붙임 3 **과제 분석 계획서 작성 양식**

접수번호	※작성하지 않음

「제3회 대구 빅데이터 분석 경진대회」 과제 분석 계획서

신청자명(팀명)	GoDART
분석주제명	경제 변수를 활용한 초개인화 서비스 제안

※ **5장 내외** 자유형식으로 작성, 박스의 목차는 준용하되, 필요시 변경 가능

1. 기획 배경 및 개요

1) 기획 목적/배경 및 필요성

4차 산업혁명의 영향으로 각 산업에서의 디지털 전환속도는 점점 빨라지고 있고 은행업 또한 이러한 추세에 발맞춰가고 있다. 은행들은 지점 축소, 비대면 어플 활용 등을 통해 체제의 전환을 이루고 있으며, 지난해 전 세계로 확산된 코로나19는 이러한 추세를 가속화시켰다. 한편, 저성장으로 인한 장기간 저금리 환경과 최근 두드러진 집값 폭등으로 인해 정기 예금 이용률은 점점 감소하고 있다. 이자가 적은 예금을 대신하여 주식, 대출을 비롯한 금융상품 활용을 통해 금융소득을 늘리려는 수요가 많아졌기 때문이다. 자본주의 시대에 금융에 대한 관심 증대는 분명 시민들이 더 계획적이고 현명한 삶을 사는데 도움을 줄 것이다. 다만, 금융상품 활용 기간이 길지 않은 사람들에게 시중에 존재하는 다양한 금융상품은 오히려 혼란을 줄 수 있다. 이러한 상황에서 금융과 관련된 비대면 맞춤서비스가 활성화된다면 금융상품을 찾아보고 선택하는데 있어 소비자의 혼란과 기회비용을 줄여 시민들의 편의를 증가시켜 줄 것이다. 더욱이 올해 8월 시행될 마이데이터 사업은 금융시장 참여자에 대한 방대한 데이터를 제공하여 고객 맞춤 서비스를 더욱 활성화시킬 수 있는 환경을 마련해줄 것이다.

한편, 글로벌화의 심화로 인해 전 세계 경제의 공동화 현상은 심화되고 있다. 황상연(2010)은 경기도 지역을 대상으로 지역 및 국가적 요인을 고려한 FAVAR 분석을 통해 금리, 환율, 유가 등 대내외 경제 충격이 다양한 지역경제변수에 영향을 미침을 보였다. 또한 변창욱 외 3인(2017)은 대내외 경제변수의 지역경제 영향 및 파급경로를 분석하며 경제의 연결성이 강화되는 상황에서 지역경제의 안정적인 운용을 위해서는 대내외 환경변화를 주시해야한다고 하였다. 이러한 선행연구들을 통해 시민들의 금융 활동에는 지역 대내외 경제변수가 많은 영향을 미친다는 것을 알 수 있다. 따라서 거시경제 변화와 금융시장 참여자 행동 사이를

분석한다면 생활 방식에 따른 금융 행동 양상을 예측할 수 있을 것으로 생각하였다. 특히 지역거점 은행으로서 대구은행의 고객 데이터는 대구시민들의 금융 활동 특성을 잘 대변해 줄 것으로 예상된다. 따라서 거시경제 변수와 지역경제 변수를 포함한 금융 데이터 활용을 통해 개인별 맞춤 서비스를 제공한다면 금융 서비스의 질적 수준을 높여 대구시민의 금융 활용 능력 제고에도 도움이 될 것으로 생각하여 본 기획을 구상하게 되었다.

2) 기획의 독창성 및 차별성

• 주로 금리라는 경제변수에 초점을 맞추는 기존의 은행 산업의 특성과 달리 다양한 경제변수를 활용하였다.
• 경제변수와 딥러닝 모델을 활용하여 은행 고객들의 행동을 사전에 예측함으로써 적절한 서비스 추천을 할 수 있다.
• 마이데이터 사업 시행으로 인해 더 많은 고객의 다양한 금융 데이터를 활용한다면 모델의 정확도를 높여 더 정밀한 추천을 할 수 있다.
• 코로나19와 같은 예상치 못한 경제충격이 발생하더라도 결국 거시경제 변수의 변화로 나타나기 때문에 이에 따른 시민들의 금융 행동 패턴을 예측하고 분석할 수 있다.

2. 분석 내용 요약

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

3. 분석 방법 및 계획

1. 활용데이터

우리가 분석에 활용데이터는 다음과 같다.

