6조 : 배상원, 신영수, 육태경,
최무성, 함혜민, 홍영화

Contents

프로젝트 기획

배경

계절 및 장소 등에 관계 없이 발생하는 김해시의 화재 문제 해결

목적

소방 및 건물관련 정보를 융합하여 지역 내 화재 위험도에 대해 분석 및 예측

화재에 대한 집중적이고 적극적인 예방 활동 수행

분석 과제

경상남도의 소방 및 건물 관련 데이터를 활용하여
김해시 내 건축물의 화재 위험도 분석 및 예측 모델을 제시

- 화재 발생에 가장 큰 영향을 미치는 소방 및 건물 변수는 무엇인가?
- 화재 발생 예측 성능이 우수한 모델은 무엇인가?

데이터 설명

데이터 출처

김해시 화재발생 예측모델 개발
(https://compas.lh.or.kr/subj/pas/info?subjNo=SBJ_1920_002#)

분석 대상

모델 성능의 검증을 위하여 화재 발생 여부(Y/N)가 포함된**Train, Validation Set**을 사용
Validation Set을 5:5로 나누어**Validation Set과 Test Set을 생성**하여 분석

데이터	지역	관측값 수(N)	변수
Train	경상남도	59,199	179
Validation	김해시	6,898	179
Test	김해시	2,957	179

변수 목록

종속 변수: 화재 발생 여부(fr_yn)

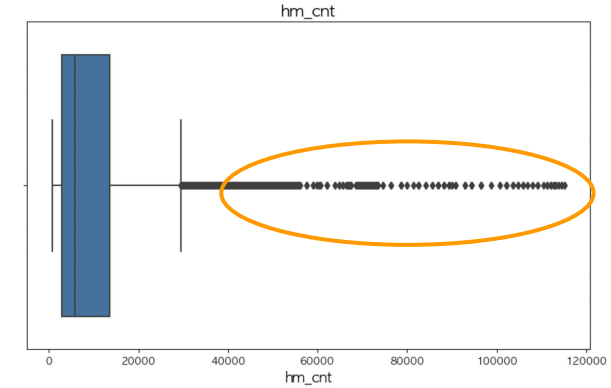
설명 변수:

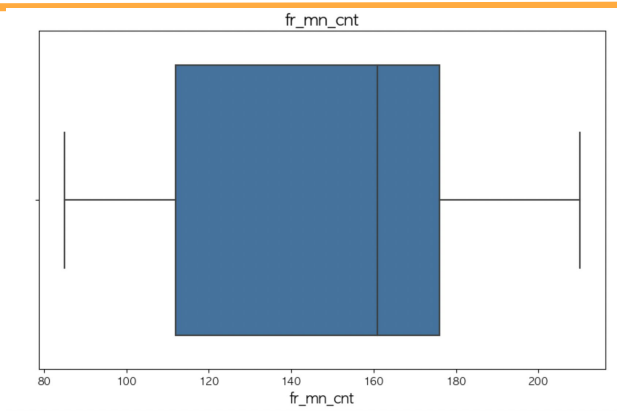
1	dt_of_fr	화재발생 일시	16	hmdt	습도
2	bldng_us	건물용도	17-76	gas_engry_us	가스 에너지 월별 사용량(2014년 - 2018년)
3	bldng_archtctr	건물구조	77-136	ele_engry_us	전기 에너지 월별사용량(2014년 - 2018년)
4	bldng_cnt	건물채수	137	lw_13101010	복도/계단/출입구의 성능 유지여부(0~5)
5	bldng_ar	건물건출면적	138	lw_13101110	옥상광장의 피난성능 유지여부(0~5)
6	ttl_ar	건물연면적(건물층별합계전체 면적)	139	lw_13101210	방화문/방화셔터 등의 성능 유지여부(0~5)
7	lnd_ar	토지면적	140	lw_13101211	방화구획 적합 여부(0~5)
8	dt_of_athrztn	건물승인일자	141	lw_13101310	경계벽 및 칸막이벽의 변경 등 방화성능 유지여부(0~5)
9	ttl_grnd_flr	건물들의 지상 층수의 합	142	lw_13101410	배연설비의 성능 유지여부(0~5)
10	ttl_dwn_flr	건물들의 지하 층수의 합	143	lw_13111010	내화구조의 성능 유지여부(0~5)
11	bldng_us_clssfctn	건물용도분류명	144	lw_13111110	방화벽의 성능 유지여부(0~5)
12	tmprtr	온도	145	lw_13121010	외벽의 성능 유지여부(0~5)
13	prcpttn	강수량	146	lw_13121011	창호의 성능 유지여부(0~5)
14	wnd_spd	풍속	147	lw_13131010	내부마감의 방화성능 유지여부(0~5)
15	wnd_drctn	풍향	148	lw_13131110	외부마감의 노후화 및 마감재 탈락 여부(0~5)

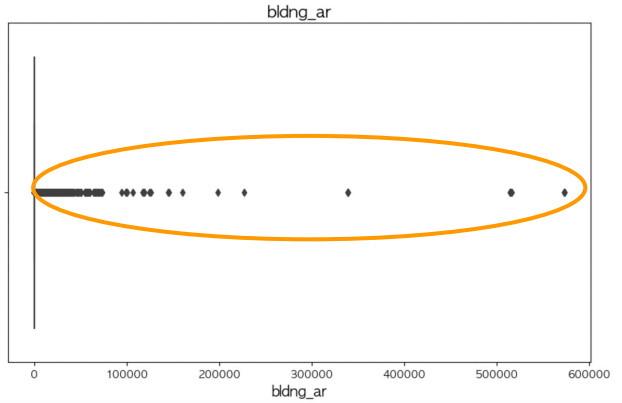
변수 목록

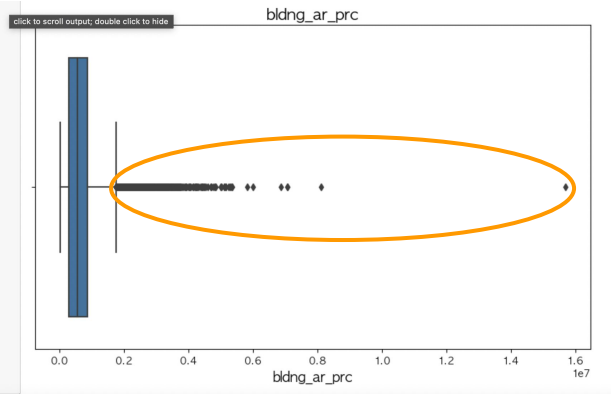
149	lw_13141010	지하층의 소방설비 성능 유지여부(0~5)	164	fr_wthr_fclt_in_100m	반경 100m 이내 소방용수 시설 수
150	lw_13141011	지하층 피난구/피난계단의 성능 유지여부(0~5)	165	cctv_in_100m	반경 100m 이내 공공 CCTV
151	jmk	지적상 지목	166	tbc_rtl_str_dstnc	담배 소매점과의 최소 거리
152	rgnl_ar_nm	용도지역지구명	167	sft_emrgnc_bll_dstnc	안전 비상벨과의 최소 거리
153	rgnl_ar_nm2	용도지역지구명2	168	ahsm_dstnc	자동 심장 충격기와의 최소 거리
154	lnd_us_sttn_nm	토지이용상황명	169	no_tbc_zn_dstnc	금연구역과의 최소 거리
155	rd_sd_nm	도로측면명	170	trgt_crtr	소방관리대상물기준
156	emd_nm	행정구역명	171	fr_fghtng_fclt_spcl_css_5_yn	소방시설특례5호여부
157	hm_cnt	행정구역 인구	172	fr_fghtng_fclt_spcl_css_6_yn	소방시설특례6호여부
158	fr_sttn_dstnc	119 안전센터와의 거리	173	us_yn	사용여부
159	bldng_ar_prc	단위 면적당 건물 가격(2019년)	174	dngrs_thng_yn	위험물대상여부
160	fr_wthr_fclt_dstnc	소방용수시설(소화전 등)과의 최소 거리	175	slf_fr_brgd_yn	자체소방대여부
161	fr_mn_cnt	관할 소방서 인원	176	blk_dngrs_thng_mnfctr_yn	대량위험물제조소등여부
162	mlt_us_yn	다중이용시설 포함여부	177	cltrl_hrtg_yn	문화재여부
163	cctv_dstnc	공공 CCTV와의 최소 거리	178	bldng_cnt_in_50m	50m내 건물채수

