



농산물 가격 예측

7, 14, 28일 후의 농산물 가격을 예측하라!

선형대수학팀
이재현 이정우 조혜현 김규범 김민지

INDEX

01

분석배경

02

데이터 소개

03

데이터 전처리


04


시각화 및 EDA


05


감성분석



 분석배경

 데이터 소개

 데이터 전처리

 시각화 및 EDA

 감성분석



분석 목표

기존 농산물 가격 예측 모형을 개선할 수 있는
새로운 아이디어와 알고리즘을 개발해보자!

분석의 의의

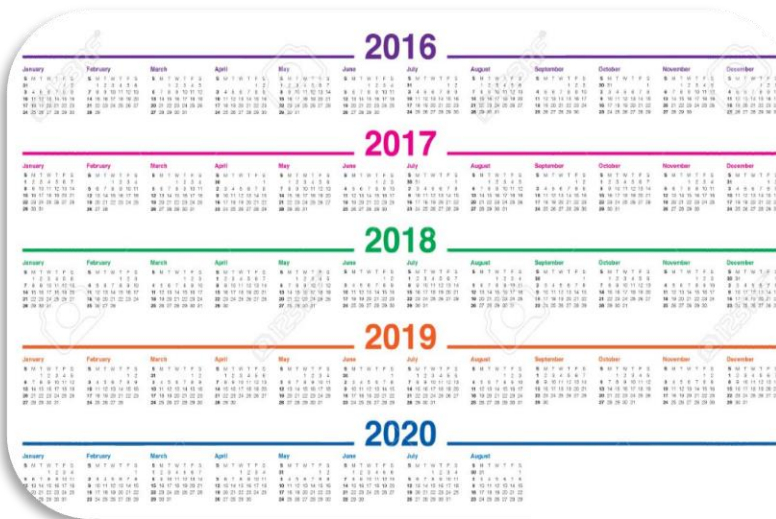
1. 기존 가격예측 서비스는 제공된 지 1년이
조금 넘었기에 모형 개발이 더욱 이루어져야 함
2. 농산물 수급 안정 및 가격 지원에 기여할 수 있음
3. 소비자들의 공정한 거래를 도모할 수 있음



02 데이터 소개

21-2 선형대수학팀 |

Train & Test set



<예측할 날짜>

2020 - 09 - 08

~

2020 - 10 - 26

<Train set>

2016 - 01 - 01 ~ 2020 - 08 - 31

분석배경

데이터 소개

데이터 전처리

시각화 및 EDA


감성분석


■ 변수 전처리


결측치 처리


날짜	요일	가격
2016-01-01	금요일	NA
2016-01-02	토요일	723
2016-01-03	일요일	NA
2016-01-04	월요일	794
2016-01-05	화요일	763
...

가격(Y)의 경우,
공휴일과 일요일에
결측치 존재

 분석배경

 데이터 소개

 데이터 전처리

 시각화 및 EDA

 감성분석

■ 변수 전처리

결측치 처리

날짜	요일	가격
2016-01-01	금요일	NA
2016-01-02	토요일	768
2016-01-03	일요일	NA
2016-01-04	월요일	794
2016-01-05	화요일	763
...

Y값에 결측치가 존재하므로 해당 행 삭제 (Y)의 경우 공휴일과

일요일에 결측치 존재

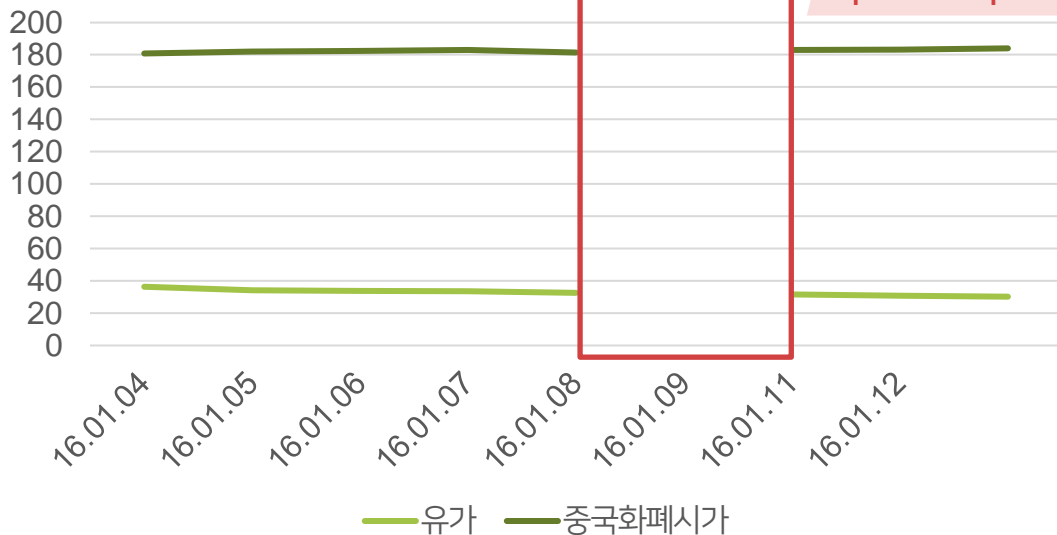
- 🍏 분석배경
- 🍏 데이터 소개
- 🍏 데이터 전처리
- 🍏 시각화 및 EDA
- 🍏 감성분석

