기계 학습 기반의 고려대학교 대학원생의 졸업 시기 예측 및 요인 분석

과제 수행 기간

2021년11월 ~ 2022년 2월

분석 데이터

1999~2019년
융합 전공 졸업생
학적, 수강 이력 데이터

분석 방법

기계학습
설명가능한 인공지능

참여 부서 및 연구자

고려대학교 전기전자공학부

고려대학교 대학원생의 성공적인 유치와 적응을 돕기 위해 기계학습 모델과 설명가능한 인공지능 기법을 사용하여 졸업 시기를 예측해보았다.

통계적으로 졸업 시기 예측에 도움을 줄 수 있는 변수들을 선택하고, 입력 변수를 구성하였다.

구성한 변수들을 사용하여 가장 안정적인 성능을 보이는 기계 학습 모델 두 개를 사용하여 졸업 시기를 예측하였고, 뛰어난 예측 성능을 보였다.

추가로, 설명가능한 인공지능 기법을 사용하여 각 입력 변수가 예측 결과에 미친 영향을 분석하였다.

INTRODUCTION

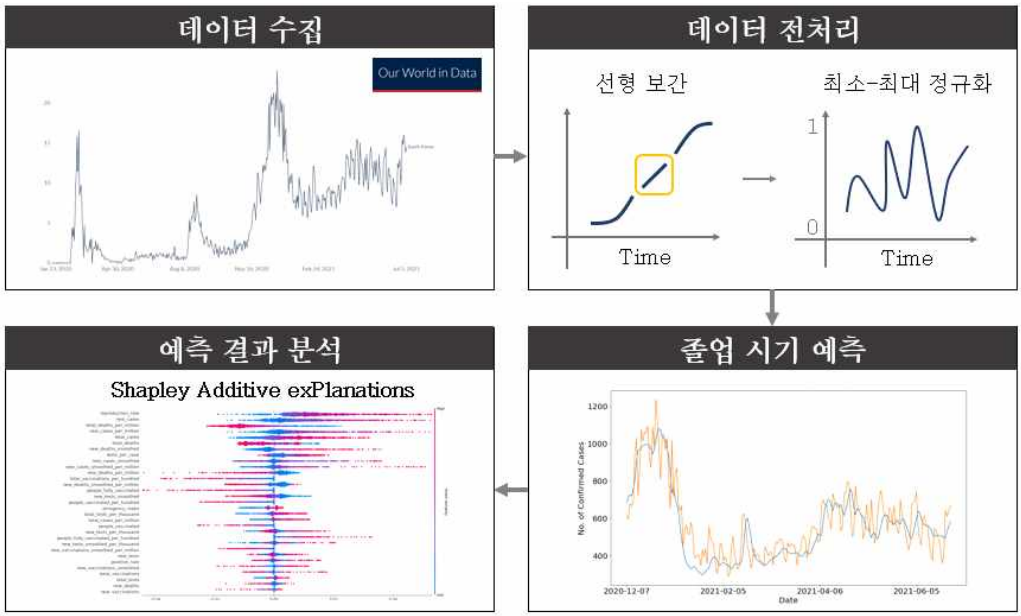
□ 연구 배경

대학원생들에게 졸업 기간은 가장 중요하고 흥미 있는 관심사이다. 이때, 학생들의 데이터를 기반으로 졸업 기간을 분석하고 예측하는 서비스를 제공한다면 대학원생들의 성공적인 유치와 적응에 큰 도움이 될 것으로 기대한다. 따라서, 본 연구팀은 고려대학교 대학원에 재학중인 학생들의 졸업 기간을 예측하고, 졸업에 영향을 미치는 요인들을 분석하였다. 이를 통해 대학원생들에게 졸업에 영향을 미치는 요인을 알려줄 수 있고, 자신의 정보를 기입하여 졸업 시기를 예측해볼 수 있다.

과거에는 각 입력 독립 변수들에 따라 종속 변수의 값을 예측하는 단순 회귀 모형(Simple Linear Regression)이나 회귀 분석(Regression Analysis)과 같은 통계 기반의 예측 기법들을 사용해서 예측 모델을 구성하였으나, 최근 인공지능을 기반으로 하는 기술들의 발전으로 기계학습과 심층학습을 활용한 데이터 예측 기법이 여러 산업 분야에서 탁월한 성능을 보여주고 있다. 또한, 현업에서는 다양한 기계학습 기반 예측 모델을 실제로 이용하고, 그 결과를 신뢰하고 있어, 본 연구에서도 기계학습 모델을 활용하였다. 먼저 고려대학교 대학원에 입학했던 학생들의 데이터 중에서 졸업 기간 예측에 적합한 변수들을 찾고, 이에 적합한 기계학습 모델을 적용하여 가장 우수한 예측 정확도를 보이는 변수 및 모델을 선정한다.

예측 모델의 학습이 끝나면, 기계학습 모델이 입력 변수 중에서 어떤 변수에 가장 영향을 받았는지 분석해보거나, 설명 가능한 인공지능(XAI) 기법을 통해 기계학습 모델의 예측 결과 도출과정을 시각화하여, 각 입력 변수가 예측 결과에 미친 영향을 분석한다. 이 과정을 통해 졸업 기간 예측에 가장 큰 영향을 미친 변수를 찾을 수 있으므로 대학원 재학생이 스스로 어떤 변수에 대응하여 학업 계획을 조정할지 알려줄 수 있다. 또한, 기계 학습 모델의 예측 결과 도출과정을 설명하는 것은 기계학습 모델의 예측 결과를 신뢰하는 데에 도움을 줄 수 있다.

본 프로젝트의 전체 연구 순서도는 아래 그림 1과 같다.



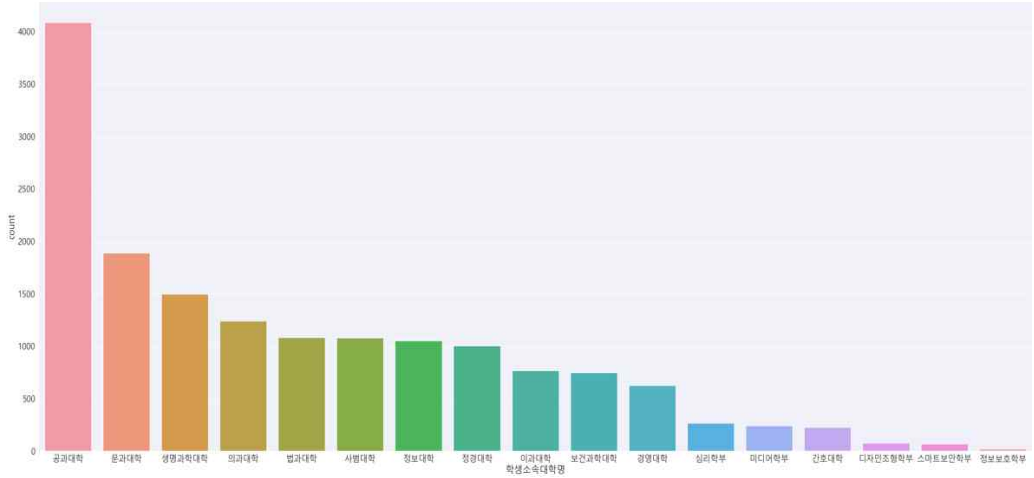
<그림 1> 전체 연구 순서도

□ 데이터 수집 및 전처리

이번 파트에서는 입력 변수 선정, 그리고 예측 모델의 입력으로 사용하기 위한 전처리 과정에 대하여 설명한다.

