Journal of the Korean Institute of Industrial Engineers
Vol. 50, No. 3, pp. 173-188, June 2024.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2024.50.3.173

공공데이터와 설명가능한 AI 기법을 활용한 자영업 상권 분석과 장기 생존 예측

장호준 · 이기환 · 이희상†

성균관대학교 시스템경영공학과

Small Business Trade Area Analysis and Survival Prediction Using Public Data and Explainable AI Techniques

Ho-jun Jang · Gi-hwan Lee · Heesang Lee

Department of Systems Management Engineering, Sungkyunkwan University

Korea has a high proportion of self-employed individuals, and their frequent businesses closures are causing serious social problems. Although there are many previous studies on commercial district analysis and long-term survival prediction for self-employed businesses, this area lacks systematic analysis. This study utilized public data and machine learning to perform commercial district analysis and long-term survival prediction for closed restaurants across Seoul. XGBoost (eXtreme Gradient Boosting) was finally selected among seven proposed models. We also utilized SHAP (SHapley Additive exPlanations), one of the explainable artificial intelligence (XAI) techniques, to investigate the influence of variables affecting the proposed model and explore the main factors affecting business closures. As a result, we found that some demographic and geographic factors are important factors for small business survival. Finally, we applied the proposed model to predict new restaurant locations across Seoul, demonstrating its potential usefulness in the decision-making process for self-employed entrepreneurs.

Keywords: Public Data, Survival Analysis, Trade Area Analysis, Machine Learning, SHAP

1. 서 론

자영업은 다수의 일자리를 제공하여 국가 경제 발전을 촉진시키고(Hatfield, 2015; Yang, 2016), 다양한 산업군에 종사하며 여러 지역에 분포해 사회경제적인 측면에서 매우 중요한 역할을 수행한다(Jeon and Oh, 2023). 자영업이 대한민국 경제에서 차지하는 비중은 국내 경제 규모에 비해 매우 높은 편인데(Kahn *et al.*, 2020), 국내 고용 규모 지표를 살펴보면 전체 경제활동 인구 2,900만 명 중 자영업자 수는 571만 명으로, 전체 취업자 수의 20.1%를 차지한다. 자영업자와 사실상 동업하는 무급 가족 종사자까지 합치면 국내 자영업의 규모는 661만 명(23.5%)으로 추산되며(Statistics Korea, 2023), 해당 수치는 OECD 38개국 중 8위이다. 이는 EU 27개국의 평균(14.5%)보다 10%가량 높은 수치이고, 한국과 유사한 제조업 기반 국가인 독일(8.7%)과 일본(9.6%)의 자영업 고용 비중은 한국의 절반에 미치지 못한 실정이다(OECD, 2023).

국내 자영업자들은 2017년 사드 사태, 2019년 코로나-19 발생, 2022년 러시아-우크라이나 전쟁 등 일련의 글로벌 경기 침체와 함께, 지속해서 줄어드는 매출과 상승하는 부채의 영향으로 심각한 상황을 맞고 있다(Kahn *et al.*, 2020). 한편 2010년대 중반부터 국내 자영업자들의 영업이익률은 주변 업체들과의 경쟁으로 인해 제한된 매출 상승과 임대료를 포함한 고정

본 논문은 21년도 정부(과학기술정보통신부)의 재원으로 한국연구재단 지원을 받아 수행되었음 (NRF-2021R1F1A1063690).
† 연락저자 : 이희상 교수, 16419 경기도 수원시 장안구 서부로 2066, Tel: 031-290-7628, E-mail: leehee@skku.edu
2023년 12월 14일 접수; 2024년 2월 10일 수정본 접수; 2024년 3월 11일 게재 확정.

지출 부담의 증가로 감소하는 추세이다(Lee, 2021). 자영업자들은 창업을 위한 초기 투자 비용의 많은 부분을 대출로 감당하는 것으로 알려져 있는데(Nam, 2017), 몇몇 자영업자들의 비은행 대출은 이들이 금리 인상과 같은 외부 충격에 더 취약해지는 결과를 초래한다(Kim, 2017).

악화하는 경영난으로 인해 국내 자영업체 폐업률은 높은 수준으로 나타난다. 자영업체의 1년 생존율은 61.7%로 창업 1년 만에 40%에 달하는 업체가 폐업하며, 5년 생존율은 26.9%로 전체 자영업자의 1/4만이 살아남는 수준이다(Kim, 2018). 특히 자영업에서 높은 비중을 차지하는 요식업 분야의 신생 기업은 2년 안에 폐업하는 업체가 절반 이상에 달하며(KOSIS, 2020), 이들의 5년 생존율은 22.8%에 불과하다(Statistics Korea, 2021). 자영업자들의 잦은 폐업은 고용의 불안정, 가계 · 기업 부채의 증가 그리고 사회경제적 양극화 등의 사회적 문제를 야기하며(Nam, 2017; Sung and Kim, 2020), 심각한 경우 우리 경제 하부구조의 붕괴를 초래할 수 있다(Noh and Chung, 2016). Nam(2015)에 따르면 1개의 자영업체 폐업 시 최대 6,500만 원의 총 사회적 비용이 발생하며, 이는 우리나라 전체적으로 최대 30조 원에 달하는 것으로 추산된다.

공공정부와 지방자치단체에서도 자영업 폐업의 막대한 사회적 비용을 문제로 인지하고 다양한 소상공인 지원 정책을 꾸준히 실시하고 있다(Kim and Shin, 2015; Kim, 2018; Jeon and Oh, 2023). 일례로 몇몇 지자체들과 공공기관은 데이터 기반의 상권분석시스템을 제공하여 소상공인들의 입지 선정과 경쟁력 강화를 지원하고 있다(Kim, 2008; Lee, 2019). 이러한 상권분석시스템은 상권 정보를 체계적으로 제공하고 예상 매출액을 제시하여 자영업자들의 의사결정을 지원한다. 그러나 계속해서 변화하는 상권 환경을 감안할 때 단기적인 매출 예측만으로는 점포의 영업 지속성을 설명하기 어려운 한계점을 가진다.

학계에서 점포의 영업 지속성을 설명하고자 하는 연구는 중소기업관리론이나 통계학적 생존분석의 형태로 다수 수행되었다(Kim *et al.*, 2018; Kim *et al.*, 2019; Lee *et al.*, 2022; Kim *et al.*, 2023). 그러나 기존 선행 연구 검토 결과, 폐업한 점포의 특징이나 실패 원인에 대한 체계적인 분석은 부족한 실정이다. 국내 자영업의 높은 폐업률 개선을 위해선 폐업 점포에 대한 실증적인 데이터 분석을 통한 실패 요인 규명이 필요하다고 판단하여 본 연구는 서울시 폐업 자영업체 중 창업이 집중된 요식업을 대상으로 점포의 장기 생존, 단기 생존, 기타 등 3가지로 분류하였으며, 설명 가능한 인공지능(XAI: eXplainable Artificial Intelligence; 이하 XAI)을 활용해 단기 생존 점포의 실패 원인을 입수 가능한 데이터를 통해 분석 및 설명하고, 장기 생존을 위한 함의를 도출하였다.

본 논문의 구성은 다음과 같다. 제2장에서는 자영업 폐업 예측과 상권 요인 규명에 관한 선행연구를 고찰하였고, 제3장과 4장에서는 데이터와 방법론을 설명하였다. 제5장에서는 실험 결과를 제시하고 제6장에서는 결론을 서술하였다.

2. 선행연구

2.1 선행연구 검토

자영업 폐업 예측과 상권 분석에 관한 선행 연구를 검토한 결과, 통계모형을 활용한 생존 분석(Survival Analysis)과 기계학습(Machine Learning)을 활용한 연구들이 다수 진행되었음을 알 수 있다. 통계모형의 경우 그 해석력이 우수해 상권의 성공/실패 요인 분석에 강점을 가진 방법론이므로 이를 활용해 서울시 일반음식점을 대상으로 상권 요인을 규명한 연구가 존재하며(Kim *et al.*, 2019; Kim *et al.*, 2023), 서울시 강남구 일반음식점 중 장기 생존 업체를 대상으로 장기 생존을 위한 공간적 특성을 연구한 사례도 발표되었다(Kim *et al.*, 2018). Lee *et al.*(2022)는 코로나19 시기 서울시 1,011개 상권에 대해 개발특성, 수요자 특성, 점포 특성의 범주에 속하는 변수들의 영향력을 파악하는 연구를 수행했다. 그러나 이들 연구에서 사용한 통계모형은 통계적 가정에 의존적이기에 광범위한 변수를 고려하기 어려우며 복잡한 비선형 관계를 제대로 포착하지 못하는 단점이 존재한다(Kim *et al.*, 2022). 또한 이들의 연구는 자영업 생존에 영향을 미치는 요인 탐색의 목적을 띄고 있기에, 점포의 장기 지속성 예측에 어려움이 존재한다.

기계학습은 복잡한 비선형 구조를 포착할 수 있고, 광범위한 변수를 사용한 분석에 강점이 있다(Bzdok *et al.*, 2018). Jang(2021)은 기계학습을 사용해 서울시 25개 자치구에 속하는 서비스업과 요식업 점포의 생존 기간을 예측하고, 주요 상권 요인을 분석하였다. Bang *et al.*(2018)은 서울시 치킨업을 대상으로 공간 정보를 활용해 3년 이내 폐업 여부를 분류하는 연구를 진행하였다. 이들의 연구는 기계학습을 사용해 점포의 폐업 예측을 시도한 점에서 기존 연구와 차별되나, Jang(2021)의 연구는 62.2%의 Accuracy를, Bang *et al.*(2018)은 61.3%의 Balanced Accuracy를 기록해 실제 의사결정에 사용되기에는 예측력이 약하다는 약점이 있다.

해외의 경우 온라인으로 생성되는 소비자 리뷰 데이터를 활용한 자영업 폐업 예측 연구들이 활발하게 진행되는 추세이다. 소비자 리뷰 데이터와 시계열 분석(Time-Series Analysis)을 사용해 음식점의 폐업을 예측하는 연구가 수행되었으며(Tao and Zhou, 2020), 소비자 별점 데이터와 가격대, 임대료 등의 점포 특화 데이터를 결합해 요식업 폐업 예측을 수행한 연구가 존재한다(Naumzik *et al.*, 2022). 소비자 생성 리뷰 데이터와 점포 이미지 데이터를 동시에 사용하여 음식점의 생존 기간을 예측한 연구도 존재한다(Zhang and Luo, 2023). Li *et al.*(2023)은 온라인 리뷰 데이터를 5가지 범주로 나눈 후 랜덤 생존 포레스트(Random survival forest)를 활용해 음식점 생존 예측을 진행하였다. 검토한 모든 해외의 연구사례는 Area under the curve(이하 AUC) 0.73 이상의 우수한 예측 성능을 보였으나 이들의 연구는 점포의 공간적 특성에 관한 정보를 반영하지 않아 생존업체의 상권 요인 규명에 어려움을 겪는다.

