2015 춘계공동학술대회 KORMS/KIIE/ESK/KSIE/KSS
일시: 2015.4.8(수)~4.11(토) / 장소: 라마다프라자 제주호텔

빅데이터 분석기법을 통한 N-스크린 서비스 활용에 영향을 미치는 요인 분석

오유리, 김도훈
경희대학교 경영학과
{yuri562, dyohaan}@khu.ac.kr

Abstract

스마트 디바이스의 확산으로 인해 정보통신 서비스의 다양성과 양이 크게 증가하고 있다. 이에 따라 사용자의 니즈에 따른 디바이스의 선택의 폭도 넓어지고 디바이스 간 상호연동도 확장 되었다. 이러한 현상은 사용자의 콘텐츠 소비방식에 영향을 줄 것인데, 그 중심에 있는 것 중의 하나가 N-스크린 서비스이다. 본 연구에서는 KISDI에서 제공한 미디어패널 데이터를 바탕으로, N-스크린 서비스 이용에 영향을 미치는 핵심 요인들을 찾는다. 미디어 패널 데이터로부터 본 연구의 목적에서 사용가능한 7,128의 레코드를 추출하였으며, 종속변수를 포함하여 총 54개의 변수로 구성된 빅데이터를 구축하였다. 정확한 예측력을 위해 학습용 집합(train data set)과 비교용 집합(test data set)을 나누어 분석하여 비교한다. 그 전에 데이터를 분류하기 위해 K-fold Cross validation과 Bootstrap을 이용한다. 로지스틱 회귀분석(Logistic Regression, LR)을 적용한 결과, LR에서는 23개의 속성이 유의한 것으로 분석되었다.

따라서 향후 N-스크린 가입자 특성이나 행태적 요인 등을 규명하거나 이에 관련된 마케팅 전략이나 정책 개발에서 이들 요인을 적극적으로 고려할 필요가 있을 것이다.

1. 서론

스마트 폰의 보급과 더불어, 스마트 미디어의 다양성과 양도 증가하고 있다. 사용자의 다양한 니즈를 반영할 수 있는 디바이스 선택의 폭도 넓어졌고, 디바이스 연동도 확장되었다. 이러한 현상은 사용자의 콘텐츠 소비 방식에도 영향을 주고, 또 다시 스마트 미디어의 확산을 촉진하는 자기강화(self-reinforcing) 사이클을 창출한다. 단말기와 서비스 간의 경계가 옅어지면서 사용자의 스마트 미디어의 선택권과 니즈는 더욱 다양해지고 있으며, 미디어 사업자들은 네트워크 효과(network effect)를 기반으로 한 차별화된 비즈니스 모델 개발에 더 많은 관심을 보이게 된다.

또한 Chang(2014)에서는, 이러한 스마트폰의 보급이 ICT(Information Communication Technologies) 시장에도 변화를 일으키고 있다고 하였다. ICT 시장은 스마트폰을 기반으로 하여 변화하고 있으며, PC와 TV를 기반으로 한 홈서비스 시장은 통합된다고 하였다. 이러한 변화를 'multi-screen Service Convergence' 또는 'ICT Convergence' 라고 한다. 앞으로는 스마트폰을 이용한 multi-screen Service Convergence 또는 ICT Convergence를 많이 보게 될 것이다.

방송을 비롯한 스트리밍 동영상(streaming video)이나 주문형 비디오(on-demand video) 등을 볼 수 있는 대표적인 모바일 어플리케이션에는 스마트 DMB가 있다. 모바일로 실시간 방송을 볼 수 있다는 장점을 가지지만, 애플의 iOS와 같은 특정 운영체제에서는 사용에 제한이 따른다는 문제가 있다. 반대로 애플의 iOS기반 클라우드 서비스는 주문형 비디오와 개인 및 단체가 소장하는 비디오를 언제 어디서나 보거나 업로드 할 수 있다는 장점이 있지만, 공중파와 같은 실시간 방송은 매우 제한적이다. 사용자들은 다양한 콘텐츠를 방식과 디바이스에 구애받지 않고 소비하는 것을 원한다.

이러한 니즈를 충족시키기 위해 개발된 서비스가 'N-스크린 서비스'이다. 'N-스크린'에서 'N'은 여러 개를 의미하는데, 보다 구체적으로 언제, 어디서나, 사용자가 원하는 디바이스로 원하는 콘텐츠를 유형을 자유롭고 끊임없이(seamless) 이용하는 것을 말한다. 실시간 방송도 볼 수 있으며, 유/무선 접속 환경과 관계없이 스마트 DMB보다 다양한 콘텐츠를 즐길 수 있다. 다양한 무선통신 네트워크에 적합한 형태로 서비스를 이용할 수 있기 때문에 서비스를 사용할 수 있는 지역에 대한 제한이 적다는 것도 장점이다. 예를 들어, 지하철에서 이동을 하면서 보던 드라마를 집에 도착하여 TV로 이어서 보는 것이 그 예이다.

N-스크린 서비스 개념은 오래 전에 등장하였지만, 디바이스의 사양이나 이를 보유하고 있는 사용자 규모, 콘텐츠 보급률, 네트워크 환경 등의 문제로 서비스 이용에 제약이 있었다. 그러나 스마트 디바이스의 보급이 급증하고 사용자의 니즈가 다양화되며 네트워크 환경이 개선되면서 N-스크린 서비스는 앞으로 크게 발전할 것으로 전망된다. 정보통신정책연구원(KISDI)이 발표한 KISDI STAT 리포트(정용찬, 2014)에서는 2013년 방송프로그램 시청이 가능한 디바이스와 그 이용 패턴을 2011년 자료와 비교·분석하였다. 이에 따르면([표 1]도 참조), TV나 데스크탑 PC와 같은 비이동형 디바이스의 보유는 감소하는 반면에, 개인형 디바이스인 스마트폰, 태블릿 등은 증가추세를 보이고 있다고 한다.

[표 1] 가구단위별 디바이스 보유 환경

구 분		2011년	2013년
비이동형 / 가족형	TV	97.5	96.9
	데스크탑 PC	69.5	62.8
	컴퓨터	75.1	74.0
이동형 / 개인형	노트북/넷북	20.2	29.3
	태블릿	1.9	8.0
	스마트폰	38.2	73.0

2011년 3,413 가구; 2013년 3,434 가구; 단위: %, KISDI-STAT Report(정용찬, 2014)

또한 TV만 보유한 가구는 19.0%로 2개 이상의 디바이스를 보유한 가구(80.4%)에 비해 크게 적었다. 황주성(2012)에서는 보유한 미디어 디바이스의 수가 많을수록 서비스를 연계하여 사용하는 경향이 증가한다고 한다. 김윤화(2014)는 N-스크린 이용 행태 및 추이에 대한 조사·연구에서, N-스크린 이용률이 2011년에는 15.9%, 2013년에는 18.4%로 2.5% 증가했다고 한다. 큰 증가폭을 보이지는 않지만, N-스크린 서비스가 활용 가능한 디바이스인 스마트폰(73.0%), 태플릿(8.0%), 노트북/넷북(29.3%)의 활용도가 커지고 있기 때문에 향후 N-스크린 서비스에 대한 전망치를 높게 보고 있다.

N-스크린 서비스에 대한 낙관적 기대감에도 불구하고, 서비스의 활성화에 미치는 요인을 찾는 많은 연구들이 작은 수의 표본이나 편향된 표본에 대한 설문조사 등에 의존하고 있어서 크게 신뢰하기 어렵다. 본 연구에서는 KISDI에서 제공하는 방대한 양의 '미디어패널 데이터'를 활용하기 때문에, 표본의 규모나 대표성에 있어서 신뢰할 만한 결과를 얻을 수 있을 것으로 기대된다. 특히, N-스크린 서비스 이용에 영향을 미치는 사용자 특성이나 행태적 요인 등을 규명하기 위하여 로지스틱 회귀분석과 같은 빅데이터 방법론을 적용한다는 점에서 의의를 가진다. 또한 로지스틱 회귀분석의 타당성을 위해 K-fold Cross validation과 Bootstrap을 이용하여 분류를 한 뒤 차별화된 결과와 시사점을 얻고자 한다.

