HR 데이터 기반의 퇴사 예측 모델 개발

박연정 · 이도길*

고려대학교

Development of a Resignation Prediction Model using HR Data

YUNJUNG PARK · Do-Gil Lee*

Korea University

E-mail : pyj9403@korea.ac.kr / motdg@korea.ac.kr

요 약

대부분의 기업에서는 우수한 인적 자원의 유출을 방지하기 위해 직원들이 이직 및 퇴사하는 이유를 연구한다. 이에 기업은 직원이 퇴사하기 전에 면담을 하거나 설문조사를 통해서 연구에 필요한 데이터를 얻는다. 하지만 설문조사에서는 직원들이 직장 생활을 하는 데에 불리할 수도 있는 의견을 드러내려고 하지 않아 정확한 결과를 얻기 힘든 것이 현실이다. 한편, 한국노동연구원에서 발표한 자료에 따르면 기업이 요구하는 최소 학력 수준과 직원의 학력 수준 간의 차이가 클수록 이직 경향이 커진다. 따라서 본 연구에서는 한국노동연구원의 자료에 착안하여, 직원이 가지고 있는 객관적 데이터인 전공, 교육수준, 재직 중인 회사 유형 등의 데이터를 기반으로 직원의 퇴사 여부를 예측하고자 한다. 퇴사 예측 모델을 생성하기 Decision Tree, XGBoost, kNN, SVM을 활용하였으며 각각의 성능을 비교했다. 이 결과, 지금까지 설문조사로 진행되었던 연구에서 파악하지 못한 다양한 요인을 알아낼 수 있었다. 이를 통해 기업이 퇴사 예측 모델을 이용하여 직원이 퇴사하기 전에 미리 이를 인지하고 방지하는 데에 도움을 줄 수 있을 것으로 예상된다.

ABSTRACT

Most companies study why employees resign their jobs to prevent the outflow of excellent human resources. To obtain the data needed for the study, employees are interviewed or surveyed before resignation. However, it is difficult to get accurate results because employees do not want to express their opinions that may be disadvantageous to working in a survey. Meanwhile, according to the data released by the Korea Labor Institute, the greater the difference between the minimum level of education required by companies and the level of employees' academic background, the greater the tendency to resign jobs. Therefore, based on these data, in this study, we would like to predict whether employees will leave the company based on data such as major, education level and company type. We generate four kinds of resignation prediction models using Decision Tree, XGBoost, kNN and SVM, and compared their respective performance. As a result, we could identify various factors that were not covered in previous study. It is expected that the resignation prediction model help companies recognize employees who intend to leave the company in advance.

키워드

Resignation Prediction, HR Data, XGBoost, kNN, SVM

Ⅰ. 서 론

인적 자본 이론에서 직원의 교육훈련에 대한 투자는 기술과 능력을 향상시키고 개발시켜 직원의 생산성을 증가시킨다고 주장한다.[1][2] 기업이 인적 자본을 축적하기 위해서는 비용이 수반되며, 축적된 인적자본이 생산을 위해 사용될 때 그 효과가 나타나는 것이다. 하지만 직원의 이직이나 퇴사

* corresponding author

는 기업의 입장에서는 투자된 인적자본이 생산 활동에 사용되지 않게 되어 비용이 발생될 뿐만 아니라, 결과적으로 성과 창출에 부정적인 영향을 미치게 된다.[3] 최근, 대부분의 기업에서 인적 자본 이론이 주장하는 바와 같이 인적 자원의 유출이 조직에 미치는 부정적인 영향을 인지하고 이를 방지하기 위해 직원들이 이직 및 퇴사하는 이유를 연구한다. 기업에서는 직원이 퇴사하기 전에 면담을 하거나 설문조사를 통해 연구에 필요한 데이터를 얻는다. 하지만 설문조사에서는 직원들이 직장생활을 하는 데에 불리할 수도 있는 의견을 드러내려고 하지 않아 정확한 결과를 얻기 힘든 것이 현실이다.

