디지털산업정보학회 논문지 제19권 제3호-2023년 9월
http://dx.doi.org/10.17662/ksdim.2023.19.3.177

부스팅 기계 학습과 SHAP를 이용한 설명 가능한 소프트웨어 분야 대졸자 취업 모델 개발

권 준 희* · 김 성 림**

Explainable Software Employment Model Development of University Graduates using Boosting Machine Learning and SHAP

Kwon Joonhee · Kim Sungrim

〈Abstract〉

The employment rate of university graduates has been decreasing significantly recently. With the advent of the Fourth Industrial Revolution, the demand for software employment has increased. It is necessary to analyze the factors for software employment of university graduates. This paper proposes explainable software employment model of university graduates using machine learning and explainable AI. The Graduates Occupational Mobility Survey(GOMS) provided by the Korea Employment Information Service is used. The employment model uses boosting machine learning. Then, performance evaluation is performed with four algorithms of boosting model. Moreover, it explains the factors affecting the employment using SHAP. The results indicates that the top 3 factors are major, employment goal setting semester, and vocational education and training.

Key Words : Machine Learning, Explainable AI, SHAP, Software Employment, University Graduates

Ⅰ. 서론

4차 산업 혁명을 이끄는 핵심 기술은 지능형 정보 기술이다. 지능형 정보 기술은 인공지능으로 구현되는 지능과 빅데이터에 기반한 정보가 결합된 소프트웨어 중심 기술로 정의할 수 있다[1]. 4차 산업 혁명이 도래하면서 최근 소프트웨어 분야 인력 수요가 크게 증가하고 있다.

우리나라 대학 교육 이수율은 경제협력개발기구(OECD) 회원국에서 매우 높은 수준이지만, 대졸자 취업률은 OECD 평균에도 미치지 못할 정도로 낮은 수치를 보이고 있다[2]. 이에 따라, 대졸자의 취업률을 높일 수 있는 방안 중 하나로, 최근 수요가 크게 증가한 소프트웨어 분야에 대졸자가 취업할 수 있는 요인을 분석할 필요성이 있다.

본 연구에서는 소프트웨어 분야 대졸자 취업 모델을 개발한다. 또한 개발된 대졸자 취업 모델을 설명함으로써 취업 요인을 분석한다.

* 경기대학교 AI컴퓨터공학부 교수
** 서일대학교 소프트웨어공학과 교수(교신저자)

취업 모델을 개발하고 이를 설명하기 위해 본 논문에서는 다음과 같은 접근 방법을 취한다. 첫째, 대졸자의 최근 취업 공공 데이터를 사용한다. 이를 위해, 한국고용정보연구원에서 제공하는 대졸자 직업 이동 경로 조사 데이터 중 가장 최근의 데이터를 사용한다. 둘째, 인공지능 기법을 활용하여 데이터 기반 기계 학습을 이용하여 모델을 개발한다. 이를 위해 뛰어난 성능을 보이는 기계 학습 기법인 부스팅 기계 학습을 사용한다. 셋째, 모델에서 취업에 영향을 미치는 요인을 분석할 수 있도록 설명 가능한 인공지능 기법을 사용한다. 이를 위해 SHAP를 이용하여 모델을 설명한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구의 이론적 배경이 되는 부스팅 기계 학습, 설명 가능한 인공지능, 대졸자 직업 이동 경로 조사 데이터에 대해 살펴본다. 3장에서는 기계 학습을 이용한 소프트웨어 분야 대졸자 취업 모델을 제안하고, 4장에서는 SHAP를 이용하여 소프트웨어 분야 대졸자 취업 모델을 설명한다. 마지막으로 5장에서 본 논문의 결론을 맺는다.

II. 이론적 배경

2.1 부스팅 기계 학습

기계 학습(machine learning)은 컴퓨터가 명시적인 프로그래밍 없이, 학습할 수 있는 능력을 부여하는 분야로 정의되고, '데이터라는 형태로 얻어지는 경험으로부터 특정한 목표 작업에 대한 성능을 향상시키는 일련의 과정'이라고도 정의된다[3].

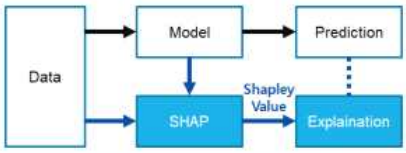
컴퓨터가 스스로 규칙을 발견하기 위해서는 대용량의 데이터를 통한 학습이 수반되어야 한다. 기계 학습을 통해 풀 수 있는 문제는 크게 주어진 데이터의 클래스를 구분해야 하는 분류(classification) 문제와 연속적인 어떤 값을 추정해야 하는 회귀(regression) 문제로 나눌 수 있다. 또한 학습 방법에 따라 지도 학습(supervised learning), 비지도 학습(unsupervised learning), 반지도 학습(semi-supervised learning), 강화 학습(reinforcement learning)으로 나눌 수 있다. 즉, 기계 학습은 사람이 풀 수 있지만 학습하기에 너무 문제의 규모가 크거나 또는 수학적으로 문제를 정의할 수 있지만 문제가 복잡하여 수학적으로 깨끗하게 정의를 하는 것이 어려울 때에 유용하게 쓰일 수 있다[4].

기계 학습에서 사용하는 기법 중 하나인 앙상블 기법은 트리 기반 알고리즘으로, 다수 알고리즘의 학습 결과를 이용하여 보다 좋은 성능을 내고자 하는 방법으로써 과적합을 막아주고 최적 해를 찾기가 쉽다. 앙상블 기법을 활용하면 모델의 예측 오류를 줄여 전체 성능을 향상시킬 수 있는데, 이는 단일 모델을 사용하는 경우보다 효과적이다. 앙상블 기법은 정형 데이터의 예측과 분석 영역에서 매우 뛰어난 성능을 보여준다[5, 6].

앙상블의 대표적인 방법으로는 보팅(Voting), 배깅(Bagging), 부스팅(Boosting), 스태킹(Stacking) 등의 방법이 있다[7]. 이 중, 부스팅 방법은 캐글 등의 경진대회에서 뛰어난 성적을 보이고 있어 최근 크게 주목받고 있다. 부스팅 알고리즘은 여러 개의 분류기를 차례대로 학습하면서 잘못 예측한 데이터에 가중치를 부여해 오류를 개선해 나가는 학습방식이다. 가중치를 부스팅하면서 학습을 진행하므로 부스팅 방법으로 부른다[7].

부스팅 방법에는 GBM(Gradient Boosting Machine), XGBoost, LightGBM, Catboost 등이 대표적이다. GBM은 기울기를 부스팅하여 알고리즘을 학습하고, XGBoost는 GBM의 단점을 개선한 알고리즘이다. LightGBM은 XGBoost를 발전시켜 더 큰 데이터셋에 대해 더 빠른 속도로 작동하도록 설계된 알고리즘이다[8]. CatBoost는 범주형 변수 전처리와 오버

피팅 문제를 해결하는 데 초점을 둔 순서대로 진행되는 부스팅 알고리즘이다. 순서가 있는 부스팅은 기존의 부스팅 모델과는 다르게 모든 잔여 오차를 차례로 학습하는 대신, 일부 데이터의 잔여 오차를 계산하여 모델을 구축하고, 이 모델을 통해 남은 데이터의 잔여 오차를 추정하는 방식으로 작동하는 기법이다[9].



<그림 1> SHAP 알고리즘 작동 방식

2.2 설명 가능한 인공지능

설명 가능한 인공지능(Explainable AI; XAI)은 주어진 데이터를 분류하고 예측할 뿐만 아니라, 결정에 대한 인과 관계를 분석하여 논리적인 근거를 제시할 수 있으며, 다양한 분야의 지식과 기술을 융합하여 개발되는 인공지능 기술이다. 이를 통해 사용자 수준에서 인공지능 모델의 의사 결정 과정을 설명할 수 있다[10].