■ 변수 전처리

결측치 처리

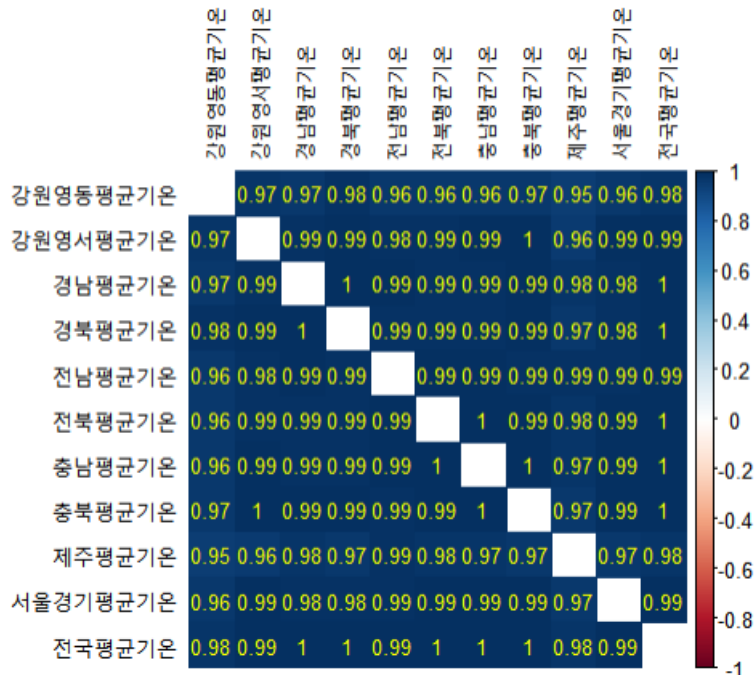
유가, 중국화폐시가

spline Imputation 사용



- 🍏 분석배경
- 🍏 데이터 소개
- 🍏 데이터 전처리
- 🍏 시각화 및 EDA
- 🍏 감성분석

설명변수들 간의 상관관계 분석

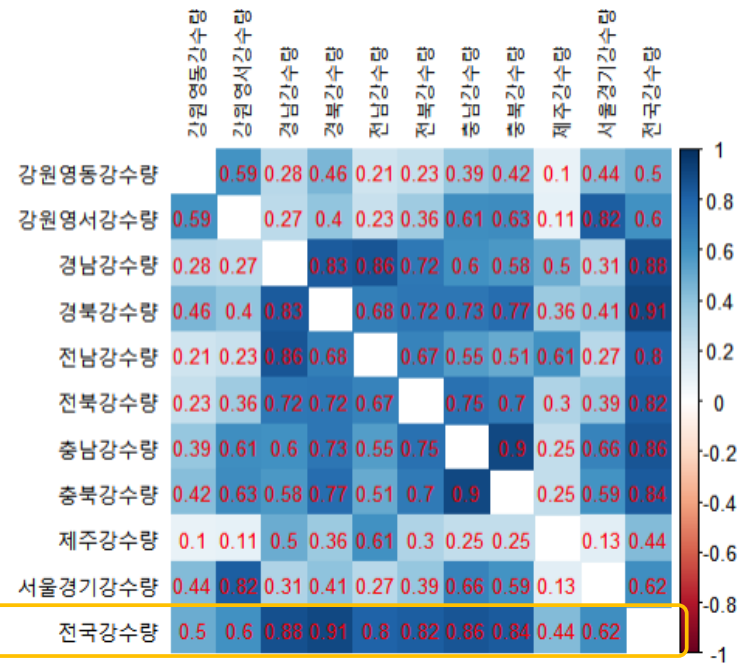


지역별 평균기온 사이의 상관계수가
0.95 이상임을 확인함!



전국 평균기온으로 11개의 변수를 대체

설명변수들 간의 상관관계 분석



지역별 강수량 사이의
상관계수 차이가 큼을 확인!