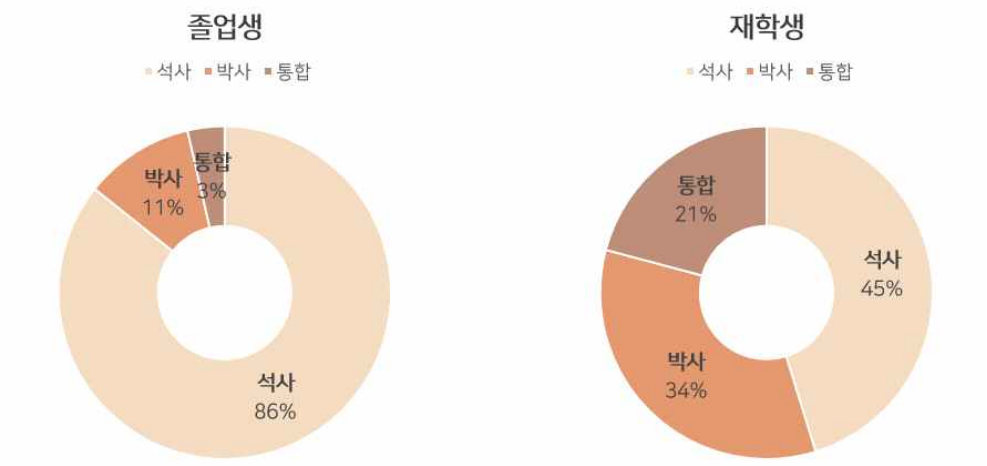
- 변수 설정

본 연구팀은 디지털 정보처에서 제공하는 데이터를 수집한 후, 기계학습 모델의 입력으로 사용하기 위해 선별과정을 거쳤다. 졸업 예측 결과에 도움이 되는 데이터를 선별하기 위해 통계적 분석 및 데이터 시각화를 사용하였다. 제공받은 데이터는 크게 졸업생과 재학생 데이터로 나뉜다. 졸업생의 데이터를 입력으로 기계학습 모델을 학습하고 재학생의 데이터로 졸업을 예측한다. 입력 데이터 분석을 위해, 우선 졸업생과 재학생 전체 데이터를 그림 2에 시각화하여 소속 단과 대학 분포를 봤을 때 공과 대학생이 가장 많은 비율을 차지하는 것을 볼 수 있었다.

