양파 도매 가격 예측을 위한 12가지 모델 성능 및 지역별 결과 비교 분석

박제인[1], 정수진[2]
[1]광운대학교 정보융합학부 학부생
[2]광운대학교 정보융합학부 학부생
jane08710@naver.com, marcusss@naver.com

A Comparative Analysis of Performance and Regional Results of 12 Models for Wholesale Onion Price Forecast

Jane Park[1], Sujin Jung[2]
[1]Dept. of Information Convergence, Kwangwoon University
[2]Dept. of Information Convergence, Kwangwoon University

요 약

한국의 주요 농산물인 양파 도매가격을 예측하기 위해 12가지 모델(SARIMA, ARIMA, Lasso Regression, Linear Regression, Ridge Regression, ElasticNet, LSTM, LightGBM, XGBoost, Random Forest, Gradient Boosting, Prophet)의 예측 성능을 비교 분석하며, 다섯 개 지역(광주, 대구, 대전, 부산, 서울)에서 모델의 성능을 평가한다. ARIMA와 SARIMA는 특히 대구와 부산에서 우수한 성과를 보였으며, Prophet과 LightGBM 모델은 상대적으로 낮은 정확도를 나타냄을 발견하였다. 다양한 모델의 성능 차이를 분석하고, 지역별 데이터 특성에 따른 맞춤형 예측 접근의 필요성을 강조한다.

1. Intro

농산물의 가격 변동은 농업 종사자, 소비자, 정책 입안자에게 중대한 영향을 미치는 요소[1]로, 특히 양파와 같은 주요 농산물[2]은 도매가격 예측이 매우 중요하다. 양파 가격은 기후 변화, 생산지 환경[3], 수요와 공급의 변동 등 단기적인 수급 변화에 민감하며, 변동성을 예측하는 데 어려움이 많다[4]. 그러나, 농업인들의 소득 안정화와 소비자 물가 안정을 도모하고, 정책 수립에도 유용한 지표로 활용[1]하기 위해서는 정확도 높은 가격 예측이 중요하다.

기존 연구들은 특정 모델에 대한 예측 성능을 단일 사례로 비교하는 경우[7, 8]가 많았으나, 본 연구는 SARIMA, ARIMA, Lasso Regression, Linear Regression, Ridge Regression, ElasticNet, LSTM, LightGBM, XGBoost, Random Forest, Gradient Boosting, Prophet 총 12개의 모델을 지역별로 비교하여 종합적으로 분석한다는 점에서 차별점이 있다. 이를 통해, 각 모델의 특성을 이해하여, 다양한 이해 관계자들[3]이 신뢰할 수 있는 정보를 얻을 수 있을 것이라 기대한다. 또한, 시계열 분석을 통한 양파 가격 예측을 진행하여, 효율적인 농업 경제 관리를 지원하는 데 이바지하고자 한다.

2. Literature review

농산물 가격 예측을 위해 선행 연구에서는 대표적인 시계열 모델로 ARIMA(Autoregressive Integrated Moving Average)[5, 6], SARIMA (Seasonal ARIMA), 그리고 지수평활법(Exponential Smoothing)이 주로 활용되었다. [7,8,9] 해당 모델들은 과거의 패턴을 분석하여 미래 가격을 예측하는 데 사용되며, 특히 SARIMA는 농산물의 계절적 특성을 효과적으로 고려할 수 있다는 점에서 자주 채택되었다. [10,11] BATS와 TBATS[12] 모델 또한 복잡한 계절성을 예측하는 선행 연구에서 활용되었지만 본 연구에서는 시계열 예측의 대표적인 모델인 ARIMA와 SARIMA를 중심으로 비교 분석을 진행하고자 한다. 다양한 예측 모델들이 특성에 따라 농산물 가격 예측의 정확도를 높이기 위해 활용되었으며, 본 연구에서는 이러한 선행 연구의 결과를 바탕으로 양파 도매가격 예측에 최적인 모델을 찾고자 한다.

