한국빅데이터학회지
제6권 제2호, 2021, pp. 169-179 https://doi.org/10.36498/kbigdt.2021.6.2.169

정형 및 비정형 데이터를 이용한 농산물 구매량 예측: 파프리카를 중심으로*

Prediction of Agricultural Purchases Using Structured and Unstructured Data: Focusing on Paprika

Somakhamixay Oui[1] · 이경희[2] · 라형철[3] · 최은선[4] · 조완섭[1†]

충북대학교 경영정보학과[1], (주)빅데이터랩스[2], 충북대학교 수의학연구소[3], 충북대학교 빅데이터 협동과정[4]

요 약

소비자의 식품소비행동은 소비자 패널 데이터와 같은 정형 데이터 뿐 아니라 매스미디어와 소셜미디어(SNS) 등 비정형 데이터로부터 영향을 받을 가능성이 높아지고 있다. 본 연구에서는 식품소비 관련된 정형 데이터와 비정형 데이터를 연계한 융합데이터 셋에 대하여 딥러닝 기반의 소비예측 모델을 생성하고 이를 검증한다. 연구의 결과는 정형 데이터와 비정형 데이터를 결합할 때 모델 정확도가 향상되었음을 보여주었다. 또한 비정형 데이터가 모델 예측 가능성을 향상시키는 것으로 나타났다. 변수들의 중요도를 식별하기 위해 SHAP 기법을 사용한 결과 블로그 및 비디오 데이터 관련 변수가 상위 목록에 있었고, 파프리카 구매 금액과 양의 상관관계가 있음을 알 수 있었다. 또한 실험 결과에 따르면 머신러닝 모델이 딥러닝 모델보다 높은 정확도를 보였고, 기존의 시계열 분석 모델링에 대한 효율적인 대안이 될 수 있음을 확인하였다.

■ 중심어 : 정형 데이터, 비정형 데이터, LSTM, CNN, SVR, Random Forest, XGBoost, SHAP

Abstract

Consumers' food consumption behavior is likely to be affected not only by structured data such as consumer panel data but also by unstructured data such as mass media and social media. In this study, a deep learning-based consumption prediction model is generated and verified for the fusion data set linking structured data and unstructured data related to food consumption. The results of the study showed that model accuracy was improved when combining structured data and unstructured data. In addition, unstructured data were found to improve model predictability. As a result of using the SHAP technique to identify the importance of variables, it was found that variables related to blog and video data were on the top list and had a positive correlation with the amount of paprika purchased. In addition, according to the experimental results, it was confirmed that the machine learning model showed higher accuracy than the deep learning model and could be an efficient alternative to the existing time series analysis modeling.

■ Keyword : structured data, Unstructured data, LSTM, CNN, SVR, Random Forest, XGBoost, SHAP

2021년 11월 29일 접수; 2021년 12월 09일 수정본 접수; 2021년 12월 10일 게재 확정.
* 본 논문은 정부(식품의약품안전처)의 출연연구사업 지원을 받아 수행된 연구임 (과제고유번호: KMDF-RnD 21163수입안 517-1)
† 교신저자 (wscho@cbnu.ac.kr)

Ⅰ. 서 론

식품에 대한 소비예측은 수급조절을 통해 가격 폭등의 혼란을 방지할 수 있으며, 생산자와 소비자의 만족도 제고와 국가정책 수립에 중요한 자료가 될 수 있다. 식품에 대한 소비예측은 주로 생산량이나 판매/유통량 등 정형 데이터를 중심으로 시계열 분석 등 기존 통계분석 기법들을 사용하여 이루어져 왔으나 최근들어 빅데이터와 인공지능을 이용한 기법들이 제안되고 있다[1-5]. 소비자의 식품소비 행동은 생산량이나 소비자 패널 데이터와 같은 정형 데이터 뿐 아니라 매스미디어와 소셜미디어(SNS) 등 비정형 데이터로부터 더 큰 영향을 받고 있기 때문이다.

본 연구에서는 정형데이터와 비정형데이터를 융합한 데이터셋에 대하여 머신러닝과 딥러닝을 활용한 수요예측 모형을 개발하고 평가한다. 다양한 식품 종류 중 파프리카를 선택하여 연구했으나 기본적인 연구방법은 다른 종류의 식품에 대해서도 동일하게 적용될 수 있다. 사용된 정형 데이터로 농촌진흥청(RDA) 농산물소비자패널 자료, 도매시장 자료, 소매가격, 통계정보시스템(KOSIS)에서 제공하는 파프리카 생산자 관련 데이터가 포함된다. 연구에서 사용한 비정형 데이터는 방송 뉴스, TV 프로그램/쇼, 블로그 및 비디오 데이터와 소셜 미디어(SNS) 데이터로 구성된다. 방송 데이터의 경우 음성을 텍스트로 변환한 후 자연어 처리 기법으로 처리하여 사용하였다.

