Machine Learning IC-PBL

최종 발표

TEAM 6

조원

로봇공학과 2015041694 신상혁

로봇공학과 2016007183 조정용

산업경영공학과 2017010146 이시현

로봇공학과 2018043490 함서연

CONTENTS

01 전처리

큰 틀

1. 결측치 값 처리
2. 정규화
3. 변수 축소
4. 이상치 보정

결측치 값 처리

1. 결측치 값 대체

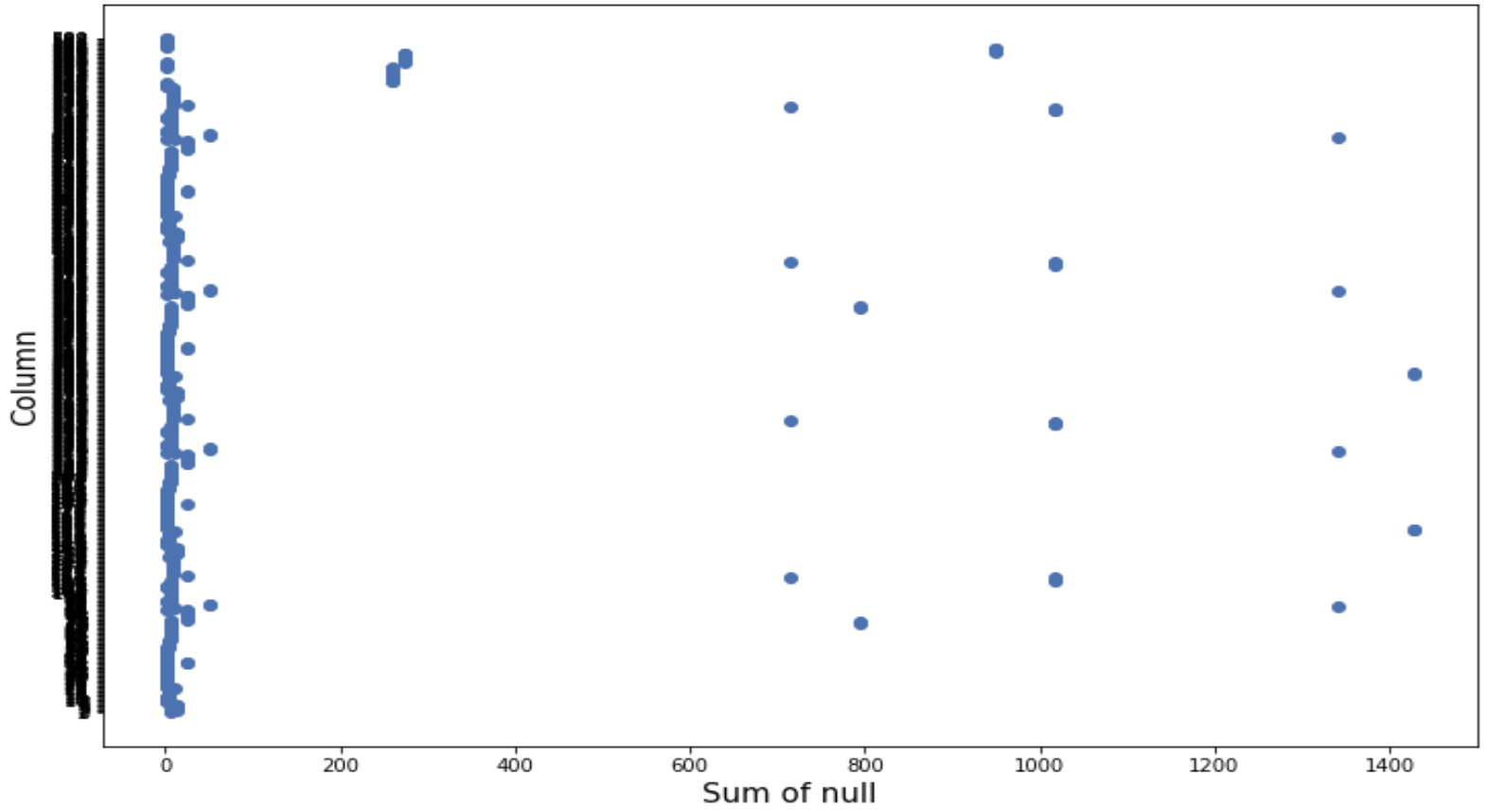
- 결측치의 값을 대체하여 처리

2. 결측치가 많은 변수 삭제

- 결측값 대체가 의미 없을 정도로 결측치가 많음 변수 제거

01 전처리

결측치 값 처리



- 일부 변수에서 비약적으로 많은 결측치를 가지고 있음

01 전처리

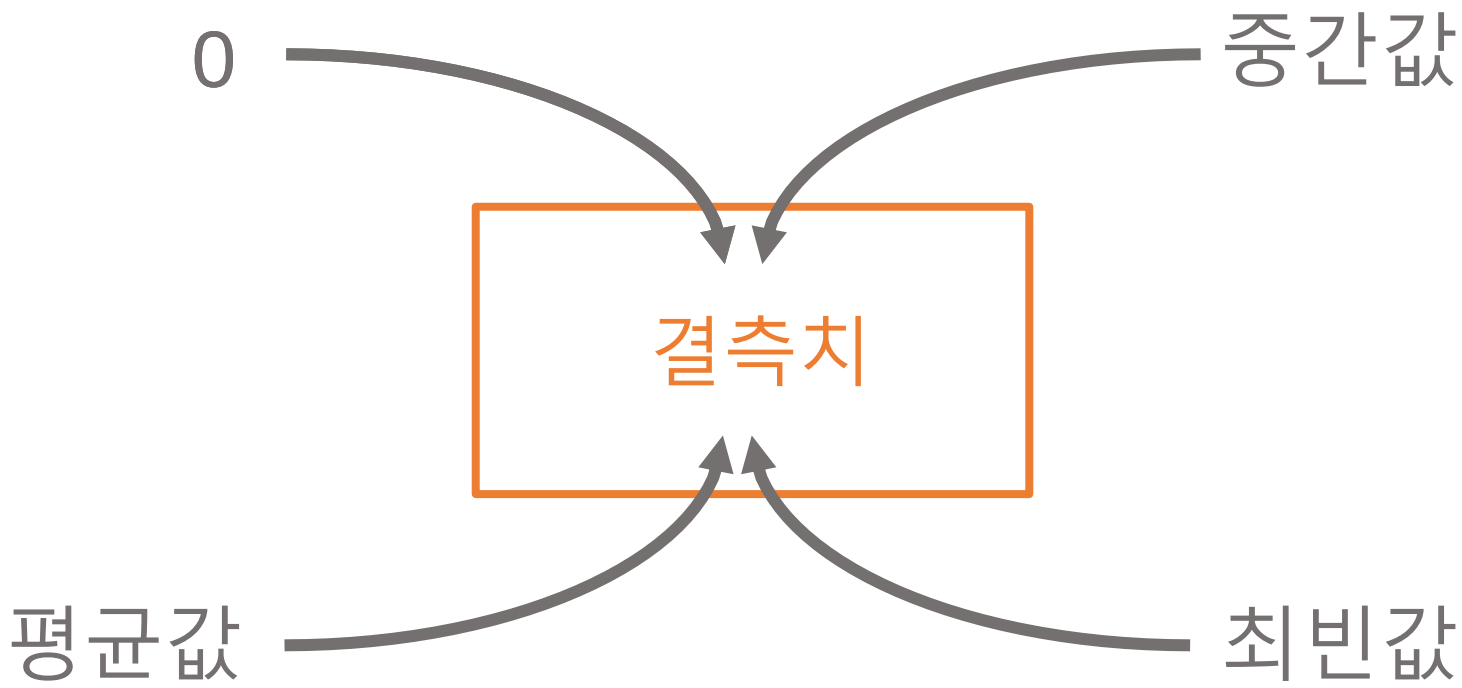
결측치 값 처리

결측치 값 개수의 유일값	해당 feature	비율
0	20, 86, 87, 88, 113, 114, 115, 116, 117, 119, 120, 156, 221, 222, 223, 248, 249, 250, 251, 252, 254, 255, 291, 359, 360, 361, 386, 387, 388, 389, 390, 392, 393, 429, 493, 494, 495, 520, 521, 522, 523, 524, 526, 527, 570, 571, 572, 573, 574, 575, 576, 577, 590	0%
1	32, 33, 34, 35, 36, 37, 38, 39, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 83, 170, 171, 172, 173, 174, 175, 176, 177, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 218, 305, 306, 307, 308, 309, 310, 311, 312, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 356, 441, 442, 443, 444, 445, 446, 447, 448, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 490, 558, 559, 560, 561, 582, 583, 584, 585, 586, 587, 588, 589	0.06%
