디지털산업정보학회 논문지 제17권 제3호-2021년 9월
http://dx.doi.org/10.17662/ksdim.2021.17.3.001

대학 빅데이터 기반 학생 취업 로드맵 추천에 관한 연구

박 상 성*

A Study on University Big Data-based Student Employment Roadmap Recommendation

Park Sangsung

〈Abstract〉

The number of new students at many domestic universities is declining. In particular, private universities, which are highly dependent on tuition, are experiencing a crisis of existence. Amid the declining school-age population, universities are striving to fill new students by improving the quality of education and increasing the student employment rate. Recently, there is an increasing number of cases of using the accumulated big data of universities to prepare measures to fill new students. A representative example of this is the analysis of factors that affect student employment. Existing employment-influencing factor analysis studies have applied quantitative models such as regression analysis to university big data. However, since the factors affecting employment differ by major, it is necessary to reflect this. In this paper, the factors affecting employment by major are analyzed using the data of University C and the decision tree model. In addition, based on the analysis results, a roadmap for student employment by major is recommended. As a result of the experiment, four decision tree models were constructed for each major, and factors affecting employment by major and roadmap were derived.

Key Words : University Student Employment Roadmap, Big Data Analysis, Decision Tree, Machine Learning, Institutional Research

Ⅰ. 서론

지속적인 대입가능자원의 감소에 따라 많은 국내대학의 입학 인원이 감소하는 추세이다[1, 2]. 특히, 대학 운영에 있어서 등록금 의존도가 높은 사립대는 학생 충원 미달에 따른 존폐 위기에 처해있다. 이러한 상황 속에서 대학들은 교육의 질 향상, 학생 취업률 제고 등 대책을 마련하여 신입생 충원에 힘쓰고 있다. 이와 같은 대책 마련에는 최근 축적된 대학의 빅데이터를 활용하는 사례가 증가하고 있다[3-6]. 대학 빅데이터는 학생의 기본정보, 학업 성취 정보, 강의 정보 등 다양한 데이터를 포함하고 있으므로, 적절한 분석 기법의 적용을 통해 새로운 가치 창출이 가능하다.

* 청주대학교 빅데이터통계학과 조교수 (단독저자)

대학 빅데이터를 활용한 대표적인 연구는 대학생의 취업에 영향을 미치는 요인을 분석하는 것이다. 기존 연구들은 대학 내에서 수집된 다양한 정보에 회귀분석 등 정량적 모형을 적용하여 취업에 영향을 미치는 요인을 분석하였다. 그러나 대학생 취업은 전공 계열별로 취업자 수, 시기, 영향요인 등이 차이가 발생하므로, 이를 반영할 필요가 있다. 본 논문에서는 대학생의 전공 계열별로 취업에 영향을 미치는 요인을 정량적으로 분석하고 로드맵을 추천하는 방법을 제안한다. 제안하는 방법은 C 대학의 취업 및 미취업 학생 데이터와 의사결정 나무 모형을 활용한다. 또한, 교육부 계열별로 모형을 구축하여 전공에 따른 취업 영향 요인 분석과 로드맵 추천을 수행한다.