본 논문의 구성은 다음과 같다. 먼저, 다음 절에서는 N-스크린에 대한 최근의 문헌을 중심으로 서비스 현황과 연구 동향을 파악한다. 3절에서는 K-fold Cross validation과 Bootstrap에 대해 설명하고, 4절에서는 로지스틱 회귀분석 모형들을 소개한 뒤, 2013년 미디어패널 데이터를 바탕으로 N-스크린 서비스에 미치는 요인들을 분석한다. 5절에서는 연구모형의 예측력을 비교하고 평가한다. 마지막으로 6절에서는 본 연구의 한계와 앞으로의 진행방향에 대해 소개하면서 논문을 마무리한다.

2. N-스크린 서비스: 정의와 문헌연구

스마트 디바이스의 확산은 사용자들의 ICT 서비스에 대한 이용 편의성을 높였고, 이로 인하여 스마트 디바이스를 대상으로 한 콘텐츠 개발과 사용도 증가하고 있다. 특히 최근 몇 년에 걸쳐서 스마트폰(73.0%)과 태블릿 PC(8.0%), 노트북/넷북(29.3%)의 증가는 괄목

할 만하다(김윤화, 2014). 강종구(2014)는 이러한 현상에 대해 TV 중심의 영상 콘텐츠 소비가 스마트 디바이스 중심으로 바뀌는 것이라고 평한다. TV 외의 스마트 디바이스 사용이 TV 이용시간을 감소시키는 것이 아니라, 다른 디바이스와의 연계와 연동을 통해 오히려 동영상 콘텐츠 사용이 증가한다고 것이다. 예를 들어, ComScre(2014)에서는 2013년 통계자료에서는 2010년에 비해 컴퓨터와 스마트폰, 테블릿의 사용율이 각각 증가했으며 전체적으로도 크게 증가했다고 했다. 또한 소비자는 자신에게 편한 디바이스를 사용할 것이라고 했다. 이는 앞으로의 소비자 활동이 여러 디바이스의 연동을 활발하게 할 것이라고 볼 수 있다. 이러한 이유에서 여러 디바이스 간 호환과 연동이 동영상 콘텐츠의 수요를 증가시킬 것으로 전망된다. 이러한 현상은 콘텐츠 개발자 및 공급자뿐만 아니라 각종 매체와 플랫폼사업자들로 하여금 N-스크린 서비스에 대한 관심을 더 높일 것이다.

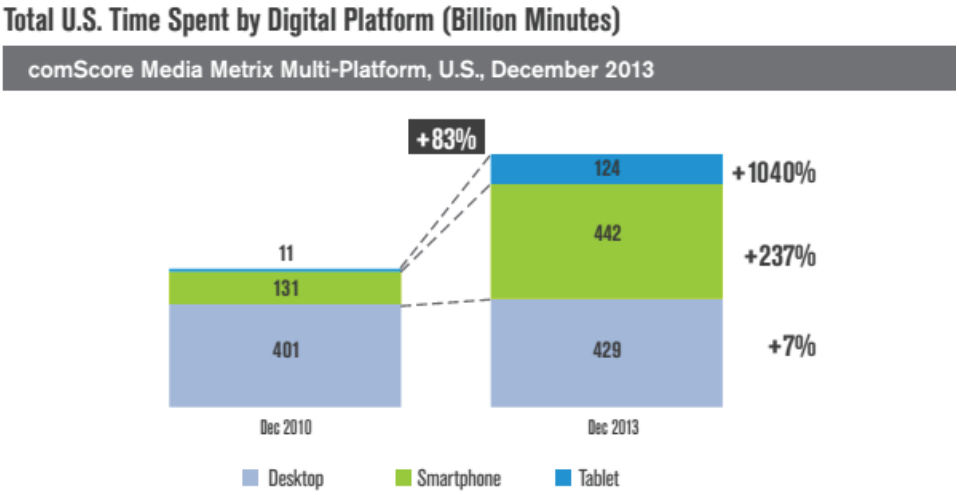

그림 1 미국 Digital platform으로 Media 총 사용 시간(1억 분)
(단위: 분) ComScore Single Source Multi-Platform Study (KPCB, 2013)

N-스크린 서비스에 대한 정의는 다양하다. 먼저, 오경수(2012)는, N-스크린 서비스란 "인터넷 기반으로 공통의 소프트웨어 플랫폼이 탑재된 TV, PC, 모바일 등 여러 스크린을 통해 다양한 콘텐츠를 제공하는 서비스"로 정의한다. 이헌주 외(2012)에서는 하나의 콘텐츠를 다양한 콘텐츠 이용환경에 최적화시켜 끊임없이(seamless) 제공하는 서비스라고 하였다. 조성수 외(2013), 이종근(2011), 최진원(2011) 등에서는 "다양한 단말기에서 같은 콘텐츠를 이용할 수 있는 서비스로서, 특정할 수 없는 다수의 단말기에서 이용이 가능한 서비스"로 N-스크린을 정의한다. 김동우(2013)에서는 N-스크린을 "여러 개(n)의 스크린을 통해 언제, 어디서나 콘텐츠를 공유하고 실행할 수 있고, TV, PC, 태블릿, 스마트폰 등 다양한 디바이스에서 콘텐츠를 끊임없이 이용 가능한 서비스"라고 하였다. 이광훈(2014)에서는 "하나의 콘텐츠를 다양한 네트워크를 통하여 여러 가지 단말기를 이용하는 서비스"로 정의하였다.

N-스크린 서비스에 대한 다양한 정의에서 공통적으로 발견되는 N-스크린 서비스의 요소는, 인터넷 사용이 가능한 디바이스들을 이용하여 언제, 어디서나 끊임없이 사용할 수 있다는 점이다. 따라서 본 연구에서는 이러한 정의를 종합하여 "동영상 미디어 서비스를 언제 어디서나 끊임없이 사용할 수 있는 서비스"로 N-스크린서비스를 정의한다.

스마트 디바이스의 확산으로 인해 N-스크린 서비스를 활용할 수 있는 환경이 조성되

면서 이에 대한 연구도 증가하는 추세이다. Kim(2014)에서는 N-스크린 서비스가 OSMU(One Source Multi-Use)에서 ASMD(Adaptive Source Multi-Device) 메소드로 발전하였다고 했다. 이는, 다양한 종류의 디바이스가 증가하면서 N-스크린 서비스와 밀접히 관련되는 디바이스 간 연동 및 연계가 가능해졌기 때문이다. 이러한 새로운 미디어 이용 행태에 대한 연구가 활발히 진행되고 있다. 김동우, 이영주(2013)는 N-스크린 서비스 만족에 영향을 주는 요인들에 대해 연구하였다. N-스크린 유료서비스를 이용한 경험이 있는 사용자들을 대상으로, 서비스 이용 행태(디바이스 간 교차 이용, 보유 디바이스당 서비스 지불액, 이용시간 등), 콘텐츠 특성(채널 다양성, VOD 다양성 등), 기능적 요인(시스템 품질, 결제시스템 등), 비용 등을 요인변수로 사용하였다. 이러한 요인들이 이용자 만족도에 어떠한 영향을 미치고 이용 만족도가 다시 서비스를 지속할 의사에 어떠한 영향을 미치는지 알아보기 위하여 계층적 다중회귀분석(hierarchical multiple regression)을 실시하였다.

