Journal of the Korean Institute of Industrial Engineers
Vol. 49, No. 1, pp. 37-45, February 2023.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2023.49.1.037

머신러닝 알고리즘을 이용한 커피 프랜차이즈 실패 요인 분석에 관한 연구

안예린[1] · 이현희[1] · 유성민[1] · 김서연[2] · 박민서[2†]

[1]서울여자대학교 데이터과학전공/ [2]서울여자대학교 데이터사이언스학과

Analysis of Coffee Franchise Failure Factor using Machine Learning Algorithms

Yelyn Ahn[1] · Hyunhee Lee[1] · Sungmin Ryu[1] · Seoyeon Kim[2] · Minseo Park[2]

[1]Major of Data Science and Engineering, Seoul Women's University
[2]Department of Data Science, Seoul Women's University

The restaurant franchise headquarters and the franchisees are organically connected, so the failure of the franchisee affects the franchise headquarters. Their sales affect the unemployment rate, causing social and economic crises. This study aims to analyze factors that affect low sales stores that are closely related to franchise failure. To meet the objectives, we first collect Gangnam-gu POS(Point of Sales) data and then analyze them with various machine learning algorithms. Our results show that LGBM(Light Gradient Boosting Machine) has the highest performance (accuracy 0.908). We apply the results with SHAP(Shapley Additional exPlanations), which is an explainale AI, to visualize the positive and negative effects of variables. In the near future, this study is expected to be utilized in suggesting a store operation strategy that can reduce the probability of franchise closure.

Keywords: Coffee Franchise, Sales Failure Factors, Machine Learning, XAI(eXplainable AI), SHAP(SHAPley Additive Explanations)

1. 서 론

프랜차이즈는 가맹본부가 본부의 상품과 서비스를 효과적으로 판매하기 위해 일정 금액을 받고 가맹점에 브랜드와 노하우를 제공하는 사업형태이다(Kim *et al.*, 2014). 이러한 사업 형태로 인해 사업에 대한 경험이나 지식이 부족한 신규창업자들이 프랜차이즈 창업을 선호한다. 실제 공정거래위원회가 발표한 2021년도 가맹사업 현황에 따르면 현재 등록된 가맹점은 270,485개로 전년 대비 4.5% 증가한 것으로 나타났다. 그러나, 전체 가맹점 평균 매출액은 반대로 전년 대비 5.5% 감소하였고, 특히 외식업종의 경우 평균 매출액이 9.0% 감소하였다(Korea Fair Trade Commission, 2022).

외식업은 타 업종에 비해 쉽게 대체재를 찾을 수 있고, 규모가 작은 업종이 많아 고객 이탈이 쉬워 매출액의 변화가 클 수밖에 없다(Kim *et al.*, 2021). 또한 타 업종에 비해 경쟁이 치열하고, 매장 주변의 상권특성과 같은 외부적인 요인으로 인해 폐업이 증가하는 특성을 갖고 있다(Lee *et al.*, 2021).

이처럼, 외식업종의 프랜차이즈는 사업에 대한 지식이나 경험이 부족하더라도 쉽게 창업이 가능하여 창업율이 높게 나타나지만, 생존하는 것은 쉽지 않다. 특히, 외식업 프랜차이즈의 경우 가맹점과 가맹본부가 유기적으로 연결되어 있어(Kim *et al.*, 2014) 가맹점의 생존이 가맹본부의 생존에 직접적인 영향을 미친다. 가맹점의 매출이 불안정하면 비용 절감을 위해 종업원 수를 조정(Statistics Korea, 2020)하기 때문에 가맹점의 위

† 연락저자 : 박민서 교수, 01797 서울특별시 노원구 화랑로 621 서울여자대학교 데이터사이언스학과, Tel : 02-970-7504, E-mail : mpark@swu.ac.kr
2022년 8월 23일 접수; 2022년 10월 11일 수정본 접수; 2022년 11월 15일 게재 확정.

기는 실업률에도 영향을 미친다. 이처럼, 가맹점의 생존 여부는 가맹본부를 넘어 사회적, 경제적 영향을 동반하기 때문에 실패할 확률이 상대적으로 높은 매출 하위 영향 요인 분석이 필요하다.

기존의 매출 영향 요인에 관한 연구는 전체 매장을 대상으로 긍정적인 요인에 대한 분석을 중심으로 제공되고 있다. 하지만, 폐업할 확률이 높고, 폐업과 실질적으로 연관이 있는 매장은 매출이 낮은 하위 매출 매장이기 때문에 해당 매장에 집중한 영향 요인 분석이 반드시 필요하다. 따라서, 본 연구에서는 상대적으로 생존이 어려운 매출 하위 프랜차이즈 매장을 분류하고, 해당 매출에 대한 영향 요인을 분석하고자 한다. 강남구에 존재하는 다수의 프랜차이즈 매장 중 M커피 프랜차이즈를 매출 하위 매장과 그 외 매장으로 분류하는 모형을 구축하였다. 독립변수는 기존 연구 고찰 결과 시간대, 입지, 생활인구 수가 매출에 영향을 미치는 것으로 나타나 세 요인을 채택해 매출 하위 매장 분류에 활용하였다. 또한, 머신러닝 알고리즘을 사용하였고, 설명 가능한 인공지능 기법 중 하나인 SHAP(SHapley Additional exPlanations) 기법을 활용하여 시각화 하였다.

본 논문의 구성은 다음과 같다. 제2장에서는 매출 영향 요인과 매출 예측에 대한 선행 연구 고찰을 수행하고, 제3장에서는 매출 하위 매장 분석에 효과적인 머신러닝 알고리즘을 기술한다. 제4장에서는 본 논문에서 제안하는 알고리즘을 설명하며, 제5장과 제6장에서는 실험 결과 및 결론을 서술한다.

