Predicting Students' Dropout

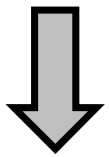
3기 이성식

Introduction

- 학생의 지속적인 학습과 성공은 교육 기관에 있어 매우 중요한 요소

- 학업 중단 위험이 있는 학생을 식별, 도움을 주는 것은 졸업률을 향상시키고 학 업적 성공을 보장하는 데 크게 기여

- 학생의 중도 탈락에 영향을 미칠 수 있는 요인에는 어떤 것들이 있는지 파악하는 것 역시 중요



학생의 다양한 배경 데이터를 이용하여 학생의 중도 탈락 여부를 예측하는 Task를 선정

Dataset

Predict students' dropout and academic success

Investigating the Impact of Social and Economic Factors

Kaggle에서 students' dropout과 관련한 데이터 수집

4424명의 학생 데이터

학생들의 인구통계학적 데이터, 사회·경제적 요인, 학업 성취 정보 등이 포함

이를 통해 학업 중도 탈락 및 성공의 가능성 있는 예측 요인을 분석

주요 변수

학생 정보
- 지원 방식
- 결혼 여부
- 선택한 과정

성적데이터
- 수강·평가·승인된 교과목 및 성적

경제적 요인
- 지역의 실업률
- 인플레이션율
- GDP

타겟 변수
- 중도 포기 유무

https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data

Preprocessing & Feature engineering

총 34개의 Columns로 이루어진 데이터

Column을 범주형 데이터, 수치형 데이터로 나누어 분석 및 EDA 과정을 진행

- 범주형 데이터 : 17개
- 수치형 데이터 : 17개
- 4424개의 데이터 중, 결측값 X

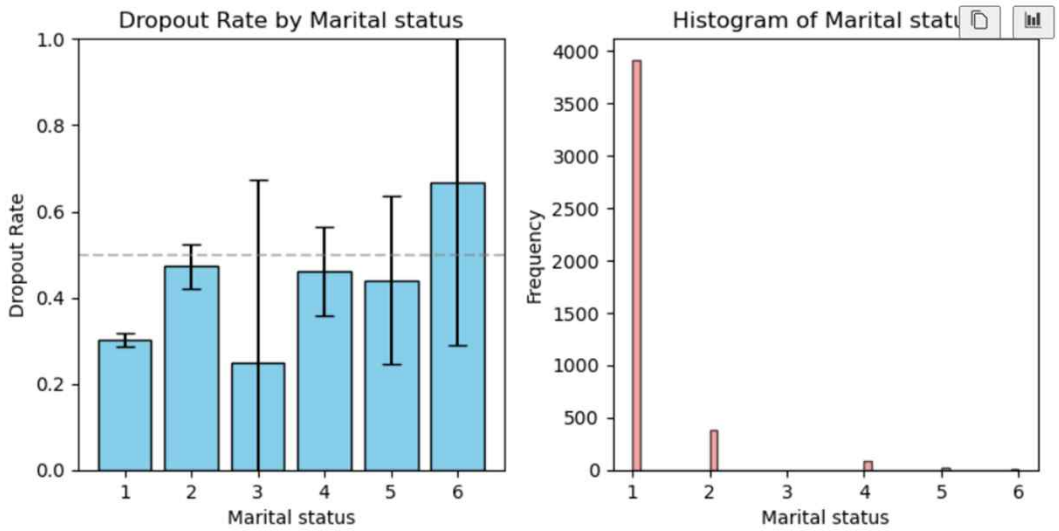
```
['Marital status',
 'Application mode',
 'Course',
 'Daytime/evening attendance',
 'Previous qualification',
 'Nacionality',
 "Mother's qualification",
 "Father's qualification",
 "Mother's occupation",
 "Father's occupation",
 'Displaced',
 'Educational special needs',
 'Debtor',
 'Tuition fees up to date',
 'Gender',
 'Scholarship holder',
 'International']
```

```
['Application order',
 'Age at enrollment',
 'Curricular units 1st sem (credited)',
 'Curricular units 1st sem (enrolled)',
 'Curricular units 1st sem (evaluations)',
 'Curricular units 1st sem (approved)',
 'Curricular units 1st sem (grade)',
 'Curricular units 1st sem (without evaluations)',
 'Curricular units 2nd sem (credited)',
 'Curricular units 2nd sem (enrolled)',
 'Curricular units 2nd sem (evaluations)',
 'Curricular units 2nd sem (approved)',
 'Curricular units 2nd sem (grade)',
 'Curricular units 2nd sem (without evaluations)',
 'Unemployment rate',
 'Inflation rate',
 'GDP']
```

