Journal of the Korean Institute of Industrial Engineers
Vol. 42, No. 2, pp. 112-121, April 2016.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2016.42.2.112

<Application Research>

데이터 사이언티스트의 역량과 빅데이터 분석성과의 PLS 경로모형분석 : Kaggle 플랫폼을 중심으로

한경진 · 조근태†

성균관대학교 기술경영학과

PLS Path Modeling to Investigate the Relations between Competencies of Data Scientist and Big Data Analysis Performance : Focused on Kaggle Platform

Gyeong Jin Han · Keuntae Cho

Management of Technology, Sungkyunkwan University

This paper focuses on competencies of data scientists and behavioral intention that affect big data analysis performance. This experiment examined nine core factors required by data scientists. In order to investigate this, we conducted a survey to gather data from 103 data scientists who participated in big data competition at Kaggle platform and used factor analysis and PLS-SEM for the analysis methods. The results show that some key competency factors have influential effect on the big data analysis performance. This study is to provide a new theoretical basis needed for relevant research by analyzing the structural relationship between the individual competencies and performance, and practically to identify the priorities of the core competencies that data scientists must have.

Keywords: Big Data, Data Scientist, Behavioral Intention, PLS-SEM, SmartPLS

1. 서 론

빅데이터 시대를 맞이하여 최근 다양한 분야에서 빅데이터 자원 확보, 빅데이터 관련 기술개발, 전문 인력 발굴 등을 통해 새로운 비즈니스가 창출되고 있다. 최근 글로벌 리서치 기관과 컨설팅 그룹 등에서 차세대 키워드로 '빅데이터'를 선정함에 따라 이를 통한 경제적 가치창출이 주목받고 있다. 특히 Fenn(2011)의 유망 기술 보고서에 따르면 빅데이터를 위한 고급분석 등의 관련기술은 현재 기술 발생단계이며 향후 2~5년 후에 성숙될 것으로 평가된다.

이와 더불어 빅데이터의 다각적 분석을 통해 조직의 전략 방향을 제시하는 기획자이자 전략가인 데이터 사이언티스트의 위상도 높아지고 있다. CNN은 2012년 유망 신규 직종으로 데이터 사이언티스트를 선정하였으며, Harvard Business Review (Davenport, 2012)도 21세기 '가장 매력적인' 직종으로 데이터 사이언티스트를 선정하였다. 이처럼 각종 매체에서 데이터 사이언티스트에 대한 언급이 쏟아지는 이유는 이들이 방대한 데이터 속에서 유의미한 결론을 도출하고 그것을 비즈니스 가치로 연결하는 사람들이기 때문이다.

이에 많은 연구자들은 데이터 사이언티스트의 역할과 역량에 대해 논의하고 있다. Patil(2012)은 기술적 숙련도, 호기심, 스토리텔링, 영리함을 갖춰야만 좋은 데이터 사이언티스트가 된다고 말한다. Rauser(2011)는 데이터 사이언티스트가 단순히 공학과 수학 전공자가 아니라 의사소통, 회의적 시각, 호기심, 성공과 행복에 대한 확신을 가진 사람이 되어야 한다고 주장한다. 이와 같이 데이터 사이언티스트는 어느 한 관점에만

† 연락저자 : 조근태 교수, 16419 경기 수원시 장안구 서부로 2066 성균관대학교 시스템경영공학과 제2공학관 26413호, Tel : 031-290-7602, Fax : 031-290-7610, E-mail : ktcho@skku.edu
2015년 7월 16일 접수; 2015년 12월 9일 수정본 접수; 2015년 12월 24일 게재 확정.

치우지지 않고 다양한 차원의 역량이 합쳐졌을 때 최적의 성과를 낸다는 의견이 대세를 이룬다. 하지만 이러한 연구는 데이터 사이언티스트가 되기 위해 갖추어야할 복합적인 역량을 제시하고 있다는 점이 특징이다. 즉, 위에서 연구자들이 제시한 각각의 변수간의 상호관계는 규명하지는 못하고 있는데 이것은 데이터 사이언티스트 연구에 대표적인 한계점이다.

이러한 한계점을 극복하고자 본 연구는 데이터 사이언티스트의 역량과 성과 간의 관계를 계량적으로 분석한 최초의 시도를 하였다. 첫째, 빅데이터 분석에 종사하는 데이터 사이언티스트의 역량에 대한 기존 연구의 고찰을 통해 각 역량의 특징을 분석하고, 이들이 빅데이터 분석성과에 어떠한 영향을 미치는지 경로분석을 통해 알아본다. 둘째, 도출된 요인과 실제 측정 지수 값으로 메트릭스를 구성하여 빅데이터 분석 성과를 증진시키기 위한 차별적 우선순위를 도출한다.

이후의 논문은 다음과 같이 구성된다. 제 2장에서는 선행연구 고찰을 통해 기존에 제시된 데이터 사이언티스트가 갖추어야 할 핵심 역량을 검토한다. 이를 토대로 제 3장에서는 본 연구에서 활용할 PLS 모형을 제시하고, 제 4장에서는 이를 적용하여 빅데이터 분석성과에 영향을 주는 데이터 사이언티스트의 역량요인을 도출한다. 마지막으로 구조방정식 분석결과에 따른 결론과 연구의 의의, 한계점, 그리고 추후 연구를 제시한다.

2. 이론적 고찰

2.1 빅데이터와 데이터 사이언티스트

최근 정보기술의 일상화가 이루어지는 스마트 시대에 도래하면서 소셜, 사물, 라이프로그 데이터 등이 결합되며 전 세계의 데이터는 기하급수적으로 증가하고 있다. 데이터의 양이 2011년 1.9제타바이트(1조 8천억 기가바이트)를 넘어섰고, 향후 5년 후 약 9배 가까이 증가하면서 제타바이트 시대의 도래를 전망하고 있다(Hollis, 2011).

빅데이터란 “기존의 컴퓨팅 기술로는 저장, 관리, 분석이 불가능할 정도로 큰 데이터의 집합 및 관련 기술과 인력(Manyika, 2011)”을 의미한다. 최근에는 빅데이터 플랫폼, 분석기법, 관련 도구까지 포괄하는 용어로 변화하고 있다. 적용분야로는 의료, 제조, 교육 및 서비스를 들 수 있으며 스마트폰의 보급, 지리 정보시스템의 발달 및 멀티미디어의 증가에 따라 더욱 다양한 분야로 확산될 것으로 예측하고 있다(Chiang *et al*., 2014). 빅데이터의 특징은 “3V를 갖는, 즉, 거대한 규모(Volume)와 다양한(Variety) 형태의 데이터를 빠른 속도(Velocity)로 처리(Laney, 2012) 하는 것”으로 설명된다. 이와 같은 데이터양의 기하급수적인 증가로 정보의 생산 주체가 개인 중심으로 변화되면서 “빅데이터 시대”를 맞이하고 있다.

글로벌 리서치 기관인 가트너의 보고서(Kart, 2013)는 차세대 키워드로 ‘빅데이터’를 설정하고 경제적 가치에 주목하고 있다. 또한 향후 정보기술을 이끌 핵심적인 4요소인 클라우드, 소셜, 모바일, 정보의 통합은 빅데이터를 기반으로 가능하다고 할 정도로 중요성이 강조되고 있다. 이에 따라 수요자가 원하는 데이터를 선별 및 재가공을 통해 전략적으로 활용하려는 경향이 대두될 전망이다. 또한 국가에서도 사회 현안 해결을 위한 기반으로 빅데이터를 활용하여 새로운 경제적 가치를 창출하려는 노력이 지속될 것이다.

