한국빅데이터학회지
제6권 제1호, 2021, pp. 83-89 https://doi.org/10.36498/kbigdt.2021.6.1.83

주부의 연령대별 농식품 소비 특성 비교*

Comparison of Housewives' Agricultural Food Consumption Characteristics by Age

홍준호[1] · 김진실[1] · 유연주[1] · 이경희[2†] · 조완섭[3]

충북대학교 대학원 빅데이터학과[1], ㈜힐링소프트[2], 충북대학교 경영정보학과[3]

요 약

라이프스타일이 빠르게 변화하고 있고, 식생활과 식품가공 기술의 발전에 따라 가구별로 식품 소비패턴이 매우 다양하다. 본 논문은 가구 단위의 농식품 구매 정보를 담고 있는 농촌진흥청이 구축하고 있는 소비자 패널 데이터의 식품군을 재분류하고 농식품 소비행위 주체인 패널 대표자의 연령대별로 그룹화하여 농식품 소비 특성 비교를 하였다. 연령대 구분의 기준은 대사질환 유병률로 20% 이상인 60대 이상 그룹과 10% 미만인 30~40대 그룹으로 나누었다. LightGBM 알고리즘을 사용하여 30~40대와 60대 이상의 식품 소비패턴의 차이를 분류 분석한 결과 정밀도는 0.85, 재현율은 0.71, F1_score는 0.77로 나타났다. 변수중요도의 결과는 과자류, 엽경채나물류, 조미채류, 과채류, 수산물류 순이었으며, SHAP 지표의 상위 5개 값은 과자류, 수산물류, 조미채류, 과채류, 엽경채나물류 순이었다. 이상치에 민감한 평균을 대신한 중앙값으로 소비패턴을 이진 분류한 결과 과자류의 경우 30~40대가 60대보다 두 배 이상 높은 것을 알 수 있었다. 이외의 변수에서도 30~40대와 60대 이상 사이에서 유의미한 차이를 보였다. 연구 결과 30~40대는 60대보다 과자류를 두 배 이상 소비하는 패턴을 보였으며, 60대의 경우 30~40대보다 수산물, 조미채류, 과채류, 엽경채나물류를 두 배 이상 섭취하였다. 상위 5개 품목 외에도 밀가공식품인 과자, 빵류, 면류에서 30~40대의 소비가 높았으며, 이는 60대의 식품 소비패턴과 차이를 보였다.

■ 중심어 : 농식품 소비, 소비패턴, 패널데이터, LightGBM

Abstract

Lifestyle is changing rapidly, and food consumption patterns vary widely among households as dietary and food processing technologies evolve. This paper reclassified the food group of consumer panel data established by the Rural Development Administration, which contains information on purchasing agricultural products by household unit, and compared the consumption characteristics of agricultural products by age group. The criteria for age classification were divided into groups in their 60s and older with a prevalence of 20% or more metabolic diseases and groups in their 30s and 40s with less than 10%. Using the LightGBM algorithm, we classified the differences in food consumption patterns in their 30s and 50s and 60s and found that the precision was 0.85, the reproducibility was 0.71, and F1_score was 0.77. The results of variable importance were confectionery, folio, seasoned vegetables, fruit vegetables, and marine products, followed by the top five values of the SHAP indicator: confectionery, marine products, seasoned vegetables, fruit vegetables, and folio vegetables. As a result of binary classification of consumption patterns as a median instead of the average sensitive to outliers, confectionery showed that those in their 30s and 40s were more than twice as high as those in their 60s. Other variables also showed significant differences between those in their 30s and 40s and those in their

2021년 07월 14일 접수; 2021년 08월 17일 수정본 접수; 2021년 08월 23일 게재 확정.

* 본 연구는 농촌진흥청 연구사업(농식품 소비, 유전체 특성 및 질병의 연관성 분석 (과제번호: PJ01538032020)) 지원에 의해 이루어졌습니다.

† 교신저자 (lee.kyunghee@gmail.com)

60s and older. According to the study, people in their 30s and 40s consumed more than twice as much confectionery as those in their 60s, while those in their 60s consumed more than twice as much marine products, seasoned vegetables, fruit vegetables, and folioce or logistics as much as those in their 30s and 40s. In addition to the top five items, consumption of 30s and 40s in wheat-processed snacks, breads and noodles was high, which differed from food consumption patterns in their 60s.

