SAMSUNG SDS Brightics Academy 공모전
Analysis Report
[팀명 : 명량대첩]

김민기
박준수
손희헌

목차 CONTENTS

Part1

1. EDA
2. 모형선정 및 설명

Part2

1. 문제 & 데이터설명
2. 전처리
3. 추가데이터 설명
4. 모형선정 및 설명
5. 한계점 및 향후 개선방향

Part1

1. EDA

	Time	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18
1	1	0.028836...	-0.64982...	0.363123...	-0.75214...	-0.63131...	-0.58881...	-0.90420...	-0.57702...	0.715938...	0.873630...	-0.06328...	1.386802...	1.369457...	-1.41596...	0.705533...	0.577331...	0.577288...	-1.40662...
2	2	0.053882...	-0.64982...	0.347376...	-0.49959...	-0.52866...	-0.48718...	-0.90420...	-0.57696...	0.702009...	-0.04018...	0.564461...	1.386802...	1.368857...	-1.78020...	0.810727...	0.577331...	0.577355...	-1.21264...
3	3	0.065778...	-0.64982...	0.347376...	-0.73279...	-0.58618...	-0.57857...	-0.90420...	-0.57784...	0.703789...	-0.423391...	0.077637...	1.386802...	1.396297...	-1.39084...	0.747610...	0.577331...	0.57732985	-1.29886...
4	4	0.653069...	-1.33202...	1.150375...	-1.70934...	0.473909...	1.220859...	-0.03815...	1.636807...	0.448838...	-0.73930...	-0.071174...	1.733044...	1.38040185	-0.14479...	0.402767...	-0.03245...	0.781490...	4.302893...
5	6	0.033532...	-0.64982...	0.389368...	-0.44059...	-0.60565...	-0.54563...	-0.90420...	-0.57731...	0.706494...	0.932585...	-0.20420...	1.386802...	1.396174...	-1.65460...	1.000076...	0.577331...	0.577345...	-1.39225...
6	8	0.031498...	-0.64982...	0.331628...	-0.67437...	-0.60653...	-0.38303...	-0.90420...	-0.57759...	0.705007...	-0.04018...	-0.33231...	1.386802...	1.396269...	-1.20244...	0.516184...	0.577331...	0.577324...	-1.27730...
7	9	0.274842...	-1.05675...	0.853002...	-1.43813...	0.135310...	-0.22689...	0.011793...	0.055596...	-0.04622...	-0.32666...	0.108882...	1.258454...	0.545066...	-0.05428...	-0.04550...	0.003739...	0.220095...	3.403590...
8	11	0.058577...	-0.64982...	0.357874...	0.113732...	-0.55344...	-0.45933...	-0.90420...	-0.57701...	0.703722...	-0.423391...	-0.114529...	1.386802...	1.368843...	-1.61692...	1.000076...	0.577331...	0.57728339	-1.40662...
9	12	0.041359...	-0.64982...	0.347376...	0.601130...	-0.55875...	-0.46927...	-0.90420...	-0.57748...	0.702843...	0.343030...	-0.37075...	1.386802...	1.396251...	-1.21500...	0.558261...	0.577331...	0.577335...	-1.37788...
10	13	0.068751...	-0.64982...	0.310632...	0.600227...	-0.71007...	-0.43120...	-1.41766...	-0.57742...	0.697794...	-0.12861...	0.000770...	1.386802...	1.396406...	0.782042...	0.495145...	0.577331...	0.577319...	-1.27730...
11	14	0.054664...	-0.64982...	0.310632...	0.883023...	-0.67556...	-0.45923...	-1.41766...	-0.57652...	0.699597...	-1.10137...	-0.42199...	1.386802...	1.395938...	0.957883...	0.242680...	0.577331...	0.577309...	-1.29167...
12	15	-0.05275...	-1.00201...	1.211025...	-1.65470...	-0.07578...	0.309228...	-0.04649...	0.104515...	-0.00791...	-0.42232...	0.134369...	1.297619...	0.41704008	-0.15407...	-0.115995...	1.052354...	0.075531...	3.301423...
13	16	0.021166...	-0.64982...	0.32637986	1.038880...	-0.64193...	-0.28131...	-1.41766...	-0.57649...	0.700521...	0.726241...	-0.05047...	1.386802...	1.395678...	0.782042...	0.368912...	0.577331...	0.577335...	-1.23420...
14	17	0.037289...	-0.64982...	0.357874...	-0.38233...	-0.61892...	-0.37285...	-1.41766...	-0.57641...	0.701580...	1.197885...	0.269804...	1.386802...	1.396006...	0.719242...	0.558261...	0.577331...	0.577324...	-1.19109...
15	18	0.022262...	-0.64982...	0.321130...	-0.63525...	-0.66140...	-0.37280...	-1.41766...	-0.57780...	0.698786...	0.726241...	0.551650...	1.386802...	1.396387...	0.543402...	0.705533...	0.577331...	0.577366...	-1.24857...
16	20	0.68592459	-1.27408...	0.354887...	-2.05606...	0.263241...	1.399410...	-0.24959...	0.539882...	-0.18593...	-0.42661...	0.22204864	1.648682...	1.074885...	-0.49777...	1.009736...	-0.06682...	0.342655...	3.749011...
17	21	0.029776...	-0.64982...	0.321130...	0.03633569	-0.69414...	-0.60914...	-0.91805...	-0.57694...	0.709627...	0.755719...	-0.29388...	1.386802...	1.376202...	0.468042...	0.474106...	0.577331...	0.577324...	-1.60060...
18	22	0.03415901	-0.64982...	0.342127...	0.035804...	-0.68706...	-0.4617511	-0.91805...	-0.57753...	0.709469...	-1.16033...	-0.24264...	1.386802...	1.396283...	0.380121...	1.273580...	0.577331...	0.577335...	-1.62215...

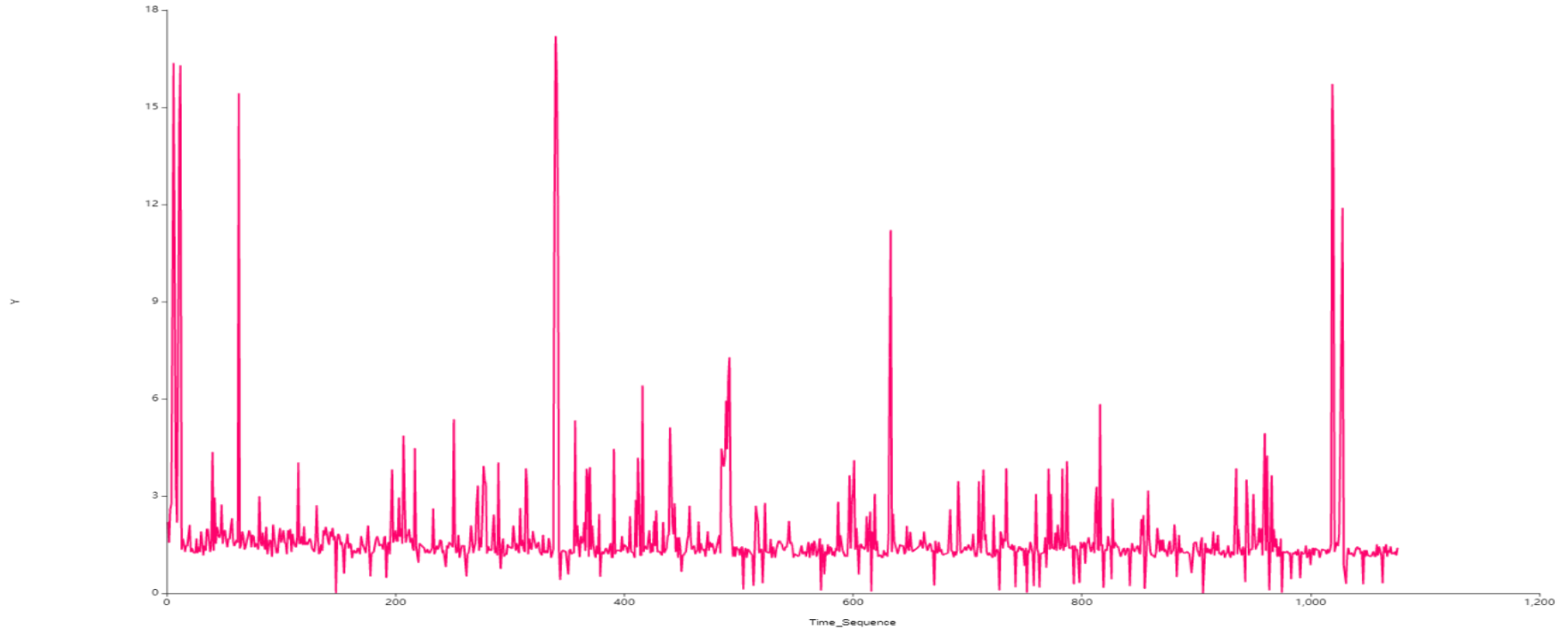
1. EDA

[제품 불량률 예측]

제품을 생산하는 과정에서 발생하는 설비의 센서 측정 데이터와 해당 제품의 불량률 사이의 유의미한 연관성 분석 및 불량률 예측

- 삼성 SDS에서 제공하는 임의의 제품 생산시의 센서 데이터와 수율 데이터 간의 관계를 분석하고 예측 모델링 개발

1. EDA



평가 기준인 WMAE를 고려하여 이상치를 제거 하지 않은 Raw Data를 분석에 사용

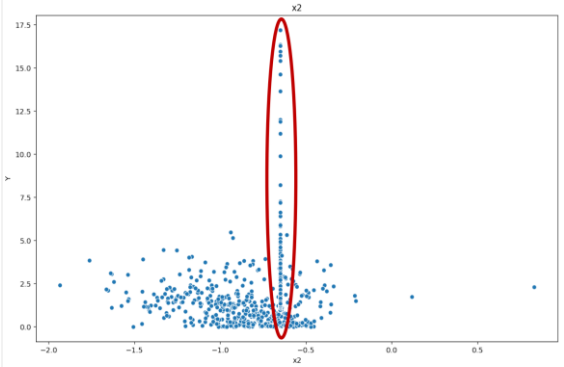
1. EDA

센서 값과 불량률의 산점도를 관찰하여 아래와 같이 3가지 case로 나눌 수 있었다.