일본 노무라 연구소(2012)는 빅데이터 활용을 위한 전략의 3요소를 데이터의 자원, 데이터를 가공하고 처리하는 기술, 데이터의 의미를 통찰하는 인력으로 분류하였고, 이 3가지 분야에 대한 전략 수립이 필수적이라고 주장한다. 최근 들어 한국에서도 정부 3.0을 계기로 공공정보의 개방과 공유가 활발하게 이루어지고 있으며, 빅데이터를 활용한 자원발굴과 기술개발이 이루어지고 있다. 이처럼 현재 빅데이터의 자원과 기술에 대한 발전은 지속적으로 이루어지고 있다.

이렇게 빅데이터가 세상에 소개된 후 학계에서는 다양한 방면에서 빅데이터에 대한 연구가 진행되어왔다. Wamba(2015)의 계량서지학 분석방법을 사용한 빅데이터 분야의 종단연구는 관련 논문의 시계열적인 추이를 보여준다. 이는 2011년 이후 빅데이터 관련 연구가 눈에 띄게 증가하면서 2012년에는 전체 출간된 논문의 70%를 차지할 정도로 활발한 연구가 진행되는 것을 알 수 있다. 연구 방법별 분류에서는 Review가 절반 이상을 차지하였고, 데이터분석, 실험, 설문조사, 사례연구가 그 뒤를 이었다. 하지만 설문조사의 경우, 데이터 사이언티스트가 비교적 신규 직종이기 때문에 표본 수집에 한계가 있는 것으로 나타난다. 본 논문에서는 이를 보완하기 위해 전 세계의 데이터 사이언티스트 플랫폼인 Kaggle 이용자를 대상으로 설문을 시행한다.

데이터 사이언티스트란 “빅데이터에 대한 이론적 지식과 분석 기술에 대한 숙련을 바탕으로 통찰력, 전달력, 협동 능력을 발휘할 수 있는 전문 인력”이다(Manyika, 2011). 따라서 빅데이터의 가치를 충분히 이끌어내기 위해서는 데이터 이면의 의미를 해석해 내는 인재, 즉 데이터 사이언티스트의 역할이 중요해지고 있다. 미국의 경우 2018년까지 14～19만 명의 고급 분석 인력과 150만 명의 데이터 관리자가 부족할 것으로 전망된다. 2015년까지는 빅데이터 분야에서의 약 190만 개 일자리가 창출될 것이므로 빅데이터로 인한 사회 전반적인 고용 창출의 효과가 미국 내에서만 약 600만 개에 달할 것으로 예측된다(Manyika, 2011). 즉, 빅데이터의 발전으로 관련 인력에 대한 수요가 폭발적으로 증가하며 신규 일자리 창출 효과 또한 기대되고 있다.

Vidgen(2014)은 조직이 빅데이터 기반의 연구에서 어떻게 가치를 창출하고, 어떻게 위기를 극복하는지에 대한 연구를 수행하였다. 이 연구에서는 사용데이터, 인력, 가치창출, 조직관리, 위기극복에 대한 사례연구를 진행하였다. 빅데이터를 기반으로 한 비즈니스 분석은 단순히 정보기술 분야에 국한되어서는 안 되며, 정보 역설(Thorp, 1998)을 극복해야 한다고 주장한다. 특히, 호기심 및 문제해결에 뛰어난 감각과 데이터를

다루는 테크닉을 동시에 갖추는 것을 데이터 사이언티스트의 핵심요소로 제안하였다.

Patil(2011)은 보다 구체적으로 데이터 사이언스팀을 결성할 때 개인의 역량, 팀의 조건, 데이터 기반 시설, 경쟁우위를 확보하는 전략 등에 대한 연구를 하였다. 비즈니스 네트워크 인맥 사이트인 Linkedin에서 데이터 사이언스팀을 결성한 케이스를 바탕으로 데이터 사이언티스트 인력 양성 전략을 제시한다. 특히, 데이터 사이언스팀은 독자적으로 데이터 분석을 수행하는 것이 아니라 조직 내 타부서와의 조직문화 공유하는 개개인의 협업 및 의사소통 역량을 강조한다.

한국의 경우 Cho(2013)은 빅데이터 시대를 이끌어갈 가장 중요한 요소를 데이터 과학자의 양성으로 보고 사례연구를 진행하였다. 데이터 과학자가 갖추어야 할 역량에 대한 해외사례를 소개하였고, 대학별 빅데이터 프로그램 교육과정을 제시하는 등 정책적인 방향을 제시하였다.

숙련된 데이터 사이언티스트는 단기간에 육성되기가 어렵고, 지속적인 현장 경험을 통해 전문 지식과 노하우를 축적하여 만들어지는 고급 인적자원이다. 이에 본 연구는 데이터 사이언티스트가 갖추어야할 역량 요인과 조건을 실증적으로 분석하였다.

2.2 Kaggle 플랫폼

빅데이터 분석 플랫폼인 Kaggle(www.kaggle.com)은 기업의 실제 데이터를 분석하여 그 결과에 따라 포상하는 데이터 사이언티스트 경진대회 중개 사이트다. 본 사이트에서는 실제 데이터를 활용하여 실전 경험을 제공하기 때문에 우수한 역량을 갖춘 데이터 사이언티스트들의 발굴과 지속 훈련이 가능하다. 운영방법은 기업이나 기관들이 해결하려는 문제를 Kaggle에 경쟁과제로 등록하고, Raw data를 제공해주며, 데이터와 목표, 마감기간, 보상조건 등을 제시하여 경쟁자들을 유도한다(Martinez, 2014). 주요 고객으로는 MS, GE, Facebook, NASA 등 다양한 글로벌 기업과 공공기관이 포함되고, 참여자는 현재 2015년 4월을 기준으로 약 25,000명의 데이터 사이언티스트들이 경쟁에 참여하여 목표에 달성하면 3,000달러~25만 달러의 상금과 라이선스비, 기업입사 등 다양한 혜택을 준다. 뿐만 아니라 각 참가자는 우승 횟수에 따라 랭킹이 매겨져 순위가 공개되며, 데이터 사이언티스트들의 구인이나 참여자간 네트워킹도 가능하다. 이는 인터넷과 소셜 네트워크를 통해 대중에게 문제 해결을 맡기는 전형적인 크라우드 소싱이라 할 수 있다.

3. 연구모형 및 가설의 설정

3.1 연구모형 설정

연구 모형은 <Figure 1>에서 보는 바와 같이 데이터 사이언티스트의 역량과 빅데이터 분석성과 간의 인과관계를 규명한다. 이를 위해 PLS-SEM을 이용하여 경로계수를 추정한다. 데이터 사이언티스트의 역량을 기술역량, 경영역량, 현업역량으로 구분하여 구조방정식 모형을 구성한다. 기술역량은 해킹, 시각화, 머신러닝, 통계 변수로, 경영역량은 경영분석, 문제해결 변수로, 현업역량은 의사소통, 창의성, 협동변수로 설명된다. 행위의도는 실제 빅데이터 분석을 하는 분석도구의 사용의도에 관한 변수이고, 분석성과는 데이터 사이언티스트가 실제 빅데이터 Kaggle 분석 경쟁에 참가하여 얻은 실제 점수이다.