2. 선행 연구 고찰

2.1 매출 예측에 관한 선행연구

본 연구는 커피 프랜차이즈 매장을 타겟팅으로 하고 있으며, 매장 매출을 바탕으로 실패 요인을 분석하고자 한다. 따라서 식음료 업종을 대상으로 매출 예측 및 매장 매출 영향 요인을 분석한 선행연구를 살펴보았다.

경주소재 특1등급 호텔 5곳의 식음료 부문의 매출액을 예측한 연구에서는 1995년 1월부터 2004년 12월까지 식음료 월별 매출액 자료를 바탕으로 2005년의 매출액을 예측하였다. ARIMA(Auto-Regressive Integrated Moving Average)의 계절변동 모형을 사용하여 4가지 모델을 설계하고, 아카이케 정보기준과 베이지안 정보기준에 따라 ARIMA(0,1,1)×(0,1,1)모형을 가장 적합한 모형으로 선정하여 매출을 예측하였다. 호텔 식음료부서의 성수기 및 비수기, 주말과 주중에 회전율 차이가 있다고 보고 이에 예측 데이터를 바탕으로 비수기에 대한 대응전략이 필요함을 시사하였다(Son *et al.*, 2005).

소상공인의 매장 운영 성과에 영향을 미칠 수 있는 요인을 분석한 연구에서는 실무적인 측면에서 실제 성공 사례를 바탕으로 매장 운영 성과에 미치는 요인을 제시하였다. 베이커리, 레스토랑 등을 운영하는 소상공인의 실제 사례를 분석함으로써 소셜커머스를 활용한 홍보, 공동 인프라 활용 등 매장 운영의 성공가능성을 높일 수 있는 시사점을 제공하였다(Hong *et al.*, 2019).

식음료업의 매출 예측에 대한 연구는 대부분 특정 음식점, 호텔 내 레스토랑 등 개인 창업 음식점과 성공요인에 초점이 맞춰져 있으며 프랜차이즈 브랜드 및 실패요인을 대상으로 진행한 연구는 부족한 실정임을 확인할 수 있다. 앞 장에서 언급한 프랜차이즈 브랜드의 특성상 매출 하위 매장을 고려한 분석이 필요하다. 프랜차이즈 하위 매장의 매출에 미치는 요인을 머신러닝 알고리즘을 활용하여 분석하고자 한다.

2.2 매출에 영향을 미치는 요인

매출에 영향을 미치는 요인에 대한 선행연구를 검토한 결과 시간대, 인구특성, 입지특성, 매장내부 특성 등 다양한 요인이 매출에 영향을 미치는 것으로 보고되었다.

커피를 소비하는 서울시민을 대상으로 커피 전문점 이용 특성을 분석한 연구에서는 선호하는 커피 전문점 형태와 이용시간대에 따라 이용 빈도가 다르게 나타났다. 커피전문점 이용 빈도는 오전(9-12시) 시간대는 18.7%, 오후(14-17시) 시간대는 21.7%, 저녁(17-20시) 시간대는 41.4%로 시간대에 따라 이용 빈도가 상이한 것으로 나타났다(Jung *et al.*, 2012)

서울특별시 강남구의 K프랜차이즈 치킨 가맹점 33개의 매출 데이터를 기반으로 매장의 성패를 예측한 연구에서는 머신러닝 기법을 활용하여 성공 및 실패를 분류하였다. 성공 예측에는 Random Forest가 정확도 0.92, 실패 예측에는 Decision Tree가 정확도 0.92로 가장 우수한 모델로 선정되었다. 특히 예측 성능에 영향을 주는 요인으로 매장 반경 500m의 상업 지역 면적이 두 모델에서 모두 유의미한 요인으로 나타났다(Ahn *et al.*, 2022).

서울시 6개구의 14개 세부 음식업종을 대상으로 매출액에 미치는 영향을 비교한 연구도 있다. 공간회귀분석을 활용하여 매장의 입지요인이 매출액에 미치는 영향을 비교 분석하였다. 유동인구, 주변 사업체 종사자 수, 집계구 면적 등을 설명 변수로 설정하여 연구를 진행하였다. 특히 유동인구는 음식업종의 매출액에 정의 영향을 미치는 것으로 나타났다(Noh and Lee, 2018).

검토한 선행연구를 참고하여 본 연구의 종속변수를 시간대, 유동인구, 상업지역 면적으로 구성하였다. 커피전문점의 입지특성과 매출액 간의 관계를 분석한 연구(Shin and Shin, 2010; Shin and Moon, 2011)와 도보로 이동 가능한 거리를 고려하였을 때, 업소 반경 300m 이내의 특성이 유의미하다고 판단하여 유동인구 및 상업지역 면적은 업소 기준 반경 300m의 특성만 반영하였다.

3. 매출 하위 매장 분류 및 분석에 사용한 알고리즘

본 장에서는 머신러닝 기법 중 대표적인 분류 모델 알고리즘과 XAI(eXplainable AI)기법 중 하나인 SHAP기법에 대해 기술한다.

3.1 Decision Tree

Decision Tree는 의사결정규칙 트리를 활용하여 데이터를 여러 개의 유사한 소집단으로 세분화하여 분류하는 알고리즘이다. 분류 과정이 트리 모양 구조로 시각화되기 때문에 신경망과 같은 타 알고리즘에 비해 과정을 이해하기 쉽고, 설명이 가능하다는 장점을 가지고 있다(Song *et al.*, 2009). 하지만, 하나의 트리를 이용해서 분석을 하기 때문에, 구조가 복잡할수록 해석이 어렵고 과적합이 발생할 가능성이 높아 가지치기 같은 조기 종료 조건을 설정해야 한다는 단점이 있다(Bramer, 2007).

