기계 학습 방법을 이용한 직장 생활 프로파일 기반의 퇴직 예측 모델 개발

윤유동† · 이설화† · 지혜성† · 임희석††

요 약

최근 대부분의 기업에서 인적 자원의 유출이 조직에 미칠 부정적인 영향을 인지하게 되면서 조직 구성원의 이직 및 퇴직의도에 대해 많은 연구가 이루어졌다. 그러나 대부분 설문조사의 형태로 이루어지며, 직장 생활 데이터를 기반으로 이직 또는 퇴직의도를 살펴본 연구는 아직까지 미비했다. 이에 본 연구에서는 직장 생활 프로파일을 기반으로 직원의 퇴직 여부에 영향을 미치는 요인에 대한 분석을 실시하고, 기계 학습 방법을 활용하여 퇴직 예측 모델을 생성했다. 이 결과, 기존의 설문조사를 중심으로 수행되었던 연구에서 접근하지 못했던 다양한 요인들을 파악할 수 있었다. 또한, 우수한 성능의 퇴직 예측 모델 생성을 통해 기업의 인적 자원 유출에 대한 해결방안을 제시할 수 있는 연구의 발판을 마련했다.

주제어 : 직장 생활 프로파일, 기계 학습, 연관성 분석, 지도 학습, 분류 알고리즘

Development of Retirement Prediction Model based on Work Life Profile Using Machine Learning Method

You-Dong Yun† · Seol-Hwa Lee† · Hye-Sung Ji† · Heui-Seok Lim††

ABSTRACT

Recently, much research has been done on the turnover and retirement intentions of the organization members as many companies recognize the negative impact of the human resource outflow on the organization. However, most of the studies are conducted in the form of questionnaires, and there is still a lack of studies on the turnover and retirement intentions based on the work life data. In this study, we analyzed the factors affecting the retirement of employees based on the work life profile, and created a retirement prediction model using the machine learning method. As a result, we could identify various factors that were not covered in previous researches. In addition, we have established a basis for research that can provide a solution for the problem of human resource outflow by generating a good performance retirement prediction model.

Keywords : Work Life Profile, Machine Learning, Association Analysis, Supervised Learning, Classification Algorithm

† 정 회 원: 고려대학교 정보대학 컴퓨터학과
†† 종신회원: 고려대학교 정보대학 컴퓨터학과 교수(교신저자)
논문접수: 2016년 12월 19일, 심사완료: 2017년 1월 18일, 게재확정: 2017년 1월 23일
* 본 연구는 문화체육관광부 및 한국콘텐츠 진흥원의 2016년도 문화기술 연구개발 지원 사업으로 수행되었음. (과제번호: R1610671)

1. 서론

조직에서의 구성원들은 조직의 중요한 인적 자원이며, 조직이 필요로 하는 다른 자원을 소유한 주체이기도 하다. 현대사회에서는 과거보다 조직의 인적자원의 중요성이 더욱 증가함에 따라, 인적자원을 조직 내에 유지하는 것이 그 언제보다 중요해졌다. 경쟁이 심화되고 시장의 변화가 가속화되는 등의 급격한 변화에 대응하는 조직변화의 과정에서 조직 구성원의 이직은 활발해질 수 있으며, 노동시장이 점차 유연해짐에 따라 이직 요인은 더욱 증가할 것으로 예상된다. 따라서 조직에서는 중요한 인적 자원을 확보하거나 유지하기 위해 더 많은 노력을 기울일 필요가 있다[1].

인재 확보에 대한 문제는 중소기업뿐만 아니라 대기업 및 세계적 기업들의 경우도 마찬가지이다. 구체적으로 인재확보의 중요성과 함께 많은 기업들이 인재확보를 위해 상당한 노력 및 투자를 해오고 있다. 그럼에도 불구하고 수많은 기업들이 인재 부족 문제로 곤경을 겪고 있다. 이러한 문제의 이유로 우수한 인재 확보의 어려움을 배제할 수는 없으나, 기업 측에서는 확보하고 육성한 우수한 인재들이 기업을 떠나는 것을 더 큰 문제로 보고 있다[2]. 그 이유로는, 인적 자원의 유출에 따른 대체인력 채용과 훈련을 위한 추가비용발생과 함께 구성원의 이직으로 발생하는 직무공백으로 인한 업무차질과 직장 내 사기 저하, 기업 이미지의 실추, 정보 기술의 유출 등의 암묵적인 비용도 포함된다[3]. 실제로 2013년 전국 490개 회사의 직장인들을 대상으로 조사를 실시한 결과, 평균인 이직률이 15.8%로 나타났으며, 기업은 직원들의 이직으로 인하여 이직 근로자 1인 당 약 1,284만원 정도의 금전적 손실을 입고 있다고 밝혔다[4]. 기업은 직원들에게 자원을 투입하여 직원 개인의 발전은 물론 기업의 성장을 추구하게 되는데, 인적 자원의 유출이 발생하게 되면 기업의 비용이 증가하고, 대체인력에 대한 고용과 재적응 등 다양한 문제가 발생하게 된다[5]. 따라서 인적 자원의 유출을 막기 위해 기업은 직원들의 근무 환경을 개선하거나 성과에 기반한 보상 체계를 구축하고, 적성에 맞는 업무를 부여하는 등의 다양한 노력을 기울이게 된다.

