XGBoost 기반 온라인 거래 가격 급등락 탐지 모델

하주형, 신예준, 홍용근*, 노학균*

대전대학교

{juhyeongh68, tlsdPwns97}@gmail.com, *{yghong, hg.roh}@dju.kr

Price Fluctuation Detection via XGBoost

Ju Hyeong Ha, Ye Jun Shin, *Yong-Geun Hong, *Hakgyun Roh

Daejeon University

요 약

온라인 유통 시장에서는 동적 가격 결정(Dynamic Pricing)으로 인해 동일 상품이라도 구매 시점에 따라 가격이 달라지는 현상이 빈번하게 나타난다. 그러나 소비자는 가격 변동의 양상과 향후 변화를 체계적으로 파악하기 어려워 최적 구매 시점 판단에 한계가 있다. 본 연구는 특정 상품군의 과거 가격 데이터를 기반으로 XGBoost 모델을 학습하여, 가격 급등 및 급락과 같은 이상 변동(Anomaly) 패턴을 탐지하고 예측하는 방법을 제안한다. 실험 결과, 제안 모델은 약 88.77%의 정확도를 기록하였으며, 특히 가격 급등 패턴 탐지에서 0.93의 높은 재현율을 보였다.

Ⅰ. 서 론

온라인 유통 시장의 급격한 성장과 함께 도입된 동적 가격 결정(Dynamic Pricing) 알고리즘은 기업에게는 이익 극대화를 제공하지만, 소비자에게는 구매 결정의 복잡성을 가중시키는 요인이 되고 있다. 소비자는 단순히 현재 가격과 품질만을 비교하는 것이 아니라, 시간적 흐름, 이벤트 유무, 수급 불균형 등 다양한 변수에 의해 실시간으로 변동하는 가격 추이를 고려하여 비용 대비 가치를 극대화하고자 한다. 그러나 대다수의 유통 플랫폼은 현재 가격 정보만을 제공할 뿐, 향후 발생할 가격 변동의 방향성이나 시점에 대한 정보는 제공하지 않는다. 이러한 정보의 비대칭성으로 인해 소비자는 최적의 구매 시점을 판단하는 데 명확한 한계를 가진다. 따라서 과거의 대규모 가격 데이터를 분석하여 변동의 패턴을 사전에 학습하고, 특히 소비자의 지출 계획에 치명적인 영향을 줄 수 있는 가격 급등이나 급락과 같은 이상 징후(Anomaly)를 사전에 탐지할 수 있는 모델의 필요하다.

본 연구에서는 과거 가격 데이터로부터 변동 패턴을 학습하고, 급등·급락과 같은 주의 구간을 탐지하여 소비자 의사결정을 지원하는 모델을 구축하고자 한다. 이를 위해서 대규모 데이터에서도 효율적으로 학습 가능한 XGBoost를 활용해 가격 변동 탐지 모델을 구축하고, 데이터 기반의 정확한 예측 환경을 조성하고 소비자의 합리적 의사결정 지원 가능성을 확인하고자 한다.

본 논문의 기여는 다음과 같다. 첫째, 온라인 거래 데이터에서 가격 급등·급락(극단 변동) 구간을 선별하는 전처리 기준을 제시하였다. 둘째, 텍스트/범주형 특징을 결합한 XGBoost 기반 급등락 탐지 모델을 설계하였고, 모델 성능을 정량 지표와 혼동행렬로 확인하였다.

Ⅱ. 본론

1. 모델 선정

가격 변동 예측을 위해 본 연구에서는 트리 기반의 앙상블 학습 기법인 Random Forest와 Boosting 기법인 XGBoost의 성능 및 효율성을 비교 분석하였다. Random Forest는 배깅(Bagging) 방식을 사용하여 과적합을 방지하는 장점이 있으나, 대규모 데이터 학습 시 트리의 깊이가 깊어짐에 따라 메모리 사용량이 급증하고 훈련 속도가 저하되는 단점이 존재한다. 반면, XGBoost(eXtreme Gradient Boosting)는 직렬적인 부스팅 구조를 통해 이전 트리의 오류(Residual)를 순차적으로 보정해 나가는 방식을 취한다[1][2]. 또한, 과적합 방지를 위한 규제(Regularization) 항이 목적함수에 포함되어 있어 모델의 복잡도를 효과적으로 제어할 수 있다. 특히 본 연구에서 다루는 데이터는 고차원 희소 데이터(Sparse Data)의 특성을 가지는데, XGBoost는 병렬 처리에 최적화된 블록 구조와 결측치를 자동으로 처리하는 희소성 인식(Sparsity-aware) 알고리즘을 내장하고 있어 대규모 데이터 환경에서도 빠르고 안정적인 학습 성능을 보장한다. 이러한 Level-wise(깊이 중심 성장) 방식의 효율성을 근거로 본 연구에서는 학습 모델로 XGBoost를 선정하였다.

