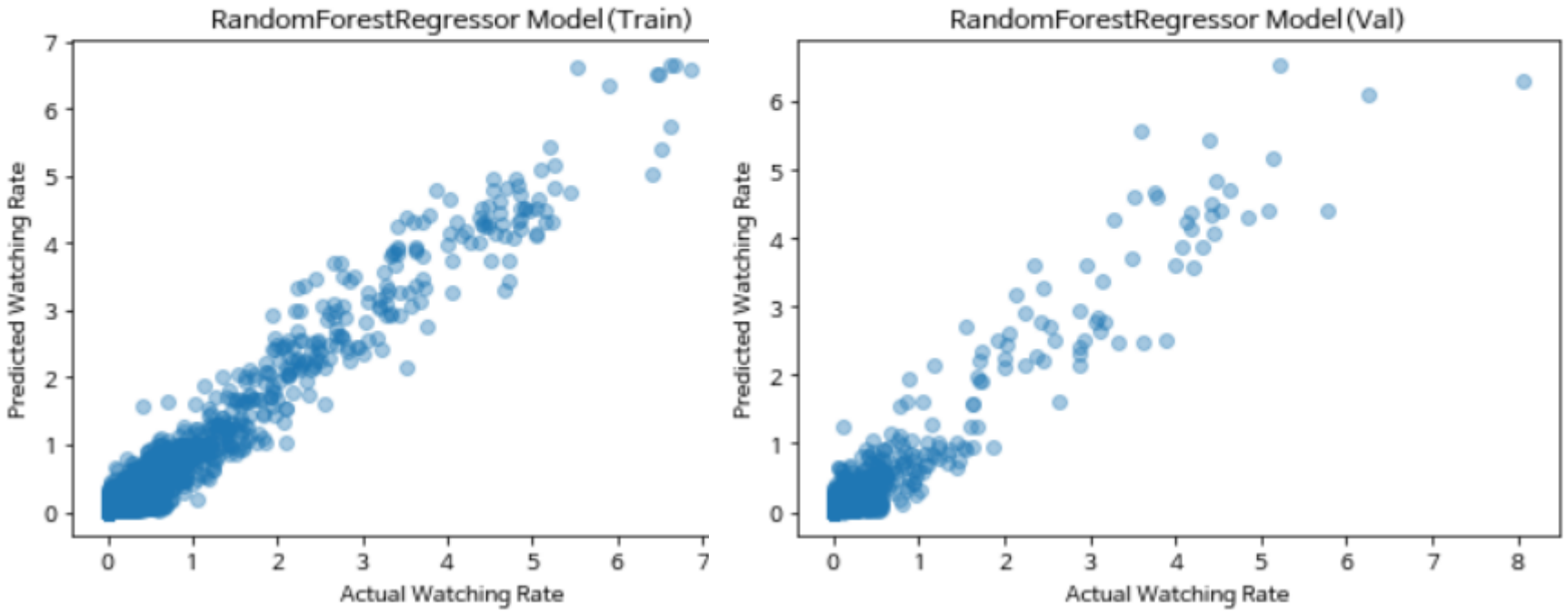
록 정확도 높음

기준모델 평가지표
r2= -0.00003 MSE= 0.309

2.모델설정및평가

```
RandomForestRegressor_훈련 세트 r2: 0.9481815248
RandomForestRegressor_훈련 세트 조정된 r2: 0.948
RandomForestRegressor_훈련 세트 MSE: 0.013970867
```

```
RandomForestRegressor_검증 세트 r2: 0.9160551462
RandomForestRegressor_검증 세트 조정된 r2: 0.9158452841
RandomForestRegressor_검증 세트 MSE: 0.0258928109
```



기준모델 평가지표

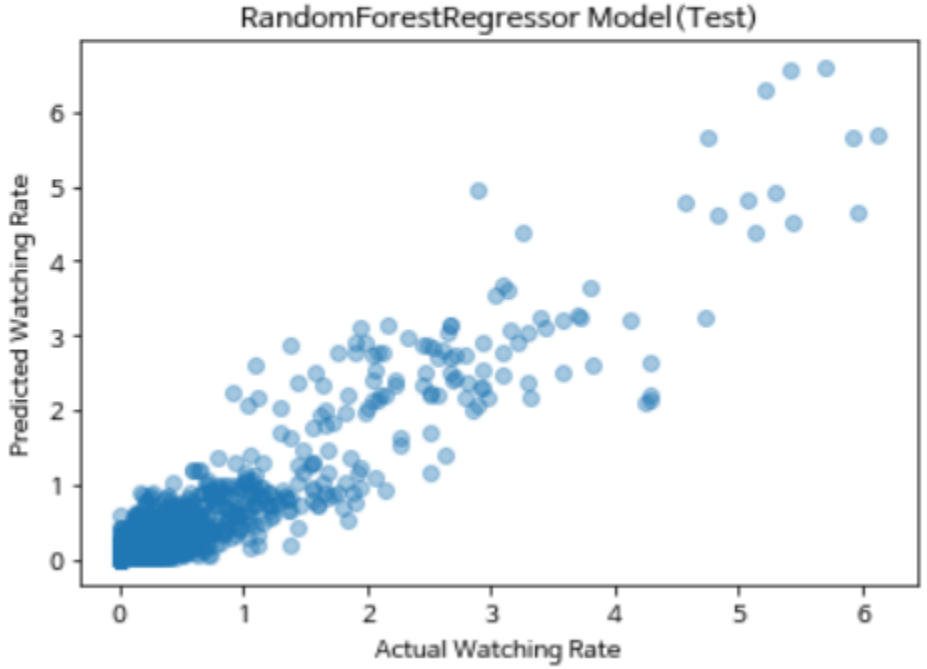
r2= -0.00003 MSE= 0.309

2.모델설정및평가

```
RandomForestRegressor_훈련 세트 r2: 0.9481815248
RandomForestRegressor_훈련 세트 조정된 r2: 0.948
RandomForestRegressor_훈련 세트 MSE: 0.013970867
```

```
RandomForestRegressor_검증 세트 r2: 0.91605
RandomForestRegressor_검증 세트 조정된 r2:
RandomForestRegressor_검증 세트 MSE: 0.0258
```

```
RandomForestRegressor_테스트 세트 r2: 0.8724048656
RandomForestRegressor_테스트 세트 조정된 r2: 0.8722365660
RandomForestRegressor_테스트 세트 MSE: 0.0297324055
```



기준모델 평가지표
r2= -0.00003 MSE= 0.309

2.모델설정및평가

Model 3:Multiple Linear 모델

다중선형회귀 모델
: 선을 기반으로 만든 모델

평가지표:

r2: 모델의 설명력, 1에 가까울수록 설명력 높음

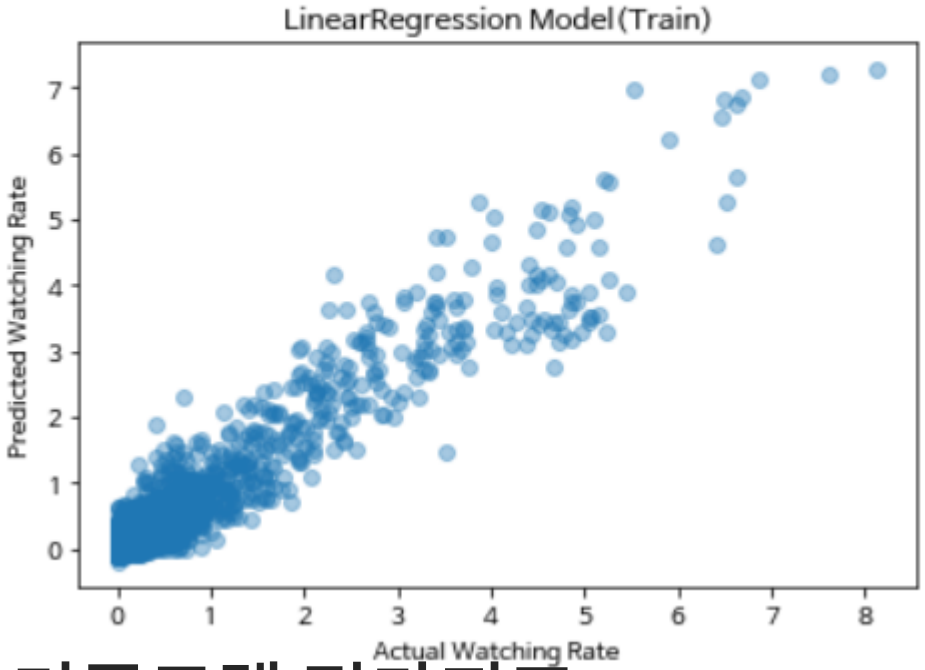
MSE: 모델의 오차, 0에 가까울수록 정확도 높음

기준모델 평가지표
r2= -0.00003 MSE= 0.309

2.모델설정및평가

Model 3:Multiple Linear 모델

```
LinearRegression_훈련 세트 r2: 0.9043877854
LinearRegression_훈련 세트 조정된 r2: 0.9023442320
LinearRegression_훈련 세트 MSE: 0.0257781727
------------------------------------------------------------
```



