산업경제연구 제29권 제5호 2016년 10월(pp.1607～1628)

데이터 마이닝 기법을 이용한 환율예측 : GARCH와 결합된 랜덤 포레스트 모형

서종덕
대전대학교 경제학과 조교수
(jjongduk@gmail.com)

실시간으로 발생하는 거대한 양의 빅 데이터를 활용할 수 있는 데이터 마이닝 기법은 기계학습모형들 중 한 가지로 아직 경제학 분야에서는 활발하게 논의되고 있지 않은 분야이다. 특히 자료의 변동성이 큰 금융시장에선 더욱 그러하다. 본 연구는 기계학습모형 가운데 예측성이 우수한 것으로 알려진 데이터 마이닝 기법의 한 종류인 랜덤 포레스트 모형과 금융시장 시계열자료의 변동성을 반영할 수 있는 GARCH 모형을 결합하여 우리나라의 환율에 대한 예측력을 향상시킬 수 있는가를 분석하고 있다. 아울러 어떤 경제변수들이 환율의 변화에 영향을 미치고 있는가를 선별하고 이들 변수들의 중요도에 대해서 분석하고 있다. 분석 결과 기계학습모형과 계량경제모형이 결합된 혼합모형이 기존의 계량경제모형보다 예측력이 향상되었으며, 외환시장에서는 미국달러지수가 시장에 가장 큰 영향을 미치고 있으며 그 다음으로 KRX100 지수와 KOSPI 지수가 영향을 미치는 것으로 나타났다.

핵심주제어 : 기계학습모형, 데이터 마이닝, 랜덤 포레스트, 환율예측

▷ 논문접수(2016. 8. 24), ▷ 심사완료(2016. 10. 6), ▷ 게재확정(2016. 10. 31)

Ⅰ. 서 론

계량경제모형을 이용한 금융시장 분석과 예측은 시장에 존재하는 많은 변수들의 변동성(volatility)문제로 인해 어려움을 겪는다. 또한 시간의 흐름에 따라 항상 변화하는 금융상품들의 이면에 여러 가지 보이지 않는 중요한 요인들이 존재하며, 이들 요인들을 계량모형에 명시적으로 포함시켜 분석하기가 용이하지 않다는 것이다. 이에 대해 Mees & Rogoff(1983)는 거시경제적 기초 요인을 고려한 임의보행모형(Random Walk model)을 지금의 것 이상으로 개선시킨다는 것은 상당히 어렵다고까지 말하고 있다.

한편 이러한 문제점들을 고려하면서 금융시장에서 발생하는 금융시계열 관련 빅 데이터를 분석하고 활용할 수 있는 새로운 계량경제학적 기법에 대한 연구의 필요성이 대두되고 있다. Varian(2014)은 첫째, 가공되지 않은 엄청난 양의 자료를 다룰 수 있는 강력한 도구가 경제학에서 필요하게 되었고 둘째, 경제변수들에 대한 예측을 수행함에 있어 기존의 방법이 아닌 새로운 예측인자들을 선별할 수 있어야 한다는 것 그리고 거대한 양의 자료를 다루기 위해선 단순한 선형형태에서 벗어나 함수의 형태에 제약받지 않는 신축성 있는 함수형태가 필요하다는 것을 강조하면서 향후 계량경제학 분야에서 기계학습모형(machine learning model)에 대한 필요성을 강조하고 있다.

기계학습모형은 앞에서 Varian(2014)이 말한 기존 계량모형이 갖는 제약성을 보완할 수 있다는 점에서는 장점을 갖고 있으나 역시 변수들의 분포가 일정한 형태를 띄어야 한다는 한계성을 갖고 있다. 그러나 금융시계열자료는 심한 변동성을 나타내는 것이 특징이므로 대용량의 금융시계열자료를 처리하기 위해서는 이러한 변동성을 고려한 기계학습모형의 필요성이 자연스럽게 대두된다. 그러나 기계학습모형은 의학이나 생리학 그리고 사회학 등 다른 분야에서는 연구가 활발하게 이루어지고 있으나 아직 경제학 분야에선 연구결과를 찾아보기 힘들다.

본 연구에서 사용되고 있는 모형은 환율예측을 위한 기존의 계량경제모형[1]과는 달리 데이터 마이닝 기법중의 하나인 랜덤 포레스트 모형(Random Forest model, 이하 RF 모형)에 금융시계열자료를 분석하는데 일반적으로 사용되고 있는 GARCH모형을 결합함으로서(이하 RF-GARCH 모형) 앞에서 Varian(2014)에 의해 제기된 계량경제모형의 한계점을 극복하면서 차별화를 시도하고 있다. 이 과정에서 자기회귀모형(이하 AR모형이라 함)을 기준모형으로 삼아 두 모형을 비교해봄으로써 새로운 모형이 기준모형의 환율예측능력을 향상시키는 효과가 있는지 분석하는데 초점을 맞추고 있다. 또한 환율을 예측함에 있어 환율의 변화에 영향을 미치는 중요한 요인들이 무엇인가를 선별하고 그 중요도에 대해서 살펴보고 있다.

1) 김창범, 모수원(2001), 김창범(2007) 참조

본 연구와 관련하여 기존 연구는 국외뿐만 아니라 국내에서도 아직 활발하지 못한 상태이다. 한태경(2010)은 시계열분석과 인공지능학습모형을 활용하여 환율예측을 하였다. ARIMA와 VAR모형을 활용하여 설명변수가 될 환율변수를 선정하고, 이후 예측하고자 하는 원-달러 환율변수와 선정된 입력변수들을 인공지능모형에 학습시켜 진단하고 예측하도록 설정되어있다. 그리고 인공지능모형인 인공신경망이나, SVM 방법을 활용한 예측방법이 기존의 통계분석 기법인 판별함수나 로지스틱회귀를 활용한 예측방법보다 나은 성능을 보여주었음을 확인하고 있다. 김경태 외(2015)는 다중회귀모형을 이용한 환율예측모형을 구성 후 회귀분석에서 사용하고 있는 변수선택법을 이용하여 환율의 결정에 어떤 변수들이 유의한 영향을 미치고 있는가를 분석하고 있다. 그러나 앞의 두 분석에서 모두 환율이 갖는 변동성을 고려하지 않고 있다는 한계점을 띄고 있다.

본 연구의 구성은 다음과 같다. 제 Ⅱ 절에서는 RF-GARCH모형을 구성하기 전에 본 연구에서 사용되고 있는 시계열모형에 대해서 간략하게 설명한 후 RF모형에 대해서 개괄하고 있다. 제 Ⅲ 절에서는 실증분석 및 그 결과를 설명하고 있으며, 마지막으로 제 Ⅳ 절에서는 결론 및 향후 연구방향을 정리하고 있다.