2	8, 9, 10, 11, 12, 21, 22, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 92, 93, 103, 104, 144, 145, 146, 147, 148, 159, 160 161, 162, 163, 164, 165, 167, 168, 169, 227, 228, 238, 239, 279, 280, 281, 282, 283, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 365, 366, 376, 377, 417, 418, 419, 420, 421, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 499, 500, 510, 511, 542, 543, 544, 545	0.13%
3	13, 14, 15, 16, 17, 18, 149, 150, 151, 152, 153, 154, 284, 285, 286, 287, 288, 289, 422, 423, 424, 425, 426, 427	0.19%
4	53, 54, 55, 56, 57, 58, 190, 191, 192, 193, 194, 195, 326, 327, 328, 329, 330, 331, 462, 463, 464, 465, 466, 467	0.26%
5	135, 270, 408	0.32%
6	0, 60, 61, 62, 66, 67, 68, 69, 70, 71, 74, 91, 94, 95, 96, 97, 98, 99, 100, 101, 102, 105, 106, 107, 108, 136, 197, 198, 199, 203, 204, 205, 206, 207, 208, 209, 226, 229, 230, 231, 232, 233, 234, 235, 236, 237, 240, 241, 242, 243, 271, 333, 334, 334, 339, 340, 341, 342, 343, 344, 347, 364, 367, 368, 369, 370, 371, 372, 373, 374, 375, 378, 379, 380, 381, 409, 469, 470, 471, 475, 476, 477, 478, 479, 480, 481, 498, 501, 502, 503, 504, 505, 506, 507, 508, 509, 512, 513, 514, 515	0.38%
7	1, 59, 63, 64, 65, 137, 196, 200, 201, 202, 272, 332, 336, 337, 338, 410, 468, 472, 473, 474	0.45%
8	132, 133, 134, 267, 268, 269, 405, 406, 407, 539, 540, 541	0.51%
9	7, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 143, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 278, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 416, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538	0.57%
10	19, 155, 290, 428	0.64%
12	84, 219, 357, 491	0.77%
14	2, 3, 4, 6, 138, 139, 140, 141, 142, 273, 274, 275, 276, 277, 411, 412, 413, 414, 415	0.89%
24	40, 41, 75, 76, 77, 78, 79, 80, 81, 82, 118, 178, 210, 211, 212, 213, 214, 215, 216, 217, 253, 313, 314, 348, 349, 350, 351, 352, 353, 354, 355, 391, 449, 450, 482, 483, 484, 485, 486, 487, 488, 489, 525	1.53%
51	89, 90, 224, 225, 362, 363, 496, 497	3.25%