■ Keyword : Agri-food consumption, consumption pattern, panel data, LightGBM

Ⅰ. 서 론

전 세계적으로 식품 가공기술 발달과 라이프스타일의 변화에 따라 가공식품의 소비가 증가하고 있다. 한국농수산식품유통공사의 2020 식품 외식산업 주요 통계자료(2019)의 국내 식품산업 성장 추이에 따르면 국내 식품 산업 시장 규모는 2017년 기준 218.02조 원으로 2007년 107.52조 원보다 약 110조 증가하였다. 전년도인 2016년 대비 3.8% 증가율을 보였다.

식품 산업 규모의 증가에 따라 식품 소비 트렌드도 변화되어, 다양한 연구가 진행되었다. 주로 마케팅 측면에서 연령별 절임 식품 소비나 라이프스타일에 따라 소비자를 유형화하여 소비자 집단 간 차이 분석 등의 연구가 이루어졌다. 본 논문에서는 위 마케팅적 측면의 소비 트렌드 연구와 달리 보건 의료 측면에서 기초자료를 목적으로 하므로 농식품 소비 특성을 결정하는 다양한 요인 가운데 연령대별 대사질환 유병률에 주목하였다. 주요 식품 구매자인 주부를 패널 대표자로 설정하고 패널 대표자의 연령을 대사질환 유병률 차이로 구분하여 그룹별 식품 소비 간의 관계를 파악고자 한다.

본 논문에서는 가구 단위의 농식품 구매 정보를 담고 있는 농촌진흥청의 농식품 소비자 패널 데이터를 이용한다. 소비자 패널 데이터의 식품군을 재분류하고 농식품 소비행위 주체인 패널 대표자의 연령대를 대사질환을 가지고 있을 확률이 높은 60대 이상과 30~40대 가구로 나누었다. 각 연령층에서 주로 구매하는 식품군을 머신러닝 분류 알고리즘인 LightGBM을 이용하여 분류하여 그룹별 농식품 소비 특성과 차이에 관하여 연구하였다.

Ⅱ. 선행연구

2.1 소비 특성 관련 연구

소비식품 차이에 대한 연구는 마케팅 전략을 위한 표적 시장 분석을 위하여 활발하게 이루어져 왔다. 김혜영(2019)의 연구에서는 FGI 조사로 설문지를 보완하고 맛, 브랜드, 식재료, 원산지 4가지 속성을 도출하여 컨조인트 분석을 실시하고 통계분석을 통해 연령별 절임 식품 선호도에 대한 연구를 진행하였다[2]. 박명은 외(2019)는 기능성 식품 소비자의 라이프스타일에 따라 소비자를 유형화하기 위하여 탐색적 요인분석과 군집분석을 실시하여 3개의 소비자 유형으로 분류, 소비자 집단 간 차이 분석을 진행하였다[3]. 본 논문도 집단의 소비 특성 비교 분석을 위하여 소비자 유형을 새로운 방식으로 나누어 연구를 진행한다.

마케팅적 관점 외에 질병 관리 측면에서도 소비자 유형별 식품 소비에 대한 연구가 활발히 이루어졌다. 송혜영 외(2015)는 국민건강영향조사 제5기 자료를 이용하여 노인의 식습관 식사

횟수, 성별, 흡연, 배우자 유무에 대한 차이 비교로 노인의 비만 특성을 파악하고, 노인 비만을 해결하기 위한 방안을 제시하였고[4], Sonia S. Anand 외(2015)는 글로벌 식량 시스템 및 글로벌 식이 패턴에 대하여 연구하였는데, 세계화된 식량시스템의 발전이 식량 공급에 미치는 영향과 식품 섭취가 심혈관 뇌 질환 및 관련 동반질환에 영향을 주는지에 대하여 연구하였다[5]. 본 논문은 이러한 질병 관리 측면에서 연령과 밀접한 관계가 있는 대사증후군 유병률 차이로 소비자를 그룹화하여 식품 소비에 대한 연구를 진행한다.

