kaggle 데이터를 활용한 대학생 학업 중단 예측 및 요인 분석

학 과	: 컴퓨터정보공학과(심화)
학 번	: 202547003
이 름	: 고태경
제 출 일 자	: 2025.12.04

인하공업전문대학
INHA TECHNICAL COLLEGE
仁荷工業專門大學

목 차

CHAPTER1 프로젝트 개요 및 연구 배경

1.1 연구 요약 (Summary)

본 연구는 UCI의 Predict Students' Dropout and Academic Success 데이터를 활용하여 대학생의 학업 중단(Dropout)을 조기에 예측하고 주요 영향 요인을 규명하는 것을 목표로 하였다. 데이터는 학업 성과(신청·이수·인정 학점, 성적, 성적 변화량), 개인 특성(성별, 나이, 혼인 상태), 가정·경제적 요인(채무, 장학금 수혜, 등록금 납부 여부), 외부 환경 변수(실업률, 물가상승률 등)로 구성된다.

학생 개인 정보 EDA에서 성별에 따른 분석 결과, 남학생(56%)이 여학생(30%)보다 중단율이 높았다. 입학 시 나이가 많을수록 중단 위험이 증가하며, 특히 20대 중반에서 뚜렷하게 나타났다. 혼인 상태는 미혼 외 집단에서 중단율이 높았으나 표본이 적어 '미혼 여부' 이진 변수로 단순화하였다. 지원 순위는 비선형 패턴을 보였으며, 0지망은 중단율이 거의 0%로 낮고 1순위에서 급격히 상승(35%)하였다. 전공별로는 간호학과가 낮은 중단율(18%)을 보인 반면, 소규모 전공(바이오연료, 말산업학 등)에서 높게 나타났다. 주간/야간, 유학생, 이주 여부, 특수교육 필요 여부는 중단과 큰 관련이 없었다.

가정·경제적 EDA에서는 재정 변수가 가장 강한 영향을 보였다. 등록금 미납 학생의 중단율은 94%로 압도적이었고, 채무가 있는 경우 76%, 장학금 미수혜 시 48%로 나타났다. 반면 장학금 수혜 학생은 14%로 현저히 낮았다. 부모 학력은 무학/문해 그룹에서 70%대, 정보 없음 그룹에서 77%로 높았으며, 중등 이상부터는 35~50%로 안정화되었다. 부모 직업은 무응답 그룹에서 가장 높았고(70% 이상), 전문직·사무직 등 안정적 직종에서 35% 수준으로 낮았다. 정보 미기재 자체가 위험 신호임을 확인하였다.

학업 관련 EDA에서는 입학 성적과 이전 학력 성적은 전체 분포에서 유지·중단 간 차이가 거의 없었으나, 입학 성적 110점 이하 극단적 저성적대에서는 중단율이 높았다(66%). 1·2학기 인정 학점, 평균 성적은 중단 여부와 강한 관련성을 보였다. 특히 성적 변화의 '방향'이 중요한 신호로 확인되었는데, 인정학점 감소 시 중단율 44%, 평균성적 '유지' 그룹에서 52%로 가장 높았다. 이는 낮은 성적을 그대로 유지하는 정체 상태가 위험 신호임을 시사한다.

사회/경제적 요인 EDA에서 거시경제 지표(실업률, 물가상승률, GDP)는 학업 중단과 뚜렷한 관계를 보이지 않았다. 경제 상황의 좋고 나쁨에 따라 중단율이 크게 달라지지 않았으며, 개인 수준 변수(학업 성과, 재정 상태)가 더 직접적인 영향을 미치는 것으로 판단되어 최종 모델에서 제외하였다.

본 연구에서는 EDA 결과를 바탕으로 성적 변화량, 학점 변화 방향 등 다양한 파생변수를 생성하고, 비효율적 변수를 제거하는 Feature Engineering을 수행하였다. 이후 Logistic Regression, Random Forest, XGBoost, LightGBM 등 다양한 머신러닝 모델을 비교하였으며, 교차검증을 통해 모델 안정성을 확보하였다. 모델 비교 결과 Logistic Regression이 가장 높은 F1 Score와 해석 가능성을 보여 최종 모델로 선정되었으며, PCA 및 성적 변수 제거 실험을 통해 학업 성과 변수가 예측에서 핵심적 역할을 한다는 점이 재확인되었다. 본 연구는 이를 바탕으로 학업 중단 위험 학생을 조기에 식별하고, 맞춤형 학업 및 재정 지원 정책 수립에 활용할 수 있는 정량적 근거를 제시하였다.

1.2 프로젝트 개요

본 프로젝트는 학생의 학업 중단(Dropout)을 예측하여 대학 차원의 학사 관리 및 학생 지원 정책에 활용할 수 있는 모델을 구축하는 것을 목표로 한다.

분석에는 UCI Machine Learning Repository의 "Predict Students' Dropout and Academic Success"[1] 데이터를 사용하였다. 데이터에는 학업 성과(성적, 신청·이수 학점, 평가·인정 학점), 개인 특성(나이, 성별, 혼인 상태 등), 경제적 요인(장학금, 채무, 등록금 납부 여부), 외부 환경 변수(실업률, 물가상승률 등)가 포함되어 있다.

1.3 분석 기획 및 문제 정의

학업 중단은 단순히 개인의 성취 저하뿐 아니라 학교 운영, 재정, 교육 품질에도 직접적인 영향을 미치는 중요한 지표이다. 따라서 본 프로젝트는 다음 핵심 질문을 중심으로 분석을 설계하였다.

1. 어떤 요인이 학업 중단과 가장 밀접하게 관련되어 있는가?
 - 학업 성취도(평균 성적, 인정 학점, 평가 학점 등)
 - 학기 간 성적 변화량
 - 경제적 변수(장학금 여부, 등록금 납부 여부, 채무 여부)
 - 개인·학적 특성(성별, 혼인 상태, 입학 자격 등)

2. 중단 학생과 유지 학생의 패턴은 어떻게 다른가?
 - 학업 성과 분포 차이
 - 학기 간 성취도의 변화 방향(증가/유지/감소)
 - 재정적 부담과 학업 지속 간의 관계

3. 예측 모델을 통해 중단 위험 학생을 조기에 식별할 수 있는가?
 - Logistic Regression, Random Forest, XGBoost 등 다양한 알고리즘 비교
 - 교차 검증(K-Fold)을 통한 모델의 안정성 평가
 - 주요 예측 변수를 기반으로 한 중단 위험 지표 제안

본 프로젝트는 이러한 문제 정의를 바탕으로, 학업 중단과 관련한 인사이트 발견과 학업 중단 예측 모델 구축을 최종 목표로 한다.

---

1) M.V.Martins, D. Tolledo, J. Machado, L. M.T. Baptista, V.Realinho. (2021) "Early prediction of student's performance in higher education: a case study" Trends and Applications in Information Systems and Technologies, vol.1, in Advances in Intelligent Systems and Computing series. Springer. DOI: 10.1007/978-3-030-72657-7_16

CHAPTER2 데이터 수집 및 전처리

2.1 데이터 출처 및 구조

본 프로젝트에서는 UCI Machine Learning Repository에 공개된 "Predict Students Dropout and Academic Success" 데이터셋을 사용하였다. 해당 데이터셋은 포르투갈의 한 고등교육기관에서 수집한 학생 등록 정보를 기반으로 구성되어 있으며, 학생 개인의 특성, 가정 및 사회·경제적 배경, 입학 전 학업 수준, 그리고 1·2학기 동안의 학업 성과(이수 학점, 평가 횟수, 평균 성적 등)을 포함하고 있다.

데이터셋은 총 4,424개의 관측치와 36개의 설명 변수, 1개의 목적 변수로 이루어져 있으며, 목적 변수로 학생 개개인의 학업 결과를 Graduate(졸업), Enrolled(재학), Dropout(중퇴)의 세 가지 범주로 구분하여 제공한다. 특히 학업 관련 변수들은 학생의 실제 성취도를 반영하는 지표로서 Dropout 예측에 중요한 역할을 할 것으로 예상된다.

□ 목적 변수 (1개)

	영문 컬럼명	한글 컬럼명	설명	비고
1	Target	학업 상태	1년의 이수 기간 후 학생의 상태	Graduate : 졸업 Dropout : 중퇴 Enrolled : 재학

□ 설명 변수 (36개)

해당 데이터셋은 범주형 변수를 포함하고 있으나, 모든 범주형 변수는 데이터 제공 단계에서 이미 정수 형태로 코딩되어 있어 전체 변수가 수치형 형태로 구성되어 있다.

	영문 컬럼명	한글 컬럼명	설명	비고
학생 개인 정보 (Student Information) - 12개				
1	Gender	성별	학생의 성별	1 - male 0 - female
2	Age at enrollment	입학 시 나이	입학 당시 학생의 나이	연속형 변수
3	Marital status	혼인 상태	학생의 혼인 상태	명목형 변수
4	Application mode	지원 방식	학생이 어떤 방식으로 대학에 지원했는지	명목형 변수
5	Application order	지원 순위	학생의 지원 순서	0~9지망
6	Course	전공 과정	학생이 선택한 전공	명목형 변수
7	Previous qualification	최종 학력	입학 이전의 최종 학력	명목형 변수
8	Nationality	국적	학생의 국적	명목형 변수

	영문 컬럼명	한글 컬럼명	설명	비고
9	Daytime/evening attendance	주간/야간 구분	주간 또는 야간 수업 참석 여부	1 - yes 0 - no
10	International	유학생 여부	학생이 유학생인지 여부	1 - yes 0 - no
11	Displaced	이주 여부	학생이 이주민인지 여부	1 - yes 0 - no
12	Educational special needs	특수 교육 필요 여부	학생이 특수 교육을 필요로 하는지 여부	1 - yes 0 - no
가정 • 경제적 변수(Socioeconomic) - 7개				
13	Mother's qualification	어머니 학력	학생 어머니의 학력 (교육 수준)	명목형 변수
14	Father's qualification	아버지 학력	학생 아버지의 학력 (교육 수준)	명목형 변수
15	Mother's occupation	어머니 직업	학생 어머니의 직업	명목형 변수
16	Father's occupation	아버지 직업	학생 아버지의 직업	명목형 변수
17	Debtor	채무 여부	학생에게 채무가 있는지 여부	1 - yes 0 - no
18	Tuition fees up to date	등록금 납부 최신 여부	등록금을 제때 납부했는지 여부	1 - yes 0 - no
19	Scholarship holder	장학금 수혜 여부	학생이 장학금을 받는지 여부	1 - yes 0 - no
학업 관련 변수 (Academic performance) - 12개				
20	Admission grade	입학 성적	입학 시 성적	연속형 변수
21	Prevuous qualification (grade)	이전 학력 성적	이전 학력에서의 성적 (등급)	연속형 변수
22	Curricular units 1st sem (credited)	1학기 이수 학점	1학기 이수 학점	연속형 변수
23	Curricular units 1st sem (enrolled)	1학기 신청 학점	1학기 수강 신청 학점	연속형 변수
24	Curricular units 1st sem (evaluations)	1학기 평가 학점	1학기 평가받은 학점	연속형 변수
25	Curricular units 1st sem (approved)	1학기 인정 학점	1학기 인정 학점 수	연속형 변수
26	Curricular units 1st sem (grade)	1학기 평균 성적	1학기 평균 성적	연속형 변수
27	Curricular units 1st sem (without evaluations)	1학기 미평가 학점	1학기 평가 제외 교과목 수	연속형 변수
28	Curricular units 2nd sem (credited)	2학기 이수 학점	2학기 이수 학점	연속형 변수
29	Curricular units 2nd sem (enrolled)	2학기 신청 학점	2학기 수강 신청 학점	연속형 변수
30	Curricular units 2nd sem (evaluations)	2학기 평가 학점	2학기 평가받은 학점	연속형 변수
31	Curricular units 2nd sem (approved)	2학기 인정 학점	2학기 인정 학점 수	연속형 변수
32	Curricular units 2nd sem (grade)	2학기 평균 성적	2학기 평균 성적	연속형 변수
33	Curricular units 2nd sem (without evaluations)	2학기 미평가 학점	2학기 평가 제외 교과목 수	연속형 변수