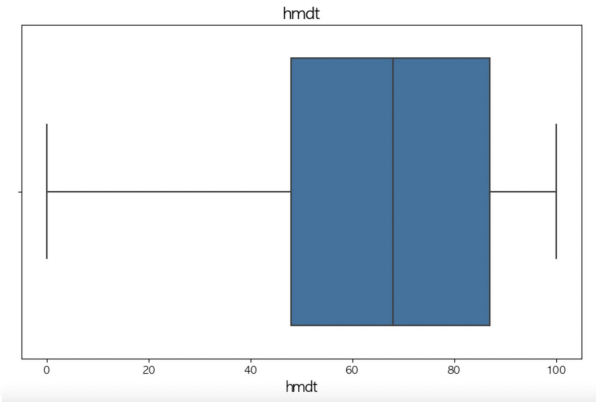
데이터 분포 파악

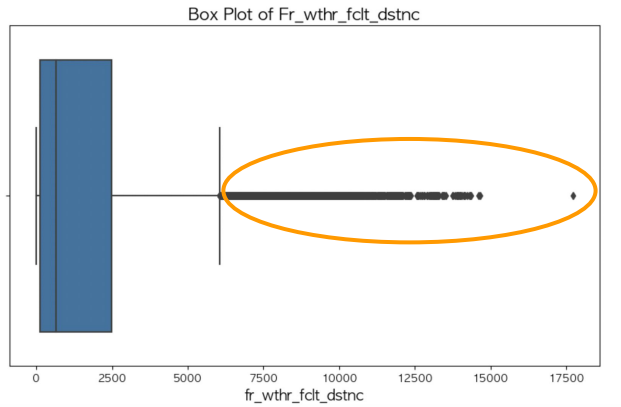

hm_cnt
0
20000
40000
60000
80000
100000
120000
hm_cnt


fr_mn_cnt
80
100
120
140
160
180
200
fr_mn_cnt

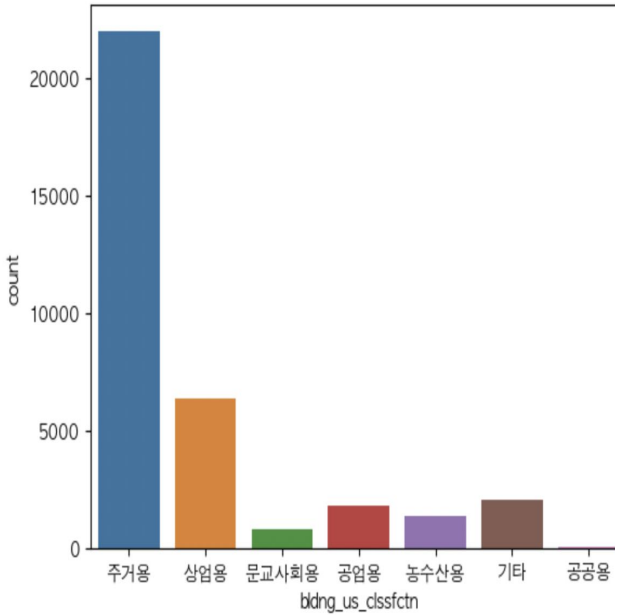

bldng_ar
0
100000
200000
300000
400000
500000
600000
bldng_ar

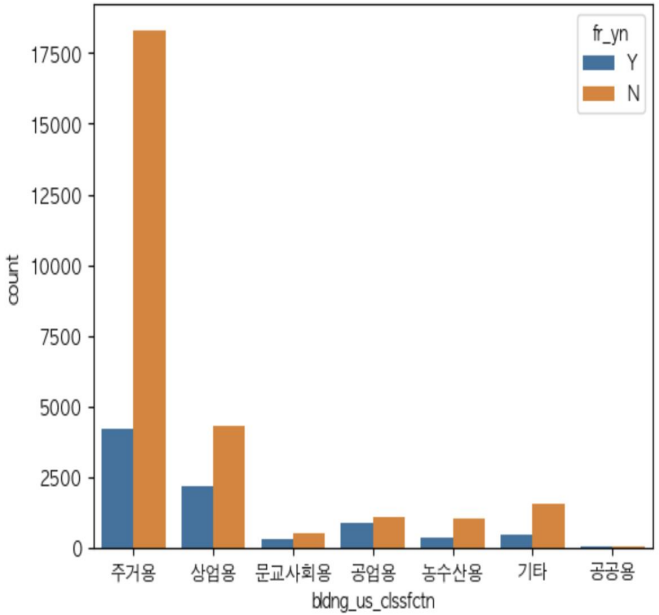

click to scroll output; double click to hide
bldng_ar_prc
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1e7
bldng_ar_prc

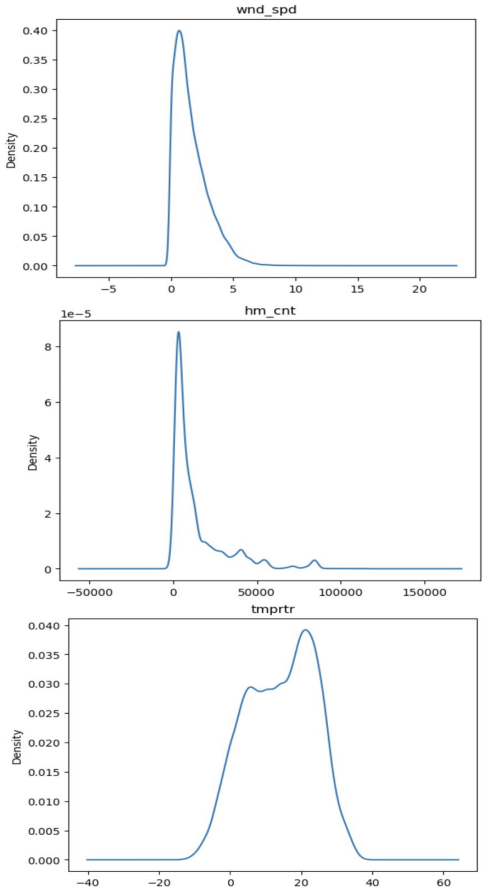

hmdt
0
20
40
60
80
100
hmdt


Box Plot of Fr_wthr_fclt_dstnc
0
2500
5000
7500
10000
12500
15000
17500
fr_wthr_fclt_dstnc