설명 가능한 인공지능에서 활발하게 연구가 진행중인 분야에는 LIME(Local Interpretable Model-agnostic Explanations), SHAP(SHapley Additive exPlanations), SA(Sensitivity Analysis), LRP(Layer-wise Relevance Propagation) 등이 있다. 설명 가능한 인공지능 알고리즘은 일반적으로 모델 자체를 해석하는 방법과, 왜 그런 결정을 내렸는지에 대해 데이터로 설명하는 방법으로 구분되며 현재에도 다양한 알고리즘이 활발히 연구되고 있는 분야이다[11].

SHAP 알고리즘은 기계학습 모델의 예측 결과를 설명하는 데 사용되는 알고리즘으로써 Shapley Values 개념을 적용하여 특성들이 모델의 예측에 어떤 영향을 미치는지 설명하려는 목적이 있다. Shapley Values는 특정한 특성의 중요도를 파악하기 위해 다양한 특성들의 조합을 고려한 후에 해당 특성의 존재 유무에 따른 평균적인 변화를 계산하여 나온 값이다[11]. SHAP 알고리즘 작동 방식은 <그림 1>과 같다. 설명 모델은 학습 데이터와 학습된 모델을 활용하여 구축되며, 새로운 입력 데이터에 대한 예측 결과에 대한 영향을 방향과 크기로 나타내는 Shapley Values를 계산하여, 입력 피처가 학습된 모델의 출력에 어떤 기여를 하는지 설명한다[11].

SHAP는 다른 방법론과 달리 견고한 이론적 기반을 가지고 가중치를 측정하기 때문에 도출되는 값에 대한 높은 신뢰도를 가지고 있으며, SHAP 값에 따라 각 특징의 영향력을 해석할 수 있다[12].

2.3 대졸자 직업 이동 경로 조사 : GOMS

대졸자 직업 이동 경로 조사(GOMS, Graduates Occupational Mobility Survey)는 매년 신규로 노동시장에 진입하는 전문대학 이상의 대학 졸업자를 대상으로 교육 과정, 재학 중 경력 개발과 취업 경험, 직업과 임금, 노동 시장 이동, 진로 탐색, 직업 훈련, 가계 배경 등의 정보를 수집하는 조사이다. GOMS는 매년 전년도 2~3년제 대학, 4년제 대학, 교육대학 졸업자를 모집단으로 하여, 약 4%에 해당하는 1만 8천 명을 표본 추출하여 졸업년도 다음 해 9월부터 3개월간 조사를 실시하여 구축하는 통계청으로부터 공식 승인을 받은 정부 승인 데이터이다. 현재, 2005년 졸업생부터 2019년 졸업생까지 매년마다 데이터가 구축되어 있다[13].

GOMS는 대학 졸업자의 경력 개발 및 직업 이동 경로를 추적 조사하여 구축된 신뢰성 있는 데이터이다. 이에 따라, 해당 데이터를 분석함으로써 교육과 노동 시장과의 관계에 대한 정보를 제공하여, 고학력 청년 실업 문제 극복을 위한 효과적인 정책 수립을

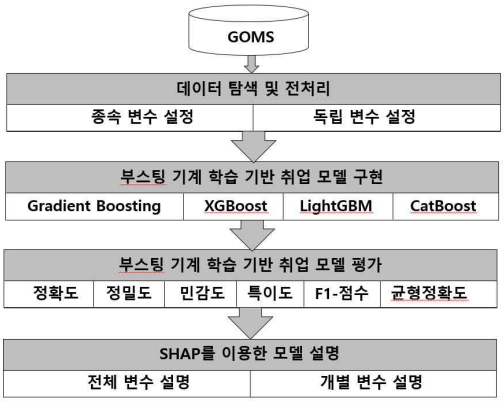
위한 기초 자료로 널리 활용되고 있다[13].

GOMS 데이터를 활용한 기존 연구들을 살펴보면 다음과 같다. 성별과 전공 계열에 따른 노동 시장 진출 결정 요인, 가족의 사회 경제적 배경과 대학원 진학 확률의 성별 격차, 다전공 선택 요인 및 다전공자 대학 생활 실태 분석, 대졸자의 정규직 취업에서 첫 직장의 중요성 등에 대한 연구들이 대표적이다. GOMS 데이터를 활용하여 연구할 때 사용한 방법으로는 선형 확률 모델(Linear Probability Model), 회귀 모형, 이항 로지스틱 회귀 분석 등의 방법을 활용하였다. 그러나, 소프트웨어 분야 대졸자 취업 관련으로는 기계 학습과 설명 가능한 인공지능을 이용한 연구는 없었다[14].

Ⅲ. 기계 학습을 이용한 소프트웨어 분야 대졸자 취업 모델 개발

3.1 모델 개발 전체 과정 및 분석 자료

본 연구에서는 기계 학습을 이용하여 소프트웨어 분야 대졸자 취업 모델을 개발하고, 설명 가능한 인공지능 기술을 이용하여 취업 요인을 분석한다. <그림 2>는 본 논문에서 제안하는 모델의 전체 개발 과정이다.



<그림 2> 모델 개발 전체 과정

본 연구에서는 한국고용정보연구원에서 제공하는 2020년 GOMS의 데이터를 사용한다. 해당 데이터는 2018년 8월 및 2019년 2월 대학교 졸업자를 대상으로 구축된 데이터로, GOMS 데이터 중 가장 최근의 데이터이며, 18,271명을 대상으로 1,332개의 항목으로 구성된 정형 데이터의 형태를 가진다.

본 논문은 청년층의 4년제 일반 대학 졸업 이후 취업 모델을 목표로 한다. 이에 따라, 40세 이상, 2~3년제 대학 졸업자와 교육대학 졸업자는 제외하면 분석 대상이 된 대학 졸업자의 수는 13,917명이다.

3.2 변수 구성

3.2.1 종속 변수

본 연구의 종속 변수는 소프트웨어 분야 취업 중 괜찮은 일자리에 취업했는지 여부이다. 우선, 소프트웨어 분야 취업 여부는 한국고용직업분류(KECO) 2018의 세분류 직업 코드[15] 중 소프트웨어 분야 직업 세분류 코드값인 '1312'~'1350', '2142', '137'을 GOMS 항목 '직장 직업 세분류 코드'값으로 가지는 경우 소프트웨어 분야에 취업한 것으로 본다.

괜찮은 일자리 여부를 결정하는 기준에 대한 그동안의 취업 연구에서 공통되게 나타나는 요인은 임금과 안정성이다[2]. 이 중 본 연구에서는 안정성을 중요 요인으로 보고 종속 변수를 다음과 같이 구성한다. 즉, 대졸자의 첫 번째 직장과 현재 일자리(2020년 8월25일~8월31일 기간에 근무한 일자리)에서 상용근로자이면서 정규직으로, 소프트웨어 분야의 직업에 취업한 경우 종속변수는 1의 값, 그렇지 않으면 0의 값을 갖는다.

상용근로자이면서 정규직인지 여부를 판단하기 위

한 GOMS 항목은 '종사상 지위'와 '정규직 여부'이다. GOMS 에서 '종사상 지위' 항목은 상용근로자, 임시근로자, 일용근로자, 무급가족종사자 중 1개의 값을 가지는데, 상용근로자란 '고용계약기간이 1년 이상인 정규직원 또는 특별한 고용계약 없이 기간이 정해져 있지 않더라도 소정의 채용절차에 의해 입사하여 인사관리 규정을 적용받는 사람'으로 정의된다. '정규직 여부' 항목은 예와 아니오 중 1개의 값을 가진다.

