「제3회 대구 빅데이터 분석 경진대회」
분석 결과 보고서

접수번호	※ 작성하지 않음

성명(팀명)	GoDart
분석과제명	경제변수를 활용한 초개인화 서비스 제안

※ 20장 내외 자유 형식으로 작성, 목차는 필요시 변경 가능

Ⅰ. 분석개요

▢ 분석목적

딥러닝을 통해 경제변수 변화에 따른 금융시장 참여자의 행동 변화를 예측·분석하여 초개인화된 금융상품과 서비스를 추천 및 개발할 수 있도록 한다

▢ 배경 및 필요성

-금융상품 활용 증가

금융위기 이후 각국 중앙은행들은 저성장 국면을 탈피하기 위한 방안으로 기준금리를 낮추어 투자를 촉진하였다. 그럼에도 불구하고 선진국을 위주로 경제성장률은 여전히 낮은 수준을 기록하였고, 10년이 넘는 장기간동안 저성장-저금리 환경이 고착화되었다. 안전자산의 한 종류로 많이 활용되었던 예금, 적금 상품들은 기준금리가 낮아짐에 따라 점차 매력이 떨어지게 되었고 부의 양극화 현상이 심화됨에 따라 소비자들은 더 높은 수익률을 가진 금융상품들을 찾기 시작했다. 오픈서베이에서 발표한 '금융 트렌드 리포트(2020.11)'에 따르면 보통예금, 정기예금을 포함한 예금, 적금 상품의 이용률은 감소한 반면, 주식, 채권, 부동산 소액투자와 같은 금융상품에 대한 이용률은 전년 대비 상승한 것을 볼 수 있다. 이는 단순 예금, 적금 상품보다 기대수익률이 높은 금융상품으로 이동하는 소비자들의 행동변화가 있음을 보여준다.

작년 3월 전 세계적으로 유행한 코로나 19는 금융상품 활용 현상을 가속화 시켰다. 실물경제 회복을 위해 실시한 막대한 유동성 정책은 부동산을 포함한 자산가격의 급격한 상승을 유발하여 부의 양극화 현상을 더 심화시켰으며 더 높은 수익률을 가진

금융상품에 대한 소비자들의 수요를 가속화시킨 계기가 되었다. 이러한 상황에서 금융상품에 관한 비대면 맞춤서비스가 활성화된다면 상품을 찾아보고 선택하는데 있어 소비자의 혼란과 기회비용을 줄여 시민들의 편의를 증가시켜 줄 것이라 생각하였다.

-마이데이터 사업

마이데이터 사업이란 본인 정보에 대한 개인의 권리를 보장하고, 정보 주체인 개인의 동의에 따라 본인의 데이터를 개방, 활용하는 것을 의미한다. 마이데이터 사업이 본격적으로 시행된다면 오픈API를 기반으로 방대한 양의 금융 고객데이터 활용이 가능하며 마이데이터 사업을 실시함으로써 새로운 비즈니스와 혁신의 기회가 도래할 것이다.

금융 분야에서 마이데이터는 분산되어있는 개인금융정보의 통합조회 및 관리, 맞춤형 데이터 분석과 이에 근거한 금융상품 자문, 추천서비스 등 데이터 기반의 금융서비스 창출을 가능케 한다. 내년 1월 시행될 마이데이터 사업은 금융시장 참여자에 대한 방대한 데이터를 제공하여 고객 맞춤 서비스를 더욱 활성화시킬 수 있는 환경을 마련해줄 것이다.

-금융생활에 있어 거시경제의 영향력

자본시장이 고도화되고 기술이 성장함에 따라 세계경제의 공동화 현상이 심화되고 있다. 황상연(2010, 경기연구원)의 연구는 금리, 환율, 유가 등 대내외 경제충격이 다양한 지역경제변수에 영향을 미침을 보였으며 대내외 경제변수의 지역경제 영향 및 파급경로를 분석한 변창욱 외 3인(2017, 산업연구원)의 연구는 경제의 연결성이 강화되는 상황에서 지역경제의 안정적인 운용을 위해서는 대내외 환경변화를 주시해야한다는 점을 시사하였다.

이러한 선행연구들을 통해 지역 대내외 경제변수가 시민들의 금융 활동에 영향을 준다는 것을 파악할 수 있다. 따라서 거시경제 변화와 금융시장 참여자 행동 사이를 분석한다면 생활 방식에 따른 금융 행동 양상을 예측할 수 있을 것으로 생각하였다. 특히 지역거점 은행으로서 대구은행데이터는 대구시민들의 금융 활동 특성을 잘 대변해줄 것으로 생각하였다. 따라서 대구시 대내외 거시경제 변수와 대구은행 데이터를 통해 개인별 맞춤 서비스를 제공한다면 금융 서비스의 질적 수준을 높여 대구시민의 금융 활용 능력 제고에도 도움이 될 것으로 생각하여 본 기획을 구상하게 되었다.

▢ 분석요약

-전체 프로세스

○ 활용데이터

1.1. 주최 측 제공 데이터

데이터명	형식	데이터기간	사용변수	출처
대구은행 고객데이터	csv	2018.01~2020.12	전체 칼럼	대구은행

1.2. 분석 시 추가 활용할 공공・민간데이터

데이터명	형식	데이터기간	사용변수	출처
원/달러 환율	csv	2018.01~2020.12	원/달러 환율	한국은행
원/위안화 환율	csv	2018.01~2020.12	원/위안화 환율	한국은행
국고채(3 년)	csv	2018.01~2020.12	국고채(3 년)	한국은행
회사채(3 년,AA-)	csv	2018.01~2020.12	회사채(3 년,AA-)	한국은행
대구 고용지표	csv	2018.01~2020.12	고용률, 실업률	통계청
대구시 기업경기실사지수(BSI)	csv	2018.01~2020.12	제조업, 비제조업 업황	한국은행
대구 소비자물가지수	csv	2018.01~2020.12	전체 칼럼	통계청

○ 분석도구 및 분석기법 :

- Windows 10 및 Linux Ubuntu 18.04 환경에서 분석 실시
- 분석도구는 Python 3.7.11을 사용하였으며, 사용한 패키지는 다음과 같다.
- numpy, pandas : 데이터 핸들링 및 n차원 array의 벡터연산.
- matplotlib : 분석 결과 시각화
- sklearn, XGBoost, pytorch … : ML 모델 작성을 위한 패키지

ㅇ 분석결과
- 경제변수에 민감하게 반응했다고 판단 가능한 고객 라벨링
- 향후 1개월간 경제변수에 민감하게 반응할 고객 분류

ㅇ 활용방안
-고객별 맞춤 상품 추천
-고객별 맞춤 상품 개발

▢ 독창성 및 차별성

금융상품에 대한 소비자들의 관심이 늘어나는 상황에서 개별 소비자의 욕구를 충족시키기 위해서는 적절한 금융상품 추천 서비스가 필요하다. 이는 금융상품을 찾고 선택하는데 있어 소비자의 기회비용을 줄여줄 수 있기 때문이다. 하지만 소비자 특성에 대한 적절한 고민이 없다면 상품추천 서비스는 오히려 소비자에게 혼란만 가져올 수 있다. 따라서 정교한 상품추천을 위해서는 그 이전 단계에서 고객의 특성을 잘 분류하는 것이 중요하다.