3.2 Random Forest

Random Forest는 여러 개의 Decision Tree 조합을 만드는 배깅(Bagging)을 통해 확장한 분류 알고리즘이다. 데이터 내에서 복원 추출하는 부트스트랩(Bootstrap)을 통해 만들어진 여러 개의 부분집합(Subset)을 사용하여 수많은 트리 분류기를 생성한다. 이후, 학습과 예측을 진행하며 집계(Aggregation)단계를 통해 예측 값들의 평균을 최종 예측 값으로 결정한다. 데이터의 전체집합을 학습에 사용하는 Decision Tree에 비해, 무작위의 종속변수로 이루어진 부분집합들을 학습에 사용하기 때문에 결측치가 많은 데이터에도 좋은 성능을 보이는 장점을 가지고 있다(Breiman, 2001).

3.3 XGBoost(eXtreme Gradient Boosting)

앙상블(Ensemble)은 하나의 트리 분류기를 사용하는 Decision Tree와 달리 수많은 성능이 약한 분류기를 결합하여 더 강력한 분류기를 만드는 기법이다. 그중 GBM(Gradient Boosting Machine)은 약한 트리 분류기들을 순차적으로 실행하며 경사하강법(Gradient descent)을 통해 학습 오차가 작아지는 지점에 가중치를 부여해 보완하며 더 강력한 분류기로 개선하는 알고리즘이다(Natekin and Knoll, 2013). XGBoost는 학습 시간이 오래 걸리는 GBM의 단점을 GPU 병렬 처리를 통해 해결한 모델이다. 또한, 최대한 균형 잡힌 트리를 유지하는 깊이중심(Level-wise) 분할 방법을 통해 모델의 과적합 문제를 해소하는 등 큰 규모의 데이터에서 안정적이고 훈련 속도가 빨라 분류와 회귀에 많이 사용되는 알고리즘이다(Chen and Guestrin, 2016).

3.4 LGBM(Light Gradient Boosting Machine)

LGBM은 같은 GBM 기반의 모델로 깊이중심 분할 방법을 사용하는 XGBoost와 모델과 달리, 최대 손실을 가지는 리프 노드를 지속적으로 분할하는 리프 중심(Leaf-wise) 분할 방법을 사용한다. 이를 통해 균형 잡힌 트리를 최대한 유지하기 위한 XGBoost보다 훈련 속도가 더 빠른 장점을 가지고 있다(Ke *et al.*, 2017).

3.5 SVM(Support Vector Machine)

SVM는 학습을 진행하며 데이터 분류 기준을 선형 회귀식으로 정의하고, 해당 수식을 바탕으로 최적의 분류 경계인 초평면(Hyper-plane)을 찾아 어떤 카테고리에 분류될지 판별하는 하는 이진 선형 분류 알고리즘이다. 데이터 중 초평면과 가장 가까이 위치한 것을 서포트벡터(Support Vector)라고 하며 초평면과 서포트벡터 간의 거리(Margin)가 최대인 초평면을 찾는 방향으로 학습을 진행한다(Cortes and Vapnik, 1995). 또한, 비선형의 형태를 가진 데이터도 고차원으로 사상(Mapping)하여 선형의 형태로 만드는 커널 기법을 활용한다. 대중적으로 사용되는 커널 함수의 종류로는 다항식(Poly nominal), RBF(Radial Basis Function), 다층 퍼셉트론(Multi-Layer Perceptron) 커널 함수 등이 있다(Hastie *et al.*, 2009). 이를 통해 저차원이나 고차원의 데이터에서 모두 좋은 성능을 보이는 장점을 가지고 있다.

3.6 SHAP(SHapley Additional exPlanations)

SHAP은 데이터에 변형을 주어 설명이 불가능한 예측 모델의 가중치(Weight)를 계산하는 LIME(Ribeiro *et al.*, 2016)과 게임이론을 바탕으로 여러 특성 조합을 구성하여 특정 변수를 활용한 예측치와 평균 예측치 차를 이용해 예측 기여도를 파악하는 Shapley value를 연결한 이론이다. 각 독립변수의 예측 기여도 계산을 통해 Feature Importance(변수 중요도)를 파악할 수 있으며, SHAP value가 양수이면 예측 결과에 양(+)의 영향, 즉 긍정적인 영향을 미쳤다고 해석할 수 있으며, 음수이면 부정적인(-) 영향을 미쳤다고 해석할 수 있다. 또한 사람의 직관과 부합하게 독립변수가 예측 결과에 미치는 영향 요인을 이해할 수 있으며, 이를 통해 예측 모델의 판단 근거를 제시해 모델의 투명성을 확보할 수 있는 장점을 가지고 있다(Lundberg and Lee, 2017).

4. 제안하는 프랜차이즈 실패 요인 분석 알고리즘

프랜차이즈 실패 요인을 분석하기 위해서는 우선 매출 하위 매장을 구분해야 한다. 적절한 기준을 도출하여 하위 매장을 분류(Classification)를 한 후, 하위 매장과 그 외 매장의 두 데이터셋의 특성을 비교한다. 이때, 머신러닝의 지도학습(Supervised Learning)을 적용하였다. 특히, 본 연구에서는 Optuna를 사용하여 파라미터 최적화를 수행하였으며, 예측 모델의 결과 도출 과정을 직관적으로 이해할 수 있게 하는 설명 가능한 인공지능(eXplainable AI, XAI)기법을 활용하여 해당 매출에 영향을 미치는 요인을 시각화하였다.

프로세스의 자세한 설명은 다음과 같다.

4.1 데이터 수집

프랜차이즈 매출 데이터를 분석하기 위해 서울특별시 강남구에 위치한 18개 프랜차이즈 M커피 가맹점에 대한 한 달 간의 매출 데이터(2019년 2월 POS 데이터)를 수집하였다. 영수증 단위 데이터로 매장코드, 매장명, 영업일자, 영수증번호, 결제일시, 테이블번호, 총판매금액, 소분류명, 상품명, 상품코드, 판매수량, 판매단가 정보를 담고 있으며, 70,786개 행으로 구성되어 있다.