Ⅱ. 시계열모형과 랜덤 포레스트(Random Forest)모형

2.1. 자기회귀(AR)모형

일반적으로 시계열자료를 갖는 계량모형을 추정할 때 기준이 되는 모형으로 AR모형이 자주 이용된다. 차수(order)가 p인 AR모형은 다음과 같다.

$$Y_t = c_t + \sum_{i=1}^{p} \phi_i Y_{t-i} + u_i \tag{1}$$

여기서 $\phi_1, \phi_2, ..., \phi_p$는 상수이고 u_i는 정규분포를 따르는 백색잡음(white noise)을 나타낸다. 가장 중요한 가정은 Y_{t-i}가 t기에 Y_t의 변화를 설명한다는 것이다. 그리고 이 모형이 안정성(stationarity)을 띄기 위해선 $|\phi_i| < 1$인 조건을 충족해야 한다. 이 모형이 갖고 있는 한계는 선형인 형태(linear form)를 전제로 한다는 것이다. 그러니 복잡한 구조를 띄는 현실

경제를 선형으로 근사화 하여 접근한다는 것은 만족스런 결과를 주지 못한다(Kumer & Thenmozhi, 2007) 이런 유형의 모형은 계량분석에서 자주 다른 모형과의 비교를 위해 기준이 되는 모형으로 이용되고 있으며, 본 연구에서도 식 (1)의 AR모형을 기준으로 해서 RF 모형 및 RF-GARCH 모형의 예측력을 비교·분석할 것이다.[2)]

2.2. ARCH모형과 GARCH 모형

자기회귀적 조건부 이분산(ARCH)모형은 Engle(1982)에 의해 개발된 분석 기법으로 모형이 단순하고 금융시계열의 특징인 변동성에 나타나는 밀집현상(volatility clustering)을 잘 반영한다는 점에서 매우 널리 사용되고 있다(남준우, 이한식, 2013). 다음의 단순모형을 보자.

$$Y_t = X_t^{'}\xi \; + \; \varepsilon_t \tag{2}$$

여기서 X_t는 개수가 k인 설명변수 벡터로 시차를 갖는 종속변수들로 이루어져 있으며, ξ는 k 차원의 회귀모수 벡터이다. 이때 가장 기본적인 ARCH 모형은 오차항 ε_t에 대해 다음과 같이 가정한다.

$$\varepsilon_t \; |\psi_{t-1} \sim \; N(0, \; h_t) \tag{3}$$

여기서 $\psi_{t-1} = [Y_{t-1,}X_{t-1,}Y_{t-2,}X_{t-2,}.....]$이고 $h_t = \omega + \sum_{i=1}^{q} \alpha_i \, \varepsilon_{t-i}^2$로 조건부 분산을 나타낸다. 여기서 α_i는 자료를 통해서 추정된다. 그리고 이 ARCH모형을 일반화시킨 것이 Bollerslev(1986)에 의해 소개된 GARCH 모형이다. 일반적으로 차수가 (p, q) 인 GARCH모형은 다음과 같이 나타낸다.

$$h_t = \omega + \sum_{i=1}^{q} \alpha_i \, \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \, h_{t-j} \tag{4}$$

따라서 ARCH(q)모형은 GARCH(0,q)모형과 같다. 본 연구에서는 단순화를 위해

2) AR(p)모형의 보다 일반적인 특성에 대해서는 Asteriou and Hall(2011)을 참고할 것.

GARCH(0,1)모형을 사용하였다.

2.3. 랜덤 포레스트모형

RF모형은 Breiman(1996, 2001)에 의해 제시된 앙상블 학습(ensemble learning)방법 모형으로 부트스트랩(bootstrap)방식을 이용하여 다수의 표본을 생성하고, 의사결정나무(decision trees) 모형을 적용하여 그 결과를 종합하는 방법이다. 이때 부트스트랩 표본을 생성하는 과정에서 무작위성이 도입된다. 또한 의사결정나무 모형을 적합(fitting)하는 과정에서 각 마디(node)에서의 설명변수를 선택하는 과정에서도 무작위성이 더해진다. 즉 랜덤 포레스트 모형은 무작위성을 최대로 주기 위해 부트스트랩 표본을 생성할 뿐만 아니라 각 의사결정마디에서의 설명변수에 대해 무작위로 표본을 수집하는 방식이다(유진은,2015). Breiman(2001)은 이러한 무작위성이 최대로 되면 의사결정나무들 간에 상관관계가 줄어들게 되고 이로 인해 예측오차가 줄어든다. Breiman(2001a)에 따르면 의사결정나무의 수가 증가할수록 예측오차가 작아지며, 의사결정나무가 수가 많아도 RF모형은 과적합문제가 발생하지 않는다는 것이다.

RF모형의 또 다른 장점으로는 예측의 우수성을 들 수 있다. 특히 설명변수가 다수일 때 예측력이 매우 뛰어나며 매우 안정적인 모형을 제공한다(박창이 외, 2013; Siroky, 2009). 의사결정나무에서처럼 성장, 가지치기 등과 같은 조율모수(tuning parameter)가 없다는 점 또한 커다란 장점이다. 그러나 부트스트랩 표본을 몇 개로 할 것인지, 각 마디에서 설명변수 개수를 몇 개로 할 것인지, 결과 종합 시 선형결합을 어떻게 할 것인지 등은 연구자가 선택해야 할 사항이다(유진은, 2013).

그러나 RF모형은 다양한 우수성에도 물구하고 알고리즘은 여전히 수학적인 관점에서 볼 때 여전히 명확하게 밝혀지고 있지는 않다[Breiman(2002), Lin & Jeon(2006), Biau & Devroye & Lugosi(2008)]. 그리고 대부분의 RF모형은 기존의 알고리즘을 이용하고 있는 정도에 머물고 있다.[3] 따라서 본 연구에서도 다른 연구들과 마찬가지로 기존의 정형화된 알고리즘을 소개하는 수준에서 머물 것이다.

2.3.1. 의사결정나무(decision tree)

앞에서 언급했듯이 RF모형은 의사결정나무에 기반을 둔 앙상블 기법이다. 즉 RF모형을 구성하는 많은 의사결정나무들의 집합체인 숲(forest)은 스스로의 학습을 통해 성장해나간

3) RF모형에 대한 가장 최근의 깊이 있는 연구는 Biau(2010)을 참고할 것.