3. Data

3.1 Data Collection

한국농수산식품유통공사(KAMIS) API에서 제공하는 '일별 품목별 도소매가격정보' 데이터를 활용하여 지역별 양파(1kg)의 일간 도매가격을 예측하고, 사용된 12개 예측 모델의 성능을 평가하여 비교하고자 한다. 해당 데이터는 2023년 9월 13일부터 2024년 6월 24일까지의 서울, 부산, 대구, 광주, 대전에서 각각 수집된 총 940개의 일간 가격 데이터(지역별 188개)를 포함한다. API 요청 시에는 부류 코드(100), 품목 코드(245), 품종 코드(00), 등급 코드(04), 도매(02) 등의 정보를 입력하였다. 그 결과 품목명, 품종명, 시군구, 마켓명(도매 시장명), 연도, 날짜, 가격을 문자열(string)로 얻을 수 있었다.

3.2 Data Preprocessing

수집한 기간의 데이터 중 일부 누락된 부분이 존재하여, 이전 시점의 가격 데이터를 반영한 'lag_1', 'lag_2', 'lag_3'과 같은 이동 평균 변수를 추가로 생성했다. 'price' 열에서는 쉼표(',')를 제거하고 결측값('-')을 NaN으로 변환한 뒤, float 형식으로 지정하였다. 'regday' 열에서는 슬래시('/')를 제거하고 문자열로 변환하여 날짜 형식을 통일하였다. 이후, 'yyyy'와 'regday' 열을 합쳐 새로운 'date' 열을 생성하고 날짜 형식으로 변환한 후, 불필요한 열들을 삭제하였다. 이와 같이, 결측값이 포함된 행은 삭제하여 데이터의 완전성을 확보하였으며, 데이터의 형식을 정제하여 일관성을 유지하였다. 정제된 데이터를 바탕으로 입력 변수(x)와 타겟 변수(y)를 정의하고, 8:2 비율로 훈련 데이터와 테스트 데이터를 분할하였다.

모델링 과정에서는 12가지 모델을 사용하여 학습을 진행하였으며, 각각의 성능은 평균 제곱 오차(MSE), 평균 절대 오차(MAE), 평균 절대 백분율 오차(MAPE), 평균 제곱근 오차(RMSE), 그리고 결정계수(R^2)를 사용하여 평가하였다.

4. Result

[표1]은 지역(광주, 대구, 대전, 부산, 서울)별로 12가지 예측 모델

에 대한 성능을 요약하여 비교하기 위해, 5가지의 평가 지표 중 결정계수(R^2)를 기준으로 나타내고 있다.

지역 / 모델	광주	대구	대전	부산	서울
ARIMA	0.61	0.98	0.76	0.96	0.79
SARIMA	0.59	0.98	0.76	0.96	0.78
Linear Regression	0.56	0.97	0.72	0.95	0.77
Ridge Regression	0.56	0.97	0.72	0.95	0.77
Lasso Regression	0.56	0.97	0.72	0.95	0.77
ElasticNet	0.56	0.97	0.72	0.95	0.77
XGBoost	-1.16	0.49	0.51	0.28	0.60
LSTM	0.24	0.89	0.57	0.82	0.56
Random Forest	-1.17	0.46	0.49	0.29	0.56
Gradient Boosting	-1.19	0.45	0.42	0.29	0.60
LightGBM	-1.52	0.45	0.56	0.26	0.30
Prophet	-8.72	-6.06	-4.56	-6.57	-8.58

[표1] 평가 지표 결정 계수(R^2)로 지역별 모델 성능 비교

특정 지역마다 모델의 성능을 비교해봤을 때, ARIMA와 SARIMA는 전반적으로 모든 지역에서 안정적인 예측 성능을 보였다. 특히 그중에서도 대구와 부산 지역에서 높은 결정 계수(0.98, 0.96)를 기록하였다. Linear Regression, Ridge Regression, Lasso Regression도 대부분 양호한 성능을 나타내었으며, 지역 간 성능을 비교해보면, 대구 지역에서 0.97로 가장 높은 예측 정확도를 보였다. 반면, ElasticNet과 XGBoost는 일부 지역(광주, 대전 등)에서 낮은 결정계수를 기록하였고, LSTM은 대구에서 비교적 높은 성능(0.89)을 보였으나 다른 지역에서는 성능이 저조하였다. 트리 기반 모델인 Random Forest, Gradient Boosting, LightGBM은 광주와 대전에서 상대적으로 낮은 성능을 보였으며, Prophet 모델은 모든 지역에서 전반적으로 저조한 성능을 나타내었다.

