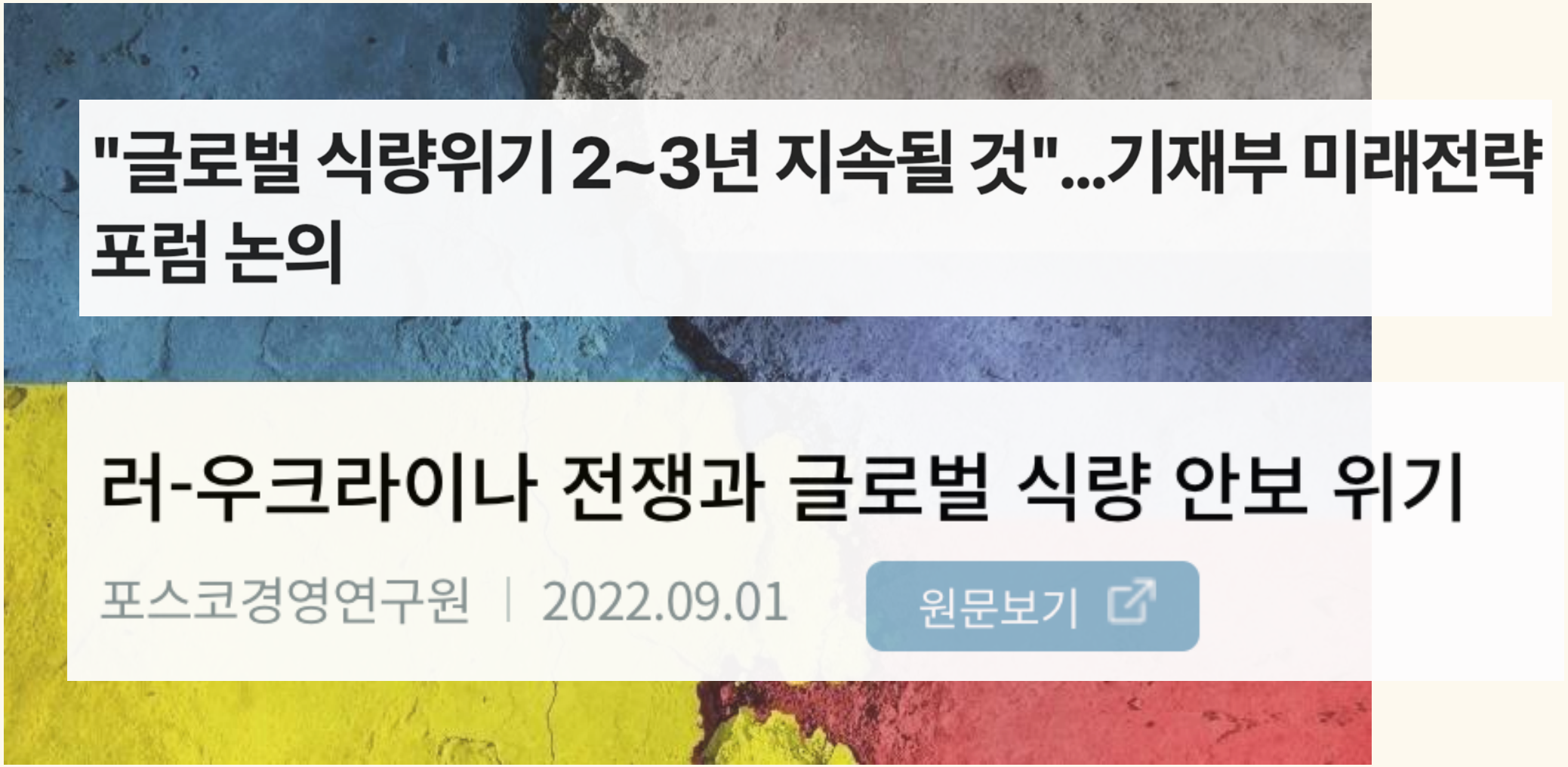
Let's eat rice!