데이터 분포 파악


count
20000
15000
10000
5000
0
주거용
상업용
문교사회용
공업용
농수산용
기타
공공용
bldng_us_clssfctn

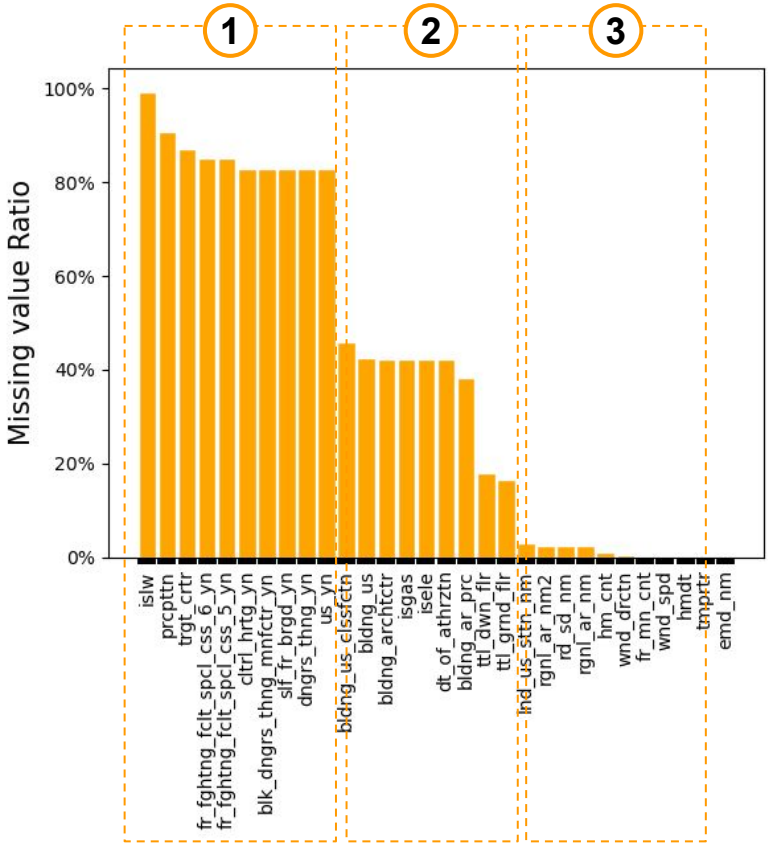

fr_yn
Y
N
count
17500
15000
12500
10000
7500
5000
2500
0
주거용
상업용
문교사회용
공업용
농수산용
기타
공공용
bldng_us_clssfctn


wnd_spd
Density
hm_cnt
1e−5
Density
tmprtr
Density

데이터 전처리 과정

결측치 처리	➔ 결측치가 많은 변수 정리
중복변수 처리	➔ 유사한 의미를 갖는 변수 정리
최종변수 도출	➔ 파생 변수를 생성하고 변수 결합을 통해 최종 변수 도출
샘플링	➔ 화재 발생 비중이 낮은 불균형 데이터 문제 해결 위한 샘플링 ➔ 언더샘플링 (RUS) 와 오버샘플링 (SMOTE) 시도, 오버샘플링 채택

결측치 처리 방향



1 **대부분 결측인 변수 (80% 이상)** ⇒ 분석 부적합

⇒ 분석 부적합, 변수 제거
단, 결측치의 화재 발생 빈도가 전체 화재 발생 확률과 차이가 나는 경우, 비 무작위 결측으로 판단 후 이진 분류

2 **상당수 결측인 변수 (10 ~ 50%)** ⇒ 결측치 처리 필요

⇒ [범주형]의 경우, 비 무작위 결측 판단 시 이진 분류
⇒ [연속형]의 경우, 편향 고려해 결측치를 평균값으로 대체

3 **소수 결측인 변수 (3% 이하)** ⇒ 결측치 대체

⇒ [연속형]의 경우, 변수 특성 고려해 동일 지역 평균값 등 합리적 방식으로 대체함
(예) 풍향의 경우 동일한 지역의 평균값으로 대체값 삽입 후 결측치를 동/서/남/북으로 범주화함

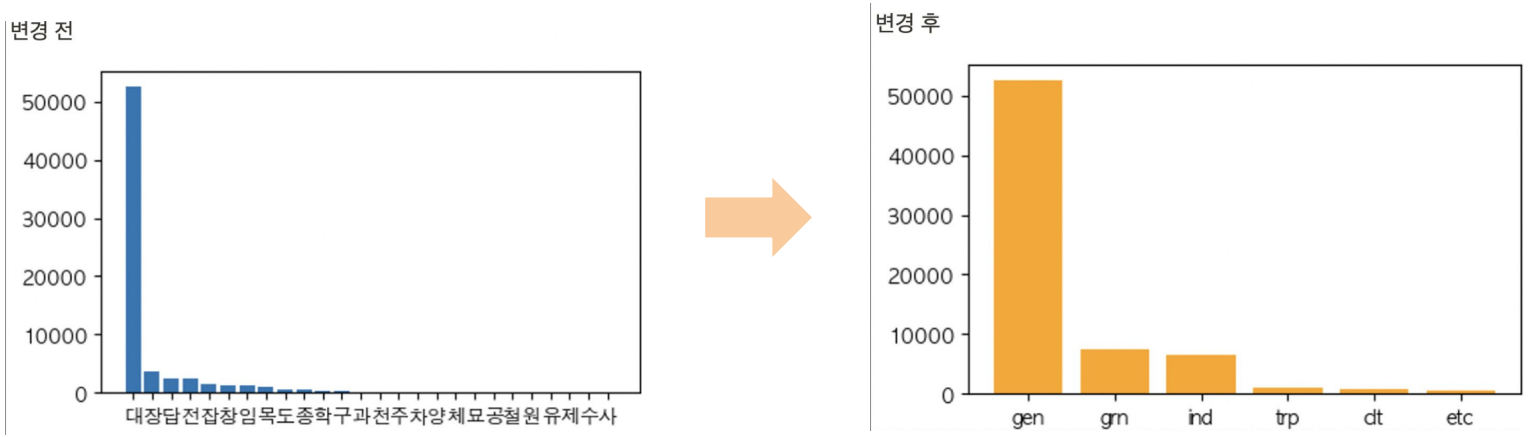
중복 변수 처리

중복 변수	변수 선정	선정 이유
건물 용도 (bldng_us) 건물 용도 분류명 (bldng_us_clssfctn)	**건물 용도 분류명** (bldng_us_clssfctn)	∵ 대분류 용도분류 활용 (주거용/상업용/공업용/공공용/농수산용/문화교육사회용 등)
지목 (rmk) 토지이용상황명 (lnd_us_sttn_nm) 용도지역지구명 (rgnl_ar_nm) 용도지역지구명2 (rgnl_ar_nm2)	**지목** (rmk)	∵ 토지이용 정보를 나타내면서, 결측치가 없는 변수를 선택함
50m이내 건물채 수 (bldng_cnt_in_50m) 건물채 수 (bldng_cnt)	**건물채 수** (bldng_cnt)	∵ 직접적인 건물 밀도를 나타내는 건물채 수 활용

파생변수

토지 관련 변수

- **지목(jmk) 을 지목대분류(jmk_grp) 로 더 큰 범주화 수행**
  지목 종류가 총 28개로 변수 개수 과다하여 더 큰 6가지 종류로 범주화, 변경 전과 변경 후의 분포는 비슷함
- **지목 대분류(jmk_grp)**
  gen(대지): 대; **grn**(녹지): 전,답,과,목,임,공,체,묘; **ind**(공업용지): 장,창,잡;
  clt(문화교육 용지): 학,원,종,사; **trp**(교통시설 용지): 차,주,도,철; **etc**(그 외): 구,수,양,유,제,천

