SAMSUNG SDS
Realize your vision

Foresee

Techtonic 2019

Disrupt

Partner

2019.11.14 • SAMSUNG SDS Tower B1F
Magellan Hall / Pascal Hall

Deep Dive Session

Brightics Academy 공모전 2019

데이터 분석 부문 수상팀 노하우 공개

501호 사람들(김미소 김재원 민경하 손진석 조예린) / 연세대학교

데이터 분석 부문

제공된 데이터만을 활용하여 제품 불량률 예측 모델과 취식 브랜드 수요량 예측 모델 개발

제품 불량률
예측 모델링 개발

검증 데이터의
실제 불량률과 예측 불량률간
오차에 대한 WMAE

취식 브랜드 수요량
예측 모델링 개발

검증 데이터의
브랜드별 MAE 평균

1

제품 불량률 예측

AGENDA

1. 문제 정의 및 데이터 설명
2. EDA(탐색적 데이터 분석)
3. 불량률 예측 알고리즘

1. 문제 정의 및 데이터 설명

문제

제품 불량률 예측 모델

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

제공 데이터

- 학습/검증 데이터는 공정 불량 전체데이터에서 sampling을 통해 선정됨
- 학습 데이터 (3507개)
  : 불량률 분석을 위한
  공정제조순서, 공정 센서, 불량률(Y값)
- 검증용 데이터 (300개)
  : 모델 정확도 측정을 위한 검증 데이터로
  불량률(Y값)이 빈 칸인 데이터

1. 문제 정의 및 데이터 설명

불량률 데이터

	Time_Sequence	x1	x2	x3	x4	x5	x6	x7	x8	x9	...	x77	x78	x79	x80	x81	x82	x83	x84	x85	Y
0	1	0.028837	-0.649824	0.363124	-0.752145	-0.631316	-0.588815	-0.904200	-0.577024	0.715939	...	1.394348	1.350135	1.774360	1.741823	1.705609	-1.175603	0.705534	1.350135	0.625413	2.185269
1	2	0.053882	-0.649824	0.347376	-0.499595	-0.528668	-0.487181	-0.904200	-0.576967	0.702009	...	1.277810	1.404929	1.701051	1.469954	1.705365	-2.588321	0.810728	1.404929	0.050509	1.572313
2	3	0.065778	-0.649824	0.347376	-0.732795	-0.586186	-0.578574	-0.904200	-0.577844	0.703790	...	1.428878	1.573875	1.758069	1.741823	1.705609	-1.371211	0.747611	1.573875	0.625413	2.581797
3	4	0.653070	-1.332022	1.150375	-1.709346	0.473910	1.220860	-0.038151	1.636807	0.448839	...	-0.833474	-1.135911	-0.813025	1.966284	1.263878	-0.059706	-0.511038	-0.885362	0.791103	2.758224
4	5	0.039497	-0.649824	0.326514	-0.361767	-0.602523	-0.636483	-0.904200	-0.576836	0.705902	...	1.426222	1.409495	1.936643	1.741823	1.705609	-2.196550	0.757861	1.409495	0.625413	NaN

Time_Sequence

- 제품 제조 순서
- (분석 시, 독립변수로 활용 가능)

x1 ~ x85

- 공정 센서 파라미터
- 제품 생산 시 파생되는 센서 변수 값

Y

- 각 제품별 불량률 지표
- 값이 클수록 불량도 심각

데이터 분석 방법론 SEMMA

1 Sample

- 전체 데이터를 학습, 테스트, 예측 데이터로 분리
- 학습과 테스트 데이터의 불량률의 분포가 유사하게 분리

2 Explore

- 전체 센서 값에 대해서 불량률과의 관계 확인 (산점도, 상관계수)
- 전체 센서 값에 대해서 Time_Sequence와의 관계 확인(산점도, 상관계수)

3 Modify

- 변수 선택
  - PCA
  - VIF
  - Top-down Elimination[1]
  - Bottom-up Elimination[1]

4 Model

- 모델 선정
  - KNN
  - XGBoost
  - LightGBM
  - NN

5 Assess

- 모델 하이퍼파라미터 선택
  - GridSearch
- 모델 평가
  - K-Fold Cross Validation
- 결과
  - 불량률 예측

참고: [1] p9 자체적으로 만든 변수 선택 프로세스

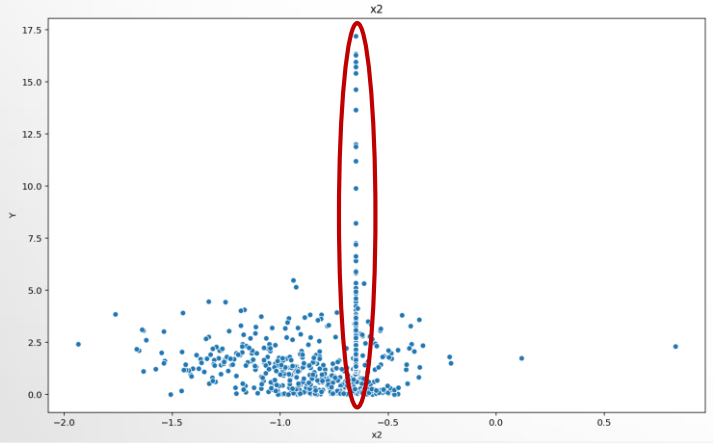
2. EDA (1) 센서값 vs 불량률

센서 값과 불량률의 산점도로 아래와 같은 3가지 특징적인 case를 확인할 수 있었다.

Case 1

각 센서의 특정 값에 데이터가 1 자 모양으로 형성

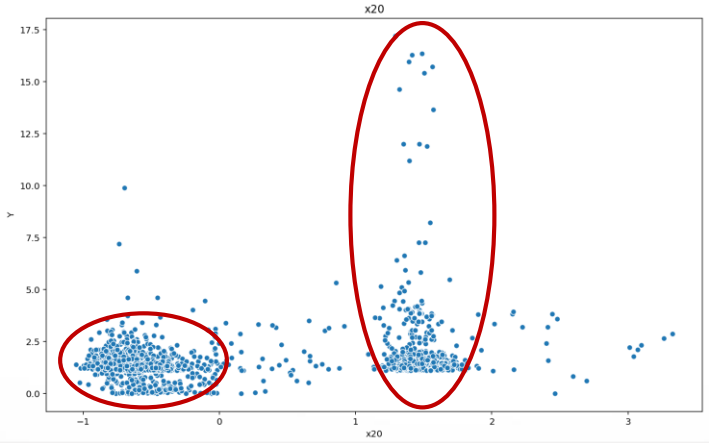
- 특징 : 전체 데이터의 약 87%가 특정 센서 값에 존재
  불량률 이상치가 이러한 특정 센서 값에서만 존재
- 해당 센서 : x2, x8, x12, x13, x16, x17, x25, x28, x29, x32, x33, x34, x37, x46, x47, x49, x50, x53, x54, x55, x61, x74, x81



