Journal of Economic Theory and Econometrics, Vol. 29, No. 3, Sep. 2018, 23–57

Determinants of Bank and Non-bank Household Loans and Short- and Long- Horizon Forecast*

Chang-hoon Lee† Kyu-Ho Kang‡ Junghwan Mok §

Abstract The instability of the financial system is likely to occur when particular types of loans surge rather than all types of loans surge at the same time. A preemptive policy response requires a monitoring system based on forecasts by different loan types. The purpose of this study is to forecast household loans by categorizing into four types : bank mortgage loan, bank credit loan, non-bank mortgage loan, and non-bank credit loan. Given the fact that there are numerous determinants and forecasting models for household loans, and that the determinants differ depending on the type of household loans, this study sets out the density forecasting algorithm based on Bayesian Machine Learning. which consists of a variable learning process, a model learning process, and a forecasting combination process. We find bank mortgage loans are largely predicted by the loan rates, the volume of apartments to be moved in, and the number of apartment units to be sold. while the key determinants of bank credit loans are the employment rate and Jeon-se price index. On the other hand, the non-bank mortgage loans are largely determined by the loan rates and the ratio of apartment sales prices relative to Jeon-se prices. The non-bank credit loans are also influenced by not only the employment rate and the Jeon-se price index but also stock returns.

Keywords Bank Loan, Non-bank Loan, Variable Selection, Model Selection

JEL Classification F31, F32, C53

*Authors gratefully acknowledges financial support from Financial Stability Department of the Bank of Korea. The views in this paper are those of the authors, and do not necessarily reflect those of the Bank of Korea

†Department of Economics, Korea University, E-mail: rollin0807@korea.ac.kr, Tel: 02-3290-5132

‡Corresponding Author, Department of Economics, Korea University, E-mail: kyuho@korea.ac.kr, Tel: 02-3290-5132

§Financial Stability Department, The Bank of Korea, E-mail: jhmok@bok.or.kr, Tel: 02-750-6858

Received April 4, 2018, Revised May 31, 2018, Accepted June 5, 2018

Journal of Economic Theory and Econometrics, Vol. 29, No. 3, Sep. 2018, 23–57

은행권 및 비은행권 가계대출 결정요인 분석과 장단기 예측*

이창훈† 강규호‡ 목정환§

Abstract 거시건전성 불안은 모든 형태의 대출이 동시에 급증하기 보다는 일부 형태의 대출이 급증할 경우에 야기될 가능성이 높다. 선제적인 정책적 대응을 위해 대출 형태별 예측을 통한 모니터링 시스템 구축이 요구된다. 본 연구의 목적은 전체 가계대출을 은행권 주택담보대출(주담보), 은행권 마이너스 통장대출(마통), 비은행권 주담보, 비은행권 마통 등 네 개 유형으로 구분하여 예측하는 것이다. 잠재적인 가계대출 결정요인과 모형이 다수이고 가계대출의 형태에 따라 결정요인이 상이하다는 점을 감안하여 본 연구는 베이지안 머신 러닝 기반 가계대출 유형 별 분포예측 알고리즘을 제시한다. 본 연구의 베이지안 머신 러닝 알고리즘은 변수 학습과정, 모형 학습과정, 예측 조합과정으로 이루어진다. 예측 결과, 은행권 주택담보대출은 주로 대출금리, 아파트 입주물량, 분양물량 등에 의해 예측가능하며, 은행권 마통은 취업률과 전세가격지수가 주요 예측변수로 작용하였다. 반면, 비은행권 주담보는 대출금리와 아파트 매매전세가비율 등에 의해 주로 결정되며, 비은행권 마통도 취업률, 주가 수익률과 더불어 전세가격지수의 영향을 많이 받는 것으로 추정되었다. 각 형태별 가계대출은 높은 지속성으로 인해 향후에도 현재와 유사한 수준의 증가율을 보일 것으로 예측된다. 다만 예측치에 내재된 불확실성이 상당하기 때문에 정책당국은 이러한 점을 반드시 고려하여야 한다.

Keywords 은행권 대출, 비은행권 대출, 변수 선택, 모형 선택

JEL Classification F31, F32, C53

*본 연구는 한국은행의 재정 지원으로 진행되었다. 본 연구의 내용은 집필자의 개인의견이며, 한국은행의 공식견해와는 무관하다.

†고려대학교 경제학과, 박사과정, E-mail: rollin0807@korea.ac.kr, Tel: 02-3290-5132

‡고려대학교 경제학과 부교수, E-mail: kyuho@korea.ac.kr, Tel: +82-2-3290-5132

§한국은행, 금융안정국 과장, E-mail : jhmok@bok.or.kr, Tel: +82-2-750-6858

Received April 4, 2018, Revised May 31, 2018, Accepted June 5, 2018

1. 도입

2015년 이후 우리나라 가계대출은 은행권 주택담보대출(이하 주담대)과 마이너스 통장대출(이하 마통)을 중심으로 급등했다. 강종구(2017)의 연구에 의하면 가계대출 증가는 유동성 제약 완화를 통해 경기활성화에 단기적인 도움을 주지만, 누적된 가계대출은 차입가계의 부채상환부담 가중과 금리변동 위험 확대 등을 통해 소비와 경제성장을 저해하기도 한다. 또한 과도한 가계대출 수준은 금융시스템을 외부충격에 취약하게 만들어 금융위기 발생 가능성을 높인다. 이러한 금융시스템 불안을 사전에 효과적으로 방지하고 선제적인 정책대응을 하기 위해서는 정확한 가계대출 전망이 필요하다.

가계대출은 크게 은행권 주담대, 은행권 마통, 비은행권 주담대, 비은행권 마통으로 분류할 수 있는데, 유형별 가계대출 증가율 동태성이 대단히 상이하다. 특히, 2015년 이후 전년동월대비 은행권 주담대 증가율은 15% 에서 10% 내외로 감소한 반면, 은행권 마통 증가율은 2% 에서 9% 가까이 상승하였다. 반대로 비은행권 주담대 증가율은 2016년 이후 급증하여 0% 에서 15% 까지 상승하였으며, 비은행권 마통 증가율은 2016년 말 이후 5%p 이상 하락하였다. 형태별 가계대출이 중요한 이유는 총 가계대출 변동의 요인이 주대담이냐 마통이냐, 또는 은행권이냐 비은행권이냐에 따라서 정부의 가계부채 안정화 정책이 다르기 때문이다. 예를 들어, 가계부채 총량이 증가하더라도 그 요인이 주담대가 아니라 마통이라면 LTV나 DTI 규제는 오히려 주택 실수요자의 후생을 감소시킬 수 있다. 더불어 은행권의 가계대출 급등이 비은행권에 비해 경제적 파급효과가 훨씬 더 크기 때문에, 은행권과 비은행권을 구분해야만 정책당국이 규제의 적용범위와 강도를 적절히 선택할 수 있을 것이다. 뿐만 아니라, 가계대출의 형태에 따라서 각각 결정요인이 상이할 수 있으므로 전체 가계대출 총량을 예측하는 것보다 유형 별로 예측할 필요가 있다. 예를 들자면, 주담대는 주택매매가격지수 또는 입주물량이 주요 결정요인일 가능성이 높은 반면, 마통은 상대적으로 취업률이나 경기 등에 민감할 수 있다.

본 연구의 목적은 전체 가계대출을 은행권 주담대, 은행권 마통, 비은행권 주담대, 비은행권 마통 등 유형 별로 구분하여 예측하는 것이며, 이를 위해 본 연구자는 베이지안 머신러닝(Bayesian Machine Learning)을 이용하여 가계대출 유형별 분포예측(density forecast) 알고리즘을 개발하고 실제 예측결과를 산출하고자 한다.

본 연구의 가계부채 예측을 위한 계량 기법은 강규호(2018)가 제시한 베이지안 방법론을 일부 수정 및 보완한 것이다. 따라서 본 연구에서 논의되는 예측기법 상 세부내용은 기본적으로 강규호(2018)의 연구를 차용한 것이며,

수정 또는 보완된 부분은 논의 과정에서 명시적으로 언급할 것이다. 우선 본 연구의 가계부채 예측은 점예측(point forecast)이 아닌 분포예측을 사용하는데, 이는 분포예측이 점예측에 비해 위험관리에 보다 유용하기 때문이다. 구체적으로 설명하자면, 특정 유형의 가계대출 증가율이나 수준이 금융시스템 불안을 야기할 수 있는 임계치가 존재할 경우, 분포예측을 사용하면 예측기간 별로 임계치를 상회할 확률을 산출할 수 있다. 이러한 확률은 금융불안 조기경보지수로도 유용하게 사용될 것으로 기대된다. 참고로 강규호(2018)에서 지적된 바와 같이, 김우영, 김현정(2009), 김승욱, 남영우(2012), 김경아(2011), 김영일, 변동준(2012)와 같은 가계부채 관련 국내 연구는 총량 예측보다는 미시자료를 기반으로 한 가계부채 결정요인에 관한 연구들이다. 또한 놀랍게도 가계부채 예측을 시도한 해외 연구 사례는 저자들이 아직 발견하지 못하였다.

한편 가계대출 예측은 변수와 모형의 불확실성이 높아서 기술적인 어려움이 존재한다. 가계대출의 결정요인이 유형 별로 상이할 뿐만 아니라 잠재적인 설명변수의 수가 대단히 많기 때문이다. 특히 장기와 단기 예측에 유용한 설명변수가 다를 수 있으며, 설명변수 별로 가계대출에 영향을 미치는 시차가 일정하지 않을 수도 있다. 뿐만 아니라, 설명변수별로 가계대출에 미치는 영향이 시점별로 일정하지 않으며, 동일한 설명변수라고 하더라도 직접 예측과 간접 예측의 상대적인 예측력이 예측대상 시점에 따라 변동할 수 있다. 따라서 정확한 예측결과를 얻기 위해서는 예측시계(forecast horizon) 별로 설명변수와 모형이 실시간(online) 최적화되어야 한다. 이런 측면에서 본 연구자는 강규호(2018)가 제시한 베이지안 머신러닝 알고리즘을 수정 및 보완한 후, 이를 적용하여 변수 및 모형 불확실성을 반영한 최적 예측분포를 시뮬레이션하고자 한다.

예측 결과, 은행권 주담대는 주로 대출금리, 아파트 입주물량, 분양물량 등에 의해 예측가능하며, 은행권 마통의 주요 예측변수는 취업률과 전세가격지수이다. 반면, 비은행권 주담대는 대출금리와 아파트 매매전세가비율 등에 의해 주로 결정되며, 비은행권 마통은 취업률, 전세가격지수와 더불어 주가수익률의 영향도 많이 받는 것으로 추정되었다. 각 유형 별 가계대출은 높은 지속성으로 인해 향후에도 현재와 유사한 수준의 증가율을 보일 것으로 예측된다. 한편 정책당국은 예측치에 내재된 불확실성이 상당하다는 점을 반드시 고려하여야 한다. 2017년 5월 현재 545.4조인 은행권 주담대는 향후 1년 내에 575조를 넘어설 가능성이 90.8% 이고, 현재 303조인 비은행권 마통도 6개월 이내에 320조를 상회할 확률이 52.7%에 이르기 때문이다.

본 논문의 구성은 다음과 같다. 2절에서는 예측 모형을 소개하고. 3절에서는 베이지안 머신러닝 알고리즘을 설명한다. 4절은 가계부채 예측결과와 주요

예측변수를 분석한다. 마지막으로 5절은 논문의 전반적인 내용을 요약하고 정책적 시사점에 대해서 논의하였다.

