



머신러닝 기법을 이용한 조기상환 규모 예측

민병철

국민의 주거 행복을 책임지는 지속가능 주택금융의 선도기관

주택금융

머신러닝 기법을 이용한 조기상환 규모 예측

2022. 12



-
- 저 자 : 민병철 연구위원(051-663-8177 / bcmin@hf.go.kr)
 - 본고의 내용은 필자의 개인 의견으로 한국주택금융공사의 공식적인 견해와 다를 수 있습니다.
-

목 차

I. 연구배경 및 목적	1
II. 모형설정	3
III. 사용자료	9
IV. 예측성능	14
1. BSS등급 사용하지 않은 모형	14
2. BSS등급을 사용한 모형	15
3. 두 모형 장단점 비교	16
V. 결론 및 연구의 한계	18
참 고 문 헌	19

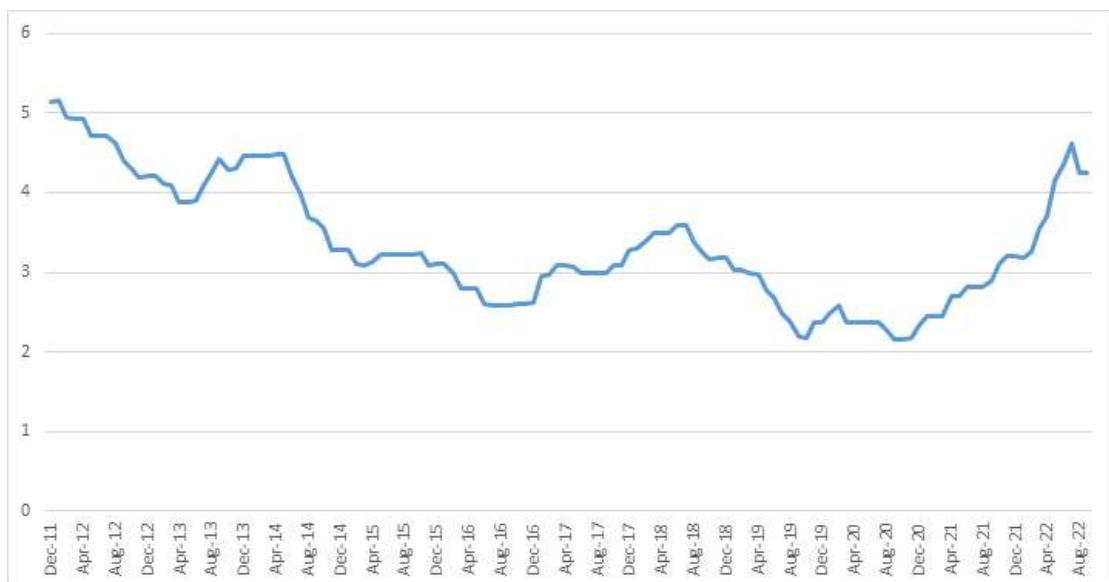
요 약

- 이 연구에서는 공사 자료를 이용하여 향후 1년 조기상환율을 예측하는 모형을 구축하고자 하였음
 - 공사는 장기 고정금리 대출을 양수하여 MBS를 발행하는데, 금리 변동으로 인해 조기상환율이 달라지면 대지급 혹은 재투자를 해야하는 상황에 처함. 따라서 경영 계획 수립 시 조기상환율에 대한 예측률을 높인다면 도움이 될 수 있음
- 이를 위해 개별 차주 단위 데이터를 활용하여 딥러닝 모형을 훈련시킴
 - 딥러닝 모형은 개별 변수의 영향력을 식별해내기 어렵다는 단점이 있으나, 이 연구에서는 예측을 보다 더 잘하는 모형을 만들하고자 하였으므로 큰 문제가 되지 않음
 - 다양한 차주 단위 정보와 거시변수를 이용함
- 훈련시킨 모형의 예측력을 시험군에서 검증한 결과, 평균 절대오차(MAE)는 0.35%~0.4%, 평균 절대비율오차(MAPE)는 7.1%~8.3% 수준임
 - 정책모기지의 특성상 금리가 정책적으로 결정되어 조기상환을 예측하는데 한계가 있다는 점을 고려하면 예측력이 아주 낮지는 않으나, 예측력 개선의 여지는 충분히 있는 것으로 보임
- BSS등급 정보를 이용하였을 시 다소 높은 예측능력을 보였으나, 현재는 BSS등급 정보를 이용한 모형이 불완전한 상태이므로 추후 보완이 이루어진다면 더 높은 예측능력을 보일 것으로 기대됨

I. 연구배경 및 목적

- (배경) 최근 15년간 금리 인하(글로벌 금융위기, Covid-19) 및 인상 요인(경기 소폭 회복, 고물가)들이 번갈아가며 나타나고 있음

〈그림 1〉 보금자리론 가중평균금리 수준



※ 자료: 공사 내부 자료

- 공사는 장기 고정금리 대출을 양수하여 MBS를 발행하는데, 금리 변동으로 인해 조기상환율이 달라지면 대지급 혹은 재투자를 해야하는 상황에 처함