Case 2

각 센서 값의 특정 범위에 따라 두 군집으로 형성

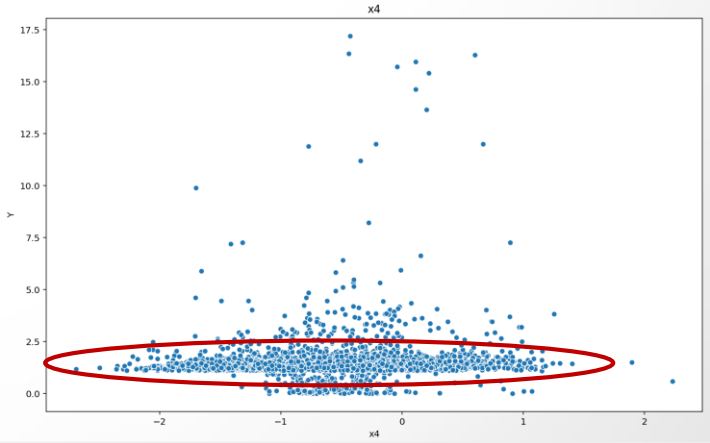
- 특징 : 대부분의 이상치는 둘 중 하나의 군집에서 발생,
  각각의 군집은 음수 영역과 양수 영역으로 구분,
  불량률과의 상관 계수 절댓값이 높음
- 해당 센서 : x2, x8, x12, x13, x16, x17, x25, x28, x29, x32, x33, x34, x37, x46, x47, x49, x50, x53, x54, x55, x61, x74, x81



Case 3

데이터가 가로로 길게 늘어진 모양을 형성

- 특징 : 데이터가 각 센서 값의 범위에 고르게 퍼짐,
  불량률과의 상관 계수 절댓값이 낮음
- 해당 센서 : x4, x10, x11, x14, x19, x23, x51, x58, x72, x82, x83



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

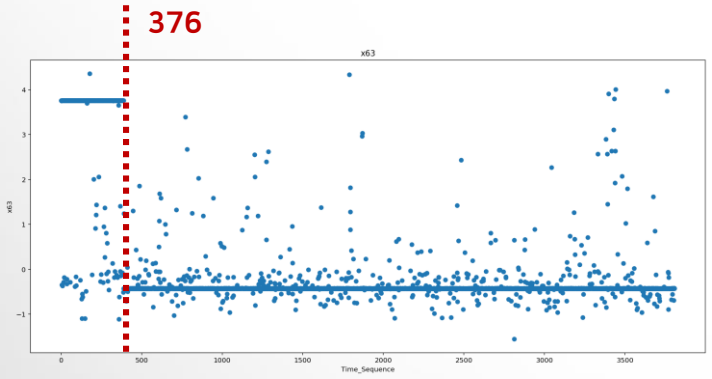
2. EDA (2) Time_sequence vs 센서 값

Time_Sequence와 센서 값의 산점도로 아래와 같은 3가지 특징적인 case를 확인할 수 있었다.

Case 1

Time_Sequence가 '376' 기점으로 센서 값이 양수에서 음수로 급격하게 바뀌는 경우

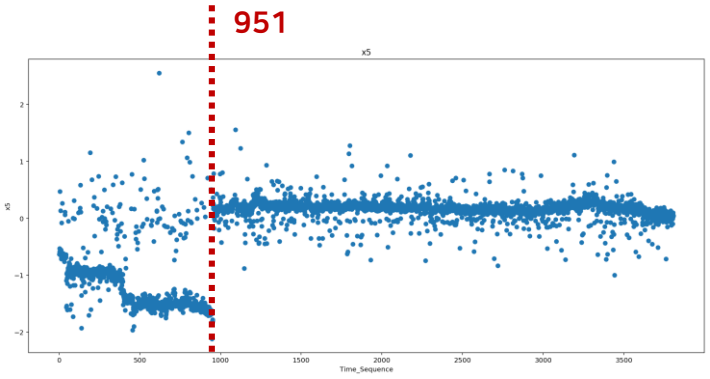
- 해당 센서 : x62, x63



Case 2

Time_Sequence가 '951' 기점으로 센서 값이 음수에서 양수로 급격하게 바뀌는 경우

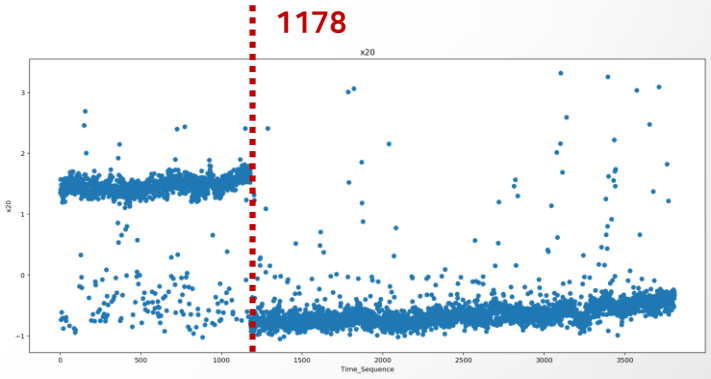
- 해당 센서 : x5, x64



Case 3

Time_Sequence가 '1178'기점으로 센서 값이 양수에서 음수로 급격하게 바뀌는 경우

- 해당 센서 : x9, x15, x18, x19, x20, x21, x22, x24, x26, x35, x38, x39, x67, x68, x69, x71, x75, x77, x78, x79, x82, x83, x84



관련 변수 생성 or 특정 기점으로 데이터셋 분리 등의 방법으로 위 특징을 모델에 적합 시켰을 경우 **유의미하지 않음**

2. EDA (3) 각 센서(독립변수) 간 상관관계 및 VIF

독립변수 간 correlation이 높은 경우가 다수 존재 '다중공선성(multicollinearity)' 문제가 의심

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

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

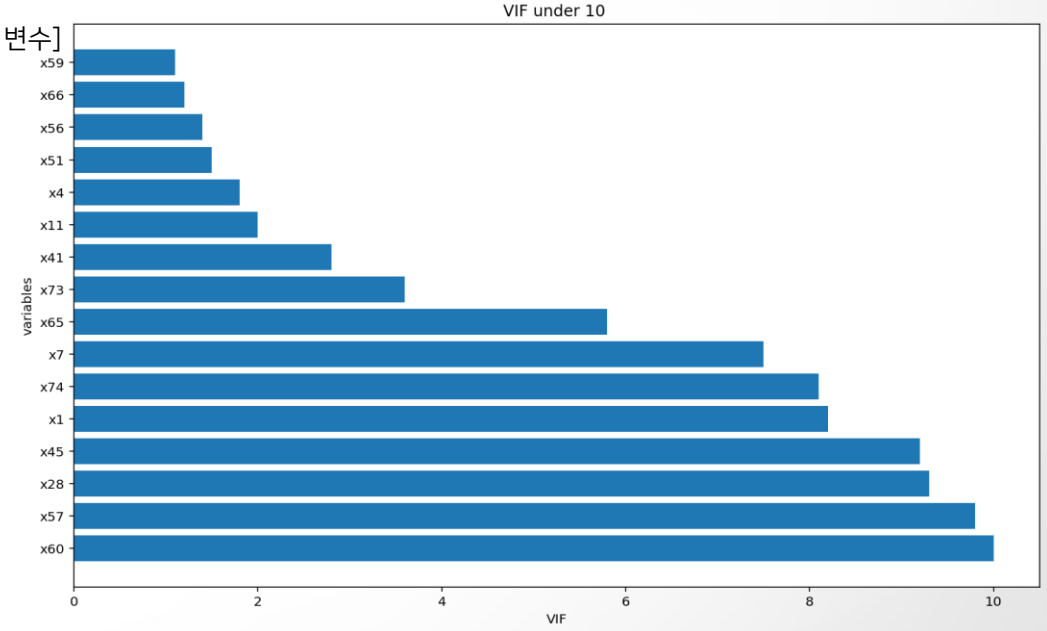
var1	var2	상관계수 절댓값
x22	x26	0.932857
x26	x75	0.932819
x18	x24	0.932318
x38	x75	0.931957
x68	x84	0.931734
x26	x79	0.931046
x18	x21	0.930620
x22	x84	0.929488
x79	x84	0.928410
x26	x77	0.927418
x38	x77	0.926881
x20	x77	0.926329
x18	x22	0.926027
x18	x38	0.925600
x26	x71	0.924730
x71	x84	0.923100
x78	x83	0.923096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