2. 예측 모형

편의상 은행권 주담대를 대상으로 예측모형에 대한 논의를 진행한다. 여타 가계대출에 대한 예측 모형은 은행권 주담대와 동일하다. 우선 $Level_t$는 t 시점의 주담대 수준이며, $y_t = \ln Level_t - \ln Level_{t-12}$은 로그 주담대를 차분하여 계산한 전년동월대비 주담대 증가율이다. 본 연구의 예측 모형은 주택담보대출 증가율을 대상으로 한다. 모형의 추정결과로부터 h기 이후 증가율 예측치, y_{t+h}가 도출되면, $t+h$ 시점 이후 주택담보대출의 예측치는

$$Level_{t+h} = \exp\left(y_{t+h} + \ln Level_{t+h-12}\right)$$

이다.

앞서 언급한 바와 같이 본 연구에서 사용되는 베이지안 머신 러닝 예측 알고리즘은 강규호(2018)에 제안된 기법을 확장한 것이다. 이 장에서는 강규호(2018)의 예측 기법을 설명하고 그 과정에서 본 연구에서 보완된 부분을 언급하고자 한다.

본 연구의 예측 모형은 크게 $AR(p)$, $ADL(p,q)$, $VAR(p)$로 분류된다. 사용되는 예측기법은 모형 군에 따라 다른데, $AR(p)$와 $ADL(p,q)$ 모형은 직접 예측기법에 기반하고, $VAR(p)$ 모형은 간접 예측기법을 통해 예측한다. 여기서 t 시점까지의 정보가 주어져 있을 때 h 기 이후의 $t+h$ 시점을 t 시점까지의 정보만을 이용해서 예측하는 것이 직접 예측이고, $t+1$ 시점부터 시작해서 $t+h$ 시점까지 1기씩 반복적으로 예측하는 것이 간접 예측이다. 예를 들어, $t+h$ 간접 예측시 $(t+h-1)$ 시점의 예측치가 주어진 설명변수로 이용된다.

2.1. 개별 예측 모형

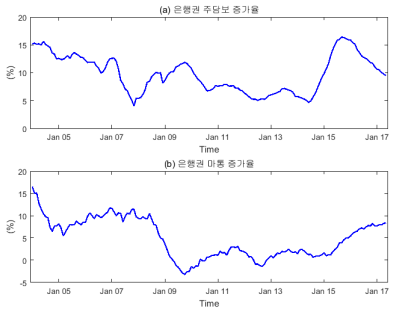
2.1.1 직접예측모형: $AR(p)$와 $ADL(p,q)$

$h(=1,2,\ldots,12)$ 기 직접 예측(direct forecasting)을 위한 $ADL(p,q)$ (autoregressive distributed lag) 모형은 다음과 같이 표현된다.

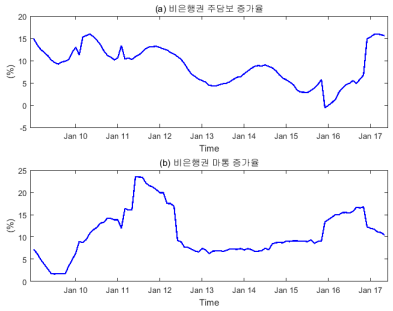
$$y_{t-1+h}|\mathscr{F}_{t-1},\theta \sim \mathscr{N}\left(\mu + \sum_{i=1}^{p} y_{t-i}\beta_i + \sum_{i=1}^{q} x_{t-i}\alpha_i, \sigma^2\right)$$

단, $p \in \{1,2,..,P\}$이고 $q \in \{1,2,..,Q\}$이다. 또한 x_t는 강외생적인(strictly exogenous) 예측변수이고 $\mathscr{F}_t = \{y_i, x_i\}_{i=1}^{t}$는 t 시점까지의 모든 관측자료이다.

1: **은행권 주담대 및 마통 증가율** 이 그림은 2004년 1월부터 2017년 5월까지의 은행권 주담대 및 마통 전년동월대비 증가율을 나타낸 것이다.



2: **비은행권 주담대 및 마통 증가율** 이 그림은 2009년 1월부터 2017년 5월까지의 비은행권 주담대 및 마통 전년동월대비 증가율을 나타낸 것이다.



예측변수는 각 ADL 모형 별로 하나씩만 포함된다. 그림 1과 그림 **??**에서 보여지듯이, 모든 행태의 가계대출이 강한 지속성을 갖기 때문에 자기 시차 외에 두 종류 이상의 예측변수를 사용한 경우가 한 종류의 예측변수를 사용했을 때보다 예측력이 떨어진다. 강규호(2018)의 연구와 큰 차이점 중 하나는 본 연구는 p와 q 모두에 대해서 최적화를 하는 반면, 강규호(2018)의 연구에서는 $p = q$ 라는 제약을 부여한다는 것이다. 예측변수의 최적 시차가 주담보의 최적 시차와 사전적으로 항상 같다고 보기는 힘들며, 실제로 본 연구의 추정결과도 이를 반영하는 것으로 나타났다.

$AR(p)$ 모형은 $ADL(p,q)$ 모형에서 x_t의 시차항을 제거함으로써 단순화된 모형이다.

$$y_{t-1+h}|\mathscr{F}_{t-1},\theta \sim \mathscr{N}\left(\mu + \sum_{i=1}^{p} y_{t-i}\beta_i, \sigma^2\right),\ p = 1,2,\ldots,P.$$

2.1.2 간접예측모형: $VAR(p)$

$VAR(p)$ 모형은 다음과 같이 표현 된다.

$$Y_t|\mathscr{F}_{t-1},\theta \sim \mathscr{N}\left(\mu + \sum_{i=1}^{p} \Phi_i \times (Y_{t-i} - \mu), \Sigma\right),$$

단, $Y_t = (y_t,\ x_t)'$, $p = 1,2,\ldots,P$. Y_t에는 ADL 모형에서와 같이 하나의 예측변수만을 포함한다.

2.1.3 사전분포

각 모형의 오차항의 분산(σ^2) 및 공분산행렬(Σ)에 대해선 무정보 사전분포(non-informative prior or diffuse prior)를 가정한다. $AR(p)$와 $ADL(p,q)$ 모형의 상수항 μ 및 β_i와 α_i에 대해서도 평균이 0이고 분산 4인 정규 분포를 따른다고 가정한다. 단, β_1에 대해서는 주택담보대출 증가율의 높은 지속성을 반영하기 위해 사전평균을 0.8로 하고 사전분산은 0.1로 설정한다. 마찬가지로 1차 자기회귀계수인 Φ_1 의 (1,1) 요소에도 $\mathscr{N}(0.8,0.1)$을 가정한다. 나머지 계수들에 대해서는 자료의 정보가 충분히 반영될 수 있도록 $\mathscr{N}(0,4)$을 사전 분포로 설정하였다.

고려되는 모든 모형에서 종속변수 y_t가 정상과정(stationary process)을 따른다고 가정하고 있으므로 추정 과정에서 안정성 조건을 위한 제약이 필요하다. 예를 들어, $AR(p)$와 $ADL(p,q)$ 모형의 경우에 안정성 조건이 만족되기 위해서는 깁스 샘플링 반복시행에서 행렬

$$\begin{bmatrix} \beta_1\ \beta_2\ \cdots & \beta_p \\ I_{(p-1)} & \mathbf{0}_{(p-1)} \end{bmatrix}$$

의 특성근의 최대 절대값이 1보다 작아야 한다. 구체적으로 깁스 샘플링의 특정 반복 시행에서 β가 샘플링되면 위의 행렬을 만들고 안정성 조건을 만족하는지 여부를 판단한다. 만약 조건을 만족하지 못하면 이전의 반복시행에서 저장된 파라메터 값이 다시 저장된다. $VAR(p)$ 모형의 경우에는 행렬

$$\begin{bmatrix} \Phi_1\ \Phi_2\ \cdots & \Phi_p \\ I_{(p-1)k} & 0_{(p-1)k\times k} \end{bmatrix}$$

의 특성근의 최대 절대값이 1보다 작아야 한다.

2.2. 모형의 구분

전체 모형군은 $AR(p)$, $ADL(p,q)$, $VAR(p)$이며, 시차 p와 q, ADL과 VAR 모형에서 사용되는 예측변수에 따라 모형이 세분화 된다. AR 모형은 시차 p에 의해서만 모형이 구분된다. ADL과 VAR 모형은 시차와 사용되는 예측변수에 의해 모형이 구분된다. 예를 들어, 동일한 시차라도 사용되는 예측변수가 다르면 서로 다른 모형이고, 반대로 사용하는 예측변수가 같아도 시차가 다르다면 서로 다른 모형이다. 따라서, K를 잠재적으로 가능한 예측변수의 수라고 할 때, 베이지만 머신러닝 알고리즘의 변수선택과 모형선택을 위해 추정해야 하는 모형의 수는 예측시계(h)별로 AR 모형이 P개, ADL 모형이 $Q \times K$개, VAR 모형이 $P \times K$개이다. 본 연구의 최대 예측시계는 12개월이므로 결국 비교 대상 모형의 총 수가

$$12 \times (P(1+K)+QK)$$

개가 되는데, 예를 들어, 예측변수의 수가 30개이고 최대 시차가 6이라면 총 모형의 수는 무려 4,392개이다.

변수선택과 모형선택 단계에서 예측시계별 최적시차를 찾고, 불필요한 예측변수를 제외함으로써 모형의 수가 축소된다. 예를 들어, AR 모형의 경우 h별로 P개의 모형이 존재하는데($H \times P$개) 모형선택 단계에서 각 예측시계별 최적시차를 찾으면 고려해야할 모형의 수가 최대 예측시계의 크기인 H개로 축소된다. ADL과 VAR 모형의 경우에는 각 $H \times QK$와 $H \times PK$개에서 $H \times$(선택된 변수의 수)로 축소된다. 따라서, 머신 러닝 알고리즘의 예측조합 단계에서 예측시계별로 예측조합에 사용되는 모형의 수는 $(1+2\times$(선택된변수의 수)) 개로 축소된다. 즉, 머신 러닝 예측분포는 예측시계별로 $(1+2\times$(선택된변수의 수))개의 모형으로부터 도출된 예측분포를 조합하여 유도된다.

2.3. 표본 외 예측

본 연구의 주 관심은 베이지안 머신 러닝 알고리즘을 통해 모형 별 예측력을 평가한 후(변수선택, 모형선택) 모형과 예측변수의 차원을 축소하고, 선택된 모형의 예측력을 기준으로 가중치를 설정하여 예측조합을 하는 것이다. 사후적 예측력 평가와 베이지안 머신 러닝 알고리즘에 대해 설명하기에 앞서 위의 모형들에 기반해서 표본 외 예측 실험과 실제 예측이 예측시계별로 어떻게 실시되는지를 설명한다.