- 공사는 기초자산으로부터 받는 원리금 상환액 및 예상되는 조기상환액을 MBS투자자에게 전달하기로 약속함. 만약에 조기상환액이 줄어들면 MBS투자자에게 전달해야 할 돈이 부족하게 되어 대지급을 해야 하고, 조기상환액이 너무 많아지면 잉여자금을 재투자 해야 하는 입장임
- 금리가 높아지면 저금리 시절에 고정금리로 받은 고객은 조기상환을 할 유인이 줄어들어, 공사는 대지급을 해야 할 가능성이 높아짐
- 금리가 낮아지면 낮은 금리로 갈아타고자 하는 고객이 늘어 조기상환이 증가할 수 있으며, 이 때 공사는 재투자를 해야 하는 상황이 될 수 있음

- 따라서 조기상환율을 합리적으로 예측하는 것은 매우 중요함. 비록 현재 공사는 자체적으로 조기상환을 예측하는 모형을 구축하여 운영하고 있으나, 모델 리스크를 고려한다면 다양한 모형을 이용하여 예측하는 것이 필요할 수도 있음
 - 모델 리스크란 모형이 틀렸을 시 발생하는 리스크로, 모델 하나에만 의존할 경우 해당 모델이 부정확한 상황이 닥쳤을 때 대응하기 어려움
 - IMF에서는 스트레스 테스트 시 모델 리스크를 줄이기 위한 방법으로 Bayesian Model Averaging방법을 이용해 다양한 모델의 예측치를 조합하도록 권장(Adrian et. al., 2020)
- 이 연구에서는 딥러닝 모형을 차주 단위의 미시데이터로 훈련시켜 조기상환율을 예측하는 모형을 개발함
 - 공사가 현재 사용하고 있는 조기상환모형은 비공개이므로, 여기서는 기존 모형과의 구체적인 비교는 하지 않음
 - 또한, 분석에 사용한 개별 차주의 데이터도 비공개이므로 이 연구에서는 사용 표본 수 외의 자세한 기초통계량 또한 소개하지는 않음

II. 모형설정

- 학술적으로 발표된 선행연구들은 대부분 예측력 자체보다는 조기상환율에 영향을 미치는 요인들을 식별하는데 초점을 맞췄으며, 따라서 모형 결과의 해석가능성이 중요함
 - 이자율의 영향: Green and Shoven(1986), Richard and Roll(1989) 등
 - 거시변수: Cunningham and Capone(1990), Schwartz and Torous(1993), Deng(1997) 등
 - 기타 요인: Deng, Quigley, and Van Order(2000), Stanton and Wallace(2011), Chernov, Dunn, and Longstaff(2018)
- 이를 위해 사용한 모형은 주로 로지스틱 회귀분석과 콕스 비례 모형으로 나뉨
 - 로지스틱 회귀분석: Campbell and Dietrich(1983), Cunningham and Hendershott(1986), Elul et al.(2010), Agarwal et al. (2011) emd
 - 콕스 비례모형: Green and Shoven(1986), Deng, Quigley, and Van Order(2000), Stanton and Wallace(2011)
- 로지스틱 회귀분석과 콕스 비례모형의 경우, 개별 차주의 조기상환 여부를 예측하는 구조이므로 부분 조기상환이 가능할 경우 평균 조기상환 금액은 별도로 예측해야 함
 - 두 모형의 종속변수는 이항변수로, 조기상환 여부를 예측하게 됨
 - 조기상환 금액은 조기상환 여부·조기상환 시 조기상환 금액 으로 계산됨
- 따라서 실무에서는 편의상 MBS 풀/코호트(발행연도)별로 연령별(대출 실행 이후 경과기간) 잔액 대비 조기상환율을 특정 함수로 묘사하되, 위 연

구들에서 발견한 주요 변수들의 영향을 반영하는 형태로 조기상환율을 예측하는 경우가 많음

- 통계적으로 조기상환율은 대출 실행 이후 일정 시점까지 증가하다가 이후 감소하는 경향을 보임

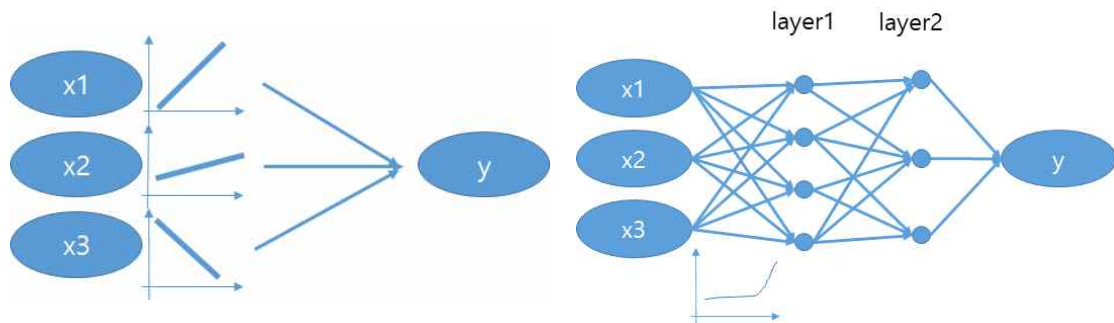
□ 이와 달리, Sadhwani et. al.(2020)은 딥러닝 모델을 이용하여 모기지 관련 위험을 분석함