우리는 대구은행의 고객데이터를 기반으로 고객을 분류하고자 한다. 여기에 더해 주로 금리라는 경제변수에 초점을 맞추는 기존의 은행 산업의 특성과 달리 다양한 경제변수를 활용하였다. 코로나19와 같이 예상치 못한 경제충격이 발생하더라도 결국 거시경제 변수의 변화로 나타나기 때문에 다양한 거시경제변수의 사용은 시민들의 금융 행동 패턴에 대한 예측과 분석을 가능케 할 것이다. 또한, 모델링 과정에서 대구경북지역과 밀접한 관련이 있는 경제변수를 추가로 고려하여 선택하였기에 대구은행 고객데이터만이 갖는 특징인 대구시민들의 금융행동 변화에 보다 집중할 수 있을 것이다.

우리는 이번 분석을 통해 경제변수마다 민감하게 반응할 은행 고객들을 사전에 예측함으로써 알맞은 금융상품 서비스 추천을 하는데 도움이 되고자 한다. 그리고 경제변수 민감도에 따른 고객들을 분류한 후 해당 군집의 특성을 분석한다면 특정 고객군이 필요로 하는 상품이 무엇인지 파악하는데도 도움이 될 것이다. 또한 향후 마이데이터 사업 시행으로 인해 활용 가능할 것으로 기대되는 대구시민들의 방대한 금융 데이터는 모델의 정확도를 높여 보다 유의미한 추천을 가능하게 할 것이다.

Ⅱ. 분석방법

▢ 활용데이터

1.1. 주최 측 제공 데이터

데이터명	형식	데이터기간	사용변수	출처
대구은행 고객데이터	csv	2018.01~2020.12	전체 칼럼	대구은행

1.2. 분석 시 추가 활용할 공공·민간데이터

데이터명	형식	데이터기간	사용변수	출처
원/달러 환율	csv	2018.01~2020.12	원/달러 환율	한국은행
원/위안화 환율	csv	2018.01~2020.12	원/위안화 환율	한국은행
국고채(3 년)	csv	2018.01~2020.12	국고채(3 년)	한국은행
회사채(3 년,AA-)	csv	2018.01~2020.12	회사채(3 년,AA-)	한국은행
대구 고용지표	csv	2018.01~2020.12	고용률, 실업률	통계청
대구시 기업경기실사지수(BSI)	csv	2018.01~2020.12	제조업, 비제조업 업황	한국은행
대구 소비자물가지수	csv	2018.01~2020.12	전체 칼럼	통계청

▢ 분석과정

○ 데이터 전처리

시계열적인 변화로부터 경제변수에 대한 민감성을 판별할 수 있으므로 대구은행 고객 데이터에서 시계열 칼럼을 획득하고자 하였다. 따라서 연도별로 시계열 칼럼을 생성하기 위해 다음과 같은 전처리 절차를 거쳤다.

삼개월
삼개월급여이체실적금액
삼개월신용카드사용금액
삼개월체크카드금액
삼개월수신평균잔액
삼개월현금서비스금액
삼개월대출평균잔액

전월
전월급여이체실적금액
전월신용카드사용금액
전월체크카드금액
전월수신평균잔액
전월현금서비스금액
전월대출월평균잔액

현월
급여이체실적금액
신용카드사용금액
체크카드거래금액
수신잔액
현금서비스이용금액
가계자금대출잔액
주택담보대출잔액

위와 같이 전체 칼럼 중 ‘삼개월’, ‘전월’ 데이터가 존재하는 칼럼에 집중하였다. 해당 칼럼들을 활용한다면 시계열 칼럼을 생성할 수 있다. 예를 들어, ‘삼개월신용카드사용금액’ 칼럼에서 ‘전월신용카드사용금액’과 ‘신용카드사용금액’을 빼면 2개월 전의 월단위의 신용카드사용 금액을 구할 수 있다. 즉 데이터 기준 시점이 2019년 12월일 때, 전월 데이터는 11월 데이터를 의미하고 기존에 존재하는 삼개월 데이터를 활용한다면 10월 데이터를 구할 수 있다.

	대출-2	대출-1	대출-0	신용-2	신용-1	신용-0
count	5.000000e+05	5.000000e+05	5.000000e+05	5.000000e+05	5.000000e+05	5.000000e+05
mean	3.183177e+07	2.261979e+07	1.427833e+07	3.470278e+05	3.240125e+05	3.373513e+05
std	3.754073e+08	1.977700e+08	8.536451e+07	1.614006e+06	1.403046e+06	1.428156e+06
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
75%	0.000000e+00	0.000000e+00	0.000000e+00	3.000000e+05	3.000000e+05	3.000000e+05
max	7.660000e+10	3.830000e+10	1.500000e+10	5.940000e+08	2.000000e+08	2.000000e+08

해당 표는 2019년의 시계열 칼럼 일부의 기초 통계량이다. 데이터 기준 시점이 2019년 12월이므로 '대출-2'의 경우 2달 전인 10월의 월 대출 데이터로 정의했다. 마찬가지로 '대출-1'의 경우 11월의 월 대출 데이터이며 '대출-0'의 경우 12월의 월 대출 데이터로 정의하였다.

	1 월	2 월	3 월	4 월	5 월	6 월	7 월	8 월	9 월	10 월	11 월	12 월
2019										생성	생성	기준
2020										생성	생성	기준
2021		생성	생성	기준								

최종적으로 대구은행 고객 데이터로부터 연도별로 3시점, 6종류 도합 18개의 시계열 칼럼을 획득하여 활용하였다.