Ⅱ. 관련 연구

2.1 대학생 취업 분석에 관한 연구

대학생의 취업에 영향을 미치는 주요 요인을 분석하는 것은 대학 빅데이터를 활용한 대표적인 연구이다. 해당 연구에서는 학점, 어학, NCS 등급 등 학생 정보에 정량적 분석 기법을 적용하고 도출된 결과를 기반으로 대학 정책 수립에 활용한다[7-9]. 윤수경・한유경[7]은 대학생의 개인 배경에 따른 취업 성과 영향요인을 분석하기 위해 한국교육고용패널 자료에 다중 회귀분석 및 로지스틱 회귀분석을 활용하였다. 조장식[8]는 대학 졸업생들의 취업여부에 미치는 영향 요인을 분석하기 위해 입학, 재학 및 개인특성 관련 변수와 로지스틱 회귀분석을 사용하였다. 염동기 외[9]는 대학 행정시스템의 데이터를 이용하여 졸업생들의 취업 현황과 취업 성과에 영향을 미치는 요인을 회귀분석과 의사결정 나무로 모형으로 파악하였다. 이와 같은 기존 연구들은 대학 졸업자의 취업에 영향을 미치는 요인을 정량적으로 분석하고 전공 계열, 학점 등 주요한 요인을 도출하였다. 일반적으로 전공 계열에 따른 일자리 수, 취업형태 등은 상이하므로, 이를 고려하여 모델을 구축할 필요가 있다. 본 연구에서는 전공 계열에 따라 모델을 개별로 구축하고 대학생의 취업에 영향을 미치는 주요 요인을 파악 및 로드맵을 추천하는 방법을 제안한다.

2.2 의사결정 나무

의사결정 나무(Decision tree) 모형은 분리 기준을 통해 데이터를 나뭇가지가 성장하는 형태로 분류 또는 예측을 수행한다[10, 11]. 또한, 의사결정 나무 모형은 분류 또는 예측 과정을 시각화할 수 있다. 의사결정 나무 모형은 특정 변수에 대해 분리 기준을 설정하고 이를 통해 데이터를 분할 한다. 이때, 분리 기준에 적합한 데이터들이 포함된 곳을 노드(Node)라 한다. 노드는 상위에서부터 하위로 이어지며 나뭇가지 형태를 구성한다. 또한, 더 이상 분리 기준이 적용되지 않는 노드를 터미널 노드(Terminal node)라한다. 의사결정 나무 모형은 분리 기준에 따라 다양한 형태로 구분되며, 지니계수(Gini index)가 가장 널리 사용되는 분리 기준이다. 아래의 <식 1>은 지니계수 산출 식이다.

$$Gini \in dex = 1-(\frac{n}{N})^2-(\frac{m}{N})^2, N=n+m \qquad \text{<식 1>}$$

위 <식 1>에서 N은 노드에 포함된 데이터의 수이며, n은 특정 분류에 속한 데이터의 수이다. m은 다른 분류에 속한 데이터의 수이다. 또한, 하나의 노드에 모두 동일한 분류에 속한 데이터가 포함되면 지니계수는 0의 값을 가진다. 취업 유무의 예측 가능성, 모형 구축 과정의 효과적인 시각화 등의 이유로, 본 논문에서는 의사결정 나무 모형을 사용하여 대학생의 취업 로드맵 추천 연구를 수행한다.

Ⅲ. 연구 방법

3.1 실험 데이터

본 논문에서는 취업 대학생의 학습 패턴을 분석하고 전공 계열별로 로드맵을 추천하기 위해 C 대학의 데이터를 사용한다. 해당 데이터는 2010년부터 2020년까지 입학한 학생을 대상으로 학내의 여러 부서로부터 수집하고 병합한다. 최초 수집된 데이터는 총 36,031건이다. 해당 데이터에서 '취업자' 또는 '구직활동중'으로 구분된 것을 바탕으로 취업 학생을 분류한다. 또한, 데이터에 결측이 있는 학생들을 제외한 총 1,824명을 모형 개발에 사용한다. 취업학생은 1,389명, 미취업학생은 435명이다. 아래의 <표 1>은 수집된 변수와 취업 여부에 따른 차이를 확인하기 위해 Chi-square 검정을 실시한 결과이다.