농민을 위한 쌀 단수 예측
- 머신러닝 기법을 이용한 접근 -

5조
신민건 권도윤 김정헌
이지우 추하연 홍정환

1 산업문제

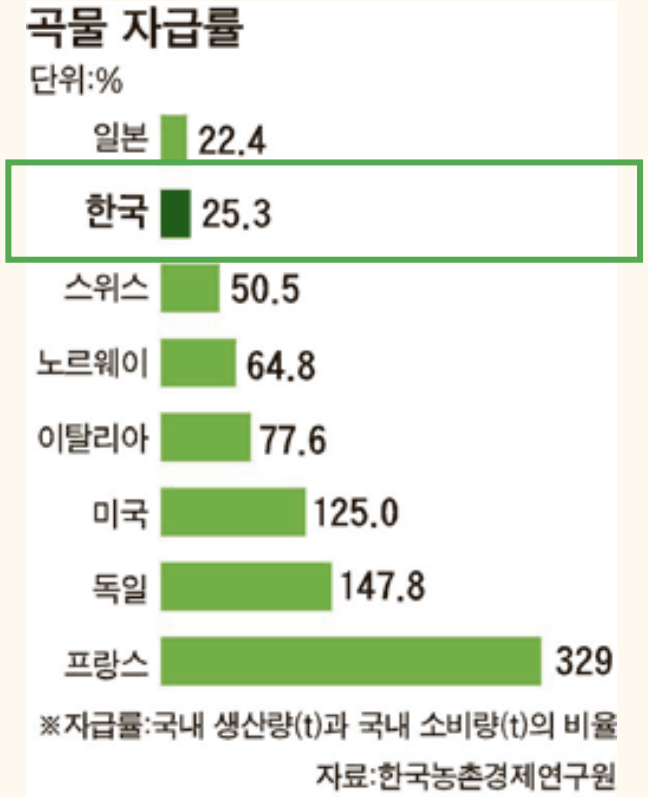
2022년 2월 24일 러시아 우크라이나 전쟁 발발



출처)
한국경제 -
https://www.hankyung.com/economy/article/202211034023Y

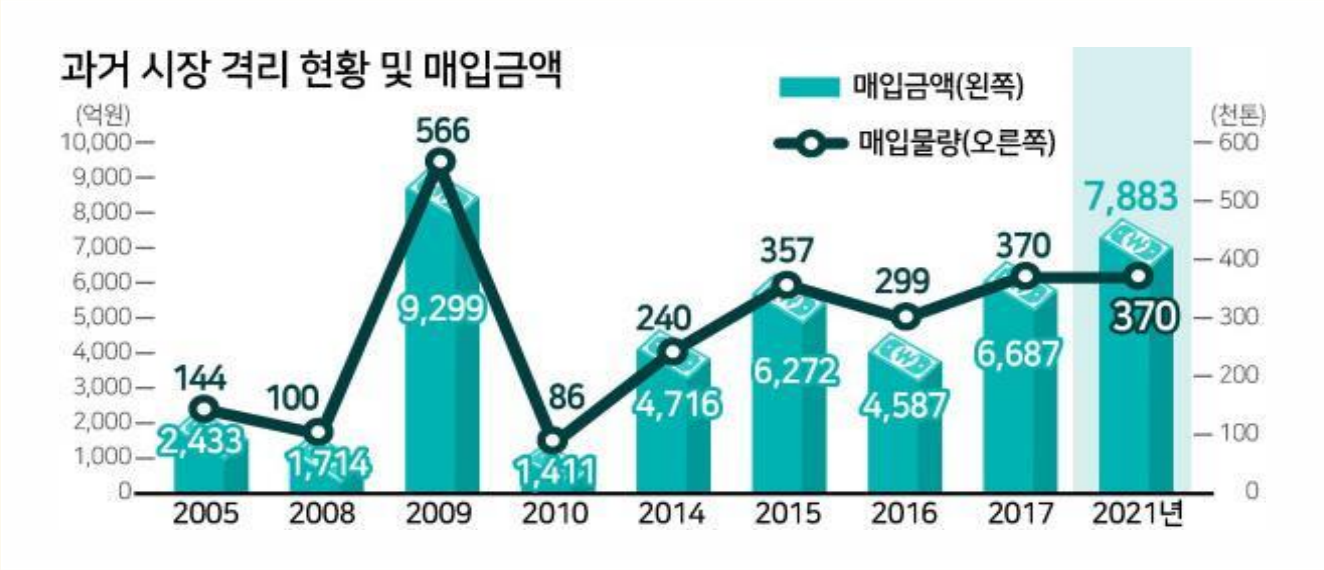
1 산업문제

우리나라 곡물자급률



- 다른나라와 비교 시 하위권

정부의 대응



- 농민 보호차원에서 가격 폭락 방지 위해 과잉 공급량 매입

출처)
한국농촌경제연구원
헤럴드경제 – http://mbiz.heraldcorp.com/view.php?ud=20221005000508

1 산업문제

출처) '쌀 예상 생산량 추정방법에 대한 연구', 강창완(동아대), 김대학(대구가톨릭대)

기존 통계청의 예측방법

통계적추정
: 각 시도별의 대표지역을 표본으로 추출

□ 본 보도자료는 2021년 논벼, 밭벼의 예상생산량을 표본조사하여 추정한 결과입니다.

○ 동 조사는 9월 15일 기준으로 실시되어, 조사 이후 기상여건에 따라 수치가 변동될 수 있음

○ 표본조사 결과에는 표본오차와 비표본오차가 포함되어 있음

새로운 접근방법

머신러닝 기법을 이용한 예측
① 쌀 생산 지역 데이터 최대한 반영하도록
② 과거 누적 데이터 사용

