Journal of the Korean Data Analysis Society (February 2019)
Vol. 21, No. 1, pp. 95-106

LASSO 방법을 이용한 프라이버시 침해 우려 행태 분석*

엄지은[1], 전승환[2], 전수영[3]

요 약

다양한 미디어 채널 사용이 급격히 증가함과 동시에 각종 서비스와 미디어 콘텐츠를 이용하기 위해서는 개인정보를 기재해야한다. 하지만 요즈음 급증하고 있는 개인정보의 유출에 많은 문제점이 대두되고 있다. 따라서 본 연구는 2014년-2017년 기간 동안의 미디어패널 자료를 이용하여 프라이버시 침해 우려 정도에 대해 다방면으로 분석을 진행하였다. 보통 설명변수가 많을 경우 다중공선성 및 과적합 문제가 발생하게 되는데 이를 해결하기 위해 벌점화 모형 중에서 LASSO 방법이 주로 사용된다. 이에 본 논문에서는 LASSO 회귀모형으로 연도별 프라이버시 우려 정도에 미치는 변수를 선택한 후 의사결정나무 분석을 이용하여 프라이버시 침해 우려에 관한 이용자들의 특징을 분류해보았다. 분석 결과, 개인 프라이버시 침해에 대해 2014년-2017년 각각 SNS 활동, 스마트폰 이용시간, 와이파이존 이용시간, 인터넷 카페 활동 빈도가 가장 큰 영향을 주는 것으로 나타났다. 또한 성별과 핸드폰 제조사 브랜드 별로 나누어 분석해본 결과, 남성들은 카오디오, 문서작업, 가구원 수에 영향을 많이 받고, 여성들은 월평균 소득, 수면시간, 애플리케이션 관련 지출금액, 음악/음원 청취 시간에 영향을 많이 받는 것으로 나타났다. 핸드폰 브랜드별 분석 결과에서 삼성과 LG는 연도 추세를 따라가는 공통점도 보였지만 세 브랜드 모두 각자의 특징을 보였고, 특히 애플의 경우 삼성, LG와는 확연히 다른 형태를 보였다.

주요용어 : LASSO 회귀, 의사결정나무, 미디어패널, 다중공선성, 과적합.

1. 소개

요즈음 다양한 형태의 미디어 채널이 급증하면서 많은 양의 정보를 전송하고 전달받을 수 있게 되었고, 인터넷을 통한 미디어 접속과 SNS 이용이 활성화되면서 더욱 많은 양의 개인 정보가 공개되고 있어 이용자들이 불안감을 느끼고 있다. SNS 이용 빈도는 전통적인 매체 이용 행위와는 상충관계에 있으며, 인터넷 등의 신규 매체를 이용한 활동들과는 동일한 흐름을 보이고 있다는 연구(Lee, 2017)가 인터넷 이용과 SNS 이용이 크게 늘어나는 추세임을 보여준다. 또한 각종 서비스 및 콘텐츠를 이용하기 위해서는 사이트 등록 및 회원가입이 필수적인데 이 과정에서 개인정보가 유출되는 문제가 발생하고 있다.

이렇게 유출된 신상정보에는 이름, 주소, 주민등록번호, 휴대폰 번호 등 상세한 정보가 포함되어 있으며 이를 이용해 개인에게 물리적, 정신적 피해를 주고 있어 더 큰 문제가 되어가고 있다. 몇 년 전, 유명 SNS 개인정보가 유출되어 사용되고 있다는 사실이 밝혀져 논란이 일어나기도 했다.

*이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2016R1D1A1B03933876).
[1]30019 세종시 세종로 2511, 고려대학교 공공정책대학 응용통계학과 대학원 석사과정.
E-mail : eomji9689@korea.ac.kr
[2]30019 세종시 세종로 2511, 고려대학교 공공정책연구소 연구교수. E-mail : bighumanities@korea.ac.kr
[3](교신저자) 30019 세종시 세종로 2511, 고려대학교 공공정책대학 경제통계학부 빅데이터전공 부교수.
E-mail : scheon@korea.ac.kr
[접수 2019년 1월 20일; 수정 2019년 2월 10일, 2019년 2월 17일; 게재확정 2019년 2월 20일]

최근 많은 연구와 신문 기사들은 SNS의 부정적인 측면으로 특정인 신상 털기와 같은 프라이버시에 대한 침해 우려를 거론하고 있다. SNS 사용 시 인지되는 프라이버시 침해 우려의 역할에 대해 검증한 연구(Kim et al., 2012)가 SNS와 프라이버시 침해 우려의 심각성에 대해 보여준다. 이에 앞서 소셜 네트워크 서비스에서는 많은 수의 계정이 노출되었고 이를 이용해 다른 정보와 조합하여 개인을 유추할 수 있는 가능성이 높다는 연구(Choi et al., 2013)가 있었으며 인터넷 이용자의 개인정보 유출 가능성에 대해 성별, 인터넷 사용량, 사이트 가입 수, 전자상거래 빈도에 따라 심리적 불안 요소를 분석한 연구도 있었다(Jin, Kim, 2011).

본 연구의 목적은 어떤 온라인 활동으로 인해 프라이버시 침해 우려 정도가 달라지는지 알아보고자 하는 것이다. 이를 위해 한국미디어패널자료를 이용한 다양한 연구(Lee, 2017; Cheon, 2018)를 기반으로 개인 및 다이어리 자료 4년치(2014년~2017년)를 사용하여 연도별로 온라인 활동으로 인한 개인 프라이버시 우려 정도에 영향을 미치는 미디어 사용 형태와 관련 있는 변수들의 변화 추이를 알아보았다.

