2024.07.10

학생 학업 중단 예측 모델 구현

빅데이터 8기 | 김솔미, 방희수, 이수민

00.

목차

00.

타임스탬프

7월 5일	7월 6-7일	7월 8일	7월 9일	7월 10일
• 데이터 세트 찾기	• 데이터 전처리 • 초기모델 훈련	• 특성중요도 확인 • 하이퍼파라미터 튜닝	• 손실 함수 • 하이퍼파라미터 튜닝	• 최종 모델 선택 • 시각화 • PPT 제작

00.

역할

김솔미	학습 데이터셋 선정 GradientBoost 모델 구축 및 하이퍼 패러미터 튜닝 손실함수 추적
방희수	프로젝트 목표 설정 프로젝트 결과 분석 XGBoost 모델 구축 및 하이퍼 패러미터 튜닝
이수민	데이터 전처리 RandomForest 모델 구축 및 하이퍼 패러미터 튜닝 특성 중요도 확인 및 시각화 자료 제작

01.

사용한 데이터

- 학생 등록 시 알려진 정보가 포함된 데이터 세트

이 데이터에는 교육 기관에 등록한 학생과 관련된 다양한 인구 통계학적, 사회 경제적 및 학업 성취 요인이 포함

학업 경로의 초기 단계에서 위험에 처한 학생을 식별하여 이를 지원하는 전략을 마련할 수 있도록 함으로써
학업 중퇴 및 실패를 줄이는 데 기여하는 것을 목표로 만들어진 데이터

- 목적

고등교육의 학업 중단 혹은 졸업하는 학생들의 여러 특성들을 분석하여
학업을 마치는데 위험이 생긴 학생을 미리 식별한다.

01.

사용한 데이터

Input	
Marital status	학생의 결혼 상태
Application mode	학교에 지원한 방식
Application order	지원 순서
Course	수강하는 과정
Daytime/evening attendance	주간/야간 수업 참석 여부
Previous qualification	이전 학력
Previous qualification (grade)	이전 학력의 성적
Nacionality	국적
Mother's qualification	어머니의 학력
Father's qualification	아버지의 학력
Curricular units 2nd sem (credited)	2학기 이수 학점
Curricular units 2nd sem (enrolled)	2학기 등록 학점
Curricular units 2nd sem (evaluations)	2학기 평가 받은 학점
Curricular units 2nd sem (approved)	2학기 승인된 학점
Curricular units 2nd sem (grade)	2학기 성적
Curricular units 2nd sem (without evaluations)	평가 없이 이수한 2학기 학점
Unemployment rate	실업률
Inflation rate	인플레이션율
GDP	국내총생산

Input	
Mother's occupation	어머니의 직업
Father's occupation	아버지의 직업
Admission grade	입학 성적
Displaced	이주 여부
Educational special needs	특수 교육 필요 여부
Debtor	채무 여부
Tuition fees up to date	등록금 납부 여부
Gender	성별
Scholarship holder	장학금 수여 여부
Age at enrollment	입학 나이
International	국제 학생 여부
Curricular units 1st sem (credited)	1학기 이수 학점
Curricular units 1st sem (enrolled)	1학기 등록 학점
Curricular units 1st sem (evaluations)	1학기 평가 받은 학점
Curricular units 1st sem (approved)	1학기 승인된 학점
Curricular units 1st sem (grade)	1학기 성적
Curricular units 1st sem (without evaluations)	평가 없이 이수한 1학기 학점

Target	
Dropout	학교를 중퇴한 학생
Graduate	졸업생

02.

데이터 전처리

재학중인 학생 데이터 삭제

- 중퇴, 졸업 학생들을 분류하는 목적이기 때문에 삭제처리

라벨 인코딩

- 머신러닝 모델에 적용시키기 위해 타겟 데이터 라벨 인코딩

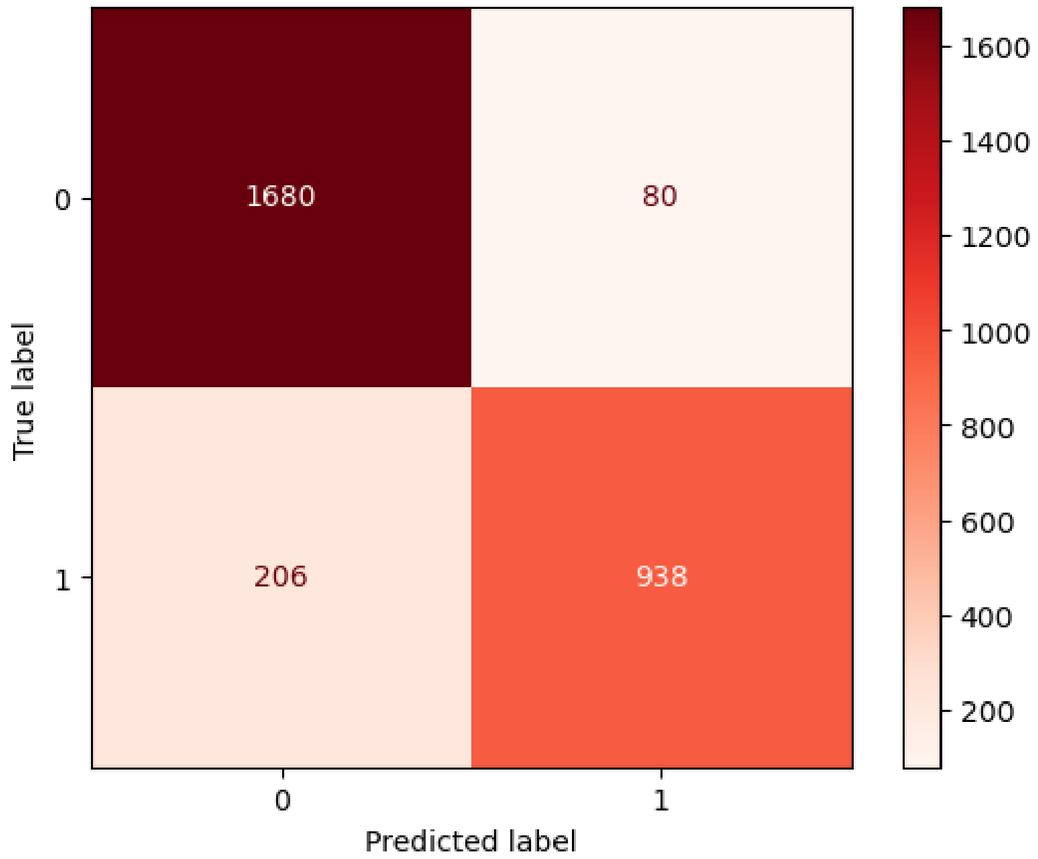
데이터 스케일링

- 모델 성능 향상을 위해 데이터 스케일링

03.

초기 모델

1. RandomForestClassifier 모델

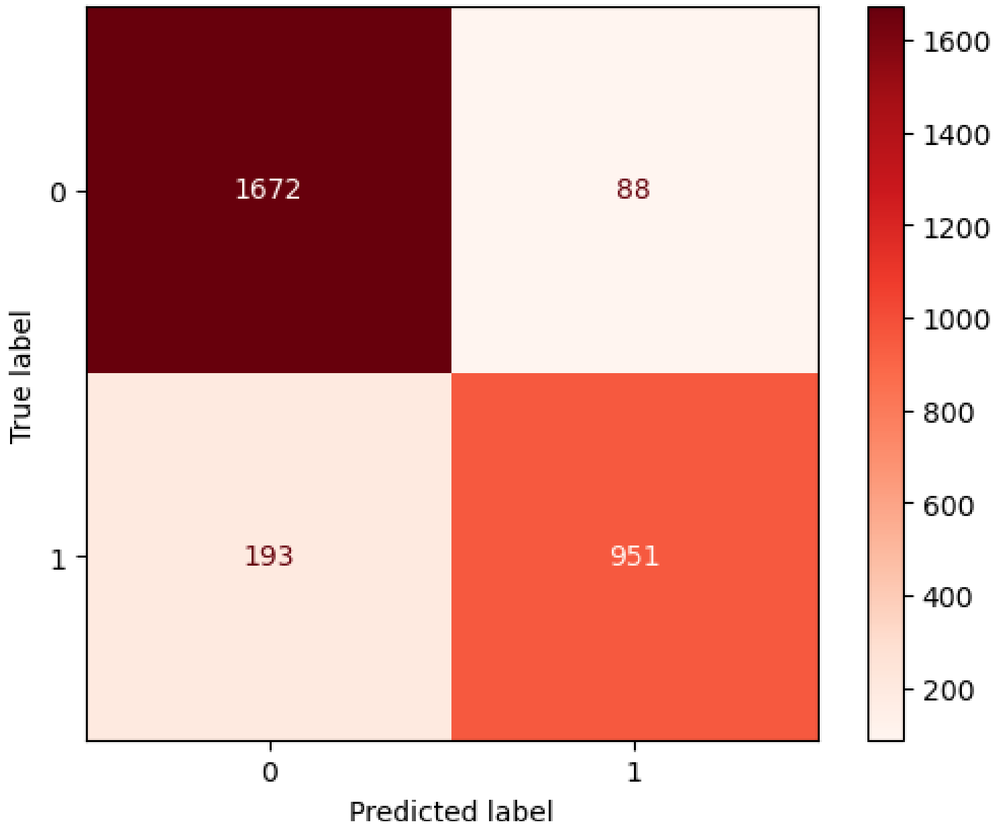


Internal	1.0
CV 평균	0.9015
Log Loss	0.2875

03.

초기 모델

2. GradientBoostingClassifier 모델

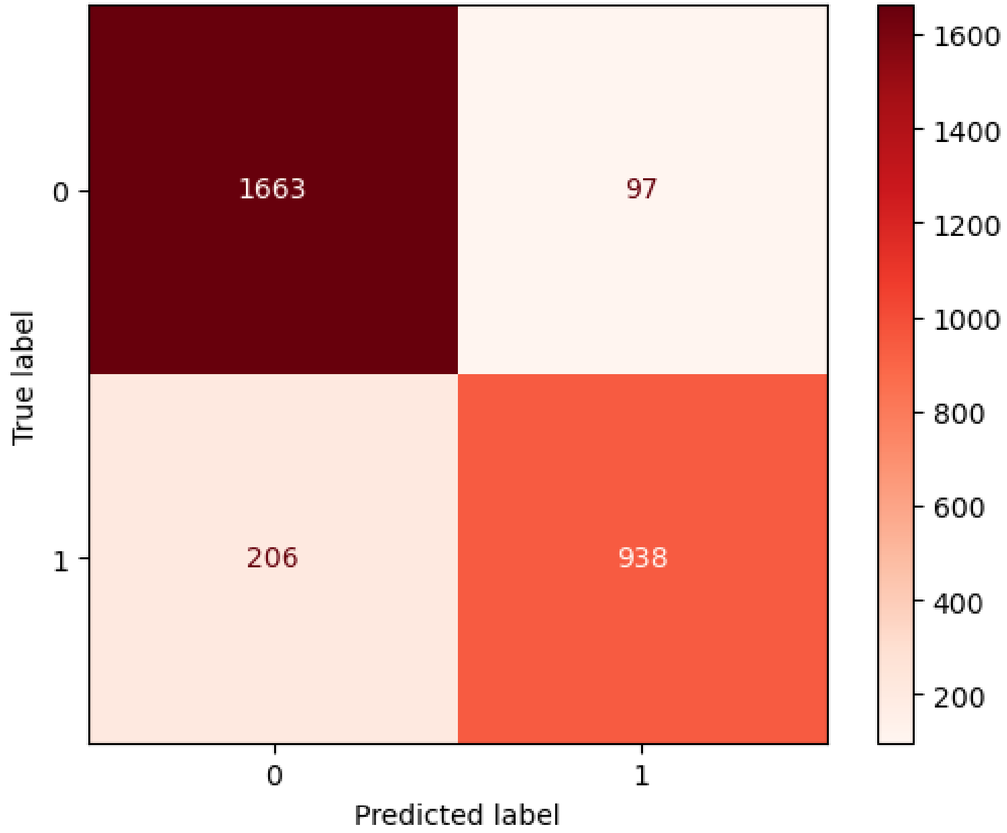


Internal	0.9443
CV 평균	0.9032
Log Loss	0.2517

03.