○ 데이터 분석 내용

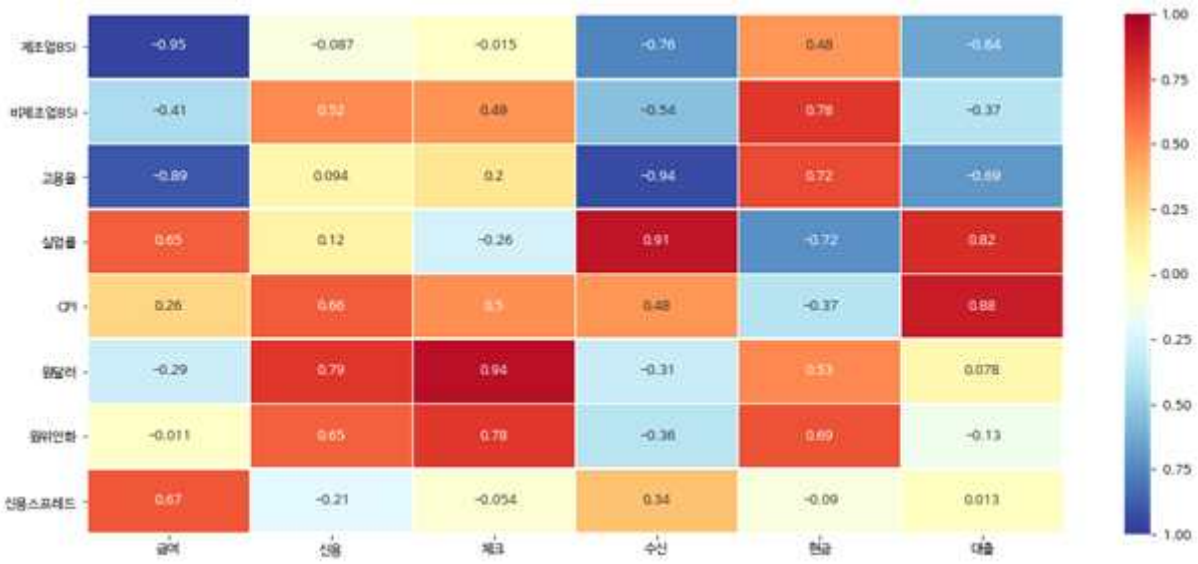
1. 라벨링

먼저, 전체 데이터로부터 경제변수에 민감하게 반응하였다고 할 수 있는 고객들을 분류하는 절차가 필요하다. 초기 계획으로는 그랜저 인과관계를 통해 경제변수와 고객의 행동 패턴 간의 인과성을 검증하려 하였다. 하지만 고객 행동 패턴을 유추할 시계열이 2019년 10~12월, 2020년 10~12월, 2021년 2~4월의 불연속적인 9개의 월 단위 시점 밖에 존재하지 않았다. 불연속적인 9개의 시점만으로는 시계열 분석하기에 부적합하였기 때문에 다른 방법을 고안해보았다.

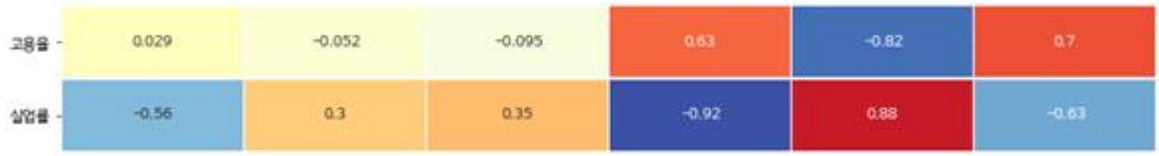
대안으로 경제 변수의 월간 변화와 고객 행동 변수의 월간 변화 평균 간의 상관관계를 활용하였다. 여기서 경제 변수의 변화와 높은 상관관계(상관계수 절대값 0.5 이상)을 보이는 시계열칼럼(이하 행동 변수)를 통해 민감도를 라벨링 하였다.
상관관계의 이점은 방향성을 알 수 있다는 점이다. 예를 들어 CPI와 신용카드 사용량의 월별 변화 간 높은 양의 상관관계가 존재한다면, CPI가 증가함에 따라 신용

카드 사용량이 높게 증가한 사람들이 CPI에 민감하다고 판단할 수 있다. 다만 대부분 경제변수의 현 시점의 변화가 행동변수에 즉각적으로 영향을 미친다고 가정한 것과 달리 고용률과 실업률의 경우 월급 등을 고려하여 1달의 시차를 가지고 영향을 미친다고 가정하였다.

-경제변수와 시계열 칼럼 간 상관관계



-고용률, 실업률 경제변수의 1달 시차 상관관계



아래 표는 각 경제변수와 유의미한 상관관계(상관계수 0.5이상)가 존재하는 칼럼들과 상관관계의 방향성(+,-)을 나타낸 표이다.

신용스프레드	CPI	원달러 환율	원위안화 환율
금여 (+)	대출(+) 체크카드(+) 신용카드(+)	신용카드(+) 체크카드(+) 현금서비스(+)	신용카드(+) 체크카드(+) 현금서비스(+)

제조업BSI	비제조업BSI	실업률	고용률
대출(-) 수신(-) 금여(-)	신용카드(+) 수신(-) 현금서비스(+)	금여(-) 수신(-) 현금서비스(+) 대출(-)	수신(+) 현금서비스(-) 대출(+)

상관관계는 인과성을 대변해주지는 않으며 제 3의 요인에 의해 상관관계가 있는 것처럼 보이는 허위 상관관계가 존재할 수 있기 때문에 주의가 필요하다. 따라서 각 칼럼의 특성을 유의 깊게 보았다. 그 결과 '급여이체실적금액'의 경우 아래와 같이 10, 11, 12월로 감에 따라 지속적으로 증가하는 계절성을 가지고 있는 것을 파악할 수 있었다.

따라서 획득 가능한 시계열 데이터가 10, 11, 12월에 밀집되어 있다는 점에서 허위 상관성이 나타났을 가능성을 고려하여, 급여 관련 칼럼은 분석에서 제외하기로 결정하였다.