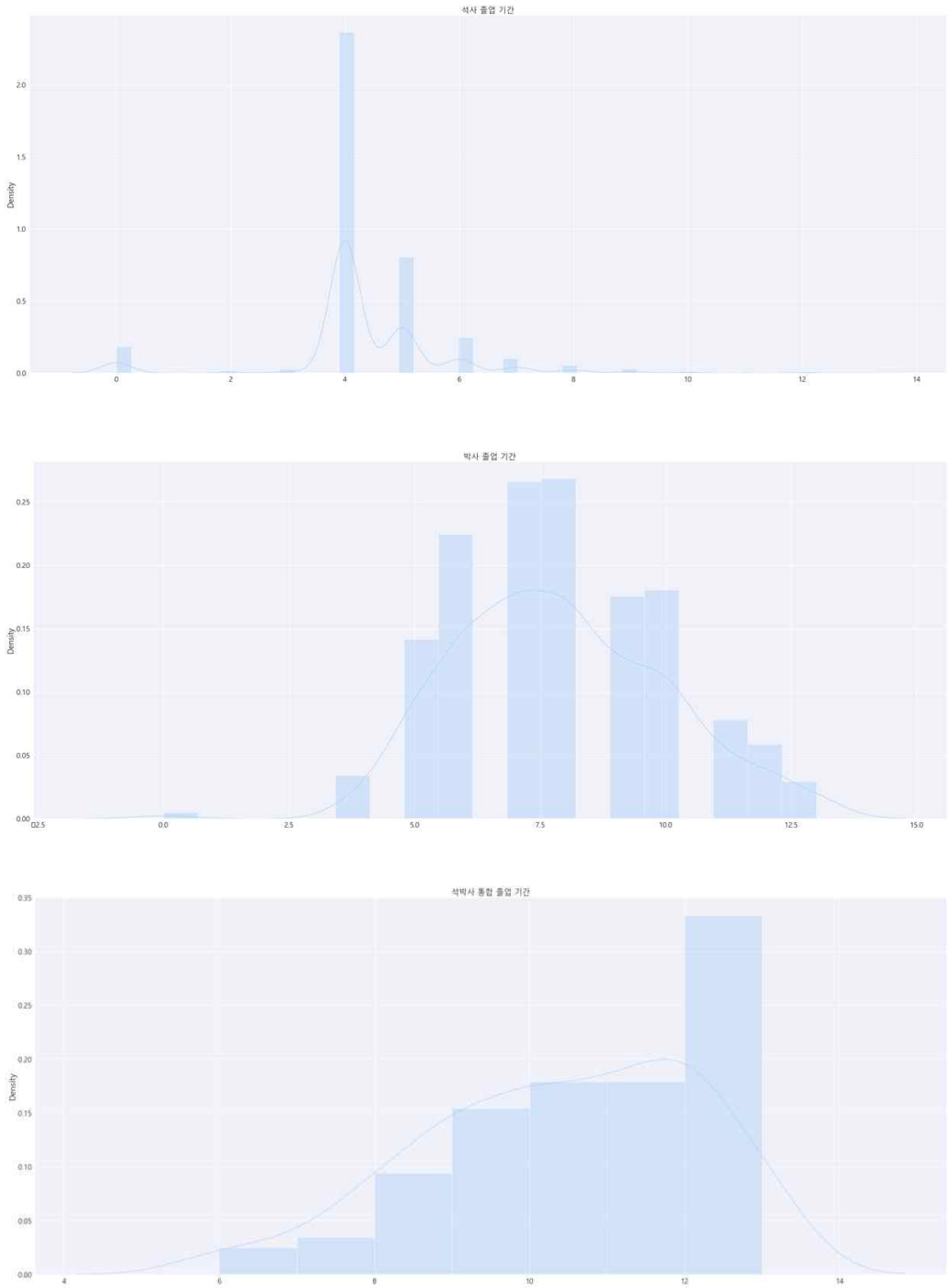


<그림 2> 소속 단과 대학 분포

이에, 졸업 예측을 진행할 때 공과 대학과 공과 대학을 제외한 다른 대학의 졸업 예측을 각각 진행하였다. 또한, 졸업생과 재학생의 각 데이터 분포를 시각화하여 다음의 인사이트를 얻을 수 있었다. 졸업생 데이터의 경우 석사 데이터가 86%를 차지하고, 석박사 통합과 박사 데이터는 각각 3%와 11%를 차지하였다. 이에 반해, 재학생 데이터의 경우 석사, 석박사 통합, 박사 데이터가 각각 45%, 21%, 34%로 어느 정도 균일하게 분포하는 것을 확인할 수 있었다. 이를 시각화한 결과는 아래 그림 3 과 같다.



<그림 3> 졸업생과 재학생에서 재학 과정의 분포 비율



<그림 4> 각 과정 별 졸업 기간 분포

또한, 석사, 석박사 통합, 박사 졸업생들의 졸업 기간을 시각화하여 분포로 그림 4에 나타냈을 때, 석사 과정에서 석박사 통합 과정으로 갈수록 졸업 기간이 길어지고 다양한 분포를 보이며 더 불확실해지는 것을 확인할 수 있었다. 이를 통해, 졸업 기간이 거의 일정한 석사보다 졸업 기간이 다양한 분포를 보이는 석박사 통합과정생의 졸업 기간을 정확하게 예측하는 것이 가장 중요하다는 것을 확인할 수 있었다.

위에서 분석한 것과 같이 훈련 데이터의 분포와 예측할 테스트 데이터의 분포가 매우 다르므로 입력 변수를 적절하게 중요한 변수들로 구성하는 것이 매우 중요하다고 판단하였다. 이에, 입력 변수 선별을 위한 통계적 분석을 시행하여 각 입력 변수와 졸업 기간 변수 간의 상관관계를 Pearson Correlation Coefficient를 사용하여 유사성을 분석하고 이 값이 거의 0에 근접한 매우 작은 입력 변수들은 졸업 기간에 큰 영향을 주지 않는 것으로 판단하여 제외하였다. 이를 통해 생일, 학번, 졸업 년도, 졸업 시기, 전공 교 수, 출신 국가 데이터를 입력 변수에서 제외시켰다.

	수료년도	생일	졸업년도	학점	인건비총액	서류점수	면접점수	school term
수료년도	1.000000	-0.140716	0.289606	-0.070210	-0.241422	-0.037390	0.182328	0.367325
생일	-0.140716	1.000000	0.111374	-0.031430	0.151609	0.124352	-0.047238	0.000296
졸업년도	0.289606	0.111374	1.000000	-0.001664	0.287679	0.071538	0.041821	0.121653
학점	-0.070210	-0.031430	-0.001664	1.000000	-0.000089	0.000739	0.192992	-0.384344
인건비총액	-0.241422	0.151609	0.287679	-0.000089	1.000000	-0.051029	-0.068811	0.057463
서류점수	-0.037390	0.124352	0.071538	0.000739	-0.051029	1.000000	-0.578041	0.004162
면접점수	0.182328	-0.047238	0.041821	0.192992	-0.068811	-0.578041	1.000000	0.121231
school term	0.367325	0.000296	0.121653	-0.384344	0.057463	0.004162	0.121231	1.000000

<그림 5> 입력 변수들과 출력 변수 사이의 피어슨 상관계수

최종적으로 위에서 기술한 입력 변수 선별 과정을 거쳐 학생 소속 단과 대학, 학생 교육부 계열, 성별, 입학 과정, 학과, 학점, 학부 출신, 면접 점수, 서류 점수, 인건비 총액 입력 변수를 선별하였다.

- 데이터 전처리

이렇게 선별한 데이터를 실제로 기계학습 모델의 입력으로 사용하여 올바른 예측 결과를 얻기 위해서는 몇 가지 전처리 과정이 필요하다. 먼저 데이터가 누락된 결측치의 경우, LightGBM 모델에서 missing = False로 설정하여 training loss에 기반한 결측치를 알아서 처리하도록 하였다. 또한, 기계학습 모델의 학습을 돕기 위해 각 입력 변수들의 스케일 범위를 맞추는 최대-최소 정규화를 진행하였다.

1차 회의 이후 전공과 같은 입력 변수의 경우 너무 많은 범주를 가지고 있기 때문에 모델에서 원-핫 인코딩을 하게 되면 각 범주마다 샘플의 수가 너무 적어져서, 중요도나 의미가 많이 떨어지고, 차원이 너무 커져서 예측에 방해가 된다고 판단하였다. 이에, 전공 입력 변수의 경우 더 큰 범주인 학생 소속 대학과 학생 교육부 계열 변수로 묶어서 재 범주화 처리해주었다. 이를 통해, 실제로 더 우수한 예측 결과를 보이는 것을 확인할 수 있었다.

마지막으로 예측할 졸업 기간의 경우 원래 제공된 데이터에서는 수료 기간만 나와 있었기 때문에 졸업 년도와 졸업 학기 수에서 수료 년도와 수료 학기를 뺀 값을 더해서 최종 졸업 기간을 추출하여 예측할 라벨로써 사용하였다. 또한, 학생의 인건비 총액은, 학생 한 명이 대학원 재학기간 동안 받은 인건비를 의미하는데, 총액 값 자체는 다닌 학기 수가 늘어날수록 함께 증가할 수밖에 없다. 따라서, 다닌 학기 수를 예측하기 위해서는 인건비 총액보다는, 학생이 받은 월 평균 인건비를 입력 변수로 사용하는 것이 합리적이다. 이를 위해서, 인건비 총액을 재학 개월 수로 나누어 월 평균 인건비를 산출하고 이를 입력 변수로 활용했다.