2.2 연구의 차별성

선행연구 검토 결과 국내 자영업의 생존에 관한 기존 연구는 장기 지속성의 예측이 어려우며, 해외의 연구 사례는 공간정보를 사용하지 않는 경우가 많아 상권 분석에 적합하지 않다는 한계점을 도출하였다. 특히 해외에서 활발한 소비자 리뷰 데이터 활용 시도는 영세 요식업 점포 및 신규 업종 등 소비자 생성 데이터가 충분하지 않은 경우 적용하기 어렵다는 한계를 가지며, 이는 요식업종뿐만 아니라 다양한 업종으로의 연구 범위 확장 시 리뷰데이터가 충분하지 않은 업종에 대해 분석이 어려워진다는 한계점을 유발한다. 또한 짧은 창업 준비 기간과 이에 따른 과도한 경쟁으로 인해 높은 폐업률과 낮은 생존율의 생존 특성을 띠는 국내 자영업 환경(Kim, 2018)에 적용하기 위해서는 국내에 축적된 방대한 자영업 폐업 점포의 특징을 반영하는 데이터 과학적인 분석이 필요하다. 따라서 본 연구는 2017년 이후 폐업한 자영업 점포를 대상으로 장기 생존 점포 예측과 주요 상권 요인 분석을 진행하였으며, 이는 기존 연구들과 다음과 같은 차별점을 가진다. 첫째로 방대한 규모의 데이터 수집이 어려운 점포 특화 데이터의 사용을 지양하고, 접근성 높은 공공데이터를 사용해 자영업자들이 의사결정에 쉽게 활용할 수 있게끔 하였다. 둘째로 국내에 발표된 기존 방법론 대비 정확한 예측 성능을 갖춘 모델을 구축하였으며, 셋째로 제안하는 AI모델이 Black-Box 모델에 그쳐 점포의 폐업에 대한 설명력을 결여하지 않도록 XAI를 사용해 상권 요인별 분석을 수행하여 자영업 실패 요인을 규명하였다.

3. 데이터

3.1 분석 범위 설정

행정안전부는 228개 시군구의 전국 자치단체를 통해 수집된 196종의 인허가 데이터를 제공하고 있으나 전국의 자영업체를 모두 고려하는 것은 지역 특성과 업종에 따른 이질성으로 인해 자영업체 생존 예측과 폐업 요인에 대한 연구의 설명력이 낮아질 가능성이 있다. 이에 본 연구의 공간적 분석 범위는 가장 많은 가용 데이터가 존재하는 서울시 전역의 일반음식점과 휴게음식점으로 설정하였고 시간적 범위는 데이터의 수집 시점을 일치시키기 위해 개업일을 기준으로 2017년부터 2022년 3분기까지로 설정하였다. 다만 2020년과 2021년은 코로나-19로 인해 국내 자영업의 매출 상황이 악화한 시기로(Lee and Yoo, 2022) 해당 기간 개・폐업한 점포는 분석 대상에서 제외하였다.

본 연구의 내용적 범위는 장기 생존 점포 예측과 상권 데이터를 활용한 공간적 실패 요인 분석이다. 이를 위해 2023년 08월 01일 기준으로 영업 중인 점포를 제외하고 분석 범위를 설정했으며, 폐업 점포의 장기 생존 기준은 기존 연구 사례를 바탕으로 5년으로 설정하였다. 장기 생존 기준이 5년 이상 생존인 이유는 다음과 같다. 미국의 경우 자영업의 과반이 5년 안에 폐업하는 것으로 나타났으며(Muller and Woods, 1991; Healy and Mac Con Iomaire, 2019), 자영업체들이 가지는 경제적 중요성을 고려할 때 자영업체의 5년 생존율을 끌어올리는

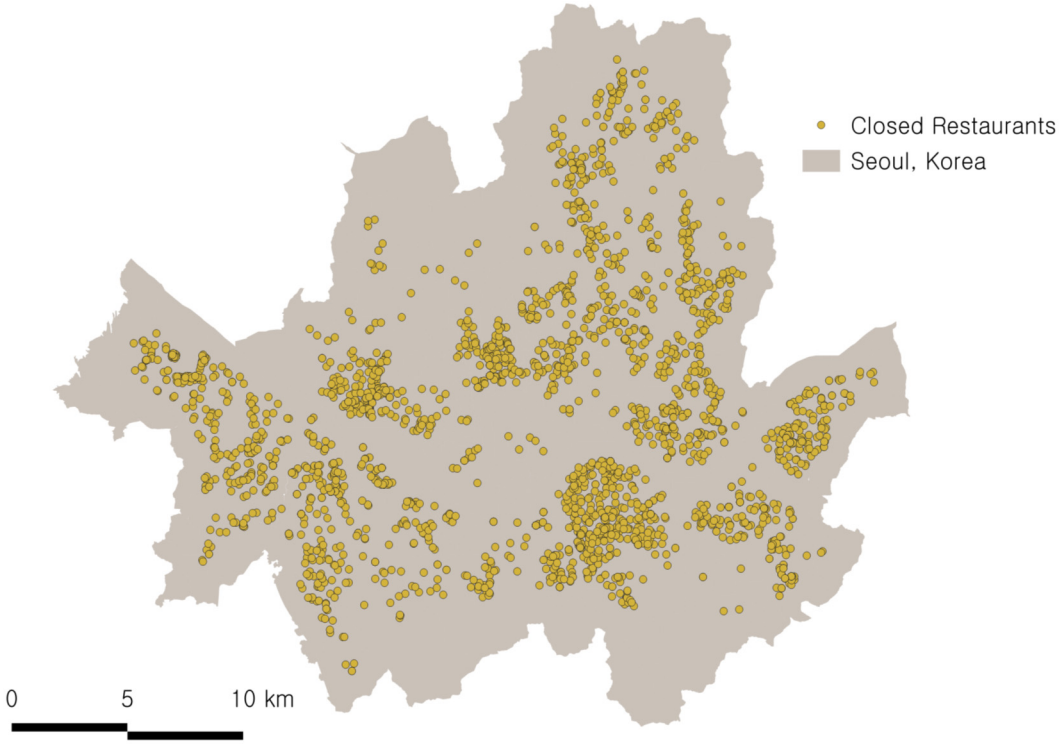


Figure 1. Restaurant Location in Study Area

것이 정책적으로 중요하다는 기존 연구들이 있다(Yang, 2016; Lum, 2017). 국내 연구에서도 자영업의 생존 특성은 5년을 기준으로 안정세에 접어드는 폐업률로 나타나는데, 개업 후 6개월 이후부터 두 자릿수씩 상승하여 3년까지 높은 수준을 유지하다 5년 이후로 안정세에 접어든다고 알려져 있다(Kim, 2018). 이러한 맥락에서 본 연구는 5년 이상 생존 후 폐업한 업체를 장기 생존업체로 지정하고 6개월 안에 폐업한 단기 생존업체들을 실패 업체로 정의하여 이들의 비교를 통해 자영업 폐업의 공간적 요인 규명과 장기 생존을 위한 입지 예측을 하였다. <Figure 1>은 본 연구의 분석 대상으로 총 2,768개의 폐업 점포 위치를 나타낸다.

3.2 변수 설계

인구밀도와 자영업 생존율이 음의 상관관계를 가진다는 연구결과(Fritsch, 2006)와 상당한 수의 일시적인 인구수가 요식업 성공에 영향을 미친다는 연구 결과가 발표되었으며(Parsa, 2015), 소매업의 경우 소비자와 점포 간 거리가 점포를 선택하는 소비자의 행동에 영향을 미친다는 연구 결과가 존재한다(Prasad, 2010). 이에 본 연구는 거리에 따라 상권을 1차와 2차로 구분하여 인구 요인을 반영한 변수를 설계하였다. AppleBaum(1966)은 사업장 이용고객의 60~70%를 포함하는 범위를 1차 상권으로 정의하고 점포로부터 반경 500m로 지정했으며, 사업장 이용고객의 15~25%를 포함하는 2차 상권은 점포 반경 1,000m로 설정하고 그 이상을 3차 상권으로 정의하였다. 그러나 본 연구의 분석 범위인 서울은 인구밀도가 매우 높은 도시 중 하나이며, 유연한 구역법으로 인해 주거지역과 상업지역이 혼재되어 발전된 형태를 띤다(Jun *et al.*, 2013). 이러한 서울의 경제지리학적인 특성을 고려하여 1차 상권은 점포 위치로부터 반경 300m로 축소하였으며, 2차 상권과 3차 상권은 2차 상권으로 통합하여 점포 반경 1,000m로 정의하였다. 거리뿐만 아니라 연령 역시 소비 행태에 주요한 원인을 미치는 요인으로 알려져 있으며(Fareed and Riggs, 1982; Fernández-Villaverde and Krueger, 2007), 수도권과 도시지역에 한해 체류 외국인 규모의 증가가 지역경제에 긍정적인 효과를 미치는 연구결과가 존재한다(Kim and Byeon, 2022). 이러한 맥락에서 인구 특성 변수의 설계는 1차 상권과 2차 상권의 내국인, 장기체류 외국인, 단기 체류 외국인 생활 인구수와 1차 상권의 직장 인구수로 구성하고, 국내 외국인 체류 인구의 40%에 가까운 중국인(Ministry of Justice, 2022)은 전체 외국인 인구수로 나눈 변수로 반영하여 중요성을 반영하고자 하였다. 마지막으로 연령별 정보가 제공되는 내국인 생활인구는 연령대를 20대부터 60대까지 세분화하고 전체 인구수로 나누어 연령별 비율로 반영하였다.

잘못된 가격정책 혹은 판촉정책과 달리 열악한 매장 위치는 수년 동안 소매업체에 부정적인 영향을 미친다(Fox *et al.*, 2007). 즉, 점포의 지리적 위치는 점포 생존에 유의미한 영향을 미치며(Haapanen and Tervo, 2009; Parsa, 2011), 이를 활용해 자영업 생존 예측을 진행한 연구도 존재한다(Karamshuk, 2013; Georgive *et al.*, 2014). 따라서 본 연구 역시 상권의 지리적 특성을 반영하기 위해 다수의 선행연구를 검토하여 지리적 특성 변수를 설계하였다. Kang(2016)은 보행자 접근성이 좋을수록 모든 소매 유형의 판매 실적이 향상된다고 발표하였고 Oh *et al.*(2015)은 서울시 패밀리 레스토랑의 매출이 인접 지하철 정거장의 개수에 유의미한 영향을 받는다고 발표하였다. 이에 보도 접근성과 대중교통 접근성 모두 요식업 생존에 유의한 영향력을 미칠 것이라 판단하여 최근접 지하철 정거장까지의 유클리드 거리, 점포 반경 1,000m내의 지하철 정거장 개수, 최근접 버스 정거장까지의 유클리드 거리, 점포 반경 500m 내의 버스 정거장 개수, 점포 반경 100m 내의 횡단보도와 도로의 수 그리고 도로 길이까지 총 7가지를 접근성 측면을 반영하기 위한 변수로 사용하였다. 점포 인접 구역의 물리적 특성 또한 소매업의 생존에 영향을 줄 것이라 판단하였다. Sung(2022)에 따르면 인접 건축물 연면적을 비롯한 점포 주변의 물리적 환경들이 소매업의 매출에 영향을 준다. 이에 점포 반경 100m 내 건축물의 연면적 총합과 반경 500m 내의 초, 중, 고교 수 및 반경 250m 내의 공공, 문화 체육시설 수를 변수로 사용하여 물리적 특성을 반영하였다. Lee *et al.*(2014)에 따르면 대학가 상권의 경우 기존의 대형 상권과 입지 요인이 상이한 특수한 경우이다. 이 점을 고려하여 대학상권의 범위를 대학부지 최외각으로부터 500m로 정의하고 개별 점포의 대학상권 포함 여부를 이진변수로 반영했으며, 대학상권에 포함된 점포 간의 차이를 반영하기 위해 점포와 대학교 최근접 출입구까지의 유클리드 거리를 변수로 사용하였다. 또한 상권의 업종 다양성은 소매업체의 생존에 영향을 미치며(Kim *et al.*, 2019; Kim *et al.*, 2022; Lee *et al.*, 2022), Kim *et al.*(2018)에 따르면 동종업체의 수와 밀집도는 요식업 장기 생존에 영향을 미치기에 1차 상권내의 유사업종 다양성과 유사업체 및 경쟁업체수 그리고 경쟁업체수를 유사업체수로 나눈 경쟁업체비를 변수로 설정하여 상권특성을 공간적 정보로서 반영하였다.