이 연구의 첫 번째 목적은 예측모델을 개발하기 위해 매스 미디어와 소셜 미디어(SNS)를 비정형 데이터로 사용할 가능성을 체크하는 것이다. 둘째, 파프리카 소비에 높은 영향을 미치는 가장 관련성 있는 요인을 파악하는 것이다. 셋째, 다양한 기법들로 생성된 예측모델 중에서 가장 적합한 모델을 평가하고 선택하는 것이다. 궁극적으로 본 연구는 파프리카 소비를 촉진하고 긍정적인 농업정책의 방향을 제시하는 것을 목적으로 한다.

연구의 결과를 요약하면 다음과 같다. 정형 데이터와 비정형 데이터를 결합할 때 모델 정확도가 향상되었음을 보여주었다. 또한 비정형 데이터가 모델 예측 가능성을 향상시키는 것으로 나타났다. SHAP 기법을 사용하여 변수들의 중요도를 확인한 결과 블로그 및 비디오 데이터 관련 변수가 상위 목록에 있었고, 이들이 파프리카 구매 금액과 양의 상관관계가 있음을 알 수 있었다. 또한 실험 결과에 따르면 머신러닝 모델이 딥러닝 모델보다 높은 정확도를 보였고, 기존의 시계열 분석 모델링에 대한 효율적인 대안이 될 수 있음을 확인하였다.

논문의 구성은 다음과 같다. 제2절에서는 연구의 방법론과 선행 연구들에 관해 소개한다. 제3절에서는 연구 데이터 수집, 변수 탐색 및 실험 과정에 대해 소개한다. 제4절에서는 연구 결과를 비교하고, 변수의 중요성을 평가한다. 제 5절에서는 연구의 결론과 한계를 기술한다.

Ⅱ. 관련연구

본 연구에서는 블로그, 식품 소비 관련 TV 프로그램 등 다양한 비정형 빅데이터를 머신러닝 및 딥러닝 모델에 적용하여 파프리카 가격 예측 모형을 구축하고 파프리카 가격 예측에 영향을 미치는 주요 변수를 확인하고자 한다. 기존의 농산물 가격 예측 연구는 전통적인 방법으로 주로 분석되어왔다. 하지만 비정형 데이터를 포함한 분석과 데이터의 양이 증가함에 따라 머신러닝과 딥러닝을 사용한 분석의 중요성이 증가되고 있다. 머신러닝과 딥러닝 기법을 적용한 분류 및 예측 모델은 복잡성을 바탕으로 뛰어난 분류 및 예측이 가능하도록 하지만 예측결과에 대한 각 입력변수들의 기여도를 알 수 없기에 신뢰성의 문제를 가지고 있다. 따라서 우리는 설명 가능한 인공지능(eXplainable Artificial Intelligence, XAI)을

구축한 모델에 적용하고, 시각화를 통해 결과에 대한 해석력을 높이고자 한다. 또한 변수 중요도를 도출하고 각 입력변수들이 결과에 얼마나 영향을 미치는지를 확인한다.