전년동월대비 주담대 증가율 자료의 표본크기가 T이고 표본 외 예측구간의 크기(out-of-sample size)를 OSS라고 할 때, $T_0 = T - OSS$를 training 표본

의 크기로 정의한다. 표본 외 예측에서는 각 모형별로 $T_0+1, T_0+2, \ldots, T(=T_0+OSS)$ 시점에 대해서 예측을 실시한다. 실제 예측에서는 T 시점 이후, $h=1,2,\ldots,H$에 대해 $T+h$ 시점의 주담보 증가율을 예측한다. 여기서 주의할 것은 표본 외 예측과 실제예측에서 예측시계(h)의 역할이다. 표본 외 예측에서 모든 h에 대해 모든 모형의 표본 외 예측시점은 항상 동일하다. 단, h는 표본 외 예측에서 사용되는 정보의 양에 영향을 준다. 다시 말해, 항상 동일한 시점을 예측하지만 예측시계에 따라 사용되는 정보의 양이 달라지는 것이다. 예를 들어, 직접 예측에 기반하는 $AR(p)$ 모형과 $ADL(p,q)$ 모형에 대해 T_0+1 시점을 예측한다고 하자. $h=1$인 경우에는 예측모형에 따라 T_0 시점까지의 정보를 이용하여 예측을 실시한다. 따라서, y_{T_0+1}의 예측분포는 다음과 같다.

$$y_{T_0+1}|\mathscr{F}_{T_0}$$

반면, $h=2$인 경우에는 T_0-1 시점까지의 정보를 이용해서 T_0+1 시점을 예측한다. 따라서 예측분포는

$$y_{T_0+1}|\mathscr{F}_{T_0-1}$$

과 같다. 이와 같이 우리는 표본 외 예측에서 예측시계에 관계없이 동일한 시점을 예측하지만 h는 사용되는 정보의 양에 영향을 주며 특히, h가 클수록 정보의 양이 줄어든다.

한번의 표본 외 예측이 완료되면 자료를 하나씩 늘려서 동일한 방법으로 T_0+2 시점을 예측한다. 예를 들어, $h=1$ 이면 T_0+1까지의 자료를 사용하고 $h=2$이면 T_0까지의 자료를 이용해서 T_0+2 시점을 예측한다. 이 과정을 총 OSS번 반복함으로써 표본 외 예측이 완료된다.

다음으로 VAR 모형을 이용한 간접예측 과정을 설명한다. 예를 들어, T_0+1 시점을 대상으로 $h=2$기 이후 예측하는 경우를 고려해보자. 이 경우, T_0-1 시점까지의 자료를 이용하여 1기 이후인 T_0 시점을 예측한 다음, 이 예측치를 기반으로 T_0+1을 예측한다. 마찬가지로 T_0+1 시점이 $h=3$기 이후 예측대상인 경우, T_0-2 시점까지의 자료를 이용해서 T_0-1, T_0, T_0+1 시점을 순차적으로 예측하고, 마지막 시점인 T_0+1의 예측결과만 취한다. 일반적으로 표현하면, T_0에서 $(h-1)$만큼의 자료를 제외한 뒤에, T_0-h+2, T_0-h+3, .., T_0+1 시점까지 간접예측하는 것이다. 이를 모든 $h=1,2,\ldots,H$에 대해서 반복함으로써 T_0+1 시점을 여러 예측시계에 걸쳐서 예측할 수 있다. 이러한 과정을 T_0+1 시점뿐 만 아니라 나머지 표본외 시점인 T_0+2, T_0+3, T_0+4, .., T에 대해서 동일하게 반복하면 표본외 예측이 완료된다.

2.4. 표본 외 예측력 평가

변수선택과 모형선택 단계에서 어떤 변수를 선택할 것인지는 표본 외 예측력을 기준으로 한다. 표본 외 예측력은 베이지안 방법론에서 가장 표준적으로 사용되는 사후예측우도(posterior predictive likelihood, *PPL*)에 근거한다. 사후예측우도는 값이 클수록 더 좋은 예측력을 의미한다. 변수선택 단계에서는 *P*개의 *AR* 모형 중에서 가장 큰 1기 예측($h=1$) 로그 *PPL*값을 기준으로 예측변수를 선택한다. 즉, 1기 예측에서 특정 예측변수를 사용한 *ADL* 모형이 예측력이 가장 좋은 최적시차 *AR* 모형보다 뛰어난 예측력을 가지고 있어야 예측변수로 선택될 수 있는 조건을 만족하는 것이다.

모형선택 단계에서는 1기 예측 뿐만 아니라 모든 예측시계($h=1,2,\ldots,H$)와 시차($p=1,2,\ldots,P$), 선택된 예측변수에 대해 $AR(p), ADL(p,q), VAR(p)$ 모형을 추정하고 표본 외 예측을 실시하여 로그 *PPL*을 계산한다. 각 모형의 로그 *PPL*은 예측시계별 예측조합 가중치를 계산하는데 사용된다. 다시 말해, 각각의 예측시계 별로 표본 외 예측력이 더 좋은 모형에 더 큰 가중치가 부여되는 것이다. 각 인덱스(*h*, *p*, 선택된 예측변수)별로 다양한 경우의 수가 생기는데, 시차는 최적시차 하나로 축소시킨다. 예를 들어, 예측시계와 예측변수가 같은 경우 *ADL*과 *VAR* 모형은 시차에 의해 모형이 구분되는데 모든 시차에 대한 로그 *PPL* 값을 저장하지 않고, 예측력이 가장 좋은 모형의 시차(최적시차)와 그때의 로그 *PPL*값을 저장한다. 로그 *PPL*은 표본 외 예측구간에 대한 1기 예측 로그 사후예측밀도(posterior predictive density, *PPD*)의 합으로 정의되며, 모형(M)의 로그 *PPL*은

$$\ln PPL(M,h=1)=\sum_{t=T_0+1}^{T}\ln f(y_t|\mathscr{F}_{t-1},M)$$

으로 계산된다. $\ln f(y_t|\mathscr{F}_{t-1},M)$가 t 시점의 *PPD*이며, 이 값은 특정 표본 외 예측구간 시점의 예측분포에 그 시점의 실현치를 대입하여 계산한다. *PPD*는 특정 표본 외 예측구간 시점의 예측분포에 그 시점의 실현치를 대입하여 계산한다. 만약 실현치가 예측분포에서 높은 빈도로 나타나는 값이라면 *PPD*값이 더 클 것이다. 따라서, 표본 외 예측구간에 속하는 시점별 로그 *PPD*을 모두 합하면 모형별 로그 *PPL*이되고, 이 값이 클수록 해당 모형이 상대적으로 높은 분포예측력을 나타낸다.

2.4.1 모형별 *PPD* 계산

$\pi(\theta|\mathscr{F}_{t-1},M)$가 모형별 모수의 사후밀도라고 했을 때, *PPD*는 정의 상

$$f(y_{t-1+h}|\mathscr{F}_{t-1},M)=\int f(y_{t-1+h}|\mathscr{F}_{t-1},\theta,M)\pi(\theta|\mathscr{F}_{t-1},M)d\theta$$

으로 계산되지만, 적분이 해석적으로 계산되지 않는 경우가 대부분이다. 이 때문에 다음과 같이 수치적으로 근사한다.

$$f(y_{t-1+h}|\mathscr{F}_{t-1},M) = \frac{1}{n_1}\sum_{j=1}^{n_1} f(y_{t-1+h}|\mathscr{F}_{t-1},\theta^{(j)},M) \tag{1}$$

단, $\theta^{(j)} \sim \theta|\mathscr{F}_{t-1},M$는 사후 분포에서 추출된 샘플이며, n_1은 burn-in을 제외한 시뮬레이션의 크기이다.

AR과 ADL 모형 *AR*과 *ADL* 모형의 파라메터들은 모든 선형회귀모형의 깁스 샘플링 방법으로 간단하게 샘플링될 수 있으므로 구체적인 추정방법은 생략한다. $AR(p)$ 모형의 *PPD*에서 모수가 주어졌을 때, y_{t-1+h}의 밀도는

$$f(y_{t-1+h}|\mathscr{F}_{t-1},\theta^{(j)},M) = \mathscr{N}\left(y_{t-1+h}|\mu^{(j)} + \sum_{i=1}^{p} y_{t-i}\beta_i^{(j)}, \sigma^{2(j)}\right)$$

이다. 마찬가지로 $ADL(p,q)$ 모형의 경우에는

$$f(y_{t-1+h}|\mathscr{F}_{t-1},\theta^{(j)},M) = \mathscr{N}\left(y_{t-1+h}|\mu^{(j)} + \sum_{i=1}^{p} y_{t-i}\beta_i^{(j)} + \sum_{i=1}^{q} y_{t-i}\alpha_i^{(j)}, \sigma^{2(j)}\right)$$

이다. 단, $(\mu^{(j)}, \{\beta_i^{(j)}\}_{i=1}^{p}, \{\alpha_i^{(j)}\}_{i=1}^{q}, \sigma^{2(j)})$는 j번째 반복시행에서 추출된 사후 샘플이다.

VAR 모형 *VAR* 모형에서도 위의 두 모형과 마찬가지로 $\mathscr{F}_{t-1},\theta,M$이 주어져 있을 때, y_{t-1+h}의 조건부 분포는 정규분포이다. 따라서, *PPD*를 계산하기 위해서는 조건부 평균과 분산을 알아야 한다. 이를 계산하기 위해 먼저 *VAR* 모형을 상태 공간 형태로 변환한다. $VAR(3)$ 모형을 예로 설명하면 $VAR(3)$ 모형은

$$\begin{bmatrix} Y_t-\mu \\ Y_{t-1}-\mu \\ Y_{t-2}-\mu \end{bmatrix} = \begin{bmatrix} \Phi_1 & \Phi_2 & \Phi_3 \\ I_k & 0 & 0 \\ 0 & I_k & 0 \end{bmatrix} \begin{bmatrix} Y_{t-1}-\mu \\ Y_{t-2}-\mu \\ Y_{t-3}-\mu \end{bmatrix} + \begin{bmatrix} u_t \\ 0 \\ 0 \end{bmatrix}$$

로 표현될 수 있다. 위 식을 다시

$$\hat{Y}_t = G\hat{Y}_{t-1} + \hat{u}_t$$

로 표현하면 조건부 평균 $\mathbb{E}\left(\hat{Y}_t|\mathscr{F}_{t-1},\theta,M\right) = G\hat{Y}_{t-1}$이며, 조건부 분산-공분산 행렬

$$Var(\hat{Y}_t|\mathscr{F}_{t-1},\theta,M) = \begin{bmatrix} \Sigma & 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & 0_{k\times k} & 0_{k\times k} \end{bmatrix}$$

와 같다. 따라서, $h=1$일 때, 조건부 평균 $\mathbb{E}(y_t|\mathscr{F}_{t-1},\theta,M) = (1,1)\ element\ of\ G\hat{Y}_{t-1} + \mu_1$과 같다. 또한, 조건부 분산

$$Var(y_t|\mathscr{F}_{t-1},\theta,M) = (1,1)\ element\ of \begin{bmatrix} \Sigma & 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & 0_{k\times k} & 0_{k\times k} \end{bmatrix}$$

로 계산된다. $h=2$인 경우에는 위와 동일한 방법으로 $\hat{Y}_t = G\hat{Y}_{t-1} + \hat{u}_t$ 으로부터 조건부 평균과 분산은 각 각

$$\begin{aligned} & \mathbb{E}(y_{t+1}|\mathscr{F}_{t-1},\theta,M) \\ = \ & (1,1)\ element\ of\ G^2\hat{Y}_{t-1} + \mu_1, \\ & Var(y_{t+1}|\mathscr{F}_{t-1},\theta,M) \\ = \ & (1,1)\ element\ of\ G \begin{bmatrix} \Sigma & 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & 0_{k\times k} & 0_{k\times k} \end{bmatrix} G' + \begin{bmatrix} \Sigma & 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & 0_{k\times k} & 0_{k\times k} \end{bmatrix} \end{aligned}$$

로 계산된다. 따라서, 일반적으로 임의의 h에 대해 조건부 기대값과 분산은 각각

$$\begin{aligned} \mathbb{E}(y_{t-1+h}|\mathscr{F}_{t-1},\theta,M) &= (1,1)\ element\ of\ G^h\hat{Y}_{t-1} + \mu_1 \\ Var(y_{t-1+h}|\mathscr{F}_{t-1},\theta,M) &= (1,1)\ element\ of\ \textstyle\sum_{i=0}^{h-1}\Sigma_i \end{aligned}$$

이다. 단,

$$\Sigma_i = G^i \begin{bmatrix} \Sigma & 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & 0_{k\times k} & 0_{k\times k} \end{bmatrix} G^{i\prime}$$

는 $(i+1)_{i=0,1,..,h-1}$기 이후 충격의 분산이다.

