데이터기반학습 Project #1 Neural Network 보고서

5조 문현지, 정유민, 한태구

본 프로젝트의 모델링을 시작하기 전에 3가지 부분을 신경썼다.

* Baseline model: 임의 모델의 예측성능에 대한 기준이 필요하다 판단하여 전날과 동일한 값으로 예측(즉, y\_(t-1))하는 모델을 baseline 모델로 설정했다. 최근 환율변동이 비교적 안정화되어 baseline 모델 역시 높은 예측력을 보였다.
* Learning curve: 모델이 주어진 train dataset에만 과적합되는 현상을 막기 위해 validation dataset에 대한 loss로 학습곡선을 그리며 loss가 수렴하는 과정을 모니터링했다. 이를 위해 sklearn model에 내장된 partial\_fit 함수를 이용하였다.
* Validation method: 시계열은 시간적 순서를 가지므로 k-fold cross-validation for autoregression이나 walk forward방식과 같이 특수 validation 방식을 이용한다. 그러나 본 프로젝트에서는 window방식으로 데이터를 가공하였으므로 해당하는 순서적 성질이 비교적 적게 작용한다고 판단하여 일반적인 cross validation방식을 이용했다.

1. Data preprocessing

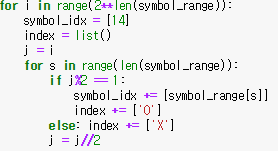
(1) Drop & Imputation

2018년 6월을 기준으로 USD\_Price의 데이터 형태가 변한다. 2018년 6월 이전에는 월요일에서 금요일까지 5일/주 형태의 데이터였다면, 그 이후에는 6일/주 형태의 데이터가 된다. 이에 input 데이터 의미를 통일하기 위해 2018년 6월 이전 비어있는 일요일에 해당하는 값을 금요일과 월요일의 선형가중평균으로 채웠다. Fillna를 이용해 앞날을 p, 뒷날을 1-p 비율로 채우며 이 방식을 적용하면 연속적인 빈 날짜 3일에 대해 imputation이 가능하다. 이때 p역시 hyperparameter로 설정하여 추후 최적화 대상에 포함시켰다. 토요일에 해당하는 데이터는 없으므로 초기에 date index에서 제외하였다.

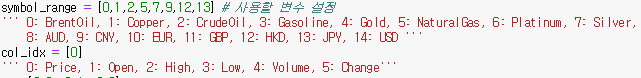
(2) Scaling

Scaling은 scaling을 하지 않는 경우, 최소.최대값을 이용하는 MinMaxScaler, StandardScaler 세 가지 방식에 대해 각각 실험을 진행한 후 가장 예측성능이 좋은 MinMaxScaler방식을 선택했다.

2/3. Feature/Model Parameter Selection (0406\_LFD\_pj1\_cross\_validation.ipynb 파일 참조 )

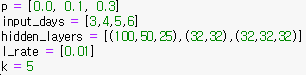
(1) Feature Selection 방법

Feature Selection의 경우 Heuristic하게 각 symbol을 설정하는 것은 인위적이라고 생각해 모든 가능한 경우의 수에 대해 왼쪽과 같은 코드로 grid search를 실시하였다.



사전 grid search 결과 위와 같은 symbol\_range가 14(USD)와 같이 사용했을 때 성능이 좋고, Column의 경우 0(=Price)만이 유의미하다고 판단되어 위와 같은 parameter 범위를 사용하였다.

(2) Model Parameter Selection 방법

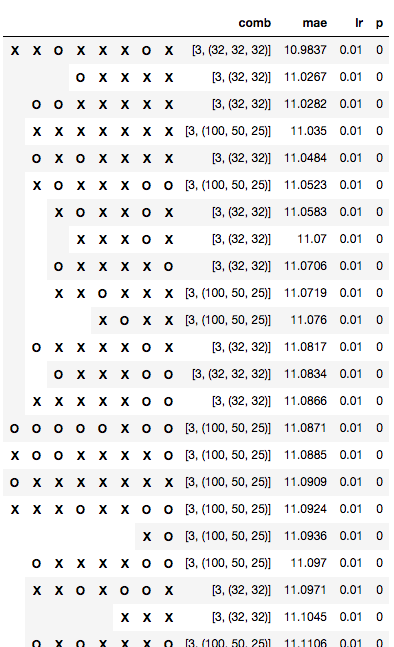


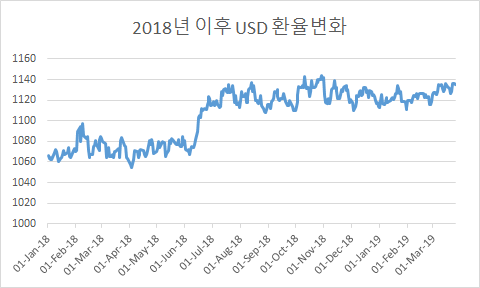
Parameter Selection의 경우에도 마찬가지로 grid search를 실시하였다. learning rate와 cross validation할 fold의 개수 k의 경우 사전에 테스트해 본 결과 최종성능에 영향을 미치지 않는다고 판단하여 고정하였고, p(), input\_days(예측시 input으로 사용할 과거 데이터의 기간), hidden layers는 위와 같은 범위에서 grid search를 실시하였다. 위 범위는 사전 grid search 테스트에 의해 나온 것으로, input\_days가 7이상인 경우와 p가 0.5이상인 경우에는 성능이 잘 나오지 않아 이와 같이 결정하였다.

https://lh6.googleusercontent.com/AtihRA5uVvpPC5SefRW53NvdCVWs5sEp1rFsWDUaVUrPSJJXE5hIGkDArZxZ3Ve6kMNfAxGKNuIwHZF3y6K7QRdM4fx6QLHbP6to9DuxJrBlqgeYC1MeWvd70QQNIN_b10qsZ91BhLM

Feature와 Parameter selection 모두 grid search시 기준이 되는 scoring은 위와 같이 cross validation 한 뒤 각 mae의 평균값으로 하였다. 또한 안정성을 위해 7개의 모델을 ensemble하여 사용하였다.

(3) Selection 결과

오른쪽은 각 symbol의 조합에 대해 가장 좋은 mae를 나타내는 parameter조합을 mae에 대해 오름차 순으로 정렬한 것이고, 이 때 comb = [input\_size, (hidden\_layers)]이다. 보다시피 input\_days = 3, p = 0 인 조합이 상위권에 압도적으로 많고, hidden\_layers의 경우 차이가 그닥 없기 때문에 최종적으로 parameter는 input\_days = 3, p = 0, hidden\_layers = (32,32)로 결정하였다.



Symbol(feature)의 경우 선택가능한 feature들이 여러 종류 있었는데, 위와 같이 환율 추세가 최근 크게 변동했다는 점에 근거해 최근 데이터에 대해 다시 test하여 가장 성능이 좋은 feature 2(=CrudeOil)을 선택하기로 하였다.

4. Result

Feature를 CrudeOil, USD의 price로 선택하여 MinMaxScaling을 이용해 Data Normalization을 하였고, hyper parameter들을input days: 3일, hidden\_layers: (32,32), p: 0.0, learning\_rate: 0.01로 설정하여 최근 10일간의 데이터에 대해 Test를 해본 결과 MAE는 약 3.141691이 나왔다.