2.1 비정형데이터

인터넷 확산으로 SNS 데이터나 인터넷 검색 데이터가 실제 경제활동보다 선행하여 일어나고, 이들 데이터를 활용하여 경제활동을 예측하는 연구가 활발하게 일어나고 있다. 농식품의 생산과 소비분야에서도 SNS데이터나 인터넷 검색 데이터를 활용한 다양한 사례가 있으나[1-4], 소비촉진과 관련된 사례는 흔하지 않다[11]. Meza 등(2016)은 Twitter 데이터를 활용하여 유기농 농산물에 대하여 멕시코와 한국 이용자의 입소문(Word-of-Mouth)과 관련된 연관어를 분석하였다[5]. 유도일(2016)은 건고추, 마늘, 양파 등의 농산물 가격 예측에 인터넷 검색 데이터와 SNS 데이터에 대해 국내 빅데이터 분석 상용플랫폼 텍스톰(Textom)을 사용하여 추출한 키워드 빈도를 활용하였다[6]. 이승용(2014)는 구제역에 관한 언론보도매체가 소비자 소비에 미친영향을 분석하였다[7]. 최경덕 등(2016)은 소셜메트릭스(SOCIALmetricsTM)를 통해 수집한 SNS 데이터를 농진흥청의 농식품 소비자 패널 조사 데이터의 소비 패턴 정보와 연계하여 일본 후쿠시마 원전 사고가 소비패턴에 어떻게 영향을 미치는지를 연구하였다[8]. 채광석 등(2017)은 소셜메트릭스 (SOCIALmetricsTM)를 이용하여 SNS 데이터에서 다양한 밭작물에 대한 연관어 분석과 감성 분석 수행하여 소비 패턴을 파악하고, 이에 따른 농산물 공급 대응 전략을 제시하였다[9]. 김재우 등(2017)은 인도네시아 지역에서 생성된 Twitter 데이터를 수집분석하여, 인도네시아의 주요 소비재에 대한 실시간 시장물가를 단기 예측하는 연구를 발표하였다[10]. 최근에는 조용빈 등[11]은 ARX (Autoregressive exogenous) 및 VECM (Vector error correlation)을 사용하여 파프리카 소비 예측 모델을 생성하였다. 이 연구는 ACF (Autocorrelation function)를 사용하여 파프리카 소비와 비정형 데이터 간의 관계를 분석하였다. 연구는 파프리카 소비와 텔레비전 프로그램/쇼, 블로그 사이에 상관관계가 있음을 확인하였다.

2.2 머신러닝 및 딥러닝

머신러닝과 딥러닝 기법을 적용한 모델은 전통적인 알고리즘 보다 뛰어난 성능을 보이기에 가격예측에 머신러닝 및 딥러닝을 적용하여 예측의 정확도를 향상시키고자 하는 다양한 연구들이 진행되고 있다. 농산물 가격 예측을 위해 LSTM, 모델을 사용하여 가격 예측 분석 모델을 구축하고, 모델 별 성능을 비교한 연구 결과가 있다[15-17]. 김미혜 등(2018)은 복숭아 가격 및 거래량 예측을 위해 랜덤포레스트와 그래디언트부스팅(gradient boosting machine), XGboost을 사용하여 예측모델을 구축하고 복숭아 거래량 예측에 영향을 미치는 주요 변수를 도출하였다[18]. 주정민 등(2020)은 아파트 가격 예측 모형 생성을 위해 Xgboost, Lightgbm, Catboost 등의 머신러닝 알고리즘을 사용하였고, RMSE를 사용하여 각 예측 모형 간의 성능 비교를 수행하였다[19].

본 연구에서는 구축한 모델의 블랙박스 문제를 해결하기 위해 설명가능한 인공지능 중 하나인 SHAP를 적용하여, 모델의 신뢰성을 보완하고자 한다. 노윤아 등(2021)은 Attention LSTM(Long Short-Term Memory) 모델을 사용하여 COVID-19 확진자 수를 예측하고, 그 결과를 SHAP (SHapley Additive exPlanations)을 통하여 예측 결과에 대한 설명 가능성을 입증했다[20]. 임도현 등(2021)은 LightGBM을 이용하여 학업 중단의 예측 정확도를 도출하고 해석가능한 인공지능 기법 중 하나인 SHAP를 통해 학업 중단에 영향을

미치는 요인에 대한 해석했다[21].

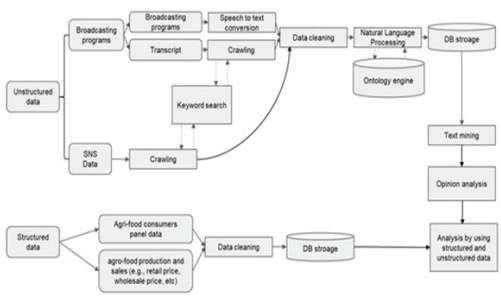
Ⅲ. 연구방법

본 장에서는 연구방법에 대한 내용으로 예측모형 개발에 사용되는 데이터셋을 설명한 후, 다양한 예측모형의 개발 과정을 설명한다.