Case 1

각 센서의 특정 값에 데이터가 1자 모양으로 형성

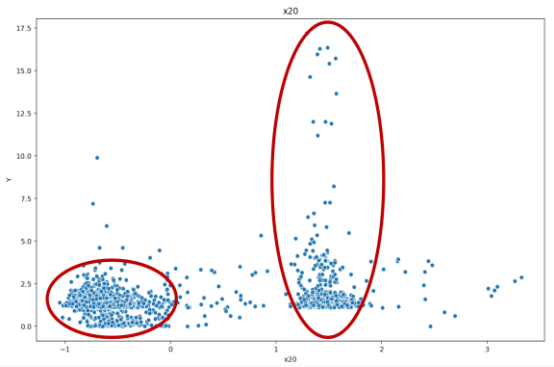
특징 : 전체 데이터의 약 87%가 특정 센서 값에 존재하고 불량률 이상치가 이러한 특정 센서 값에서만 존재



Case 2

각 센서 값의 특정 범위에 따라 두 군집으로 형성

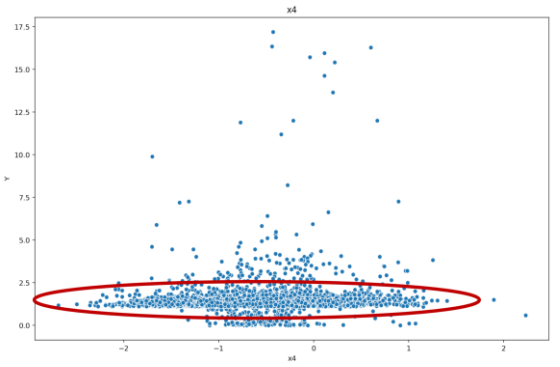
특징 : 대부분의 이상치는 둘 중 하나의 군집에서 발생, 각각의 군집은 음수 영역과 양수 영역으로 구분, 불량률과의 상관 계수 절댓값이 높음



Case 3

데이터가 가로로 길게 늘어진 모양을 형성

특징 : 데이터가 각 센서 값의 범위에 고르게 퍼짐, 불량률과의 상관 계수 절댓값이 낮음



같은 케이스에 속한 변수끼리 서로 연관이 있을 가능성이 높음 → 상관 분석 필요

1. EDA

다음과 같이 독립변수 간 correlation이 높은 경우가 다수 존재한다. 이처럼 독립변수 간에 높은 상관관계가 존재하여 ' 다중공선성(multicollinearity)'이 의심된다.

(총 85개의 센서 중 2개의 조합 $C\binom{85}{2} = 3570$가지의 케이스 중 345개의 조합이 상관계수의 절댓값이 0.7 이상임)

var1	var2	상관계수 절댓값
x22	x68	0.995580
x38	x79	0.983583
x15	x83	0.983574
x20	x78	0.983552
x20	x84	0.982117
x39	x69	0.979501
x62	x63	0.974786
x20	x21	0.974038
x21	x84	0.973898
x78	x84	0.970541
x21	x78	0.970196
x67	x78	0.967133
x38	x68	0.966321
x69	x84	0.965388
x38	x78	0.965121
x38	x67	0.963555
x20	x38	0.963351
x22	x38	0.963031

var1	var2	상관계수 절댓값
x20	x67	0.962882
x77	x84	0.961509
x26	x67	0.961208
x20	x69	0.961171
x71	x78	0.961014
x68	x79	0.960879
x10	x58	0.960021
x21	x67	0.959807
x21	x69	0.959769
x78	x79	0.959640
x18	x71	0.958789
x67	x71	0.958059
x22	x79	0.957506
x21	x77	0.957291
x22	x71	0.956777
x39	x78	0.956170
x71	x79	0.956083
x23	x70	0.956055

var1	var2	상관계수 절댓값
x38	x71	0.954803
x67	x68	0.954775
x39	x84	0.954510
x67	x79	0.954344
x67	x84	0.954056
x68	x71	0.953916
x21	x39	0.953561
x26	x84	0.952751
x22	x67	0.952517
x20	x79	0.951691
x21	x38	0.951414
x38	x84	0.951028
x20	x26	0.950268
x5	x64	0.949696
x39	x77	0.949294
x69	x78	0.948889
x20	x39	0.948202
x20	x71	0.946798

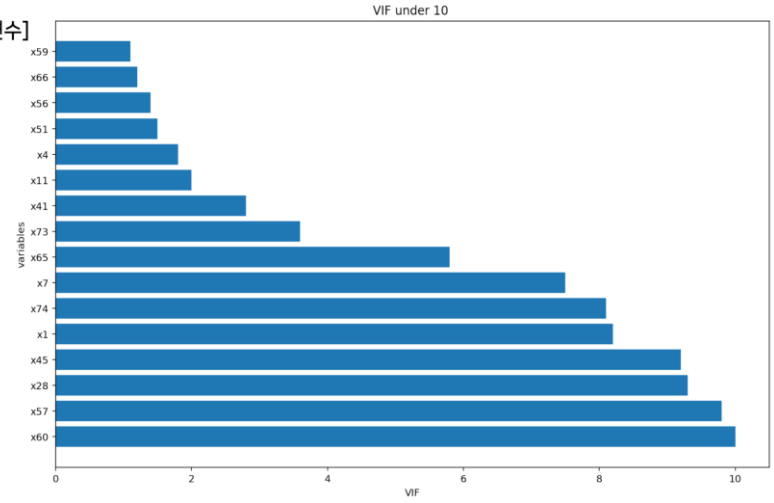
var1	var2	상관계수 절댓값
x22	x78	0.946585
x21	x71	0.946495
x20	x68	0.946136
x26	x38	0.945214
x68	x78	0.945098
x77	x78	0.944535
x21	x22	0.944409
x21	x79	0.943301
x26	x78	0.943242
x21	x68	0.941364
x20	x22	0.940893
x21	x26	0.938770
x26	x68	0.937063
x69	x77	0.936914
x18	x67	0.936642
x67	x75	0.936638
x18	x78	0.936199
x67	x77	0.933138

var1	var2	상관계수 절댓값
x22	x26	0.932857
x26	x75	0.932819
x18	x24	0.932318
x38	x75	0.931957
x68	x84	0.931734
x26	x79	0.931046
x18	x21	0.930620
x22	x84	0.929488
x79	x84	0.928410
x26	x77	0.927418
x38	x77	0.926881
x20	x77	0.926329
x18	x22	0.926027
x18	x38	0.925600
x26	x71	0.924730
x71	x84	0.923100
x78	x83	0.923096
x39	x67	0.922984

85개의 각 센서(x1~x85)를 독립 변수로 다중선형회귀 결과 대부분의 독립변수의 VIF(분산팽창인자)가 매우 높았다. VIF가 10보다 큰 변수는 85개 중 69개, 30보다 큰 변수는 48개로 나타났고, 이는 대부분의 독립변수가 다중공선성이 있는 변수임을 보여준다.

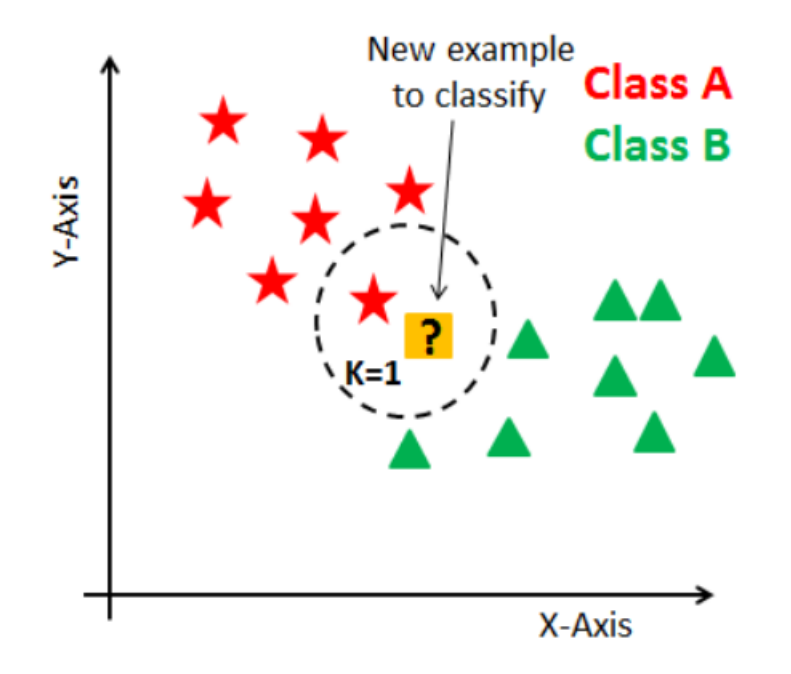
[VIF가 10 이하인 변수]

features	VIF Factor
x60	10.0
x57	9.8
x28	9.3
x45	9.2
x1	8.2
x74	8.1
x7	7.5
x65	5.8
x73	3.6
x41	2.8
x11	2.0
x4	1.8
x51	1.5
x56	1.4
x66	1.2
x59	1.1



다중공선성 문제를 고려하여 선형 모델이 아닌 비선형 모델을 채택하였다.