또한 많은 설명변수들로 인해 과적합문제가 발생하고 자료의 특성상 설명변수들 간의 다중공선성이 발생하기 때문에 이에 대한 대안으로 LASSO 회귀 모형을 통하여 유의미한 변수들이 어떠한 형태로 프라이버시 침해 우려 여부에 영향을 미치는지 의사결정나무 모형을 통하여 살펴보고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 미디어 패널자료, 3장에서는 LASSO 회귀모형에 대해 설명한다. 4장에서는 연도별 LASSO 회귀 모형 결과에 따른 의사결정나무 결과를 설명하고, 5장에서는 특정 관심 있는 집단별로 나눈 자료에서의 LASSO 회귀 모형 결과와 의사결정나무 결과를 설명한다. 마지막으로 6장에서는 결론을 다룬다.

2. 미디어 패널 자료

본 논문에서 분석을 위해 사용된 자료는 정보통신정책연구원에서 주관한 한국미디어패널조사의 원시자료인 국가승인통계이며, 2011년 전국 16개 광역시도(제주 포함) 5,085가구 및 해당 가구원의 만 6세 이상 가구원을 대상으로 구축된 패널을 대상으로 2014년부터 2017년까지 총 4개 년도의 패널 자료이다(“일반 사용이 허용된/공개된 자료이다”). 조사대상은 4년 동안 추적조사가 완료된 9,788명 중에서 NA값(무응답, 모름)이 포함되거나 이상치를 제거한 4,352명이다.

본 논문에서는 개인용 설문지로부터 얻어진 여러 조사항목 중 대상자의 일반적 특성을 알아보기 위해 ID, 연도, 성별, 핸드폰 제조 브랜드, 나이, 소득, 가구원 수, 직업 유무, 최종 학력 일곱 가지 변수를 고려하였다.

프라이버시 침해에 대한 우려 정도는 성별을 제외한 모든 일반적 특성에서 유의한 차이가 있었다(Table 1). 또한 개인 설문지 중 방송통신 서비스 가입 및 지출에서 11가지 정보, 미디어 이용행태에서 23가지 정보, 다이어리 자료에서는 모든 매체, 연결방법, 행위를 이용하여 총 143가지 변수를 분석에 사용하였고, 미디어 행태의 프라이버시 침해에 대한 우려 정도를 종속변수로 사용하였다. 개인용 자료에 포함되어 있는 가중치를 LASSO 회귀에 있어서 모든 독립변수에 적용시켜 분석하였다.

분석 초기에 LASSO회귀와 전체적인 의사결정나무 구축 시에는 앞서 언급한 4,352명에 대한 자료를 표본으로 사용하였지만, 성별은 개인용 설문지로부터 얻어진 gender 변수에서 둘로 나누었고, 핸드폰 제조 브랜드는 삼성, LG, 애플 3사를 사용하는 사람들만 대상으로 하여 분류한 표본으로 의사결정나무를 구축하였다.

Table 1. General characteristics of participants: TS means test statistics

Characteristics		2014			2015			2016			2017		
		Mean	TS	p-value	Mean	TS	p-value	Mean	TS	p-value	Mean	TS	p-value
Gender	Male	3.443	-1.348	0.178	3.255	-0.393	0.694	3.401	1.633	0.103	3.636	-2.125	0.034
	Female	3.485			3.267			3.357			3.697		
Job	Employed	3.509	3.190	0.001	3.314	4.384	<0.001	3.442	5.890	<0.001	3.710	3.692	<0.001
	Unemployed	3.408			3.185			3.280			3.602		
Age	≤19	3.271	23.728	<0.001	3.100	27.779	<0.001	3.219	26.138	<0.001	3.583	16.121	<0.001
	20-29	3.666			3.443			3.480			3.791		
	30-39	3.589			3.389			3.491			3.713		
	40-49	3.572			3.372			3.513			3.791		
	50-59	3.337			3.173			3.292			3.601		
	60-69	3.024			2.816			3.019			3.370		
	70-79	2.344			2.460			2.737			3.227		
	≥80	3.466			2.139			2.227			2.577		
Education	No formal education	3.673	63.554	<0.001	3.596	58.963	<0.001	3.515	50.853	<0.001	3.756	37.233	<0.001
	Middle school	2.737			2.602			2.720			2.986		
	High school	3.129			2.961			3.065			3.342		
	University	3.423			3.192			3.339			3.640		
	Graduate school	3.665			3.443			3.519			3.788		
Household	One-person (single)	3.216	16.384	<0.001	3.012	41.222	<0.001	3.231	20.998	0.001	3.525	13.947	<0.001
	Two people	3.278			2.951			3.166			3.492		
	Three or more people	3.505			3.319			3.415			3.702		
Phone-brand	Samsung	3.441	7.108	0.001	3.247	7.189	0.001	3.340	13.292	<0.001	3.654	7.328	0.001
	LG	3.523			3.253			3.456			3.648		
	Apple	3.733			3.492			3.594			3.858		

3. LASSO 회귀 모형

회귀 모형을 만드는 경우 다중 회귀 모형이 흔한데, 일반적으로 설명 변수의 수가 증가하면 설명 변수 간 강한 상관관계로 인한 다중공선성과 과적합 문제가 존재하게 되어 회귀계수 추정량의 분산이 커지게 되며, 추정 회귀식의 예측 정확도가 떨어지게 되고 변수에 대한 해석력이 저하되는 문제점이 발생한다. 이런 문제점을 해결하기 위해 일반적으로 사용하는 방법이 벌점화 방법이다.

LASSO(Tibshirani, 1996) 방법은 벌점화 방법 중에 하나로 다른 벌점화 방법인 능형회귀와 달리 변수선택 방법으로 일반적으로 많이 이용되고 있다. LASSO 회귀는 능형 회귀의 장점인 축소를 사용하면서 동시에 변수선택을 통해 예측력을 향상시키고, 그에 따라 최종 모형에 대한 해석을 용이하게 하는 방법이다. 예측 정확도를 높이기 위해 영향력이 적은 회귀 계수 값을 0으로 만듦으로써 변수 선택을 하는데 이것은 고차원 자료의 최종 모형에 대한 해석력을 높여준다.