N-스크린과 같이, 기존의 서비스를 대체・보완하는 새로운 기술-서비스에 대해서는 Davis가 제안한 기술수용모형(Technology Acceptance Model, 이하 TAM)을 적용하는 연구가 많다. TAM은 사용의 편의성(ease-of-use)과 유용성(effectiveness)에 미치는 요인들을 규명하고, 편의성과 유용성에 의해 기술-서비스가 수용되고 확산된다는 프레임워크를 전제로 한다. 조성수 외(2013)는 TAM을 바탕으로 N-스크린 서비스의 이용 특성(이용시간, 실시간 채널 이용정도, 디바이스 개수 등)이 디바이스 간 교차 이용, 연계 및 동시 이용, 단순 동시 이용, 이용자 인터페이스 등에 영향을 주고, 인지된 유용성과 편의성에 어떠한 영향을 미치는지 분석하였다. 양명자(2013)도 TAM을 바탕으로, N-스크린 서비스의 구매의도에 관한 연구를 하였다. 미디어 선택과 관련된 요인(서비스의 기능적 특성, 미디어 레퍼토리 이용자 특성 등)을 이용하여 확장된 미디어 선택 모형을 제안하였으며, 각 특성이 갖는 영향력을 계층적 다중회귀분석으로 분석하였다. Shin(2013)에서도 TAM과 IDT(Innovation Diffusion Theory)를 통합하여 N-스크린 서비스의 수용에 대해 분석하였다. IDT는 1995년에 Rogers에 의해 제안되었다. 소비자에 의해 얼마나 새로운 기술과 의사소통하고, 평가하고, 채택하고, 거절되고, 다시 평가되는가하는 정보를 제공한다.

박현정(2013)은 N-스크린 서비스 이용 유/무를 중심으로, 서비스 수용에 소비자 특성, 라이프 스타일, 기존 미디어 활용 행태가 어떤 영향을 미치는지 검증하였다. 또한 사용확산모형을 이용하여 N-스크린의 이용자를 유형화 하였다. 즉, N-스크린의 이용자와 비이용자의 차이를 분석한 뒤, N-스크린 서비스 가입자만을 대상으로 이 서비스를 이용한 콘텐츠, 미디어 활용 행동을 분석하고 미디어 이용 유형을 세분화 하였다. 세분화된 이용자 유형별로 온라인 활동 참여, 라이프 스타일, 미디어 레퍼토리 등에서 차이가 있는지도 분석하였다.

임소혜, 이영주(2013)에서는 N-스크린 서비스 사용자의 이용 동기와 불만족 요인을 조사하고 이를 지속적인 사용의사와 연결시키고 있다. 이용 동기를 요인별로 추출하기 위해서 오블리민 회전(Oblimin rotation)을 통한 주성분분석(Principle Component Analysis, 이하 PCA)을 적용하였다. 그 결과, 매체 활용성, 사회성, 휴대성, 콘텐츠 속성, 습관성의 5개의 요인을 구성하였다. 마찬가지로 오블리민 회전과 PCA를 통해 불만족 요인으로 고객서비스 불만족, 콘텐츠 불만족의 2개의 요인을 구성하였다. 마지막으로 이들 동기 요인과 불만족 요인이 N-스크린 서비스를 지속적으로 사용하려는 의사에 미치는 영향을 계층적 회귀분석을 이용하여 분석하였다.

이광훈(2014)에서는 KISDI에서 제공하는 한국미디어패널의 2011년과 2012년 다이어리 자료를 이용하여 프로빗(Probit) 분석을 하였다. 여기서 사용한 다이어리 자료는 3일간 개인별 미디어 이용에 대한 시간 다이어리 자료이다. 매 15분 단위로 개인이 사용한 미디어를 어떤 매체를, 어떤 방식의 연결을 통해 어떤 행위를 하였는가를 기록하도록 한 자료이다. 이러한 자료를 토대로 실제로 N-스크린 서비스 이용유무를 종속변수로 두고, 주로 개인 특성(나이, 성별, 학력, 뉴스 선호, 드라마 선호 등)이나 미디어 사용유무(스마트폰 보유, 홈 TV 보유, 노트북 보유 등)를 독립변수로 두어 프로빗 분석을 하였다.

한윤, 이상우(2012)에서는 N스크린 서비스인 CJ헬로비전의 티빙(tving) 가입자를 대상으로 총 유효설문 132개를 설문하여 티빙의 이용이 홈TV의 이용을 대체 또는 보완 관계에 있는지 실증 분석을 실시하였다. 본 연구에서는 도구변수를 사용하여 2단계 회귀분석, 즉

2SLS(Two-Stage Least squares)를 사용하였다.

Kim(2014)에서는 SEM(Structural Equation Model) 중 PLS 분석 툴(Tool)을 사용하여 소비자 행동의 관점에서 N-스크린 서비스 사용에 미치는 영향을 분석하였다. 본 연구에서는 개인 경향(자기 효능감, 혁명을 받아들이는 정도, 정보를 받아들이는 정도) , 기술 경향(보안 위험, 통제 편리), 사회 영향(집단 성향, 준거집단 영향)이 N-스크린 서비스 사용에 수반하고 영향을 미치는지 분석하였다.

이상의 문헌연구에서 보는 바와 같이, N-스크린 서비스에 대한 계량적 분석을 위해 계층적 다중회귀분석(김동우, 이영주, 2013; 임소혜, 이영주, 2013)이나 TAM(조성수 외, 2013; 양명자, 2013; Shin, 2013)을 도입하는 경우가 많다. 계층적 다중회귀분석은 연구자의 판단에 의해 변수를 하나씩 추가함으로써 인과적 설명력을 제공할 수 있다는 장점을 가지고 있지만, 베타값이나 결정계수(R^2)의 설명력이 낮다는 단점도 보인다. TAM의 경우 프레임워크의 강건성(robustness)에 대해서는 비교적 좋다고 알려져 있지만, 구조적 단순성과 미시적인 특성에 초점을 맞추기 때문에 새로운 기술-서비스가 받아들여지는 다양한 요인과 거시적 환경을 도외시한다는 비판도 존재한다.

또한 위 문헌을 포함한 많은 연구에서 표본의 규모와 대표성에 대해서는 신뢰성 측면에서 크게 작은 문제점들을 안고 있다. 조성수 외(2013)에서는 483명을 웹서베이를 통해 조사하였으며, 양명자(2013) 및 임소혜, 이영주(2013)에서는 스마트폰 이용자들로만 구성된 700명 및 558명의 표본을 사용하여 설문조사와 웹서베이를 실시하였다.

최근에 쓰여진 이광훈(2014)에서는 로지스틱 회귀분석과 유사한 프로빗 분석을 사용하였다. 이 분석으로 얻어진 모형의 차이점은 로지스틱 분포의 양쪽 끝이 더 두껍기 때문에 로지스틱 모형에서 확률이 0이나 1로 수렴하는 속도가 프로빗 모형보다 느리다는 것이다. 그러나 수학적 편의성으로 인하여 프로빗 모형보다는 로짓 모형을 더 많이 이용한다.