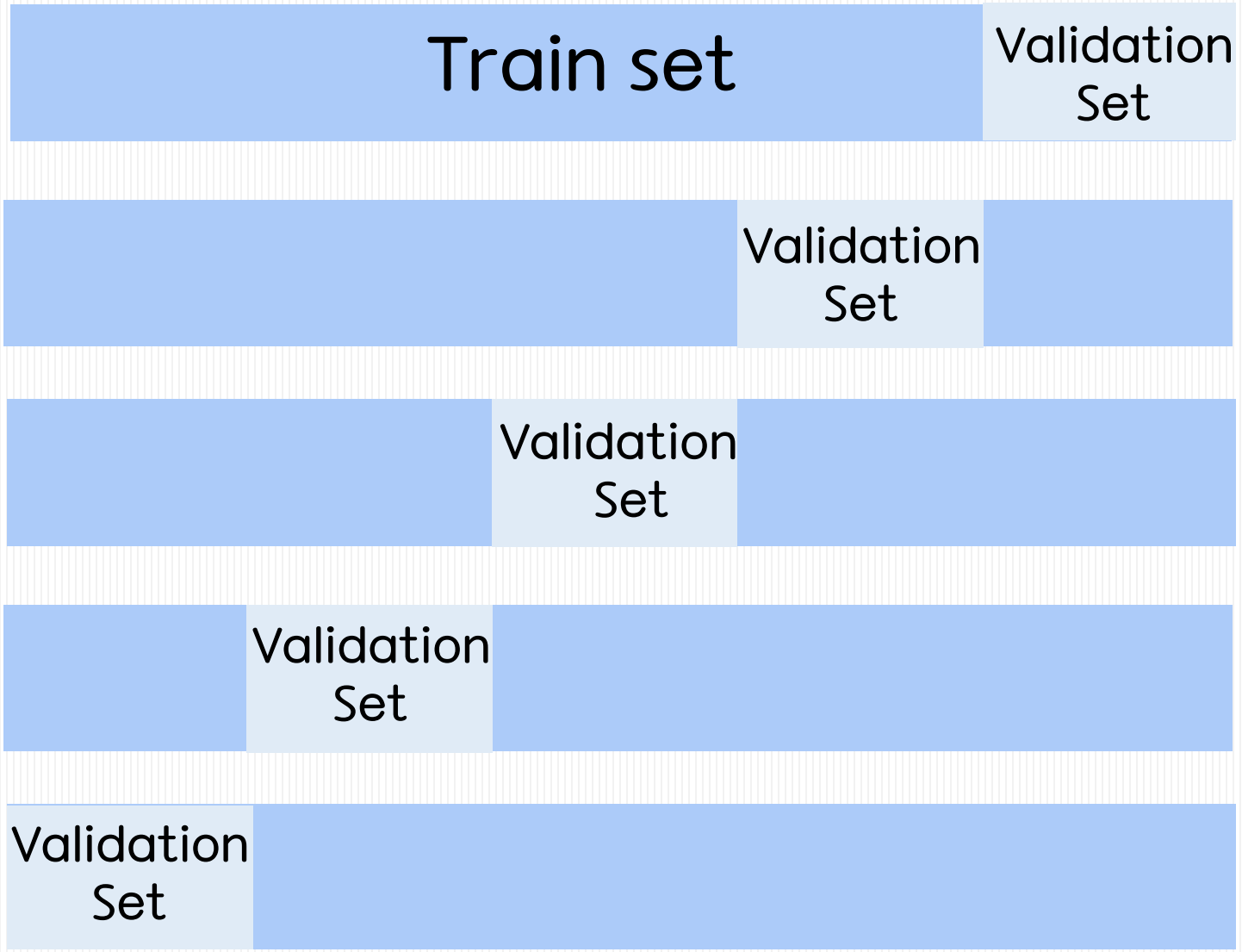
2. 모형선정 및 설명 [KNN]



KNN 사용이유

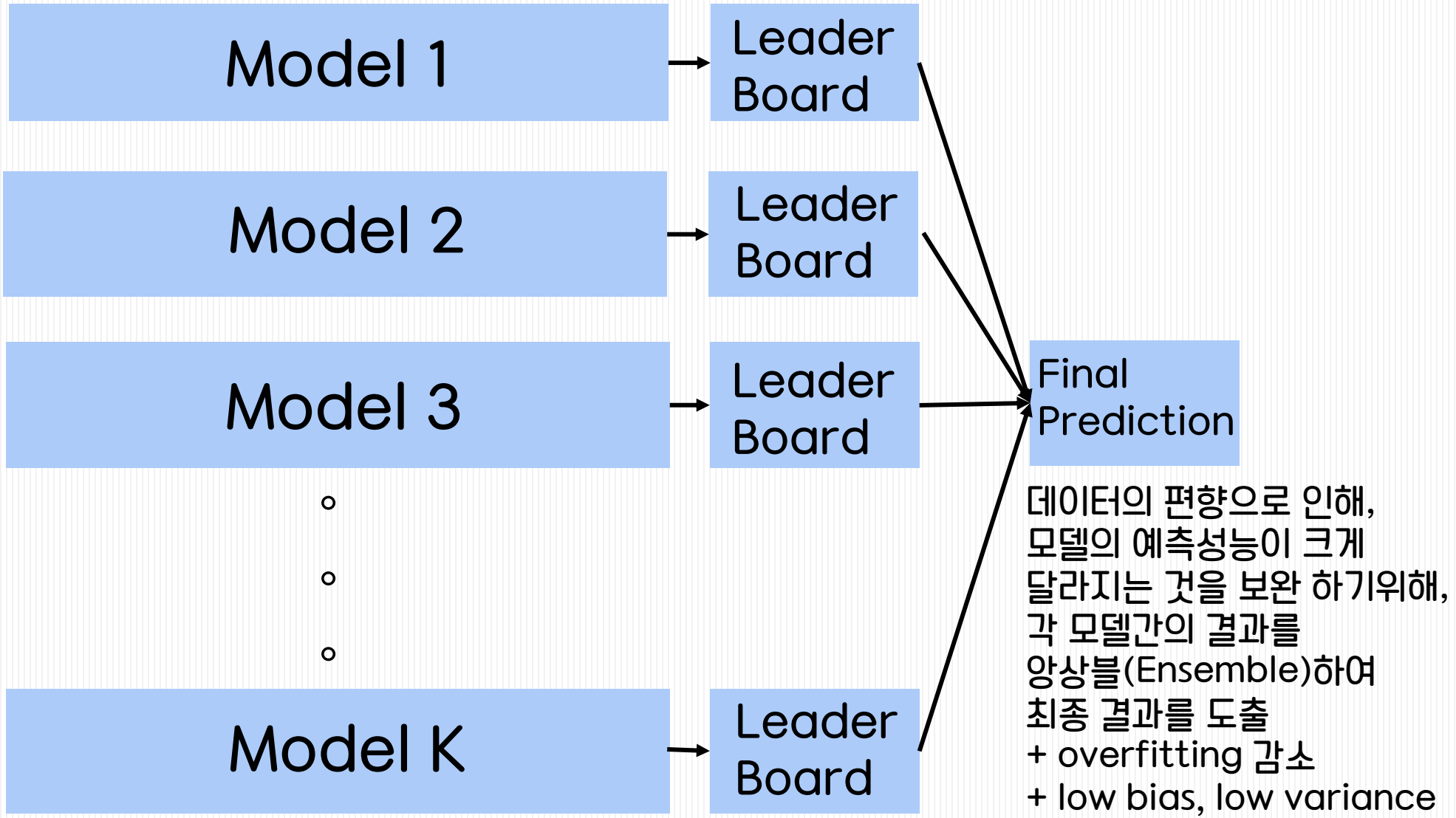
1. 훈련데이터에 잡음이 있는 경우에도 적용가능
2. 임의의 K값에 대해 베이즈 오차율에 항상 근접
3. 사례기반으로 높은 정확성

2. 모형선정 및 설명



Hyper-Parameter 튜닝과 Train Set의 소 표본의 편향성을 고려한 중첩 5-fold CrossValidataion 으로 평가

2. 모형선정 및 설명



데이터의 편향으로 인해,
모델의 예측성능이 크게
달라지는 것을 보완 하기위해,
각 모델간의 결과를
앙상블(Ensemble)하여
최종 결과를 도출
+ overfitting 감소
+ low bias, low variance