<표 1> 취업 여부에 따른 변수들의 통계적 검정 결과

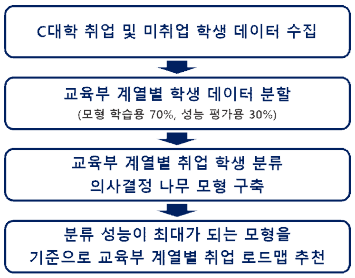
변수명	미취업학생 (n=435)		취업학생 (n=1,389)		p-value
성별					0.941
여성	244	(56.09%)	784	(56.44%)	
남성	191	(43.91%)	605	(43.56%)	
교육부 계열					
공학계열	55		189		< 0.001
예체능계열	60		121		
인문사회계열	265		609		
자연과학계열	55		470		
NCS 등급					
최우수	3	(0.69%)	1	(0.07%)	< 0.001
우수	16	(3.68%)	104	(7.49%)	
보통	369	(84.82%)	1190	(85.67%)	
미흡	45	(10.35%)	90	(6.48%)	
매우미흡	2	(0.46%)	4	(0.29%)	
졸업연령(Age)					
22세이하	127	(29.19%)	514	(37.09%)	< 0.001
23세	88	(20.23%)	230	(16.56%)	
24세	97	(22.30%)	376	(27.07%)	
25세	81	(18.62%)	175	(12.60%)	
26세	23	(5.29%)	70	(5.04%)	
27세이상	19	(4.37%)	24	(1.73%)	
평점평균					
1.9<≤2.9	100	(22.99%)	263	(18.94%)	0.163
2.9<≤3.9	304	(69.88%)	1,012	(72.85%)	
3.9<≤4.5	31	(7.13%)	114	(8.21%)	
어학능력*					
≤282.5	384	(88.27%)	1,230	(88.54%)	0.727
282.5<≤450	7	(1.61%)	20	(1.44%)	
450<≤720	17	(3.91%)	67	(4.83%)	
720<≤990	27	(6.21%)	72	(5.19%)	
자격증현황					
0개	358	(82.29%)	1,149	(82.72%)	0.230
1개	24	(5.52%)	101	(7.27%)	
2개	22	(5.06%)	70	(5.04%)	
3개이상	31	(7.13%)	69	(4.97%)	
수상경력					
0개	289	(66.43%)	757	(54.49%)	< 0.001
1개	80	(18.39%)	323	(23.26%)	
2개	33	(7.59%)	186	(13.39%)	
3개이상	33	(7.59%)	123	(8.86%)	

* 어학능력은 토익점수 기준이며 제2사분위수는 282.5점, 중앙값이 450점이며 최대값은 990임. 이때 중앙값과 최대값의 차이가 커서 450점 990점 평균인 720점을 기준점으로 추가함

상기 <표 1>과 같이 '교육부 계열', 'NCS 등급', '졸업연령', '수상경력'은 취업 여부와 관계가 있다. 본 논문에서는 <표 1>에서 명시한 변수들을 사용하여 교육부 계열별 취업 로드맵 추천 모형을 구축한다.

3.2 제안한 연구방법

본 논문에서는 취업 대학생의 학습 패턴에 따른 전공 계열별 로드맵을 추천하기 위하여 의사결정 나무 모형을 활용한다. 아래의 <그림 1>은 본 논문에서 제안한 연구방법을 도식화한 것이다.



<그림 그림 1> 제안한 연구방법

<그림 1>에서 제안한 연구방법의 절차는 다음과 같다. 먼저, C 대학의 취업 및 미취업 학생의 데이터를 수집한다. 수집된 데이터를 교육부 계열별로 구분하고 모형 학습용(70%)과 성능 평가용(30%)으로 분할한다. 모형 학습용 데이터를 사용하여 교육부 계열별 취업 학생 분류를 위한 의사결정 나무 모형을 구축한다. 의사결정 나무 모형 구축시 분리 기준은 지니계수를 사용한다. 또한, 노드를 최대로 성장시키나, 효과적인 로드맵 제시를 위해 4번째 노드까지만을 시각화한다. 다음으로, 성능 평가용 데이터를 사용하여 구축된 모델의 취업 학생 분류 성능을 측정한다. 마지막으로, 분류 성능이 최대가 되는 모형을 통해 교육부 계열별 학생 취업 로드맵을 추천한다. 본 논문에서 제시한 방법은 통계 소프트웨어인 R을 통해 구현한다.