- 딥러닝 모델이란 인공 신경망(Artificial Neural Network)의 일종으로, 인공 신경망의 은닉층(hidden layer)이 2개 이상일 때 이를 심층 신경망(Deep Neural Network)라 하는데, 이를 딥러닝 시스템(Deep Learning System), 혹은 딥러닝 모델이라고도 부름

□ 일반 회귀모형들은 개별 설명변수의 선형결합으로 종속변수를 설명하는데 반해, 딥러닝 모형은 개별 설명변수의 비선형적 결합이 새로운 설명변수를 만들어 내고, 이렇게 만들어진 새로운 설명변수들의 비선형적 결합으로 종속변수를 설명함(<그림 2>)

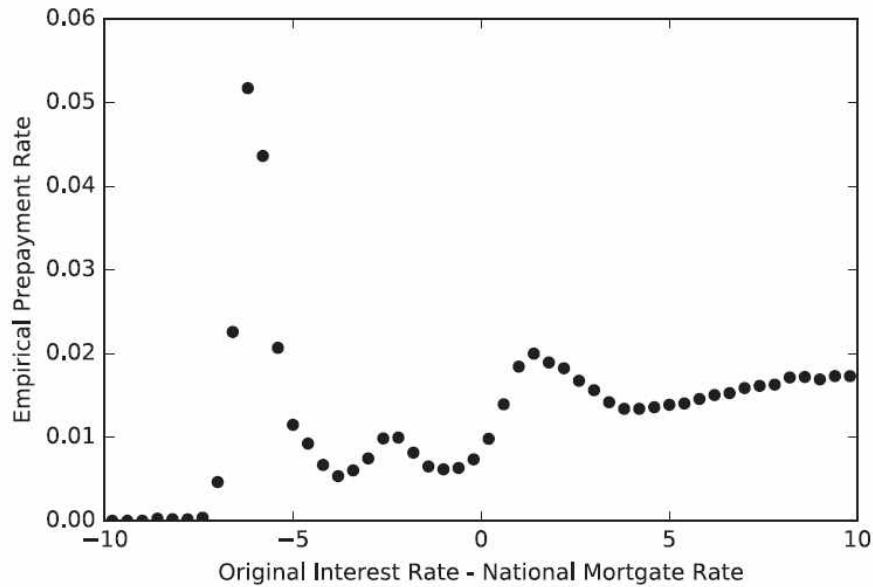
- 딥러닝 모형은 설명변수들의 조합으로 여러 새로운 설명변수(node, 노드)를 만들어 냄. 이렇게 만들어진 새로운 설명변수들의 집합을 첫 번째 레이어(layer)라 할 수 있음. 첫 번째 레이어에 있는 새로운 설명변수(노드)들의 조합으로 또 다시 새로운 설명변수들을 만들어 내는데, 이렇게 만들어진 새로운 설명변수(노드)들의 집합이 두 번째 레이어임. 이런 식으로 여러 레이어와 노드들을 만들어 내어 이를 이용해 결국 종속변수를 설명함
- 이 과정에서 개별 변수값에 상수값을 곱하는 것이 아닌(선형 관계 가정), 비선형 함수의 함수값(주로 ReLU함수 사용)을 이용하여 비선형적 영향력을 모델링함
- 이 설명은 딥러닝 모형에 대해 모형 사용자 입장에서 개념적으로 풀어쓴 것이며, 엄밀한 정의 및 설명은 Géron, A. (2017) 참조

〈그림 2〉 일반 회귀모형과 딥 러닝 모형의 비교



- 이렇게 복잡한 과정을 거치는 이유는, 충분히 많은 레이어와 노드가 있다면, 이론적으로 딥러닝 시스템은 모든 형태의 관계를 묘사하는 것이 가능하기 때문임
 - 다만, 설명변수(노드)들이 많아지는 만큼 충분히 많은 데이터와 연산능력이 필요함
 - 또한, 과최적화의 문제를 해결할 필요가 있음
- 딥러닝 모형은 일반적으로 매우 높은 예측력을 보이는 반면, 개별 변수들의 영향력을 정량적으로 측정하기가 매우 어렵다는 단점이 있음
 - 따라서 이를 블랙박스(black box) 라고 부르기도 함
 - 최근에는 이러한 한계를 넘어서고자 설명 가능한 AI(Explainable AI, XAI)에 관한 연구가 이루어지고 있음
- Sadhwani et. al.(2020)에서는 훈련시킨 모형에, 기타 변수들은 고정시킨 채 하나의 변수만 변화시킨 인공 자료를 입력시켜 예측치를 관찰하는 방식으로 개별 변수의 영향력을 살폈음(<그림 3>)

〈그림 3〉 최초 대출이자율 - 현재 대출 이자율 수준과 조기상환율 간 관계



※ Sadhwani et. al.(2020)

□ **딥러닝 모형으로 조기상환율의 예측 자체에 초점을 맞춘 연구로는 Amar(2020)을 들 수 있는데, 이 연구에서는 개별 대출을 포트폴리오 단위로 묶어서 조기상환율을 예측하는 방법을 제시함**