기존의 연령별 대사질환자 연구를 보면, 임현정 외(2017) 연구에서 전체 성인을 연령대별 세 그룹으로 나누고 유병률을 분석하여 연령대별 그룹에 따른 대사 증후군 유병률을 연구하였다[6], 전나미 외(2018) 연구에서 5년간 여성 패널 데이터를 활용하여 여성의 대사증후군 유병률을 분석하였는데, 20대 1.3-2.1%, 30대 2.2-3.8%, 40대 5.7-7.6%, 50대 10.9-14.0%, 60대 20.3-22.4%, 70대 이상 23.8-29.8%로 나타났다[7]. 여성의 경우 대사질환 유병률이 연령이 증가함에 따라 함께 증가하였다. 이를 참고하여 패널 대표자의 연령을 기준으로 대사질환 유병률이 20% 이상인 60대 이상과 10% 미만인 30~40대 두 집단으로 그룹화하여 연구를 진행하고자 한다.

2.2 LightGBM 알고리즘

본 논문에서는 분류 분석을 위하여 다른 분류 알고리즘과의 비교를 통해 가장 빠르고 적은 손실률을 보이는 LightGBM 알고리즘을 선택하였다. 일반적인 트리 알고리즘은 분류 분석에서 시각화가 용이하여 설명력이 좋고 비전문가도 쉽게 이해할 수 있다는 장점이 있어 널리 사용된다. 그 중 LightGBM(Light Gradient Boosting Machine)은 Microsoft사에서 개발한 기계학습 무료 오픈소스 알고리즘으로 결정 트리 알고리즘을 기반으로 하여 순위지정이나 분류 작업에 사용된다.

LightGBM은 트리 기반의 부스팅 알고리즘 중에도 좋은 성능을 보여주는 것으로 알려진 XGBoost의 단점을 개선하였다. XGBoost와 같은 기존의 부스팅 알고리즘은 level-wise 분석으로 레벨 단위로 트리를 수평적으로 확장한다. 반면 LightGBM은 leaf-wise 분석으로 max delta loss를 가지는 잎을 기준으로 수직적으로 트리를 확장하는 알고리즘이다[8][9]. 따라서 두 알고리즘으로 동일한 잎을 생성한다고 가정할 때, 손실률과 관계없이 대칭적으로 확장하여 시간이 오래 걸리는 level-wise 분석 알고리즘보다 빠르고 손실이 적은 분류 모델을 얻을 수 있다.

Ⅲ. 연구방법

본 장에서는 연구에 사용된 데이터 세트를 설명하고, 다변량 자료의 차원축소를 위해 주성분 분석과 요인분석한 결과를 기술한 후, 두 기법을 비교한다.

3.1 데이터 세트

본 연구에 사용된 데이터는 2015년부터 2019년까지 조사된 5,829,475건의 농촌진흥청 농식품 장바구니 패널 데이터로, 가구 단위 패널의 농식품 구매 정보를 담고 있다. 60대는 493가구 30~40대는 936가구로 해당 데이터의 소비식품은 기준에 따라 대분류, 중분류, 소분류로 나누어져 있는데, 원활한 분류를 위하여 <Table 1>을 기준으로 식품군을 재분류하여 진행하였다.

〈Table 1〉 식품 재분류표

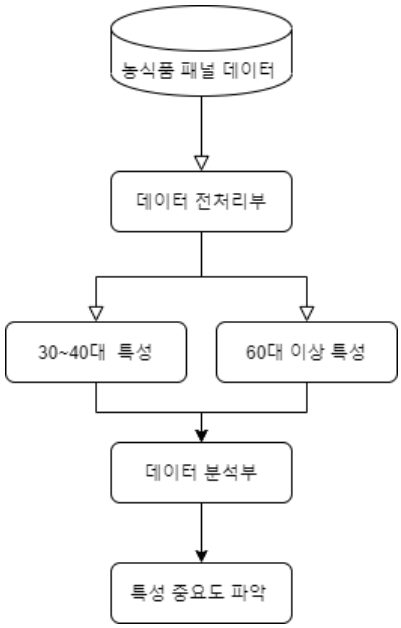
구분	변경 전		변경 후	
	분류	%	분류	%
1	가공식품	58%	과자류	6%
			빵류	5%
			면류	4%
			수산,축산물 가공식품	8%
			유제품	12%
			음료	5%
			주류	3%
			간식류	5%
			기타가공식품	9%
2	축산물	8%	축산물	8%
3	수산물	4%	수산물	4%
4	과일류	6%	과채류	12%
5	과채류	5%		
6	기타채소류	1%		
7	엽경채류	4%	엽경채, 나물, 근채, 서류	12%
8	나물류	3%		
9	근채류	2%		
10	특작류	2%		
11	서류	1%		
12	조미채류	6%	조미채류	6%
13	곡물류	1%	기타	1%
14	건과_견과류	0.1%		
15	과일과채혼합	0.1%		
	합계	**100%**	**합계**	**100%**