이와 함께, 매장 주변의 상업지역 면적과 생활인구가 매출 예측에 영향을 미친다는 선행연구(Ahn *et al.*, 2022; Noh and Lee, 2018)를 기반으로 상업지역 면적과 생활인구를 독립변수를 선정하고 공공 데이터를 수집하였다. 점포의 실제 좌표 정보를 도출하기 위해 서울시 열린데이터광장(https://data.seoul.go.kr)의 강남구 휴게음식점 인허가정보(2019) 데이터를 수집하였다. 2019년 기준 강남구의 다류 및 아이스크림류를 조리하여 판매하는 업소의 관리번호, 인허가일자, 영업상태, 전화번호, 소재지 면적 및 주소, 사업장명, 좌표정보(X, Y), 종사자수, 총규모 등 업소에 관한 정보를 담고 있으며, 중부원점TM(EPSG:2097) 좌표계를 기준으로 업소의 좌표정보가 표기되어 있다.

또한 실제 소비하는 인구수를 분석하기 위해 특정 시점에 강남구에서 생활하는 생활인구수를 도출하고자 동사이트의 행정동별 서울생활인구(2019) 데이터를 수집하였다. 2019년 2월 기준 강남구의 행정기관별 인구수, 구성비, 성비, 세대수, 세대당 인구 정보를 담고 있으며, 23개 행으로 구성되어 있다. 상업지역 면적을 추출하기 위해 환경부 환경공간정보서비스(https://egis.me.go.kr)의 세분류 토지피복지도(2019)를 사용하였다. UTM-K(GRS80타원체)(EPSG:5179) 좌표계를 기반으로 해상도 1M 급으로 이루어진 지도로 지형지물을 41개 항목으로 분류한 공간정보를 담고 있다. 더불어, 매장이 위치한 행정동 정보를 추출하기 위해 통계청 통계지리서비스(https://sgis.kostat.go.kr)의 2019년 기준 센서스용 행정구역경계(시군구) 및 센서스용 행정구역경계(읍면동) 데이터를 획득하였다. UTM-K(GRS80타원체)(EPSG:5179) 좌표계를 기반으로 제작된 지도로, 각 행정구역 경계 정보가 담겨있다.

4.2 데이터 정제

POS(Point of Sales) 데이터 중 데이터 분석에 필요한 매장코드, 매장명, 영업일자, 결제일시, 총판매금액을 제외한 나머지 데이터는 삭제하였다. 강남구는 오피스 밀집 지역으로 유입 인구의 대부분이 통근 인구이며, 대부분의 소비가 주중에 몰리는 경향이 있다(KOSIS, 2020). 따라서 해당 지역의 특수성을 고려하여 주말을 제외한 주중 매출 데이터만 사용하였으며, 데이터의 규칙성을 위해 명절의 매출 데이터는 제거하였다. 또한 8자리 숫자 형태로 표현되었던 결제일시를 시계열 데이터 형태로 변경하여 결제 시간대를 추출하였다.

강남구 휴게음식점 인허가 정보 중 사업장명을 기준으로 M커피 가맹점 이외의 데이터는 모두 삭제하였다. 좌표 정보를 토지피복지도 및 행정동 경계 지도와 병합하기 위하여 EPSG:5179 - Korea 2000 / Unified CS 좌표계로 변환하였다. 또한 세분류 토지피복지도의 상업지역만을 사용하기 위해 환경공간정보서비스에서 상업지역으로 정의하고 있는 상업 및 업무시설(131), 혼합지역(132) 면적 이외의 데이터는 모두 제거하였다.

매장 좌표 정보, 토지피복지도, 행정동 경계 지도를 병합하였으며, 소비자가 커피전문점에 접근하는데 도보를 선호하기 때문에 도보 5분 거리인 매장 반경 300m의 특성이 매출에 유의미하다는 선행연구(Shin and Shin, 2009; Shin and Moon, 2011)를 기반으로 매장 반경 300m 이외의 데이터는 매출 예측에 영향이 미미할 것으로 판단하여 제거하였다.

4.3 변수 추출 및 가공

정제된 데이터 안에서, 매출과 연관된 독립변수를 추출하였다. 시간대, 상업지역 면적, 생활인구수가 독립변수로 사용되었다.

시간대의 경우, 시간대별로 커피전문점 이용 빈도가 상이하다는 선행연구(Jung *et al.*, 2012)에 따라 결제 시간별로 나누어서 분석하였다. 추출한 결제 시간대 별 결제된 가장 이른 시간부터 가장 늦은 시간 까지를 영업시간으로 하였을 때, 선행연구에서 설정한 시간보다 M커피 가맹점의 영업시간이 긴 것을 고려하여 본 연구에서는 세 가지 범주가 아닌 오픈(1), 미들 상(2), 미들 하(3), 마감(4)의 네 가지 범주로 설정하였다. 상업지역 면적이 매출 예측에 영향을 끼친다는 연구결과(Ahn *et al.*, 2022)를 기반으로 점포 반경 300m의 상업지역 면적 합을 QGIS(ver 3.22 Biatowiza)를 활용하여 점포별로 계산하였다. 생활인구수에 따라 매출이 변동된다는 선행연구(Noh and Lee, 2018)를 바탕으로 점포가 위치한 행정동에 따른 생활인구수를 활용하려 했으나, 반경 300m이내에 2개 이상의 행정동이 걸쳐있는 경우가 다수 존재하였다. 따라서 점포별로 반경 300m 내에 존재하는 각 행정동의 비율을 QGIS로 추출한 뒤, 행정동의 비율만큼 생활인구수를 곱한 값을 변수로 활용하였다.

