LFD Project#1: Exchange rate prediction

Team 11 : 김 석, 이 은, 한창진

1. Objective and Dataset

본 프로젝트의 목표는 환율, 원자재의 데이터로 미래의 미국달러 환율을 예측하는 Neural network 모델을 만드는 것이다. 주어진 데이터셋은 2010.01.01~2020.04.06의 환율 및 원자재에 관한 daily data이다. Train set과 Test set 분리는 주어진 뼈대코드대로 진행하였다. 제출 모델을 위한 성능 평가로는 TimeSeries를 위한 cross validation 방법인 TimeSeriesSplit의 평균 MAE, Train 평균 MAE, Test MAE을 이용하였다.

2. Feature preprocessing

Feature Selection 이전에 전처리는 다음과 같이 진행을 했다. 우선 후보 feature에 모든 환율과 원자재의 데이터를 활용하였고, 토요일과 일요일은 비어있는 데이터의 수가 많아 Business day로 월~금의 주중 데이터만을 활용하였다. 원자재의 Volume data에는 – 로 입력되어 있는 feature들의 경우 null으로 치환하고, K와 M은 각 단위에 맞게 변환해주었다. 모든 값이 비어있는 column은 drop해주었다.

Null 처리의 경우 해당 row를 drop하는 방법, 직전 데이터로 채우는 방법, 해당 column를 drop하는 방법에 대해 test를 진행했다. 그러나 기본적으로 월-금의 주중 데이터를 유지하고 feature selection의 경우 이후의 feature selection에서 진행하는 것이 좋다고 판단하여 직전 데이터로 채우는 방법을 사용하였고 실제로도 이들 중 이 방법이 결과가 좋았다. 그렇기에 Null 처리는 기본적으로 직전 데이터로 채운 후 시작 데이터가 비어있을 경우에 직후 데이터로 입력해주었다.

3. Feature selection

전처리한 데이터를 바탕으로 Vector Autoregression (VAR) 방식과 ExtraTreeRegressor (ETR)의 두 가지 방식으로 Feature Selection을 시도하였고, 이 중 최종 결과가 가장 좋은 ETR의 방식으로 추출된 Feature들을 선택하였다. 이 때 VAR 구현을 위해 최종 모델에서는 사용하지 않았으나 'statsmodels' package를 이용하였다.

3-1. VAR

Ince and Trafalis (2006)은 환율 예측에서 MLP 모델의 feature selection에 VAR과 ARIMA 를 활용한 hybrid model를 제시하고, VAR+MLP 모델이 ARIMA+MLP 모델보다 더 우수함을 보였다. 이 선행 연구를 바탕으로 각 항목의 Price만을 추출하여 VAR을 실시했고, 최종적으로 probability가 0.05보다 낮은 USD, AUD, CNY, EUR, GOLD price가 feature 로 선택되었다.

	coefficie	std. erro	t-stat	prob
L1.USD_Price	1.372947	0.246749	5.564	0
L1.AUD_Price	-0.12309	0.025698	-4.79	0
L1.CNY_Price	1.113883	0.367049	3.035	0.002
L1.EUR_Price	-0.0596	0.020551	-2.9	0.004
L1.Gold_Price	0.029581	0.012252	2.414	0.016
L1.HKD_Price	-3.75104	1.95304	-1.921	0.055
L1.Silver_Pric	-0.73129	0.383901	-1.905	0.057
L1.Platinum_Pr	-0.01847	0.011108	-1.663	0.096
L1.JPY_Price	-2.30596	1.885655	-1.223	0.221
L1.Gasoline_Pr	-4.62929	4.03542	-1.147	0.251
L1.Copper_Pric	-3.35505	3.492327	-0.961	0.337
L1.BrentOil_Pr	0.171029	0.191849	0.891	0.373
L1.NaturalGas_	0.371092	1.263774	0.294	0.769
L1.CrudeOil_Pr	-0.03092	0.1819	-0.17	0.865
L1.GBP_Price	0.002555	0.016358	0.156	0.876

그림 1. VAR 결과

3-2. ETR

Random Forest 의 변형 형태인 ExtraTreesRegressor 를 활용하여 한 시점에서 다른 feature 들로 USD price 를 예측하는 regressor 을 생성했다. 이 때 feature importance 를 계산한 후에 importance 값이 높은 상위 5 개 feature 를 선정했다. 이 때, target value 인 USD_Price 는 제외하고 계산하였으나 우리는 이전 time step 의 USD_Price 는 다음 USD_Price 와 가장 연관이 클 것으로 가정하고 USD_Price 또한 사용하기로 결정했다. 그리고 추가적인 feature 를 만들어 성능을 높일 수 있는 방법을 고안했는데 USD_Price 의 제곱 및 제곱근 값을 새로운 feature 로 추가했을 때 모델의 성능이 약간 좋아지는 것을 확인하여 함께 feature set 을 구성하기로 했다.

