2021 금융 데이터 경진대회

(주관: 금융보안원)

지역 및 업종 Embedding을 활용한

새로운 지역 경제 지표 제시

: NMF, NCF, GMF 모델을 활용한 Representation Learning

팀명: 범내려온다

이성범

김동현

이수민

목차

2021 금융 데이터 경진대회 결과 보고서

□ 요약

최근 들어 지역별 경제 상황에 따른 차별화된 정책 및 대출 정책 시행과 더불어 지역별 금융 위험도 측정에 대한 중요성이 대두되고 있다. 또한, 코로나19 장기화에 따라 은행별 취약 업종에 대한 여신 건전성에도 변화가 생기는 등 업종별 금융 위험도를 평가하는 것이 이전보다 중요한 과제로 자리 잡고 있다. 이러한 상황 속에서, 정부의 정책 시행과 금융기관의 고객 유치 및 개인의 신용평가에 활용할 수 있는 새로운 지역경제 지표를 제안한다.

국가 전체의 경제 상황을 대변하는 지표는 지역의 재정과 경제시책 수립에 필요한 자료를 제공할 수 없다는 한계점이 있다. 따라서, 이를 해결하기 위해 한 나라의 지역별 경제 상황을 파악하는데 대표적으로 지역내총생산(GRDP) 지표가 활용된다. 다만 GRDP는 완결성과 활용도가 현저하게 떨어져 지역 경제 정책 수립에 활용하는 데에 있어서 한계를 보인다. 따라서 기존 지역 경제 지표의 단점 및 한계점을 보완하면서 기업의 여신 심사에 도움이 될 수 있는 새로운 지역 경제 지표를 개발하였다.

새롭게 제안하는 지역경제 지표에는 지역별 통계 데이터와 신한카드에서 제공하는 광역시도별 업종별 가맹점 데이터가 사용되었다. 딥러닝을 활용하여 새로운 지역경제 지표를 제시하고 각 기업의 영업 이익, 매출액을 기반으로 금융 위험도를 반영한 target 변수를 생성한 뒤 임베딩을 학습시켰다. 지역별 통계 데이터에서는 시군구명, 업종대분류명, 업종중분류명, 업종소분류명의 임베딩을 학습시켰고, 광역시도별 업종별 가맹점 데이터를 활용할 때는 광역시도명, 업종대분류, 업종중분류, 업종소분류의 임베딩을 학습시켜 임베딩 값 자체적으로 Target을 반영하는 그 지역, 업종의 새로운 지표로써 활용할 수 있도록 하였다.

이때 target 변수는 두 가지 방법으로 생성되었다. 먼저 군집화를 활용한 Target 생성 방법을 통해 매출, 영업 이익 관련 변수를 활용하여 계급화를 진행한 뒤 각각 가맹점과 매출을 반영하는 Label 값을 만들어 최종 Target 변수를 생성했다. 다만 이를 위험도 지표로 사용하기에는 편향성이 크다고 판단되어 결과적으로 사용하지 않고 위 방식을 개선하고 변동성을 고려할 수 있는 CV 값을 이용하는 새로운 방법을 고안해냈다. 최종적으로 변동성을 고려한 방법을 활용하여 만든 target 값을 가지

고 임베딩을 진행하였다.

변동성(변동계수(CV))을 고려한 Target 변수 생성 방법에서는 각 데이터를 기준년월별로 분리하고 시군구별 및 기업구분별 변동계수와 업종별 및 기업구분별 변동계수를 활용하여 cv를 산정했다. 이때 기업 구분에 따른 영업이익중위액을 활용하였으며 이를 기반으로 최종 target 값을 생성하였다. 광역시도별 업종별 가맹점 데이터를 활용하여 생성한 Label도 위와 동일한 방법으로 진행되었다. Target 값은 점당 매출금액과 광역시도별 CV, 업종별 CV 값을 모두 활용하여 만들어졌으며 광역시도와 업종 Label을 최종적으로 더한 뒤 정렬하여 최종 Label을 완성했다.

분석에 활용한 모델은 다음과 같다. MF(Matrix Factorization)의 방식을 활용하여 Target을 예측하는 지역과 업종 간의 상호작용을 나타내는 Latent Space를 구하고 지역의 Latent Factor을 지역경제 지표로 활용하는 방식을 사용하였다. 다만 MF(Matrix Factorization)의 방식을 활용하면 변수 간의 복잡한 상호작용을 저차원 공간에 표현하기 어렵기 때문에 MF 방식 대신 변수 간의 복잡한 상호작용을 표현하기에 표현할 수 있는 DNN 기반의 MF 방식을 사용하였다.

3가지 이상의 임베딩을 학습할 수 있는 GMF(General Matrix Factorization) 방식을 활용하였으며, GMF에 비선형성이 추가된 형태인 NCF 모델 또한 복잡한 상호작용도 학습할 수 있기에 본 분석에 적합한 모델이라고 판단하여 활용하였다. 추가로 안정적인 학습이 이루어지는 NMF 모델을 사용해보았지만, GMF와 NCF에 들어가는 임베딩을 다르게 설정하여 학습시킴으로써 하나의 변수에 2개의 임베딩을 얻게 되어 하나의 지역경제 지표가 필요한 본 Task에는 부합하지 않은 모델이라고 판단하였다. 결과적으로 NCF 모델 사용 시 가장 좋은 성능을 얻을 수 있었다.

생성한 지표에 대한 검증을 위해 여러 외부 데이터를 활용하였다. 먼저 전력 사용량은 그 지역의 공장들이 얼마나 가동되는지 대변해줄 수 있는 지표로써, 새로운 지역 경제 지표와 전력 사용량에 대한 상관관계를 파악하였고, 결과적으로 0.47이라는 높은 상관성을 발견했다. 두 번째로 가장 대표적인 지역경제 지표인 지역내총생산(GRDP)의 성장률과 생성한 지표 사이에 상관관계가 있다면 생성한 지표가 타당하다고 판단할 수 있을 것이라 가정했다. 결과적으로 GRDP의 성장률 역시 생성한 지역경제 지표와 0.41로 높은 상관성을 나타내는 것을 확인했다. 따라서 새로운 지역 경제 지표가 지역의 경제성과 지역의 성장률을 모두 반영한다고 해석할 수 있다.

생성한 Target 값이 지역 및 업종별 위험도를 반영한다는 것을 설명하기 위해 생성한 지표를 활용하여 각 데이터셋에 대해서 시각화를 진행하였다. 광역시도별 업종별 가맹점 데이터(신한카드)로 생성한 Target 을 검증하기 위해 지역별로 뚜렷한 차이를 보일 것으로 예상되는 업종인 '면세점'의 Target 값들을 살펴보았을 때, 면세점 소비 매출이 높은 지역의 Target 은 높게 나타난다는 결과를 통해 생성한 Target 값이 지역 및 업종별 위험도를 반영한다는 것을 설명할 수 있다. 지역별 통계 데이터를 활용하여 생성한 Label의 경우, Target 값이 가장 큰 인천의 경우 위험도가 가장 낮은 기업 구분에 해당하는 대기업의 비중이 전체 매출 중 가장 크다는 것을 알 수 있다. 본 Target 을 신한은행 서울시 데이터와 연관 지어 보다 세세한 분석을 진행했다. 결과적으로 위험도가 낮다고 판단한 지역들에서 총입금은 높지만, 소비금액은 낮다는 특징을 확인할 수 있었으며, 이는 생성한 Target 이 지역의 금융 위험도를 구 단위에서도 반영한다는 것을 나타낸다.

새로운 지역경제 지표는 정부, 공공기관, 금융기관 등 다양한 곳에서 활용될 수 있다. 특히 기업의 위치와 업종을 임베딩 값으로 표현하여 재무 정보가 존재하지 않은 기업에 대한 여신 평가가 제대로 이루어지기 힘든 경우에도 지역, 업종 등 비재무적 정보를 통해 위험성 지표를 생성할 수 있다는 장점을 가진다.

정부는 각 지역에 대한 위험성과 재무 건전성 등을 파악할 수 있게 됨에 따라 예산 분배 전략을 비롯하여 정책 수립에 대한 방향성을 설정하고, 지역 경제 상황을 파악하는 지표로 활용할 수 있다. 또한, 지역별 업종에 따른 위험도를 반영하고 있는 본 지표를 활용한다면 업종에 대한 세부적인 정책을 수립하는 데에도 참고할 수 있다는 점에서 다양한 방면에서 활용할 수 있다. 마지막으로 미래의 경기 변동성을 예측해 지역 경제에 영향을 미치는 주요 요인 파악이 보다 쉬워진다면 더 나은 지역 발전 방향을 설정하고 구체적인 전략 수립이 가능해진다. 은행에서는 본 지표를 기반으로 위험 지역을 판단하여 지역에 자금 조달과 대출과 관련된 마케팅, 지역 은행에 대한 이자율 산정 등에 활용할 수 있다. 더불어 개인의 신용평가에 생성한 경제 지표를 활용한다면 거주지역이나 회사가 있는 지역의 위험성 및 재정건전성과 같은 정보를 반영할 수 있게 되면서, 기존 개인 신용평가 모형보다 더 신뢰도 높고 정교한 모형 수립이 가능하다.

1. 주제

기업의 신용리스크 파악을 위해 많은 금융기관들은 기업의 재무 정보를 활용한다. 하지만 모든 기업에 대한 재무 정보를 가지고 있는 금융 기관은 한정적이다. 또한, 실제 업무에서 활용되는 한국기업데이터를 통해서도 기업의 재무 정보에 대한 데이터가 대부분 결측치로 저장되어 있다는 것을 확인할 수 있다. 따라서 재무정보가 존재하지 않는 기업의 신용리스크 파악을 위한 새로운 변수가 필요하다.

우리는 이를 해결하기 위한 새로운 지역경제지표를 제시한다. 실제 한국기업데이터에서도 기업의 재무정보는 결측치이지만 지역, 업종 등의 정성적 지표는 결측치가 아닌 경우가 대다수이다. 따라서 우리는 지역, 업종 등의 정성적 지표를 실제 기업들의 신용리스크 파악을 위한 변수로써 활용하기 위해 정량적 지표로 만드는 과정을 진행했다.

영업이익중위액, 점당매출액 등의 정량적 지표를 활용해 Target 변수를 만들었고, Target 변수는 전력사용량, GRDP와의 상관관계가 존재하였으며, EDA를 통해서도 각 지역의 위험도를 대체적으로 반영하고 있다. 이러한 Target 변수를 지역, 업종 등의 임베딩을 통하여 예측하는 GMF, NCF, NMF 모델을 구축하였다. 각 모델의 임베딩은 서로의 복잡한 상호작용을 표현하는 형태로 학습된다. 지역의 임베딩은 수치로써 표현되기 때문에 정량적 지표로 활용될 수 있고, 이는 그 지역의 위험도를 반영하는 지역경제지표로써 활용이 가능하다.

우리가 만든 지역경제지표는 재무정보가 없는 기업들의 정량적 평가지표로써 활용될 수 있다. 또한, 지역의 경제 상황을 한눈에 보여주는 지표로 활용되어 정부와 공공기관에서 지역 경제 상황을 파악하는 데 도움을 주어 지역 균등 발전을 도모할 수 있을 것이다. 그리고, 기존의 존재하는 지역 경제지표나 은행의 여신 평가 모델 등에 활용되어 더 정교한 지역 및 개인의 위험성을 판단할 수 있다.