ᄂ록 설명력 높음

수록 정확도 높음

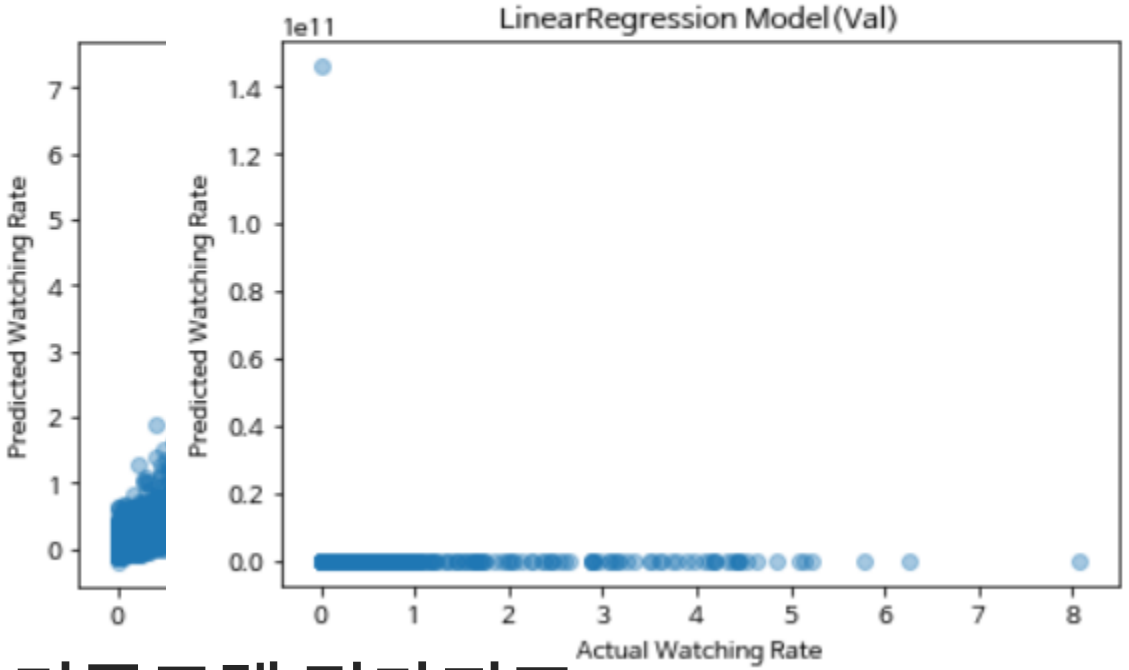
기준모델 평가지표

r2= -0.00003 MSE= 0.309

2.모델설정및평가

Model 3:Multiple Linear 모델

```
LinearRegr LinearRegression_검증 세트 r2: -2461909222624991232O.0000000000
LinearRegr LinearRegression_검증 세트 조정된 r2: -2686850384101224857G.0000000000
LinearRegr LinearRegression_검증 세트 MSE: 7,593,765,076,998,982,656.0000000000
---------- ------------------------------------------------------------
```



높음

높음

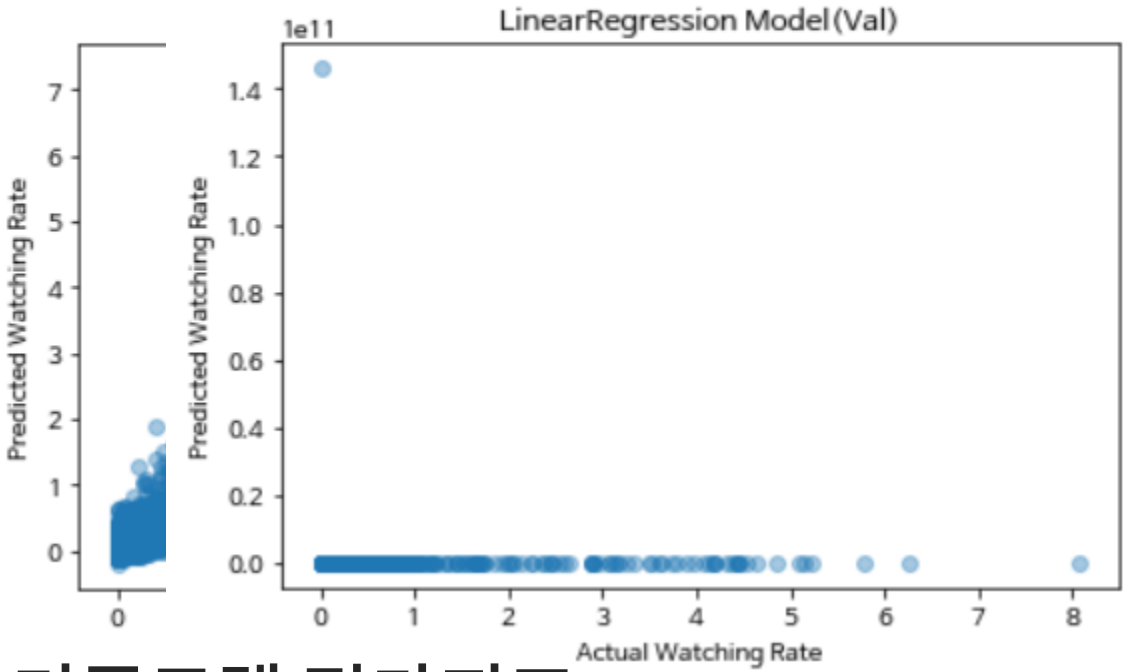
기준모델 평가지표

r2= -0.00003 MSE= 0.309

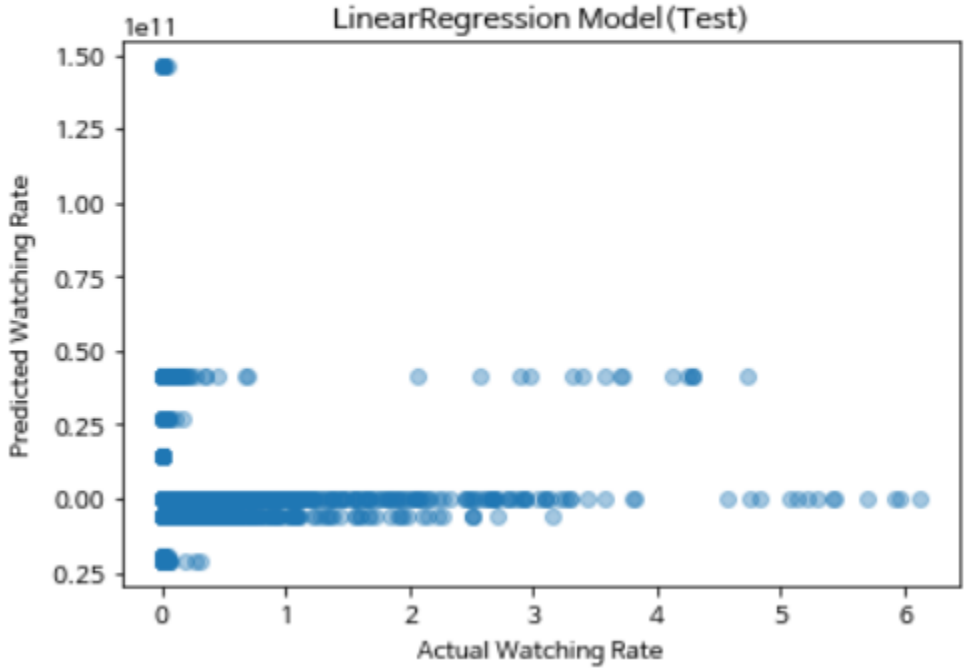
2.모델설정및평가

Model 3:Multiple Linear 모델

```
LinearRegr LinearRegression_검증 세트 r2: -2461909222624991232O.0000000000
LinearRegr LinearRegression_검증 세트 조정된 r2: -2686850384101224857G.0000000000
LinearRegr LinearRegression_검증 세트 MSE: 7,593,765,076,998,982,656.0000000000
---------- ------------------------------------------------------------
```




```
LinearRegression_테스트 세트 r2: -4743436497462243491840.0000000000
LinearRegression_테스트 세트 조정된 r2: -4962910326346596352O0.0000000000
LinearRegression_테스트 세트 MSE: 110,532,253,508,689,444,864.0000000000
------------------------------------------------------------
```



기준모델 평가지표

r2= -0.00003 MSE= 0.309

2.모델설정및평가

Model 4:XGBoost 모델

XGBoost 모델
: 랜덤포레스트와 비슷하지만, 오류에 가중치를 반영

평가지표:
r2: 모델의 설명력, 1에 가까울수록 설명력 높음
MSE: 모델의 오차, 0에 가까울수록 정확도 높음

기준모델 평가지표
r2= -0.00003 MSE= 0.309

2.모델설정및평가

Model 4:XGBoost 모델

XGBoost 모델
: 랜덤포레스트와 비슷하지만, 오류에 가중치를 반영

평가지표:
r2: 모델의 설명력, 1에 가까울수록 설명력 높음
MSE: 모델의 오차, 0에 가까울수록 정확도 높음

기준모델 평가지표
r2= -0.00003 MSE= 0.309

2.모델설정및평가

XGBoostRegression_훈련 세트 r2: 0.9456443486
XGBoostRegression_훈련 세트 조정된 r2: 0.9444825861
XGBoostRegression_훈련 세트 MSE: 0.0146549201

류에 가중치를 반영

록 설명력 높음

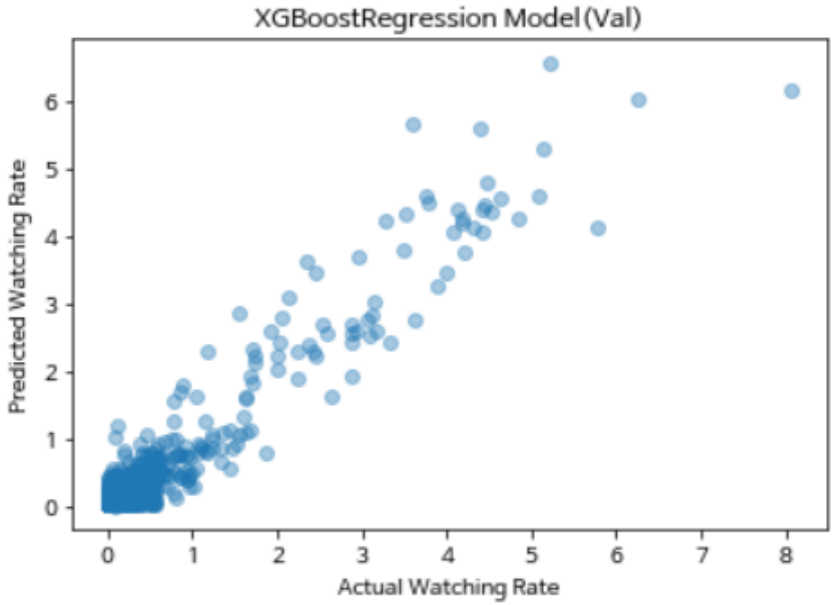
수록 정확도 높음

기준모델 평가지표
r2= -0.00003 MSE= 0.309

2.모델설정및평가

XGBoostRegression_훈련 세트 r2: 0.9456443486
XGBoostRegression_훈련 세트 조정된 r2: 0.94448258
XGBoostRegression_훈련 세트 MSE: 0.0146549201