- 개별 차주 단위로 분석할 때는 조기상환 여부와 조기상환 시 금액을 각각 예측해야 하나, 대출을 특성별로 포트폴리오 단위로 묶을 경우 해당 포트폴리오의 전체 금액 대비 조기상환 금액을 예측하면 되므로 하나의 모형으로 예측가능
- 예를 들어, $LTV(50) > LTV$, $50 \leq LTV$ 와 상환회차($30 \leq$ 상환회차, 상환회차 < 30)를 기준으로 대출을 분류하면 월별로 4개의 포트폴리오가 만들어짐 (〈그림 4〉). 이렇게 묶인 포트폴리오의 기준시점 잔액 대비 향후 1년 조기상환액을 예측할 수 있음

〈그림 4〉 Amar(2020)의 포트폴리오 구성 방식

계좌번호	연월	LTV	상환회차	기준시점 잔액	향후 1년 조기상환액
1	202001	70	5	1억 원	1천만 원
2	202001	20	10	2억 원	0
3	202001	15	60	3억 원	0
4	202001	65	40	2억 원	1억 원
5	202001	10	50	1억 원	1천만 원

연월	LTV<50	LTV>=50	상환회차<30	상환회차>=30	변수 평균값	기준시점 잔액	1년 조기상환율
202001	1	0	1	0		2억 원	0
202001	1	0	0	1		4억 원	1천만 원/4억 원
202001	0	1	1	0		1억 원	1천만 원/1억 원
202001	0	1	0	1		2억 원	1억 원/2억 원

- 이 연구에서는 Amar(2020)의 방식을 채택하여 1년 조기상환율 예측모형을 만들었는데, 주요 이유는 Sadhwani et. al.(2020)의 방식에 비해 연산 능력을 덜 요구하지만 예측능력에는 큰 차이가 없기 때문임
- Amar(2020) 이 시도한 바에 따르면 조기상환 여부와 조기상환 시 금액을 각각 예측하여 결합한 모형의 MAE(Mean Absolute Error)는 0.005334, 개별 대출을 포트폴리오 형식으로 묶어 하나의 모형으로 예측한 경우의 MAE는 0.005681로, 전자의 예측능력이 더 높기는 하지만 그 차이는 크지 않음
- 또한, 과최적화(overfitting)을 방지하기 위해 훈련군(training set)과 시험군(test set)을 나누고, 훈련 과정에서 dropout방식을 사용함
 - Dropout이란 개별 노드에 대한 의존도를 줄이기 위해, 매 iteration마다 임의로 일정 비율의 노드를 배제하고 훈련시키는 방식을 말함
- 하이퍼파라미터(hyperparameter, 여기서는 노드 및 레이어 숫자) 결정은 일반적인 하이퍼파라미터 튜닝 방식을 사용하기 어려워 trial and error 방식을 사용함
 - 파이썬(Python)의 kerastuner함수 등을 이용하여 노드 및 레이어 수준별로 모형의 예측능력을 대략적으로 파악하여 최적 범위를 좁힐 수 있으나, 이 연구에서 사용한 방식은 모형의 예측능력이 최종 조기상환 예측능력으로 직결되지 않기 때문에 사용하지 못함

- <그림 4>의 예시에서, kerastuner함수는 모든 포트폴리오의 예측오차를 동일한 중요도로 평가하여 최적의 모형을 제시하나, 실제로는 대출잔액이 가장 많은 $LTV < 50$, $30 \leq$ 상환회차인 포트폴리오에 대한 예측능력이 가장 중요함

□ 모형 훈련은 파이썬의 케라스(Keras)패키지를 활용하였음

III. 사용자료

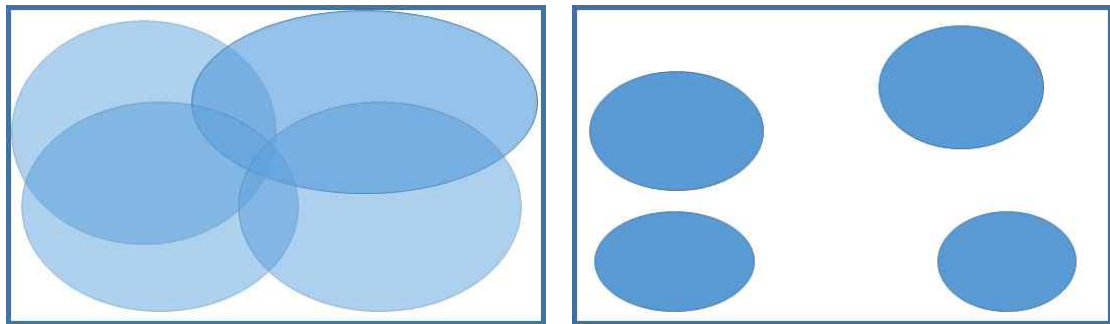
- 이 연구에서 사용한 자료는 크게 발행 시점의 대출정보, 월별 정보, 그리고 거시 변수 정보로 나뉨
- (대출정보) 2004년 11월부터 2022년 2월까지 대출 실행된 3,286,852 건에 대하여 자료 정제하여 3,163,969건 사용
 - 분석에 사용하는 변수의 값이 모두 존재하는 건만 사용함
 - 대출실행일, 대출실행액, 대출종류, 대출목적, LTV, DTI, 거치기간, 상환방식, 소득, 주택종류, 생년월일 주소
 - 값이 존재하는 건을 대상으로 조기상환율을 추정하고, 이를 전체의 조기상환율과 같다고 가정함
 - 이는 결측치는 단순 실수 및 착오로 발생한 것으로, 결측치의 존재유무는 조기상환율과 무관하거나, 아니면 최소한 전체적으로 봤을 때 큰 차이를 주지 않는다는 가정을 전제로 함
- (월별 정보) 2004년 11월부터 2022년 2월까지의 상환회차, 잔액, 조기상환액, BSS등급 정보를 활용하였으며, CLTI 정보를 별도로 계산함
 - $CLTI(\text{Current Loan to Income}) = \text{현재 대출잔액} / \text{가구 연소득}$
 - 단, BSS등급 정보는 2011년 9월부터 존재함
- (거시 정보) 실업률, 전산업생산지수, 소비자물가지수, 주택가격지수(지역별, 종합/아파트), 보증자리론 금리, 국고채 (1, 3, 5, 10)년 금리를 사용하였으며, 추가로 금리갭(대출 이자율-현재 보증자리론 금리) 변수를 계산하여 사용함
 - 기준 월 및 6개월 전 값을 그대로 사용함
 - 6개월 전 값이 포함되므로, 차주 정보가 존재하는 2004년 11월보다 6개월 전인 2004년 5월부터의 값을 사용함