대부분의 기업에서 인적 자원의 유출이 조직에 미칠 부정적인 영향을 인지하게 되면서 조직 구성원의 이직 및 퇴직의도에 대한 연구가 많이 이루어졌다. 그동안 이직 및 퇴직의도와 관련한 선행연구들에서는 직무에 대한 만족도, 직무에 대한 스트레스, 정서적 소진, 직무 환경, 보상, 조직몰입, 조직갈등, 조직운영의 합리성, 직장의 유연성, 조직정치 등이 이직 및 퇴직의도에 영향을 미치는 요인으로 검증된 바 있다[6][7][8][9][10][11]. 그러나 대부분의 연구가 주로 설문조사의 형태로 데이터를 수집하여 직장인들의 직장 생활 데이터에 따른 이직 또는 퇴직의도를 살펴본 연구는 아직까지 미비하다고 볼 수 있다. 직장인들의 이직 및 퇴직의도는 직장인들이 설문지를 통해 표출하는 의견을 통해서도 연구가 이루어질 수 있지만, 직원이 수행하는 프로젝트의 수, 직원의 근무 시간, 직원이 직장에서 보내는 시간과 같이 직장인이 실제로 직장 생활에서 겪는 프로파일 데이터를 기반으로도 직장인들의 이직 및 퇴직의도를 살펴볼 수 있다. 그럼에도 불구하고 아직까지 이직 및 퇴직의도를 살펴보는 대부분의 연구에서는 설문조사 데이터에 의지하는 경향이 강하다.

이에 본 연구에서는 직장 생활 프로파일 데이터를 기반으로 직원의 퇴직 여부에 영향을 미치는 요인에 대한 분석을 실시하여 어떠한 요인이 직원의 퇴직 여부에 어떠한 영향을 미치는지를 확인하고, 기계 학습 방법을 활용하여 퇴직 예측 모델을 생성할 수 있도록 한다. 인적 자원의 유출은 기업의 입장에서 금전적인 문제 뿐만 아니라 암묵적인 다양한 문제를 야기한다. 이러한 이유에서 다양한 분석 단계를 거쳐 직원의 퇴직 여부에 영향을 미치는 요인이 무엇인지 확인하고, 이를 기반으로 조직 구성원들의 이직 및 퇴직을 예방할 수 있는 해결방안에 대해 논의하도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 직장 생활 프로파일 데이터의 소개와 함께 연관성 분석과 지도 학습을 어떻게 수행하는지 소개한다. 그리고 3장에서는 연관성 분석과 지도 학습 수행을 통해 어떤 결과가 도출되었는지 확인하고, 4장에서 분석 결과를 기반으로 결과 해석 및 조직 구성원들의 이직 및 퇴직을 예방할 수 있는 해결방안에 대해 논의하여 5장에서 결론을 맺는다.

2. 연구 방법

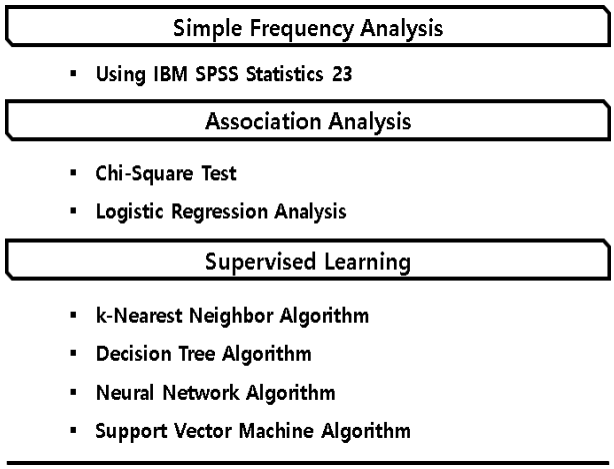
2.1 연구 자료

본 연구에서는 kaggle의 “Human Resources (HR) Analytics 2016” 데이터셋을 활용하였으며, “Human Resources Analytics” 데이터는 직원의 만족도 (Employee satisfaction level), 직원의 마지막 평가 (Last evaluation), 직원의 수행 프로젝트 수 (Number of project), 직원의 월간 평균 근무시간 (Average monthly hours), 직원이 회사에서 보낸 시간 (Time spent at the company), 직원의 업무적 사고 여부 (Whether they have had a work accident), 직원의 지난 5년 동안 승진 여부 (Whether they have had a promothon in the last 5 years), 직원의 부서 (Sales), 직원의 봉급 (Salary), 직원의 퇴직 여부 (Whether the employee has retirement)의 총 10개의 요인과 14,999개의 표본 수로 이루어져 있다.

2.2 연구 방법

본 연구는 “Human Resources Analytics” 데이터의 다양한 변인 중에서 직원의 퇴직 여부에 영향을 미치는 요인을 데이터 분석을 통해 확인한다. “Human Resources Analytics” 데이터에 대한 분석 과정은 [그림 1]과 같이 단순빈도분석, 연관성 분석, 지도 학습의 세 가지 단계로 진행된다.

첫 번째로, 단순빈도분석 (Simple frequency analysis)을 통해 데이터의 구성 및 직원의 퇴직 여부에 따른 군집별 특징을 살펴보도록 한다. 두 번째로, 연관성 분석 (Association analysis) 에서는 우선 카이제곱검정 (Chi-square test)을 통해 다른 요인들과 직원의 퇴직 여부와의 관계가 연관성이 있음을 확인한다. 그리고 카이제곱검정을 통해 직원의 퇴직 여부와 연관성이 검증된 요인들을 대상으로 로지스틱 회귀분석 (Logistic regression analysis)을 통해 직원의 퇴직 여부에 어떠한 영향을 미치는지 확인한다. 세 번째로, 지도 학습 (Supervised learning) 에서는 kNN, Decision Tree, Neuram Network, SVM의 4가지 분류 알고리즘을 활용하여 직장 생활 프로파일 기반의 퇴직 예측 모델을 생성하도록 한다.