개별 점포 특화 정보가 분류에 핵심적인 변수로 작용하여 상권 요인 분석이 어려워짐을 방지하기 위해 개별 점포 특화 정보는 최소한의 변수로 구성하였다. 점포 총 규모는 점포 생존에 주요한 영향을 미치기에(Parsa, 2015) 점포 면적을 변수로 사용하고, 서울시 공시지가 데이터를 활용해 점포의 총 지가를 계산해 임대료를 대신하여 반영하였으며, 개별 점포의 구분을 주류 판매 허가에 따른 일반음식점과 휴게음식점의 이진변수로 반영하였다. 해당 변수들은 개별 점포 특화 정보 중 점포 개업 전 수집이 용이한 데이터로, 학습된 모델이 추후 자영업자들의 개업 전 입지 예측을 수행하는데 어려움이 없도록 하였다. 이로써, 인구 특성, 지리적 특성 그리고 점포 특성에 걸쳐 총 40가지의 독립변수 설계를 완료했으며 이는 <Table 1>과 같다.

Table 1. Independent Variables Description

	Variable	Description
Demographic Features	1A_Total	Total residential population within 300m radius
	1A_20	20s age group ratio of 1A_Total
	1A_30	30s age group ratio of 1A_Total
	1A_40	40s age group ratio of 1A_Total
	1A_50	50s age group ratio of 1A_Total
	1A_60	60s age group ratio of 1A_Total
	2A_Total	Total residential population within 1,000m radius
	2A_20	20s age group ratio of 2A_Total
	2A_30	30s age group ratio of 2A_Total
	2A_40	40s age group ratio of 2A_Total
	2A_50	50s age group ratio of 2A_Total
	2A_60	60s age group ratio of 2A_Total
	1A_Long_Total	Total long-term foreign resident population within 300m radius
	1A_Long_CN	Long-term Chinese resident ratio of 1A_Long_Total
	2A_Long_Total	Total long-term foreign resident population within 1,000m radius
	2A_Long_CN	Long-term Chinese resident ratio of 2A_Long_Total
	1A_Temp_Total	Total short-term foreign resident population within 300m radius
	1A_Temp_CN	Short-term Chinese resident ratio of 1A_Temp_Total
	2A_Temp_Total	Total short-term foreign resident population within 1,000m radius
	2A_Temp_CN	Short-term Chinese resident ratio of 2A_Temp_Total
	Working_Pop	Total working population within 300m radius
Geographic Features	Sub_C	Total number of subway stations within 1,000m radius
	Sub_M	Euclidean distance to the nearest subway station
	Bus_C	Total number of bus stations within 500m radius
	Bus_M	Euclidean distance to the nearest bus station
	Road_L	Total road distance within 100m radius
	Road_C	Total number of roads within 100m radius
	Crosswalk_C	Total number of crosswalks within 100m radius
	TF_Area	Total gross floor area of all buildings within 100m radius
	UniDist_YN	Binary variable for university trade area (N:0 / Y:1)
	UniEnt_M	Euclidean distance to the nearest university entrance
	School	Sum of primary, secondary and high schools within 250m radius
	PubBuilding	Sum of public buildings within 250m radius
	Competitor_C	Total number of competitors within 300m radius
	Competitor_R	Competitors ratio within 300m radius
	Business_D	Business diversity within 300m radius
	Adjacent_BIZ	Total number of restaurants within 300m radius
Store-Specific Features	Total_LV	Total land value of restaurants
	Service	Binary variable for liquor sales permit (N:0 / Y:1)
	Area	Size of restaurants

3.3 데이터 수집 및 전처리

본 연구에 사용한 모든 데이터는 공공데이터로 구성할 수 있었다. 서울시 열린 데이터 광장에서 생활인구, 지하철 정거장 위치 정보, 버스 정거장 위치 정보, 도시계획시설, 학교 기본정보 그리고 공시지가 데이터를 수집했으며, 스마트치안 빅데이터 플랫폼에서 직장인구 데이터를, 서울시 도로 데이터와 건축물 정보는 국가공간정보포털에서 수집하였다. 분석 대상

Table 2. Dependent Variable Description

	Count	Mean	Standard deviation	Min	Max
Short-term survival restaurant (0)	1757	104.52	44.23	31	180
Long-term survival restaurant (1)	1011	2002.22	136.64	1825	2394

인 서울시 일반음식점과 휴게음식점 데이터는 지방행정 인허가데이터에서 수집하였다. 데이터 수집 과정에서 원본 데이터의 업데이트 주기가 데이터마다 상이한 관계로 독립변수의 반영 시점 역시 기준을 다르게 하여 처리하였다. 즉, 생활인구는 폐업 점포의 개업 월을 기준으로 결합하였고, 직장인구는 분기를 기준으로 일치시켜 처리하였다. 공시지가 데이터의 경우 1년 주기로 제공되기에, 점포의 개업 연도를 기준으로 결합하였으며 지리적 특성 변수는 데이터 제공 주기가 일정치 않아 한 시점으로 통일하여 처리하였다. 지하철 정거장 위치 데이터와 도로 데이터는 2018년을, 버스 정거장 위치 데이터는 2019년을, 서울시 학교 정보는 2023년을 사용했으며, 그 외의 모든 변수는 2021년을 기준으로 사용했다.

생활인구와 직장인구 데이터는 원본 데이터가 각각 행정동과 법정동을 기준으로 제공되어, 정확한 상권별 인구수 반영을 위해 건축물 연면적을 기준으로 서울시 전역에 재분배하여 사용하였다. 해당 과정에 지리공간데이터 분석기능을 제공하는 QGIS(ver 3.32.2-Lima)를 활용하였으며 방법은 다음과 같다. 우선, 서울시 전역에 10m 등간격으로 좌표를 생성 후 건축물과 대조하여 겹치는 좌표만 추출하였다. 만약 개별 건축물이 하나의 좌표도 가지지 않는다면 해당 건축물의 중심에 좌표를 생성하여 모든 건축물이 최소 1개 이상의 좌표를 갖게 처리하였다. 생성된 좌표들을 각각이 포함된 건축물의 연면적을 해당 건축물에 포함된 좌표 개수로 나눈 값으로 가중치로 부여한 후 법정동 단위로 재분배해 각각의 좌표가 속해있는 법정동의 연면적 합의 일정 비율을 갖게 하였다. 이를 생활 및 직장인구 데이터와 곱하여 좌표별로 가중치를 부여했으며, 상권 범위 안에 포함되는 가중치의 총합으로 인구수를 산출하였다. 이때 서울시 경계 너머까지 2차 상권 범위가 설정되는 경우는 제거하였다.

점포 총 지가는 서울시 공시지가 데이터를 점포 주소와 일대일로 대응시킨 후 각 점포 크기와 곱해 산출했으며, 공시지가가 확인되지 않는 점포는 제거하였다. 서울시 도시계획 시설 데이터는 학교, 공공청사, 연구시설, 문화시설 등의 위치 정보를 담은 데이터로 대학교와 그 외 시설을 분리하여 사용하였다. 대학 출입구는 별도 표기하여 사용했으며, 점포의 생존기간이 30일 이하인 경우와 점포 크기 혹은 인접 건축물 연면적이 3.3㎡ 미만이면 이상치로 간주해 제거하였다. 상권특성 반영을 위한 경쟁업체는 2023년 8월 27일을 기준으로 지방행정 인허가 데이터에 기록된 서울시의 모든 일반음식점 및 휴게음식점 중 세부업종이 일치하는 점포의 개·폐업 시기를 확인하여 반영했으며 유사업체 수와 유사업종 다양성은 1차 상권내의 모든 요식업체수와 서로 다른 세부업종의 개수로 반영하였다. 동일 데이터에서 폐업일이 기록되어 있는 455,872개의 점포 데이터를 선별하여 분석 대상으로 삼고 전처리를 거쳐 필요한 모든 변수를 결합해 구축한 데이터는 11,910개의 데이터 포인트와 40개의 변수로 구성된다. 이 중 6개월 이하 생존업체 1,757개와 5년 이상 생존업체 1,011개 등 총 2,768개를 선별하여 각각 단기 생존업체(Short-term survival restaurant, 0)와 장기 생존업체(Long-term survival restaurant, 1)로 정의하여 종속변수를 생성하였으며, 각 그룹의 기술 통계는 <Table 2>와 같다.

4. 분석 방법론

4.1 기계학습 이진분류 모델

데이터 전처리 후 높은 정확도 달성을 위해 기계학습 기반의 일곱 가지 분류 모델(Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, XGBoost, LightGBM)을 사용하여 학습을 진행하였다. Logistic Regression은 선형회귀모형의 한 종류로서 독립변수와 종속변수 간 오차를 최소화하는 회귀선을 추정해 분류를 진행하는 모델이다(Hastie *et al.*, 2009). K-Nearest Neighbors(이하 KNN)는 개별 데이터의 예측 시 최근접 한 k개의 데이터 포인트의 레이블을 결합하여 예측을 수행하는 비모수적 모델이며(Cover, 1967), Support Vector Machine(이하 SVM)은 데이터들을 고차원 공간으로 사상한 후 이를 분류하는 하나의 초평면을 찾아 분류를 수행하는 모델이다(Cortes and Vapnik, 1995). Decision Tree(이하 DT)는 if-else 구조의 결정 트리를 학습하여 데이터 분류를 수행하는 모델이며(Song and Ying, 2015), Random Forest(이하 RF)는 여러 개의 DT의 조합을 통해 분류를 수행하는 모델이다. 일반적으로 RF는 DT보다 정확한 예측을 수행하며, 일반화 성능이 뛰어나다고 알려져 있다(Breiman, 2001). XGBoost는 그라디언트 부스팅(Gradient Boosting, 이하 GB)의 한 종류로서, 다수의 약 분류기를 순차적으로 학습하여 오차를 최소화하는 하나의 강 분류기를 만드는 모델이다(Chen and Guestrin, 2016). LightGBM은 GB의 한 종류이나 기존 GB 모델들과 달리 개별 약 분류기의 학습이 Leaf-wise 방식으로 이루어진다. 이는 수직적으로 트리를 전개하는 방식으로 트리 전개 시 손실 값이 가장 큰 노드를 우선으로 전개해 학습 속도가 빠르다는 장점이 있다(Ke *et al.*, 2017).