농민신문 - https://www.nongmin.com/opinion/OPP/SNE/CJE/361793/view
한국농정 - http://www.ikpnews.net/news/articleView.html?idxno=47294
머니투데이 - https://news.mt.co.kr/mtview.php?no=2022101112054620322

2 데이터

독립변수

기상관련

월별 평균기온

월별 평균일교차

월별 평균강수량

월별 누적적산온도

년도별 누적일조시간

+

농업관련

농기계대수

농가인구수

농가수

농업용수

논 경지면적

종속변수

단수
(10a 당 생산량)

행정구역별	2020		
	재배면적(ha)	생산량 (톤)	단수 (10a당 생산량: kg)
부산	2169	958	512

2 데이터 – 기상관련 변수 수집

- 기온, 강수량, 일조시간, 일교차 DATA

지점	지점명	일시	기온 (°C)	강수량 (mm)	일조(hr)
159	부산	1999-05-01 00:00:00	14.4	0.0	0.0

- 적산온도 DATA

지점명	년도	일수	값
부산	1996-05-01	122	40.3

기상관련 변수들의
반영기간 및 반영방법

변수	기간	내용
기온	5월 ~ 9월	기간평균
일교차	5월 ~ 9월	기간평균
강수량	5월 ~ 9월	기간평균
적산온도	5월 ~ 8월 중순	누적합계
일조시간	5월 ~ 9월	누적합계

출처) 이동필, "기상요인을 고려한 단수예측모형 개발 연구", 정책연구보고, (2011), 38.