문헌연구의 결과를 정리하면 N-스크린 서비스에 관한 많은 연구들이 소규모 표본을 이용하고 (학문적 엄밀성을 담보하기 어려운) 웹서베이에 의존한다는 문제점들을 노출하고 있다. 또한 방법론적 측면에서도 새로운 기술-서비스가 수용・확산되는 총체적 요인들을 포괄하지 못한다는 한계를 보인다.[1] 본 연구에서는 이러한 문제점을 보완하기 위하여 KISDI에서 제공한 빅데이터인 미디어패널 데이터를 사용한다. 이 미디어패널 데이터는 4년간의 조사를 거친 1만 명 이상으로 구성된 표본을 제공한다. 이러한 표본은 보다 정확하고 객관적인 분석을 가증하게 할 것이다. 또한 이 표본은 사회과학 기준으로 볼 때 빅데이터에 해당하는 규모이기 때문에, 위에서 소개한 연구들이 적용한 방식과는 잘 맞지 않는다. 즉, 빅데이터에 적합한 분석방법론을 적용해야 할 것이다. 본 연구에서는 N-스크린 서비스를 가입자와 비가입자를 대상으로 이들의 의사결정에 미친 핵심 변수를 찾고자 로지스틱 회귀분석(logistic regression)과 의사결정나무(decision tree)를 확장한 CART(Categorical Analysis Regression Tree)를 적용한다.(김수영;2006)

본 연구가 취하는 전체적인 접근과 맥락은 박현정(2013) 및 김수진, 김보영(2013)과 유사하다. 이들의 연구에서도 로지스틱 회귀분석을 사용하여 이진값(binary value)을 갖는 종속변수에 대해 특화된 회귀분석을 통해 비교적 정확한 판단력과 예측력을 제공하였다. 특히 박현정(2013)의 경우, N-스크린 서비스 이용자와 비이용자 간의 차이를 알아보기 위하여 t-검정과 χ^2-검정을 사용하였으며, 보다 상세하게 영향력 있는 변수를 찾기 위해 로지스틱 회귀분석을 적용하였다. 이를 바탕으로, 사용확산모형을 적용하여 N-스크린 서비스의 이용자 유형을 구성하였다. 이용자를 유형화하는 작업은 유의미한 결과를 보여주었지만, 로지스틱 회귀분석의 한계에 따라 유형을 해석하는데 어려움이 따랐다. 이는 선형회귀분석과는 달리, 로지스틱 회귀분석을 통해 선정된 핵심 변수가 종속변수의 특정 카테고리에 미치는 영향력을 평가・비교하는 것이 직관적일 수 없기 때문이다. 이에 뒷받침하는 연구로 김완섭(2012)이 있다. 학생들의 교육성과의 영향력을 분석하기 위해 로지스틱 회귀분석을 실시하였다. 하지만 회귀분석만으로는 다양한 요

1) 이는 총체적으로 모든 변수가 미치는 영향을 고려해야 한다는 것이 아니다. 과학적 방법론은 핵심 변수를 추출하고 이를 통해 대상을 분석하는 것이기 때문에 총체적 변수 모두를 한 모형에서 동시에 고려하는 것은 불가능할 뿐만 아니라 타당하지도 않다. 그렇지만, 핵심 변수를 추출하기 위해서 최소한 초기 단계에서는 총체적이고 포괄적인 접근을 시도해야 한다는 의미이다.

인들 간의 계층적인 관계를 규명하기 어렵다는 단점을 보완하기 위하여 의사결정나무를 사용한다.

이에 반해 CART는 의사결정나무를 통한 분석의 틀은 유지한 채 정확성을 높인 것이므로, 선형회귀분석보다 오히려 핵심 변수를 찾아내고 그 의미를 해석하는 것이 용이하다는 장점을 가진다. 이는 특히 종속변수가 이진값을 가지는 분류형 나무(classification tree)로 CART를 이용할 때 더 큰 장점을 발휘한다. 또한 분석 결과를 시각적으로 표현하여 제시하기 때문에 직관적 이해가 쉽고 해석과 예측을 위해 활용하기도 편리하다.

최근 발표된 Ngo et. al(2014)에서는 예측력의 정확성을 알아보기 위하여 로지스틱 회귀분석과 CART, CHIAD(Chi-Squared Automatic Interaction Detection), 다층 퍼셉트론 신경회로망(multi-layer perceptron neural network ;MLPNN)을 분석하였다. 그 결과 학습용 집합에서 로지스틱 회귀분석보다 다층 퍼셉트론 신경회로망이 좋은 예측력의 결과를 도출하였다.

본 연구에서는 또한 로지스틱 회귀분석 방법론을 적용한 뒤 두 모형의 예측력을 비교・평가한다. 이를 위해 미디어패널 데이터를 모형에 적합 시키기 위한 학습용 집합(training data set)과 모형의 정확성과 예측력을 평가하기 위한 비교용 집합(test data set)으로 사전에 나누어 놓는다. 학습용 집합과 비교형 집합을 보다 정확하게 분류하기 위해 K-fold-Cross validation과 Bootstrap을 이용한다. 그 뒤 로지스틱 회귀분석 결과를 종합하여 N-스크린 서비스에 영향을 미치는 보다 정확한 모형을 개발하기 위한 기반을 제공하고, 서비스 확산에 관한 시사점을 얻고자 한다.

3. 타당성 확인 방법

본 연구에서는 모형을 측정하기 위해 학습용 자료를 이용한다. 이러한 학습용 자료의 예측치가 너무 유연하게(flexible) 적합화되어, 구조적인 특성뿐만 아니라 우연적인 특성까지 모두 반영하게 되어 문제가 발생할 수 있다. 이러한 문제를 과적합(Over-fitting) 현상이라고 부른다. 또한, 모형을 측정하는데 관여하지 않는 독립적인 데이터 집합인 비교용 자료를 이용한다.

과적합 현상의 문제를 막기 위해서는 현대 통계학에서의 방법 중에 리샘플링(Resampling)이 있다. 적합화된 모델에서 추가적인 정보를 얻기 위하여, 반복하여 계층화된(Stratified) 샘플로 학습용 자료를 만든다.

주어진 데이터를 리샘플링하는 방식을 기반으로 분류의 정확도를 평가하는 방법에는 일반적으로 K-fold Cross validation과 Bootstrap이 사용된다. 이러한 방법에 대해 조금 더 설명하려고 한다.

3.1. K-Fold Cross Validation(CV)

계층화된 샘플을 추출하기 위하여 많은 샘플링 방법이 고안되었다. 그 중에 반복적으로 임의 추출(random sampling)을 하였을 때, 데이터 수에 제한 받지 않고 반복하여 여러 번 사용할 수 있다는 장점이 있다. 하지만 반복하였을 때 사용되지 않는 데이터가 생겨, 일부분만을 사용할 수 있다는 단점이 있다. 이러한 단점을 보완한 방법이 모든 데이터를 골고루 비교용 집합으로 사용하는 'K-Fold Cross Validation'이다.

K-Fold Cross Validation의 접근은 먼저 무작위로(randomly) 대략 동일한 크기로 K개의 집단으로 임의 분할한다. 그 중에 하나를 비교용 집합으로 사용하고, 나머지 K-1개를 학습용 집합에 사용하는 과정을 순차적으로 진행된다. 이러한 과정은 비교용 집합과 학습용 집합이 역할을 번갈아 가며 실시하게 된다. 구체적으로 살펴보면, K개의 집단을 G_k라고 할 때, G_1을 비교용 집합으로 사용하면 $G_2, G_3 \cdots G_k$의 집단을 학습용 집합으로 사용한다. 다음에 G_2를 비교용 집합으로 하면, $G_1, G_3 \cdots G_k$의 집단이 학습용 집합으로 검증을 실시하게 된다. K번의 검증으로 얻은 결과는 K개의 추정오차를 평균한 값으로 추정성능을 최종평가하게 된다.

K-Fold Cross Validation의 측정은 아래와 같이 계산된다.

$$CV = \frac{1}{K}\sum_{i=1}^{k} MSE_i$$

3. 2. Bootstrap

Bootstrap은 균등한 기회로 무작위 복원 표본추출하고, 추정치들의 분산, 평균 등을 이용하는 비모수적 통계 방법이다. 한 개의 데이터가 추출될 때마다 다시 재추출될 가능성이 동일하고 학습용 집합에 추가될 가능성도 같다. 즉, 샘플링은 대체(replacement)로 이루어지는데 같은 관찰치가 한번이상 Bootstrap 데이터에서 사용되어진다는 것이다. 주어진 추정치나 통계적 학습방법에 관련된 불확실성을 정량화하는데 사용 될 수 있다. 또한 선형 회귀분석의 적합으로부터 계수의 표준 오차를 추정하기 위해 사용할 수도 있다. 선형 모델 적합에서 회귀분석과 관련된 가변성을 평가하기 위해 Bootstrap의 사용되어진다.