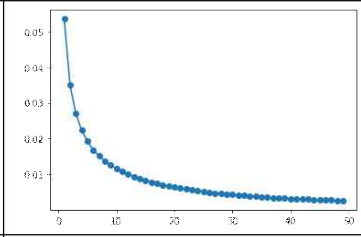
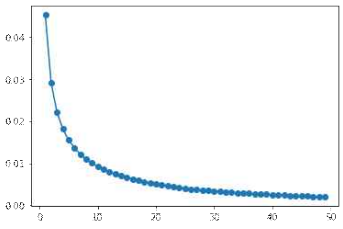
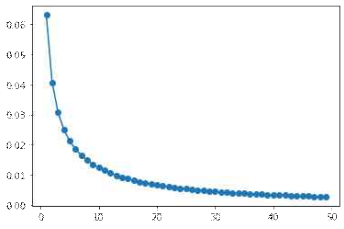
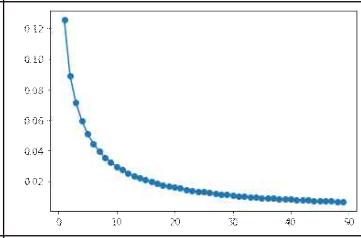
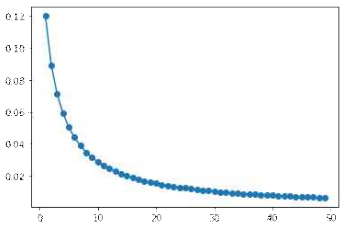
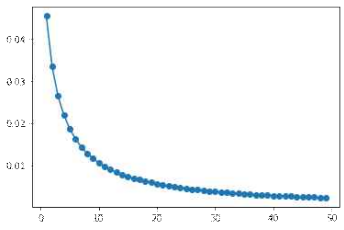
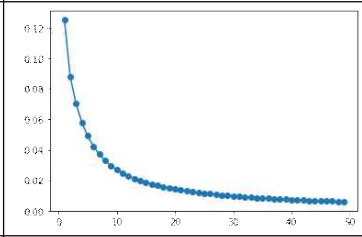
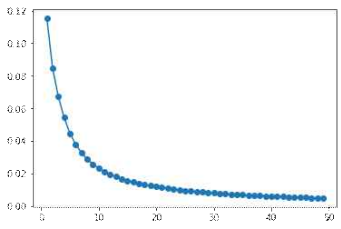
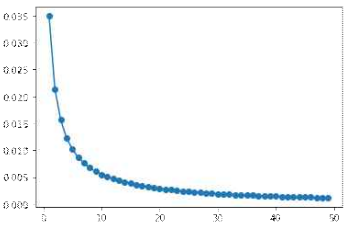
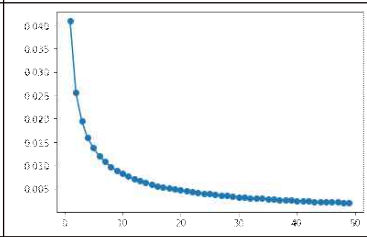
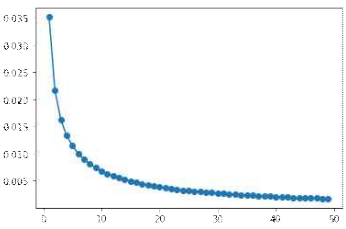
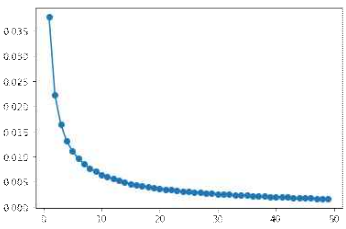
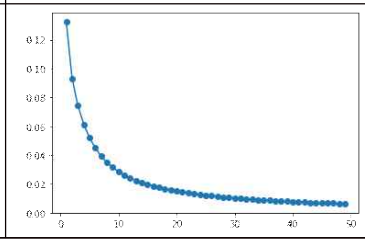
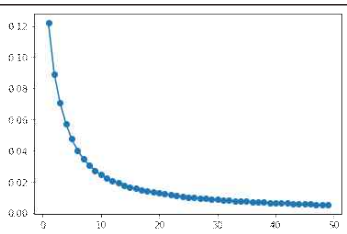
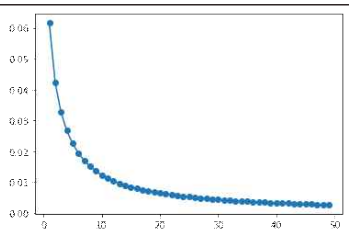
급여 칼럼을 제외하고는 위와 같은 기준으로 변화율 순위를 통해 민감한 사람을 라벨링 하였다. 단, 원달러 환율과 원위안화 환율에 대해 고객 행동 패턴의 방향성이 같게 나왔으므로 '환율'이라는 범주로 묶어서 분류하기로 하였으며, 실업률과 고용률의 경우도 고객 행동 패턴의 방향성만 다르므로 '고용'이라는 범주로 묶어 하나의 경제변수로 분류하였다.

민감하게 반응하는지의 유무를 파악하기 위해서는 상위 몇 %의 사람을 민감한 사람이라고 볼 수 있는가에 대한 기준을 세울 필요가 있다. 따라서 상위 변화율 n% 집단의 변화율 평균을 그래프로 나타내 민감한 기준을 선택하기로 결정하였다. 예를 들어 '환율'에 민감한 집단을 라벨링하기위해서 우선 지난 달 대비 환율의 변화량을 구한다. 다음으로 '환율'과 상관관계가 있는 행동변수인 '신용카드, 체크카드, 현금서비스 금액'의 지난 달 대비 변화량을 구한 후 세 변수의 단위 차이를 고려하여 Min-Max Normalization을 하였다. 그 후 세 변수의 변화량 평균을 개인 별로 구하고 최종적으로 상위 n%씩 끊어서 민감한 집단의 변화량 평균을 나타내었다.

-Min-Max Normalization

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

-민감한 정도를 파악하기 위한 경제 변수 별 Elbow method

	2019년	2020년	2021년
환율			
CPI			
제조업			
비제조업			
고용			

Elbow method으로 판단하였을 때 각 경제변수 별로 약 10%에서 elbow 지점이 발생함을 확인할 수 있다. 따라서 변화량 상위 10%의 사람들을 민감한 집단으로 레이블링 해주었다.

2. 모델링

1) 학습 데이터셋의 생성

모델로 표현하고자 하는 결과에 알맞은 학습 데이터셋을 생성한다. 모델이 예측해야 할 것은 현재까지의 고객의 금융 상태 정보들을 통한 앞으로의 행동 양상이다. 우리는 이를 위해 기준월(-0)을 기준으로, 전월(-1)을 현재 시점으로 한 학습용 데이터 셋을 생성하여 활용하였다. 이는 기준월(-0)과 전월(-1)의 변화로 고객별 민감도 라벨링을 진행하여 예측해야 할 결과를 만들고, 전월(-1)시점에서 획득할 수 있는 특징들을 입력 특징으로 사용한다.

칼럼명	급여-2	급여-1	수신-2	수신-1	...	전문직여부	급여이체여부	수신좌수	...
비고	2 달 전	전월	2 달 전	전월	...	변화없음 추정	변화없음 추정	변화없음 추정	

위의 표는 전월(-1) 시점에서 획득할 수 있다고 판단한 특징들 중 일부이다. 전월(-1)과 2달 전(-2)의 데이터와 함께 기준월과 전월에서 차이가 없을 것으로 판단되는 전문직여부, 자동이체건수와 같은 특징을 추가하여 사용하였다.

예를 들어, 2019년 데이터의 경우 12월(-0)이 기준월이다. 기준월과 전월의 변화인 11월(-1)~12월(-0)의 변화로 민감도 라벨링을 진행한다. 이후 입력 특징으로는 전월(-1) 시점에서 획득할 수 있는 특징들인 10월(-2), 11월(-1) 데이터 및 12월(-0)의 데이터이지만 변화가 없을 것으로 추정되는 데이터들을 추가하여 사용한다.

또한, 모델이 현 시점의 경제 상황에 대해 함께 학습할 수 있도록 하기 위해 현 시점의 데이터에 대한 정보들을 추가하여 준다. 이는 같은 시점의 샘플들에는 차이가 없고, 시점이 달라질 경우에만 차이가 발생한다.

2) train set과 test set의 분리

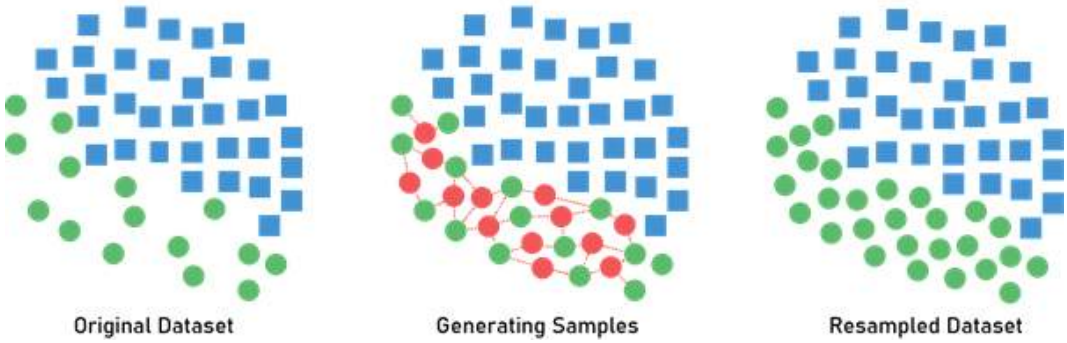
시계열 데이터를 활용함에 따라, 시간에 따른 데이터 축적 및 미래참조 문제를 고려하기 위해 2019, 2020년 데이터를 train set으로 두고 2021년 데이터를 test set으로 두고 학습 및 평가를 진행하였다.