Ⅳ. 실험 및 결과

모형 학습용 데이터에 속한 학생 수는 총 1,276명이며, 취업학생은 75.24%이다. 또한, 성능 평가용 데이터에 속한 학생 수는 총 548명이며, 취업학생은 78.28%이다. 아래의 <표 2>는 의사결정 나무 모형 구축을 위한 교육부 계열별 데이터 분할 결과이다.

<표 2> 교육부 계열별 데이터 분할 결과

교육부계열	구분	미취업학생		취업학생		총합
공학	학습용	44	(26.19%)	124	(73.81%)	168
	평가용	11	(15.28%)	61	(84.72%)	72
	총합	55	(22.92%)	185	(77.08%)	240
예체능	학습용	41	(32.54%)	85	(67.45%)	126
	평가용	19	(34.55%)	36	(65.45%)	55
	총합	60	(33.15%)	121	(66.85%)	181
인문사회	학습용	189	(30.93%)	422	(69.07%)	611
	평가용	76	(28.90%)	187	(71.10%)	263
	총합	265	(30.32%)	609	(69.68%)	874
자연과학	학습용	39	(10.63%)	328	(89.37%)	367
	평가용	16	(10.13%)	142	(89.87%)	158
	총합	55	(10.48%)	470	(89.52%)	525
전체	학습용	316	(24.76%)	960	(75.24%)	1,276
	평가용	119	(21.72%)	429	(78.28%)	548
	총합	435	(23.85%)	1,389	(76.15%)	1,824

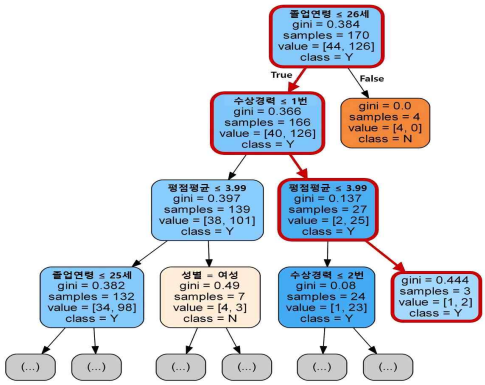
아래의 <표 3>은 교육부 계열별로 구축된 의사결정 나무 모형에 성능 평가용 데이터를 사용하여 취업 학생 분류 성능을 측정한 것이다. <표 3>에서 자연과학계열에 대한 의사결정 나무 모형 성능이 가장 우수하다.

<표 3> 교육부 계열별 구축된 의사결정 나무 모형 성능

교육부 계열	Accuracy	Precision	Recall	Specificity	F1-score
공학	0.76	0.87	0.84	0.27	0.85
예체능	0.56	0.66	0.69	0.32	0.68
인문사회	0.66	0.74	0.82	0.28	0.77
자연과학	0.86	0.91	0.94	0.13	0.92

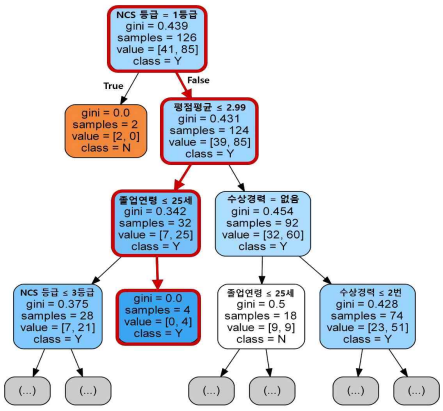
교육부 계열별로 구축된 의사결정 나무 모형을 시각화하여 취업 학생에 대한 로드맵을 추천한다. 먼저, 공학계열에 대한 취업 학생 로드맵은 <그림 2>와 같다. <그림 2>와 같은 결과로, 공학계열은 졸업연령이 26세 이하, 수상경력 2번 이상, 평점평균이 4.0 이상

인 학생이 취업에 유리한 것으로 도출되었다. 따라서 공학계열 학생은 전공 학점을 체계적으로 관리할 필요가 있으며, 교내외 대회에 적극적으로 참여할 필요가 있다.