2. 모형선정 및 설명

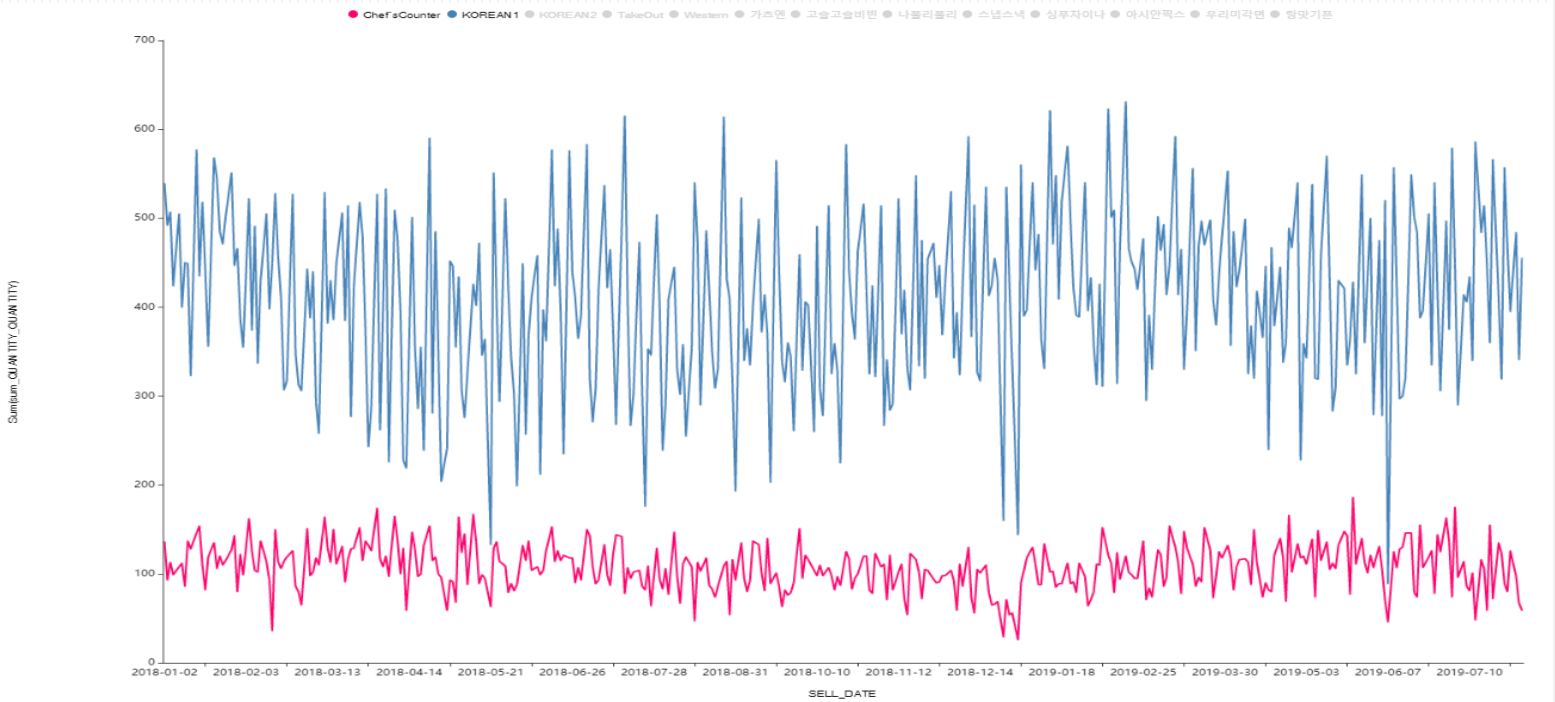
Leader Board

제품 불량율 예측 TOP 20

순위	신청구분	팀명	제출자ID	WMAE
1	팀	Sindorim	LHJ3890	0.61102
2	팀	KUZI	vx0918	0.71106
3	팀	501호 사람들	miso1178	0.72764
4	팀	명량대첩	bjsax	0.73244
5	팀	bigku	noonetwo	0.75897
6	팀	모희또	zizon233	0.77333
7	팀	B.A.F	surfing2003	0.77731

Part2

1.문제 및 데이터 설명



Chef's Counter
Korean1의
수요량 그래프

1.문제 및 데이터 설명

[취식 브랜드 수요량 예측]

사내 식당 취식 데이터를 활용하여
지정된 기간의 13개 브랜드 별 일일 취식 수요량 예측
- 삼성 SDS에서 제공하는 사내 브랜드 별 취식 데이터를 활용하여 브랜드 별 수요 예측 모델링 개발

2. 전처리 [메뉴이름 통일]

KOREAN1	나가사키 부대찌개
KOREAN1	나가사키부대찌개
KOREAN1	사골떡만두국
KOREAN1	사골떡만둣국
KOREAN1	대패삼겹된장찌개
KOREAN1	대패삼겹살된장찌개
TakeOut	HOT TO GO
TakeOut	HOTTOGO
Western	돈가스오므라이스
Western	돈까스 오므라이스

같은 메뉴인데도, 띄어쓰기,
자음 차이 등으로
메뉴이름이 일치가
안되는 것을 모두 통일

2. 전처리 [Customer Data 이용]

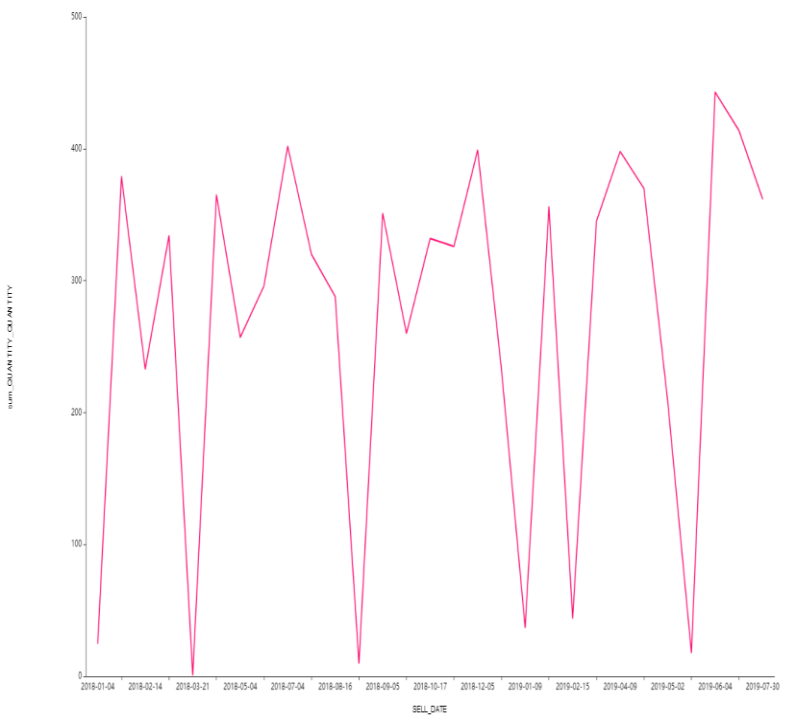
사원 정보가 없는 취식데이터
145개 제거

2. 전처리 [서브메뉴 통합]

	SELL_DATE	BRAND	MENU	sum_QUANTITY_QUANTITY	is_outlier_sum_QUANTITY_QUANTITY↓
1	2018-06-08	KOREAN2	갈치구이	18	out
2	2018-10-31	KOREAN2	갈치구이	9	out
3	2019-04-23	KOREAN2	갈치구이	517	out
4	2018-01-04	KOREAN2	고등어구이	25	out
5	2018-03-21	KOREAN2	고등어구이	1	out
6	2018-08-29	KOREAN2	고등어구이	10	out
7	2019-05-21	KOREAN2	고등어구이	18	out
8	2018-01-29	KOREAN2	돈육강정	408	out
9	2018-11-27	KOREAN2	돈육강정	307	out

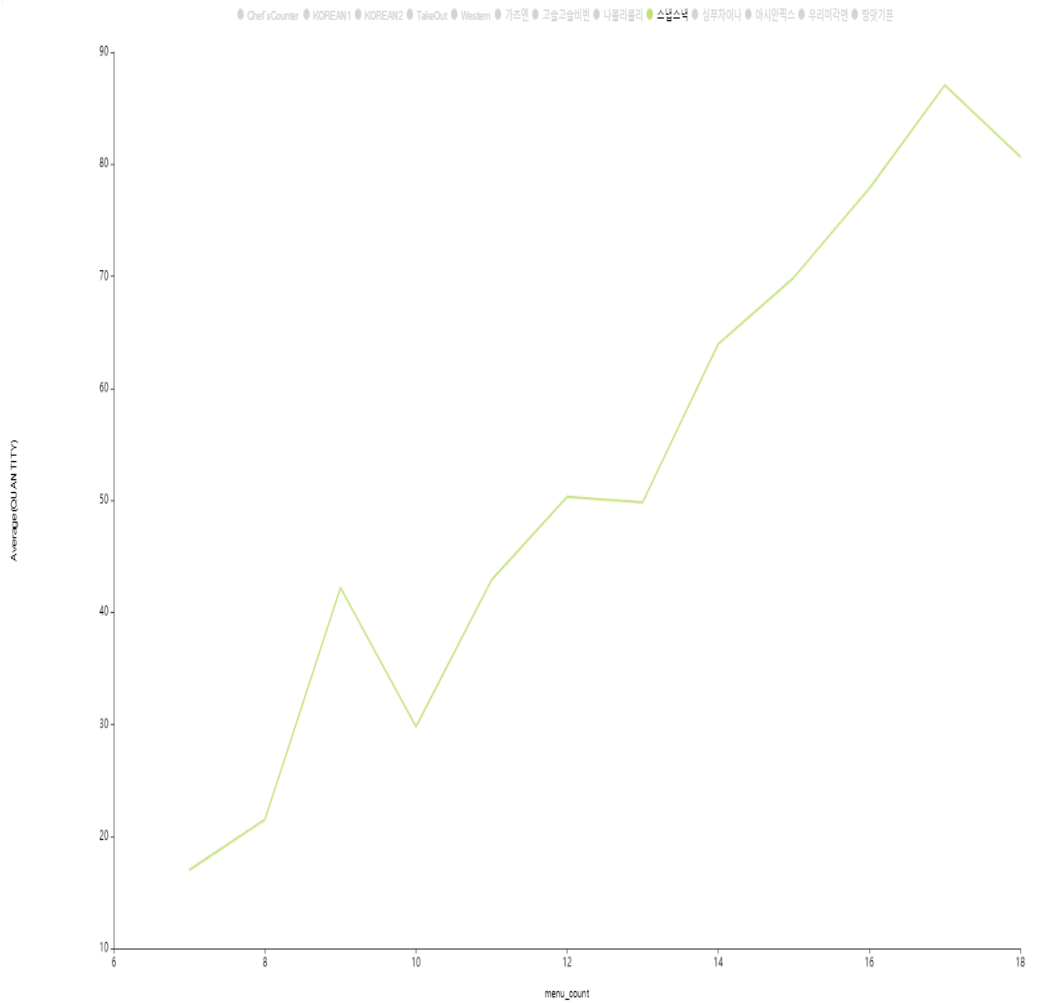
KOREAN1,2 의 메뉴들을 Outlier Detection을 통해 확인 해본 결과, 같은 메뉴임에도 불구하고, 수요량이 확연히 작은 메뉴를 발견하여 그 메뉴들의 수요량을 확인