2. 배경 및 필요성

[기존 경제 지표의 문제점]

한 국가의 경제는 경제성장률, GDP(Gross Domestic Product: 국내총생산), 물가 상승률, 수출 및

수입액, 환율 등 다양한 주요 경제 지표를 통해 나타낼 수 있다. 국내총생산(GDP)은 일정 기간 (1년) 동안 한 국가 안에서 생산한 최종재, 즉 최종 생산물의 시장 가치를 합한 것으로 국가의 전반적인 생산활동 수준과 경제 규모를 나타낸다. 가장 널리 쓰이는 경제 지표인 GDP는 여러 경제학자들이 국가의 경제 성장의 척도로 삼을 만큼 중요한 지표지만, 국가의 세부적인 요소들을 반영하지는 못한다는 한계점이 있다. 우선 GDP는 총합계의 관점이기 때문에 세부적으로 국가 내의 어떠한 계층이 얼마만큼의 가치를 생산하느냐를 반영하지 못한다. 예를 들어 대기업이 50을 생산하고 중소기업이 50을 생산하는 경우와 대기업이 90을 생산하고 중소기업이 10을 생산하는 경우, 편중이 심해진 후자의 경우를 GDP에서는 반영하지 못하게 되면서 삶의 질이나 성장률 측면에서는 앞으로 분명히 영향을 미치게 될 이 격차를, GDP 지표 하나만으로는 알 수가 없다.

GNI(Gross National Income: 국민총소득)는 국민소득을 보다 정확하게 반영하기 위해 나온 경제 지표로, 일정 기간 동안 가계, 기업, 정부 등 국민 경제가 국내외 생산활동에 참여한 대가로 벌어들인 소득을 모두 포함한다. GNI가 GDP와 가장 크게 다른 점은 자국민이 국외로부터 받은 소득, 즉 국외수취요소소득을 지표 계산에 포함시킨다는 점이다. 다만 GNI는 전체 국민소득의 크기를 나타내기 때문에 한 나라의 경제 규모를 파악하는데 유용하나, 국민들의 평균적인 생활수준을 알아보는 데는 적합하지 못하다는 한계점이 있다. 개인들의 실제 소득 상황이나 실질 구매력을 나타내지 못하고 계층 간 소득 분배 상태를 보여주지 못한다는 한계도 지니고 있다. 이는 명목 GNI를 인구수로 나눠 구하는 1인당 GNI를 통해 어느 정도 해결이 가능하지만, 여전히 국민들의 실제 생활 수준을 정확히 파악하기에는 지표가 일반 국민들이 체감하는 소득 수준이나 계층별 소득 분배 상태 등을 대변해 주지는 못한다.

GDP과 GNI는 국가 전체의 총체적인 경제 상황을 알려줄 뿐 지역별 경제 수준을 나타낼 수 없다는 한계점이 있다. 다만 한 나라의 지역별 경제 상황을 파악하고 비교하는 것과 비롯하여 지역의 재정과 경제 정책 수립에 필요한 자료를 제공하는 것의 중요성이 갈수록 높아지고 있기 때문에 국가 전체에 관한 내용보다는 국가 내 지역별 자료의 활용에 있어서 지역별로 나라의 경제 성장과 현황을 보여 줄 수 있는 지표의 필요성이 대두되고 있다.

[기존의 여신 평가 모델]

과거에는 기업 신용리스크를 파악하기 위해서 주로 특정 기업의 재무 현황이나 매출 현황 등을 통해 회사의 정량적 정보만을 파악하는 데 집중을 하였다. 그동안 기업 여신 심사 과정은 개별 심사역이 협의체를 구성해 진행해왔다. 재무제표를 통해 해당 기업의 현황을 파악하는 업무 외에도 경기 동향 및 업황, 인허가 제한 여부, 환경 이슈 등 정량적으로 다루기 쉽지 않은 정보들을 종합적으로 검토할 필요가 있기 때문이다. 따라서 이로 인한 비용이 발생하고, 기업의 정량적 데이터 이외의 법적, 환경적 문제 등 정성적 데이터에 대한 부분을 제대로 파악할 수 없었기 때문에 정교한 신용리스크 판단 모델을 구축하기 힘들었다.

그러나 데이터가 방대해지고 인공지능 시스템이 발달함에 따라 정성적 데이터에 대한 활용 가능성이 높아졌고, 이에 따라 많은 은행에서 기업 여신 자동심사 시스템 개발을 위한 투자를 진행하고 있다. 자동심사 시스템 도입하게 된다면 시스템이 기업 여신 심사 과정 상당 부분을 대신하게 된다. 빅데이터, AI 등 최신 기술들을 활용해 종합적인 판단이 필요하지 않은 부분을 시스템화하고 효율적이고 신속한 의사결정을 가능하게 할 수 있다.

신한은행은 2019년 2월 기업 여신 자동심사모형 CSS 도입을 완료했다. 우리은행은 2019년 9월 무렵 내부신용등급을 보유한 기업을 대상으로 일정 대출금액 이하 여신 심사와 여신승인을 자동화했다. KB국민은행은 지난해 3월 산업 및 업황 정보와 기업 재무 및 비재무 정보를 자동으로 수집, 심사하는 기업 여신 자동심사 지원시스템(Bics)을 오픈했다.

이러한 상황 속에서 기업의 재무제표 너머에 있는 정보 중 대표적으로 활용할 수 있는 요인은 지역에 대한 지표이다. 지역은 실제로 돈이 움직이는 공간이고, 기업이 있는 지역의 물가, 토지, 경제 상황에 따라 기업의 위험성이 변동될 수 있기 때문이다. 지역에 대한 위험성을 스코어링 한다면 기업의 환경적 요소를 파악할 수 있고, 그 지역의 영업점에서 미리 위험성을 판단할 수 있어 여신에 대한 대비가 가능하다. 또한, 기업의 건전성 수준을 판별할 때 스코어링으로 반영할 수 있으며, 기존에 구축된 신용평가 등급 체계를 토대로 자산규모, 업종 등 개별 기업 특성에 따라 심사지표 적용 여부 및 지표별 스코어링 적용 수준 등을 차별화할 수 있다.

[지리적 위치의 중요성]

미국 IT산업의 메카 실리콘 밸리, 미국 영화 산업의 메카 할리우드, 대한민국 IT산업의 메카 판교, 대한민국 방송 산업의 메카 상암 등 우리는 지역을 통해서 해당 지역의 특성을 파악할 수 있다. 이처럼 하나의 지역에 비슷한 기업들이 모이는 것을 산업 클러스터라고 한다.

Michael Porter는 기업들이 산업 클러스터를 형성함으로써 생산성을 높일 수 있고, 현장에서 혁신을 주도할 수 있으며, 현장에서 새로운 비즈니스를 자극받을 수 있다고 했다. 이렇듯 많은 기업들에게 지리적 위치는 지속 가능한 경쟁우위를 달성하기 위한 중요한 요소 중 하나이다.

또한, 투자자의 입장에서도 기업의 지리적 위치는 매우 중요한 요소 중 하나이다. 기업의 위치와 실제 이익조정과의 관계와 기업의 지리적 위치가 배당 의사결정에 미치는 영향이라는 논문들을 보면 기업의 지리적 위치가 기업의 투자, 회계정보에 대한 수요 등에 영향을 미친다는 결과를 보여준다. 따라서 기업의 위험도를 평가할 때 지리적 요소를 반영하여 더 타당성 있는 평가를 할 수 있다.

[현재 지역 경제 지표에 대한 문제점]

최근 들어 지역 경제 간의 격차 해소는 국민의 삶의 질 향상과 균등한 정책환경 조성과 같은 복합적 차원의 국가 문제와 직결되기 때문에 반드시 극복하여야 할 과제로 대두되고 있다. 이에 다양한 방법으로 지역의 경제 수준이 표현되고 있는데, 보편적으로 KOSIS 국가통계포털에서 제공하는 지역경제상황판을 통해 지역별 생산과 소비, 물가, 고용 지수를 파악할 수 있으며, 지역 내의 경제활동 수준을 평가하기 위한 가장 대표적인 지표로는 지역내총생산(GRDP)이 활용되고 있다.

지역내총생산(GRDP)은 일정 기간 동안 특정 지역 내에서 경제활동별로 얼마만큼의 부가가치가 발생되었는가를 나타내는 경제 지표로, 우리나라에서는 광역자치단체 및 기초자치단체 단위의 GRDP가 작성 및 공표되고 있다. GRDP는 각 지역에서 경제 활동별로 부가가치를 반영하기 때문에 지방자치단체의 경제성과 등을 평가하고 지역 간 비교를 가능하게 함으로써 국민 경제의 지역적 분석과 지역개발시책 수립에 활용되고 있다. 뿐만 아니라 지역 경제구조를 개량적으로 파악하여 해당 지역의 개발계획을 수립하거나 지역경제의 장기 예측 등에 사용될 수 있는 지표다.

이렇듯 GRDP가 지역경제를 평가하는 데에 있어서 중요한 역할을 하고 있음에도 불구하고, 실제로는

여러 한계점으로 인해 활용도가 높지 못하다는 점이 문제점으로 지적되고 있다. 첫 번째로 GRDP 공표의 시의성 문제가 있다. 현재 통계청의 광역자치단체 GRDP(확정치)는 작성 기준 연도 18개월, 기초자치단체 GRDP는 작성 기준 연도 24개월 이후 공표되고 있는데, 특히 후자의 경우 공표에 있어서 작성 시점과 공표 시점 간의 차이가 커 이를 실제로 활용하는 것이 타당한지에 대한 의문이 제기되고 있다. 2년의 시차 안에서 분기별 GRDP를 알 방법조차 존재하지 않아 시의성 있고 지속적인 통계 모형 개발 및 수정에 어려움이 발생하며, 결국 실질적인 활용도가 떨어진다는 치명적인 단점이 지적되고 있다. 뿐만 아니라 기초자치단체 단위로 생산되는 통계량이 충분하지 않아 기초자료가 부족하고, 특정 인력에게 업무가 몰리기 때문에 통계 자료 자체의 정확성에 대한 신뢰도가 떨어진다는 평을 받고 있다. 추가로 GRDP가 생산 측면에서만 작성되고 분배 및 지출 계정은 작성되지 않다는 점 또한 지표 자체에 대한 완결성과 활용도를 떨어뜨리는 요인이 되며, 지역 경제 정책 수립에 활용하기에는 한계가 있다.

[지역 경제 지표와 관련된 최근 연구 동향]

최근에는 기존의 지역 경제 지표에 대한 문제점을 해결하기 위해 다양한 연구가 진행되고 있다. 현재 생산 측면에서만 측정되고 있는 시·군의 1인당 GRDP는 분배(소득)와 지출 측면을 대변하기 어렵다는 점과 직주분리가 상당한 상황에서 GRDP를 (생산에 기여한 종사자와 다를 수밖에 없는) 거주자 수로 나누어 계산한다는 점이 가장 큰 문제점으로 나타나고 있다. 그 결과 현행 1인당 GRDP는 생산 측면의 경제력뿐만 아니라 분배와 지출 측면의 경제력도 제대로 대변하지 못하는 결과를 초래하고 있다.

이러한 문제점을 해소하기 위하여 지역 생산과 분배(소득)에 대한 경제력 지표를 새롭게 제시하였다. 1인당 지역 생산의 지표로는 종사자 기준의 1인당 GRDP를 제시하여 지역 산업의 생산성을 판단할 수 있게 하였고, 1인당 지역 분배(소득)의 지표로는 거주자 1인당 지방소득세 종합소득분을 제시하여 개별 시·군의 소득 수준을 가늠할 수 있게 하였다.

경기도 31개 시군의 자료를 분석한 결과, 기존 지표인 1인당 GRDP는 일반인들의 인식과 상당한 차이를 보였으나, 새로운 지역 생산과 소득의 지표는 지역의 산업 활동과 거주 특성 등을 상당 부분

반영하는 것으로 나타났다. 그리고 새로운 지표는 기존 지표와 별도로 측정할 필요가 있다는 것을 일반 상관계수 분석과 순위 상관계수 분석을 통하여 확인하였고, 회귀 분석을 통하여 새로운 지표가 기존 지표보다 개별 시·군의 경제 상황을 더 잘 설명하고 있다는 것을 확인하였다.

