업종별 데이터 변동성에 따른 최적 머신 러닝 모델 선택 전략 : 서울시 상권 예측 사례

이하나
고려대학교 SW·AI 융합대학원
hanalee11@korea.ac.kr

Optimal Machine Learning Model Selection Strategy Based on Industry-Specific Data Volatility: A Case Study of Seoul Commercial District Prediction

Lee Ha Na
Korea University Graduate School of SW·AI convergence

요 약

본 연구는 업종별 데이터 변동성에 따른 최적 머신러닝 모델 선택 전략을 제안한다. 서울시 상권분석 데이터를 음식점업, 소매업, 개인서비스업으로 세분화하고, AutoML 기법을 활용하여 각 업종별로 8 개의 머신러닝 알고리즘 중 최적 모델을 탐색하였다. 2023 년부터 2024 년까지의 실험 결과, 데이터 변동성에 따라 최적 알고리즘이 체계적으로 달라짐을 확인하였다. 저변동성 업종(개인서비스업, σ=0.879)에서는 ExtraTrees 가, 중변동성 업종(음식점업, σ=2.253)에서는 VotingRegressor 가, 고변동성 업종(소매업, σ=4.622)에서는 Ridge 가 각각 최적 성능을 보였다. 이를 통해 **변동성-복잡도 역상관 법칙과 유사업종 변화 지배 법칙**을 발견하였으며, 실무에서 활용 가능한 모델 선택 가이드라인을 제시하였다.

I. 서론

본 논문에서는 도시 상권의 변화 예측은 창업자의 사업 계획 수립, 임대업자의 투자 결정, 그리고 지방자치단체의 정책 수립에 중요한 정보를 제공한다. 특히 서울시와 같은 대도시에서는 상권의 급속한 변화와 업종별 상이한 특성으로 인해 정확한 예측이 더욱 중요해지고 있다.

기존 연구들은 주로 전체 상권을 통합적으로 분석하는 접근법을 사용해왔으나, 이는 업종별 고유한 특성과 영향 요인의 차이를 충분히 반영하지 못하는 한계가 있다. 예를 들어, 음식점업은 식사 시간대 유동인구와 밀접한 관련이 있는 반면, 개인서비스업은 거주 인구의 안정성에 더 영향을 받는다.

본 연구의 목적은 AutoML 기법을 활용하여 업종별 데이터 특성에 따른 최적 머신러닝 모델 선택 전략을 제시하는 것이다. 특히 데이터 변동성과 최적 알고리즘 간의 관계를 체계적으로 분석하여, 실무에서 활용 가능한 모델 선택 가이드라인을 개발하고자 한다.

II. 본론

1. 관련 연구

1.1 상권 분석 연구

강민경(2022)은 서울시 상권변화 유형도출 및 예측모델 연구에서 상권의 시계열적 변화 패턴을 분석하였으나, 업종별 특화 접근보다는 전체적인 변화 유형 분류에 중점을 두었다.[1] 김지연 등(2024)은 딥러닝 기반 서울시 행정동별 외식업종 상권 변화 예측 연구를 통해 외식업종에 특화된 딥러닝 모델을 제안하였으나, 단일 업종에 국한된 접근이었다.[2] 기존 상권 분석 연구들은 대부분 전체 업종을 통합하거나 단일 업종에만 집중하는 접근법을 취해왔다. 이러한 접근법들은 각각의 장점이 있으나, 업종 간 특성 차이를 고려한 선택적 모델링 전략은 제한적으로만 시도되었다.

1.2 머신러닝 기반 도시 예측

도시 데이터 분석 분야에서 Random Forest 와 XGBoost 는 높은 예측 성능으로 널리 활용되고 있다. 특히 공간 정보와 시계열 특성이 복합된 도시 데이터에서 앙상블 기법의 효과가 입증되었다.

1.3 업종별 특화 모델링

소매업 분야에서는 이미 고객 세분화 기반 개별 모델링의 효과가 검증되었으나, 상권 예측 분야에서는 업종별 최적 알고리즘 탐색이 제한적으로만 시도되었다. 특히 AutoML 을 활용한 체계적 알고리즘 비교를 통해 데이터 특성과 모델 성능의 관계를 분석한 연구는 부족한 실정이다. 본 연구는 이러한 연구 공백을 채우고자 한다.