- 변화율 값을 따로 계산하여 넣지 않았는데, 이는 딥러닝 모형이 최근 추세를 반영할 모양까지 결정하도록 한 셈임
- 자료출처: 보금자리론 금리→공사 내부자료, 주택가격지수→한국부동산원, 기타 거시변수→한국은행 ECOS

□ (다중공선성 문제) 일반 회귀분석에서는 서로 상관관계가 있는 변수들을 같이 포함시키는 것을 피하기도 하나, 예측 목적의 모형 설정 시 다중공선성으로 인한 문제보다 최대한 많은 변수를 포함시키는 것의 이득이 더 커 최대한 많은 정보를 포함시킴

- 다중공선성이 존재하면 개별 계수의 영향력이 제대로 파악되지 않을 수 있으나, 전체적인 설명력에는 차이가 없음
- 아래 그림과 같이 서로 정보가 겹치지 않는 변수들만을 사용하면 개별 변수의 영향력은 명확히 알 수 있지만(<그림 5> 오른쪽), 전체적인 설명력은 떨어질 수 있음. 반대로 서로 정보가 일부 겹치더라도 최대한 많은 정보를 담고 있는 변수들을 사용한다면(<그림 5> 왼쪽) 대상을 더 잘 설명할 수 있음
- 회귀분석의 목적이 경제적 이론 및 가설의 검증이라면 개별 변수의 영향력을 검증하는 것이 중요하므로 다중공선성이 문제가 되겠으나, 이 연구와 같이 예측 자체가 목적이라면 다중공선성은 큰 문제가 되지 않음
- 심지어 선행연구에는 다중공선성이 존재한다면 중요 변수가 누락되었을 우려를 덜 수 있다고 언급함
 - Mankiw and Weil(1998): (239p) indeed, multicollinearity may be a strength, for it acts to eliminate any worry over the role of omitted variables
- 물론, 모형 훈련에 사용한 변수 값 범위를 벗어난 상황에 대한 예측 시에는 문제가 될 수 있음. 예를 들어, 자료에 포함된 GDP증가율이 -10%~+10%이었는데, 이 때 훈련된 계수값들을 이용해 GDP변화율이 -20%이거나 +20%일 때를 예측하려 한다면, 개별 변수의 영향력이 제대로 식별되지 않아 부정확한 예측을 할 위험이 있음

〈그림 5〉 다중공선성 예시



- (포트폴리오 설정) 개별 대출 단위로 분석하는 것이 아닌 포트폴리오 단위로 분석을 하게 되는데, 이때 포트폴리오를 얼마나 세분화하느냐는 선택이 필요한 문제임
 - 포트폴리오를 너무 세분화하면 일부 포트폴리오에는 포함된 대출의 숫자가 너무 적어져 조기상환율 값이 이상치의 영향을 크게 받을 수 있음
 - 포트폴리오를 너무 크게 나누면 결국 모형에 들어가게 되는 표본의 숫자가 너무 적어져 훈련이 제대로 되지 않을 수 있음
- 분류하였을 때 각 포트폴리오의 크기, 그리고 분류하였을 때 조기상환율의 차이 등을 고려하여 나눔
 - 개별 대출 정보가 아닌 포트폴리오 정보로 모형을 훈련시키는 만큼, 포트폴리오가 최대한 전체 정보 집합의 구석구석을 세밀하게 다루도록 나누어야 함
- 여러 시험 결과, 최종적으로 대출종류, 상환방식, 상환목적, 주택종류, 지역, 상환회차, CLTI, 이자율갭을 기준으로 포트폴리오를 나눔
 - 대출종류: 보증자리론, 디딤돌, 적격대출, 안심전환(4가지)
 - 상환방식: 원리금균등, 원금균등, 체증식(3가지)
 - 대출목적: 구입, 보전/상환(2가지)
 - 주택종류: 아파트, 기타(2가지)

