3대 조미채소 가격 예측

-트렌디한 지표를 활용하여

DATA MINING TEAM

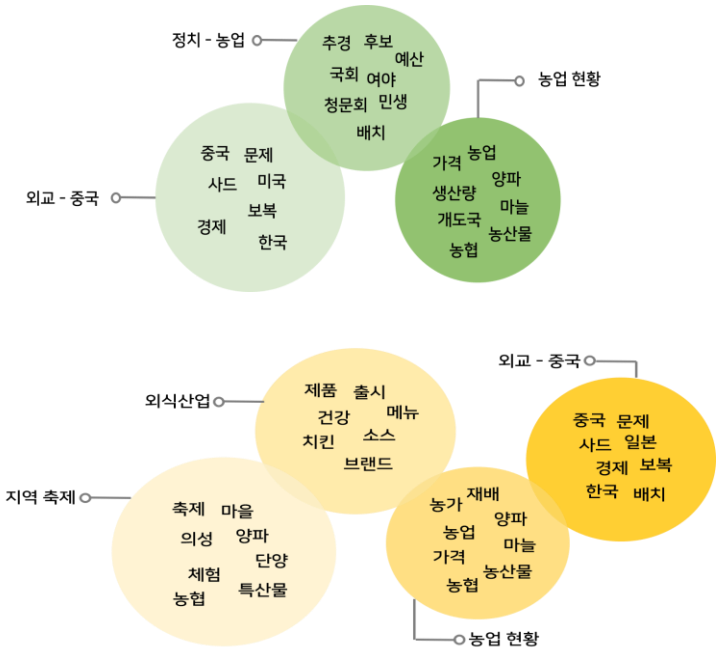
김동영
강수경
김재희
유경민
최윤혜

동영이와 아이들

01 지난주 복습

토픽모델링

LDA(Latent Dirichlet Allocation)



각 토픽들에 대한 특성 파악

특성을 활용한 파생변수 생성

동영이와
아이들

02 파생변수

파생 변수 생성

토픽모델링에서 나온 토픽들로

새로운 변수를 만들어보자!

