「정보시스템연구」 제30권 제4호
한국정보시스템학회
2021년 12월, pp. 21~42

http://dx.doi.org/10.5859/KAIS.2021.30.4.21

XGBoost와 SHAP 기법을 활용한 근로자 이직 예측에 관한 연구*

이재준** · 이유린*** · 임도현**** · 안현철*****

〈목 차〉

Ⅰ. 서 론

기업이 지속적으로 생존하기 위해서는 다른 기업과의 경쟁에서 우위를 점해야 하고, 꾸준히 이윤이 발생해야 하며, 이를 위해서는 기업 경쟁력의 핵심으로 인식되고 있는 인적자원을 유지하고 관리하는 것이 필요하다. 그렇기 때문에 기업들은 인적자원을 관리하고 개발하는 데 적극적으로 투자하는 등 많은 노력을 기울이고 있다(강만수, 윤상용, 2018; 고건우 등, 2020). 따라서 기업에서 많은 투자를 통해 육성한 내부직원이 이직을 하게 될 경우, 이는 부정적인 신호로 인식된다.

이직이란 일반적으로 조직의 구성원이 근무 조직을 이탈하는 것으로서, 어떤 주체가 의사결정을 내렸는지에 따라서 자발적 이직(voluntary turnover)과 비자발적 이직(involuntary turnover)으로 구분한다. 자발적 이직이란 조직원 스스로 이직을 선택하는 것을 말하며, 보다 좋은 조건의 조직으로 이동하는 것을 의미하는

* 본 논문은 교육부 및 한국연구재단의 4단계 두뇌한국21 사업(4단계 BK21 사업)으로 지원된 연구입니다.
** 국민대학교 비즈니스IT전문대학원, jrack707@naver.com(주저자)
*** 국민대학교 비즈니스IT전문대학원, urin4377@naver.com
**** 국민대학교 비즈니스IT전문대학원, ehgus7011@naver.com
***** 국민대학교 비즈니스IT전문대학원, hcahn@kookmin.ac.kr(교신저자)

전직과 개인적인 사유로 직장을 그만두는 사직을 포함하는 개념이다. 반면, 비자발적 이직이란 해고, 정년퇴직, 계약기간 만료 등을 이유로 근로자의 계속적인 근무의지와 상관없이 고용관계가 끝나는 것을 의미한다(이만기, 2013; 김재신, 2017; 강윤경 등, 2020).

이 중에서 기업의 성과에 부정적인 영향을 주는 이직은 우수 인력의 자발적인 이직이다(고건우 등, 2020). 노동시장에 참여한 지 얼마 되지 않은 청년 취업자가 이직할 경우 초기 경력을 형성하기 어렵고, 재취업할 때까지 추가적인 비용이 발생하며, 국가적 인적 자원의 축적이 저해되는 등 사회·경제적 비용을 야기한다(정인호 등, 2018). 또한, 청년 취업자가 입사 후 얼마 되지 않아 이직할 경우 기업은 이미 투입한 교육비용 등 투자비용을 회수하지 못한다는 문제가 발생한다. 기업에 오래 근무한 숙련자의 이직은 주요 인력의 이탈로 인한 기술 유출, 생산성 저하, 새로운 사업기회 상실, 다른 직원의 사기 저하, 고객 이탈이나 잠재적 고객 감소 등의 성과 저해뿐만 아니라, 해당 인력을 대체하기 위한 새로운 인력을 채용하고 교육함으로써 신규 인력의 업무 능력을 개발하는 데 추가적인 비용이 발생하는 등 기업에 비용 부담을 가중시킬 수 있다(정인호 등, 2018; 고건우 등, 2020). 또한 이직은 동료 직원들에게 부정적인 신호로 작용하여 지속적으로 이직을 발생시킬 수 있다. 강만수와 윤상용(2018)에 따르면, 당기 이직은 차기뿐만 아니라 차차기의 이직에도 영향을 미치며, 채용 비용에 있어서도 당기 채용 비용이 차기와 차차기의 채용 비용을 증가시킴으로써 지속적으로 기업에 부정적인 영향을 끼칠 수 있다.

따라서 이직에 영향을 미치는 요인들을 분석하고, 이를 기반으로 이직 여부를 미리 예측하여 대응하는 것은 매우 중요하다고 할 수 있다. 만약 인적 자원 관리 부서에서 이직 예측 모형을 만들어 운영한다면 실제 이직이 발생하기 전에 미리 각 직원들의 이직 가능성을 파악할 수 있을 것이고, 이직 가능성이 높은 것으로 파악된 직원들의 이직을 예방하기 위한 적절한 사전 대응이 가능해 질 것이다. 특히 최근 기업에 빅데이터 환경이 조성되면서 데이터에 기반해 효과적인 인적자원관리(HRM, Human Resource Management)를 수행하고자 하는 이른바 피플 애널리틱스(people analytics)가 주목을 받고 있는데, 기업이 보유하고 있는 직원들의 다양한 데이터를 활용하여 이직 예측 모형을 개발한다면 효과성과 효율성이 담보된 이직 관리가 가능하다. 이러한 배경에서 본 연구는 이직을 예측할 수 있는 데이터 기반 기계학습(machine learning) 모형을 제안하고, 해당 모형을 통해 직원의 이직에 영향을 미치는 요인들은 무엇인지 분석해 보고자 한다.

일반적으로 이직을 예측하는 모형을 개발하기 위해서는 종속변수로서 이직 행동(이직 여부)을 활용하여야 하지만, 이직 행동이 자발적이었는지 비자발적이었는지 판단하기에는 데이터가 충분하지 않고, 설문을 통해 확보한 데이터의 경우 불리한 점은 숨기려는 사람의 특성상 해당 데이터를 신뢰하기도 어렵다. 따라서 기존에 수행된 이직에 관한 연구들은 이직의도를 종속변수로 한 경우가 많았다(이만기, 2013; 김정은 등, 2017; 정인호 등, 2018; 엄남현, 2019). 이직의도는 당장 이직을 하지는 않더라도 이직을 할 의사가 있다는 것을 의미하며, 자

발적 이직을 예측할 수 있는 주요 요인 중 하나로서 이직과 밀접한 관련이 있다. 이직의도는 이직에 영향을 주는 가장 직접적이면서도 중요한 선행요인이므로 이직의 대체 개념으로 사용하는 것이 가능하다(이만기, 2013; 전희주, 2015; 김정은 등, 2017; 정인호 등, 2018; 엄남현, 2019). 이에 본 연구에서도 이직의도를 종속변수로 하는 기계학습 모형을 구축하고자 한다.

한편 기존에 수행된 이직의도와 관련한 연구들은 광고대행사 직원이나 보험 설계사, 공항 근무자와 같은 특정 직종에 한정하여 이직의도에 영향을 주는 요인을 연구한 경우가 많았다. 전반적인 직종에 대하여 수행된 연구의 경우도 있으나 청년 취업자를 대상으로 이루어진 연구가 대부분이었다. 그러나 이직에 따른 비용의 발생은 입사한 지 얼마 안 된 취업자뿐만 아니라 오랫동안 한 기업에서 근무한 근로자에 의해서도 발생하므로 보다 포괄적인 근로자를 대상으로 한 연구가 필요한 상황이다. 또한, 실력이 우수한 직원이 이직을 고민할 경우 미리 해당 상황을 인식하고, 어떤 이유에서 이직을 고민하는 지 파악할 수 있다면 기업은 빠르게 선제적 조치를 취함으로써 추가적인 비용이 발생할 가능성을 줄일 수 있을 것이다. 따라서 이직에 전반적으로 영향을 미치는 요인뿐만 아니라 개개인에게 있어 이직에 영향을 미치는 요인을 파악하는 것도 필요하다.

아울러 기존의 이직의도 관련 연구들은 이직의도에 영향을 미치는 요인들을 탐색한 연구가 대부분이었다. 주로 설문조사를 통해 데이터를 수집하였으며, ANOVA 분석, 독립표본 T-test, 로지스틱 회귀분석이나 포아송 회귀분석, 구조방정식 모형 등을 사용하여 어떤 요인들이 이직의도에 영향을 미치는지 확인하였다. 그리고 이직의도 예측 연구들은 영향 요인을 파악하기 위하여 상대적으로 예측력이 떨어지는 의사결정나무나 로지스틱 회귀분석을 사용한 경우가 많았다. 이들 기법들은 어떤 변수가 이직의도에 영향을 주는지 직관적으로 설명할 수 있다는 장점이 있지만 예측력이 다른 기계학습 기법에 비해 떨어진다는 한계점을 갖는다(천예은 등, 2021).

