DNA

농산물 가격예측문제에 대한 RF모델 시계열 분석 시도

Price Predict Of Agriculture

박예진, 서형준, 양정윤, 조성민

데이터 설명. 2022 농넷 농산물 가격예측 대회 데이터

농산물 품목에 대한 도매, 소매, 수입 ,수출, 날씨 데이터.

4개년에 대한 시계열 데이터.

"해당일자_전체평균가격" 을 예측하는 시장 데이터.

데이터 전처리. 통상적인 Table 데이터와 시계열 데이터의 차이

통상적인 데이터 전처리 방법론

Feature Engineering
-Feature Selection
-Feature Add

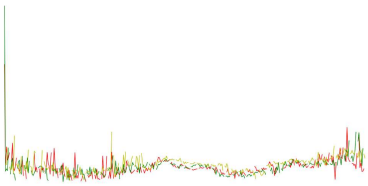
Scaling
-standard scaler
-Min-Max Scaler

Fillnan
- spline interpolation
- fillna(method='ffill')

시계열 데이터의 패턴

시계열 데이터의 4가지의 구성요소

추세, 순환, 계절성, 불규칙



다중공선성문제(Multicollinearity)
- 회귀분석에서 사용된 일부의 입력변수가 다른 입력변수와 상관정도가 높아 데이터 분석에 악영향을 끼치는 현상.

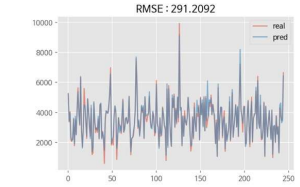
>>Feature Selection & Add
- **Wrapper Methods:** 가장 야만적인(?) method, 직접 넣어보면서 validation을 통해 최적의 집합을 찾기.

조건감소
- 분석결과의 안정성을 확보하기 위해 입력변수의 조건수를 감소시켜야 함

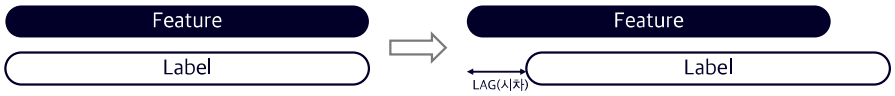
>>Min-Max scaler
- 최소~최대 값이 0~1 또는 -1~1 사이로 변환(정규 분포 가정을 안함)

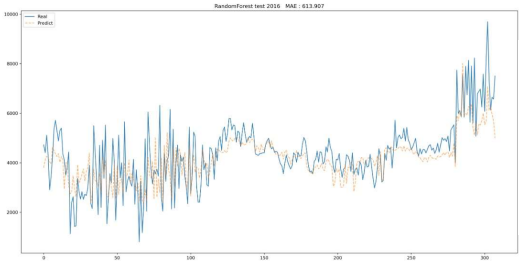
Validation. 4개년의 데이터 중 1개년을 validation set으로 사용 >> work-forward validation

전처리 방법론 적용 후 RMSE 변화
480.9154 >> 292.7518 >> 291.2092



Lag Data 생성
- 회귀 모델에 들어가는 예측변수의 영향이 단순하거나 즉각적이지 않을 수 있다. 이러한 상황에서 입력변수에 lag(시차)를 보정하여 회귀모델에서 시계열 데이터를 분석 할 수 있다.





$$MAE = \frac{1}{n}\sum \left| y - \hat{y} \right|$$

Divide by the total number of data points
Predicted output value
Actual output value
Sum of
The absolute value of the residual

결론

시계열 예측 모델은 시계열에 영향을 주는 변수들을 다양하게 고려하는 방법으로 발전해왔다. 그 기법들이 고도화 되어 공모전에서 는 ARIMA, LSTM 기반의 모델들이 널리 쓰이고 있다.

우리팀은 기존의 머신러닝 파이프라인에 데이터만 바꾸어 적용하는 가설을 세웠으나, 그 결과는 처참했다.

대회 일정 이후, 모델과 이론을 보강하예측 모델이 추세를 따라가는 결과를 만들었지만, RF모델의 한계로 시장 데이터의 잔차까지 예측하는데 좋은 결과를 만들지 못하였다.

결국 이러한 이유로 시계열 데이터는 일반적인 table데이터와 다른 관점에서 문제를 해석하고 데이터에 맞는 모델링 해야 한다