연구방법으로는 구조방정식 모델인 PLS-SEM(Partial Lease Square Equation Modeling) 접근 방식을 따랐다. PLS를 분석 방법으로 사용한 이유는 첫째, CB-SEM을 위한 maximum likelihood 추정방식과는 달리 이론적 연구들이 부족한 상황에서 사용되는 연구방법이다(Hair *et al*., 2012). 즉, 데이터 사이언티스트의 정립된 이론들이 확고하지 않은 본 연구에 탐색적 연구로 적합한 모형이라 할 수 있다. 둘째, PLS 구조방정식은 복잡한 문제를 간단하게 시각화할 수 있는 이상적인 도구다. 다시 말해, 요인을 구성하는 독립변수들과 종속변수들 간의 상호의존성 분석결과물을 시각화하여 전략적 지침을 제공하는 데 용이한 분석이라 할 수 있다(Hair, 2013). 몇몇 연구(예 : Dino, 2015; 2012; Venkatesh, 2003)에서도 표본의 크기가 작은 연구, 이론적 기반이 명확하지 않아 불확실한 결과를 예측한 연구, 지표모델을 구체화하기 어려운 연구의 경우 PLS 분석 기법을 활용하였다. 또한 분석 도구는 다양한 기능을 오픈소스로 제공하는 SmartPLS 3 V.3.2.0(2015)을 사용하여 부분최소제곱 알고리즘과 부분 최소제곱 부트스트래핑을 활용하였다.

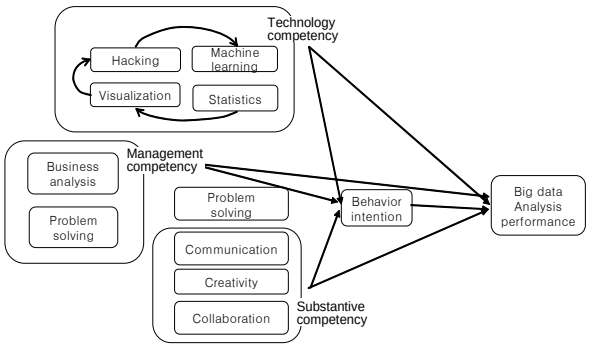

Figure 1. Research Model

3.2 가설 설정

우수한 데이터 사이언티스트는 다방면에 걸쳐 복합적이고 고도화된 지식과 능력을 갖추는 것이 필수적이다. 특히, 여러 분야에 걸친 전문성과 이를 복합적으로 활용하기 위한 도구인 수학, 통계학, 컴퓨터 공학 등 다양한 분야에 걸친 심도 있는 지식이 필수적이다. 이는 기존에 조직에서 데이터 분석을 수행하던 데이터 분석가(Data Analyst)와 차이가 있으며, 이들과 비교하여 한층 높은 수준의 전문성과 다양성을 요구받고 있다(Rahul, 2012).

Conway(2010)는 명령어의 텍스트를 조작하고 벡터화된 작업을 이해하며 알고리즘적으로 생각할 수 있는 능력인 해킹 기술을 가진 자를 우수한 데이터 사이언티스트로 제시하였다. LaValle *et al*.(2013)는 시각화기술이 빅데이터 행위의도와 빅데이터 분석성과에 영향을 미친다고 주장하였고, 그 중에서 머신러닝이 인터넷 기반의 데이터 중심 경영모델을 분석하는 빅데이터 분석에 중심역할을 한다(Davenport, 2012). Rauser (2011)는 거대 데이터 세트를 획득하고 가치를 추출해 내는 수학 및 통계능력이 데이터 사이언티스트의 필수역량이라고 주장하였다. 마지막으로 Dhar(2013)는 머신러닝은 데이터 홍수 속에서 자동적으로 결정 시스템을 개발하기 때문에 데이터 사이언티스트가 필수적으로 갖추어야할 요건으로 본다. 또한 이는 통계, 컴퓨터 사이언스, 상관관계와 인과관계의 이해와 같은 기본적인 지식이 기반이 되어야 한다고 주장한다. 이처럼 빅데이터 분석을 하는데 있어서 공학적인 기술을 갖추는 것이 매우 중요하다. 따라서 데이터사이언티스트의 기술 변수인 해킹스킬, 머신러닝, 시각화, 통계능력에 대해 다음과 같이 가설을 설정한다.

[가설 1] 기술역량은 빅데이터 행위의도 및 분석성과에 정(+)의 영향을 미칠 것이다.

[가설 1a] 해킹 능력은 빅데이터 행위의도에 정(+)의 영향을 미칠 것이다.

[가설 1b] 해킹 능력은 빅데이터 분석성과에 정(+)의 영향을 미칠 것이다.

[가설 1c] 시각화 능력은 빅데이터 행위의도에 정(+)의 영향을 미칠 것이다.

[가설 1d] 시각화 능력은 빅데이터 분석성과에 정(+)의 영향을 미칠 것이다.

[가설 1e] 머신러닝 능력은 빅데이터 행위의도에 정(+)의 영향을 미칠 것이다.

[가설 1f] 머신러닝 능력은 빅데이터 분석성과에 정(+)의 영향을 미칠 것이다.

[가설 1g] 통계 능력은 빅데이터 행위의도에 정(+)의 영향을 미칠 것이다.

[가설 1h] 통계 능력은 빅데이터 분석성과에 정(+)의 영향을 미칠 것이다.

[가설 1i] 통계 능력은 시각화 능력에 정(+)의 영향을 미칠 것이다.

[가설 1j] 시각화 능력은 해킹능력에 정(+)의 영향을 미칠 것이다.

[가설 1k] 해킹 능력은 머신러닝 능력에 정(+)의 영향을 미칠 것이다.

빅데이터 분석에 있어 데이터 기반의 의미 있는 함의를 도출하여 경영 전략을 수립하는 것이 빅데이터의 핵심 가치로 부각되면서, 복합적이고 고도화된 사고능력에 대한 요구는 더욱 증대 되고 있다. Chiang *et al*.(2012)은 예측모델의 생성과 성과 증진을 위한 경영 절차의 최적화와 같은 데이터 기반의 트렌드 분석이 빅데이터 분석성과에 영향을 미친다고 주장하였다. 경영분석은 세 가지 범주 즉, 기술범주, 예측범주, 처방범주로 나뉜다. 서술범주(descriptive)는 데이터를 사용하여 과거에 일어난 일을 알아내는 것, 예측범주(predictive)는 데이터를 사용하여 앞으로 일어날 일을 알아내는 것, 마지막으로 처방범주(prescriptive)는 데이터를 사용하여 최적의 선택을 하는 것이다. Dhar(2013)는 문제 식별과 형식화는 데이터 사이언티스트들이 갖춰야할 기술 중에 핵심적인 것이라고 주장하였다. 따라서 경영변수인 경영분석 능력과 문제해결 능력에 대해 다음과 같이 가설을 설정한다.

[가설 2] 경영역량은 빅데이터 행위의도 및 분석성과에 정(+)의 영향을 미칠 것이다.

[가설 2a] 경영분석 능력은 빅데이터 행위의도에 정(+)의 영향을 미칠 것이다.

[가설 2b] 경영분석 능력은 빅데이터 분석성과에 정(+)의 영향을 미칠 것이다.