[그림 1] 연구 단계

3. 연구 결과

3.1 단순빈도분석

다음 <표 1>은 10개 요인의 단순빈도분석에 대한 결과를 표로 나타낸 것이다. 직원 만족도 (satisfaction_level)에서는 만족도 수치가 높은 직원 (0.60~0.79)이 4,239(28.27%)명으로 가장 많았으며, 만족도 수치가 매우 낮은 직원 (0.00~0.19)이 1,409(9.4%)명으로 가장 적게 나타났다. 마지막 평가 (last_ evaluation)에서는 만족도 수치가 높은 직원 (0.60~0.79)이 5,740(38.3%)명으로 가장 많았으며, 만족도 수치가 매우 높은 직원 (0.80~1.00)이 0(0%)명으로 가장 적게 나타났다. 직원이 수행하는 프로젝트 수 (number_project)에서는 프로젝트를 4개 수행하는 직원이 4,365(29.1%)명으로 가장 많았으며, 프로젝트를 7개 수행하는 직원이 256(1.7%)명으로 가장 적게 나타났다. 직원의 월간 평균 근무 시간 (average_montly_hours)에서는 월간 평균적으로 120~159시간 근무하는 직원이 3,815(25.4%)명으로 가장 많았으며, 월간 평균적으로 120시간 미만으로 근무하는 직원이 389(2.6%)명으로 가장 적게 나타났다. 직원이 회사에서 보낸 시간 (time_spend_company)에서는 회사에서 3년을 일해온 직원이 6,443(43.0%)명으로 가장 많았으며, 회사에서 8년을 일해온 직원이 162(1.1%)명으로 가장 적게 나타났다. 직원의 업무적 사고 여부 (work_accident)에서는 업무적 사고를 내지 않은 직원이 12,830(85.5%)명으로 업무

적 사고를 낸 직원보다 많은 것으로 나타났다. 직원의 지난 5년 동안의 승진 여부 (promotion_last_5years)에서는 승진하지 못한 직원이 14,680(97.9%)명으로 승진한 직원보다 많은 것으로 나타났다. 부서 (sales)에서는 sales 부서에 속한 직원이 4,140(27.6%)명으로 가장 많은 것으로 나타났으며, management 부서에 속한 직원이 630(4.2%)명으로 가장 적게 나타났다. 직원의 봉급 (salary)에서는 낮은 봉급 (low)을 받는 직원이 7,316(48.8%)명으로 가장 많았으며, 높은 봉급 (high)을 받는 직원이 1,237(8.2%)명으로 가장 적게 나타났다. 마지막으로 직원의 퇴직 여부 (retirement)에서는 퇴직을 하지 않은 직원이 11,428(76.2%)명으로 퇴직 한 직원보다 많은 것으로 나타났다.

<표1> 단순빈도분석 결과

Feature	Explanation	Frequency(%)
satisfaction_level	① 0.80~1.00	4,224(28.2)
	② 0.60~0.79	4,239(28.3)
	③ 0.40~0.59	3,621(24.1)
	④ 0.20~0.39	1,506(10.0)
	⑤ ~0.19	1,409(9.4)
last_evaluation	① 0.80~1.00	0(0)
	② 0.60~0.79	5,740(38.3)
	③ 0.40~0.59	4,517(30.1)
	④ 0.20~0.39	4,563(30.4)
	⑤ 0.00~0.19	179(1.2)
number_project	① 2	2,388(15.9)
	② 3	4,055(27.0)
	③ 4	4,365(29.1)
	④ 5	2,761(18.4)
	⑤ 6	1,174(7.8)
	⑥ 7	256(1.7)
average_montly_hours	① 280~	591(3.9)
	② 240~279	3,710(24.7)
	③ 200~239	3,232(21.5)
	④ 160~199	3,262(21.7)
	⑤ 120~159	3,815(25.4)
	⑥ ~119	389(2.6)
time_spend_company	① 2	3,244(21.6)
	② 3	6,443(43.0)
	③ 4	2,557(17.0)
	④ 5	1,473(9.8)
	⑤ 6	718(4.8)
	⑥ 7	188(1.3)
	⑦ 8	162(1.1)
	⑧ 10	214(1.4)
work_accident	① 0	12,830(85.5)
	② 1	2,169(14.5)
promotion_last_5years	① 0	14,680(97.9)
	② 1	319(2.1)
sales	① accounting	767(5.1)
	② hr	739(4.9)
	③ IT	1227(8.2)
	④ management	630(4.2)
	⑤ marketing	858(5.7)
	⑥ product_mng	902(6.0)
	⑦ RandD	787(5.2)
	⑧ sales	4,140(27.6)
	⑨ support	2,229(14.9)
	⑩ technical	2,720(18.1)
salary	① high	1,237(8.2)
	② medium	6,446(43.0)
	③ low	7,316(48.8)
retirement	① 0	11,428(76.2)
	② 1	3,571(23.8)

3.1.1 군집별 특징

단순빈도분석 결과를 기반으로 직원의 퇴직 여부에 따라 퇴직하지 않은 직원과 퇴직한 직원이 각 요인에서 어떠한 특징을 갖는지 살펴보았다.