한편, 한국노동연구원에서 발표한 자료에 따르면 현재 직장에서 이직을 준비하는 취업자 비율은 2014 ~ 2016년은 15% 내외이지만 2017 ~ 2018년은 22.5%로 증가하여 직장 이동 경향이 최근에 활발해진 것을 알 수 있다. 한국노동연구원에서는 일자리 미스매치 관점에서 기업이 요구하는 최소 학력 수준과 직원의 학력 수준 간의 차이가 증가한 것을 이직 경향이 활발해진 이유로 들고 있다. 자신의 교육 수준보다 낮은 일자리를 얻은 직원은 더 높은 수준의 일자리로 이동하기 위해 이직을 고려할 것이다. 반면에 자신의 교육수준보다 높은 수준의 일자리에 취업한 경우, 업무에 적응하기 어려울 수 있으며 이는 이직의 원인이 될 수 있다.[4]

따라서 본 연구에서는 한국노동연구원의 자료에 착안하여, 직원에 대한 객관적 데이터인 전공, 교육수준, 재직 중인 회사 유형 등의 데이터를 기반으로 직원의 퇴사 여부를 예측하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 연구에 사용한 데이터를 소개한다. 3장에서는 어떠한 알고리즘을 사용하여 퇴사 예측 모델을 개발했는지 소개한다. 4장에서는 모델의 성능을 확인하고 각각의 성능을 비교한다. 5장에서는 도출된 결과를 해석하고 6장에서 결론을 짓는다.

II. 데이터

본 연구에서는 Kaggle의 “HR Analytics: Job Change of Data Scientists” 데이터셋을 사용했다.[5] 이 데이터는 데이터 분석가의 이직 여부를 예측하기 위한 용도로, 직원 ID, 직원이 거주 중인 도시 코드, 직원이 거주하는 도시 발전 정도, 성별, 직무 관련 경험 여부, 졸업한 대학 유형, 교육 수준, 전공, 총 경험 년수, 현직장의 직원 수, 현직장의 유형, 전직장과 현직장 사이의 공백기, 교육 이수 시간, 퇴사 여부의 총 14개의 항목에 대해 19,158건으로 구성되어 있다. 전체 데이터에서 재직자는 14,381명이고 퇴사자는 4,777명으로, 각각 75%와 25%를 차지하고 있다.

직원 ID와 직원이 거주 중인 도시 코드를 제외한 나머지 12개 항목에 대한 단순 빈도 분석 결과를 표 1에 정리했다. 직원이 거주 중인 도시 코드 항목은 123개의 코드 정보로 구성되어 있어 불필요한 변수로 판단했다. 이에 직원이 거주 중인 도시 코드는 직원 ID와 같이 분석 대상에서 제외되었다.

III. 연구방법

본 연구에서는 “HR Analytics: Job Change of Dta Scientists” 데이터를 분석하여 직원의 이직 및 퇴사에 영향을 미치는 항목을 알아낸다. Decision Tree, XGBoost, kNN, SVM의 4가지 알고리즘을 활용하여 퇴사 예측 모델을 생성한다. 각 알고리즘의 특징은 다음과 같다.

Decision Tree는 전형적인 분류 모델이며, 의사결정 규칙과 그 결과를 트리구조로 표현한다. Decision Tree 알고리즘은 계산 비용이 적은 것에 비해 괜찮은 성능을 얻을 수 있다. 하지만 과적합(Overfitting) 되기 쉽다는 단점이 있다.[7][8]

XGBoost는 Gradient Boosting 알고리즘의 속도 문제를 해결하기 위해 전산 속도와 모델의 성능에 초점을 맞춘 알고리즘으로 직관적인 모델이다. 과적합(Overfitting)을 방지하기 위해 변수의 regularized를 사용하여 정확도를 높였다.

kNN은 새로 입력된 데이터와 기존 데이터를 비교하여, 새로운 데이터를 유사하게 판단되는 기존 데이터로 분류하는 알고리즘이다. 데이터가 많을수록 높은 정확도를 보이지만, 그만큼 분석 속도가 느리다.