YouTube

Google Trends

동영이와
아이들

03 Catboost

모델 선정 이유

: Category + Boosting

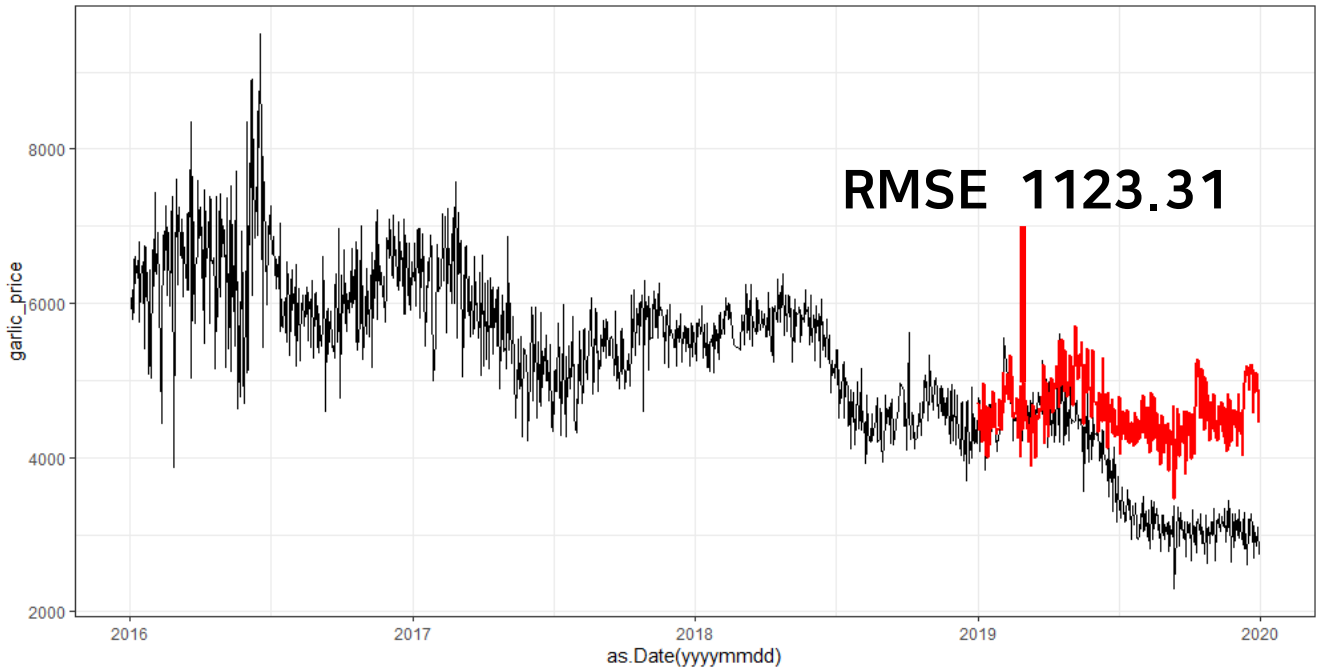
범주형 변수가 많은 데이터에서 좋은 성능을 보이는 모델

→ 도매데이터에 다양한 범주형 변수가 존재하므로 Catboost 사용

```
>glimpse(도매 데이터)
$ 시장명 <fct>"가락동농수산물시장" "강릉농산물도매시장" "강서농수산물도매시장" …
$ 법인명 <fct> "강릉농산물" "강서청과" "경기청과" "경인농산" "광주원협(공)"…
$ 포장상태명 <fct> "PE대" "PP대" "그물망" "단" "봉지"…
$ 등급명 <fct> "4등" "5등" "8등" "등외" "보통" "상" "특"
$ day.y <fct> "금" "목" "수" "월" "일" "토" "화"
$ 출하구분명 <fct> "개별" "계통" "상인" "수입" "협동"
      ⋮
```

03 Catboost

모델링 결과



예측 값이 실제 값의 큰 트렌드를 따라가지 못하는 문제

시계열 요소 고려 필요

04 시계열 모델링

시계열 데이터

1차 차분을 통해 데이터 정상화시킨 결과

H_0 : *non-stationary*

KPSS test　　ADF test　　PP test

P-value <=0.01 이므로 H_0 기각

차분을 통해 정상 시계열 데이터 만들기 성공!!

04 시계열 모델링

ARIMAX

ARIMAX

Auto-Regressive Integrated Moving Average
with explanatory variable or transfer function

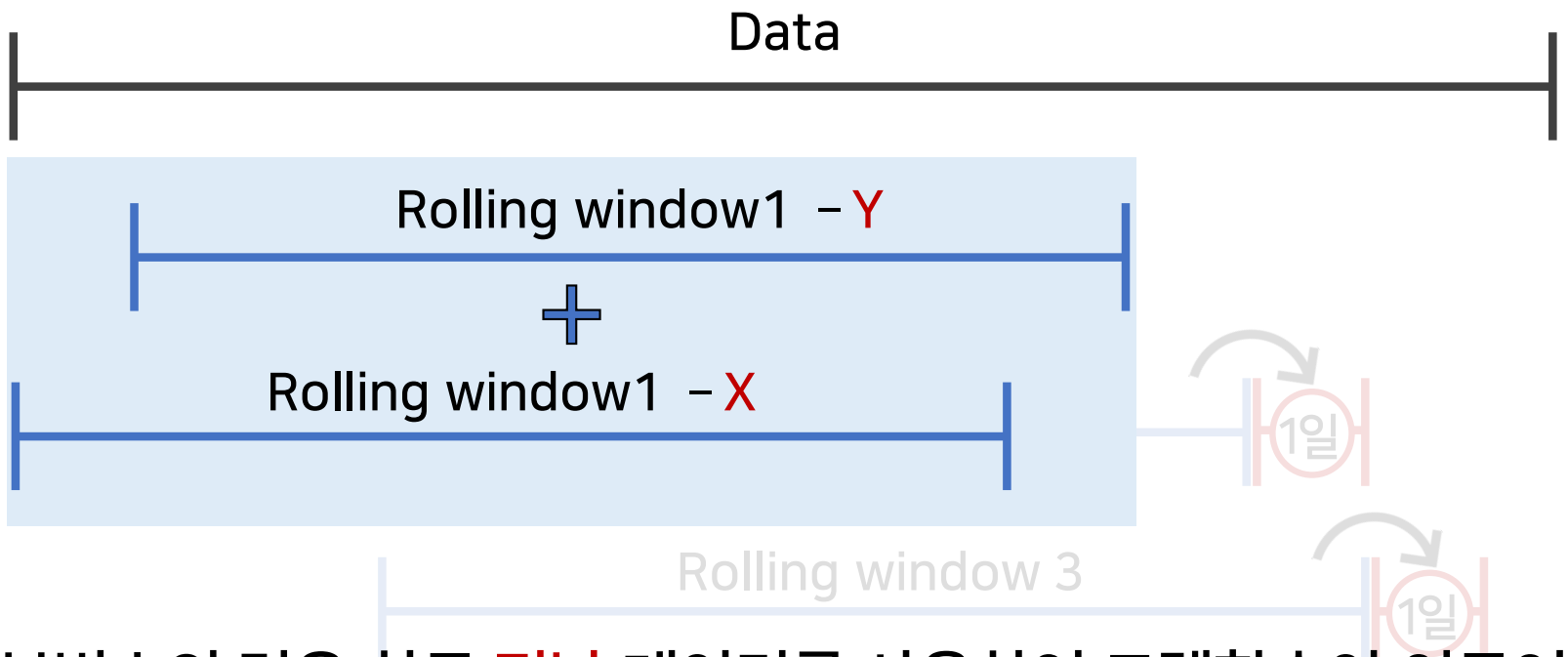
비정상 데이터를 정상화하기 위하여 **차분**하고
현재시점의 상태를 과거시점 **상태**, **오차**들과 **외부변수**로 설명하는 모형