1.1. 주최 측 제공 데이터

데이터명	형식	데이터기간	사용변수	출처
대구은행 고객데이터	csv	2018.01~2020.12	전체 컬럼	대구은행
유동인구 소블럭 기준	csv	-	전체 컬럼	SK 텔레콤
통신사 생활인구	csv	2017.01~2018.12	전체 컬럼	SK 텔레콤
통신사 유동인구	csv	2019.01~현재	전체 컬럼	SK 텔레콤

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 컬럼	통계청

2. 데이터 확보 계획

: 주최 측이 제공하는 데이터에 더하여, 빅데이터 활용센터, 한국은행, 통계청의 데이터를 이용한다.

3. 데이터 분석 환경 및 도구

1) 분석에 사용하는 언어 : 파이썬 3.7 이상(sklearn, tensorflow, pytorch 등)
2) 데이터 전처리 및 시각화 및 결과 보고서 : pandas, numpy, matplotlib 등
3) 데이터분석 최종 목표 : 고객별 경제변수에 따른 민감도 측정

4. 데이터 전처리

ㄱ. 대구은행 고객데이터

: 대구은행 고객데이터는 대부분 NUM형태의 값을 가지고 있다. 하지만 일부는 VARCHAR형이다. VARCHAR형 데이터 중 대부분은 고소득 고객 여부, 고수신 고객 여부, 전문직 여부 칼럼처럼 '여부'란 이름을 가지는 칼럼들은 True/False 형태로 추정되어 One-hot encoding 하여 활용할 계획이다.

하지만, 최다거래점번호, 실적기준고객우대구분코드는 따로 처리할 필요가 있다고 판단된다. 실적기준고객우대구분코드는 범주형 데이터지만, 등급 간에 상하관계가 존재한다고 판단되어 임의의 수치형 변수로 변환하여 활용 예정이다.

최다거래점번호는 개인별 주요 활동 지역을 나타내는 중요한 변수라 생각되지만, One-hot encoding하여 활용하기엔 차원의 수가 필요 이상으로 늘어날 것으로 예상된다. 그러므로 변수의 특성을 살리면서 차원을 줄일 방법이 필요하다. 따라서 유동인구 소블럭 기준, 통신사 생활인구, 통신사 유동인구 데이터를 활용하여 최다거래점번호와 연결시켜 개인별 주요 활동 지역의 유동인구로 치환하여 활용할 계획이다.

ㄴ. 유동인구 소블럭 기준, 통신사 유동인구, 통신사 생활인구

: 유동인구 소블럭 기준은 시군구/행정동/법정동에 대한 대구 소지역 위치정보데이터를 가지고 있다. 먼저 소지역_코드를 주키로 하여 레코드를 결합하여 대구시 각 지역들에 대한 상세한 정보를 획득한다.

이후 위에서 생성한 릴레이션을 활용하여 통신사 유동인구, 통신사 생활인구 릴레이션을 결합한다. 통신사 유동인구의 시계열은 17년~18년, 통신사 생활인구의 시계열은 19년~현재이므로 겹치는 부분이 존재하지 않는다. 중심이 되는 대구은행 고객 데이터의 시계열은 18년~20년으로 두 데이터 셋을 이용하여 전체 시계열 구간에 대한 최다거래점번호를 치환이 가능할 것으로 기대한다.

그리고 인구 관련 데이터는 시간대별로 획득할 수 있으므로, 단순히 평균으로 활용하는 것은 물론 오전, 오후, 야간으로 그룹화하여 활용하는 방안도 고려한다.

ㄷ. 추가 활용할 공공・민간데이터(거시경제 변수, 대구지역 경제변수)

: 거시경제 변수로 원/달러 환율, 원/위안화 환율, 국고채(3년), 회사채(3년, AA-) 금리를 사용한다.

대구지역 경제변수로는 대구 고용지표의 실업률과 고용률, 대구 기업경기실사지수(BSI)의 제조업과 비제조업 업황, 대구소비자물가지수를 사용한다.

이들은 모두 일정한 주기로 발표되는 수치형 시계열 데이터로, 발표일을 고려해 미래 참조 문제를 해결하고 대구은행 고객 데이터와 시계열만 일치시킨 후 사용한다.

거시경제 변수는 매일 발표되므로, 1일 lagging하여 활용한다. 대구지역 경제변수는 변수마다 발표일이 다르며 대구 고용지표는 내월 중순, 기업경기실사지수(BSI)는 당월 말, 소비자물가지수 내월 초 등 다양하다. 이들은 1달 lagging하여 활용한다.

단, 국고채(3년)과 회사채(3년, AA-)의 경우 개별 지표로서 사용하기 보다는 이들의 금리차를 사용할 것이다. 즉 무위험채권인 국고채와 회사채의 차이를 통해 신용스프레드를 만들어 경기 파악을 위한 지표로 활용한다.