3. 베이지안 머신 러닝 알고리즘

베이지안 방법론으로부터 유도되는 예측분포는 모형 불확실성과 파라메터 불확실성을 반영한다.

본 연구에서 주택담보대출 예측분포를 도출하기 위한 베이지안 머신 러닝 알고리즘은 (1) 변수선택, (2) 모형선택, (3) 예측조합의 3단계로 구성된다.

변수선택과 모형선택은 잠재적으로 가능한 수많은 경우의 수에 대한 차원 축소를 위한 것이며, 변수 및 모형 학습의 결과를 이용하여 예측조합을 통해 미래 시점의 예측분포를 유도한다.

위의 3단계로 이루어진 알고리즘을 N번 반복해서 미래의 모든 예측시계에 대한 예측분포를 유도한다. 구체적으로 표본크기가 T인 경우, T까지의 표본을 사용하여 머신 러닝 알고리즘을 실시한다. 한번의 알고리즘이 완료되면 $T-1$까지의 표본만을 이용하여 동일한 알고리즘을 다시 실시한다. 즉, 표본의 크기가 $T-N$이 될 때까지 동일한 알고리즘을 N번 반복하는 것이다. 이렇게 반복적으로 미래의 예측시계를 다르게 하여 반복해서 예측하는 것은 베이지안 머신 러닝의 예측력을 평가하기 위한 것이다. 특정 예측시계에 대해 머신 러닝의 예측력이 좋더라도 또 다른 예측시계에 대해서는 예측력이 떨어질 수 있다. 따라서, 정확한 예측력 평가를 위해서는 다양한 예측시계에 대한 예측을 통한 비교 평가가 필요하다.

우리의 주 관심은 실현되지 않은 미래 시점들에 대한 예측이지만, 미래의 예측은 현 시점에서 예측력을 평가할 근거가 없기 때문에 표본 외 예측을 실시한다. 구체적으로 우리에게 주어진 자료의 표본크기가 $T+12$라고 할 때, 마지막 12개를 제외한 T까지의 표본만을 사용하여 알고리즘을 실시하고 이 과정에서 마지막 12개의 종속변수는 실현되지 않은 것으로 간주한다. 최종적으로 예측이 이루어진 뒤 마지막 12개 시점의 종속변수 실현치를 이용하여 예측력을 평가하는 것이다.

포괄적으로 3 단계로 구성된 베이지안 머신 러닝 알고리즘은 세부적으로 6 단계로 구분하여 실시한다.

[변수선택]

1 단계: 주어진 자료에 대해서 training 표본, 표본 외 예측구간, 최대 예측시계(H), 최대 시차(P, Q)의 크기를 설정한다.

2 단계: 모든 예측시계($h=1,2,\ldots,H$)에 대한 $AR(p)$ 모형의 로그 PPL 값을 계산하여 저장한다. 각 예측시계별로 P개 로그 PPL 값이 계산되는데 이 중에서 가장 큰 값을 저장하고, 그에 해당하는 시차를 각 예측시계별 AR 최적시차로 저장한다.

3 단계: 고려되는 모든 설명변수에 대해서 $ADL(p,q)$ 모형을 추정하고, 1기 예측($h=1$) 로그 PPL의 최대값을 계산하여 저장한다. 여기서 p는 $h=1$일 때, AR 최적시차로 고정시킨다. 따라서, 설명변수에 따른 모형별로 총 Q개의 로그 PPL값이 계산되며 그 중에서 가장 큰 값을 저장한다. 이후, 2

단계에서 저장한 *AR* 최적시차 모형의 로그 *PPL*값과 비교하여 더 큰 값을 갖는 모형의 설명변수를 선택하고 나머지 변수는 제외한다. 즉, 표본 외 예측력이 최소한 *AR* 모형보다 좋은 변수들만 선택한다.

[모형선택]

4 단계: 3 단계에서 선택된 예측변수들만 이용하여 모든 예측시계($h = 1, 2, \ldots, H$)에 대해 $ADL(p,q)$ 모형과 $VAR(p)$ 모형을 추정하고 로그 *PPL*값과 최적시차를 저장한다. $ADL(p,q)$ 모형의 경우에 시차 p는 각 예측시계별 *AR* 최적시차로 고정시키고 예측변수의 시차 Q에 대해서만 추정한다. 여기서 각 예측시계에 대해서 모형별, 시차별로 상이한 로그 *PPL*값이 계산되는데, 가장 큰 값의 로그 *PPL*과 그에 해당하는 최적시차를 저장한다. 또한, 최적시차에 해당하는 모형별 실제 예측분포를 저장한다.

5 단계: 각 예측시계에 대해 앞서 계산된 *AR*, *ADL*, *VAR* 모형들의 로그 *PPL* 값을 이용하여 모형별 가중치를 계산한다. 즉, 표본 외 예측력이 좋을수록 더 큰 가중치를 부여하는 것이다. 특정 예측시계 h에 대한 모형별 가중치는 각 모형의 $PPL(M,h)$ 를 모든 모형의 *PPL*값들의 합으로 나눈 것이다.

[예측조합]

6 단계: 각각의 예측시계에 대해서 모형별 예측분포로부터 가중치만큼의 예측치들을 예측조합 분포로 할당한다. 예를 들어, $h = 1$일 때, *AR* 최적시차 모형의 가중치가 0.1이고, 예측분포의 크기가 n_1이라면 *AR* 최적시차 모형의 예측분포에서 $n_1 \times 0.1$개를 예측조합 분포로 할당한다. 이와 같은 방법이 Eklund and Karlsson(2007)이 제시한 조건부 베이지안 모형평균(conditional Bayesian model averaging)의 개념이다.

4. 예측 결과

표 8은 예측에 사용된 설명변수의 목록이다. 은행권 가계대출의 표본기간은 2003년 1월부터 2017년 5월까지이며, 비은행권은 자료추계 시작 시점인 2008년 1 월부터 2017 년 5월이다. 표 1는 유형별 가계대출의 기초통계량이며, 은행권 주담대를 제외한 나머지 세 변수는 5% 유의수준에서 단위근이 존재한다는 귀무가설을 기각하지 못하였다. 자료의 강한 지속성에도 불구하고 본 연구에서는 안정계열을 가정한다. 우선은 높은 지속성으로 인해 가계대출 순환이 명확히 관찰될 만큼 시계열이 충분히 길지 않다고 판단되기 때문이다.

또한 단위근을 가정하게 되면 공적분에 대한 고려가 필요하며, 이로 인해 오차수정모형이 추가되어 추정대상 모형의 수가 지나치게 많아질 수 있는 우려가 있기 때문이다. 마지막으로 강규호(2018)의 연구에서와 같이 LTV, DTI 규제는 규제 대상 범위나 대상이 시기 별로 다르기 때문에 강화시기는 1, 완화시기는 0이라는 더미로 처리한다.

1: **가계대출 유형별 기초 통계량**

	은행권 주담대	은행권 마통	비은행권 주담대	비은행권 마통
평균값	10.36	5.06	9.32	11.45
중위값	10.42	3.99	9.46	9.51
최대값	17.82	17.96	17,32	26.55
최소값	4.10	-3.14	-0.61	1.69
표준편차	3.70	4.67	4.55	5.93

4.1. 은행권 주담대 예측

먼저 은행권 주담대 예측 결과부터 논의하도록 한다.[1] 그림 3은 2017년 6 월부터 2018년 5월까지의 은행권 주담대 전년동월대비 증가율 및 수준을 예측한 결과이다. 은행권 주담대 증가는 9% 내외로 지속될 것으로 추정된다. 예측분포 추정결과로부터 표 2은 향후 12개월 동안 주담대 수준이 임의의 임계치를 상회할 가능성을 나타낸 것이다. 현재 545.4조인 은행권 주담대는 향후 6개월 이내에 91.6% 의 확률로 550 조를 넘어설 것이며, 1년 내에 575조를 상회를 가능성이 90.8% 이다.

그림 4는 본 연구의 머신 러닝 알고리즘을 이용한 최근 12 개월(2016년 6 월-2017년 5월) 은행권 주담대에 대한 표본외 예측결과이다. 그림 5는 2015 년 12월부터 2016년 11월의 주담대를 2015년 11 월까지를 사용하여 예측한 결과이다. 이 두 표본 외 예측결과의 의하면 본 연구의 예측이 상당히 정확한 것으로 판단된다. 실현치와 예측치 사이의 거리가 아주 좁음에도 불구하고 신용구간의 폭은 상당히 넓어보인다. 이는 실현치가 신용구간의 상단 또는 하단에 위치할 가능성을 반영하기 때문이다. 그림 6 의 결과에서 보듯이 2015 년 6월-2016년 5월 기간의 경우, 주담대가 예상 외로 상승하면서 신용구간의

[1] 베이지안 머신러닝을 통한 예측분포 계산 전체 과정은 2시간 30분 내외의 시간이 요구된다. 이 때 사용된 컴퓨터의 주요사양은 Intel(R) Core i7-3690OX CPU @3.30GHz, 64비트, 16GB 램이다.

2: **은행권 가계대출 위험예측: 2017년 6월 - 2018년 5월** 이 표는 각각의 예측시계에 대해 가계대출이 임계치를 상회할 확률을 나타낸 것이며, 단위는 (%)이다.

	은행권 주담보				은행권 마통			
예측시계	임계치(조)				임계치(조)			
(h)	550	575	600	650	180	190	200	210
1	46.7	0.0	0.0	0.0	56.2	0.0	0.0	0.0
2	62.2	0.6	0.0	0.0	59.1	0.0	0.0	0.0
3	68.7	4.3	0.2	0.0	72.6	0.2	0.0	0.0
4	78.3	13.1	0.3	0.0	75.6	0.9	0.0	0.0
5	84.4	26.3	1.1	0.0	87.1	6.9	0.0	0.0
6	91.6	43.5	4.0	0.0	95.9	25.5	0.0	0.0
7	93.9	53.9	7.4	0.0	94.2	26.2	0.2	0.0
8	94.5	56.9	10.7	0.0	90.7	22.3	0.2	0.0
9	94.2	61.9	16.2	0.0	92.4	31.5	0.8	0.0
10	96.2	70.1	22.6	0.1	92.2	36.2	1.4	0.0
11	98.9	85.3	34.2	0.1	94.8	48.7	3.2	0.0
12	99.7	90.8	47.9	0.3	98.7	76.3	12.2	0.1

상단에서 실현될 수 있다. 따라서 본 연구의 예측 분포는 주담대의 예측치와 더불어 불확실성을 적절히 산출하고 있다고 볼 수 있다.