□ 졸업시기 예측 및 결과 분석

- 예측 모델 구성 및 설명

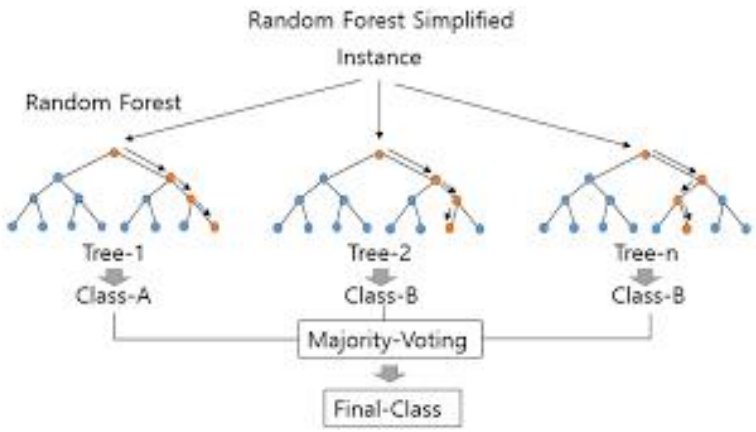
데이터 전처리가 끝나면 외부 데이터와 과거 졸업 관련 데이터를 입력 변수로, 예측 대상 변수인 졸업 기간 데이터를 출력 변수로 나눈다. 따라서, 기계학습 모델은 주어진 입력 변수를 통해 출력 변수인 졸업 기간 데이터를 출력하도록 학습한다. 이때 본 연구에서는 대표적인 기계학습 모델로는 랜덤 포레스트(Random Forest), 의사 결정 나무(Decision Tree), 지지 벡터 회귀모형(Support Vector Regressor) 등이 있다. 이때 예측 모델의 성능을 평가하기위해 데이터를 학습용과 평가용으로 분리한다. 예를들어, 2015년부터 2019년까지의 데이터를 학습용으로, 2020년 데이터를 평가용으로 사용하는 것이다. 데이터의 분리가 끝나면 학습용 데이터를 사용해 다양한 기계학습 모델을 학습시키고, 평가용 데이터를 사용하여 가장 좋은 성능을 나타내는 기계학습 모델을 선정한다. 이때 평가하는 식은 평균적인 오차를 나타내는 MAE(Mean Absolute Error)와 MSE(Mean Squared Error)를 사용한다.

석사, 박사, 석박통합과정의 졸업생 데이터를 이용해 각 재학 과정에 대한 기계학습 모델의 훈련 및 성능 평가를 진행하고, 학습한 모델을 이용해서 재학생들의 졸업 기간

을 예측하는 방식으로 연구를 진행했다. 또한, 특정 단과대학의 학생들의 졸업 기간이 전체 학생들의 졸업 기간과 다른 경향을 보이는지 확인하기 위해 공과대학 학생들만 따로 추출하여 위와 동일한 연구 과정을 진행했다. 예측에 사용한 기계학습 모델은 부스팅 계열의 알고리즘인 LightGBM을 사용하였다.

1. 랜덤포레스트

구성된 입력변수를 활용하여 다음 날의 특정 지역의 감염병 발생 수를 예측하도록 랜덤 포레스트 모델을 학습시킨다. 랜덤포레스트는 의사결정나무(Decision Tree)의 확장 형태로서, 데이터 집합에 대하여 다양한 모델을 종합하여 결과를 내는 방식인 앙상블 학습법을 사용한다. 주어진 전체 데이터의 부분집합으로 학습된 의사결정나무를 다수 구성한 후, 각 의사결정나무에서 얻어진 결과의 가중 평균이나 최빈값 등을 최종 예측값으로 결정한다. 이러한 방식은 입력 변수의 노이즈에 취약한 의사결정나무의 단점을 보완하여 예측 오차를 줄인다. 더하여, 학습된 랜덤포레스트 모델은 각 입력변수의 변수 중요도를 구할 수 있다는 장점이 있다. 랜덤포레스트의 하이퍼 파라미터로는, 의사결정 나무의 개수, 의사 결정 나무에서 사용할 특징 등이 있다.



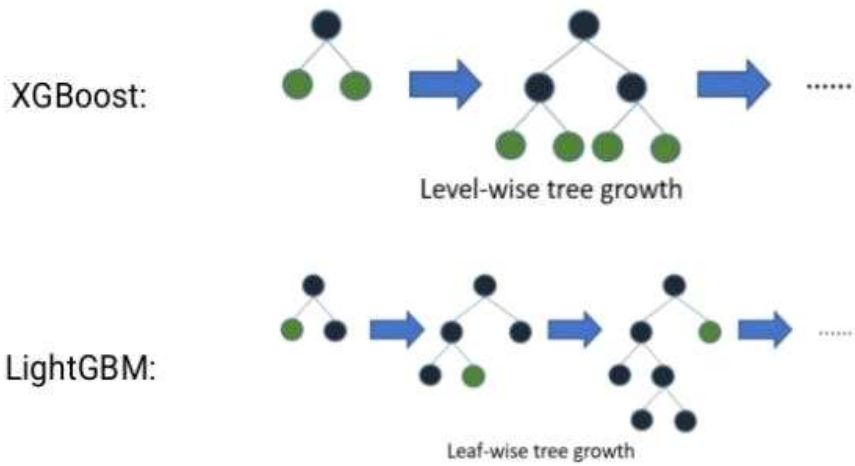
<그림 6> 랜덤 포레스트 개념도

2. LightGBM

LigthGBM은 부스팅 알고리즘을 사용하는 앙상블 모델로, 높은 정확도를 보여줌과 동시에 빠른 학습 속도를 보유하고 있어서 최근 많이 사용되고 있다. 부스팅 알고리즘은 비교적 단순하고 약한 모델들을 여러 개 결합하여 더욱 정확하고 강력한 모델을 만드

는 방법이다. 여러 분류기를 차례대로 학습하며 이전 모델에서 잘못 예측한 데이터에 대해서 가중치를 부여하여 오류를 개선하고 최선의 결과를 도출한다. 부스팅 계열 알고리즘인 XGBoost가 Level-wise tree growth 방식을 사용하는 것과는 달리, LightGBM은 Leaf-wise tree growth 방식을 사용한다.

LightGBM은 기존의 부스팅 알고리즘이 사용된 다른 모델들의 단점인 긴 데이터 처리 시간을 GOSS (Gradient-based One-side Sampling) 기술과 EFB (Exclusive Feature Bundling) 기술을 사용하여 보완하였다. GOSS 기술을 사용하여 데이터 중 기울기가 큰 부분 만을 사용하여 정보를 얻고, EFB 기술을 사용하여 상호배타적 변수들을 묶어서 처리하기 때문에 XGBoost와 같은 기계학습 모델과 비교하였을 때 월등하게 빠른 속도를 보여준다.