5. 경제변수 선정사유

: 한국은행 대구경북본부에서 발간한 '최근 대구경북지역 주력산업의 수출 동향 및 시사점(2020.3)'에 따르면 대구경북지역의 주요 수출국(2019년 기준)은 중국(26.8%), 미국(16.8%) 순이며 이 두 국가에 대한 수출 비중(43.6%)은 전국 평균(38.6%)에 비해 높은 편이다. 그만큼 다른 국내 지역에 비해 미국, 중국 통화의 영향을 많이 받는다는 것을 추론할 수 있었고 따라서 대구지역에 영향을 주는 대외변수로서 원/달러, 원/위안화 환율을 선정하였다. 또한 금융 활동에 가장 큰 영향을 미치는 금리 또한 대외변수로 선정하였는데, 변동이 적은 한국은행 기준금리 대신 시중금리를 사용하기로 결정하였다. 그 중 대출과 밀접한 관련이 있는 COFIX(자금조달비용지수) 금리를 활용하려 하였으나, 이전 달의 금리를 다음달 15일에 공시해주는 COFIX금리의 특성상 딥러닝에 활용할 시 미래 참조 이슈가 발생하기 때문에 매일 발표해주는 금리들 중에서 선정하기로 결정하였다. 최종적으로 국고채와 회사채 차이를 통해 간접적으로 경기상황을 알려주는 신용스프레드를 활용하기로 결정하였다.

대구지역 내 경제변수 선정을 위해서 대구의 경제를 크게 고용, 산업, 소비로 나누어 보는 것이 중요하다고 판단하였고, 이에 따라 각 항목에 맞게 지표를 선정하였다. 고용 관련 지표에는 대구 고용지표를, 생산 관련 지표에는 대구 기업경기실사지수(BSI)를, 마지막으로 소비 관련 지표로서 대구소비자물가지수를 활용하여 대구 경제를 반영하는 지표로 활용할 예정이다.

6. 모델 학습을 위한 데이터 셋 생성

ㄱ. 입력변수(x)

: 대구은행 고객 데이터의 대부분을 위 ㄱ, ㄴ의 전처리 과정을 거친 후 대부분 활용한다. 이는 고객 개개인을 대변하는 특징이며 결과적으로 이를 통해 딥러닝 모델을 학습시키고 또한 미리 개인들의 행동 방향을 예측할 중요한 지표로 활용된다.

ㄴ. 레이블(y)

: 우리의 최종적인 목표는 고객 개인별로 경제변수에 대한 민감도를 측정하는 것이다. 이에 앞서서 개인별로 경제변수에 대한 민감도를 레이블링 할 필요가 있다.

우리는 이를 위해 대구은행 고객데이터의 기간 변화율에 거시경제 변수, 대구지역 경제변수와 그랜저 인과검정을 수행하여 각 경제변수별로 인과관계를 갖는 칼럼을 도출할 것이다. 이 과정에서는 인과검정으로 1차적인 선택 후, 각 변수별로 정성적인 판단 기준을 도입해 인과관계를 갖는 칼럼을 최종 결정할 것이다.

이후 기간 변화율과 각 변수별로 인과관계를 갖는 칼럼들이 어느 정도 변화하였는지 관측하여 고객마다 경제변수에 대한 민감도를 레이블링할 수 있게 된다.

예를 들자면, 환율이 하락하였을 경우 외화잔고가 n% 이상 증가한 고객들은 환율에 민감한 고객들로 구분할 수 있을 것이다.

이러한 레이블링 결과는 개인의 가치관이나 경제 상황, 시장에 대한 관심 등으로 비롯된 것이라 추정할 수 있고, 이는 현 시점의 개인의 재무상태로 드러날 것이라 기대한다.

그렇다면 개인의 재무상태 및 특성들을 가진 대구은행 데이터에 딥러닝 모델을 도입하여 개인별로 경제변수들에 대한 민감도를 예측할 수 있을 것이다.

7. 모델링

위에서 생성한 학습용 데이터 셋을 통한 지도학습을 수행하며, 개인 고객마다 경제 변수에 갖는 민감도를 예측하기 위해 분류 문제로 접근한다. 베이스라인 모델로 MLP 분류 모델을 사용할 것이며, 랜덤 포레스트를 위시한 다양한 머신러닝 분류 모델도 적용하여 성능을 비교하여 본다. 또한 생성한 학습용 데이터 셋 역시 시계열을 활용할 수 있다는 특징이 있으므로, 시퀀스 데이터를 분류에 활용할 수 있도록 RNN기반 알고리즘을 적용하는 방안 역시 고려한다.

기본적으로 각 경제변수에 대한 이진 분류 모델을 고려하지만 고객들을 좀 더 세분화하기 위해 고객들에 대한 민감도 수준을 상/중/하 와 같이 분류하는 멀티 레이블 분류 문제로 확장시킬 수 있을 것이다. 이 경우 판단 기준을 좀 더 세분화하여 레이블링 후 Softmax로 고객 민감도를 예측하게 된다.