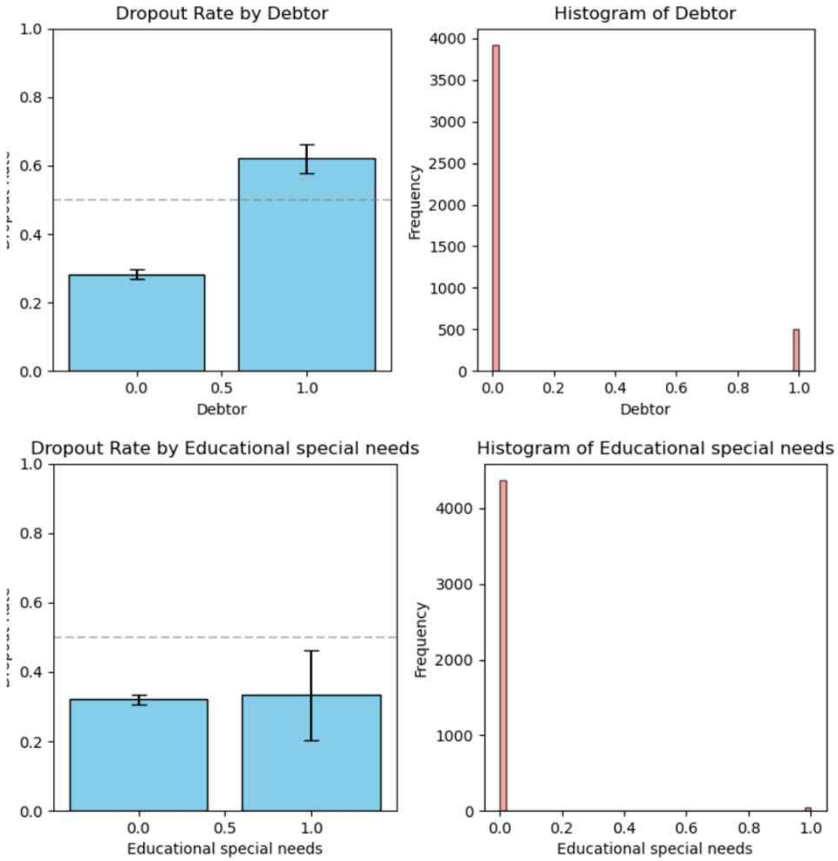
EDA – Categorical Data

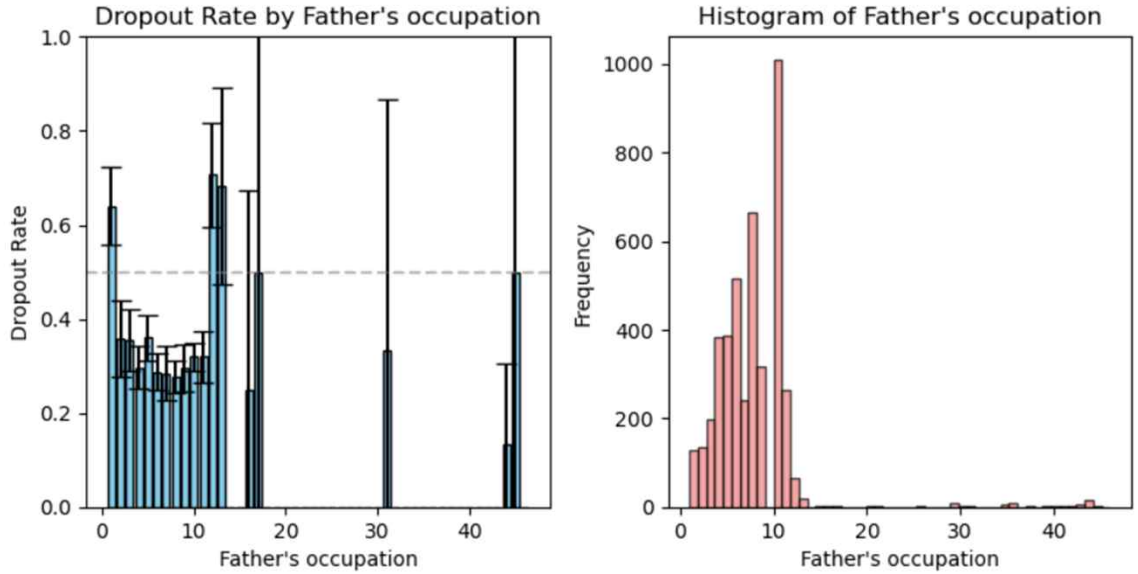
각 범주형 데이터에 대하여 범주마다 Dropout의 비율을 신뢰도 95% 신뢰구간과 함께 시각화를 진행

각 범주 마다 신뢰구간이 겹치지 않을수록 타겟 변수를 설명하는 능력이 높다고 판단



EDA – Categorical Data





EDA – Categorical Data

각 범주마다 타겟의 비율의 신뢰구간을 시각화한 결과,