종속변수는 앞서 정의한 네 가지 시간대를 기준으로 M커피 가맹점의 매출을 시간대별 평균 매출로 재가공하여 분류하였다. 이때, 각 점포를 동일한 조건으로 비교하기 위해 시간대별 평균 매출을 평균 생활인구로 나누어 1만 생활인구 당 평균 매출을 산출하여 하위 매장을 구분하였다. <Figure 1>은 산출한 1만 생활인구 당 시간대별 평균매출 데이터의 분포를 나타낸다. 큰 원 그래프는 1000원 단위 별 평균매출 데이터 분포를 나타내고, 작은 원 그래프는 매출 금액을 5,000원 이하인 데이터를 하위(Lower), 그 외 데이터(Other)를 나머지로 정의하였을 때의 분포를 나타낸다. 해당 그래프를 살펴보았을 때, 5,000원 이하인 데이터의 개수가 절반 가까이의 높은 비중을 차지한다

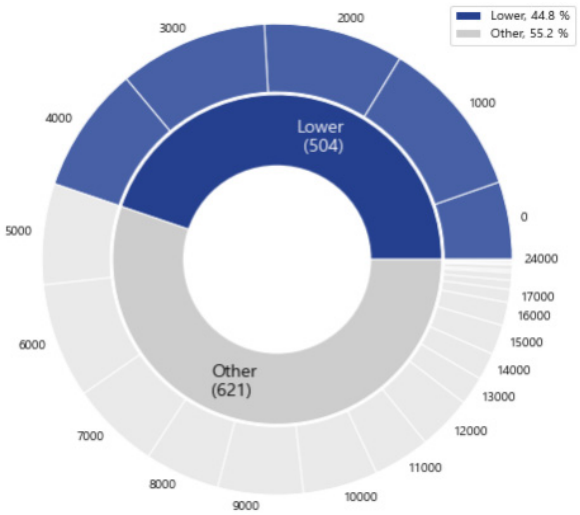

Figure 1. Distribution of Average Sales per De Facto Population in "M" Coffee

Table 1. Description of Variables

Variables		Description	Unit	Type	Range
Dependent Variable	LOWER_ GROUP	Under 5,000's Average sales by time per 10,000 de facto population	Lower group(1)	Category	
		Over 5,000's Average sales by time per 10,000 de facto population	Other group(0)		
Independent Variables	TIME	Open(6 am to 10am)	1	Category	
		Middle Ⅰ(11am to 2pm)	2		
		Middle Ⅱ(3pm to 6pm)	3		
		Closing(7pm to 11pm)	4		
	COMMERCIAL _AREA	Sum of commercial area with a radius of 300m	㎢	Numeric	(269935.2, 45141.24)
	DE_FACTO_ POPULATION	(Ratio of admin with a 300m radius of the store)* (Average daily active population by admin)		Numeric	(194179.53, 34224.16)

는 것을 알 수 있다. 따라서, 1만 인구 당 평균매출이 5,000원 이하이면 매출 하위 매장(Lower group, 1)으로, 그 외의 경우는 나머지(Other group, 0)로 정의하여 변수로 사용하였다. 위 과정을 통해 모델 학습에 사용된 데이터는 최종 1,125개의 행으로 구성되었으며, 추출된 변수는 <Table 1>과 같다.

4.4 모델링

앞서 데이터 정제와 변수 추출 및 가공 과정을 거친 최종 변수를 다섯 가지 분류 모델(Decision Tree, Random Forest, XGBoost, LGBM, SVM)을 사용하여 학습을 진행하였다. 최종 도출된 데이터는 7.5:2.5 비율로 훈련 데이터와 검증 데이터로 나누었으며 843개의 훈련 데이터는 예측모형 구축에 사용되었고, 282개의 검증 데이터는 예측모델 검증에 사용되었다. 각 모델의 하이퍼파라미터 최적화를 위해 최적화 프레임워크인 Optuna를 사용하였다. 또한 모델의 과적합을 방지하기 위해 5-fold cross validation을 실행하였다. 다양한 성능 지표를 활용하여 최종 제안 모델을 선정하였으며, SHAP기법을 활용하여 각 독립변수의 영향력을 시각화 하여 확인하였다. <Figure 2>는 매출 하위 매장 분석 모델링의 전체 프로세스를 나타낸다.

본 연구에서 사용한 하이퍼파라미터 최적화 프레임워크 Optuna는 사용자가 API 형식으로 목표 접근 방법을 직접 정의할 수 있으며, 최고 또는 최적의 하이퍼파라미터 조합을 자동으로 찾아 가지치기(Pruning)와 검색 절차(Searching procedures)에 모두 효율적인 오픈소스이다 (Akiba *et al.*, 2019). Optuna의 최적화 알고리즘은 반복(Trial) 동안 Loss function으로 평가된 하이퍼파라미터의 기록을 사용하여 베이지안 최적화 방법으로 평가될 하이퍼파라미터의 Sample을 제안하는 TPE(Tree-Structured Parzen Estimator)알고리즘을 사용하였다 (Bergstra el al., 2011). <Table 2>는 Optuna의 검색 파라미터와 Optuna를 통해 최종 도출된 예측 모델의 하이퍼파라미터를 보여준다.