그리고 경제 변수별 민감도에 대한 고객 수의 분포가 정규분포를 따른다고 가정하면, 민감도가 높아질수록 확률밀도는 줄어들 것이다. 모델의 목적은 이 소수의 사람들을 보다 정확하게 구분해 내는데 있기에 모델의 성능은 정밀도와 F1-Score를 이용하여 평가한다.

4. 분석 결과 활용 및 시사점

코로나 팬데믹 상황 속에서 거시경제지표는 크게 변화되었고, 그에 따라 사람들의 행동에서도 특정한 패턴이 두드러지게 나타난다. 한국은행이 발표한 2021년 1분기 중 가계신용(잠정)에 따르면, 올 1분기 가계대출은 1년 새 144.2조가 증가하였으며 역대 최대치를 기록하였다고 한다. 또한 경기 회복세에 따른 인플레이션이 우려되는 시점에서, 금리가 빠르게 인상된다면 고객들의 부채 상환 부담이 크게 증대될 수 있다. 따라서 리스크관리 측면에서도 금융기관은 거시경제 흐름에 따른 고객 행동 패턴을 파악하는 것이 그 어느 때보다 중요해 보인다.

이러한 시점에서 우리는 거시경제 변수에 따른 고객의 행동 패턴을 분석하여 금융기관에서의 고객 맞춤 상품 개발 및 리스크 관리를 돕고, 시민들은 필요한 금융상품을 추천받아 편리한 금융 생활을 할 수 있도록 돕고자 한다. 이번 분석과 관련하여 금융기관과 시민, 대구광역시 모두에게 이익을 가져다줄 구체적 활용방안과 기대효과는 다음과 같다.

첫째로, 초개인화된 금융서비스의 발전이 가능해진다. 위의 분류 모델을 통해 각 거시경제 변수에 민감하게 반응하는 고객들을 예측할 수 있고, 나아가 거시경제 상황에 따른 행동 패턴을 예측할 수 있다. 예를 들어, 금리에 민감한 집단은 금리가 하락함에 따라 요구불예금 잔액이 줄고, 가계자금대출 잔액이 늘며, 수익증권 좌수가 증가한다는 특성을 가진다고 가정하자. 이러한 상황에서 다음 달 금리가 하락할 것이라고 예상된다면, 금리에 민감하다고 예측된 집단에게 가계대출 상품과 수익증권 상품을 소개하는 맞춤형 금융서비스를 제공할 수 있을 것이다. 또한 환율에 민감하다고 예측된 집단에겐 환전서비스 혹은 외화자유적금 상품을 추천해줄 수 있을 것이고 더 나아가면 고객 맞춤형 상품 개발 또한 가능할 것이다. 시민들은 이러한 상품을 추천받음으로써 편의성이 증대되고 개인 맞춤 혜택을 누릴 수 있다.

특히 마이데이터 사업 시행이 예정됨에 따라 앞으로 활용 가능한 데이터의 범위는 증대될 것이다. 여수신, 금융투자 범위를 넘어 보험, 카드, 전자금융, 통신 등 여러 분야의 정보가 활용 가능해진다면 기존의 모델에 확장된 데이터들을 기반으로 더욱 다양하고 정교한 고객

행동 패턴을 찾아내어 초개인화된 금융서비스를 제공할 수 있을 것이다.

둘째로, 지역 경제 발전에 이바지할 수 있다. 분석 결과는 대구광역시 경제 현안 파악과 경제 정책 수립에 활용이 가능할 것이다. 거시경제 변수별로 예측된 고객의 비율을 통해 대구시민들이 민감하게 반응하는 거시경제 지표에는 어떤 것이 있는지 파악할 수 있다. 또한 고용 정책 수립 시 대구시 고용 지표에 민감하게 반응하는 집단의 특성을 참고하여 정책 수립이 가능할 것이다.

이 외에도 거시경제 및 지역경제 흐름에 따른 시민들의 행동 패턴을 파악함으로 경제 위기 대응 시에도 활용할 수 있을 것이며. 금융취약계층 파악에도 활용이 가능할 것이다. 특히 대구시에서 금융취약계층을 위한 교육을 위해 DGB금융그룹과 협약을 맺고 지원에 힘쓰고 있는 것으로 알고 있다. 이번 분석에서 금융취약계층의 특성을 파악할 수 있다면, 금융투자상품에 대한 이해가 부족한 취약계층을 도와 지역경제 불평등 해소에도 기여할 수 있을 것으로 기대된다.

이처럼 대구시민과 금융기관과 대구시 모두의 차원에서 유익을 가져다줄 수 있기에, 이번 기획을 성공적으로 마무리하여 지역경제 발전에 이바지할 수 있기를 바란다.

5. 참고문헌 출처 등

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)’