	영문 컬럼명	한글 컬럼명	설명	비고
학생의 입학 시점 경제/사회 환경 변수 (Regional Socioeconomic Indicators) - 3개				
34	Unemployment rate	실업률	실업률 (%)	연속형 변수
35	Inflation rate	물가상승률	물가상승률 (%)	연속형 변수
36	GDP	GDP	GDP	연속형 변수

2.2 데이터 전처리

□ 결측치 확인

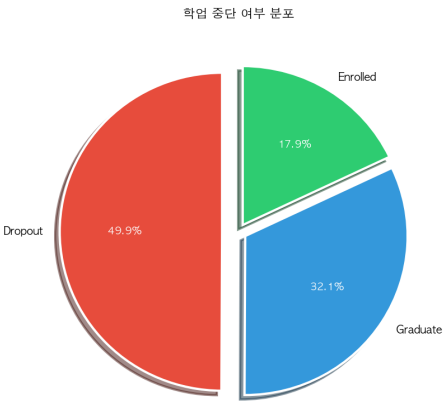
해당 데이터는 결측치가 존재하지 않는 데이터이므로, 따로 결측치 처리를 하지 않았다.

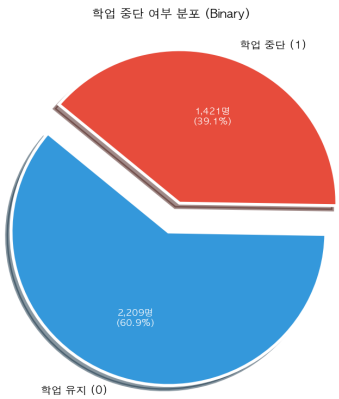
□ 한글 컬럼 매핑

원본의 영어 컬럼명을 앞서 설명했던 표의 한글 컬럼명으로 변경하여 해석이 용이하도록 했다.

□ 목적 변수 생성

본 프로젝트의 목적 변수는 학생의 최종 학업 상태로, 원본 데이터에서는 Graduate(졸업), Enrolled(재학), Dropout(중퇴)의 세 가지 범주로 제공된다. 이 중 Enrolled(재학)은 아직 학업 결과가 확정되지 않은 상태이다. 이를 '학업 유지'로 분류할 경우, 향후 중퇴할 가능성이 있는 학생까지 유지 그룹에 포함되어 노이즈가 발생할 수 있다. 이는 마치 아직 배송 중인 택배를 '배송 완료'로 처리하는 것과 같아, 모델 학습에 혼란을 줄 수 있다. 따라서 본 분석에서는 학업 결과가 확정된 Graduate(졸업)와 Dropout(중퇴)만을 분석 대상으로 하였으며, 재학생 794명을 제외한 3,630명의 데이터로 이진 분류 모델을 구축하였다. 이를 통해 타겟 정의를 명확히 하고, 보다 신뢰성 있는 예측 모델을 학습하고자 하였다.





□ 부모 변수 값 범주화

원본 데이터의 부모 학력 및 직업 변수는 매우 세분화된 숫자 코드로 되어 있으며, 각 코드가 실제 어떤 의미를 갖는지 직관적으로 해석하기 어렵다. 이렇게 세분화된 상태에서는 각 범주의 표본 수가 너무 작아져 중단률에서 의미를 찾을 때 어려움이 있을 것이라고 생각했다.

```
부모 관련 컬럼 유니크 값 개수
어머니 학력                    :  29개
아버지 학력                    :  34개
어머니 직업                    :  29개
아버지 직업                    :  42개
```

의미적으로 유사한 값을 묶어 학력을 6개의 수준(level)으로 범주화 하였다.

코드	분류명	설명
0	무학/문해	정규 교육 미이수 또는 기초 문해 수준
1	기초·중등 미완료	초등~중학교 수준, 고교 미졸업
2	중등(고교)	고등학교 졸업 또는 동등 학력
3	고등교육	대학교 학사 학위 취득
4	석·박사	석사 또는 박사 학위 취득
9	정보 없음	학력 정보 미기재 또는 불명

어머니/아버지의 직업은 12개로 분류 하였으며, 어머니 직업의 범주화 결과는 다음과 같다.

코드	분류명	설명
0	학생	현재 학업 중
1	관리자/임원	기업 임원, 고위 공무원 등
2	전문직	의료, 교육, ICT 등 고급 전문 직종
3	중급 전문/기술직	준전문가, 기술직 등
4	사무/행정직	일반 사무 및 행정 업무 종사자
5	서비스/판매/돌봄	서비스업, 판매직, 돌봄 노동 등
6	농림어업	농업, 임업, 어업 종사자
7	산업/건설/생산직	제조업, 건설업, 기계 조작 등
8	비숙련 노동자	단순 노무직

코드	분류명	설명
9	군 관련 직종	군인 및 관련 직종
90	기타	위 분류에 해당하지 않는 직종
99	무응답	직업 정보 미기재

아버지 직업 역시 동일한 기준으로 12개 그룹으로 재분류하였다. 어머니 직업과의 차이점은 서비스/판매/돌봄 범주에 보안직이 추가로 포함된 점이다.

코드	분류명	설명
0	학생	현재 학업 중
1	관리자/임원	기업 임원, 고위 공무원 등
2	전문직	의료, 교육, ICT, 금융 전문가 등
3	중급 전문/기술직	준전문가, 기술직 등
4	사무/행정직	일반 사무 및 행정 업무 종사자
5	서비스/판매/돌봄/보안	서비스업, 판매직, 돌봄, 보안직 등
6	농림어업	농업, 임업, 어업 종사자
7	산업/건설/생산직	제조업, 건설업, 기계 조작 등
8	비숙련 노동자	단순 노무직
9	군 관련 직종	군인 및 관련 직종
90	기타	위 분류에 해당하지 않는 직종
99	무응답	직업 정보 미기재

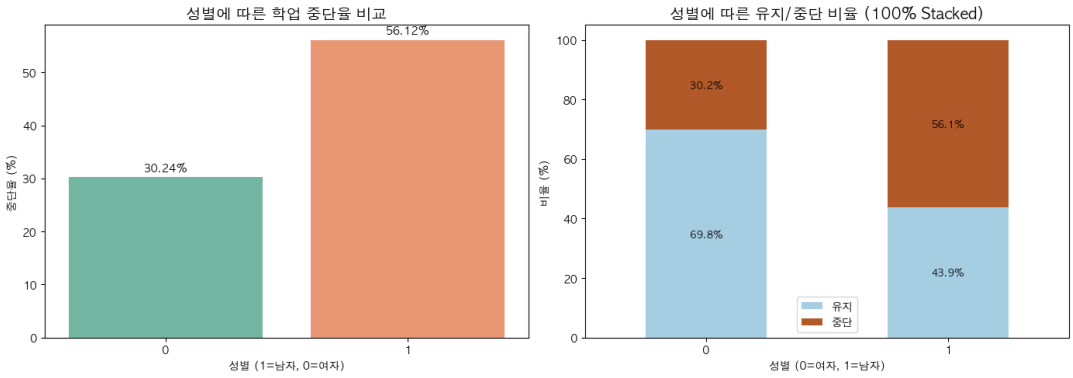
CHAPTER3 데이터 분석 및 피쳐 엔지니어링

3.1 탐색적 데이터 분석 (EDA)

□ 학생 개인 정보

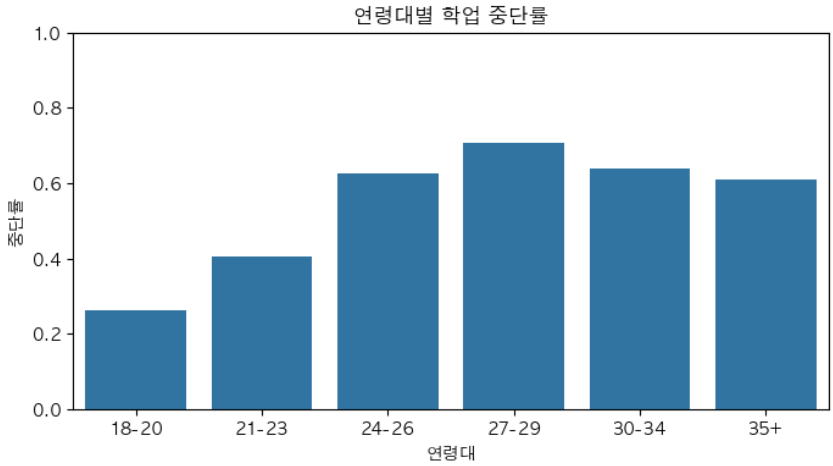
1. 성별(Gender)에 따른 학업 중단 분석

성별에 따른 학업 중단률을 시각화 한 결과, 여학생(0)의 중단률은 약 25% 수준으로 나타난 반면, 남학생(1)은 약 45%에 달해 상대적으로 높은 중단률을 보였다. 100% 누적 비율 그래프에서도 동일한 경향이 확인되었으며, 남학생 집단은 유지 비율이 낮고 중단 비율이 높게 나타났다. 다만, 성별은 학업·경제적 요인과 결합하여 간접적으로 영향을 받을 가능성이 높다.