이처럼 지역 경제 지표에 관한 연구는 활발히 일어나고 있으며 우리는 이러한 기존의 지역 경제 지표를 보완하면서 기업의 여신을 심사하는 데 도움이 될 수 있는 지역 경제 지표를 개발하고자 한다.

3. 아이디어 제안 및 분석 결과

3-1. 데이터 설명 및 전처리

(3-1-1) 데이터 설명

- 지역별통계데이터(한국기업데이터)
  - ✓ 기업의 영업 이익을 바탕으로 Target 변수를 생성했다.
  - ✓ Target 변수를 바탕으로 시군구명, 업종대분류명, 업종중분류명, 업종소분류명의 임베딩을 학습 진행했다.
  - ✓ 임베딩 그 자체가 Target을 반영하는 그 지역, 업종의 새로운 지표로써 활용이 가능하다.
- 광역시도별 업종별 가맹점 데이터(신한카드)
  - ✓ 기업의 매출을 바탕으로 Target 변수를 생성했다.
  - ✓ Target 변수를 바탕으로 광역시도명, 업종대분류, 업종중분류, 업종소분류의 임베딩을 학습을 진행했다.
  - ✓ 임베딩 그 자체가 Target을 반영하는 그 지역, 업종의 새로운 지표로써 활용이 가능하다.

(3-1-2) 데이터 전처리

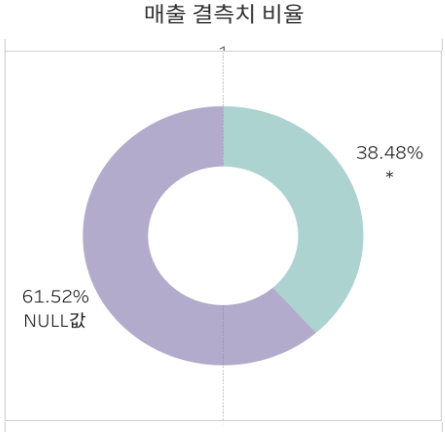
우리의 목적은 Target에 대한 정보가 없을 때 단순히 광역시도명, 시군구명, 업종명 등의 텍스트를 바탕으로 해당 지역 업종의 위험도를 예측하는 것이다. 따라서 우리는 과거의 데이터를 바탕으로 미

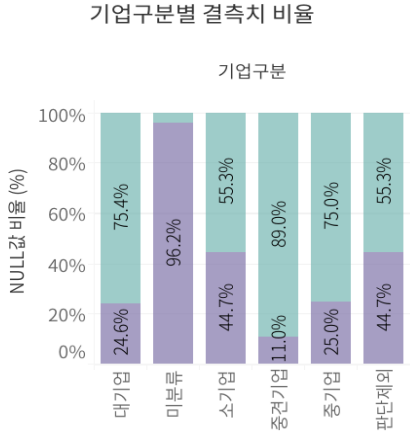
래를 예측하는 방식으로 데이터 셋을 분리했다.

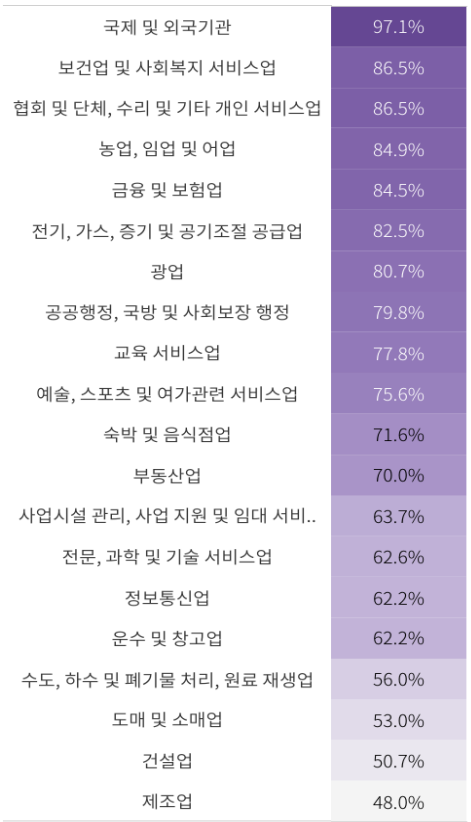
- 지역별통계데이터(한국기업데이터)
  - ✓ 가장 많은 결측치가 존재하는 데이터로, 주요 가설 증명에 활용되었다. 대부분의 현실 속의 데이터는 이처럼 그 기업에 대한 수치적 정보를 알 수 없는 결측치가 매우 많이 존재하는 데이터가 비일비재하다. 따라서 우리가 제시한 임베딩을 활용하는 지역경제지표는 이러한 문제를 해결하는 것에 도움을 줄 수 있다.
  - ✓ 결측치를 Test로, 가장 최신의 데이터인 2020-12를 Val로 나머지 데이터를 Train으로 활용했다.
- 광역시도별 업종별 가맹점 데이터(신한카드)
  - ✓ 가장 최신의 데이터인 202009를 Test로, 202003을 Val로, 나머지 데이터를 Train으로 활용했다.

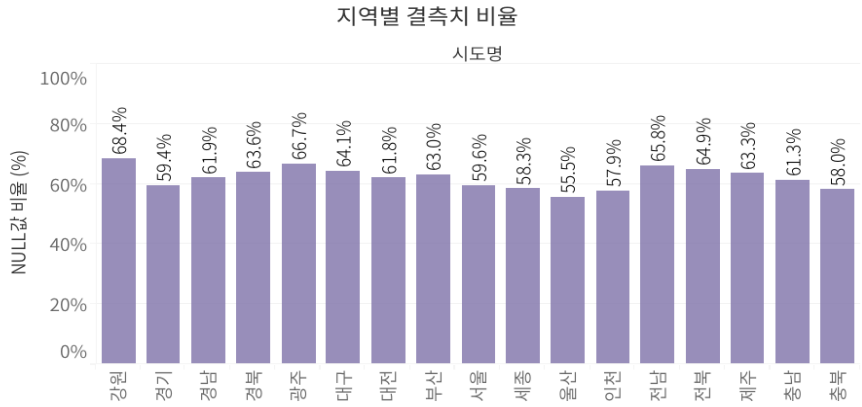
3-2. 탐색적 데이터 분석 (EDA)

[수집되지 않는 기업 및 개인의 재무 정보]









[그림 1. 제공받은 지역별 통계 데이터 결측치 현황]

이번 경진 대회에서 제공한 지역별 통계 데이터에서는 전체 약 104만개의 데이터 중에서 약 60만개의 데이터가 기업에 대한 기본적인 정보만 있을 뿐 영업이익과 같은 재무적 요소에 대한 값이 비어있는 것을 확인할 수 있다.

기업 재무 정보가 없는 기업들은 주로 미분류로 파악된 기업들이었으며, 그다음으로 소기업과 판단제외가 많은 것을 볼 수 있다. 또한, 업종별로는 국제기구, 사회복지 서비스업, 협회, 농업 등이 결측치가 많은 것을 볼 수 있다. 이를 종합해보면, 주로 소규모 기업이나 정보가 파악되지 않은 기업 혹은 업종의 규모가 작은 기업들이 결측치가 많다는 것을 알 수 있다.

이처럼 기업의 재무적 정보를 받을 수 없는 상황에서는 은행에서의 기업에 대한 여신 평가가 제대로 이루어지기 힘들다. 하지만 그렇다고 해서 재무적 정보를 파악할 수 없는 기업에 대해 단순히 무시하고 넘어가기에는 위에서 보았듯이 전체 기업의 60%를 고객 자원으로 확보할 수 없게 된다. 따라서, 이러한 재무적 정보를 확인할 수 없는 기업에 대한 위험성 지표를 지역, 업종 등 비재무적 정보를 통해 만들어 더 많은 고객을 확보하기 위한 노력이 필요하다.

3-3. Target 변수 생성 및 타당성

EDA를 통해 살펴본 것처럼 재무 정보가 결측치인 데이터가 많기 때문에 지역과 업종 데이터 만으로 결측치를 가장 잘 대변해 줄 수 있는 Target 변수를 선정하였다. 재무적 변수인 영업이익, 매출금액, 해지가맹점비, 대출금액 등 다양한 변수를 반영하여 Target에 따른 지역 및 업종을 비교해본 결과 지역별 통계 데이터에서는 영업이익중위액이, 광역시도별 업종별 가맹점 데이터에서는 점당매출금액이 가장 유의미한 결과를 나타났다. 따라서 각 데이터별로 영업이익중위액과 점당매출금액을 활용하여 군집화를 통한 방법과 변동계수(CV)를 고려한 방법을 통해 Target 변수를 생성 및 검증을 진행했다.

(3-3-1) 군집화를 활용한 Target 생성

- 지역별통계데이터(한국기업데이터)

✓ 각 데이터를 기준년월별로 분리했다. (년도마다의 영향력을 반영하기 위해)

✓ 데이터의 첨도가 심해서 classing 기법을 사용했다. (관련 자료: 빅데이터 분석 기법을 이용한 소상공인 신용평가 모형 구축 연구)

✓ Target 생성

- 매출, 영업 이익 관련 변수를 모두 계급화한다. (10개 그룹)

- 구해진 계급화 변수를 모두 더한다.

- 더해진 값을 100점 만점 기준으로 정렬한다.

● 광역시도별 업종별 가맹점 데이터(신한카드)

✓ 각 데이터를 기준년월별로 분리했다. (년도마다의 영향력을 반영하기 위해)

✓ k-means를 시행하기 전 로그 변환과 정규화를 시행했다. 로그 변환의 경우 데이터의 분포가 한쪽으로 치우쳐져 있어 이상치의 영향력을 줄이기 위해 사용했다. (실제로 로그변환시 정규분포와 비슷한 형태로 변화함) 정규화의 경우 k-means 알고리즘은 거리 기반이기 때문에 알고리즘을 제대로 실행하기 위해 사용했다.

✓ 가맹점을 반영하는 Label

- 해지가맹점수 / 매출가맹점수를 통해서 해지가맹점비를 구했다.

- 업종중분류 별로 각각 해지가맹점수, 매출가맹점수의 평균을 구하고, 두 값을 서로 나눠 업종별해지가맹점비를 구했다.

- 해지가맹점비 / 업종별해지가맹점비를 구해 해지가맹점의 비율을 반영하는 변수로 사용했다.

- 해지가맹점수, 매출가맹점수, 해지가맹점의 비율을 반영하는 변수를 가지고 k-means 알고리즘을 실행해 5개의 군집으로 나누었다.

- 군집을 각 군집 분석에 활용한 변수의 특성을 바탕으로 임의로 정렬을 했다.

✓ 매출을 반영하는 Label

- 카드매출건수, 카드매출금액, 점당매출금액, 건당매출금액을 가지고 k-means 알고리즘을 실행해 5개의 군집으로 나누었다.

- 군집을 각 군집 분석에 활용한 변수의 특성을 바탕으로 임의로 정렬을 했다.

✓ 구해진 두 Label을 서로 곱해 최종 Target 변수를 생성했다. 본 방식의 단점은 단순히 좋은 기업일 수록 좋게 평가하는 편향적인 추세를 보이기 때문에 위험도로 쓰기에는 편향성이 많이 포함되어 있어 사용하지 않았다. 이러한 방식을 개선하고 변동성을 고려할 수 있는 CV 값을 이용하여 Target 새롭게 변수를 생성했다.