퇴직하지 않은 직원은 대부분 직원 만족도가 좋은 것으로 나타났으며 (0.6 이상 65.5%), 퇴직한 직원은 직원 만족도가 비교적 낮은 것으로 나타났으나 (0.4 미만 43.9%), 직원의 마지막 평가에서는 두 군집 모두 대부분 평균 이상의 평가를 한 것으로 나타났다 (퇴직하지 않은 직원 0.6이상 72.6%, 퇴직한 직원 0.6 이상 55.0%). 직원이 수행하는 프로젝트 수에서는 퇴직하지 않은 직원은 3-5개로 너무 많거나 적지 않은 프로젝트를 수행하는 반면 (3-5개 88.3%), 퇴직하는 직원은 너무 적은 프로젝트를 수행하거나 너무 많은 프로젝트를 수행하는 것으로 나타났다 (2개 43.9%, 6-7개 25.5%). 직원의 월간 평균 근무 시간에서 퇴직하지 않은 직원들은 120~279시간 안에서 다양하게 분포되었으나 (120~159시간 23.0%, 160~199시간 25%, 200~239시간 27.4%, 240~279시간 20%) 퇴직하는 직원은 240~279시간과 120~159시간에 편중된 것으로 보아 (120~159시간 30.1%, 240~279시간 42.8%), 평균 근무 시간이 너무 많거나, 너무 적은 경우 퇴직할 가능성이 높아질 수 있다고 해석할 수 있었다. 퇴직하는 직원과 퇴직하지 않는 직원 두 군집 모두 회사에서 보낸 시간이 적은 사람이 많았으며 (퇴직하지 않은 직원 2-4년 85.0%, 퇴직한 직원 2-4년 70.8%), 두 군집 모두 지난 5년 동안의 승진 여부 역시 승진하지 못한 사람이

많았으나 (퇴직하지 않은 직원 97.4%, 퇴직한 직원 99.5%), 퇴직하는 직원의 승진 여부가 더 적은 것으로 나타났다. 직원의 업무적 사고 여부에서는 퇴직하는 직원이 퇴직하지 않는 직원보다 더 많은 것으로 나타났다 (퇴직하지 않은 직원 82.5%, 퇴직한 직원 95.3%). 마지막으로 직원의 봉급에서는 퇴직하는 직원의 봉급이 적은 사례가 더 많은 것으로 나타났으며 (퇴직하지 않은 직원의 low salary 45.0%, 퇴직한 직원의 low salary 60.8%), 직원의 부서에 따른 군집별 비율의 차이는 크게 드러나지 않았다.

3.2 연관성 분석

본 연구에서는 연관성 분석을 위하여 카이제곱 검정을 통해 연관성을 확인하고 로지스틱 회귀분석을 통해 영향의 정도를 확인한다. 카이제곱검정은 카이제곱 분포에 기초한 통계적 방법으로서, 관찰된 빈도가 기대되는 빈도와 의미 있게 다른지의 여부를 검증하기 위해 사용되는 검증방법이며, 로지스틱 회귀분석은 종속 변수와 독립 변수 간의 관계를 구체적인 함수로 나타내어 향후 예측 모델에 사용하기 위한 분석방법이다[12][13].

3.2.1 카이제곱검정

다음 <표 2>는 직원의 퇴직 여부에 대한 카이제곱검정 결과를 표로 나타낸 것이다. 카이제곱검정은 카이제곱 분포를 기반으로 수행하는 통계적 방법으로, 관찰된 빈도가 기대되는 빈도와 의미 있게 다른지에 대한 여부를 검증하기 위해 사용하는 검증방법이다. 카이제곱검정 결과, 직원의 만족도 (p<.001), 직원의 마지막 평가 (p<.001), 직원이 수행하는 프로젝트 수 (p<.001), 직원의 월간 평균 근무 시간 (p<.001), 직원이 회사에서 보낸 시간 (p<.001), 직원의 업무적 사고 여부 (p<.001), 직원의 지난 5년 동안의 승진 여부 (p<.001), 직원의 부서 (p<.001), 직원의 봉급 (p<.001)까지 모든 요인이 직원의 퇴직 여부와 매우 높은 연관성을 가지고 있는 것으로 나타났다.

<표 2> 카이제곱검정 결과

Dependent Variable	Independent Variable	x^2	p-value
Employee Satisfaction Level	satisfaction_level	7937.744	.000***
	last_evaluation	2534.842	.000***
	number_project	5373.586	.000***
	average_montly_hours	3623.054	.000***
	time_spend_company	2110.080	.000***
	work_accident	358.594	.000***
	promotion_last_5years	57.263	.000***
	sales	86.825	.000***
	salary	381.225	.000***

* p < .05, ** p < .01, *** p < .001

3.2.2 로지스틱 회귀분석

다음 <표 3>은 카이제곱검정 결과를 기반으로 직원의 퇴직 여부와의 연관성이 검증된 요인에 대해 로지스틱 회귀분석을 실시한 결과이다. 로지스틱 회귀분석은 종속 변수와 독립 변수간 관계를 구체적 함수로 나타내어 예측 모델에 사용하기 위한 분석방법이다[14][15]. 분석 결과 생성된 로지스틱 회귀모형은 각 요인이 직원 만족도에 어떠한 영향을 미치고 있는지를 제시하고 있다.

로지스틱 회귀모형에 의하면 직원의 만족도가 높아질수록 (OR=0.016, p<.001), 직원의 마지막 평가가 낮아질수록 (OR=2.065, p<.001), 직원이 수행하는 프로젝트 수가 적을수록 (OR=0.732,

<표 3> 로지스틱 회귀분석 결과

Whether the employee has retirement			
Feature	p-value	OR	95% CI
satisfaction_level	.000***	0.016	0.013~0.020
last_evaluation	.000***	2.065	1.511~2.762
number_project	.000***	0.732	0.702~0.763
average_montly_hours	.000***	1.004	1.003~1.005
time_spend_company	.000***	1.293	1.255~1.332
work_accident	.000***	0.215	0.180~0.256
promotion_last_5years	.000***	0.221	0.133~0.365
sales	.138	1.012	0.996~1.028
salary	.000***	2.018	1.874~2.174
Cox & Snell R-square	0.209		
Nagelkerke R-square	0.313		
Chi-square	3513.074		

* p < .05, ** p < .01, *** p < .001

p<.001) 직원이 퇴직할 확률이 높은 것으로 나타났다. 그리고 직원의 월간 평균 근무시간이 낮아질수록 (OR=1.004, p<.001), 직원이 회사에서 보낸 시간이 많을수록 (OR=1.293, p<.001), 직원이 업무적 사고를 치지 않았을수록 (OR=0.215, p<.001) 직원이 퇴직할 확률이 높은 것으로 나타났다. 마지막으로 직원이 지난 5년 동안 승진을 하지 못할수록 (OR=0.221, p<.001), 직원의 봉급이 낮아질수록 (OR=2.018, p<.001) 직원이 퇴직할 확률이 높은 것으로 나타났다. 그러나, 직원이 속한 부서는 직원의 퇴직 여부에 영향을 미치지 못하는 것으로 나타났다.