XGBoostRegression_검증 세트 r2: 0.9156379361
XGBoostRegression_검증 세트 조정된 r2: 0.9079298937
XGBoostRegression_검증 세트 MSE: 0.0260214995



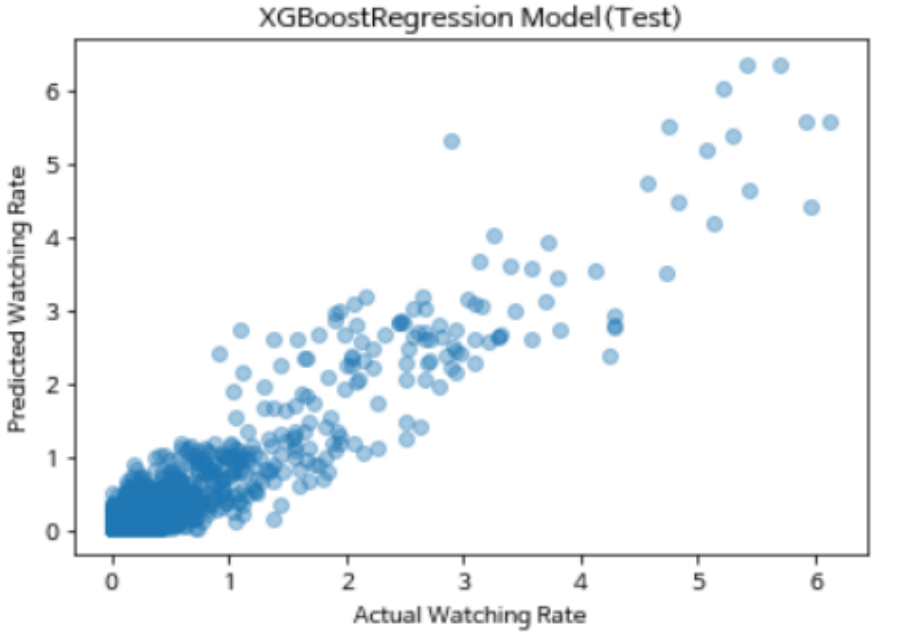
기준모델 평가지표
r2= -0.00003 MSE= 0.309

2.모델설정및평가

XGBoostRegression_훈련 세트 r2: 0.9456443486
XGBoostRegression_훈련 세트 조정된 r2: 0.944482581
XGBoostRegression_훈련 세트 MSE: 0.0146549201

XGBoostRegression_검증 세트 r2: 0.9156379361
XGBoostRegression_검증 세트 조정된 r2: 0.9079298937
XGBoostRegression_검증 세트 MSE: 0.0260214995

XGBoostRegression_테스트 세트 r2: 0.8825560181
XGBoostRegression_테스트 세트 조정된 r2: 0.8771220083
XGBoostRegression_테스트 세트 MSE: 0.0273669691



기준모델 평가지표
r2= -0.00003 MSE= 0.309

2.모델설정및평가

Model 5:XGBoost 모델(Hyper Parameter Tuning)

HP 튜닝: RandomizedSearchCV

최적의 Hyper Parameter 튜닝 값:
n_estimators: 125
Max_depth: 7
subsample: 0.89
lamda: 0.4

기준모델 평가지표
r2= -0.00003 MSE= 0.309

2.모델설정및평가

r Parameter Tuning)

nCV

l 값:

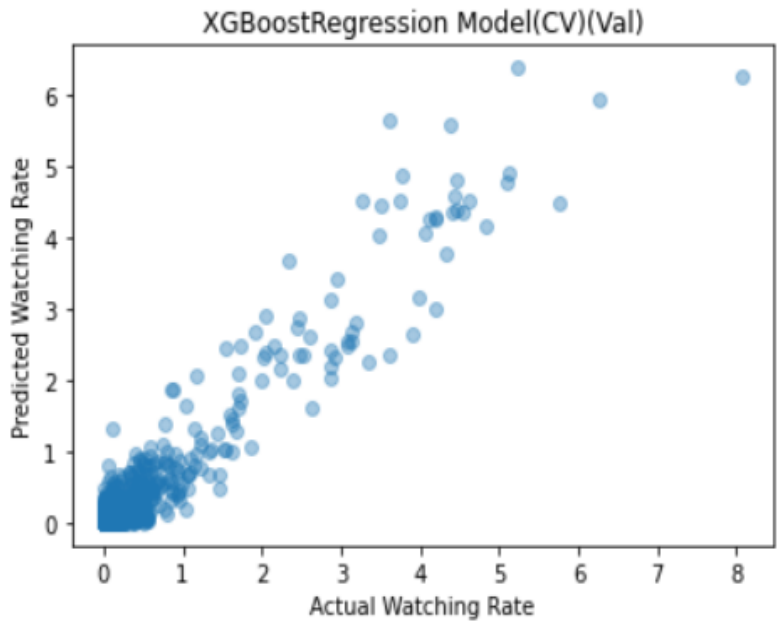
```
XGBoostRegression(CV)_훈련 세트 r2: 0.9832444154
XGBoostRegression(CV)_훈련 세트 조정된 r2: 0.9828862924
XGBoostRegression(CV)_훈련 세트 MSE: 0.0146549201
----------------------------------------------------------
```

기준모델 평가지표
r2= -0.00003 MSE= 0.309

2.모델설정및평가

XGBoostRegression(CV)_훈련 세트 r2: 0.9832444154
XGBoostRegression(CV)_훈련 세트 조정된 r2: 0.9828
XGBoostRegression(CV)_훈련 세트 MSE: 0.014654920

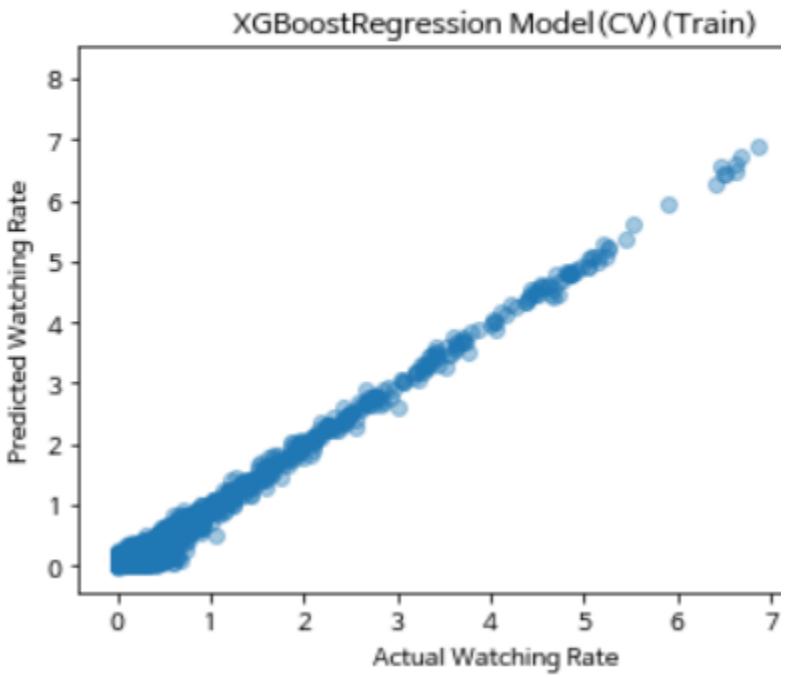
XGBoostRegression(CV)_검증 세트 r2: 0.9128948236
XGBoostRegression(CV)_검증 세트 조정된 r2: 0.9049361469
XGBoostRegression(CV)_검증 세트 MSE: 0.0268676132



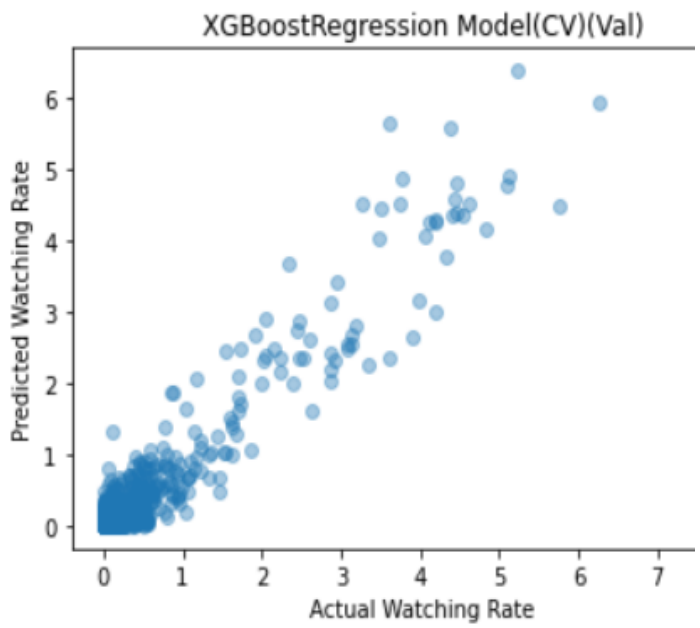
기준모델 평가지표
r2= -0.00003 MSE= 0.309

2.모델설정및평가

XGBoostRegression(CV)_훈련 세트 r2: 0.9832444154
XGBoostRegression(CV)_훈련 세트 조정된 r2: 0.9828
XGBoostRegression(CV)_훈련 세트 MSE: 0.014654920



XGBoostRegression(CV)_검증 세트 r2: 0.9128948236
XGBoostRegression(CV)_검증 세트 조정된 r2: 0.9049361469
XGBoostRegression(CV)_검증 세트 MSE: 0.0268676132



XGBoostRegression(CV)_테스트 세트 r2: 0.8778539292
XGBoostRegression(CV)_테스트 세트 조정된 r2: 0.8722023587
XGBoostRegression(CV)_테스트 세트 MSE: 0.0284626567