다. 이제 p-차원 벡터 $X_i=[X_1,\ldots..,X_P\ ]$는 $X_i\in R^P$로 설명변수를 나타내고, $Y\in R$는 종속변수를 나타낸다. 즉 주어진 독립변수 벡터를 이용해서 반응변수(response variable) Y를 예측하는 것이다. 이제 $P_{XY}(X,Y)$를 X와 Y의 결합분포라 하고 분포형태는 알려져 있지 않다고 하자. RF모형의 목적은 Y를 예측하기 위한 예측함수를 찾아내는 것이다. 이것은 다음의 조건부 기댓값을 구함으로서 얻어진다.

$$f(X)=E(Y|X=x)\Rightarrow Y=f(X)+\varepsilon \tag{5}$$

이러한 함수를 회귀함수(regression function)라 한다. 일반적으로 앙상블 모형은 "기초분류기(base learners)"라 불리는 $h_1(x),h_2(x),\ldots h_j(x)$들의 집합체들로서 다음과 같이 앙상블 예측인자(ensemble predictor)의 결합 형태로 나타낼 수 있다.

$$f(x)=\frac{1}{J}\sum_{j=1}^{J}h_j(x) \tag{6}$$

RF에서는 j번째의 기초분류기를 하나의 회귀나무(regression tree) $h_j(X,\Theta_j)$라고 표기한다. 여기서 Θ_j는 확률변수들의 집합이고 $j=1,2,\ldots.,J$ 이며 서로 독립적이다. 한편 의사결정나무의 분기는 모든 변수들에 대해서 이루어지며 특정 기준에 따라서 반응변수를 가장 잘 설명하는 의사결정나무를 선택하여 가지(branch)를 형성하며 성장해 나간다. 이제 분기가 이루어지는 각 마디(node)의 반응변수를 $y_1,y_2,\ldots y_n$라 하면 각 마디의 분기기준(splitting criterion)은 다음과 같다.

$$Q=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{7}$$

여기서 $\bar{y}$는 각 마디에서 예측한 값들의 평균을 나타낸다. 위의 분기기준 Q는 적합도(goodness of fit)를 나타내주는 것으로 값이 클수록 적합도가 낮고 낮을수록 적합도가 좋다는 것을 의미한다. 이제 좌측 분기기준을 Q_L, 우측 분기기준을 Q_R이라 하자. 그리고 각각의 경우에 대한 표본의 개수를 n_L과 n_R이라 하자. 그러면 분기는 다음의 값을 최소화시키는 방향으로 이루어질 것이다.

$$Q_{split} = n_L Q_L + n_R Q_R \tag{8}$$

모든 설명변수들을 대상으로 하여 이 조건을 만족하는 값들, 즉 가장 우수한 값들이 구해지면 자료는 다시 두 개의 공간으로 분리되고 위의 과정이 다시 반복되다가 특정 기준을 충족하게 되면 과정이 멈추어진다. 이때 이 과정이 멈추었을 때 분기가 발생하지 않은 마디를 끝마디(terminal node)혹은 잎(leaf)이라 한다. 이때 반응변수의 예측값은 이 끝마디의 값들을 평균해서 얻어진다.

2.3.2. 랜덤 포레스트

앞에서 언급했듯이 RF모형에서는 기초분리기 $h_j(X,\Theta_j)$를 의사결정나무로 사용한다. 이제 훈련자료를 $\Xi = \{(x_1,y_1),....,(x_N,y_N)\}$라고 하자. 이때 $x_i = (x_{i,1},x_{i,2},....x_{i,p})^T$는 p개의 설명변수이고 y_i는 반응변수 그리고 θ_j는 확률적 요인인 Θ_j의 임의값이라 하자. 그러면 선택된 의사결정나무(fitted tree)는 $\hat{h_j}(x,\theta_j,\Xi)$가 된다. 이 방법이 Breiman(2001)이 제시한 방법이다. 그러나 θ_j는 관찰할 수 없는 변수이므로 이것의 무작위성을 확보하기 위해서 일반적으로 사용하는 다음의 두 가지 방법을 이용한다. 첫째, 모든 의사결정나무를 부트스트랩 방식으로 추출된 표본에 적합시켜 의사결정나무 숲(forest)을 구성한다. 두 번째는 마디를 분기하는 과정에서 변수에 무작위성을 부여한다. 즉 모든 분기과정에서 변수를 선택할 때 p개의 모든 변수를 고려하는 것이 아니고 m개의 변수를 무작위로 사용한다. 이 과정에서 전체 훈련자료 중에서 N개의 부트스트랩 표본을 얻은 후 남는 표본이 있는데 이것을 'OOB(out-of-bag)'자료라 하며, 이 데이터를 이용해서 과적합 문제를 방지하기 위해서 분류오차(generalization error)와 변수 중요도 지수(variable importance)를 구한다.

2.3.3. 조율(tunning)

일반적으로 RF모형의 효율성을 향상시키기 위해 사용하는 조율모수(tunning parameter)에는 다음 세 가지가 이용된다. 첫째, 각 마디에서 무작위로 이용되는 설명변수의 개수(m), 둘째, 성장하는 의사결정나무의 개수(J). 셋째. 끝마디의 최대 개수로 측정되는 나무의 크기를 들 수 있다. 이 가운데 RF모형에 민감하게 영향을 주는 것은 m으로 보통 반응변수가 연속형 자료인 경우 $p/3$개(p는 전체 설명변수의 개수), 범주형 자료인 경우 $\sqrt{p}$개가 이용된다(Zhang & Ma, 2012, p.167)[4] 한편 조율을 하는 과정에서 가장 문제가 되는 것은 과적

합문제이나 Diaz-Uriatre & Alvarez de Andres, 2006)에 따르면 과적합은 아주 미미한 것으로 나타났다.

2.3.4. 변수 중요도

RF모형은 학습과정을 통해 각 변수의 중요도를 계산한다. 이는 분기되는 각 의사결정나무 마디마다 변수 값들이 분류의 정확성에 미치는 정도를 기초로 하여 측정되는 값으로, 변수의 중요도를 파악하는데 유용하게 이용된다. 어떤 변수들이 종속변수를 예측하는데 있어서 중요한 변수로 고려되어야 할 것인가를 알아보기 위해서 OOB(out-of - back)자료로 실제 관측값과 예측된 값과의 차이를 이용해서 오차를 구하는 방법으로 정확도(Mean Decrease Accuracy)로 나타내진다. 즉 원래의 OOB 자료의 표준오차와 수정된 OOB자료의 분류오차 차이 평균을 표준오차로 나눈 값을 그 변수에 대한 중요도 지수로 사용하게 된다(Hastie er al.,2001; Strobl et al., 2009). 이때 그 변수의 영향력이 클수록 중요도 지수가 커진다(Strobl et al., 2009).

Ⅲ. 실증분석

3.1. 자료

본 연구에서 사용하고 있는 변수를 선택함에 있어 환율에 영향을 줄 수 있는 가능한 변수를 최대한 고려하면서 RF모형의 학습과정을 통해 그 중요도가 결정되도록 하는 방법을 채택하고 있다. 자료는 주로 KRX Index와 Quandl을 이용하고 있으며 국내부문 변수는 대미달러 환율, 금리, KRX 섹터지수, 주요국 환율을 이용하고 있고, 해외부문에 대해서는 원자재 가격, 주요 선진국 주가지수, 주요 선진국 금리 그리고 주요 선진국의 대미달러 환율을 이용하였다.[5] 자료는 김경태, 안정국, 김동현(2015)에서 제공하고 있는 2010년 7월 1일부터 2014년 6월 30일 까지 일별자료를 이용하였다.[6]

모형에 사용된 변수와 전체 관찰 값의 개수는 각각 65개와 1096개로 모형의 구성을 위해

4) 본 연구에서 사용하고 있는 통계프로그램인 R에서는 Breiman(2001)이 제시한 값들을 기본값으로 사용하고 있다.
5) 자세한 변수들의 출처 및 데이터에 대한 설명 등은 <부록>을 참조할 것.
6) R을 이용하여 데이터 마트를 구성하는 자세한 방법은 김경태 외(2015) 참조할 것.