로지스틱 회귀모형에서 직원이 속한 부서를 제외하고는 모든 요인이 직원의 퇴직 여부에 영향을 미치는 것으로 나타났다. 이 모형을 검증하기 위해 Cox&Snell의 R-sqare 수치와 Nagelkerke R-sqare 수치를 사용하였다. Cox&Snell R-sqare 수치는 0.209, Nagelkerke R-sqare 수치는 0.313으로 전체 반응 변수의 변동 중 31.3%을 모형이 설명하고 있다고 해석할 수 있다.

3.3 지도 학습

지도 학습 (Supervised Learning)은 테스트 데이터로부터 하나의 함수를 예측하기 위한 기계학습의 한 방법이다[16]. 지도 학습에서는 직원의 퇴직 여부를 제외한 9개의 요인들을 기반으로 기계 학습 (Machine Learning)방법의 일종인 4가지의 분류 알고리즘 (Classification Algorithm)을 이용하여 데이터를 학습시켜 직원의 퇴직 여부를 예측할 수 있는 모델을 생성한다. 분류 알고리즘은 데이터의 크기나 품질, 특성에 따라서 다른 결과가 나타날 수 있다. 따라서 본 연구의 퇴직 예측 모델은 kNN, Decision Tree, Neural Network, SVM의 네 가지 알고리즘을 사용하여 분류 정확도를 비교분석하여 정확도가 가장 높은 알고리즘을 확인할 수 있도록 한다. 분류 알고리즘을 수행하기에 앞서 테스트 데이터는 전체 데이터의 10% 수준인 1,500명으로 설정하였으며, 더욱 신뢰도 있는 분석 결과 도출을 위해 직원의 퇴직 여부(퇴직하지 않는 직원, 퇴직하는 직원)에 따라 각각 10% 수준으로 테스트 데이터를 구성하였다.

3.3.1 k-Nearest Neighbor

kNN (k-Nearest Neighbor) 알고리즘은 훈련 데이터와 같은 속성을 가졌으나 분류되지 않은 데이터에서 유클리디안 거리 (Euclidean distance)를 활용하여 훈련 데이터와 가장 가까운 곳에 위치한 데이터 k개를 추출하고, 추출된 데이터의 클래스를 통해 분류되지 않은 데이터의 범주를 지정해 주는 방법이다[17]. kNN 알고리즘에서는 직원의 퇴직 여부에 따라 퇴직하지 않은 직원과 퇴직한 직원을 각각 1,100명, 400명으로 총 1,500명의 테스트 데이터를 구성하였다. 그리고 k 이웃의 수는 총 표본 수인 14,999명의 제곱근에 가장 근접한 수인 122로 설정하였다.

분석 결과, 퇴직을 하지 않은 군집에서는 1,100명 중 1,035명을 일치하게 분류하여 94.1%의 분류 정확도를 보였으며, 퇴직을 한 군집에서는 400명 중 351명을 일치하게 분류하여 87.7%의 분류 정확도를 보여, 전체적인 분류 정확도는 92.4%로 나타났다. kNN 알고리즘 분석 결과에 대한 자세한 내용은 다음 <표 4>에 나타나 있다.

<표 4> kNN 알고리즘 분석 결과

Test Data Correct Set	Non-Resign (%)	Resign (%)	Row Total (%)
Non-Resign	1,035 (0.941)	65 (0.059)	1,100 (0.733)
Resign	49 (0.122)	351 (0.877)	400 (0.267)
Column Total	1,084	416	1,500
Accuracy	92.4 %		

3.3.2 Decision Tree

Decision Tree 알고리즘은 의사결정규칙을 나무구조로 도표화하여 분류 및 예측을 수행하는 분석 방법으로, 입력 변수를 기반으로 목표 변수의 값을 예측하는 모델을 생성할 수 있다. 특히 Decision Tree 알고리즘은 저렴한 계산에 비해 합리적인 분류 정확도를 얻을 수 있기 때문에 데이터 마이닝 영역에서 매우 유용한 알고리즘으로서 활용되고 있다[18][19]. Decision Tree 알고리즘에서는 직원의 퇴직 여부에 따라 퇴직하지 않은 직원을 1,139명, 퇴직한 직원을 361명으로 총 1,500명의 테스트 데이터를 구성하였다.

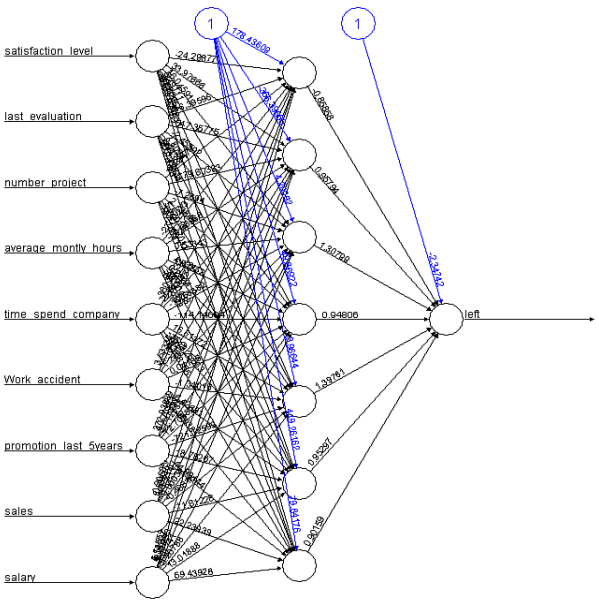
분석 결과, 트리의 크기 (Tree Size)는 45로 나타났으며, 퇴직을 하지 않은 군집에서는 1,139명 중 1,132명을 일치하게 분류하여 99.4%의 분류 정확도를 보였으며, 퇴직을 한 군집에서는 361명 중 334명을 일치하게 분류하여 92.5%의 분류 정확도를 보여, 전체적인 분류 정확도는 97.7%로 나타났다. Decision Tree 알고리즘 분석 결과에 대한 자세한 내용은 다음 <표 5>에 나타나 있다.