기준모델 평가지표
r2= -0.00003 MSE= 0.309

2. 모델 설정 및 평가

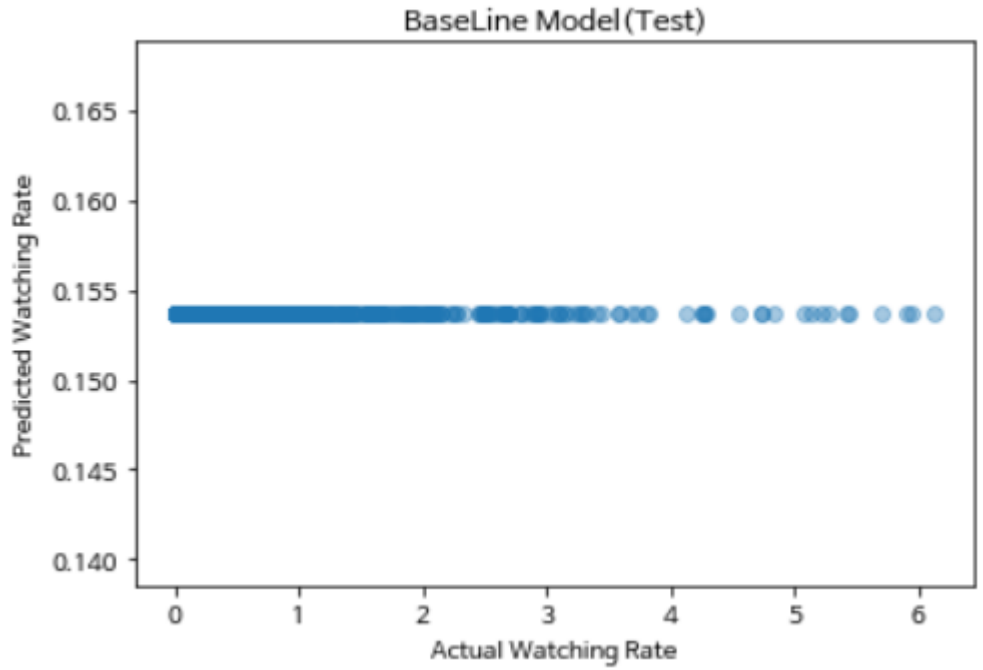
모델 비교

평가지표|모델	기준모델	랜덤포레스트	다중선형회귀	XGBoost(HP 튜닝)
r2	-0.0000344486	0.872	-47 * 10^19	0.877
조정된 r2		0.872	-49 * 10^19	0.872
MSE	0.3085190294	0.030	11 * 10^19	0.028

2. 모델 설정 및 평가

모델 비교

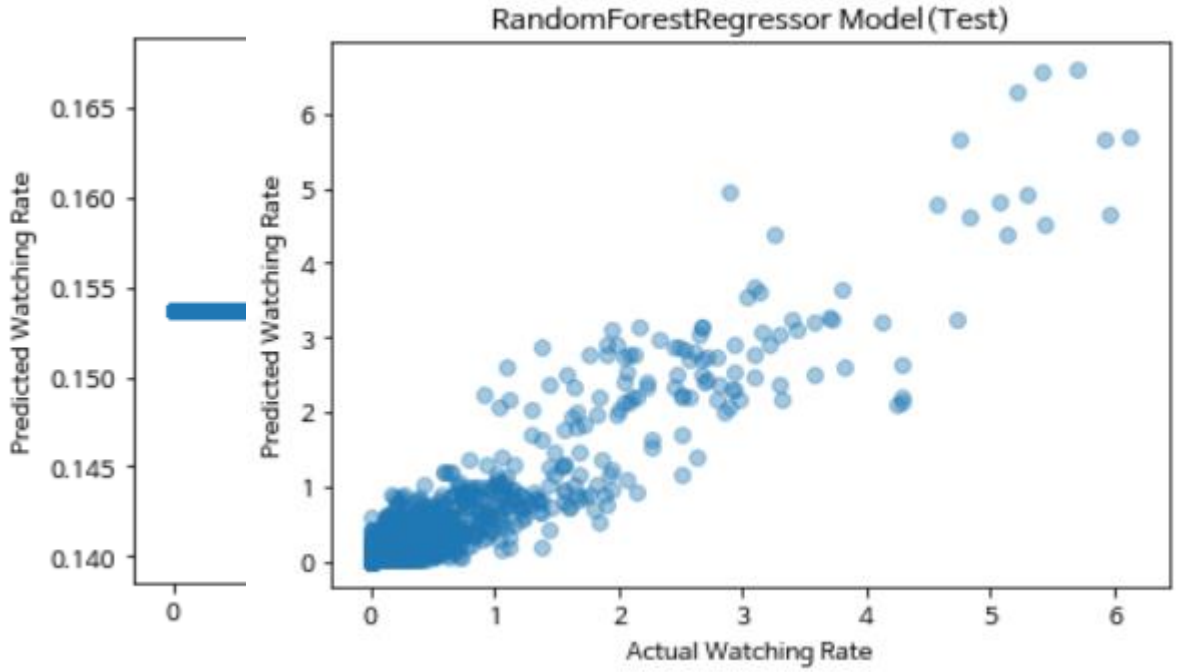
평가지표|모델	기준모델	랜덤포레스트	다중선형회귀	XGBoost
r2	-0.0000344486	0.872	-47 * 10^19	0.877
조정된 r2		0.872	-49 * 10^19	0.872
MSE	0.3085190294	0.030	11 * 10^19	0.028



2. 모델 설정 및 평가

모델 비교

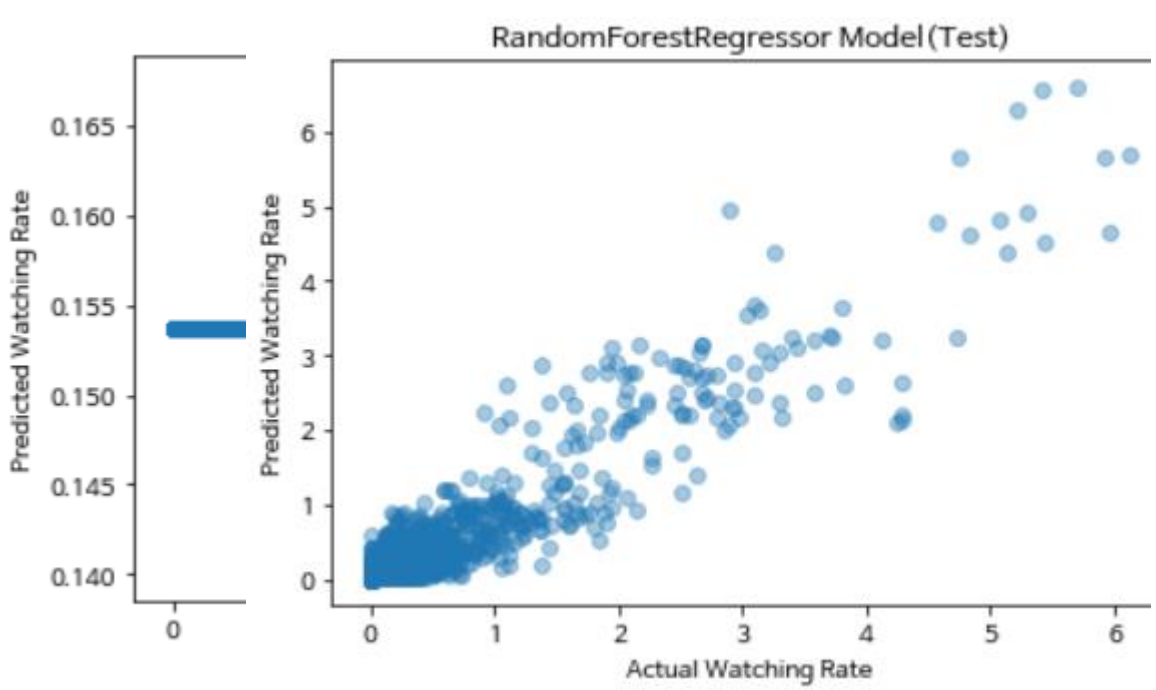
평가지표 | 모델	기준모델	랜덤포레스트	다중선형회귀	XGBoost
r2	-0.0000344486	0.872	-47 * 10^19	0.877
조정된 r2		0.872	-49 * 10^19	0.872
MSE	0.3085190294	0.030	11 * 10^19	0.028

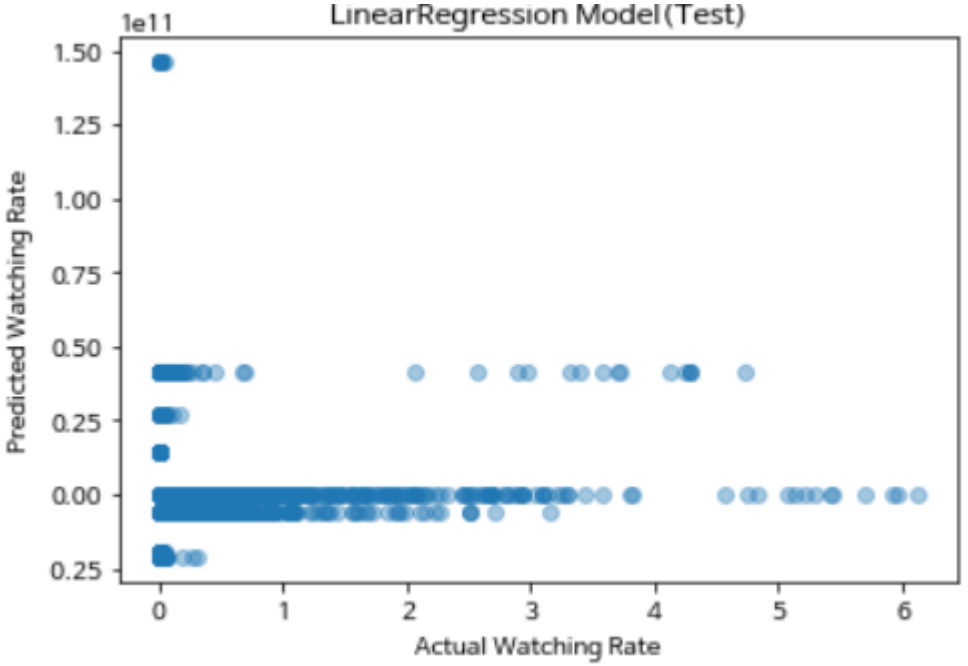


2. 모델 설정 및 평가

모델 비교

평가지표|모델	기준모델	랜덤포레스트	다중선형회귀	XGBoost
r2	-0.0000344486	0.872	-47 * 10^19	0.877
조정된 r2		0.872	-49 * 10^19	0.872
MSE	0.3085190294	0.030	11 * 10^19	0.028

