LFD Project#2: Forecasting gold price with HMM

Team 10 : 김 석, 이 은, 한창진

1. Objective and Dataset

본 프로젝트의 목표는 환율, 원자재의 데이터로 미래의 금값을 예측하는 Hidden markov model (HMM)을 만드는 것이다. 주어진 데이터셋은 2010.01.01~2020.04.18의 환율 및 원자재에 관한 daily data이다. Train set과 Test set 분리는 주어진 뼈대코드대로 진행하였다. 제출 모델을 위한 성능 평가로는 k-cross validation 방법으로 평균 MAE와 test MAE를 이용하였으며 CV MAE에 좀 더 높은 가중치를 두었다.

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 Pearson Correlation 방식과 ExtraTreeRegressor(ETR) 방식, 그리고 이들에 각각 Technical Indicator(TI)를 함께 이용하는 방식으로 Feature Selection을 시도하였고, 이 중 최종 결과가 가장 좋은 ETR의 방식만으로 추출된 Feature들을 선택하였다.

3-1. Pearson Correlation

Pandas dataframe 의 corr()함수를 이용해 pearson correlation을 계산했고 gold price를 포함하여 correlation이 0.5가 넘지않는 feature set을 만들었다. 이렇게 선정된 feature는 총 14개였으며 그 feature set은 {'CNY_Price', 'Gold_Change', 'Gasoline_Volume', 'Gold_Volume', 'USD_Change', 'Brentoil_Change', 'USD_Price', 'Copper_Change', 'AUD_Change', 'BrentOil_Volume', 'NaturalGas_Volume', 'NaturalGas_Change', 'Gold_Price', 'Copper_Volume'} 이다. 아래 그림 1은 feature들간에 correlation을 나타낸 heatmap이다.

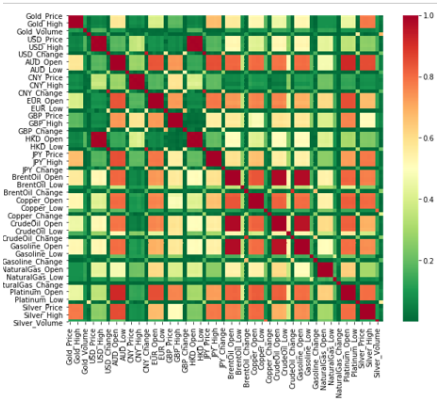

그림 1. Features Correlation heatmap

3-2. ETR

Random Forest 의 변형 형태인 ExtraTreesRegressor 를 활용하여 한 시점에서 다른 feature 들로 금값을 예측하는 regressor 을 생성했다. 이 때 feature importance 를 계산한 후에 importance 값이 높은 상위 10 개 feature 를 선정했다. 최종적으로 선택된 feature set 은 { 'Gold_High', 'Gold_Open', 'Gold_Price', 'Gold_Low', 'Silver_Low', 'Silver_Price', 'Silver_Open', 'AUD_Low', 'Silver_High', 'AUD_Price' } 이다. 그리고 이 10 개 안에서 1-10 개를 사용할 지 성능을 비교해본 결과 top 3 인 'Gold_High', 'Gold_Open', 'Gold_Price'를 이용하는 것이 가장 성능이 좋았다. 아래 그림 2 는 ExtraTreesRegressor 을 이용하여 추출한 top 10 의 feature importances 이다.

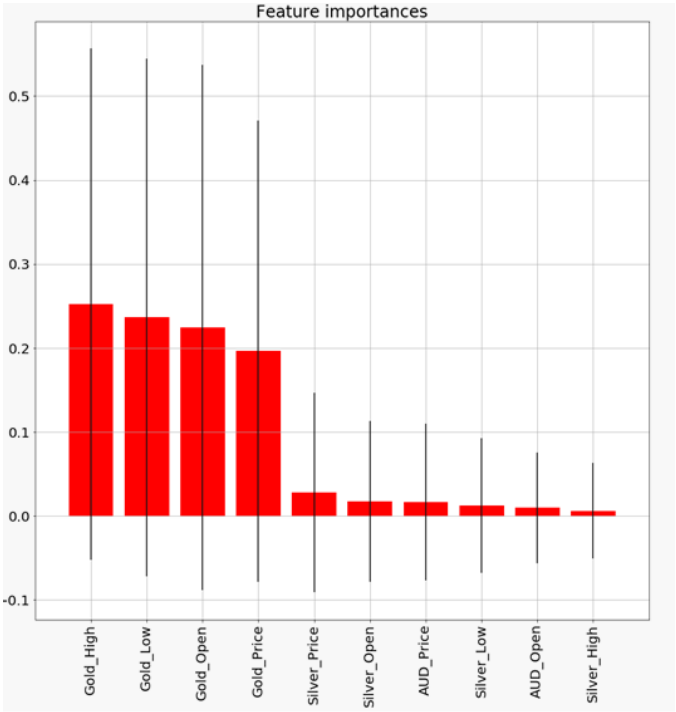

그림 2. ETR로 추출한 feature importances

3-3. TI

Gold price 를 예측하는 technical indicator 에는 Trendlines, Rate Of Change(ROC), Stochastic Oscillator 등이 있다.[1, 2] 본 프로젝트에서는 ROC 를 추가적인 feature 로 채택해 사용하였고 그 값을 구하는 방법은 아래와 같다.

$$ROC_n(x) = \frac{P(x) - P(x-n)}{P(x-n)}$$

n=3, 5 인 ROC 값을 각각 3-1, 3-2 의 feature 들과 결합해 사용해 봤으나 유의미한 결과를 얻지 못해 최종적으로 선택하지는 않았다.

4. Model - ETR

Gold price 를 예측하는 선행연구에서 Gold price 가 non-stationary 해서 이를 해결하기 위한 방법으로 price 의 first difference 를 제안했으며[3], 모든 input data 를 first difference 를 사용한 연구[4]도 있기 때문에 input data 모두 first difference 를 이용하였다. 최적의 조합을 찾기 위해 테스트한 parameter 는 다음과 같다. Input days(1~10 일), ETR - topk(1~10), scaler(Standard, MinMax, None)를 테스트했으며, GaussianHMM model 의 parameter 로는 n_components(1~27), covariance_type(spherical, diag, full, tied)에 대해 test 를 진행했다.

ETR 최종 모델은 TOP3('Gold_High', 'Gold_Open', 'Gold_Price')들의 first diffrence, Input days=1 일, Standard scaler 로 feature 를 구성했으며, n_components=11, covariance_type='tied'로 model 의 parameter 를 구성했을 때 k-cross validation 의 MAE 결과가 가장 좋았으며 test MAE 역시 성능이 크게 나쁘지 않아 최종 모델로 채택하였다.

5. Result

최종적으로 제출하는 모델의 K=3, time_step= 100 을 이용한 Cross Validation 의 평균 MAE 는 15.646484, test MAE 는 23.329342 였다.

6. Reference

[1] Jan Ivar Larsen. (2010). Predicting stock prices using technical analysis and machine learning. NTNU

[2] Vatsal H. Shah. (2007). Machine learning techniques for stock prediction. NYU

[3] Shafiee, S., & Topal, E. (2010). An overview of global gold market and gold price forecasting. Resources policy, 35(3), 178-189.

[4] Parisi, A., Parisi, F., & Díaz, D. (2008). Forecasting gold price changes: Rolling and recursive neural network models. Journal of Multinational financial management, 18(5), 477-487.