<표 5> Decision Tree 알고리즘 분석 결과

Test Data Correct Set	Non-Resign (%)	Resign (%)	Row Total (%)
Non-Resign	1,132 (0.994)	7 (0.006)	1,139 (0.759)
Resign	27 (0.075)	334 (0.925)	361 (0.241)
Column Total	1,159	341	1,500
Accuracy	97.7 %		

3.3.3 Neural Network

Neural Network 알고리즘은 인간의 뇌에 존재하는 생물학적 신경세포를 모방해 프로그램화 한 수학 모델로, 일반적으로 과거에 수집된 데이터로부터 반복적인 학습과정을 통해 패턴을 발견하여 새로운 데이터에 대한 목표값을 예측하는데 사용된다[20][21]. Neural Network 알고리즘에서는 13,499명의 훈련 데이터, 1,500명의 테스트 데이터



[그림 2] Neural Network 알고리즘 분석 결과

통해 분류를 수행하였다. 은닉 노드의 수는 7개를 초과하였을 때 과도한 은닉 노드의 수로 인해 오히려 정확률이 감소하는 양상을 확인하였으며, 연구 데이터에 과적합 될 가능성을 고려하여 은닉 노드의 수를 7개로 설정하여 분석을 진행하였다. 은닉 노드를 7개로 설정하여 Neural Network 알고리즘을 사용하여 분석한 결과, 전체적인 분류 정확도는 92.8%로 나타났다. Neural Nework 알고리즘 분석 결과에 대한 내용은 다음 [그림 2]에서 자세하게 나타나 있다.

3.3.4 Support Vector Machine

SVM 알고리즘은 간단하게 말해서, 특징 공간에서 트레이닝 샘플의 두 클래스 사이에서 최상의 분리 초평면 (hyperplane)을 찾는 것이다. SVM 알고리즘은 커널 기반의 학습 알고리즘으로 적은 학습 데이터를 통해서도 높은 차원의 공간을 일반화할 수 있다. 즉, 학습과정에서 이용되지 않은 새로운 데이터 표본에 대해서도 올바른 분류가 가능하다는 것이다[22][23]. SVM 알고리즘에서는 직원의 퇴직 여부에 따라 퇴직하지 않은 직원과 퇴직한 직원을 각각 1,100명, 400명으로 총 1,500명의 테스트 데이터를 구성하였다. 그리고 SVM 알고리즘의 분류 능력의 대부분은 커널의 선택으로부터 나타나게 되는데, 본 연구에서는 SVM 분석을 위해 Radial Basis 커널 (Kernel)을 사용하여 분석을 실시하였다[24].

<표 6> SVM 알고리즘 분석 결과

Test Data Correct Set	Non-Resign (%)	Resign (%)	Row Total (%)
Non-Resign	1,078 (0.980)	22 (0.020)	1,100 (0.733)
Resign	28 (0.019)	372 (0.930)	361 (0.241)
Column Total	1,106	394	1,500
Accuracy	96.7 %		

분석 결과, 퇴직을 하지 않은 군집에서는 1,100명 중 1,078명을 일치하게 분류하여 98.0%의 분류 정확도를 보였으며, 퇴직을 한 군집에서는 400명 중 372명을 일치하게 분류하여 93.0%의 분류 정확도를 보여, 전체적인 분류 정확도는 96.7%로 나

타났다. SVM 알고리즘 분석 결과에 대한 자세한 내용은 상단의 <표 6>에 나타나 있다.

kNN, Decision Tree, Neural Network, SVM의 4가지의 분류 알고리즘을 기반으로 분석을 수행한 결과, 모든 알고리즘에서 90% 이상의 높은 분류 정확도를 보여주었다. 그 중에서도 Decision Tree 알고리즘이 97.7%의 분류 정확도로 다른 알고리즘에 비해 가장 높은 분류 정확도를 보이는 것으로 나타났다. 이에 대한 내용은 다음 <표 7>에 자세하게 나타나 있다.

<표 7> 4가지 분류 알고리즘 분석 결과

No.	Algorithm	Accuracy
1	k-Nearest Neighbor	92.4 %
2	Decision Tree	97.7 %
3	Neural Network	92.8 %
4	Support Vector Machine	96.7 %
Average Accuracy = 94.9 %		

4. 결과 해석 및 제언

본 연구에서는 직장 생활 프로파일 데이터를 기반으로 직원의 퇴직 여부에 영향을 미치는 요인을 탐색하기 위하여 카이제곱검정, 로지스틱 회귀분석을 활용하였으며, 퇴직 예측 모델 생성을 위해 4가지 분류 알고리즘을 사용하였다. 알고리즘 분석에 대한 결과는 다음과 같다.

첫 번째로, 카이제곱분석을 통해 직원의 퇴직 여부에 연관성이 있는 변수에 대해 분석하였을 때 모든 요인이 직원의 퇴직 여부와 연관성이 있다는 것을 확인하였다.