<표 1>은 본 연구에서 사용한 종속 변수이다. 여기서 변수 이름과 변수 값은 모델 개발을 위해 생성한 종속 변수 이름과 각 종속 변수가 가질 수 있는 모든 값이며, GOMS 항목은 모델의 종속 변수 생성을 위해 사용한 원시 데이터인 GOMS 항목의 이름이다.

<표 1> 종속 변수

변수 이름	변수 값	GOMS 항목
소프트웨어 분야 직업 상용근로자/정규직 취업 유무	1 (취업)	직장 직업 세분류 코드
		직장 정규직 여부
	0 (미취업)	직장 종사상 지위

3.2.2 독립 변수

GOMS 항목은 크게 경제 상황, 일자리 경험, 학교생활, 외국어 수준, 졸업 전 취업 목표, 직업 능력 향상 교육 및 훈련, 출신 고등학교, 졸업 후 진학 경험, 시험 준비, 신체 및 정신 건강 정도, 인적 사항 및 가족으로 크게 구분된다.

본 연구에서는 독립 변수로써 다음과 같은 항목은 고려하지 않는다. 첫째, 경제 상황, 신체 및 정신 건강 정도, 인적 사항 및 가족 항목을 고려하지 않는다. 둘째, 시험 준비 항목을 고려하지 않는다. 이는 소프트웨어 분야 취업의 경우 고시, 회계사 등과 같은 전문 시험을 통해 취업하는 경우가 많지 않기 때문이다. 셋째, 외국어 수준 항목을 고려하지 않는다. 이는 대부분의 4년제 대학교 졸업 요건으로 외국어 수준이 명기되어 있어 4년제 대학교를 졸업했다면, 일정 수준 이상의 외국어 능력을 보유하고 있다고 볼 수 있기 때문이다. 또한 소프트웨어 분야 취업의 경우 외국어 능력이 취업을 결정하는데 주요 요인이 되지 않는다고 판단하였기 때문이다. 넷째, 출신 고등학교와 졸업 후 진학 경험과 같은 대학교 입학 전 정보와, 졸업 후 취업과 직접적인 관련이 없는 진학 정보는 제외한다. 넷째, 결측치값을 추정하기 어려운 항목은 제외한다. 예를 들면, 졸업 평점 점수와 같은 항목의 경우 결측치값에 대해 임의 추정이 어렵다. 이에 따라, 최종적으로 선정된 독립 변수는 <표 2>와 같다.

<표 2> 독립 변수

변수 이름	변수 값	GOMS 항목
전공	0 (비전공)	전공 소분류 코드
	1 (유사 전공)	
	2 (소프트웨어 전공)	
직업교육훈련 (횟수)	0~3	직업 교육/ 훈련 분야
인턴경험	0 (인턴경험 없음)	인턴경험 직장 직업 세분류 코드
	1 (소프트웨어 비관련)	
	2 (소프트웨어 분야)	
졸업유예기간 (학기)	0~12	졸업유예 이유, 유예 학기 수
휴학기간 (학기)	0~16	휴학 이유, 휴학 기간
취업목표설정 시기(학기)	0~12	졸업 전 취업목표 설정시기
취업준비시작 시기(학기)	0~13 (13: 졸업 후 취업 준비)	취업준비 시작시기

'전공' 변수는 GOMS 항목 '전공 소분류 코드'값이 한국교육개발원의 중분류 학과코드[16]값 중 '0408'인 경우 소프트웨어 전공 변수 값 2를 부여하고, 코드값 '0404', '0405', '0409', '0504', '0201'에 대해서는 소프트웨어 유사 전공 변수 값 1을 부여한다. 그 이외의 값에 대해서는 비전공 변수 값 0을 부여한다.

'직업교육훈련(횟수)' 변수는 GOMS 항목 '직업 교

육/훈련 분야'의 값이 '컴퓨터 교육' 분야인 경우에 대해서만 1회의 훈련을 받은 것으로 한다. GOMS 항목 '직업 교육/훈련 분야'는 최대 3회까지의 교육에 대해서만 응답하도록 구성되어 있어, '직업교육훈련(횟수)'의 가능한 최대값은 3이다. GOMS 항목값이 결측치값인 경우에는 0으로 처리한다.

'인턴경험' 변수값은 인턴을 경험한 직장의 직업 세분류 코드값이 소프트웨어 분야 직업 세분류 코드값인 '1312'~'1350', '2142', '137'을 가지는 경우 소프트웨어 분야 인턴경험 값 2를 부여한다. 인턴 경험은 있으나, 소프트웨어 분야 직업에서 근무하지 않은 경우에는 1값을, 인턴 경험이 없거나 결측치값인 경우는 0값을 부여한다.

'졸업유예기간(학기)' 변수값은 GOMS 항목 '졸업 유예 이유'가 '취업'과 관련된 경우에만 유예인 것으로 보고, 해당 GOMS 항목 '졸업 유예 학기수'를 졸업 유예 기간으로 설정한다. 졸업 유예를 하지 않았거나, 결측치값인 경우는 0 값을 가진다.

'휴학기간(학기)' 변수값은 GOMS 항목 '졸업유예 이유'가 '취업'과 관련된 경우에만 휴학인 것으로 보고, 해당 GOMS 항목 '휴학 기간' 값으로 설정한다. 휴학을 하지 않았거나, 결측치값인 경우 변수값은 0이다.

'취업목표설정시기(학기)' 변수값은 GOMS 항목 '졸업 전 취업 목표 설정 시기'값으로 한다. 취업목표를 설정하지 않거나 결측치값인 경우 변수값은 0이다.

'취업준비시작시기(학기)' 변수값은 GOMS 항목 '취업준비 시작시기'로 한다. 취업 준비를 시작한 시기가 없거나 결측치값을 가지는 경우에는 0 값을, 졸업 후 취업 준비를 한 경우는 13 값을 가진다.

3.3 부스팅 기계 학습 기반 취업 모델 구현

소프트웨어 분야 대졸자 취업 모델은 개발 단계별로 다음과 같은 방법을 사용하여 구현하였다. 첫째, 데이터 탐색과 변수 구성을 위한 데이터 전처리 단계에 파이썬을 사용했다. 둘째, 부스팅 기계 학습을 이용한 취업 모델을 구현하기 위해 파이썬에서 제공되는 부스팅 기계 학습 모델 라이브러리와 사이킷런(sckit-learn)을 사용했다. 본 연구에서는 정형 데이터 기계 학습 모델에서 뛰어난 성능을 보이는 것으로 알려진 부스팅(boosting) 모델 중 Gradient Boosting, XGBoost, LightGBM, CatBoost 모델의 분류(classification) 모델을 사용한다. 셋째, 범주형 변수인 '전공'과 '인턴경험' 변수를 레이블 인코딩(label encoding) 방식으로 처리하여 사용했다. 로지스틱 회귀 등과 같은 선형 계열 모델에서는 변수값이 0과 1이 아닌 다른 값을 가지게 되면 값의 차이가 모델에 영향을 미치기 때문에 원 핫 인코딩(one-hot-encoding)에 의한 이진법 값 변환 등의 기법을 사용해야 올바른 결과를 얻을 수 있다. 그러나, 본 연구에서 사용한 부스팅 모델과 같은 트리 기반 모델에서는 숫자의 차이가 모델에 영향을 주지 않아 서로 다른 범주값에 임의의 숫자값을 부여하는 레이블 인코딩 방식을 사용할 수 있으며, 이진법 값 변환을 앙상블 모델에서 사용하는 경우 비효율적이다[17].