4.2. 은행권 마통 예측

다음으로 은행권 마통에 대한 예측 결과를 살펴보도록 하자. 그림 7 이 2017년 6월 이후 12개월간 마통 예측분포 추정결과이며, 위험에 대한 예측 결과는 표 2에서 볼 수 있다. 마통은 전년동월대비 7 내지 8% 내외에서 증가할 것으로 추정되었다. 그러나 신용구간의 폭이 예측시계에 따라 증가하면서, 올 연말의 증가율은 낮으면 2.5%, 높으면 11% 에 가까울 것으로 예측된다. 특히 2017년 5월 현재 178.5조인 은행권 마통은 올해 말에 190조를 상회할 확율이 20% 이상이다.

은행권 주담대와 마찬가지로 최근 12개월과 2015년 12월-2016년 11월기간의 은행권 마통 표본외 점예측 결과는 상당히 정확한 반면, 2015 년 6월-2016 년 5월 기간의 실현치는 사후 평균과 상당한 격차를 두고 있다.(그림 8, 9, 10) 그럼에도 신용구간이 충분히 넓게 추정되어 마통의 증가율 및 수준이 신용구간 내에서 실현되었다.

4.3. 비은행권 주담대 예측

2016년 말부터 급증한 비은행권 주담대 증가율은 15% 에서 차츰 하락하여 1년 뒤에는 10% 이하로 감소할 것으로 예상된다.(그림 11) 표 3에 있는 위험예측 결과에 의하면, 2017년 5월 현재 159.6조인 비은행권 주담대가 올해 중에 10조 가량 증가할 가능성이 90% 에 이른다. 또한 180조를 돌파할 가능성도 18.7% 이다.

3: **비은행권 가계대출 위험예측: 2017년 6월 - 2018년 5월** 이 표는 각각의 예측시계에 대해 가계대출이 임계치를 상회할 확률을 나타낸 것이며, 단위는 (%)이다.

	비은행권 주담보				비은행권 마통			
예측시계	임계치(조)				임계치(조)			
(h)	170	180	190	200	310	320	330	350
1	0.0	0.0	0.0	0.0	20.2	0.2	0.0	0.0
2	1.0	0.0	0.0	0.0	43.3	4.5	0.1	0.0
3	6.8	0.0	0.0	0.0	50.8	16.4	2.0	0.0
4	7.5	0.1	0.0	0.0	62.8	28.5	7.4	0.0
5	22.2	0.5	0.0	0.0	72.0	43.1	17.4	1.1
6	43.7	2.4	0.1	0.0	77.2	52.7	26.1	2.9
7	68.8	10.0	0.3	0.0	81.3	61.0	36.1	6.9
8	77.0	18.7	1.2	0.0	82.5	63.2	40.1	7.9
9	85.2	28.1	2.7	0.0	84.4	68.0	46.7	11.5
10	88.6	35.3	4.4	0.1	85.6	70.3	50.0	13.2
11	89.5	42.6	6.8	0.5	92.4	82.1	62.6	20.0
12	78.7	35.4	8.0	1.1	95.6	87.6	71.9	25.4

그림 12는 2016년 6월-2017년 5월 기간의 비은행권 주대담 표본외 예측실험 결과이다. 단기적으로는 상당히 정확한 예측이 이루어졌으나, 2016 년 12월 이후 주담대가 급등하면서 실현치가 신용구간을 벗어날 정도로 예측오차가 커졌다. 이와 같은 단기적 급등세는 빈번히 또는 주기적으로 반복되는 현상이 아니기 때문에 본 연구의 머신 러닝 기법으로 예측하기가 어렵다는 한계가 있다. 특히 급락과 비교하여 급등세는 거시건정성 측면에서 더 큰 위험요인이 될 수 있다는 측면에서 정책당국은 기술적 분석 결과에 내포된 불확실성을 항상 고려해야 한다.

4.4. 비은행권 마통 예측

마지막으로 비은행권 마통의 예측결과를 그림 13을 통해 분석하도록 하자. 예측결과, 마통은 향후 12개월 동안은 전년동월대비 10% 수준으로 증가할 것

으로 추정된다. 2017년 12월 증가율 예측 신용구간이 음의 값을 포함할 정도로 신용구간이 넓게 추정되었는데, 이는 마통의 증가율 자체가 대단히 지속성이 강하면서도 동시에 변동성 또한 높기 때문이다. 이로 인해 현재 303조인 비은행권 마통이 6개월이내에 320조를 넘어설 확률이 52.7%로 추정된다. 또한 1년 내에 350조를 상회할 확률 또한 25% 에 해당한다. 특히 은행권 마통에 비해서 비은행권 마통은 대출규모가 크고 예측의 불확실성까지 높아서 여타 가계대출에 비해 보다 면밀한 모니터링이 요구된다. 그림 14에서 보여지는 바와 같이, 비록 최근 12개월을 대상으로 한 표본외 예측결과는 상당히 정확한 것으로 보일 수도 있으나, 넓은 신용구간의 폭을 감안하면 예외적인 상황으로 받아들여야 한다.

4.5. 가계대출 주요 예측변수

총 40개의 예측변수집합과 예측 모형 중 대출유형 및 예측 시계별로 학습된 변수와 모형를 설명하고자 한다. 우선 2017년 6월 이후 가계대출 예측을 위한 변수와 모형에 최근 12개월 동안의 평균적인 예측력을 기준으로 계산된 가중치가 부여된다. 표 4와 5는 각각 은행권과 비은행권 대출의 주요결정요인을 나타낸 것이다. 은행권 주담대의 경우 단기예측에는 주택담보대출금리, 장기예측은 아파트 분양물량이 주요 결정요인인 것으로 나타났다. 반면 은행권 마통은 아파트 및 주택 매매가격지수를 이용한 ADL 모형의 예측력이 높았다. 비은행권 주담대는 은행권 주담대와 달리 입주물량을 이용한 VAR모형이 장단기예측에 우월하였으며 마통은 취업률과 전세가격지수가 주요 예측변수 역할을 하였다. 결과적으로 가계대출 행태에 따라서 예측력 향상에 기여하는 예측변수가 상이하다는 것을 알 수 있다.

뿐만 아니라, 동일한 대출 유형이더라도 시기에 따라서 예측변수가 다를 수 있다. 표 6와 7은 최근 12개월 표본외 예측시 선택된 주요 예측변수를 나타낸 것다. 표 4과 6을 비교했을 때, 은행권 주담대 단기예측에는 주택담보금리가 공통적인 주요 예측변수인 것으로 보인다. 하지만 장기 예측의 경우에는 각각 아파트 입주물량과 분양물량이 선택되었다. 은행권 마통의 경우에도 시기에 따라 아파트 매매가격지수가 중요하기도 하고, 소비자물가지수나 금리가 주요 예측변수이기도 하다.

예측변수와 더불어 예측방법도 시기와 대출 유형별로 상이하다는 점도 주목할 만하다. 표 4-5에 의하면, 2017년 6월-2018년 5월 간 은행권 주담대와 마통, 비은행권 마통 예측에는 직접예측이 우월한 반면, 비은행권 주담대는 간접예측이 우월하였다. 그러나 표 6-7에서 나타나듯이. 2016년 6월-2017년 5월 간 비은행권 주담대 장기예측에는 직접예측이 우월하였으며, 비은행권 마통

4: **은행권 가계대출 결정요인: 2017년 6월 - 2018년 5월 예측** 이 표는 사후 분포 예측시 표본외 예측을 통해 선택된 예측변수 중 가장 큰 가중치가 부여된 예측변수와 해당 예측 모형을 나타낸 것이다. (간접)은 VAR 모형, (직접)은 ADL 모형을 의미한다.

	은행권 주담대	은행권 마통
	선택된 예측변수의 수=11개	선택된 예측변수의 수=4개
예측시계	주요 예측변수와 모형	주요 예측변수와 모형
1	주택담보대출금리(직접)	아파트 및 주택 매매가격지수(직접)
2	주택담보대출금리(직접)	아파트 및 주택 매매가격지수(직접)
3	주택담보대출금리(직접)	아파트 및 주택 매매가격지수(직접)
4	아파트매매가격지수(직접)	아파트 및 주택 매매가격지수(직접)
5	아파트매매가격지수(직접)	아파트 및 주택 매매가격지수(직접)
6	아파트매매가격지수(직접)	아파트 및 주택 매매가격지수(직접)
7	주택 및 아파트매매가격지수(직접)	아파트 및 주택 매매가격지수(직접)
8	아파트 입주물량(전국, 직접)	아파트 및 주택 매매가격지수(직접)
9	아파트 입주물량(전국, 직접)	미국 채권금리(6개월물, 직접)
10	아파트 입주물량(전국, 직접)	미국 채권금리(6개월물, 직접)
11	아파트 입주물량(전국, 직접)	미국 채권금리(6개월물, 직접)
12	아파트 입주물량(전국, 직접)	미국 채권금리(6개월물, 직접)

단기예측에는 간접예측이 선호되었다.

비은행권 주담대의 경우, 최근에는 아파트 입주물량이 주요 결정요인인 반면, 표본외 예측에서는 주택담보대출금리 또는 아파트 매매전세가 비율이 선택되었다. 이는 아마도 최근 문제가 되고 있는 갭투자의 영향을 반영한 결과로도 볼 수 있을 것이다. 비은행권 마통은 전반적으로 주식시장, 취업률, 전세가격지수, 주택 매매가격지수 등에 의해서 예측력 향상이 가능한 것으로 나타났다. 이를 통해 주택 가격 및 전세가 상승 등 부동산 시장의 상황 변화가 비은행권 주담대 뿐만 아니라 마통에도 파급될 수 있음을 알수 있다.

향후 은행권 주담대를 중심으로 10% 내외의 가계부채 증가율이 예상되는 상황에서 최근 정부의 재건축 초과이익 환수제, 다주택자 양도세 중과 시행, 투기과열지구, 투기지역, 조정대상지역별 차등화된 LTV, DTI 규제 강화, 부동산 세제 개편 등을 통한 실수요자 중심 시장형성 정책기조는 바람직한 방향이라고 할 수 있다. 특히 동일한 분양 또는 입주물량이더라도 지방에 비해 수도권 지역의 중도금대출액이 훨씬 더 크기 때문에 구로구, 금천구, 동작구 등 14개 구와 과천을 투기과열지구로, 강남구, 서초구 등 11개 구와 세종시를 투기지역으로 지정하여 분양권 전매제한, 조합원 지위 양도 금지, 자금조달계획 신고의무 등을 부과하는 것은 일부 효과가 있을 것으로 기대된다. 하지만 이러한

5: **비은행권 가계대출 결정요인: 2017년 6월 - 2018년 5월 예측** 이 표는 사후 분포 예측시 표본외 예측을 통해 선택된 예측변수 중 가장 큰 가중치가 부여된 예측변수와 해당 예측 모형을 나타낸 것이다. (간접)은 VAR 모형, (직접)은 ADL 모형을 의미한다.

	비은행권 주담대	비은행권 마통
	선택된 예측변수의 수=11개	선택된 예측변수의 수=19개
예측시계	주요 예측변수와 모형	주요 예측변수와 모형
1	아파트 입주물량(간접)	산업생산지수(직접)
2	아파트 입주물량(간접)	종합주가지수 및 산업생산지수(직접)
3	아파트 입주물량(간접)	종합주가지수(직접)
4	아파트 입주물량(간접)	취업율(직접)
5	아파트 분양물량, 아파트 입주물량(간접)	아파트 전세가격지수 및 취업율(직접)
6	아파트 분양물량, 아파트 입주물량(간접)	아파트 전세가격지수(직접)
7	아파트 입주물량(간접)	취업율(직접)
8	아파트 입주물량(간접)	아파트 전세가격지수 및 취업율(직접)
9	아파트 입주물량(간접)	아파트 전세가격지수 및 취업율(직접)
10	아파트 입주물량(간접)	아파트 전세가격지수 및 취업율(직접)
11	아파트 입주물량(간접)	아파트 전세가격지수 및 취업율(직접)
12	아파트 입주물량(간접)	아파트 전세가격지수 및 취업율(직접)

정책의 효과를 정량적으로 판단하거나 통계적 유의성 여부를 분석하기 위해서는 향후 충분한 자료를 바탕으로 면밀한 연구가 진행되어야 할 것이다.

