[가설 설정]

1. 기온이 low -> 난방 사용 증가-> 도시가스 공급량 증가 (채택)

2. 국제 천연가스 가격 상승 ⇒ 가스 공급에 차질 발생 가능 ⇒ 가스 공급량 감소 (채택X)

* 분석 내용
* 1분기와 4분기의 공급량 합계가 ‘상대적으로’ 2분기와 3분기의 공급량 합계보다 크다
* 공급사별 공급량의 차이가 존재하기 때문에 공급량 예측의 중요 변수로 작용한다는 사실 파악 가능
* IF 기온 그래프 a자 => 가설 신빙성

[대한민국 평균 기온 분석]

평균 기온 데이터 출처

* U자 모양이었던 분기별 가스 공급량 합계와 반대의 모양인 A자의 모습을 확인 가능

[가설 채택하지 않은 이유]

2-1) 천연 가스 공급은 중장기계약으로 가져오는 것

∴ 매일 달라지는 국제천연가스 가격과는 큰 관련 x

설정한 가설을 토대로 머신러닝을 진행하기 앞서

[기온 데이터 수집하기]

* 나주시 농업기상정보시스템

[데이터 수집 및 전처리 과정 중 겪은 문제점]

1. 웹크롤링

: 날짜, 기온, 시간 데이터가 없어서/ 리스트로 만들어 없는 날짜 채우고, 기온이 빈 값은 NAN값으로 채웠음

1. 웹크롤링

: XPATH로 수집 X , BEAUTIFULSOUP으로 담아서 수집했고 인덱스 규칙에 따라 함수를 작성하고 조건문으로 적용

1. 데이터 전처리

: 특정 날짜의 기온이 없는 경우 같은 날짜 년도별 기온의 평균값으로 처리,

매년 12/31 자정 기온 데이터 x => 같은 해 바로 전 시간 기온 데이터로 처리

[머신 러닝 모델 비교]

기온 데이터 없이 모델 만들어서 비교

1. 선형회귀, 라쏘, 릿지 => 정확도가 3%대로 낮았음
2. Knn은 77.3%의 정확도
3. 그 외 모델들은 97~99% 등 비교적 높은 정확도를 보였음

[기온 데이터를 추가한 머신 러닝 모델 비교]

2018년까지의 데이터를 모델 학습 -> 2019년 데이터 추정치로 입력 -> 입력한 데이터로 모델 작성해 정확도 판단

1. 선형회귀, 라쏘, 릿지의 정확도 향상 (3.8%->26.8%)
2. KNN은 (77.3% -> 61.3%)로 정확도 감소
3. 그 외 모델들은 큰 변화가 없었음

데이콘에 입력해본 게 아니고 내부적으로 검증해본 것임을 참고

[하이퍼파라미터 튜닝 후]

랜덤포레스트 기준으로 하이퍼파라미터 튜닝

(선택 이유: 앙상블 모델 중 가장 기본적인 모델부터 해보자는 의도)

1. 튜닝을 안 했을 때의 정확도(99%, 98.4%)가 했을 때의 정확도(92.3%) 보다 더 높음
2. 내부적으로 검증했을 때에는 xgboost의 정확도가 98.6%, randomforest로 하이퍼파라미터 튜닝한 결과 정확도가 92.3%로 xgboost의 정확도가 높은 것을 확인했는데,

정확도가 낮음에도 점수가 높은 이유는 기존에 시도하던 머신러닝 모델들이 과적합되어있다는 결론

[결정계수 히트맵]

**과적합 해결 위해 특성 선택 변경**

가장 좋은 수치 보여준 4개 특성 – month 시간 구분 기온을 선택하여 랜덤포레스트리그레서 모델을 돌려본 결과 0.118점으로 가장 높은 제출 점수를 받은 것을 확인할 수 있었습니다.

자체결정계수는 떨어졌지만 제출 점수가 상승하며

적절한 특성 선택이 과대적합을 해소한 것으로 판단할 수 있습니다.

기말 발표