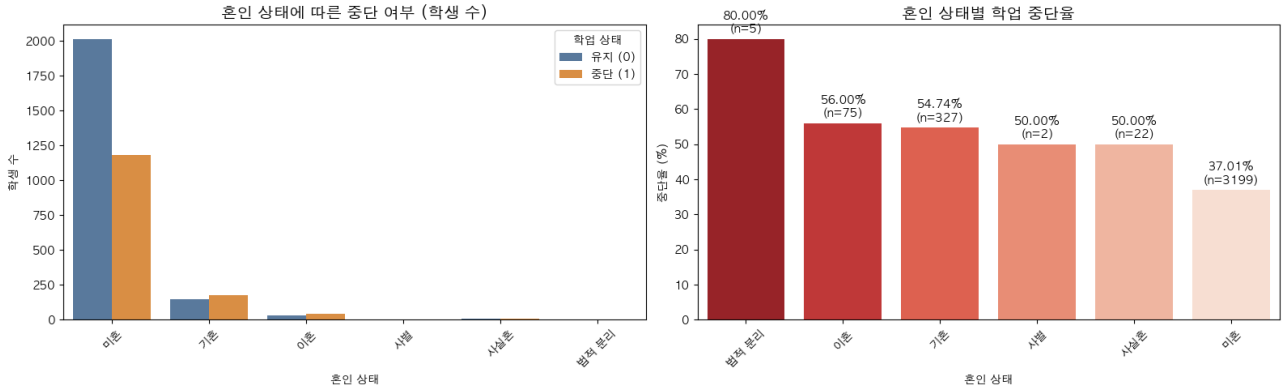
2. 입학 시 나이(Age at enrollment)에 따른 학업 중단 분석

연령대를 구간화하여 중단률을 분석한 결과, 18-20세에서는 약 27%로 가장 낮았으나 21-23세부터 중단률이 상승하기 시작하였다. 특히 24-29세 구간에서는 중단률이 60~70% 수준으로 가장 높아, 이 연령대가 학업 중단의 핵심 위험군인 것으로 나타났다. 30세 이상에서도 중단률이 여전히 높은 수준을 유지하여 성인 학습자의 직장/가정 부담이 학업 지속에 영향을 미치는 것으로 해석된다. 이러한 결과는 연령대별 맞춤형 학업 지원 정책이 필요함을 시사한다.



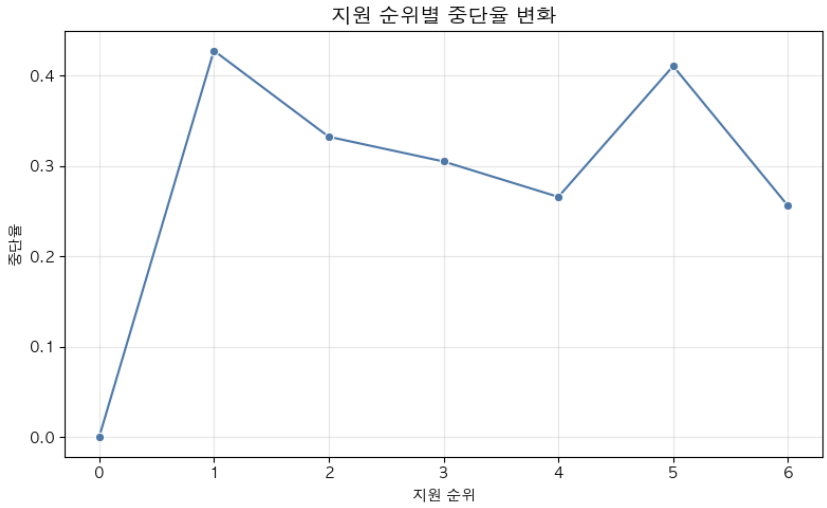
3. 혼인 상태(Marital status)에 따른 학업 중단 분석

혼인 상태와 중단 여부 간의 관계를 단순 빈도로 비교한 결과, 미혼 학생이 90% 이상으로 절대적으로 많아 단순 countplot은 해석적 의미가 제한적이었다. 이에 따라 혼인 상태별 중단률(%)을 산출하여 비교한 결과, 미혼보다 미혼이 아닌 집단에서 더 높은 중단률이 관찰되었다. 그러나 미혼을 제외 하고는 표본 수가 매우 적어 중단률의 통계적 신뢰성이 낮다고 판단하였다. 이에 따라 본 분석에서는 혼인 상태를 범주형 전체로 활용하기보다, 해석 가능성을 높이기 위해 '미혼 여부(이진 변수)'로 단순화하여 활용하는 것이 바람직하다고 판단하였다.



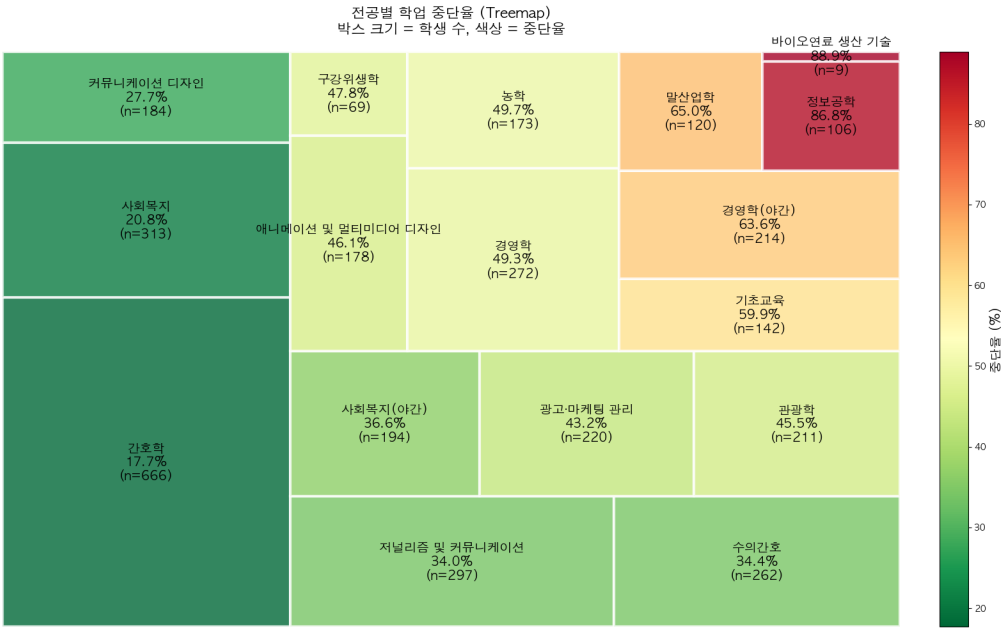
4. 지원 순위(Application order)에 따른 학업 중단 분석

지원 순위별 중단률을 분석한 결과, 비선형적 패턴이 관찰되었다. 지원순위 0은 중단률이 거의 0%로 매우 낮았으며, 이는 0지망의 학교를 온 만큼 만족도가 높은 것으로 보인다. 1순위에서는 중단률이 급격히 상승(약 35%) 한다. 이것은 0지망을 아쉽게 못갔다고 생각하는 학생들의 결과일 것으로 보이며, 이후 순위별 중단률은 20~30%대로 일관된 감소 혹은 증가 패턴을 보이지 않는다. 5순위에서도 다시 한 차례 상승이 관찰됨에 따라, 원하는 전공이 아니거나, 원하지 않은 학교 등 어쩔 수 없이 들어온 케이스로 보인다. 이는 지원 순위가 단순한 선호도를 넘어, 학생의 전공 매칭도, 입학 동기, 적성 부합 여부 등을 반영하기 때문으로 해석할 수 있다.



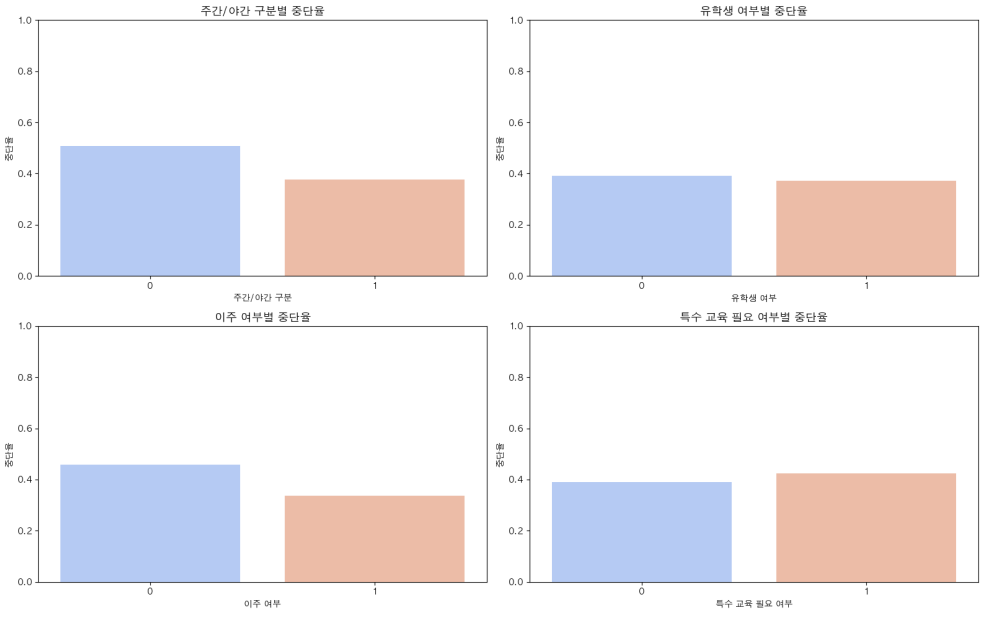
5. 전공(Course)에 따른 학업 중단 분석

학생 규모가 가장 큰 간호학과는 매우 낮은 중단률을 보인 반면, 바이오 연료 생산 기술이나 말산업학처럼 소규모 전공에서는 중단률이 상대적으로 높게 나타났다. 이는 전공 규모에 따라 학업 지원 체계나 행정적 자원이 균등하게 배분되지 않았을 가능성을 시사하며, 소형 전공에 대한 학습 지원이나, 커리큘럼 개편 등 개선이 필요함을 보여준다.



6. 학생의 배경 및 상황에 따른 학업 중단 분석

학생 개인 배경 및 상황과 관련 있는 변수 (주간/야간, 유학생 여부, 이주 여부, 특수 교육 필요 여부)들은 전반적으로 학업 중단 여부에 큰 영향을 주지 않는 것으로 나타났다. 일부 차이가 존재하지만 그 크기는 크지 않다.



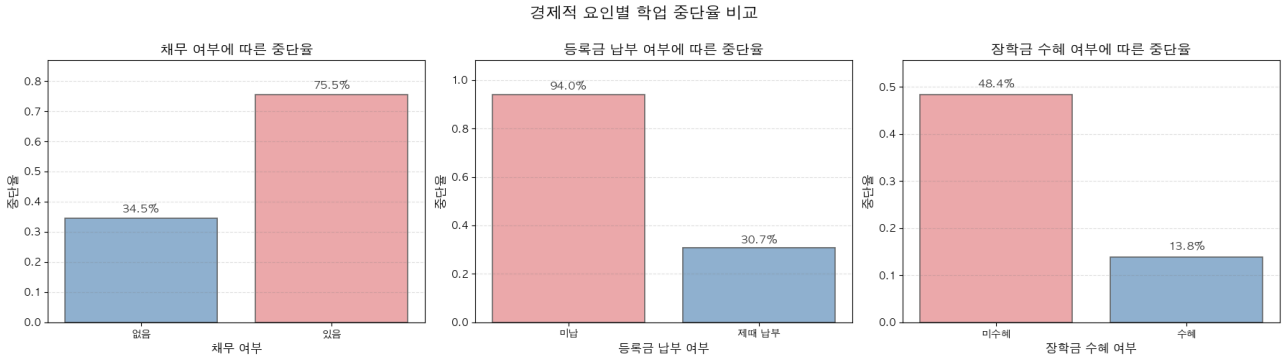
□ 가정・경제적 변수

1. 재정 상태에 따른 학업 중단 분석

채무가 있는 학생의 학업 중단률이 두 배 이상 높게 나타나며, 재정적 어려움이 학업 지속에 직접적인 부담으로 작용했을 것으로 보인다.