이러한 기존 연구들의 한계를 극복하기 위해, 본 연구에서는 보다 예측력을 높이기 위하여 여러 의사결정나무를 부스팅 방법을 통해 결합한 앙상블(ensemble) 방법인 XGBoost (eXtreme Gradient Boosting)를 제안한다. 그러나 XGBoost는 의사결정나무에 비하여 높은 예측 성능을 보이지만 여러 개의 의사결정나무를 혼합하는 방법이기 때문에 예측결과가 도출된 이유에 대하여 설명하기 힘든 블랙박스 모델이라는 한계가 있다(김성훈 등, 2021). 따라서 예측에 대한 유의미한 분석을 도출하기 위하여 설명가능한 AI(XAI, eXplainable AI) 기법 중 하나인 SHAP(SHapley Additive exPlanations)을 적용하고자 한다. SHAP은 Shapley Value를 이용하여 예측에 영향을 미치는 변수들을 파악할 수 있는 방법으로, 종속변수에 긍정적인 영향을 미치는 변수뿐만 아니라 부정적인 영향을 미치는 변수도 파악할 수 있는 기법이다. 또한 SHAP은 변수의 전역적인 영향력뿐만 아니라, 분석에 사용된 개개인에 대해서도 어떤 변수가 이직의도에 긍정적인 영향을 끼쳤는지, 부정적인 영향을 끼쳤는지 파악할 수 있는 지역적인 영향력 정보도 제공해 준다. 이에 본 연구에서는 XGBoost와 SHAP을 적용하여, 보다 우수한

성능을 가지는 예측 모형을 구축하는 동시에 구축된 모형에 대한 유의미한 해석을 제공하는 것을 목표로 한다.

Ⅱ. 이론적 배경

2.1 이직 관련 연구 동향

기존 이직의도 영향 요인 연구는 특정 직종에 한정한 연구가 많았다. 엄남현(2019)은 국내 광고대행사 직원들을 대상으로 이직 원인을 밝혀내고자 했다. 데이터는 이직 경험이 있는 광고업계 종사자를 대상으로 심층 인터뷰를 통해 수집하였으며 질적 연구 방법을 통해 인터뷰 내용을 해석하였다. 해당 연구에서는 이직의도에 영향을 미치는 요인들을 직무관련 요인, 조직관련 요인, 작업환경 요인의 세 분야로 나누어 설명하였다. 첫째로 직무관련 요인에서는 과도한 업무 및 스트레스, 업무 고가 평가에 대한 불만, 광고주와의 갑을 관계에서 오는 정신적 스트레스 등이 이직의도에 영향을 미치는 요인으로 나왔다. 다음으로 조직관련 요인에서는 보수, 승진, 회사의 비전, 체계적인 관리 시스템의 부재 등이 이직의도에 영향을 미치는 요인으로 나왔으며, 마지막으로 작업환경 요인에서는 팀워크, 조직문화, 대인관계, 성차별 등이 이직의도에 영향을 미치는 요인으로 나왔다.

전희주(2016)와 안철경 등(2017)은 보험설계사를 대상으로 분석을 진행하였다. 두 연구 모두 국내 보험설계사를 대상으로 설문조사를 진행하여 데이터를 수집하였다. 전희주는 포아송 회귀분석을 사용하여 분석을 진행하였으며, 현재 속한 회사가 어디인지, 나이, 입사 동기, 월 평균 소득, 월 평균 신계약 건수, 최종학력 등이 보험설계사의 이직의도에 영향을 미치는 것으로 나왔다. 안철경 등은 ANOVA를 통해 설계사 유형 간 이직의도의 차이를 확인하였으며, 독립표본 T-test를 통해 유의한 변수를 추출한 후 해당 변수들을 가지고 로지스틱 회귀분석을 진행해 이직의도에 영향을 미치는 변수를 분석하였다. 분석 결과 수수료 수준 만족도, 분급 만족도가 공통적으로 이직의도에 영향을 주는 변수인 것으로 나왔으며, 대리점 설계사는 평판, 전속설계사는 신계약 건수와 근무경력, 텔레마케터 설계사는 보유고객 수가 이직의도에 영향을 미치는 요인인 것으로 나타났다.

이영석 등(2016)은 공항에서 근무하는 지상서비스직종을 대상으로 설문조사를 진행하였다. 변수들을 직무관련 요인, 조직관련 요인으로 분류하여 분석을 진행하였으며, 두 분야의 변수들이 이직행동에 미치는 영향 변수를 확인함에 있어 조직헌신의 매개효과를 검정하였다. 연구 결과 직무관련 요인 중 역할 갈등이 조직헌신을 매개로 이직에 영향을 주는 것으로 나왔으며, 조직관련 요인 중에서는 보상과 의사결정 요인이 조직헌신을 매개로 이직에 영향을 주는 것으로 나왔다.

청년 취업자를 대상으로 한 연구에는 이만기(2013)의 연구와 정인호 등(2018)의 연구가 있다. 이만기(2013)는 30세 미만의 초기 경력자 중 이직 경험이 없는 임금근로자를 대상으로 분석을 진행하였다. 데이터는 2008년 대학 졸업생을 대상으로 한 한국고용정보원의 대졸자 직업이동경로조사(GOMS) 자료를 사용하였으며, 일자리나 전공의 적합도, 임금과 복지의 수

준, 직무만족도 등으로 분류된 18개의 변수 중 이직의도에 영향을 미치는 변수가 어떤 것이 있는지 구조방정식 모형을 사용해 확인하였다. 분석 결과 일자리적합은 이직의도에 직・간접적으로 영향을 미치는 것으로 나왔으며 전공적합과 임금복지수준은 간접적으로만 이직의도에 영향을 미치는 것으로 나타났다. 추가적으로 기업규모별 분석을 진행한 결과 중소기업 근로자의 경우 일자리적합, 전공적합 그리고 임금복지수준이 대기업 근로자에 비해 이직의도에 더 큰 영향을 미치는 것으로 나왔다. 정인호 등(2018)은 2015년 대학 졸업자를 대상으로 한 GOMS 자료를 사용하였다. 분석 대상은 35세 미만 임금근로자로, 그중에서도 상용근로자와 임시근로자를 대상으로 분석을 진행하였다. 개인요인, 직무요인, 외부환경요인으로 분류된 12가지 변수를 사용해 분석하였으며, 분석에는 의사결정나무를 사용하였다. 분석 결과 직무기술적합도, 복리후생 만족도, 임금 만족도 그리고 교육수준이 이직의도에 크게 영향을 미치는 변수인 것으로 나타났다. 그 중에서도 일의 기술수준이 낮고, 복리후생제도에 불만족하며, 4년제 대학을 나온 그룹의 경우 이직의도가 가장 높게 나왔다. 또한, 복리후생제도와 임금 만족도가 이직의도를 낮출 수 있다는 것을 확인하여 이직 문제를 해결하는 방안으로 양질의 일자리를 만들 것을 제안하였다.

김정은 등(2017)과 고건우 등(2020)은 보다 넓은 나이대를 대상으로 이직의도 연구를 진행하였다. 김정은 등(2017)은 17차년도 한국노동패널조사 자료를 이용하여 회귀분석을 통해 분석을 수행하였으며, 청년층, 중년층, 장년층으로 분류한 60세 미만 임금근로자를 대상으로 연구를 진행하였다. 종속변수로는 리커트 5점 척도로 조사한 이직의사의 정도를 사용하였으며, 독립변수로는 직무만족과 생활만족으로 구분한 14개의 변수와 7개의 인구통계학적 변수, 5개의 직무특성 변수, 2개의 조직특성 변수 등 총 28개의 변수를 사용하였다. 성별, 임금, 직종이 이직의사에 영향을 미치는 것으로 나왔다. 직무만족과 생활만족 또한, 각각 높은 만족도를 느낄수록 이직의사가 낮아진다는 것을 확인하였다. 연령층을 구분하여 이직의사에 영향을 미치는 요인을 분석한 결과 취업안정 만족도가 모든 연령층에서 유의미한 변수인 것으로 나왔으며, 직무내용과 발전가능성 변수는 중・장년층에서, 근무환경 변수는 청・중년층에서 의미있는 변수인 것으로 나왔다. 고건우 등(2020)은 2017년 대학 졸업자를 대상으로 한 GOMS 데이터를 사용하여 의사결정나무 모형을 통해 분석을 진행하였다. 수습 기간이 지났다고 볼 수 있는 4개월 이상 근속자를 4~6개월, 7~12개월, 13~24개월, 25개월 이상 근속자로 분류하여 분석대상으로 사용하였다. 이직 준비 여부를 종속변수로 설정하였고, 인구통계학적, 근로 관련, 근로 자아정체성, 근로 만족도, 구직경험, 4대 보험 가입 여부로 분류된 33개의 변수를 독립변수로 설정하였다. 의사결정나무를 사용한 예측모형을 구축하였는데, 연령대별 예측모형의 평균 예측 정확도는 68% 정도를 기록하였다. 분석 결과 교육수준과 일의 수준의 일치 정도, 개인 발전 가능성, 직무관련 교육 및 훈련, 일의 수준 대비 본인의 수준 일치 정도, 승진제도, 임금, 일자리에 대한 사회적 평판, 고용 안정성이 이직의도에 영향을 미치는 것으로 나타났다. 근속기간별로는 4~6개월 근속자 그룹은 고용 안

정성, 직장 내 인간관계, 복리후생제도가 이직의도에 영향을 주는 것으로 나왔고, 13~24개월 근속자 그룹에서는 업무와 자신의 적성・흥미 일치 만족도, 업무에 대한 사회적 평판이 영향을 주는 변수로 나왔으며, 25개월 이상 근속자 그룹에서는 직무관련 훈련 또는 교육, 승진제도, 교육수준, 일-수준 일치 정도가 영향을 주는 변수인 것으로 나왔다.