각 센서(x1~x85)의 VIF(분산팽창인자)가 매우 높았음 대부분의 독립변수가 다중공선성이 있는 변수임

VIF가 10보다 큰 변수는 85개 중 69개
30보다 큰 변수는 48개로 나타남

[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



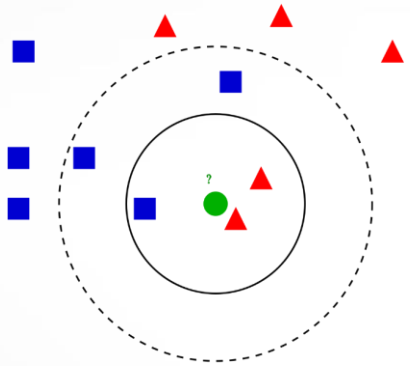
다중공선성 문제를 고려하여 선형 모델이 아닌 **비선형 모델**을 채택하였다.

3. 불량률 예측 알고리즘 (1) 최종 모델

- **최종 모델 : KNN (K-Nearest Neighbor) Regression**

→ 새로운 데이터가 주어졌을 때 기존 데이터 가운데 가장 가까운 k개 이웃의 정보로 새로운 데이터를 예측 하는 방법

→ <u>불량률 예측의 경우 '회귀'의 문제이므로 이웃들의 종속변수(불량률)의 '평균'이 예측 값</u>이 된다.



ex) 왼쪽의 예시 그림에서 k=3 이라면 초록색 데이터의 예측 값은 거리가 가장 가까운 진한 선 원 안의 3개 데이터의 평균값으로 예측된다.

- **최적 파라미터: p=2 , n_neighbors=1, weights=uniform**

KNN 모델은 hyper parameter로 사용자가 탐색할 이웃의 수(n_neighbors)와 거리측정방법(p)를 지정해주어야 한다. 추가적으로 데이터에 가중치를 주고 싶다면 weights 파라미터를 변경할 수 있다.

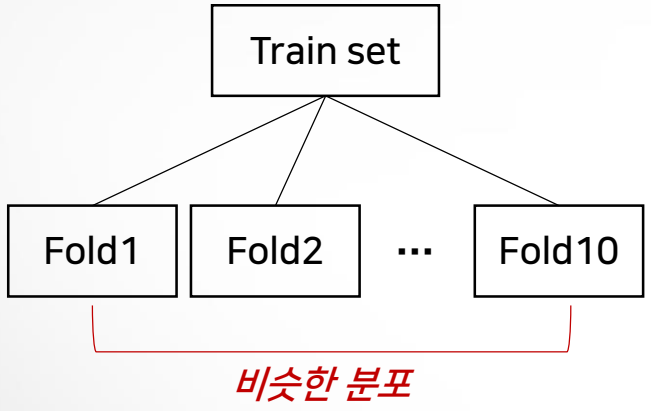
우리는 <u>Grid Search를 이용하여 우리의 데이터에 맞는 최적의 파라미터 조합을 찾았는데</u>, 그 결과 우리는 KNN 분석 시 탐색할 이웃의 수는 1, 거리 측정 방법은 유클리디안, 가중치는 모든 데이터가 동등하도록 설정하였다.

*Grid search: 각각의 하이퍼 파라미터 설정에서 K-fold Cross-Validation을 학습 데이터 셋에 적용하고 여러 개의 모델을 만들어 성능을 추정하는 방식

3. 불량률 예측 알고리즘 (2) 변수 선택(Feature Selection)

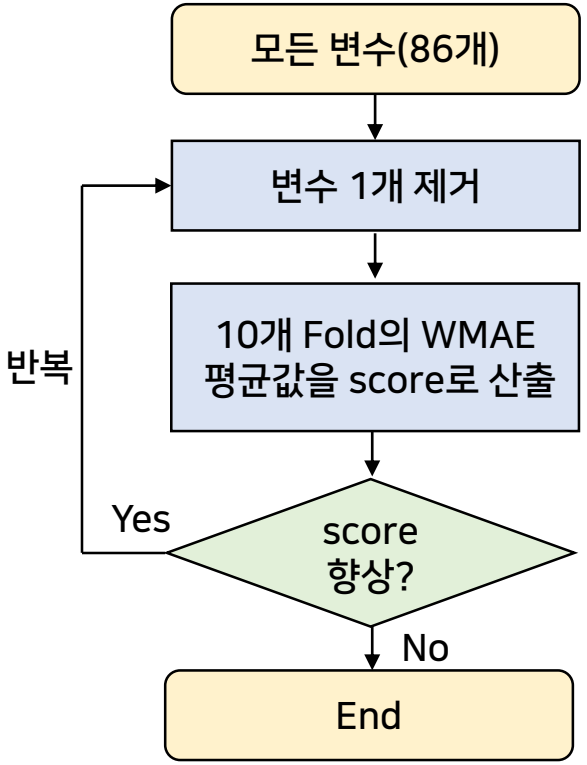
데이터의 편향을 최소화 하기 위해 **문제 상황에 알맞은 Feature Selection 프로세스를 만들어 적용**하였다.

① 분포가 최대한 같은 Fold 찾기



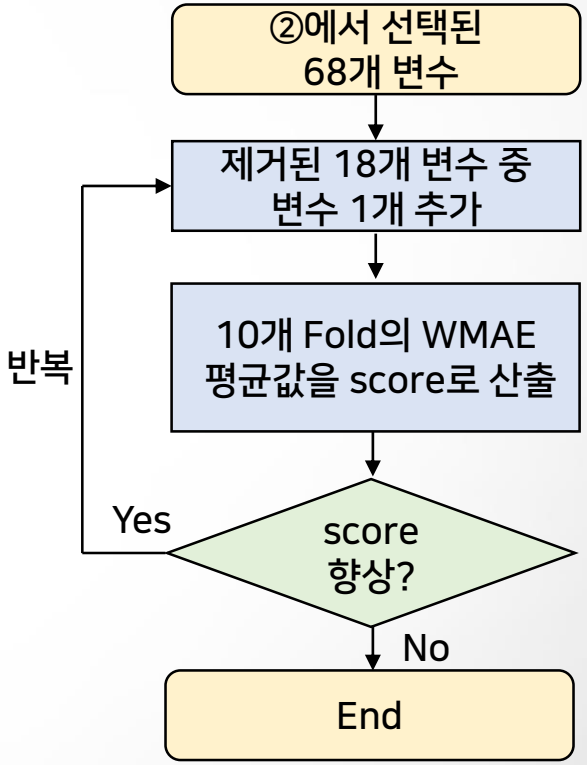
Train set에서 Fold간 분포가 최대한 같은 10개의 Fold를 찾는다.

② Top down Elimination



68개 변수가 선택되었고 18개 변수가 제거되었다.

③ Bottom up Elimination



1개의 변수가 추가되어 최종적으로 69개의 변수가 선택되어 이를 분석에 사용한다.

3. 불량률 예측 알고리즘 (2) 변수 선택(Feature Selection)

1. 모든 86개의 변수를 학습하여 fold=10의 cross-validation WMAE값 계산 → CV: 0.71

2. 총 86개의 변수에서 어떤 변수를 제거해야 WMAE가 낮아질까?