(3-3-2) 변동성(변동계수(CV))을 고려한 Target 변수 생성

- 지역별통계데이터(한국기업데이터)
  - ✓ 각 데이터를 기준년월 별로 분리했다. (년도마다의 영향력을 반영하기 위해)
  - ✓ 대기업 / 중견기업 / 소기업 등 기업구분에 따른 영업이익중위액에 대한 분산이 커 기업구분에 따라 영업이익중위액을 나누었다.
  - ✓ 변동계수(CV)를 구할 때 시군구별 및 기업구분별 변동계수와 업종별 및 기업구분별 변동계수로 나누어 CV를 산정했다.
  - ✓ Target 생성
    - 영업이익중위액과 시군구별 CV, 업종별 CV 변수를 모두 계급화 했다. (10개 그룹)
    - 구해진 계급화 변수를 모두 더하고, MinMax Scaling을 진행 했다.
    - 더해진 값을 100점 만점 기준으로 정렬했다.
- 광역시도별 업종별 가맹점 데이터(신한카드)
  - ✓ 각 점포별 매출금액을 기준으로 Target을 생성하기 위해 점당매출금액을 활용한다.
  - ✓ 기준년월에 따른 점당매출금액에 대한 분산이 커 년도마다의 영향력을 반영하기 위해 기준년월에 따라 점당매출금액을 나누었다.
  - ✓ 변동계수(CV)를 구할 때 광역시도별 및 기준년월별 변동계수와 업종별 및 기준년월별 변동계수로 나누어 CV를 산정했다.
  - ✓ Target 생성
    - 점당매출금액과 광역시도별 CV, 업종별 CV 변수를 모두 계급화 했다. (10개 그룹)

- 구해진 계급화 변수를 모두 더하고, MinMax Scaling을 진행 했다.

- 점당매출금액과 CV에 대한 계급화 점수를 곱하여 광역시도와 업종에 대한 Label을 구했다.

- 광역시도와 업종 Label을 최종적으로 더하여 더해진 값을 100점 만점 기준으로 정렬한 했다.

3-4. 모델(NMF, NCF, GMF) 설명 및 비교

기존의 Matrix Factorization 방식은 어떠한 변수 간 관계의 상호작용을 내적을 통해서 Latent Space로 나타내어 Latent Factor를 찾는 것이다. 이러한 MF의 방식은 대부분 추천시스템에서 많이 활용되며 아이템과 유저 간의 상호작용을 나타내는 Latent Space를 구해서 추천에 활용된다.

우리는 이러한 MF의 방식을 적용하여 Target을 예측하는 지역과 업종 간의 상호작용을 나타내는 Latent Space를 구하고 지역의 Latent Factor를 역경제지표로 활용하는 방식을 제안한다.

하지만 단순한 MF 방식은 단점이 존재한다. 변수 간의 상호작용은 복잡한 구조로 이루어져 있는데 단순히 선형을 바탕으로 결합하는 내적은 이러한 복잡한 구조를 저차원 공간에 표현하기 어렵다는 점이다.

따라서 우리는 복잡한 구조를 더 잘 나타낼 수 있는 DNN 기반의 MF 방식을 사용하였다. DNN 기반의 MF 방식은 활성화 함수를 통해서 비선형 변환이 가능하기 때문에, 어떠한 변수 간의 복잡한 상호작용을 더욱더 표현하기 쉽다는 장점이 존재한다.

우리는 DNN 기반의 MF 방식을 Neural Collaborative Filtering(2017) 논문에서 제시한 GMF, NCF, NMF을 우리의 방식에 맞게 구현해 각각의 성능을 비교했다.

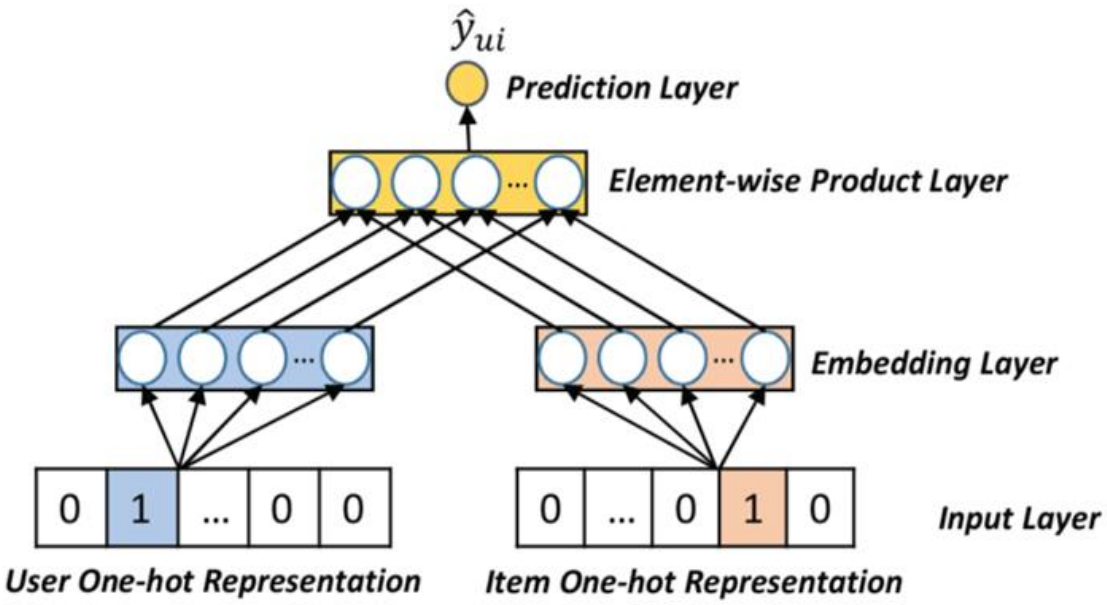
- 지역별통계데이터(한국기업데이터)
  - ✓ 텍스트인 시군구명, 기업구분, 업종대분류명, 업종중분류명을 64차원의 Embedding Layer로 나타냈다.
  - ✓ 각각의 Embedding Layer를 서로 element-wise product를 통해서 상호작용을 나타냈다.
  - ✓ 나온 값을 Output Layer에 통과시켜서 Target 과의 MSE가 최소가 되도록 각각의

Embedding Layer를 학습시켰다.

- ✓ 학습된 시군구명의 Embedding Layer를 Latent Space로 칭하고, Latent Space 속에 64차원의 Latent Factor를 지역경제지표로 활용했다.

- 광역시도별 업종별 가맹점 데이터(신한카드)
  - ✓ 텍스트인 광역시도명, 업종대분류, 업종중분류, 업종소분류를 64차원의 Embedding Layer로 나타냈다.
  - ✓ 각각의 Embedding Layer를 서로 element-wise product를 통해서 상호작용을 나타냈다.
  - ✓ 나온 값을 Output Layer에 통과시켜서 Target 과의 MSE가 최소가 되도록 각각의 Embedding Layer를 학습시켰다.
  - ✓ 학습된 광역시도명의 Embedding Layer를 Latent Space로 칭하고, Latent Space 속에 64차원의 Latent Factor를 지역경제지표로 활용했다.

(3-4-1) General Matrix Factorization(GMF)



[그림 2. GMF 모델 구조]

우리의 데이터는 학습해야 할 임베딩이 3가지 이상이기 때문에 기존의 MF 방식은 활용할 수가 없다. 하지만 GMF의 경우에는 element-wise product를 통해서 3가지 이상의 임베딩을 학습할 수 있기 때문에 본 방식을 활용했다. 또한 본 방식은 Latent Space를 찾음과 동시에 학습된 임베딩을 바탕으

로 Linear Layer를 통과시켜 Target을 예측한다는 측면에서 기존 MF 방식과 큰 차이가 있다. 따라서 GMF는 Latent Space로 나타낼 수 있는 지역경제지표를 찾고, 더불어 Target를 같이 예측한다는 측면에서 우리의 Task에 부합한다고 생각되어 본 모델을 구현했다.

(3-4-2) Neural Collaborative Filtering(NCF)

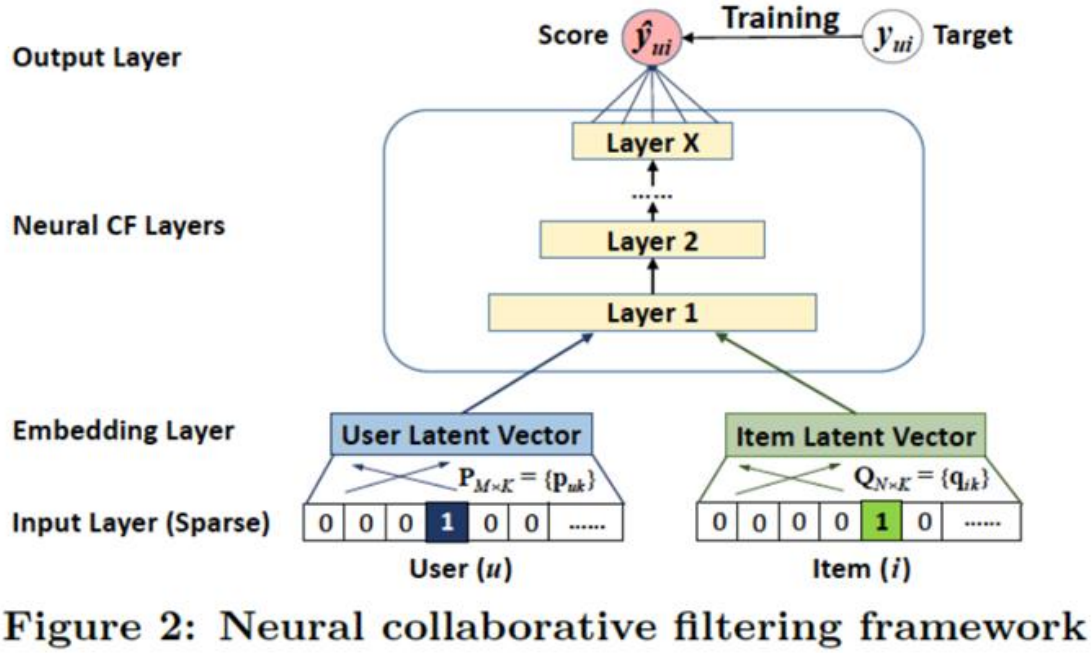


Figure 2: Neural collaborative filtering framework

[그림 3. NCF 모델 구조]

NCF 모델은 이전에 소개된 GMF에 비선형성이 추가된 형태의 모델이다. 우리가 표현하고자 하는 변수 간의 상호작용은 매우 복잡한 관계를 가지고 있다. 이전의 GMF 방식은 활성화 함수를 활용하지 않기 때문에 복잡한 상호작용을 capture 하기에는 어려움이 존재한다. 하지만 본 NCF 모델 같은 경우에는 예측을 는 과정에서 활성화 함수를 거치기 때문에 비선형성을 가질 수 있으며, 이러한 비선형성을 바탕으로 복잡한 상호작용을 조금 더 잘 capture 할 수 있다. 따라서 NCF 모델은 하나의 변수에 하나의 임베딩을 가지고 있으며, 복잡한 상호작용도 학습할 수 있기에 본 분석에 가장 부합되는 모델이다. 실제로 분석 결과 본 NCF 모델이 우리의 Task에서 가장 좋은 성능을 보였다.