적산온도 : (생육 일수) × (일평균기온)

① 작물의 생육에 필요한 열량을 나타내기 위한 지표
② 일평균기온은 해당 작물이 활동할 수 있는 최저 온도(기준 온도라고 한다) 이상의 것만을 택함 (벼의 경우: 기준 10도)

2 데이터 – 농업관련 변수 수집

- 농기계 대수 DATA

행정구역별	특성별	2020 (경운기 대수)	2020 (콤바인 대수)	2020 (건조기 대수)	2020 (이앙기 대수)	2020 (굴착기 대수)
부산	논벼	1	2	1	3	2

- 논 경지면적 DATA

행정구역별	1996 (논)	1997 (논)	1998 (논)	1999 (논)	2000 (논)	2001 (논)
부산	8425	7851	7474	7310	7147	6694

- 농가인구수, 농가수 DATA

행정구역별	2020 농가 (가구)	2020 농가인구 (명)
부산	8457	18659

- 농업용수 DATA

년도	계	논용수	밭용수	축산용수
2020	4,289.4	2,871.1	1,414.7	3.6

3 분석계획 – 데이터 전처리 - ① 기상데이터

1) QC 플래그 체크

QC 플래그 = 1 → 오류 | QC 플래그 = 9 → 결측치

지점	지점명	일시	기온 (°C)	기온 QC플래그
159	부산	2020-05-02 00:00:00	0.8	1
159	부산	2020-05-02 01:00:00	NaN	9
⋮	⋮	⋮	⋮	⋮

2) 한 컬럼 이상이 비어있다면 지역 제외

지점	지점명	일시	일조시간(hr)
257	양산	2020-05-01 00:00:00	NaN
257	양산	2020-05-01 01:00:00	NaN
⋮	⋮	⋮	⋮
257	양산	2020-09-30 23:00:00	NaN

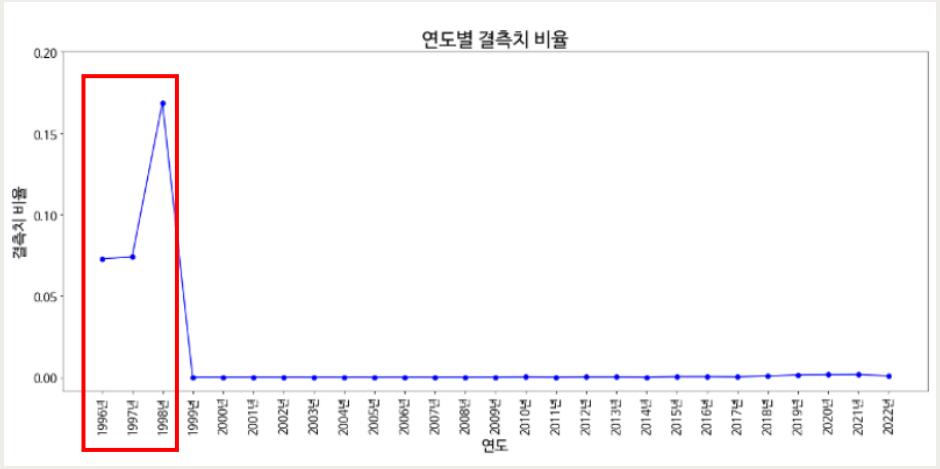
3) 전체의 1/3 이상이 결측치인 컬럼이 있을 경우 해당 지역을 제외

지점	지점명	일시	일조(hr)	강수량(mm)
159	부산	2020-05-01 00:00	NaN	0.3
159	부산	2020-05-01 01:00	NaN	0.3
⋮	⋮	⋮	⋮	⋮
159	부산	2020-05-01 08:00	0.1	0.9
159	부산	2020-05-01 09:00	1	0.9
⋮	⋮	⋮	⋮	⋮
159	부산	2020-05-01 17:00	0.9	1.2
159	부산	2020-05-01 18:00	0.1	1.2
⋮	⋮	⋮	⋮	⋮
159	부산	2020-05-02 00:00	NaN	NaN
159	부산	2020-05-02 01:00	NaN	NaN