3.1 데이터셋 생성

본 연구에서 사용하는 데이터는 정형 데이터와 비정형 데이터로 구성된다. 정형데이터로는 농촌진흥청(RDA) 농산물소비자패널 자료, 농촌경제연구원 전망 및 농업통계정보(OASIS) 도매시장 자료, 한국농업마케팅정보원(KAMIS) 소매가격, 파프리카 생산관련 데이터는 통계정보시스템(KOSIS)의 데이터를 사용하였다. 이와 함께 사용되는 비정형 데이터로는 파프리카를 언급하는 방송 뉴스, TV 프로그램/쇼, 블로그, 동영상 등이 있다. 비정형 데이터는 비디오와 오디도 등의 정보를 포함하므로 이를 텍스트로 변환한 후, 자연어 처리기법으로 언급하는 빈도를 계산하고, 감정단어를 구분해서 긍부정 데이터로 활용한다. 각 데이터에 대한 상세한 항목과 설명은 참고문헌[12]에 있다.

그림 1은 정형 데이터와 비정형 데이터의 수집과 전처리를 통한 데이터셋 생성 방법을 도식화



〈그림 1〉 데이터셋 수집과 전처리 과정

한 것이다. 그림에서 보는 바와 같이 방송 뉴스, 텔레비전 프로그램/쇼, 비디오 데이터를 텍스트로 변환하거나 방송 스크립트를 수집하여 활용한다.

기본적으로 언급된 식품의 종류를 인식하기 위해 온톨로지를 사용하며, 감정을 포함한 단어를 인식하여 긍정적인 단어와 부정적인 단어로 구분하고, 빈도수를 계산한다. 정형데이터와 비정형 데이터는 별도의 데이터셋으로 활용할 수도 있고, 식품유형이나 시간으로 양자를 결합하여 융합데이터셋으로 활용할 수도 있다.

본 연구는 Socialmetrics(http://www.some.co.kr/) 와 Google Trend (https://trends.google. com/) 등 여러 소스로부터 파프리카 소비 관련 키워드를 검색하여 비정형 데이터를 수집하였다. 파프리카는 다이어트 하는 사람들 사이에서도 사용된다. 관련 키워드 검색 결과를 분석한 결과 “파프리카”, “파프리카와 효능”, “파프리카와 음식”, “파프리카와 다이어트(체중 감량)”와 같은 파프리카 관련 키워드를 연구 대상으로 선정하였다. 이러한 키워드는 매스미디어 및 소셜미디어(SNS)에서 데이터를 크롤링하는 데 사용되는 파프리카 소비 관련 키워드로 사용된다. 정형데이터와 비정형 데이터의 수집 기간은 2010년부터 2017년까지 8년간이다.

3.2 예측모형 개발

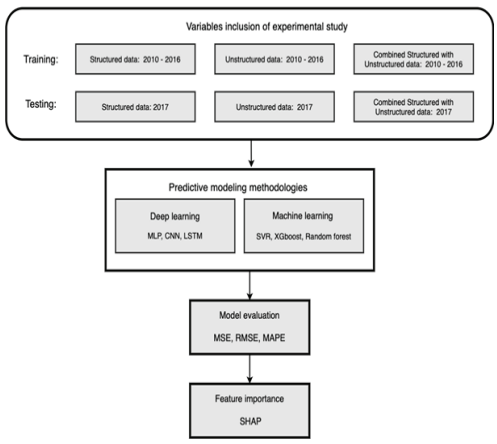
본 연구의 모델링 과정은 <그림2>와 같이 진행되었으며, 크게 4단계로 설계하였다. 첫 번째 단계에서는 분석 목적에 맞게 데이터셋을 생성하는 단계이다. 훈련 데이터셋은 2010년부터 2016년까지 정형데이터/비정형데이터/융합데이터셋 세 가지로 구성하고, 테스트 데이터셋은 이들 각각에 대하여 2017년 데이터로 구성한다. 두번째 단계에서, 1단계에서 생성한 3가지 데이터셋에 대하여 인공지능 알고리즘으로 딥러닝 기법(MLP, CNN, LSTM)과 기계학습 기법(SVR,