5. 결론

본 연구는 가계대출 유형을 은행권 주담대 및 마통, 비은행권 주담대 및 마통 등 네 가지로 분류하고 베이지안 머신러닝 기법을 적용하여 가계대출 유형별 추이를 예측하였다. 예측 결과, 유형 별로 주요 예측변수와 모형이 대단히 상이하였다. 은행권 주담대는 주로 대출금리, 아파트 입주물량, 분양물량 등에 의해 예측 가능하며, 은행권 마통의 주요 예측변수는 취업률과 전세가격지수이다. 반면, 비은행권 주담대는 대출금리와 아파트 매매전세가비율 등에 의해 주로 결정되며, 비은행권 마통은 취업률, 전세가격지수와 더불어 주가 수익률의 영향도 많이 받는 것으로 추정되었다. 이러한 대출 유형 별 예측변수와 모형은 예측 대상 시점에 따라서 시변하는 경향이 있는데, 최근에는 아파트 입주 물량이 은행권 주담대 직접 예측에 주요한 반면, 작년 은행권 주담대는 아파트 분양 물량이 높은 간접 예측력을 가졌다. 결과적으로 강규호(2018)의 연구에서 주장된 바와 같이 가계대출 총량 예측력 향상을 위해서 예측 변수와

6: **은행권 가계대출 결정요인: 2016년 6월 - 2017년 5월 예측** 이 표는 사후 분포 예측시 표본외 예측을 통해 선택된 예측변수 중 가장 큰 가중치가 부여된 예측변수와 해당 예측 모형을 나타낸 것이다. (간접)은 VAR 모형, (직접)은 ADL 모형을 의미한다.

	은행권 주담대	은행권 마통
	선택된 예측변수의 수=4개	선택된 예측변수의 수=13개
예측시계	주요 예측변수와 모형	주요 예측변수와 모형
1	아파트 분양물량, 산업생산지수(직접)	소비자물가지수, CD 금리(직접)
2	주택담보대출금리(직접)	소비자물가지수, CD 금리(직접)
3	주택담보대출금리(직접)	소비자물가지수, CD 금리(직접)
4	주택담보대출금리(직접)	소비자물가지수, CD 금리(직접)
5	주택담보대출금리(직접)	소비자물가지수, CD 금리(직접)
6	주택담보대출금리(직접)	소비자물가지수, CD 금리(직접)
7	산업생산지수(간접)	소비자물가지수, CD 금리(직접)
8	아파트 분양물량, 산업생산지수(간접)	소비자물가지수, CD 금리(직접)
9	아파트 분양물량, 산업생산지수(간접)	소비자물가지수, CD 금리(직접)
10	아파트 분양물량, 산업생산지수(간접)	소비자물가지수, CD 금리(직접)
11	아파트 분양물량, 산업생산지수(간접)	소비자물가지수, CD 금리(직접)
12	아파트 분양물량, 산업생산지수(간접)	소비자물가지수, CD 금리(직접)

모형이 실시간으로 학습되어야 한다. 본 연구에서 제시된 예측 기법과 예측 결과는 가계부채 급등 조기경보 시스템 구축과 선제적인 정책대응에 도움이 될 것이다.

본문에서 언급된 바와 같이 본 연구의 예측기법은 기본적으로 강규호(2018)의 연구에서 차용된 것이며, 추정기법 상 일부 보완에도 불구하고 본 연구의 계량경제학적 기여도는 크지 않다. 또한 향후 후속 연구에서 추정방법 상 개선되어야 할 점들이 많이 남아있다. 그 중에서도 현재 $ADL(p,q)$ 모형의 최적시차가 동시에 결정되는 것이 아니라 순차적으로 결정된다는 점은 반드시 개선될 필요가 있다. 즉, 현재 알고리즘 상으로 p가 먼저 선택된 후, 주어진 p에서 표본외 예측력을 극대화시키는 q가 선택되고 있는데, 이러한 방식은 전역적 최적화를 보장하지 못한다는 심각한 한계가 있다. (p, q)를 동시에 최적화할 경우, 모형의 수가 $12 \times (PQK)$로 급격히 늘어나 계산비용이 크게 상승할 수 있지만, 추후 연구에서 Mitchell and Beauchamp(1988)이 제시한 변수선택 알고리즘으로 K의 크기를 줄이는 등의 방식을 도입해볼 여지가 있을 것이다. 다음으로 $ADL(p,q)$과 $VAR(p)$ 모형에 가계대출 이외에 하나의 변수만 각 모형에서 고려된다는 점도 큰 한계점이다. 이는 모형의 수가 과도하게 늘어나는 경우를 방지하기 위해 불가피한 선택일 수 있으나, 앞서 언급한 사전적인 변수선택과 더불어 컴퓨터 및 프로그램 언어의 계산능력이 향상되면 보다 많은

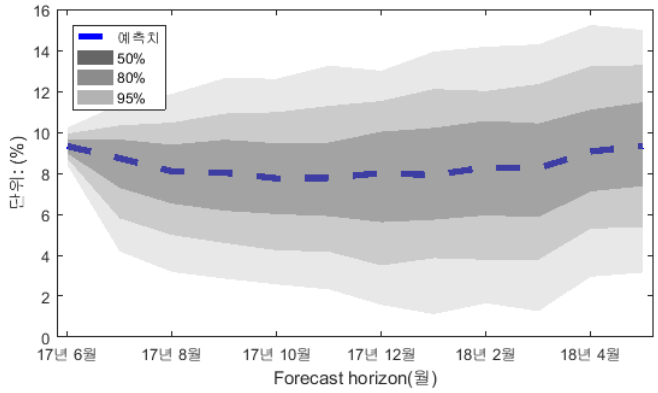
7: **비은행권 가계대출 결정요인: 2016년 6월 - 2017년 5월 예측** 이 표는 사후 분포 예측시 표본외 예측을 통해 선택된 예측변수 중 가장 큰 가중치가 부여된 예측변수와 해당 예측 모형을 나타낸 것이다. (간접)은 VAR 모형, (직접)은 ADL 모형을 의미한다.

	비은행권 주담대	비은행권 마통
	선택된 예측변수의 수=4개	선택된 예측변수의 수=13개
예측시계	주요 예측변수와 모형	주요 예측변수와 모형
1	주택담보대출금리, 국고채 5년금리(간접)	취업률(간접)
2	주택담보대출금리, 국고채 5년금리(간접)	아파트매매가격지수(간접)
3	주택담보대출금리, 국고채 5년금리(간접)	아파트 및 주택전세가격지수(간접)
4	주택담보대출금리, 국고채 5년금리(간접)	아파트 및 주택전세가격지수(간접)
5	아파트 매매전세가비율(직접)	아파트 및 주택전세가격지수(간접)
6	아파트 매매전세가비율(직접)	주택전세가격지수(직접)
7	아파트 매매전세가비율(직접)	주택 매매가격지수(직접)
8	아파트 매매전세가비율, 국고채 5년금리(직접)	주택 매매가격지수(직접)
9	아파트 매매전세가비율, 국고채 5년금리(직접)	주택 매매가격지수(직접)
10	아파트 매매전세가비율, 국고채 5년금리(직접)	주택 매매가격지수(직접)
11	아파트 매매전세가비율, 국고채 5년금리(직접)	주택 매매가격지수(직접)
12	아파트 매매전세가비율, 국고채 5년금리(직접)	주택 매매가격지수(직접)

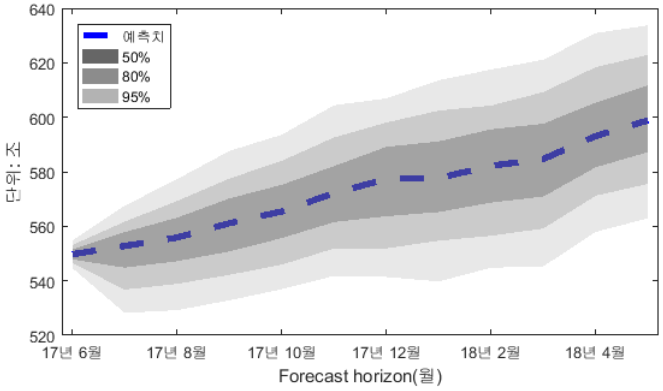
수의 변수를 통한 예측을 충분히 시도해 볼 수 있을 것이다.

3: **은행권 주담대 예측분포: 2017년 6월-2018년 5월** 점선은 사후 예측 분포의 평균이며, 50%, 80%, 및 95% 신용구간은 음영으로 표시된다.

(a) 증가율 예측분포

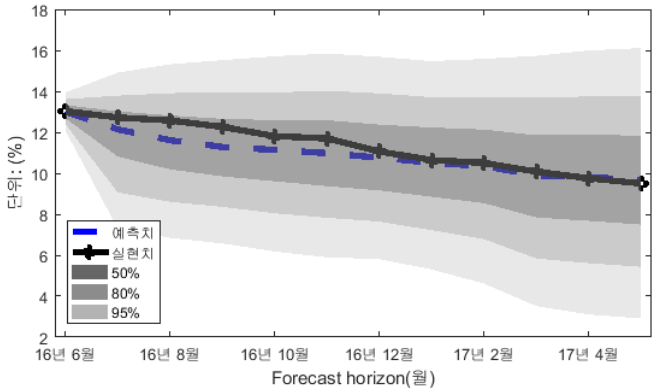


(b) 수준 예측분포

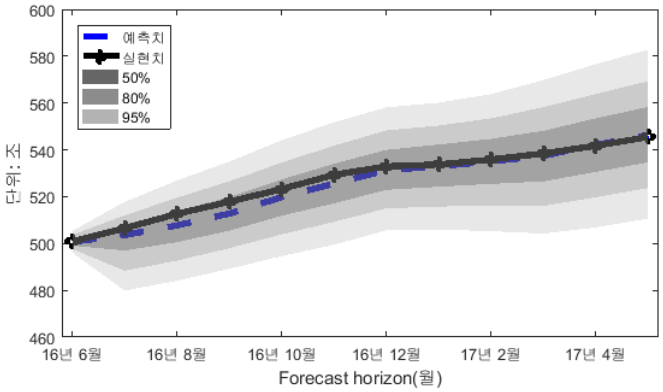


4: **은행권 주담대 표본외 예측분포: 2016년 6월-2017년 5월** 점선은 사후 예측 분포의 평균이고, 실선은 실현치이며, , 50%, 80%, 및 95% 신용구간은 음영으로 표시된다.