기온 변수만큼 타지역과의 상관관계가
높지 않아 한 개의 변수로
대체하기 어렵다고 판단

설명변수들 간의 상관관계 분석

시도별	대파 재배면적	시도별	토마토 재배면적	시도별	당근 재배면적
전남	28.1%	강원	16.4%	제주	59%
경기	20%	전남	15.8%
강원	9.2%	충남	13.2%		
전북	8.4%	경남	11.3%		
충북	7.5%	전북	9%		
...		

재배지역이 고루 분포돼 있거나
한 지역에 치우쳐 있어
지역을 선정하기가 **까다로움**



지역별 강수량 변수의 차원 축소
방향 논의 예정

- 분석배경
- 데이터 소개
- 데이터 전처리
- 시각화 및 EDA
- 감성분석

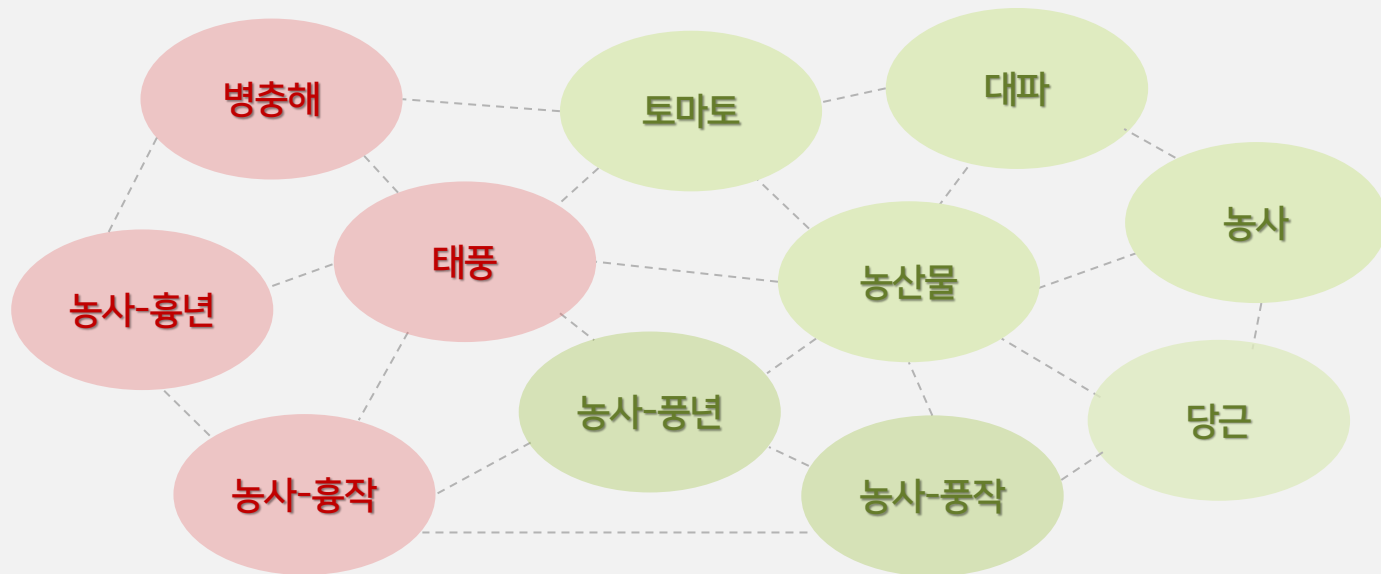
■ 감성분석 FLOW



- 🍏 분석배경
- 🍏 데이터 소개
- 🍏 데이터 전처리
- 🍏 시각화 및 EDA
- 🍏 감성분석

크롤링(Crawling)

키워드 구체화



- Konlpy (형태소 분리 패키지)

Okt 분석기

"걱정"	"면적"
"품종"	"풍작"
"갱신"	"다시마"
"밀식"	"연근"
"재배"	"콩"

Kkma, Hannanum보다
Okt가 단어가 깔끔하게 분리됨
Okt를 사용해서 형태소 분리



앞서 크롤링한 뉴스 기사의 문장을 형태소로 분리해,
추후 사전과 비교하여 분류할 예정

- KNU + 수작업 감성사전

감성 변수


$$\text{Sent1 : } \frac{\text{긍정어} - \text{부정어}}{\text{단어수}}$$


$$\text{Sent2 : } \frac{\text{긍정어}}{\text{단어수}}$$


$$\text{Sent3 : } \frac{\text{긍정어} - \text{부정어}}{\text{긍정어} + \text{부정어}}$$


$$\text{Sent4 : } \frac{\text{긍정어}}{\text{긍정어} + \text{부정어}}$$


여러 선행 논문을 참고하여 4개의 변수 수식 완성

 분석배경

 데이터 소개

 데이터 전처리

 시각화 및 EDA

 감성분석

■ KNU + 수작업 감성사전

감성 변수

날짜	Sent1	Sent2	Sent3	Sent4
2016-01-04	-0.500000	-0.019608	0.250000	0.009804
2016-01-05	0.142857	-0.019608	0.250000	0.009804
2016-01-06	0.142857	0.014085	0.571429	0.056338
2016-01-07	0.142857	0.014085	0.571429	0.040404
2016-01-08	-0.250000	-0.012658	0.375000	0.018987
⋮	⋮	⋮	⋮	⋮