1) 86개에서 x1을 제거해서 train -> 10개 fold의 cv값 0.72
2) 86개에서 x2를 제거해서 train -> 10개 fold의 cv값 0.68 → 이 경우 CV WMAE값이 가장 낮음

⋮

85) 86개에서 time_sequence를 제거해서 train -> 10개 fold의 cv값 0.73

x2 제거

3. 총 85개의 변수에서 어떤 변수를 제거해야 WMAE가 낮아질까?

1) 85개에서 x1을 제거해서 train -> 10개 fold의 cv값 0.68
2) 85개에서 x3를 제거해서 train -> 10개 fold의 cv값 0.64 → 이 경우 CV WMAE값이 가장 낮음

⋮

85) 85개에서 time_sequence를 제거해서 train -> 10개 fold의 cv값 0.72

x3 제거

4. 총 84개의 변수에서 어떤 변수를 제거해야 WMAE가 낮아질까?

1) 84개에서 x1을 제거해서 train -> 10개 fold의 cv값 0.65
2) 84개에서 x4를 제거해서 train -> 10개 fold의 cv값 0.66

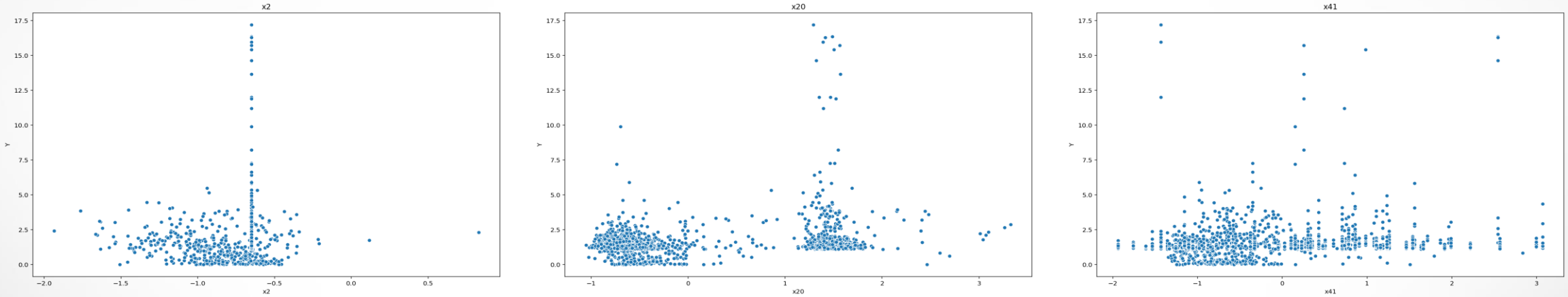
85) 84개에서 time_sequence를 제거해서 train -> 10개 fold의 cv값 0.70

cv값이 0.64보다 낮아진 경우가 없음
해당 84개의 변수가 최적의 조합이라 판단하여
더 이상 변수 제거하지 않고 종료

3. 불량률 예측 알고리즘 (3) 선정 이유

- **이유1: 데이터가 비선형적 분포 양상을 띄며 독립변수 간 상관관계가 매우 높다.**

앞선 탐색적 데이터 분석(EDA) 시 각각의 독립변수와 종속변수의 관계는 아래의 예시 그림과 같이 비선형적 양상을 띄는 것을 확인해 볼 수 있었다.



선형 회귀 모델의 경우 모수적 기법(parametric method)으로 1차 선형식을 가정하게 되는데
선형 모델의 경우, 위와 같이 비선형적 특징을 갖는 우리의 데이터와 맞지 않는다고 판단했다.
또한, 독립변수 간 상관관계가 매우 높은 경우가 대다수였기 때문에 선형 회귀 모델을 사용하게 되면 다중공선성 문제가 발생될 수 있다.

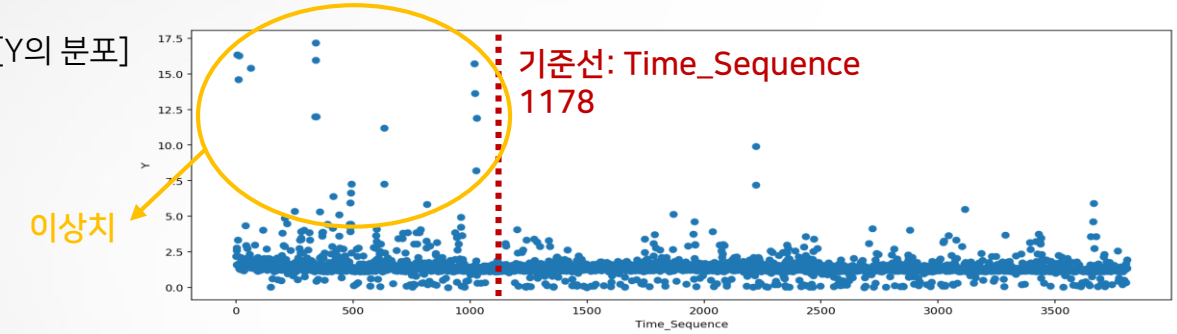
따라서, 각 독립변수마다 데이터의 분포 양상이 가지각색이어서 데이터의 분포 양상을 명시적인 함수 형태의 모델로 가정하기가 어렵기 때문에 **비모수적 기법(non-parametric method)이면서 비선형 모델 중 하나인 KNN을 최종 모델로 선정**하였다.

3. 불량률 예측 알고리즘 (3) 선정 이유

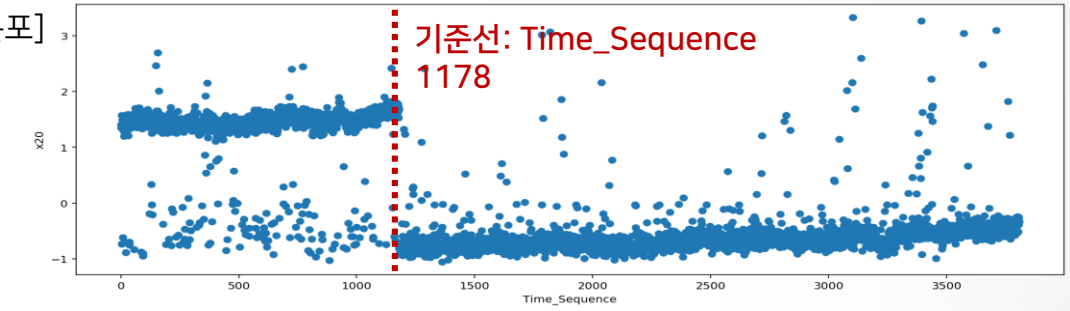
- **이유2: 이상치 예측에 효과적이다.**

① 전체 데이터에서 Time_Sequence에 따른 데이터 분포

[Y의 분포]

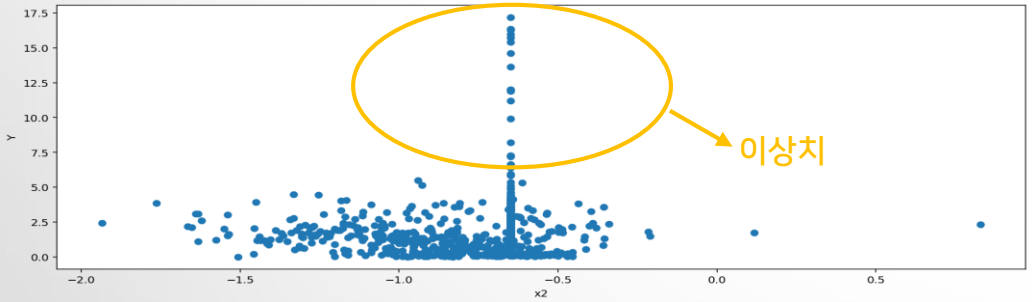


[X의 분포]



전체 데이터에서 Y의 분포를 보면 Y의 평균은 약 1.45 정도지만 값이 5가 넘는 이상치가 존재함을 알 수 있다. 위의 [Y의 분포] 그래프에서 이러한 이상치들은 대부분 Time_Sequence가 1178 이전에 발생하며, 이상치가 발생하는 Time_Sequence는 대부분 비슷한 영역대의 Sequence에서 발생한다. 또한, 위의 [X의 분포] 그래프에서 몇 몇 특정 센서들은 Time_Sequence 1178 이전으로 값이 급격하게 바뀐다.