취업 모델을 구현하는데 있어, 전체 데이터 중 훈련 데이터의 비율은 80%로 설정하고, 각 모델에서 사용한 하이퍼패라미터 값은 <표 3>과 같다. 하이퍼패러미터는 StratifiedK 폴드 교차검증과 그리드 서치(grid search) 기법을 이용하여 설정하였다.

<표 3> 하이퍼패라미터

모델	패라미터	값
GradientBoosting	learning_rate	0.1
	max_depth	3
XGBoost	n_estimators	100
	learning_rate	0.3
	max_depth	3
	reg_alpha	0.1

	reg_lambda	0.2
LightGBM	gamma	0.2
	learning_rate	0.1
	max_depth	-1
	n_estimators	100
	num_leaves	31
CatBoost	depth	6
	iterations	1000
	learning_rate	0.03

3.4 부스팅 기계 학습 기반 취업 모델 평가

본 연구의 모델은 종속 변수의 값이 취업과 미취업으로 분류되는 이진 분류 모델이다. 또한, 모델 개발에 사용한 데이터는 전체 데이터 개수 13,917개 중 종속 변수의 값이 취업인 경우 579개, 미취업인 경우 13,338개로 미취업인 학습 데이터의 개수가 지나치게 많은 불균형 데이터(imbalanced data)이다. 불균형 데이터를 사용한 이진 분류 모델의 평가 지표는 정확도(accuracy) 외에도, 다른 평가 지표인 정밀도(precision), 민감도(sensitivity), 특이도(specificity), F1-점수(F1-score), 균형 정확도(balacned accuracy)를 함께 사용하는 것이 바람직하다[18].

평가 지표는 혼동 행렬(Confusion Matrix)의 TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative)를 기반으로 계산된다. 여기서 TP는실제값과 예측치가 모두 참, TN은 실제값과 예측치 모두 거짓, FP는 실제값은 거짓인데 예측값은 참, FN은 실제값은 참인데 예측값이 거짓임을 의미한다. 본 논문에서 사용한 평가 지표는 식 (1) ~ 식(6)에 의해 계산된다.

$$\text{정확도} = \frac{TN+TP}{TN+FP+FN+TP} \quad (1)$$

$$\text{정밀도} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{민감도} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{특이도} = \frac{TN}{TN+FP} \quad (4)$$

$$F1-\text{점수} = 2 \times \frac{\text{정밀도} \times \text{민감도}}{\text{정밀도} + \text{민감도}} \quad (5)$$

$$\text{균형 정확도} = \frac{\text{민감도} + \text{특이도}}{2} \quad (6)$$

<표 4>는 모델별 성능 평가 결과이다. 정확도는 XGBoost 모델이 가장 높고, GradientBoosting 모델과 CatBoost 모델이 가장 낮다. 정밀도는 XGBoost 모델이 가장 높고, CatBoost 모델이 가장 낮다. 민감도는 XGBoost 모델, LightGBM 모델, CatBoost 모델이 가장 높고, GradientBoosting 모델이 가장 낮다. 특이도는 GradientBoosting 모델이 가장 높고, CatBoost 모델이 가장 낮다. F1-점수와 균형 정확도는 XGBoost 모델이 가장 높고, GradientBoosting 모델이 가장 낮다.

<표 4> 모델별 성능 평가 결과

성능지표 \ 모델	Gradient Boosting	XGBoost	Light GBM	Cat Boost
정확도(%)	96.52	96.59	96.55	96.52
정밀도(%)	60	61.11	59.46	57.89
민감도(%)	17.48	21.36	21.36	21.36
특이도(%)	99.55	99.48	99.44	99.4
F1-점수(%)	27.07	31.65	31.43	31.21
균형정확도(%)	58.51	60.42	60.4	60.38

IV. SHAP를 이용한 소프트웨어 분야 대졸자 취업 모델 설명

본 장에서는 개발된 모델에서 취업에 영향을 미친 독립 변수와 각 변수의 영향 정도를 통해 취업 요인을 설명한다. 이를 위해 파이썬의 SHAP 라이브러리를 이용하여 구현하였다. 본 연구에서는 SHAP 값 계산 이외에도 이를 시각화함으로써 개발된 취업 모델

을 보다 이해하기 쉽게 설명한다.

4.1 전체 변수 설명

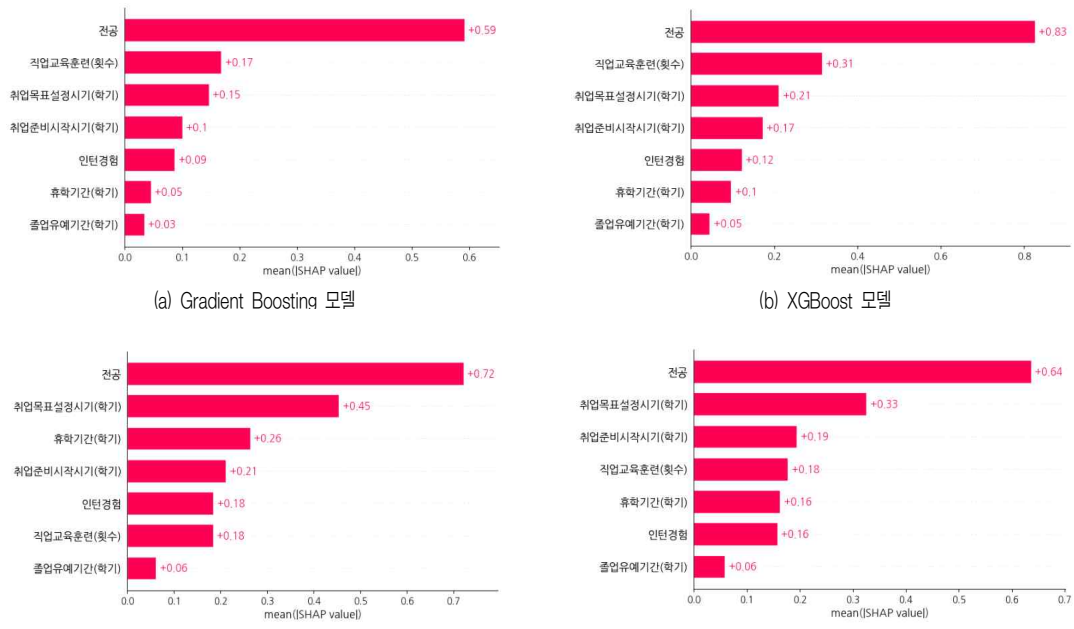
본 절에서는 취업에 영향을 미친 독립 변수들의 중요도와 영향을 설명한다. 이를 위해, 파이썬의 SHAP 라이브러리에서 제공하는 global bar plot과 beeswarm plot을 이용하여 시각화한다.

<그림 3>은 SHAP의 global bar plot을 이용한 취업 요인 설명 결과이다. SHAP의 global bar plot은 각 영향 변수의 SHAP 값의 절대값 평균을 막대 그래프 형태로 큰 값에서 작은 값 순으로 보여준다. 즉, 중요한 변수일수록 막대 그래프의 길이가 길게 나타난다.

<그림 3>에서 각 모델별로 중요 변수의 순위는 서로 다르게 나타난다. 모든 모델에 대한 전체적인 중요 변수를 설명하기 위해 중요 변수의 모든 모델에 대한 평균 중요 순위를 계산해보면, 전공은 1위, 취업목표설정시기(학기)는 2.5위, 직업교육훈련(횟수)는 3.5위, 취업준비시작시기(학기)는 3.75위, 휴학기간(학기)는 5위, 인턴경험은 5.25위, 졸업유예기간(학기)는 7위이다. '전공' 변수는 모든 모델에서 공통적으로 가장 중요한 변수로 나타났고, SHAP 값도 다른 변수에 비교해 매우 높은 수치로 나타나, 소프트웨어 분야 취업을 위해 대학의 '전공'이 미치는 영향이 매우 크다는 것을 알 수 있다.