초기 모델

3. XGBClassifier 모델



Internal	1.0
CV 평균	0.8597
Log Loss	0.3287

03.

초기 모델

- 3개의 초기 모델 비교

	Internal	CV_score
RandomForest	1.0	0.9015
Gradient Boosting	0.9443	0.9032
XGBoost	1.0	0.8957

04.

하이퍼파라미터 튜닝

하이퍼파라미터		특징
n_estimators	결정 트리의 개수	개수가 많을수록 성능이 좋아질 수 있지만, 과적합 가능성, 계산 시간 ↑
max_depth	결정 트리의 최대 깊이	깊어질수록 복잡한 패턴 학습할 수 있지만, 과적합 가능성 ↑
min_samples_split	노드를 분할하기 위한 최소 샘플 수	값이 클수록 과적합을 방지할 수 있지만, 모델의 복잡도 ↓
learning_rate	학습률	값이 클수록 빠르게 합습할 수 있지만, 과적합 가능성 ↑
subsample	훈련 데이터 샘플링 비율	과정합 방지 위해 사용 너무 높은 값은 과적합 위험 ↑ 너무 낮은 값은 충분한 학습 ↓
colsample_bytree	각 트리마다 사용하는 feature 비율	모델의 다양성을 높이며 과적합을 방지
gamma	가지치기를 결정하기 위한 최소 손실 감소 값	값이 클수록 복잡도 ↓

04

하이퍼파라미터 튜닝

1. RandomForestClassifier 모델

	n_estimators	max_depth	min_samples_split	Best training score	Best validation score
1	[100, 200, 300]	[5, 10, 15, 20, 25]		0.9660	0.9067
2	[250, 300, 330, 350]	[5, 10, 15, 17, 20]	[2, 4, 6, 8, 10]	0.9628	0.9091
3	[330, 350, 360, 370]	[8, 9, 10, 11, 12, 13]	[2, 3, 4, 5, 6]	0.9630	0.9094
4	[360, 370, 380, 390, 400]	[8, 9, 10, 11, 12, 13]	[2, 3, 4, 5, 6]	0.9630	0.9094

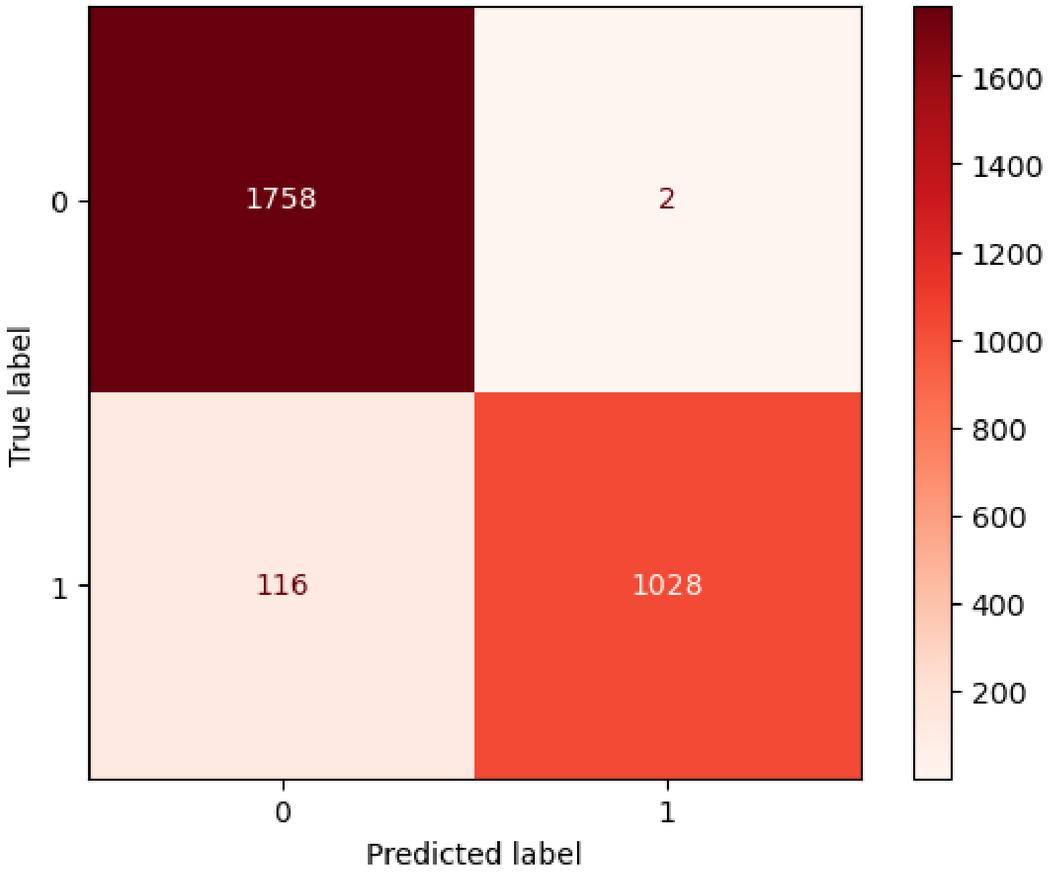
04

하이퍼파라미터 튜닝

1. RandomForestClassifier 모델

- Best parameters for rfc
  : 'max_depth': 10,
  'min_samples_split': 4,
  'n_estimators': 370

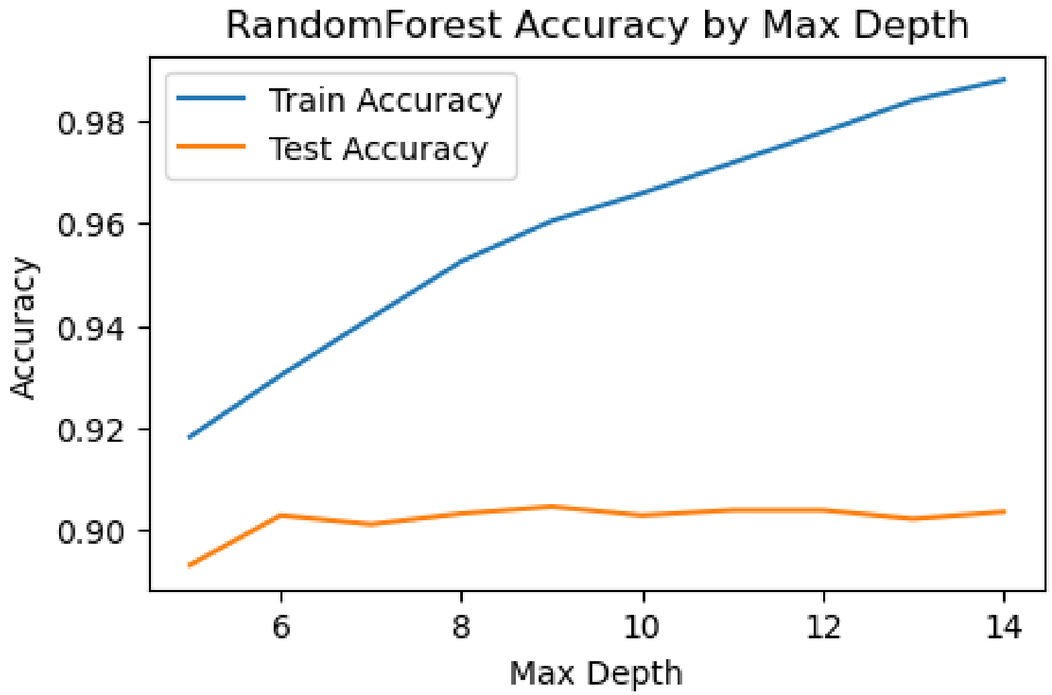
- Best training score for rfc: 0.9630
- Best validation score for rfc: 0.9094
- Log Loss for training set: 0.1476