Train		Test
2019	2020	2021

현재는 3개의 분리된 시점의 데이터만 주어졌지만, 더 많은 시점의 데이터를 활용한다면 모델의 성능은 개선 될 수 있을 것으로 기대한다.

3) 오버샘플링

Train set에서 민감도 별 0의 비율은 90%, 1의 비율은 10%으로, 매우 불균형한 데이터 셋이다. 따라서 학습이 제대로 이루어지지 않을 가능성이 높다. 이를 해결하기 위해 over sampling 기법을 활용하여 1과 0의 비율을 맞추어 학습을 시켜주었다.



SMOTE(synthetic minority oversampling technique)는 데이터의 개수가 적은 클래스의 표본을 가져온 뒤 KNN기법을 통해 임의의 값을 추가하여 over sampling하는 기법이며, Python의 SMOTE package를 통해 구현하였다,

4) 모델 선택

민감도를 분류하기 위해 5가지 형태의 모델을 활용해 비교분석하였다. 딥러닝을 제외한 4가지 분류 모델은 각각의 민감도에 대하여 분류 모델을 만들어 5개 모델을 학습하였으며, 딥러닝 기반 모델은 Sigmoid 레이어를 통한 멀티 레이블 분류를 구현하여 단일 모델이 5가지 민감도를 한번에 분류해낼 수 있도록 구성하였다.

1. 로지스틱 회귀분석

로지스틱 회귀분석은 시그모이드(sigmoid)함수를 기초로 구성된다(Shalev-Shwartz and Ben-David, 2014). 선형회귀와 달리 0과 1로 분류되므로 이진분류 알고리즘으로 유용하게 사용 된다. 시그모이드 함수는 식은 다음과 같다.

$$\Phi_{sig}(z) = \frac{1}{1+\exp(-z)}$$

로지스틱 회귀는 독립변수( )와 회귀계수( )에 관한 선형 예측함수를 기초로 한다. 이에 따라 독립변수( )가 주어졌을 때 종속변수( )가 1의 범주에 속할 확률을 계산한 로지스틱 회귀는 아래와 같다.

$$P(Y=1|X_1^{i)} = \frac{1}{1+\exp(B_0 + B_1X_1 + \cdots + B_iX_i)}$$

분류 모델에서 아주 간단한 형태의 모델로, 타 모델과의 벤치마크로 활용하기 위해 해당 모델을 선택하였다.

2. XGboost

XGboost는 트리 기반 앙상블 기계학습 알고리즘이다(T. Chen and C. Guestrinm, 2016). 그래디언트 부스팅(Gradient Boosting)이라는 기술을 사용한다. 부스팅은 약한 분류기(weak learner)를 세트로 묶어 정확도를 예측하는 기법이다. 경사하강법을 적용하여 약한 예측 모형들의 학습 오차에 가중치를 두고, 순차적으로 다음 학습 모델에 반영하여 손실을 최소화하는 강한 예측모형을 만드는 것이다. 과적합(Overfiiting)을 방지하는 규제 기능이 있어 강한 내구성을 갖으며, 조기 종료(Early Stopping)을 제공한다.

트리 기반 앙상블 모델로, 정형 데이터 분야에서 높은 성능을 기대할 수 있으므로 해당 모델을 선택하였다.

3. 랜덤포레스트

랜덤포레스트는 다수의 의사결정나무가 모여 숲을 이룬 형태로 모델을 학습하는 방법이다(Breiman, 2001). 숲을 구성 하는 방법은 배깅(Bootstrap Aggregating, Bagging)을 기초로 한다. 배깅은 부트스트랩(Bootstrap)을 통해 조금씩 다른 Train 데이터셋을 여러 개 생성하여 훈련하고 결합시키는 방법이다.

XGBoost와 같은 트리 기반 앙상블 모델이지만, Gradient Boosting이 아닌 배깅을 사용하는 모델로, 앙상블 방법에 따른 모델 성능 차이를 관측하기 위해 선택하였다.

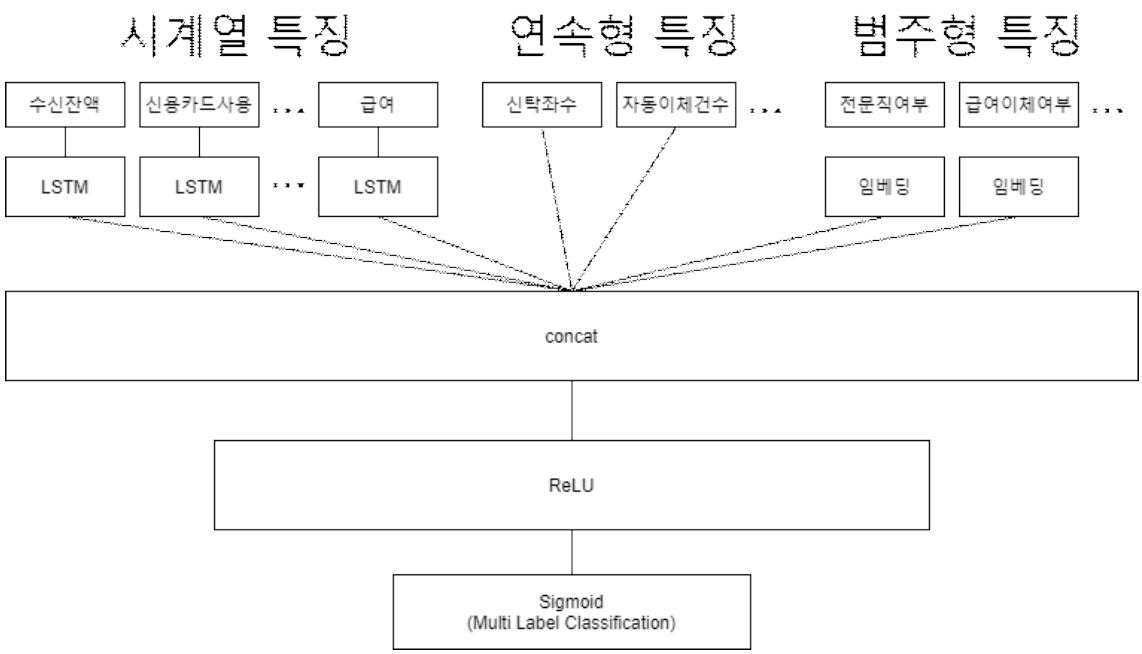
4.LightGBM

LightGBM 알고리즘은 의사 결정 트리 알고리즘에 기반한 고성능의 알고리즘으로 순위 또는 분류를 위한 기계학습 작업에 사용되고 있다. LightGBM은 트리의 깊이(depth wise)나 균형 트리(level wise)로 분할하는 다른 부스팅 알고리즘과 달리, 가장 잘 맞는 트리의 리프중심(leaf wise)으로 분할한다. 따라서 LightGBM에서 동일한 리프(leaf)에서 성장할 때 리프(leaf)방식 알고리즘은 균형 레밸(level)방식 알고리즘보 다 손실을 더 줄일 수 있으므로 기존 부스팅 알고리즘으로는 거의 달성하기 어려울 만큼의 정확도를 달성 하면서도 매우 빠른 수행이 가능하다.