강수량, 일조시간의 경우
측정 방법 특성 상,
높은 결측치 비율 ≠ 데이터의 incompleteness

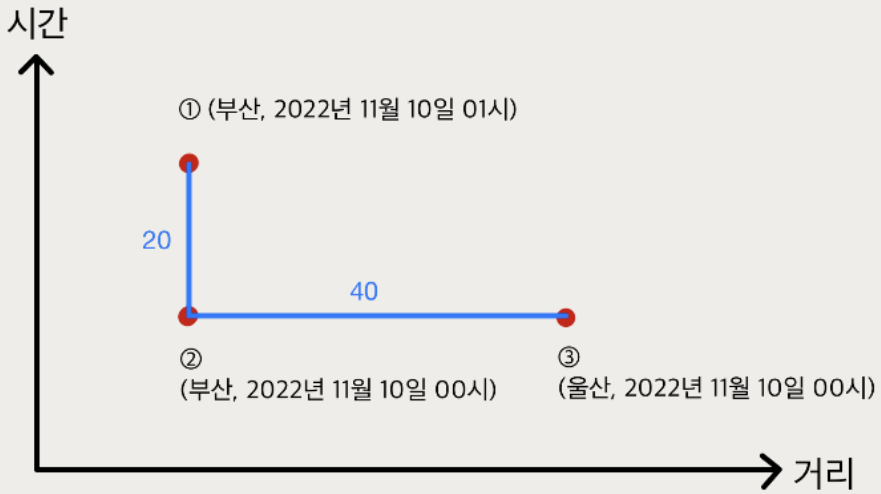
3 분석계획 – 데이터 전처리 - ① 기상데이터

기온, 일교차



96, 97, 98년도에 결측치가 많음을 알 수 있음
➞ KNN 사용

나머지
➞ Interpolate



KNN 거리계산 예시

강수량

비가 온 날 : fillna(0)
비가 오지 않은 날 : Interpolate

일조시간

해가 떠있는 시간 : Interpolate
그 외의 시간 : fillna(0)

적산온도

결측치 X

3 분석계획 – 데이터 전처리 - ② 농업데이터

농기계

Linear
interpolate

농가인구, 농가수

Linear
interpolate

농업용수

Linear
interpolate

논 경지면적

결측치 X

3 분석계획 - 최종데이터프레임

독립변수 / **종속변수**

기상관련 변수: 평균기온, 평균일교차, 평균강수량, 누적일조시간, 누적적산온도

농업관련 변수

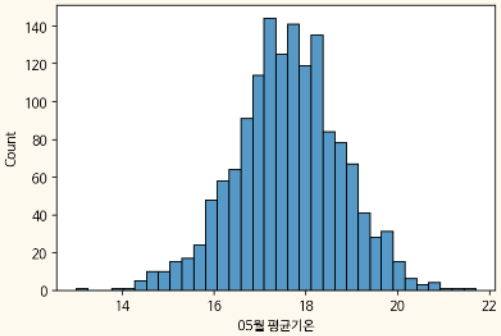
year	지역	5월 평균 기온	6월 평균 기온	7월 평균 기온	8월 평균 기온	9월 평균 기온	5월 평균 일교차	6월 평균 일교차	7월 평균 일교차	8월 평균 일교차	9월 평균 일교차	5월 평균 강수량	6월 평균 강수량	7월 평균 강수량	8월 평균 강수량	9월 평균 강수량	누적 일조 시간	5월 적산 온도	6월 적산 온도	7월 적산 온도	8월 적산 온도	농기계 대수	농가수	농가 인구	농업 용수	경지 면적	단수 (10a당 생산량)
1996	부산	17.94194	20.50472	23.84691	26.20282	22.51556	7.712903	5.006667	5.709677	6.203226	7.213333	0.063038	0.454306	0.382796	0.18871	0.037083	312.4	5802.3	15679.3	28187.95	20826.9	13138	10078	18896	132,638.20	8425	493