훈련용 자료와 시험용 자료를 각각 70%와 30%로 구분한다. 훈련용 자료는 모형을 구성하기 사용하는 자료에 해당하고, 시험용 자료는 훈련용 자료와 동일 기간의 검증용 데이터를 말한다. 한편 E_t 를 t 시점에서의 환율이라고 한다면 환율의 변화율은 다음과 같이 나타낼 수 있다.

$$r_t = 100^* \ln\left(\frac{E_t}{E_{t-1}}\right) \tag{9}$$

아래의 <표 1>은 이렇게 변환된 환율의 변화율에 대한 기본적인 통계적 특성을 나타내고 있다.

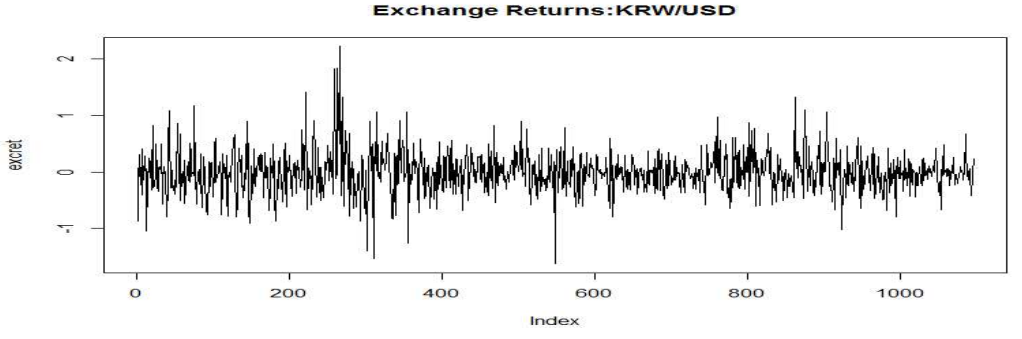
<표 1> 환율의 변화율에 대한 요약

구분	최소값	최대값	평균	중앙값	왜도	첨도	표준편차
USD/KRW	-1.6290	2.224	-0.0006	-0.0109	0.5442	4.1157	0.3568

<표 1>에서 보는 것처럼 대미달러 환율의 첨도값은 4.0을 넘고 있어 분포의 형태가 정규분포 형태를 띄고 있지 않다는 것을 나타내주고 있다. 또한 환율의 평균 수익률은 제로(0)임을 나타내주고 있다. 왜도의 경우 양(+)의 값을 나타내고 있어 오른쪽 꼬리 모양을 갖는 분포모습을 띄고 있다.

<그림 1>은 <식 9>를 이용하여 1차 차분 변환된 환율의 일별 자료를 그림으로 나타낸 것이다. 그림에서 보는 것처럼 환율은 평균을 중심으로 변동성이 아주 강하게 나타나고 있다. 본 연구에서는 이렇게 구한 원-달러 환율의 1일 변화율 r_t 를 반응변수로 이용하고 있다.

<그림 1> 원-달러 환율의 변화율



3.2. 모형의 추정

3.2.1. 변수의 안정성

예측 모형의 기준으로 삼을 AR모형을 추정하기 이전에 목표변수로 삼고 있는 환율의 변화자료에 대한 안정성(stationarity)에 대해 먼저 살펴본다. 변수가 안정적인지를 판단하기 위해선 다양한 방법을 사용할 수 있다. 그 가운데 가장 일반적으로 사용되는 ADF 검정(ADF: Augmented Dickey-Fuller Test)[7]을 이용한 단위근 검정(unit root test)결과를 <표 2>에 제시해놓고 있다.

<표 2> 단위근 검정 결과

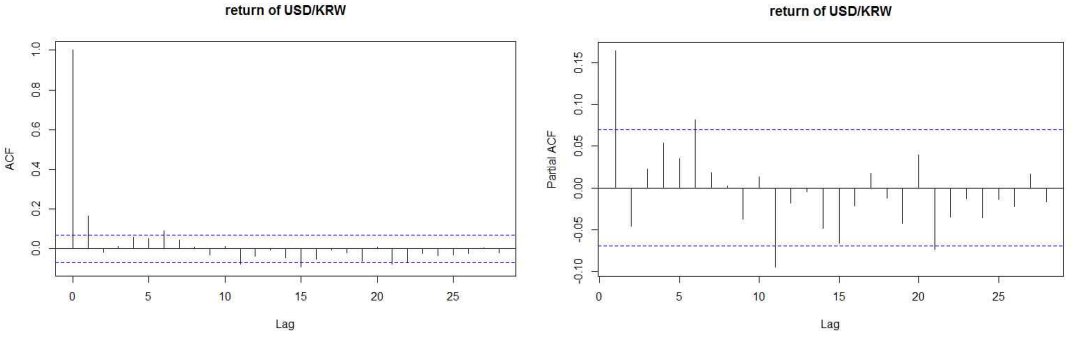
변수명	t-통계값		임계치
환율 변화율	수준변수	1차 차분 변수	10% : -2.57 5% : -2.86 1% : -3.43
	-2.2139	-18.976	

이 결과에 따르면 환율의 변화율은 1차 차분을 하게 되면 단위근이 존재한다는 귀무가설을 기각하고 안정적 자료로 전환됨을 알 수 있다. 다른 변수들에 대해서도 안정성을 확보하기 위해 1차 로그 차분한 자료들을 사용하고 있다. AR모형의 차수를 결정하기 위해 환율의 변화율에 대한 ACF와 PACF에 대한 상관도표(correlogram)를 통해 차수를 살펴보자.

우선 <그림 2>의 ACF 그림을 보면 ACF가 급격하게 줄어들고 있어 환율의 변화율 자료가 안정적임을, 즉 자기회귀 과정이 단기에 머물고 있음을 알 수 있고, PACF 의 경우 첫 번째 시차 이후 급격하게 줄어들며 나머지는 표준오차 범위안에 머물고 있다. 이 두 가지 결과를 종합해서 볼 때 환율의 변화율은 안정적이며 AR(1)과정을 따른다고 볼 수 있다.

7) ADF 테스트는 시계열 자료가 단위근(unit root)을 갖는지 여부를 검증하는 방법으로 시계열 자료가 안정적이지 않다(단위근이 존재한다)는 것을 귀무가설로 하고 대립가설인 안정적이라는 것(단위근이 존재하지 않는다)을 적극적으로 채택하려 한다.