〈표 1〉 추출된 정형 및 비정형 데이터 현황

데이터 종류	데이터 명	변수 명	데이터 수
정형 데이터	농식품 소비자 패널 조사 데이터	주 단위 파프리카 구매금액	423
	도매시장 데이터	파프리카 도매가격	423
		파프리카 도매 반입량	423
		파프리카 소매가격	423
	파프리카 생산량 데이터	단위 면적당 연간 수율(ha)	423
		연간생산량(톤)	423
		작년생산량(톤)	423
비정형 데이터	방송 뉴스	파프리카 언급 기사 수	379
		이모티콘과 댓글 작성 빈도	105
		파프리카 효과에 대한 키워드가 언급된 주간뉴스 댓글	97
		파프리카 및 효과 언급 기사 수	36
		파프리카 및 건강 언급 기사 수	10
		파프리카 및 요리 언급 기사 수	129
		파프리카 및 요리 언급 기사 댓글 수	25
		파프리카 요리 긍정적 언급	119
	뉴스 외 방송 프로그램	파프리카 언급된 주간 시청률 순위	230
		파프리카에 대한 긍정 키워드 언급	230
		파프리카에 대한 부정 키워드 언급	197
		파프리카 효과에 대한 긍정적 키워드가 언급	52
		파프리카 요리 긍정적 언급	213
		파프리카 요리 부정적 언급	175
		파프리카 및 건강에 대한 긍정적 언급	93
		파프리카 및 건강에 대한 부정적 언급	82
	블로그	파프리카 언급 블로그 수	423
		파프리카 언급 블로그 댓글 수	361
		파프리카 언급 블로그 댓글의 긍정이모티콘 수	319
		긍정적인 단어 언급 블로그 수	421
		부정적인 단어 언급 블로그 수	395
		네이버 블로그 언급 빈도	385
		네이버 블로그 댓글 빈도	320
		네이버 블로그 댓글의 긍정이모티콘 수	312
		긍정적인 단어 언급 네이버 블로그 수	384
		부정적인 단어 언급 네이버 블로그 수	354
		파프리카 효과 언급 네이버 블로그 수	253
		파프리카 효과 언급 네이버 블로그 댓글	111
		파프리카 효과에 대한 네이버블로그 좋아요 수	216
		파프리카 효과에 대한 긍정적 언급 네이버블로그	253

데이터 종류	데이터 명	변수 명	데이터 수
비정형 데이터	블로그	파프리카 효과에 대한 부정적 언급 네이버블로그	205
		네이버블로그 파프리카 요리 언급	343
		네이버블로그 파프리카 요리 댓글수	232
		네이버블로그 파프리카요리 좋아요 수	278
		네이버블로그 파프리카 요리 긍정단어	343
		네이버블로그 파프리카 요리 부정단어	298
		네이버블로그 파프리카 건강언급	298
		네이버블로그 파프리카 건강 댓글수	150
		네이버블로그 파프리카건강 좋아요 수	250
		네이버블로그 파프리카 건강 긍정언급	298
		네이버블로그 파프리카 건강 부정언급	248

XGBoost, Random Forest)을 사용하여 총 18개의 예측모델을 생성한다(데이터셋 3가지 * 6가지 AI알고리즘 = 18가지 예측모형). 세 번째 단계는 모델 평가단계이다. 테스트 데이터셋에 대한 예측 결과는 MSE(Mean Square Error), RMSE(Root Mean Square Error) 및 MAPE(Mean Absolutes Percentage Evaluation)를 사용하여 평가한다. 네 번째 단계는 변수들의 기능 중요도를 식별하는 단계이다. 단일 변수가 목표 변수인 소비량에 미치는 영향을 탐색하기 위해 SHAP(Shapley) 기법을 사용하여 산출한 결과를 통해 모델을 해석함으로써 모델의 신뢰성을 높이고자 한다.



〈그림 2〉 예측모형의 개발

본 연구에서는 Python 프로그래밍 언어를 사용하여 Jupyter Notebooks에서 모델을 구축하였다. 오픈소스 인공지능 라이브러리인 케라스(Keras)와 사이킷런(Scikit-learn)을 사용하였다.

Ⅳ. 분석결과

4.1 예측모형 평가

본 연구에서 그림 2에서 보는 바와 같이 6개의 예측모형을 생성하여 평가한다. 각 실험은 모델 정확도를 비교하기 위해 세 가지 다른 데이터 세트와 하나의 알고리즘을 사용한 것이다. 실험 결과는 정형데이터, 비정형데이터, 정형 데이터와 비정형 데이터가 결합된 융합데이터 각각에 대하여 딥러닝과 머신러닝을 적용한 18가지 모형의 평가지표는 표 2와 같이 정리되었다. 표에서 보는 바와 같이 MAPE 값은 전반적으로 10%~15% 정도로 유사한 정확도를 보였다. 특히 XGBoost에 융합데이터를 적용한 모델의 평가 결과가 가장 뛰어난 것을 확인했다. MLP 모델에 정형데이터를 적용한 모델의 MS와 RMSE가 가장 낮았으며, CNN에 비정형데이터를 적용한 모델의 MAPE가 가장 낮았다.

