DAV 6150 모듈 11 할당

결정 트리 및 랜덤 포레스트

***** 이 과제를 위해 3명 이하의 소규모 그룹으로 작업할 수 있습니다 ***.**

의사 결정 트리와 랜덤 포레스트 모델은 분류 문제에 적용할 때 모두 매우 효과적일 수 있으며, 부트스트랩 집계를 통해 많은 수의 개별 의사 결정 트리를 구성하고 이들의 집합적 결과를 사용하여 예측 데이터 값에 도달함으로써 단일 개별 트리의 성능을 향상시킬 수 있다는 것을 알게 되었습니다. 일반적으로 랜덤 포레스트 모델의 성능이 단일 의사 결정 트리의 성능보다 우수할 것으로 예상되지만, 단일 의사 결정 트리와 랜덤 포레스트 중 어느 것을 구현할지 결정할 때 평가해야 하는 복잡성 대 성능의 상충 관계는 분명합니다. 랜덤 포레스트는 개별 의사 결정 트리보다 훨씬 더 계산적으로 복잡하고 일반적으로 설명/해석하기가 더 어렵다는 점입니다.

모듈 11 과제의 과제는 의사 결정 트리와 랜덤 포리스트의 성능을 비교하는 것입니다. 여러분이 작업하게 될 데이터 세트는 프로젝트 1에서 처음 사용한 데이터 세트의 하위 집합입니다. 데이터 파일에 포함된 속성을 설명하는 데이터 사전이 아래에 제공됩니다.

데이터 집합 속성	**설명**
보고_학교_연도	고등학교 졸업 정보가 제공되는 학년도를 나타냅니다. 보고
집계_인덱스	고등학교 졸업 데이터의 방식을 식별하는 숫자 코드 집계됨
집계 유형	고등학교 졸업 데이터가 집계된 방식에 대한 텍스트 설명
nrc_code	지표인 "필요/자원 용량"을 식별하는 숫자 코드입니다. 학군 유형
nrc_desc	학군 유형에 대한 텍스트 설명
카운티_코드	카운티 이름의 숫자 코드
카운티_이름	해당 뉴욕주 카운티의 전체 이름
nyc_ind	학군이 다음 경계 내에 있는지 여부를 나타냅니다. NYC
멤버십_설명	학생이 고등학교에 처음 등록한 학년을 나타냅니다.
subgroup_code	학생 하위 그룹을 식별하는 숫자 코드
서브그룹_이름	학생 하위 그룹에 대한 텍스트 설명입니다. 한 학생이 **둘 이상의** 하위 그룹(예: '여성', '히스패닉', '영어가 아님', '영어 없음 언어 학습자" 등)
enroll_cnt	표시된 하위 그룹에 속한 학생 중 해당 기간 동안 등록한 학생 수 주어진 학년도
grad_cnt	표시된 하위 그룹의 재적 학생 중 졸업한 학생 수 해당 학년도 말
grad_pct	표시된 하위 그룹에 등록한 학생의 비율은 몇 퍼센트입니까? 해당 학년도 말에 졸업한 경우

reg_cnt	표시된 하위 그룹에 등록된 학생 중 얼마나 많은 학생이 "리전트" 디플로마
reg_pct	표시된 하위 그룹에 등록한 학생의 비율은 몇 퍼센트입니까? "리전트" 졸업장 수여
dropout_cnt	표시된 하위 그룹에 등록한 학생 중 중단된 학생 수 학년도 중 고등학교 등록 여부
dropout_pct	표시된 하위 그룹에 등록한 학생의 비율은 몇 퍼센트입니까? 학기 중 고등학교 등록을 중단한 경우

아시다시피, 데이터 집합은 73,000개 이상의 관측치로 구성되어 있으며, 각 관측치는 2018-2019학년도 말 기준으로 4년 이상 재학한 고등학생의 특정 뉴욕주 교육구 및 관련 하위 그룹/분류에 관한 것입니다. 모델링할 응답 변수는 데이터 집합의 **reg_pct** 속성에서 파생된 범주형 지표 변수입니다. 이 새 지표 변수(생성해야 하는 변수)는 세 가지 가능한 값으로 구성됩니다:

A. "**낮음**": 특정 학군/학생 하위 그룹에서 수여된 리전트 졸업장의 비율이 전체 리전트 졸업장(즉, 모든 학군/학생 하위 그룹에서)의 중간 비율의 ½ 미만인 경우를 나타냅니다;

B. "**중간**": 특정 학군/학생 하위 그룹에 대해 수여된 리전트 졸업장의 비율이 0.5 * 수여된 모든 리전트 졸업장의 중앙값 비율(즉, 모든 학군/학생 하위 그룹에서, 모든 학군/학생 하위 그룹에서) 및 1.5 * 수여된 모든 리전트 졸업장의 중간 비율(즉, 모든 학군/학생 하위 그룹에서), 즉 (0.5 * 중간 비율) < 주어진 학군의 리전트 졸업장 수여 비율 <= (1.5 * 중간 비율) 사이입니다.

C. "**높음**": 특정 학군/학생 하위 그룹에서 수여된 리전트 졸업장의 비율이 1.5 * 수여된 모든 리전트 졸업장의 중간 비율(즉, 모든 학군/학생 하위 그룹에서)을 초과하는 것을 나타냅니다.

따라서 의사 결정 트리와 랜덤 포레스트 모델은 세 가지 필수 새 지표 값 중 주어진 관측값에 적용될 가능성이 가장 높은 값을 예측하기 위한 목적으로 설계해야 합니다.

과제를 시작하려면 다음과 같이 하세요:

1) 제공된 **M11_Data.csv** 파일을 DAV 6150 Github 리포지토리에 로드합니다.

2) 그런 다음, 주피터 노트북을 사용해 Github 리포지토리에서 데이터 세트를 읽고 판다스 데이터 프레임에 로드합니다. 데이터 프레임 내에서 데이터 속성이 올바르게 레이블이 지정되었는지 확인합니다.

3) Python 기술을 사용하여 기본적인 탐색적 데이터 분석(EDA)을 수행하여 각 변수(응답 변수 포함)의 특성을 이해합니다. (참고: 이미 프로젝트 1의 고품질 EDA를 가지고 있다면 여기에 관련 구성 요소를 통합할 수 있습니다. 프로젝트 1 EDA에서 M11 데이터 세트에 다시 나타나는 속성에 결함이 있는 경우, EDA 작업을 반복하고 프로젝트 1에서 확인된 관련 결함을 해결해야 합니다. 수정되지 않은 결함은 이 과제에서 해당 점수가 감점된다는 점에 유의하세요.)

   EDA 작성에는 속성에 대한 통계 분석과 함께 작성한 탐색 그래픽(예: 막대형 차트, 박스형 차트, 히스토그램, 선형 차트 등)에서 도출할 수 있는 모든 인사이트가 포함되어야 합니다. 또한 몇 가지 예비 예측 추론(

예: 설명 변수 중 응답 변수를 상대적으로 더 '예측'하는 것으로 보이는 변수가 있습니까?)을 확인해야 합니다. 설명 변수와 응답 변수 간의 이러한 관계를 잠재적으로 식별할 수 있는 다양한 방법이 있습니다. 계산할 적절한 통계 메트릭과 사용할 탐색 그래픽 유형을 선택하는 등 EDA를 수행하는 방법을 결정하는 것은 데이터 과학 실무자의 몫입니다. 독자가 흥미를 잃을 정도로 상세하지 않으면서도 철저하고 간결한 EDA를 제공하는 것이 목표여야 합니다.

4) 데이터 준비 작업의 첫 번째 단계로, 위에서 설명한 접근 방식을 사용하여 **reg_pct** 속성의 콘텐츠에서 파생된 새 범주형 지표 변수를 만들어야 합니다. EDA 결과를 사용하여 위에서 설명한 세 가지 가능한 분류(즉, "**낮음", "중간**", "**높음**")를 가진 "**reg_pct_level**"이라는 새 지표 변수를 생성합니다.

   데이터 세트에 포함된 모든 관측값에 대해 적절한 "**reg_pct_level**" 값이 계산되었는지 확인합니다.

   reg_pct_level 표시기를 **생성한** 후에는 **데이터프레임에서 "reg_pct" 및 "reg_cnt" 속성을 제거해야** 합니다. 이는 속성 컬렉션에 "**reg_pct_level**" 지표가 추가될 때 발생할 수 있는 선형성을 제거하기 위해 반드시 수행해야 합니다.