<그림 2> 환율의 변화율에 대한 ACF와 PACF



3.2.2. 자기회귀(AR)모형

기준 모형인 AR모형을 추정하기 위해 자료를 훈련용 자료와 시험용 자료로 70:30의 비율로 구분한다. 그리고 훈련용 자료를 이용해서 모델을 만든 후 시험용 자료를 이용해서 예측을 수행하기로 한다. 그 결과 회귀계수 값은 0.1652로 PACF 상관도표에서 반영되고 있음을 확인할 수 있다. 또한 훈련용 자료를 이용하여 1단계 전진 방식(one-step-ahead)으로 모형을 예측해 본 결과 모형의 예측력을 나타내주는 RMSE 값은 0.3785로 나타났다. 앞으로 모형의 예측력을 비교할 때 이 값을 기준으로 삼을 것이다. 즉 RF모형이나 RF-GARCH 모형의 RMSE값이 이 값보다 적을 값을 나타낼 경우 모형의 예측력이 향상되었다고 판단할 수 있다.

3.2.3. 랜덤 포레스트

GARCH 모형이 도입되지 않은 순수한 RF 모형을 추정하기 위해 결정나무의 개수는 기본값인 J=500으로 선택하였고, 변수의 개수 m은 모든 변수인 65개를 선택하는 대신 Breiman(2001a)의 방법을 따라 기본값인 $p/3$를 적용하여 m=21로 한 후 이 값의 절 반 즉 m=11 그리고 두 배인 m=42을 각각 적용하여 가장 좋은 결과를 채택하였다.

3.2.4. GARCH로 확장된 랜덤 포레스트

시계열자료를 예측하는 경우 해당 시점까지의 모든 정보들을 이용하여 예측을 행하게 된

다. 따라서 t시점까지의 이용 가능한 정보들 즉, X_t에 근거를 둔 반응변수 Y_t의 분포형태에 대해서 알아야 하나 아직까지 종속변수의 분포, 특히 이분산 형태를 갖는 변수를 고려한 데이터 마이닝 모형 연구결과가 국내의 경우 거의 없는 실정이다. 본 연구에서는 이러한 점을 고려하여 RF모형에 GARCH 모형을 결합하여 반응변수를 예측하는 방법을 택하고 있다. 다음과 같은 함수를 고려하자.

$$\hat{Y}_t = f(\Theta; X_t) \tag{10}$$

$$h_t = \gamma_0 + \gamma_1 \varepsilon_{t-1}^2 \tag{11}$$

여기서 Θ는 정보 집합 이고 $\varepsilon_t = Y_t - \hat{Y}$ 그리고 h_t는 Y_t의 분산을 나타낸다. 이 과정에서 회귀계수 γ를 구하기 위해 계량경제학에서 사용하고 있는 Cochran-Orcutt 반복 추정법을 사용하고 있다. 추정과정을 간단하게 설명하면 다음과 같다. 첫 단계, 식 (10)을 RF모형을 통해 추정을 한다. 두 번째, 첫 단계에서 구한 잔차(residual)를 이용해서 GARCH(0,1)모형을 통해 식(11)의 분산을 구한다. 이를 통해 $\sqrt{h_t}$, 즉 조건부 표준편차가 추정된다. 세 번째 단계에선 $\sqrt{h_t}$를 이용해서 Y_t의 변동성 군집효과(volatility clustering effect)를 제거한다. 즉

$$Y_t^* = Y_t / \sqrt{h_t} \tag{12}$$

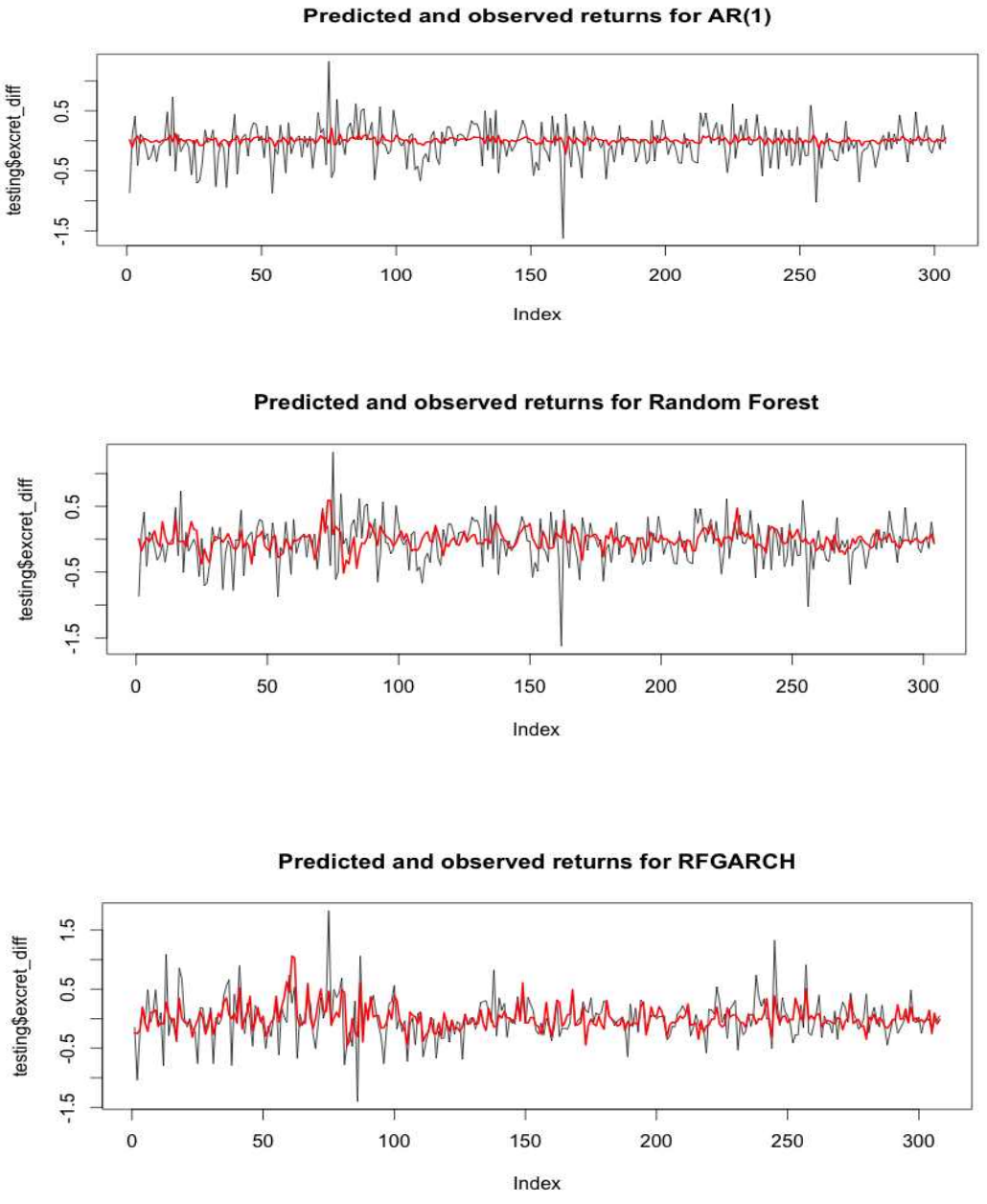
마지막 단계로 변동성 군집효과를 제거하여 구한 Y_t^*를 이용하여 수렴할 때까지 첫 단계에서 부터 세 번째 단계를 반복적으로 수행한다. 이렇게 해서 얻어진 예측 값을 원래의 값으로 변환하여 모형의 예측능력을 평가한다.