(a) 증가율 예측분포

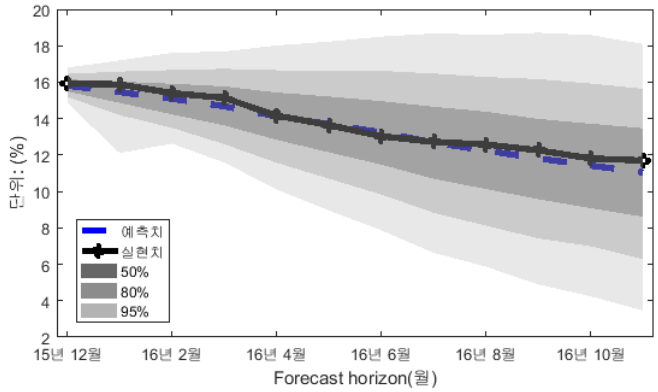


(b) 수준 예측분포

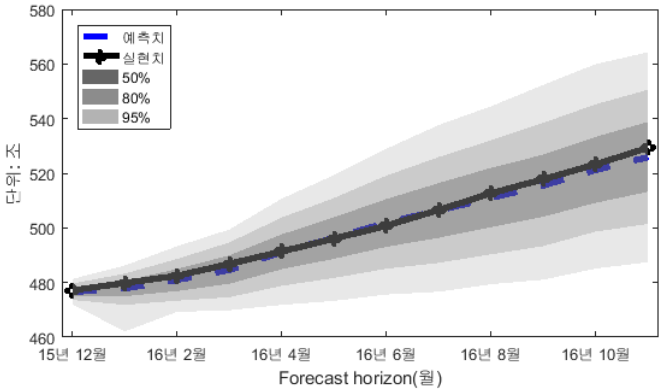


5: **은행권 주담대 표본외 예측분포: 2015년 12월-2016년 11월** 점선은 사후 예측 분포의 평균이고, 실선은 실현치이며, , 50%, 80%, 및 95% 신용구간은 음영으로 표시된다.

(a) 증가율 예측분포

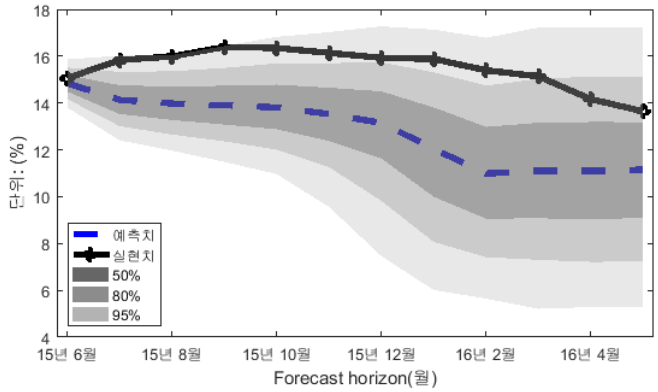


(b) 수준 예측분포

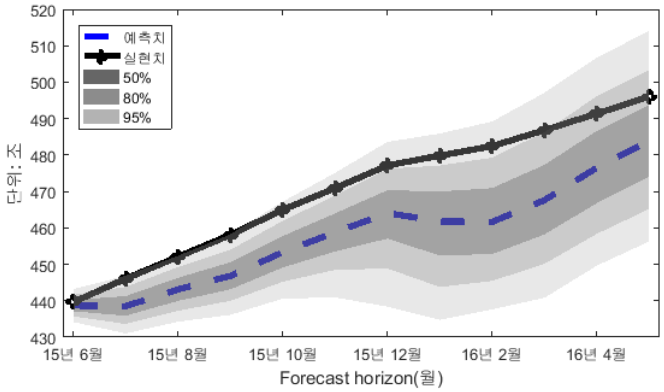


6: **은행권 주담대 표본외 예측분포: 2015년 6월-2016년 5월** 점선은 사후 예측 분포의 평균이고, 실선은 실현치이며, , 50%, 80%, 및 95% 신용구간은 음영으로 표시된다.

(a) 증가율 예측분포

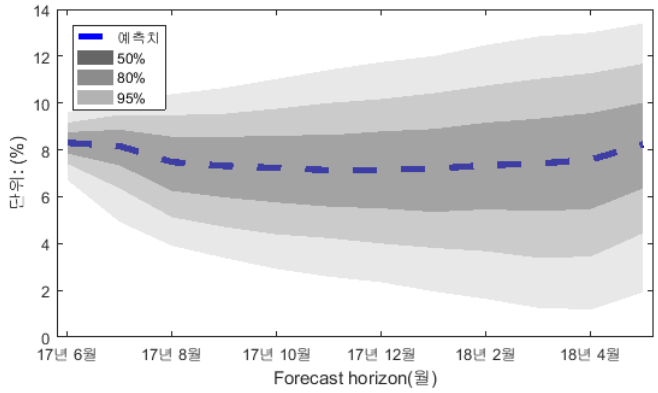


(b) 수준 예측분포

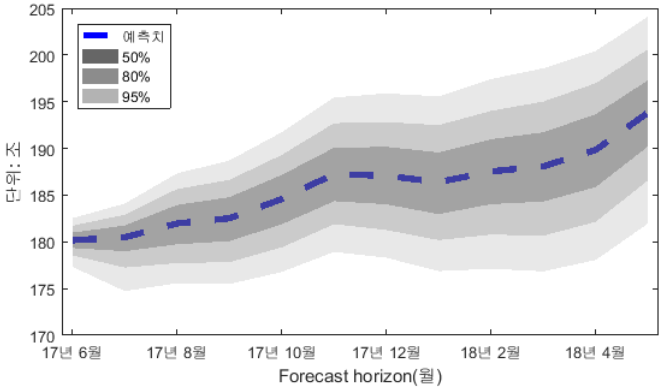


7: **은행권 마통 예측분포: 2017년 6월-2018년 5월** 점선은 사후 예측 분포의 평균이며, 50%, 80%, 및 95% 신용구간은 음영으로 표시된다.

(a) 증가율 예측분포

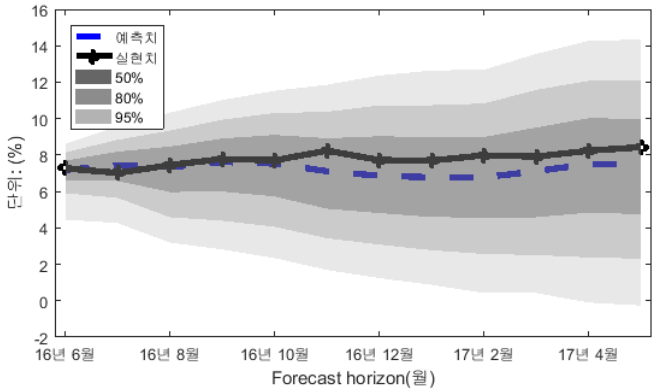


(b) 수준 예측분포

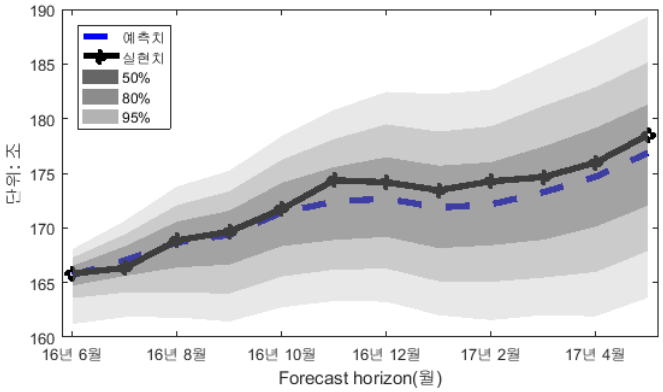


8: **은행권 마통 표본외 예측분포: 2016년 6월-2017년 5월** 점선은 사후 예측분포의 평균이고, 실선은 실현치이며, , 50%, 80%, 및 95% 신용구간은 음영으로 표시된다.

(a) 증가율 예측분포

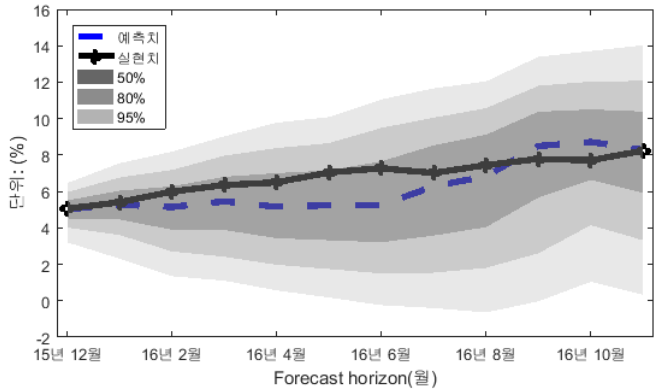


(b) 수준 예측분포

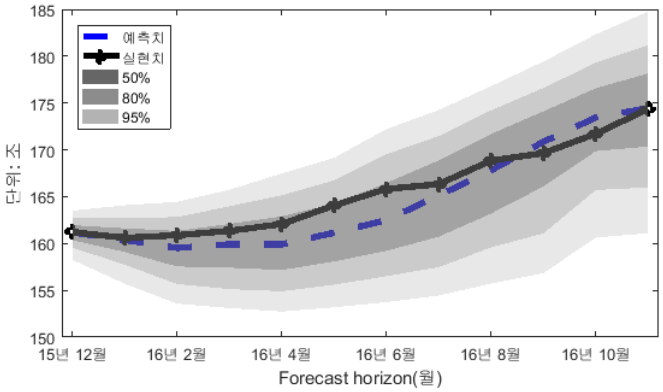


9: **은행권 마통 표본외 예측분포: 2015년 12월-2016년 11월** 점선은 사후 예측 분포의 평균이고, 실선은 실현치이며, , 50%, 80%, 및 95% 신용구간은 음영으로 표시된다.

(a) 증가율 예측분포

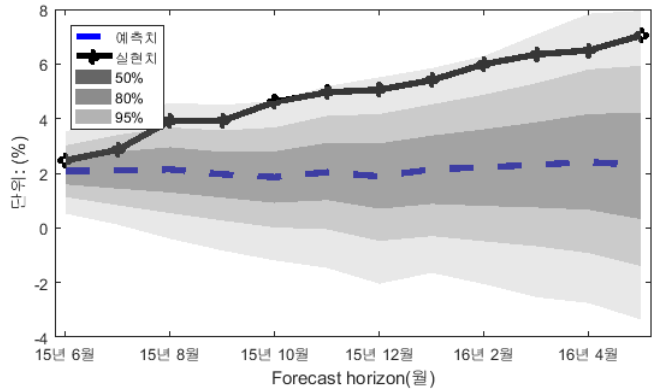


(b) 수준 예측분포

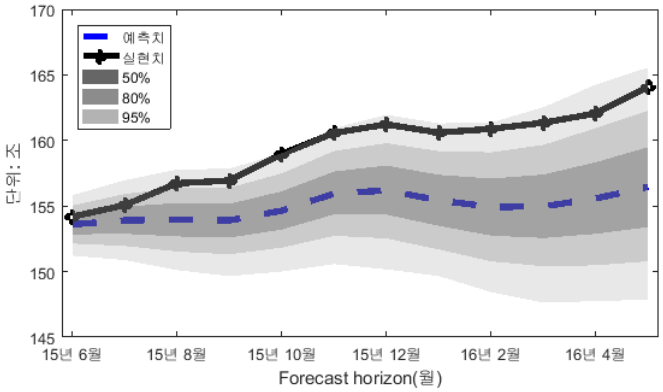


10: **은행권 마통 표본외 예측분포: 2015년 6월-2016년 5월** 점선은 사후 예측 분포의 평균이고, 실선은 실현치이며, , 50%, 80%, 및 95% 신용구간은 음영으로 표시된다.