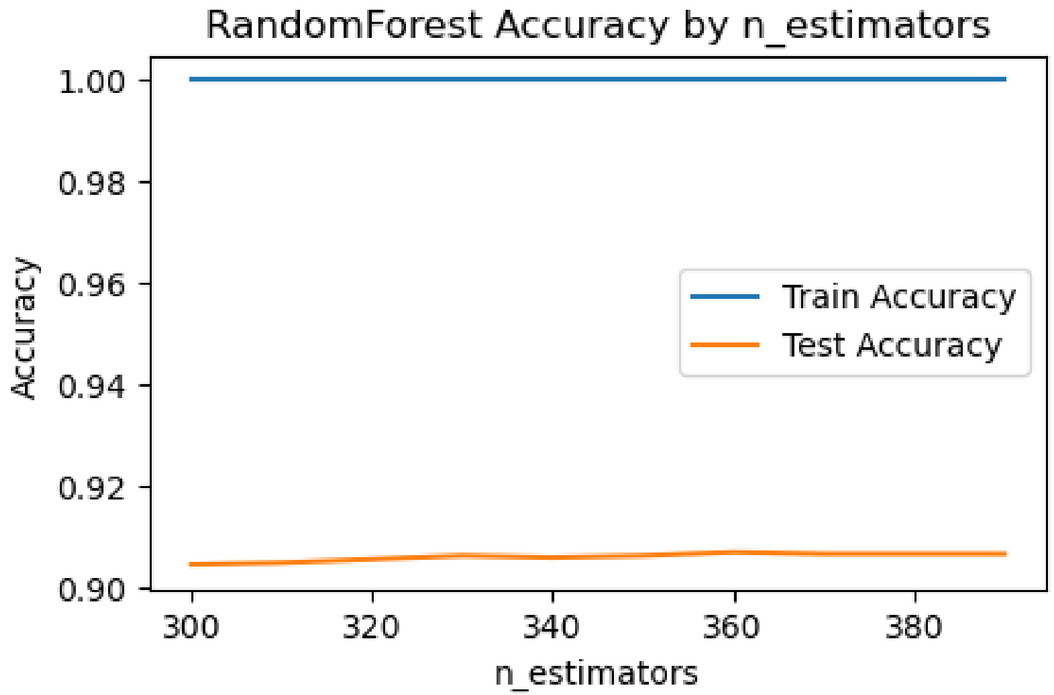
하이퍼파라미터 튜닝

1. RandomForestClassifier 모델

- 하이퍼파라미터에 따른 변화



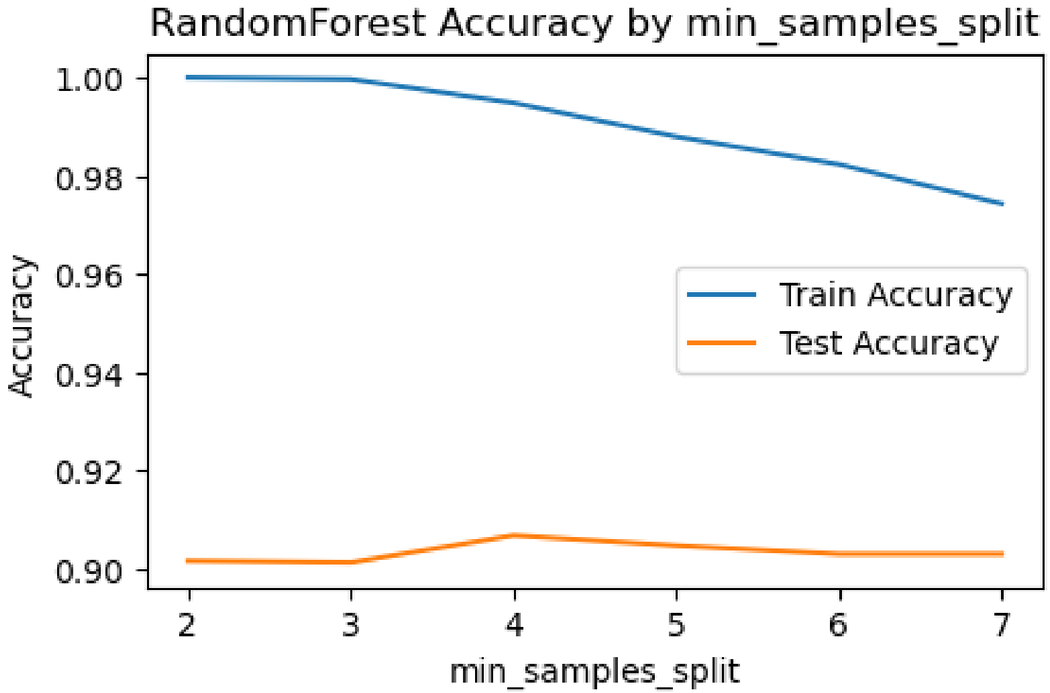
< Max_Depth >



< n_estimators >

하이퍼파라미터 튜닝

- 하이퍼파라미터에 따른 변화



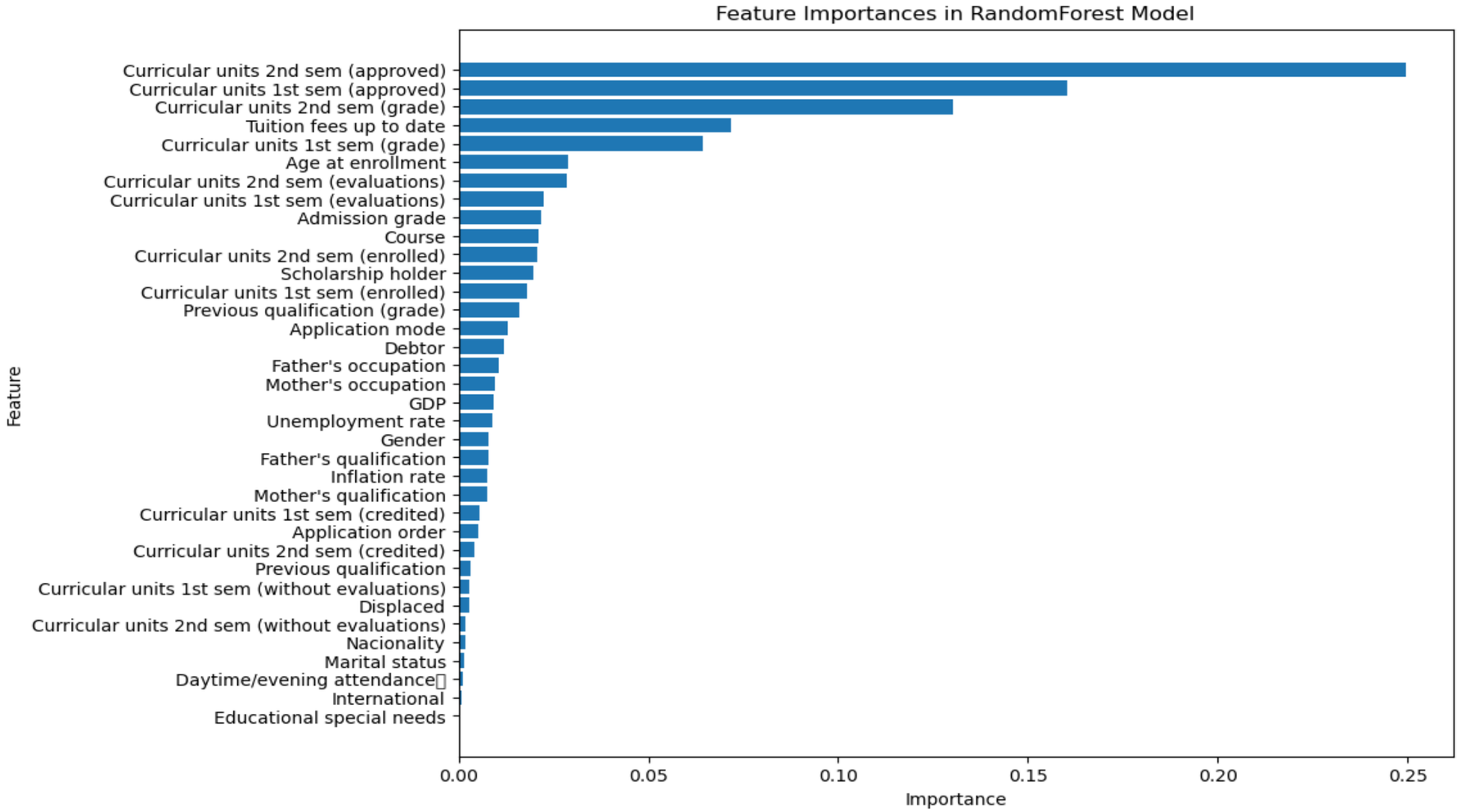
< min_samples_split >

04

하이퍼파라미터 튜닝

1. RandomForestClassifier 모델

- 특성 중요도



04

하이퍼파라미터 튜닝

1. RandomForestClassifier 모델

- 특성 중요도

Feature	Importance
Curricular units 2nd sem (approved)	0.249401
Curricular units 1st sem (approved)	0.160454
Curricular units 2nd sem (grade)	0.130415
Tuition fees up to date	0.071852
Curricular units 1st sem (grade)	0.064391
...	
Daytime/evening attendance₩t	0.001313
International	0.000811
Educational special needs	0.000540

하이퍼파라미터 튜닝

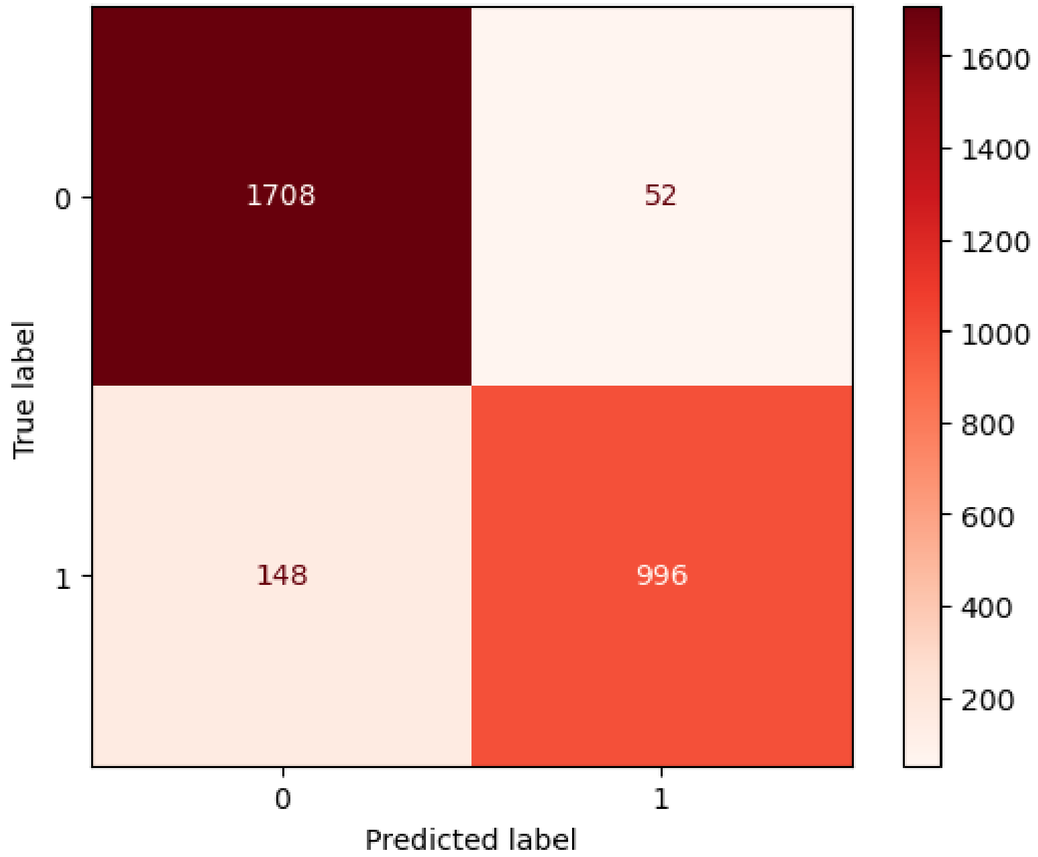
2. GradientBoosting Classifier 모델

	n_estimators	learning_rate	max_depth	Best training score	Best validation score
1	[100, 200, 300, 400]	[0.05, 0.1, 0.2, 0.3]	[3, 5, 7, 9]	0.9888	0.9039
2	[350, 400, 450, 500]	[0.05, 0.1, 0.15]	[1, 2, 3, 4]	0.9351	0.9118
3	[380, 400, 420]	[0.01, 0.03, 0.05, 0.07]	[2, 3, 4, 5]	0.9371	0.9125
4	[400, 420, 430, 440, 450]	[0.04, 0.05, 0.06, 0.07]	[2, 3, 4, 5]	0.9371	0.9125

하이퍼파라미터 튜닝

2. GradientBoosting Classifier 모델

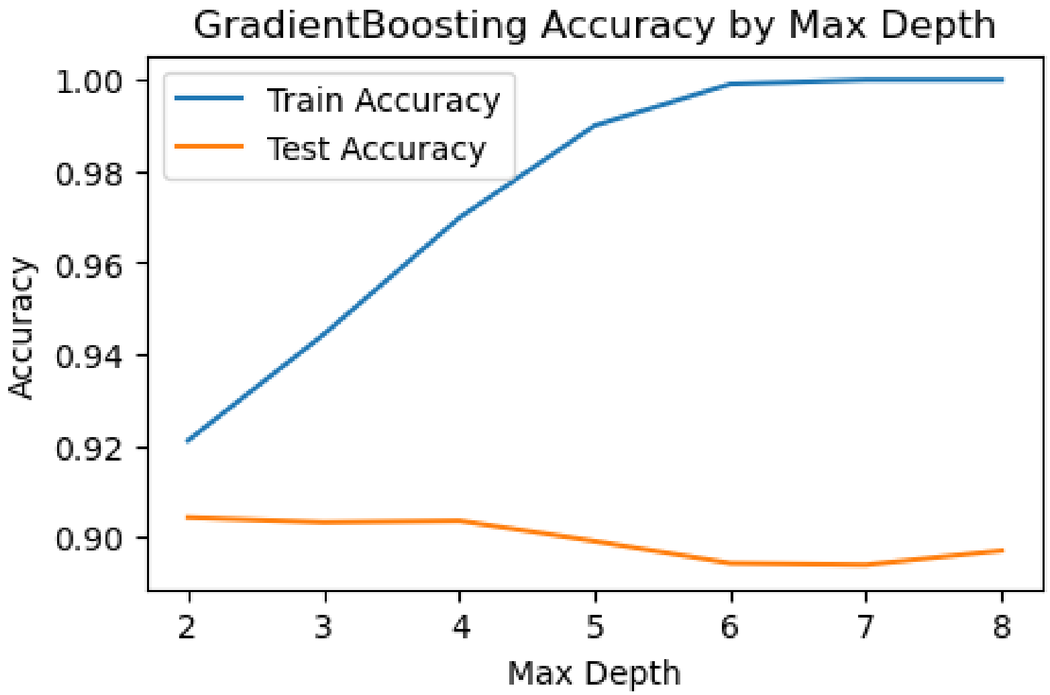
- Best parameters for gbc
  : 'learning_rate': 0.05,
  'max_depth': 2,
  'n_estimators': 420
- Best training score for gbc: 0.9371
- Best validation score for gbc: 0.9125