결측치가 전체의 16%이상 차지하는 변수 삭제

01 전처리

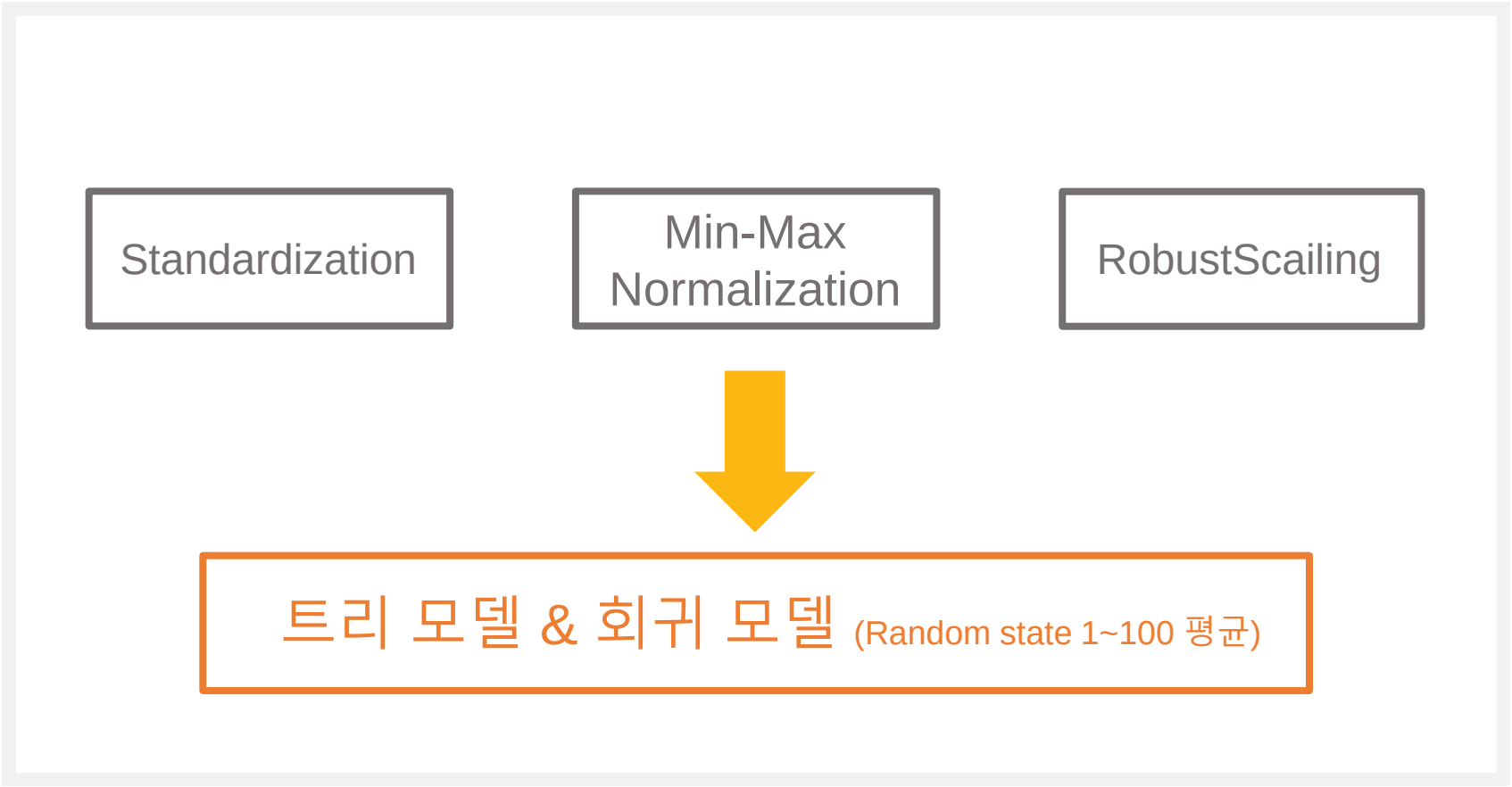
결측치 값 처리



- 최빈값은 보통 범주형 변수일 때 대체하는 방식이라 사용하지 않음
- 0은 변수의 특성을 반영하지 못하기 때문에 사용하지 않음
- 평균값과 중간값은 변수의 특성을 반영하고 수치형 변수일 때 대체할 수 있는 값이므로, 대체했을 때의 성능 비교를 통해 성능이 높게 나온 중간값 대체를 채택

01 전처리

정규화



- 다음의 3가지 정규화 방법을 트리 모델과 회귀 모델에 모두 적용해본 결과, 0.02정도 앞서는 성능으로 Min-Max Normalization을 사용

변수 축소

1. ‘Time’ 변수 제거

- 반도체 공정 데이터는 시간에 따른 변화가 없어, 시계열적 특성은 없다고 판단하여 ‘time’변수를 제거

2. 데이터 값의 변동이 없는 변수 제거

- 데이터가 하나의 값으로만 구성되어 있는 변수는 제거.
- 변수의 특성을 반영하고 있지 않아 성능의 영향을 주지 못함
- 변수 선택과정에서 탈락하기에 알고리즘의 경제성을 위해 변수 제거 진행