총 17개의 범주형 데이터 중 9개의 데이터가 통계적으로 유의미하게 타겟 변수를 설명할 수 있다고 판단

'Application mode'

'Course'

'Daytime/evening attendance'

'Displaced'

'Educational special needs'

'Debtor'

'Tuition fees up to date'

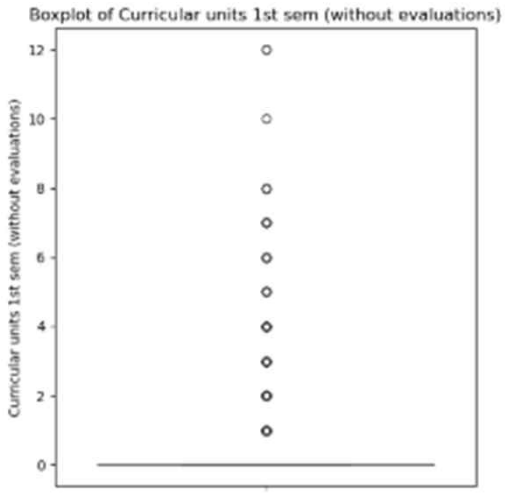
'Gender'

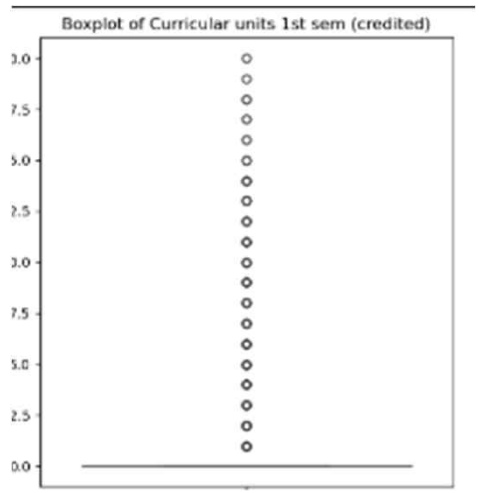
'Scholarship holder'