또한 등록금을 제때 납부하지 못한 학생의 중단률이 압도적으로 높았다. 학업 유지에 있어 재정적 안정성이 얼마나 중요한지를 명확하게 보여주기도 하지만, 학업을 중단 할 것이기 때문에 일부러 납부하지 않았을 가능성도 고려해 볼 필요가 있다.

장학금을 받는 학생의 중단률은 현저히 낮은데, 이는 장학금이 학생의 학업 의지 및 지속 가능성을 실질적으로 높여주는 역할을 함을 보여준다.

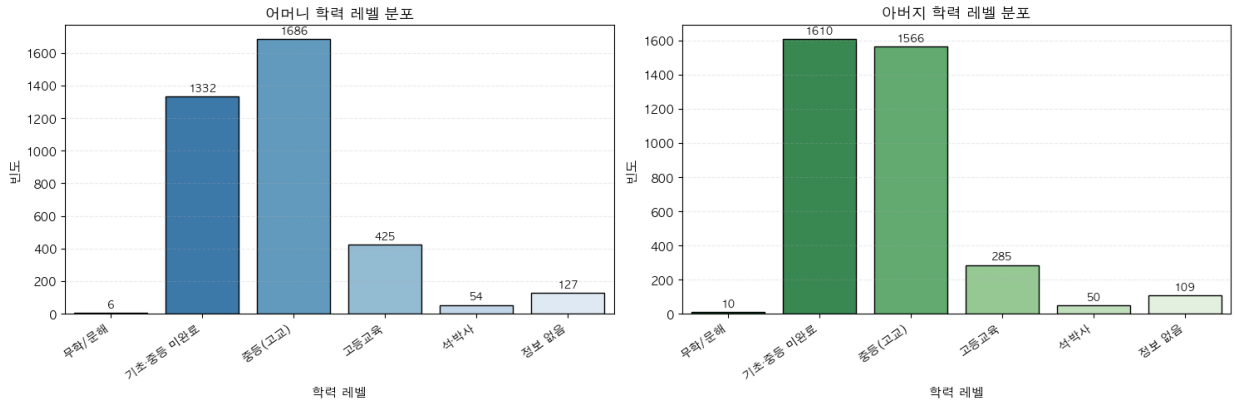


특히 최근 등록금을 납부 하지 않은 학생은 94%의 높은 학업 중단률을 보였으며, 채무가 있는 학생, 장학금 수혜를 받지 않는 학생이 뒤를 이었다.

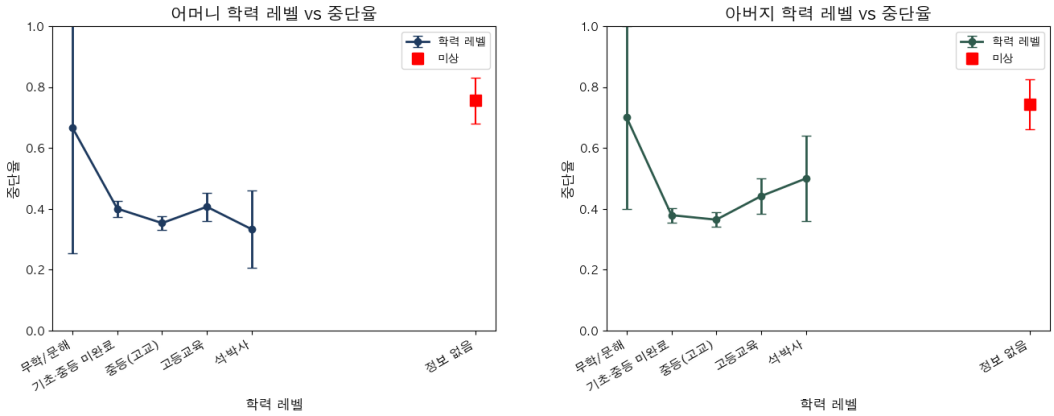
	Variable	Category	Count	DropoutRate	RiskLevel
0	등록금 납부 최신 여부	0	486	0.940	HIGH
1	채무 여부	1	413	0.755	HIGH
2	장학금 수혜 여부	0	2,661	0.484	MEDIUM
3	채무 여부	0	3,217	0.345	MEDIUM
4	등록금 납부 최신 여부	1	3,144	0.307	MEDIUM
5	장학금 수혜 여부	1	969	0.138	LOW

2. 부모님의 최종 학력 및 직업에 따른 학업 중단 분석

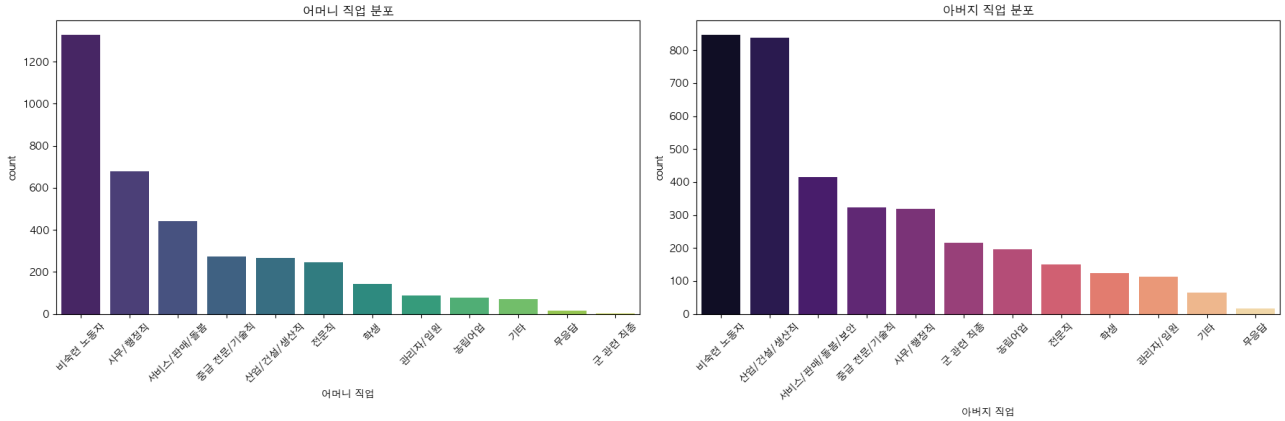
부모 학력 분포를 보면, 어머니와 아버지 모두 중등(고교) 학력이 가장 많고, 그 다음으로 기초·중등 미완료 순이다. 석·박사 학력은 전체의 약 1.5% 수준으로 소수에 해당한다.



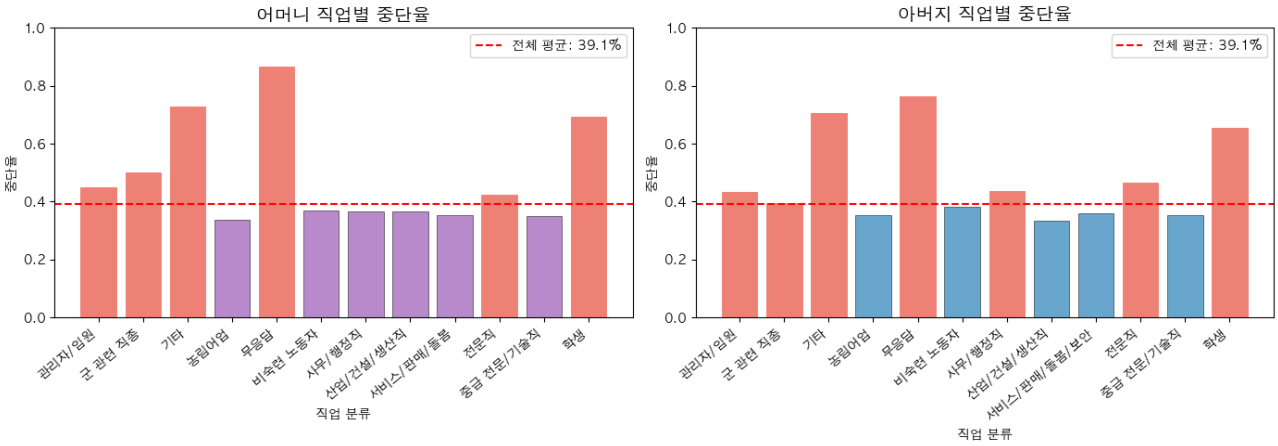
학력 레벨별 중단율을 살펴보면, 무학/문해 그룹에서 약 70%대로 가장 높게 나타났으며, 기초·중등 미완료 이상부터는 35~50% 수준으로 낮아져 안정화되는 경향을 보였다. 특히 정보 없음 그룹은 약 77%의 높은 중단율을 보이는데, 이는 학력 정보 미기재 자체가 사회적, 경제적 취약 계층의 특성을 반영할 가능성이 있을 것으로 보인다. 또는 부모님이 계시지 않는 경우도 고려할 수 있을 것이다.



직업 분포에서 어머니는 비숙련 노동자가 가장 많고, 아버지는 비숙련 노동자와 산업/건설/생산직이 비슷한 비율로 높게 나타났다. 전문직, 관리자/임원 등 상위 직종은 상대적으로 소수이다.



직업별 중단율을 보면, 무응답 그룹이 어머니, 아버지 모두 가장 높았으며, 학생 그룹도 약 70% 정도로 평균(39.1%)보다 눈에 띄게 높다. 반면 중급 전문/기술직, 전문직, 사무/행정직, 서비스/판매/돌봄 그룹은 35% 수준으로 상대적으로 낮은 중단율을 보였다. 이는 안정적인 직장을 가진 부모님이 계실수록 학업 중단이 줄어듦을 나타내는 것으로 보인다.