(a) 증가율 예측분포

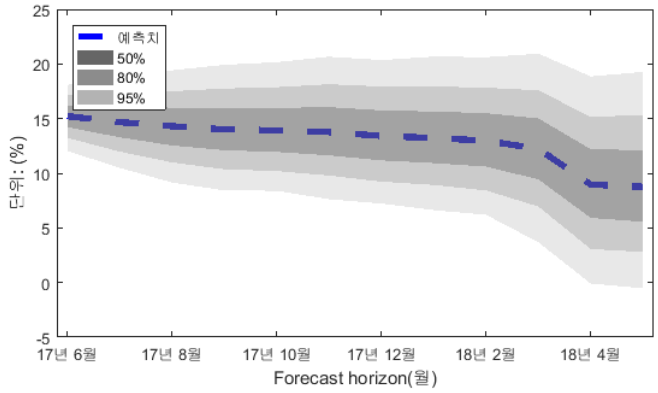


(b) 수준 예측분포

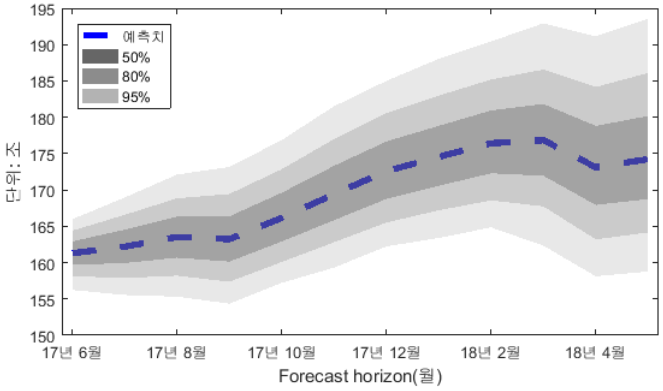


11: **비은행권 주택담보대출 예측분포: 2017년 6월-2018년 5월** 점선은 사후 예측 분포의 평균이며, 50%, 80%, 및 95% 신용구간은 음영으로 표시된다.

(a) 증가율 예측분포

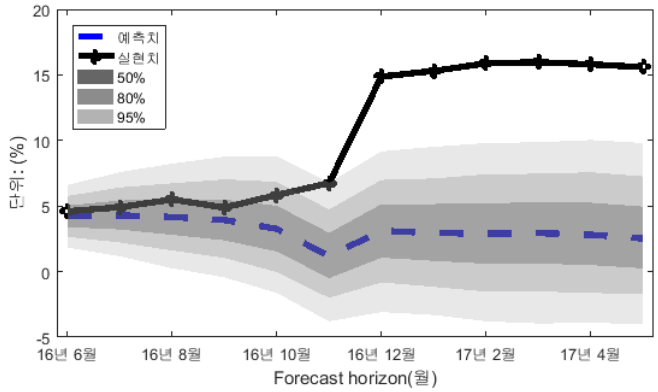


(b) 수준 예측분포

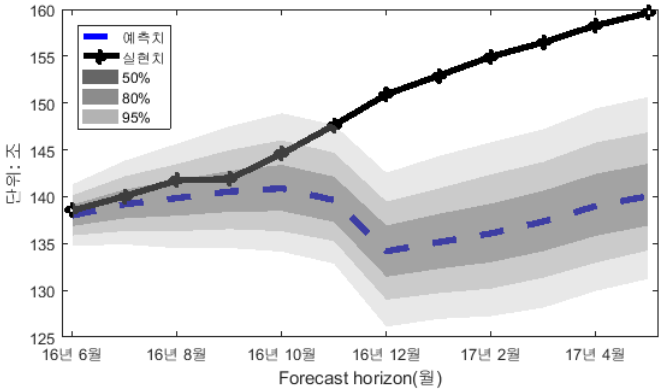


12: **비은행권 주담대 표본외 예측분포: 2016년 6월-2017년 5월** 점선은 사후 예측 분포의 평균이고, 실선은 실현치이며, , 50%, 80%, 및 95% 신용구간은 음영으로 표시된다.

(a) 증가율 예측분포

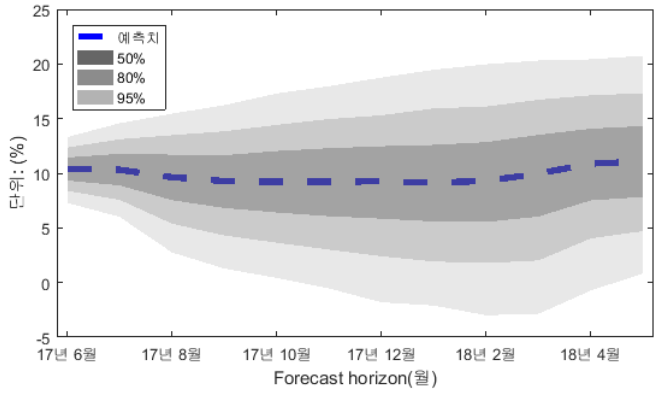


(b) 수준 예측분포

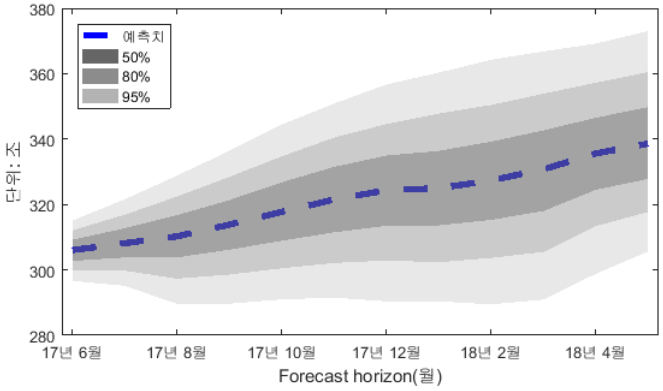


13: **비은행권 마통 예측분포: 2017년 6월-2018년 5월** 점선은 사후 예측 분포의 평균이며, 50%, 80%, 및 95% 신용구간은 음영으로 표시된다.

(a) 증가율 예측분포

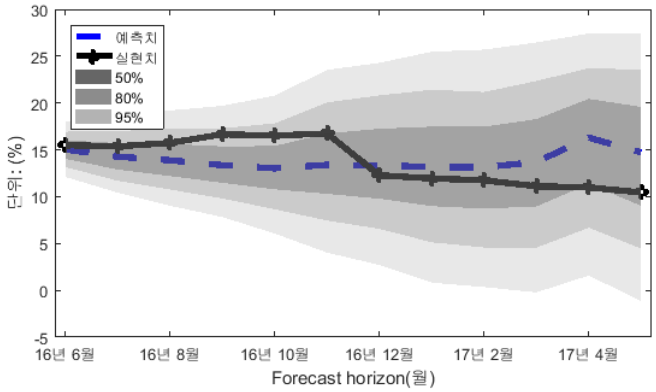


(b) 수준 예측분포

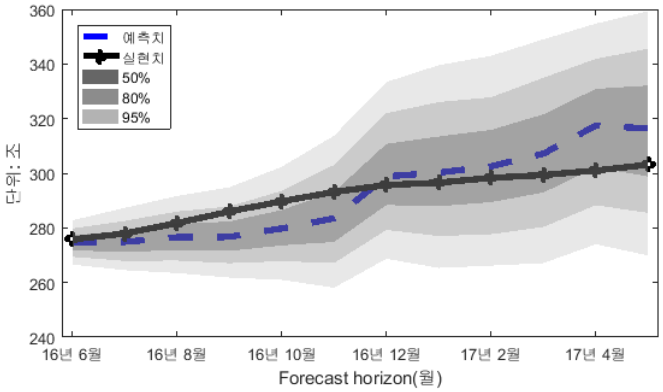


14: **비은행권 마통 표본외 예측분포: 2016년 6월-2017년 5월** 점선은 사후 예측 분포의 평균이고, 실선은 실현치이며, , 50%, 80%, 및 95% 신용구간은 음영으로 표시된다.

(a) 증가율 예측분포



(b) 수준 예측분포



8: **예측변수 목록**

자료 번호	자료명	출처
1	아파트 분양물량(전국, 전년동월대비 증가율)	부동산114
2	아파트 분양물량(수도권, 전년동월대비 증가율)	
3	아파트 분양물량(서울, 전년동월대비 증가율)	
4	아파트 입주물량(전국, 전년동월대비 증가율)	
5	아파트 입주물량(수도권, 전년동월대비 증가율)	
6	아파트 입주물량(서울, 전년동월대비 증가율)	
7	주택 매매가격지수(전국, 전년동월대비 증가율)	KB부동산
8	주택 매매가격지수(수도권, 전년동월대비 증가율)	
9	주택 매매가격지수(서울, 전년동월대비 증가율)	
10	아파트 매매가격지수(전국, 전년동월대비 증가율)	
11	아파트 매매가격지수(수도권, 전년동월대비 증가율)	
12	아파트 매매가격지수(서울, 전년동월대비 증가율)	
13	주택 전세가격지수(전국, 전년동월대비 증가율)	
14	주택 전세가격지수(수도권, 전년동월대비 증가율)	
15	주택 전세가격지수(서울, 전년동월대비 증가율)	
16	아파트 전세가격지수(전국, 전년동월대비 증가율)	
17	아파트 전세가격지수(수도권, 전년동월대비 증가율)	
18	아파트 전세가격지수(서울, 전년동월대비 증가율)	
19	소비자물가지수(전국, 총지수, 전년동월대비 증가율)	한국은행
20	소비자물가지수(전국, 주택임차료, 전년동월대비 증가율)	
21	전산업생산지수(농림어업제외, 2010=100)	
22	취업률	통계청
23	고용율	
24	종합주가지수	한국은행
25	미국 채권금리(6개월물)	
26	미국 채권금리(10년물)	
27	건설경기실사지수(CBSI)	한국건설산업연구원
28	금리(CD 91일)	한국은행
29	평균 주택담보대출금리	
30	국고채 5년 만기수익률	
31	아파트 매매전세가비율(전국, %)	KB부동산
32	아파트 매매전세가비율(수도권, %)	
33	아파트 매매전세가비율(수도권, %)	
34	로그 아파트 분양물량(전국, 호)	부동산114
35	로그 아파트 분양물량(수도권, 호)	
36	로그 아파트 분양물량(서울, 호)	
37	로그 아파트 입주물량(전국, 호)	
38	로그 아파트 입주물량(수도권, 호)	
39	로그 아파트 입주물량(서울, 호)	
40	지가변동률	통계청
41	LTV, DTI 규제	금융연구원

강규호 (2018). 베이지안 머신 러닝을 이용한 은행권 주택담보대출 예측, *금융안정연구*, 게재예정.

강종구 (2017). 가계부채가 소비와 경제성장에 미치는 영향: 유량효과와 저량효과 분석, *BOK 경제연구* 제2017-1호, 한국은행.

김우영, 김현정 (2009). 가계부채의 결정요인 분석, *금융경제연구* 제380호, 한국은행.

김경아 (2011). 국내가계의 부채증가 추세 및 요인에 관한 연구 -미시자료에 대한 분석을 중심으로, *응용경제* 13(1), 209-237.

김영일, 변동준 (2012). 우리나라 가계부채의 주요현황과 위험도 평가: 차주단위 자료를 중심으로, *KDI 정책연구시리즈* 2012-06, 한국개발연구원.

김승욱, 남영우 (2012). 주택가격변화에 따른 가계부채의 위험증가에 대한 연구, *부동산학보* 15, 240-251.

Eklund, Jana and Sune Karlsson (2007). Forecast combination and model averaging using predictive measures, *Econometric Reviews* 26, 329–363.

Mitchell, T. J., and J. J. Beauchamp (1988). Bayesian Variable Selection in Linear Regression, *Journal of the American Statistical Association* 83(404), 1023-1032.