[가설 2c] 문제해결 능력은 빅데이터 행위의도에 정(+)의 영향을 미칠 것이다.

[가설 2d] 문제해결 능력은 빅데이터 분석성과에 정(+)의 영향을 미칠 것이다.

Rauser(2011)는 이상적인 데이터 사이언티스트는 알고리즘을 설명하는 모호한 사색가가 아닌 다양한 세계와 소통하는 능력을 갖추어야 한다고 주장한다. Patil(2012)는 표면적 문제해결을 넘어서 핵심 문제를 던지고 가설검정을 통해 명백한 결론을 이끌어 내는 창의적인 사고가 빅데이터 분석을 활발하게 한다고 주장하였고, Dinter *et al*.(2014)은 빅데이터 연구에 있어 학계와 산업사이의 공동 노력이 필요하다고 주장하였다. 따라서 다음과 같이 가설을 설정한다.

[가설 3] 현업역량은 빅데이터 행위의도 및 분석성과에 정(+)의 영향을 미칠 것이다.

[가설 3a] 의사소통 능력은 빅데이터 행위의도에 정(+)의 영향을 미칠 것이다.

[가설 3b] 의사소통 능력은 빅데이터 분석성과에 정(+)의 영향을 미칠 것이다.

[가설 3c] 창의적 능력은 빅데이터 행위의도에 정(+)의 영향을 미칠 것이다.

[가설 3d] 창의적 능력은 빅데이터 분석성과에 정(+)의 영향을 미칠 것이다.

[가설 3e] 협동 능력은 빅데이터 행위의도에 정(+)의 영향을 미칠 것이다.

[가설 3f] 협동 능력은 빅데이터 분석성과에 정(+)의 영향을 미칠 것이다.

Venkatesh *et al*.(2003)은 성능기대, 노력기대, 사회적 영향과 촉진조건은 도입 행위의도에 영향을 주고 사용하게 하며, 행위 의도는 사용 행동에 영향을 준다고 하였다. Vidgen(2014)은 빅데이터를 사용하는 행위의도가 조직 내/외부의 성과에 긍정적인 영향을 미친다고 주장하였다. 행위의도가 빅데이터 분석 성과에 영향을 주는 매개변수로 사용된다. 따라서 다음과 같이 가설을 설정한다.

[가설 4] 빅데이터 행위의도는 분석성과에 정(+)의 영향을 미칠 것이다.

3.3 변수의 설정 및 변수의 조작적 정의

본 연구에서는 데이터 사이언티스트의 핵심역량이 빅데이터 분석 행위의도와 성과에 미치는 영향을 분석하기 위해 구조방정식 모형을 사용했다. 연구 모형의 각 변수들을 측정하기 위한 변수의 조작적 정의와 연구자를 <Table 1>에 제시하였다.

설문지는 기술역량(해킹 4, 시각화 3, 머신러닝 3, 통계 3) 13개, 경영역량(경영분석 3, 문제해결 3) 6개, 협업역량(의사소통 3, 창의성 3, 협동 3) 9개, 행위의도 3개, 일반항목 10개로 총 41개 항목으로 구성되었다. 데이터 사이언티스트 역량 요인은 '매우 동의함(1점)'부터 '매우 동의하지 않음(7점)'까지로 구성된 리커트 7점 척도를 활용하였다.

기술역량의 해킹은 명령어의 텍스트를 조작하고 벡터화된 작업을 이해하며 알고리즘적으로 생각할 수 있는 정도로, 시각화는 분석된 데이터 결과를 인포그래픽 등을 사용하는 정도로 측정한다. 또한, 머신러닝은 클라우드 기반의 데이터를 분석 및 예측 하는 기술로 기계가 학습하듯 다양한 데이터를 수집 분석하여 패턴을 만들어내는 정도로, 통계는 빅데이터 규모의 데이터를 조작하고 종합적으로 분석하는 정도로 측정한다.

경영역량의 경영분석은 예측 모델의 생성과 성과 증진을 위해 데이터 기반의 트렌드를 분석하는 정도로, 문제해결은 미가공 데이터로부터 직면한 문제를 이끌어내는 정도로 측정한다.

현업역량의 경우 의사소통은 알고리즘을 설명하는 차원을 넘어서 동료 및 다양한 부서와 소통하는 정도, 창의성은 내부에 숨겨진 것을 알고자하는 욕구와 문제해결을 위해 명확한 가설 집합을 만드는 정도로 측정한다. 마지막으로, 협동은 학계 및 산업 사이의 공동 노력과 협업하는 정도로 측정한다.

행위의도는 데이터 사이언티스트의 빅데이터 사용계획과 빅데이터 추천 정도로 측정하고, 분석성과는 Kaggle에 실제로 빅데이터 분석 경쟁에 참여하여 얻은 성과 점수를 활용한다. 하나의 경쟁에 참여할 때마다 개인 혹은 팀의 순위, 팀원 수, 참여 수 등의 정보를 기반으로 하여 Kaggle 연구진이 개발한 식을 사용한다. 본 식은 다음과 같다(Cukierski, 2015).

$$\left[\frac{100{,}000}{N_{teammates}}\right]\left[Rank^{-0.75}\right]\left[\log_{10}(1+\log_{10}(N_{teams}))\right]\left[e^{-t/500}\right]$$

Table 1. Operational Definition and Construct from Literature Review

Variables		Operational definition	Relevant studies
Technology Competency			
Hacking	TH1	Computer programming language	Conway, 2010
	TH2	Network system	
	TH3	System/server	
	TH4	Hacking/security system	
Visualization	TV1	Information structuring; information visualization; visual information representation; infographics	LaValle *et al*., 2013
	TV2	Data visualization methods and tools	
	TV3	Library-based visualization	
Machine learning	TM1	Basic grounding concepts of machine learning	Davenport, 2012
	TM2	Supervised learning	
	TM3	Unsupervised learning	
Statistics	TS1	Mathematical knowledge needed for statistics	Rauser, 2011
	TS2	Statistical analysis tools	
	TS3	Operating data and analyze it comprehensively	
Management Competency			
Business analysis	MB1	Identifying of business trends	Chiang *et al*., 2012
	MB2	Designing potentially promising business models	
	MB3	Creating optimized business process	
Problem solving	MP1	Solving problem through trial and error or insight learning	Dhar, 2013
	MP2	Identifying the problems arising from unprocessed data	
	MP3	Using tool for formulizing problem	
Substantive Competency			
Communication	SM1	Building fellowship with co-workers	Rauser, 2011
	SM2	Employee or organization that can help with	
	SM3	Delivering the results analyzed	
Creativity	SR1	Dealing with challenging issues	Patil, 2012
	SR2	Identifying hidden issues	
	SR3	Making a clear hypothesis group to solve problem	
Collaboration	SL1	Working in a multi-functional strategy team with multiple departments	Dinter, 2014
	SL2	Carrying out a collaboration project with people in other fields	
	SL3	Collaboration with university of company	
Behavioral Intention			
BI	BI1	Planning to use big data	LaValle *et al*., 2013
	BI2	Degree of using big data	
	BI3	Recommendation using big data to others	
Big data analysis performance			
Performance score	Pfm	The formula of big data performance score in Kaggle competition $\left[\frac{100{,}000}{N_{teammates}}\right]\left[Rank^{-0.75}\right]\left[\log_{10}(1+\log_{10}(N_{teams}))\right]\left[e^{-t/500}\right]$	Kaggle (www.kaggle.com)

4. 실증분석

4.1 자료의 수집

설문조사는 2015년 3월 23일부터 4월 20일까지 30일간 빅데이터 분석 플랫폼인 Kaggle에 등록된 데이터 사이언티스트 중 3월 22일을 기준으로 10,000points 이상 점수를 얻은 데이터 사이언티스트를 대상으로 수행하였다. 기준에 부합하는 데이터 사이언티스트 1864명 중 연락처가 등록된 데이터 사이언티스트를 추출하여 280명에게 설문지를 배포하여 유효응답 103개(유효응답률 36.7%)를 분석에 활용하였다. 수집된 자료를 통계프로그램인 SPSS for win. 22.0과 SmartPLS 3.0을 사용하여 다음과 같이 처리하였다.