② 특정 센서의 특정 범위의 값이 이상치를 유발함



이상치들은 특정 센서에서의 값이 유사함을 알 수 있다.

이상치의 값이 보통의 데이터 값의 평균인 1.45보다 10배가 넘는 값을 가지는 경우가 존재하므로, 이러한 이상치들의 값을 얼마나 더 근사하게 예측하느냐에 따라 WMAE의 값이 크게 좌우된다.

EDA 결과, **이상치를 유발하는 Time_Sequence의 각 센서들은 값이 유사한 특징을 가지고 있고 KNN은 데이터간의 거리, 즉 유사도를 기반으로 하는 모델이기 때문에**

KNN은 이러한 이상치를 잘 예측할 수 있는 모델이다.

2

취식 브랜드 수요량 예측

AGENDA

1. 문제 정의 및 데이터 설명
2. 활용 데이터 정의
3. 데이터 전처리 및 파생변수 생성
4. 수요량 예측 알고리즘

1. 문제 정의 및 데이터 설명

문제

취식 브랜드 수요량 예측

Samsung SDS 사내 식당 **취식 데이터**를 활용
지정된 기간(2019.08.01~09.11)의
13개 브랜드별 **일일 취식 수요량 예측**

제공 데이터

- 브랜드별 취식 데이터
  '18년 1월 ~ '19년 7월까지의
  13개 브랜드 별 중식 취식 데이터
- 브랜드별 8월 식단 데이터
  '19년 8월 한 달동안 13개 브랜드의 메뉴 식단
- 이용객 데이터
  '18년 1월 ~ '19 7월까지의
  식당 이용객에 대한 기본 데이터

데이터 분석 방법론 SEMMA

1 Sample

- 제공된 데이터를 학습, 테스트, 예측 데이터로 분리
- 논리적 추론을 통해 외부 데이터 수집
  - 날씨
  - 대기정보
  - 사원 연차 경향

2 Explore

- 제공된 데이터와 외부 데이터 병합 후 메뉴 수요량과의 관계 확인
- 수요량의 외부 요인과 내부 요인을 파악
  - 외부요인 : 날씨, 대기정보
  - 내부요인 : 사원 연차 경향

3 Modify

- **변수 생성**
  - 수요량이 많아지는 메뉴
  - 과거 수요량
- **변수 선택**
  - Top-down Elimination[1]
  - Bottom-up Elimination[1]

4 Model

- **모델 선정**
  - Scaled KNN[2]
  - XGBoost
  - LightGBM
  - Linear Regression

5 Assess

- **Ensemble 모델 선택**
  - Scaled KNN
  - XGBoost
  - LightGBM
- **모델 평가**
  - K-Fold Cross Validation
- **결과**
  - 취식 브랜드 수요량 예측

1. 문제 정의 및 데이터 설명 (1) 브랜드별 취식 데이터

브랜드별 취식 데이터

	SELL_DATE	CUSTOMER_ID	BRAND	MENU	PRICE	QUANTITY
0	2018-01-02	2718043	Chef'sCounter	헬스기빙	6500	1
1	2018-01-02	2271646	우리미각면	의령온메밀국수	6000	1
2	2018-01-02	2577432	나폴리폴리	이탈리안포크라이스	6000	1
3	2018-01-02	2048428	우리미각면	의령온메밀국수	6000	1
4	2018-01-02	2554917	TakeOut	쌈밥	5500	1

8월 식단 데이터

	SELL_DATE	BRAND	MENU
0	2019-08-01	Korean 1	사리곰탕
1	2019-08-01	Korean 2	돈육깻잎볶음
2	2019-08-01	탕맛기픈	개성왕만둣국
3	2019-08-01	탕맛기픈	밀양돼지국밥
4	2019-08-01	가츠엔	탄탄멘

이용객 데이터

	CUSTOMER_ID	GENDER	BIRTH_YEAR
0	2111697	남	1969
1	3571091	남	1964
2	2111628	남	1967
3	2259400	남	1974
4	2010653	남	1975

2. 활용 데이터 정의 (1) 기본 제공 데이터

Samsung SDS에서 제공된 '브랜드별 취식 데이터', '식단 데이터', '이용고객 데이터' 중,
향후 데이터 확보의 안정성과 결측치 문제를 고려하여 '브랜드별 취식 데이터'와 '식단 데이터'를 택하여 활용

브랜드별 취식 데이터

요약: 일별 발생한 취식 데이터
기간: 2018년 1월 2일 ~ 2019년 07월 31일

컬럼명
- SELL_DATE(날짜)
- CUSTOMER_ID(고객식별자)
- BRAND(브랜드명)
- MENU(메뉴명)
- PRICE(가격)
- QUANTITY(수량)

	SELL_DATE	CUSTOMER_ID	BRAND	MENU	PRICE	QUANTITY
0	2018-01-02	2718043	Chef'sCounter	헬스기빙	6500	1
1	2018-01-02	2271646	우리미각면	의령온메밀국수	6000	1
2	2018-01-02	2577432	나폴리폴리	이탈리안포크라이스	6000	1
3	2018-01-02	2048428	우리미각면	의령온메밀국수	6000	1
4	2018-01-02	2554917	TakeOut	쌈밥	5500	1

식단 데이터

요약: 날짜/브랜드별 식단 메뉴 데이터
기간: 2019년 8월 한달 간 주별 식단 데이터

'19.08 이전까지의 식단 데이터 보완
- '브랜드별 취식 데이터'로부터 날짜/브랜드별 식단 메뉴를 추출하여 기존 식단 데이터 보완

'19.09 이후의 식단 데이터는 제공되지 않음
- 업계 관련자의 interview를 통해, 향후 식단 데이터를 확보하는 것이 안정적일 것이라 예상
- 1)취식 수요량 예측의 key data라는 점, 2)식단을 고려한 수요량 예측 시, 식단 재료 수급의 효율적인 예산 운영이 가능하다는 이점을 고려하여 활용 방법 고안

이용 고객 데이터

요약: 사내 식당을 이용한 고객 데이터
기간: 2018년 1월 2일 ~ 2019년 7월 31일

데이터 제외 이유
- 검증 데이터 기간인 8, 9월 데이터 제공되지 않아 예측 시 활용할 수 없음
- 학습 데이터 기간 동안의 이용 고객 변동이 가변적 (Ex. 입사, 퇴사, 해외 출장 등)
- 이용 고객 데이터에서 이용 고객 정보(성별, 생년) 결측치 존재
- 왜곡된 예측 모델 훈련 우려