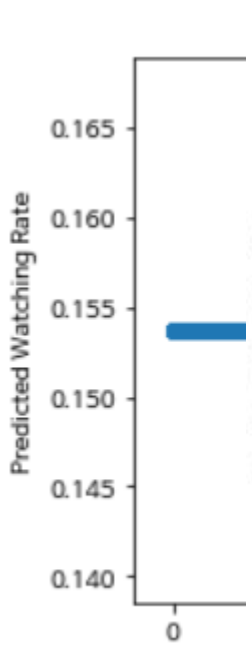


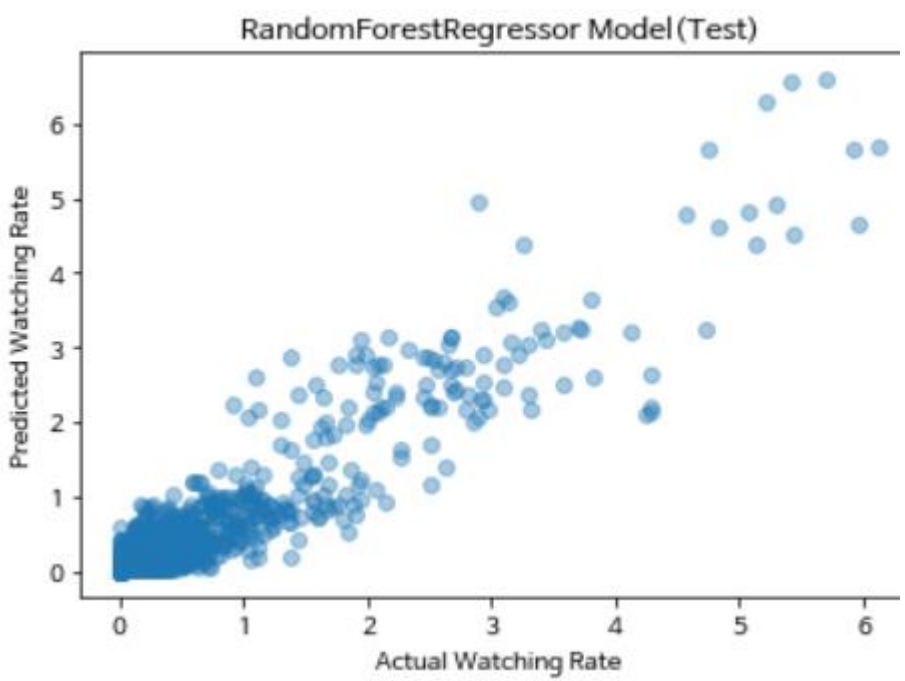


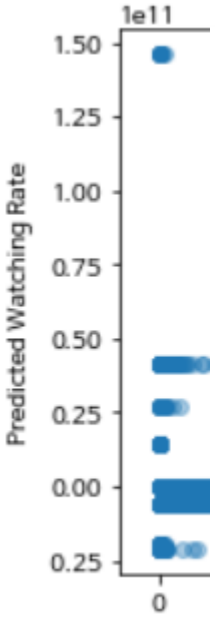
2.모델설정및평가

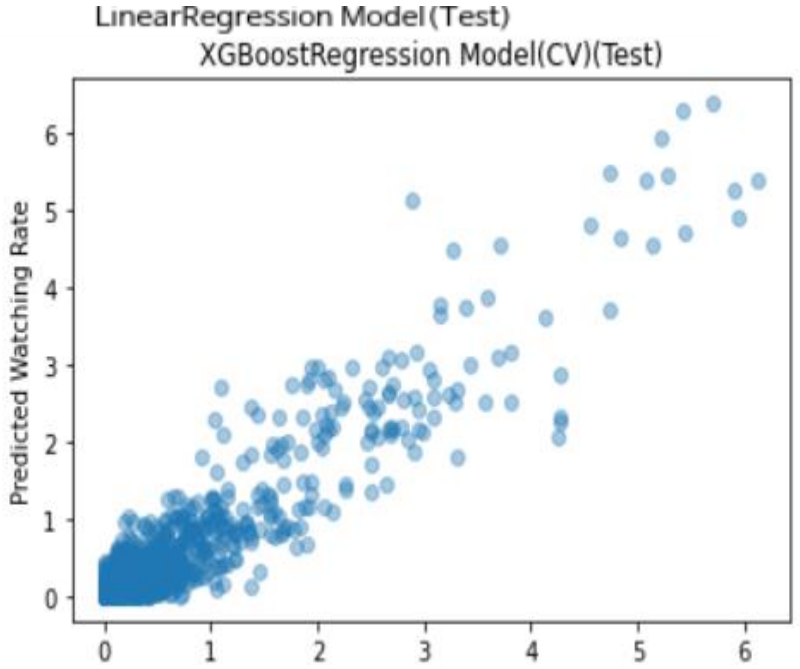
모델 비교

평가지표|모델	기준모델	랜덤포레스트	다중선형회귀	XGBoost
r2	-0.0000344486	0.872	-47 * 10^19	0.877
조정된 r2		0.872	-49 * 10^19	0.872
MSE	0.3085190294	0.030	11 * 10^19	0.028









3. 모델 설명

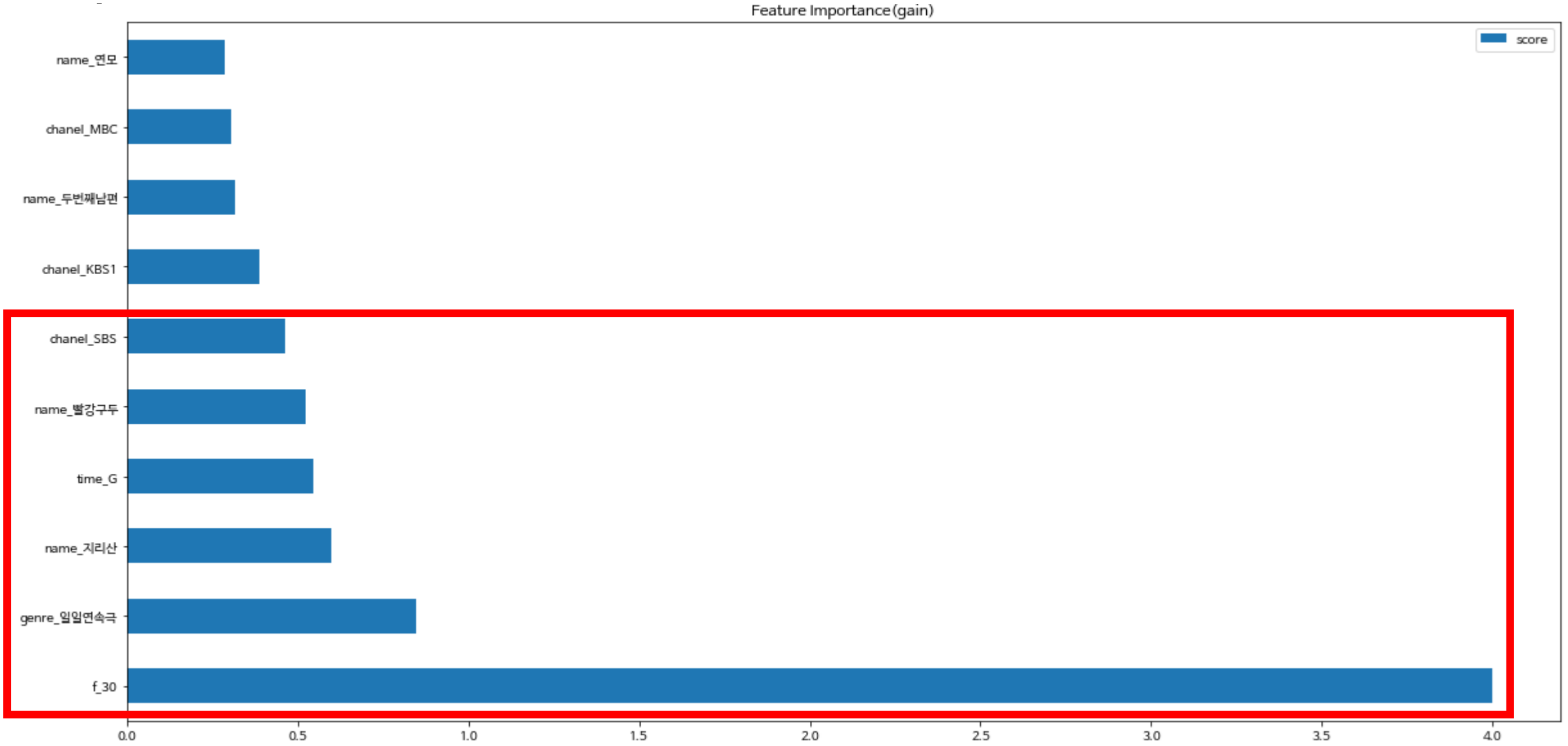
특성 중요도		PDP 그래프		SHAP 그래프
Gain 기준 Weight 기준 Cover 기준	>>	F_30 Time_G Chanel_SBS	>>	SHAP 그래프