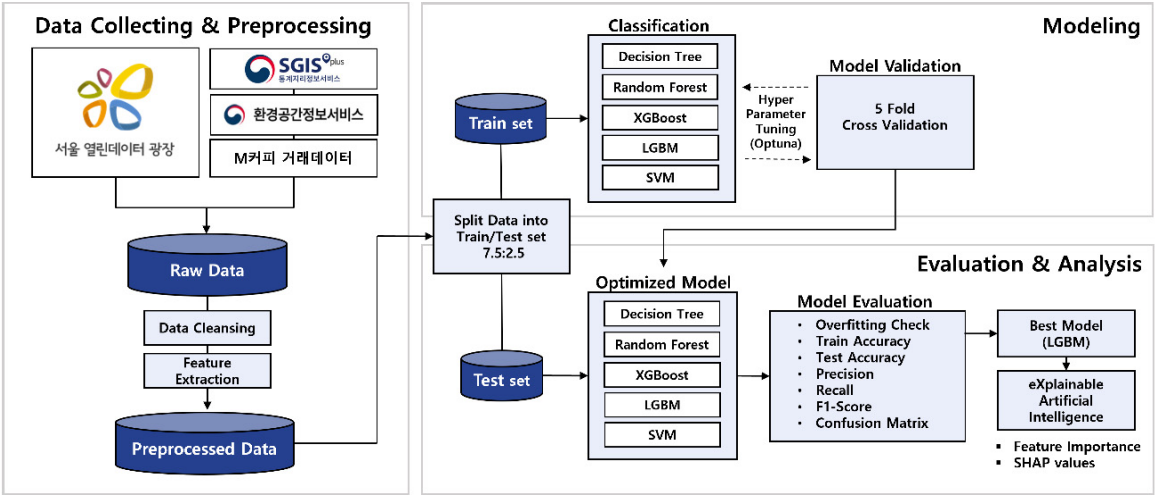

Figure 2. Flow Chart of Low Sales Franchise Store Classification Model

Table 2. Parameters of Optuna and Derived Hyperparameters of Five Models (Decision Tree, Random Forest, XGBoost, LGBM, SVM)

Parameters of Optuna					
Trial	100	Cross Validation	5 fold	Optimization Algorithm	TPE Algorithm
Model	Hyperparameters				
Decision Tree	max_depth	min_samples_split		min_samples_leaf	
	5	43		2	
Random Forest	max_depth	min_samples_leaf	n_estimators		max_features
	4	7	327		0.120802808
XGBoost	max_depth	min_child_weight	n_estimators		learning_rate
	5	7	115		0.485036366
LGBM	max_depth	min_child_samples	n_estimators	num_leaves	learning_rate
	5	10	143	6	0.067009560
SVM	kernel	gamma	C		
	RBF	2.797443056	2.771604125		

5. 실험 결과

<Table 3>은 Optuna를 활용하여 튜닝을 진행하기 전, 후의 5개의 모델(Decision Tree, Random Forest, SVM, XGBoost, LGBM)의 성능 지표 비교를 나타낸다. 모델 성능 평가 시 가장 중요한 요소인 정확도를 비교한 결과, SVM의 경우 튜닝 전, 후 모두 학습 정확도가 검증 정확도 보다 낮아 학습 데이터의 패턴을 잡아내지 못하는 과소 적합이 발생하였다. XGBoost의 경우, 튜닝 전 모델은 학습 정확도가 검증 정확도 보다 낮아 과소 적합이 발생하였고, 튜닝 후 모델은 0.931로 가장 높은 학습 정확도를 보였으나 검증 정확도가 학습 정확도와 큰 차이를 내며 과적합이 발생하는 등 튜닝 전, 후 모두 불안정한 성능을 보였다. 이후 3개 모델(Decision Tree, Random Forest, LGBM)의 성능은 다음과 같다.

튜너를 사용하지 않은 3개 모델들의 학습 정확도와 검증 정확도 차는 평균 약 10%로 큰 차이가 나 실제로 과적합이 발생하였으나, 튜너를 사용한 모델들의 경우 약 1%로 과적합이 발생하지 않은 것으로 나타났다. Decision Tree의 경우, 검증 정확도, 정밀도, 재현율, F1-Score 모두 3개 모델 중 가장 낮은 성능을 보였으며 학습 정확도와 검증 정확도의 차이 또한 2%로 가장 큰 차이를 보였다. Random Forest의 경우, 학습 정확도와 검증 정확도의 차이는 크게 없으나 학습 정확도, 검증 정확도, 정밀도, 재현율, F1-Score 모두 0.9 미만으로 낮은 성능을 보였다. LGBM의 경우, 0.911의 학습 정확도와 0.908의 검증 정확도로 가장 차이가 적어 안정적인 성능을 보였으며 0.937의 가장 높은 재현율을 보였다. 또한, 정밀도와 재현율의 조화평균인 F1-Score도 0.901로 가장 높은 값을 보이기 때문에 타 모델에 비해 안정적이며 우수한 성능을 보인다고 할 수 있다. 따라서, LGBM을 최종 제안 모델로 선정하였다.

예측 모델의 결과 도출 과정을 직관적으로 이해할 수 있도록 설명 가능한 인공지능(eXplainable AI, XAI)기법인 SHAP (SHapley Additional exPlanations)기법을 활용하여 하위 매출에

Table 3. Performance Comparison Between Five Models (Decision Tree, Random Forest, XGBoost, LGBM, SVM)

Model	Tuning	max_depth	Overfitting	Train_acc	Test_acc	Precision	Recall	F1-Score
Decision Tree	X	20	O	0.988	0.865	0.873	0.817	0.844
	O	**5**	**X**	0.894	0.872	0.857	0.857	0.857
Random Forest	X	Until the end	O	0.988	0.872	0.875	0.833	0.854
	O	**4**	**X**	0.891	0.883	0.860	0.881	0.871
XGBoost	X	Until the end	-	0.904	0.908	0.868	0.937	0.901
	O	5	O	0.931	0.894	0.881	0.881	0.881
LGBM	X	Until the end	O	0.945	0.883	0.872	0.865	0.869
	O	**5**	**X**	0.911	**0.908**	0.868	**0.937**	**0.901**
SVM	X	-	-	0.864	0.901	0.883	0.897	0.890
	O	-	-	0.899	0.908	0.868	0.937	0.901

영향을 주는 요인을 시각화 하였다. <Figure 3>는 최종 학습 모델(LGBM)의 Feature Importance 계산 결과를 나타낸다. 그 결과, 미들 하(Middle Ⅱ), 생활인구수(DE_FACTO_POPULATION), 마감(Closing), 상업지역 면적(COMMERCIAL_AREA)순으로 매출 하위 매장 분류에 기여도가 높은 것을 확인하였다.