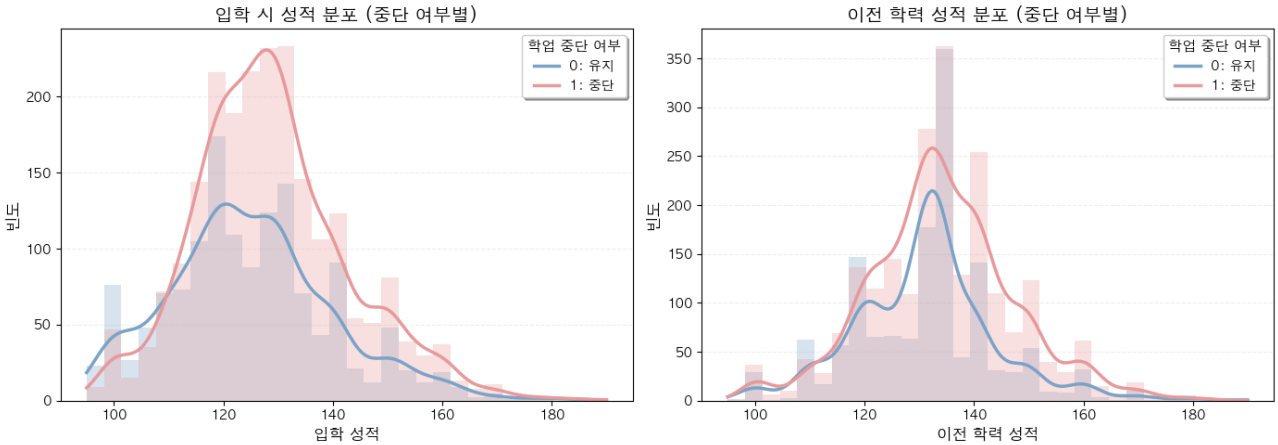


□ 학업 관련 변수

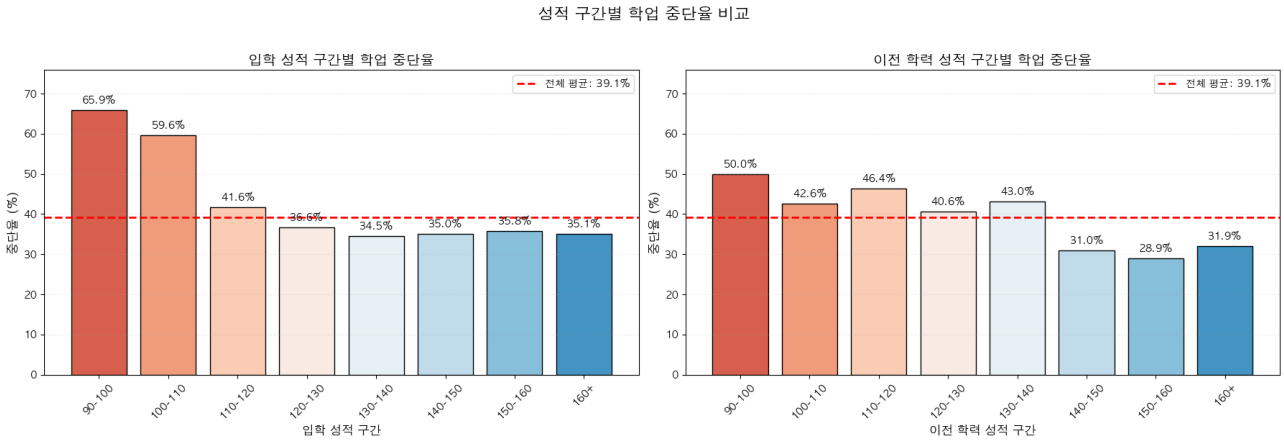
1. 입학 성적과 이전 학력 성적과 학업 중단 분석

히스토그램에서 입학 시 성적과 이전 학력 성적 분포 모두 학업 중단의 경우에서, 왼쪽에 치우쳐져 있을 것이라고 예상 했던 것과 다르게, 유지와 중단은 비슷한 분포 형태를 보여주었다.

입학 시 성적에서 학업 유지 학생(0)은 약 120~130점 구간에서 가장 높은 빈도를 보이고 학업 중단 학생(1) 도 거의 동일한 구간에 집중되어 있다. 이전 학력 역시 학업 유지/중단 두 그룹 모두 130점 후반에 밀집되어 있다.

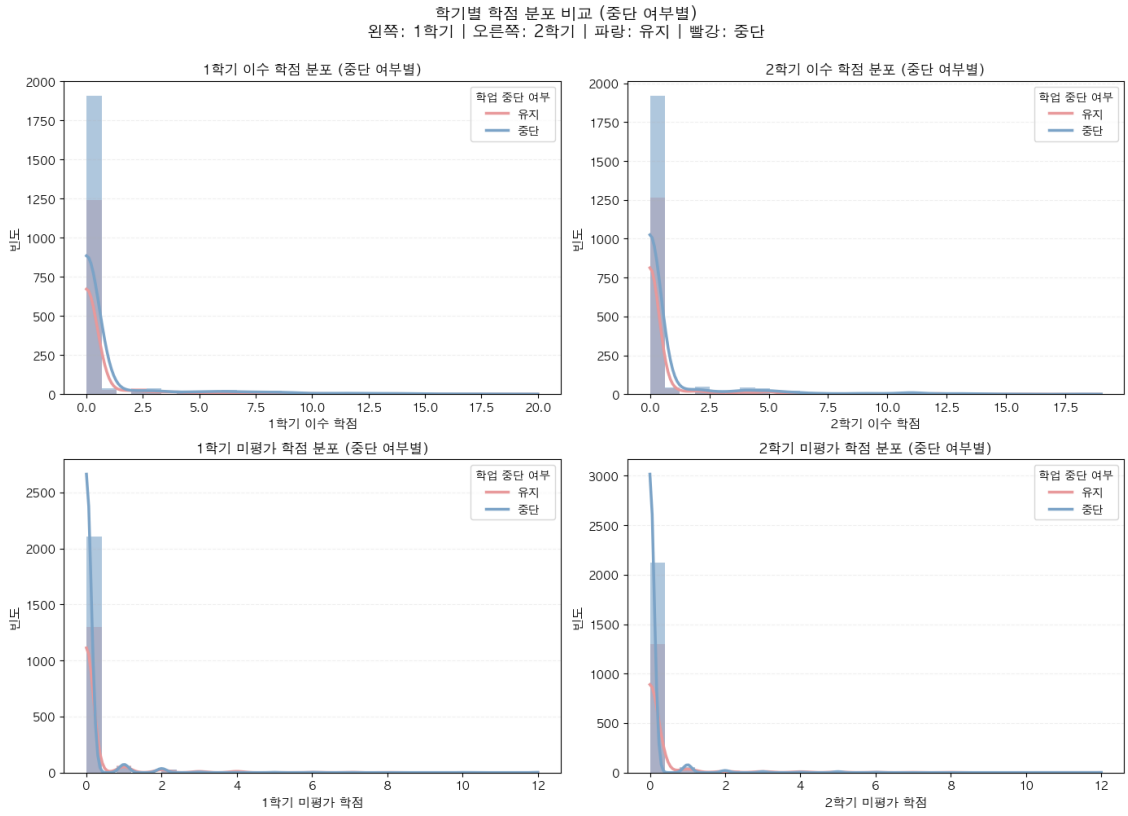


성적 구간별 학업 중단률을 비교해 볼 때, 입학 성적은 큰 폭은 아니지만 낮을수록 높은 경향을 보였다. 이전 학력의 성적 구간은 큰 영향을 주지 않는 것으로 보인다.

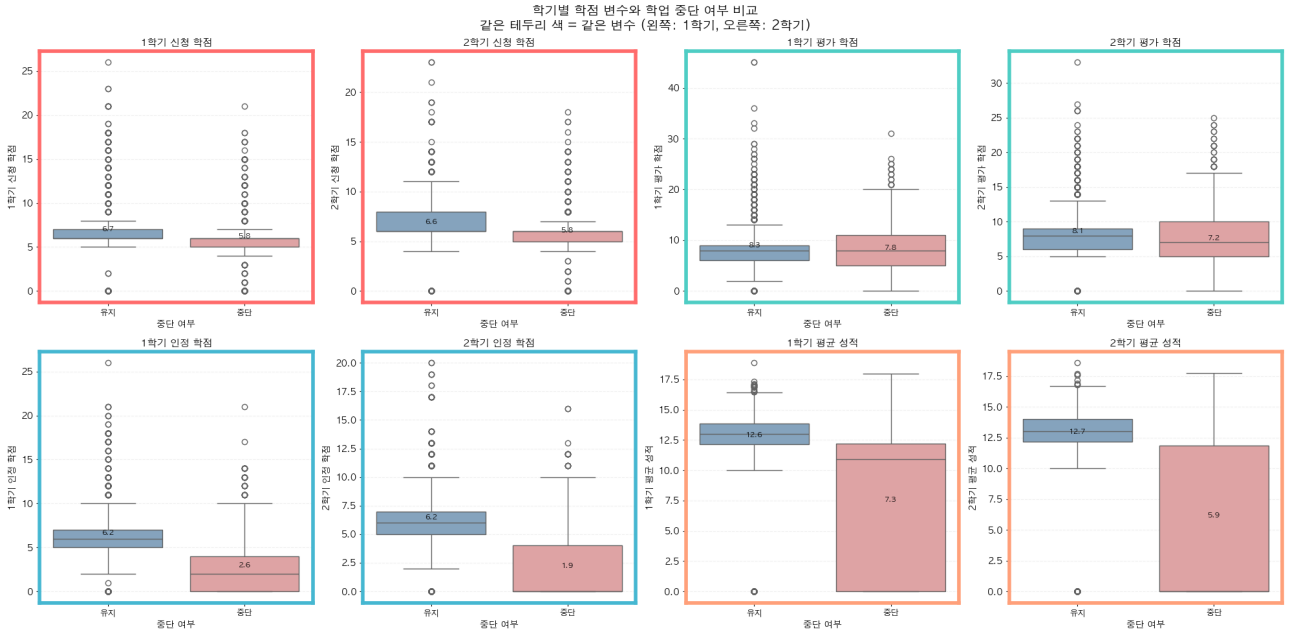


2. 1학기 / 2학기 학업 정보에 따른 학업 중단 분석

이수 학점 / 미평가 학점은 분포가 0 근처 구간에 매우 치우쳐 있고, 분포의 형태도 비슷해서 학업 중단 여부와 거의 무관하다고 판단했다.



전반적으로 학업 유지 학생들은 신청 학점, 인정 학점, 평균 성적이 약간 더 높게 나타났는데, 특히 평균 성적의 경우, 중단 학생이 더 낮은 점수대에 넓게 분포함을 알 수 있다.

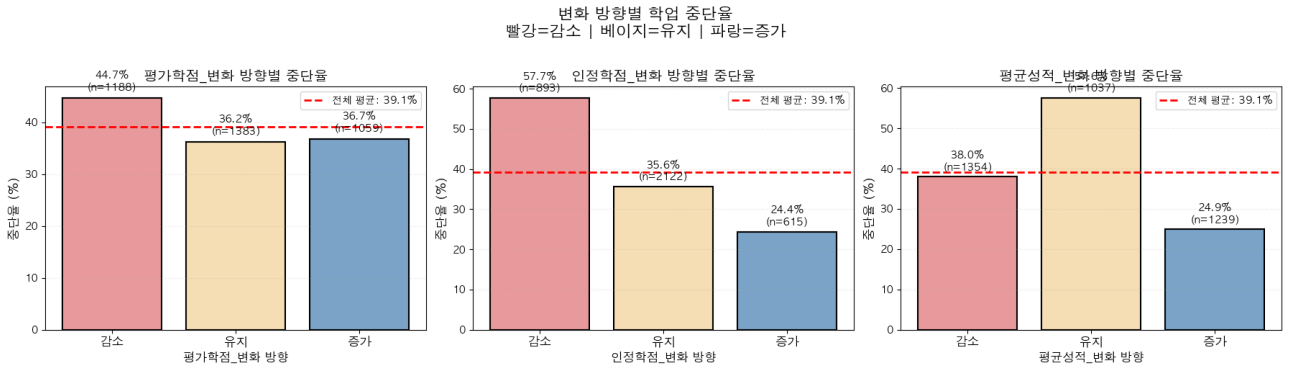


학기 간 성적, 학점 변화가 학업 중단 여부에 어떤 영향을 미치는지 확인하기 위해, 변수를 감소/유지/증가의 세 방향으로 구분하여 중단률을 비교했다. 그 결과, 변화의 '크기'보다 방향 자체가 중단 위험을 파악하는 데 더 중요한 신호임이 확인되었다.