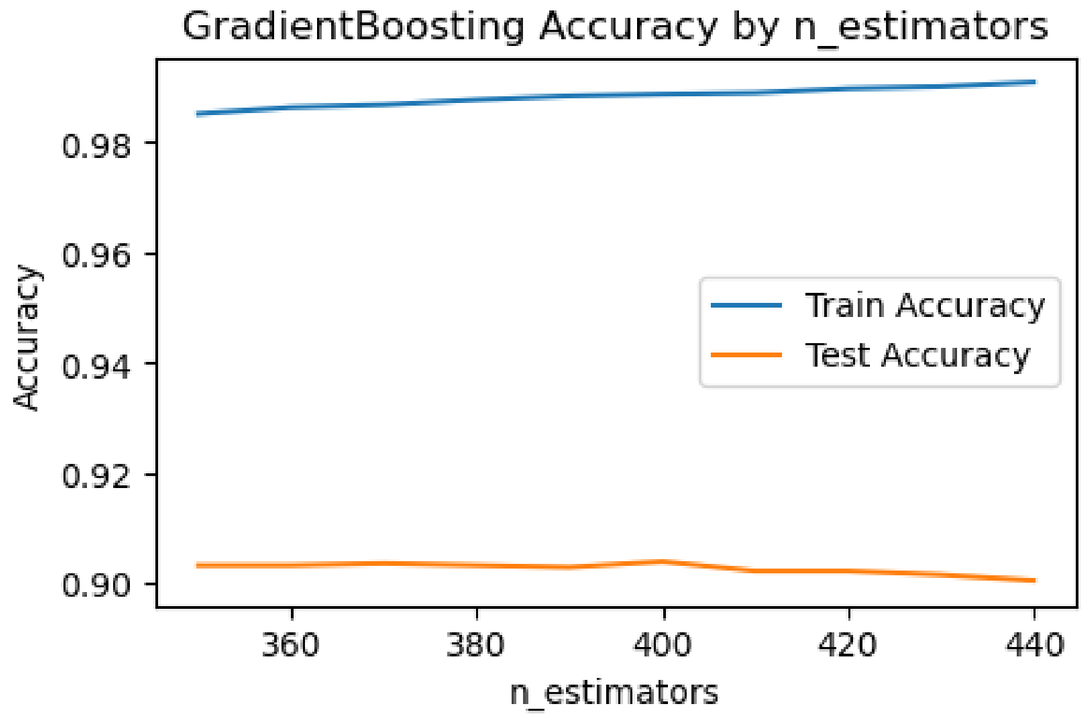
하이퍼파라미터 튜닝

2. GradientBoosting Classifier 모델

- 하이퍼파라미터에 따른 변화



< Max_Depth >

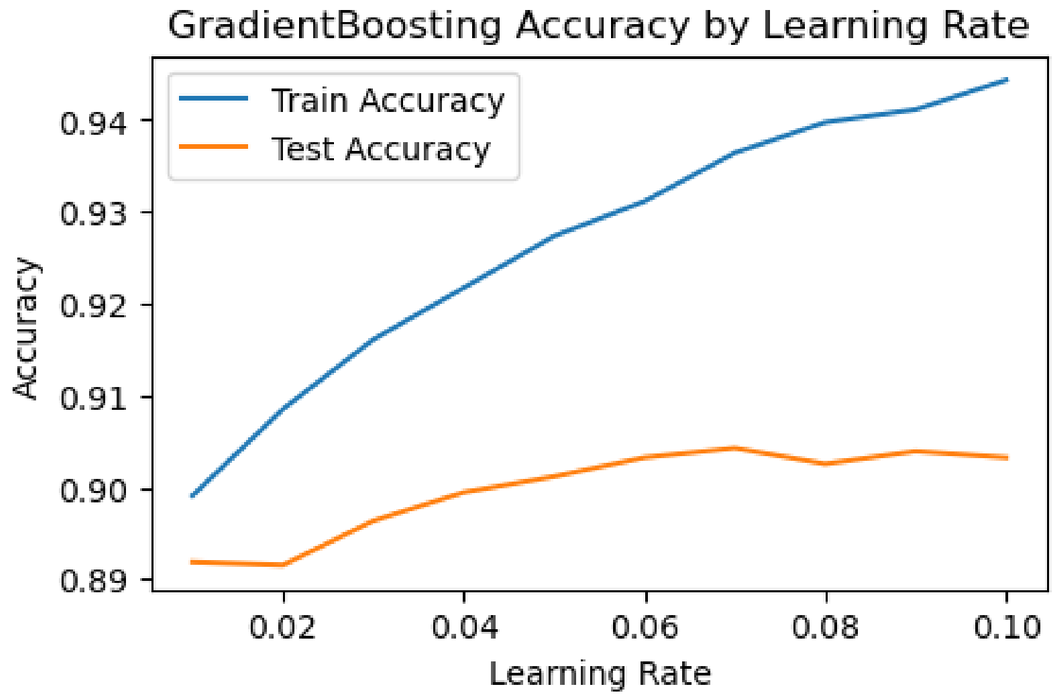


< n_estimators >

하이퍼파라미터 튜닝

2. GradientBoosting Classifier 모델

- 하이퍼파라미터에 따른 변화



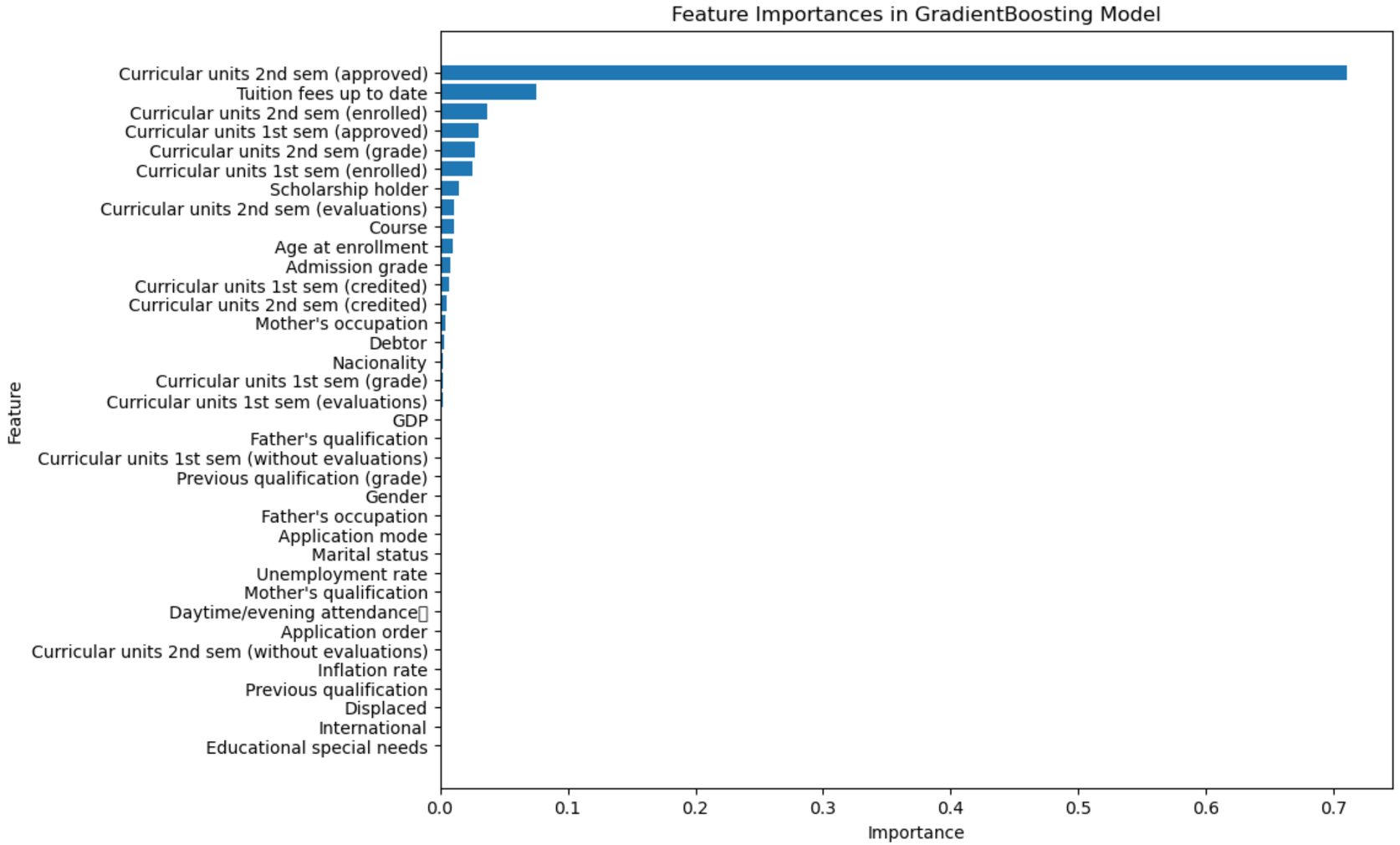
< Learning Rate >

04

하이퍼파라미터 튜닝

2. GradientBoosting Classifier 모델

- 특성 중요도



04

하이퍼파라미터 튜닝

2. GradientBoosting Classifier 모델

- 특성 중요도

Feature	Importance
Curricular units 2nd sem (approved)	0.709796
Tuition fees up to date	0.075397
Curricular units 2nd sem (enrolled)	0.037061
Curricular units 1st sem (approved)	0.030300
Curricular units 2nd sem (grade)	0.027750
...	
Displaced	0.000162
International	0.000059
Educational special needs	0.000000

하이퍼파라미터 튜닝

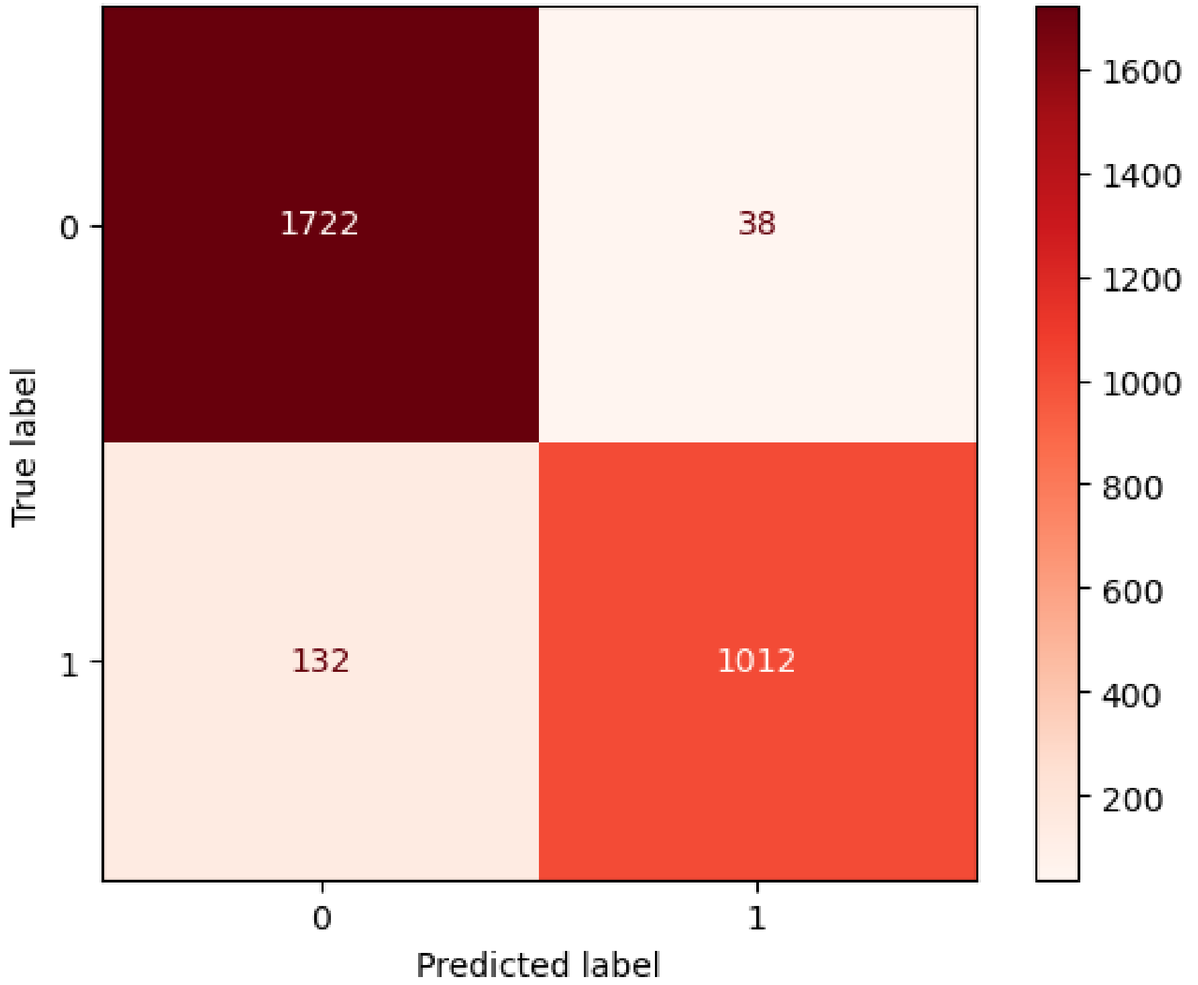
3. XGBoostClassifier 모델

	1	2	3	4	5
n_estimators	[50, 100, 200]	[150, 200, 250]	[190, 200, 210]	[195, 200, 205]	[198, 200, 202]
max_depth	[3, 5, 7]	[3, 4, 5]	[3, 5]	[2, 3, 4]	[3]
learning rate	[0.1, 0.01, 0.05]	[0.03, 0.05, 0.07]	[0.1, 0.05, 0.07, 0.09]	[0.06, 0.07, 0.08]	[0.07]
subsample	[0.7, 0.8, 0.9]	[0.7, 0.8, 0.9]	[0.6, 0.7, 0.8]	[0.6, 0.7, 0.8]	[0.7]
colsample_bytree	[0.7, 0.8, 0.9]	[0.6, 0.7, 0.8]	[0.6, 0.7]	[0.6, 0.7]	[0.6]
gamma	[0, 1, 5]	[0, 1, 3]	[0, 1, 2]	[0, 1, 2]	[1]
Best training	0.9369	0.9466	0.9466	0.9466	0.9461
Best validation	0.9112	0.9125	0.9125	0.9125	0.9129