2. 연구 방법론

2.1 데이터 수집 및 전처리

본 연구는 서울시에서 제공하는 상권분석 공공데이터를 활용하였다. 2023 년부터 2024 년까지 2 년간의 분기별 데이터를 수집하여 총 8 개 시점을 분석하였다.

업종 분류:

- 음식점업: 한식, 중식, 일식, 서양식 음식점 (4 개 하위 업종)
- 소매업: 편의점, 슈퍼마켓, 의복소매, 화장품소매 (4 개 하위 업종)
- 개인서비스업: 미용업, 세탁업, 수선업, 애완동물서비스 (4 개 하위 업종)

2.2 업종별 특화 변수 설계

각 업종의 특성을 반영한 맞춤형 변수를 다음과 같이 설계하였다:

공통 특화 변수:

- 프랜차이즈비율 = 프랜차이즈점포수 / 전체점포수
- 회전율 = (개업점포수 + 폐업점포수) / 전체점포수
- 업종집중도 = 해당업종점포수 / 행정동내전체점포수
- 시계열변화량 = 현재분기 - 이전분기 점포수 변화

유사업종 변화 변수의 정의와 중요성:

유사업종 변화는 본 연구의 핵심 발견 중 하나로, 해당 행정동 내에서 동일한 업종군에 속하는 모든 점포의 분기별 변화량을 의미한다. 예를 들어, 특정 동의 음식점업 예측 시, 해당 동 내 모든 음식점(한식, 중식, 일식, 서양식)의 총 점포수가 이전 분기 대비 얼마나 증감했는지를 나타낸다.

이 변수가 중요한 이유는 다음과 같다.

1. **상권 생태계 효과**: 동일 업종 점포들은 고객층과 입지 조건을 공유하므로, 한 지역의 업종 전체 트렌드가 개별 점포에 직접적 영향을 미침
2. **시장 포화도 신호**: 유사업종 증가는 시장 기회를, 감소는 시장 위축을 신호함
3. **정책/환경 변화 반영**: 지역 개발, 규제 변화 등이 업종 전체에 동시에 영향을 미치는 양상을 포착

2.3 AutoML 기반 모델 설계

2.3.1 포괄적 알고리즘 탐색 및 입출력 구성 각 업종별로 8 개의 머신러닝 알고리즘을 체계적으로 비교 평가하였다. 모든 알고리즘은 동일한 입출력 구성을 갖는다:

공통 입력 구성(X):

- 10 개 특성 변수: [점포수, 유사 업종 점포수, 프랜차이즈점포수, 프랜차이즈_비율, 개업비율, 폐업비율, 회전율, 업종집중도, 점포수 변화, 유사업종 변화]
- 데이터 형태: 연속성 수치 데이터 (정규화 미적용)
- 입력 차원: (n_samples, 10) - 각 업종별 샘플 수는 약 13,000 개

공통 출력 구성(Y):

- 목표 변수: 순개업점포수 (연속형 회귀 문제)
- 값 범위: 음식점업(-86~43), 소매업(-268~131), 개인서비스업(-11~13)

2.3.2 업종별 최적 모델 선택 각 업종의 데이터 특성에 맞는 최적 알고리즘을 자동으로 탐색하고, 3-fold 시계열 교차검증을 통해 성능을

검증하였다. 모든 알고리즘은 동일한 평가기준(RMSE)으로 비교되며, 최저 RMSE 를 달성한 모델을 해당 업종의 최적 모델로 선정하였다.

2.3.3 데이터 변동성 기반 모델 선택 전략 업종별 목표변수의 표준편차를 기준으로 다음 가설을 검증하였다:

- 저변동성($\sigma < 2.0$): 복잡한 비선형 모델이 효과적
- 중변동성($2.0 \leq \sigma < 4.0$): 앙상블 기법이 적합
- 고변동성($\sigma \geq 4.0$): 정규화된 단순 모델이 안정적

3. 실험 결과

3.1 포괄적 AutoML 기반 최적 모델 탐색 결과

8 개의 최신 머신러닝 알고리즘을 포함한 포괄적 AutoML 탐색을 실시하였다. 다중 평가지표를 통해 모델 성능을 종합적으로 평가하였다.