이와 같은 감성사전은 문맥을 고려하지 못한다는 단점이 존재

문맥까지 고려한 감성분석을 진행해보자!


- 분석배경
- 데이터 소개
- 데이터 전처리
- 시각화 및 EDA
- 감성분석


- 문맥 고려 감성분석


SNU 감성사전


단어	값
가/JKC	NEG
가/JKS	POS
가/JKC;되/VV	NEG
가/JKC;되/VV;ㄴ/ETM	NEUT
가/JKC;아니/VCN;면/EC	POS
...	...


문장 안에서 단어의 위치에 따라
각 단어를 분류한 Lexicon 사전

 분석배경

 데이터 소개

 데이터 전처리

 시각화 및 EDA

 감성분석

■ 문맥 고려 감성분석

SNU 감성사전

단어	값
가/JKC	NEG
가/JKS	POS
가/JKC;되/VV	NEG
가/JKC;되/VV;ㄴ/ETM	NEUT
가/JKC;아니/VCN;면/EC	POS
...	...

해당 형태소 단위들은
Kkma에서 지원하므로
SNU 감성사전에서는 Kkma 사용

🍏 분석배경

🍏 데이터 소개

🍏 데이터 전처리

🍏 시각화 및 EDA

🍏 감성분석



농산물 가격 예측

7, 14, 28일 후의 농산물 가격을 예측하라!

선형대수학팀
이재현 이정우 조혜현 김규범 김민지

INDEX

01

변수 수집

02

파생변수 생성

03

변수 검증

04

변수 선택

05

모델링

06

결론



■ 농산물 ETF 변수

인플레이 우려 커져

인플레이션(물가 상승)하는 상장지수펀드(ETF)라는 전망이 이어지자 따르면 미국 시장에 우크리움 워트(Teuc)치다.

뒤를 이어 옥수수에 우크리움 소이빈 ETF 쓴 것이다.



김진영 키움증권 연구원은 "올해는 수확 전망치가 하향 조정되고 있고 농작물 공급난 우려가 심화되면서 관련 대체자산 ETF들이 상승세를 보이고 있다"고 설명했다. 최근 밀 주요 생산자인 미국, 캐나다 등에서 기록적인 폭염이 이어지면서 밀 수확량이 감소할 것으로 예상되고 있다. 이에 따라 밀 가격 역시 급격히 치솟고 있다. 밀 ETF 중 가장 큰 수익을 올린 미국산 강맥은 최근 가격이 전년 동기 대비 40% 상승했다.

TIGER 농산물 선물 ETF 데이터 수집

🍏 변수 수집

🍏 파생변수 생성

🍏 변수 검증


🍏 변수 선택

🍏 모델링

🍏 결론


플레이션에
되면서
세를 보임

■ 농산물 소매가격 변수

 변수 수집


 파생변수 생성

 변수 검증


 변수 선택

 모델링


 결론



date	mean
2016-01-04	2,460
2016-01-05	2,460
2016-01-06	2,447
...	...



date	mean
2016-01-04	4,285
2016-01-05	4,344
2016-01-06	4,365
...	...



date	mean
2016-01-04	3,451
2016-01-05	3,417
2016-01-06	3,406
...	...

대형마트나 전통시장 등 전국 주요 시장에서 조사된 **소매 평균 가격** 데이터 수집

02 파생변수 생성


장마·폭염 지속시간 변수


가설 장마와 폭염이 일시적인 경우와 장기적으로 이어지는 경우의 영향력은 차이가 있을 것이다.


$$\frac{X_t + X_{t-1} + \cdots + X_{t-N}}{N + 1}$$




이동 평균에 착안하여 지속 영향 변수를 생성
t시점에서 과거 N시점까지 강수량의 평균을 냄
장마나 폭염이 오랜 기간 지속되었다면
이를 반영할 수 있음