강만수와 윤상용(2018)은 한국직업능력개발원에서 조사한 인적자본 기업패널 조사 자료 중, 1차에서 6차까지의 조사에 모두 참여한 기업을 대상으로 이직이 채용 비용에 어떠한 영향을 미치는지를 연구하였다. 회귀분석을 사용해 연구를 진행하였으며, 연구 결과 당기 이직은 차기 이직을 증가시키는 것으로 나타났고, 당기 이직이 대체로 차기 채용비용도 증가시키는 것으로 나타남으로써 이직이 지속적으로 다음 이직과 채용 비용에 영향을 미치는 것을 보였다. 이를 통해 기업은 인사관리에 더 신경을 써야 하며, 이직을 최소화하는 노력을 기울여야 한다고 주장했다.

은화리 등(2018)은 직무배태성 개념을 기초로 어떤 요인이 이직의도를 높이고, 어떤 요인이 이직의도를 낮추는지를 연구했다. 직무배태성이란 개인이 조직이나 직무에 깊은 소속감을 느끼는 것을 의미하며, 개인이 조직에 머무르도록 하는 영향력을 의미한다. 해당 연구는 질적 연구로서 5년 이하 경력의 정규직 20명을 대상으로 심층 면접을 통해 데이터를 수집하였다. 질적 연구는 깊이 있고 풍부한 자료를 제공하는 것에 중점을 둔 연구로서 표본의 일반화 가능성은 낮다고 볼 수 있다. 연구 결과 업무 불만, 향후 진로 방향 부재, 새로운 일에 대한 욕구, 과도한 업무량과 야근, 조직의 구조적 한계에 대한 회의감, 조직 문화에 대한 불만, 비전 없는 회사, 연봉 불만족, 직속상사에 대한 불만, 동료 관계의 어려움 등이 이직의도를 높이는 것으로 분석되었으며, 업무에 대한 만족, 장기적 경력관리 고려, 조직에 대한 만족, 동료들과의 관계에 대한 만족, 불확실한 미래에 대한 불안, 취업상태에 대한 만족 등이 이직의도를 낮추는 것으로 분석되었다.

2.2 XGBoost(eXtreme Gradient Boosting)

XGBoost는 성능이 약한 분류기를 여러 개 결합하여 보다 강력한 분류 성능을 제공하는 방법인 부스팅 알고리즘 중 하나이다(하대우 등, 2019). 부스팅 알고리즘은 분류기를 수정해 가면서 성능을 개선한다(천예은 등, 2021). 우선 하나의 트리에 의해 분류가 이루어지고 나면, 오분류된 사례에 더 큰 가중치를 부여한다. 다음에 만들어지는 트리는 큰 가중치가 적용된 오분류 사례에 집중하여, 해당 사례를 더욱 잘 분류할 수 있게끔 만들어진다.

개별 사례들에 대해 나무의 결정 규칙을 사용하여 leaf들로 분류한다. 그리고 해당 leaf들의 점수를 합산하여 최종 예측을 계산한다. 분류 모델을 학습하기 위하여 목적 함수 식 (1)을 최소화한다.

$$L(\Phi) = \sum_i l(\hat{y_i}, y_i) + \sum_k \Omega(f_k) \ where \ \Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \quad (1)$$

이 때, l은 예측값 $\hat{y_i}$와 실제값 y_i의 차이를 계산하는 미분 가능한 convex loss function이다. 두 번째 항은 패널티를 나타낸다. T는 모형에 있는 leaf node의 수이며, w는 leaf의 가중치를 의미한다. f_k는 독립적인 tree 구조를 의미한다. $\Omega(f_t)$는 트리의 비중을 조절하는 조정 함수이다.

식 (1)을 최적화하기 위해서는 추가적인 변형이 필요하다. t 번째 반복 시 i 번째 사례에 대한 예측을 $\widehat{y_i^{(t)}}$라고 하자. 목적 함수 $L^{(t)}$를 최소화하기 위해 수식을 변형하면 다음과 같다.

$$L^{(t)} \simeq \sum_{i=1}^{n}[l(y_i,\widehat{y_i^{(t-1)}}) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t)$$
$$where\ g_i = \partial_{\widehat{y^{(t-1)}}} l(y_i,\widehat{y^{(t-1)}}) \text{ and } h_i = \partial^2_{\widehat{y^{(t-1)}}} l(y_i,\widehat{y^{(t-1)}}) \quad (2)$$

g_i와 h_i는 각각 목적 함수에 대한 1차 편미분 함수와 2차 편미분 함수이다. I_j를 가지 j에 의해 분류된 사례들의 집합이라고 할 때, 우리에게 필요한 식만 남기기 위하여 상수항을 제거한 후 정리하면 최종적으로 계산해야 할 목적 함수를 다음과 같이 구할 수 있다.

$$\tilde{L} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$$
$$= \sum_{j=1}^{T}\left[(\sum_{i\in I_j} g_i)w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)w_j^2\right] + \gamma T \quad (3)$$

여기서 g와 h의 최소점에서의 w와 최소값은 다음과 같이 구할 수 있다.

$$\text{argmin}_x\left(\sum_{i\in I_j} g_i w + \frac{1}{2}\sum_{i\in I_j}(h_i+\lambda)w^2\right) = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j}(h_i+\lambda)} \quad (4)$$

$$\text{min}_x\left(\sum_{i\in I_j} g_i w + \frac{1}{2}\sum_{i\in I_j}(h_i+\lambda)x^2\right) = -\frac{\left(\sum_{i\in I_j} g_i\right)^2}{2\left(\sum_{i\in I_j}(h_i+\lambda)\right)} \quad (5)$$

보통 생성 가능한 모든 트리 구조를 열거하는 것은 불가능하므로, 단일 leaf에서 시작하여 가지를 반복적으로 추가하는 Greedy Algorithm이 사용된다. XGBoost는 Greedy Algorithm을 사용해 다양한 분류기들을 발견하고, 분산 처리를 사용해 각 분류기에 적용할 최적의 가중치를 찾아낸다(안재현, 2020; 천예은 등, 2021). I_L과 I_R이 분할 후 왼쪽과 오른쪽 node에 분류된 사례들의 집합이라고 할 때, $I = I_L \cup I_R$이라고 하면 분할 후 감소하는 손실은 다음과 같이 계산할 수 있다.

$$L_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \quad (6)$$

이 값은 일반적으로 분할이 잘 되었는지 평가하는 데 사용된다(Chen and Guestrin, 2016; 안재현, 2020).

2.3 SHAP(SHapley Additive exPlanations)

SHAP 기법은 전체 성과를 만들어내는 데 각 변수가 얼마나 기여했는지를 수치로 표현하는 Shapley Value를 이용하는 기법이다. 특정 변수의 기여도는 모든 변수를 조합하였을 때 나오는 성과와 해당 변수를 제외한 변수들을 조합해 나오는 성과의 차이를 각각 계산하고, 해당 값들을 가중평균하여 측정한다(나광택 등, 2020). 즉, 특정 특성을 제외하였을 때 전체 성과의 변화 정도로 해당 특성의 기여도를 나타낼 수 있다(김성훈 등, 2021).

SHAP 기법은 Shapley Value와 학습에 사용된 복잡한 모델인 $f(x)$를 해석하기에 좋은 간단한 모델인 $f_x(z')$으로 변환하여 해석을 진행하는 Addictive Feature Attribution Method를 이용한다. 객체 x는 함수를 통해 0과 1로 표현되는 z'으로 변환된다. 이 때, Local Accuracy, Missingness, Consistency의 세 가지 특성을 모두 만족하는 Addictive Feature Attribution Method를 적용한 설명 모델이 SHAP이다. SHAP 기법에서 사용되는 Shapley Value를 구하는 식은 다음과 같다.

$$\Phi_i(f,x) = \sum_{z'\subseteq x'} \frac{|z'|!(n-|z'|-1)!}{n!}\left[f_x(z') - f_x(z'\setminus i)\right] \quad (7)$$

여기서 Φ_i는 특성 i의 Shapley Value이며, n은 전체 특성의 수, $f_x(z')$는 모든 특성의 기여도, $f_x(z'\setminus i)$는 특성 i를 제외한 나머지 특성들을 이용해 구한 기여도를 의미한다. 즉, 특성 i의 기여도는 전체 기여도에서 특성 i를 제외한 기여도를 뺀 값이라고 할 수 있다(Lundberg and Lee, 2017, 안재현, 2020).