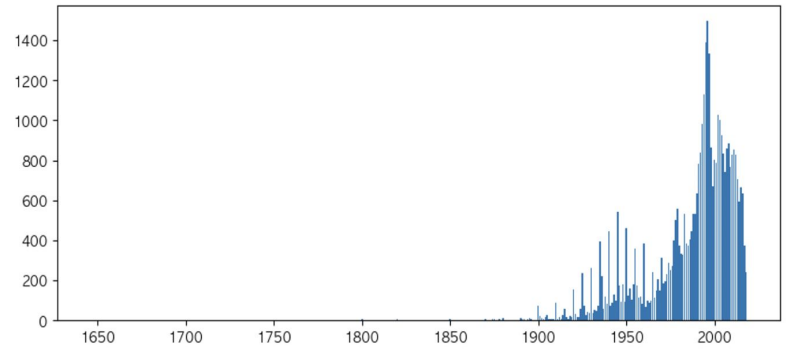


파생변수

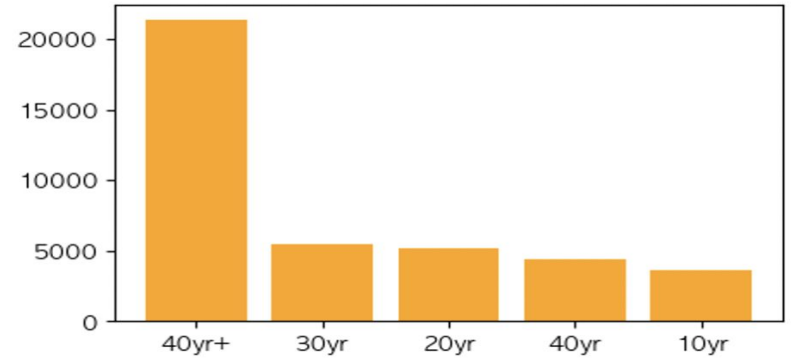
건물 관련 변수1

- **건물승인일자(dt_of_athrztn) → 건물 연령(bld_athr_grp)**
  단순히 건물승인일자로만 보기에는 분포가 넓고, 결측치가 많아 단순 평균으로 처리하기에 어려움을 겪음
  건물의 연령을 40년 초과, 40년, 30년, 20년, 10년으로 범주화
  이후, 건물연령에 따른 평균값을 구하여 결측치를 대체

변경 전



변경 후



파생변수

건물 관련 변수2

- **평균 층수(floar) = (지상층수(ttl_grnd_flr) + 지하층수(ttl_dwn_flr)) / 건물 채수(bldng_cnt)**
  지상층수(ttl_grnd_flr)와 지하층수(ttl_dwn_flr)의 결측치를 대체한 뒤 발생하는 이상치(ex. 100층 단독주택)를 해결하기 위해
  평균 층수(floar)라는 파생변수를 생성

건물 관련 변수3

- **건폐율(flr_area_rat) = 연면적(ttl_ar) / 토지면적(lnd_ar)**
- **용적률(bldng_cov_rat) = 건물건축면적(bldng_ar) / 건축면적(lnd_ar)**
  건물의 면적과 관련된 변수들을 새롭게 파생변수로 생성

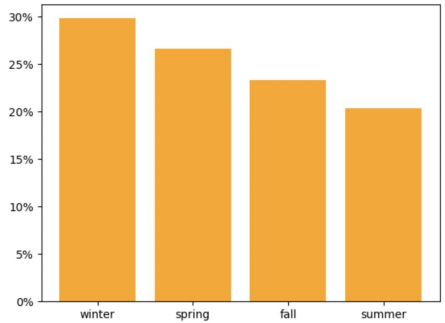
파생변수

날짜 관련 변수

- 계절(season) : 봄(3, 4, 5월) , 여름(6, 7, 8월), 가을(9, 10, 11월), 겨울(12, 1, 2월)
- 평일/주말(weekend) : 평일, 주말
- 시간대(time_of_day) : 낮(6시~17시), 밤(18시~23시), 새벽(자정~5시)

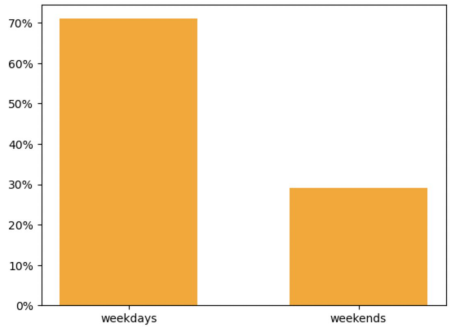
화재발생일시(dt_of_fr) 에서 **(1) 계절별 (2) 평일/주말별 (3) 시간대별** 파생변수를 생성

(1) 계절별 화재 비율



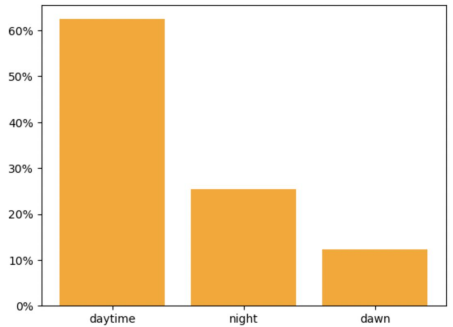
봄, 여름, 가을, 겨울 중 겨울이 가장 큰 비중을, 여름이 가장 작은 비중을 차지함

(2) 평일/주말별 화재 비율



주말보다 평일의 화재 발생 비율이 높음

(3) 시간대별 화재 비율

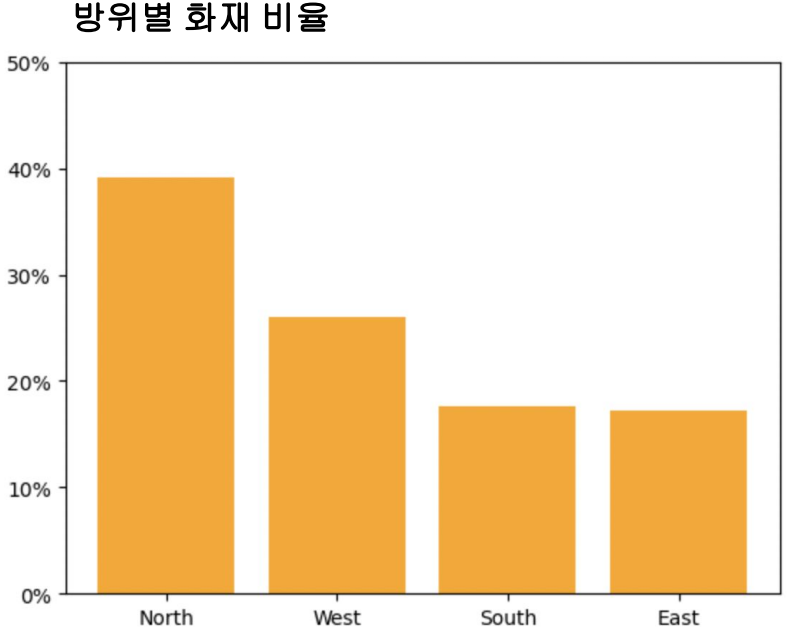


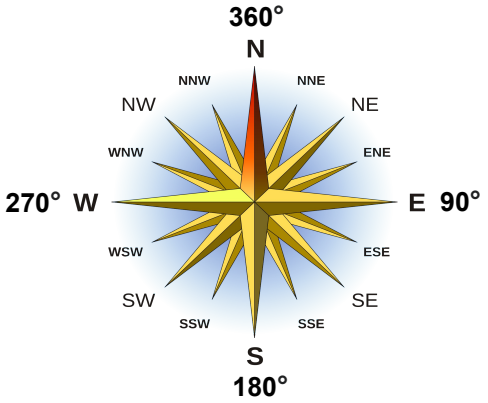
낮, 밤, 새벽 중 낮이 가장 큰 비중을, 새벽이 가장 작은 비중을 차지함

파생변수

풍향

- 풍향(wnd_drctn) 을 **동, 서, 남, 북**으로 그룹화하여 파생변수 생성
- 결측치의 경우 해당 지역의 평균 풍향(360도)을 산정하여 4방위로 변환





16방위를 참고하여 4방위로 그룹화
동, 서, 남, 북 중 북쪽의 화재 발생 비율이 가장 높음.