따라서 최종적으로 선택된 feature set 은 {'USD_High', 'HKD_Low', 'HKD_Price', 'USD_Open', 'USD_Low', 'USD_Price', 'USD_Price_square', 'USD_Price_root'} 이다.

이들에 대해 Standard Scaler, MinMax Scaler, Non-scale 을 실험해본 결과 StandardScaler 를 채택하여 Train set 과 Test set 을 train set 의 mean 과 std 를 기준으로 z-normalization 해주었다. 또한 y 에 대해서도 안정적인 학습을 위해 StandardScaler 로 정규화를 해준 후 이후 MAE 산출에는 inverse_transform 을 이용하여 산출하였다.

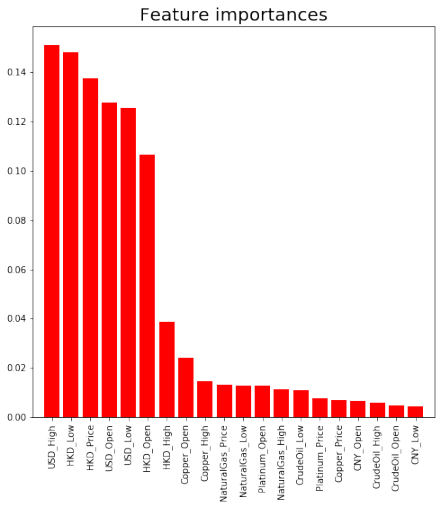

그림 2. ETR로 추출한 feature importances

4. Model

최종 모델을 위해 search해본 parameter들은 다음과 같다. Input으로는 시작 년도(2010년~2018년), input days(3,5,10,20), VAR 방식과 ETR 방식, ETR에 대해선 topk(1,3,5,6,10,20)에 대해서 test를 진행했다. Model의 parameter로는 activation function(relu,tanh), solver(adam,sgd), hidden_layer_sizes((64,32), (64,16), (32,8))에 대해 test를 진행했다.

최종 모델은 ETR의 topk=5 feature와 USD price(self, square, root), start_date='2012-01-02', input_days=5로 설정해서 feature을 구성하였다. 이에 대해서 학습시킨 MLPRegressor는 activation function으로는 tanh, solver는 sgd, 그리고 hidden layer는 각각 64, 32개의 node를 가진 두 층으로 이루어졌으며 그 외의 구체적인 parameter 들의 값은 아래와 같다.

{'activation': 'tanh', 'alpha': 0.0001, 'batch_size': 'auto', 'beta_1': 0.9, 'beta_2': 0.999, 'early_stopping': True, 'epsilon': 1e-08, 'hidden_layer_sizes': (64, 32), 'learning_rate': 'adaptive', 'learning_rate_init': 0.001, 'max_iter': 2000, 'momentum': 0.9, 'n_iter_no_change': 10, 'nesterovs_momentum': True, 'power_t': 0.5, 'random_state': 0, 'shuffle': False, 'solver': 'sgd', 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': False, 'warm_start': False}

5. Result

최종 모델의 TimeSeriesSplit MAE는 scaled y 기준으로 0.26599, Train MAE는 9.568, **Test MAE는 6.506** 이 나왔다. 학습 시 Loss는 그림 3와 같고, Predict한 값과 y를 plot한 결과는 그림 4과 같다. 이 때 값으로는 10개 output의 평균을 기준으로 사용하였다.

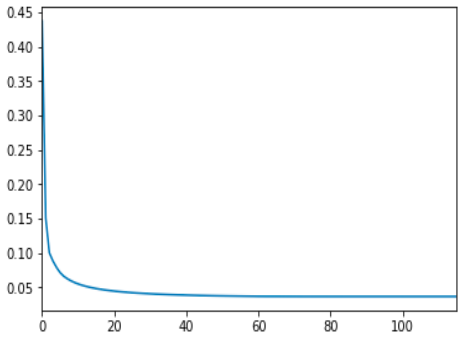

그림 3. Loss

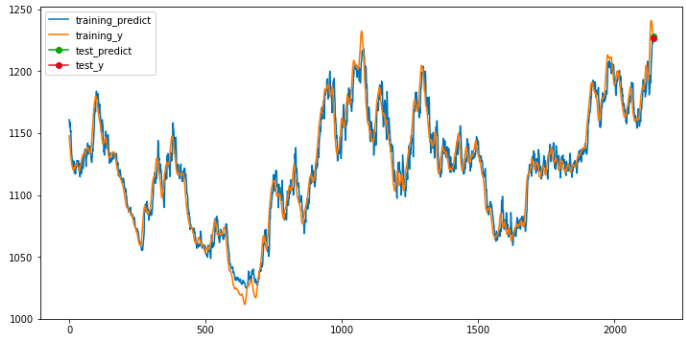

그림 4. predict와 y 비교

6. Reference

Ince, H., & Trafalis, T. B. (2006). A hybrid model for exchange rate prediction. Decision Support Systems, 42(2), 1054-1062.