하이퍼파라미터 튜닝

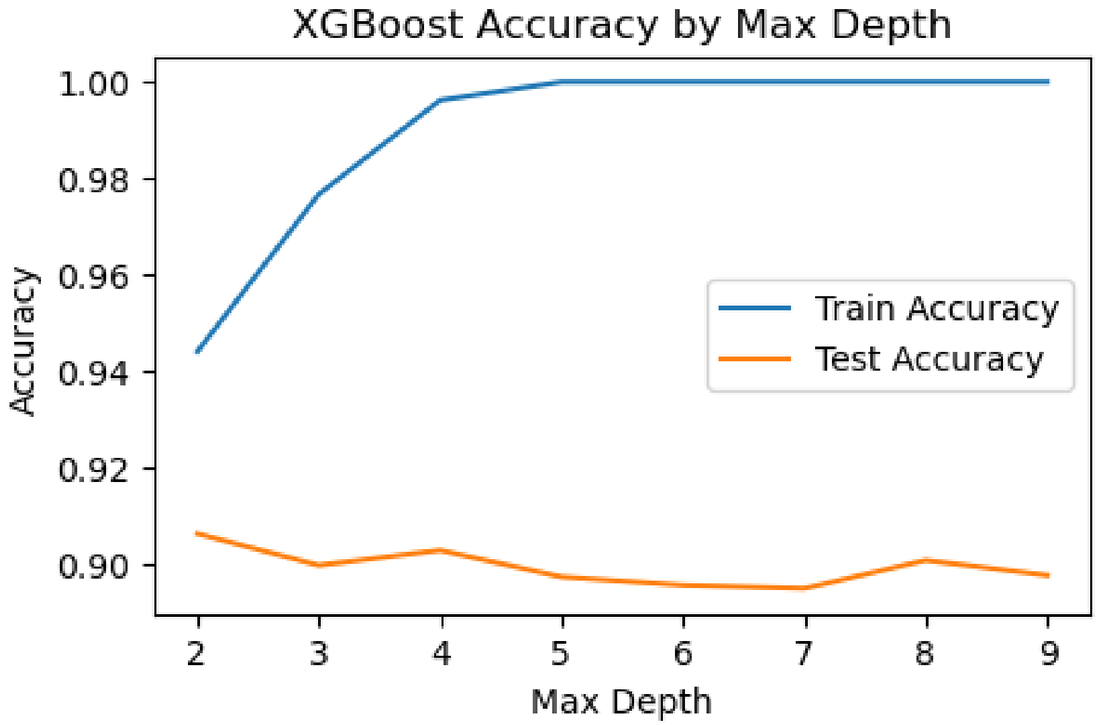
3. XGBoostClassifier 모델

- Best parameters for xgb
  : 'colsample_bytree': 0.6,
  'gamma': 1,
  'learning_rate': 0.07,
  'max_depth': 3,
  'n_estimators': 198,
  'subsample': 0.7
- Best training score for xgb: 0.9461
- Best validation score for xgb: 0.9129