4.2 하이퍼파라미터 탐색

정확한 일반화 성능평가를 위해 전체 데이터 세트를 8:2의 비율로 훈련 데이터 세트와 평가 데이터 세트로 분할해 사용하였으며, 모델 학습 시 과적합 방지를 위해 10-fold cross validation을 실시하였다. 본 연구에서 사용한 데이터는 단기 생존업체가 장기 생존업체보다 7 : 4 정도로 많은 불균형 데이터로, 모델 학습 시 다수 범주에 편향되어 학습될 여지가 존재한다. 이는 점포들의 수명이 지수분포의 형태를 띠는 국내 자영업 환경에서 적절하게 사용되기 위해서 필수적으로 해결되어야 하는 문제이다. 이에 본 연구는 불균형 데이터 학습에 효과적인 비용민감학습(Elkan, 2001)을 도입하여 소수 범주와 다수 범주 간 분류성능의 균형이 유지되도록 조정하였다. Logistic Regression, SVM, KNN 모델의 경우 데이터의 스케일에 민감하게 반응하기에 학습 시 데이터 표준화를 고려하였다. 비용민감학습, 모델별 하이퍼파라미터, 데이터 표준화 여부를 반영한 모델별 탐색 범위는 <Table 3>와 같으며, 탐색은 그리드 서치(Grid Search)로 진행하였다.

4.3 모델 평가

불균형 데이터 세트에서의 모델 평가는 소수 범주에 대한 분류가 간과되지 않게 하는 것이 중요하며, 이를 위해 F-Measure, Balanced Accuracy, AUC 등이 사용된다(Bekkar *et al.*, 2013). 본 연구는 Accuracy, Balanced Accuracy, F1 score, 그리고 AUC를 평가지표로 사용하였다. 각 평가지표는 <Table 4>와 같은 혼동행렬을 기반으로 표현되며 AUC를 제외한 지표들의 식은 식 (1)~식 (3)과 같다.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \quad (1)$$

Table 3. Model Hyperparameter Search Space

Algorithm	Cost sensitive learning weights		Data Scaling	Hyperparameter	Search space
	Majority	Minority			
K-Nearest Neighbors	N/A		None / Standard	n_neighbors	1, 3, 5, 7, 9, 11
				weights	uniform, distance
				p	1, 2
Logistic Regression	0.10	0.10, 0.12, 0.14, 0.16, 0.18, 0.20, 0.22, 0.24, 0.26, 0.28		C	0.001, 0.01, 0.1, 1, 10, 100
				penalty	l1, l2, elasticnet
Support Vector Machine				C	0.1, 1, 10
				gamma	0.01, 0.1, 1
				kernel	linear, rbf
				max_iter	10000
Decision Tree			N/A	max_depth	3, 4, 5, 6, 7, 8
				min_samples_leaf	1, 3, 5, 7
Random Forest				n_estimators	200, 300, 400
				max_depth	3, 4, 5, 6, 7, 8
				min_samples_leaf	1, 3, 5, 7
XGBoost				n_estimators	400, 500, 600
				max_depth	5, 6, 7, 8
				colsample_bytree	0.8, 0.9, 1.0
				eta	0.05, 0.07, 0.1
				tree_method	gpu_hist
LightGBM				n_estimators	100, 200, 300, 400, 500
				max_depth	3, 4, 5, 6
				colsample_bytree	0.6, 0.7, 0.8, 0.9, 1.0
				learning_rate	0.01, 0.03, 0.05, 0.1
				tree_method	gpu_hist

Table 4. Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

$$Balanced\ Accuracy = \frac{TP}{TP+FP} + \frac{TN}{TN+FN}. \quad (2)$$

$$F1\ Score = \frac{TP}{TP+\frac{1}{2}(FP+FN)}. \quad (3)$$

$$True\ positive\ rate = \frac{TP}{TP+FN}. \quad (4)$$

$$False\ positive\ rate = \frac{FP}{TN+FP}. \quad (5)$$

AUC는 ROC Curve의 면적으로 계산된다. ROC Curve란 모델의 임계값에 대한 분류 감도를 수식(4)와 (5)의 함수로 나타내 수치들을 연결한 곡선이며(Bekkar *et al.*, 2013), AUC는 ROC Curve의 요약 지표로 ROC Curve의 하단 면적을 의미한다(Bradley, 1997). 따라서 AUC는 0과 1 사이의 값을 갖게 되며, 일반적으로 0.7 이상인 경우를 납득 가능한 분류기로 간주한다(Hosmer *et al.*, 2013; Lavazza *et al.*, 2023).

4.4 SHAP

대부분의 기계학습 모델은 Black-Box 모델로 예측 결과에 대한 해석의 제공에 어려움을 갖는다(Ribeiro *et al.*, 2016). 따라서 Black-Box 모델을 그대로 사용할 경우 본 연구의 목표인 점포 실패 요인 규명에 한계점을 가지며 이를 해결하기 위해 XAI를 도입하였다. XAI는 분류 모델에 따라 사용할 수 있는 종류가 달라지는데, 특정 모델에 적용할 수 있는 모델 특정 기법과 별도의 제한 없이 적용 가능한 모델 불특정 기법이 존재하며(Gohel *et al.*, 2021), 본 연구는 모델 불특정 기법 중 전역 해석과 국소 해석을 모두 제공할 수 있는 SHAP(SHapley Additive exPlanations)을 사용해 단기 생존 점포의 공간적 실패 요인 분석을 시도하였다.

SHAP은 Shapley(1953)의 게임이론을 바탕으로 Lundberg and Lee(2017)가 고안한 방법론이다. SHAP은 해석이 어려운 기존 모델을 대신하여 해석할 수 있는 근사치로 정의되는 대리모델을 구축해 변수별 영향력을 설명한다. 대리모델은 모든 변수의 예측치에 대한 기여도의 가중합으로 표현되며 이들의 합은 Black-Box 모델의 예측치에 근사된다.

$$f(\boldsymbol{x}) \approx g(\boldsymbol{z}) = \phi_0 + \sum_{i=1}^{M} \phi_i z_i. \quad (6)$$

식 (6)에서 $f(\boldsymbol{x})$는 학습된 모델 f의 $\boldsymbol{x}$값에 대한 예측치이며 이는 대리모델 $g(\boldsymbol{z})$로 근사된다. $g(\boldsymbol{z})$는 ϕ_i와 z_i의 선형결합으로 구해지는데, $\phi_i \in \mathbb{R}$는 변수 i가 가지는 예측치에 대한 기여도를 의미하며 $z_i \in \{0,1\}^M$는 변수 i의 사용 여부를 나타내는 지시변수이다(Lundberg and Lee, 2017).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)]. \quad (7)$$

이때 F는 전체 변수의 부분 집합, S는 변수 i가 제거된 부분집합을 의미한다. 식 (7)과 같이 변수 i의 기여도 ϕ_i는 한계기여도 $[f(S \cup \{i\}) - f(S)]$의 가중합으로 구해진다. 산출된 기여도 집합은 지역 정확성(Local accuracy), 없어짐(Missingness), 일치성(Consistency)을 모두 만족하는 유일한 값이며, 인간 직관과 유사하게 작동하여 해석이 쉬운 장점이 존재한다(Lundberg and Lee, 2017).

5. 실험 결과

5.1 분류기 예측 성능

<Table 5>는 모델별 하이퍼파라미터 탐색 결과를, <Table 6>은 그에 따른 성능평가 결과를 나타낸다. 검증데이터 세트에 대한 평가는 Balanced Accuracy로 진행했으며 일반화 성능 평가는 4개(Accuracy, Balanced Accuracy, F1 Score, AUC)의 지표를 모두 사용하여 평가를 진행하였다. 실험 결과 다수의 평가지표에서 좋은 성능을 기록한 XGBoost모델의 일반화 성능이 가장 우수하다고 판단하여 본 연구의 제안모델로 선정하였다.

Table 5. Best Settings of 7 Models

Algorithm	Data Scaling	Model		Minority-Weight
K-Nearest Neighbors	Standard	n_neighbors	5	N/A
		weights	distance	
		p	1	
Logistic Regression	Standard	C	1	0.22
		penalty	l2	
Support Vector Machine	Standard	C	10	0.16
		gamma	0.01	
		kernel	rbf	
		max_iter	10000	

Table 5. Best Settings of 7 Models(Continued)

Algorithm	Data Scaling	Model		Minority-Weight
Decision Tree	N/A	max_depth	8	0.18
		min_samples_leaf	1	
Random Forest		n_estimators	400	0.20
		max_depth	8	
		min_samples_leaf	3	
XGBoost		n_estimators	400	0.24
		max_depth	6	
		colsample_bytree	1.0	
		eta	0.07	
		tree_method	gpu_hist	
LightGBM		n_estimators	400	0.28
		max_depth	4	
		colsample_bytree	0.8	
		learning_rate	0.05	
		tree_method	gpu_hist	

Table 6. Score Comparison of 7 Models

Algorithm	Validation Score	Test Score			
	Balanced Accuracy	Accuracy	Balanced Accuracy	F1 Score	AUC
K-Nearest Neighbors	0.6293	0.6750	0.6315	0.5135	0.7132
Logistic Regression	0.6452	0.6354	0.6740	0.6203	0.7386
Support Vector Machine	0.6549	0.6444	0.6537	0.5852	0.7310
Decision Tree	0.6316	0.6336	0.6473	0.5814	0.6533
Random Forest	0.6956	0.6805	0.6863	0.6177	0.7624
XGBoost	**0.7096**	**0.7004**	0.6893	0.6121	**0.7669**
LightGBM	0.7054	0.6967	**0.7054**	**0.6394**	0.7054

Table 7. Selected Features

	Variable	
Demographic Features	1A_Total	2A_20
	1A_20	2A_30
	1A_30	2A_40
	1A_40	2A_50
	1A_50	2A_60
	1A_60	2A_Temp_Total
	1A_Temp_CN	2A_Temp_CN
	1A_Long_Total	2A_Long_Total
	Working_Pop	2A_Long_CN
Geographic Features	PubBuilding	School
	Competitor_C	Competitor_R
	Adjacent_BIZ	Business_D
Store-Specific Features	Area	Total_LV
	Service	

모델간 성능이 동일하다면 더 적은 변수를 사용해 낮은 복잡도를 가진 모델이 더 효율적이다. 따라서 제안모델의 복잡도 감소 및 데이터 수집 비용 감소를 위해 변수제거를 실시하였다. XGBoost모델의 변수중요도를 기반으로 중요도 최하위 변수를 하나씩 제거하는 후진제거법을 실시했으며 각 단계마다 검증데이터 평가 결과를 비교해 Balanced Accuracy 0.7128로 가장 우수한 성능을 보인 최종 변수집합을 선택하였고 선정된 27개 변수는 <Table 7>과 같다. 변수 제거 후 제안 모델의 성능 평가는 평가 데이터의 구성을 변경하며 30번 반복 실험하였고 각 지표의 평균과 표준편차를 <Table 8>로 나타냈다.