<그림 4>는 SHAP의 beeswarm plot을 이용한 취업 요인 설명 결과이다. SHAP의 beeswarm plot은 각 변수가 모델 예측에 어떤 방향으로 영향을 미치는지를 보여주는 밀도를 SHAP의 평균 절대값 순서로 정렬하여 보여준다. 이 때, SHAP 값이 음수라는 것은 예측값을 감소시켰다는 것이고, 양수는 예측값을 증가시켰다는 것을 의미한다. 또한, 각 변수값이 높을수록 붉은 색으로 표시되고, 변수값이 낮을수록 푸른 색으로 표시된다.

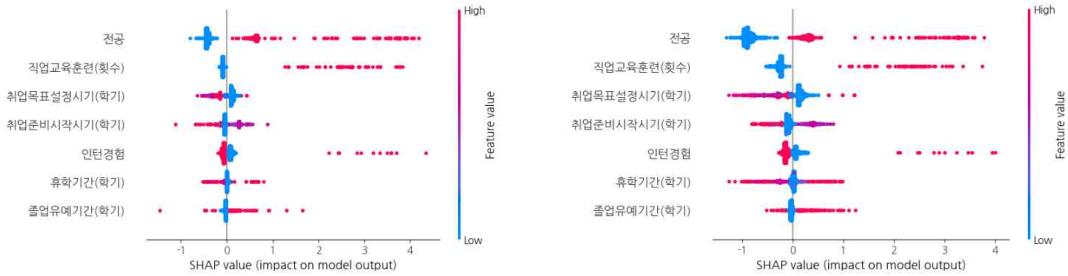
<그림 4>로부터 다음을 알 수 있다. '전공', '직업교



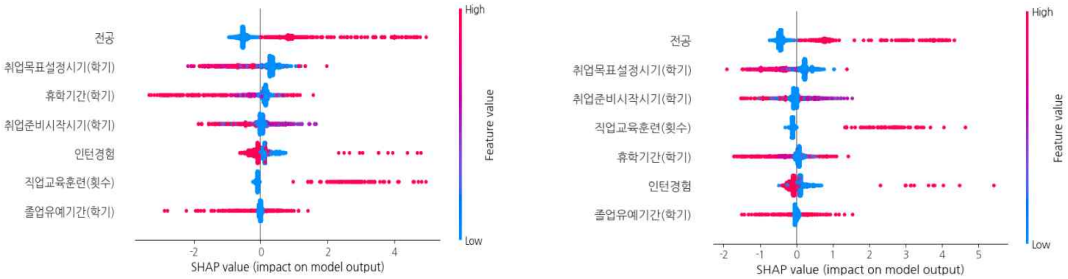
(a) Gradient Boosting 모델 (b) XGBoost 모델

(c) LightGBM 모델 (d) CatBoost 모델

<그림 3> SHAP global bar plot

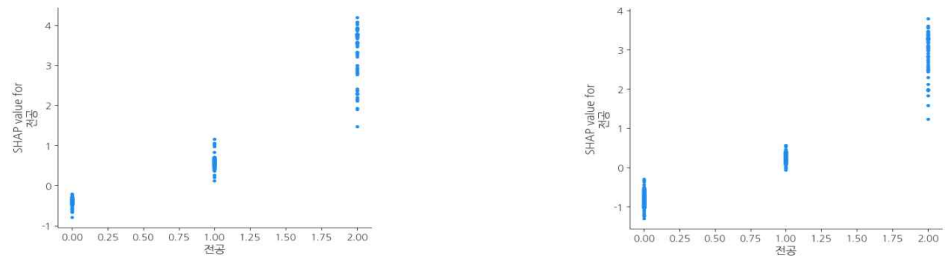


(a) Gradient Boosting 모델 (b) XGBoost 모델

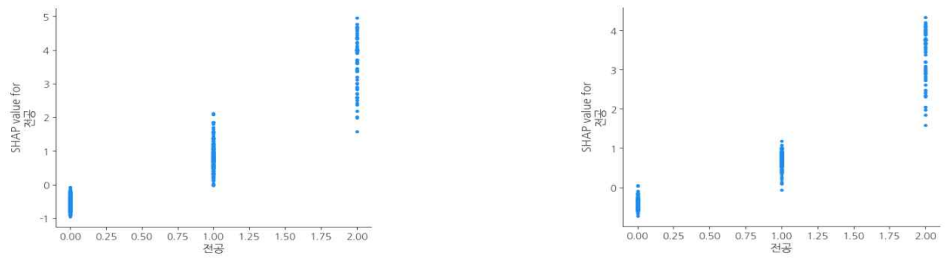


(c) LightGBM 모델 (d) CatBoost 모델

<그림 4> SHAP beeswarm plot

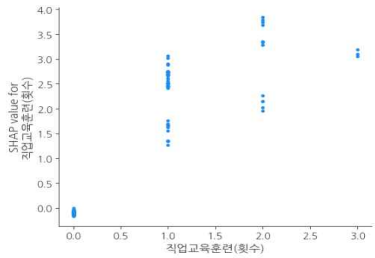


(a) Gradient Boosting 모델 (b) XGBoost 모델

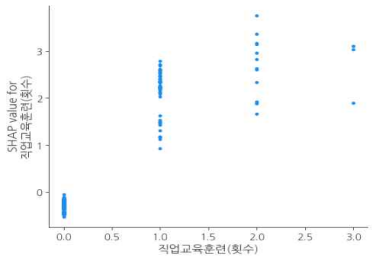


(c) LightGBM 모델 (d) CatBoost 모델

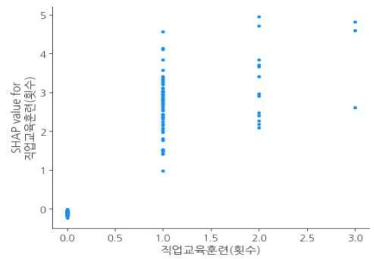
<그림 5> SHAP dependence plot : 전공



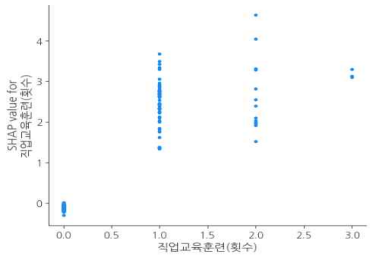
(a) Gradient Boosting 모델



(b) XGBoost 모델

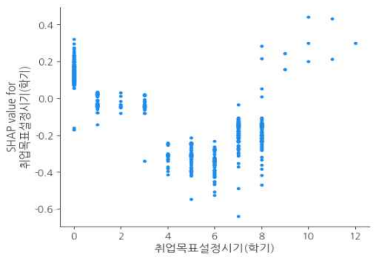


(c) LightGBM 모델

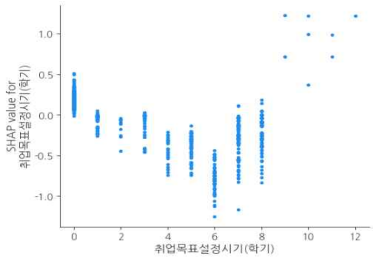


(d) CatBoost 모델

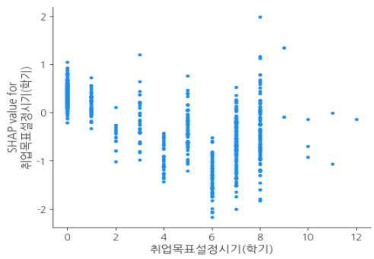
<그림 6> SHAP dependence plot : 직업교육훈련(횟수)