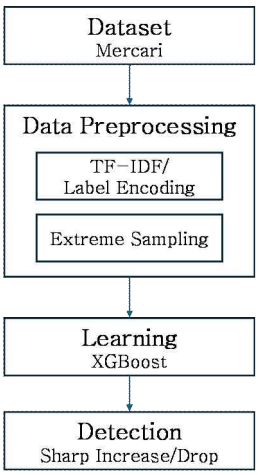

Fig. 1. 제안 모델의 전체 파이프라인

2. 데이터 셋 구성

실험 데이터는 Kaggle의 Mercari Price Suggestion Challenge 데이터를 활용하였다[3]. 중고 물품의 적정 가격을 예측하는 정적 데이터이고, 약 140만 건으로 구성되며, 해당 데이터 셋은 텍스트 및 범주형 변수를 포함하고 있다. 텍스트 및 범주형 변수는 가격 변동의 맥락을 제공하므로, 본

연구에서는 상품명(name), 상품 설명(item_description), 브랜드(brand_name), 카테고리(category_name) 정보를 결합하여 특징을 구성하였다. 텍스트 필드는 TF-IDF 기반 벡터화를 통해 고차원 희소 특징으로 변환하였고, 범주형 변수는 Label Encoding으로 수치화하였다. 최종 입력은 (1) 텍스트 기반 TF-IDF 벡터, (2) 카테고리/브랜드 기반 범주 특징을 결합한 형태이며, 이를 통해 동일/유사 상품명이라도 카테고리 차이에 따른 가격 양상이 반영되도록 설계하였다.

3. 전처리 구성

본 연구의 목표는 단순한 가격 예측이 아닌, 가격의 급격한 변동의 탐지이다. 이를 위해 기준이 되는 시점을 이전 시간 대비 상승 또는 하락으로 판정하여 가격 상승을 1, 하락을 0으로 두는 시계열화된 이진 분류 문제로 설정하였다. 또한 변동 폭이 작은 구간은 특성이 불명확해 성능에 불리할 수 있으므로, 극단 구간 데이터를 선별하여 주의가 필요한 급등락 사례 탐지에 초점을 맞추었다.

사전 데이터 분석 결과, 대다수의 데이터는 변동 폭이 미미하여 학습 효율을 저하시키는 것으로 나타났다. 이를 위해서 전체 데이터 중 가격 변동 폭이 상위 20%(변동 폭 34.0 이상)와 하위 20%(변동 폭 10.0 이하)에 해당하는 극단적인 구간만을 선별하였다. 가격 변동폭은 기준가격인 동일 상품에 대한 평균값(P_{ref}) 대비 편차로 정의하였다: $\Delta p = p - p_{ref}$. 레이블은 $\Delta p > 0$ 이면 가격 상승(1), $\Delta p \leq 0$ 이면 가격 하락(0)으로 정의하였다. 또한 $|\Delta p|$ 의 상하위 20% 구간별 선별하여 급등락 탐지에 초점을 맞추었다. 이 과정을 통해 중간 변동 구간의 불명확성을 제거하고, 총 40만 개의 데이터를 샘플링하여 학습 및 테스트에 활용하였다.

III. 실험결과

1. 실험환경 및 성능지표

원본 데이터 약 148만 건 중 상·하위 20%에 해당하는 극단 구간을 기준으로 필터링하였다. 하위 20% 경계는 10.0, 상위 20% 경계는 34.0으로 산정했으며, 필터링 후 약 68만 건 중 40만 건을 샘플링하였다. 레이블 분포는 가격 상승(1) 클래스가 약 37만 6천 개, 가격 하락(0) 클래스가 약 31만 개로 균형을 확보하였다. 이를 통해 전처리된 40만 개의 데이터셋을 8:2 비율로 학습과 테스트 데이터 세트를 분류하였다.