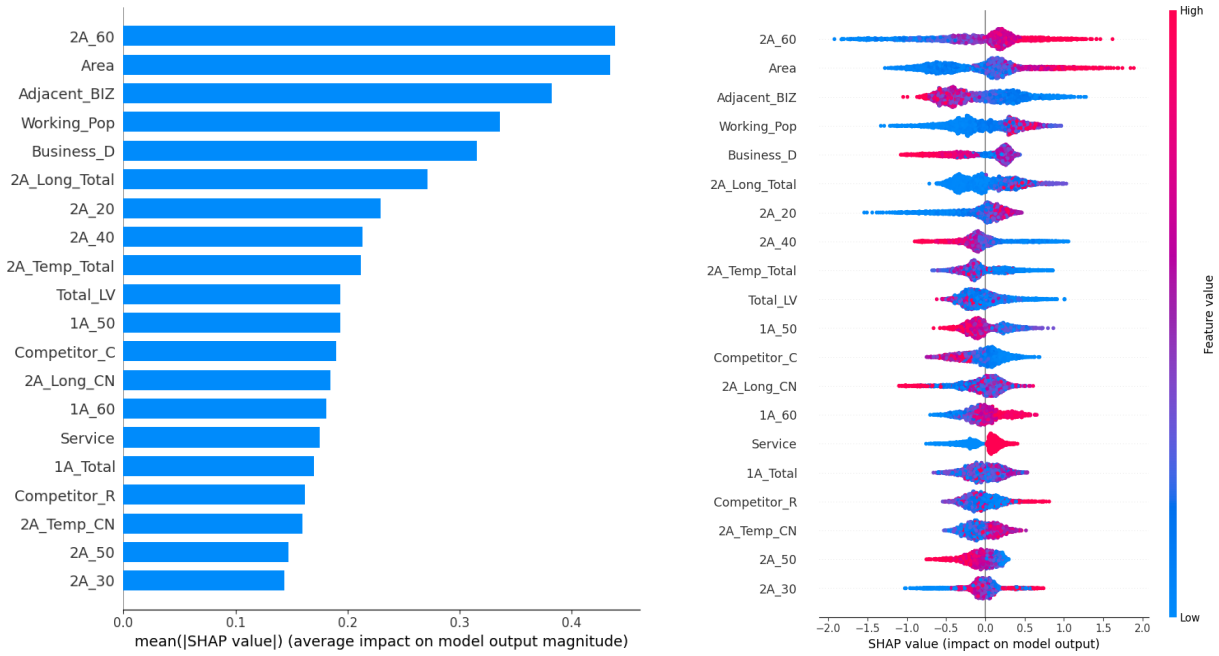
Table 8. Test Score Mean(Std) after Feature Selection

Model	Accuracy	Balanced Accuracy	F1 score	AUC
XGBoost (27 variables)	0.7091 (0.0179)	0.7030 (0.0206)	0.6300 (0.0257)	0.7794 (0.0171)

5.2 SHAP 분석

모든 데이터의 변수별 영향력을 SHAP을 통해 파악하여 주요 상권 요인 분석을 진행하였다. 계산된 SHAP value는 양수일 경우 양의 영향력을, 음수일 경우 음의 영향력을 가진다. 즉, 변수의 SHAP value가 양의 값을 가질 경우 5년 이상 생존하는 점포로 분류될 확률을 높이는 것으로 해석 가능하다. <Figure 2(a)>는 제안 모델의 변수별 영향력의 절대값의 평균을 계산하여 상위 20개 변수를 내림차순으로 나타낸 그림이다. 이에 따르면 2차 상권 60대 내국인 생활 인구비(2A_60)와 점포의 면적(Area)이 점포 생존에 있어 가장 큰 영향력을 행사함이 확인된다. 다음으로는 인접 유사업체수(Adjacent_BIZ), 직장인구수(Working_Pop), 유사업종 다양성(Business_D), 2차 상권 20대 내국인 생활 인구비(2A_20), 2차 상권 40대 내국인 생활 인구비(2A_40)등이 주요 요인으로 판별되며, 2차 상권 장기 체류 외국인 생활 인구수(2A_Long_Total) 및 2차 상권 단기 체류 외국인 생활 인구수(2A_Temp_Total)와 같은 국내 체류 외국인 수가 생존에 영향을 미치는 것으로 확인된다.

<Figure 2(b)>는 제안 모델의 SHAP Summary Plot을 나타낸다. 이는 모든 데이터의 변수별 영향력을 변수의 값에 따라 색을 달리해 표시한 산점도로, 변수의 값이 클수록 붉게, 작을수록 파랗게 표시된다. 2차 상권 60대 내국인 생활 인구비율 변수의 경우, 개별 점포의 크기가 클수록 SHAP value가 증가하고, 작을수록 SHAP value가 감소하는 경향을 확인할 수 있다. 즉, 2차 상권 60대 내국인 생활 인구비와 점포의 장기 생존은 양의 상관관계를 가지며, 이는 점포 면적과 직장인구수도 마찬가지이다. 반면 인접 유사업체수와 유사업종 다양성은 반대의 양상을 보이는데, 인접 유사업체수가 적을수록, 그리고 유사업종 다양성이 낮을수록 점포의 생존에 긍정적 영향을 미친다. 2차 상권 40대 내국인 생활 인구비와 2차 상권 50대 내국인 생활 인구비(2A_50)의 증가는 점포의 장기 생존에 부정적 영향을 미치나, 2차 상권 20대 내국인 생활 인구비(2A_20)와 2차 상권 30대 내국인 생활 인구비(2A_30)의 증가는 긍정적인 영향을 미치는 것으로 확인된다. 경쟁업체수(Competitor_C)는 많을수록 장기 생존에 부정적인 영향을 미치나 경쟁업체비(Competitor_R)는 증가할수록 양의 값을 기록해 반대의 양상이 확인된다. 일반음식점 여부(Service)의 경우 SHAP value의 절대적인 크기가 크지 않지만, 일반음식점과 휴게음식점 간의 방향성이 상반되게 나타나 주류를 판매할 수 있는 일반음식점의 경우가 장기 생존에 도움이 됨을 알 수 있다. 2차 상권 체류 중국인 비율의 경우 단기 체류(2A_Temp_CN)는 높은 편이 긍정적이나 장기 체류(2A_Long_CN)는 SHAP value가 분산되어



(a) SHAP Importance plot (b) SHAP Summary plot

Figure 2. Feature Importance of XGBoost by SHAP

영향력을 해석할 수 없다.

<Figure 3>은 변수의 영향력과 상호작용을 같이 나타낸 그래프로, 개별 변수에 대해 자세한 정보를 파악할 수 있다. <Figure 3(a)>는 지리적 특성 변수 중 가장 큰 영향력을 가진 인접 유사업체수를 나타낸 그래프로, 2차 상권 30대 내국인 생활 인구비(2A_30)와의 상호작용을 색을 사용해 같이 표시하고 있다. 인접 유사업체수는 약 100개를 기준으로 SHAP value의 부호가 달라져 점포 생존에 미치는 영향이 반전되며 약 200개까지 장기 생존에 미치는 긍정적인 영향력이 감소하는 것으로 확인된다. 그러나 이보다 많다면 더 이상 값이 감소하지 않고 유지됨을 확인할 수 있는데, 상권의 유사업체의 수가 일정 수준 이상이라면 점포의 장기 생존 확률에 미치는 부정적인 영향이 더 이상 심해지지 않는 것으로 이해할 수 있다. 상호작용을 같이 확인한다면 SHAP value가 양의 값을 갖는 구간은 30대 인구 비율이 낮은 것이 긍정적이며, 많은 유사업체의 수로 인해 부정적 영향을 받는 상권의 경우는 이와 반대로 30대 인구 비율이 높은 것이 장기 생존에 긍정적임을 알 수 있다. <Figure 3(b)>는 유사업종 다양성과 SHAP value의 관계를 나타낸 그래프이며 3가지 구간으로 나누어 영향력을 살펴본다. 첫째로, 상권의 세부 업종 개수가 12개 이하라면 업종의 종류가 다양할수록 SHAP value가 0 근방에서 분산이 커지며 감소하는 것이 확인된다. 둘째로, 세부 업종의 개수가 13개 이상 19개 미만인 상권의 경우 SHAP value의 값이 모두 양의 값을 가지며 장기 생존에 긍정적인 영향을 미치는 것으로 확인된다. 셋째로, 세부 업종이 19개 이상인 상권의 경우 SHAP value의 부호가 음으로 반전되며 다양성이 높아질수록 절대적인 값이 커짐을 확인할 수 있다. 즉 상권 내의 유사업종 다양성은 증가할수록 장기 생존에 부정적인 영향을 미치나 세부 업종의 개수가 13개 이상 19개 미만인 상권은 예외적으로 긍정적인 영향을 미침을 확인할 수 있다. 세부 업종이 13개 미만이면 점포 면적과 다양성 간의 상호작용 확인이 어려우나 세부 업종의 개수가 13개 이상 19개 미만인 경우는 점포 면적이 작은 편이, 19개 이상이면 점포의 면적이 큰 경우가 장기 생존을 위한 경쟁력을 갖음이 확인된다. <Figure 3(c)>는 경쟁업체수와 SHAP value간 관계를 나타내는 그래프이다. 경쟁업체수는 인접 유사업체수와 비슷한 양상으로 나타나는데, 전체적으로 경쟁업체수가 증가할수록 장기 생존 확률이 감소하는 것으로 보이나 약 50개 이상의 상권은 경쟁업체수가 증가해도 더 이상 SHAP