첫째, 조사도구의 타당성과 신뢰도 검증을 위하여 탐색적 요인분석을 실시하고, 분산추출지수를 확인하였다.

둘째, 가설검증을 위하여 PLS 구조방정식 모형을 사용하여 경로계수를 산출하였고, 각 변수의 영향력을 도출하였다.

4.2 기술통계

인구통계학적 특성에서 연령대로 보면 20대가 42.9%로 가장 많았으며, 다음으로 30대가 39%, 40대 이상이 18.1%의 순이었다. 학력은 고졸이 1.9%, 대졸이 26.7%, 석사가 46.7%, 박사가 24.8%이고, 전공별로 컴퓨터 사이언스는 43.8%, 기타 공학이 22.9%, 수학 및 통계 전공이 20%, 경영 및 경제 전공이 8.6%, 기타가 4.8%로 나타났다. 국가별로는 유럽이 45.7%, 미주가 31.4%, 아시아 및 오세아니아가 21%, 아프리카가 1.9%로 나타났다.

Table 2. Descriptive Analysis

Group	No.	Percentage	Group	No.	Percentage
Gender			**Major**		
Male	98	93.3	Computer science	46	43.8
Female	7	6.7	Engineering	24	22.9
Age					
20s	45	42.9	Mathematics and statistics	21	20
30s	41	39	Business and economics	9	8.6
40 or older	19	18.1	Etc.	5	4.8
Education level			**Country**		
High school diploma	2	1.9	Europe	48	45.7
Bachelors'	28	26.7	America	33	31.4
Masters'	49	46.7	Asia and Oceania	22	21
Doctoral	26	24.8	Africa	2	1.9

4.3 요인분석

본 연구의 측정변수는 척도 순화과정을 통하여 일부항목을 제거하였다. 먼저, 타당도 검증을 하기 위하여 탐색적 요인분석을 실시하였다. 모든 측정변수는 구성요인을 추출하기 위해서 주성분 분석을 사용하였으며, 요인 적재치의 단순화를 위하여 직교회전 방식을 채택하였다. 요인 적재치는 각 변수와 요인간의 상관관계의 정도를 나타낸다. 그러므로 각 변수들은 요인적재치가 가장 높은 요인에 속하게 된다. 또한 고유값은 특정 요인에 적재된 모든 변수의 적재량을 제곱하여 합한 값을 말하는 것으로, 특정 요인에 관련된 표준화된 분산을 가리킨다. 일반적으로 사회과학 분야에서 요인과 문항의 선택 기준은 고유값은 1.0 이상, 요인 적재치는 0.4 이상이면 유의한 변수로 간주하며 0.5가 넘으면 아주 중요한 변수로 본다. 따라서 본 연구에서는 이들의 기준에 따라 고유값이 1.0 이상, 요인 적재치가 0.4 이상을 기준으로 하였다.

<Table 3>은 데이터 사이언티스트의 역량에 대한 요인분석 결과이다. 설명된 총 분산은 70.738%로 나타났다. 기존의 연구 설계에서 제시되었던 요인들 중 적재치가 0.4 이하인 SL요인(협업역량의 협동스킬)을 제거하였고, 변수 TH1 또한 이론 구조에 맞지 않게 적재되어 제거하였다. 그 결과 독립변수로는 기술역량의 해킹, 시각화, 머신러닝, 통계, 경영역량의 경영분석과 문제해결, 협업역량의 의사소통과 창의성이, 매개변수로는 행위의도가 분석요인으로 도출되었다.

Table 3. Factor Analysis and Reliability Analysis

Variables		Factor loading	Communality	Eigen value	Variance explanation power	Alpha if item deleted	Cronbach's α
Business analysis	MB2	0.918	0.859	3.020	10.068	0.913	0.913
	MB3	0.897	0.884			0.874	
	MB1	0.868	0.845			0.883	
Hacking	TH2	0.884	0.879	2.345	7.816	0.813	0.813
	TH4	0.823	0.762			0.681	
	TH3	0.808	0.741			0.541	
Visualization	TV1	0.879	0.831	2.399	7.996	0.811	0.811
	TV3	0.824	0.787			0.644	
	TV2	0.737	0.756			0.609	
Problem solving	MP2	0.790	0.746	2.536	8.452	0.765	0.765
	MP3	0.781	0.683			0.698	
	MP1	0.728	0.643			0.712	
Machine learning	TM1	0.819	0.687	1.925	6.415	0.764	0.764
	TM2	0.616	0.766			0.584	
	TM3	0.478	0.645			0.626	
Behavior intention	BI1	0.967	0.791	2.655	8.850	0.746	0.746
	BI3	0.920	0.628			0.719	
	BI2	0.897	0.747			0.573	
Statistics	TS2	0.746	0.685	1.822	6.074	0.698	0.698
	TS1	0.523	0.742			0.640	
	TS3	0.491	0.632			0.570	
Creativity	SR2	0.835	0.748	2.207	7.355	0.670	0.670
	SR1	0.807	0.722			0.608	
	SR3	0.585	0.594			0.700	
Communication	SM1	0.756	0.623	2.322	7.740	0.660	0.660
	SM2	0.743	0.675			0.566	
	SM3	0.739	0.619			0.601	
					70.738		

4.4 모형 적합성 평가

평가를 위한 첫 번째 기준은 일반적으로 사용되는 내적 일관성 신뢰도이다. 내적 일관성에 대한 전통적인 기준은 Cronbach's alpha 이며 이것은 관찰된 측정 변수 간 상관관계를 기반으로 신뢰도를 평가한다. 그러나 이는 척도에 속하는 항목의 수에 민감하게 반응하며 일반적으로 내적 일관성 신뢰도가 저평가되는 경향이 있다. 따라서 Cronbach's alpha는 일반적으로 보수적인 내적 일관성 신뢰도 측정법으로 사용된다. 이러한 한계점으로 인하여 내적 일관성 신뢰도 평가에 좀 더 적합한 다른 평가방법인 구성개념 신뢰도(composite reliability-CR(p_c))를 사용한다. 구성개념 신뢰성은 0과 1사이의 분산을 가지며 값이 높을수록 높은 신뢰도를 나타낸다. 구체적으로 탐험적 연구에서 구성개념 신뢰성의 값이 0.708 이상인 경우 수용가능하다(Nunnally and Bernstein, 1994). 본 연구의 측정 변수는 <Table 4>에서와 같이 모두 0.708 이상이므로 내적 일관성을 확보한 것으로 나타났다.