Bootstrap은 표본을 모집단으로 간주하며, B개의 Bootstrap 표본을 만든다. 각 Bootstrap에 대해서 예측모형을 구축한다. 구축된 B개의 예측 모형을 결합하여 최종 모형을 만든다. 일반적으로 사용되는 Bootstrap 방법으로는 '0.632 Bootstrap'이 있다. Jiawei Han, Micheline Kamber(2007)에서는 무작위 표본추출을 할 때 훈련용 집합에 포함되지 않은 표본들은 비교용 집합에 형성되게 된다. 이러한 과정이 반복된다고 할 때, 평균적으로 초기 표본의 63.2%가 학습용 집합에 포함되고, 36.8%가 비교용 집합에 형성된다. 이러한 통계적 이유로 '0.632 Bootstrap'이라고 불린다. Bootstrap 샘플수를 B, 예측모형을 M_i, k번 추출과정을 거친다고 할 때 모형의 전제 정확도의 식은 아래와 같다.

$$\text{Acc(M)} = \sum_{i=1}^{k}(0.632 \times Acc(M_i)_{test} + 0.368 \times Acc(M_i)_{train})$$

여기서 ACC(M_i)는 I번째 Bootstrap 표본이 초기 데이터에 적용될 때 얻어진 정확도를 의미한다.

4. 데이터와 연구 모형

4.1 데이터 소개

KISDI에서 주관하여 시행된 미디어패널 조사는, 2005년 인구주택 총조사 결과를 표본추출(sampling)의 베이스로 삼고 층화 2단계 확률비례계통추출법을 통해 충분한 수의 표본을 추출하므로 대표성과 객관성이 확보된다. 본 논문에서는 2013년의 미디어패널 데이터를 사용하였는데, 전체 패널 수는 10,464명에 이르며, 개인용/가정용/미디어 다이어리 등으로 나누어 조사하였다. 개인용 설문의 조사항목도 122개 문항에 이르며, 미디어 다이어리로부터 얻을 수 있는 변수도 2,689개에 이른다. 본 연구에서는 개인용 조사항목 전체와 미디어 다이어리 조사항목 중 일부를 사용하였다. 패널(개인)의 인구통계학적 구성은 [표 2]와 같다.

항목	구분	빈도(%)	항목	구분	빈도(%)
성별	여자	4809(46%)	소득	소득 없음	4903(46.9%)
	남자	5655(54%)		100만원 미만	1451(13.8%)
연령대	10대 미만	441(4.2%)		100-300만원 미만	2915(27.9%)
	10대	1400(13.4%)		300-500만원 미만	979(9.3%)
	20대	887(8.5%)		500만원 이상	215(2.1)
	30대	1498(14.3%)		무응답	1(0.0%)
	40대	2073(19.8%)	최종 학력	미취학	52(0.5%)
	50대 이상	4165(39.8%)		초졸 이하	2497(23.9%)
직업	임금 금로자	3323(31.8%)		중졸 이하	1269(12.1%)
	고용주	107(1.0%)		고졸 이하	3511(33.6%)
	단독 자영업자	1266(12.1%)		대졸 이하	2986(28.5%)
	무급 가족 종사자	304(2.9%)		대학원 재학 이상	149(1.4%)
	무응답	5464(52.2%)	합계	10,464 명	

표 4 표본의 인구통계학적 특성

이론적으로 조사항목들이 하나의 변수로 산정될 수 있지만, 질문 내용에 따라서는 묶어서 처리해야 할 항목들도 존재하며, 순수하게 인구통계학적인 데이터(성별 등)나 표본 식별 번호 등도 존재하기 때문에, 변수 후보로 사용될 조사항목들을 그대로 사용할 수는 없으며 정제(data cleaning)와 가공이 필요가 있다. 특히 조사항목들을 묶는 과정에서 상당한 수의 항목들을 하나의 변수로 가공할 수 있었다. 예를 들어, 본 연구의 목적에 비추어 볼 때, 사용하는 디바이스를 나열하는 복수의 조사항목들은 디바이스 개수라는 하나의 변수로 요약하여 다루는 것이 보다 적합하다. 또한 N-스크린 서비스를 선택하는 의사결정과는 직접적인 관계가 매우 약해 보이는 조사항목들도 제거하였다.

이러한 사전처리를 거쳐서 53개의 변수를 산출하였다([부록 1] 참조). 그런데 변수별로 결측치(무응답 및 모름 등의 답변이 이에 해당함)가 존재하여 이에 대한 처리가 필요했다. 결측치를 처리하는 대표적인 방식은 해당 변수의 평균으로 대체하는 것이지만, 본 논문에서는 53개 변수에 대하여 하나 이상의 결측치가 포함된 응답자의 수가 많지는 않기 때문에 해당 레코드를 삭제하였다. 또한 명백하게 N-스크린 서비스를 사용할 수 없는 환경에 놓인 응답자 레코드도 삭제하여, 전체 10,464개의 레코드 중 3,336개가 제거된 7,128의 표본을 대상으로 연구를 진행하였다.

결과적으로, 본 연구에서는 사용한 기초 데이터는 1개의 종속변수와 52개의 독립변수로 구성된다([부록 2] 참조).

4.2 연구 모형과 접근법

4.2.1 로지스틱 회귀분석

2013년 미디어패널 자료를 가공하여 얻은 기초 자료([부록 2])에 대하여 로지스틱 회귀분석과 CART를 적용하여 분석하고 서로 비교한다. 이들 방법론을 구현하기 위해 R 소프트웨어(버전 3.0.2)를 사용하였다. 먼저, 로지스틱 회귀분석(이하 LR로 부름)은 종속변수가 이진값(binary value)을 갖는 범주형 자료(categorical data)에 회귀분석방법을 응용한 것이다. LR은 두 유형으로 분리되는 모집단에서 각 유형을 결정하는데 핵심적인 역할을 하는 속성(독립변수)을 찾기 위해 많이 사용되는 방법론으로, '지도형 학습(supervised learning)'의 빅데이터 분석에서 가장 많이 사용되는 방법 중의 하나이다. 본 연구에서는 기초 자료의 7,128명의 모집단을 N-스크린 서비스를 한 번 이상 이용한 경험이 있는 사람과(종속변수 = 1)과 이용하지 않는 사람(종속변

수 = 0)으로 나누고, 이들의 서비스 선택에 미치는 영향력을 LR을 적용하여 분석하였다.

지도형 학습을 위해 먼저, 기초 자료를 5,128개의 레코드를 가지는 학습용 집합(training data set)과 2,000개의 레코드를 가지는 비교용 집합(test data set)으로 나누었다. 두 자료는 기초 자료로부터 임의로 생성되었다(threshold = 약 0.28). 두 자료군은 레코드만 다를 뿐, 변수는 모두 동일하다(독립변수 47개 및 종속변수 1개). 본 절에서 학습용 집합에 대해 LR을 적용하여 가장 적합한 모형을 구축한 뒤, 다음 절에서 비교용 집합을 가지고 예측의 정확성을 평가할 것이다.