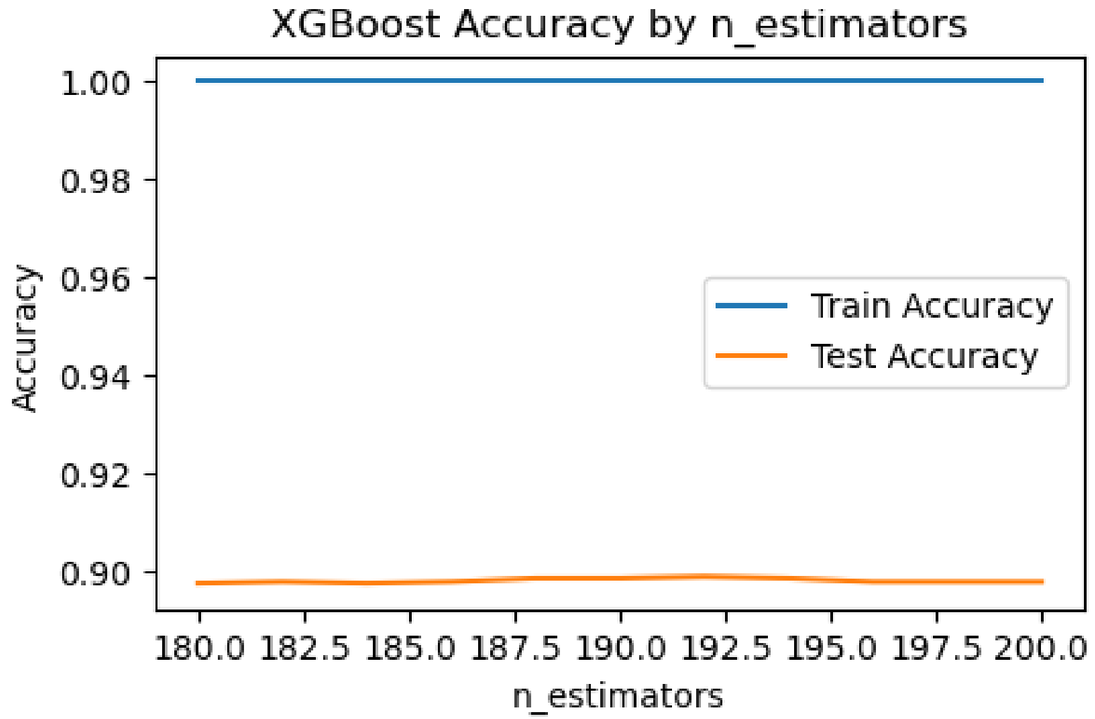


하이퍼파라미터 튜닝

- 하이퍼파라미터에 따른 변화



< Max_Depth >

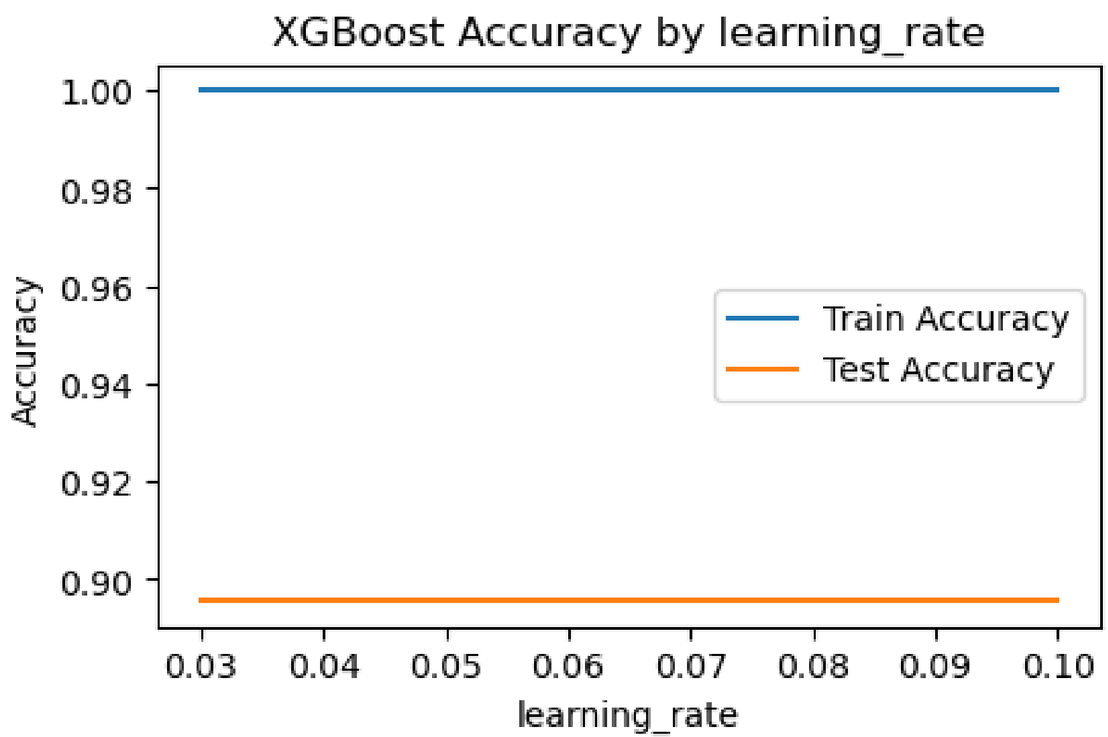


< n_estimators >

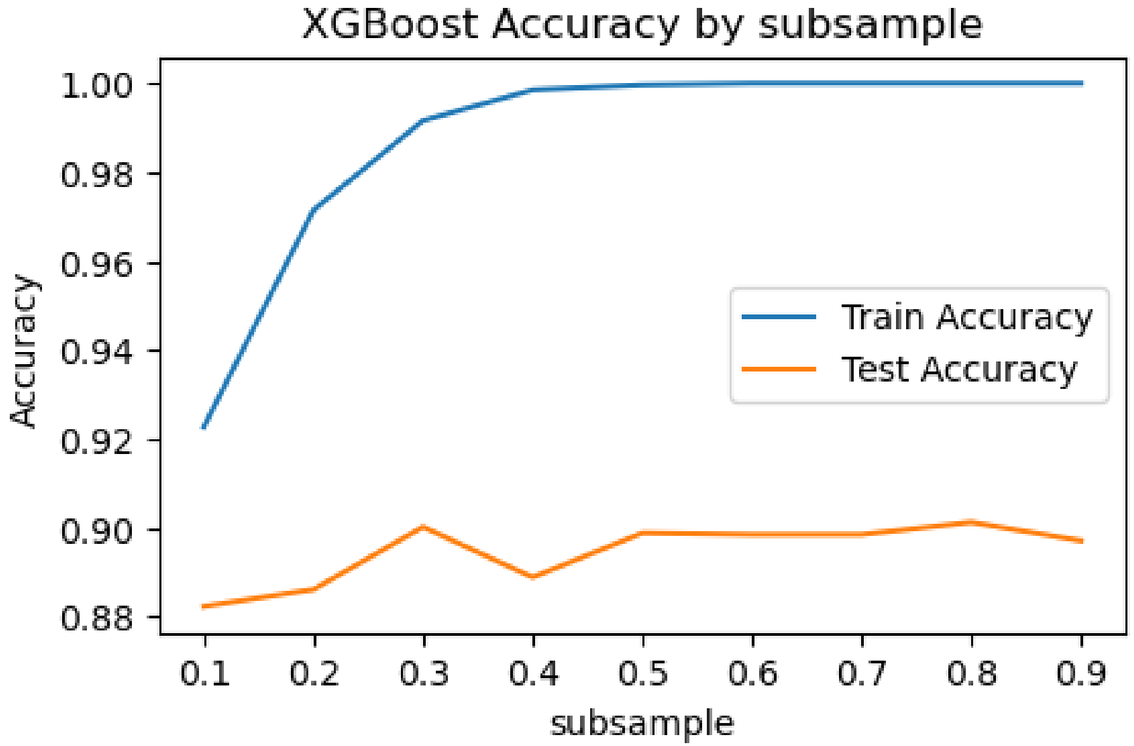
하이퍼파라미터 튜닝

3. XGBoostClassifier 모델

- 하이퍼파라미터에 따른 변화



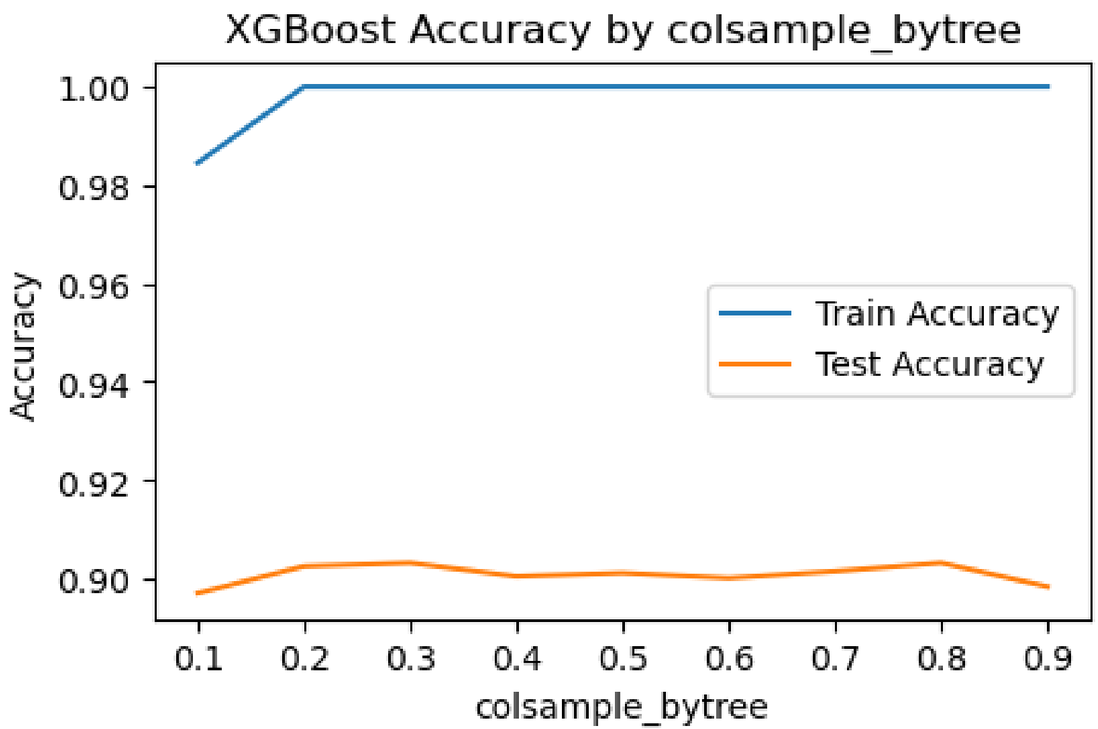
< learning rate >



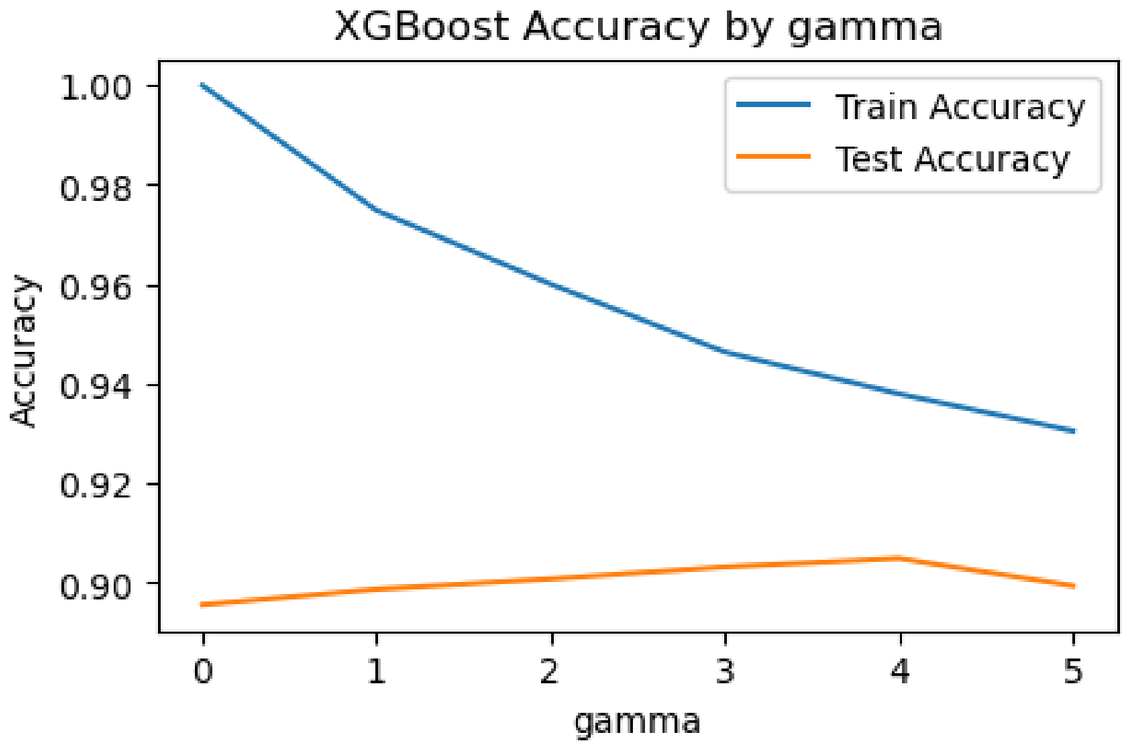
< subsample >

하이퍼파라미터 튜닝

- 하이퍼파라미터에 따른 변화



< colsample_bytree >



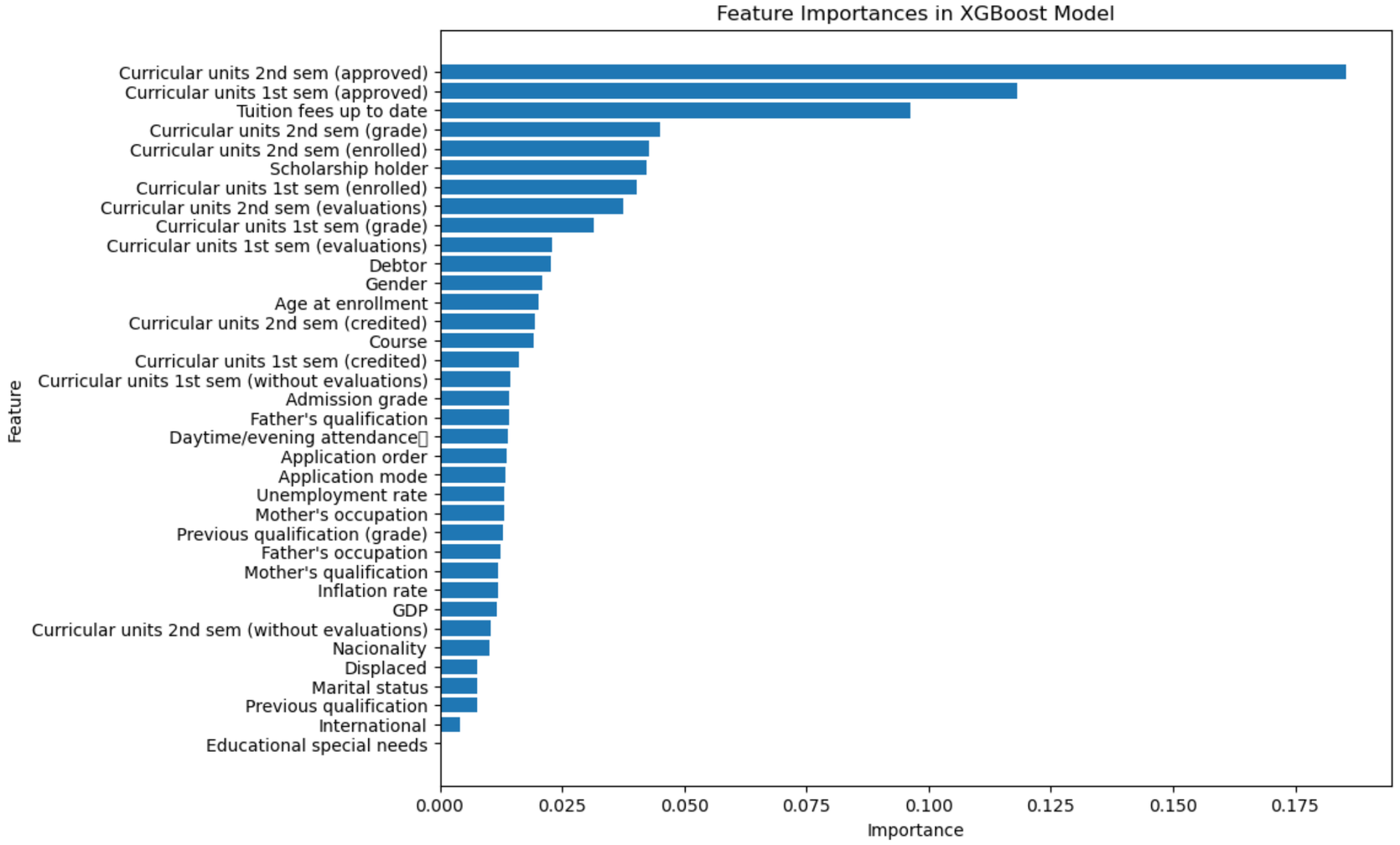
< gamma >

04

하이퍼파라미터 튜닝

3. XGBoostClassifier 모델

- 특성 중요도



04

하이퍼파라미터 튜닝

3. XGBoostClassifier 모델

- 특성 중요도

Feature	Importance
Curricular units 2nd sem (approved)	0.185312
Curricular units 1st sem (approved)	0.118055
Tuition fees up to date	0.096272
Curricular units 2nd sem (grade)	0.045129
Curricular units 2nd sem (enrolled)	0.042695
...	
Previous qualification	0.007565
International	0.004222
Educational special needs	0.000000

05.