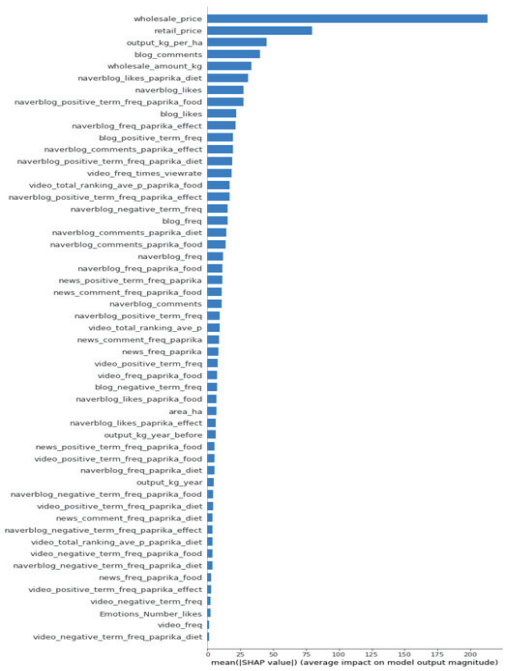
〈표 2〉 모델평가

		Structured data	Un structured data	Combined structured with unstructured data
Experiment1	MLP	MSE: 266631.44	MSE: 186797.76	MSE: 264047.49
		RMSE: 516.36	RMSE: 432.20	RMSE: 513.85
		MAPE: 11.63	MAPE: 12.63	MAPE: 14.72
Experiment2	CNN	MSE: 236306.95	MSE: 243068.60	MSE: 262251.92
		RMSE: 486.11	RMSE: 493.01	RMSE: 512
		MAPE: 15.05	MAPE: 15.15	MAPE: 16.13
Experiment3	LSTM	MSE: 236997.96	MSE: 177564.14	MSE: 220585.07
		RMSE: 486.82	RMSE: 421.38	RMSE: 469.66
		MAPE: 15.04	MAPE: 10.60	MAPE: 11.62
Experiment4	XGboost	MSE: 139660.9	MSE: 171854.10	MSE: 134078.00
		RMSE: 373.71	RMSE: 414.55	RMSE: 366.16
		MAPE: 10.42	MAPE: 10.62	MAPE: 9.30
Experiment5	SVR	MSE: 196306.19	MSE: 237425.56	MSE: 159472.75
		RMSE: 443.06	RMSE: 487.26	RMSE: 399.34
		MAPE: 12.46	MAPE: 15.06	MAPE: 11.14
Experiment6	RF	MSE: 164642.55	MSE: 229873	MSE: 172630.12
		RMSE: 405.76	RMSE: 479.41	RMSE: 415.48
		MAPE: 11.44	MAPE: 12.64	MAPE: 11.09

4.2 결과 시각화

여기서는 SHAP를 사용하여 예측모델의 정확도에 영향을 미치는 변수의 중요도를 살펴본다. 그림 3의 막대 그래프는 SHAP 분석결과를 보여주는 것으로 위쪽에 위치한 변수의 중요도가 높다는 의미이다. 그림에서는 복잡도를 피하기 위해 전체 변수중에서 상위 10개만 표시하였다. 가장 중요한 변수는 정형 데이터와 비정형 데이터에서 나온 것으로 도매가격, 소매가격, 단위면적당 연간 수율(ha), 블로그 댓글, 파프리카 도매 반입량, 네이버블로그 파프리카건강 좋아요 수로 나타났다.

그림 4는 SHAP를 모델에 적용한 결과로 모델의 해석력에 기여하는 자료이다. 여기서는 도매가격, 소매가격, 단위 면적당 연간 수율(ha), 파프리카 언급 블로그 댓글 수, 파프리카 도매 반



〈그림 3〉 Bar Plot of Importance Features

입량, 네이버블로그 파프리카건강 좋아요 수 같은 정형 데이터의 변수가 상위권에 위치한 주요 변수로 나타났다. 도매가격과 소매가격, 단위 면적당 연간 수율, 네이버블로그 파프리카 건강 긍정언급, 파프리카 및 건강 언급 기사 댓글 수는 파프리카 구매량에 강한 영향력을 미치는 데이터임을 확인했다.