<그림 7> LightGBM 개념도

- 실험 결과

실험 결과 도출에 앞서, 우리 LightGBM 모델에 적용할 최적의 Hyperparameter를 탐색하였다. 딥러닝 모델에서는 반복적인 학습을 통해서 값이 결정되는 일반적인 Parameter들과는 달리, Hyperparameter는 모델 구성 단계에서 사용자가 직접 결정해주는 변수이다. 모델의 과적합을 방지하거나 연산 시간이 과도하게 증가하는 것을 방지하기 위하여 최적 Hyperparameter 값을 찾는 것이 중요하다. LightGBM에는 learning_rate, num_iterations, max_depth, num_leaves, boosting 등의 다양한

Hyperparameter들이 있다. 우리는 이중에서 num_iterations, max_depth, num_leaves의 최적의 변수를 찾기 위해서 Grid Search를 활용하였다.

Grid Search는 모델의 Hyperparameter로 넣을 수 있는 값들을 순차적으로 입력한 뒤 학습해보고, 그중에서 가장 높은 성능을 보이는 최적의 Hyperparameter를 찾는 탐색 방법이다. Grid Search 통해서 선정된 최적인 Hyperparameter는 다음 표와 같다.

Grid Search	**Hyperparameter 후보군**			
num_iterations	30	50	**100**	200
max_depth	3	**5**	7	9
num_leaves	10	20	**30**	40

<표 1> Grid Search를 통하여 선정된 최적의 Hyperparameter

예측 성능은 MAE(Mean Absolute Error)와 MSE(Mean Squared Error)의 두 가지 평가지표를 이용해 평가했다. MAE는 평균절대오차로써, 예측값과 실제값의 차이의 절대값을 평균 내기 때문에, 오차에 대한 직관적인 해석이 가능하며, MSE는 평균제곱오차로써, 예측값과 실제값의 차이의 제곱을 평균 내기 때문에, 큰 오차에 대해 큰 페널티를 부여한다.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

전체학생	석사	박사	석박통합과정
MAE	0.557	1.37	1.52
MSE	0.802	3.21	3.66

<표 2> 전체 학생에 대한 예측 평가 지표

공과대학 학생	석사	박사	석박통합과정
MAE	0.436	1.21	1.47
MSE	0.734	3.02	3.38

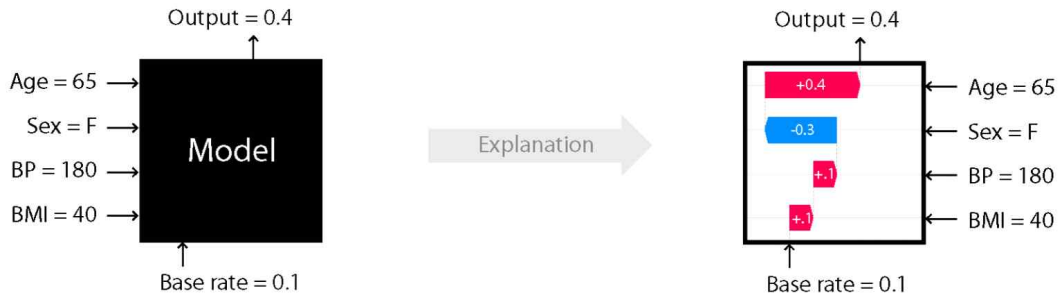
<표 3> 공과대학 학생에 대한 예측 평가 지표

훈련한 기계학습 모델을 통해 평가한 성능을 전체 학생들과 공과대학 학생으로 나누어 나타낸 표이다. 학기 단위로 예측하기 때문에, 오차의 단위는 학기 수 이다. 우선, 석사가 다른 두 재학 과정에 비해 압도적으로 적은 오차를 보여주었다. 이는 석사가 다른 두 과정에 비해 실제로 더 적은 재학 기간을 가지며, 대부분의 경우 4학기에 졸업하게 되기 때문으로 보인다. 또한, 가장 긴 재학 기간을 갖는 석박통합과정이 가장 높은 오차를 보인다. 공과대학 학생들에 대해서 적용했을 때는, 모든 재학 과정에 대해 더 적은 오차를 보인다. 전체적인 결과로 봤을 때, 제한된 입력 변수만을 사용하였음에도 매우 우수한 예측 성능을 보였다.

- 실험 결과 해석

우수한 예측 성능에 더하여, 예측 결과의 도출 과정을 분석하여 의사 결정자들에게 예측 성능에 대한 신뢰를 주기 위해 설명 가능한 인공지능 (XAI, eXplainable AI) 기법을 사용하게 되었다. 아래의 그림에서 왼쪽 그림과 같이 기존에는 블랙박스와 같이 모델 예측 결과에 대해 얼마나 정확한 성능을 보이는지에만 초점이 맞춰져 있었다. 하지만 실제 산업에서는 모델링을 하면서 결과에 대한 원인 인자를 찾고, 얼마나 결과에 영향을 주었는지를 파악해야 할 때가 많다.

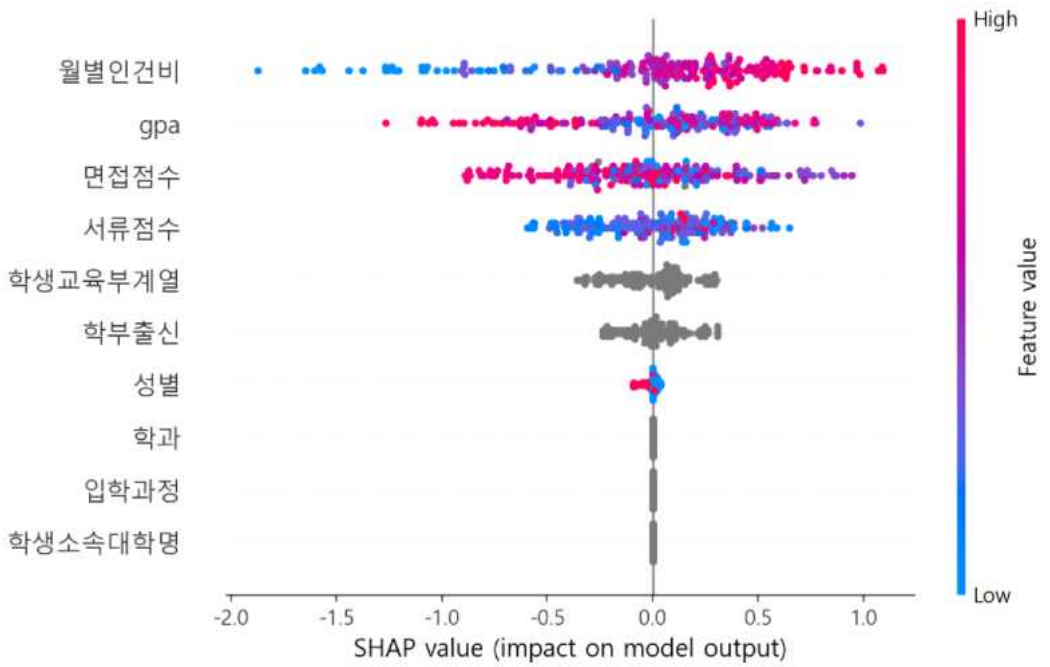
예를 들어, 이번에 진행한 데이터 분석 연구 과제 프로젝트의 경우에도 재학생들의 졸업 기간을 예측함에 있어서 단순히 예측 성능을 확인하고 그치는 것이 아니라 어떤 입력변수들이 졸업 기간에 얼마만큼 긍정적 혹은 부정적 영향을 끼치는지를 분석할 필요가 있다. 이를 통해, 어떤 입력변수가 졸업에 중요한지에 대한 정보를 바탕으로 학교 차원에서는 학생들에게 졸업을 빠르게 대비할 수 있도록 정보를 알려주고 학생들의 졸업을 효율적으로 점검하고 관리할 수 있다. 학생들의 경우에는 본인이 부족한 부분을 명확하게 분석하여 졸업을 제때 할 수 있도록 대비할 수 있다.