SVM은 결정 경계(Decision Boundary), 즉 분류를 위한 기준 선을 정의하는 알고리즘이다. SVM 알고리즘은 학습과정에서 보지 못한 새로운 데이터도 분류가 가능하지만, 학습 속도가 느리다.[9][10]

표 1. 단순 빈도 분석 결과

항목	설명	재직자수 (%)	퇴사자수 (%)	전체수 (%)
직원이 거주하는 도시 발전정도	0.875~1.0	26(0.2)	37(0.8)	63(0.3)
	0.75~0.875	1,384(9.6)	**1,980(41.4)**	3,364(17.6)
	0.625~0.75	1,152(8.0)	425(8.9)	1,577(8.2)
	0.5~0.625	2,082(14.5)	459(9.6)	2,541(13.3)
	~0.5	**9,737(67.7)**	**1,876(39.3)**	**11,613(60.6)**
성별	Male	**10,209(71.0)**	**3,012(63.1)**	**13,221(69.0)**
	Female	912(6.3)	326(6.8)	1,238(6.5)
	공란	3,260(22.7)	1,439(30.1)	4,699(24.5)
직무관련 경험여부	Has relevent experience	**10,831(75.3)**	**2,961(62.0)**	**13,792(72.0)**
	No relevent experience	3,550(24.7)	1,816(38.0)	5,366(28.0)
졸업한 대학유형	Full time course	2,326(16.2)	1,431(30)	3,757(19.6)
	no_enrollment	**10,896(75.8)**	**2,921(61.1)**	**13,817(72.1)**

	Part time course	896(6.2)	302(6.3)	1,198(6.3)
	공란	263(1.8)	123(2.6)	386(2)
교육 수준	Graduate	**8,353(58.1)**	**3,245(67.9)**	**11,598(60.5)**
	High School	1,623(11.3)	394(8.2)	2,017(10.5)
	Masters	3,426(23.8)	935(19.6)	4,361(22.8)
	Phd	356(2.5)	58(1.2)	414(2.2)
	Primary School	267(1.9)	41(0.9)	308(1.6)
	공란	356(2.5)	104(2.2)	460(2.4)
전공	Arts	200(1.4)	53(1.1)	253(1.3)
	Business Degree	241(1.7)	86(1.8)	327(1.7)
	Humanities	528(3.7)	141(3)	669(3.5)
	No Major	168(1.2)	55(1.2)	223(1.2)
	Other	279(1.9)	102(2.1)	381(2)
	STEM	**10,701(74.4)**	**3,791(79.4)**	**14,492(75.6)**
	공란	2,264(15.7)	549(11.5)	2,813(14.7)
총 경험 년수	>20	2,783(19.4)	503(10.5)	3,286(17.2)
	15~20	1,324(9.2)	258(5.4)	1,582(8.3)
	10~15	2,288(15.9)	541(11.3)	2,829(14.8)
	5~10	**3,750(26.1)**	1,261(26.4)	5,011(26.2)
	1~5	**3,909(27.2)**	**1,954(40.9)**	**5,863(30.6)**
	<1	285(2)	237(5)	522(2.7)
	공란	42(0.3)	23(0.5)	65(0.3)
현직장의 직원 수	10000~	1,634(11.4)	385(8.1)	2,019(10.5)
	5000~9999	461(3.2)	102(2.1)	563(2.9)
	1000~4999	1,128(7.8)	200(4.2)	1,328(6.9)
	500~999	725(5)	152(3.2)	877(4.6)
	100~500	2,156(15)	415(8.7)	2,571(13.4)
	50~99	**2,538(17.6)**	**545(11.4)**	**3,083(16.1)**
	10~49	1,127(7.8)	344(7.2)	1,471(7.7)
	~10	1,084(7.5)	224(4.7)	1,308(6.8)
	공란	**3,528(24.5)**	**2,410(50.5)**	**5,938(31)**
현직장의 유형	Early Stage Startup	461(3.2)	142(3)	603(3.1)
	Funded Startup	861(6)	140(2.9)	1,001(5.2)
	NGO	424(2.9)	97(2)	521(2.7)
	Other	92(0.6)	29(0.6)	121(0.6)
	Public Sector	745(5.2)	210(4.4)	955(5)
	Pvt Ltd	**8,042(55.9)**	1,775(37.2)	**9,817(51.2)**
	공란	3,756(26.1)	**2,384(49.9)**	6,140(32)
전직장과 현직장 사이의 공백기	1	**5,915(41.1)**	**2,125(44.5)**	**8,040(42)**
	2	2,200(15.3)	700(14.7)	2,900(15.1)
	3	793(5.5)	231(4.8)	1,024(5.3)
	4	801(5.6)	228(4.8)	1,029(5.4)
	>4	2,690(18.7)	600(12.6)	3,290(17.2)
	never	1,713(11.9)	739(15.5)	2,452(12.8)
	공란	269(1.9)	154(3.2)	423(2.2)
교육 이수 시간	280~350	**9,405(65.4)**	**3,217(67.3)**	**12,622(65.9)**
	210~280	3,321(23.1)	1,078(22.6)	4,399(23)
	140~210	1,080(7.5)	336(7)	1,416(7.4)
	70~140	350(2.4)	81(1.7)	431(2.2)
	0~70	225(1.6)	65(1.4)	290(1.5)
퇴사 여부	0	14,381(100)	0(0)	**14,381(75.1)**
	1	0(0)	4,777(100)	4,777(24.9)