파생변수 정의

생성 변수	원 변수	정의
islw	lw_13101310 ~ lw 13141011	총 14개의 화재 시설 성능 점수를 하나의 변수로 통합함. 1은 성능 유지 여부가 있고, 0은 결측치를 나타냄.
jmk_grp	jmk	10 개 이상으로 세분화되어 있던 지적상 지목 분류를 {녹지, 대지, 공업용지, 문화교육 용지, 교통시설 용지, 기타}로 구분함.
floar	ttl_grnd_flr, ttl_dwn_flr	층수 합을 건물 채수로 나눠서 올림한 "평균 층수". 단독주택 층수 합이 이상하여 조정한 값.
bld_athr_grp	dt_of_athrztn	건물 승인 일자를 기반으로 건물 연령 도출.
flr_area_rat	ttl_ar, lnd_ar	연면적
bldng_cov_rat	bldng_ar, lnd_ar	건폐율

최종 변수 선택

변수 목록	결측 과다 / 중복 변수 삭제	범주화 또는 파생변수 변환
lw_	lw_	**trgt_crtr** 결측여부 이진분류
prcpttn	prcpttn	
trgt_crtr	trgt_crtr	
fr_fghtng_fclt_spcl_css_6_yn	fr_fghtng_fclt_spcl_css_6_yn	
fr_fghtng_fclt_spcl_css_5_yn	fr_fghtng_fclt_spcl_css_5_yn	
cltrl_hrtg_yn	cltrl_hrtg_yn	
blk_dngrs_thng_mnfctr_yn	blk_dngrs_thng_mnfctr_yn	
slf_fr_brgd_yn	slf_fr_brgd_yn	
dngrs_thng_yn	dngrs_thng_yn	
us_yn	us_yn	**us_yn** 이진분류

최종 변수 선택

변수 목록	결측 과다 / 중복 변수 삭제	범주화 또는 파생변수 변환
bldng_us_clssfctn	bldng_us_clssfctn	**bldng_us_clssfctn** 용도 범주화
bldng_us	bldng_us	
bldng_archtctr	bldng_archtctr	**bldng_archtctr** 건축 구조 범주화
gas_energy_us	gas_energy_us	
ele_energy_us	ele_energy_us	
dt_of_athrztn	dt_of_athrztn	**dt_of_athrztn** [파생]노후도 범주화
bldng_ar_prc	bldng_ar_prc	**bldng_ar_prc** 범주화
ttl_dwn_flr	ttl_dwn_flr	**floar** [파생]평균 층수 범주화
ttl_grnd_flr	ttl_grnd_flr	: 지상+지하 층수 합 / 건물채수
lnd_us_sttn_nm	lnd_us_sttn_nm	
emd_nm	emd_nm	

최종 변수 선택

변수 목록	결측 과다 / 중복 변수 삭제	범주화 또는 파생변수 변환
rgnl_ar_nm2	rgnl_ar_nm2	
rd_sd_nm	rd_sd_nm	**rd_sd_nm 범주화**
rgnl_ar_nm	rgnl_ar_nm	
hm_cnt	hm_cnt	**hm_cnt**
wnd_drctn	wnd_drctn	**wnd_dir [파생]4방위 범주화**
fr_mn_cnt	fr_mn_cnt	**fr_mn_cnt**
wnd_spd	wnd_spd	**wnd_spd**
hmdt	hmdt	**hmdt**
tmprtr	tmprtr	**tmprtr**
dt_of_fr	dt_of_fr	**[파생] season(사계절)/ weekend (주말) / time of day밤-낮 파생변수 범주화**

최종 변수 선택

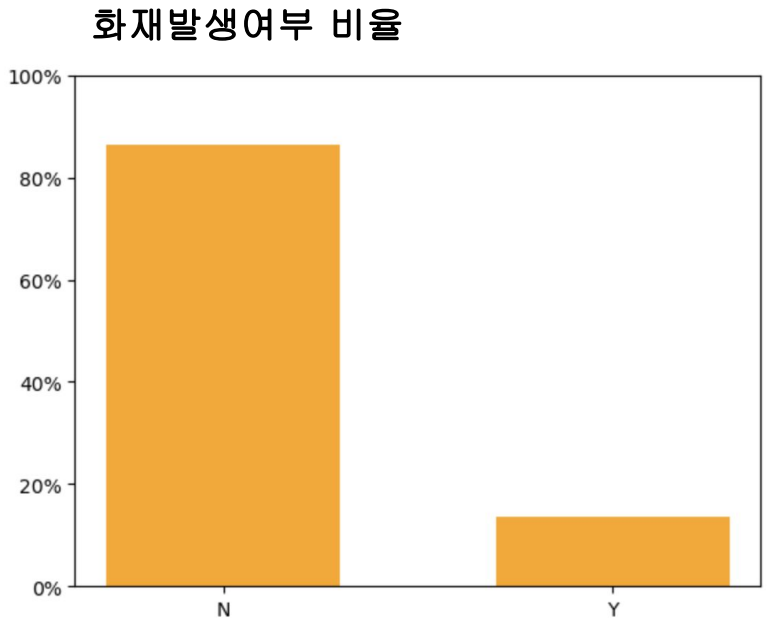
변수 목록	결측 과다 / 중복 변수 삭제	범주화 또는 파생변수 변환
ahsm_dstnc	ahsm_dstnc	**ahsm_dstnc**
bldng_cnt	bldng_cnt	**bldng_cnt [파생]** 평균 층수 변수 추가
bldng_ar	bldng_ar	**bldng_ar [파생]** 건폐율 변수 추가
cctv_dstnc	cctv_dstnc	**cctv_dstnc**
cctv_in_100m	cctv_in_100m	**cctv_in_100m**
fr_sttn_dstnc	fr_sttn_dstnc	**fr_sttn_dstnc**
fr_wthr_fclt_dstnc	fr_wthr_fclt_dstnc	**fr_wthr_fclt_dstnc**
fr_wthr_fclt_in_100m	fr_wthr_fclt_in_100m	**fr_wthr_fclt_in_100m**
jmk	jmk	**jmk_grp** 대분류 범주화
lnd_ar	lnd_ar	**lnd_ar [파생]** 건폐율/용적률 변수 추가

최종 변수 선택

변수 목록	결측 과다 / 중복 변수 삭제	범주화 또는 파생변수 변환
mlt_us_yn	mlt_us_yn	**mlt_us_yn 범주화**
no_tbc_zn_dstnc	no_tbc_zn_dstnc	**no_tbc_zn_dstnc**
sft_emrgnc_bll_dstnc	sft_emrgnc_bll_dstnc	**sft_emrgnc_bll_dstnc**
tbc_rtl_str_dstnc	tbc_rtl_str_dstnc	**tbc_rtl_str_dstnc**
ttl_ar	ttl_ar	**ttl_ar [파생] 용적률 변수 추가**

결측 과다 변수 및 중복 변수 처리, 파생변수 변환 등 데이터 전처리를 통해
최초 179개의 변수에서 최종 93개 변수로 정리

오버 샘플링



1. 화재 발생과 관련된 데이터는 실제 발생 건수보다 발생하지 않은 건수가 많은 불균형한 구조로, 샘플링을 통해 균형을 맞춰줄 필요가 있음.