평가학점이 감소한 학생들의 중단률은 35.8%로 전체 평균(32.1%)보다 높게 나타났다. 평가받은 과목 수가 줄어드는 것은 학업 참여도 감소를 의미한다고 판단된다.

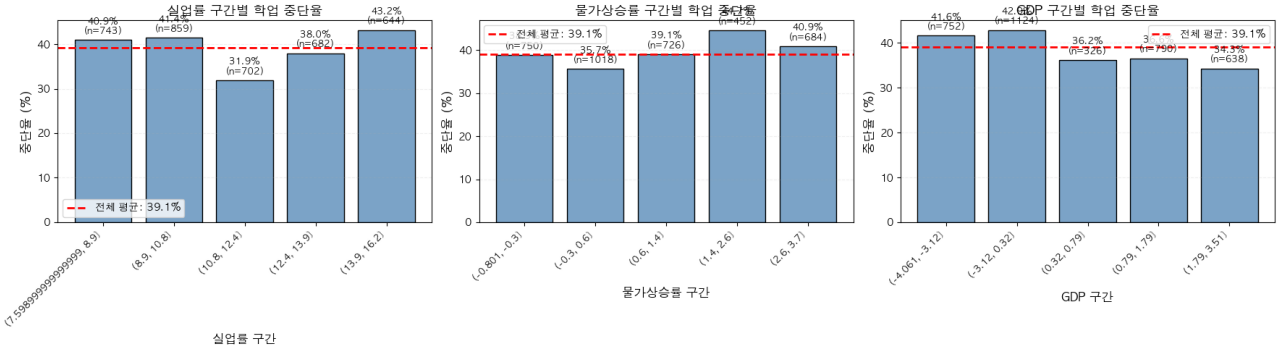
인정학점이 감소한 학생들의 중단률은 43.6%로 높은 비율을 가지며, 인정학점 감소는 학사 관리 미흡, 출석 문제 등 학교 생활의 적극성과 연관이 있을 것으로 판단된다.

평균성적 감소 그룹의 중단률(29.9%)은 전체 평균 대비 낮지 않지만, 성적이 '유지된' 학생들의 중단률이 51.8%로 가장 높게 나타났다는 점을 볼 때, 단순히 "성적이 유지되었다"가 아니라, 낮은 성적을 그대로 유지하거나 성장이 정체된 학생들이 중단 위험이 높다는 신호로 해석될 가능성이 크다.



□ 입학 당시 사회·경제 지표

거시경제 지표인 실업률, 물가상승률, GDP가 학업 중단률에 미치는 영향을 분석한 결과, 세 변수 모두 학업 중단률과 뚜렷한 관계를 보이지 않았다. 경제 상황이 좋거나 나쁨에 따라 중단률이 크게 달라지지 않는 것으로 보인다. 결론적으로 거시경제 환경보다는 개인의 학업 성과나 재정 상태 등 개인 수준 변수가 학업 중단에 더 직접적인 영향을 미치는 것으로 판단된다.



3.2 feature engineering

탐색적 데이터 분석(EDA)을 수행한 후, 학업 중단 여부를 보다 효과적으로 예측하기 위해 Feature Engineering을 수행하였다. 특히 개인 특성, 학업 성과 변화, 학업 패턴의 방향성을 반영하기 위한 변수를 중심으로 Feature Engineering을 수행하였으며, 학업 중단에 영향을 많이 안 주는 것으로 보였던 변수들은 삭제하였다.

① 학생 개인 특성 관련 파생변수

원본 데이터의 '혼인 상태' 변수는 여러 범주로 구성되어 있었으나, 학업 중단과의 관계를 해석하기에는 복잡하였다. 이에 따라, 혼인 상태가 1(미혼)인 경우를 1(미혼), 그 외 범주를 모두 0(미혼 아님)으로 단순화하여 이진 파생변수 '미혼 여부'를 생성하였다.

② 학업 성과 변화량 기반 파생변수 (연속형 변화 Feature)

학생의 학업을 중단한 학생들이 모두 학점이나 성적이 유지 학생보다 낮았던 EDA 결과에 따라, 1학기와 2학기 성적 차이를 이용하여 다음과 같은 변화량 변수를 생성하였다.

```
평가학점 변화량 = '2학기 평가 학점' - '1학기 평가 학점'
인정학점 변화량 = '2학기 인정 학점' - '1학기 인정 학점'
평균성적 변화량 = '2학기 평균 성적' - '1학기 평균 성적'
```

③ 학업 성과 변화 방향 파생변수 (감소/유지/증가 범주형 Feature)

EDA 분석에서 성적의 변화량보다 성적 변화의 '방향'이 학업 중단 위험을 더 명확히 구분하는 것으로 나타났다. 예를 들어, 평균 성적이 그대로 '유지'된 학생의 중단율이 오히려 더 높게 나타나는 등 방향성 자체가 중요한 위험 요인임이 확인되었다. 이를 반영하여, 변화량을 기반으로 다음과 같은 범주형 파생변수(감소/유지/증가)를 생성하였다. 감소는 0, 유지는 1, 증가는 2로 처리하였다.

```
평가학점 변화방향 ∈ {감소 / 유지 / 증가}
인정학점 변화방향 ∈ {감소 / 유지 / 증가}
평균성적 변화방향 ∈ {감소 / 유지 / 증가}
```

④ 변수 삭제

EDA 수행 결과, 학업 중단에 영향을 많이 안 주는 것으로 보였던 변수들은 삭제하였다.

```
del_col =[
"혼인 상태", "주간/야간 구분", "최종 학력", "유학생 여부", "이주 여부",
"이전 학력 성적", "1학기 이수 학점", "2학기 이수 학점", "1학기 미평가 학점", "2학기 미평가 학점",
"1학기 평가 학점", "2학기 평가 학점", "실업률", "물가상승률", "GDP"
]
```

3.3 학습용 데이터 구성 (train/test 분리)

모델 학습 및 평가를 위해 전체 데이터를 학습용(Train)과 테스트용(Test)으로 분리하였다. 분리 비율은 8:2로 설정하였으며, 학업 중단 여부(Target)의 클래스 비율을 유지하기 위해 층화추출(Stratified Sampling)을 적용하였다. 이를 통해 학습 데이터와 테스트 데이터 모두에서 중단/유지 비율이 원본 데이터와 동일하게 유지되도록 하였다.

분리 결과, 학습 데이터는 2904개, 테스트 데이터는 726개의 관측치로 구성되었다.

```
Train / Test 데이터셋 분할
총 데이터 수 :  3630
  - Train: 2,904개 (80.0%)
  - Test:  726개 (20.0%)

타겟 분포 (Train):
학업 중단 여부
0    0.608471
1    0.391529
Name: proportion, dtype: float64

타겟 분포 (Test):
학업 중단 여부
0    0.608815
1    0.391185
Name: proportion, dtype: float64
```

```
X : 27 개의 컬럼

Index(['지원 방식', '지원 순위', '전공 과정', '국적',
'어머니 학력', '아버지 학력', '어머니 직업', '아버지 직업',
'입학 성적', '채무 여부', '등록금 납부 최신 여부',
'성별', '장학금 수혜 여부', '입학 시 나이','1학기 신청
학점', '1학기 인정 학점', '1학기  평균 성적', '2학기
신청 학점', '2학기 인정 학점', '2학기 평균 성적', '미혼
여부', '평가학점 변화량', '인정학점 변화량','평균성적
변화량', '평가학점 변화방향', '인정학점 변화방향',
'평균성적 변화방향']
,dtype='object')
```

CHAPTER4 모델링 및 검증 평가

4.1 모델 학습 및 파라미터 설정

학업 중단 예측을 위해 다양한 특성의 분류 모델을 비교하고자 5개 모델을 선정하였다. 모델 선택은 선형 기반 모델(Logistic Regression), Bagging 기반 모델(Random Forest), Boosting 기반 모델(XGBoost, LightGBM)을 균형 있게 포함하도록 구성하였다.

Logistic Regression은 이진 분류의 기본 모델로, 변수별 계수를 통해 예측 결과를 해석할 수 있어 중단 요인 분석에 적합하다. Decision Tree는 단일 트리 기반 모델로, 분류 규칙이 직관적이며 앙상블 모델과의 성능 비교 기준으로 활용하였다. Random Forest는 여러 Decision Tree를 병렬로 학습하는 Bagging 앙상블 기법으로, 단일 트리 대비 과적합을 줄이고 일반화 성능을 높일 수 있다. XGBoost와 LightGBM은 Boosting 기반 앙상블 모델로, 이전 모델의 오류를 순차적으로 보완하며 학습한다.

모델의 일반화 성능을 정량적으로 비교하기 위해 전체 데이터를 학습용(Train)과 평가용(Test)으로 분리한 뒤, 학습 데이터에 대해 5-Fold 교차검증을 적용하여 각 모델의 성능을 안정적으로 추정하였다. 모델 학습 시 하이퍼파라미터는 기본값을 기반으로 안정적인 설정을 사용하였다. 이는 본 연구의 목적이 모델 간 상대적 성능 비교에 있으며, 과도한 튜닝을 통해 특정 모델이 유리하게 되는 것을 방지하기 위함이다. 각 모델별 주요 설정은 다음과 같다.

코드	분류명
Logistic Regression	max_iter=1000, random_state=42
Decision Tree	random_state=42
Random Forest	n_estimators=100, random_state=42
XGBoost	n_estimators=100, random_state=42, eval_metric='logloss'
LightGBM	n_estimators=100, random_state=42, verbose=-1

4.2 모델 평가 지표

1. 기본 모델 성능 비교

학습된 각 모델은 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 Score, ROC-AUC 지표를 기준으로 평가하였다. 특히 F1 Score는 학업 중단(Positive Class)이라는 비대칭적 문제 특성을 고려할 때 중요한 판단 기준으로 활용하였다.

모델	F1 Score (평균)	표준편차
Logistic Regression	0.8699	±0.0151
LightGBM	0.8632	±0.0103
XGBoost	0.8588	±0.0103
Random Forest	0.8588	±0.0103
Decision Tree	0.7923	±0.0103

평가 결과, Logistic Regression이 가장 높은 F1 Score와 안정적인 재현율을 보여 최종 모델로 선정되었으며, Boosting 기반 모델인 LightGBM이 가장 낮은 표준편차로 안정적인 성능을 나타냈다.