동영이와 아이들

04 시계열 모델링

예측 방법

: Rolling Window



외부변수의 경우 하루 전날 데이터를 사용하여 모델학습이 이루어짐

04 시계열 모델링

ARFIMAX

ARFIMAX

Auto-Regressive Fractionally Integrated Moving Average
with explanatory variable or transfer function

비정상 데이터를 정상화하기 위하여 **분수**형태로 **차분**하고
현재시점의 상태를 과거시점 **상태**, **오차**들과 **외부변수**로 설명하는 모형

동영이와 아이들

04 시계열 모델링

예측 결과

ARIMAX	ARFIMAX
272.28	292.02

ARIMAX의 성능이 ARFIMAX보다 좋음!

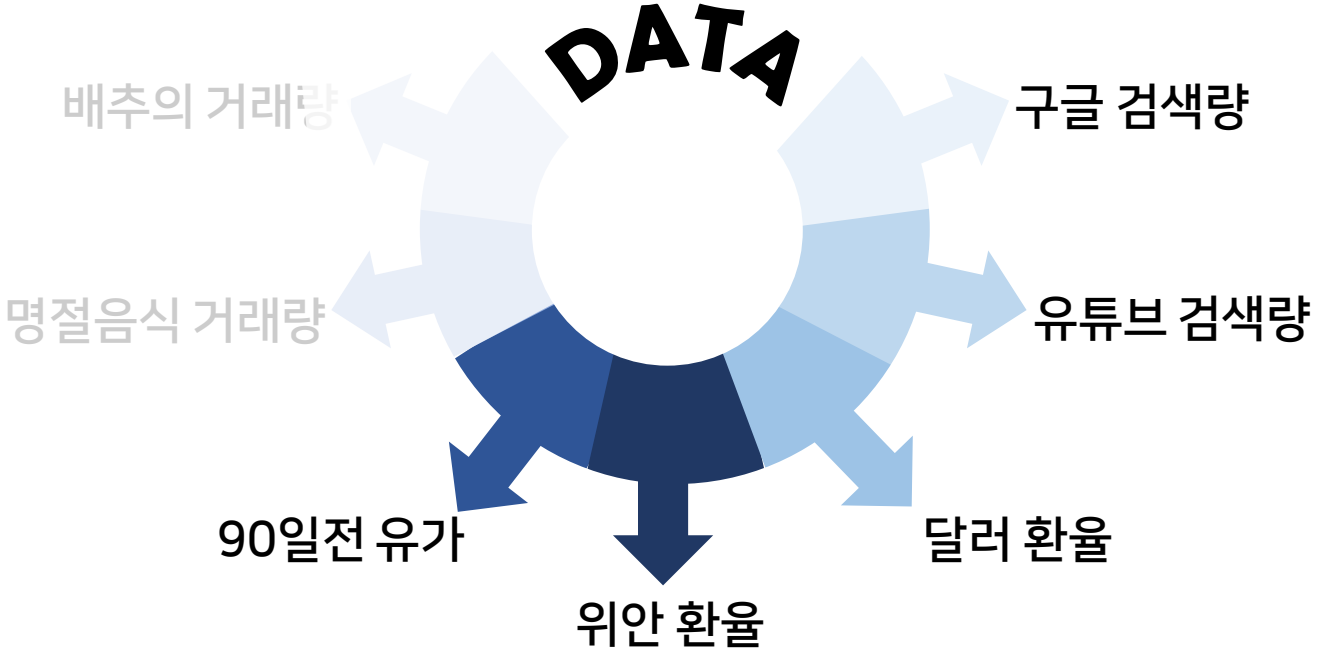
하지만 우리의 데마는 여기서 멈추지 않고

변수 선택을 해볼거예요~

04 시계열 모델링

변수 선택

Best subset selection



변수들의 **모든 조합**을 고려하여 가장 낮은 CV error가 나오는 변수 조합을 선택하는 방법

04 시계열 모델링

변수 선택

ARIMA	ARIMAX
265.59	263.69

ARIMAX의 성능이 ARIMA보다 좋음!