2. 전처리 [서브메뉴 통합]



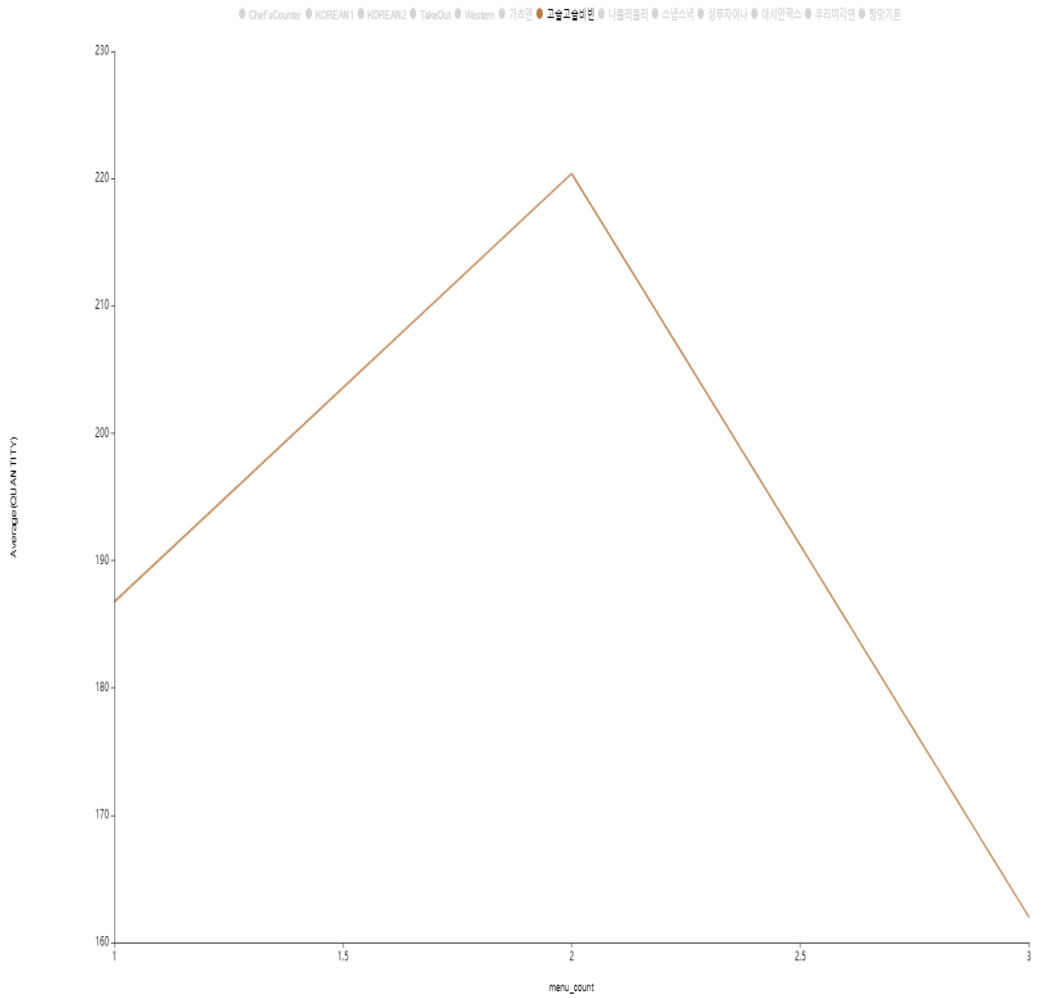
그림에서와 같이 확연하게 차이가 나는 날의 메뉴를 보니,
그 메뉴들이 서브메뉴임을 알 수 있었고,
본 메뉴와 서브메뉴의 수요량을 통합

3. 추가데이터 설명 [메뉴 개수]



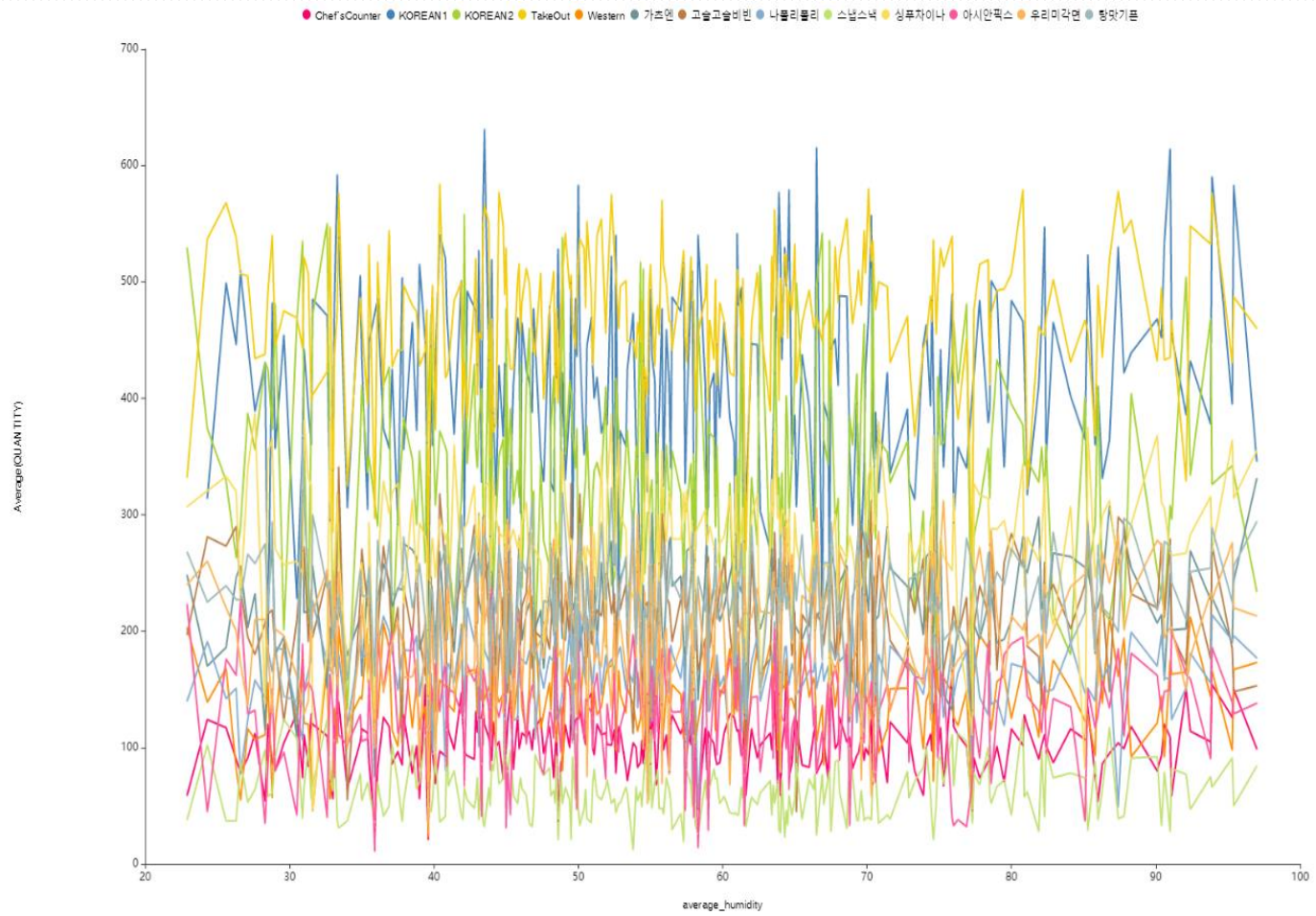
같은 브랜드 인데도, 메뉴의 개수에 따라 수요량이 달라지는 것을 발견
→ 메뉴개수를 변수에 추가

3. 추가데이터 설명 [메뉴 개수]