⋮

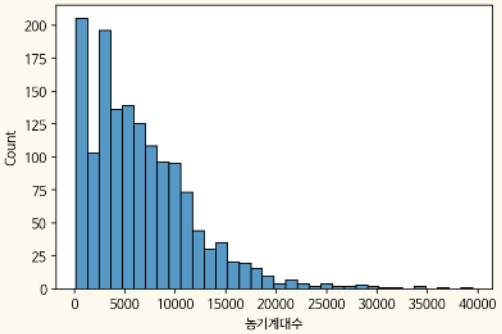
최종데이터프레임에 포함된

년도 수	26년 (1996~2021년)		
지역 수	66개 ('강릉', '거제', … , '해남', '홍천')		
독립변수	25개	종속변수	1개

정규분포를 따르는 독립변수 예시



정규분포를 따르지않는 독립변수 예시



Min-max (Normalization)로
독립변수 정규화 후 모델링 예정

모델 후보

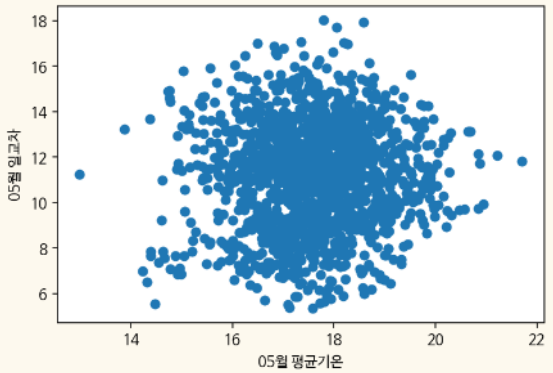
선형회귀

다중선형회귀모형
(Multi linear regression)

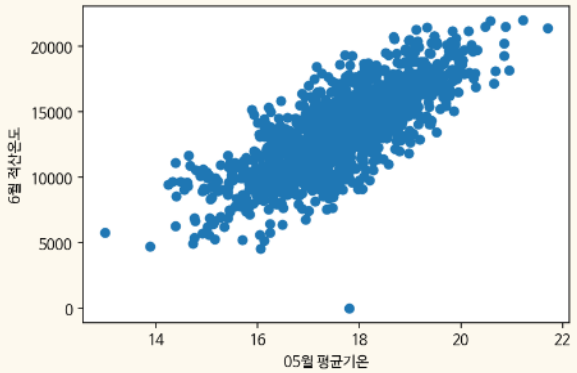
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1,2,\ldots,n$$