3.3. 추정 결과

3.3.1. 모형의 평가

<그림 3>은 각각 환율의 변화율에 대한 AR(1)모형의 추정결과와 RF모형 그리고 RF-GARCH 모형을 이용한 예측값(predicted value)과 실제값(actual value)을 비교해서 나타내주고 있다. 붉은 선이 예측 값을 나타내고 흑백 선이 실제 값을 나타낸다.

<그림 3> 원-달러 환율 변화율에 대한 실제값과 예측값



이 그림을 통해서 알 수 있듯이 환율의 변동성이 심하게 나타남에도 불구하고 RF모형과 RF-GARCH모형은 실제값에서 나타나고 있는 변동성을 반영하면서 변화하고 있다. 그러나 AR(1)모형의 경우 훈련용 데이터가 지속적으로 업데이트됨에도 불구하고 변동성을 반영하지 못한 체 평균수익률인 제로(0)수준에서 거의 변화가 없어 RF모형과 RF-GARCH모형이 더 효율적임을 나타내고 있다.

한편 모형의 예측효율성을 나타내는 RMSE값을 계산해본 결과 AR(1) 모형은 0.3785로 이는 AR(1) 모형을 이용하여 예측을 할 경우 대략 0.3785% 포인트만큼 평균적으로 오차가 발생한다는 것이다. RF모형은 각 마디에서 선택하는 설명변수의 개수(m)가 두 배인 42개로 했을 경우 0.3545로 가장 작은 값을 나타내고 있으나 설명변수를 다른 값으로 변화시켜도 거의 같은 값을 나타내고 있다. 그리고 AR(1) 모형과 비교해보았을 때 RF모형은 AR(1)모형에 비해 예측력이 개선되고 있음을 알 수 있다. RF-GARCH 모형의 경우 설명변수를 각각 달리 했을 경우 RMSE값은 차이가 거의 없으나 AR(1)모형이나 RF모형에 비해 예측력이 상당부분 개선되고 있음을 알 수 있다. 따라서 환율 변화율에 대한 예측의 효율성은 기존의 시계열분석 모형보다 데이터 마이닝을 통한 RF모형이나 RF-GARCH모형이 더 높은 것으로 나타났다.

<표 3> 모형의 예측능력 비교

모형	옵션	RMSE
AR(1)		0.3785
Random Forest	mtry=half(10)	0.3548
	mtry=default(21)	0.3760
	mtry=double(42)	0.3545
Random Forest-GARCH	mtry=half	0.3085
	mtry=default	0.3078
	mtry=double	0.3057

3.3.2. 모형의 평가

다음으로 <표 4>는 앞의 2.3절에서 제시한 기준에 따라 환율의 변동에 있어 영향을 미치는 중요한 변수들을 나타내고 있다. 두 번째 열(%IncMSE)은 해당 변수를 모형에서 제외시켰을 경우 MSE로 측정되는 예측오류가 얼마나 증가하는가를 나타낸 것으로 이 값이 클수록 변수의 중요도가 크다고 볼 수 있다. 세 번째 열 노드의 순도증가(IncNodePurity)는 각 변수들이 노드의 불순도(node impurity)를 개선하는데 얼마만큼 기여하는가를 나타내는 것

으로 Gini Index를 이용한다. 즉 지니 지수 값이 적을수록 불순도가 감소하므로 노드의 순도는 증가한다. 이렇게 구해진 변수들은 다른 모형들을 구성함에 있어 사용할 변수들을 선택하거나 모형의 예측력을 향상시키는데 중요한 역할을 한다. 이것을 나타낸 것이 <그림 5>이다.

<표 4>와 <그림 4>에서 보는 것처럼 환율의 변화에 가장 큰 영향을 미치는 변수는 미국 달러지수로 나타나고 있다. 그 뒤를 이어 KRX 100과 KOSPI 지수가 환율의 변화에 영향을 미치는 변수로 작용하고 있음을 확인할 수 있다. 아울러 유럽과 일본의 주가지수, 유럽과 중국의 환율도 중요한 요인으로 작용하고 있다.

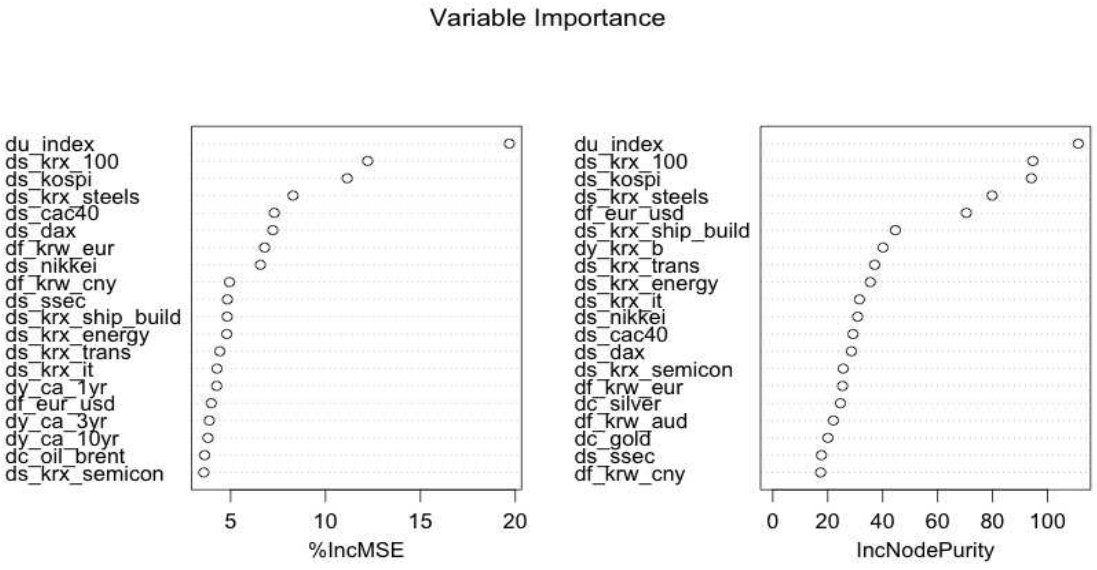
중요한 변수를 10개만 열거한 것은 10개 이후에는 중요도 값이 4 이하로 급격하게 감소하기 때문에 나머지 변수들을 모두 열거할 필요가 없기 때문이다. 이 결과들은 김경태, 안정국, 김동현(2015)에서 제시하고 있는 변수들과 대부분 일치하고 있다.