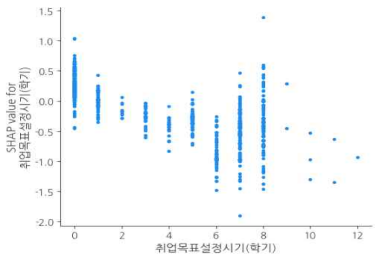
(a) Gradient Boosting 모델



(b) XGBoost 모델

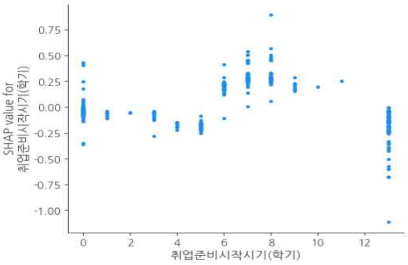


(c) LightGBM 모델

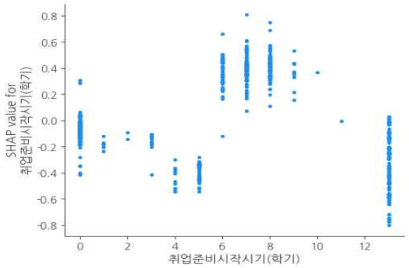


(d) CatBoost 모델

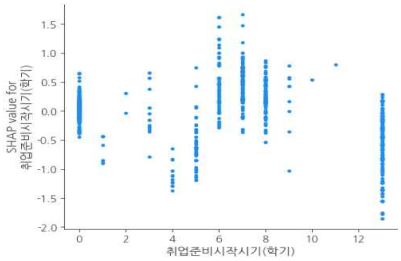
<그림 7> SHAP dependence plot : 취업목표설정시기(학기)



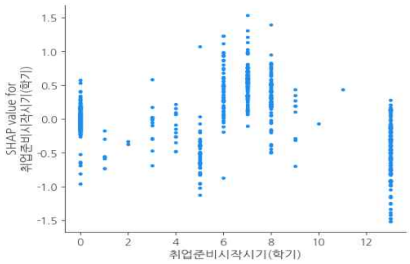
(a) Gradient Boosting 모델



(b) XGBoost 모델

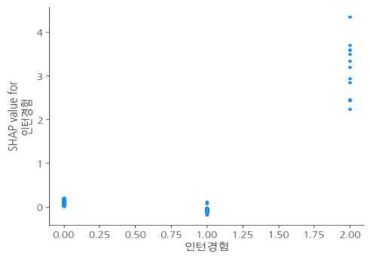


(c) LightGBM 모델

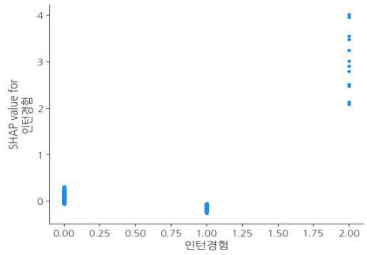


(d) CatBoost 모델

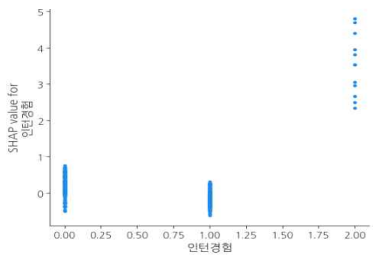
<그림 8> SHAP dependence plot : 취업준비시작시기(학기)



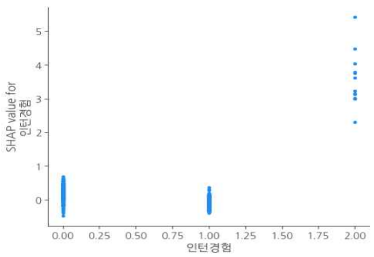
(a) Gradient Boosting 모델



(b) XGBoost 모델



(c) LightGBM 모델



(d) CatBoost 모델

<그림 9> SHAP dependence plot : 인턴경험

육훈련(횟수)', '인턴경험' 변수의 경우, SHAP값이 양수인 곳에서 붉은색 점들이 많은 것을 볼 수 있는데, 이는 해당 변수의 값이 높을 때 취업에 긍정적인 영향을 미치는 정도가 크다는 것을 의미한다. 또한, 해당 SHAP 최대값이 모든 모델에서 4 이상으로, 해당 변수의 값이 높을 때 취업에 긍정적인 영향을 미치는 정도가 매우 크다는 것을 알 수 있다. 그 외의 변수는 취업에 긍정적인 영향과 부정적인 영향을 모두 비슷한 수준으로 미치는 것으로 보인다.

4.2 개별 변수 설명

본 절에서는 각 변수별로 변수값의 범위에 따른 영향을 보다 자세하게 살펴본다. 이를 위해, 파이썬의 SHAP 라이브러리에서 제공하는 dependence plot을 이용하여 시각화한다. <그림 5>부터 <그림 11>은 변수별 SHAP의 dependence plot을 이용한 취업 요인 설명 결과를 보여준다.

<그림 5>의 '전공' 변수는 모든 모델에서 변수값이 커질수록 취업에 긍정적인 요소임을 알 수 있다. '전공' 변수는 값이 높을수록 소프트웨어 전공과 유사함을 의미하기 때문에, 대학에서 소프트웨어를 전공한 것이 취업에 긍정적인 영향을 주는 것을 알 수 있다. 특히, 소프트웨어 전공을 의미하는 변수값 3일 때, SHAP 최대값이 모든 모델에서 4 이상으로 나타나, 취업에 긍정적인 영향을 주는 정도가 매우 큰 요소임을 알 수 있다.

<그림 6>은 '직업교육훈련(횟수)' 변수를 설명하는데, 모든 모델에서 횟수가 1 이상인 경우 취업에 긍정적인 영향을 주고 있다. 이 중, 훈련 횟수가 2 인 경우 가장 높은 SHAP 값을 보이고 있다. 또한, SHAP 최대값이 모든 모델에서 3.5 이상인 것으로 볼 때, 직업교육 훈련을 1회 이상 하는 것이 취업에 꽤 긍정적인 영향을 미친다는 것을 알 수 있다.

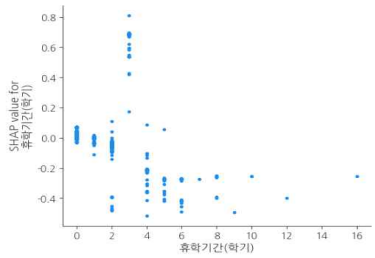
<그림 7> '취업목표설정시기(학기)' 변수는 값의 분포가 높은 8학기까지의 구간을 중심으로 살펴보면, 취업 목표 설정 시기가 늦어질수록 SHAP 값이 0 이하인 경우가 대부분이며, SHAP 최소값이 더 작아지는 경향을 보인다. 이를 통해, 취업 목표 설정 시기가 늦어질수록 취업에 부정적인 영향을 미치는 것을 알 수 있다.

<그림 8> '취업준비시작시기(학기)' 변수는 모든 모델에서 6, 7학기에 준비를 시작한 경우 대부분의 데이터가 SHAP값이 0 이상이고, SHAP의 최대값도 커지는 것을 관찰할 수 있다. 이를 통해, 6, 7학기에 취업 준비를 시작하는 경우, 다른 시기에 비해 취업에 긍정적인 영향을 미치는 정도가 크다는 것을 알 수 있다. 하지만, 졸업 후 준비를 시작한 13값을 가지는 경우, 취업에 부정적인 영향을 미치는 정도가 가장 크다.