LASSO 추정치는 설명변수 x로 반응변수 y를 추정할 때 제약조건 $\sum_{j=1}^{p}|\beta_j| \le t$ 하에서 $\widehat{\beta^{lasso}} = \mathrm{argmin}(\beta)\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j)^2\right\}$이며, 라그랑지 승수법에 의하면

$$\widehat{\beta^{lasso}} = \arg\min(\beta)\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

와 동치가 되는 벌점모수 λ가 존재한다.

이와 같이 λ값에 따라 축소추정을 어느 정도로 할 것인지 결정된다. 즉, λ가 클수록 계수가 0에 수렴을 하고, λ가 작아질수록 최소제곱법의 추정치와 가까워진다. 따라서 벌점모수 λ를 크게 하여 유의하지 않는 계수의 추정값을 0으로 만들어 변수선택이 이루어지는데 λ의 크기는 일반적으로 CV(cross-validation)를 이용한다(Hastie et al., 2009).

본 연구에서는 고차원인 유한표본 자료에서의 예측력이 다른 방법에 비해 좋은 LASSO 회귀 모형을 데이터 마이닝 기법으로 이용하였다. 최근 다양한 미디어 이용의 증가로 개인 정보와 프라이버시가 화두로 떠오르고 있는 점을 감안하여, 개인이 프라이버시 침해에 대해 우려하고 있는 정도와 관련 있는 설명변수를 탐색하고자 한다. 미디어를 사용하는 사람들의 개인정보 우려 정도를 파악, 그리고 어떤 요인이 프라이버시 우려에 영향을 주고 있는지 분석하고, 그것을 이용하여 사람들이 미디어를 좀 더 경계심 없이 사용할 수 있도록 할 것이다.

4. LASSO회귀 모형을 통한 의사결정나무 분석

미디어 패널 자료 중 설명변수로 다이어리 자료의 각 행위, 매체 종류, 연결 방법을 이용한 시간, 그리고 개인 자료 중에서 2014년에서 2017년까지 공통으로 있는 변수들을 모두 이용하여 총 143개의 변수를 선택하였다. 종속변수는 개인 미디어 이용행태의 프라이버시 침해에 대한 우려 정도 중 온라인 활동에 관한 항목 4가지만 가져와서 이용하였다. ‘온라인 사이트에 가입할 때 개인정보를 너무 많이 요구하는 것이 걱정스럽다.’, ‘내 온라인 아이디를 도용당할까 걱정스럽다.’, ‘일반적으로 인터넷을 사용할 때 나의 프라이버시에 대해 걱정스럽다.’, ‘온라인에서 자기가 누구인지 밝히지 않은 사람들은 의심스럽다.’ 이러한 4가지 항목에 대해 1번 ‘전혀 그렇지 않다’부터 5번 ‘매우 그렇다’의 응답을 평균 낸 뒤, 그 평균이 4 이상이면 비교적 걱정을 많이 한다는 의미의 1, 4미만이면 그렇지 않다는 의미의 0 두 가지로 코딩하였다. 또한 본 조사는 복합표본추출 방법에 의해 표본이 추출되었으므로, 추정의 정확도를 높여주기 위해 가중치를 고려하였다.

R의 glmnet 패키지를 이용하여 각 연도별로 LASSO 회귀방법을 적용하여 총 143개의 독립변수들 중 유의한 변수들이 2014년에는 81개, 2015년에는 45개, 2016년에는 51개, 2017년에는 58개로 나타났다(Table 2). 가장 유의한 변수들을 살펴보면, 2014년에는 TV 방송 프로그램 보기 지출 금액, 가구원 수, 지난 3개월 동안 인터넷 뉴스/토론 게시판 댓글, 글쓰기 활동 빈도가 계수가 높게 나타났다. 2015년에는 가구원 수, 음악 지출 금액, 게임 지출 금액, 지난 3개월 동안 인터넷 온라인 추천, 평점 주기 기능 활동 빈도가 높을수록 개인 프라이버시 침해에 대해 더욱 우려하는 것으로 나타났다. 2016년에는 블로그 사용/운영 여부와 학력, 가구원 수, 인터넷 동호회/카페/클럽 운영 여부, 지난 3개월 동안 인터넷 동호회/카페/클럽 글쓰기 활동 빈도가 유의한 가장 유의한 변수이다. 마지막으로, 2017년은 성별, 가구원 수, 인스턴트 메신저 사용 여부, 지난 3개월 동안 인터넷 동호회/카페/클럽 게시글 스크랩 활동 빈도수가 유의하게 나타났는데, 여자일수록 그리고 가구원 수가 많을수록, 인스턴트 메신저를 사용하고, 게시글 스크랩 활동 빈도수가 많을수록 개인 프라이버시에 대해 우려하는 것을 알 수 있었다.

지도학습 문제에서 최종 모형의 예측력과 해석력이 중요한 만큼, 어떤 사회 현상에 대해서 예측하는 자살 생각 예측 요인 연구(Kwon, 2010; Kim, An, 2010)에 의사결정 나무가 주된 방법론으로

분석되었다. 의사결정 나무는 지도학습 기법 중 하나로 각 변수의 영역을 반복적으로 분할함으로써 전체 영역에서의 새로운 규칙을 만든다. 앞서 LASSO 회귀로 만든 모형의 예측이 잘 들어맞는지 확인하기 위해 모든 변수들을 가지고 의사결정 나무를 그려 변수들의 영향력을 확인해보았다. 그 결과, 프라이버시 침해와 관련이 크지 않은 변수들로 의사결정 나무가 나뉘었고 나뉜 결과 또한 우려하는 사람들과 우려하지 않는 사람들의 비율이 5:5로 제대로 분류되지 않는 것을 볼 수 있었다. 모든 형태의 자료들이 의사결정나무로 좋은 결과가 나오지 않아, 의사결정나무 분석 전에 다른 분류 방법을 접합하는 방법을 제안하는 경우가 있는데, 계수형 자료 분석을 위한 의사결정나무에서의 변수 선택 방법을 제안한 연구(Lee, Jo, 2012)도 있다. 이 연구에서는 143개의 변수들 중 유의한 변수들만 선택하여 의사결정 나무 분류를 하는 것이 효율적이라고 판단하여, 앞서 LASSO 회귀 모형 분석을 통해 유의하게 나온 변수들만 가지고 의사결정 나무 분석을 진행했다.