SHAP 기법은 예측 모델의 출력 결과를 해당 예측에 사용된 특성들의 기여도로 분해한다. 이 때 Shapley Value는 음수일 수 있으며, Shapley Value가 음수라는 것은 해당 특성이 예측에 음(-)의 영향을 미친다고 해석할 수 있다. SHAP는 특성 사이에 서로 영향을 미칠 가능성인 의존도도 고려하여 모델의 영향력을 계산한다(안재현, 2020).

SHAP 기법은 종속변수 예측에 영향을 미치는 특성들의 평균 영향도를 제공한다. SHAP는 특성 간의 의존도를 고려하고 음의 영향력도 반영하여, 예측에 영향을 미치는 특성을 그 정도에 따라 나타내 준다. 따라서 특성의 부정적인 영향이 반영되지 않는 특성 중요도 기법에 비해 보다 정확한 영향력을 표시해 준다. 또한, 각각의 사례에 대해 예측에 긍정적 또는 부정적인 영향을 준 특성들을 제시해 줌으로써 개

별 사례들의 예측 결과에 대한 설명이 가능하다는 장점이 있다. 다만, 대용량의 데이터를 다루는 데에는 연산량이 크고, 학습된 모델에 대해서만 설명이 가능해 특성이 자주 추가되고 제거되는 경우에는 적용하기 부적합하다는 단점이 있다(안재현, 2020; 천예은 등, 2021).

Ⅲ. 제안 모형

3.1 제안 모형

본 연구에서는 기존 예측 연구에 사용된 의사결정나무의 낮은 예측 성능을 개선하기 위하여 여러 의사결정나무를 결합한 앙상블 방법인 XGBoost 기법을 사용하여 예측 모델을 구성하고자 한다. 또한 의사결정나무, 로지스틱 회귀분석을 비교 모델로 제시하여 XGBoost 기법의 예측 성능이 개선되었음을 확인하려고 한다. 그리고 예측 성능은 개선되었지만 설명력이 사라진 블랙박스 모델에 해당하는 XGBoost에 설명력을 부여하기 위해 XAI 기법 중 하나인 SHAP 기법을 사용하려고 한다.

3.2 분석 데이터

본 연구는 한국고용정보원에서 2018년 졸업자를 대상으로 2019년 9월 1일에 조사한 2019 대졸자직업이동경로조사(GOMS) 자료를 사용하였다. 해당 데이터는 전국 전문대 및 대학 졸업자 중 3~4%를 선정하여 대졸자의 직업 이동 경로를 조사한 자료이다. 졸업 후 18개월에서 24개월 정도 지난 2017년 8월 대졸자와 2018년 2월 대졸자 18,164명에 대해 설문을 진행한 것으로, 응답자의 이직준비 여부, 현재 경제활동 상황, 현재 일자리, 인적사항 및 가족에 관한 정보들을 담고 있다. 전처리 과정을 거쳐 총 5,747명의 표본을 확보하였다. 분석 대상은 임금근로자 중 상용근로자와 임시근로자로 한정하였다. 이직준비 여부를 종속변수로 하였으며 이직을 준비 중이면 이직의도가 있는 것으로 보았다. 분석에 사용할 독립변수는 기존 연구들에서 사용한 변수들을 종합적으로 고려하여 총 55개의 변수를 사용하였다. <표 1>은 본 연구에서 사용된 변수들의 목록이다. 이 중 jobseeking 변수는 현 직장을 얻기 위해 구직활동을 하였는지 여부를 나타내는 변수로, 여기서 구직활동이란 공·사립 직업소개소에 등록, 사업주와 면담 또는 전화상담, 광고게재, 광고에 응모 또는 구인광고란 탐색, 원서접수 또는 취직시험 응시, 무보수 견습 또는 직장훈련, 노동관련 기관에 알선 의뢰, 일자리를 찾아 사업장·농장·건설공사장 등 방문, 친구·친지에게 부탁을 하는 행위, 창업 준비활동 등을 의미하며, 구직활동을 안 하고 취업한 경우는 다른 구직 노력 없이 가족 또는 지인의 소개로 바로 취업된 경우, 헤드헌터가 먼저 제안하여 취업한 경우를 의미한다. 본 논문에서 사용한 데이터에서는 설문에 참여한 18,164명 중 13,296명이 해당 문항에 대해 답을 하였는데, 응답자의 42.9%에 해당하는 5,699명이 구직활동을 하지 않고 현 직장에 다니게 되었다고 응답하였다. 이처럼 구직활동을 하지 않고 현 직장에 다니는 사람의 비율이 높은 이유는 조사 대상들에 20대 젊은 청년들뿐만 아니라 이미 직장 생활을 해 오고 있던 40~60대까지 포함되어 있기 때문인 것으로 보

인다. 학점의 경우 4.5점 만점을 기준으로 환산하여 사용하였으며, 혼인 여부는 기혼, 이혼, 사별은 기혼으로 처리하여 사용하였다. 대학과 회사 소재지의 경우 5도를 기준으로 서울과 제주도를 추가하여 총 7개의 값을 갖도록 변환하였다. 확보된 표본 중 이직의도를 가지고 있다고 응답한 1,539명에 맞춰 이직의도가 없는 표본 1,539명을 무작위로 추출하여 분석에 투입하였다. 총 3,078명의 표본 중 2,462명은 학습 데이터로 사용하고, 나머지 616명은 검정 데이터로 사용하였다. 학습 데이터와 검정 데이터의 비율은 8:2로 맞추었다.

<표 1> 실험에 사용한 변수 정보

변수명	설명	변숫값	변수명	설명	변숫값
univ	국공사립	국공립, 사립	univ_type	대학 유형	전문대, 4년제
univ_loc	대학 위치	서울, 경기, 강원, 충청, 경상, 전라, 제주	comp_loc	사업장 위치	서울, 경기, 강원, 충청, 경상, 전라, 제주
comp_size	사업체 종사자 수	10명 미만, 100명 미만, 1000명 미만, 1000명 이상	emp_cont	근로계약기간 지정 여부	예, 아니요
full_part	근로시간 형태	전일제, 시간제	univ_comp_loc	대학 위치와 사업장 위치 일치 여부	일치, 불일치
comm_temp	현 직장 종사상 지위	상용근로자, 임시근로자	retir_allow	법정퇴직금	제공, 미제공
paid_vac	유급휴가	제공, 미제공	parant_leav	육아휴직	제공, 미제공
overt_pay	시간외 수당	제공, 미제공	bonus	상여금	제공, 미제공
regu_nonregu	정규직 여부	정규, 비정규	lab_week_h	주당 근로시간	숫자값
lab_week_d	주당 정규 근로일	숫자값	now_intern	현 직장 인턴 여부	예, 아니요
holi_work	월 평균 휴일 근로일	숫자값	earn_month	월 평균 근로소득 (만원)	숫자값
sati_money	급여 만족도	1~5점	sati_safe	고용 안정성 만족도	1~5점
sati_work	직무내용 만족도	1~5점	sati_workenv	업무환경 만족도	1~5점
sati_workh	근로시간 만족도	1~5점	sati_rel	인간관계 만족도	1~5점
sati_dev	개인 발전가능성 만족도	1~5점	sati_bene	복리후생제도 만족도	1~5점
sati_auth	일의 자율성과 권한 만족도	1~5점	sati_repu	일의 사회적 평판 만족도	1~5점
sati_apti	일과 적성 일치 만족도	1~5점	sati_jobrepu	일자리 사회적 평판 만족도	1~5점
sati_prom	승진제도 만족도	1~5점	sati_train	직무교육 만족도	1~5점
equl_edu	일-교육수준 일치도	1~5점	equl_ablt	일-본인기술 일치도	1~5점

변수명	설명	변숫값	변수명	설명	변숫값
equl_maj	일-전공 일치도	1~5점	asst_maj	전공의 업무 도움 정도	1~5점
help_forelen	업무에 외국어 필요 정도	1~5점	annuity	국민연금	가입, 미가입
health_insu	건강보험	가입, 미가입	hire_insu	고용보험	가입, 미가입
ind_acc_insu	산재보험	가입, 미가입	lab_uni	노동조합	있다, 없다
jobseeking	현 직장 얻기 위한 구직활동 경험 여부	있다, 없다	und_gra_job	재학 중 일자리 경험	있다, 없다
GPA	학점	숫자값	abroad_stdy	어학연수 경험	있다, 없다
eng_test	최근 2년 내 영어 시험 응시 여부	있다, 없다	license	자격증 소지 여부	있다, 없다
num_licen	자격증 개수	숫자값	married	결혼 여부	결혼, 미혼
sex	성별	여, 남	age	나이	숫자값
child	자녀 유무	있다, 없다	target	이직의도	있다, 없다