전체 52개의 독립변수를 가지고 학습용 집합에 대해 LR을 적용한 결과 상당한 개수의 변수들이 유의하지 않은 것으로 나타났다. 가장 유의하지 않은 것으로 보이는 몇 개의 변수를 제거하고 LR을 적용하는 과정을 반복함과 동시에 독립변수들 사이의 상관관계를 지속적으로 검토하면서 최종적으로 모든 변수들이 유의한 LR 모형으로 축약(reduction)하였다. 그 결과 다음의 수식과 표와 같은 LR 모형을 얻을 수 있었다. 아래에서 β_j (j=0,…,14)는 LR 모형에 의한 최우추정량(maximum likelihood estimator)이며, Y는 종속변수, X_j (j=1,…,14)는 독립변수이다(변수에 대한 설명은 [표 3]을 참조).

$$Y = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_{14} X_{14}}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_{14} X_{14}}} \qquad [식\ 1]$$

변수	변수 이름	설명
Y	종속변수	N-스크린 사용 경험 유무
X_1	휴대폰 구분	사용하고 있는 휴대폰의 구분
X_2	MP3 기능	휴대폰의 MP3 재생 기능 가능한지의 여부
X_3	휴대폰 부담자	휴대폰 요금을 부담하는 사람
X_4	유료 어플 다운	유료 애플리케이션 다운로드 경험여부
X_5	영화 횟수	극장에서의 영화 관람 회수
X_6	영화 지출	극장에서의 영화 관람 지출 금액
X_7	공연 횟수	공연 관람 회수
X_8	TV 방송 장르	좋아하는 TV 방송 프로그램의 장르
X_9	신문 구독	신문 구독 여부
X_{10}	이메일 계정	이메일 계정 사용 여부
X_{11}	블로그 사용/운영	블로그 사용/운영 여부
X_{12}	SNS 사용	SNS 사용 여부
X_{13}	클라우드	클라우드 서비스 사용 여부
X_{14}	취미 활동 글 읽기	지난 3개월 동안 인터넷 동호회/카페/클럽 글 읽기 활동 빈도
X_{15}	온라인 추천, 평점	지난 3개월 동안 온라인 추천, 평점 주기 기능 활동 빈도

X_{16}	지식 서비스 질문	지난 3개월 동안 인터넷 지식 서비스 질문 글 쓰기 활동 빈도
X_{17}	미디어 이용 능력	미디어 이용할 수 있는 능력
X_{18}	TV 방송 채널	좋아하는 TV 방송 채널(지상파, 비지상파, 종합 편성)
X_{19}	TV 이용	가정용 TV 이용 하루 평균 빈도
X_{20}	스마트폰 이용	스마트폰 이용 하루 평균 빈도
X_{21}	지역	거주하고 있는 지역
X_{22}	N 스크린 지출	영상 콘텐츠 N스크린 전용 서비스 월 평균 지출 요금
X_{23}	방송 통신 지출	방송통신(신문/동영상/tv/음악) 월 평균 지출 요금

표 5 LR 모형에서 사용된 변수들

[식 1]의 LR 모형의 적합도와 계수별 추정량 등은 [표 4]와 같다.

	Estimate	Std. Error	Z-value	Pr(>|Z|)	
Intercept	-6.557043	0.630105	-10.406	< 2e-16	***
휴대폰 구분	0.220510	0.063296	3.484	0.000494	***
MP3 기능	-0.610644	0.189877	-3.216	0.001300	**
휴대폰 부담자	0.094073	0.029037	3.240	0.001196	**
유료 어플 다운	0.475716	0.156295	3.044	0.002337	**
영화 횟수	-0.397260	0.175187	-2.268	0.023352	*
영화 지출	0.125288	0.018834	6.652	2.88e-11	***
공연 횟수	0.727208	0.401633	1.811	0.070198	.
TV 방송 장르	0.032350	0.009353	3.459	0.000543	***
신문 구독	-0.367908	0.104432	-3.523	0.000427	***
이메일 계정	0.456507	0.162205	2.814	0.004887	**
블로그 사용/운영	0.525554	0.143176	3.671	0.000242	***
SNS 사용	0.270340	0.095009	2.845	0.004435	**
클라우드	3.012620	0.180537	16.687	< 2e-16	***
취미 활동 글 읽기	0.152664	0.034546	4.419	9.91e-06	***
온라인 추천, 평점	0.167531	0.056480	2.966	0.003015	**

지식 서비스 질문	0.214375	0.079216	2.706	0.006806	**
미디어 이용 능력	0.638455	0.110196	5.794	6.88e-09	***
TV 방송 채널	-0.159634	0.048354	-3.301	0.000962	***
TV 이용	-0.012178	0.005999	-2.030	0.042345	*
스마트폰 이용	0.015385	0.008476	1.815	0.069520	.
지역	0.021343	0.009884	2.159	0.030830	*
N 스크린 지출	1.679599	0.377739	4.446	8.73e-06	***
방송 통신 지출	0.264392	0.040457	6.535	6.35e-11	***

표 6 LR 모형의 추정치(β_j (j=0,…,14))

[식 1]에 [표 4]의 LR 결과를 도입하여 정리하면 [식 2]와 같다.

$$\ln\left(\frac{p}{1-p}\right) = -6.557043 + 0.220510X_1 - 0.610644X_2 + 0.094073X_3 + 0.475716X_4 - 0.397260X_5 - 0.125288X_6 + 0.727208X_7 + 0.032350X_8 - 0.367908X_9 + 0.456507X_{10} + 0.525554X_{11} - 0.270340X_{12} + 3.012620X_{13} + 0.152664X_{14} + 0.167531X_{15} + 0.214375X_{16} + 0.638455X_{17} - 0.159634X_{18} - 0.012178X_{19} + 0.015385X_{20} + 0.021343 X_{21} + 1.679599 X_{22} + 0.264392X_{23} \quad [식\ 2]$$

[식 2]에서 p는 N-스크린 서비스에 가입하여 사용할 확률을 나타낸다(즉, $p \equiv P(Y=1)$). 적합된 LR 모형에 따를 때, N-스크린 서비스 이용에 가장 큰 영향을 미치는 요인은 '클라우드'과 'N 스크린 지출 금액' (각각 X_7 및 X_{22})이다. 또한 공연 횟수(X_1)와 미디어 이용 능력(X_{17}) 등도 영향력을 가지는 변수이다. 반면에 MP3 기능(X_2)는 N-스크린 서비스 가입을 주저하게 만드는 요인으로 해석된다. 이는 유료 정보미디어 서비스와 클라우드 등을 많이 사용하고 인터넷 활동 및 문화 활동이 활발한 사람들이 N-스크린 서비스에 가입할 가능성이 높다는 것을 의미한다. 반면에 휴대폰에 MP3 기능이 가능한지의 여부는 N-스크린 이용에 영향을 주지 않는 것으로 해석된다.

5. 분석 결과의 비교

학습용 집합으로부터 구축된 LR 모형의 정확성을 평가하기 위하여 [식 2]의 LR 모형에 비교용 집합을 적용하였다. 빅데이터 분석에서 어떤 모형의 예측력은 보통 Confusion Matrix(이하 CM)를 구성하여 평가한다. CM은 실제 값(종속변수의 실제 유형)과 예측된 값(모형에 의해 예측된 유형)의 차이를 비교하여 정리한 것이다. 이때 가능한 경우는 총 4가지인데, 각각 TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative)로 불린다. TP는 실제 값과 예측 값에서 모두 유형 1로 평가한 경우이며, TN는 실제 값과 예측 값 모두 유형 0으로 평가한 경우이다. 따라서 TP와 TN는 모형의 예측이 실제 데이터와 정확히 맞는 경우에 해당한다. 이에 반하여 FP는 예측은 유형 1로 평가했지

만, 실제로는 유형 0에 해당하는 경우이며, FN는 이와 반대되는 경우이다.

[표 7]는 LR 모형의 CM을 정리·비교하여 보여준다. 학습용 집합으로 모형을 구축한 뒤, 비교용 집합으로부터 예측된 값을 구한다. LR의 경우 모형이 예측한 유형 1에 속할 기준확률(reference threshold)을 0.5로 하였을 때의 결과이다(즉, $P(Y=1|x_1,\cdots,x_{14})\geq 0.5$ 이면 유형 1에 속한다고 예측한다). 예측력의 정확성은 보통 전체 데이터 크기(N)에 대한 TP+TN의 비율로 평가되는데, LR의 경우 1679/2000 = 0.8395이다.