Table 4. Reliability and Convergent Validity

	AVE	Composite Reliability	R square	Cronbach's alpha	Communality	Redundancy
BI	0.719	0.836	0.470	0.610	0.719	0.169
MB	0.853	0.946		0.914	0.853	
MP	0.564	0.788		0.598	0.564	
Pfm	1.000	1.000	0.678	1.000	1.000	1.000
SM	0.636	0.840		0.715	0.636	
SR	0.618	0.826	0.016	0.702	0.618	0.291
TH	0.777	0.912	0.032	0.854	0.777	0.298
TM	0.895	0.953	0.032	0.941	0.895	0.586
TS	0.604	0.749		0.363	0.604	
TV	0.728	0.889	0.093	0.813	0.728	0.304
평균			0.220		0.739	

집중 타당성을 측정하기 위한 일반적인 방법은 평균분산팽창(average variance extracted(AVE))이다. 이 기준은 construct에 해당하는 indicator들의 loading 제곱의 전체평균을 의미한다. 그러므로 AVE는 construct의 공통성(communality)과 동일하다. 개별 indicator들에 사용된 기준값과 동일하게 사용하여 평균적으로 AVE 값이 0.5 이상이면 적합하게 분산을 설명한다고 할 수 있다. 이를 근거로 본 연구에서는 집중 타당도는 모두 기준치를 만족하는 있는 것으로 나타났다.

본 연구에서는 Fornell-Larcker 기준으로 판별타당성을 검증하는 보수적인 방법을 사용한다. 이 방법은 각 construct AVE 값이 제곱근과 잠재변수 상관관계를 비교하는 방법이다. 구체적으로 각 construct AVE 값의 제곱근은 다른 construct와의 가장 높은 상관관계보다 커야만 한다. <Table 5>에서 construct BI를 살펴보면, BI의 AVE 제곱근은 0.847로 나머지 잠재변수들 간의 상관관계 값보다 크기 때문에 판별타당성을 충족한 것으로 나타났다.

Table 5. Correlations and Discriminant Validity

	BI	MB	MP	Pfm	SM	SR	TH	TM	TS	TV
BI	**0.847**[a]									
MB	0.328	**0.924**								
MP	0.224	0.050	**0.751**							
Pfm	0.535	0.020	0.079	**1.000**						
SM	-0.339	0.056	-0.102	-0.687	**0.798**					
SR	0.150	0.165	0.008	0.087	-0.126	**0.786**				
TH	0.101	-0.047	0.000	0.271	-0.231	0.206	**0.881**			
TM	0.602	0.082	0.174	0.676	-0.403	0.141	0.179	**0.946**		
TS	0.035	0.290	-0.007	-0.067	0.117	0.192	0.068	0.051	**0.777**	
TV	-0.004	0.135	0.133	0.041	0.000	0.249	0.178	0.050	0.304	**0.853**

[a] : Square root AVE.

구조 모형의 적합도 지표는 Stone-Gisser Q^2 test 통계량인 교차 검증된 Redundancy 지표가 있다. 이 지표는 구조 모형의 통계 추정량으로 구조 모형의 적합성을 나타내며, 그 값이 양수여야 한다(Chin, 1998; Tenenhaus *et al*., 2005). 본 연구에서는 <Table 4>에서와 같이 종속변수를 중심으로 모두 양의 값을 보이고 있으므로 구조 모형의 예측 적합성을 확인할 수 있다. 또한 PLS 구조 모형에 대한 평균적인 적합도 평가는 우선 각각의 종속 변수 경로 모형에 대한 평가를 고려해야 하는데 해당 종속변수의 R^2 값으로 평가하게 된다. Cohen(1977)에 의하면 R^2 값이 효과 정도는 상(0.26 이상), 중(0.13～0.26), 하(0.02～0.13)으로 구분하고 있다. 이를 근거로 본 연구에서 설정한 연구모형의 종속변수의 적합도는 기준치를 대부분 만족하고 있는 것으로 나타났다. R^2 값은 종속변수에 직접적으로 영향을 미치는 값만 나타나기 때문에 간접적으로 영향을 미치는 경영분석(MB), 문제해결(MP), 의사소통(SM), 통계(TS) 값은 제외한다.

마지막으로 PLS 경로 모형 전체의 적합도는 모든 종속 변수의 R^2 평균 값과 Communality의 평균값을 곱한 후, 이를 다시 제곱근한 값으로 정의된다(Chin, 1998; Tenenhaus *et al*., 2005). 이 적합도의 크기는 최소 0.1 이상이어야 하며, 그 크기에 따라서 상(0.36 이상), 중(0.25～0.36), 하(0.1～0.25)로 구분된다. 본 연구의 PLS 경로 모형 전체 적합도를 측정해 본 결과 모든 종속 변수의 R^2 평균 값은 0.22이며, Communality 평균 값은 0.739이고, 이 둘을 곱한 값의 제곱근은 0.4(= $\sqrt{0.22\times0.739}$)로 나타나서 모델의 전체 적합도가 매우 높은 것으로 보여 진다. 따라서 본 연구모형의 적합도가 확인되어 가설의 검증 및 결과 해석이 가능한 것으로 나타났다.

4.5 경로분석

데이터 사이언티스트의 역량이 빅데이터 분석성과에 미치는 영향에 대한 가설 검증은 연구모형에 대한 경로계수를 통하여 <Figure 2>와 <Table 6>과 같이 검증하였다. 경로계수의 통계적 유의수준은 C.R.(Critical Ratio)값을 우선적으로 고려하는데, C.R. 값이 -1.96～+1.96의 범위를 벗어나면 95% 신뢰

수준에서 귀무가설이 기각되므로 통계적으로 유의하다고 판단하였다.

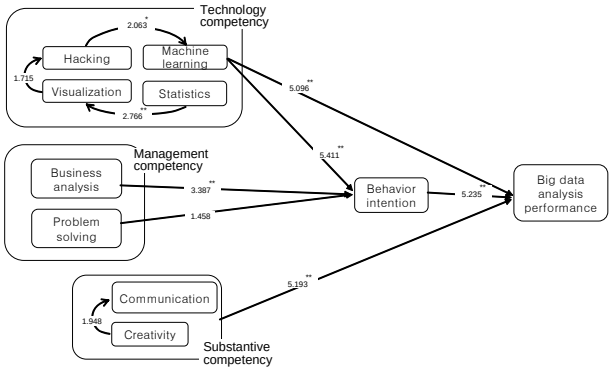

Figure 2. The Results of PLS-SEM

Table 6. The Results of PLS-SEM

	Original Sample (O)	Sample Mean (M)	Standard Error	T Statistics (Critical Ratio)	P Values
TM → BI	0.503	0.497	0.093	5.411	0.000**
TM → Pfm	0.396	0.388	0.078	5.096	0.000**
TS → TV	0.304	0.350	0.110	2.766	0.006**
TV → TH	0.178	0.187	0.104	1.715	0.087
TH → TM	0.179	0.184	0.087	2.063	0.040*
MB → BI	0.289	0.290	0.085	3.387	0.001**
MP → BI	0.107	0.118	0.074	1.458	0.145
SR → SM	-0.141	-0.150	0.073	1.948	0.052
SM → Pfm	-0.464	-0.471	0.075	5.193	0.000**
BI → Pfm	0.423	0.425	0.084	5.235	0.000**

Note) * $p < 0.05$; ** $p < 0.01$.