4.3 향후과제

파프리카에는 칼륨, 카로틴, 식이섬유, 칼슘, 비타민A, 비타민C 등 풍부한 영양소가 포함된 채소로 국내 소비량도 많지만 신선채소 중 수출량이 가장 많은 채소로 알려져 있다. 파프리카는 일정한 고온이 유지되는 재배환경이 필요하기 때문에 노지보다는 온실재배가 주를 이루며, 볶음 등 익혀먹는 경우보다는 생으로 섭취하는 경우가 많아 잔류농약관리가 식품안전 관리에서 이슈이다. 향후 연구로 채소 등 신선식품의 잔류농약, 유통이나 보관에 필요한 약품처리와 관련된 식품안전 위해관리를 융합한 분석이 필요하다. 연구팀에서는 파프리카 등 온실재배 농산물 재배 농가 및 위해 환경 데이터를 수집, 연계하여 식품위해여부를 예측하는 연구를 수행하고 있다.

〈그림 4〉 Summary Plot of Importance Features

Ⅴ. 결 론

식품 수요예측 모형을 생성할 때 정형 데이터뿐 아니라 비정형 데이터를 함께 활용하는 예측모형을 개발하고, 어떤 변수의 영향도가 높은지를 분석하는 연구를 수행하였다. 본 연구에서는 정형 데이터셋, 비정형 데이터셋, 정형과 비정형 데이터 융합데이터 셋에 대하여 딥러닝과 머신러닝을 적용하여 정확도를 비교하였다. 연구결과를 요약하면 정형 데이터와 비정형 데이터를 결합할 때 모델 정확도가 더욱 향상되었고, 비정형 데이터와 관련된 변수들이 중요한 변수로 나타났다. 또한, 실험 결과에서 머신 러닝은 딥러닝보다 더 효율적으로 예측 모델을 생성할 수 있으며, 만족스러운 성능을 보였다. 분석결과로부터 파프리카 소비 촉진을 위해 언론, 특히 블로그와 동영상의 영향력이 높다는 점이며, 소비촉진을 위해 이들 매체를 활용하는 것이 효율적인 방안이 될 것이다. 향후 과제로는 소셜 미디어 중에서 Youtube가 빠르게 확산되고 있으므로 이에 관한 데이터 수집과 활용이 필요할 것이다. 수많은 유튜버가 음식에 관한 콘텐츠를 만들고 있기 때문이다.

참 고 문 헌

[1] Prabhu, C. S. R., Sreevallabh Chivukula, A., Mogadala, A., Ghosh, R., & Livingston, L. M.

J. “Predictive Modeling for Unstructured Data. Big Data Analytics: Systems, Algorithms, Applications”, (2019). 167-194.

[2] Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. “The power of prediction with social media. Internet Research”,(2013). 23(5), 528-543.

[3] Schoen, H.; Gayo-Avello, D.; Metaxas, P.T.; Mustafaraj, E.; Strohmaier, M.; Gloor, P. “The power of prediction with social media”. Internet Research 2013.

[4] Bahceci, O.; Alsing, O. Stock Market Prediction using Social Media Analysis. 2015.

[5] Artola, C.; Pinto, F.; de Pedraza García, P. Can internet searches forecast tourism inflows? International Journal of Manpower 2015, 36, 103-116.

[6] Cho, W.-S.; Cho, A.; Kwon, K.; Yoo, K.-H. Implementation of smart chungbuk tourism based on SNS data analysis. Journal of the Korean Data and Information Science Society 2015, 26, 409-418.

[7] Meza, X.V.; Park, H.W. Organic Products in Mexico and South Korea on Twitter. Journal of Business Ethics 2016, 135, 587-603.

[8] Yoo, D.-i. Vegetable Price Prediction Using Atypical Web-Search Data. In Proceedings of 2016 Annual Meeting, July 31-August 2, 2016, Boston, Massachusetts.

[9] Lee, S.Y. Analysis on how media report regarding FMD(Foot and mouse disease) affects house-holds’ consumption of meat product. Sogang University, Seoul, 2014.

[10] Choi, K.D.; Kang, H.-G.; Joo, H.H. Does the Harmful Information Regarding Food Safety Affect the Consumption Pattern of Consumers? - Focusing on Fukushima Nuclear Accident. Journal of Korean Economics Studies 2016, 34, 41-83.

[11] Kim, J.; Cha, M.; Lee, J.G. A Model for Nowcasting Commodity Price based on Social Media Data. Journal of Korean Institute of Information Scientists and Engineers 2017, 44, 1258-1268.