Table 2. Variables selected by LASSO regression model: CVM means mean cross-validated error - length(λ)

Year	2014	2015	2016	2017
CVM	0.2197	0.1925	0.1981	0.2360
$\min_{\lambda} CVM$	0.0060	0.0094	0.0085	0.0097
Number	81	45	51	58
Most significant variables	The expenditure amount on TV program show, the number of household members, the frequency of writing on Internet news/debate board during the last three months	The number of household members, the amount spent on music, the amount spent on game, the frequency of online recommendation/ rating during the last three months	Whether or not to use/ manage blog, highest level of education, number of household members, Whether or not to manage Internet, frequency of writing in Internet cafe/club	Sex, number of household members, Whether or not to use instant messages, frequency of Clipping postings in Internet cafe/club/society during the last three months

Table 3에서 각 연도별로 의사결정 나무 분석 결과 유의하게 갈라진 요인들을 차례로 작성한 것이며, 각 연도 아래의 백분율은 의사결정 나무 분류표의 정확도(accuracy)를 나타낸다. 각 연도별 프라이버시 침해에 영향을 미치는 요인에 대하여 구체적으로 살펴보면, 2014년에는 페이스북, 트위터 등의 SNS의 인기가 높아지고 이용자들이 많아짐으로써 가장 유의한 변수가 SNS 이용시간이다. 2014년 한국에 페이스북 이용자가 확연히 늘면서, 처음 접하는 SNS에 대해 개인 정보 침해를 우려하고 경계했던 추세를 읽을 수 있다. 하지만 2015년에서는 SNS와 관련된 유의한 변수를 찾아볼 수 없는데, 1년 동안 이용자들이 SNS를 통한 개인 정보에 대해 많이 무뎌지고 둔해졌다는 것을 짐작할 수 있다. 2016년은 2015년과 비슷한 결과를 보여준다. 즉, 가장 유의한 변수로 스마트폰 이용시간과 와이파이존 이용 시간으로 그룹이 나누어졌고, 정보 콘텐츠 이용시간이 다음으로 유의한 변수로 새로이 등장했다. 또한, 과거 2개년에 비해 프라이버시를 우려하는 의미 있는 그룹들이 다양해졌다. 스마트폰 사용 시간이 비교적 많고, 정보 콘텐츠는 적게 쓰지만 유선 전화 연결 이용 시간이 긴 사람들 집단이 프라이버시 침해에 대해 걱정하는데, 스마트폰과 사무실 전화를 많이 사용하는 직장인들 집단이라고 예측하였다. 스마트폰을 비교적 적게 사용하고 와이파이존, 정보 콘텐츠, 그리고 문서 작업 프로그램 사용 시간이 긴 집단 또한 프라이버시에 대한 걱정이 크고, 이 집단은 카페에서 와이파이를 이용하여 노트북으로 문서 작업프로그램과 인터넷을 많이 쓰는 프리랜서라고 짐작된다. 다른 직군의 사람들에 비해 직장인들과 프리랜서들이 와이파이나 스마트폰을 이용하면서 개인 프라이버시 침해에 대해 비교적 우려를 많이 하고 있음을 볼 수 있다.

2017년에는 다른 3개년에서 볼 수 없었던 인터넷 동호회/카페/클럽 활동이 가장 유의하게 나타나는 것을 볼 수 있다. 본인이 관심이 있는 분야에 대해 인터넷에서 검색해서 카페를 찾아 가입하

는 것이 예전의 카페 접근 방법이었다면, 최근 들어 카카오톡이나 다른 메신저 창에도 카페 인기 게시글이 뜨면서 카페 활동을 하지 않았던 사람들의 카페 접근이나 가입이 용이해졌다. 그 추세가 반영되어 인터넷 동호회/카페/클럽 활동이 예전보다 활발해졌고, 마음에 들거나 공감하는 게시글을 스크랩 하는 활동 빈도가 가장 유의하게 나타난 것으로 생각된다. 이전에는 SNS 사용시간 자체가 유의한 변수로 나타났다면, 좀 더 적극적인 SNS 활동을 의미하는 SNS 정보 공유 활동 빈도가 유의한 변수로 새로이 나타났다.

Table 3. Most significant factors using decision trees by year

Year	2014	2015	2016	2017
Accuracy	65.7%	71.6%	70.5%	60.0%
Most significant variables	• SNS using time • Music listen time • Wi-fi zone time • Sleeping time • Radio broadcast / music channel program(real-time) listening time • Call time	• Wi-fi zone time • Smartphone time • Sleeping time • Movie/video watching time • Self-function/save file using time • Document work program using time	• Smartphone using time • Wi-fi zone using time • Information contents using time • Text messaging time • Document work program using time	• The frequency of clipping postings in Internet cafe/ club/society • Wi-fi zone time • Non-ground-wave TV broadcast real-time watch • The frequency of sharing information in SNS(3 months) • Phone bill • Smartphone time

5. 성별과 핸드폰 브랜드별 LASSO회귀 모형을 통한 의사결정나무 분석

위의 의사결정 나무 분석 결과 외에 본 연구는 성별과 핸드폰 브랜드별로 개인정보 유출에 대한 우려의 정도를 비교해보기 위한 분석을 진행하였다. 앞의 분석과 마찬가지로 143개의 모든 변수들을 가지고 LASSO 회귀를 진행한 뒤, 유의한 변수들만을 가지고 의사결정 나무 분석을 했다.