먼저 연구가설 첫 번째 가설인 기술역량을 살펴보면, 머신러닝이 행위의도에 미치는 영향은 C.R. 값 5.411, 유의수준 0.000으로 나타나, H1e "머신러닝은 빅데이터 행위의도에 정(+)의 영향을 미칠 것이다"는 채택되었다. 머신러닝이 분석성과에 미치는 영향은 C.R. 값 5.096, 유의수준 0.000으로 나타나 H1f "머신러닝은 빅데이터 분석성과에 정(+)의 영향을 미칠 것이다"는 채택되었다. 통계 능력이 시각화에 미치는 영향은 C.R. 값 2.766, 유의수준 0.006으로 신뢰수준 95%에서 H1i가 채택되었다. 또한, 해킹이 머신러닝에 미치는 영향도 C.R. 값 2.063, 유의수준 0.040으로 H1k가 채택되었다. 그러나 해킹스킬과 시각화, 통계는 행위의도와 분석성과에 영향을 미치지 않는 것으로 나타나 가설이 기각되었다. 비록 이러한 기술역량은 행위의도와 분석성과에 직접적인 영향을 미치지는 못하지만 통계 → 시각화 → 해킹 → 머신러닝 사이클이 형성되어 빅데이터 분석성과를 내는 간접적인 요인으로 도출 되었다. 일반적으로 LISREL이나 AMOS와 같은 공분산 기반의 분석기법을 많이 쓰지만, 본 연구에서는 상호작용 효과를 검증해야하기 때문에 PLS 구조방정식을 사용하였다. 이는 성분(Component)을 기반으로 하므로 독립변수와 종속변수간의 한 방향의 영향력뿐만 아니라 독립변수간의 상호관계를 규명하여 사이클을 도출하였다.

두 번째 가설인 경영역량을 살펴보면 경영분석이 행위의도에 미치는 영향은 C.R. 3.387, 유의수준 0.001로 가설 H2a인 "경영 분석은 빅데이터 행위의도에 정(+)의 영향을 미칠 것이다"가 채택되었다. 문제해결이 행위의도에 미치는 영향은 C.R. 1.458, 유의수준 0.145으로 신뢰수준 90%에 조금 못 미치는 수준에서 기각 되었고, 경영분석과 문제해결능력도 분석성과에 직접적인 영향을 미치지 않는 것으로 나타나 가설 H2b와 H2d는 기각되었다.

세 번째 가설인 현업역량과 관련해서 살펴보면, H3b "의사소통 능력은 빅데이터 분석성과에 정(+)의 형향을 미칠 것이다."는 C.R. 값 5.193, 유의수준 0.000으로 채택되었다. 그러나 창의적 역량은 행위의도와 빅데이터 분석성과에 영향을 미치지 않는 것으로 나타나 가설 H3c와 H3d는 기각되었다. 마지막으로 H4 "빅데이터 행위의도는 분석성과에 정(+)의 영향을 미칠 것이다"도 C.R. 5.235, 유의수준 0.000으로 채택되었다. 이는 빅데이터를 사용하려는 행위의도가 내재되어있는 경우 빅데이터 분석성과에 긍정적인 영향을 준다는 것을 의미한다. 가설의 검증결과를 요약하면 <Table 7>과 같다.

Table 7. The summary of Hypothesis Results

Hypothesis		Accept/ Reject	Hypothesis		Accept/ Reject
1a	TH → BI(+)	Reject	2a	MB → BI(+)	Accept
1b	TH → Pfm(+)	Reject	2b	MB → Pfm(+)	Reject
1c	TV → BI(+)	Reject	2c	MP → BI(+)	Reject
1d	TV → Pfm(+)	Reject	2d	MP → Pfm(+)	Reject
1e	TM → BI(+)	Accept	3a	SM → BI(+)	Reject
1f	TM → Pfm(+)	Accept	3b	SM → Pfm(+)	Accept
1g	TS → BI(+)	Reject	3c	SR → BI(+)	Reject
1h	TS → Pfm(+)	Reject	3d	SR → Pfm(+)	Reject
1i	TS → TV(+)	Accept	3e	SL → BI(+)	Reject
1j	TV → TH(+)	Reject	3f	SL → Pfm(+)	Reject
1k	TH → TM(+)	Accept	4	BI → Pfm(+)	Accept

5. 결 론

본 연구는 데이터 사이언티스트의 주요 역량요인이 빅데이터 행위의도와 분석성과에 미치는 영향의 방향에 대해 구조방정식의 경로모형으로 분석하였다. 영향의 방향에 대한 분석결과 타당한 모형을 얻었으며 핵심역량에 대한 경로계수가 도출되었다.

연구결과 빅데이터 분석성과에 대한 영향력은 '머신러닝 > 창의성 > 경영분석 > 문제해결'순으로 나타났으며, 분석성과

에 가장 영향력이 높은 기술역량의 경우 통계 → 시각화 → 해킹 → 머신러닝으로 경로계수를 추정하여 기술역량 사이클을 얻을 수 있었다.

본 연구를 통해 빅데이터 분석성과를 높이기 위해서는 영향력을 기준으로 데이터 사이언티스트의 역량을 강화할 필요가 있음을 확인했다. 특히 데이터 사이언티스트 역량에 대한 차별적인 우선순위를 도출했다. <Figure 7>에서 가로축은 실제 응답자의 평균 점수를 지수로 환산한 값, 세로축은 경로계수 값을 도식화 한 것이다. 이는 각 요인의 지수가 한 단위 증가할 때마다 경로계수만큼 빅데이터 분석성과가 증가하는 것을 보여준다. 기술역량의 머신러닝(TM)의 경우, 평균점수는 61점이고 영향력은 0.934로 머신러닝 기술이 62점으로 한 단위 증가하면 빅데이터 분석성과는 0.934만큼 증가한다고 해석할 수 있다. 반면, 기존의 연구에서는 데이터 사이언스팀의 의사소통 역량이 빅데이터 분석력을 향상시키는데 긍정적인 영향을 준다는 주장이 있지만(Rauser, 2011), 본 연구에서는 의사소통 역량이 오히려 문제점으로 지적되고 있다.

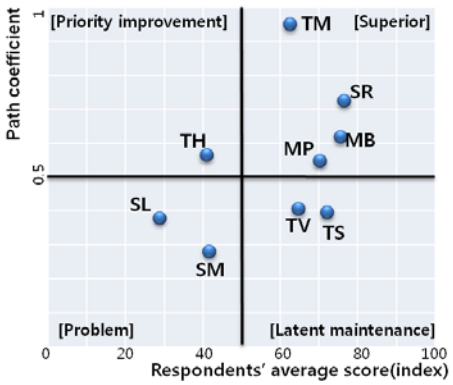

Figure 3. The Matrix of Main Competency of Data Scientist

연구 결과의 의의는 다음과 같다. 첫째, 데이터 사이언티스트 역량과 빅데이터 분석성과의 관계에 대해 정량적인 분석을 수행하였다. 데이터 사이언티스트를 대상으로 한 대부분의 연구들은 정성적인 분석이 주를 이루었다. 이는 그동안 데이터 사이언티스트의 인력풀이 확보되지 않아 표본수집의 한계 때문인 것으로 보여 진다. 이에 본 연구는 크라우드 소싱의 특징을 가지는 Kaggle 플랫폼을 활용하여 전 세계의 데이터 사이언티스트를 유인함으로써 이를 보완하였다.