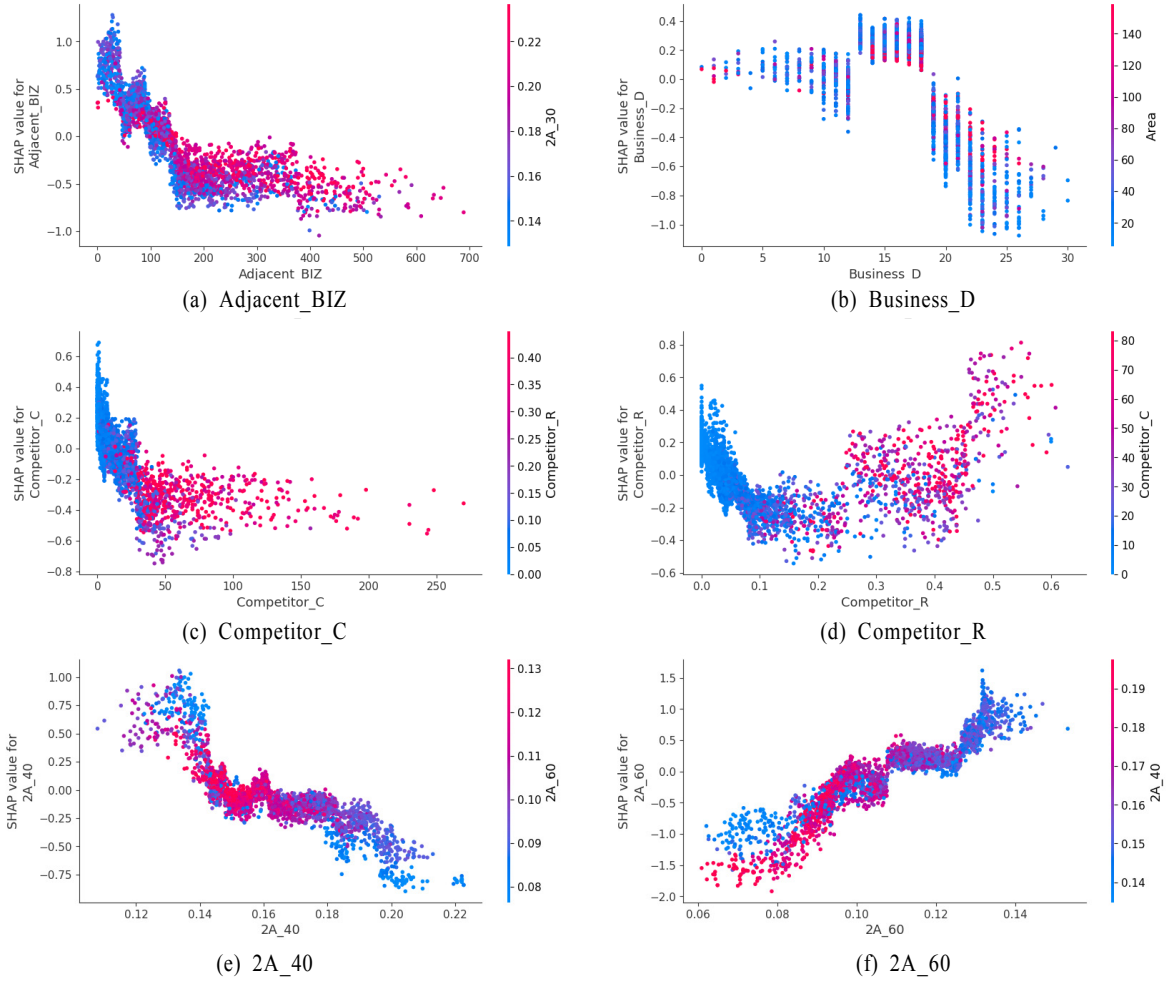


Figure 3. SHAP Dependency Plots for 6 Selected Features

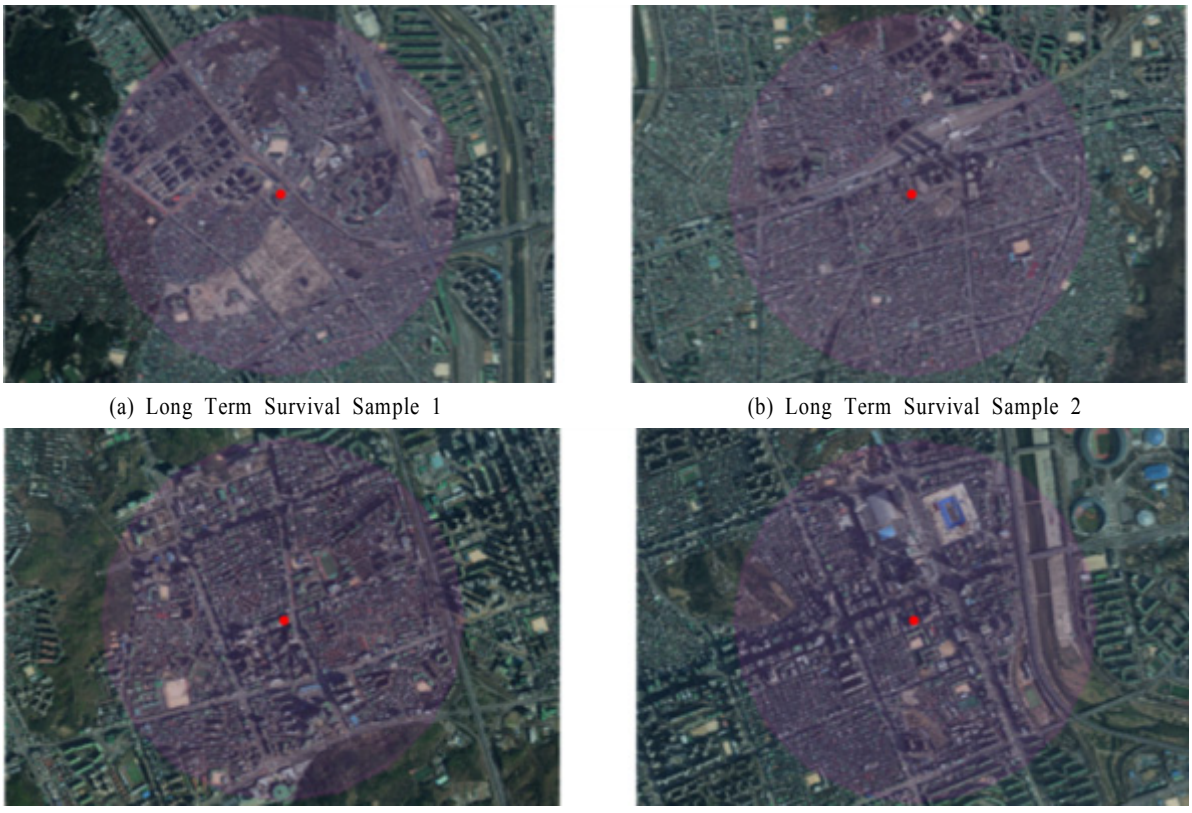
value가 감소하지 않는 것이 확인된다. <Figure 3(d)>는 경쟁업체비를 나타내는 그래프로 경쟁업체수와는 다른 양상을 보인다. 경쟁업체비가 약 10%까지는 SHAP value가 빠르게 감소하여 일정하게 유지되다 30%를 전후로 다시 증가하는 양상을 보이는데 이는 인접 유사업체수에 영향을 받은 것으로 파악할 수 있으며 이를 종합하면 경쟁업체수가 많은 상권은 장기 생존에 부정적이나 전체 유사업체 대비 경쟁업체의 비율이 높다면 장기 생존에 유리한 조건임을 알 수 있다. <Figure 3(e)>는 2차 상권 40대 내국인 생활 인구비를 나타낸 그래프이며 2차 상권 60대 내국인 생활 인구비를 나타낸 <Figure 3(f)>와 함께 확인하여 연령별 영향력을 비교한다. 각 그래프의 상호작용 요인에 따르면 40대 인구비와 60대 인구비는 음의 상관 관계에 있는 것으로 확인되며, SHAP value와의 상관관계는 서로 반대로 나타나 점포 장기 생존에 대한 영향력이 상반되게 작용하는 것으로 확인된다. 즉, 상권 인구구성에서 40대 인구 비율의 증가는 장기 생존에 부정적인 요인으로 판별되며 60대 인구 비율의 증가는 긍정적인 상권 요인으로 판단할 수 있다.

5.3 신규 점포 입지 추천

본 논문에서 개발한(장기 생존 vs. 단기 생존) 분류를 위한 기계학습 모델을 활용해 2023년 10월 26일 신규 분양 중인 서울 전역의 상가들을 대상으로 장기 생존확률 예측을 통한 입지 추천을 수행하였다. 신규 분양 상가 정보는 서울시 부동산 정보광장을 통해 수집했으며, 부동산 주소, 건축물의 연면적, 층별 점포 수를 취합하여 데이터로 구축하였다. 직장인구 데이터의 경우 2022년 3분기까지만 제공되는 관계로 직장인구를 제외한 모든 변수는 시점을 맞추어 결합하였다. 점포 면적은 건축물 연면적과 상가의 분양 점포 수를 사용해 전용면적을 계산해 사용하였으며, 경쟁업체 산출을 위해 개업일은 2023년 10월 26일, 업종은 요식업 창업이 집중되는 일반음식점-한식으로 설정하였다. 전처리 완료 후 총 186개의 신규 입지에 대한 분류를 진행하였다. 제안 모델이 산출한 확률값의 크기에 근거하여 5년 이상 생존할 것으로 예상되는 생존확률 상위 2지점과 6개월 이내에 폐업이 예상되는 생존확률 하위 2지점의 위치를 지도상에 나타냈으며, 선정된 각 지점은 2차 상권의 범위를 같이 표기하여 <Figure 4>로 나타냈고, 지점별 제안 모델 확률값은 <Table 9>와 같다.

Table 9. Model Probability for Figure 4

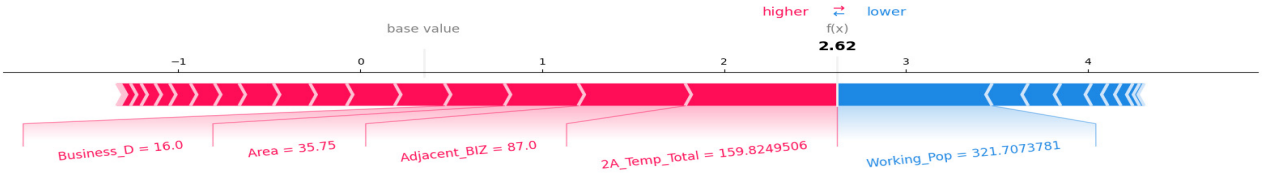
Sample	Model Probability for 5 year survival
(a)	0.9323
(b)	0.8807
(c)	0.0016
(d)	0.0016



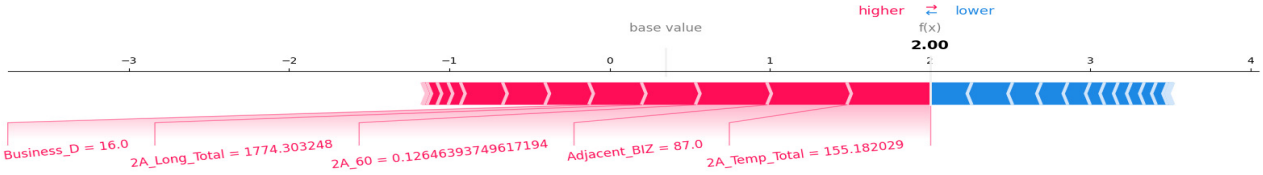
(a) Long Term Survival Sample 1 (b) Long Term Survival Sample 2