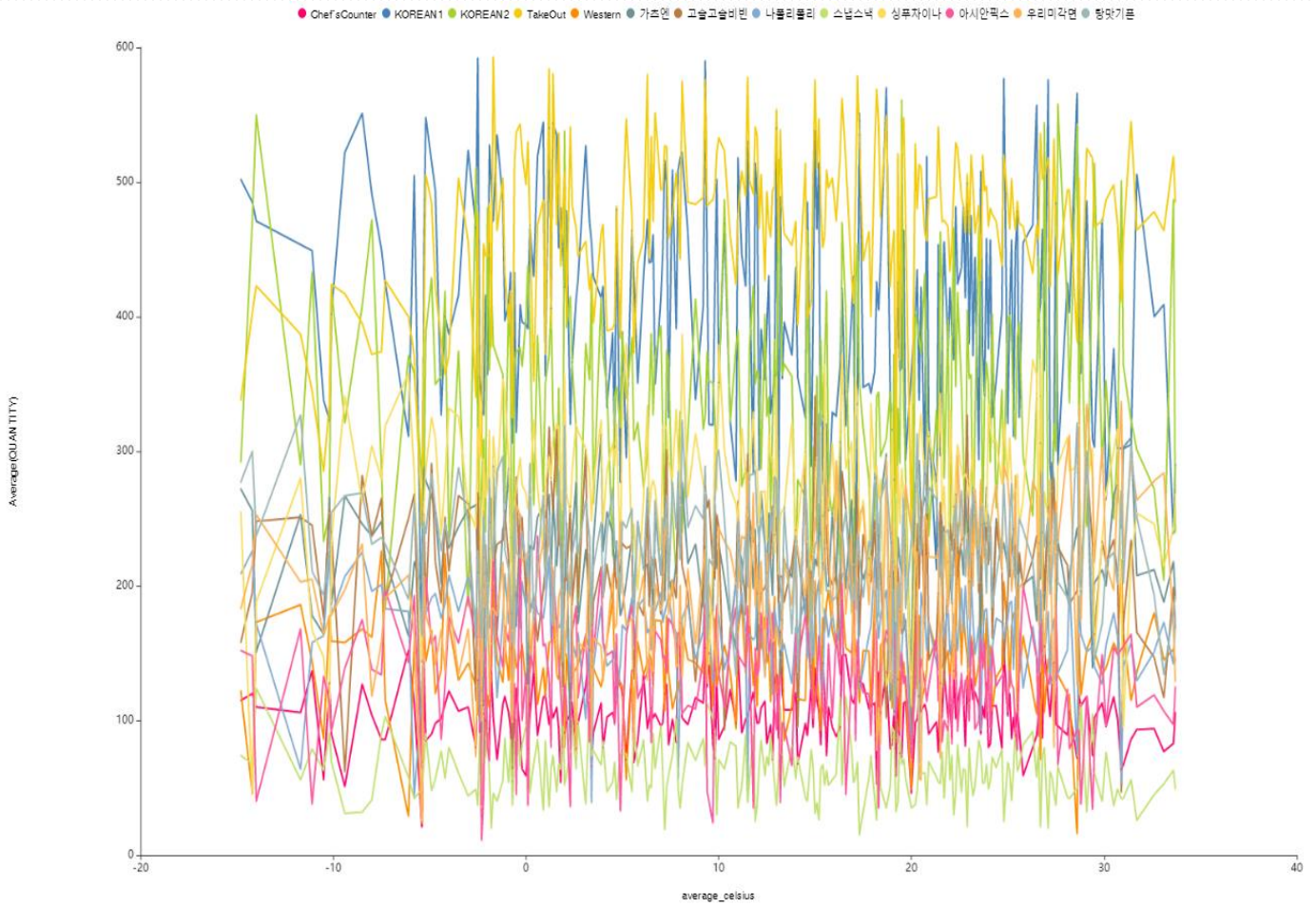
같은 브랜드 인데도, 메뉴의 개수에 따라 수요량이 달라지는 것을 발견
→ 메뉴개수를 변수에 추가

3. 추가데이터 설명 [날씨 데이터]



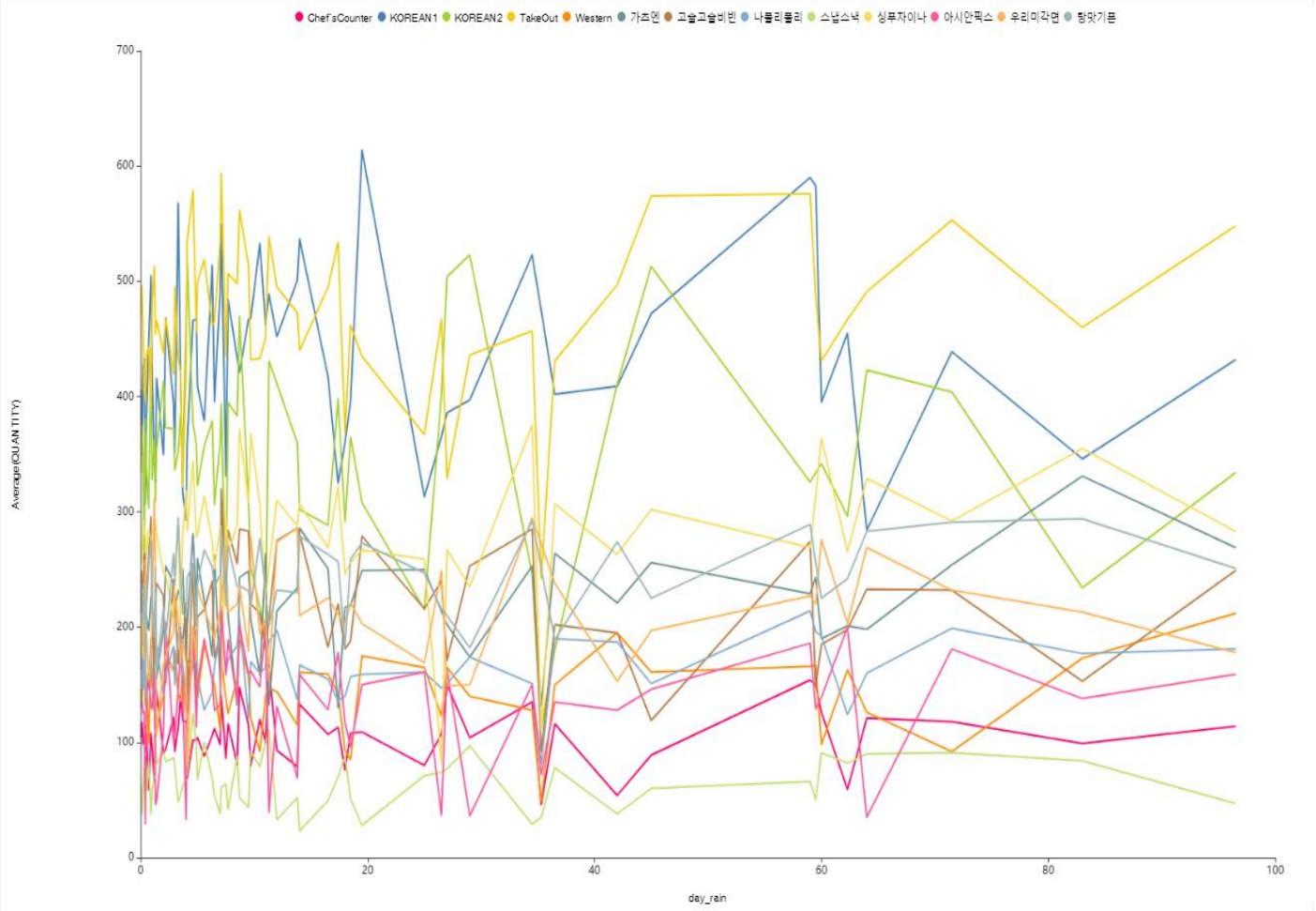
평균 상대습도

3. 추가데이터 설명 [날씨 데이터]



평균 온도

3. 추가데이터 설명 [날씨 데이터]



일강수량

3. 추가데이터 설명 [날씨 데이터]

8월 27일 ~ 9월 11일 까지의 날씨 데이터를 ARIMA 예측을 통해 써보려고 했으나, 값들이 모두 비슷한 값으로 예측(정확한 예측불가)한 것을 확인 하였음
그 이유를 그래프를 통해 살펴보니, 각 변수들이 White Noise 형태를 띠고 있었고, 예측이 불가한 날씨변수를 최종적으로 변수로 넣지 않기로 결정

3. 추가데이터 설명 [메뉴지수]

1. 나이 범주화

20대 : 나이<30
30대 : 30<나이<40
40대 : 40<나이<50
50대 : 50<나이<60
60대이상 : 나이>60

2. 연령대°성별

나이범주(5개)

* → 9개의 범주의 비율구하기

성별(남,여) (60대*여자는 재직x)

X = 9개 범주의 비율

° : X를 구하는 연산자로 정의

3. Y = 메뉴에 따라 X(연령대*성별 비율)로 분류하여 먹은 취식량

4. N = 메뉴별 영업일 수 계산

5. X*Y/N = 메뉴별 평균 가중치

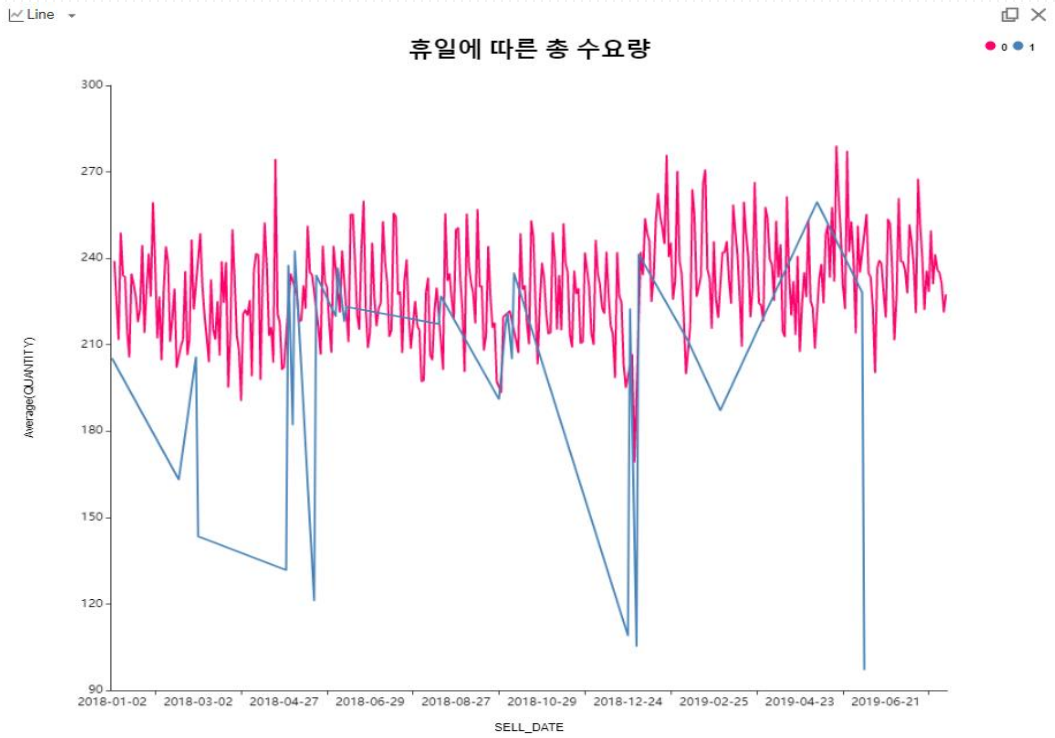
6. 메뉴 가중치를 min_maxScaler로 정규화

7. 메뉴가 2개이상 나온 날의 브랜드는 메뉴 가중치의 선형결합으로 계산

3. 추가데이터 설명 [메뉴지수] - 예시

범주	X	Y(김치찌개)	X*Y/N
20대*남	0.05	50	=0.05*50/20
20대*여	0.05	100	=0.05*100/20
30대*남	0.1	50	=0.1*50/20
30대*여	0.2	80	=0.2*80/20
40대*남	0.25	70	=0.25*70/20
40대*여	0.25	60	=0.25*60/20
50대*남	0.1	40	=0.1*40/20
50대*여	0.05	30	=0.05*30/20
60대*남	0.05	20	=0.05*20/20