<표 4> 환율 변화에 영향을 미치는 요인들의 중요도

변수명	%IncMSE	IncNodePurity
us_index (미국 달러 지수)	19.6847	111.1833
stck_krx_100(KRX100 지수)	12.2198	94.6382
stck_kospi(KOSPI지수)	11.1421	94.0881
stck_krx_steels(KRX Steel))	8.2877	79.8065
stck_cac40(프랑스 주가 지수)	7.3000	29.1884
stck_dax(독일 주가 지수)	7.2210	28.6195
fx_krw_eur(유로화 환율)	6.7915	25.3743
stck_nikkei (니케이 지수)	6.5724	30.8779
fx_krw_cny(중국 환율)	4.9393	17.4062
stck_ssec(중국 주가 지수)	4.8312	17.7236

주) 모든 변수들에 대한 중요도(%IncMSE)와 불순도(IncNodePurity)에 대한 값들은 <부록>을 참조.

<그림 4> 변수들의 중요도



Ⅳ. 결론

본 연구에서는 최근 들어 빅 데이터 분석과 관련하여 널리 이용되곤 있으나 경제학 분야에선 아직까지 활발하게 사용되고 있지 못한 데이터 마이닝 기법을 이용하는 기계학습모형에 금융시계열자료에 내재되어 있는 변동성효과를 고려한 계량경제모형을 결합하여 환율의 변화를 예측하여 보았다. 주목할 만한 결과는 GARCH 모형을 접목한 랜덤 포레스트 모형이 기존의 AR(1)모형 그리고 변동성이 고려되지 않은 RF 모형보다 변동성 군집효과를 잘 반영하며 예측능력이 보다 향상되고 있음을 나타냈다.. 이는 RF모형이 갖고 있는 예측의 우수성과 변동성이 제거된 데이터를 같이 활용한 결과인 것으로 생각된다.

또한 RF- GARCH 모형의 변수들에 대해 다양한 값을 선택하여 보았으나 큰 차이를 나타내지 않았으며, 환율의 변화에 중요한 영향을 미치는 변수들로는 미국 달러지수와 KRX 100지수 그리고 KOSPI 100 지수 순으로 나타났으며 그 외 유럽과 미국 그리고 일본의 주가지수가 중요 변화 요인으로 관측되었다. 이는 금융시장에서 경제주체들이 환율의 변화를 예측하는데 있어 중요한 고려요인으로 작용하고 있다는 것을 시사한다.

본 연구는 빅 데이터를 처리할 수 있는 기계학습모형이 기존의 계량경제모형보다 예측 능

력이 뛰어나다는 것을 보이는 데에만 중점을 두고 있다. 즉 본 연구에서 이루어진 모형의 신뢰성에 대한 검증은 본 연구에서 이루어지고 있지 않아 향후 연구과제로 남겨두고 있다. 또한 선형회귀분석의 변수 선택에 있어서 RF 모형을 통한 변수의 선택이 선형회귀분석의 예측능력을 향상시킬 수 있을 것인가 하는 것도 향후 연구과제로 남겨둔다. 또한 GARCH 모형뿐만 아니라 다양한 시계열 모형들과 기계학습모형과의 결합을 통해 금융시장에 존재하는 다양한 자료들의 예측력을 향상시킬 수 있을 것으로 기대한다.

참 고 문 헌

김경태 · 안정국 · 김동현(2015), “Big Data: 빅 데이터 활용서”, 시대에듀, 제6장.

남준우 · 이한식(2010), “계량경제학(3판):이론과 EViews/Excel 활용”, 홍문사.

박창이 · 김용대 · 김진석 · 송종우 · 최호식(2013). “R을 이용한 데이터 마이닝(개정판), 교우사.

유진은(2013), “랜덤 포레스트: 의사결정나무의 대안으로서 데이터 마이닝 기법”, **교육평가연구**, 제28권 제2호, 427-448

한태경(2010), “환율예측을 위한 합성모형 연구 : 시계열분석방법과 기계학습방법을 이용한 합성모델”, **KAIST 석사학위논문**

김창범 · 모수원(2001),“공적분과 오차수정모형을 이용한 환율의 추정과 예측”, **산업경제연구**, 제13권 제6호, 479-489

김창범(2007), “몬테카를로 시뮬레이션을 이용한 환율 예측”, **산업경제연구**, 제20권 제5호, 2075-2093

Asteriou, D. and Hall, S.G.(2011), “Applied Econometrics”, Palgrave

Biau, G., Devroye, L. and Lugosi, G.(2008), “Consistency of Random Forest and Other Averaging Classifiers,” *Journal of Machine Learning Research,* 9, 2015-2033

Bollerslev, T.(1986), “Generalized Autoregressive Conditional Heteroscedasticity,” *Journal of Econometrics,* 31, 307-327

Breiman, L.(2001a). “Random Forests,” *Machine Learning* 45(1), 5-32

Breiman, L.(2001b). “Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author),” *Statistical Science* 16(3), 199-231

L. Breiman.(2002) Manual on Setting Up, Using, and Understanding Random Forests vol. 3.1

Breiman, L.(1996). "Bagging Predictors," *Machine Learning* 24(2), 123-140.

Diaz-Uriarte, R. and Alvarez de Andres, S.(2006). “Gene Selection and Classification of

Microarray Data Using Random Forest," *BMC Bioinformatics* 7(1) 3.

Engle, R.F.(1982), "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica,* 50(4), 987-1008

Hastie T, Tibshirani R and Friedman J(2001), "The Elements of Statistical Learning; Data Mining, Inference, and Prediction," Springer.

Kumar, M. and Thenmozhi, M.(2007), "Predictability and Trading Efficiency of S&P CNX Niftyindex Returns using Support Vector Machines and Random Forest Regression," *Journal of Academy of Business and Economics,* 7(1), 150-164.

Lin, Y. and Jeon, Y.(2006), "Random Forest and Adoptive Nearest Neighbor", *Journal of the American Statistical Association,* 101, 578-590.

Meese, A.R. and Rogoff, K.(1983), "Empirical Exchange Rate Models of the Seventies : Do They Fit out of Sample?," *Journal of International Economics* 14, 3-24.

Siroky, D.S.(2009), "Navigating and Random Forest and Related Advances in Algorithmic Modeling," *Statistics Survey,* 3, 147-163.

Strobl, C. and James, M. and Gerhard, T.(2009), "An Introduction tot the Recursive Partitioning: Rationale, Application, amd Characteristics of Classification and Regression Trees, Bagging and Random Forests," *Psychological Methods.* 14(4), Dec. 323-348

Varian, H.(2014), "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives,* 28(2), Spring, 3-28

Zang, C. and Ma, Y.(2012), "Ensemble Machine Learning: Methods and Application", Springer

Foreign Exchange Rate Forecasting Using the GARCH extended Random Forest Model

Jong-Duk Suh*

Abstract

This paper focused on investigating the predictability of exchange rate returns on daily frequency using the Random Forest model that have been mostly developed in the machine learning field especially data mining. A financial returns data is the presence of volatility clustering but machine learning models assume a constant variance. We therefore extend Random Forest model with the most widely used model for volatility clustering, the GARCH process. This GARCH extended Random Forest model is applied to show whether to improve the predictability of exchange rate return or not. In addition using this model, which variables are important for volatility in exchange market.