[12] Cho, Y.; Oh, E.; Cho, W.-S.; Nasridinov, A.; Yoo, K.-H.; Rah, H. Relations Between Paprika Consumption and Unstructured Big Data, and Paprika Consumption Prediction. International Journal of Contents 2019, 15, 113-119.

[13] Rah, H.; Oh, E.; Yoo, D.-i.; Cho, W.-S.; Nasridinov, A.; Park, S.; Cho, Y.; Yoo, K.-H. Prediction of Onion Purchase Using Structured and Unstructured Big Data. The Journal of the Korea Contents Association 2018, 18, 30-37.

[14] Som Akhamixay, O. Predictive Modeling of the Amount Purchased Paprika Using Deep Learning and Machine Learning. Chungbuk National University, Cheongju, 2021.

[15] Seungwon Oh, Namhui Im,Sang-Hyun Lee, Min Soo Kim. “Long-term Price Prediction and Trend Analysis of Garlic Using Prophet Model.” Journal of the Korean Data Analysis Society 22.6 (2020): 2325-2336.

[16] Shin, S., Lee, M., & Song, S. (2018). A Prediction Model for Agricultural Products Price with LSTM Network. The Journal of the Korea Contents Association, 18(11), 416-429.

[17] Im, J., Kim, W.-Y., Byoun, W.-J., & Shin, S.-J. (2018). Fruit price prediction study using artificial intelligence. The Journal of the Convergence on Culture Technology, 4(2), 197-204.

[18] Mi hye Kim, Sung min Hong,Yoon Sanghoo . (2018).The Comparison of Peach Price and Trading Volume Prediction Model Using

Machine Learning Technique, .Journal of The Korean Data Analysis Society, 20(6), 2933-2940.

[19] Jeong-min Ju, Sun-mee Kang, Ji-wung Choi, Youngwoo Han. “A Study on the Prediction of Apartment Sale Price Using Machine Learning : Focused on the Collection of Internal and External Data and Price Prediction of Korean Apartments.” Proceedings of the Korea Information Processing Society Conference 27.2 (2020): 956-959.

[20] Yoona Noh, Seungwon Jung, Jaeuk Moon, Eenjun Hwang. “Explainable COVID-19 Forecasting Scheme Using Attention LSTM and SHAP.” SIGDB 37.2 (2021): 37-51.

[21] Do Hyeon Lim, Yu-rin Lee, Jaejun Lee, Kee-Young Kwahk, Hyunchul Ahn. “LightGBM-based Dropout Prediction and Its Interpretation using SHAP.” Proceedings of KIIT Conference. 2021.11 (2021): 91-93.

[22] Hyerin Jeong, Park Jung hoon, Yung-Seop Lee, Changwon Lim. (2020). Visualization of Explainable Artificial Intelligence Techniques Using Variable Importance with Its Applications to Health Information Data. Journal of Health Informatics and Statistics, 45(4), 317-334.

저 자 소 개

Somakhamixay Oui

·2018년: National University of Laos, Computer Science (학사)
·2018년～2021년: 충북대학교 빅데이터협동과정(석사)
·관심분야 : 빅데이터, 머신러닝

이 경 희(Kyung-Hee Lee)

·2004년 : 충북대 컴퓨터과학과 (박사)
·2016년～2020년 : 충북대 빅데이터학과 초빙교수
·2020년～현재 : ㈜빅데이터랩스
·관심분야 : 빅데이터, 알고리즘

라 형 철(HyungChul Rah)

·1996년 2월 : 건국대학교 수의학과(학사)
·2000년 2월 : 고려대학교 농학 (석사)
·2001년 9월 : 미국 Univ. of California, Davis 예방수의학 (석사)
·2006년 12 월 : 미국 Univ. of California, Davis 비교병리학(박사)
·2017년 8월～현재 : 충북대학교 대학원 빅데이터 협동과정 초빙교수
·관심분야 : 농축산 및 보건의료 빅데이터 분석

최 은 선(Eun-Seon Choi)
·2019년 : 충북대학교 경영정보학과(학사)
·2021년 : 충북대학교 빅데이터학협동과정(석사)
·2021년~현재 : 충북대학교 빅데이터학협동과정(박사)
·관심분야 : 빅데이터, 머신러닝

조 완 섭(Wan-Sup Cho)
·1987년 : KAIST 전산학과(박사)
·1996년~현재 : 충북대학교(교수)
·관심분야 : 빅데이터, 빅데이터 거버넌스, 블록체인