3. 모델 설명

특성 중요도: Gain

Gain이란?
해당 특성이 모델 예측에 얼마나 영향을 미쳤는가

3. 모델 설명

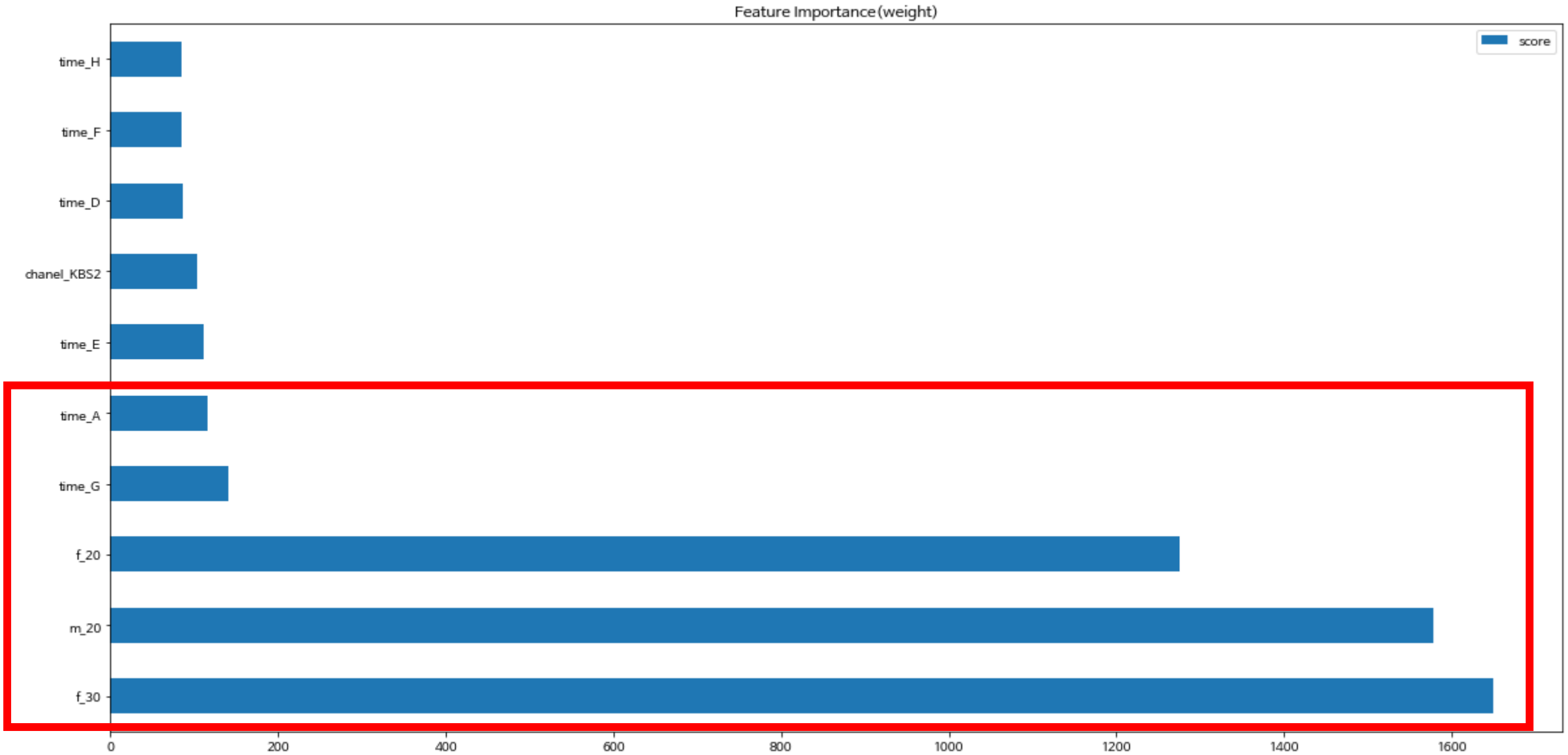


3. 모델 설명

특성 중요도: Weight

Weight이란?
해당 특성과 관련된 샘플의 상대적인 개수

3. 모델 설명



3. 모델 설명

특성 중요도: Cover

Cover란?
해당 특성이 모델의 의사결정에 사용된 횟수

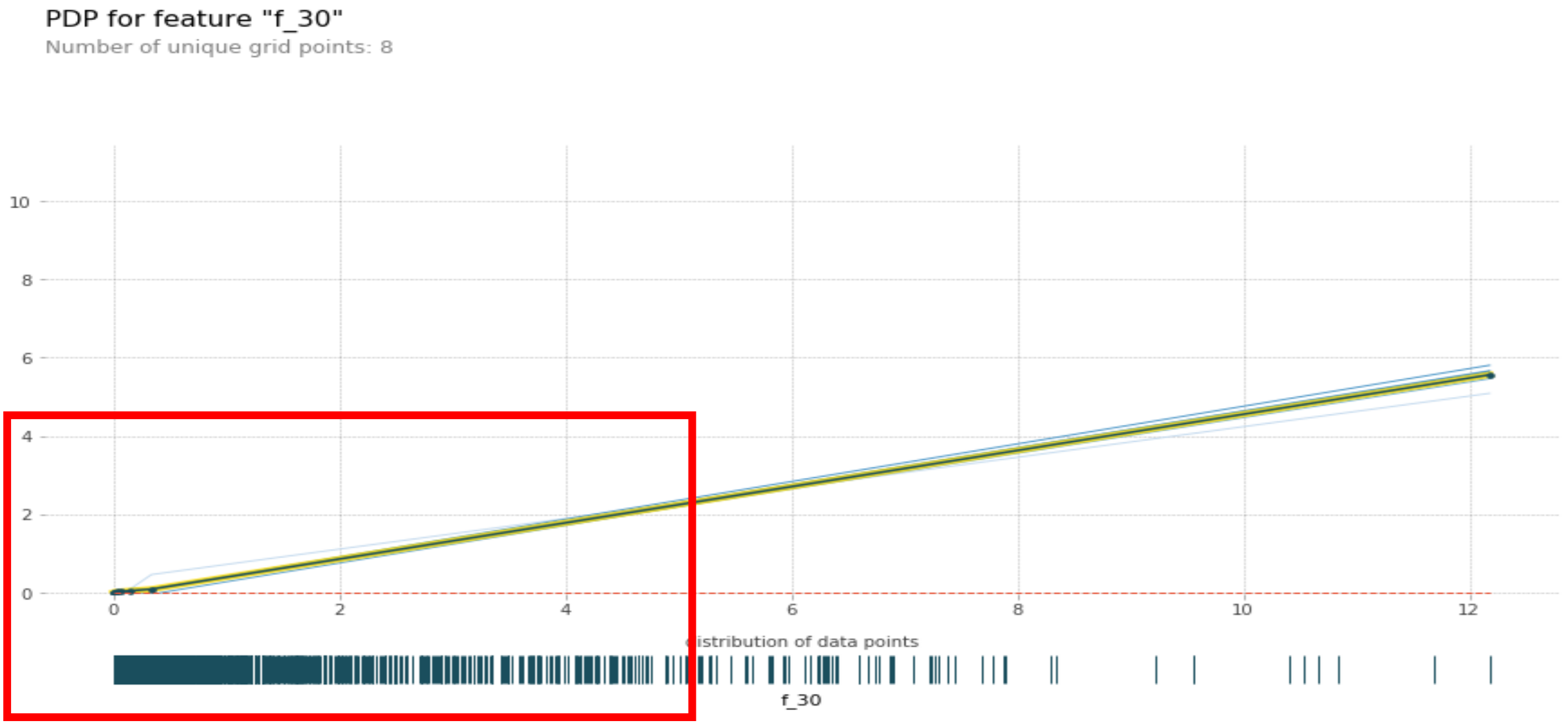
3. 모델 설명

3. 모델 설명

PDP 그래프: F_30

PDP란?
해당 특성이 모델의 분석에 어떤 영향을 주었는지

3. 모델 설명

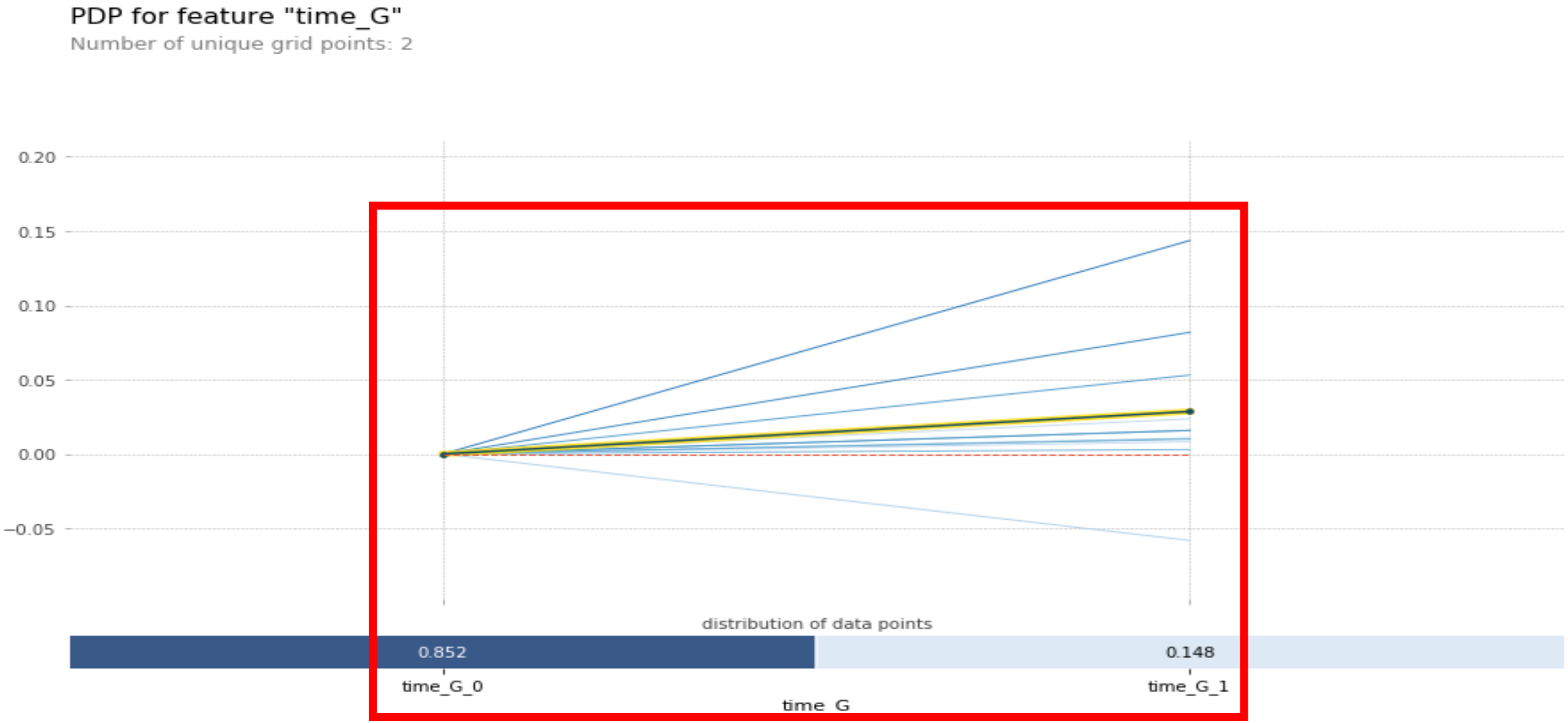