Ⅳ. 연구 결과

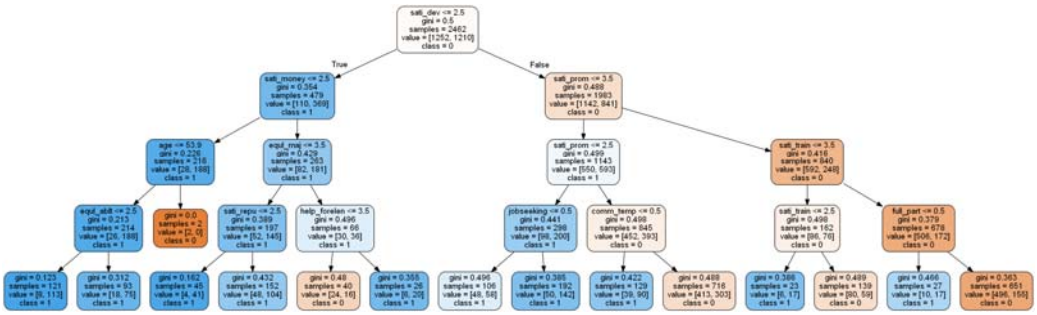
XGBoost 기법을 적용하기에 앞서, 예측 성능의 개선을 확인하기 위해 전통적인 로지스틱 회귀모형과 의사결정나무를 적용해 보았다. 로지스틱 회귀는 종속변수가 0또는 1로 이루어진 범주형 데이터에 사용하는 예측 모형으로, 다양한 독립변수와 하나의 종속변수 간의 상관관계를 분석하여 사건의 발생 가능성을 예측하는 기법이다. 여러 개의 독립변수와 종속변수 간의 인과관계를 표현할 수 있으므로, 여기서 구해진 회귀식을 통해 모델이 수행한 예측에 대한 설명이 가능하다(정동균 등, 2021). 의사결정나무는 IF-THEN 형식으로 표현 가능한 의사결정규칙을 이용하여 예측을 수행하는 방법이다. 의사결정규칙을 통해 나무 모양으로 도표화하며, 각 의사결정규칙에 따른 선택 사항에 대한 확률을 할당함으로써 예측을 수행한다. 의사결정나무는 추론규칙에 의해서 형성이 되므로 예측 과정을 쉽게 이해할 수 있고, 설명이 가능하다는 장점이 있다(김덕현 등, 2019; 이동훈, 김태형, 2020). 각 모델은 예측정확도, 정밀도, 재현율, AUC(Area Under the ROC Curve) 등을 비교하여 전반적인 성능 향상이 있는지를 확인하였다.

검정 데이터에 로지스틱 회귀모형을 적용한 결과, 예측정확도는 0.69, 정밀도는 0.73, 재현율은 0.64, AUC는 0.69가 나왔다. <표 2>는 로지스틱 회귀분석 결과 p-value가 0.05보다 작은 변수들의 목록이다. p-value가 0.05보다 작다는 것은 해당 모델에서 이직의도에 유의하게 영향을 미치는 변수라는 의미이다.

분석 결과 로지스틱 회귀분석에서는 국민연금 가입여부(annuity)와 현 직장을 얻기 위한 구직활동 여부(jobseeking)가 이직의도에 정(+)의 영향을 미치는 것으로 나타났으며, 현 직장 종사상 지위(comm_temp), 근로시간형태(full_part)가 이직의도에 부(-)의 영향을 미치는 것으로 나타났다.

<표 2> 로지스틱 회귀분석 결과 유의미한 변수 목록

variables	coef	std.err	z	p-value	0.025	0.975
comp_loc	-0.1049	0.0306	-3.4293	0.0006	-0.1649	-0.045
comm_temp	-0.5541	0.1859	-2.9809	0.0029	-0.9184	-0.1898
emp_cont	0.4891	0.1525	3.207	0.0013	0.1902	0.788
full_part	-0.5805	0.2441	-2.3779	0.0174	-1.0589	-0.102
lab_week_d	-0.2264	0.1008	-2.2458	0.0247	-0.4239	-0.0288
lab_week_h	0.0179	0.0068	2.6311	0.0085	0.0046	0.0312
sati_money	-0.2367	0.0515	-4.5956	0	-0.3377	-0.1358
sati_work	-0.186	0.0685	-2.7167	0.0066	-0.3201	-0.0518
sati_dev	-0.317	0.0585	-5.4194	0	-0.4317	-0.2024
sati_prom	-0.2159	0.0575	-3.7536	0.0002	-0.3286	-0.1032
asst_maj	-0.1305	0.0482	-2.7055	0.0068	-0.2251	-0.036
annuity	0.6507	0.1982	3.2839	0.001	0.2624	1.0391
jobseeking	0.568	0.093	6.1061	0	0.3857	0.7504
eng_test	0.3326	0.0945	3.5189	0.0004	0.1473	0.5178



<그림 1> 의사결정나무

다음으로 의사결정나무를 적용해 보았다. 분류 기준은 지니 계수(Gini Index)를 사용하였으며, 나무의 최대 깊이는 4로 설정하였다. 검정 데이터에 의사결정나무를 적용한 결과, 예측 정확도는 0.65, 정밀도는 0.74, 재현율은 0.52, AUC는 0.66이 나왔다. 전반적으로 로지스틱 회귀분석보다 성능이 약간 떨어지는 것을 확인할 수 있었다. <그림 1>은 의사결정나무를 적용하였을 때 생성된 트리이다. 의사결정나무는 표본을 분류하는 데 우선적으로 사용된 변수가 예측에 보다 유의미한 변수라고 할 수 있다. 생성된 나무를 살펴보면, 개인 발전가능성 만족도(sati_dev), 급여 만족도(sati_money), 승진제도 만족도(sati_prom), 나이(age), 일-전공 일치도(equl_maj), 직무교육 만족도(sati_train) 등이 이직의도를 예측하는 데 유의미한 변수인 것을 확인할 수 있다.

마지막으로 제안모델인 XGBoost 기법을 적

용해 보았다. 생성되는 나무의 최대 깊이는 이전에 적용한 의사결정나무와 같은 4로 설정하였다. 검정 데이터에 XGBoost 기법을 적용한 결과, 예측 정확도는 0.74, 정밀도는 0.77, 재현율은 0.72, AUC는 0.74가 나와 로지스틱 회귀모형이나 의사결정나무보다 전반적으로 성능 개선이 있는 것을 확인할 수 있었다.

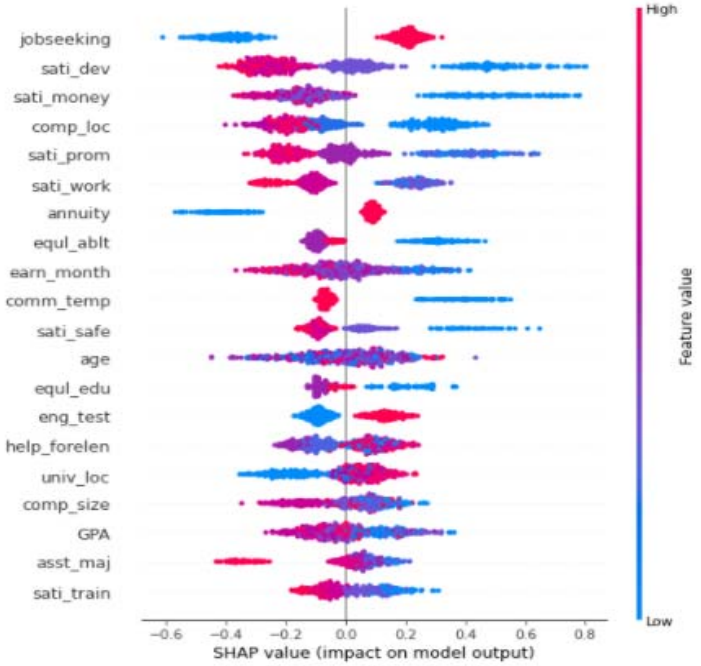
각 모델을 학습 데이터와 검증 데이터에 적용한 결과는 <표 3>과 같다.

블랙박스 모델인 XGBoost에 SHAP 기법을 적용하여 어떤 변수가 이직의도 예측에 영향력을 미치는지 확인해 보았다. <그림 2>는 SHAP 기법을 적용하여 도출한 전역적인 변수 영향도 그래프이다.