- 지역: 서울, 경기/인천, 5대광역시, 기타지역(4가지)
- 상환회차: 20미만, 20-40, 40초과(3가지)
- CLTI: 2 이상, 2 미만(2가지)
- 금리갭: -0.5%이하, -0.5% ~ 0, 0~0.5%, 0.5% 초과(4가지)
- BSS등급: 2등급 이하, 2-5등급, 5등급 초과(3가지)
- 가능한 총 포트폴리오(표본) 수: $4 \times 3 \times 2 \times 2 \times 4 \times 3 \times 2 \times 4 \times 3 = 13,824$
 - BSS등급을 포함하지 않는 모형의 경우는 최대 4,608개임 ($4 \times 3 \times 2 \times 2 \times 4 \times 3 \times 2 \times 4 = 4,608$)
 - 개별 포트폴리오에 포함된 대출 수가 30건 미만일 경우(BSS등급을 포함하지 않는 모형은 100건 기준) 사용하지 않아, 실제 월별 표본의 개수는 이보다 적음

□ 포트폴리오 분류 기준 정보를 포함하여 총 84개의 설명변수를 포함시킴

- 상기 언급한 포트폴리오 분류 기준을 모두 설명변수로 포함시키되, CLTI, 금리갭, 상환회차는 포트폴리오 평균값을 사용하고 나머지는 더미변수로 포함시킴
- 차주정보 변수: LTV, DTI, 나이, (포트폴리오별)거치기간 중인 비중
- 거시정보 변수: 보금자리론금리, 실업률, 가계 평균 주담대금리, 전산업생산지수, 경제심리지수, 소비자물가지수, 주택가격지수(아파트/종합, 전국/수도권/지방), 국고채금리(1, 3, 5, 10년)
 - 기준시점 값 및 6개월 전 값 사용
- 직전 조기상환 추이 변수: 직전 1, 3, 6, 12개월 조기상환율
 - 전체 조기상환율이 아닌 각 포트폴리오 별 조기상환율 사용
- 서민형안심전환대출 시기 더미: 서민형 안심전환대출 취급 시기에는 거시 변수나 차주 특성과 무관하게 조기상환율이 증가하여, 이를 통제하기 위해 해당 시기에 대하여 더미변수를 포함시킴

- 1차 서안심: 201503-201515
 - 2차 서안심: 201910-201912
- 2004년 11월-2019년 12월을 훈련군으로, 2020년 1월-2021년 6월까지를 시험군으로 사용함
- 1년 조기상환율을 예측하는 모형을 구축하므로, 2021년 6월에 대한 표본은 2021년 6월의 정보를 가지고 2021년 7월-2022년 6월의 조기상환율을 예측하는 것임
 - 훈련군에서 다시 20%는 validation set으로 사용함
- BSS등급 정보를 사용하는 모형의 경우, 2011년 9월부터 정보가 존재하므로 2011년 9월-2019년 12월을 훈련군으로, 2020년 1월-2021년 6월까지를 시험군으로 사용함
- BSS등급을 사용하지 않은 모형과의 비교를 위해 시험군을 동일하게 설정

IV. 예측성능

- 이 장에서는 훈련시킨 모형의 예측성능을 평가하되, MAE(Mean Absolute Error: 평균 절대 오차)와 MAPE(Mean Absolute Percentage Error: 평균 절대비율 오차)를 기준으로 사용함

- $MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$

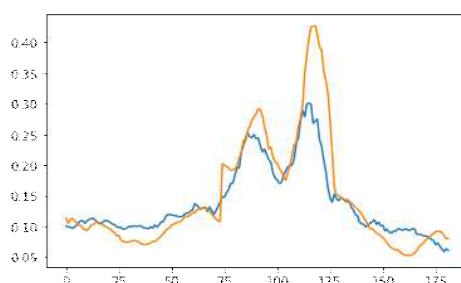
- $MAPE = \frac{100}{n} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right|$

1. BSS등급 사용하지 않은 모형

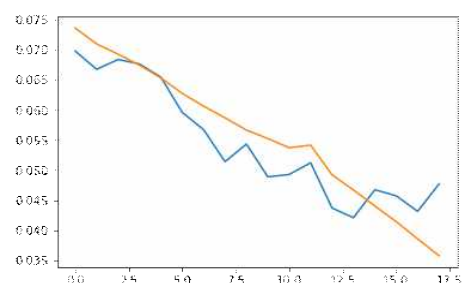
- 다양한 시도 결과, 2개의 layer에 node개수는 각각 20개, dropout 비율은 0.2일 때 최적의 예측능력을 보임
- Optimizer는 RMSPROP(0.01)를, 활성화 함수는 ReLU를 사용
 - 일반적으로는 optimizer로 ADAM에 더 나은 성능을 낸다고 알려져 있으나, 이 연구에서는 RMSPROP을 활용하였을 때 예측력이 더 나았음
- 물론, 이는 trial and error를 통해 찾아낸 것이므로 더 나은 예측력을 가진 조합이 있을 가능성은 얼마든지 존재함