XGBoost와 비교하여 훨씬 빠른 학습 속도로 비슷한 수준의 성능을 달성할 수 있는 모델로, 빠른 학습 속도가 필요한 경우 활용을 고려할 수 있도록 하기 위해 선택하였다.

5.딥러닝

시계열을 포함한 다양한 형태로 존재하는 개인의 금융 상태를 잘 표현할 것으로 기대되는 딥러닝 모델을 구현하여 학습하였다.



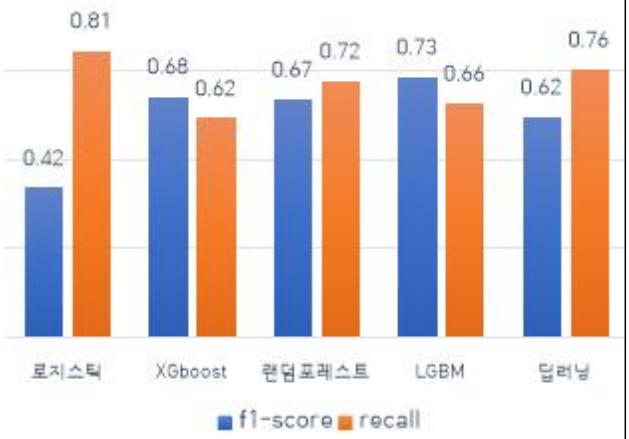
위는 분석에서 활용한 딥러닝 모델의 구조이다. 개인의 금융 상태를 나타내는 입력 데이터들을 크게 시계열, 연속형, 범주형 3가지로 나누어 처리하였다.

시계열 특징들은 각기 단방향 LSTM을 통과시키고, 범주형 특징들은 원 핫 인코딩하여 사용한다. 연속형 특징들은 아웃라이어들을 윈소라이즈 시켜준 후, 그대로 사용한다. 이후 모두 이어 붙인 후 분류를 위한 레이어를 통과시켜 예측한다.

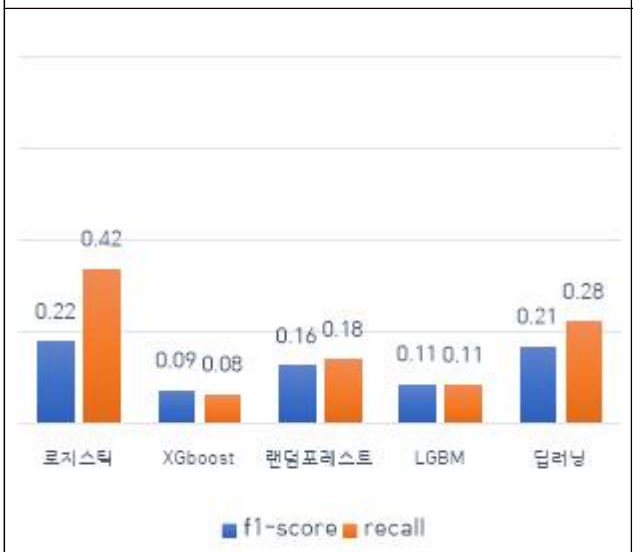
개인의 금융 상태 정보는 다양한 형태의 데이터로 존재하고 있으므로, 각기 다른 데이터의 특성을 보존한 모델을 적용하기 위해 선택하였다.

Ⅲ. 분석결과

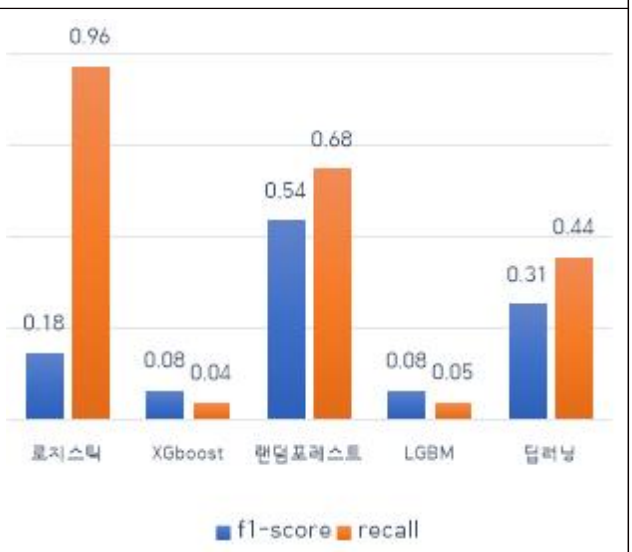
CPI



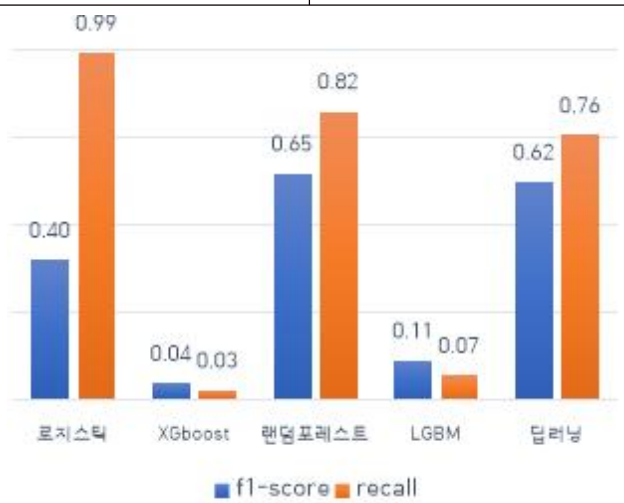
고용



제조업



비제조업



환율

CPI와 고용률에 대해선 모델 전체적으로 잘 분류하는 모습을 보이나, 제조업 BSI는 대체로 잘 분류하지 못하고, 비제조업BSI와 환율은 모델별 성능 편차가 발생하는 모습을 확인할 수 있었다.

제조업의 경우 현재의 입력만으로는 분류하기 힘든 최적화를 위한 데이터가 부족했을 가능성이 있다. 더 다양한 시점의 데이터를 이용할 수 있다면 문제점을 확인하여 성능을 개선시킬 수 있을 것이다.

비제조업 및 환율의 경우 XGBoost와 LGBM, 공통적으로 Gradient Boosting방식을 이용하는 모델의 성능이 좋지 않게 나타났다. 모델에 앙상블 기법을 적용할 시 부스팅 방식은 피하는 것이 좋을 것으로 사료된다.