모델 학습에 필요한 데이터에서 NaN 값은 최빈값으로 대체했다. 성별, 직무 관련 경험, 교육 수준 등과 같이 범주형인 항목은 데이터 전처리를 과정에서 숫자 형으로 매핑 했다. 또한 재직자와 퇴사자의 비율이 불균형하므로 SMOTE 알고리즘을 적용하여 균형을 맞췄다.[6]

Ⅳ. 실험 결과

본 연구에서는 Decision Tree, XGBoost, kNN, SVM까지 4가지 알고리즘을 활용하여 퇴사 예측 모델을 생성하고 각각의 알고리즘을 활용한 모델의 성능을 비교했다.

Scikit-Learn에서 제공하는 메서드를 이용하여 정확률(Precision), 재현율(Recall), F1-score를 계산했고, 이를 각 모델의 평가 척도로 사용했다. 정확률과 재현율의 수식은 다음과 같으며, 수식에 사용된 용어의 의미는 표 2에 정리했다.

$$\mathrm{Precision} = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

표 2. Precision과 Recall수식에 사용된 용어 의미

	정답	
시스템 예측 (판별 결과)	퇴사자를 퇴사자로 판별 (TP)	재직자를 퇴사자로 판별 (FP)
	퇴사자를 재직자로 판별 (FN)	재직자를 재직자로 판별 (TN)

모든 모델에서 학습 데이터 23,009건과 테스트 데이터 5,753건을 사용하여 실험했다. 각각의 모델에 대한 실험 결과는 표 3에 정리했다.

Decision Tree와 XGBoost, kNN, SVM 알고리즘을 기반으로 모델을 생성ㅇ하여 실험한 결과, 모든 모델이 F1-score를 기준으로 0.7 이상의 결과를 보였다. 그 중에서도 XGBoost 알고리즘이 0.843으로 가장 높았다.

표 3. 모델 실험 결과

모델	Precision	Recall	F1-SCORE
Decision Tree	0.762	0.781	0.771
XGBoost	**0.840**	**0.847**	**0.843**
kNN	0.729	0.793	0.759

Ⅴ. 결과 해석

본 연구에서는 전공, 교육수준, 재직 중인 회사 유형 등의 데이터를 기반으로 직원의 퇴사 여부를 예측하기 위해 Decision Tree, XGBoost, kNN, SVM 알고리즘을 사용했다. 모든 모델이 F1-score를 기준으로 0.7 이상의 결과를 보였으며 그 중에서도 XGBoost를 활용한 모델이 0.843으로 높았다.

가장 좋은 성능을 보였던 XGBoost 알고리즘을 사용하여 퇴사 여부에 가장 많은 영향을 끼치는 요인을 계산했다. 도시 발전 정도가 직원의 퇴사 여부에 제일 많은 영향을 준다고 나왔으며, 그 다음으로 직무 관련 경험 여부, 교육수준, 회사 직원 수, 회사 유형, 졸업한 대학 유형, 성별, 전공, 총 경험 년수, 전직장과 현직장 사이의 공백기, 교육 이수 시간 순서로 나왔다. 이에 대한 내용은 표 4에 정리했다.