〈그림 6〉 BSS등급을 사용하지 않은 모형의 예측결과

훈련군 (200411-201912)



시험군(202001-202106)



※ 주황색: 실제 값, 파란색: 예측치

□ 훈련군의 MAE 와 MAPE는 각각 0.026, 19.0%, 시험군의 MAE와 MAPE는 각각 0.004, 8.3%로 나타남

- 훈련군에서는 1차 서안심 시기에 조기상환율이 급등하여 모형이 이를 제대로 예측하지 못했으나, 시험군에서는 1차 서안심 시기가 포함되지 않으므로 오히려 더 높은 예측능력을 보임
- 일반적으로는 시험군의 예측오차가 훈련군의 예측오차보다 큼
- 2차 서안심 시기에는 1차와 같은 조기상환률 급등은 관찰되지 않음

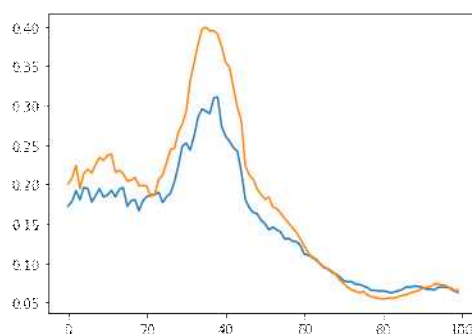
2. BSS등급을 사용한 모형

□ 다양한 시도 결과, 3개의 layer에 node개수는 각각 70개, 20개, 10개, dropout 비율은 0.3일 때 최적의 예측능력을 보임

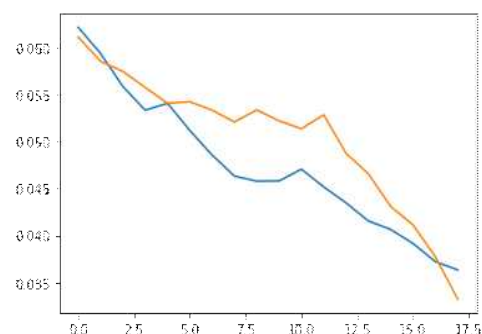
- Optimizer는 RMSPROP(0.01)를, 활성화 함수는 ReLU를 사용
- 앞선 경우와 마찬가지로, 이는 trial and error를 통해 찾아낸 것이므로 더 나은 예측력을 가진 조합이 있을 가능성은 얼마든지 존재함

〈그림 7〉 BSS등급을 사용한 모형의 예측결과

훈련군 (201109-201912)



시험군(202001-202106)



※ 주황색: 실제 값, 파란색: 예측치

□ 훈련군의 MAE 와 MAPE는 각각 0.028, 13.2%, 시험군의 MAE와 MAPE는 각각 0.0035, 7.1%로 나타남

- 훈련군에서는 1차 서안심 시기에 조기상환율이 급등하여 모형이 이를 제대로 예측하지 못했으나, 시험군에서는 1차 서안심 시기가 포함되지 않으므로 오히려 더 높은 예측능력을 보임
- 일반적으로는 시험군의 예측오차가 훈련군의 예측오차보다 큼
- 2차 서안심 시기에는 1차와 같은 조기상환률 급등은 관찰되지 않음

□ BSS등급은 대출 실행 이후 3개월 지난 후부터 산출되며, 또한 차주가 특정 상태가 되면 더 이상 산출하지 않으므로 이 모형에서 산출한 조기상환율은 전체 자산의 조기상환율과 다소 차이가 있을 수 있음

- 특히, 통계적으로 대출 초반에 조기상환 발생 비율이 높은 경우가 많은데 BSS등급은 대출 실행 후 첫 3개월은 산정되지 않으므로 조기상환율을 과소평가하게 되어 이 예측치를 직접적으로 이용하기는 어려움

□ BSS등급을 이용한 모형의 예측치를 전체 데이터의 조기상환율과 비교한 결과, 훈련군의 MAE 와 MAPE는 각각 0.033, 15.7%, 시험군의 MAE와 MAPE는 각각 0.008, 14.1%로 나타나, BSS등급을 이용한 모형의 예측치를 그대로 사용하기는 어려움

- BSS등급이 산정된 학습된 데이터로 학습한 후 BSS등급이 없는 데이터를 이용해 검증한 것이기 때문에 예측력이 다소 낮게 나옴

3. 두 모형 장단점 비교

□ BSS등급을 사용한 모형의 예측력이 더 높고, 최근의 추세 또한 실제와 더 유사하지만 이 모형의 예측치를 그대로 사용하기는 어려움

- BSS등급이 산정되지 않는 대출들이 존재함

- 추후 신용정보 등재 등으로 인해 BSS등급이 산정되지 않는 경우와 대출 실행 이후 일정 기간이 경과되지 않아 BSS등급이 산정되지 않는 경우를 구분하여 BSS등급이 없는 차주에 대해서도 모델링할 필요가 있음
- BSS등급을 사용하지 않은 모형의 경우 최근의 예측능력이나 추세는 다소 부정확하지만, 글로벌 금융위기 등을 모두 포함하는 데이터로 훈련시켰으므로, 추후 다양한 위기가 닥쳤을 때의 예측능력은 BSS등급을 사용한 모형보다 보다 높을 것으로 기대됨
- BSS등급을 사용하지 않은 모형을 기준으로 삼되, BSS등급을 사용한 모형을 참고로 하여 최종 예측치를 반영하는 것이 적절할 것으로 보임