EDA – Numerical Data

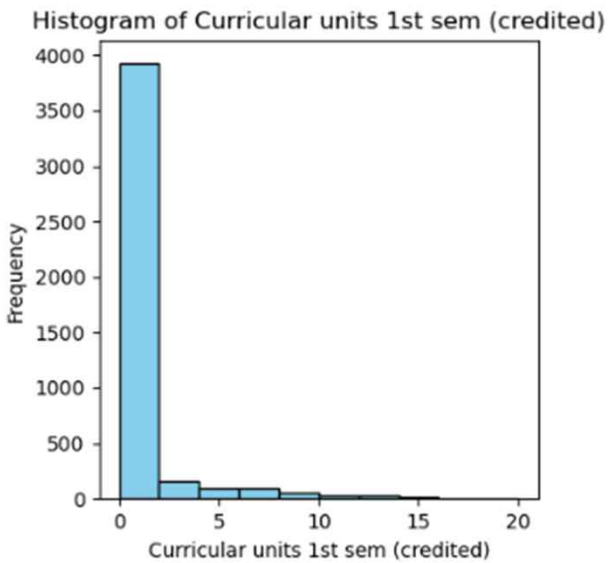
수치형 데이터의 경우, 전처리를 위해 이상치 처리를 하기 위해 Box plot 시각화를 진행

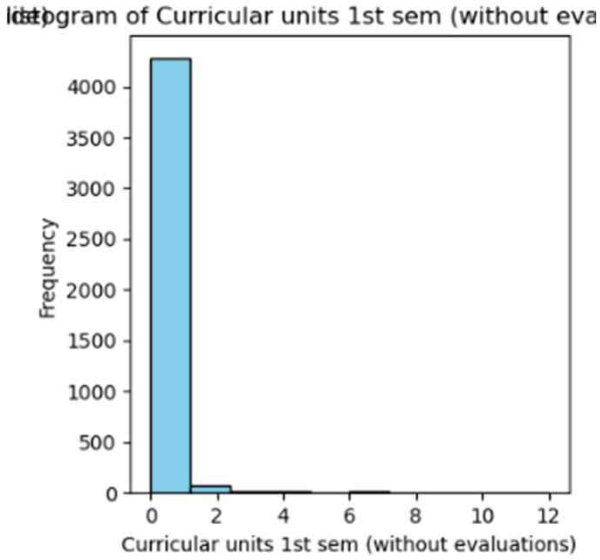
네 변수
'Curricular units 1st sem (credited)', 'Curricular units 2nd sem (credited)',
'Curricular units 1st sem (without evaluations)', 'Curricular units 2nd sem (without evaluations)'
에 대하여 아래와 같은 Box plot이 관찰됨.