데이터 전처리 후 활용

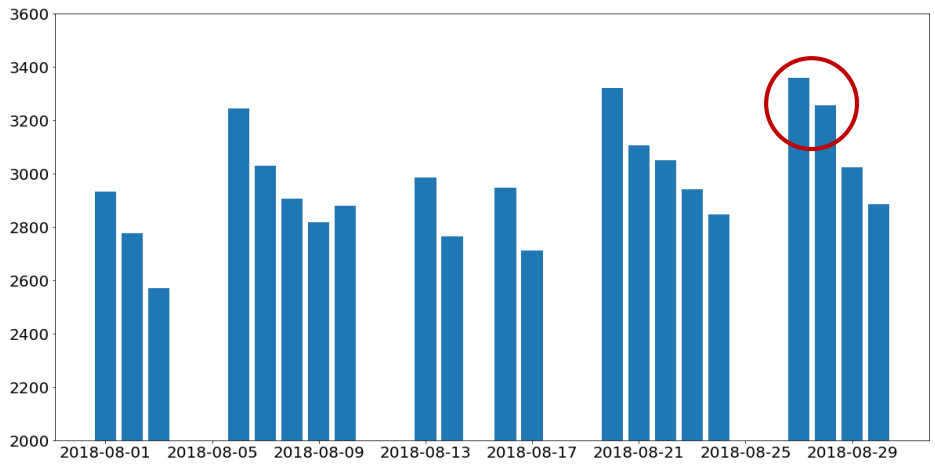
데이터 제외

2. 활용 데이터 정의 (1) 기본 제공 데이터

EDA 과정에서 1)외부 식당을 이용하기에 불편한 환경이었는지, 2)연차로 인한 사원 공석 수가 많은 날인지가 사내 식당 이용률 및 취식 수요량에 영향을 주는 중요 요인이라 분석됨

EDA 결과

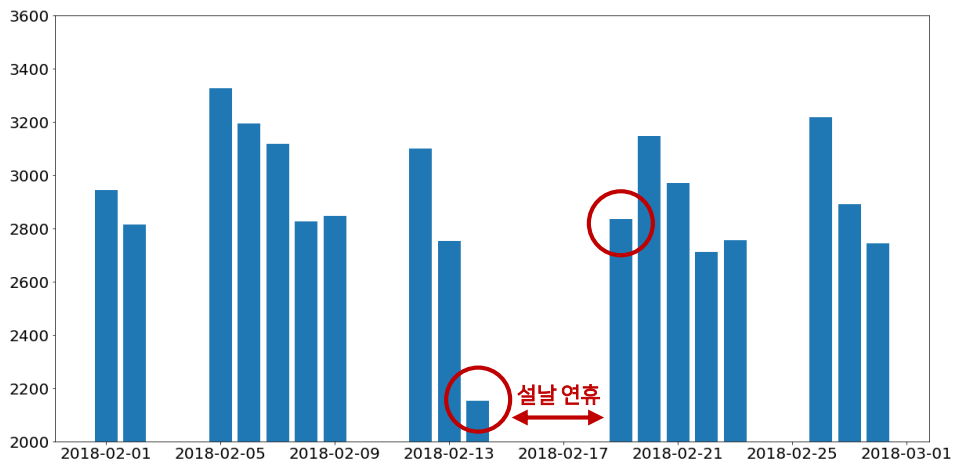
2018-08-01 ~ 2018-08-31 동안의 취식 총 수요량



수요량 peak 지점의 경우, 비가 많이 왔던 날

외부 식당을 이용하기에 불편한 환경이었는지

2018-02-01 ~ 2018-02-28 동안의 취식 총 수요량



직장인들이 연차를 주로 쓰는
설날 연휴 앞뒤 날짜로 사내 취식 수요량 저하

연차로 인한 사원 공석 수가 많은 날인지

2. 활용 데이터 정의 (2) 추가 외부 데이터

'외부 식당을 이용하기에 불편한 환경이었는지', '연차로 인한 사원 공석 수가 많은 날인지'를
1)날씨, 2)대기 정보, 3)사원 연차 추천 날짜 데이터를 통해 변수를 추가하여 예측 모델에 반영

외부 식당을 이용하기에 불편한 환경이었는지

날씨

출처: 기상자료개방포털
기간: 2018년 1월 1일 ~ 2019년 08월 30일
지역: 서울특별시

변수
- HIGHEST_TEMP(최고기온)
- LOWEST_TEMP(최저기온)
- AVERAGE_TEMP(평균기온)
- RAINING_HOUR(강수시간)
- RAIN_PER_DAY(강수량)
- AVERAGE_WIND(평균풍속)
- AVERAGE_WIND(평균습도)

결측치 처리
2019년 9월 결측치는 기상 예보값으로 대치

대기 정보

출처: 에어코리아
기간: 2018년 1월 1일 ~ 2019년 08월 30일
지역: 서울특별시 송파구

변수
- DUST(미세먼지[1])
- MDUST(초미세먼지[2])

결측치 처리
2019년 9월 결측치는 2018년 해당기간 중앙값으로 대치

연차로 인한 사원 공석 수가 많은 날인지

사원 연차 경향

출처: 연차 추천 달력
기간: 2018년 1월 1일 ~ 2019년 12월 31일

변수
- DAYOFF (연차추천날짜)
- DAYOFF_1BEFORE (연차추천날짜 하루전)
- DAYOFF_2BEFORE (연차추천날짜 이틀전)
- DAYOFF_1AFTER (연차추천날짜 하루후)
- DAYOFF_2AFTER (연차추천날짜 이틀후)

참고: [1]미세먼지: [PM-10 (μg/㎥)], [2]초미세먼지: [PM-25 (μg/㎥)]

3. 데이터 전처리 및 파생변수 생성

'브랜드별 취식 데이터'를 활용하여 1)일일 브랜드별 판매 메뉴, 2)일일 브랜드별 총수요량 데이터프레임 생성
이를 통해, 수요량이 많아지는 브랜드별 메뉴와 과거 수요량 추세를 파생 변수를 통해 반영

브랜드별 취식 데이터

요약: 일별 발생한 취식 데이터
기간: 2018년 1월 2일 ~ 2019년 07월 31일

컬럼명

- SELL_DATE(날짜)
- CUSTOMER_ID(고객식별자)
- BRAND(브랜드명)
- MENU(메뉴명)
- PRICE(가격)
- QUANTITY(수량)

	SELL_DATE	CUSTOMER_ID	BRAND	MENU	PRICE	QUANTITY
0	2018-01-02	2718043	Chef'sCounter	헬스기빙	6500	1
1	2018-01-02	2271646	우리미각면	의령온메밀국수	6000	1
2	2018-01-02	2577432	나폴리폴리	이탈리안포크라이스	6000	1
3	2018-01-02	2048428	우리미각면	의령온메밀국수	6000	1
4	2018-01-02	2554917	TakeOut	쌈밥	5500	1

일일 브랜드별 판매 메뉴 & 메뉴별 수요량

	SELL_DATE	BRAND	MENU	QUANTITY
0	2018-01-02	KOREAN1	사골떡만둣국	529
1	2018-01-03	KOREAN1	동태매운탕	492
2	2018-01-04	KOREAN1	돼지고기김치찌개	507
3	2018-01-05	KOREAN1	나가사키부대찌개	424
4	2018-01-08	KOREAN1	꽃게된장찌개	2

인기가 많은 메뉴는?

일일 브랜드별 총수요량

	SELL_DATE	BRAND	QUANTITY
0	2018-01-02	Chef'sCounter	134
1	2018-01-02	KOREAN1	529
2	2018-01-02	KOREAN2	417
3	2018-01-02	TakeOut	390
4	2018-01-02	Western	125

시계열적인 수요 변화는?