모델	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.938	0.925	0.915	0.920	0.972
LightGBM	0.937	0.922	0.915	0.919	0.970
XGBoost	0.924	0.896	0.912	0.904	0.967
Random Forest	0.923	0.925	0.873	0.899	0.967
Decision Tree	0.855	0.793	0.852	0.822	0.855

최종 평가 결과, Logistic Regression이 F1 Score 92.0%, ROC-AUC 97.2%로 가장 우수한 성능을 보여 최종 모델로 선정되었다. 이는 EDA 기반의 피처 선택을 통해 학업 중단과 선형적 관계가 강한 변수들만 남겼기 때문으로 해석된다. 이는 복잡한 모델이 항상 좋은 것이 아니라, 데이터 특성에 맞는 모델 선택이 중요함을 보여준다.

2. PCA 성능 실험 결과

학업 성적 관련 변수들은 서로 강한 상관관계를 가질 가능성이 높아, 다중공선성을 완화하기 위한 방법으로 PCA를 적용하여 성능 변화를 확인하였다. PCA 적용 시 성적 변수 4개('1학기 평가 학점', '1학기 평균 성적', '2학기 인정 학점', '2학기 평균 성적')를 2개의 주성분으로 축약하여 Logistic Regression 모델에 입력하였다.

PCA 적용 시 오히려 F1 Score가 감소하였다. 이는 성적 원본 변수들이 학업 중단 여부와 직접적으로 연결되는 중요한 정보를 담고 있어, 여러 변수를 하나의 축으로 결합하는 PCA 과정에서 설명력이 손실되기 때문이다.

```
Logistic (No PCA): F1 = 0.8784
Logistic (PCA): F1 = 0.8739
```

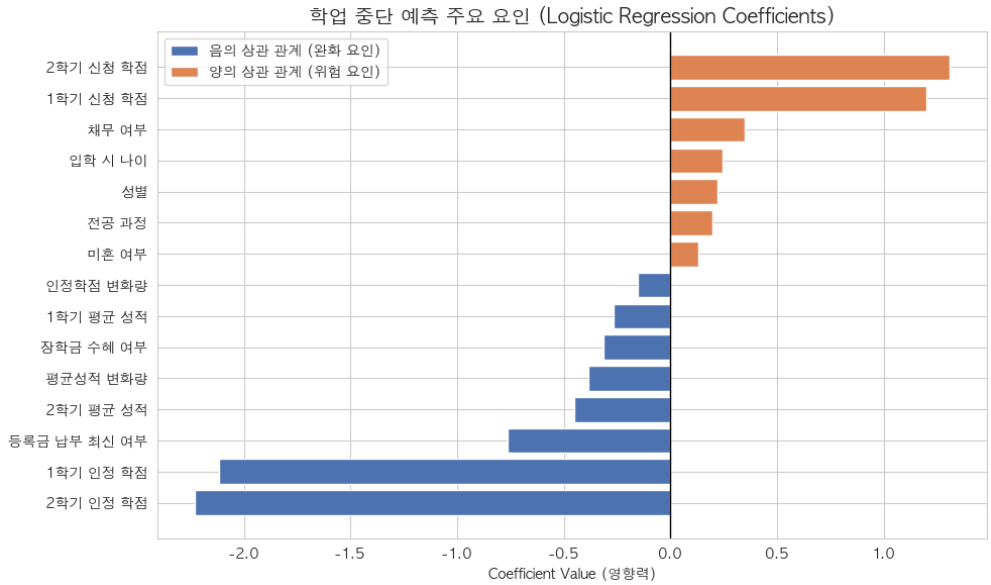
PCA 적용 시 성능이 하락한 이유는 여러 가지로 설명될 수 있다. 우선, PCA가 생성하는 주성분은 학업 성취를 구성하는 세부적 구조를 충분히 반영하지 못하며, 서로 다른 의미를 가진 성적 변수들이 하나의 축으로 결합되면서 중요한 정보가 축약되는 문제가 발생한다. 또한 성적 변수들은 본래 높은 설명력을 지니고 있어 차원 축소가 오히려 예측에 불리하게 작용할 수 있다. 더불어 이진 분류 문제에서 중요한 변수의 방향성 정보가 PCA 변환 과정에서 소실되는 점 역시 성능 저하의 원인으로 볼 수 있다.

결과적으로, 학업 성적 변수들은 해석과 예측 성능 모두에서 핵심적인 정보를 제공하기 때문에, 본 데이터에서는 PCA 기반의 차원 축소가 모델 성능 향상에 도움이 되지 않는 것으로 확인되었다.

4.3 주요 변수 중요도

1. 로지스틱 회귀 계수(Coefficients) 분석

아래 그림은 로지스틱 회귀 모델의 계수를 시각화한 것이다. 계수가 양수일수록 학업 중단 위험을 증가시키는 요인, 음수일수록 학업 중단 위험을 감소시키는 요인으로 작용함을 의미한다.



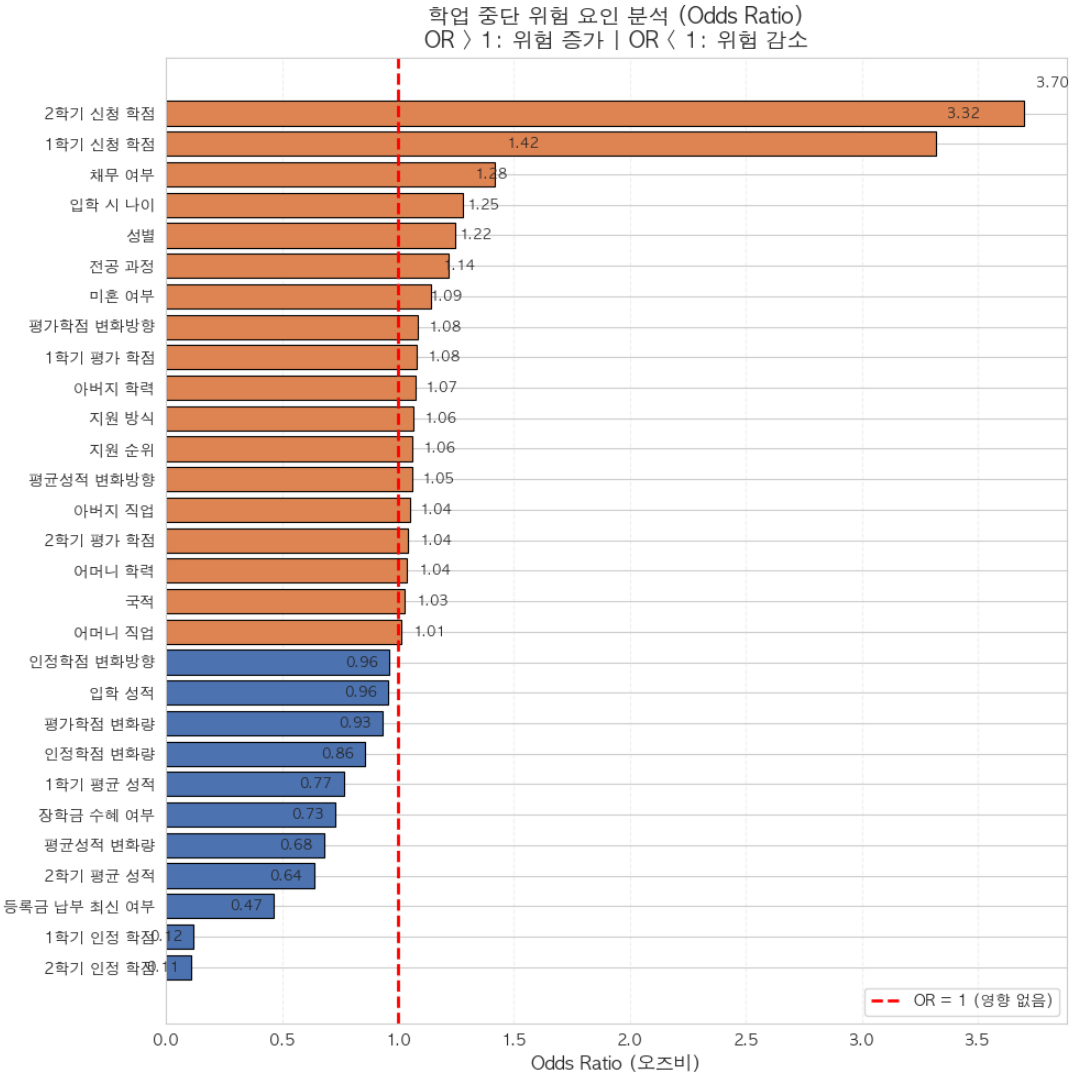
학업 중단 위험을 높이는 주요 요인은 학업 부담과 관련된 변수들로 나타났다. 특히 1·2학기 신청 학점이 많을수록 중단 위험이 크게 증가했으며, 이는 과도한 수강 신청이 학업 부담을 가중해 중도 포기로 이어질 가능성이 높음을 의미한다. 2학기 이수 학점 또한 양의 영향을 보여, 무리한 학점 이수가 학업 지속에 부정적일 수 있음을 시사한다.

학업 부담 외에도 경제적·배경적 요인이 중단 위험에 영향을 미쳤다. 채무 여부, 전공 과정, 입학 시 나이, 성별 등이 일정 수준 위험을 높이는 것으로 나타났으며, 특히 연령이 높을수록 외부 요인으로 인해 학업을 지속하기 어려울 가능성이 크다는 점이 확인되었다.

반면, 중단 위험을 가장 효과적으로 낮추는 보호 요인은 실제 학업 성취였다. 1·2학기 인정 학점이 높을수록 중단 위험이 크게 감소하였고, 이는 신청 학점보다 실제 취득한 학점이 학업 지속을 결정짓는 핵심 요소임을 보여준다. 또한 등록금 납부가 최신 상태이거나 장학금을 수혜받는 경우 중단 위험이 낮아지는 등 경제적 안정성도 중요한 보호 요인으로 확인되었다. 성적이 상승하는 학생(평균성적 변화량 양수) 역시 중단 가능성이 낮았는데, 이는 학업 성취의 개선이 지속 동기로 작용하기 때문이다.

2. 오즈비(Odds Ratio) 기반 위험 요인 해석

오즈비(Odds Ratio)는 Logistic Regression 계수에 지수(exp)를 취한 값으로, 해당 변수가 1단위 증가할 때 학업 중단 확률이 몇 배로 변하는지를 나타낸다. OR이 1보다 크면 중단 위험이 증가하고, 1보다 작으면 감소한다.



가장 강한 영향을 미치는 변수는 2학기 신청 학점(OR=3.32)과 1학기 신청 학점(OR=2.26)으로 나타났다. 이는 학점을 많이 신청했음에도 이수하지 못한 경우 중단 위험이 크게 증가함을 시사한다. 2학기 이수 학점(OR=1.70), 채무 여부(OR=1.40), 전공 과정(OR=1.38) 역시 중단 위험을 높이는 주요 요인으로 확인되었다.