[표2]에 따르면, 본 연구의 결과를 통해 예측 성능이 지역별 데이터 특성에 크게 영향을 받는다는 것을 확인할 수 있다. 대구와 부산 지역에서는 ARIMA와 SARIMA 모델이 우수한 성능을 보였는데, 이는 해당 지역의 도매 가격 변동이 다른 지역에 비해 크지 않으며, 비교적 안정적인 패턴을 나타냈기 때문으로 해석된다. 반면, 서울, 광주, 대전에서는 가격 변동이 비교적 크고 불규칙하여 예측 모델의 성능이 상대적으로 낮게 나타났다. 트리 기반 모델들은 비선형적 데이터 특성에 강점을 보이지만, 시계열 데이터의 시간적 의존성을 반영하는 데 한계가 있었으며, 이는 대전과 광주에서의 낮은 예측 성능으로 나타났다. 또한, Prophet 모델은 긴 주기의 계절적 패턴을 반영하도록 설계되었으나, 본 연구에 사용된 농산물 가격 데이터가 비교적 짧은 기간의 변동을 포함하고 있어 모든 지역에서 낮은 성능을 기록하였다.

결론적으로, 각 지역의 데이터 특성과 변동성에 따라 최적의 예측 모델을 선택하는 것이 중요하며, 해당 시계열 데이터에 대해서는 ARIMA와 SARIMA 모델이 유효한 선택지임을 본 연구를 통해 확인할 수 있었다. 이러한 결과는 향후 농산물 가격 예측이나 시계열 데이터 분석 시 지역별 특성을 고려한 맞춤형 모델 선택의 필요성을 강조하며, 시사점을 제공한다.

본 논문은 과학기술정보통신부 대학디지털교육역량강화사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고 문헌

[1] 국승용, 서홍석, 서동주, 권상욱, 김경진, 최근 농산물 가격 변동 실태와 시사점, 서울: 한국농촌경제연구원, 2021.
[2] Nam, Kuk-Hyun, and Choe, Young-Chan. “A Study on Onion Wholesale Price Forecasting Model.” Journal of Agricultural Extension & Community Development, Vol. 22, No. 4, pp. 423-434, 2015.
[3] 최현오, 여현, 이명훈, 박장우, “생산지 환경에 따른 도매시장 농산물 가격 예측 연구,” 한국지식정보기술학회 논문지, 제17권 제6호, 1,285-1,295, 2022.
[4] 신성호, 이미경, 송사광, “LSTM 네트워크를 활용한 농산물 가격 예측 모델,” 한국콘텐츠학회논문지, 제18권, 제11호, pp. 416-429, 2018.
[5] J. H. Ha, S. T. Seo, and S. W. Kim, “Evaluation on the Performance of Onion and Garlic Forecasts,” Proceedings of the Summer Academic Conference of the Korea Food Trade Association, pp. 559-572, Jul. 2019
[6] B. Choi and I. C. Choi, “Monthly Price Forecasting of Fruit-type Vegetables Using Time-Series Analyses,” Journal of Rural Development, Vol. 30, No. 1, pp. 129-148, Apr. 2007.1
[7] Ranjit Kumar Paul, Md. Yeasin, Pramod Kumar, Prabhakar Kumar, M. Balasubramanian, H. S. Roy, A. K. Paul, Ajit Gupta. Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. PLoS One, 17(7): e0270553, 2022.
[8] Paul RK, Yeasin M, Kumar P, Kumar P, Balasubramanian M, Roy HS, Paul AK, Gupta A. Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. PLoS One. 17(7): e0270553, 2022.
[9] 정민제. 농산물 도매시장 주별 가격예측 모델 및 출하 의사결정지원도구 개발 연구: Research of forecasting weekly price on agricultural wholesales market and developing decision making tool. 서울, 서울대학교 대학원, 2022.
[10] Mithiya D, Mandal K, Datta L. Forecasting of potato prices of Hooghly in West Bengal: time series analysis using SARIMA model. International Journal of Agricultural Economics. 4(3): 101-108, 2019.
[11] Luo CS, Zhou LY, Wei QF. Application of SARIMA model in cucumber price forecast. Applied Mechanics and Materials. 373: 1686-1690, 2013.
[12] De Livera AM, Hyndman RJ, Snyder RD. Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American Statistical Association. 106(496): 1513-1527, 2011.