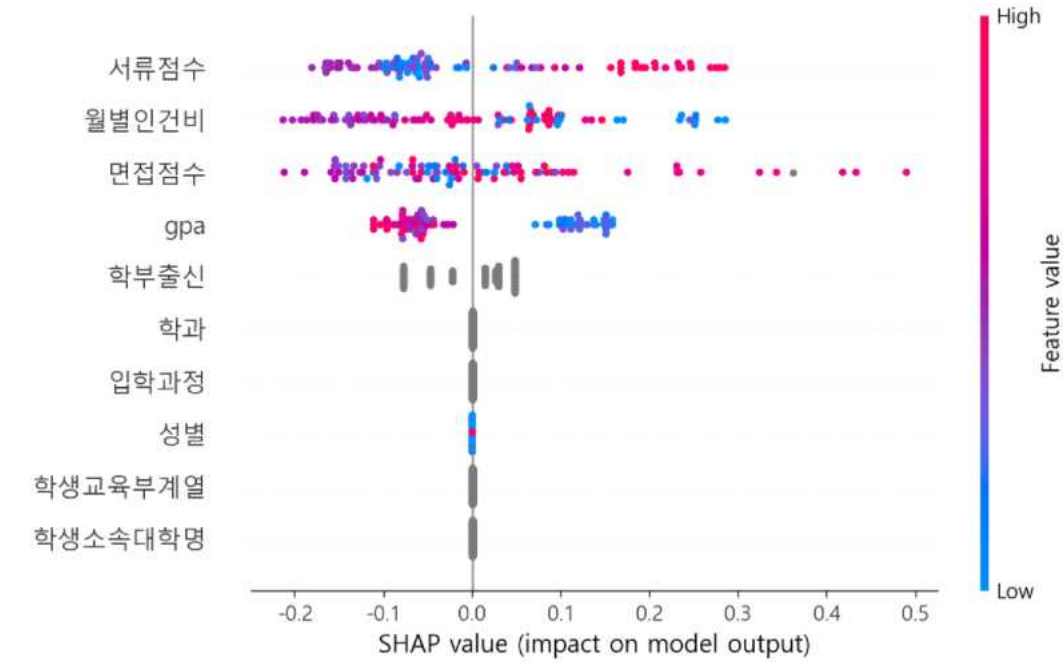
<그림 8> SHAP의 개념

이에, 수치형 입력변수의 기여도를 계산하여 모델의 예측 결괏값을 설명하기 위해서 본 연구팀은 SHAP(SHapley Additive exPlanations)을 채택하여 활용하였다. SHAP은 게임이론에 기반하여 각 수치형 입력변수에 대한 SHAP value를 계산한 뒤 입력변수와 모델의 결괏값 사이의 관계를 탐색하는 설명 가능한 인공지능 기법이다. SHAP을 사용하여 모델을 분석하면 위의 그림에서 오른쪽 부분과 같이 각 입력변수의 영향력을 알 수 있고, 각 입력변수의 값이 변할 때, 결과가 어떻게 변하는지도 해석할 수 있다. SHAP은 Local Explanation을 기반으로 하며, 데이터의 전체적인 영역에 대한 해석이 가능하다.

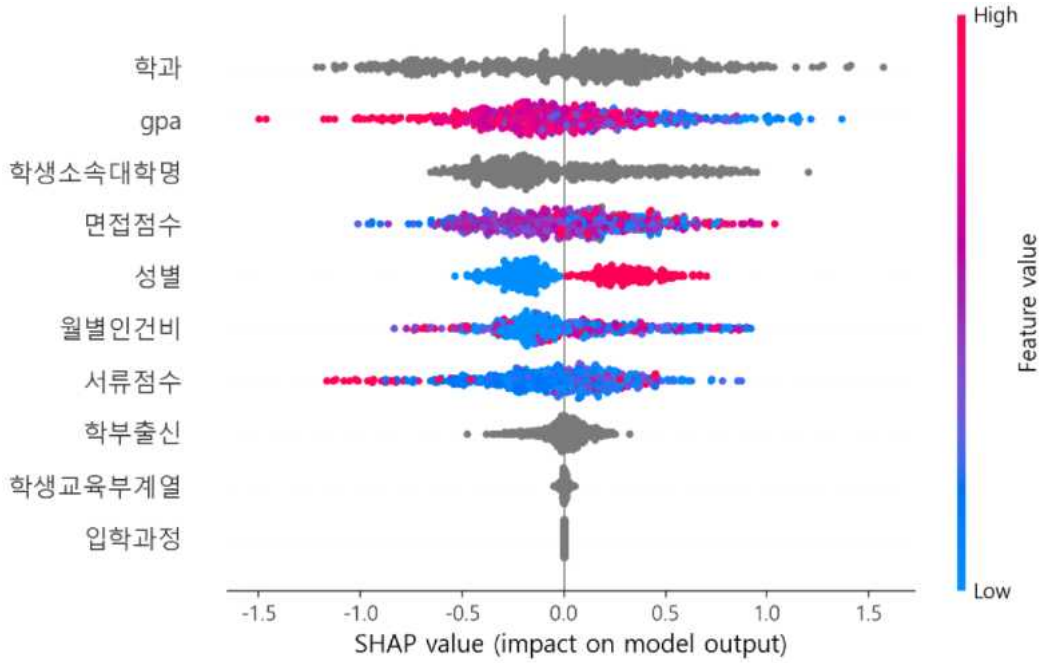
아래는 각 과정생 분류에 대한 SHAP 분석 결과를 summary plot으로 나타낸 그림으로 각 입력변수가 shapley value 분포에 어떤 영향을 미치는지 시각화한 것이다.