3. 알고리즘 내재 방법을 이용한 변수 선택

- 알고리즘 내재 방법을 통해 변수 선택하고 최종 학습시킬 변수 선정

01 전처리

변수 축소

트리기반 알고리즘 내재 방법을 이용한 변수 선택

Random Forest	LGBM
0.6550446936454	0.6798342272062

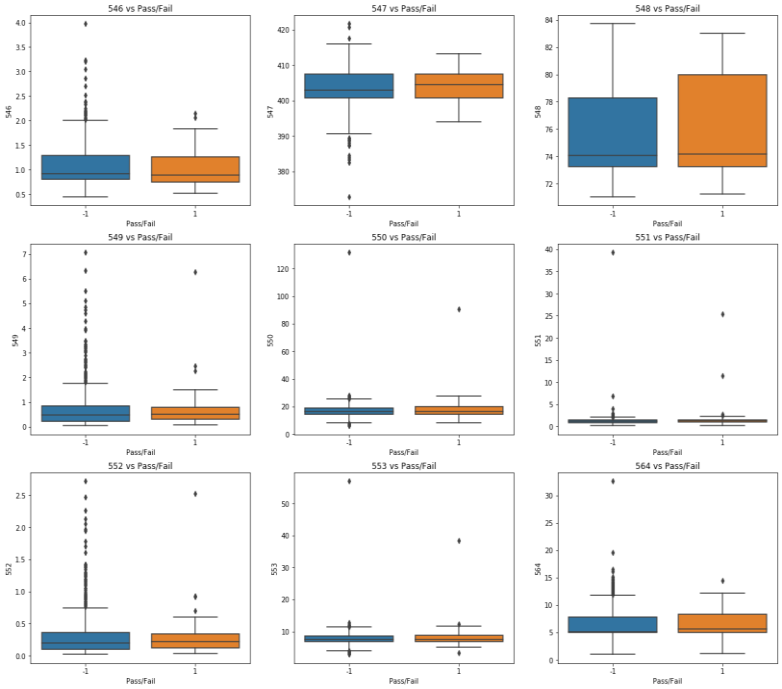
➡ 총 55개

- 트리 기반의 알고리즘은 분기할 때 분기의 기준으로 삼는 변수를 선택하기 때문에, 변수 중요도를 명확히 파악 가능한 신뢰성 있는 변수 선택 방법
- LGBM은 Random Forest의 향상된 모델로, LGBM의 성능이 더 좋았음
- 최종적으로 총 55개의 변수 선택