- 독립변수: 25개 / 종속변수: 1개
- 독립변수들 간 다중공선성 확인 필요

상관관계를 보이지 않는 두 변수



강한 상관관계를 보이는 두 변수



➞ 현재 독립변수들에는 PCA를 적용하기 어려운 상황
➞ 다중선형회귀모형 사용 보류

비선형회귀

트리기반 회귀모형

Decision Tree	Random forest
Xgboost	lightGBM

- 독립변수들 간 다중공선성 존재여부 상관 無
- 예시: Decision Tree Regressor / 독립변수 1개 – 종속변수 1개

3 분석계획

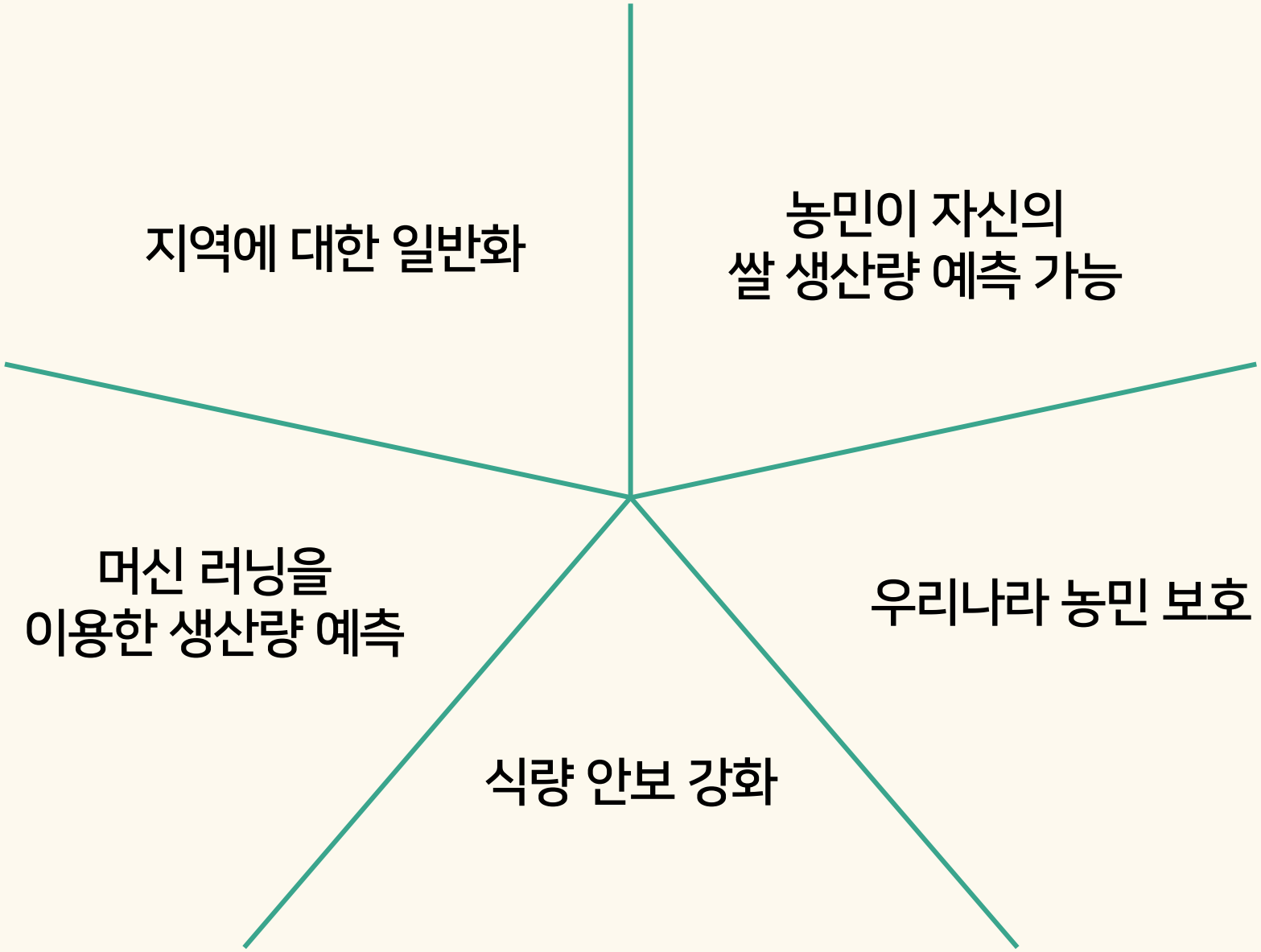
회귀모형 평가지표

평가지표	식	특성 (이점 및 단점)
MSE	$\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$	• 오차민감도 up → 파라미터에 따른 변화를 쉽게 관측 가능 • 이상치의 영향을 많이 받음
$RMSE$	$\sqrt{\sum_{i=1}^{N}\frac{(y_i-\hat{y}_i)^2}{n}}$	• 종속변수와 단위가 같음 → 오차의 해석이 쉬움
MAE	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i-\hat{y}_i\rvert$	• 종속변수와 단위가 같음 → 오차의 해석이 쉬움 • 특이값이 많은 경우 사용하기 좋음
$MAPE$	$\frac{100}{n}\sum_{i=1}^{N}\left\lvert\frac{y_i-\hat{f}(x_i)}{y_i}\right\rvert$	• 스케일의 영향이 적음 • 특이값이 많은 경우 사용하기 좋음

각 평가지표마다 특성(이점 및 단점)이 다르므로,
다양한 기준으로 종합적으로 모델성능평가 및 보완 예정

① 예측력에 따라서 변수추가 or 제거 | ② 모델 하이퍼파라미터 조정

4 기대효과 및 시사점



Q&A