이 그래프에서 X축을 기준으로 SHAP value가 0.0보다 음의 값을 가지면 결과에 부(-)의 영향을, 양의 값을 가지면 정(+)의 영향을 미친다는 것을 의미한다. 각 변수의 값이 클수록 붉은

<표 3> 모형 간 성능 비교

	로지스틱 회귀분석		의사결정나무		XGBoost	
	train	test	train	test	train	test
Accuracy	0.69	0.69	0.69	0.65	0.85	0.74
Precision	0.70	0.73	0.74	0.74	0.86	0.77
Recall	0.67	0.64	0.56	0.52	0.84	0.72
AUC	0.69	0.68	0.69	0.66	0.85	0.74



<그림 2> SHAP 기법을 적용한 전역적 변수 영향도

색을 띠고, 값이 작을수록 파란색을 보인다. 해당 그래프를 살펴보면 값이 클수록 이직의도에 정(+)의 영향을 미치는 변수는 현 직장을 얻기 위한 구직활동 여부(jobseeking), 국민연금 가입여부(annuity), 최근 2년 내 영어시험 응시 여부(eng_test)이고, 이직의도에 부(-)의 영향을 미치는 변수는 현 직장 종사상 지위(comm_temp)이다. 현 직장을 얻기 위한 구직활동 여부(jobseeking)는 값이 클수록 이직의도에 정(+)의 영향을 미치는 것으로 나왔다. 해당 값이 클수록 이직의도를 높인다는 것은 구직활동에서 성공한 경험으로 근로자가 자신감을 얻어, 보다 좋은 요건의 직장으로 이직을 도전하기 용이한 것으로 해석된다. 국민연금 가입여부(annuity) 변수에 대해서는 국민연금에 가입된 경우 이직의도가 높아지는 것으로 나왔다. 현재 우리나라에서는 만 60세 이상인 근로자, 1개월 미만의 기간을 정하고 근로하는 일용근로자, 월 소정근로시간이 60시간 미만인 근로자가 아니라면 누구나 국민연금에 의무 가입되어야 하기 때문에, 국민연금에 가입되어 있지 않은 근로자는 매우 열악한 조건에 있는 근로자로서 직장 선택의 폭이 상당히 제약되어 있을 것임을 짐작할 수 있다. 따라서 국민연금 가입이 안 되어 있는 근로자보다는 국민연금 가입이 되어 있는 근로자일수록 이직의도가 높게 나타난 것으로 해석된다. 최근 2년 내 영어 시험 응시 여부(eng_test) 변수의 경우, 한국의 취업시장에서 영어 시험이 필수라고 생각되는 만큼 영어 시험을 응시한 근로자들은 이직을 염두에 둔 경우가 많은 것이라 해석 가능하다. 현 직장 종사상 지위(comm_temp)는 상용직(1)과 임시직(0)을 나타내는 변수로, 재계약 여부가 불투명한 임시직일수록 현재 직장을 계속 다닐 수 있을지 불안감을 느끼게 되어 새로운 직장에 대한 준비를 하는 것이라 판단된다. 개인 발전가능성 만족도(seti_dev), 급여 만족도(seti_money), 일-본인 기술 일치도(equl_ablt), 고용안정성 만족도(sati_safe) 변수의 경우 SHAP value가 0.0에 걸쳐 분포하기는 하지만 해당 값이 특정 임계치보다 더 작아질 경우 이직의도를 높이는 부정적 영향을 미치는 것은 확실해 보인다. 앞에 seti_가 붙은 변수들은 만족도를 측정한 변수로, 개인의 발전가능성, 임금, 고용안정성에서 중간 이상의 만족을 느낄 때는 이직의도에 큰 영향을 미치지 않지만 낮은 만족도를 느낄 때는 이직의도에 매우 큰 영향을 미친다고 해석할 수 있다. 한편 일과 본인 기술 일치 정도를 의미하는 변수(equl_ablt)에 대해서는 값이 낮을수록 이직의도를 높이는 것으로 나와, 자신의 능력을 넘어서는 일을 담당할 경우 해당 직원이 받는 스트레스가 상당하다는 것을 알 수 있다.

다음으로 <그림 3>과 <그림 4>는 개별 표본에 대해 SHAP 기법을 적용한 결과이다. 우선 <그림 3>은 이직의도를 가지고 있는 276번 표본에 대한 결과이다. 276번 표본에 대해서는 고

<그림 3> SHAP 기법을 적용한 276번 표본의 결과

<그림 4> SHAP 기법을 적용한 153번 표본의 결과

용 안정성 만족도(sati_safe), 개인 발전가능성 만족도(sati_dev), 급여 만족도(sati_money), 현 직장을 얻기 위한 구직활동 여부(jobseeking) 변수가 특히 이직의도를 높인 것을 확인할 수 있다. 반면, 일-전공 일치도(equl_maj), 월 평균 휴일 근무일(holi_work), 사업장 위치(comp_loc), 업무에 외국어 필요 정도(help_forelen) 변수는 이직의도를 낮추는 방향으로 작용한 요인이라고 나왔다. 이 중 개인 발전가능성 만족도(sati_dev) 변수와 급여 만족도(sati_money) 변수, 고용안정성 만족도(seti_safe) 변수는 전역적 변수 영향도에서도 이직의도에 영향을 미친다고 나온 변수들이며, 그 외에 새로운 변수들도 해당 표본에 영향을 미친다는 것을 확인할 수 있다. 276번 표본의 경우 일-전공 적합도가 낮음에도 해당 직무에는 만족하고 있는 것으로 보이며, 월평균 3일 휴일근로를 하는 상황도 이직의도를 낮춘 것으로 나왔다. 또한, 사업장이 경상도에 위치하고 있는 점, 업무에서 외국어가 별로 필요하지 않은 상황도 이직의도를 낮춘 것으로 나왔다. 이 결과를 통해 해당 직원이 어떤 부분에서 현재 회사에 만족하고 있고 어떤 부분에서 불만족하고 있는지를 파악할 수 있으며, 만약 앞으로 새로운 회사를 찾는다면 현재 직무와 비슷한 일을 할 수 있으면서 경상도에 있는, 외국어의 필요성은 낮은 회사를 찾을 것이라 예상할 수 있다.

<그림 4>는 이직의도를 가지고 있지 않은 153번 표본에 대한 결과이다.

153번 표본의 경우 사업장 위치(comp_loc)와 시간외 수당(overt_pay) 변수가 이직의도를 높였으나 월 평균 근로소득(earn_month), 개인 발전가능성 만족도(sati_dev), 현 직장을 얻기 위한 구직활동 여부(jobseeking), 급여 만족도(sati_money), 직무내용 만족도(sati_work), 업무에 외국어 필요 정도(help_forelen), 직무교육 만족도(sati_train), 승진제도 만족도(sati_prom) 변수가 이직의도를 크게 낮춘 것을 확인할 수 있다. 해당 표본의 경우 월 평균 근로소득이 500만원으로 높은 월급을 받고 있고 구직활동을 하지 않고 들어온 것으로 보아 스카우트를 통해 해당 직장에 들어오게 된 것일 수도 있다.

이처럼 전역적으로 영향력을 가지는 변수와 표본 각각에 영향을 미치는 변수는 상이할 수 있다. 그렇기 때문에 특정 표본에 특화된 전략을 짜기 위해서는 특정 표본에 대한 변수별 영향도를 살펴보는 것이 필요하다. 전반적인 직원들의 이직은 전역적 영향 변수를 이용해 관리해도 되지만, 특별히 더 신경써야 하는 우수인력의 경우에는 해당 직원의 이직의도에 영향을 미치는 변수를 파악하여 그 직원에 특화된 이직의도 감소 전략을 세울 수 있어야 할 것이다. SHAP 기법을 적용하면 전역적인 특성 중요도뿐만 아니라 각각의 표본에 대한 지역적인 특성 영향도를 확인할 수 있다.

Ⅴ. 결론 및 한계점

본 연구에서는 한국고용정보원에서 조사를 진행한 2019 대졸자직업이동경로조사(GOMS) 자료를 이용하여 이직의도 예측 모형을 만들었다. 보다 높은 예측 성능을 가지는 예측모형을 만들기 위해 블랙박스 모형인 XGBoost 기법을 적용하였으며, 해당 예측 결과에 대해 어떠한 설명도 제공하지 못하는 블랙박스 모델의 한계점을 보완하기 위하여 XAI 방법 중 하나인 SHAP 기법을 적용하였다.