표 1. 업종별 포괄적 AutoML 결과 (다중 지표)

업종	최적 모델	RMSE	MAE	R^2	방향성 정확도
음식점업	VotingRegressor	1.147	0.117	0.845	0.829
소매업	Ridge	0.735	0.214	0.908	0.896
개인서비스업	ExtraTrees	0.200	0.020	0.962	0.987

특히 개인서비스업에서 가장 높은 예측 성능을 달성하였으며($R^2 = 0.962$), 방향성 예측 정확도도 98.7%로 실무적으로 매우 유용한 수준을 보였다. 반면 변동성이 높은 소매업에서도 Ridge 모델을 통해 89.6%의 방향성 정확도를 확보하였다.

3.2 데이터 변동성 구간의 객관적 설정

K-means 클러스터링 분석을 통해 업종별 데이터 변동성을 3 개 구간으로 객관적 분류하였다.

표 2. 클러스터링 기반 변동성 구간 설정

클러스터 ID	중심점(σ)	포함 업종	클러스터 특성
클러스터 2	0.879	개인서비스업	저변동성 그룹
클러스터 0	2.253	음식점업	중변동성 그룹
클러스터 1	4.622	소매업	고변동성 그룹

K-means 클러스터링 결과, 각 업종이 서로 다른 클러스터에 명확히 분류되어 **변동성 기반 구간 설정의 통계적 근거**를 확보하였다. 이는 경험적 구간 설정이 아닌 데이터 기반 객관적 분류임을 입증한다.

3.3 최적 모델별 변수 중요도 분석

최적 알고리즘으로 선정된 모델들의 feature importance 를 분석한 결과, **모든 업종에서 "유사업종_변화"가 최상위 중요도를 보이는 공통 패턴**을 발견하였다.

표 3. 업종별 최적 모델의 상위 5 개 중요변수

순위	음식점업 (VotingRegressor)	소매업 (Ridge)	개인서비스업 (ExtraTrees)
1	유사업종_변화 (0.751)	유사업종_변화 (0.916)	유사업종_변화 (0.469)
2	점포수_변화 (0.143)	폐업비율 (0.564)	점포수_변화 (0.323)
3	폐업비율 (0.031)	개업비율 (0.549)	개업비율 (0.085)
4	개업비율 (0.024)	업종집중도 (0.249)	폐업비율 (0.070)
5	프랜차이즈_점포_수 (0.023)	점포수_변화 (0.101)	회전율 (0.024)

핵심 발견인 "유사업종_변화 지배 법칙"의 실무적 의미: 모든 업종에서 유사업종_변화가 압도적 1 위를 차지하는 현상은 상권 예측에서 다음과 같은 중요한 시사점을 제공한다.

1. **생태계 관점의 예측**: 개별 점포의 성공/실패는 해당 점포의 특성보다 그 지역 업종 전체의 성장/쇠퇴 패턴에 더 크게 좌우된다. 즉, "숲을 보고 나무를 예측하는" 접근이 효과적이다.
2. **선행 지표 역할**: 유사업종의 변화는 시장 트렌드, 소비자 선호 변화, 정책 효과 등을 종합적으로 반영하는 선행 지표로 작용한다.
3. **업종별 민감도 차이**: 소매업(91.6%)이 가장 높은 중요도를 보이는 것은 소매업이 경기 변동, 온라인 쇼핑 트렌드 등 외부 환경 변화에 가장 민감하게 반응함을 의미한다.

이러한 발견은 창업자나 투자자들이 개별 점포의 내부 요인보다 **해당 지역 내 동종 업계의 전반적 동향**을 우선적으로 모니터링해야 함을 시사한다.

4. 토론 및 활용 방안

4.1 연구 결과의 의미

본 연구 결과는 다음과 같은 중요한 의미를 갖는다:

1. **업종별 최적 알고리즘의 다양성**: AutoML 탐색을 통해 업종별로 최적 알고리즘이 현저히 다름을 확인하였다. 특히 앙상블 기법(VotingRegressor), 극도 랜덤화 기법(ExtraTrees), 정규화 기법(Ridge)이 각각 다른 변동성 특성에 최적화되어 있음을 발견하였다.
2. **변동성-복잡도 역상관 법칙**: 데이터 변동성이 증가할수록 최적 모델의 복잡도가 감소하는 명확한 패턴을 발견하였다. 이는 고변동성 데이터에서 과적합 방지가 성능 향상의 핵심임을 시사한다.
3. **유사업종 변화의 지배적 영향**: 모든 업종에서 유사업종_변화가 최상위 중요도를 보여, 시계열 변화 패턴이 AI 모델 성능에 결정적 영향을 미침을 입증하였다.