<Figure 4>는 최종 학습 모델(LGBM)의 SHAP Summary Plot 이다. 각 변수의 값은 붉을수록 높고, 파랄수록 낮은 것을 의미하며 SHAP value의 경우 양수이면 양의 영향을, 음수이면 음의 영향을 미치는 것을 의미한다.

이러한 특성을 바탕으로 각 변수의 영향력을 분석하면 영업시간대의 경우, 미들 하와 마감의 양의 영역에서 SHAP value 또한 양의 값으로 확인되는 것을 보아 미들 하와 마감일 때 매출 하위 매장에 분류될 확률이 크다는 것을 알 수 있다. 따라서 해당 시간대에는 하위 매출일 가능성이 높다는 의미로 해석할 수 있다. 반대로 오픈(Open)과 미들 상(Middle Ⅰ)일 때에는 음의 영향을 미쳐 매출 하위 매장으로 분류될 확률이 적으며 해당 시간대에는 하위 매출일 가능성이 낮다는 의미로 해석할 수 있다. 상업지역 면적의 경우, 매장 300m 근방 상업지역이 많을수록 매출 하위 매장에 분류될 확률이 크다는 것을 알 수 있다. 매장 근방에 상업지역이 많을수록 하위 매출일 가능성이 크다는 의미로 해석할 수 있다. 하지만 생활인구수의 경우, SHAP value가 넓게 분산되어 있어 명확한 특징 해석이 불가능하다고 볼 수 있다.

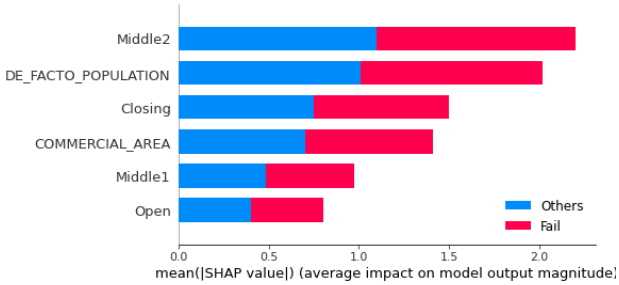

Figure 3. LGBM's Feature Importance Calculated by the SHAP Method

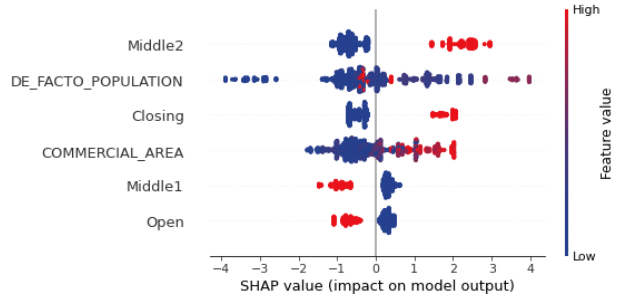

Figure 4. LGBM's Feature Importance with Interactive Impacts of Features

6. 결 론

본 연구는 머신러닝 기법을 활용하여 강남구 프랜차이즈 M커피 중 매출 하위 매장을 분류하고 원인을 분석하는 모델을 설계하였다. 시간대, 상업지역 면적, 생활인구수를 독립변수로 설정하여 Decision Tree, Random Forest, XGBoost, LGBM, SVM에 적용하였으며 각 모델의 최적화를 위해 Optuna를 활용하였다.

학습 정확도, 검증 정확도, 정밀도, 재현율, F1-score를 통해 성능 비교 및 검증하였다. 그 결과, LGBM이 학습 정확도 0.911, 검증 정확도 0.908, 정밀도 0.868, 재현율 0.937, F1-score 0.901로 가장 안정적이며 우수한 성능을 보였다.

SHAP기법을 활용하여 각 독립변수가 매출에 미치는 영향을 시각화 하였다. 이를 통해 기존 작동 과정을 파악하기 어려워 해석 불가능했던 머신러닝 모델과 달리, 모델의 작동 과정과 결과에 대한 해석력을 가지고 있어 신뢰성이 높고, 안정적인 성능 또한 갖춘 모델을 제안했다는 것에 큰 의의가 있다.

본 연구를 통해, 매출 하위 매장을 분류하고 매출에 영향을 미치는 중요 변수를 도출할 수 있었다. 같은 매장이라도 시간대별로 매출이 다르기 때문에 이를 바탕으로 매출이 적은 시간대의 근무 인력을 감축하는 등 근무 인력의 효율적인 배치

가 가능하다. 또한 매장 주변의 상업지역 면적, 매장의 영업시간대를 파악하여 매장의 운영시간을 조절하거나 매출이 낮은 시간대에 행사를 기획하여 매출 증가를 유도하는 등 실제 데이터에 기반한 매장 운영 전략을 수립할 수 있다.

매출 하위 매장의 안정화는 가맹점주 및 가맹본부 생존 등 사회 및 경제 전반에 긍정적인 효과를 가져올 수 있다. 또한, 매출 하위 매장의 실패는 경기침체에 영향을 미치고 이는 다시 소극적인 소비를 유발해 매출 하위 매장의 재생산을 야기할 수 있다. 따라서 매출 하위 매장에 집중한 요인 분석이 반드시 필요하다. 본 연구에서는 가맹점의 실제 매출 데이터를 기반으로 영향 요인을 분석했기 때문에 더 정확하고 신뢰성 있는 매장 운영 전략을 제시할 수 있을 것으로 기대한다.