3.2 연구 내용 및 방법

첫째로, 분류 과정에서 데이터의 편향을 줄 수 있는 소비의 횟수가 너무 많거나 적은 300가구를 제외한 후 2015년~2019년까지 1429가구를 30~40대(936가구), 60대 이상(493가구)으로 분류하였다.

둘째로, 30~40대와 60대 이상을 분류하는 소비 특성의 변수 중요도를 찾기 위하여 LightGBM의 분류 모델링을 통하여 중요 변수를 도출하였다.

셋째로, 변수 중요도와 주 효과를 같이 나타내주는 SHAP 지표를 가지고 30~40대와 60대를 분류하는 주요 변수의 영향을 찾았다.



〈Fig. 1〉 분석 흐름도

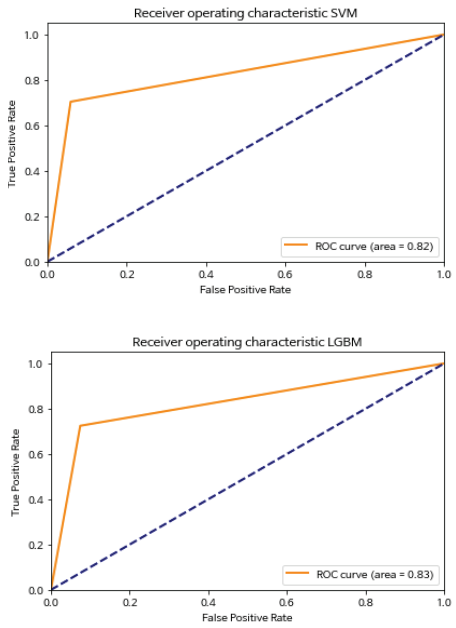
Ⅳ. 실험결과 및 해석

데이터를 트레이닝 데이터셋 80%, 테스트 데이터셋 20%로 나누어 LightGBM 분류 알고리즘을 사용하였으며, learning_rate = 0.003, max_depth = 10으로 실험한 결과 정밀도는 0.86, 재현율은 0.72, F1_score는 0.77, ROC_AUC값은 0.825로 나타났다.

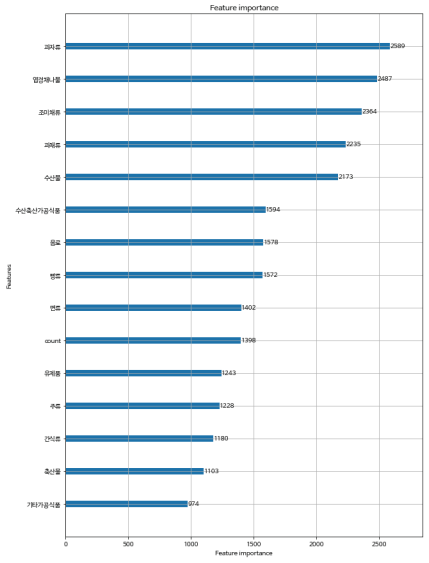
데이터 분석결과를 비교하기 위해 SVM 분류 알고리즘을 비교한 결과 큰 차이는 나지 않았지만 LightGBM 결과가 더 좋게 나타났다.

분류 분석에 사용된 변수들의 변수중요도를 살펴본 결과는 <Fig. 3>과 같다.

<Fig. 3>의 결과 과자류, 엽경채・나물류, 조미채류, 과채류, 수산물류가 중요 변수로 나타났다. 이를 impact 효과를 고려하여 시각화한 SHAP지표를 보면 <Fig. 4>와 같다.