3. 데이터 전처리 및 파생변수 생성

'브랜드별 취식 데이터'를 활용하여 1)일일 브랜드별 판매 메뉴, 2)일일 브랜드별 총수요량 데이터프레임 생성
이를 통해, 수요량이 많아지는 브랜드별 메뉴와 과거 수요량 추세를 파생 변수를 통해 반영

일일 브랜드별 판매 메뉴 & 메뉴별 수요량

수요량이 많아지는 메뉴

SPECIAL_MENU '특식메뉴'
: '복날', '초복', '중복', '말복', 'SDS', '[' 등이 들어간
특별 메뉴가 나온 브랜드에 해당 날짜 더미 변수 부여

POPULAR_MENU1 '인기메뉴1'
: 학습데이터의 브랜드별 수요량 상위 10% 이상
메뉴가 나온 브랜드에 해당 날짜 더미 변수 부여

POPULAR_MENU2 '인기메뉴2'
: 학습데이터의 브랜드별 수요량 상위 25%이상 10%미만
메뉴가 나온 브랜드에 해당 날짜 더미 변수 부여

GROUP[1] '브랜드별 판매량 기준'
: 학습데이터의 브랜드별 판매량 기준으로 메뉴들을
5개의 그룹(0 ~ 4)으로 나누어 categorical 변수 부여

일일 브랜드별 총수요량

과거 수요량 추세

MEDIAN_2MONTHS '이전 두달간 해당 요일 수요량 중앙값'
: 해당 데이터 월 기준, 이전 두달간 해당 요일의
브랜드별 수요량 중앙값

AVERAGE_2MONTHS '이전 두달간 해당 요일 수요량 평균값'
: 해당 데이터 월 기준 이전 두달간 해당 요일의
브랜드별 수요량 평균값

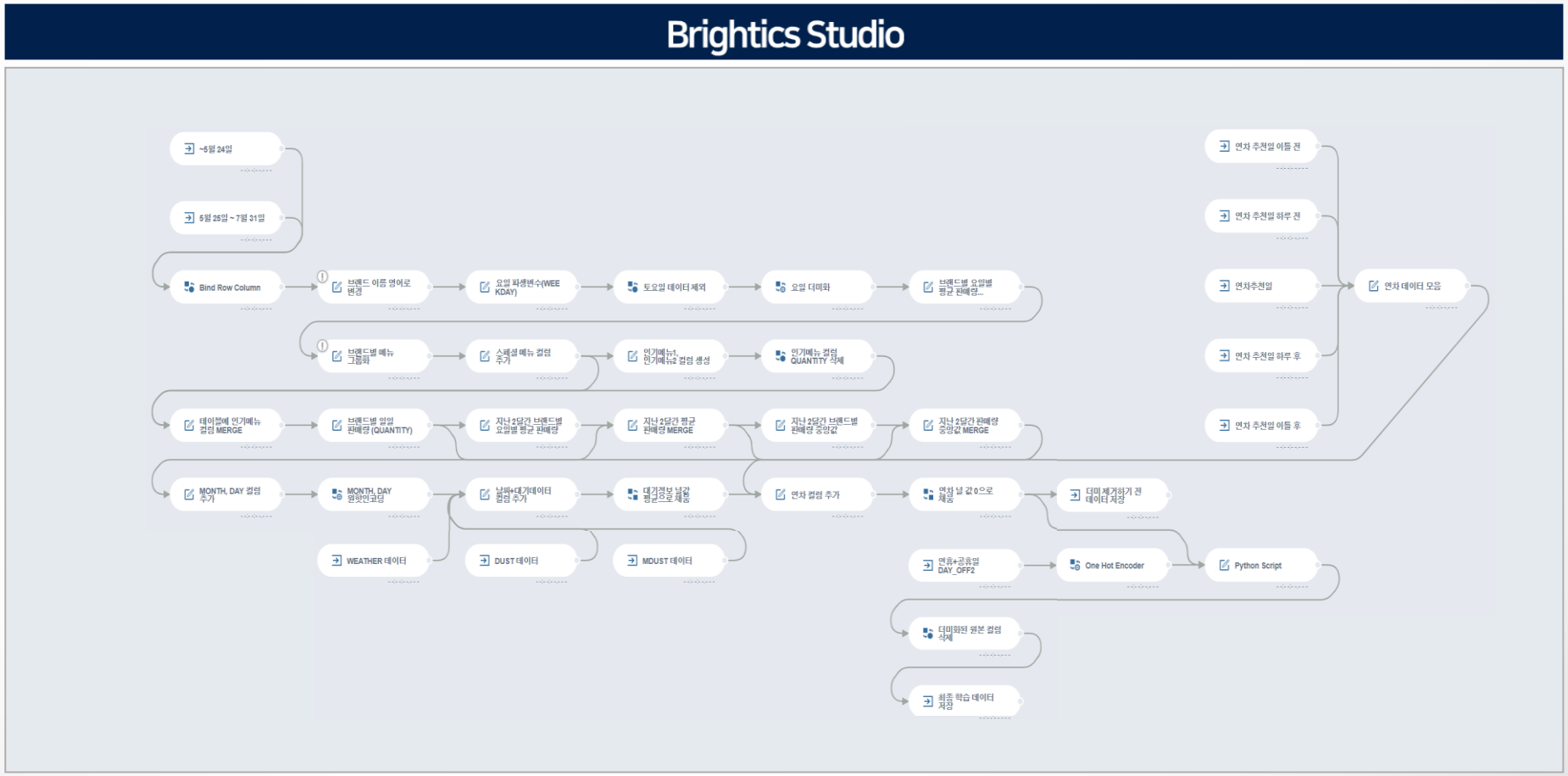
BASE_MEDIAN '각 브랜드 별 하루 수요량 중앙값'
: 학습 데이터 기간동안의 각 브랜드 별 하루 수요량 중앙값

TREND '시계열 데이터의 특성을 나타낼 수 있는 변수'
: 시작 월에 1을 부여하고, 추후 달 마다 1씩 증가하는 변수

참고: [1]0 ~ 4의 그룹(5개)으로 나누어 하루 한 개 이상의 메뉴를 다루는 브랜드는 그룹 중 최솟값(그룹 숫자가 최솟값일 수록 판매량이 많은 메뉴 그룹)으로 처리

3. 데이터 전처리 및 파생변수 생성

Brightics Studio를 통한 데이터 전처리로 예측 모델 학습에 필요한 데이터프레임 생성



4. 수요량 예측 알고리즘 (1) 전체 FLOW

'식단 데이터'의 경우 2019-08-01 ~ 2019-09-11 검증데이터 기간 중 9월 데이터가 제공되지 않았으나 추후 사내 브랜드별 식단 데이터 확보의 안정성과 수요량 예측의 Key Data라는 점을 고려하여 활용 방법을 고안

영양사 Interview

"보통 한 주 전에 식단이 확정이 되는 편인데, 삼성웰스토리나 아워홈같이 규모가 큰 위탁급식 업체는 더 일찍 식단이 확정되는 편이다. 식단을 고려한 수요량 예측의 경우 식단 재료 수급의 효율적인 예산 운영에 도움이 될 것이다."