5.1. 성별

Table 4의 결과를 보면, 2014년 남자의 경우 스마트폰 사용과 노트북 PC 사용이 비교적 많은 사람들이 프라이버시 침해에 대해 많이 우려했고, 스마트폰 사용을 비교적 적게 한 집단 중에서는 SNS 사용 시간이 길고 지상파 TV 프로그램 시청 시간이 비교적 적은 사람들이 침해 우려도가 높았다. 여자의 경우 SNS 사용 시간이 많을수록, 그리고 SNS 사용 시간이 비교적 적더라도 음악/음원 청취 시간이 많으면 프라이버시 침해에 대해 더 많이 우려하는 것을 확인할 수 있다. 또한 상대적으로 고학력자일 경우, 프라이버시 침해 우려도가 높았으며 남성에 비해 음악 관련 시간이 프라이버시 침해에 영향을 미친다는 차이점을 발견할 수 있었다.

2015년에서는 남자는 와이파이 사용시간과 스마트폰 사용시간, 그리고 문서 작업 프로그램 시간이 많을수록 개인 프라이버시 우려도가 높음을, 그리고 여자는 남자와 달리 와이파이존 이용 시간이 비교적 적고 인터넷 동호회 회원이며, 애플리케이션 관련 지출 금액이 많을수록 우려를 많이 하는 것을 확인할 수 있다. 2016년에서 남자의 경우 정보 콘텐츠 사용 시간과 스마트폰 사용 시간이 많고 일반 전화기 사용시간 또한 많은 집단이 우려도가 가장 높은 것으로 나타났다. 스마트폰과 사무실 전화기를 많이 사용하고 정보 콘텐츠 또한 많이 이용하는 직장인 그룹이 개인 프라이버시에 대해 우려를 많이 한다고 예측할 수 있다.

2017년 남자와 여자 집단의 결과(Figure 1, 2)를 살펴보면, 앞서 살펴본 2017년 전체 자료의 결과에서 가장 유의했던 카페/클럽 게시글 스크랩 활동 빈도 변수가 여자의 영향을 많이 받았음을 예

Table 4. Most significant factors using decision trees by sex

Year		2014	2015	2016	2017
Men	Accuracy	66.0%	73.0%	72.1%	61.9%
	Most significant variables	• Smartphone using time • Normal laptop using time • SNS using time • Car audio using time	• Wi-fi zone using time • Smartphone using time • Document work program using time	• Smartphone using time • Information contents using time • Normal telephone using time	• Wi-fi zone time • The frequency of recommendation in SNS during the three months. • The frequency of writing comments in Internet cafe
Women	Accuracy	66.3%	70.7%	71.5%	58.5%
	Most significant variables	• SNS using time • Music listen time • Highly educated -> the degree of privacy invasion concern ↑	• Wi-fi zone time • Members of Internet cafe/club • The amount spent in applications	• Smartphone using time	• The frequency of Clipping postings in Internet cafe • Individual monthly income • Sleeping time

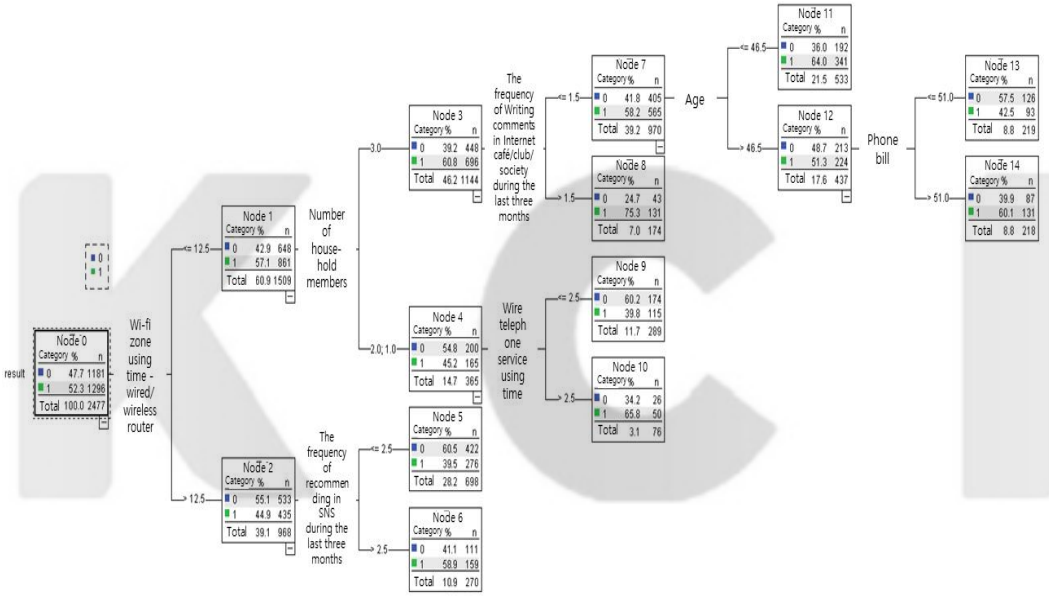

Figure 1. Decision trees for men using significant variables selected by LASSO in 2017

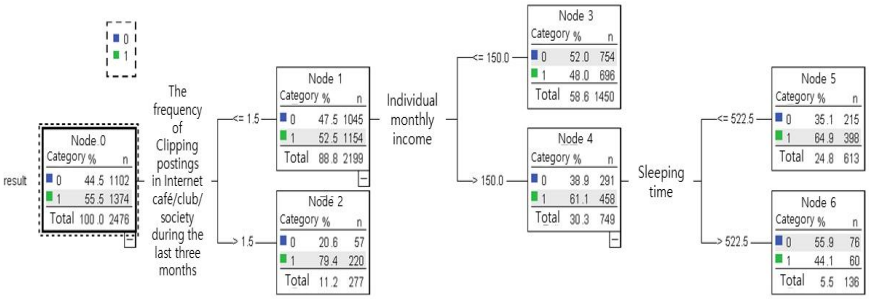

Figure 2. Decision trees for women using significant variables selected by LASSO in 2017