또한, 데이터의 시점 수가 아쉬운 현재로서는 배깅 앙상블 기법을 적용한 랜덤 포레스트 모델이 전체적으로 좋은 성능을 나타내고 있다. 따라서, 우리 분석에서는 랜덤 포레스트 모델을 최종 모델로 선정하였다.

하지만 현재 딥러닝 모델에서는 LSTM의 인풋으로 길이가 2인 아주 짧은 시계열을 사용한 점 등 개선의 여지가 많이 존재한다. 이를 실제 데이터의 많은 시점의 데이터와 긴 기간의 시계열을 활용하여 학습을 수행한다면 딥러닝 모델의 성능을 많이 개선시킬 수 있으리라 생각한다.

마지막으로, 딥러닝 모델의 경우 분류 레이어를 거치기 전의 결과는 개인의 금융상태를 표현하고 있는 벡터로서 활용할 수 있다고 있다고 기대할 수 있기에 이를 확장하여 경제 변수의 민감도 분류작업 뿐 아니라 신용평가 모델과 같이 개인의 금융상태에 대한 표현이 필요한 작업에서도 전이학습하여 적용하면 간편하게 좋은 결과를 얻을 수 있으리라 생각한다.

Ⅳ. 활용방안

▢ 적용부문

1) 고객별 맞춤 금융상품 추천 서비스

우리의 분석대상인 ‘경제변수에 따른 민감 고객 분류’ 작업이 끝나게 되면 이를 이용하여 고객별 맞춤 금융상품 추천 서비스를 진행할 수 있다. 상품 추천 프로세스는 다음과 진행된다.

<고객별 맞춤 금융상품 추천 서비스 프로세스>

1. 모델에 의한 경제변수 별 민감 고객 분류

랜덤포레스트	CPI	고용률	제조업BSI	비제조업BSI	환율
1	77007	57941	60626	75869	77439
0	432993	4421059	439374	75869	422561

	전문직여부	세분화고수신고객여부	세분화고소득고객여부	삼성기존고객우대구분코드	멀카슈랑스보유좌수	수익증권좌수	신탁좌수	연간상품가입건수	급여이체여부	자동이체건수	수신좌수
count	53919.000000	53919.000000	53919.000000	53919.000000	53919.000000	53919.000000	53919.000000	53919.000000	53919.000000	53919.000000	53919.000000
mean	0.010015	0.295925	0.654204	1.239062	0.413305	0.664775	0.246351	8.195664	0.365715	18.097016	6.669115
std	0.099574	0.456462	0.475632	1.329100	1.312997	3.454236	0.992112	5.953490	0.481635	18.743281	6.998161
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	1.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	4.000000	0.000000	8.000000	3.000000
50%	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	7.000000	0.000000	15.000000	5.000000
75%	0.000000	1.000000	1.000000	2.000000	0.000000	0.000000	0.000000	10.000000	1.000000	24.000000	8.000000
max	1.000000	1.000000	1.000000	5.000000	33.000000	242.000000	49.000000	83.000000	1.000000	1746.000000	260.000000

메인모델이었던 딥러닝 모델 대신 성능이 가장 좋았던 랜덤포레스트를 통해 민감 고객을 분류하였다. 고용률 변수를 예시로 들어 설명하자면 우선 분류 모델을 통해 고용률 경제변수에 '1'로 라벨링된, 즉 고용률에 민감하게 반응하는 고객들을 추출한다.

2. 경제변수 별 민감 고객들의 다음 시기 금융 행동변화 파악

우리는 모델링 과정에서 상관관계를 통해 설정된 방향에 따라 1로 라벨링된 고객들의 다음 시기 금융행동을 예측할 수 있다.

"고용률이 상승할 시, '수신잔액', '대출잔액' 칼럼은 상승, '현금서비스 이용금액' 칼럼은 하락"

3. 추천 상품군 정의

이후에는 정성적 판단에 따라 상품 추천 알고리즘을 설계한다. 위와 같이 고용률이 상승한다면 모델에 의해 다음 달 수신잔액과 대출잔액은 상승할 것이며, 단기대출인 현금서비스이용은 하락할 것이라는 것을 알 수 있다. 이에 의해 추천 상품군은 다음과 같이 정성적으로 정의할 수 있다.

(1) 수신잔액이 늘면 저축을 위한 여유자금이 증가하는 고객이라고 가정하여 '예금,

적금 상품’ 을 추천 상품군에 추가

(2) 고용률이 상승흐름이라면 실물경기 상태는 좋다고 판단할 수 있으며, 경기가 좋을땐 단기대출보다 미래를 구체적으로 설계하여 장기대출을 활용할 고객이 더 많을 것이라는 가정 하에 장기대출상품을 추천 상품군에 추가

물론 이는 단순 예시일 뿐이며 실무에서는 더 합리적인 사고로 다양한 상품군 추천이 가능하다.

4. 민감고객군 세분류 작업 진행

예금, 적금, 대출이라는 추천 상품군이 정의되면 이후에는 ‘고용률’ 이라는 변수에 민감한 고객들을 대상으로 세분류 작업을 진행한다. ‘세분화 고소득 고객여부’, ‘가계자금 대출잔액’, ‘외화예금잔액/해외현금서비스 금액’ 등 기존 고객의 특성을 나타내는 칼럼들을 이용하여 고객 세분류 작업을 진행하며 이후 최종적으로 고객별 특성에 맞게 상품을 추천해준다.

5. 고객별 맞춤 상품 추천

이처럼 경제변수를 통해 실물경제 상황에 적합한 금융상품을 각 경제변수에 민감한 고객 개개인의 특성에 맞추어 추천해준다면 시의적절한 금융상품 추천 서비스를 제공할 수 있을 것이다.