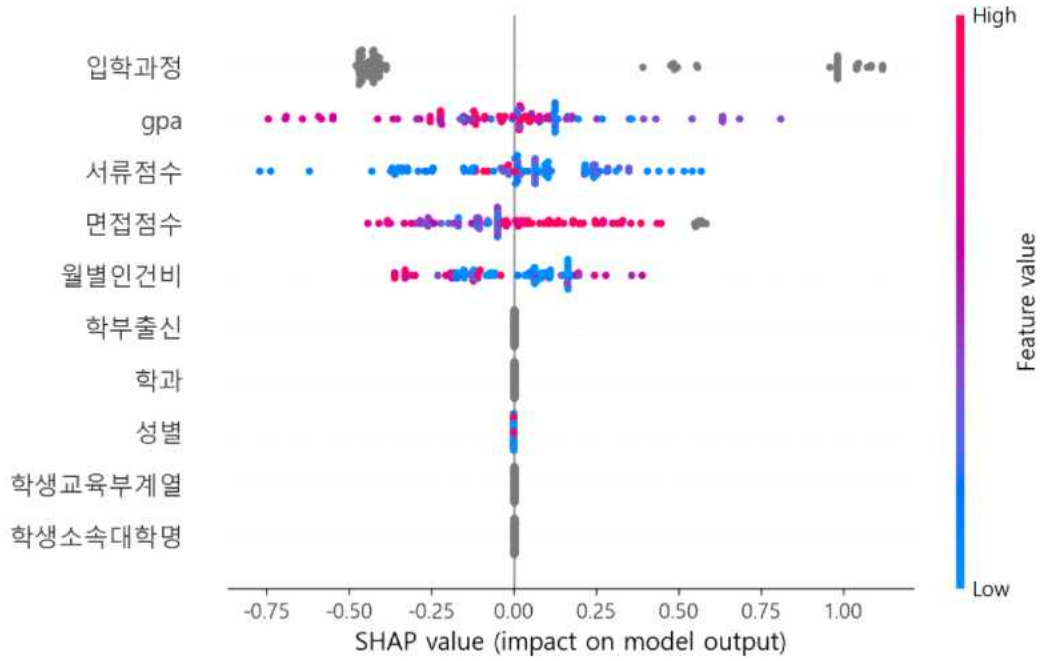
<그림 9> 석박통합과정 – 전체 학생



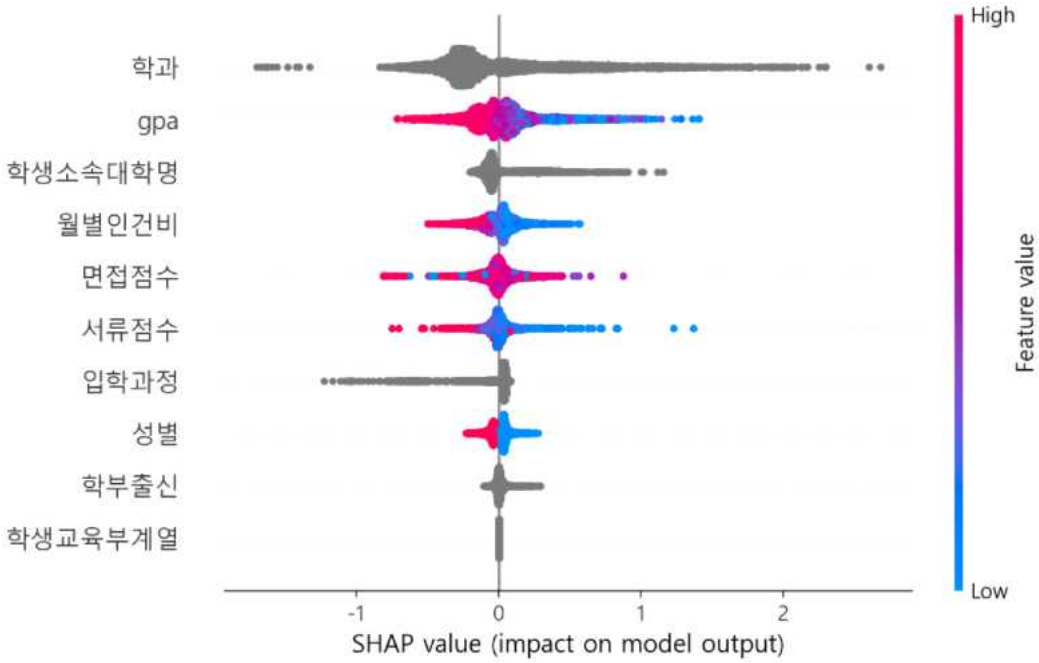
<그림 10> 석박통합과정 – 공대 학생



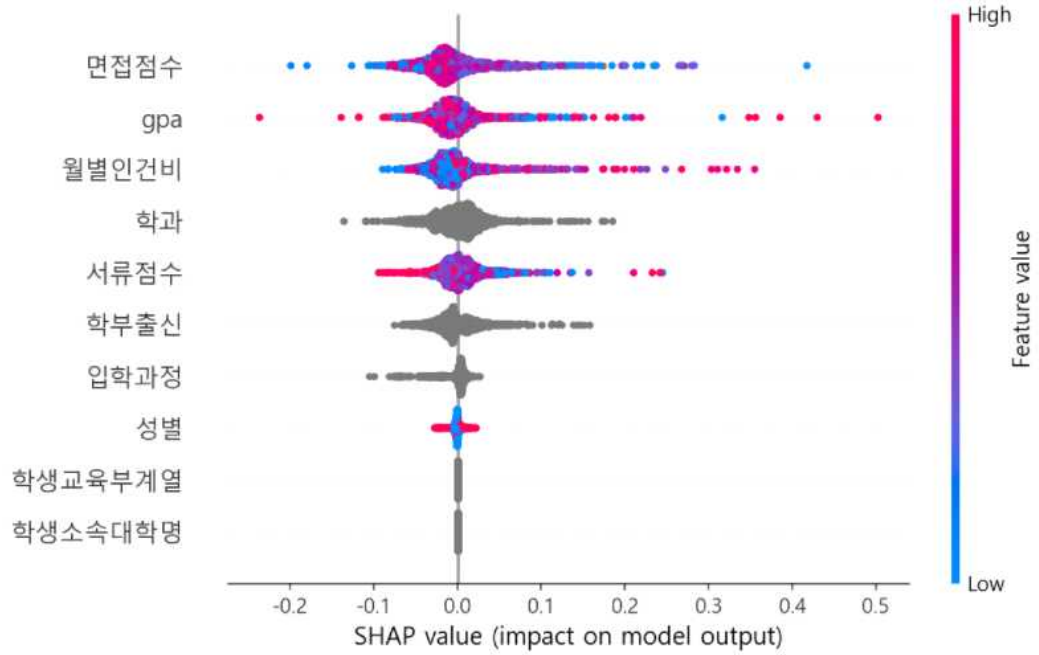
<그림 11> 박사과정 – 전체 학생



<그림 12> 박사과정 – 공대 학생



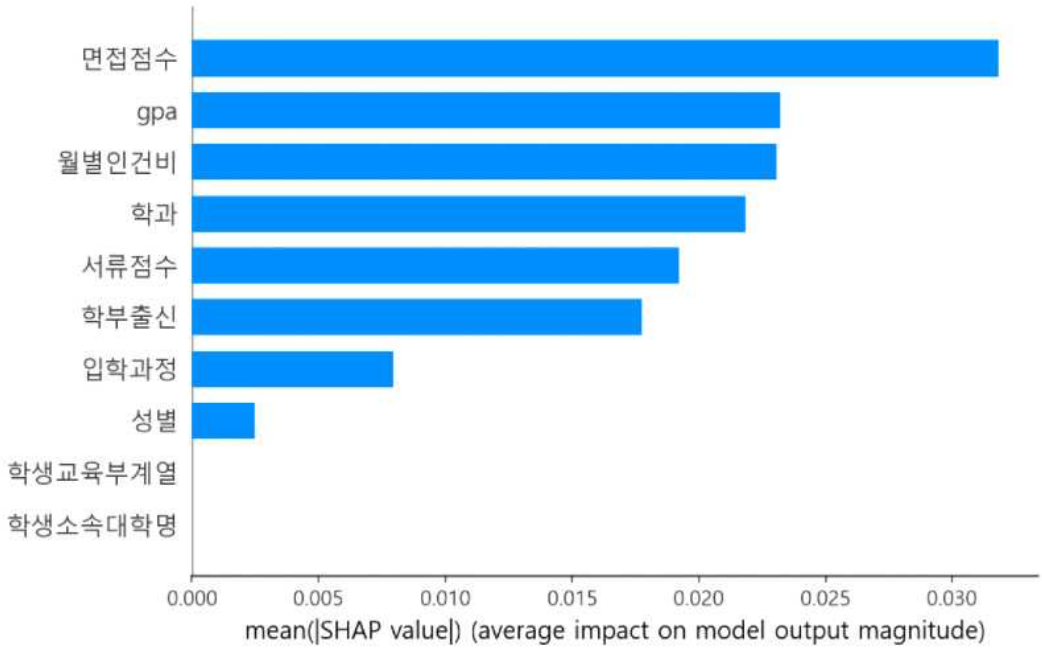
<그림 13> 석사 과정 – 전체 학생



<그림 14> 석사 과정 – 공대 학생

summary plot을 해석하는 방식은 다음과 같다. x축은 shapley value를 y축은 각 특성을 나타낸다. 각 표에서 파랑이나 빨강의 분포로 표현된 변수가 수치형 변수이고 회색으로 표현된 변수는 범주형 변수이다. 색깔은 특성값을 나타내어, 파란색은 해당 입력변수의 값이 낮은 경우, 빨간색은 해당 입력변수의 값이 큰 경우이다. summary plot 하단의 SHAP value 값은 해당 입력변수의 값이 모델의 결괏값을 높이는데에 영향을 주는지, 또는 낮추는데에 영향을 주는지 보여주는 지표이다. 그래프상에서 입력변수들은 예측에 미치는 영향력 즉, 중요도에 따라 위에서부터 정렬된다. 바로 위의 석사과정 공대생들의 졸업 예측의 SHAP 결과를 해석해보면 면접점수가 졸업 기간 결괏값 예측에 가장 큰 영향을 미친다고 해석할 수 있다.

이는 아래 그림과 같이 bar plot으로 시각화하여 석사과정 공대생들의 입력변수 중요도만을 확인할 수도 있다. 맨 밑의 학생 교육부 계열이나 학생 소속 대학명 입력변수의 경우 졸업 기간 예측에 거의 영향을 미치지 못한다고 해석할 수 있다.