3. 추가데이터 설명 [공휴일변수]



0 : 평일(공휴일과 무관)
1 : 공휴일 전,후

→

그림에서 볼 수 있는
저점들을 확인해본 결과,
휴일과 휴일 사이의 날짜인
것을 확인하여,
그 점들을 새로운 범주로
추가

3. 월별 데이터 변수 정리

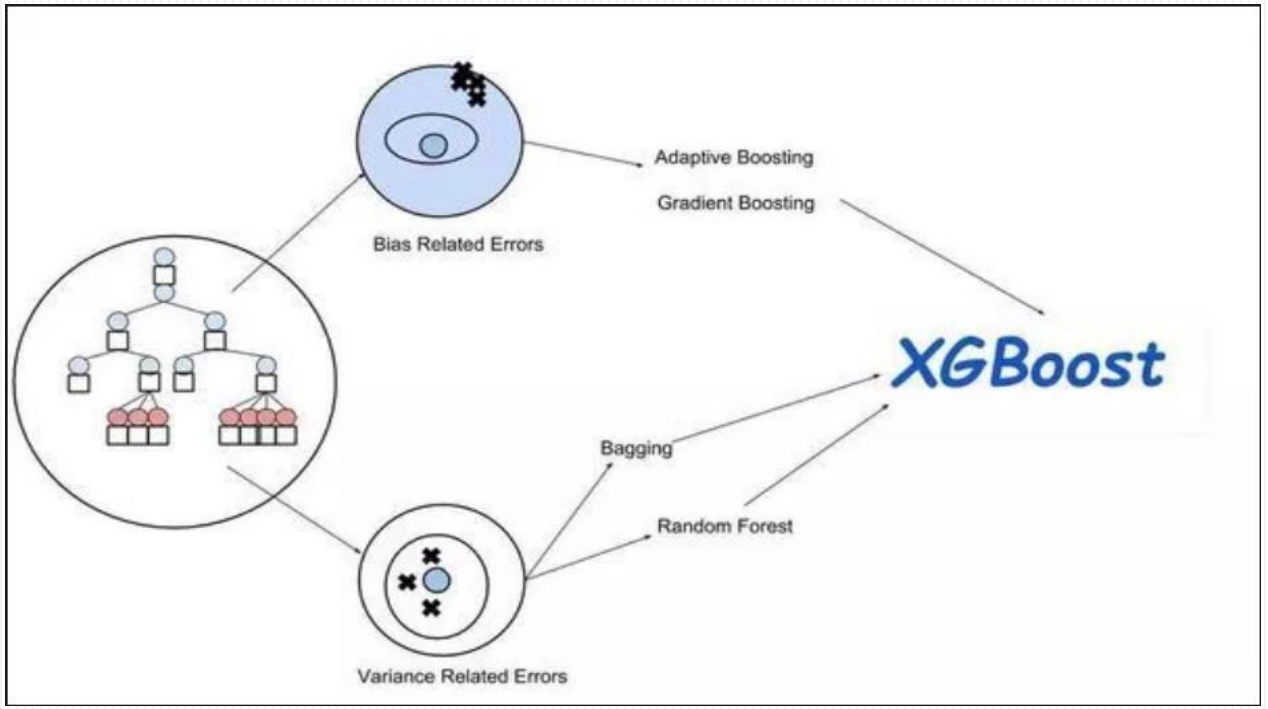
8월

Quantity	일별 수요량 (브랜드별)
Menu_count	일별 판매 메뉴 개수 (브랜드별)
sum	정규화된 메뉴 지수
Sum_sum_Menu_0~516	메뉴 더미
Dayoftheweek_0~4	요일 더미(월,화,수,목,금)
Year_0~1	연도 더미(2018,2019)
Month_0~11	월 더미(1~12월)
Day_0~30	일 더미(1~30일)
Holiday_0~2	휴일 더미

9월

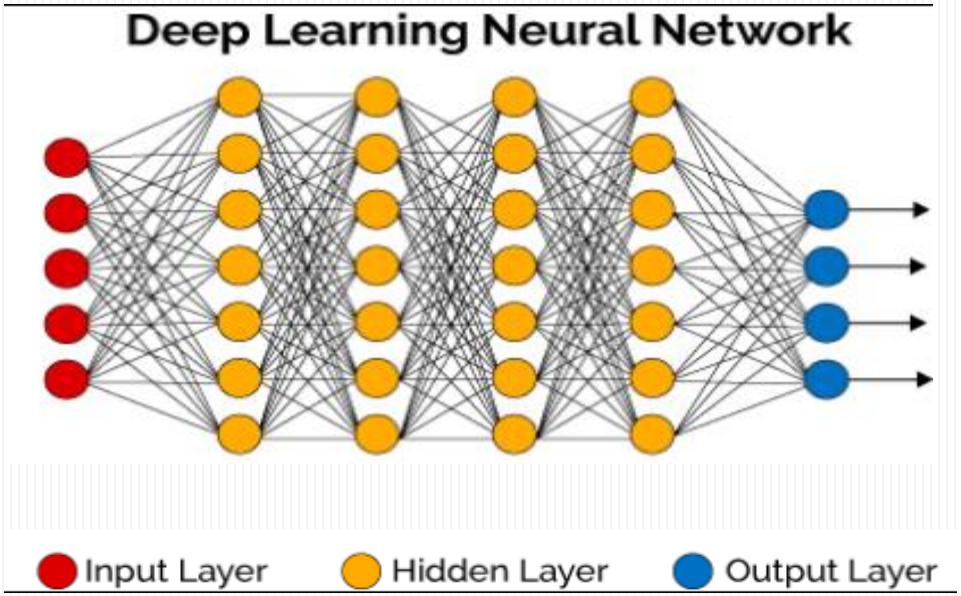
Quantity	일별 수요량 (브랜드별)
Dayoftheweek_0~4	요일 더미(월,화,수,목,금)
Year_0~1	연도 더미(2018,2019)
Month_0~11	월 더미(1~12월)
Day_0~30	일 더미(1~30일)
Holiday_0~2	휴일 더미

4. 모형선정 및 설명 [XGBoost]



Xgb 사용 이유
1. GBM에 비해 학습 속도가 빠름
2. 일반화 오차의 편향과 분산을 모두 조절하여 낮출 수 있는 가변적인 모델
3. 정규화 변수를 넣을 수 있어 과적합을 해결할 수 있음

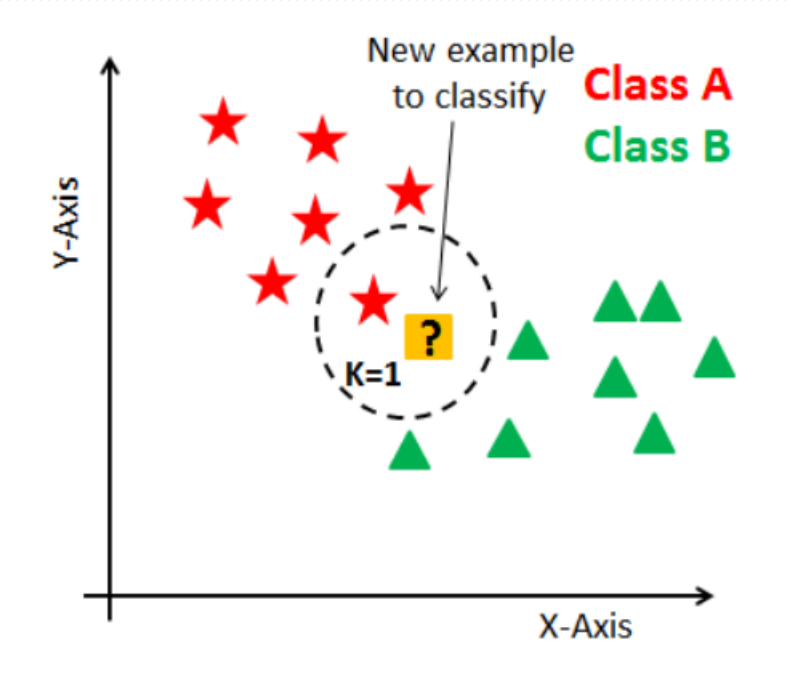
4. 모형선정 및 설명 [Neural Net]