지역	광주					대구					대전					부산					서울				
모델 / 평가지표	MSE	RMSE	MAE	MAPE	R2	MSE	RMSE	MAE	MAPE	R2	MSE	RMSE	MAE	MAPE	R2	MSE	RMSE	MAE	MAPE	R2	MSE	RMSE	MAE	MAPE	R2
ARIMA	7064.71	84.05	54.38	4.99%	0.61	568.57	23.84	14.74	1.14%	0.98	5753	75.85	46.93	3.39%	0.76	995.7	31.55	18.24	1.44%	0.96	2926.13	54.09	36.28	2.67%	0.79
SARIMA	7549.98	86.89	54.82	5.06%	0.59	577.31	24.03	16.68	1.29%	0.98	5646.9	75.15	52.69	3.74%	0.76	936.87	30.61	19.65	1.53%	0.96	3079.49	55.49	40.61	2.97%	0.78
Linear Regression	8036.54	89.65	61.57	5.71%	0.56	671.66	25.92	14.08	1.10%	0.97	6643.7	81.51	49.62	3.56%	0.72	1096.96	33.12	18.45	1.45%	0.95	3288.88	57.35	37.96	2.78%	0.77
Ridge Regression	8036.53	89.65	61.57	5.71%	0.56	671.67	25.92	14.08	1.10%	0.97	6643.6	81.51	49.62	3.56%	0.72	1096.97	33.12	18.45	1.45%	0.95	3288.89	57.35	37.96	2.78%	0.77
Lasso Regression	8036.62	89.65	61.57	5.71%	0.56	671.99	25.92	14.08	1.10%	0.97	6643.4	81.51	49.62	3.56%	0.72	1096.77	33.12	18.44	1.45%	0.95	3288.84	57.35	37.96	2.78%	0.77
ElasticNet	8036.53	89.65	61.57	5.71%	0.56	671.79	25.92	14.08	1.10%	0.97	6643.5	81.51	49.62	3.56%	0.72	1096.89	33.12	18.44	1.45%	0.95	3288.89	57.35	37.96	2.78%	0.77
XGBoost	39336.16	198.33	162.9	15.82%	-1.16	12570	112.11	89.24	7.43%	0.49	11465	107.07	76.49	5.75%	0.51	15863.7	125.95	101.24	8.70%	0.28	5602.89	74.85	58.92	4.39%	0.6
LSTM	13861.15	117.73	97.87	9.23%	0.24	2693.1	51.9	43.22	3.50%	0.89	10207	101.03	71.51	5.27%	0.57	4009.26	63.32	49.59	4.09%	0.82	6904.86	83.1	65.47	4.86%	0.56
Random Forest	39527.5	198.82	164.64	15.96%	-1.17	13507	116.22	96.65	7.95%	0.46	12007	109.57	78.46	5.83%	0.49	15655.71	125.12	100.58	8.65%	0.29	6208.29	78.79	61.36	4.60%	0.56
Gradient Boosting	39930.2	199.83	165.8	16.06%	-1.19	13661	116.88	97.39	8.01%	0.45	13751	117.26	79.96	5.96%	0.42	15779.15	125.62	100.54	8.65%	0.29	5611.24	74.91	60.41	4.55%	0.6
LightGBM	45940	214.34	177.6	17.24%	-1.52	13808	117.51	98.36	8.10%	0.45	10291	101.45	71.79	5.39%	0.56	16304.73	127.69	103.06	8.83%	0.26	9904.68	99.52	70.72	5.24%	0.3
Prophet	176969.8	420.68	414.98	36.82%	-8.72	175729	419.2	413.15	31.56%	-606	130934	361.85	349.24	24.38%	-4.56	167844	409.69	404.52	31.79%	-657	134768.7	367.11	360.94	26.58%	-8.58

[표2] 5가지 평가 지표 별 12가지 모델의 모든 지역 성능 비교