(3-4-3) Neural Matrix Factorization(NMF)

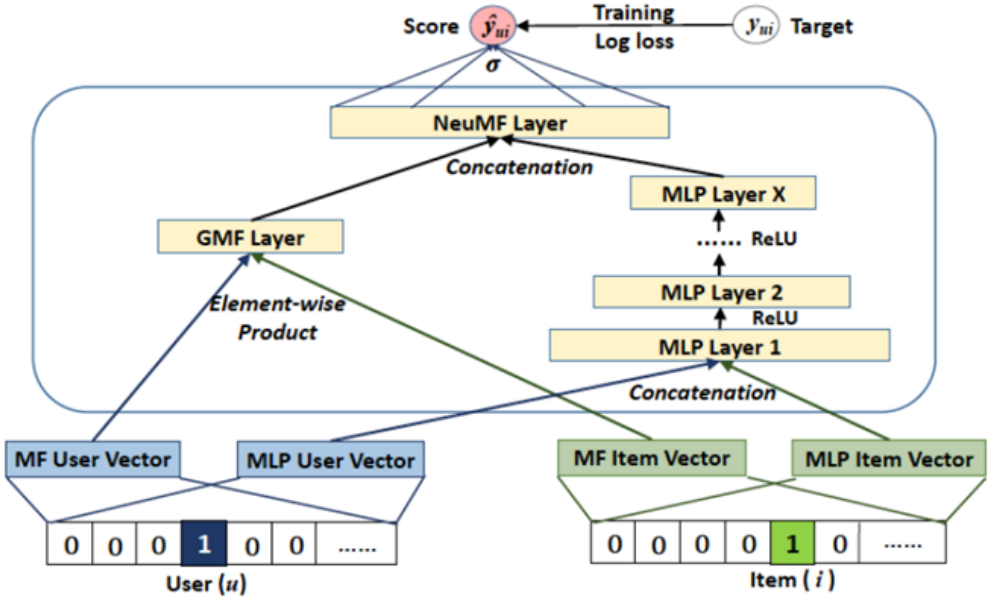


Figure 3: Neural matrix factorization model

[그림 4. NMF 모델 구조]

NMF는 학습이 안정적이지만 복잡한 상호관계를 Capture하는 것은 어려운 GMF와 학습은 안정적이지 않지만 복잡한 상호관계를 Capture할 수 있는 NCF를 서로 결합한 모델이다. 그러므로 상황에 따라서 가장 좋은 성능을 보일 수 있다. 하지만 NMF는 우리의 Task에는 부합하지 않은 모델이고 실제로도 분석 결과 좋은 성능을 보이지 않았다. 왜냐하면 GMF와 NCF에 들어가는 임베딩을 다르게 설정하여 학습시킴으로써 하나의 변수에 2개의 임베딩을 얻게 되기 때문이다. 따라서 하나의 지역경제지표가 필요한 우리의 Task에는 부합하지 않은 형태의 모델이다.

(3-4-4) 모델 결과

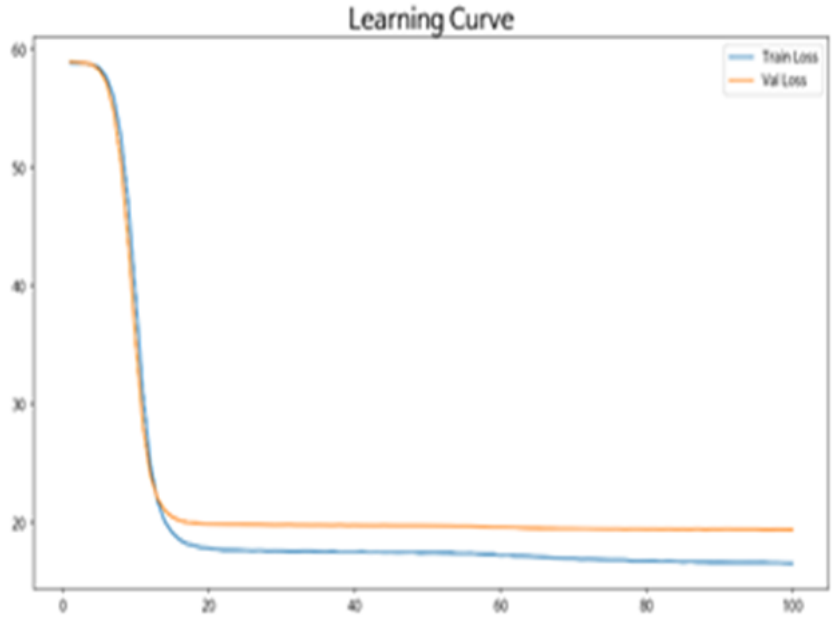
NCF, NMF 모델은 GMF와 달리 Layer에 활성화 함수를 포함하기 때문에 모델이 비선형성을 학습할 수 있다. 이에 우리가 임베딩 하고자 하는 지역, 업종 등의 복잡한 상호작용을 표현할 수 있어 NCF, NMF 모델이 GMF 모델보다 더 높은 성능을 보인다.

지역, 업종 등의 임베딩을 1개씩만 사용하는 NCF 모델과 달리 NMF 모델은 2개의 임베딩을 학습했다. 이에 NMF 모델은 각각의 임베딩이 다른 표현을 학습하게 되어 우리의 지역경제지표를

얻기 위한 우리의 Task와는 맞지 않은 형태이다. 따라서 하나의 임베딩을 통하여 지역, 업종 간의 복잡한 상호작용을 조금 더 구체적으로 학습하는 NCF 모델이 우리의 Task에 더 부합하며, 실제로도 가장 좋은 성능을 보인다는 것을 그림5를 통해 확인할 수 있다.

Model	지역별통계데이터	광역시도별데이터
General Matrix Factorization	30.72	19.38
Neural Collaborative Filtering	30.52	18.24
Neural Matrix Factorization	30.62	18.22

[그림 5. 모델별 Loss 비교]

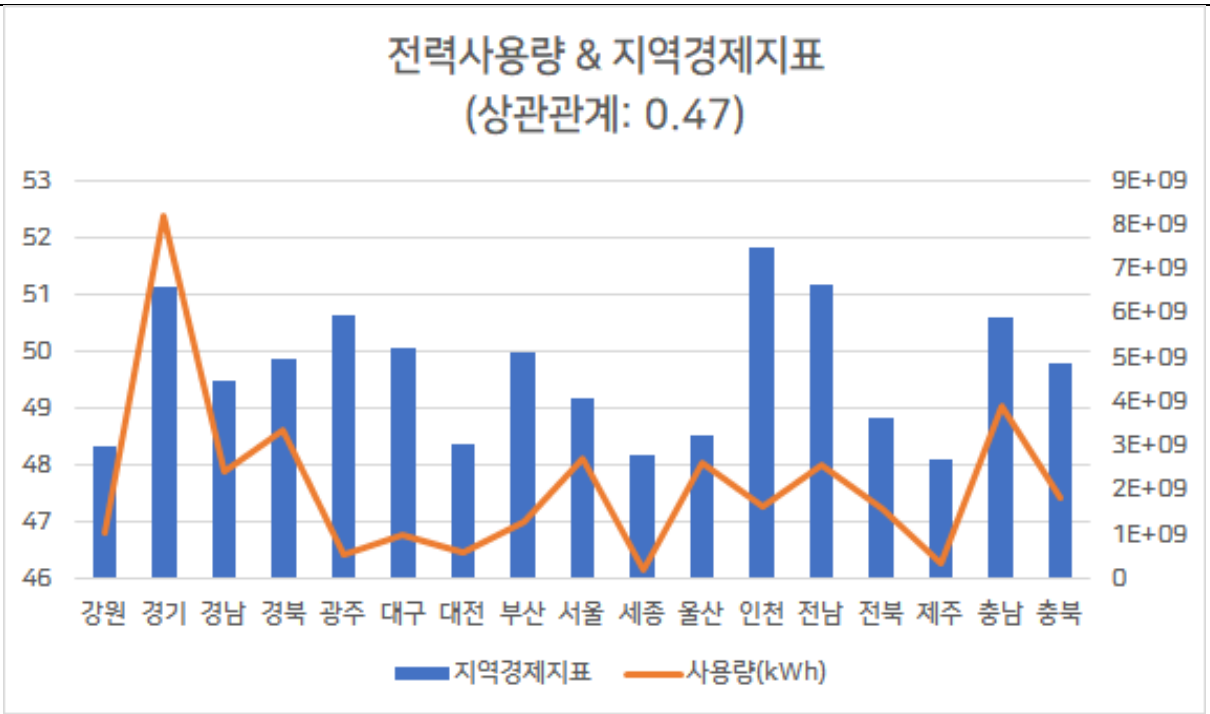


[그림 6. Learning Curve]

3-5. 가설 검증 및 타당성

[전력 사용량 데이터 (KEPCO)]

전력데이터 개방 포털시스템에서 제공하는 업종 및 지역별 전력 사용량을 통해 새로운 지역 경제 지표에 대한 검증을 진행하였다. 전력 사용량은 그 지역의 공장들이 얼마나 가동되는지 대변해 줄 수 있는 자료이며, 기업의 전력 사용량이 클수록 기업의 생산성이 높다고 해석할 수 있다. 따라서, 이를 지역의 경제성을 나타낼 수 있는 지표로서 보고 2018년도 광역시도별 전력 데이터를 수집하여 검증을 진행했다.

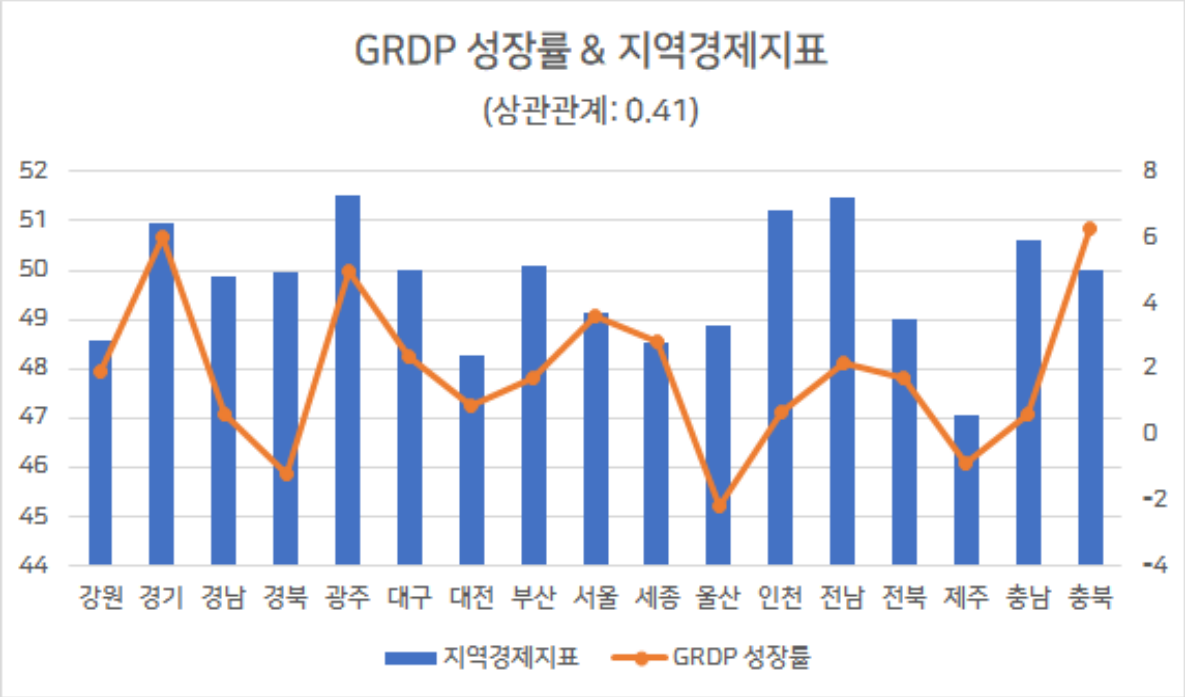


[그림 8. 전력 사용량과 지역경제지표의 상관관계]

파란색 막대그래프는 새로 도출한 지역 경제지표를 의미하며, 빨간색 선 그래프는 지역의 전력 사용량 합계를 나타낸다. 지역 경제지표가 낮을 수록 지역의 위험성을 나타내며 하위 3개 지역인 강원, 대전, 제주의 전력 사용량도 낮은 것으로 나타난다. 또한, 좋은 지역을 나타내는 상위 3개 지역인 경기, 인천, 전남의 전력 사용량이 대체적으로 높은 것으로 나타난다. 새로운 지역 경제지표와 전력 사용량에 대한 상관관계는 0.47로 높은 상관성을 나타낸다. 즉, 새로운 지역 경제지표로 좋은 지역이라고 판단한 곳일 수록 지역의 경제성을 대변해줄 수 있는 전력 사용량이 높은 것으로 나타난다. 따라서, 새로운 지역 경제지표가 지역의 경제성을 반영한다고 해석할 수 있다.