둘째, 본 연구는 정부의 데이터 사이언티스트 전담조직 구성과 학위과정 개발, 기업의 데이터 사이언티스트 팀 구축 등 관련 정책을 수립하는데 있어 지표가 될 것이다. 데이터 사이언티스트 인재를 양성하는 전략을 세우는데 있어서 기존의 비즈니스 구루, 빅데이터 전문가들 개인의 주관적인 주장을 토대로 우선순위에 따른 정량적인 역량 체계를 도출했다는데 의의가 있다.

본 연구는 데이터 사이언티스트의 역량 구성요인들이 빅데이터 분석을 극대화 하는 전략을 구축하기 위한 기반연구로 수행되었다. 하지만 빅데이터 분야에서는 현재까지 본 연구와 비교할 선행연구들이 충분하지 않아 제시된 연구결과가 특정 데이터 사이언티스트의 특성인지 빅데이터 전담 조직의 현황인지 알 수가 없다. 일반화를 위해 다양한 산업, 업종 및 기업에 종사하는 표본을 얻어 추후 다양한 방향으로 연구가 이루어 져야만 할 것이다. 또한 인구통계학적 변수로 조절효과를 분석하여 각 집단별로 빅데이터 분석성과에 미치는 영향력에 대한 차이를 비교해 볼 필요도 있다. 방법론적으로는 본 연구에서 활용한 분산기반 부분최소제곱 구조방정식(PLS-SEM)기법 뿐만 아니라 AMOS나 LISREL 등의 공분산기준 구조방정식모형의 분석을 실시하여 도출된 결과를 비교분석 할 수 있을 것이다.

참고문헌

Chiang, R. H., Goes, P., and Stohr, E. A. (2012), Business intelligence and analytics education, and program development : A unique opportunity for the information systems discipline, *ACM Transactions on Management Information Systems(TMIS)*, **3**(3), 12.

Chiang, R. M., Kauffman, R. J., and Kwon, Y. (2014), Understanding the paradigm shift to computational social science in the presence of big data, *Decision Support Systems*, **63**, 67-80.

Chin, W. W. (1998), The partial least squares approach to structural equation modeling, *Modern Methods for Business Research*, **295**(2), 295-336.

Cho, S. G., Cho, J., and Kim, S. B. (2015), Discovering meaningful trends in the inaugural addresses of United States presidents via text mining, *Journal of the Korean Institute of Industrial Engineers*, **41**(5), 453-460.

Cho, W.-S. (2013), A study on the education and training methods of Data scientist, *Science and Technology Policy*, **23**(3), 44-55.

Cohen, J. (1977), Statistical power analysis for the behavioral sciences, *Lawrence Erlbaum Associates, Inc.*

Conway, D. (2010), *The data science venn diagram*, Dataists, Retrieved February, 9, 2012 (http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram).

Davenport and Thomas, H. (2012), The human side of big data and High-performance analytics, *International Institute for Analytics* (http://www.ndm.net/datawarehouse/pdf/Research_Human_Side_of_Big_Data_and_High_Performance_Analytics.pdf).

Dhar, V. (2013), Data science and prediction, *Communications of the ACM*, **56**(12), 64-73.

Dino, M. J. S. and de Guzman, A. B. (2015), Using partial least squares (PLS) in predicting behavioral intention for telehealth use among filipino elderly, *Educational Gerontology*, **41**(1), 53-68.

Dinter, B., Douglas, D., Chiang, R. H., Mari, F., Ram, S., and Schoder, D. (2014), Big data panel at SIGDSS Pre-ICIS 2013 : A Swiss-army knife? the profile of a data scientist, *Reshaping Society through Analytics, Collaboration, and Decision Support : Role of Business Intelligence and Social Media*, **18**, 7.

Fenn, J. and LeHong, H. (2011), Hype cycle for emerging technologies, *Gartner*.

Hair, J. F., Sarstedt, M., Pieper, T. M., and Ringle, C. M. (2012), The use

of partial least squares structural equation modeling in strategic management research : a review of past practices and recommendations for future applications, *Long Range Planning*, **45**(5), 320-340.

Hair Jr, J. F., Hult, G. T. M., Ringle, C., and Sarstedt, M. (2013), A primer on partial least squares structural equation modeling (PLS-SEM), *Sage Publications*.

Hollis, C. (2011), IDC digital universe study : big data is here, now what.

Jung, H. and Song, S.-K. (2012), Strategy for cultivating talent in the world of big data, *Journal of Internet Computing and Services*, **13**(3), 45-50.

Kart, L., Heudecker, N., and Buytendijk, F. (2013), Survey analysis : big data adoption in 2013 shows substance behind the hype, *Gartner Report* GG0255160.

Kim, M. and Koo, P. (2013), A study on big data based investment strategy using internet search trends, *Journal of the Korean Operations Research and Management Science Society*, **38**(4), 53-63.

Kim, S. W., Kim, G. G., and Yoon, B. K. (2014), A study on a way to utilize big data analytics in the defense area, *Journal of the Korean Operations Research and Management Science Society*, **39**(2), 1-20.

Laney, D. and Kartpaper, L. (2012), Emerging role of the data scientist and the art of data science, *Gartner Inc, Stamford.*

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., and Kruschwitz, N. (2013), Big data, analytics and the path from insights to value, *MIT Sloan Management Review*, **21**, 20-32.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., and Roxburgh, C. (2011), Big data : The next frontier for innovation, competition, and productivity, *McKinsey Global Institute.*

Martinez, M. G. and Walton, B. (2014), The wisdom of crowds : The potential of online communities as a tool for data analysis, *Technovation*, **34**(4), 203-214.

Nomura Research Institute (2012), The era of big data, *IT Solutions Frontier*.

Nunnally, J. C. and Bernstein, I. H. (1994), Psychometric theory, *New York : McGraw-Hill*.

Pantai, K. L. (2012), PLS path model for testing the moderating effects in the relationships among formative IS usage variables of academic digital libraries, *Australian Journal of Basic and Applied Sciences*, **6**(7), 365-374.

Patil, D. J. (2011), Building data science teams, *O'Reilly Media, Inc.*

Patil, D. J. and Davenport, T. H. (2012), Data scientist, *Harvard Business Review*, **90**, 70-76.

Rahul, D. (2012), Data/Web Analyst vs. Data Scientist (http://blogs.splunk.com/2012/05/16/analytics-staffing-for-big-data/).

Rauser, J. (2011), What is data scientist? (http://www.forbes.com/sites/danwoods/2011/10/07/amazons-john-rauser-on-what-is-a-data-scientist/).

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., and Lauro, C. (2005), PLS path modeling, *Computational statistics and data analysis*, **48**(1), 159-205.

Thorp, J. (2003), *The information paradox : realizing the business benefits of information technology*, McGraw-Hill Ryerson.

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003), User acceptance of information technology : Toward a unified view, *MIS Quarterly*, **27**(3), 425-478.

Vidgen, R. (2014), Creating business value from big data and business analytics : organizational, managerial and human resource implications (http://www.nemode.ac.uk/wp-content/uploads/2014/07/Vidgen-2014-NEMODE-big-data-scientist-report-final.pdf).

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., and Gnanzou, D. (2015), How 'big data' can make big impact : Findings from a systematic review and a longitudinal case study, *International Journal of Production Economics*, **165**, 234-246.

Will Cukierski (2015), Improved Kaggle Rankings (http://blog.kaggle.com/2015/05/13/improved-kaggle-rankings/).