2. 양성 (화재 발생) 건수가 전체 데이터에 비해 적으므로 언더 샘플링은 부적합, 화재 발생의 경우 정밀도보다 재현율이 중요하기 때문에 **오버 샘플링(SMOTE)** 진행

모델 선정

1. LGBM

2. XGBoost

3. Random Forest

모델링

결측값 대체

- 과적합을 막기 위해 EDA 과정에서 평균치를 통해 결측값을 대체할 때 Train, Validation, Test에 대해 따로 진행

하이퍼파라미터 최적화

- Validation data set를 이용하여 베이지안최적화 기법으로 각 모델의 하이퍼파라미터 도출,
  모든 최적화 과정에서 f1 score를 최대화하는 방향으로 파라미터 최적화 진행

최종 모델 선택

- f1 score를 기준으로 평가하되, recall(재현율)이 높은 모델을 선택

모델 선정

1. LGBM

2. XGBoost

3. Random Forest

분류성능평가지표

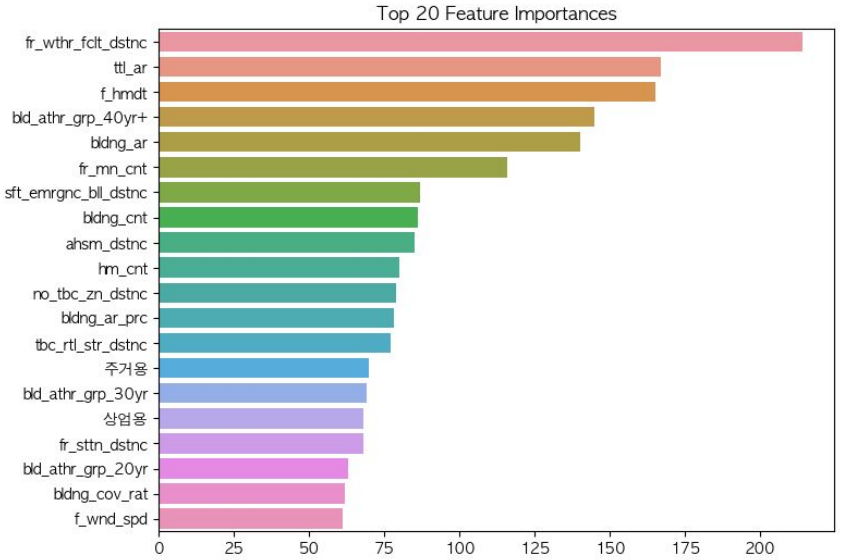
LGBM

	Accuracy	Recall	Precision	F1_score
Raw Data	0.75	0.59	0.40	0.48
SMOTE	0.74	0.68	0.39	0.50
Under Sampling	0.52	0.89	0.27	0.41

모델 선정

LGBM

```
#목적함수 생성
def lgbm_cv(learning_rate, num_leaves, max_depth, min_child_weight, colsample_bytree, feature_fraction, bagging_fra
    model = lgb.LGBMClassifier(objective='binary',
                               learning_rate=learning_rate,
                               n_estimators = 300,
                               #boosting = 'dart',
                               num_leaves = int(round(num_leaves)),
                               max_depth = int(round(max_depth)),
                               min_child_weight = int(round(min_child_weight)),
                               colsample_bytree = colsample_bytree,
                               feature_fraction = max(min(feature_fraction, 1), 0),
                               bagging_fraction = max(min(bagging_fraction, 1), 0),
                               lambda_l1 = max(lambda_l1, 0),
                               lambda_l2 = max(lambda_l2, 0)
                              )
    scoring = {'f1_score': make_scorer(f1_score)}
    result = cross_validate(model, X_train_over, y_train_over, cv=5, scoring=scoring)
    f1 = result["test_f1_score"].mean()
    return f1

# 입력값의 탐색 대상 구간
pbounds = {'learning_rate' : (0.0001, 0.05),
           'num_leaves': (300, 600),
           'max_depth': (2, 10),
           'min_child_weight': (30, 100),
           'colsample_bytree': (0, 0.99),
           'feature_fraction': (0.0001, 0.99),
           'bagging_fraction': (0.0001, 0.99),
           'lambda_l1' : (0, 0.99),
           'lambda_l2' : (0, 0.99),
          }
lgbmBO = BayesianOptimization(f = lgbm_cv, pbounds = pbounds, verbose = 2, random_state = 0 )
```



모델 선정

LGBM

```
#목적함수 생성
def lgbm_cv(learning_rate, num_leaves, max_depth, min_child_weight, colsample_bytree, feature_fraction, bagging_fra
    model = lgb.LGBMClassifier(objective='binary',
                               learning_rate=learning_rate,
                               n_estimators = 300,
                               #boosting = 'dart',
                               num_leaves = int(round(num_leaves)),
                               max_depth = int(round(max_depth)),
                               min_child_weight = int(round(min_child_weight)),
                               colsample_bytree = colsample_bytree,
                               feature_fraction = max(min(feature_fraction, 1), 0),
                               bagging_fraction = max(min(bagging_fraction, 1), 0),
                               lambda_l1 = max(lambda_l1, 0),
                               lambda_l2 = max(lambda_l2, 0)
                              )
    scoring = {'f1_score': make_scorer(f1_score)}
    result = cross_validate(model, X_train_over, y_train_over, cv=5, scoring=scoring)
    f1 = result["test_f1_score"].mean()
    return f1


# 입력값의 탐색 대상 구간
pbounds = {'learning_rate' : (0.0001, 0.05),
           'num_leaves': (300, 600),
           'max_depth': (2, 10),
           'min_child_weight': (30, 100),
           'colsample_bytree': (0, 0.99),
           'feature_fraction': (0.0001, 0.99),
           'bagging_fraction': (0.0001, 0.99),
           'lambda_l1' : (0, 0.99),
           'lambda_l2' : (0, 0.99),
          }
lgbmBO = BayesianOptimization(f = lgbm_cv, pbounds = pbounds, verbose = 2, random_state = 0 )
```

#하이퍼파라미터 최적화 결과	Validation Set
Accuracy	0.83
Recall	0.38
Precision	0.58
F1-Score	0.46