[지역내총생산(GRDP) 성장률]

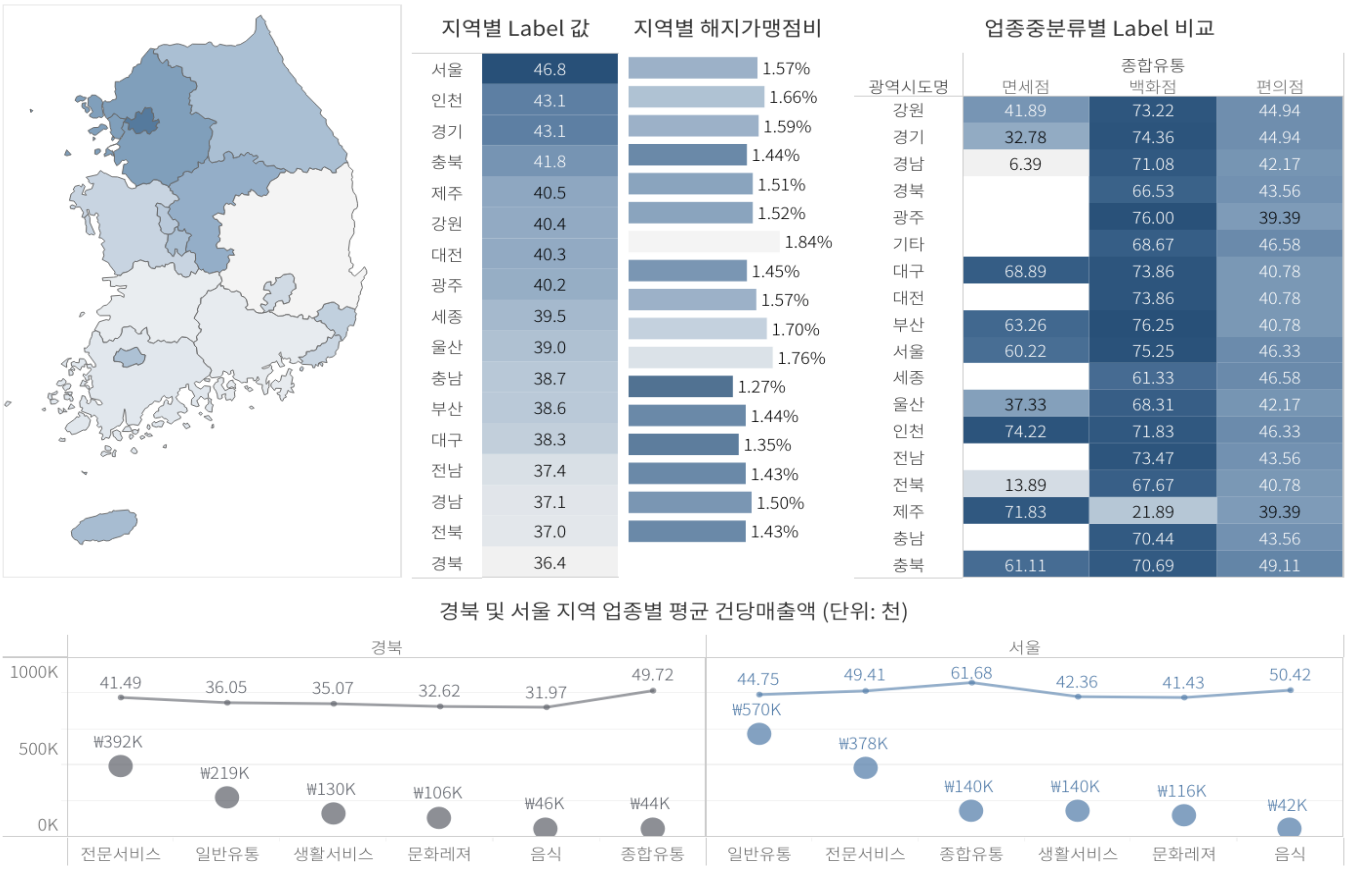
지역내총생산은 일정기간 동안 정해진 경제구역 내에서 생산된 모든 최종재화와 서비스의 시장가격 합으로 경제구조나 규모 파악에 활용하는 지표이다. 이에 대한 성장률은 2015년을 기준으로 GRDP의 증감에 대해 나타낸다. 따라서 지역의 경제가 얼마나 성장했는지 나타낼 수 있는 대표적인 지역 경제지표이다. 따라서, 통계청에서 2018년도 광역시도별 GRDP 실질성장률 데이터를 수집하여 검증을 진행했다.



[그림 9. GRDP 성장률과 지역경제지표의 상관관계]

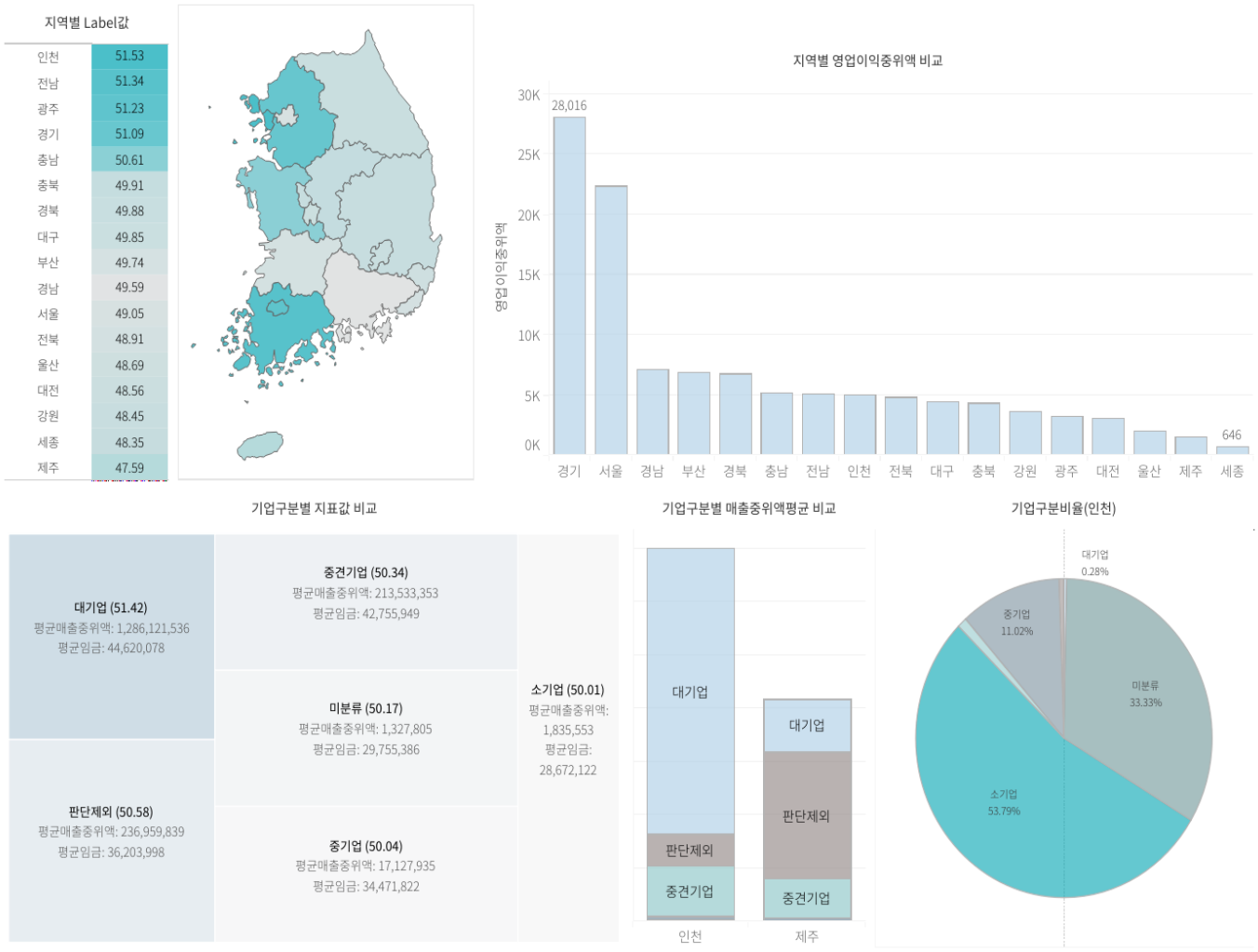
파란색 막대 그래프는 새로 도출한 지역 경제지표를 의미하며, 주황색 선 그래프는 지역의 GRDP 성장률을 나타낸다. 지역 경제지표가 낮을 수록 지역의 위험성을 나타내며 하위 3개 지역인 강원, 대전, 제주의 GRDP 성장률도 낮은 것으로 나타난다. 또한, 좋은 지역을 나타내는 상위 3개 지역인 경기, 광주, 전남의 GRDP 성장률이 대체적으로 높은 것으로 나타난다. 새로운 지역 경제지표와 GRDP 성장률에 대한 상관관계는 0.41로 높은 상관성을 나타낸다. 즉, 새로운 지역 경제지표로 좋은 지역이라고 판단한 곳일 수록 지역의 경제 성장률이 높은 것으로 나타난다. 따라서, 새로운 지역 경제지표가 지역의 성장성을 반영한다고 해석할 수 있다.