XGBoost 기법을 사용하여 예측 모형을 구축한 결과 74%의 예측 정확도를 보였으며, 전반적인 성능에 있어 기존에 사용하던 기법인 로지스틱 회귀분석이나 의사결정나무보다 개선되었음을 확인하였다. 학습된 XGBoost 모델에 대해 SHAP 기법을 적용한 결과 전역적인 변수 영향도를 확인할 수 있었다. 값이 클수록 이직의도에 정(+)의 영향을 미치는 변수는 '현 직장에 들어오기 위해 구직활동을 하였는지 여부', '국민연금 가입 여부', '최근 2년 내 영어 시험 응시 여부' 등이 있었다. 현 직장을 얻기 위한 구직활동 경험이 있는 경우 이직의도를 높였다는 것은 스스로의 노력으로 취업에 성공함으로써 자신의 능력에 대하여 자신감을 얻게 된 것으로 생각할 수 있다. 본인이 극복한 경험을 토대로 새로운 구직상태도 충분히 극복할 수 있다는 자신감을 얻음으로써 만약 현 직장이 본인의 기대에 못 미치거나 자신의 처우에 불만이 생길 경우 보다 쉽게 더 나은 직장을 찾기 위해 이직을 준비하는 것이라 생각된다. 국민연금 가입여부의 경우 기업의 문제라기보다는 개인의 특성을 나타내는 변수로 판단된다. 현재 국내에서 매우 특별한 예외조건이 아닌 경우 모든 직장에 대해 국민연금 가입을 의무화하였기 때문에, 국민연금에 가입되어 있지 않은 근로자는 연금수령 가능연령 이상의 고령자이거나 직업의 선택 폭이 제약된 일용직에 근무하는 근로자인 것으로 해석될 수 있다. 최근 2년 내 영어 시험 응시 여부의 경우, 한국의 취업시장에서 대부분 영어 시험 성적을 요구하므로 이직의도를 가진 경우 영어 시험에 응시할 가능성이 클 것이라 추측 가능하다.

한편 본 연구에서는 '종사상 지위'가 상용직(1)일수록 이직의도가 감소한다는 점 역시 확인하였다. 언제 퇴직하게 될지 예측하기 어려운, 불안정한 지위인 임시직에 종사하는 경우에는 짧은 계약 기간이나 불투명한 계약 연장 여부 등을 이유로 이직을 항상 염두에 두고 있을 것이라 추측되며, 직장을 옮기는 데 제약이 상용직에 비해 덜하므로 현 직장이 마음에 안 들 경우 다른 곳으로 옮기겠다는 결정도 보다 쉽게 이루어지는 것으로 해석된다.

또한 값이 적당히 클 경우에는 이직의도에 영향을 미치지 않지만, 값이 작을 경우에는 이직의도를 높이는 변수로 '개인의 발전가능성 만족도', '임금 만족도', '고용안정성 만족도', '일과 본인 기술의 일치 정도'가 있음을 확인할 수 있었는데, 이는 유명한 허츠버그(Fredrick Herzberg)의 두 요인 이론(two factor theory) 관점에서 해석될 수 있다. 허츠버그는 이 이론을 통해 직무만족을 통해 동기유발 정도를 높이는 동기요인(motivative factors)과 직무불만족을 통해 동기유발 정도를 낮추는 위생요인(hygine factors)이 상존함을 주장했는데, 본 연구를 통해 확인된 '개인의 발전가능성 만족도',

‘임금 만족도’, ‘고용안정성 만족도’, ‘일과 본인 기술의 일치 정도’는 이 중 위생요인으로서 작용하는 것으로 해석된다. 즉, 이러한 요인들에 대한 만족도가 높아진다고 해서 전반적인 직무만족도가 높아지고 이를 통해 이직의도가 낮아지는 것은 아니지만, 해당 요인들에 대한 만족도가 떨어질 경우에는 전반적인 직무불만족을 유발함으로써 궁극적으로 이직의도를 높이고 있음을 확인할 수 있었다. 이렇듯 직원의 이직의도에는 다양한 요인들이 복잡한 메커니즘을 통해 영향을 미치고 있기 때문에, 기업의 인적 자원 관리 부서에서는 각 요인들이 각자의 특성에 맞게 적절히 관리될 수 있도록 세심한 관심을 기울일 필요가 있음을 알 수 있다.

한편 SHAP 기법을 적용할 경우 개별 표본에 대한 변수별 영향도도 확인할 수 있었다. 구체적으로 이직의도를 가진 직원과 가지지 않은 직원을 각각 한 명씩 살펴본 결과, 전역적으로 영향을 미친다고 나온 변수 이외에도 다양한 변수가 해당 직원의 이직의도에 영향을 미치는 것을 확인할 수 있었다. 이직의도를 가지고 있는 276번 직원의 경우 추가적으로 낮은 일-전공 일치 정도, 많지 않은 월 평균 휴일 근무일수, 사업장 위치, 낮은 업무에 외국어 필요 정도가 이직의도를 소폭 낮추는 방향으로 영향을 미친 것으로 나왔다. 낮은 일-전공 일치도가 이직의도를 낮춘 것으로 보아 해당 업무를 담당하면서 자신에게 더욱 잘 맞는 분야를 발견한 것으로 보인다. 또한 이직의도가 없는 153번 직원의 경우 추가적으로 사업장 위치, 낮은 시간외 수당이 이직의도를 높였으며, 월 평균 근로소득, 높은 업무 만족도, 중간 정도의 업무에 외국어 필요 정도, 높은 직무교육 만족도와 높은 승진 제도 만족도가 이직의도를 크게 낮춘 것으로 나왔다. 이는 전역적 영향 변수를 바탕으로 전략을 세울 경우 해당 전략의 효과를 어느 정도 볼 수는 있지만 각각의 개별적인 대상에 대해서는 기대만큼의 효과가 나오지 않을 수 있다는 것을 의미한다. 따라서 개별 표본에 대해 영향력 있는 변수를 파악하는 것이 필요하고, 지역적 특성 변수를 바탕으로 전략을 수립한다면 보다 특화된 전략을 사용할 수 있으므로 비용 또한 효율적으로 지출할 수 있을 것이다. 예를 들어, 앞에서 살펴 본 이직의도가 있는 276번 직원의 경우 담당 업무를 통해 자신의 적성을 새롭게 발견한 것으로 보이므로, 해당 업무에 대해서 개인의 발전가능성 만족도를 높일 수 있도록 교육을 제공하거나 맡고 있는 업무와 유사하면서도 보다 도전적인 업무를 맡긴다면 해당 직원의 이직의도를 낮출 수 있을 것이라 기대할 수 있다.

본 연구는 높은 예측성능을 가지는 블랙박스 모델을 사용하면서도 XAI 기법을 이용해 설명 가능성을 제공할 수 있다는 것을 보였다는 점에서 학술적인 의의를 갖는다. 기존에는 원인을 파악해야 하는 문제나 설명 가능성을 필요로 하는 문제를 다룰 때, 설명 가능성을 제공하기 위해 상대적으로 예측 성능이 떨어지는 기법을 사용했다. 그러나 XAI 기법을 적용함으로써 블랙박스 모델에도 설명 가능성을 제공할 수 있으므로, 기존의 분야에서도 예측 성능이 높은 모델을 XAI 기법과 함께 사용한다면 개선된 성과를 기대할 수 있을 것이다.

그리고 본 연구는 전반적으로 적용 가능한 전략뿐만 아니라 개개인에 특화된 전략을 세울 수 있음을 보였다는 점에서 실무적인 의의를

갖는다. 전반적인 전략을 사용하면 기대에 미치지 못하는 효과를 얻을 수 있고, 개개인에 특화된 전략만을 세우면 너무 많은 비용이 들어갈 것이다. 그러나 SHAP 기법을 사용한다면 전역적 영향 변수와 지역적 영향 변수를 파악할 수 있으므로, 상대적으로 중요도가 낮은 대상들을 상대로는 전반적인 전략을 세워서 적용하고 중요도가 높은 특정 대상을 상대로는 특화된 전략을 사용한다면 보다 향상된 효과를 얻으면서도 합리적이거나 보다 효율적인 비용을 지출하는 것이 가능할 것이다.

이처럼 여러 학술적, 실무적 의의를 갖고 있지만, 본 연구는 다음과 같은 점에서 한계를 갖는다. 우선 본 연구는 성능이 우수한 여러 기법들에 대해서는 실험을 해 보지 못했다는 한계점을 갖는다. 최근 주목받고 있는 딥러닝 기법을 포함해 XGBoost 이외에도 우수한 예측 성능을 가지는 기계학습 기법들은 다양하게 존재한다. 이번에 적용한 XGBoost 기법의 예측 정확도는 74%로, 나쁘다고는 할 수 없지만 크게 우수하다고도 할 수 없는 수치이다. 예측 성능을 개선하기 위해 후속 연구에서는 여러 딥러닝 기법을 시도해 볼 필요가 있다. 다양한 딥러닝 기법에 XAI 기법을 접목시킴으로써 설명 가능성을 가지는 우수한 예측 모델들을 개발하는 후속 연구가 필요할 것이다.