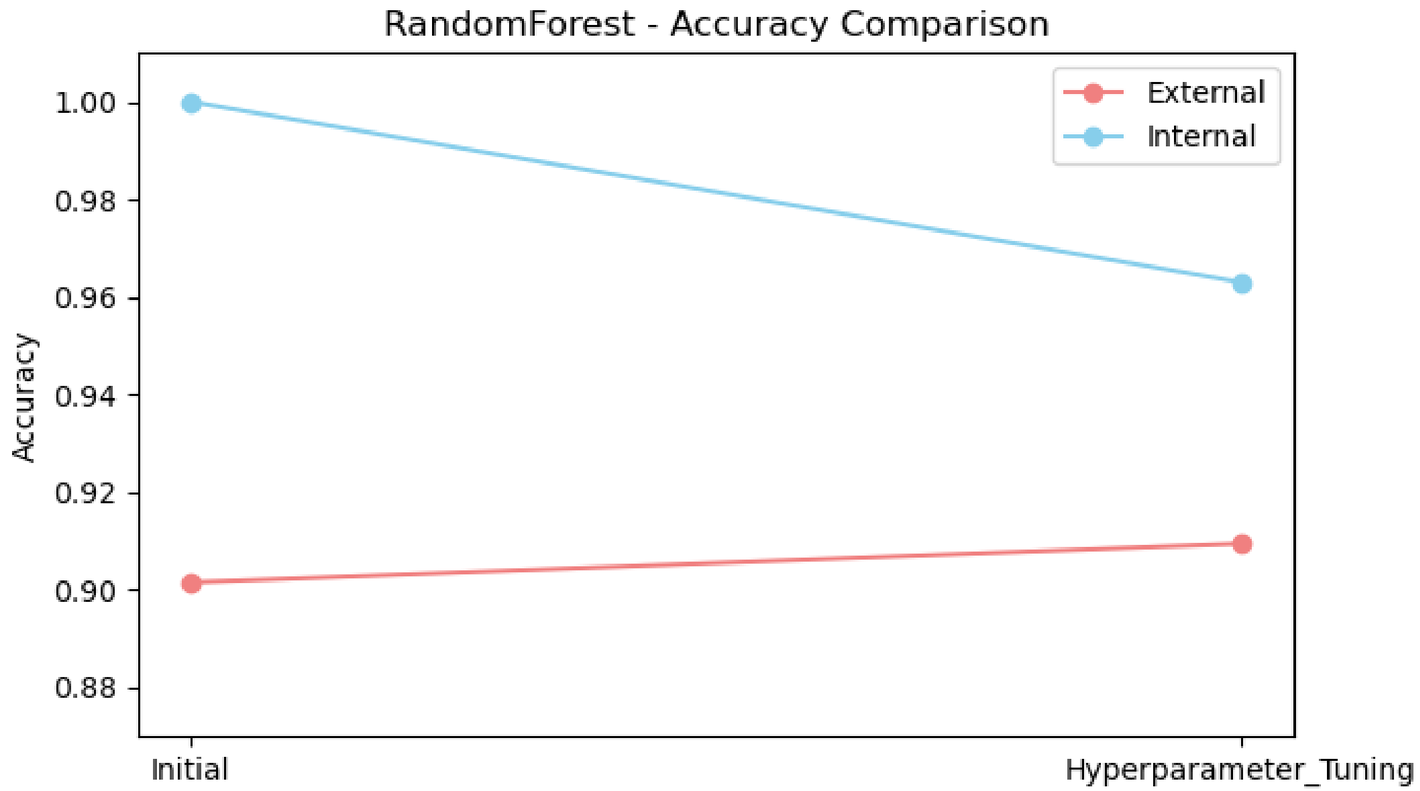
결과 및 비교

- 각 모델별 정확도, 내부 정확도 비교

Model	Initial_cv	Initial(Internal)	Hyper_external	Hyper_Internal
Random Forest	0.902	1.0	0.909	0.963
Gradient Boosting	0.903	0.944	0.913	0.937
XGBoost	0.896	1.0	0.913	0.946

05.

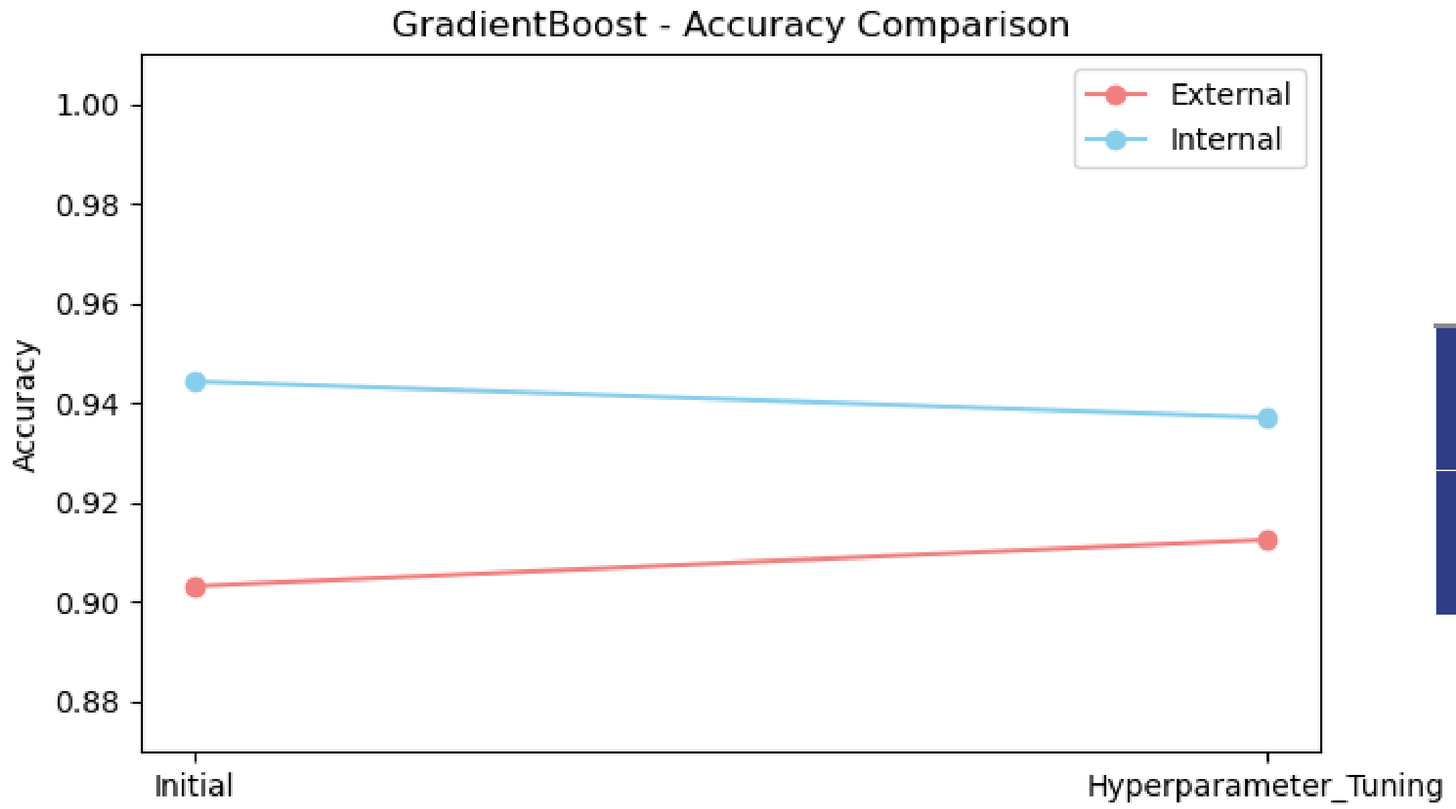
정확도 비교 - RandomForest



	Internal	External
초기 모델	1.00	0.9015
하이퍼파라미터 튜닝 후	0.9630	0.9094

05.

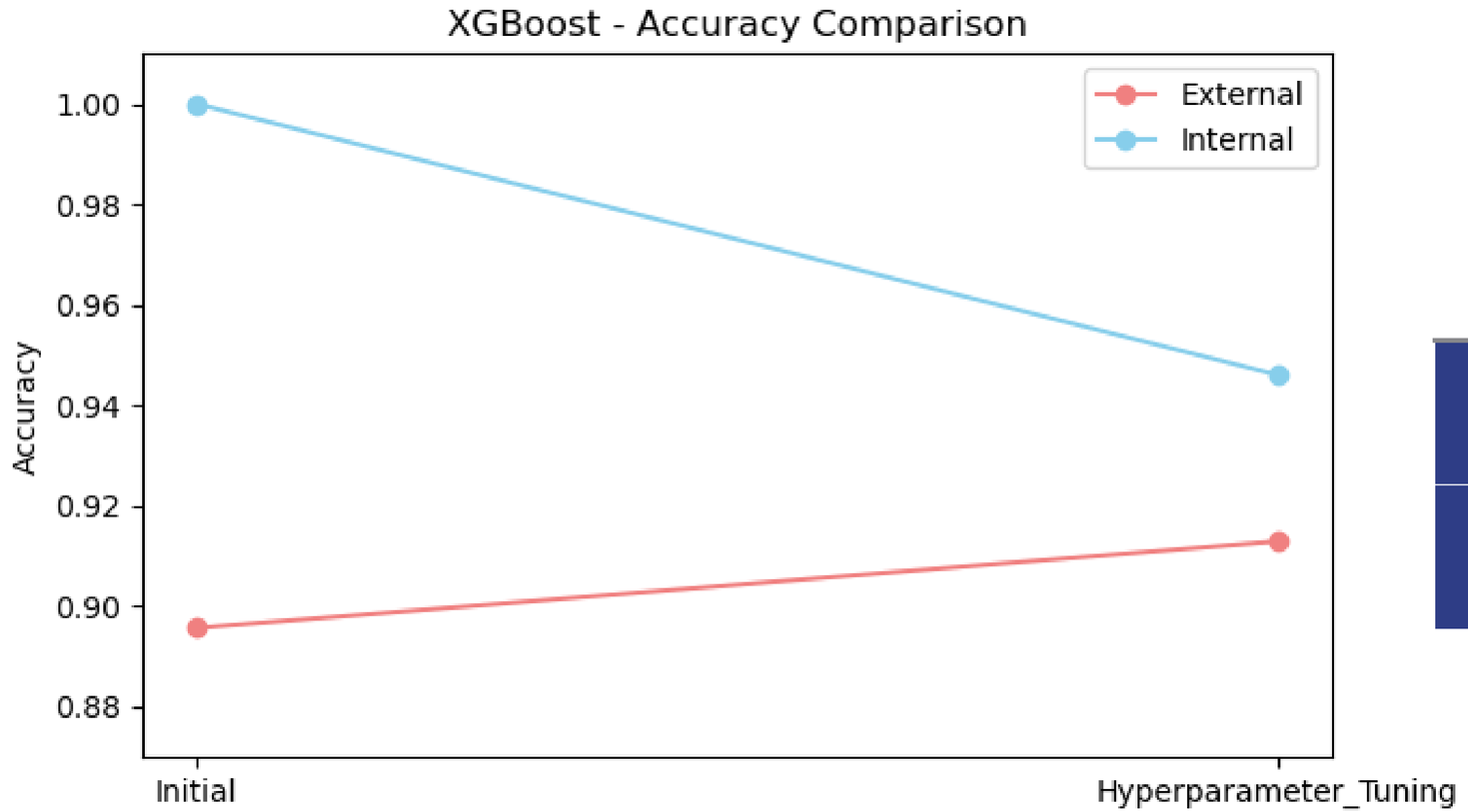
정확도 비교 - Gradient Boost



	Internal	External
초기 모델	0.9443	0.9032
하이퍼파라미터 튜닝 후	0.9370	0.9125

05.