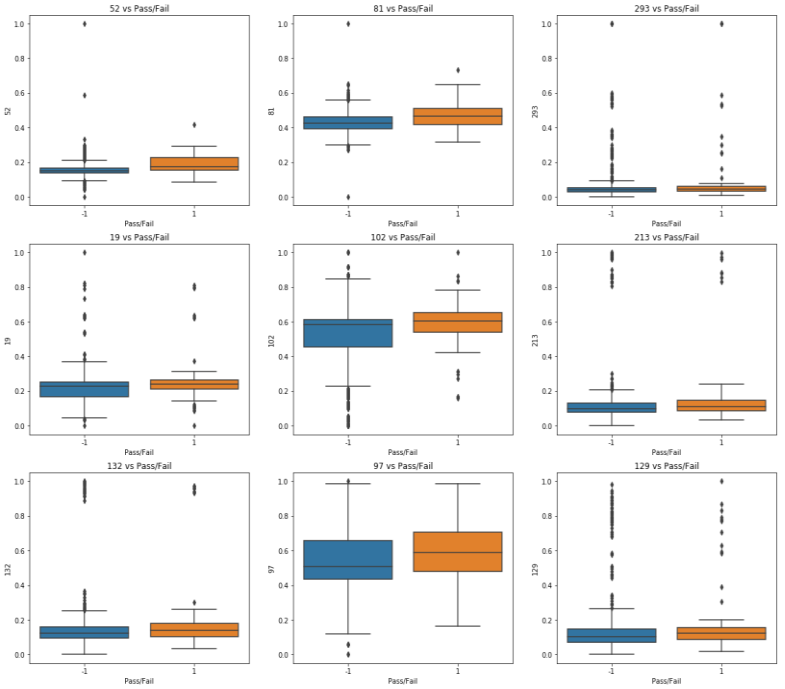
01 전처리

변수 축소

<선택 되지 못한 변수>



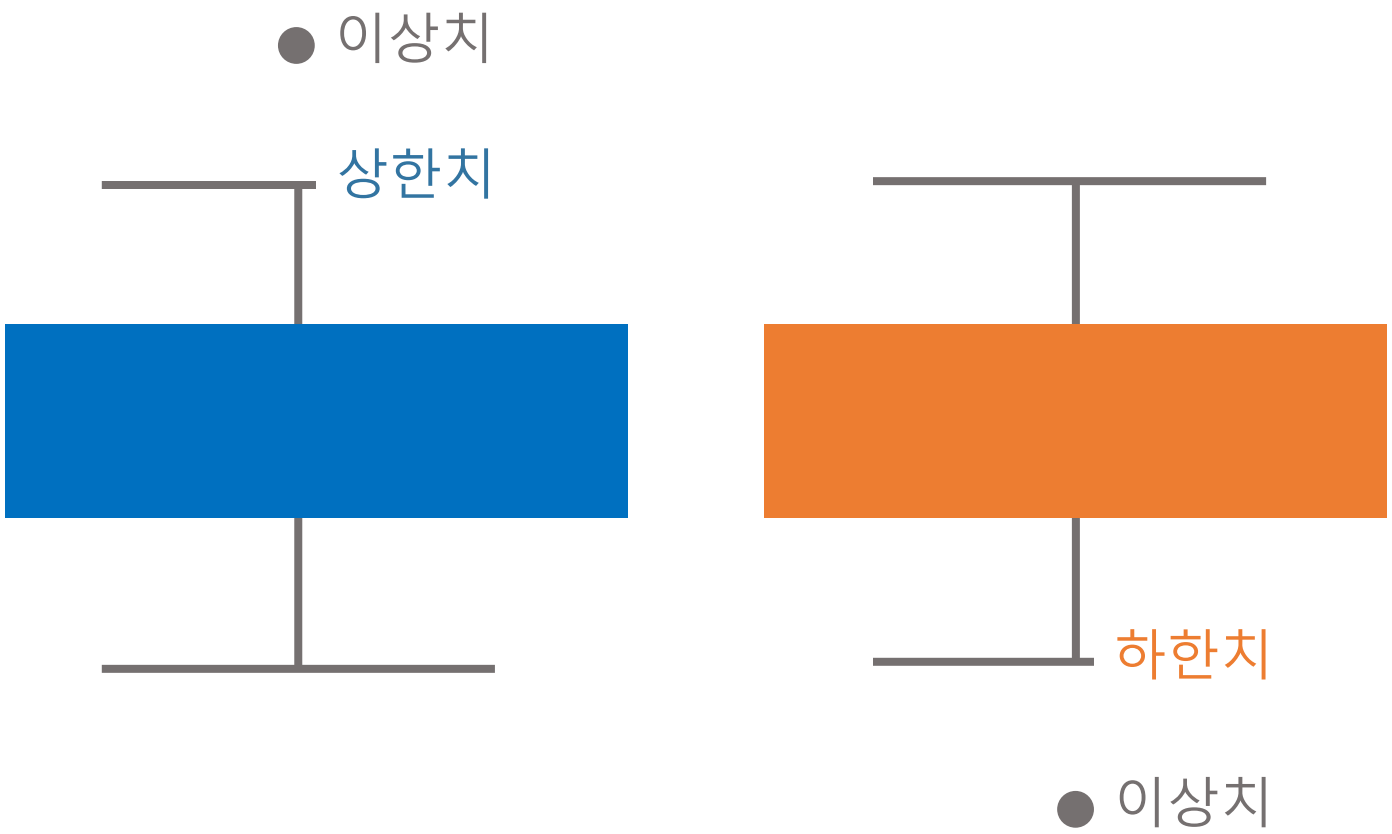
<선택된 변수>



- 선택 되지 못한 변수는 Target 값의 분포에 따른 차이가 없음을 알 수 있음. 즉, Target 변수에 대한 설명을 잘 설명해주지 못함을 의미
- 선택된 변수는 Target 값에 따른 분포에 따른 차이가 상대적으로 있음을 확인.
- Target 변수와 독립변수와의 최대 상관관계가 0.16 정도로 높지 않아 극명한 차이를 보이진 않지만 비교적 차이가 남을 알 수 있음

01 전처리

이상치 보정



- 이상치를 포함한 평균값이나 중간값은 데이터의 대표성을 반영하지 못해 이상치 보정 진행
- 각 이상치에 가까운 바운더리 값으로 대체
- 전체 셀의 5% 이내인 1837개를 보정