EDA – Numerical Data

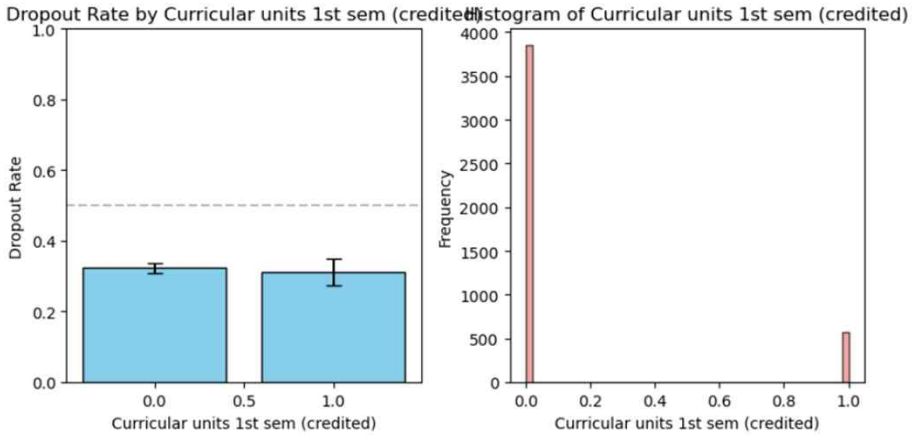


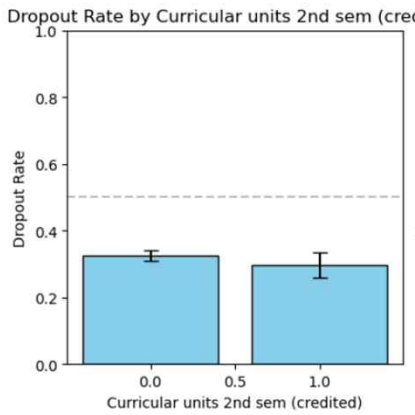


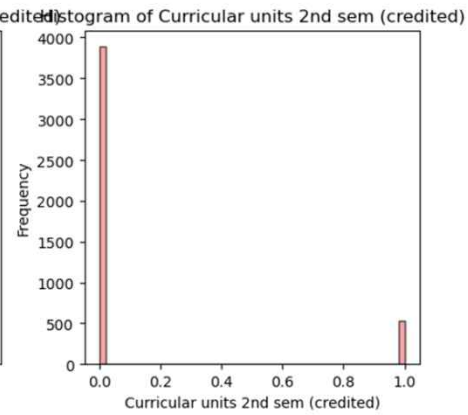
수치형 데이터이지만 분포가 편향된 경우, 이상치 처리를 어떻게 할 것인가?

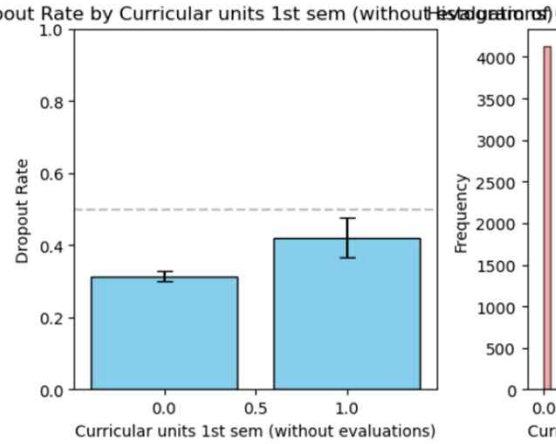
수치형 데이터이지만 대부분 0에 몰려 있기 때문에, 0이냐 아니냐로 범주형으로 만들어버리자.

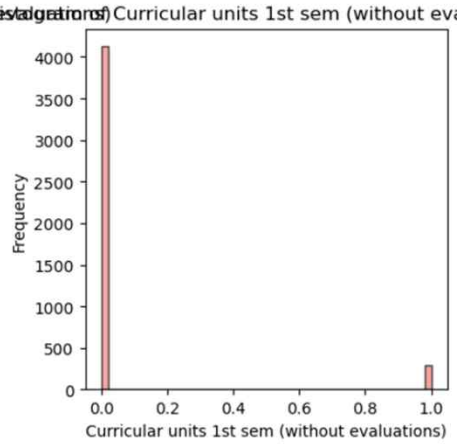
EDA – Numerical Data

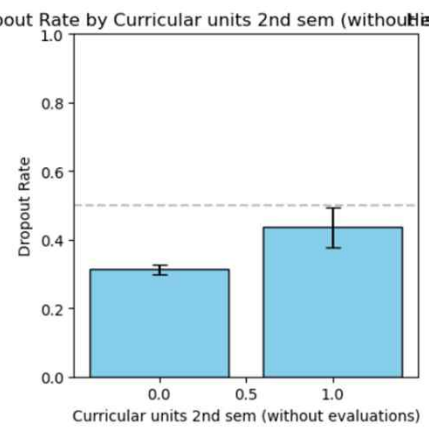


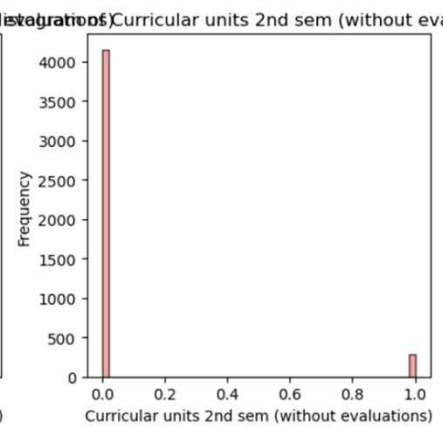












EDA – Numerical Data

'Curricular units 1st sem (credited)', 'Curricular units 2nd sem (credited)'