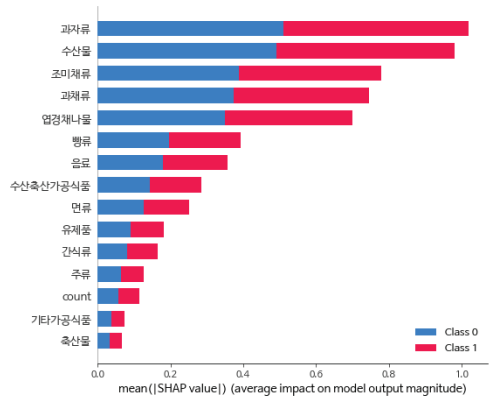


〈Fig. 2〉 SVM과 LightGBM 분석결과 비교

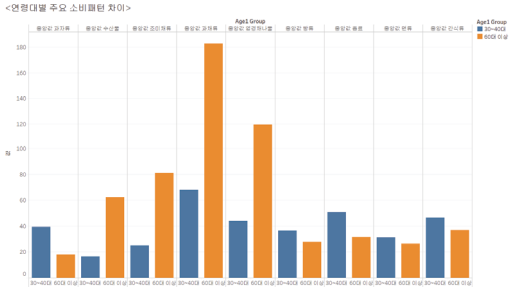


〈Fig. 3〉 feature importance

<Fig. 4>의 상위 5개의 값을 보면 클래스를 분류하는 주요 변수를 잘 나타내고 있다. 주로 과자류, 수산물류, 조미채류, 과채류, 엽경채・나물류에 의해 30~40대와 60대 이상의 가구가 분류될 수 있음을 의미한다.



〈Fig. 4〉 SHAP values



〈Fig. 5〉 주부 연령대별 주요 소비패턴 차이

결과를 해석하기 위해서 30~40대와 60대 이상의 가구의 주요 변수 소비량의 중앙값 차이를 비교하였다.

<Fig. 5>에서 평균은 이상치에 민감하기에 중앙값으로 소비패턴의 차이를 나타냈다. 위 결과에서 중앙값의 차이가 큰 변수가 이진 분류 시에 중요변수로 작용하는 것이 아님을 알 수 있다. 하지만 과자류의 경우 30~40대가 60대보다 2배 이상 높은 것을 알 수 있으며, 나머지 4개의 변수인 수산물, 조미채류, 과채류, 엽경채・나물류에서는 60대가 30~40대보다 2배 이상 높은 소비패턴을 보였다. 상위 5개의 중요변수를 제외한 값들은 크게 차이는 나타나지 않았지만, 빵류, 음료, 면류, 간식류도 30~40대의 소비가 높았다.

V. 결론

본 논문에서는 패널 대표자의 연령대를 대사증후군 유병률에 따라 60대 이상, 30～40대 두 그룹으로 나누어 비교한 결과를 제시하였다. 30～40대의 소비패턴의 경우는 과자류를 60대보다 2배 이상 소비하는 패턴을 보였으며, 60대의 경우 30～40대보다 수산물, 조미채류, 과채류, 엽경채・나물류에서 2배 이상의 소비패턴을 보여주었다. 이를 통해 30～40대보다 60대 주부의 소비패턴이 상대적으로 과자류는 적게 먹고 수산물, 조미채, 과채, 엽경채, 나물류를 많이 소비함을 알 수 있었다. 상위 5개의 품목 외에도 주로 밀가공식품인 과자, 빵류, 면류는 30～40대의 소비패턴이 높았으며, 60대의 소비의 패턴과 차이를 보여주었다.

연구의 한계점은 가구 단위의 패널데이터를 활용하면서 패널 대표자의 연령으로만 집단을 구분하여 구체적으로 각 가구원들의 나이나 질병과 같은 특성이 반영되지 않은 점에 있다. 한성희(2010)는 20～50대 기혼 여성을 대상으로 웰빙 지향 식품 구매행동을 분석하였는데 기혼 여성이 웰빙 지향 식품을 구매할 때 부모나 자녀를 위해 구매한다는 결론을 도출하였다[10]. 패널 대표자의 연령대를 유병률만으로 구분하는 것은 소비패턴의 인과성이 명확하지 않다는 한계를 보여준다. 하지만 패널 데이터의 개인정보의 문제로 인한 한계점으로 인해 각 가구원의 질병과 같은 특성을 반영하기 어려운 점을 고려해 보았을 때 한국 사회에서 주부의 연령층에 따라 주로 구매하는 농식품류가 무엇이고 그룹별 어떤 식품을 가장 많이 소비하는지에 대한 차이점을 분석하였다는 점에서 본 연구의 의의를 찾을 수 있다.