그러나 본 연구에서는 강남구에 위치한 커피전문점 데이터만을 사용하였기 때문에, 강남구 이외의 다른 지역의 특성을 반영하지 못하였다는 한계가 있다. 향후 다른 지역의 데이터에도 반영하여 적용 범위를 확대할 예정이다. 특히 안정적인 성능을 갖춘 인공지능 모델을 제안하였기 때문에, 학습 데이터의 양이 늘어나면 더욱 정확한 설명력을 가질 수 있을 것이라 기대한다. 나아가 매장의 메뉴를 고려하여 매출 수준 예측을 진행한다면 커피전문점 이외의 업종에도 구체적이고 신뢰성 있는 분석 결과 제시와 함께 매장 운영 전략 수립이 가능할 것이라고 사료된다.

참고문헌

Ahn, Y. L., Ryu, S. M., Lee, H. H., and Park, M. S. (2022), Prediction of Food Franchise Success and Failure Based on Machine Learning, *The Journal of the Convergence on Culture Technology*, **8**(4), 347-353.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019), Optuna: A Next-generation Hyperparameter Optimization Framework, In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623-2631.

Bergstra, J., Bardenet, R., Bengio, Y., and Kegl, B. (2011), Algorithms for hyper-parameter optimization, *Advances in Neural Information Processing Systems*, 24.

Bramer, M. (2007), Avoiding Overfitting of Decision Trees, *Principles of Data Mining*, 119-134.

Breiman, L. (2001), Random forests, *Machine Learning*, **45**(1), 5-32.

Chen, T. and Guestrin, C. (2016), Xgboost: A Scalable Tree Boosting System, In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785-794.

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

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Lie, T. (2017), LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Advances in Neural Information Processing Systems*, 30.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009), *The Elements of sTatistical Learning: Data Mining, Inference, and Prediction*, Springeer, New York.

Hong, H. K., Kim, N. H., and Kim, C. M. (2019), Examining the Success Factors of Start-ups of Small Businessmen in Qualitative Perspective, *The Journal of the Korea Contents Association*, **19**(12), 229-237.

Jung, J. Y., Kim, H. M., and Cha, S. B. (2012), The Differences of Sensitivity to Americano Coffee Price Based on Consumers' Demographic Characteristics and Use Patterns Using Price Sensitivity Measurement, *Tourism and Leisure Research*, **24**(5), 241-261.

Kim, H., Lee, K. -S., Lee, Y. H., and Song, Y. N. (2021), Restaurants' Survival in the Era of COVID-19 - A Case Study of Seoul, *Journal of the Korean Geographical Society*, **56**(1), 35-51.

Kim, T. H., Joo, S. H., and Kim, E. H. (2014), A Study of the Differences by Operation Style, Relation Character, and Business Outcome in Terms of Franchise Motivation and Business Category, *Korean Journal of Food Marketing Economics*, **31**(3), 107-130.

Korea Fair Trade Commission (2022), Franchise Business Status Report.

KOSIS (Statistics Korea) (2020), Seoul City (nighttime)·Daytime population (12 years old and older) statistics.

Lee, J. M., Kim, D. J., and Lee, S. I. (2021), The Effect of Density by Type of Commercial Facilities on Closure of Restaurant : Targeting Major and Side-Street Trade Areas, Seoul, *Jounal of Korea Planning Association*, **56**(1), 108-120.

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

Natekin, A. and Knoll, A. (2013), Gradient Boosting Machines, a Tutorial, *Frontiers in Neurorobotics*, **7**, 21.

Noh, E. B. and Lee, S. K. (2018), A Comparative Study on the Effects of Location Factors on Sales by Restaurant Type, *Korea Real Estate Review*, **28**(4), 37-51.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016), "Why Should i Trust You?" Explaining the Predictions of any Classifier, In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

Shin, W. J. and Moon, S. Y. (2011), A Study on the Effects of Locational Characteristics on the Sales of a Coffee Shop Franchise, *Journal of the Korea Real Estate Analysts Association*, **17**(2), 111-123.

Shin, W. J. and Shin, W. H. (2009), Spatial Patterns of Retail Stores in Seoul, Korea, *Korea Real Estate Review*, **19**(2), 279-296.

Son, E. H., Shu, J. W., and Jeong, M. B. (2005), Forecasting of Hotel Food and Beverage Sales Using ARIMA Model: In the Case of Gyeongju Deluxe Hotels, *Journal of Tourism and Leisure Research*, **17**(3), 117-132.

Song, Y. S., Cho, Y. C., Seo, Y. S., and Ahn, S. R. (2009), Development and its Application of Computer Program for Slope Hazards Prediction using Decision Tree Model, *Journal of the Korean Society of Civil Engineers*, **29**,(2C), 59-69.

Statistics Korea (2020), Franchise Survey Result.

저자소개

안예린 : 서울여자대학교 데이터과학전공에 재학 중이다. 관심 연구 분야는 데이터 분석, 머신러닝이다.

이현희 : 서울여자대학교 데이터과학전공에 재학 중이다. 관심 연구 분야는 데이터 분석, 머신러닝이다.

유성민 : 서울여자대학교 데이터과학전공에 재학 중이다. 관심 연구 분야는 데이터 분석, 머신러닝이다.

김서연 : 서울여자대학교 데이터사이언스학과에 재학 중이다. 관심 연구 분야는 데이터 분석, 머신러닝이다.

박민서 : 2009년 메사추세츠대학교 컴퓨터사이언스(머신러닝) 전공으로 박사학위를 취득하였다. 삼성 SDS Bioinformatics Lab 및 성균관대학교 삼성융합의과학원 수석연구원, SK 텔레콤 팀리더, 한화시스템 상무(AI Lab 장)를 거쳐 현재 서울여자대학교 데이터사이언스학과 교수로 재직하고 있다. 관심 연구분야는 데이터 분석, 머신러닝이다.