측할 수 있다. 남자에게서 가장 유의하게 나온 변수들을 보아, 2017년에 카페와 SNS에서 적극적인 활동이 두드려졌고 그것이 프라이버시 침해에 영향을 주었다는 것을 다시 한 번 확인할 수 있다. 또한 프라이버시 침해 우려도가 가장 크게 나온 노드9를 보면 와이파이 사용 시간은 비교적 적고, 가구원 수는 많으며 인터넷 동호회/카페/클럽 댓글 달기 활동 빈도가 높은 사람들 집단이다. 노드 11과 14를 비교해보면 두 집단 모두 와이파이 사용 시간은 비교적 적고 가구원 수는 많으며 댓글 달기 활동 빈도가 적음을 볼 수 있는데, 46세 이하인 집단과 46세 이상이라면 휴대폰 이용금액이 많은 집단이 프라이버시 침해에 대해 더 많이 우려한다. 여자의 경우, 지난 3개월 동안 인터넷 동호회/카페/클럽 게시글 스크랩 활동 빈도가 높으면 높을수록, 월평균 소득이 150만원보다 많고 수면시간이 적을수록 프라이버시 침해 우려도가 높다.

5.2. 핸드폰 브랜드별(삼성, LG, 애플)

Table 5의 결과를 보면, 2015년 삼성에서 와이파이존과 스마트폰을 적게 사용하는 사람들은 영화/동영상 시청이 많을수록 프라이버시 침해 우려가 높게 나타났는데, 이 집단은 스마트폰을 많이 사용하지 않지만 예능이나 유튜브 시청을 많이 하는 학생들 집단으로 추측할 수 있다. LG 사용자 집단에서는 와이파이존 사용 시간이 많고 남자인 사람들 중 수면시간이 적은 사람들이 프라이버시 우려를 많이 한다. 이 집단은 스마트폰뿐만 아니라 노트북을 사용할 때 와이파이를 사용하고 수면시간이 다른 나이 대에 비교적 적은 대학생 남자 집단으로 추측된다. 애플 핸드폰을 사용하는 사람들은 삼성과 LG 회사 핸드폰 사용자 집단에 비해 확연히 다른 결과를 보였다. 정보 콘텐츠를 많이 사용할수록, 정보 콘텐츠를 비교적 적게 사용한다면 문서 작업 프로그램을 많이 할수록 우려 정도가 높다. 애플 핸드폰 사용자의 경우, 삼성과 LG와 달리 매니아 층이 단단하여 애플 상품 자체에 열광하는 열성 소비자들일 가능성이 높다. 이들에게 애플의 제품은 신뢰와 애정의 대상이며 실제로 그들은 아이폰, 아이패드, 맥북 등 모든 전자기기를 애플에서 구매한다. 애플 제품끼리 연

Table 5. Most significant factors using decision trees by smartphone brand

Year		2014	2015	2016	2017
Samsung	Accuracy	66.3%	72.7%	72.3%	61.3%
	Most significant variables	• Wi-fi zone time • Members of Internet cafe • SNS time • Music listen time	• Wi-fi zone using time • Smartphone using time	• Smartphone time • Information contents time • Mobile wireless Internet time	• The frequency of clipping postings in Internet cafe • Wi-fi zone time • Cable Internet time
LG	Accuracy	66.6%	71.6%	70.5%	60.9%
	Most significant variables	• SNS using time • Sleeping time	• Wi-fi zone time • Smartphone time • sex(male,degree privacy invasion concern ↑) • Sleeping time	• Information contents time • Smartphone time • Phone bill • Age	• The frequency of Clipping postings in Internet cafe • Desktop time • Sleeping time
Apple	Accuracy	63.0%	72.1%	64.4%	65.6%
	Most significant variables	• SNS using time	• Information contents time • Document work program time	• Information contents using time • Whether to use Cloud services	• Amount spent Game • The frequency of sharing information in SNS (3 months) • The highest level of education • Smartphone time

동되는 특별한 애플리케이션을 쓰거나 그 기기 자체 기능을 많이 사용하는데, 문서 작업도 이런 애플리케이션과 기능을 이용하는 것으로 추측된다.

2016년 결과를 보면, 앞서 와이파이존 이용 시간이 유의한 변수로 자주 출현하였는데, 이동 통신 무선 인터넷이 유의한 변수로 새로이 등장한 것을 보면 2016년에 LTE 보급과 데이터 사용에 따른 요금제가 잘 정비되었다는 것을 증명한다고 볼 수 있다. 애플 사용자 집단에서 정보 콘텐츠 사용 시간과 함께 클라우드 서비스 사용 여부가 유의한 변수로 선택되었는데, 이 새로운 변수의 등장 또한 의미가 있다고 보인다. 애플의 경우, 핸드폰 자체의 아이클라우드 서비스가 잘 되어 있어 사용하는 이가 많은 만큼, 그것에 대한 개인 프라이버시 보안에 철저히 힘써야 할 것이다.

2017년 결과에서 먼저 삼성과 LG 핸드폰 사용자 집단 의사결정나무(Figure 3, 4)를 보면 2017년 전체 데이터 의사결정나무의 추세를 잘 담고 있다고 판단된다. 삼성 핸드폰 사용자 집단에서는 게시글 스크랩 활동 빈도가 비교적 낮더라도 와이파이존 이용 시간과 유선 인터넷 이용 시간이 많을수록 프라이버시에 대해 더 많이 우려하였고, LG 핸드폰 사용자 집단에서는 게시글 스크랩 활동 빈도가 비교적 낮더라도 데스크톱 PC 사용 시간이 많을수록 프라이버시 침해 우려 정도가 높았다. 애플은 삼성과 LG와 달리 게임 지출 금액이 약 5천 원 이상이면 거의 100%확률로 프라이버시 침