(c) Short Term Survival Sample 1 (d) Short Term Survival Sample 2

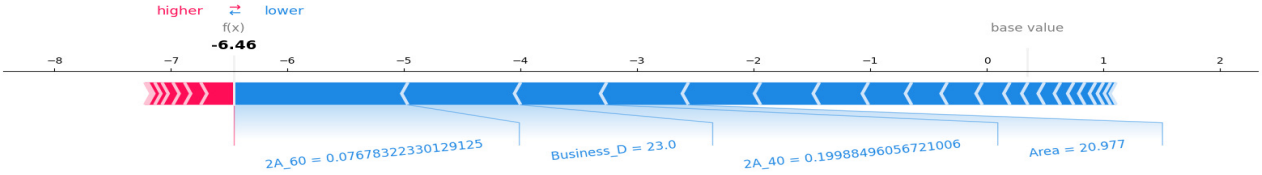
Figure 4. Locations expected to survive for more than 5 year(a,b) and survive less than 6 month(c,d)



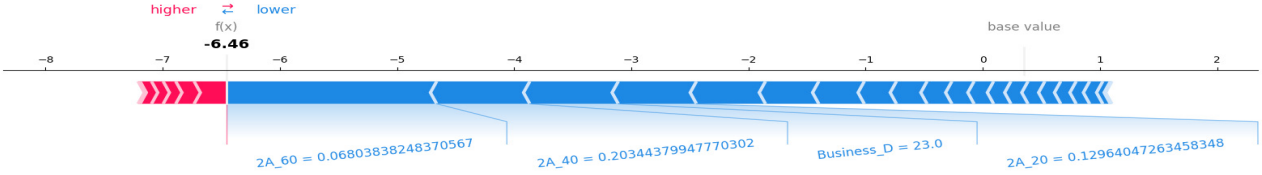
(a) Variable affecting for (a) in <Figure 5>



(b) Variable affecting for (b) in <Figure 5>



(c) Variable affecting for (c) in <Figure 5>



(d) Variable affecting for (d) in <Figure 5>

Figure 5. Variables Affecting for 4 Samples

<Figure 5>는 SHAP을 이용하여 각각의 예측 지점에 대한 변수별 영향력을 가시화한 결과이다. 붉은색은 점포가 5년 이상 생존할 것으로 예측하도록 영향을 주는 변수를 의미하며, 파란색은 6개월 내로 폐업할 것으로 예측하도록 영향을 주는 변수를 의미한다. <Figure 4(a)>는 최종모델이 5년 이상 생존할 것으로 예측하는 신규 입점 후보지 중 생존확률이 가장 높게 산출된 지점으로, <Figure 5(a)>에 따르면 2차 상권 단기 체류 외국인 생활 인구수와 인접 요식업체수 및 요식업체 다양성 그리고 점포 면적이 예측에 긍정적인 영향을 미쳤으며 직장인구의 경우 부정적인 영향을 미쳤다. <Figure 4(b)>는 <Figure 4(a)>에 이어 모델이 두 번째로 높은 장기 생존확률을 기대하는 지점이다. <Figure 5(b)>를 통해 해당 지점은 2차 상권 단기 체류 외국인 생활 인구수, 인접 요식업체수, 2차 상권 60대 내국인 생활 인구비 등의 변수가 점포 생존확률에 대해 긍정적인 영향을 주었음이 확인된다. <Figure 4(c)>와 <Figure 4(d)>는 제안 모델이 6개월 내로 폐업할 것으로 예측하는 신규 입점 후보지 중 장기 생존확률이 가장 낮게 산출된 두 지점이며, <Figure 5(c)>와 <Figure 5(d)>에 따르면 2차 상권 60대 내국인 생활 인구비, 2차 상권 40대 내국인 생활 인구비, 요식업체 다양성이 예측에 주요한 영향을 미침을 확인할 수 있었다.

6. 결 론

본 연구에서는 자영업 장기 지속성 예측과 공간적 실패 요인

규명을 위해 서울시 폐업 요식업 점포를 대상으로 장기 생존 업체를 예측하는 기계학습 모델을 구축하고 XAI를 활용하여 점포의 공간적 특성을 분석하였다. 본 논문은 기존 방법론 대비 Accuracy 0.71, Balanced Accuracy 0.70, AUC 0.78의 안정적이고 우수한 성능을 보여준 XGBoost와 비용민감학습을 활용한 최종 모델을 제안하며, 이를 바탕으로 서울 전역의 신규 분양 상가들을 대상으로 장기 생존확률 예측을 통한 입지 추천을 수행하였다.

제안한 모델은 기존 선행연구(Bang *et al.*, 2018; Jang, 2021)와 비교해 정확성이 우수할 뿐만 아니라, 개별 데이터에 대한 해석이 가능하다는 장점이 존재한다. SHAP을 사용한 주요 상권 요인 분석 결과 자영업 주요 실패 요인으로 인접 유사업체수, 유사업종 다양성, 경쟁업체수와 비율 등의 지리적 특성을 비롯해 2차 상권 60대 생활 인구비, 직장인구수 등의 인구 특성 변수와 점포 면적 등의 개별 점포 특성이 주요 요인으로 선정되었다. 이를 요약하자면 첫째로, 지리적 특성의 경우, 인접 유사업체수 및 경쟁업체수의 증가는 점포의 장기 생존에 부정적 영향을 미치나 경쟁업체 비율은 일정 수준 이상에서는 높을수록 긍정적으로 나타났는데 이는 동종업체 밀집도와 장기 생존 확률이 양의 상관관계를 갖는다는 Kim *et al.* (2018)의 연구와 일치하는 결과이다. 반면 유사업종 다양성의 증가는 기존 연구(Kim *et al.*, 2019)와 상반되는 결과로 장기 생존에 부정적인 영향을 미침이 확인되었으나 13개 이상 19개 미만의 적절한 세부 업종 개수는 반대의 영향을 주는 것이 확인되었다. 따라서 점포의 안정적 영업을 위해선 경쟁업체의 밀도 관리와 상권의 적절한 다양성 확보가 요구됨이 시사된다. 둘째로, 인구 특성의 경우 연령별로 장기 생존에 대한 영향력이 다르게 나타났으며 이는 기존 선행 연구를 통해 발표된 연령별 소비 행태(Fareed and Riggs, 1987; Fernández-Villaverde and Krueger, 2007) 및 상권의 연령대와 자영업 매출액간의 관계(Son, 2021)가 실제 자영업체 수명에 미치는 영향을 확인할 수 있는 결과이다. 이를 통해 선행연구(Nam, 2017; Kim *et al.*, 2018, Lee *et al.*, 2022)에서 제시한 인구수와 점포 생존 간의 기존 연구 결과를 연령별로 세분화하여 파악할 수 있었으며, 상권의 인구 구성에서 20대와 30대 그리고 60대는 감소할수록, 40대와 50대는 증가할수록 요식업체의 단기 폐업확률을 증가시키는 것으로 파악되었다. 특히, 가장 큰 영향력을 미친 2차 상권 60대 생활 인구비 변수는 고령층의 이동성이 크게 떨어져(Noh and Joh, 2008) 지역 상권을 더 많이 이용하는 것이 반영된 것이라고 해석이 된다. 또한 청년층인 20대와 30대는 코로나19 시기 상권의 회복탄력성을 연구한 기존 결과(Lee *et al.* 2022)와 상반되며 이를 통해 일반적인 상황에 적용하기 위한 본 연구의 필요성이 강조된다. 셋째로, 개별 점포 특화 변수 중 가장 큰 영향력을 가진 점포면적은 큰 경우가 장기 생존에 유리하며 이는 선행연구(Kim *et al.*, 2023)와 일치하는 결과이다.

본 연구는 모든 데이터를 공공데이터로 구성하고 독립 변수의 대부분을 상권 정보로 설계하여 모델의 접근성을 높였으며, 서울시 전역을 대상으로 일반화를 진행하고 신규 입지에 대한 생존 확률까지 제시하였으므로 개업을 고민하는 예비 자영업자들의 의사결정에 활용할 수 있음을 보여주었으나 다음과 같은 한계점을 가진다. 우선, 기존 선행연구와 마찬가지로 인구 특성 요인이 점포 생존의 중요한 인자로 판별됐으나, 사용한 생활 인구 데이터는 특정 구역 특정 시점에 존재하는 모든 인구를 산출한 데이터로 유동 인구와 주거 인구를 통합한 성격을 띠기에 상세한 인구 정보 반영이 이루어지지 못하는 한계점을 가진다. 또한 인구 특성 반영 시 청소년과 고령층을 배제하여 데이터를 구축하였으며 이는 이들을 주 고객층으로 하는 세부 업종에 대한 분석에 한계점을 갖는다. 둘째로, 비정형 데이터를 활용한 일부 해외의 연구 사례(Naumzik *et al.*, 2022)에 비해 낮은 정확성을 기록하였다. 이는 구축 데이터가 포착하기 어려운 요식업의 59가지 세부 업종에 따른 이질성에서 기인 된 것으로 보이며, 사업간 구별을 위한 고품질 데이터 사용이 필요함을 의미한다. 따라서 향후 연구에서는 청소년과 고령층을 포함한 유동 인구 데이터와 주거 인구 데이터의 반영 및 소비자 리뷰 데이터와 점포의 세부 업종 등 점포 특화 데이터의 사용을 통해 전반적인 데이터 품질을 향상하여 더욱 정확한 예측 성능을 가진 모델 구축이 요구된다. 셋째로, 생존 주기의 극단에 있는 그룹을 대상으로 장기 생존과 단기 생존으로 나누어 예측 모델을 구축하였기에, 그 사이에 걸쳐있는 점포를 설명하지 못하는 한계점을 가진다. 향후 연구에서는 생존 주기 전역의 점포를 대상으로 분석을 실시하여 더욱 폭넓은 상권 요인의 이해를 가능케 한다면 자영업자들의 의사결정에 유용하게 활용될 수 있을 것이다.

참고문헌

Applebaum, W. (1966), Methods for determining store trade areas, market penetration, and potential sales, *Journal of marketing Research*, **3**(2), 127-141.

Bang, J. A., Son, K. M., Lee, S. J., Lee, H. G., and Jo, S. B. (2018), A Study on Predictive Modeling of Public Data: Survival of Fried Chicken Restaurants in Seoul, *The Journal of Bigdata*, **3**(2), 35-49.

Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013), Evaluation measures for models assessment over imbalanced data sets, *Journal of Information Engineering and Applications*, **3**(10).

Bradley, A. P. (1997), The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, **30**(7), 1145-1159.

Breiman, L. (2001), Random Forests, *Machine Learning*, **45**, 5-32.

Bzdok, D., Altman, N., and Krzywinski, M. (2018), Statistics versus machine learning, *Nature Methods*, **15**(4), 233-234.

Chen, T. and Guestrin, C. (2016, August), Xgboost: A scalable tree boosting system, *KDD'16*, 785-794.

Cortes, C. and Vapnik, V. (1995), Support-vector networks, *Machine Learning*, **20**, 273-297.

Cover, T. and Hart, P. (1967), Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, **13**(1), 21-27.

Elkan, C. (2001, August), The foundations of cost-sensitive learning, *Proc. 17th Int. Joint Conf. on artificial intelligence*, 973-978.

Fareed, A. E. and Riggs, G. D. (1982), Old-young differences in consumer expenditure patterns, *Journal of Consumer Affairs*, **16**(1), 152-160.

Fernández-Villaverde, J. and Krueger, D. (2007), Consumption over the life cycle: Facts from consumer expenditure survey data, *The Review of Economics and Statistics*, **89**(3), 552-565.