-타겟 변수에 대한 충분한 설명력을 가지지 못한다고 판단하여 column 제거를 진행

'Curricular units 1st sem (without evaluations)', 'Curricular units 2nd sem (without evaluations)'

-범주화를 진행하여 새로운 column으로 추가

-이상치로 간주되어 사라졌을 데이터가 타겟 변수에 대한 설명력이 높은 변수가 될 수 있다는 사실을 알 수 있음

EDA – Numerical Data

상관계수 heatmap에서 알 수 있듯이,

Curricular units 1st sem(enrolled) & Curricular units 2nd sem (enrolled)

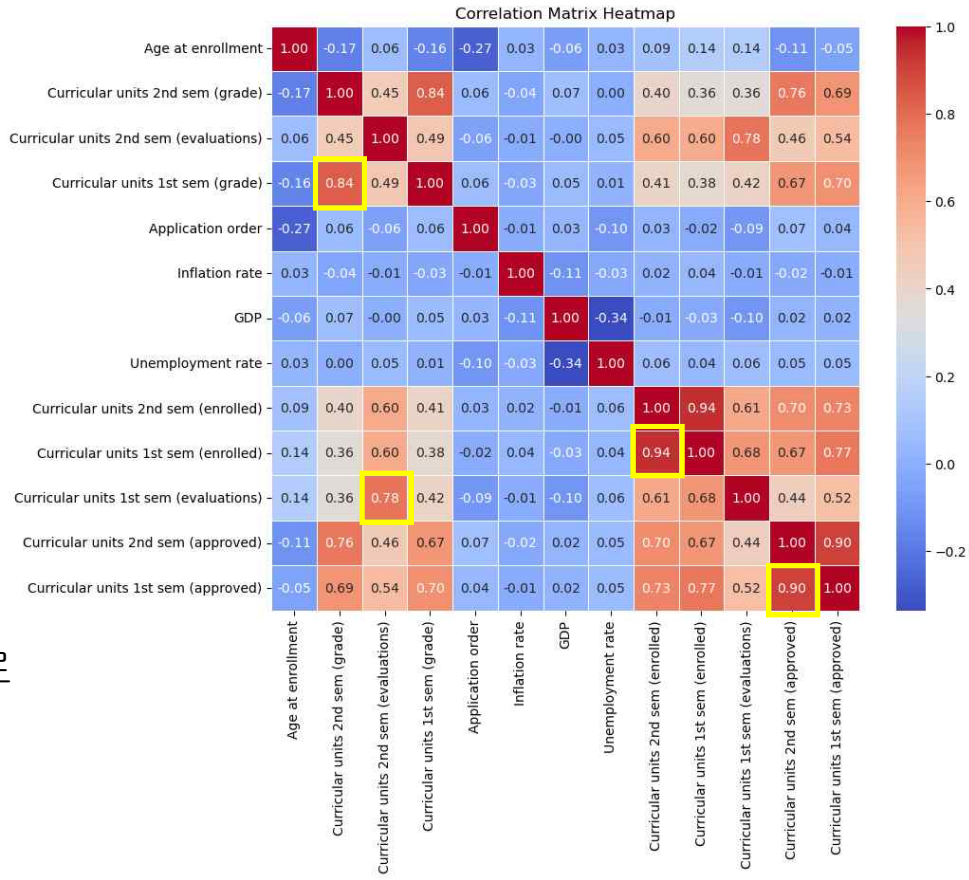
Curricular units 1st sem(approved) & Curricular units 2nd sem (approved)

Curricular units 1st sem(evaluations) & Curricular units 2nd sem (evaluations)

Curricular units 1st sem(grade) & Curricular units 2nd sem (grade)