3.모델설명

PDP 그래프: Time_G

Time_G란?
18시 ~ 21시에 시작하는 프로그램

3. 모델 설명

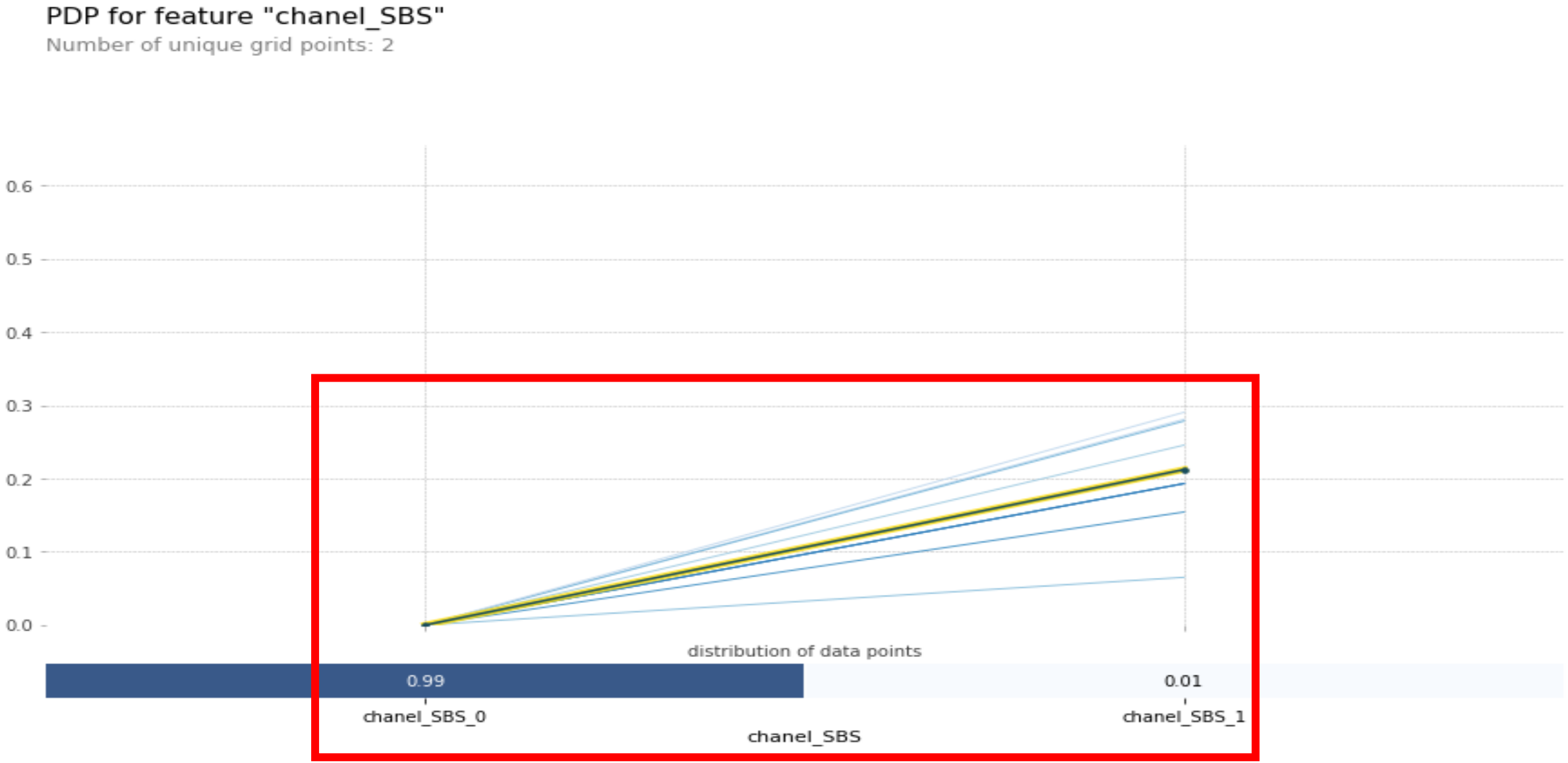


3. 모델 설명

PDP 그래프: Chanel_SBS

Chanel_SBS란?
SBS에서 방영하는 프로그램

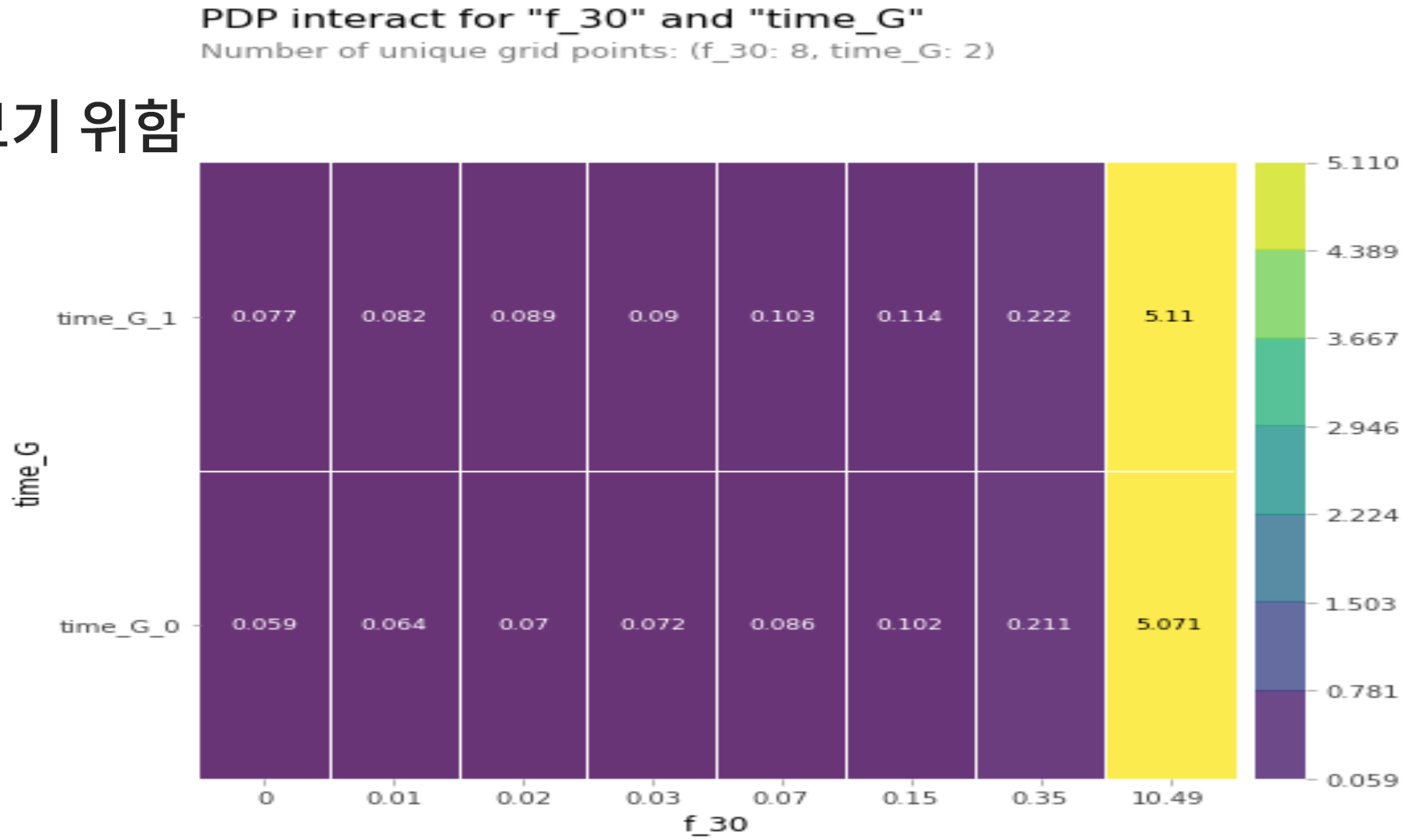
3. 모델 설명



3.모델설명

PDP 그래프: F_30 and Time_G

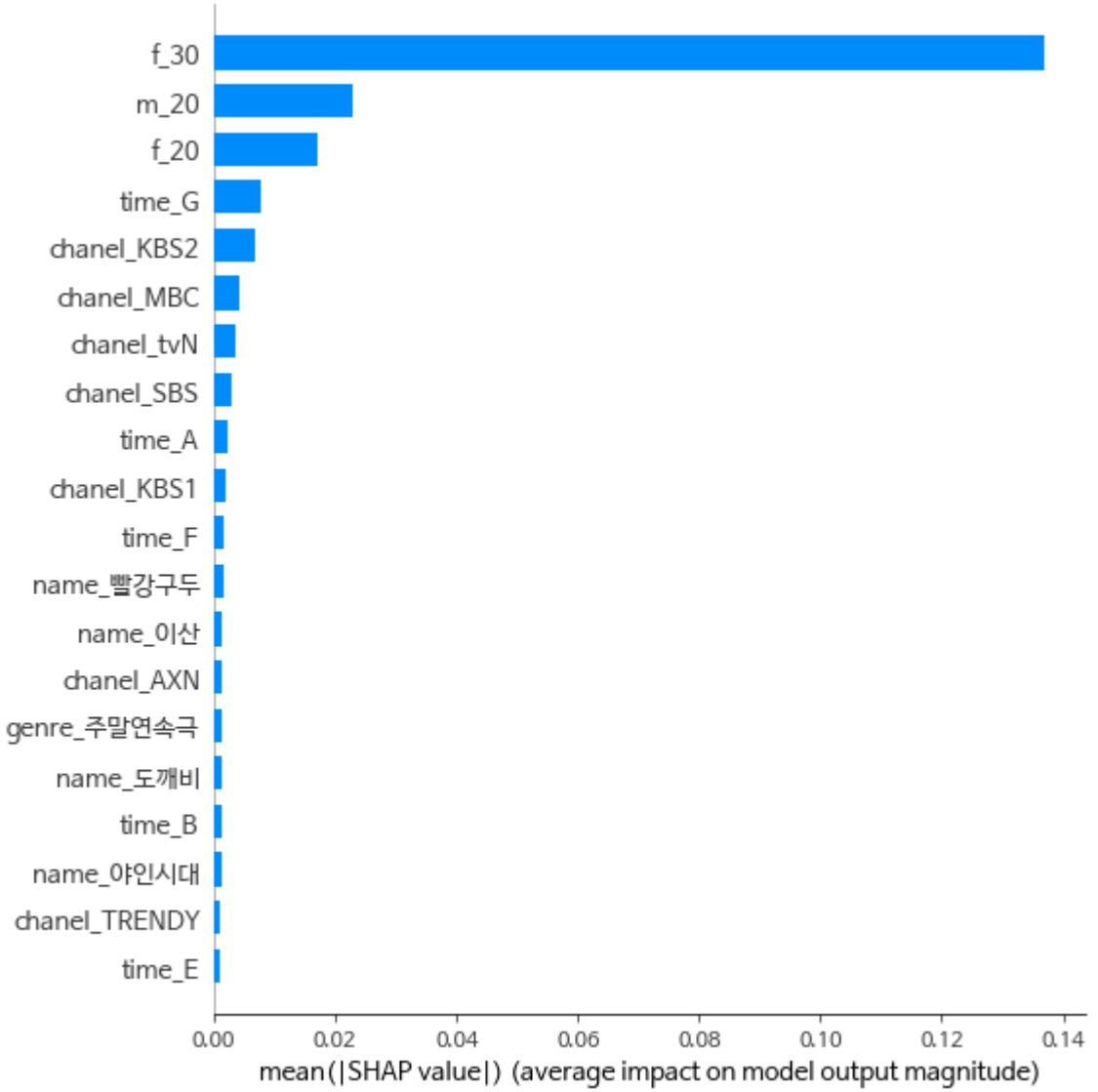
F_30 and Time_G?
두 가지 특성을 함께 보기 위함



3. 모델 설명

SHAP 그래프