3-6. 결과 시각화 및 분석



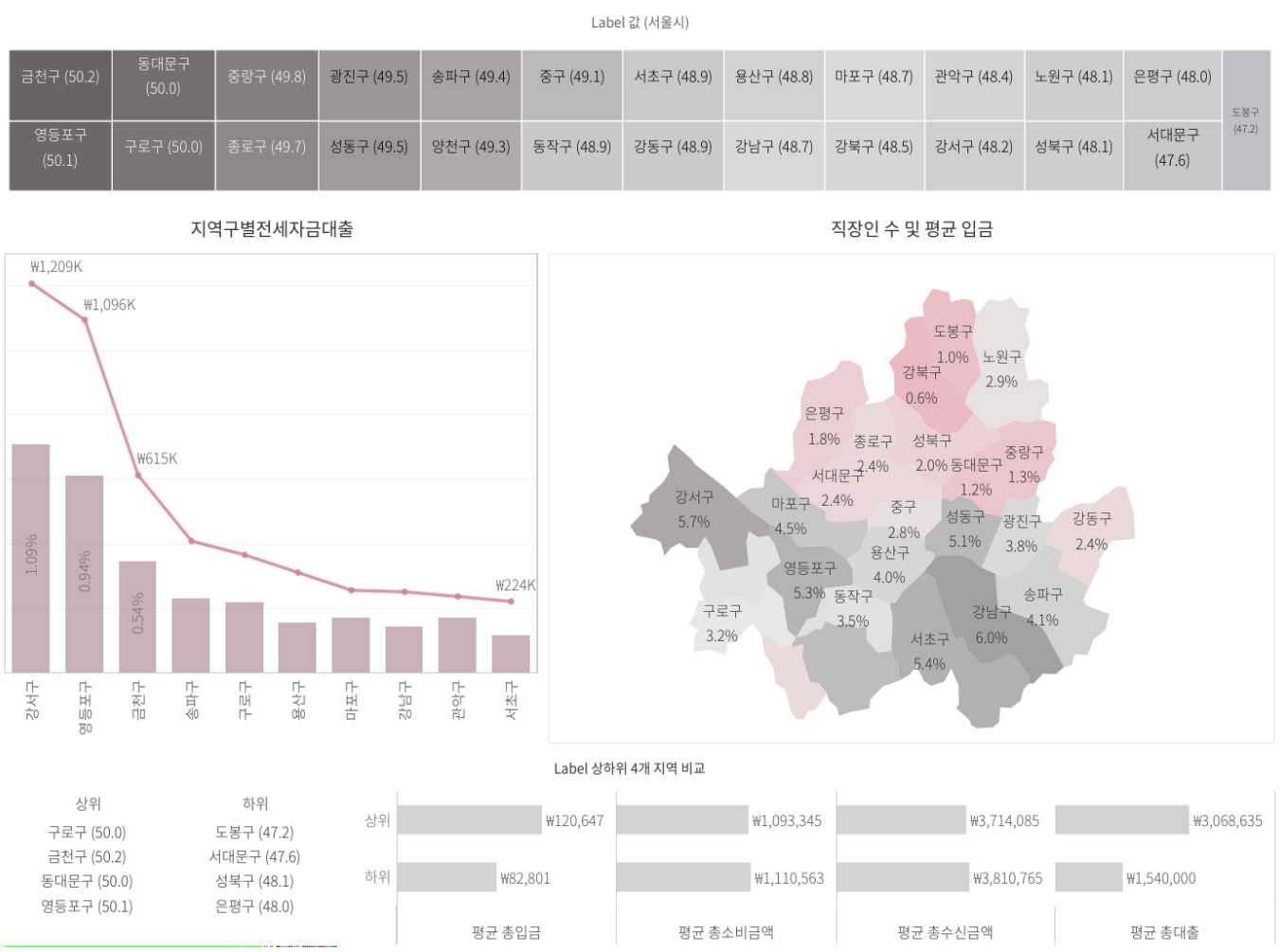
[그림 10. 새로운 지역경제지표를 반영한 광역시도별 데이터 대시보드]

광역시도별 업종별 가맹점 데이터(신한카드)를 활용하여 생성한 Label 값들을 살펴본 결과, 서울과 인천, 경기도 지역의 Label 값이 높아 가장 위험도가 높은 지역임을 확인할 수 있다. 이때 위험도 지표와 직접적인 연관이 있는 '해지가맹점비'와 Label 값들을 함께 보았을 때, Label 값이 높은 상위 지역들 중 해지가맹점비가 높은 (1.7% 이상) 지역은 없는 것으로 나타났다. 다음으로 지역별로 뚜렷한 차이를 보일 것으로 예상되는 업종인 '면세점'의 Label 값들을 살펴보았을 때, '인천' 지역과 '제주' 지역에서 Label 값이 가장 높은 것을 확인할 수 있었다. 인천에는 인천국제공항이 있어 공항에 유동인구가 많고, 제주 지역 또한 우리나라의 대표 여행지로서 공항 면세점 소비가 많아 Label 값이 위험도를 잘 대변하고 있다고 볼 수 있다. 반대로 지역별로 거의 차이를 보이지 않을 것으로 예상되는 두 업종인 '백화점'과 '편의점'을 비교해보았다. 두 업종은 실제로 어느 지역에 있어도 Label 값에 큰 차이를 보이지 않지만, 실제로 위험도가 낮은 '백화점'의 경우에 높은 Label 값을 가지며, 편의점은 그보다 낮은 Label 값을 가지는 것으로 나타났다. 따라서 광역시도별 업종별 가맹점 데이터(신한카드)를 활용하여 생성한 Label 값이 지역 및 업종 별 위험도를 반영한다는 것을 설명할 수 있다.



[그림 11. 새로운 지역경제지표를 반영한 지역별 통계 데이터 대시보드]

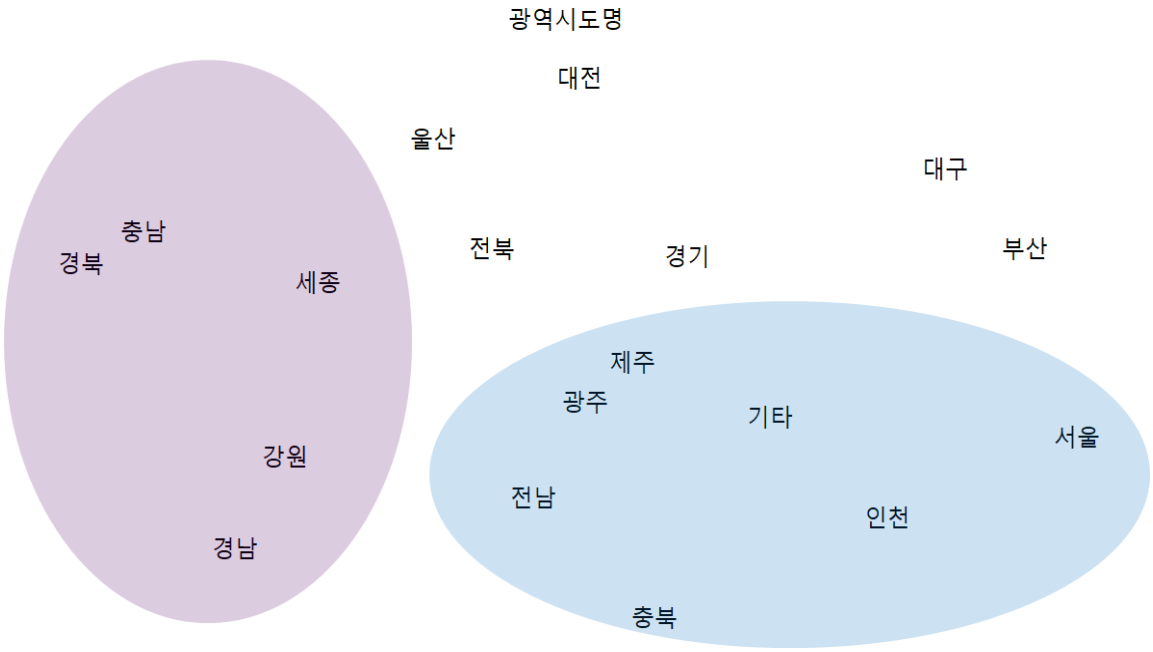
지역별통계데이터(한국기업데이터)를 활용하여 생성한 Label을 지역별로 나누어 살펴보면, 인천 지역의 Label이 가장 크고 제주가 Label이 가장 작다는 걸 확인할 수 있다. Label 값이 50 이상인 지역 중 영업이익중위액이 상위권에 속하는 지역이 80%인 것으로 보아 Label이 지역의 경제 상황을 어느 정도 반영하고 있음을 짐작할 수 있다. 두 지역의 기업구분별 매출중위액평균을 비교해보니, 인천은 평균 매출중위액의 대부분을 대기업에서 나오는 반면 제주는 매출의 대부분은 기업 구분 중 '판단제외'에 해당하는 기업들로부터 내고 있음을 볼 수 있다. 추가로 인천 지역의 기업구분비율과 기업구분별 매출총액 평균을 알아보면, 인천은 50% 이상이 소기업으로 이루어져 있고 대기업은 그 비율이 굉장히 미미함에도 불구하고, 매출총액을 보면 대기업이 상당 부분을 차지하고 있는 것을 확인할 수 있다. 실제로 Label 값이 가장 큰, 즉 위험도가 가장 낮은 기업 구분에 해당하는 대기업의 비중이 전체 매출 중 가장 크다는 것은 충분히 지역 자체의 Label이 높다는 점을 대변할 수 있다.



[그림 12. 새로운 지역경제지표를 반영한 서울시 지역단위 금융정보 데이터 대시보드]

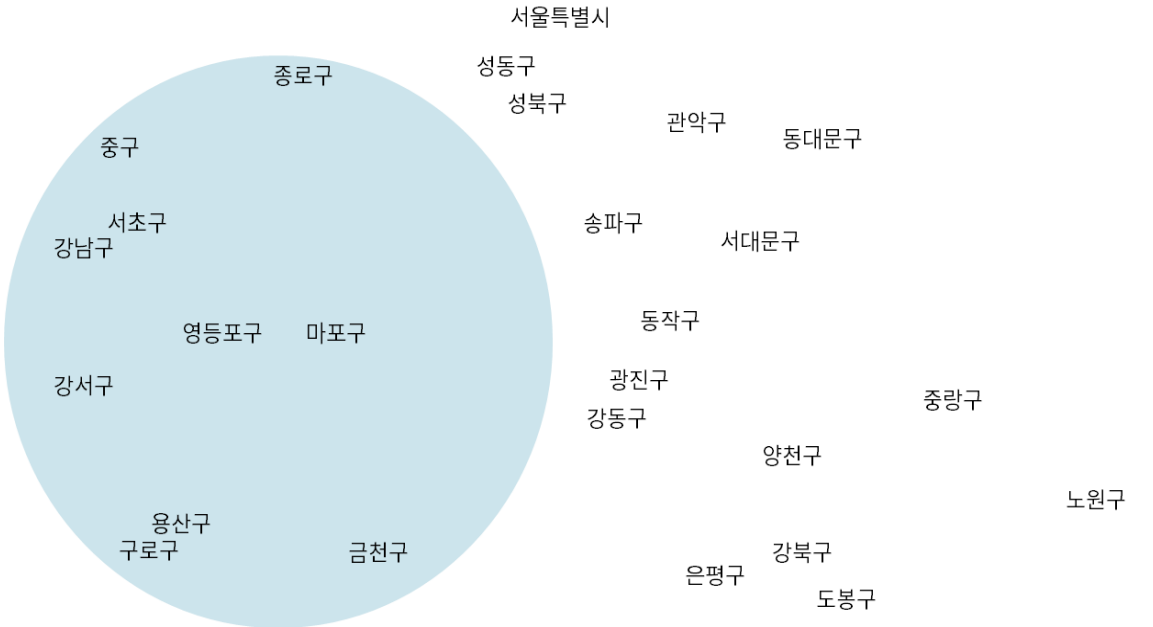
위에서 생성된 Label 값을 보다 상세하게 살펴보기 위해 서울시 각 구에 대한 Label 값을 신한은행 서울시 데이터와 엮어서 분석 결과를 도출했다. 금천구의 Label 값이 50.2로 가장 크고 도봉구의 Label 값이 47.2로 가장 낮은 것을 확인할 수 있다. 각 지역의 직장인 비율 및 평균 입금액을 알아보았을 때, 강북구 및 도봉구, 동대문구가 상대적으로 낮은 입금을 받고 있으며 강남구와 서초구, 영등포구 등은 높은 입금은 지급받고 있음을 알 수 있다. 또한 지역별 직장인의 비율도 함께 맵 상에 나타났는데, 강남구와 강서구가 대표적으로 직장인 비율이 많고, 도봉구와 강북구 등을 직장인의 비율이 적은 편이라는 것을 알 수 있다. 대출 종류 중 지역의 영향을 가장 많이 받을 것이라 예상하여 여러 대출금액 종류 중 전세자금대출을 살펴보았고, 가장 전세자금 대출을 많이 받는 강서구, 영등포구, 금천구가 실제 위험도 Label 값이 높은 것을 확인할 수 있었다. 대출을 많이 받는다는 것이 무조건 위험도가 높다고 판단할 수는 없으며, 은행 입장에서는 좋은 고객들이기 때문에 Label이 높은 지역들에 대해서는 전세자금 대출과 관련하여 주시할 필요가 있다고 생각한다. 다음으로 Label 값으로

도출한 상위권 및 하위권 지역을 비교하기 위해 4개의 지역끼리 하나의 그룹을 형성하여 주요 지표 값들을 비교해보았다. 결과적으로 Label 값을 기준으로 위험도가 상대적으로 낮다고 판단한 지역들에서는 위험도가 높다고 판단된 지역들보다 총입금은 높지만, 소비금액은 낮은 것을 확인할 수 있었다.