또한 현재 다양한 XAI 기법들이 존재하고 있는데, 그 중 SHAP 기법만을 적용했다는 것도 본 연구의 한계점이다. SHAP 기법은 다양한 블랙박스 모델에 적용할 수 있는 XAI 기법이지만, 각 블랙박스 모델마다 보다 적합한 XAI 기법이 존재할 수 있다. 각 분야에 적합한 딥러닝 기법을 찾고 그에 적합한 XAI 기법을 적용한다면 보다 효율적이고 우수한 예측 모델을 생성할 수 있을 것이다. 마지막으로 본 연구에서 사용한 데이터는 국내 대학 졸업자를 대상으로 한 설문조사 데이터로서 해외 대학을 나온 취업자나 고졸 취업자에 대한 정보는 포함하지 않고 있다. 현재 대한민국의 노동시장에 참여하고 있는 사람들의 학력은 매우 다양하며 외국에서 우수 인력을 채용하여 데려온 경우도 많을 것이다. 그러므로 이후에는 실제 기업에서 근무하고 있는 다양한 근로자를 대상으로 데이터를 수집하여 대졸자를 대상으로 한 이직 영향 요인과 차이가 있는지 확인하는 연구를 진행할 필요성이 있으며, 만약 차이가 있다면 해당 그룹을 대상으로 한 후속 연구를 진행할 필요가 있다.

참고문헌

고건우, 조현진, 이건창, “근로자들의 이직 의도에 영향을 주는 요인에 관한 실증연구: 공공 데이터베이스와 의사결정나무 기법을 중심으로”, Information Systems Review, 제22권, 제4호, 2020, pp. 41-58

강만수, 윤상용, “이직이 직원 채용 비용에 미치는 영향 연구: 인적자본기업패널을 이용하여”, 기업과혁신연구, 제41권, 제1호, 2018, pp. 17-28

강윤경, 배상영, 홍세희, “대졸자의 직장만족도 잠재계층과 이직의 관련성: 잠재전이분석을 이용한 전이형태와 영향요인 검증”, 인적자원개발연구, 제23권, 제3호,

2020, pp. 1-31

김덕현, 유동희, 정대율, “의사결정나무 기법을 이용한 노인들의 자살생각 예측모형 및 의사결정 규칙 개발”, 정보시스템연구, 제28권, 제3호, 2019, pp. 249-276.

김성훈, 김우진 장연주, 김현철, “설명 가능한 AI 학습 지원 시스템 개발”, 컴퓨터교육학회 논문지, 제24권, 제1호, 2021, pp. 107-115

김정은, 강경주, 이영면, “연령별 이직의사 결정요인에 대한 연구: 요인별 직무만족, 요인별 생활만족을 중심으로”, 노동정책연구, 제17권, 제1호, 2017, pp. 55-84

나광택, 이진영, 김은찬, 이효찬, “증권 금융 상품 거래 고객의 이탈 예측 및 원인 추론”, 한국빅데이터학회지, 제5권, 제2호, 2020, pp. 215-229

배수명, 김희경, “치과위생사의 이직실태와 이직결정 요인에 관한 연구”, 한국산학기술학회논문지, 제13권, 제12호, 2012, pp. 5986-5992

안재현, XAI 설명 가능한 인공지능, 인공지능을 해부하다, 위키북스, 2020

안철경, 정세창, “보험설계사 이직 요인과 정착률 제고 방안”, 보험금융연구, 제28권, 제4호, 2017, pp. 3-33

엄남현, “국내 광고대행사 직원들의 이직의도에 영향을 미치는 요인들에 대한 심층 인터뷰를 통한 탐색적 연구”, 광고학연구, 제30권, 제5호, 2019, pp. 109-124

은화리, 구자복, 정태연, “직무배태성에 따른 이직과 잔류의 영향 요인에 대한 질적 연구”, 한국심리학회지, 제24권, 제2호, pp. 221-250

이동훈, 김태형, “머신러닝 기법을 활용한 대졸 구직자 취업 예측모델에 관한 연구”, 정보시스템연구, 제29권, 제2호, 2020, pp. 287-306.

이만기, “대졸 초기경력자의 이직의도에 미치는 영향요인 분석”, 진로교육연구, 제26권, 제3호, 2013, pp. 61-81

이영석, 이정현, “공항 지상직 서비스 근무자의 이직행동 요인에 관한 연구 -조직헌신의 매개효과를 중심으로-”, 한국산학기술학회논문지, 제17권, 제10호, 2016, pp. 147-157

전희주, “포아송회귀 모형을 활용한 생명보험 설계사들의 이직 요인 분석”, 한국데이터정보 과학회지, 제27권, 제5호, 2016, pp. 1337-1347

정동균, 이종화, 이현규, “머신러닝을 이용한 국내 수입 자동차 구매 해약 예측 모델 연구: H 수입차 딜러사 대상으로”, 정보시스템연구, 제30권, 제2호, 2021, pp. 105-126.

정인호, 이대웅, 권기헌, “청년 취업자의 이직의사 예측모형 탐색 연구: 의사결정나무 모형을 중심으로”, 국정관리연구, 제13권, 제3호, 2018, pp. 147-174

조장식, 강창완, 최승배, “2단계 회귀모형을 활용한 이직 결정요인 분석”, 한국데이터정보과학회지, 제31권, 제1호, 2020, pp. 75-83

채주석, 박상석, 윤경희, “고용안정성 및 정보공유의 인적자원관리 정책이 경영성과에 미치는 영향: 협력적 노사관계분위기의

매개효과", 대한경영학회지, 제32권, 제8호, 2019, pp. 1371-1399

천예은, 김세빈, 이자윤, 우지환, "설명 가능한 AI 기술을 활용한 신용평가 모형에 대한 연구", 한국데이터정보과학회지, 제32권, 제2호, 2021, pp. 283-295

Chen, T., and Guestrin, C. "XGBoost: A Scalable Tree Boosting System", Knowledge Discovery and Data Mining, 2016, pp. 785-794.

Lundberg, S. M., and Lee, S.-I. "A Unified Approach to Interpreting Model Predictions", The 31st Conference on Neural Information Processing Systems, 2017, pp. 1-10.

이 재 준(Lee, Jae Jun)

동국대학교 수학과에서 학사 학위를 취득하였다. 현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학 중이다. 주요 관심분야는 데이터 마이닝, 빅데이터 분석, 신용평가 등이다.

이 유 린(Lee, Yu Rin)

극동대학교 중국항공운항서비스학과에서 항공경영학사를 취득하였다. 현재 국민대학교 비즈니스IT전문대학원에서 석사과정에 재학 중이다. 주요 관심 분야는 추천시스템, 텍스트 마이닝 등이다.

임 도 현(Lim, Do Hyun)

광운대학교 산업심리학과에서 산업심리학 학사를 취득하였다. 현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학중이다. 주요 관심분야는 머신러닝, 사기탐지, 딥러닝 등이다.

안 현 철(Ahn, Hyun Chul)

KAIST에서 산업경영학사를 취득하고, KAIST 테크노경영대학원에서 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 현재 국민대학교 비즈니스IT전문대학원 교수로 재직 중이다. 주요 관심분야는 정보시스템 수용과 관련한 행동 모형, 금융 및 고객관계관리 분야의 인공지능 응용 등이다.

<Abstract>

A Study on the Employee Turnover Prediction using XGBoost and SHAP

Lee, Jae Jun · Lee, Yu Rin · Lim, Do Hyun · Ahn, Hyun Chul

Purpose

In order for companies to continue to grow, they should properly manage human resources, which are the core of corporate competitiveness. Employee turnover means the loss of talent in the workforce. When an employee voluntarily leaves his or her company, it will lose hiring and training cost and lead to the withdrawal of key personnel and new costs to train a new employee. From an employee's viewpoint, moving to another company is also risky because it can be time consuming and costly. Therefore, in order to reduce the social and economic costs caused by employee turnover, it is necessary to accurately predict employee turnover intention, identify the factors affecting employee turnover, and manage them appropriately in the company.

Design/methodology/approach

Prior studies have mainly used logistic regression and decision trees, which have explanatory power but poor predictive accuracy. In order to develop a more accurate prediction model, XGBoost is proposed as the classification technique. Then, to compensate for the lack of explainability, SHAP, one of the XAI techniques, is applied. As a result, the prediction accuracy of the proposed model is improved compared to the conventional methods such as LOGIT and Decision Trees. By applying SHAP to the proposed model, the factors affecting the overall employee turnover intention as well as a specific sample’s turnover intention are identified.

Findings

Experimental results show that the prediction accuracy of XGBoost is superior to that of logistic regression and decision trees. Using SHAP, we find that jobseeking, annuity, eng_test, comm_temp, seti_dev, seti_money, equl_ablt, and sati_safe significantly affect overall employee turnover intention. In addition, it is confirmed that the factors affecting an individual’s turnover intention

are more diverse. Our research findings imply that companies should adopt a personalized approach for each employee in order to effectively prevent his or her turnover.

Keyword: Employee Turnover, XGBoost(eXtreme Gradient Boosting), XAI(eXplainable AI), SHAP(SHapley Additive exPlanations)

* 이 논문은 2021년 8월 5일 접수, 2021년 9월 10일 1차 심사, 2021년 10월 18일 게재 확정되었습니다.