두 번째로, 카이제곱분석을 통해 검증된 요인을 로지스틱 회귀분석을 활용하여 직원의 퇴직 여부에 영향을 미치는 공통 변수를 확인하였을 때 직원이 속한 부서를 제외한 모든 요인이 직원의 퇴직 여부에 영향을 미치는 것으로 나타났다.

세 번째로, 지도 학습의 4가지 분류 알고리즘을 활용하여 데이터를 학습시켜 직원의 퇴직 여부를 예측할 수 있는 모델을 생성하여 알고리즘별 분류 정확도를 비교하였다. 분류 정확도는 모든 알고리즘에서 90% 이상으로 높게 나타났으나, 특히 Decision Tree 알고리즘의 분류 모델에서 97.7%의 높은 분류 정확도를 보여주었다. 우수한 성능의 분류 모델은 직원들의 퇴직 여부를 예측하여, 그에 대한 해결 방안을 미리 마련할 수 있다.

이러한 결과를 바탕으로 직원의 이직 및 퇴직을 예방할 수 있는 해결방안은 다음과 같다. 첫째, 직원이 수행하는 프로젝트 수와 직원의 월간 평균 근무시간이 과도하게 적은 경우 퇴직할 확률이 증가하는 것으로 나타났다. 그렇기 때문에 직원들에게 어느 정도의 프로젝트를 수행하도록 하여 직장 내 활동에 대해 어느 정도 동기부여를 해 줄 필요가 있을 것으로 보인다. 또한, 어느 정도의 프로젝트를 담당하게 되면 어느 정도의 동기부여와 함께 자연스럽게 직원의 월간 평균 근무 시간이 증가하게 되어 퇴직할 확률이 감소할 것으로 판단된다. 둘째, 직원이 회사에서 보낸 시간이 많아질수록 퇴직할 확률이 증가하는 것으로 나타났다. 이와 같은 분석 결과를 통해 오랜 시간 회사에서 근무한 사람일수록 업무 환경의 개선과 업무에 대한 동기부여를 지속적으로 제공할 필요가 있으며, 오랜 시간 회사에서 근무한 직원에 대해 이에 대한 적절한 보상이 제공되어야 할 것으로 보인다. 셋째, 직원이 오랜 시간 동안 승진을 하지 못하거나 직원의 봉급이 낮아질수록 퇴직할 확률이 증가하는 것으로 나타났다. 이와 같은 분석 결과를 통해 직원들에게 성과에 따른 승진 및 봉급을 보장해 줄 필요가 있을 것으로 판단된다.

이 외에도 큰 차이를 보이지는 않았으나 (OR=0.016), 직원의 만족도가 높아질수록 직원이 퇴직할 확률이 높은 것으로 나타났으며, 매우 큰 차이로 (OR=2.065) 직원의 마지막 평가가 낮아질수록 직원이 퇴직할 확률이 높은 것으로 나타났다. 이는 직장인들의 감정노동 수행전략 방법 중 “표면연기 (Surdace Acting)”와 관련이 있을 것으로 판단된다. 표면연기는 자신의 내적 감정을 변화시키지 않고 외적인 표현에서 조직이 요구하는 표현규칙에 어긋나지 않도록 행동하는 것을 의미한다[25]. 이에 따라 분석된 결과를 해석하자면 다음과 같다. 직장인들이 평소 만족도 평가에서 겉으로는 직장의 만족도에 대한 부정적인 의견을 드러내지는 않으나, 퇴직에 가까워질수록 자신의 부정적인 내적 감정이 드러나 마지막 평가에서 표출된다고 해석할 수 있다. 그리고 직원의 업무적 사고 측면

에서는, 큰 차이로 업무적 사고를 치지 않았을수록 직원이 퇴직할 확률이 높은 것으로 나타났다.

5. 결론

본 연구는 “Human Resources Analytics” 데이터를 활용하여 직원의 퇴직 의도에 영향을 미치는 요인을 다양한 통계적인 분석 방법 및 기계 학습 방법을 활용하여 살펴보았다. 본 연구에서는 직장인이 실제로 직장 생활에서 겪는 프로파일 데이터를 기반으로 분석을 수행하여 기존의 설문조사를 중심으로 수행되었던 연구에서 접근하지 못했던 다양한 요인들을 파악할 수 있었다. 또한, 기계 학습 방법의 일종인 분류 알고리즘을 기반으로 우수한 성능의 퇴직 예측 모델을 생성하였다. 이러한 분류 모델은 직원들의 퇴직 여부를 미리 예측하여, 조직의 인적 자원 유출에 대한 해결 방안을 마련할 수 있도록 할 수 있다. 즉, 본 연구에서는 이러한 분류 모델을 통해 직원의 퇴직 여부를 미리 예측하여 기업의 인적 자원 유출로 인한 부정적인 영향을 사전에 방지할 수 있는 연구에 대한 발판을 마련하였다고 할 수 있다.

본 연구의 한계 및 향후 연구과제는 다음과 같다. 본 연구에서는 14,999명이라는 비교적 많지 않은직장 생활 프로파일 데이터를 기반으로 분석을 수행하여 전국에 분포하는 수많은 기업 및 직장인들에게 일반화 시키기에는 어려움이 있을 수 있다. 또한, 본 연구에서 사용한 직장 생활 프로파일 데이터 요인들의 세분화가 깊지 않아 직원의 퇴직 의도에 영향을 미치는 요인들에 대한 깊은 연구에 어려움이 존재했다. 향후 더욱 세분화된 직장 생활 요인들과 함께 더 많은 데이터를 확보하여 연구를 진행한다면, 지금보다 발전된 결과를 기반으로 퇴직 예측 모델의 일반화 및 응용 범위를 넓힐 수 있을 것으로 기대한다.