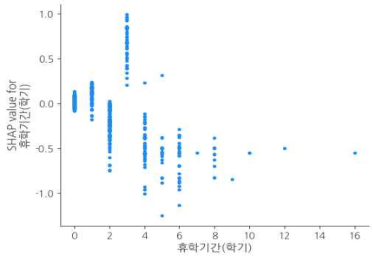
<그림 9> '인턴경험' 변수는 모든 모델에서 '소프트웨어 분야 인턴 경험'을 의미하는 2 값을 가질 때, 모든 모델에서 SHAP 최대값이 4 이상으로, 취업에 미치는 긍정적인 영향이 매우 크다는 것을 알 수 있다. 소프트웨어와 관련이 없는 인턴을 경험한 경우를 의미하는 변수 값 1일 때는 인턴 경험이 전혀 없는 경우의 변수 값 0과 비교할 때 큰 차이가 없으며, 오히려 부정적인 영향을 미치는 정도가 약간 더 높은 것을 볼 수 있다.

<그림 10> '휴학기간(학기)' 변수에 대한 모델별 분석 결과에서는 휴학을 하는 것이 취업에 긍정적인 영향과 부정적인 영향 모두를 주고 있음을 알 수 있다. 휴학을 할 때의 최대 SHAP 값은 1.5 미만으로, 긍정적인 영향을 주는 정도는 크지 않다. 또한, 휴학 기간이 4학기 이상이 될 때는 모든 모델에서 SHAP 값이 0 이하인 경우가 대부분으로, 취업에 부정적인 영향을 미치는 것을 알 수 있다.

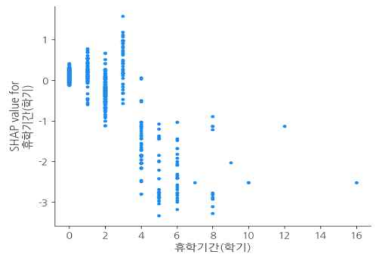
<그림 11> '졸업유예기간(학기)' 변수는 졸업 유예를 할 때 취업에 긍정적인 영향을 주기도 하지만, 부정적인 영향도 함께 주는 것을 볼 수 있다. 취업에 긍



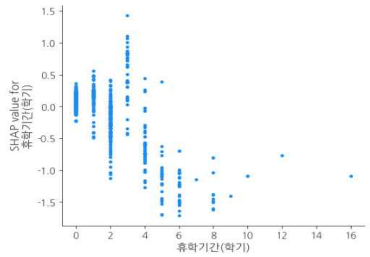
(a) Gradient Boosting 모델



(b) XGBoost 모델

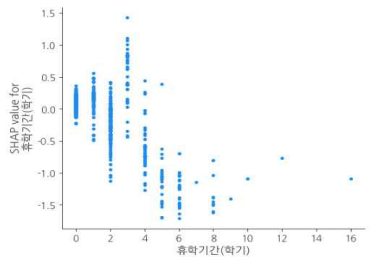


(c) LightGBM 모델

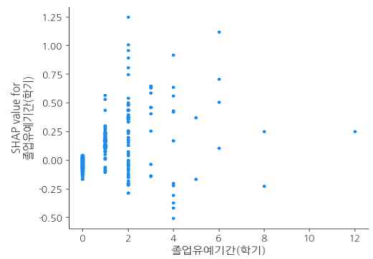


(d) CatBoost 모델

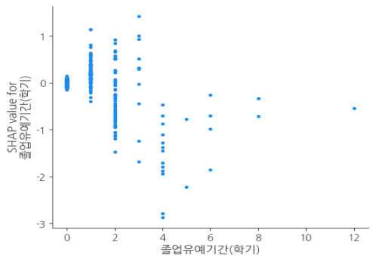
<그림 10> SHAP dependence plot : 휴학기간(학기)



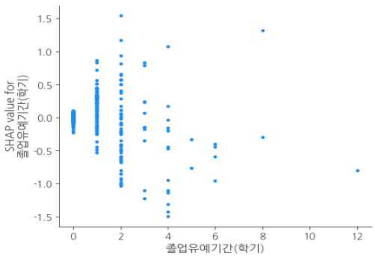
(a) Gradient Boosting 모델



(b) XGBoost 모델



(c) LightGBM 모델



(d) CatBoost 모델

<그림 11> SHAP dependence plot : 졸업유예기간(학기)

정적인 영향을 미치는 정도가 부정적인 영향을 미치는 정도보다 높거나 같은 결과를 보이는 기간을 4개 모델을 기준으로 살펴보면 1학기는 모든 모델, 2학기는 3개 모델, 그 외 학기는 2개 모델 이하에서 관찰된다. 또한, 졸업 유예를 하는 경우 긍정적인 영향을 미치는 최대 SHAP 값은 모든 모델에서 1.5 이하로 긍정적인 영향을 미치는 정도가 크지 않음을 알 수 있다.

Ⅴ. 결론

본 논문에서는 부스팅 기계 학습과 SHAP를 이용한 설명 가능한 소프트웨어 분야 대졸자 취업 모델을 제안하였다. 모델 개발을 위해 사용한 데이터는 GOMS 데이터 중 가장 최근 데이터인 2018년 8월 및 2019년 2월 대학교 졸업자를 대상으로 구축된 데이터이다.

본 논문에서는 설명 가능한 소프트웨어 분야 대졸자 취업 모델을 개발하기 위해 부스팅 기계 학습과 SHAP를 이용하였다. 첫째, 모델 개발을 위해 활용한 데이터는 정형 데이터 형태로, 본 연구에서는 정형 데이터의 예측과 분석 영역에서 뛰어난 성능을 보이는 기계 학습 기법인 부스팅 기계 학습을 사용하였다. 둘째, 개발된 모델에서 취업에 영향을 미친 요인을 설명 가능하도록 SHAP를 사용하고, SHAP 값을 그래프로 시각화함으로써 결과를 해석했다.

본 연구로 도출된 결론은 다음과 같다. 첫째, 부스팅 기계 학습 기반 취업 모델별 성능 평가 결과는 정확도, 정밀도, F1-점수, 균형 정확도는 XGBoost 모델이 가장 높았다. 민감도는 XGBoost 모델, LightGBM 모델, CatBoost 모델이 가장 높았다. 특이도는 GradientBoosting 모델이 가장 높았다.

둘째, 각 모델별로 중요 변수의 순위는 서로 다르게 나타났으나 '전공' 변수는 모든 모델에서 공통적으로 가장 중요한 변수로 나타났고, SHAP 값도 다른 변수에 비교해 매우 높은 수치로 나타나, 소프트웨어 분야 취업을 위해 대학의 '전공'이 미치는 영향이 매우 크다는 것을 알 수 있었다. '전공' 이외의 중요 변수로는 '취업목표설정시기(학기)', '직업교육훈련(횟수)', 취업준비시작시기(학기)' 변수가 다른 변수에 비해 중요한 것으로 분석된다.

셋째, '전공', '인턴경험', '직업교육훈련(횟수)' 변수 값이 클수록 취업에 긍정적인 영향을 미치는 정도가 크다는 것을 알 수 있었다. 즉, 대학에서 소프트웨어 분야를 전공할수록, 소프트웨어 분야 인턴을 경험할수록 취업에 미치는 긍정적인 영향이 매우 크다. 또한, 직업 교육 훈련을 받는 것이 취업에 꽤 유리하게 작용함을 알 수 있다.