02 학습 모델

학습 모델

Bagging Classifier

KNN

Decision tree

Logistic Regression

LDA

XG boost

QDA

LGBM

Random forest

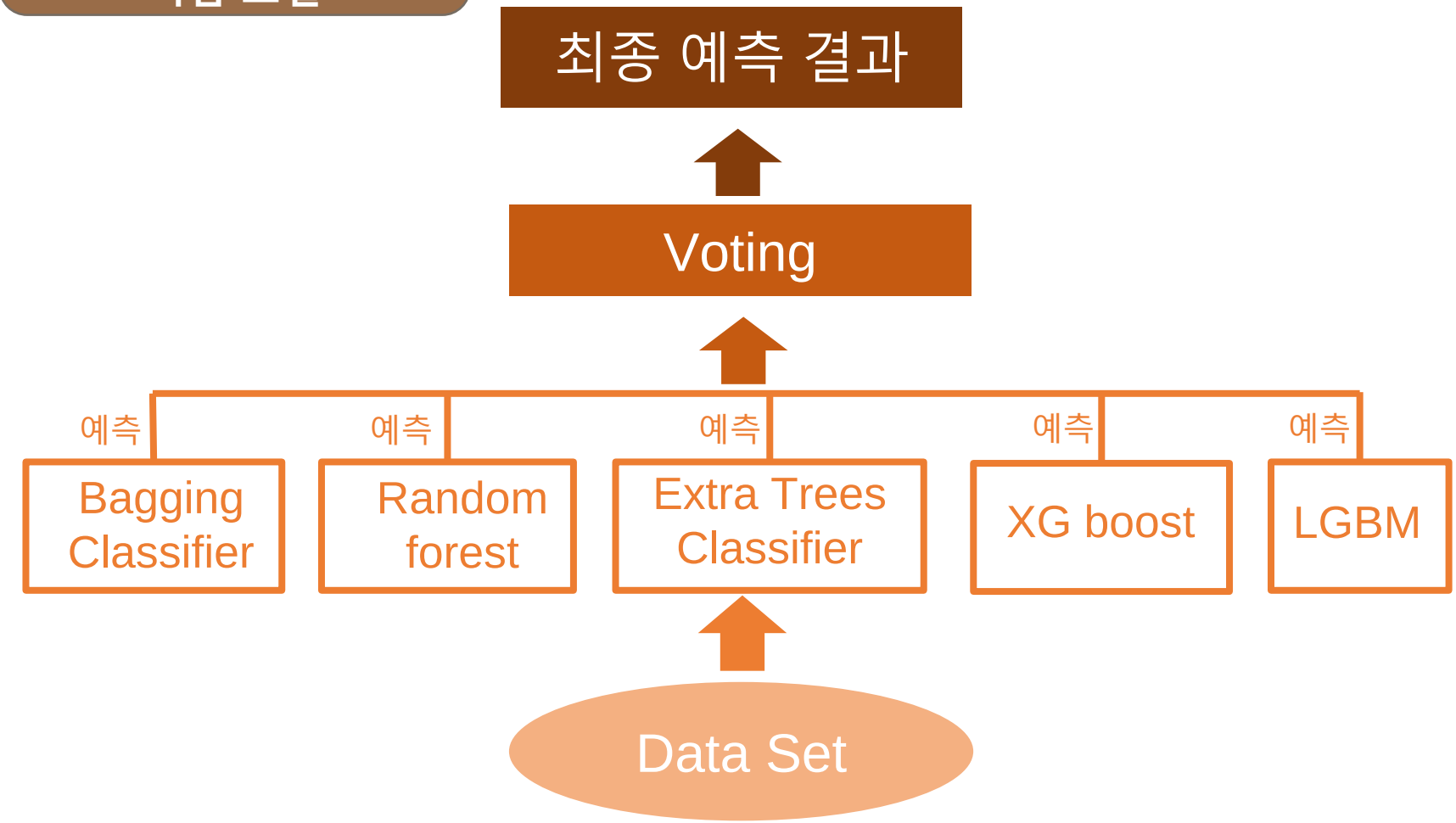
SVM

Extra Trees Classifier

- 11개의 모델 모두 시행 후, 성능이 잘 나온 상위 5개의 모델을 이용해 앙상블 기법 적용

02 학습 모델

학습 모델



- 오버비팅을 방지해 줄 수 있는 앙상블 기법 중, 소프트 보팅 분류기 적용
- 모든 분류기가 예측한 레이블 값의 결정 확률 평균을 구한 뒤 가장 확률이 높은 레이블 값을 최종 결과로 선정하는 소프트 보팅 분류기 적용

03 학습 결과

ROC AUC score

random_state= 1 auc_test = 0.952055907687307

random_state= 2 auc_test = 0.9335283601495205

random_state= 3 auc_test = 0.9440923126929953

random_state= 4 auc_test = 0.9606695920689095

⋮

random_state= 99 auc_test = 0.9522184300341296

random_state= 100 auc_test = 0.9002112790508694

0.9226897448399153

QnA