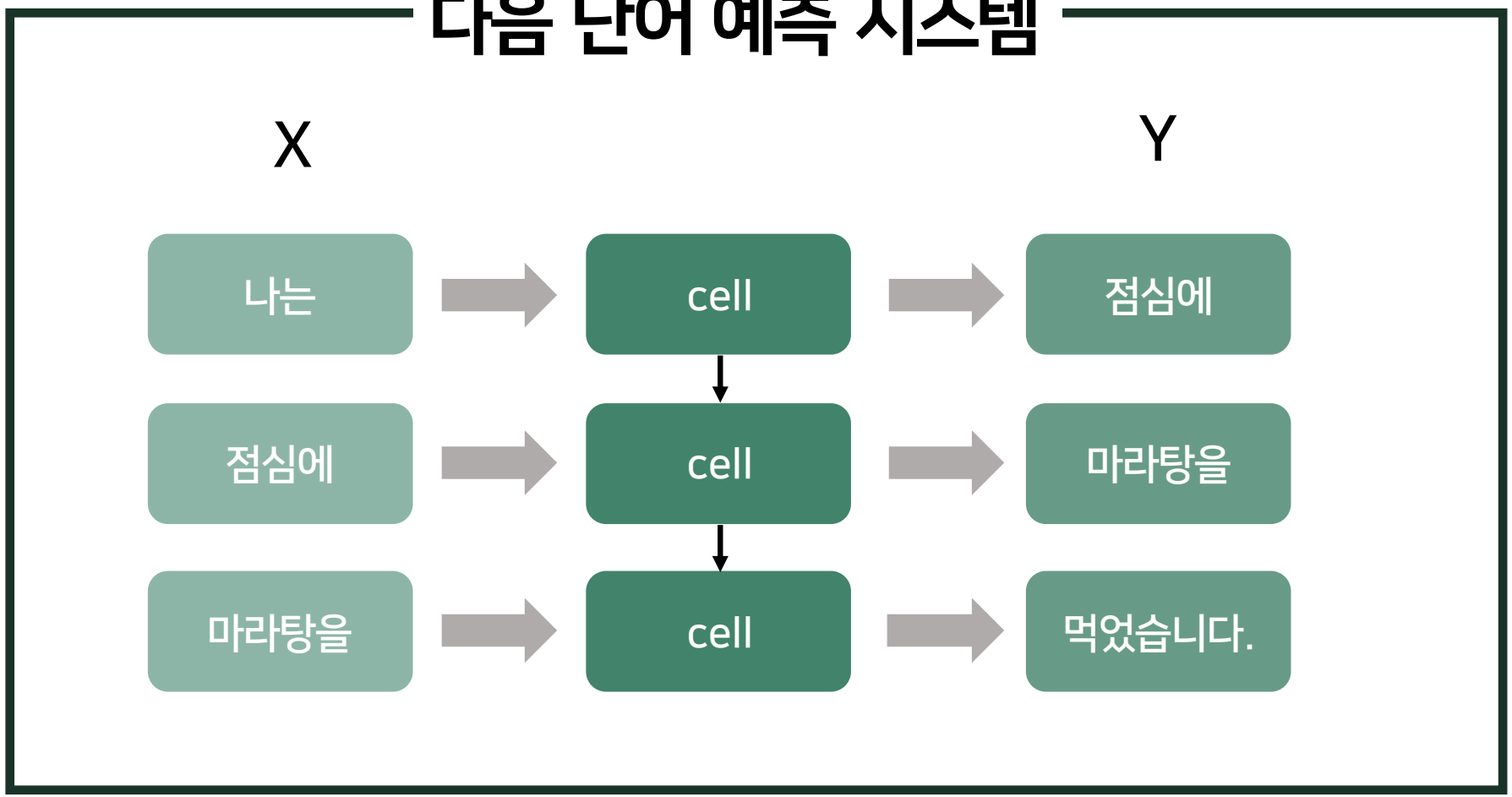
우리의 데마가 해냈습니다....!

트렌트 변수가 효과가 있어요!!!

동영이와
아이들

05 LSTM 모델링

자연어 처리



동영이와
아이들

05 LSTM 모델링

최종 결과

	Catboost	ARIMAX	LSTM
RMSE	1123.31	263.69	269.38
MAE	944.41	196.89	235.76

ARIMAX의 성능이 가장 좋게 나옴!

우리의 최종 모델은 바로 ARIMAX!!!

동영이와 아이들

05 LSTM 모델링

최종 결과

전날 가격을 다음날 가격으로 예측해도 이정도는 나오는거 아니냐고요??

	SIMPLE	ARIMAX
RMSE	376.59	263.69
MAE	290.43	196.89

ARIMAX의 성능이 훨~씬 좋게 나옴!

우리가 만든 모델이 하는 역할이 있어요!!!