#POC에 대한 자세한 설명

**1. 누락 국가의 품목에 대한 해당 2자리 류 품목들의 관세평균으로 대치하는 것에서 왜 HSCD로 안하고 해당 2자리 류로만 하였는지?**

먼저 결측치 처리에 있어서 우선순위를 HSCD 6자리 전부 쓰는 것과 앞의 4자리, 2자리를 사용하는 순으로 두었습니다. 하지만 6자리와 4자리의 경우 동일한 데이터의 개수가 제한적이어서 결측치 대치함에 있어서 대치값이 뚜렷하지 않다고 판단하였습니다. 그래서 6자리와 4자리를 쓰는 것보다 HS코드 앞 2자리만 사용하여 데이터의 개수를 늘린 후 평균대치를 하는 것이 해당 결측치인 데이터의 특성을 더 효과적으로 반영될 거라 생각하여 앞의 2자리만 사용하였습니다.

**2. 예측모형에서 영향력관측치, 이상치 제거가 제거한 데이터에서만 잘 맞는건 아닌가?**

먼저, 일반화 성능을 높이기 위해서 이상치를 제거해야 한다고 판단하였고, 제거한 뒤의 데이터가아닌 전체 데이터의 MAE를 기준으로 이상치를 얼마나 제거할지 판단하였기 때문에 이상치를 제거하더라도 최종 모형의 예측력은 더 향상되었습니다.

+) 최종 모형의 예측력 향상을 위해서 제거된 데이터들을 잘 학습하지 못한 부분은 감수했습니다. (추가 질문이 들어오면 얘기)

**3. 예측모형에서 규제나 교차검증과 같은 것을 사용하지 않고 시드를 사용한 이유?**

주어진 데이터가 국가와 품목에 의해 특정되는 데이터이므로 훈련세트와 테스트 세트를 나누는 머신러닝 기법을 통해 학습하면 훈련세트에 없는 데이터에 대해 설명력이 확연히 떨어질 것이라고 판단하여 데이터세트 분할 과정을 10개의 시드에서 진행했습니다.

**4. 파생변수에서 GDP\_RATIO를 할 때 전년도 GDP 제거하고 비율을 넣는다 해서 자기 상관이 없어지는게 맞는지? 원변수가 수식에 포함 되어있는데도??**

히트맵을 확인해보면 상관관계가 낮아진 모습을 확인할 수 있습니다.

**5. 여러개의 시드 중에 10개를 선정한 이유?**

많은 시드를 고려하면 모형성능에는 좋지만 효율면에서 분석시간이 오래걸린다는 단점때문에 그러한 영향도 고려하여 10개로 선정했습니다.

**6. 가중평균을 구체적으로 어떻게 낼 것인가?**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**7. 변수중요도에서 왜 19개만 했는지?**

20개 부터는 뚜렷한 효과를 보이지 않는다고 판단하였고 19개만 모형에 고려하였습니다.

**8. 모형마다 왜 최종변수가 다른건지? -> 트레인 세트가 달라져서 그에 맞는 최적변수가 달라짐**

그러면 왜 처음 변수중요도 19개를 시드별로 출력하지 않았는지? -> 시드별로 변수 중요도는 상위 19개 내에서 순서는 약간씩 달라지는 경우가 발생했지만 19개의 종류자체는 변함이 없었다.

**9. 해외 시장 정보에 대한 효과적인 활용을 기대한다 하였는데 구체적으로 어떤건지?**

앞서 저희가 만든 대시보드를 활용하여 해외 시장이 앞으로 어떤 전망인지에 대한 정보를 탐색할 수 있습니다. 그래서 이러한 정보를 바탕으로 기존보다 효과적으로 활용할 수 있다고 생각합니다.

즉, 분석배경에서 언급했던 해외 시장 정보에 대한 부족을 해결해줄 수 있다고 생각합니다.

+예를 들어 한국농수산식품유통공사는 중소수출업체에게 수출국 현지의 시장현황과 유통환경, 수출경쟁력 분석, 소비자 선호도 등 업체특성에 맞춘 심층정보를 받아볼 수 있다고 하였는데 유망시장에 대한 탐색이 이루어지면 이에 대한 신청을 효과적으로 활용할 수 있다고 생각합니다.

**10. 효율적인 재고관리 및 운임계약을 기대할 수 있다하였는데 구체적으로 어떻게 한다는건지?**

특정 기업의 수출품목에 대한 전망을 대시보드를 통해 파악하게 되면 전년대비 어느정도로 감소하거나 증가함을 알 수 있는데 이로 인해 기존 재고관리, 운임계약에 대한 방식을 어떻게 변화시켜야 하는지 계획할 수 있다고 생각합니다.

**11. 유망시장을 제안한다고 했는데, 그래서 유망시장이 구체적으로 무엇인가요?**

단순하게 2019년도 수입액 예측 결과를 바탕으로 판단했을 때 품목은 85류, 87류가 수입액이 많을 것이라고 예상되었고 주요 수출 국가는 중국 미국 홍콩 베트남 멕시코로 판단하였습니다.

엄밀하게 유망시장을 판단하려면 품목별 수출입 금액, 국가별 수출입 금액, 점유율 등 다양한 요인을 고려해야겠지만, 본 대회의 목적이 수입액을 예측하는 것이었으므로 유망시장에 대한 엄밀한 분석은 진행하지 못했습니다.

**12.스태킹이나 앙상블이 사용한 이유가 무엇인가?**

여러 개의 같은 종류의 모델이 예측한 결과를 평균을 내서 사용하는 배깅방식을 사용.

같은 종류의 모델을 결합해서 사용하지만, 데이터 샘플링을 다르게 적용함.