4.2 업종별 모델 적용 전략

데이터 변동성에 따른 체계적 모델 선택 가이드라인을 다음과 같이 제시한다.

표 4. 실무용 모델 선택 가이드라인 (클러스터링 검증 기반)

변동성 분류	표준편차 범위	권장 모델	예상 성능	방향성 정확도	적용 사례
저변동성	$\sigma \approx 0.9$	ExtraTrees	R^2>0.95	>98%	안정적 서비스업
중변동성	$\sigma \approx 2.3$	Voting Regressor	R^2>0.84	>82%	일반 상업 시설
고변동성	$\sigma \approx 4.6$	Ridge	R^2>0.90	>89%	변동성 큰 소매업

본 가이드라인은 사전 데이터 분석(표준편차 계산)만으로도 최적 모델 유형을 예측할 수 있어, 실무에서 효율적인 AutoML 모델 선택이 가능하다.

4.3 AI 모델 선택 방법론의 확장 가능성

- **다른 시계열 데이터:** 본 연구의 변동성 기반 모델 선택 전략은 시계열 특성을 가진 다른 데이터에서도 활용 가능성이 있음
- **AutoML 도구 개선:** 데이터 변동성을 자동 분석하여 최적 알고리즘을 추천하는 지능형 AutoML 시스템 구축에 기여
- **적응형 모델링:** 데이터 특성 변화에 따라 동적으로 모델을 전환하는 적응형 시스템 개발의 이론적 기반 제공

4.4 한계점 및 향후 연구 방향

본 연구의 한계점은 서울시 상권 데이터라는 단일 도메인에 국한되어 일반화 가능성이 제한적이라는 점이다. 향후 연구에서는 다양한 도메인의 데이터를 활용하여 변동성-복잡도 역상관 법칙의 범용성을 검증하고, 더 많은 최신 알고리즘을 포함한 확장된 AutoML 프레임워크를 개발할 예정이다.

III. 결론

본 연구는 AutoML 기법을 활용하여 업종별 데이터 변동성에 따른 최적 머신러닝 모델 선택 전략을 제시하였다. 8 개의 머신러닝 알고리즘을 체계적으로 비교한 결과, 주요 발견은 다음과 같다:

핵심 기여:

1. **변동성-복잡도 역상관 법칙**: 저변동성 업종(ExtraTrees, R^2=0.962) → 중변동성 업종(VotingRegressor, R^2=0.845) → 고변동성 업종(Ridge, R^2=0.908)으로 데이터 변동성이 증가할수록 최적 모델의 복잡도가 체계적으로 감소함을 발견
2. **유사업종 변화 지배 법칙**: 모든 업종에서 유사업종_변화가 최상위 중요도(46.9~91.6%)를 보여, 시계열 변화 패턴이 AI 모델 성능에 결정적 영향을 미침을 입증
3. **포괄적 AutoML 검증**: 전통적 방법론부터 최신 앙상블 기법까지 8 개 알고리즘 비교를 통해 업종별 최적 모델이 현저히 다름을 체계적으로 검증

이는 기존의 "단일 알고리즘 적용" 접근법의 한계를 극복하고, **데이터 특성 기반 알고리즘 선택**의 중요성을 입증한다. 특히 K-means 클러스터링으로 검증된 변동성 구간을 통해 최적 모델 유형을 사전 예측할 수 있는 실무적 가이드라인을 제공하여, AutoML 적용의 효율성을 크게 향상시켰다.

향후 연구에서는 더 다양한 시계열 데이터로 확장하여 변동성 기반 모델 선택 이론의 일반화 가능성을 검증할 계획이다.

참 고 문 헌

[1] 강민경. (2022). "서울시 상권변화 유형도출 및 예측모델을 활용한 변화예측." 석사학위논문, 한양대학교.

[2] 김지연, 오수민, 박민서. (2024). "딥러닝 기반 서울시 행정동별 외식업종 상권 변화 예측 연구." 국제문화기술진흥원, 10(2), 459-463.

[3] 서울특별시. (2024). "서울시 상권분석 서비스 데이터." Retrieved from https://data.seoul.go.kr/