DNN(Deep Neural Networks) 사용이유

1. Feature Extraction이 자동으로 수행됨
2. 입력 변수들간의 비선형 조합이 가능(유연하게 모델적합이 가능)
3. 데이터 양이 다른 머신러닝 알고리즘보다 성능이 좋을 수 있음

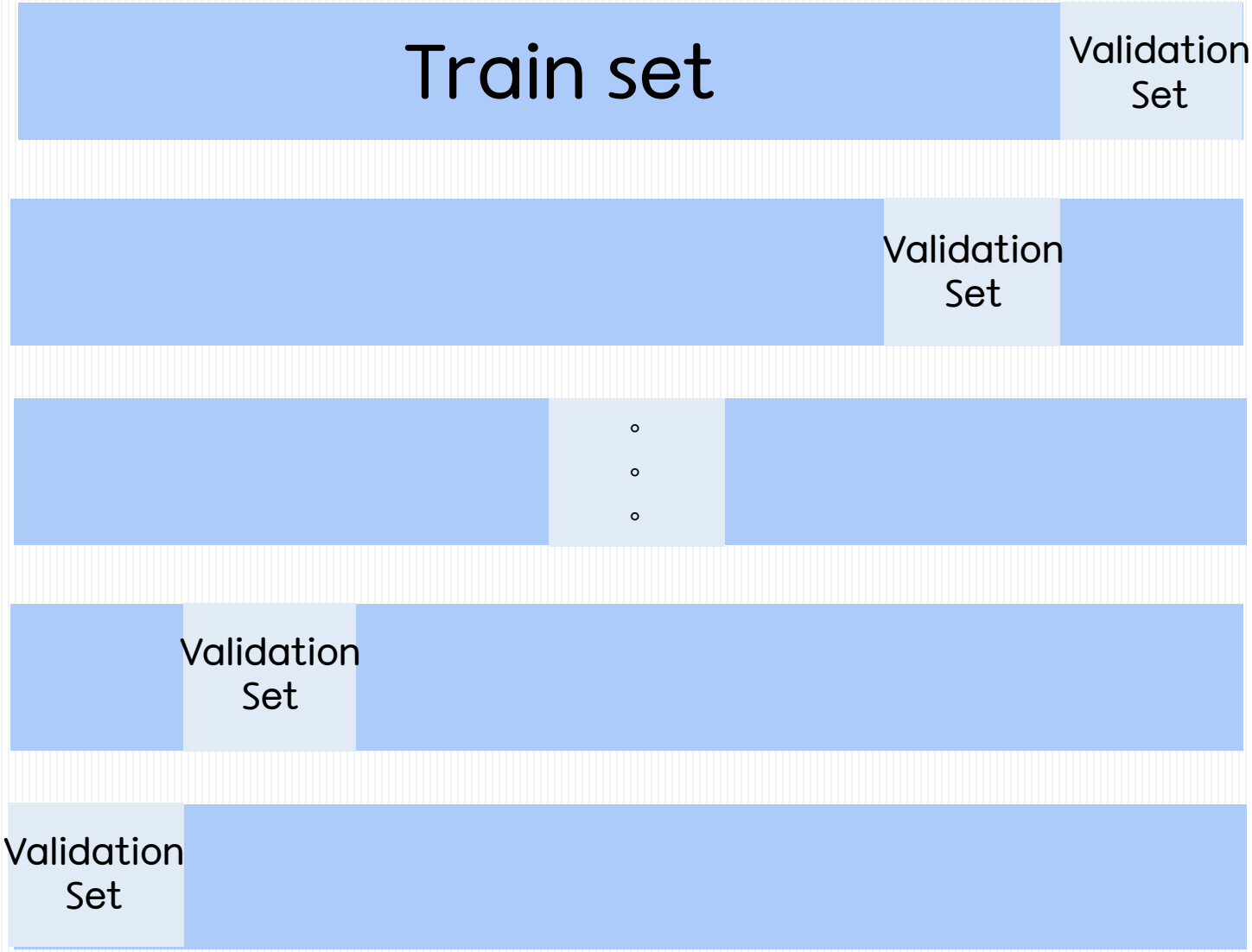
4. 모형선정 및 설명 [KNN]



KNN 사용이유

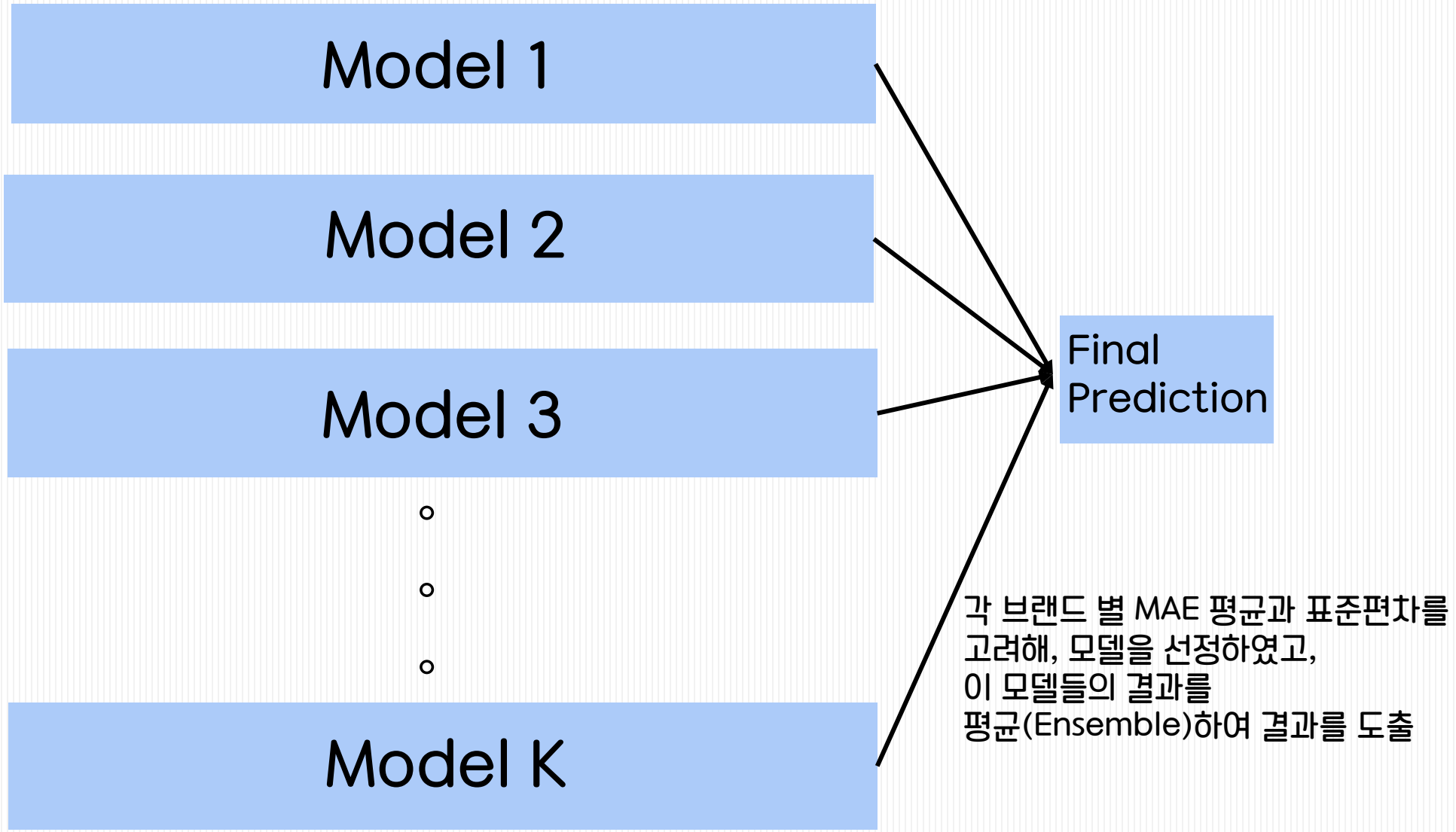
1. 훈련데이터에 잡음이 있는 경우에도 적용가능
2. 임의의 K값에 대해 베이즈 오차율에 항상 근접
3. 사례기반으로 높은 정확성

4. 모형선정 및 설명



Hyper-Parameter 튜닝과 Train Set의 소 표본의 편향성을 고려한 중첩 10-fold CrossValidataion 으로 평가

4. 모형선정 및 설명


Model 1
Model 2
Model 3
Model K
Final Prediction
각 브랜드 별 MAE 평균과 표준편차를 고려해, 모델을 선정하였고,
이 모델들의 결과를
평균(Ensemble)하여 결과를 도출

4. 모형선정 및 설명

Leader Board

취식 수요 예측 TOP 20

※ MAE(중간) : 8월 20일까지의 데이터로 채점한 중간 결과입니다.

※ MAE(최종) : 9월 11일까지의 데이터로 채점한 최종 결과입니다.

※ 리더보드의 순위는 MAE(최종)값이 기준입니다.

순위	신청구분	팀명	제출자ID	MAE(중간)	MAE(최종)
1	팀	명량대첩	bjsax	30.51555	31.32695
2	팀	데굴데굴	key1036	30.79238	33.84402
3	팀	501호 사람들	miso1178	36.64320	33.87815
4	팀	B.A.F	surfing2003	36.18343	33.92838
5	팀	미산소	ehvndfb	33.45363	33.94821

5. 한계점 및 향후 개선방향

1. 적절한 추가 외부데이터를 찾지못함
2. KOREAN1,2의 취식량에 대한 상호작용을 충분히 설명하지 못함
3. Customer Data를 메뉴지수 이외에 더 많은 파생변수를 생성하여 사용하였다면 더 좋은 결과를 얻었을 것으로 예상
4. 브랜드 메뉴별 평가(별점 등)가 있었다면 감성분석을 통해 취식량이 적은 데이터에 대해 더 좋은 예측값을 얻을 수 있었을 것으로 예상

감사합니다.