<그림 15> 입력 변수 중요도

앞의 shap summary plot들을 분석해보면, 모든 과정의 경우에서 gpa가 높을수록(빨간색) 졸업 기간이 줄어든다(shap value가 작음)는 것을 확인할 수 있다. 이는 성적이 좋은 학생들이 성적이 낮은 학생들보다 평균적으로 빨리 졸업할 것이라는 예상과 일치한다. 또한, 인건비 총액이 높을수록 졸업 기간을 높이는 것으로 확인하였는데, 이는 졸업 기간이 길어지면 자연스럽게 인건비도 비례해서 증가하기 때문으로 보인다.

입력변수 중 성별은 전체 학생, 공대 학생에 따른 분류와 과정별로 차이를 보였다. SHAP 결과에서 파란색이 남자, 빨간색이 여자이다. 전체 석·박통합과정 학생에서 성별은 졸업 기간에 영향을 주지 않는 것으로 해석되지만, 전체 석사과정 학생들에 대해서는 여자가 남자보다 졸업 기간이 짧다는 것을 확인할 수 있다. 그러나 전체 박사과정 학생에 대해서는 여자가 남자보다 졸업 기간이 긴 것을 확인할 수 있다. 이는 박사과정에 입학한 여자의 경우, 임신이나 출산 혹은 육아 등으로 인한 휴학 기간이 유효하게 작용하였을 것으로 보인다.

공대 석·박통합과정 및 박사과정에서 성별은 졸업 기간에 거의 영향을 주지 않는 것으로 보이지만, 공대 석사과정에서는 여자가 남자보다 짧은 기간 안에 졸업하는 것을 확인할 수 있다. 이는 일반적으로 공대의 성비가 남자가 더 많고, 여자 공대 학생의 경우 석사과정까지만 하는 경우가 더 많기 때문으로 예상된다. 이외 면접점수, 서류점수 변수는 SHAP 결과에서 고르게 분포되어있는 것으로 보아, 졸업 기간에 크게 영향을 주지 않는 변수로 판단된다.