V. 결론 및 연구의 한계

- 이 연구에서는 공사 자료를 이용하여 향후 1년 조기상환율을 예측하는 모형을 구축하고자 하였음
 - 다양한 차주 단위 정보와 거시변수를 이용함
- 최근 몇 년간 금융 환경이 여러 차례 변했다는 점을 감안하면 예측력이 아주 낮다고는 하기 어려우나, 이 연구에서는 아주 높은 예측력을 가진 모형을 도출해내지는 못했음
 - 시험군의 평균 절대오차(MAE): 0.35%~0.4%
 - 시험군의 평균 절대비율오차(MAPE): 7.1%~8.3%
 - 첫 번째 이유로는 정책금융상품이라는 특성상, 금리가 시장 금리를 그대로 따라가는 것이 아닌 서민 주거복지 증진을 위하여 정책적으로 결정된다는 점을 꼽을 수 있음
 - 두 번째로는, 비정기적으로 실시되는 서민형안심전환 대출의 존재가 조기상환율에 영향을 미침
 - 마지막으로, trial and error 방식으로 최적의 하이퍼파라미터 조합을 찾고자 했으나 연산능력의 한계로 충분히 다양한 시도를 하지 못함
- BSS등급 정보를 이용하였을 시 다소 높은 예측능력을 보였으나, BSS등급이 부재한 차주에 대해서는 예측하지 못하는 한계가 있어 추후 보완이 필요함
 - BSS등급이 부재한 이유로는 크게 대출 실행 이후 3개월 이내인 경우와 신용정보 등재 등 상환능력에 문제가 생긴 경우로 나뉘므로, 이를 구분하여 데이터를 분석해야 함

참 고 문 헌

- Adrian, M. T., Morsink, M. J., & Schumacher, M. B. (2020). Stress Testing at the IMF. International Monetary Fund.
- Agarwal, S., G. Amromin, I. Ben-David, S. Chomsisengphet, and D. Evanoff. 2011. The Role of Securitization in Mortgage Renegotiation. *Journal of Financial Economics* 102(3): 559-578.
- Amar, S. 2020. Modeling of Mortgage Loan Prepayment Risk with Machine Learning, Thesis in postgraduate studies of Informatics, Department of Informatics, Technical University of Munich
- Campbell, T., and J. Dietrich. 1983. The Determinants of Default on Insured Conventional Residential Mortgage Loans. *The Journal of Finance* 38(5): 1569-1581.
- Chernov, M., B. Dunn, and F. Longstaff. 2018. Macroeconomic-Driven Prepayment Risk and the Valuation of Mortgage-Backed Securities. *The Review of Financial Studies* 31(3): 1132-1183.
- Cunningham, D. F., and C. A. Capone. 1990. The Relative Termination Experience of Adjustable to Fixed-Rate Mortgages. *The Journal of Finance* 45(5): 1687-1703.
- Cunningham, D., and P. Hendershott. 1986. Pricing Fha Mortgage Default Insurance. *Housing Finance Review* 13: 373-392.
- Deng, Y., J. Quigley, and R. Van Order. 2000. Mortgage Terminations, Heterogeneity, and the Exercise of Mortgage Options. *Econometrica* 68(2): 275-307.
- Deng, Y. 1997. Mortgage Termination: An Empirical Hazard Model with a Stochastic Term Structure. *The Journal of Real Estate Finance and Economics* 14(3): 309-331.
- Elul, R., N. Souleles, S. Chomsisengphet, D. Glennon, and R. Hunt. 2010. What Triggers Mortgage Default. *The American Economic Review* 100(2): 490-494.

참 고 문 헌

- Géron, A. (2017). Hands-on machine learning with scikit-learn and tensorflow: Concepts, Tools, and Techniques to build intelligent systems.
- Green, J., and J. Shoven. 1986. The Effects of Interest Rates on Mortgage Prepayments. *Journal of Money, Credit and Banking* 18(1): 41-59.
- Mankiw, N. G., and Weil, D. N. (1989). The Baby Boom, the Baby Bust, and the Housing Market, *Regional Science and urban Economics*, 19(2): 235-238
- Richard, S., and R. Richard. 1989. Prepayments on Fixed-Rate Mortgage-Backed Securities. *The Journal of Portfolio Management* 15(3): 73-82.
- Sadhwani, A., K. Giesecke, and J. Sirignano. 2020. Deep Learning for Mortgage Risk, *Journal of Financial Econometrics*, 19(2): 313-368.
- Schwartz, E., and W. Torous. 1993. Mortgage Prepayment and Default Decisions: A Poisson Regression Approach. *Real Estate Economics* 21(4): 431-449.
- Stanton, R., and N. Wallace. 2011. The Bear's Lair: Index Credit Default Swaps and the Subprime Mortgage Crisis. *Review of Financial Studies* 24(10): 3250-3280.