Fox, E. J., Postrel, S., and McLaughlin, A. (2007), *The impact of retail location on retailer revenues: An empirical investigation. Unpublished manuscript*, Edwin L. Cox School of Business, Southern Methodist University, Dallas, TX.

Fritsch, M., Brixy, U., and Falck, O. (2006), The effect of industry, region, and time on new business survival: A multi-dimensional analysis, *Review of Industrial Organization*, *28*, 285-306.

Georgiev, P., Noulas, A., and Mascolo, C. (2014, May), Where businesses thrive: Predicting the impact of the olympic games on local retailers through location-based services data, *Proc. 8th Int. AAAI Conf. on Web and Social Media*, 151-160.

Gohel, P., Singh, P., and Mohanty, M. (2021), Explainable AI: current status and future directions, arXiv preprint arXiv:2107.07045.

Haapanen, M. and Tervo, H. (2009), Self-employment duration in urban and rural locations, *Applied Economics*, **41**(19), 2449-2461.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009), *The elements of statistical learning: data mining, inference, and prediction*, **2**, 1-758, Springer, New York.

Hatfield, I. (2015), *Self-employment in Europe, Institute for Public Policy Research*, London

Healy, J. J. and Mac Con Iomaire, M. (2019), Calculating restaurant failure rates using longitudinal census data, *Journal of Culinary Science & Technology*, **17**(4), 350-372.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013), *Applied logistic regression*, 3, John Wiley & Sons, New York

Jang, J. H. (2021), A Study on the Prediction Model for Self-employed Business Closure Using Machine Learning Techniques: Focusing on 25 Autonomous Districts in Seoul, *The Journal of Humanities and Social science (HSS21)*, **12**(1), 1081-1096.

Jeon, S. I. and Oh, J. S. (2023), An Exploratory Study on the Determinants of Business Performance in Small Business Start-ups, *The Journal of Internet Electronic Commerce Research*, **23**(2), 23-37.

Jun, M. J., Kim, J. I., Kwon, J. H., and Jeong, J. E. (2013), The effects of high-density suburban development on commuter mode choices in Seoul, Korea, *Cities*, **31**, 230-238.

Kahn, H. S., Yang, J. H., and Ahn, Y. J. (2020), A Study on the Problems of Self-employment in Korea and Marketing Strategies to Overcome, *Journal of Product Research*, **38**(3), 89-98.

Kang, C-D. (2016), Spatial access to pedestrians and retail sales in Seoul, Korea, *Habitat International*, **57**, 110-120.

Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., and Mascolo, C. (2013, August), Geo-spotting: mining online location-based services for optimal retail store placement, *KDD'13*, 793-801.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... and Liu, T. Y. (2017), Lightgbm: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems*, 30.

Kim, C. M. and Shin, S. M. (2015), A Conceptual Study of the Government Supporting Policy for the Micro Commerce and Self-employed Industries: A Critical Prospective, *The Journal of Industrial Innovation*, **31**(4), 175-205.

Kim, D. J., Kim, K. J., and An, Y. S. (2018), A Study on the Spatial Characteristics of the Long-term Survival Commercial Facilities Location - Focused on Ordinary Restaurant in Gangnam-gu, Seoul, *Journal of Korea Planning Association*, **53**(2), 161-181.

Kim, D. J., Yi, C. H., and Lee, S. I. (2019), A Study on the Survival Characteristics of the Restaurant Business in Major and Side-Street Trade Areas, Seoul, *Journal of Korea Planning Association*, **54**(5), 76-90.

Kim, D. W. and Byeon, J. W. (2022, December), Immigration and Local Economic Growth: An empirical evidence from South Korea, *Migration Research & Training Centre Reserach Report Series* 2022-01.

Kim, I. G. (2018), A Study on the Present Status of Individual Businesses and the Characteristics by Industry in Korea, *Regional Industry Review*, **41**(3), 343-364.

Kim, I. S. (2008), *Analysing Policy Effectiveness and Future Policy Directions in Trade Area Information Systems*, Korea SMEs & Startups Institute.

Kim, J-S., Seo, K-H., Lee, H-S., and Kim K-M. (2022), The Effect of Seoul Retail Area Characteristics on Its Survival, *Journal of Korea Planning Association*, **57**(1), 75-90.

Kim, S. H., Yi, C. H., and Nam, J. (2023), Survival Rate and Survival Factors of the Restaurant according to the Decline Level in Seoul, *Journal of Korea Planning Association*, **58**(3), 68-81.

Kim, Y. W., Kim, M. G., and Kim, Y. M. (2022), Prediction of patent lifespan and analysis of influencing factors using machine learning, *Journal of Intelligent Information Systems*, **28**(2), 147-170.

Kim, J. C. (2017), Self-Employed Households' Debt Structure and the Implications, *Korea Capital Market Institute*, Retrieved from https://www.kcmi.re.kr/en/publications/pub_detail_view?syear=2017&zcd=002001017&zno=1341&cno=463.

KOSIS (Statistics Korea) (2020), Survival Rate of New Businesses by Industry.

Lavazza, L., Morasca, S., and Rotoloni, G. (2023, June), On the Reliability of the Area Under the ROC Curve in Empirical Software Engineering, *Proc. 27th Int. Conf. on Evaluation and Assessment in Software Engineering*, 93-100

Lee, J. H. (2019), Reseach on Ways to Utilize the Gentrification Diagnosis System to Advance the Seoul's Trade Area Analysis System, *Korea Reseach Institute for Human Settlements*.

Lee, J. K. (2021), *Causes of Self-Employment Management Difficulties and Policy Directions*, KDI Policy Study 2021-12.

Lee, S. M. and Yoo, H. B. (2022), Research on Factors Affecting the Closure Rate of Small Businesses in Seoul during Covid-19, *The Korea Local Administration Review*, **36**(3), 57-86.

Lee, S., Kim, T. G., and Kim, K. S. (2022), A Study on Resilience and Business Crisis on Seoul's Side Street Trade Areas during the COVID-19 Pandemic, *Journal of Korea Real Estate Analysts Association*, **28**(2), 7-25.

Lee, Y. S., Park, H. S., Lew, S. H., and Kang, J. M. (2014), An analysis of the location factors that affects the sales of campus commercial district. *Seoul Studies*, **15**(1), 17-34.

Li, H., Bruce, X. B., Li, G., and Gao, H. (2023), Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews, *Tourism Management*, **96**, 104707.

Lum, B. W. (2017), *Business strategies for small business survival* (Doctoral dissertation, Walden University).

Lundberg, S. M. and Lee, S. I. (2017), A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems*, **30**.

Ministry of Justice (2022), Immigration Statistics.
Muller, C. and Woods, R. H. (1991), The real failure rate of restaurants, *Hospitality Review*, **9**(2), 7.
Nam, Y. H. (2015), A Comparative Study on Support Policy for Micro Business and Social Costs, *Journal of SME Policy*, **2015**(16), 1-133.
Nam, Y. M. (2017), Analysis on the Determinants of Exit of Self-Employed Businesses in Korea, *Bank of Korea WP*, **2017**(5).
Naumzik, C., Feuerriegel, S., and Weinmann, M. (2022), I Will Survive: Predicting Business Failures from Customer Ratings, *Marketing Science*, **41**(1), 188-207.
Noh, H. B. and Chung, N. K. (2016), Cross Country Comparison on Policy Support for Bankrupt Small Businesses: Cases of Korea, Germany and Japan, *Ordo Economics Journal*, **19**(1), 69-84.
Noh, S. H. and Joh, C. H. (2008), Travel pattern differences between age groups in Seoul Metropolitan Area, *Korean Geographical Society Conference*, 261-268.
OECD (2023), Self-employment rate (indicator). doi: 10.1787/fb58715e-en (Accessed on 27 November 2023).
Oh, S. J., Lee, J. H., Kim, H. K., and Shin, J. C. (2015), Sales determinants of restaurant chain business: Focused on family restaurants in Korea, *Indian Journal of Science and Technology*, **8**(23), 1.
Parsa, H. G., Self, J., Sydnor-Busso, S., and Yoon, H-J. (2011), Why restaurants fail? Part II-The impact of affiliation, location, and size on restaurant failures: Results from a survival analysis, *Journal of Foodservice Business Research*, **14**(4), 360-379.
Parsa, H. G., Van Der Rest, J. P. I., Smith, S. R., Parsa, R. A., and Bujisic, M. (2015), Why restaurants fail? Part IV: The relationship between restaurant failures and demographic factors, *Cornell Hospitality Quarterly*, **56**(1), 80-90.
Prasad, C. J. (2010), Effect of consumer demographic attributes on store choice behaviour in food and grocery retailing-an empirical analysis, *Management and Labour Studies*, **35**(1), 35-58.
Ribeiro, M. T., Singh, S., and Guestrin, C. (2016, August), Why should i trust you? Explaining the predictions of any classifier, *KDD'22*, 1135-1144.
Shapley, L. S. (1953), *Contributions to the Theory of Games (AM-28), Volume II*, 307-317, Princeton University Press, Princeton.
Son, K. -M. (2021), *The Effect of Spartial Distribution of De Facto Population on Commercial Sales Volumn in Seoul: Using Big Data and Panel Model* (Master's thesis). Chung-Ang University.
Song, Y. Y. and Ying, L. U. (2015), Decision tree methods: applications for classification and prediction, *Shanghai Archives of Psychiatry*, **27**(2), 130-135.
Statistics Korea (2021), Administrative statistics on enterprise birth and death for the year 2021.
Statistics Korea (2023, September), Employment Trend.
Sung, H. -G. (2022), Estimating the spatial impact of neighboring physical environments on retail sales, *Cities*, **123**, 103579.
Sung, N. -I. and Kim, J. -K. (2020), Entry and exit of small self-employed businesses in Korea's service industries, *Small Business Economics*, **54**(2), 303-322.
Tao, J. and Zhou, L. (2020), Can Online Consumer Reviews Signal Restaurant Closure: A Deep Learning-Based Time-Series Analysis, *IEEE Transactions on Engineering Management*, **70**(3), 834-848.
Yang, X. X. (2016), *Key Success Factors of Small Business in a Southern Region of California* (Doctoral dissertation). Walden University.
Zhang, M. and Luo, L. (2023), Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp, *Management Science*, **69**(1), 25-50.

저자소개

장호준: 성균관대학교 시스템경영공학과 학석사과정 재학 중이다. 관심분야는 최적화, 데이터 분석이다.

이기환: 성균관대학교 시스템경영공학과 학사과정 재학 중이다. 관심분야는 머신러닝 알고리즘, 데이터 분석이다.

이희상: 서울대학교 산업공학과에서 학사학위와 석사학위를 취득하고 Georgia Tech에서 Industrial & Systems Engineering 박사학위를 취득하였다. KT 선임연구원, 한국외국어대학교 조교수/부교수를 역임하고 2004년부터 성균관대학교 시스템경영공학과에서 교수로 재직 중이다. 연구분야는 경영과학, 비즈니스 애널리틱스, 기술경영이다.