 변수 수집

 파생변수 생성

 변수 검증

 변수 선택

 모델링

 결론

02 파생변수 생성

21-2 선형대수학팀 | 6

■ 강수량 파생 변수

date	당근 가중강수량	...
2016-01-04	0	...
2016-01-05	2.54399	...
2016-01-06	0	...
...

$$\sum \frac{\text{재배면적(\%)} \times \text{강수량(지역i)}}{100}$$

재배면적을 가중치로 하는 강수량 변수

🍏 변수 수집

🍏 파생변수 생성

🍏 변수 검증

🍏 변수 선택

🍏 모델링

🍏 결론

■ VARselect

그레인저 인과검정의
order에는 일반적으로
VARselect 함수를 통해
order값을 받은 뒤 입력함

VARselect에서는
VAR 모형 예측에 대한
최적의 order를 산출

VARselect에서
산출된 order는
우리가 필요한 7시점 이후
예측에 대한 최적의
order와 **다름!**

- 🍏 변수 수집
- 🍏 파생변수 생성
- 🍏 변수 검증
- 🍏 변수 선택
- 🍏 모델링
- 🍏 결론

- 그레인저 인과검정(Granger Causality Test)

- 🍏 변수 수집
- 🍏 파생변수 생성
- 🍏 변수 검증
- 🍏 변수 선택
- 🍏 모델링
- 🍏 결론

1차 차분을 통해
변수들을 **정상화**한 뒤
그레인저 인과검정 진행



그레인저 인과검정은
시계열 Y변수 예측에
도움이 되는 X변수를
찾아주는 검정 방법



그러나!

7시점 후 그레인저
인과검정을 시행할 때
6, 5, 4, ...시점 후의
변수가 포함되므로
**기존 함수를 그대로
사용할 수 없음**

■ 그레인저 인과검정(Granger Causality Test)

```
lag_function <- function(x, k, y, diff_data) {
  for_list <- x:k
  My_list_X = list()
  col_names_X <- c()

  list_index_X = 1

  for (i in for_list) {

    My_list_X[list_index_X] <- diff_data[, y] %>% shift(i) %>% list
    col_names_X <- c(col_names_X, paste(y, "lag", as.character(i)))

    list_index_X = list_index_X + 1
  }
  matrix_lag_X <- as.data.frame(My_list_X) %>% as.matrix
  lagX <- matrix_lag_X[complete.cases(matrix_lag_X), ]
  if (is.vector(lagX) == TRUE) {
    lagX <- lagX %>% as.matrix
  }

  colnames(lagX) <- col_names_X

  col_names_Y <- c()
  My_list_Y = list()
  list_index_Y = 1
```

```
for (i in for_list) {

  My_list_Y[list_index_Y] <- list(diff_data[, 1] %>% shift(i))
  col_names_Y <- c(col_names_Y, paste("Y : lag", as.character(i)))

  list_index_Y = list_index_Y + 1
}
matrix_lag_Y <- as.data.frame(My_list_Y) %>% as.matrix
lagY <- matrix_lag_Y[complete.cases(matrix_lag_Y), ]
if (is.vector(lagY) == TRUE) {
  lagY <- lagY %>% as.matrix
}

colnames(lagY) <- col_names_Y
y <- diff_data[-(1:k), 1]

lag_data <- cbind(y, lagX, lagY) %>% as.data.frame

fm_full <- lm(y~lagY+lagX)
fm_reduce <- lm(y~lagY)

## compare models with waldtest

rval <- waldtest(fm_full, fm_reduce)
return(rval)
}
```

원하는 시점 후의 예측에 대한 인과를 확인하기 위해
직접 그레인저 인과검정 함수를 목적에 맞게 직접 코딩

🍏 변수 수집

🍏 파생변수 생성

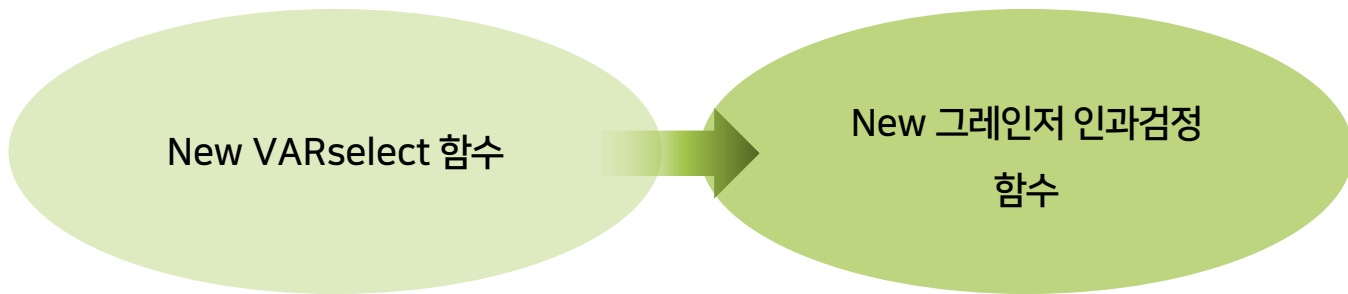
🍏 변수 검증

🍏 변수 선택

🍏 모델링

🍏 결론

- 그레인저 인과검정(Granger Causality Test)