2019-08-01 ~ 2019-08-30 (21일간): 식단 데이터 제공 O

기존 변수

식단 파생 변수 — 포함

› 모델 학습 › 검증 데이터 예측

2019-09-01 ~ 2019-09-11 (08일간): 식단 데이터 제공 X

기존 변수

식단 파생 변수 — 제외

› 모델 학습 › 검증 데이터 예측

meal_result.csv

8월 검증 데이터 예측 수요량[1]
+
9월 검증 데이터 예측 수요량

참고: [1]2019-08-16 나폴리폴리 미운영으로 해당 날짜의 수요량 0으로 변경, 그만큼 다른 브랜드로 나폴리폴리의 예측 수요량을 분산

4. 수요량 예측 알고리즘 (2) 모델 선정 및 선정 이유

수요량 예측이라는 본래의 목적달성과 사내 식당운영 인사이트 제공이라는 부가적인 목적을 위해
모델 선정 기준을 1)예측력, 2)안정성, 3)해석가능성으로 정하고, 이 기준에 적합한 '앙상블 모델'을 최종 모델로 선정

모델 선정 기준

1. 예측력
모델의 직접적인 효용성은 브랜드별 수요량 예측의 정확도에서 나오므로
높은 예측력을 지닌 모델이어야 함

2. 안정성
모든 모델에는 의도하지 않은 혹은 보완하지 못한 noise가 포함되어 있는데
이러한 noise를 상쇄시킬 수 있는 안정적인 모델이어야 함

3. 해석가능성
취식 수요량 예측이라는 목적 외에도
사내 식당 운영 시 활용할 수 있는 인사이트를 제공하기 위해서
모델의 해석 가능성은 높아야 함

앙상블 모델 적용

Model #1 (XGBoost) × weight
Model #2 (LGBM) × weight
Model #3 (Scaled KNN) × weight

→ Weighted mean
= 일별 브랜드별 수요량

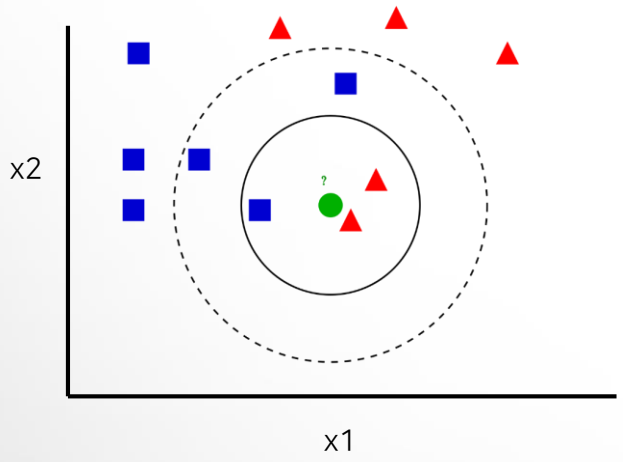
- Ensemble이란 여러개의 모델을 학습시켜
  하나의 모델보다 더 나은 값을 예측하는 방법
- 각자의 장점을 지닌 모델들에 weight을 준
  최종 Ensemble Model을 수립함

4. 수요량 예측 알고리즘 (3-1) 모델 세부 설명

앙상블 모델에 활용한 Scaled KNN은 자체적으로 만든 알고리즘으로 해석력이 높다는 장점을 지님

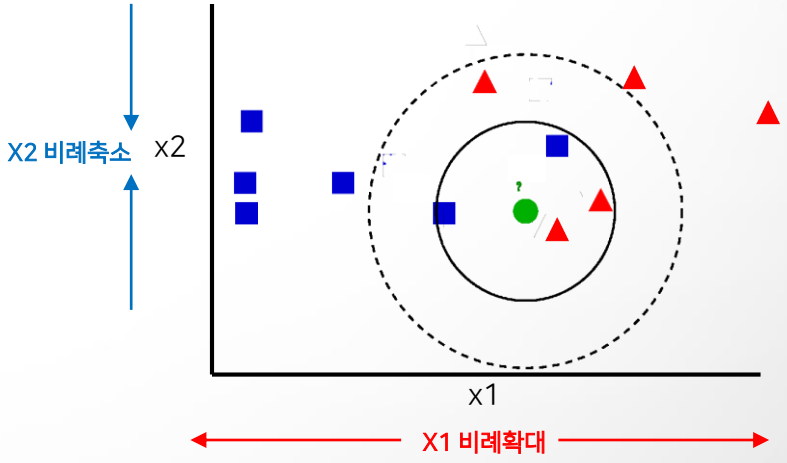
KNN

- 독립변수의 중요도를 고려하지 않음
- 종속변수와의 상관계수를 고려하지 않은 단순 독립변수 벡터와의 거리 계산



Scaled KNN

- 독립변수의 중요도를 계산하기 위해 독립변수 별로 임의의 상수 값으로 곱하거나 나눔 (비례확대, 비례축소)
- 기본, 비례확대, 비례축소 각각의 값으로 학습 하였을 때 cross validation 점수가 높은 값으로 scaling 진행



4. 수요량 예측 알고리즘 (3-1) 모델 세부 설명

모델마다 예측값을 산출하는 메커니즘이 다르기 때문에, 설명력있는 변수들의 조합이 달라질 수 있음
각 모델의 장점을 최대한 취하기 위해 모델별로 Feature selection 과정을 거쳐 최적의 변수 조합을 활용함

참고: [1] p9 자체적으로 만든 변수 선택 프로세스

4. 수요량 예측 알고리즘 (3-2) 모델 예측 성능

Ensemble Model에 활용한 모델은 1)XGBoost, 2)LightGBM, 3)Scaled KNN 세 가지로 각 모델은 아래와 같은 특징을 지니며 이 모델들을 활용한 Ensemble Model은 우수한 성능을 보임

1 XGBoost

Gradient Boosting 알고리즘을 기반으로 함

Boosting은 이전 모델의 Error 데이터에 가중치를 부여하여 다음 모델이 학습을 하므로, 예측을 잘 하지 못하는 데이터에 강점을 지닌다.

2 LightGBM

Decision Tree 알고리즘을 기반으로 함

대량의 데이터를 병렬로 빠르게 학습이 가능하며(XGBoost 대비 2~10배), 예측 정확도가 높은 특징을 지닌다.

3 Scaled KNN

CV의 점수가 높아지도록 KNN을 진행하면서 Feature들의 거리를 점수가 최대 점수가 될 때까지 scaling 한다.

모델 해석력이 우수하다는 장점을 지닌다.

각 모델별 취식 수요량 예측 성능 (KFold = 10 기준)

Model	Best MAE	Worst MAE	Mean MAE
XGBoost	26.457	32.229	29.025
LightGBM	26.321	31.869	28.899
Scaled KNN	24.972	30.604	28.134
Ensemble	25.845	30.305	28.075

5. 모델 분석 결과

Scaled KNN 모델을 기준으로 각 변수의 중요도를 분석한 결과,
사원 연차 경향을 담은 변수들과 식단 데이터 파생 변수들의 설명력이 높게 나타남

각 변수의 중요도를 설명하는 KNN weight

Scaled KNN 모델이 학습하는 과정에서
각 Feature의 값에 대하여 임의의 상수 값(=weight)으로 곱하거나 나누어 CV 점수를 높이는 스케일링 과정을 거치는데,
이 가중치인 KNN weight이 클 수록 변수의 중요도가 크므로, 변수 중요도를 판단할 수 있는 척도가 된다.

KNN weight 상위 6개 변수

변수명	KNN_weight
Dayoff 샌드위치	1723
POPULAR_MENU2	268
Dayoff	115
Dayoff-2	96
Dayoff 연휴전날	95.30
Dayofweek	54

›

사원들의 연차 경향과 식단 메뉴에 따른 수요량 변화를 고려한 식당 운영이 필요함

Thank You

Q & A

Partner Disrupt Foresee