Our results show that GARCH extended Random Forest model has a great potential for improving the predictive performance of the forecasting of exchange rate returns. In the exchange market US dolla r index is the most important variable and KRX 100 is the second and finally KOSPI index is the third variable.

Keyword: machine learning model, data mining, random forest

* Assistant Professor, Dep't of Economics, Daejeon Univ.

<부록 1> 변수 명 및 자료 출처

NO	구분	변수그룹	변수명	설명	데이터 원천
1		목표변수	fx_krw_usd	미국달러화 환율	Quandl
2		금리	yield_krx_b	KRX 채권지수	KRX
3			yield_krx_g	KRX 국고채프라임지수	KRX
4			yield_krx_t	KRX-Korea Treasury Bond Index	KRX
5		KRX섹터지수	stck_kospi	코스피지수	Quandl
6			stck_krx_100	KRX 100	KRX
7			stck_krx_autos	KRX Autos	KRX
8			stck_krx_energy	KRX Energy &Chemical	KRX
9			stck_krx_it	KRX IT	KRX
10			stck_krx_semicon	KRX Semicon	KRX
11			stck_krx_ship_build	KRX Shipbuilding	KRX
12			stck_krx_steels	KRX Steels	KRX
13			stck_krx_trans	KRX Transportation	KRX
14		주요국 환율	fx_krw_aud	호주 달러화 환율	Quandl
15			fx_krw_cny	중국 위안화 환율	Quandl
16			fx_krw_gbp	영국 파운드화 환율	Quandl
17			fx_krw_eur	유로화 환율	Quandl
18	해외	USD지수	usd_index	US Dollar Index	Quandl
19		원자재 가격	cmd_copper	동 가격	Quandl
20			cmd_com	옥수수 가격	Quandl
21			cmd_gold	금 가격	Quandl
22			cmd_oil_brent	브랜트유 가격	Quandl
23			cmd_oil_wti	서부텍사스유 가격	Quandl
24			cmd_gas	천연가스 가격	Quandl
25			cmd_silver	은 가격	Quandl
26		주요국 주가지수	stck_cac40	프랑스 주가지수	Quandl
27			stck_dax	독일 주가지수	Quandl
28			stck_nasdaq	나스닥 주가지수	Quandl

29			stck_nikkei	일본 주가지수	Quandl
30			stck_nyse	미국 나이스지수	Quandl
31			stck_snp500	미국 S&P500 지수	Quandl
32			stck_ssec	중국 주가지수	Quandl
33		금리	yield_ca	캐나다 금리(1,3,6개월, 1,3,5,10년)	Quandl
34			yield_jp	일본 금리(1,3,5,10년)	Quandl
35			yield_fr	프랑스 금리(10년)	Quandl
36			yield_nz	뉴질랜드 금리(1,3,6개월,1,5,10년)	Quandl
37			yield_uk	영국 금리	Quandl
38			yield_us	미국 금리(6개월, 1,3,5,10년)	Quandl
39		주요국 대 달러 환율	fx_aud_usd	달러/호주 달러화 환율	Quandl
40			fx_cad_usd	달러/캐나다 달러 환율	Quandl
41			fx_cny_usd	달러/ 중국 위안화 환율	Quandl
42			fx_eur_usd	달러/ 유로화 환율	Quandl
43			fx_gbp_usd	달러/ 영국 파운드화 환율	Quandl
44			fx_jpy_usd	달러/ 일본 엔화 환율	Quandl
45			fx_nzd_usd	달러/ 뉴질랜드 달로화 환율	Quandl

<부록 2> 변수별 중요도 및 불순도

	%IncMSE	IncNode Purity			%IncMSE	IncNode Purity
cmd_copper	0.2432	16.3099		yield_uk_5yr	1.4514	11.7332
cmd_corn	0.7807	2.5638		yield_uk_10yr	1.7998	10.8594
cmd_gold	1.8422	20.0583		yield_us_6m	0.1045	15.0747
cmd_oil_brent	3.6265	13.4940		yield_us_1yr	-2.5468	8.1564
cmd_oil_wti	1.9303	13.7201		yield_us_3yr	1.9093	8.2716
cmd_gas	2.1805	13.3912		yield_us_5yr	2.3567	8.3766
cmd_silver	3.5364	24.6250		yield_us_10yr	1.9197	10.1680
fx_aud_usd	2.0569	13.0108		stck_cac40	7.3002	29.1883
fx_cad_usd	0.7446	15.6230		stck_dax	7.2210	28.6194
fx_cny_usd	1.2137	13.8537		stck_nasdaq	2.7950	11.9301
fx_eur_usd	3.9785	70.4264		stck_nikkei	6.5723	30.8779
fx_gbp_usd	3.5809	11.5797		stck_snp500	0.8887	12.0812
fx_jpy_usd	2.0565	13.8080		stck_ssec	4.8312	17.7235
fx_nzd_usd	3.1487	14.4739		usd_index	19.6847	111.1883
yield_ca_1m	1.5465	10.8542		fx_krw_aud	1.5735	25.0517
yield_ca_3m	-1.9409	10.3139		fx_krw_cny	4.9392	17.4062
yield_ca_6m	1.2215	5.5960		fx_krw_gbp	2.0069	15.6953
yield_ca_1yr	4.2659	8.0288		fx_krw_eur	6.7914	25.3743
yield_ca_3yr	3.8777	9.6126		yield_krx_b	3.2835	40.1360
yield_ca_5yr	3.4083	11.9427		yield_krx_g_3yr	2.9747	15.2317
yield_ca_10yr	3.8049	11.0066		yield_krx_g_5yr	0.8641	14.9701
yield_jp_1yr	0.9833	9.6186		yield_krx_g_10yr	1.3222	13.2361
yield_jp_3yr	1.6542	14.9460		yield_krx_t	2.9397	13.0393
yield_jp_5yr	2.4564	14.3364		stck_kospi	11.1421	94.0881
yield_jp_10yr	1.5076	14.0051		stck_krx_100	12.2198	94.6381
yield_fr_10yr	3.1351	14.0716		stck_krx_autos	1.6322	12.7034
yield_nz_1m	-0.8893	7.2205		stck_krx_energy	4.7912	35.4710
yield_nz_3m	0.6004	4.3536		stck_krx_it	4.2853	31.5392
yield_nz_6m	-0.9827	12.8022		stck_krx_semicon	3.5853	25.6189
yield_nz_1yr	1.1917	3.6998		stck_krx_ship_build	4.8197	44.6129
yield_nz_5yr	1.5493	9.4784		stck_krx_steels	8.2876	79.8064
yield_nz_10yr	2.2828	10.4923		stck_krx_trans	4.4302	37.0658