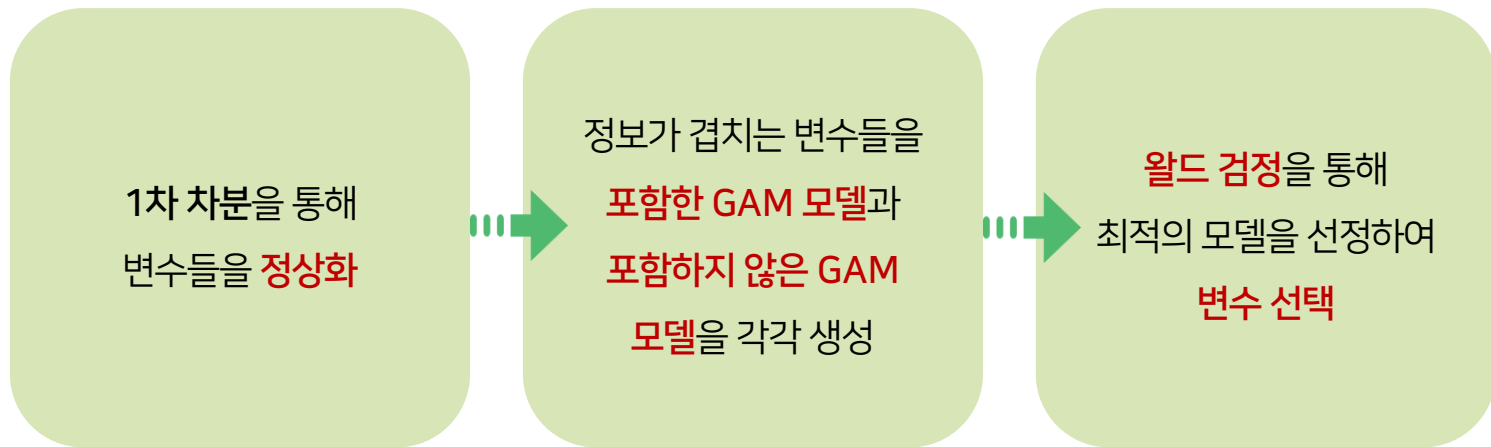


새 VARselect 함수에서 **optimal order**를 받아서
새롭게 만든 그레인저 인과검정 시행





- 🍏 변수 수집
- 🍏 파생변수 생성
- 🍏 변수 검증
- 🍏 변수 선택
- 🍏 모델링
- 🍏 결론

- 월드 검정(Wald Test)

- 🍏 변수 수집
- 🍏 파생변수 생성
- 🍏 변수 검증
- 🍏 변수 선택
- 🍏 모델링
- 🍏 결론



■ 모델링 목표

-  변수 수집
-  파생변수 생성
-  변수 검증
-  변수 선택
-  모델링
-  결론



- Forecasting in R

ARIMA

`auto.arima function`으로

모형 적합

VAR

`VAR function`으로

모형 적합



`predict(model, n.ahead = ★)`



R에서 제공하는 함수로
손쉽게 예측 가능

🍏 변수 수집

🍏 파생변수 생성

🍏 변수 검증

🍏 변수 선택

🍏 모델링

🍏 결론

- LGBM(Light Gradient Boosting Machine)

트리 기반의 알고리즘으로 그레디언트 부스팅 기법 중 하나
속도가 매우 빠르다는 장점이 있음



Hyper-parameter

n_iterators : 반복 수행 트리개수

max_depth : 트리의 최대 깊이



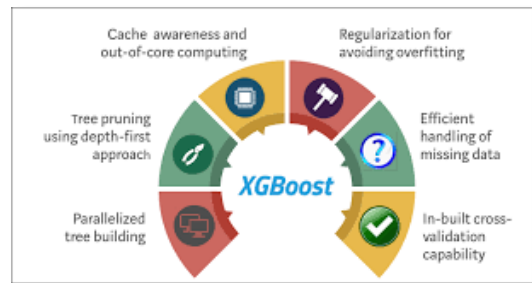
- 🍏 변수 수집
- 🍏 파생변수 생성
- 🍏 변수 검증
- 🍏 변수 선택
- 🍏 모델링
- 🍏 결론