모델 선정

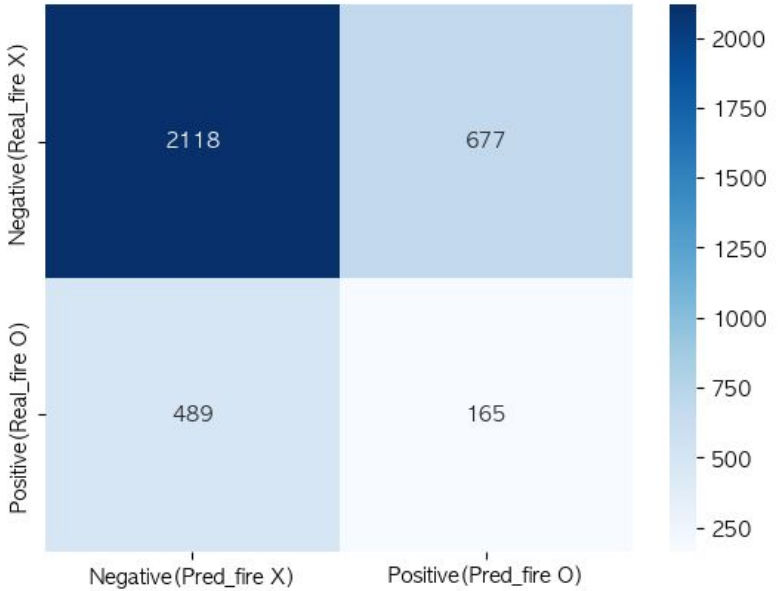
1. LGBM

2. XGBoost

3. Random Forest

분류성능평가지표

XGBoost

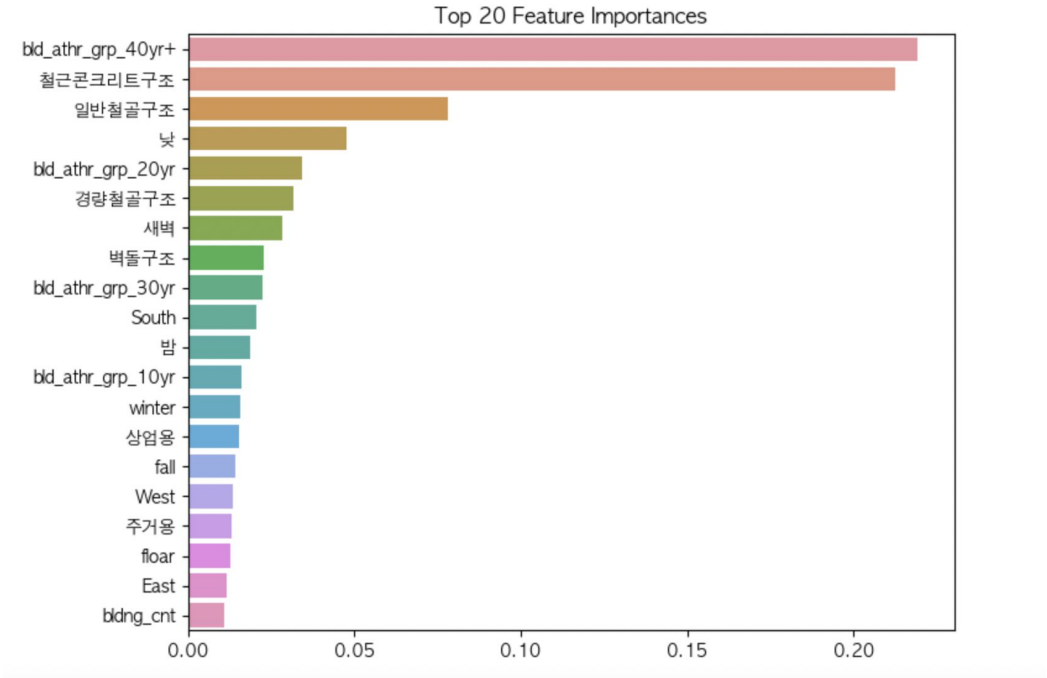


#하이퍼파라미터 최적화 결과	Validation Set
Accuracy	0.78
Recall	0.58
Precision	0.44
F1-Score	0.50

Feature Importance

XGBoost

```
#xgboosting

from xgboost import XGBClassifier

clf_xgb = XGBClassifier(n_estimators=100, learning_rate=0.1, max_depth=10)
clf_xgb.fit(X_train_over, y_train_over)
pred_xgb = clf_xgb.predict(val_x)
# 하이퍼 파라미터 최적화
# 베이지안 최적화

xgb_parameter_bounds = {'max_depth' : (5, 10),'n_estimators' : (30, 100) }
def xgb_bo(max_depth, n_estimators):
    xgb_params = {'max_depth' : int(round(max_depth)), 'n_estimators' : int(round(n_estimators))}
    xgb = XGBClassifier(**xgb_params)
    X_train, X_valid, y_train, y_valid = train_test_split(X_train_over,y_train_over,test_size = 0.2)
    xgb.fit(X_train,y_train)

    score = f1_score(y_valid, xgb.predict(X_valid))

    return score

BO_xgb = BayesianOptimization(f = xgb_bo, pbounds = xgb_parameter_bounds, random_state = 0)
BO_xgb.maximize(init_points = 5, n_iter = 5)
```



모델 선정

1. LGBM

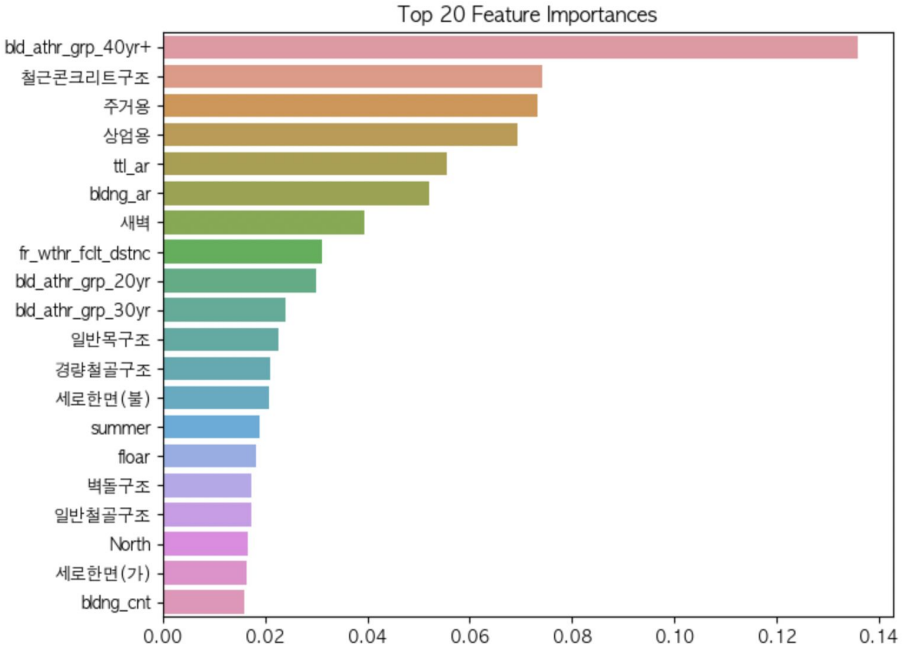
2. XGBoost

3. Random Forest

Feature Importance

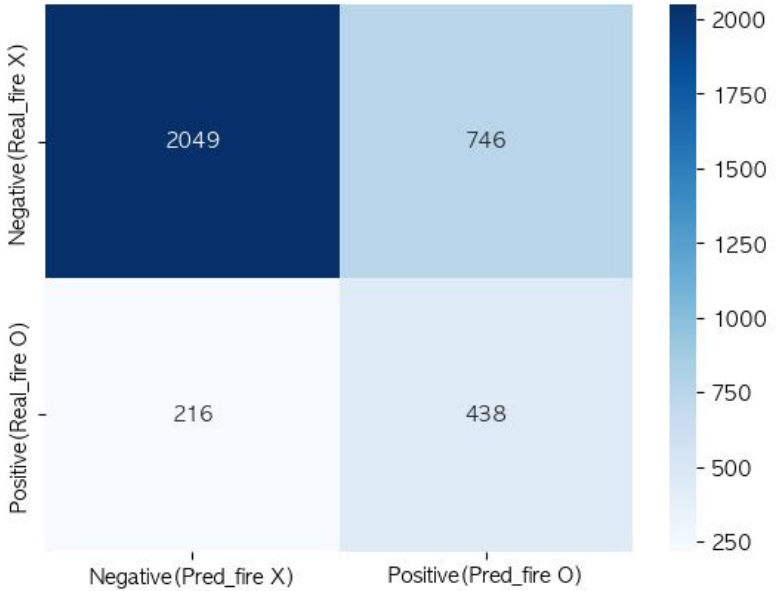
Random Forest

```
clf = RandomForestClassifier(n_estimators=100, criterion = "entropy", max_depth=10, random_state=0)
clf.fit(X_train_over,y_train_over)
pred_rf = clf.predict(val_x)

rf_parameter_bounds = {'max_depth' : (5, 10),'n_estimators' : (30, 100)}
def rf_bo(max_depth, n_estimators):
    rf_params = {'max_depth' : int(round(max_depth)), 'n_estimators' : int(round(n_estimators))}
    rf = RandomForestClassifier(**rf_params)
    X_train, X_valid, y_train, y_valid = train_test_split(X_train_over,y_train_over,test_size = 0.2)
    rf.fit(X_train,y_train)

    score = f1_score(y_valid, rf.predict(X_valid))

    return score
```