표 4. 직원의 퇴사 여부에 영향을 미치는 요인

요인	영향 미치는 정도(%)
직원이 거주하는 도시 발전 정도	**23.0**
직무 관련 경험 여부	**21.0**
교육 수준	**16.0**
현직장의 직원 수	6.9
현직장의 유형	6.5
졸업한 대학 유형	5.5
성별	4.9
전공	4.3
총 경험 년수	4.0
전직장과 현직장 사이의 공백기	3.6
교육 이수 시간	3.3

직원의 퇴사 여부에 영향을 미치는 요인 중에서 10%가 넘는 3가지가 주요 요인으로 판단된다. 표 1을 보면 재직자는 대부분 거주하는 도시의 발전 정도가 낮은(~0.5 67.7%) 반면에 퇴사자는 거주 중인 도시의 발전정도가 높은(0.75~0.875 41.4%) 곳에 거주 중이었다. 재직자와 퇴사자 모두 직무 관련 경험을 가지고 있는 사람이 많았지만 퇴사자(62.0%)가 재직자(75.3%)보다 더 적었다. 그리고 재직자와 퇴사자 모두 교육수준은 대학원 졸업이 가장 많았지만 퇴사자(67.9%)가 재직자(58.1%)보다 더 많은 것으로 나타났다. 이는 좀 더 발전된 도시에서 높은 수준의 교육을 받을 수 있는 기회가 더 많은 것으로 볼 수 있다. 또한 직무와 관련된 경험이 많을수록 이직하는 데에 수월한 것이 원인으로 예상된다.

결과적으로, 거주하고 있는 도시의 발전 정도가 높으며, 높은 수준의 교육을 받고 직무와 관련된 경험이 많을수록 퇴직할 확률이 높은 것으로 나타났다.

Ⅵ. 결 론

본 연구는 Kaggle의 “HR Analytics: Job Change of Data Scientists” 데이터를 기반으로 Decision Tree, XGBoost, kNN, SVM 알고리즘을 활용하여 직원의 퇴사 여부를 예측하는 모델을 개발했다. 또한 직원의 이직 및 퇴사에 영향을 미치는 요인에 대해서 살펴볼 수 있었다. 이를 통해 이전까지 설문조사를 중심으로 수행되었던 연구에서 도출하지 못한 다양한 요인을 파악할 수 있었다. 이를 통해 기업이 퇴사 예측 모델을 이용하여 직원이 퇴사하기 전에 미리 이를 인지하고 방지하는 데에 도움을 줄 수 있을 것으로 예상된다.

다만, 본 연구에서 사용한 데이터는 19,158건으로 비교적 많지 않다. 또한 해외에서 데이터 분석가의 이직 여부를 예측하기 위한 용도였다. 따라서 한국의 상황과 상이한 부분이 있을 것으로 예상되며, 다양한 직군의 직원에 대해서 일반화시키기에 어려움이 있다. 향후, 데이터 분석가뿐만이 아니라 다양한 직군의 데이터를 수집하고 더 많은 수의 데이터를 확보하여 활용한다면 지금보다 더 좋은 성능의 모델을 개발할 수 있을 것으로 예상된다.

References

[1] Becker. G. (1964). Human Capital. New York: Columbia University Press. Burt, R. S. (1992). Structural Holes. Cambridge, MA : Harvard University Press.

[2] Strobe, M. H.(1990). “Human Capital Theory: Implications for HR Managers.” Industrial Relations 29(2) : 214~239.

[3] 권기욱, “직원 이직률과 기업성과의 관계:고성과자와 비고성과자의 이직률을 고려한 탐색적 연구”, 노동정책연구2016, 제16권, 제1호, 1-26, 2016.

[4] 지상훈, “일자리의 인지적 숙련 수준과 이직”, 노동리뷰, 10월호, 55-72, 2020.

[5] “HR Analytics: Job Change of Data Scientists” kaggle, last modified Dec 25, 2020, accessed Feb 10, 2021, https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists

[6] Hansoo Lee, Jonggeun Kim, Sungshin Kim. (2017). Gaussian-Based SMOTE Algorithm for Solving Skewed Class Distributions. INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT SYSTEMS, 17(4), 229-234.

[7] 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..

[8] 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.

[9] Mavroforakis, M. E., & Theodoridis, S. (2006). A geometric approach to support vector machine (SVM) classification. IEEE transactions on neural networks. 17(3), 671-8682.

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

[11] 안관영, “경제적·심리적 요인과 이직 의도의 관계에 대한 연구-외식업 종사자를 중심으로.”, 경영교육연구, 48(-), 241-257, 2007.

[12] 성지미·안주엽, “일자리 만족도와 이직의사 및 이직-청년층을 중심으로.”, 한국산업노동연구, 22(2), 135-179, 2015.