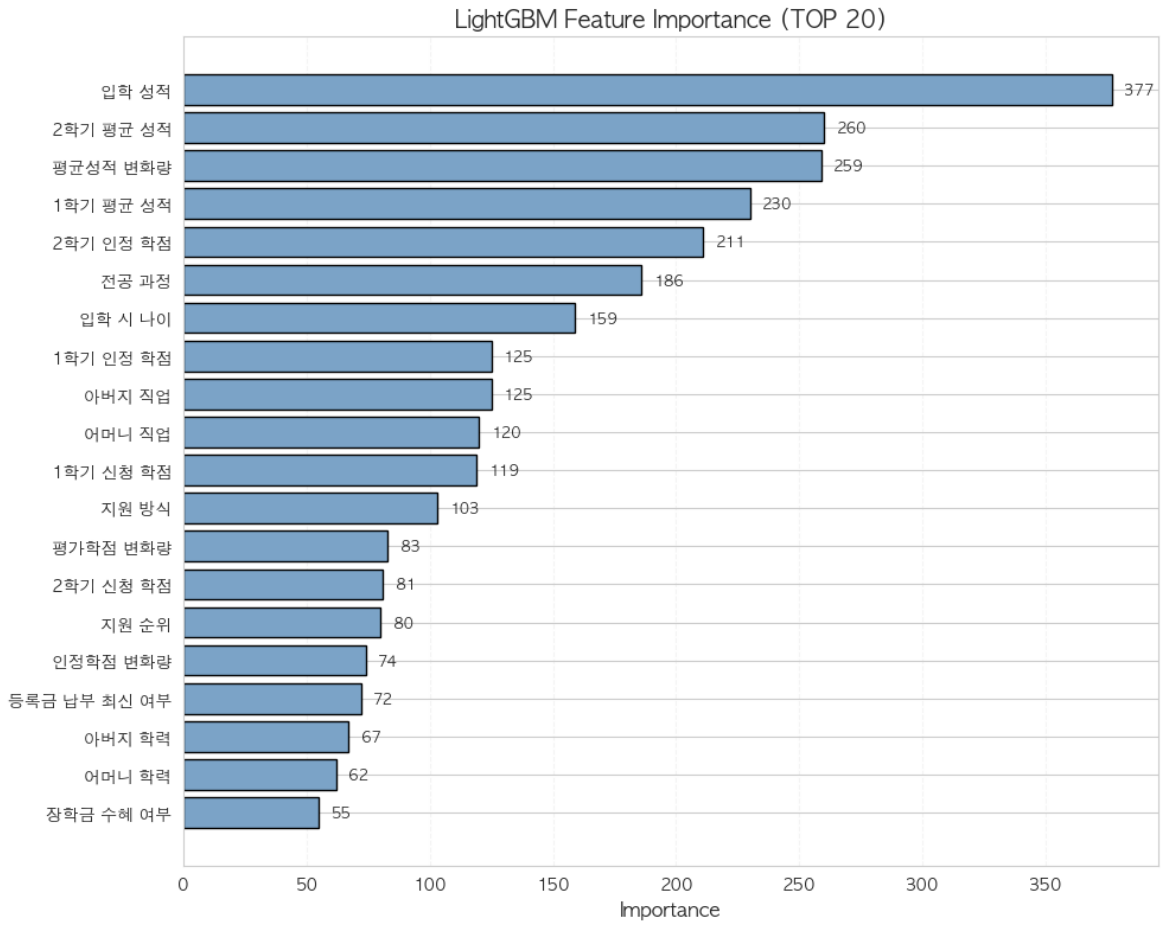
반면 2학기 인정 학점(OR=0.10)과 1학기 인정 학점(OR=0.24)은 중단 위험을 크게 낮추는 요인으로 나타났다. 이는 학점을 실제로 취득한 학생일수록 학업을 지속할 가능성이 높음을 의미한다. 등록금 납부 최신 여부(OR=0.47), 2학기 평균 성적(OR=0.74), 장학금 수혜 여부(OR=0.74) 역시 중단 위험을 낮추는 보호 요인으로 확인되었다.

3. LightGBM importance을 통한 변수 중요도

LightGBM은 비선형 구조를 반영하는 모델로, 예측 성능 향상에 기여한 변수를 Gain 기준으로 평가한다. 분석 결과, 평균성적 변화량, 입학 성적, 전공 과정, 입학 시 나이, 평가학점 변화량 등이 주요 변수로 나타났으며, 특히 평균성적 변화량이 가장 높은 중요도를 보였다. 이는 성적의 절대 수준보다 성적이 어떻게 변했는지(Trend) 가 학업 중단 여부 예측에 더 큰 영향을 미친다는 점을 의미한다.

또한 전공 과정, 지원 방식, 부모 직업 등 배경 변수도 일정한 중요도를 보이며, 학생의 환경적, 선택적 요인이 예측에 기여함을 확인할 수 있었다.

종합적으로 LightGBM 중요도는 선형 모델 결과를 보완하며, 성적 변화와 학업 관련 변수들이 학업 중단 예측의 핵심이라는 결론을 다시 한 번 뒷받침한다.



CHAPTER5 결론 및 향후 개선 방향

5.1 결과 해석 및 연구 결론

세 가지 모델(Logistic Regression, Odds Ratio, LightGBM Feature Importance)을 종합한 결과, 학업 중단에 영향을 미치는 핵심 요인은 **과도한 신청 학점, 재정 안정성, 입학 시 나이, 전공 과정의 특성**으로 일관되게 나타났다. 신청 학점은 높을수록 중단 위험을 증가시키는 가장 강력한 위험 요인이었으며, 반대로 인정 학점, 평균 성적, 성적 변화량 등 실제 성취 지표는 중단을 예방하는 데 가장 중요한 보호 요인으로 작용하였다. 재정 상태 역시 주요 변수로, 채무 여부와 등록금 체납은 위험을 증가시키는 반면 장학금 수혜는 중단을 완화하는 효과가 있었다. 또한 입학 시 나이가 많을수록 학업 외적인 부담이 커져 중단 위험이 높아졌으며, 일부 전공 과정(특히 소규모 학과)의 경우 학습 환경과 지원 체계의 차이로 인해 중단 위험이 상대적으로 높은 것으로 분석되었다.

이러한 분석 결과를 바탕으로 학교 차원의 정책적 개선 방향은 다음과 같이 정리할 수 있다.

첫째, 신청 학점 기반의 조기 개입 체계 구축이 필요하다. 과도한 학점 신청은 학업 부담 증가로 이어져 중단 가능성을 높인다. 따라서 초과 학점 신청자는 자동으로 위험군으로 분류하여 학업 계획 상담, 시간 관리 교육, 개인 맞춤형 학기 설계 지원 등을 제공할 필요가 있다. 또한 평균 학점에 따른 수강 학점에 제한을 두는 방법으로 학습의 부담을 줄이는 것이 좋을 것이다. 반대로 낮은 학점 신청은 학업 참여 부족 또는 동기 저하의 신호일 수 있으므로 일정 기준 이하 신청자 역시 조기 지원이 필요하다.

둘째, 둘째, 성취 기반 학습 지원을 강화해야 한다. 분석 결과, 입학 성적이 낮은 학생과 학기 중 성적이 정체되거나 하락하는 학생에서 중단 위험이 높게 나타났다. 이는 학업 기초 역량의 부족이 다음 학기 학습을 따라가는 데 직접적인 어려움으로 이어지기 때문이다. 따라서 방학 기간을 활용한 기초 학습 보강 프로그램, 사전 학습 과정, 복습 중심의 맞춤형 지도 등을 제공하여 학습 격차를 줄이고, 학기 초 학업 적응을 지원하는 체계가 필요하다. 이러한 예방적 학습 지원은 학생의 학업 자신감을 높이고 중단 가능성을 효과적으로 낮출 수 있을 것이다.

셋째, 재정적 부담 완화 정책의 강화가 필요하다. 등록금 체납과 채무 여부는 중단 위험을 유의하게 증가시키는 요인으로 확인되었으며, 이는 학생의 생활 안정성이 학업 지속성과 직결된다. 따라서 생활비를 지원해주는 장학금 확대, 식비 지원 프로그램 신설 등 학생의 생활 기반을 안정시키는 제도적 장치가 필요하다.

넷째, 전공 적응을 위한 맞춤형 지원 체계 마련이 중요하다. 전공 과정이 위험 요인으로 나타난 것은 일부 학과, 특히 소규모 학과에서 학습 지원 체계 부족이나 교육 만족도 저하가 발생할 가능성을 의미한다. 이에 따라 전공별 기초학습 보완 프로그램 제공, 교수-학생 정기 면담, 전공 만족도 조사 기반 개선 시스템 구축 등이 필요하다. 진로 동기가 약한 학생일수록 중단 위험이 높아지므로, 전공 선택 초기 단계에서 충분한 탐색과 상담 지원이 필수적이다.

다섯째, 연령대별 특성을 고려한 유연한 학사 운영이 필요하다. 분석 결과, 입학 시 나이가 많을수록 중단 위험이 증가하는 경향이 뚜렷했으며 이는 직장, 육아, 가사 등 외적 책임의 영향이 크다는 점을 보여준다. 따라서 성인 학습자를 위한 야간·주말·온라인 강의 확대, 유연한 수강 및 휴학 제도 마련, 성인 학습자 전용 상담센터 운영 등 현실적 제약을 고려한 학사 운영이 필요하다.

종합하면, 본 연구의 분석 결과는 학업 부담 조절, 성취 기반 지원 강화, 재정적 안정성 확보, 전공 적응 지원, 연령 맞춤형 학사 운영이라는 다섯 가지 정책 축을 중심으로 학업 중단을 예방할 수 있음을 시사한다. 이는 대학이 실제 학사 운영과 학생 지원 정책을 설계하는 데 활용할 수 있는 실질적이고 정량적 근거를 제공한다.

5.2 모델 개선 및 추가 연구 제안

본 연구를 통해 성적 정보가 학업 중단 예측에서 핵심적인 역할을 수행한다는 점을 확인하였으나, 향후 연구에서는 학업 중단을 보다 정교하게 설명할 수 있는 다양한 데이터의 확장을 고려할 필요가 있다. 특히 성적과 연관 있는 출석 현황, 과제 제출 여부, LMS 활동 기록 등과 같은 학습 행동 데이터를 추가한다면 예측 성능을 더욱 향상시킬 수 있을 것이다. 더 나아가 학생의 개인적 배경, 심리적 요인, 학교 환경과 같은 비정형 데이터를 포함한 통합 모델을 구축한다면 학업 중단을 보다 다각적으로 이해하는 데 도움이 될 것이다. 또한 본 연구는 정적 데이터에 기반하고 있으므로, 시간에 따른 성취 변화 패턴을 반영할 수 있는 시계열 기반 모델을 적용한다면 실제 교육 현장에서 활용 가능한 보다 실질적인 시사점을 제공할 수 있을 것으로 기대된다.