간의 상관계수가 각각 0.94, 0.90, 0.78, 0.84로 상당히 높음

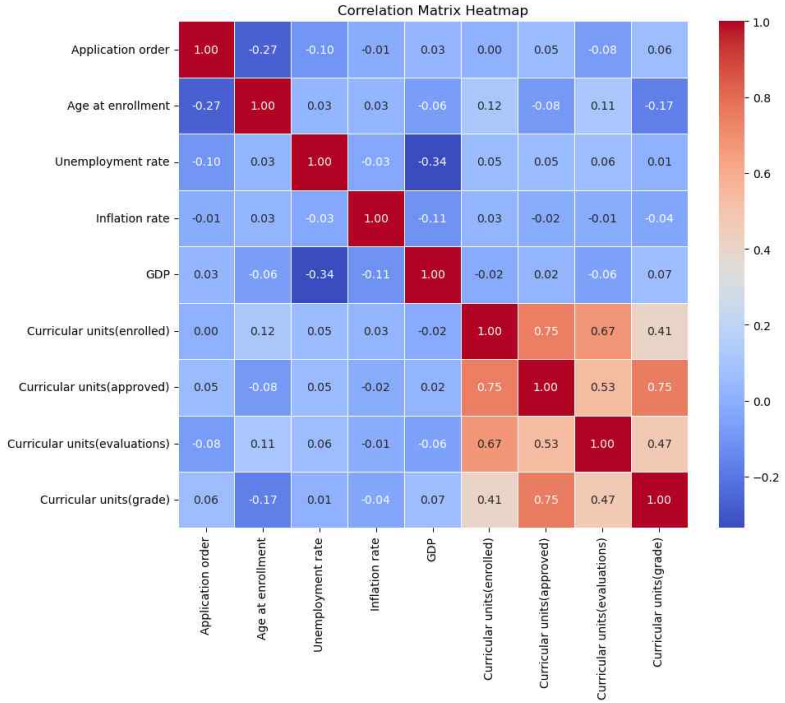
각 변수는 독립된 기간에 대한 데이터이므로 둘 중 한 feature를 제거하는 것보다는, 1년 동안의 지표라는 의미로 두 feature의 평균을 새로운 feature로 사용



EDA – Numerical Data

결과적으로 총 17개의 수치형 데이터를 총 9개의 수치형 데이터로 축소

여전히 상관계수가 0.75, 0.67 로 강한 상관관계를 가지는 변수들이 존재



Results

범주형 데이터

- 신뢰구간을 통해 타겟 변수를 잘 설명할 수 있는 정도에 따라 Feature Selection을 진행
- 총 17개의 feature 중 9개의 유의미한 feature 선택

수치형 데이터

- 분포가 편향된 데이터에 대한 이상치 처리를 새로운 파생변수 생성으로 해결
- 상관계수를 기반으로 강한 상관관계를 가지는 두 변수를 하나의 변수로 병합
- 17개의 feature에서 9개의 수치형 변수를 제작

Discussion

1. 수치형 변수에 대하여 타겟 변수를 설명하는 능력을 어떻게 측정할 것인가?

-수치형 변수에 대해서는 타겟 변수와의 관계를 직접적으로 파악하는 과정이 없었기 때문에, 설명력이 부족한 변수가 있을 수 있다.

2. 피처 엔지니어링 기법의 타당성은 어떻게 측정할 것인가?

-모델 학습 후에 만족스럽지 않으면 다시 피처 엔지니어링 파트로 돌아와서 수정 후 다시 모델 평가 이를 반복할 것인가?

3. 조금 더 많은 파생 변수를 만드는 것을 목표로 했으나, 생각만큼 잘 되지 않았음.

-도메인에 대한 지식이 부족하다는 것을 느낄 수 있었음. 이 부분은 모델 학습까지 이어져야 아이디어가 나올 듯함

4. 범주형 데이터에 대한 인코딩

-세션에서 다룬 target encoding 기법을 고려해볼 수 있을 것 같음

Finish

데이터를 이처럼 시각화하고 분석하는 경험은 처음

책에서 배운 내용을 직접 적용해보니, 처음 보는 데이터를 어떻게 이해할지에 대한 감이 잡히는 듯함

보다 엄밀하고 합리적인 분석을 위해서는 통계 공부를 더 열심히 해야 할 듯함

그래서 수리통계 같이 스터디 하실 분..?