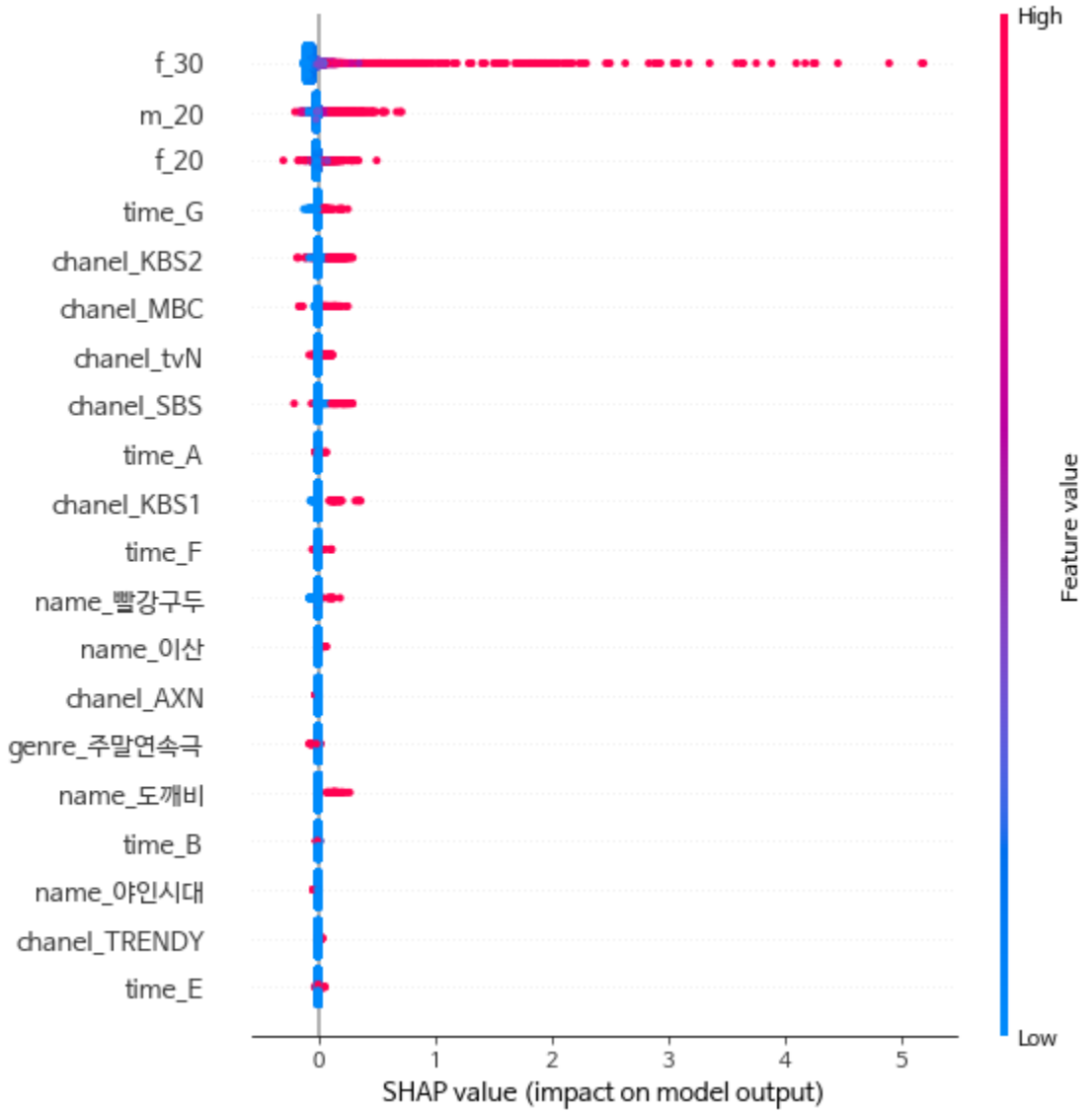
SHAP 그래프?
모델이 개별 특성 값을
어떻게 예측하였는지



3. 모델 설명

SHAP 그래프: 요약 그래프

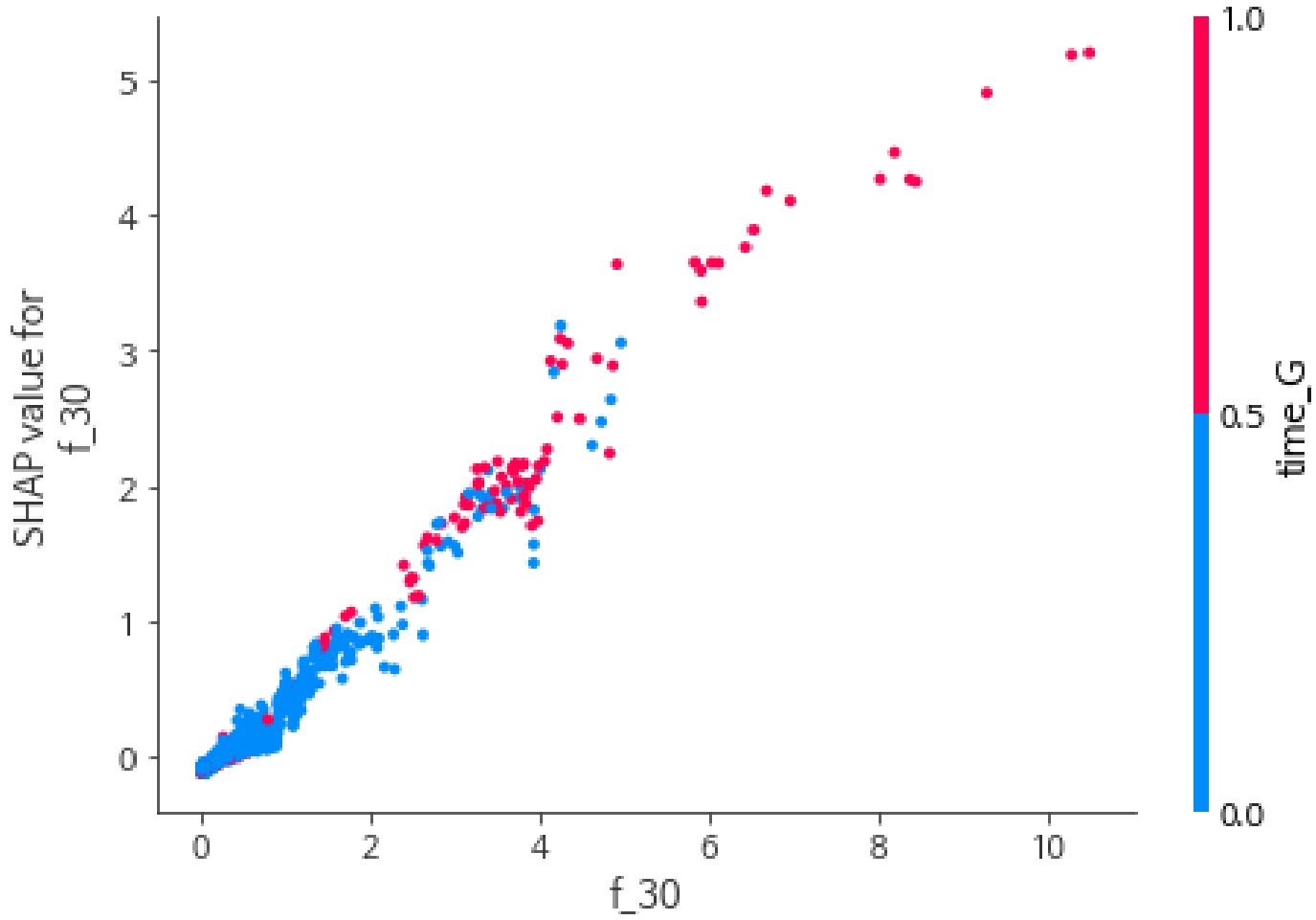
SHAP 요약 그래프?
모델이 개별 특성 값의 높낮이에 따라
어떻게 예측하였는지



3.모델설명

SHAP 그래프: F_30 and Time_G

SHAP F_30 and Time_G?
모델이 두 특성 값에 따라
어떻게 예측하였는지



4. 결론

모델

XGBoost 모델

평가지표가 비교적 우수

시간대 채널

Time_G
18 ~ 21시 시간대에 시작

Chanel_SBS
SBS

30대 여성

30대 여성의 시청률

감사합니다