[그림 13. 서울시 지역구 임베딩 벡터 시각화]

[그림 10. 새로운 지역경제지표를 반영한 광역시도별 데이터 대시보드]와 비교했을 때 Label 값이 서로 유사한 위치의 지역끼리 가깝다는 것을 확인할 수 있다. Label 값이 상위에 위치한 서울, 인천, 충북 지역이 서로 가까운 위치에 임베딩 되어 있다.



[그림 14. 서울시 지역구 임베딩 시각화]

[그림 12. 새로운 지역경제지표를 반영한 서울시 지역단위 금융정보 데이터 대시보드]와 비교했을 때 Label 값이 서로 유사한 위치의 지역끼리 서로 가깝다는 것을 확인할 수 있다. Label 값이 상위에 위치한 영등포구를 중심으로 그 인접 위치에 Label 값이 큰 지역이 서로 가깝게 임베딩 되어 있다.

4. 기대효과

[새롭게 제시한 지역 경제 지표의 유용성]

지역 경제 지표는 지역의 경제 상황을 한눈에 보여주는 수치자료로 그중 중요한 경제 현황이 전략적으로 잘 드러나는 통계를 경제지표라고 한다. 경제통계는 과거부터 현재까지의 경제 추이를 추적하는 데 유용할 뿐만 아니라, 미래의 경제 상황을 예측하는 바로미터로 사용된다. 이 때문에 경제통계는 지방정부나 지역주민에게 경제 부문의 기초정보를 제공하는 역할을 한다.

따라서 우리는 기업의 지리적 위치에 대한 정보를 반영할 수 있는 지역경제지표를 제시한다. 현재의 지역경제지표는 그 기업 자체의 위험도를 평가하기에는 많은 한계점을 가지고 있다. 하지만 우리가 제시하는 지역경제지표는 기업의 재무 정보를 바탕으로 평가된 위험도를 통해 구해진 지표이기 때문에 직접적으로 기업의 위험도와 관련이 있다고 할 수 있다.

또한, 새로운 지역경제지표는 기업의 재무정보가 없더라도 활용할 수 있다는 장점을 가진다. 금융업계에서 기업의 위험도를 평가할 때는 대부분 재무정보를 많이 활용한다. 하지만 대부분의 기업들은 이러한 재무정보가 결측치인 경우가 대다수이다. 따라서 정보를 파악할 수 없는 기업에 대해서는 위험도 평가를 하는데 있어 어려움을 겪는다. 이를 해결하기 위해 새로운 지역 경제지표는 기업의 위치, 업종을 하나의 임베딩으로 표현했기 때문에 기업의 위험도를 평가하기 위한 수치로써 사용할 수 있어 재무정보가 존재하지 않는 기업의 위험도를 평가할 수 있다.

새로운 지역경제지표가 높은 지역은 그 기업이 좋은 위치에 존재하고, 유망 업종인 회사이기 때문에 재무정보가 존재하지 않는 신생 기업의 경우 지표가 높은 지역에 입지했다면, 그 기업의 위험도를 낮게 평가할 수 있다.

[새롭게 제시한 지역 경제 지표의 기대효과 - 정부]

정부의 입장에서는 새로운 지역 경제 지표를 통해 각 지역에 대한 위험성과 재무건전성 등을 파악하여 예산 분배를 적재적소에 할 수 있으며, 정책에 대한 방향성을 수립하는 데 활용할 수 있다. 가령, A와 B 지역의 새로운 경제 지표가 각각 50, 80이라고 했을 때, 정부에서는 예산을 수립하고 분배할 때 이러한 상황을 반영하여 위험한 A 지역에 예산을 더 분배하거나 추가적인 경기 부양 정책들을 시행하여 모든 지역이 균등하게 발전할 수 있도록 한다. 또한, 새로운 지역 경제 지표는 지역별 업종에 따른 위험도를 반영하고 있어 시의회나 시청 등 지방자치단체에서는 세부적인 정책을 수립할 수 있다.

이를 활용하여 지역 경제 실태와 구조를 포괄적으로 파악할 수 있으며, 이는 재정 및 전략 수립의 근거가 될 수 있다. 또한 미래의 경기 변동성을 예측하여 지역별 재정에 대한 분배에 대한 계획을 수립하고, 지역 경제에 영향을 미치는 주요 요인들을 파악하여 지역 발전을 위해 보다 세분화된 경쟁 전략을 수립할 수 있다.

[새롭게 제시한 지역 경제 지표의 기대효과 - 금융기관]

금융기관의 측면에서는 새로운 지역 경제 지표를 활용하여 지역별 소비 패턴 및 경제 흐름을 파악하고 특정 요인을 타깃으로 삼아 마케팅 전략을 수립할 수 있다. 예를 들어, 특정 지역의 업종 중 점수가 높은 업종을 파악하여 그 업종을 대상으로 하는 마케팅을 진행할 수 있다. 더 나아가 점수가 높은 업종은 그 지역에서 소비가 많이 이루어지는 업종이기 때문에 이 업종을 대상으로 한 카드 할인 혜택을 제안할 수 있다.

또한, 지역의 경제 지표를 개인의 신용평가에 반영하여 기존보다 더 신뢰도 높고 정교한 신용평가 모형을 개발할 수 있다. 가령, 특정 개인에 대한 신용평가 혹은 기업에 대한 신용평가를 진행할 때, 거주지역이나 회사가 있는 지역의 위험성 및 재정건전성 등을 파악하여 더 정교하고 개인화된 신용평가를 할 수 있다. 그리고, 은행에서는 위험한 지역이나 재정 건전성이 불안정한 지역을 조기에 파악하여 대출이 많이 이루어질 것 같은 지역에 자금 조달과 대출과 관련된 마케팅, 지역 은행에 대한 이자율 산정 등을 할 수 있다.

한국은행에서는 지방중소기업지원 프로그램을 통해 지역 간 균형발전을 도모할 목적으로 은행의 지방중소기업에 대한 대출실적과 지역별 경제 상황 등을 감안하여 한국은행 지역본부별로 한도를 배정하였다. 지역경제 여건 및 지역 특성을 감안하여 중점 지원이 필요하다고 인정하는 부문에 은행의 대출취급실적을 기준으로 지원하는 것이 타당하다고 보면서, 지역별 경제 상황의 차이를 반영하고 있다. 뿐만 아니라 동일한 규모의 기업이라 하더라도 영업점포의 분포, 주택담보대출비율(LTV) 등 지역별로 상이하는 대출 정책, 대출 규모나 연체율에 따라 위험도는 다르게 측정될 수 있기 때문에 업종과 지역에 대해 고려하고 있는 현 상황에서 우리의 새로운 지역 경제지표는 유용하게 활용될 수 있다.

[새롭게 제시한 지역 경제 지표와 방향성 제시]

최근 들어 코로나19 피해 기업에 대한 만기 연장, 이자 상환 유예 조치 등이 행해지고 있음과 동시에 사태가 장기화됨에 따라 은행에서 기업에 지원해주는 액수와 건수가 모두 크게 증가하고 있다.

코로나19 관련 금융권 업종별 지원

업종	지원액(조원)	지원건수(만 건)
기계·금속 제조업	26.5	9.9
도매업	21.4	21.2
소매업	12.3	30.3
음식점업	11.2	35.4
섬유·화학 제조업	10.8	4.5
자동차 제조업	9.0	1.7
운수·창고업	6.4	9.4
숙박업	4.0	2.5
여행·레저업	3.4	6.7

[그림 15. 코로나19 관련 금융권 업종별 지원 현황]

이에 위험 업종에 대한 은행의 대출 심사 또한 까다로워지고 있을 뿐만 아니라, 대출 취급액 또한 줄어들고 있으며 이러한 지원 규모 자체가 해당 업종의 리스크를 평가하는 데에 중요하게 작용하고 있다. 따라서, 기존에 사용된 위험도 측정 지표와 더불어 우리의 업종 데이터를 반영한 새로운 지역 경제지표를 반영한다면, 고도화된 리스크 관리와 대출 심사를 진행할 수 있다.

새로운 지역경제 지표는 기존의 여신 평가 모델, 위험도 측정 지표, 지역내경제총생산 등을 넘어 정부와 공공기관, 금융기관 등의 널리 활용될 수 있다. 정부와 공공기관에서는 정확한 지역 경제 상황을 파악하여 지역 균등 발전을 도모할 수 있으며, 금융기관에서는 기업의 여신 평가나 지역 및 업종을 대상으로 하는 마케팅이 가능하다. 더 나아가 새로운 지역경제 지표와 더불어 코로나19 팬데믹 같은 경제에 영향을 미치는 특수 상황에 대한 지표들을 개발하여 반영한다면 더 정확하고 효율적인 지역 경제 상황에 관한 판단과 위험도 평가를 할 수 있을 것이다.

5. 활용 데이터

[1] 지역별통계데이터(기업) (출처: 한국기업데이터)

[2] 광역시도별업종별가맹점데이터 (출처: 신한카드)

[3] 서울시지역단위 '소득', '지출', '금융자산' 정보 (출처: 신한은행)

[4] 전력 사용량 데이터 (출처: KEPCO 전력 빅데이터 센터)

[5] 지역내경제총생산(GRDP) 데이터 (출처: 통계청)

6. 참고자료

[1] 박인천 외 1명, 빅데이터와 AI를 활용한 신용평가의 변화 시도, 서울신용평가(주), 2018

[2] 박주완 외 2명, 빅데이터 분석 기법을 이용한 소상공인 신용평가 모형 구축 연구, 신용보증재단중앙회, 2019

[3] 권황현, 머신러닝과 금융: 머신러닝 기반 신용평가모형, KDB산업은행, 2020

[4] 이형철, 기업의 지리적 위치와 주식가치와의 관계, 재무관리연구, 2014

[5] 남혜정, 기업의 지리적 위치가 배당의사결정에 미치는 영향, 한국세무학회, 2018

[6] 이명건 외 1명, 기업의 위치와 실제이익조정과의 관계, 지역발전연구, 2016

[7] 최성민, 기업 신용평가모형의 현황과 변화 트렌드, CIS이슈리포트, 2018

[8] Xiangnan He 외 5명, Neural Collaborative Filtering, ACM, 2017

[9] 김을식 외 1명, 시·군의 경제력 지표, '1인당 GRDP'의 보완 방안 연구: 보완 지표 개발과 타당성 검토, 한국지역경제연구, 2015

[10] 박성근 외 2명, 전력통계를 활용한 지역별 경기동향지수 개발 연구, KIET 산업연구원, 2015

※ 작성 시 유의사항

– 분량 제한 없음 /글자 폰트 크기 11 포인트(한글 및 워드로 작성)

– 도표, 이미지 등 활용 가능

– 설명을 위한 추가자료 첨부 가능

– 제출 시 표지를 함께 제출하되, 식별이 가능하도록 참가자 명(팀의 경우 팀장 및 팀원명)을 작성하여 제출

2021 년 9 월 5 일

참가자(대표자) (안○○ (인) 명)

금융보안원 귀중