XGBRegressor는 부스팅이고, 오차를 학습하며 강학 학습기를 만드는 Gradient Boosting의 일종.

앙상블 기법의 장점으로 과적합을 개선

* 각 모델별로 발생하는 편향을 종합하여 예측하기 때문에, 과적합을 완화시켜줌.

예측정확도 향상

* 다양한 모델의 결과를 종합하여 예측 결과를 만들어내어 정확도가 향상.

우리 모델은 정확히는 앙상블 중 그레이디언트 부스팅을 이용한 XGBregressor를 사용한 개별 학습기를 10개의 시드를 고려한 배깅을 통해 부스팅을 이용한 편향 감소 목적과 배깅을 이용한 분산 감소 목적을 동시에 달성함. 또한, 변수선택과정과 이상치제거를 통한 과적합을 배깅을 통해 과적합을 완화시켜주는 효과를 사용.

**13.모델 구조를 그렇게 구성한 이유가 무엇인가?**

Ppt 설명 및 다시 설명드리기. 계량 경제 + 데이터 특성 + 머신러닝 방법으로.

**14. 왜 뒤로부터 변수제거 기법을 사용하였나요? Stepwise방식도 있는데**

먼저 저희는 617개의 변수를 선택함에 있어, 특성 중요도를 통해 설명력이 있는 19개의 변수를 선정했습니다. 따라서, 19개의 변수를 baseline으로 시작하여, 매 스텝마다, 각 변수의 존재 유무에 따른 MAE의 변화를 통해 변수 선택을 하여, backward 방식을 사용하게 되었습니다.

또한, 변수선택기법은 데이터와 모형에 따라 최적의 변수 선택 과정이 다릅니다. 분석가는 여러가지 기법 중 최선의 방법을 이용함으로 모델링을 진행하는 것이고, 어떤 방식이 최적의 방법이라고 판단할 수 없을 것 같습니다.

# 앙상블 설명

일반적으로 앙상블의 결과는 원본 데이터셋으로 하나의 예측기를 훈련시킬때와 비교해 편향은 비슷하지만, 분산은 줄어든다.

앙상블은 모든 예측기의 예측을 모아서 새로운 샘플에 대한 예측을 만드는데 회귀 문제 같은 경우 평균을 계산함.(분류는 최빈값)

개별 예측기는 원본 훈련 세트로 훈련시킨 것보다 훨씬 크게 편향되어 있지만, 수집 함수를 통과하면, 편향과 분산이 모두 감소함.

**15. 이상치 데이터의 경우 그것이 유효한 데이터일 수도 있지 않느냐?**

-> 유효한 데이터일 수도 있지만 저희가 이상치나 영향력관측치로 규명하고 제거하는 과정에 있어서 제거한 배경에는 최종 예측 모형의 전반적인 성능을 향상시키는 것에 중점을 더 두는 것이었습니다. 그래서 비록 이상치인 데이터는 잘 학습되지 못한 부분이 있더라도 그 부분은 감수하여 모형의 전체적인 예측성능을 높였습니다.

Cf) 시계열 모형을 사용하지않았는데 이러한 효과는 고려하지 않았는지?

시계열 모형에 대한 충분한 학습을 시키기에는 주기적인 데이터의 폭이 좁다 판단하였습니다. 하지만 이러한 시계열성을 어느정도 고려해주기위해 파생변수와 PoC의 타입스템간의 특성을 이용해주었습니다.

**13. XGBRegressor를 사용한 이유?**

test-score, train-score, MAE를 기준으로 판단했을 때, train-score와 MAE에서 가장 좋은 성능을 나타낸 모델을 사용하였습니다.

**14. 노이즈를 제거할 때 분석자의 판단하에 반복문을 설정하였다고 보고서에 나와있는데, 어떤 기준인가요?**

영향력관측치를 잔차가 5이상인 것으로 규명하였고, 이러한 영향력 관측치들은 모형학습에 있어서 일반화 성능을 저하시키는 요인이었습니다. 그래서 제거를 반복하면서 일반화 성능이 최적으로 향상되는 지점과 제거되는 데이터의 일정 개수를 제한하기위해 조건문 및 반복문을 활용하여 했습니다.???

**16. 2017 데이터 가져올 때 규칙은 어떻게?**

다운로드 받는 과정에서 Reporter가 korea이고 Trade flows가 export인 데이터와, Partner 가 korea 이고, Trade flows가 export인 데이터가 존재했습니다. 주최측에 문의해본 결과 import인 데이터가 없으면 export로 대체하였다는 답변을 받았습니다. 그러나 출처사이트에서 2018년도 데이터를 새로 다운받아 분석용 데이터셋과 비교해보니 둘 다 일치하지 않는 경우도 존재했습니다. 이는 UN comtrade의 자료가 계속해서 최신화 되기 때문인데, 이 때문에 2017년도 데이터를 가져올 때 논리적인 문제가 생겼습니다. 최대한 동일한 데이터를 가져오기 위해 2018년도 데이터를 기준 직접 비교해보며 몇 가지 규칙을 찾아냈습니다. 규칙 적용 결과 2018년 기준, 200개의 데이터를 제외하고는 오차가 0.01% 이내였습니다.

**17. 최종 선택된 변수를 확인해보니, seed마다 다른데 이는 모형 설명력이 부족하다는 말은 아닐까요?**

발표에서 말씀드렸던 것처럼 본 데이터의 특성상 다범주인 데이터로 인해 어떤 데이터들이 학습되느냐에 따라 변동이 심하다고 판단하였고 다중 시드를 고려한 배경 중 하나였습니다. 그래서 이렇게 어떤 데이터들이 트레인세트에 있느냐에 따라 시드 별 선택된 변수가 다른것이라고 생각합니다.