학습-평가 절차는 (1) 극단 구간 샘플링, (2) 학습/테스트 분할, (3) XGBoost 학습, (4) 정확도·정밀도·재현율·F1-score 평가 순으로 수행하였다. 또한 본 문제는 특정 방향(특히 급등) 탐지 누락이 소비자 의사결정에 큰 영향을 줄 수 있으므로, 단순 정확도뿐 아니라 재현율(Recall)과 F1-score를 함께 보고 모델의 탐지 민감도를 확인하였다.

2. 결과 분석

XGBoost 모델 학습 결과, 모델은 수렴(Convergence)에 도달하였으며 테스트 데이터에 대해 약 88.77%의 정확도를 기록하였다. 상세 성능 지표를 살펴보면, 가격 하락 클래스는 정밀도 0.91, 재현율 0.84, F1-score 0.87을 나타냈다. 반면, 가격 상승 클래스는 정밀도 0.87, 재현율 0.93, F1-score 0.90을 기록하였다. 이는 본 모델이 가격이 급등하는 상황을 탐지하는 민감도(Recall)가 우수함을 보여주었고, 소비자가 가격 상승 전에 구매 결정을 내리는 데 실질적인 도움을 줄 수 있음을 확인하였다.

또한, 혼동 행렬(Confusion Matrix) 분석 결과, 하락 클래스는 약 3만 2천 건이 정확히 분류되었고, 상승 클래스에서도 오분류(FN)가 약 5천 8백 건에 그쳐 전체적으로 낮은 오분류율을 보였다. 이는 극단적 가격 변동 구간을 선별한 샘플링 전략이 모델의 학습 명확성을 높였다.

Table 1에서 가격 상승(1) 클래스의 재현율이 0.93으로 높아, 가격 급등 상황을 놓치지 않는 탐지 특성을 확인하였다. 또한 Table 2의 혼동행렬에서 오분류가 제한적으로 나타나, 극단 구간 선별이 학습 명확성 향상에 기여했음을 확인하였다. 따라서, 필터링 된 데이터 샘플링 전략이 모델 성능 향상에 긍정적인 영향을 끼친 것을 확인하였고, 테스트 데이터 셋에서의 높은 정확도와 균형 잡힌 분류 지표는 과적합 없이 일반화가 잘 이루어졌음을 보여주며, 이는 모델이 가격 변동 방향 예측에 효과적임을 보여준다.

Table 1. Performance Matrix 결과

Class	Precision	Recall	F1-Score
가격 상승(1)	0.87	0.93	0.90
가격 하락(0)	0.91	0.84	0.89
Overall Accuracy	-	-	0.89

Table 2. Confusion Matrix 결과

		예측	
		가격 상승(1)	가격 하락(0)
실제	가격 상승 (1)	30,328(TP)	5,836(FN)
	가격 하락 (0)	3,145(FP)	40,781(TN)

IV. 결론

본 연구에서는 XGBoost를 활용해 온라인 거래 가격 데이터에서 변동 패턴을 학습하고, 급등·급락과 같은 극단 변동 구간을 탐지하는 방법을 제안하였다. 극단 구간을 선별하는 샘플링 과정과 실험을 통해, 제안 모델이 가격 변동의 방향성을 89%의 정확도로 예측함을 확인하였으며, 특히 가격 급등 패턴에 대한 높은 탐지 성능을 확보하였다. 다만 본 연구는 공개 데이터의 구조적 제약으로 인해 실제 시계열 기반 동적 가격 변동을 완전하게 반영하지 못할 수 있으며, 기준가격 정의와 극단 구간 임계값 설정에 따라 탐지 결과가 달라질 가능성이 있다. 향후 연구에서는 상품군 정의를 더 정교화(브랜드·카테고리·키워드 유사도 기반 군집화)하고, 임계값을 고정값이 아닌 데이터 분포 적응형 방식으로 설정하며, LightGBM 등 다양한 부스팅 계열과의 비교 등으로 체계적으로 개선할 예정이다.

ACKNOWLEDGMENT

이 논문은 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송표준개발지원사업(RS-2024-00397768)사업으로 지원받은 연구결과임

참 고 문 헌

[1]Yoo-jin Hwang, Seung-yeon Son, and Zoon-ky Lee, "Prediction of Stock Returns from News Article's Recommended Stocks Using XGBoost and LightGBM Models," Journal of the Korea Society of Computer and Information, vol. 29, no. 2, pp. 51-59, 2024.

[2] Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," Cornell University, 2016.

[3] Mercari Price Suggestion Challenge, https://www.kaggle.com/c/mercari-price-suggestion-challenge/