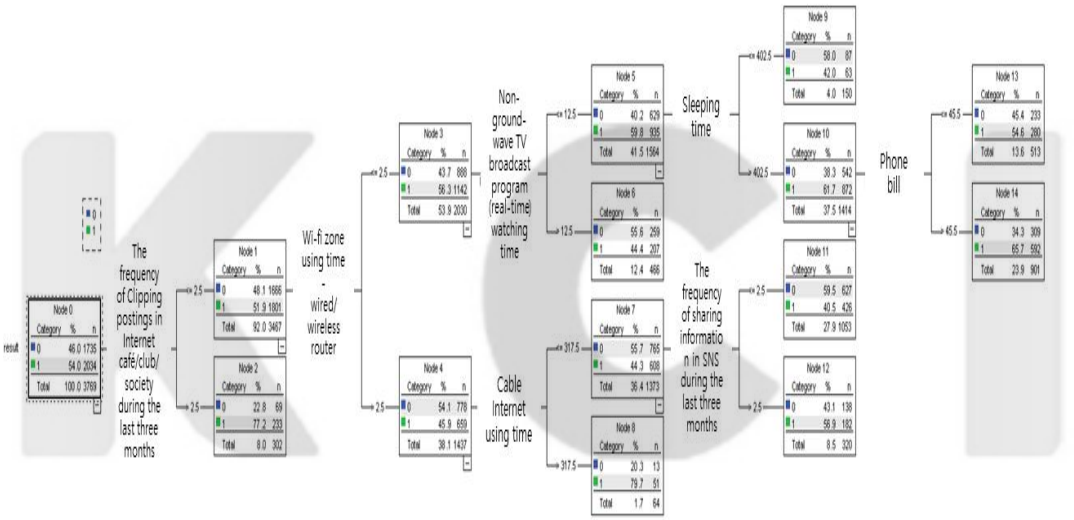

Figure 3. Decision trees for Samsung phone user using significant variables selected by LASSO in 2017

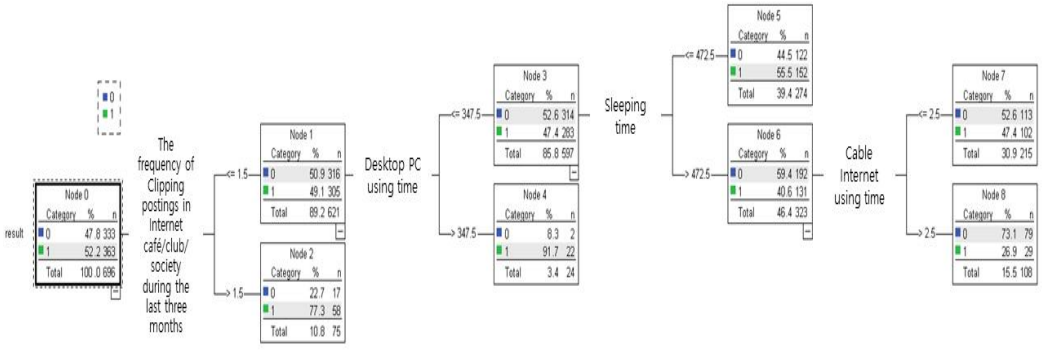

Figure 4. Decision trees for LG phone user using significant variables selected by LASSO in 2017

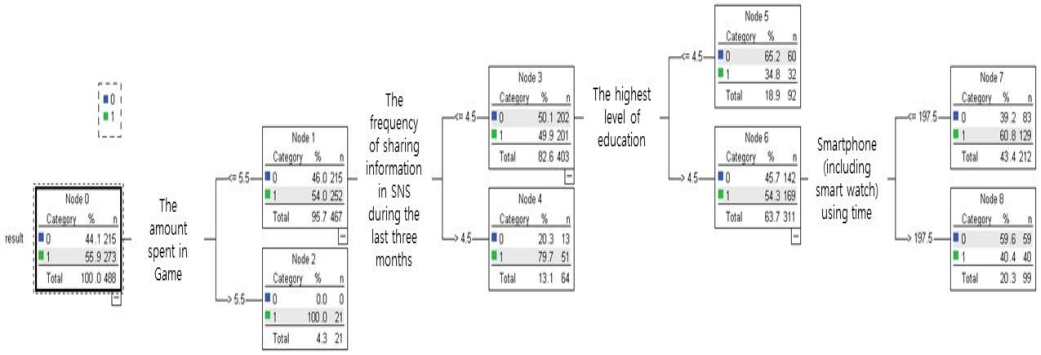

Figure 5. Decision trees for Apple phone user using significant variables selected by LASSO in 2017

해에 대해 우려한다는 결과가 나왔다(Figure 5). 게임 지출 금액은 앞서 전체 데이터나 다른 집단들에서 유의하게 나온 적이 없는 변수로, 애플 핸드폰 사용자들은 게임 지출 금액에 영향을 많이 받는다는 것을 알 수 있다. 그 다음으로 게임 지출 금액이 적더라도 SNS 정보 공유 활동 빈도가 높거나, SNS 정보 공유 활동 빈도가 낮더라도 학력이 높고 스마트폰 사용 시간이 비교적 낮은 사람들이 프라이버시 침해에 대해 크게 우려하는 것으로 나타났다.

6. 결론

미디어 매체를 통한 정보 전달 속도와 양이 급격하게 증가하면서 개인정보에 대한 우려도 급속히 증가하고 있다. 본 연구는 설명 변수의 수가 많을 경우에 회귀 모형의 구축의 방법으로 LASSO 회귀분석 방법을 이용하여 2014년-2017년 기간 동안의 연도별 미디어 패널자료를 분석하여 프라이버시 침해 우려에 미치는 영향을 살펴보았다.

연도별 LASSO 회귀분석에서 유의한 변수들로 분석한 의사결정나무 결과로는 전체적으로 와이파이 사용량이 많으며, 수면시간이 적은 이용자가 프라이버시 침해 우려가 더 높게 나타났다. 이는 공공 와이파이를 사용할 경우 개인정보를 누출 위험이 있다는 뉴스기사와 연관이 있다고 판단할 수 있으며 인터넷 이용자가 비사용자보다 수면시간이 적다는 연구결과(Lee, 2005)와 관계가 있다. 또한 2017년에 들어 카페 활동과 SNS 활동에 프라이버시 침해 우려 정도가 영향을 받는 것을 확인할 수 있었으며, 카페와 SNS 이용자가 프라이버시 침해에 더 우려하는 것으로 확인했다.