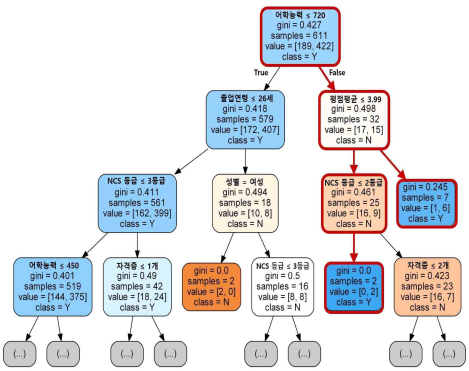
<그림 2> 공학계열 취업 학생 로드맵

예체능계열에 대한 취업 학생 로드맵은 <그림 3>과 같다. <그림 3>의 결과로, 예체능계열은 NCS 등급, 평점평균이 취업에 크게 영향을 미치지 않는다. 따라서 예체능계열 학생의 취업에는 적성이 중요한 요인일 것으로 판단된다.



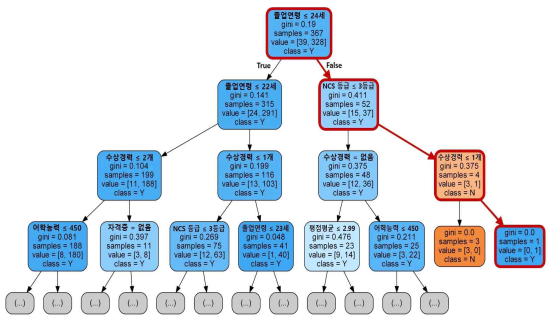
<그림 3> 예체능계열 취업 학생 로드맵

인문사회계열에 대한 취업 학생 로드맵은 <그림 4>와 같다. <그림 4>의 결과로, 인문사회계열은 어학능력(토익점수) 720점 초과, 평점평균 4.0 이상, NCS 등급이 2등급 이하인 학생이 취업에 유리한 것으로 도출되었다. 따라서 인문사회계열 학생은 전공 학점의 체계적인 관리와 기준 어학점수 획득을 위한 추가 수강, 시험 응시가 필요하다.



<그림 4> 인문사회계열 취업 학생 로드맵

자연과학계열에 대한 취업 학생 로드맵은 <그림 5>와 같다. <그림 5>의 결과로, 자연과학계열은 졸업연령 25세 이상, 수상경력 2개 이상인 학생이 취업에 유리한 것으로 도출되었다. 따라서 자연과학계열 학생은 교내외 대회에 적극적으로 참여하는 것이 중요하다.



<그림 5> 자연과학계열 취업 학생 로드맵

아래의 <표 4>는 상기 교육부 계열별로 도출한 로드맵의 내용을 정리한 것이다. 아래의 <표 4>와 같이 교육부 계열별로 취업에 유리한 요인을 식별할 수 있다. 따라서 제안한 방법을 통해 해당 대학에서는 학생들의 취업률 제고를 위한 다양한 전략 구성이 가능할 것이다.

<표 4> 교육부 계열별 도출된 로드맵

교육부 계열	로드맵 내용
공학	• 전공학점의 체계적 관리 • 수상경력 축적을 위한 교내외 대회 참여 독려
예체능	• 학생별 포트폴리오 관리 필요 • 적성에 적합한 세부 및 부전공 모색
인문사회	• 전공학점의 체계적 관리 • 어학점수 획득을 위한 관련 과목 수강 및 시험 응시 권고
자연과학	• 수상경력 축적을 위한 교내외 대회 참여 독려