정확도 비교 - XGBoost

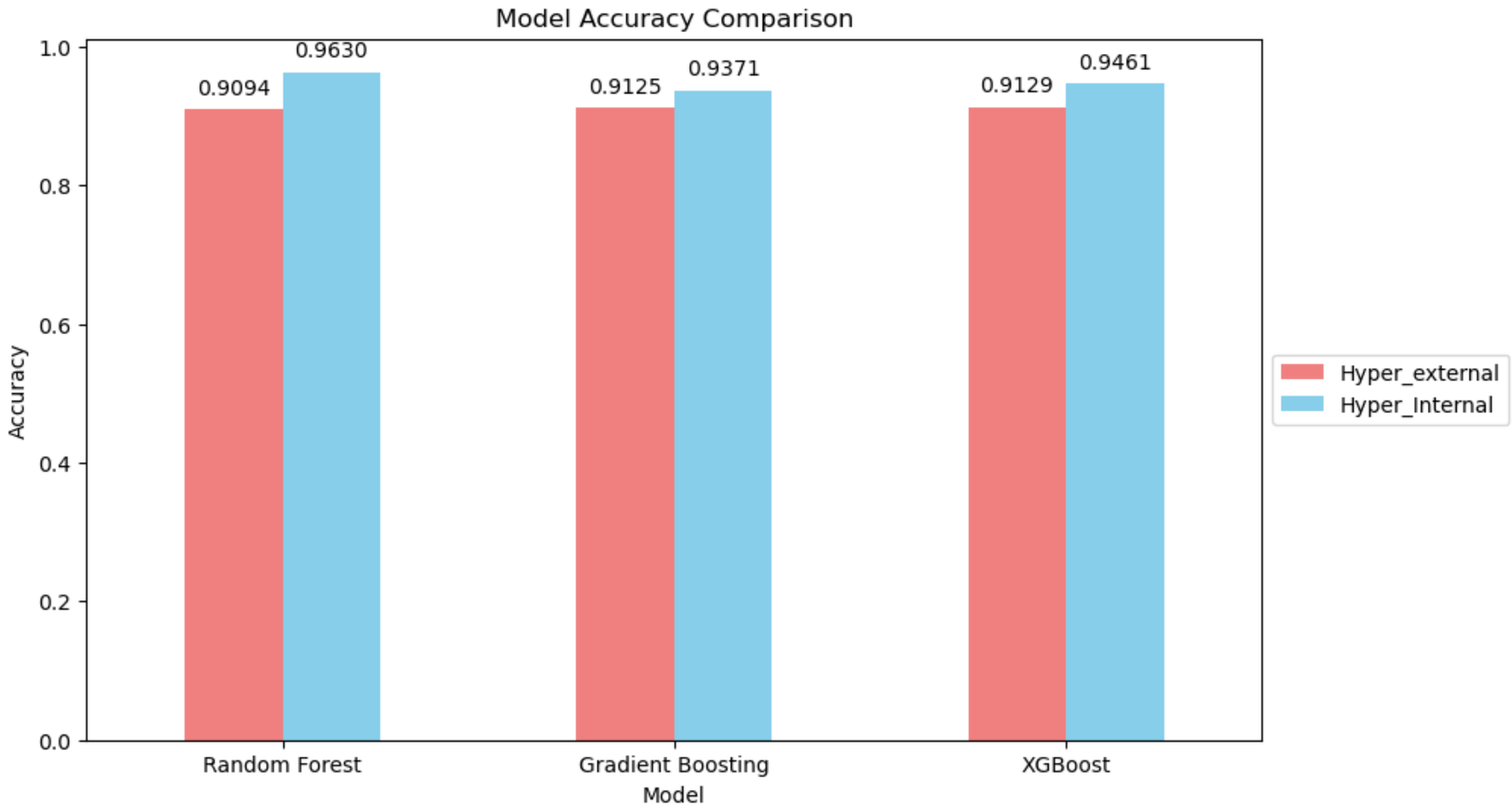


	Internal	External
초기 모델	1.00	0.8956
하이퍼파라미터 튜닝 후	0.9460	0.9128

05.

정확도 비교

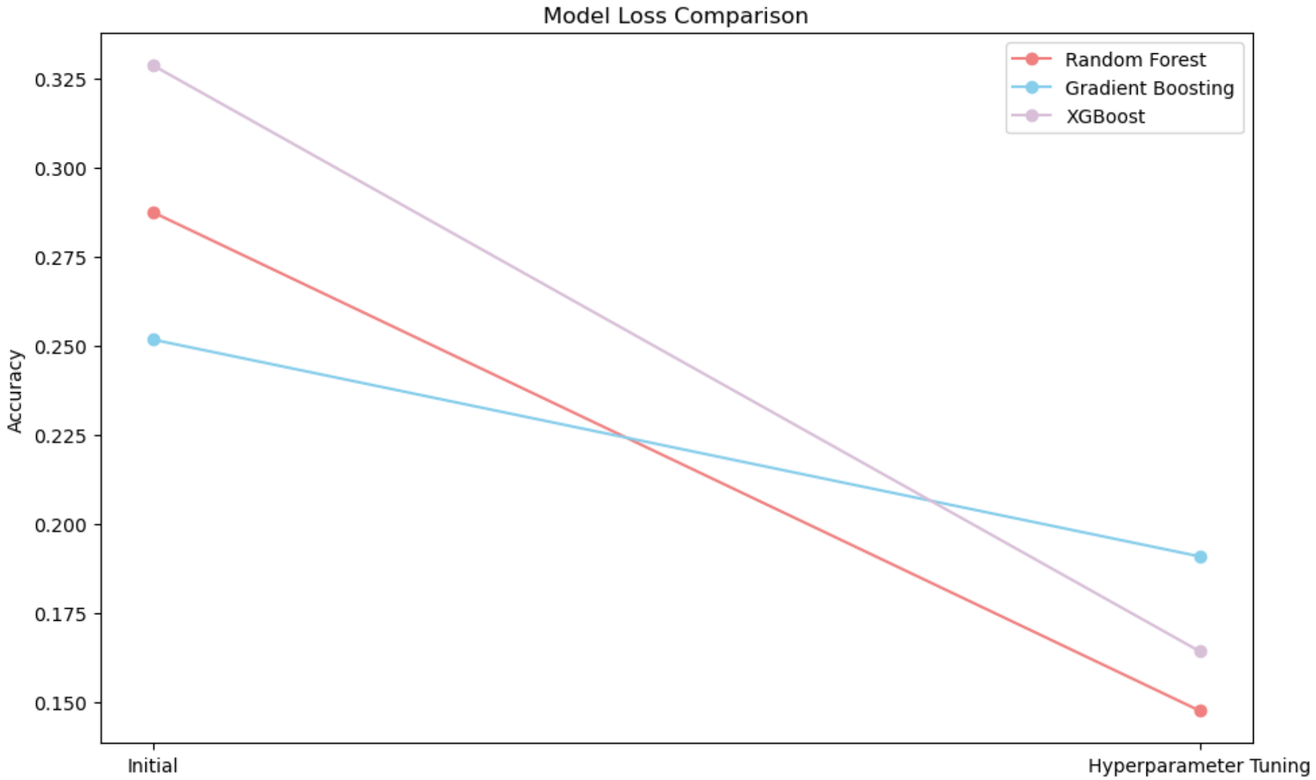
- 하이퍼파라미터 튜닝 후 정확도 비교 그래프



05.

손실 함수 비교

초기 모델과 하이퍼파라미터 튜닝 후 손실 함수 비교



Model	Initial	Hyperparameter Tuning
Random Forest	0.29	0.15
Gradient Boosting	0.25	0.19
XGBosst	0.32	0.16

손실 함수 값이 낮을수록 모델의 성능 ↑
세 모델 모두 하이퍼파라미터 튜닝 후 손실 함수 값이 낮아짐

튜닝 후 모델의 성능이 향상됨을 의미

06.

최종 모델 선택 및 적용

가장 성능이 좋았던 XGBoost에 테스트 데이터 적용

선택한 하이퍼파라미터

n_estimators : 198
max_depth : 3
learning_rate : 0.07
subsample : 0.7
colsample_bytree : 0.6
gamma : 1

테스트 데이터 정확도

- 약 90.36%

07.

결론

- Curricular units 1st, 2nd sem (approved)(승인 교과목)이 공통적으로 높은 중요도를 보였다.
수강신청을 하지 않는다는 것은 학업 중단여부와 직접적인 관련이 있고 이미 학업 중단을 결정한 경우라고도 볼 수 있다.
또한 교과목과 관련된 모든 열이 이 항목에 직접적으로 영향받기 때문에 이와 관련된 열들의 특성 중요도가 높다.

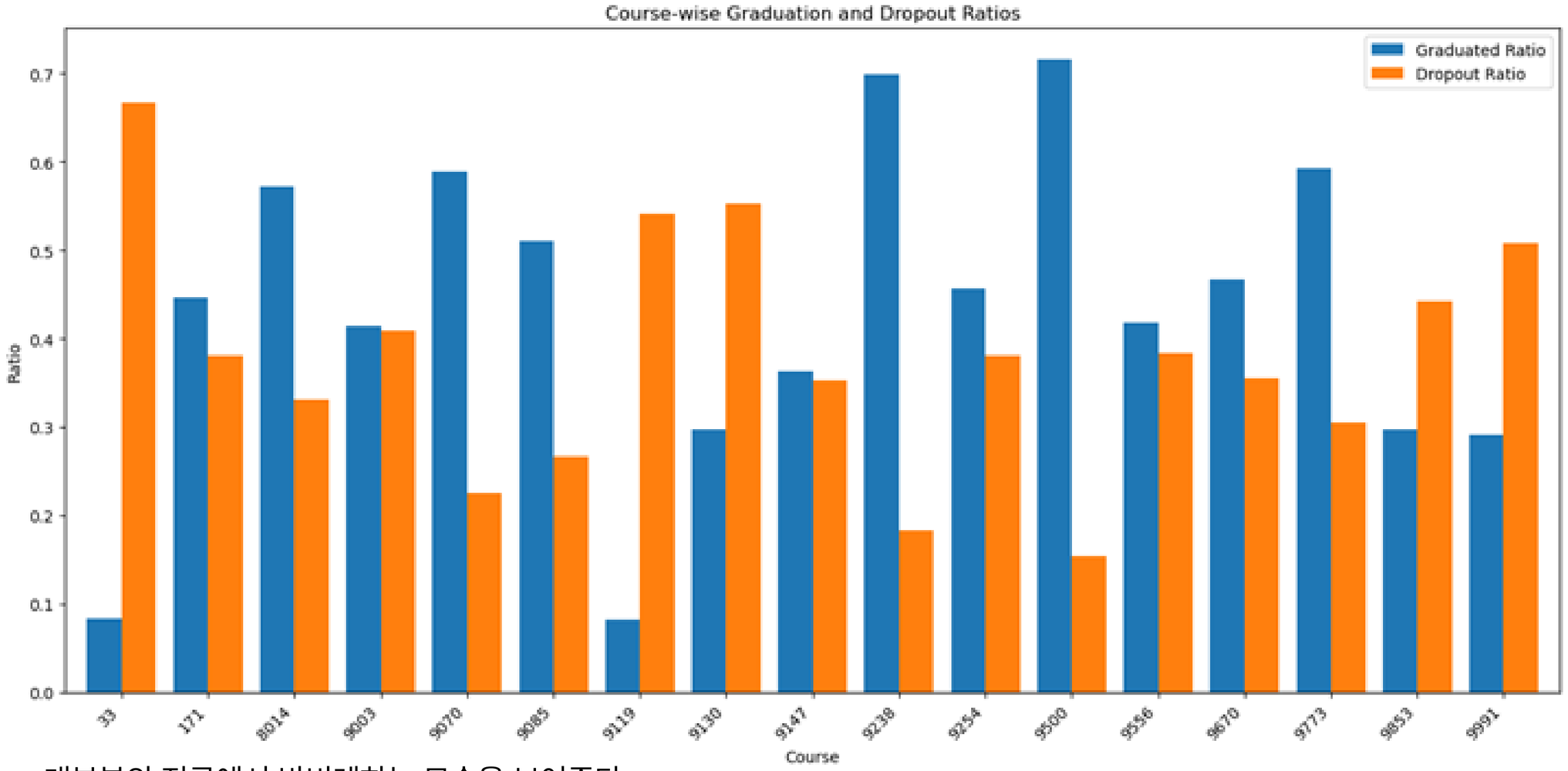
- Tuition fees up to date(수업료 납부 여부) 또한 학업 중단과 직접적인 연관이 있다.
이미 수업료를 납부하지 않은 경우는 학업이 중단된 경우이다.

- 위 항목과 달리 직접적인 연관성이 낮고 학업 중단을 조기식별할 수 있는 항목은 각 모델에서 공통적으로 높은 중요도를 보이는 Course(수강한 과목), Admission grade(입학 성적), Age at enrollment(입학 시 나이), Scholarship holder(장학금 수여 여부), Debtor(채무자 여부)이다.
즉, 학생들은 전공, 입학시의 상황, 경제적 상황에 따라 학업을 중단하는 경우가 많다.

	입학 나이 평균	입학 성적 평균	채무자 비율	장학금 수혜 비율
학업 중단	26세	125점	22%	9%
졸업	22세	129점	5%	38%

결론

- 학업중단, 졸업 별 전공 선택 비율



대부분의 전공에서 반비례하는 모습을 보여준다.

08.

개선점

- 타겟과 직접적인 관련이 있고 특성중요도가 높은 열은 학업 중단을 조기식별 하기에 적절하지 않다.

- 조기식별이라는 목적을 위해 앞서 말한 열들을 제외하고도 높은 정확도가 나오는 모델을 구축하면 더 유용한 모델이 될 것이다.

- 특성 중요도 등을 고려한 데이터 전처리를 수행하여 성능을 더욱 개선시킬 수 있을 것이다.

-전공 코드에 따른 전공명 또한 알 수 있다면 더욱 의미있는 분석이 될 것이다.