LR(Logit)	Predicted=0	Predicted=1
Actual=0	1478	261
Actual=1	60	201

표 7 LR의 CM 비교 (기준확률=0.5)

LR의 경우 기준확률을 보다 정교하게 산출하면 예측력이 좋아질 수 있다. 이를 위해 시각적으로 비교할 수 있는 ROC(Receiver Operating Characteristic) 분석을 이용하면 유용하다. 각 모형에 대한 ROC 곡선을 계산하여 평가하면 [그림 2]와 같다.

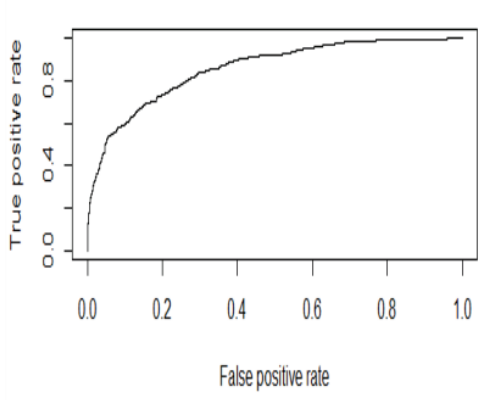

그림 2 LR의 ROC 비교

[그림 2]의 결과로부터 LR은 기준확률을 0.5~0.6 사이에서 책정하는 것이 FP와 FN의 가능성을 줄이는데 도움이 될 것으로 판단된다. 이로부터 기준확률을 0.5가 아닌 0.6로 책정하여 모형을 비교·평가한 뒤 CM을 구성하면 [표 8]과 같다.

LR(Logit)	Predicted=0	Predicted=1
Actual=0	1489	281
Actual=1	49	181

표 8 LR의 CM 비교 (기준확률=0.6)

기준확률을 0.6로 하였을 경우 LR의 예측력((TP+TN)/N)은 0.835이다. 이는 기준확률을 낮추면 일반적으로 TN은 개선되고 TP는 나빠지는데, 여기서는 TN이 개선되는 정도가 TP가 악화되는 것을 상쇄할 만큼 크지 않았기 때문이다(LR의 경우 TN은 201 → 181로 감소하였고, TP는 1478 → 1489으로 증가하였다).

LR 모형에 따를 때 N-스크린 서비스 이용에 가장 큰 영향을 미치는 요인은 '클라우드'과 'N 스크린 지출 금액' (각각 X_7 및 X_{22})이다. 또한 공연 횟수(X_1)와 미디어 이용 능력(X_{17}) 등도 영향력을 가지는 변수이다. 이로부터 N-스크린 서비스와 관련된 마케팅 및 정책적 의사결정에서 위 4가지 요인을 우선적으로 고려하는 것이 중요할 것으로 판단된다.

마지막으로, Cross validation의 결과로 랜덤으로 추출된 10개의 집단으로 분석되었다. 그 결과 내부 측정의 정확도는 85.5%, Cross-validation 측정의 정확도 85.5%로 측정되었다.

주요 속성을 기준으로 판단기준도 제공하기 때문에 향후 구체적인 마케팅 전략(소비자 프로파일 개발 등)이나 정책 개발(N-스크린 활성화를 위해서는 전반적인 방송통신 미디어 사용에 드는 비용을 낮추어야 한다는 것 등)에도 구체적인 지침을 제공할 수 있을 것으로 기대된다.

6. 결론

이상에서 KISDI 미디어패널 빅데이터를 대상으로 N-스크린 서비스에 영향을 미치는 주요 요인을, 로지스틱 회귀분석을 통해 분석하였다. 분석 결과, 서비스 확산에 영향을 미치는 주요 요인으로 클라우드이용과 N 스크린 지출 금액, 공연 횟수와 미디어 이용 능력

등을 발견할 수 있었다. 또한 예측정확도를 확인하기 위하여 Cross-validation을 이용하여 예측의 타당성을 확인하였다는 데에 의의가 있다. 조금 더 정확한 예측력을 위해서 파라미터에 대한 미세조정(fine tuning)을 통해 보다 더 나은 모형으로 발전할 수 있다. 향후 이에 대한 연구를 전개할 예정이다.

부록

	변수 이름	변수설명
1	휴대폰 구분	사용하고 있는 휴대폰의 구분
2	사진 촬영	휴대폰의 사진 촬영이 가능한지의 여부
3	동영상 촬영	휴대폰의 동영상 촬영이 가능한지의 여부
4	MP3 기능	휴대폰의 MP3 재생 기능 가능한지의 여부
5	지상 DMB 기능	휴대폰의 지상파 DMB 기능 가능한지의 여부
6	와이파이 기능	휴대폰의 와이파이 지원 가능한지의 여부
7	이동 통신사	가입한 휴대폰의 이동 통신사
8	휴대폰 이용 금액	월평균 휴대폰 이용 총 금액
9	휴대폰 할부금	월평균 휴대폰 기기 할부금
10	휴대폰 부담자	휴대폰 요금을 부담하는 사람
11	유료 어플 경험	유료 애플리케이션 다운로드 경험여부
12	유료 어플 횟수	유료 애플리케이션 다운로드 수
13	영화 횟수	극장에서의 영화 관람 회수
14	영화 지출	극장에서의 영화 관람 지출 금액
15	공연 횟수	공연 관람 회수
16	공연 지출	공연 관람 지출 금액
17	스마트 어플	자주 이용하는 스마트 기기 어플리케이션
18	TV 방송 장르	좋아하는 TV 방송 프로그램의 장르
19	신문 구독	신문 구독 여부
20	이메일 계정	이메일 계정 사용 여부
21	블로그 사용/운영	블로그 사용/운영 여부
22	SNS 사용	SNS 사용 여부
23	클라우드	클라우드 서비스 사용 여부
24	취미 활동 회원	인터넷 동호회/카페/클럽 회원 여부

25	취미 활동 운영	인터넷 동호회/카페/클럽 운영 여부
26	취미 활동 글 읽기	지난 3개월 동안 인터넷 동호회/카페/클럽 글 읽기 활동 빈도
27	취미 활동 댓글 달기	지난 3개월 동안 인터넷 동호회/카페/클럽 댓글 달기 활동 빈도
28	취미 활동 스크랩	지난 3개월 동안 인터넷 동호회/카페/클럽 게시글 스크랩 활동 빈도
29	취미 활동 글 쓰기	지난 3개월 동안 인터넷 동호회/카페/클럽 글 쓰기 활동 빈도
30	뉴스/토론 게시판 글 쓰기	지난 3개월 동안 인터넷 뉴스/토론 게시판 댓글, 글 쓰기 활동 빈도
31	뉴스/토론 게시판 스크랩	지난 3개월 동안 인터넷 뉴스/토론 게시판 게시글 스크랩 활동 빈도
32	온라인 투표	지난 3개월 동안 온라인 투표 참여 빈도
33	온라인 추천, 평점	지난 3개월 동안 온라인 추천, 평점 주기 기능 활동 빈도
34	지식 서비스 질문	지난 3개월 동안 인터넷 지식 서비스 질문 글 쓰기 활동 빈도
35	지식 서비스 답변	지난 3개월 동안 인터넷 지식 서비스 답변 글 쓰기 활동 빈도
36	유용한 정보 등록	지난 3개월 온라인 상 유용한 정보 등록 활동 빈도
37	소득	개인 월평균 소득
38	연령	연령
39	지역	거주하고 있는 지역
40	최종학력	최종 학력
41	직업	직업 유무
42	방송 통신 지출	방송통신(신문/동영상/tv/음악) 월 평균 지출 요금
43	N 스크린 지출	영상 콘텐츠 N스크린 전용 서비스 월 평균 지출 요금
44	TV 방송 채널	좋아하는 TV 방송 채널(지상파, 비지상파, 종합 편성)
45	미디어 이용 능력	미디어 이용할 수 있는 능력
46	장소 이동	사람들의 장소 이동 변경 횟수
47	TV 이용	가정용 TV 이용 하루 평균 빈도
48	데스크탑 이용	데스크탑 PC 이용 하루 평균 빈도
49	노트북 이용	일반 노트북 PC 이용 하루 평균 빈도