▪ XGBoost

트리 기반의 알고리즘으로 그레디언트 부스팅 기법 중 하나
LGBM과 비교해 학습속도가 느리지만 성능이 좋은 편



Hyper-parameter
Gamma : loss의 감소 정도
max_depth : 트리의 최대 깊이



- 🍏 변수 수집
- 🍏 파생변수 생성
- 🍏 변수 검증
- 🍏 변수 선택
- 🍏 모델링
- 🍏 결론

- 랜덤 포레스트(Random Forest)

여러 결정 트리들로 구성되어 의사결정을 하는 앙상블 머신러닝 기법
무작위적으로 최적의 변수를 찾아 과적합을 방지함



Hyper-parameter

Ntree : 랜덤 포레스트 안의 결정 트리 개수

Mtry : 무작위로 선택할 변수 개수



- 🍏 변수 수집
- 🍏 파생변수 생성
- 🍏 변수 검증
- 🍏 변수 선택
- 🍏 모델링
- 🍏 결론

- Prophet

PROPHET

추세, 계절성, 휴일 효과

세 요소로 구성된 가산회귀모델

$$y(t) = g(t) + s(t) + h(t) + \epsilon t$$

Hyper-parameter

Trend : $g(t)$, 비주기적인 트렌드

Seasonality : $s(t)$, 주기적인 패턴

Holiday : $h(t)$, 휴일과 같이 불규칙한 이벤트

🍏 변수 수집

🍏 파생변수 생성

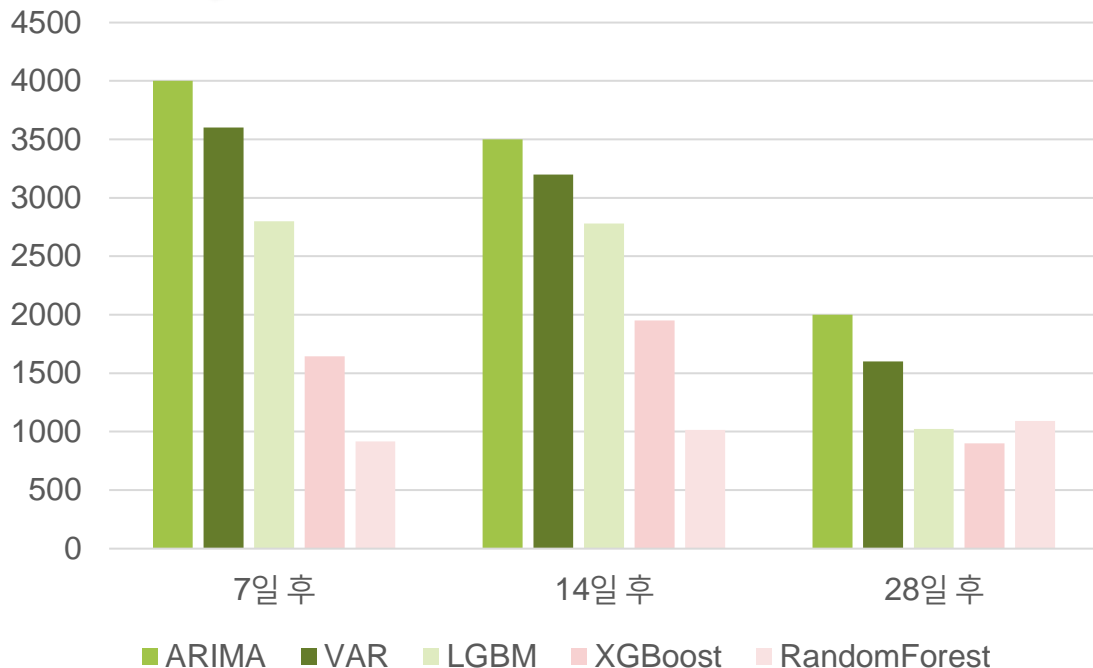
🍏 변수 검증

🍏 변수 선택

🍏 모델링

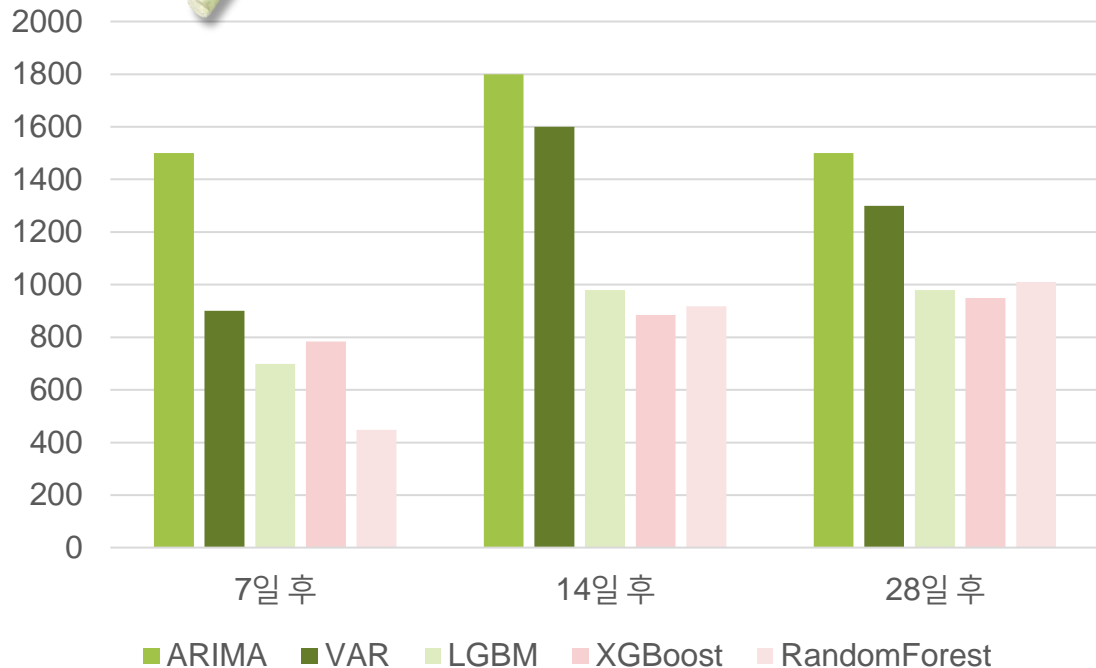
🍏 결론







■ 최종결론 - 당근



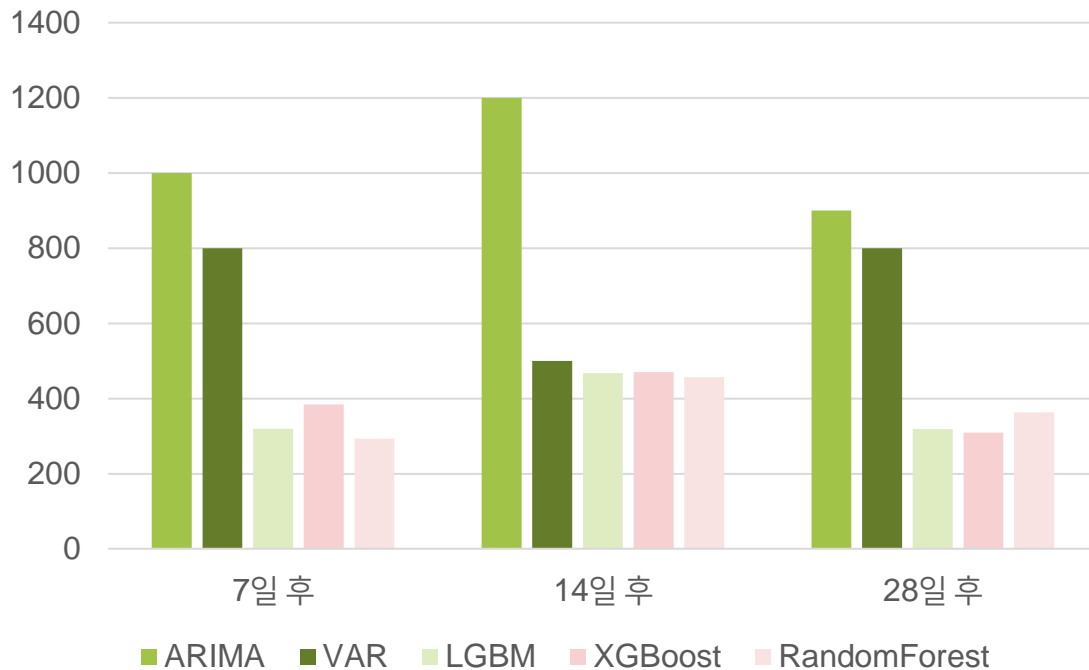
- 🍏 변수 수집
- 🍏 파생변수 생성
- 🍏 변수 검증
- 🍏 변수 선택
- 🍏 모델링
- 🍏 결론

최종결론 - 대파



-  변수 수집
-  파생변수 생성
-  변수 검증
-  변수 선택
-  모델링
-  결론

■ 최종결론 - 토마토



- 변수 수집
- 파생변수 생성
- 변수 검증
- 변수 선택
- 모델링
- 결론