성별을 고려했을 때, 남성의 경우 노트북, PC 사용과 문서작업 시간이 많을수록 우려 정도가 높았으며 여성과 비교해서는 카오디오 사용 시간이 사생활 침해 우려 정도에 영향을 미치는 변수로 나타났다. 여성의 경우, SNS 사용 시간과 인터넷 카페, 클럽 사용이 많을수록 우려를 많이 했으며, 남성에 비해 음악과 관련된 시간이 프라이버시 침해 영향을 미치는 것으로 확인하였다. 또한 2015년 이후부터는 남성과 여성 모두 와이파이 이용시간이 가장 유의한 변수로 나타났다. 핸드폰 브랜드별로 보았을 때, 운영체제가 비슷한 삼성과 LG 사용자는 와이파이 이용과 큰 연관이 있었으며, 2016년부터 이동 통신 무선 인터넷이 안정화되었음을 엿볼 수 있었다. 따라서 와이파이 이용, 그리고 이동 통신 무선 인터넷의 안정화에 따른 무선 인터넷 사용에 있어서 사용자들의 우려를 줄일 수 있도록 접속 보안을 철저히 해야 할 것이며 개인 정보와 인터넷 사용 기록 등이 어떻게 쓰이는지 확인할 수 있도록 해야 할 것으로 보인다. 애플 사용자는 정보 콘텐츠의 이용, 문서 작업 프로그램 시간, 클라우드 서비스 사용이 높을수록 프라이버시 침해를 우려하는 것으로 나타났다. 이것은 삼성, LG와 달리 애플 서비스의 고유한 특징을 반영한 것으로 보이므로 애플의 문서 프로그램,

클라우드 서비스에 대한 철저한 프라이버시 보호에 힘써야 할 것으로 보인다.

References

Cheon, S. Y. (2018). Factor analysis of media panel on mobile phone average revenue per user by telecommunication companies, *Journal of the Korean Data Analysis Society,* 20(4), 1843-1852. (in Korean).

Choi, D. S., Kim, S. H., Cho, J. M., Jin, S. H., Cho, H. S. (2013). Personal information exposure on social network service, *Journal of the Korea Institute of Information Security and Cryptology*, 23(5), 977-983. (in Korean).

Hastie, T., Tibshirani, R., Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed., Springer, New York.

Jin, S. C., Kim, I. K. (2011). A study on the psychological anxiety on private information leakage to likelihood of internet users, *The Korea Institute of Electronic Communication Sciences,* 6(5), 731-737. (in Korean).

Kim, Y. S., An, H. K. (2010). Analysis of factors related to suicidal ideation in elementary school students using decision tree analysis, *Journal of the Korean Data Analysis Society,* 12(3B), 1379-1399. (in Korean).

Kim, B. S., Sim, S. Y., Lee, S. Y. (2012). Understanding the effects of privacy concern and network externality on SNS continuance intention, *Journal of the Korean Data Analysis Society,* 14(1B), 465-476. (in Korean).

Kwon, Y. R. (2010). The comparative analysis of predictors of suicidal ideation on middle school students using decision tree and logistic regression, *Journal of the Korean Data Analysis Society*, 12(6B), 3103-3115. (in Korean).

Lee, D. H. (2017). Analysis of social network services' users by Korea media panel survey, *Journal of the Korean Data Analysis Society*, 19(3B), 1363-1378. (in Korean).

Lee, J. H. (2005). Internet, traditional media, and time-use pattern: proposal of a time reallocation hypothesis, *Korean Society For Journalism And Communication Studies,* 49(2), 224-254. (in Korean).

Lee, S. H., Jo, H. J. (2012). Variable selection in decision tree for count data, *Journal of the Korean Data Analysis Society,* 14(1B), 101-116. (in Korean).

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society, Series B,* 58, 267-288.

Analysis of Privacy Invasion Concern Using LASSO Method*

Jieun Eom[1], *Seungwhan Jeon*[2], *Sooyoung Cheon*[3]

Abstract

As the use of various media channels increase rapidly, personal information must be filled to use various services and media contents. However, there are many problems in leakage of personal information. This study analyzes various aspects of the degree of privacy invasion concern using media panel data from 2014 to 2017. When model includes many variables, multi-collinearity and overfitting problems generally arise. This paper uses LASSO to solve these problems. For privacy invasion concern, we selected variables by LASSO and then classified the characteristics of users by decision tree analysis. The results show that SNS activity, smartphone or Wi-Fi zone usage time, and frequency of internet cafe activities have the greatest impact on personal privacy invasion. Also, men are strongly influenced by car audio, document work, and number of household members, and women are more likely to be affected by average monthly income, sleeping time, amount spent in applications, and music listening time. Samsung, LG and Apple show their own characteristics, although Samsung and LG have similar trends by year. Especially, Apple has a distinctly different trend from others.

Keywords : LASSO regression, decision tree, media panel, multi-collinearity, over-fitting.

*This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1B03933876).

[1]Graduate Student, Department of Applied Statistics, College of Public Policy, Korea University, 2511 Sejong-ro, Sejong-city, 30019, Korea. E-mail : eomji9689@korea.ac.kr

[2]Reseach Professor, Institute of Public Policy, Korea University, 2511 Sejong-ro, Sejong-city, 30019, Korea. E-mail : bighumanities@korea.ac.kr

[3](Corresponding Author) Associate Professor, Big Data Science, Division of Economics and Statistics, College of Public Policy, Korea University, 2511 Sejong-ro, Sejong-city, 30019, Korea.
E-mail : scheon@korea.ac.kr

[Received 20 January 2019; Revised 10 February 2019, 17 February 2019; Accepted 20 February 2019]