본 논문에서는 식품의 소비패턴의 차이만을 규명하였지만, 식품 소비가 건강에 미치는 영향이나 연령대별 식품 소비패턴이 차이가 나는 명확한 원인을 설명하기에는 부족하다. 향후 질환정보나 건강검진 정보를 연계하여 연령별 주요 식습관이 특정 질병에 미치는 영향에 대하여 연구할 수 있을 것으로 기대한다.

참 고 문 헌

[1] 한국농수산식품유통공사,「2019년도 식품외식산업 주요통계」

[2] 김혜영 (2019). 컨조인트 분석을 이용한 소비자의 연령별 절임 식품 선호도 연구. Culinary Science & Hospitality Research, 25(8), 170-182.

[3] 박명은, 남정미, 유소이. (2019). 생태 친화적 기능성식품 선택과 관련된 소비자 특성 분석. 한국산학기술학회 논문지, 20(8), 456-471.

[4] 송혜영, 박효은. (2015). 노인의 식습관에 따른 비만도. 한국산학기술학회 논문지, 16(8), 5404-5412.

[5] Anand, S. S., Hawkes, C., De Souza, R. J., Mente, A., Dehghan, M., Nugent, R., Zulyniak, M. A., Weis, T., Bernstein, A. M., Krauss, R. M., Kromhout, D., Jenkins, D. J. A., Malik, V., Martinez-Gonzalez, M. A., Mozaffarian, D., Yusuf, S., Willett, W. C., & Popkin, B. M. (2015). Food Consumption and its Impact on Cardiovascular Disease: Importance of Solutions Focused on the Globalized Food System A Report from the Workshop Convened by the World Heart Federation. In Journal of the American College of Cardiology (Vol. 66, Issue 14, pp. 1590–1614). Elsevier USA. https://doi.org/10.1016/j.jacc.2015.07.050

[6] 임현정, 김응준. (2017). 복합표본분석을 적용한 한국 성인의 대사증후군 유병률. 한국체육측정평가학회지, 19(3), 85-97.

[7] Chun, N., & Chae, H. J. (2017). Prevalence of

Metabolic Syndrome and Its Components in Adult Women. *Journal of Korean Biological Nursing Science*, 20(4), 261-269. https://doi.org/10.7586/jkbns.2018.20.4.261

[8] Microsoft Corporation. LightGBM Release 3.2.1.99. 2021. https://lightgbm.readthedocs.io/

[9] Al Daoud, E. (2019). Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *International Journal of Computer and Information Engineering*, 13(1), 6-10.

[10] 한성희. (2010). 기혼여성의 가정생활관리행동: 웰빙지향 식품 구매행동 및 식생활 행동과 소비만족도, 14(2), 127-152.

저 자 소 개

홍 준 호 (Jun-Ho Hong)
·2015년 : 고려대학교 응용통계학과 (학사)
·2020년~현재 : 충북대학교 빅데이터협동과정 석사
·관심분야 : 빅데이터, 머신러닝

김 진 실 (Jin-Sil Kim)
·2017년 : 청주대학교 광고홍보학과 (학사)
·2020년~현재 : 충북대학교 빅데이터협동과정 석사
·관심분야 : 빅데이터, 머신러닝

유 연 주 (Yeon-Ju Yu)
·2021년 : 한국교통대 영어영문학과(학사)
·2021년~현재 : 충북대학교 빅데이터협동과정 석사
·관심분야 : 빅데이터, 머신러닝

이 경 희 (Kyung-Hee Lee)
·2004년 : 충북대 컴퓨터과학과 (박사)
·2016년~2020년 : 충북대 빅데이터학과 교수
·2020년~현재 : ㈜힐링소프트
·관심분야 : 빅데이터, 알고리즘

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