참 고 문 헌

[1] 조경순 (2006). 조직구성원의 이직의도에 대한 변화몰입의 효과 : 국내금융기관의 인수합병 상황에 대한 분석. **인적자원관리연구. 13**(1), 167-182.

[2] 안관영 (2007). 경제적·심리적 요인과 이직의도의 관계에 대한 연구-외식업 종사자를 중심으로. **경영교육연구. 48**(-), 241-257.

[3] 성지미·안주엽 (2016). 일자리 만족도와 이직의사 및 이직-청년층을 중심으로. **한국산업노동연구. 22**(2), 135-179.

[4] 윤명숙·이희정 (2015). 직장인의 분노가 이직의도에 미치는 영향. **인적자원관리연구. 22**(1), 249-269.

[5] 이수연·양해술 (2008). 콜센터 근로자의 감정노동과 감정소진 및 이직의도의 관계에 대한 연구. **한국콘텐츠학회논문지. 8**(4), 197-210.

[6] 서종수 (2016). 조직몰입이 이직의도와 사업성과에 미치는 영향. **벤처창업연구. 11**(4), 215-225.

[7] 현선해·윤기혁·최세경 (2016). 직무만족과 처우불공정 지각이 조직구성원의 이직의도에 미치는 영향. **한국조직학회보. 13**(3), 1-20.

[8] 김양신·이영민 (2015). 대졸 여성 초기경력자의 직무스트레스, 정서적 소진, 조직사회화가 이직의도에 미치는 영향. **경영컨설팅연구. 15**(1), 109-121.

[9] 양현철·정현선·박동건 (2013). 직장 유연성이 신입사원급 직장인들의 이직의도와 혁신적 업무행동에 미치는 영향. **한국심리학회지 산업 및 조직. 26**(1), 149-176.

[10] 장진혁·유태용 (2013). 조직 내 정치적 행동 지각과 이직의도 간의 관계-스트레스, 조직몰입의 매개효과와 정직성의 조절효과. **한국심리학회지 산업 및 조직. 26**(3), 413-436.

[11] 강광석·박계홍·최영근 (2012). 조직 내 부정적인 행태들과 구성원의 이직의도 간 관계에서 상사-부하 교환관계의 질과 조직지원인식의 조절효과에 관한 연구. **상업교육연구. 26**(2), 155-181.

[12] 윤유동·지혜성·임희석 (2016). 청소년 시기의 인터넷 사용에 영향을 미치는 요인 분석 연구, **컴퓨터교육학회논문지. 19**(5), 55-71.

[13] 정재근 (2011). 부모의 사회경제적 지위와 청소년의 인터넷 이용행태 : 생활시간조사의 활용. **한국사회학. 45**(5), 197-225.

[14] 김명종 (2012). 로지스틱 회귀분석과 인공신경망을 적용한 내부회계관리제도 평가모형의 성과비교. **국제회계연구. 46**(-), 1-30.

[15] 권영란 (2010). 의사결정나무분석과 로지스틱 회귀분석을 이용한 중학생 자살생각 예측요인 비교연구. **한국자료분석학회. 12**(6), 3103-3115.

[16] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. **In Proceedings of the 22nd international conference on Machine learning**, 625-632. ACM.

[17] Islam, M. J., Wu, Q. J., Ahmadi, M., & Sid-Ahmed, M. A. (2007). Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. **In Convergence Information Technology, 2007. International Conference on**, 1541-1546. IEEE.

[18] Du, W., & Zhan, Z. (2002). Building decision tree classifier on private data. **In Proceedings of the IEEE international conference on Privacy, security and data mining. 14**(-), 1-8. Australian Computer Society, Inc..

[19] Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. **Remote sensing of environment. 86**(4), 554-565.

[20] Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. **Neural networks. 21**(2), 427-436.

[21] Saxena, A., & Saad, A. (2007). Evolving an artificial neural network classifier for condition monitoring of rotating mechanical systems. **Applied Soft Computing. 7**(1), 441-454.

[22] Mavroforakis, M. E., & Theodoridis, S. (2006). A geometric approach to support vector machine (SVM) classification. **IEEE transactions on neural networks. 17**(3), 671-682.

[23] Muller, K. R., Mika, S., Ratsch, G., Tsuda, K., & Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. **IEEE transactions on neural networks. 12**(2), 181-201.

[24] Moreno, P. J., Ho, P. P., & Vasconcelos, N. (2003). A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. **In Advances in neural information processing systems**.

[25] Zapf, D. (2002). Emotion work and psychological well-being: A review of the literature and some conceptual considerations. **Human resource management review. 12**(2), 237-268.

윤 유 동

2015 목원대학교 마케팅 정보컨설팅학과(경제학사)
2015～현재 고려대학교 컴퓨터 학과 석사과정
관심분야: 학습 분석, 데이터마이닝, e-learning
E-Mail: 2015010492@korea.ac.kr

이 설 화

2015 백석대학교 소프트웨어 학과(이학학사)
2015～현재 고려대학교 컴퓨터 학과 석사과정
관심분야: 자연어처리, 데이터마이닝, 컴퓨터교육
E-Mail: whiteldark@korea.ac.kr

지 혜 성

2009 한신대학교 소프트웨어 학과(이학학사)
2011 고려대학교 컴퓨터교육 학과(이학석사)
2011～현재 고려대학교 컴퓨터교육과 박사과정
관심분야: 정보검색, 자연어처리, 컴퓨터교육
E-Mail: hyesung84@korea.ac.kr

임 희 석

1992 고려대학교 컴퓨터학과(이학학사)
1994 고려대학교 컴퓨터학과(이학석사)
1997 고려대학교 컴퓨터학과(이학박사)
2008～현재 고려대학교 사범대학 컴퓨터교육과 교수
관심분야: 자연어처리, 뇌신경 언어 정보 처리
E-Mail: limhseok@korea.ac.kr