Ⅴ. 결론 및 향후 연구

대학생 취업에 영향을 미치는 요인을 분석하는 연구는 대학 빅데이터를 활용한 대표적인 사례이다. 기존 취업 영향 요인 분석 연구는 전공 계열을 구분하지 않고 학생의 개인 특성 정보 및 정량적 모형을 활용하였다. 그러나 실제 전공 계열에 따라 채용 인원수, 시기, 영향요인이 상이하므로 이를 반영하여 모형 구축 및 연구를 수행할 필요가 있다. 이를 위해 본 연구에서는 C 대학 데이터와 의사결정 나무 모형을 활용하여 전공 계열별 취업 영향 요인을 분석하였다. 분석 결과를 바탕으로 대학생의 취업을 위한 로드맵을 추천하였다. 이는 대학의 데이터를 활용하여 학생들의 취업률 제고에 도움을 주는 방향을 제시한 것이다. 또한, 해당 연구를 통해 향후 대학에서 데이터의 수집 및 관리 방안을 모색할 수 있다. 다만, 본 논문에서 수행한 연구는 다음과 같은 한계점이 존재한다. 먼저, 본 논문에서는 특정 대학의 데이터만을 활용하여 타 대학의 학생에게 분석 결과를 그대로 적용하는 것에는 어려움이 있다. 다음으로, 전공을 교육부에서 명시한 계열로만 구분하고 보다 세분화하지 못한 점이다. 이는 같은 계열에 속한 전공일지라도 학과 커리큘럼, 교육방식 등이 상이하므로 취업 영향 요인에 차이가 발생할 수 있다. 따라서 향후 연구에서는 이와 같은 문제점을 보완하여 정교한 모형이 구축되어야 할 것이다.

참고문헌

[1] 김기환·이창호·최보승, “학령인구 감소에 따른 지역별 대입지원자 감소에 대한 예측연구,” 한국데이터정보과학회 논문지, 제26권, 제6호, 2015, pp.1175-1188.

[2] 김명용·박근영, “학령기 인구 감소 극복을 위한 모집 활성화 방안 – D대학 중심으로,” 산업기술연구논문지, 제24권, 제3호, 2019, pp.21-27.

[3] Y. Kim and J. Ahn, “A Study on the Application of Big Data to the Korean College Education System,” Procedia Computer Science, Vol.91, 2016, pp.855-861.

[4] M. Nie, L. Yang, J. Sun, S. Han, X. Hu, D. Lian and K. Yan, “Advanced forecasting of career choices for college students based on campus big data,” Frontiers of Computer Science, Vol.12, 2018, pp.494-503.

[5] 권영옥, “빅데이터를 활용한 맞춤형 교육 서비스 활성화 방안연구,” 지능정보연구, 제19권, 제2호, 2013, pp.87-99.

[6] 박상성, “앙상블 기법을 활용한 대학생 중도탈락 예측 모형 개발,” 디지털산업정보학회 논문지, 제17권, 제1호, 2021, pp.109-115.

[7] 윤수경·한유경, “대학생의 취업성과 영향 요인

분석," 교육재정연구, 제23권, 제4호, 2014, pp.131-160.

[8] 조장식, "학생정보를 이용한 대졸 취업에 미치는 영향력 분석," 한국데이터정보과학회지, 제22권, 제5호, 2011, pp.849-856.

[9] 염동기·문상규·박성수, "대학졸업자의 취업성과 결정요인에 관한 실증분석," 취업진로연구, 제7권, 제4호, 2017, pp.45-68.

[10] S. Safavian and D. Landgreb, "A survey of decision tree classifier methodology," IEEE Transactions on Systems, Man, and Cybernetics, Vol.21, No.3, 1991, pp.660-674.

[11] W. Loh, "Classification and regression trees," WIREs Data Mining and Knowledge Discovery, Vol.1, No.1, 2011, pp.14-23.

■ 저자소개 ■

박 상 성
(Park, Sangsung)

2019년~현재
청주대학교 빅데이터통계학과 조교수
2015년~2019년
고려대학교 기술경영전문대학원
조교수
2006년~2015년
고려대학교 산업경영공학부 연구교수
2006년 고려대학교 산업시스템정보공학과
(공학박사)

관심분야 : Patent Analysis, Data Mining, Management of Technology, Technology Evaluation
E-mail : hanyul@cju.ac.kr

논문접수일 : 2021년 7월 20일
수 정 일 : 2021년 8월 2일
게재확정일 : 2021년 8월 26일