50	넷북 이용	넷북 이용하루 평균 빈도
51	태블릿 이용	태블릿 PC 이용 하루 평균 빈도
52	스마트폰 이용	스마트폰 이용 하루 평균 빈도
53	종속 변수	N스크린 사용 경험 유무

참고문헌

강중구 (2014), “통계로 보는 콘텐츠산업”, 제14-15호, 통권 86호, 1-11
고윤석 (2011), “CART 알고리즘 기반의 의사결정트리 기법을 이용한 규칙기반 전문가 시스템 구축방법론”, 한국전자통신학회논문지, 6권, 849-855
김남진, 지수정, 조남훈(2012), “다중겹 교차검증 기법을 이용한 증기세관 결함크기 예측을 위한 신경회로망 성능 향상”, 조명·전기설비학회논문지 26권, 9호, 73-79
김동우, 이영주 (2013), “N스크린 서비스의 이용행태, 콘텐츠, 기능, 비용이 이용 만족도와 지속이용의사에 미치는 영향에 관한 연구”, 방송공학회논문지, 18권, 749-757
김미정, 염동문, 이경은 (2013), “CART 분석을 활용한 아동학대 예측요인에 관한 연구”, 피해자학연구, 21권, 293-315
김수영(2006), ″다변량 판별분석과 로지스틱 회귀분석, 인공신경망 분석을 이용한 호텔 도산 예측″, 한국관광학회, 30(2), 53-75
김수진, 김보영 (2013), “로지스틱 회귀분석과 의사결정나무 분석을 이용한 일 대도시 주민의 우울 예측요인 비교 연구”, 한국콘텐츠학회논문지, 12권, 829-839
김완섭(2012), ″로지스틱 회귀분석과 데이터마이닝 분석을 이용한 컴퓨터 교양교육 성과의 요인에 대한 연구″, 한국교양교육학회, 6(3), 743-767
김윤화 (2010), “N 스크린 전략 및 추진 동향 분석”, 정보통신정책연구, 22권, 1-23
김윤화 (2014), “N스크린 이용행태 및 추이”, KISDI-STAT Report, 14-02호, 11-15
김종하 (2013), “N스크린 환경에서 다중미디어를 활용한 TV프로그램 이용행태 연구”, 만화애니메이션연구, 31호, 177-208
김태호 외 (2007), “CART분석을 이용한 지하철 소음모형 개발 및 특성 연구”, 한국철도학회, 10권, 480-486
김형도(2013), “불균형 신용평가 데이터의 분류 향상을 위한 균형 교차검증”, 한국정보기술학회논문지, 11권, 4호, 169-175
김형준, 하규수 (2013), “지상파 방송사의 채널 이미지와 N-스크린 서비스 운영 전략”, 디지털정책연구, 11권, 43-55
노진수, 백승현, 전상길 (2013), “데이터마이닝을 활용한 HR제도들의 상대적 중요도 평가: 제조업을 중심으로, 한국시뮬레이션학회논문지, 22권, 55-69
박용석, 박세호, 이경택 (2012), “디지털 모바일 방송 기반 N-스크린 콘텐츠 제공”, 방송공학회지 17권, 87-92
손기철, 신임희(2012), “잭 나이프 및 붓스트랩 방법을 이용한 임상자료의 회귀계수 타당성 확인”, 한국데이터정보과학회지, 23권, 4호, 643-648
손지은, 김성범 (2014), “의사결정나무 모델에서의 중요 룰 선택기법, 대한산업공학회지, 40권, 375-381
신지형 (2014), “1. 동영상 콘텐츠 소비와 디바이스”, KISDI Report, 14-07호, 1-6
양명자 (2013), “N-스크린 서비스 구매의도에 관한 연구”, 한국방송학보, 27권, 131-166
오경수 (2012), “N스크린 서비스 잠재적 수용자의 수용의도 영향요인 연구”, 한국콘텐츠학회논문지, 12권, 80-92
이광훈(2014), ″다이어리 자료를 이용한 N스크린 방송서비스 이용 행태 분석″, 정보통신정책연구, 21(3), 1-21
이청호, 신현식 (2013), “광산업 제품의 품질시험인증서비스 만족도에 관한 연구”, 산업경제연구, 26권, 1715-1737
이헌주 외 (2012), “N-스크린 서비스를 위한 주요기술 및 콘텐츠의 발전 방향”, 정보처리학회지, 19권, 9-18
임소혜, 이영주 (2013), “N스크린 서비스 이용자의 이용 동기와 불만족 요인에 관한 연구”, 한국콘텐츠학회논문지, 13권, 99-108
임준 (2011), “N 스크린 서비스 활성화 방안”, KISDI Premium Report, 11-08호, 1-20
정용찬 (2014), “1. 방송프로그램 시청 가능 기기 보유와 이용특성”, KISDI-STAT Report, 1-9

조성수, 김동우, 이영주 (2013), “멀티 디바이스의 융합 이용을 토대로 한 N스크린 서비스 이용자의 연계이용/동시이용과 인지된 유용성 및 인지된 용이성의 관계에 관한 연구”, 애니메이션연구, 9권, 55-73
최민재 (2013), “스마트 미디어 시대 지상파 방송사의 N-스크린 추진 전략 및 광고 플랫폼 전략연구”, 한국광고홍보학보, 15권, 192-222
최세경 (2010), “N스크린 시대에 TV 비즈니스의 전망과 대응 전략: 콘텐츠 유통과 소비 패러다임의 변화를 중심으로”, 방송문화연구, 22권, 7-36
하형석 (2015), “2. 멀티미디어 시대의 N스크린 이용”, KISDISTAT Reprot, 11-18
한윤, 이상우(2012), "N 스크린 서비스와 홈TV간 대체 및 보완관계에 대한 실증적 연구:국내 대표 N스크린 서비스인 티빙을 중심으로", 한국콘텐츠학회논문지, 12(5), 144-153
황주성 (2012), “멀티디바이스 환경에서 디바이스 간 연계이용”, 사이버커뮤니케이션학보, 29권, 131-171
Borra, S., & Di Ciaccio, A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Computational statistics & data analysis, 54(12), 2976-2989.
Chang, S. G. (2014). A structured scenario approach to multi-screen ecosystem forecasting in Korean communications market. Technological Forecasting and Social Change.
ComScore(2014), *U.S. Digital future in Focus 2014*
Jiawei Han, Micheline Kamber(2007), *“Data mining”*
Kim, H., You, Y., & Kim, K. (2014). Impacts of Personal, Social and Culture Factors on User Acceptance of ASMD N-Screen Cloud Contents and Services.*Int. J. Advance Soft Compu. Appl*, *6*(3).
Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).
Li, J. (2012). Applications of the Bootstrap in ROC Analysis. Communications in Statistics-Simulation and Computation, 41(6), 865-877.
Ngo, F. T., Govindu, R., & Agarwal, A. (2014). Assessing the Predictive Utility of Logistic Regression, Classification and Regression Tree, Chi-Squared Automatic Interaction Detection, and Neural Network Models in Predicting Inmate Misconduct. *American Journal of Criminal Justice*, 1-28.
Shin, D. H. (2013). N-SCREEN: How multi-screen will impact diffusion and policy?. *Information, Communication & Society*, *16*(6), 918-944.
Steyerberg, E. W., Harrell, F. E., Borsboom, G. J., Eijkemans, M. J. C., Vergouwe, Y., & Habbema, J. D. F. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of clinical epidemiology, 54(8), 774-781.