분류성능평가지표

Random Forest



#하이퍼파라미터 최적화 결과	Validation Set
Accuracy	0.72
Recall	0.67
Precision	0.37
F1-Score	0.48

Feature Importance

LGBM	Random Forest	XGBoost
fr_wthr_fclt_dstnc	bld_athr_grp	bld_athr_grp
ttl_ar/bldng_ar	bldng_archtctr	bldng_archtctr
f_hmdt	bldng_us_classfn	bldng_us_clssfctn
bld_athr_grp	ttl_ar/bldng_ar	time_of_day
fr_mn_cnt	time_of_day	wnd_drctn

세 모델에서 TOP 20 Feature Importance를 확인한 결과
특정 변수들이 화재 발생에 있어 높은 설명력을 가진다는 것을 확인

세가지 모델 모두 **건물 연령(bld_athr_grp)**이 중요 변수로 나타나 건물 노후화가 화재 발생과 밀접한 관계임을 확인
그 외, 2개 모델에서 **건물 구조(bldng_archtcr), 토지 및 건물 면적(ttl_ar/bldng_ar), 시간대(time_of_day)**도
화재 발생과 연관성이 높음을 알 수 있음

기존 모형과의 차별점

기존 모형	accuracy	recall	precision	f1_score
로지스틱	0.74	0.38	0.41	0.39
의사결정나무	0.76	0.11	0.12	0.02
Random Forest	**0.45**	**0.50**	**0.20**	**0.29**
XGBoost	0.71	0.40	0.61	0.48

현재 모형	accuracy	recall	precision	f1_score
LGBM	0.83	0.38	0.58	0.46
Random Forest	**0.72**	**0.67**	**0.37**	**0.48**
XGBoost	0.78	0.58	0.44	0.50

최종 모델 (RF)	Test Set
Accuracy	0.70
Recall	0.70
Precision	0.34
F1-Score	0.46

의의 및 한계점

데이터 정교화:

- 결측치가 과다하고 라벨링 오류도 포함된 공공 데이터의 처리
  ⇒ 변수별 결측치 빈도와 결측치의 화재 발생 빈도를 고려하여 변수를 제한적으로 제외함
- 범주가 다양하고 소방 및 건축물 자료가 체계없이 뒤섞인 원데이터의 가공
  ⇒ 유사한 변수의 축소, 1차 가공을 통한 카테고리화 및 파생변수 변환으로 핵심 변수 도출

모델 성능 개선:

- Random Forest, XGB 모두 모형 성능 지표가 소폭 개선됨
- 성능 향상을 위해 건물 용도에 따라 다르게 모델링을 진행하자는 의견이 있었음
  ⇒ 건물 용도 뿐만 아니라 다양한 변수들에 대해 세분화하여 모델링을 진행하면 성능 향상에 도움이 될 것

하이퍼 파라미터 최적화 수행:

- LGBM의 경우, 베이지안 최적화 시 f1 score가 오히려 낮아짐
  ⇒ Oversampled된 Train Set에 비해 Validation, Test Set은 불균형한 데이터

발표 후 피드백 사항

Train / Val / Test EDA 관련:

- 데이터의 전처리 과정에서 Test Set의 경우 알지 못하는 데이터이기 때문에 결측치 처리 과정에서 데이터 사용 불가
  ⇒ Train Set에서 실행한 전처리 과정으로 대체하려 했으나, 해당 데이터에 결측이 있는 열들이 너무 많고 결측값 자체가 많다는 점, Train Set에서 기존 데이터를 활용해서 지역별 평균을 구하는 방법으로 전처리하여 따로 값을 저장해서 사용하기 어렵다는 점에서 수정에 어려움이 있었음

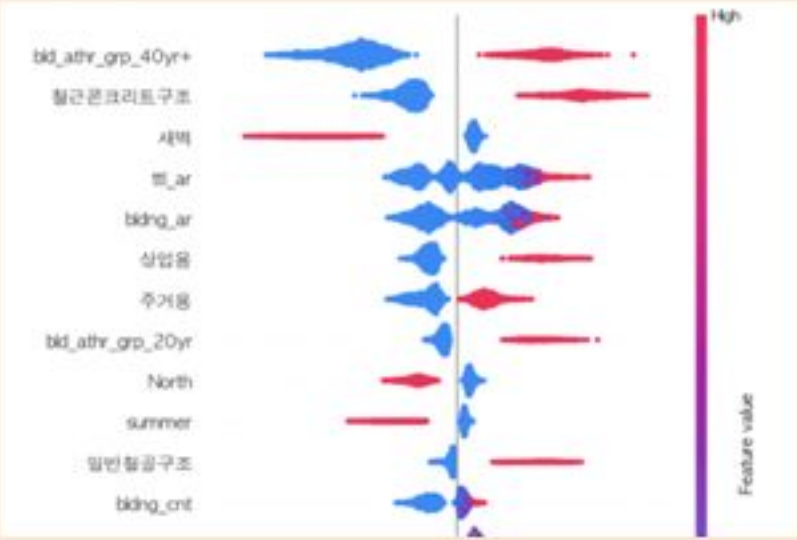
모델 성능 개선:

- LGBM의 경우, 베이지안 최적화 시 f1 score가 오히려 낮아짐
- 이를 개선하기 위해 다른 팀에서 사용한 Optuna 사용
  ⇒ 사용 결과 기존 방식보다 준수한 결과 도출, 그러나 여전히 f1_score와 recall이 낮아지는 결과, RF 성능또한 떨어져 기존 최종 모델 사용

LGBM	accuracy	recall	precision	f1_score
최적화 진행 X	0.74	0.68	0.39	0.50
기존 방식 최적화	0.83	0.38	0.58	0.46
Optuna 사용	0.76	0.59	0.41	0.49

발표 후 피드백 사항

SHAP Value : Random Forest 모델에 대한 해석 추가



SHAP Dot Plot : SHAP Value Feature 간의 상관관계와 Feature들이 예측에 어떤 방향으로 영향을 미쳤는지 알 수 있음

특성(y축)은 예측에 미치는 영향력(중요도)에 따라 정렬됨
- 40년 이상된 건물, 건물의 자재, 새벽시간 … 순으로 결과값 예측에 큰 영향을 끼침.

SHAP Value가 양수 = 예측값을 증가시킴
- 40년 이상인 건물일수록, 철근콘크리트구조일수록 예측값을 증가시킴.
  => 전체적으로 볼 때,40년 이상인 건물변수와 철근콘크리트구조는 양의 상관성으로 영향을 미침
- 새벽이 아닌 시간대에 예측값을 증가시킴
  => 전체적으로 볼 때, 새벽은 음의 상관성으로 영향을 미침

감사합니다