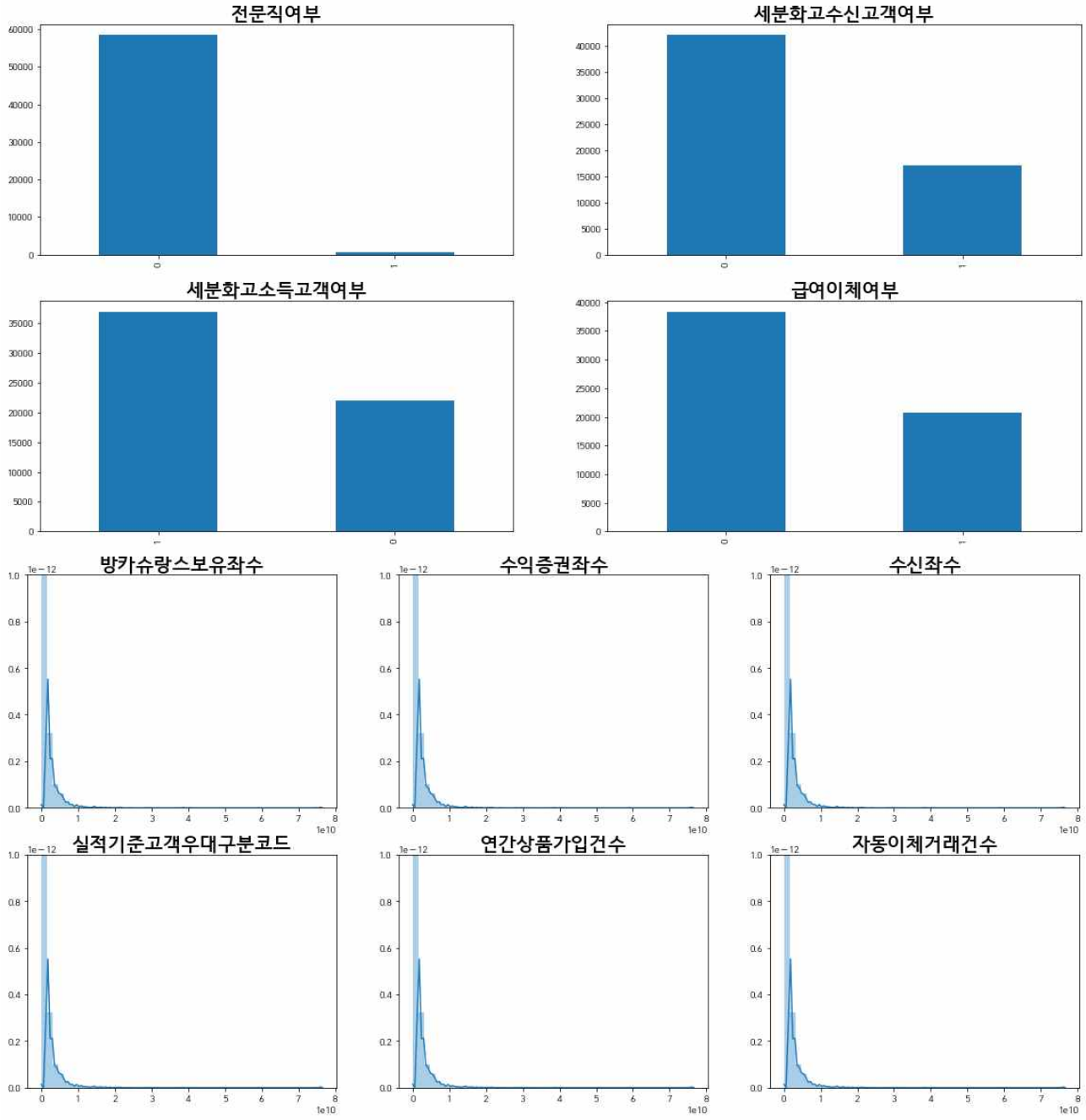
▢ 정책 활용 방안

1) 맞춤형 정책 제안 및 홍보

우리의 분석모델은 경제변수마다 민감한 고객들을 분류하여 그들의 특성을 파악할 수 있도록 한다. 공공영역에서는 ‘목표 시민 계층 맞춤 정책 제안’ 이나 ‘특정 계

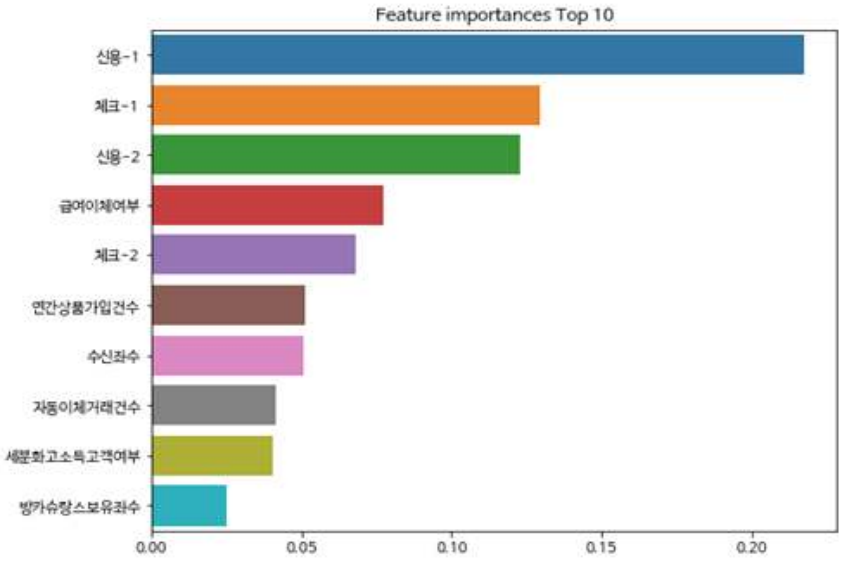
층 맞춤 홍보방안 수립’ 과 같은 분야에서 활용할 수 있을 것으로 기대된다.

‘목표 시민계층 맞춤 정책 제안’ 을 위해 우선 목표 계층 시민들의 특성을 파악함으로 분석을 시작할 수 있다. 주어진 데이터셋인 대구은행고객데이터를 통해 고용에 민감하게 반응하는 고객들에 대한 특성을 살펴보면 다음과 같다.



고용에 민감한 고객들은 대부분 전문직이 아니며 수익증권좌수 및 상품가입이 없는 경우가 많다. 이를 통해 해당 계층은 금융상품 활용이 부족하다는 추론을 할 수 있다. 이를 타개하기 위한 정책을 고안한다면 이들을 위한 금융투자 교육을 집중적으로 실시하는 방안이 있을 것이다. 물론 이는 데이터셋 부족에 기인하여 성급하게 일반화한 추론이지만 데이터분석을 위한 표본과 특성을 나타내는 칼럼이 많을수록 목표 계층에 대한 정교한 분석과 정책제안이 가능할 것으로 기대된다.

-고용률 민감도 모델의 Feature importance



또한 모델에 영향을 준 변수들을 확인하는 방법인 feature importance를 활용한다면 경제 정책 수립 과정에서 시민들의 금융행동에 주로 영향을 주는 변수들이 무엇인지 파악함으로써 효과적이고 유효성이 있는 적절한 정책을 제안할 수 있을 것이다.

Ⅴ. 참고자료

한국은행 대구경북본부, '최근 대구경북지역 주력산업의 수출 동향 및 시사점(2020.3)'

오픈서베이.'금융 트렌드 리포트(2020.11)'

KIET(산업연구원), 변창욱 외 3인,'대내외 경제변수의 지역경제 영향 및 파급경로 분석(2017.12)'

경기연구원, 황상연,'경기도 단기지역경제전망 모형 구축에 관한 연구(2010)'

대구경북연구원, '대구경제동향(2021.04)'

이현미 외 2인, 'LightGBM 알고리즘을 활용한 고속도로 교통사고심각도 예측모델 구축.', 한국전자통신학회 논문지, (2020): 1123-1130

윤우진 외 3인, 'RandomForest와 XGBoost를 활용한 유방암 종양 분류.', 한국통신학회 학술대회논문집 . (2021): 113-114.

이형탁 외 3인, '머신러닝 분류 알고리즘을 활용한 선박 접안속도 영향요소의 중요도 분석', 해양환경안전학회지, (2020)

김은빈 외 3인, 'Two-dimensional attention-based multi-input LSTM for time series prediction', Communications for Statistical Applications and Methods, (2021)