네째, '취업목표설정시기(학기)', '취업준비시작시기(학기)', '휴학기간(학기)', '졸업유예기간(학기)' 변수에 대해서는 다음과 같이 설명할 수 있다. 취업 목표 설정 시기는 시기가 늦어질수록 취업에 부정적인 영향을 미치며, 취업 준비 시작 시기가 6, 7 학기인 경우 취업에 다소 긍정적인 영향을 미친다. 휴학은 취업에 긍정적인 영향과 부정적인 영향을 모두 주는데, 휴학 기간이 4학기 이상일 때는 오히려 취업에 부정적인 영향을 미친다. 졸업 유예 기간은 1, 2학기 정도의 단기의 경우에는 취업에 다소 긍정적인 영향을 줄 수 있으나, 큰 영향을 미치지는 못하는 것으로 분석된다.

본 연구는 다음과 같은 시사점을 가진다. 첫째, 실무자의 경우에는 본 연구를 통해 개발된 모델을 활용함으로써 대학교 재학생 혹은 졸업생들과의 상담에 활용할 수 있다. 즉, 취업을 희망하는 피상담자의 현재 상태를 개발된 취업 모델에 입력하여 취업 가능 여부를 예측해보고, 각 취업 항목별로 취업에 필요한 수준을 제시함으로써 취업에 도움을 줄 수 있다. 둘째, 연구자의 경우에는 최신 기술 중 하나인 부스팅 기계 학습과, 설명 가능한 인공지능 기술인 SHAP를

함께 이용하여 모델을 개발하고 영향 요소를 분석하는 연구 사례를 제시한다.

본 연구는 다음과 같은 한계점이 있다. 개발된 모델을 통해 제시된 중요 변수들은 본 논문의 모델을 적용하지 않고도 예측이 어느 정도 가능하다. 이는 본 논문에서 사용된 데이터가 소프트웨어 분야를 중심으로 구축된 데이터가 아닌, 대졸자 직업 이동 경로 파악을 목표로 구축된 데이터를 사용한 것이 가장 직접적인 이유로 볼 수 있다. 본 논문에서 사용한 데이터는 우리나라 전체 대졸자들의 취업 요인을 파악할 수 있는 데이터 중 가장 신뢰성 있는 방대한 양의 데이터 중 하나로 많은 연구들에서 꾸준히 사용된 데이터이다. 하지만, 소프트웨어 분야 취업을 목표로 구축된 데이터는 아니기 때문에, 소프트웨어 분야 취업에서 특수하게 고려할 항목이 충분히 검토되어 있지 않다. 이에 따라, 해당 데이터를 통해 개발된 모델에서 소프트웨어 분야 취업에서만 나타나는 특수 변수들이 발견되지 못했다는 문제점이 있다. 차후 연구에서는 소프트웨어 분야 중심으로 구축된 데이터와 연계 보완을 통해, 기계 학습과 설명 가능 인공지능 기술을 활용에 의해서만 알 수 있는 보다 의미있는 결과가 도출될 수 있는 연구로 확대하는 것이 필요하다.

참고문헌

[1] 민윤지 · 최창열, “4차 산업혁명과 ICT 산업 및 정책 동향분석,” e-비즈니스연구, 제21권, 제2호, 2020, pp.103-118.

[2] 조윤서, “4년제 대학 졸업자의 취업준비행동이 취업 및 괜찮은 일자리에 미치는 영향,” 학습자중심교과교육연구, 제21권, 제1호, 2021, pp.133-161.

[3] 김창식 · 김남규 · 곽기영, “머신러닝 및 딥러닝 연구동향 분석: 토픽모델링을 중심으로,” 디지털산업정보학회 논문지, 제15권, 제2호, 2019, pp.19-28.

[4] 문성은 · 장수범 · 이정혁 · 이종석, “기계학습 및 딥러닝 기술동향,” 한국통신학회지(정보와통신), 제33권, 제10호, 2016, pp.49-56.

[5] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma, “A survey on ensemble learning,” Frontiers of Computer Science Vol.14, 2020, pp.241–258.

[6] 박수연, “앙상블 학습의 부스팅 방법을 이용한 악의적인 내부자 탐지 기법,” 정보보호학회논문지, 제32권, 제2호, 2022, pp.267-277.

[7] Ammar Mohammed and Rania Kora, “A comprehensive review on ensemble deep learning: Opportunities and challenges,” Journal of King Saud University - Computer and Information Sciences, Vol.35, Issue 2, 2023, pp.757-774.

[8] 정재원, 김진율, 오성권, “부스팅 알고리즘을 사용한 폐플라스틱 데이터 패턴 분류기의 비교연구,” 한국지능시스템학회 논문지, 제33권, 제3호, 2023, pp.242~248.

[9] 전민종 · 최혜진 · 박지웅 · 최하영 · 이동희 · 이욱, “Catboost 알고리즘을 통한 교통흐름 예측에 관한 연구,” 한국산학기술학회논문지, 제22권, 제3호, 2021, pp.58-64.

[10] 최재식, “설명가능 인공지능 연구동향,” 정보과학회지, 제37권, 제7호, 2019, pp.8-14.

[11] 정찬일 · 이후진, “프로세스 분석을 위한 설명 가능한 인공지능 기법 비교 연구,” 전자공학회논문지, 제57권 제8호, 2020, pp.51-59.

[12] 김홍비 · 심산신, “악성 사이트 탐지를 위한 설명 가능한 인공지능(XAI) 기반 기계학습 특징 선별에 관한 연구,” 2022년도 한국통신학회 추계종합학술발표회 논문집, 2022, pp.411-412.

[13] 대졸자직업이동경로조사(GOMS), https://survey.keis.or.kr/goms/goms01.jsp
[14] 대졸자 직업이동경로조사(GOMS) 데이터활용 주요 연구리스트, https://survey.keis.or.kr/openresearch/researchlist/Read.jsp?ntt_id=5553
[15] 고용노동부, 한국고용정보원, "한국고용직업분류 2018 해설서 수정판," https://www.keis.or.kr/user/extra/main/3875/publication/publicationList/jsp/LayOutPage.do?categoryIdx=125&publIdx=6132&onlyList=N
[16] 교육부, 한국교육개발원, "2020 학과(전공) 분류 자료집," https://www.kedi.re.kr/khome/main/research/selectPubForm.do
[17] Rakesh Rav, "One-Hot Encoding is making your Tree-Based Ensembles worse, here's why?," Towards Data Science, https://towardsdatascience.com/one-hot-encoding-is-making-your-tree-based-ensembles-worse-heres-why-d64b282b5769
[18] Nur Hanisah Abdul Malek, Wan Fairos Wan Yaacob, Yap Bee Wah, Syerina Azlin Md Nasir, Norshahida Shaadan, and Sapto Wahyu Indratno, "Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data," Indonesian Journal of Electrical Engineering and Computer Science, Vol.29, No.1, 2023, pp. 598~608.

■ 저자소개 ■

권 준 희
(Kwon Joonhee)

2003년 3월~현재 경기대학교 AI컴퓨터공학부 교수
2002년 8월 숙명여자대학교 컴퓨터과학과 (이학박사)
1994년 8월 숙명여자대학교 전산학과 (이학석사)
1992년 2월 숙명여자대학교 전산학과(학사)

관심분야 : 데이터베이스, 빅데이터, 인공지능, 정보검색
E-mail : kwonjh@kyonggi.ac.kr

김 성 림
(Kim Sungrim)

2004년 3월~현재 서일대학교 소프트웨어공학과 교수
2002년 2월 숙명여자대학교 컴퓨터과학과 (이학박사)
1997년 8월 숙명여자대학교 전산학과 (이학석사)
1994년 2월 숙명여자대학교 전산학과(이학사)

관심분야 : 데이터베이스, 빅데이터, 인공지능, 정보검색
E-mail : srkim@seooil.ac.kr

논문접수일 : 2023년 8월 16일
수정접수일 : 2023년 8월 28일
게재확정일 : 2023년 9월 1일