5) 준비된 데이터 검토에서 새로 생성된 "**reg_pct_level**" 지표 값의 분포를 분석해야 합니다. 분석 결과 새로 생성된 지표 변수의 분포에 대해 무엇을 알 수 있나요?

6) Python 기술을 사용하여 기능 선택 및 차원 축소 지식을 예상 설명 변수에 적용하여 모델 내에서 상대적으로 유용할 것으로 생각되는 변수를 식별합니다. 여기에는 EDA 작업을 통해 얻은 지식이 반영되어야 합니다. 기능을 선택할 때는 모델 성능과 모델 간소화 간의 상충 관계(예: 모델의 복잡성을 줄이는 경우 정확도(또는 다른 성능 지표)를 너무 많이 희생하고 있지는 않은지)를 고려해야 합니다. 기능 선택 및/또는 차원 축소 결정을 구현하는 방식은 데이터 과학 실무자가 결정해야 합니다. 필터링 방법을 사용할 것인가? PCA? 단계별 검색을 사용할 것인가? 선호하는 접근 방식을 결정하는 것은 사용자의 몫입니다. 의사 결정 과정을 정당화할 수 있는 설명적인 내러티브를 반드시 포함해야 합니다.

7) 데이터를 훈련 및 테스트 하위 집합으로 분할한 후, 훈련 하위 집합을 사용하여 서로 다른 설명 변수(또는 다른 변환 방법을 통해 변환된 경우 동일한 변수)의 조합을 사용하여 **최소 두 개의 서로 다른 의사 결정 트리** 모델과 **최소 두 개의 서로 다른 랜덤 포레스트 모델을** 구성합니다. **모델에는 각각 최소 4개의 설명 변수가 포함되어야 합니다.**

8) 다양한 모델을 훈련한 후, 구축한 분류 모델 중에서 "최상의" 분류 모델을 어떻게 선택할지 결정하세요. 예를 들어, 해석이 더 쉽거나 구현이 덜 복잡하다면 성능이 약간 낮은 모델을 선택할 의향이 있나요? 모델을 비교/대조하기 위해 어떤 메트릭을 사용하시겠습니까? 학습 데이터 세트를 사용하여 교차 검증을 통해 모델의 성능을 평가하세요. 그런 다음 선호하는 모델을 테스트 하위 집합에 적용하고 이전에 볼 수 없었던 데이터에서 얼마나 잘 작동하는지 평가하세요.

이 과제의 결과물은 Jupyter 노트북입니다. 이 노트북에는 적절한 형식의 마크다운 셀에 포함된 Python 코드 셀과 설명 내러티브의 조합이 포함되어야 합니다. 노트북에는 최소한 다음 섹션이 포함되어야 합니다(각 섹션의

관련 Python 코드 포함):

1) **소개(5점)**: 문제 요약 + 문제 해결을 위해 취할 단계를 설명합니다.

2) **탐색적 데이터 분석(15점)**: 예비 예측 추론을 포함하여 분석에서 도출한 결론을 포함하여 EDA 작업을 설명하고 제시합니다. 이 섹션에는 EDA에 사용된 Python 코드가 포함되어야 합니다.

3) **데이터 준비(10점)**: 데이터 세트에 적용한 기능 엔지니어링 기법을 포함하여 EDA에서 식별한 데이터 무결성 + 사용성 문제를 해결하기 위해 수행한 단계를 설명하고 보여주세요. 이 섹션에는 데이터 준비에 사용된 Python 코드가 포함되어야 합니다.

4) **준비된 데이터 검토(5점)**: 데이터 준비 후 EDA 분석을 설명하고 제시합니다. 이 섹션에는 데이터 준비 작업 중에 조정한 변수에 대해 EDA를 다시 실행하는 데 사용된 Python 코드가 포함되어야 합니다.

5) **의사 결정 트리 + 랜덤 포리스트 모델링(40점)**: 기능 선택/차원 축소 결정 및 모델의 하이퍼파라미터를 선택한 과정을 포함하여 의사 결정 트리 및 랜덤 포레스트 모델링 작업을 설명하고 제시합니다. 이 섹션에는 기능 선택, 차원 축소 및 모델 구축에 사용된 Python 코드가 포함되어야 합니다.

6) **모델 선택(15점)**: 모델 선택 기준을 설명합니다. 선호하는 모델을 식별합니다. 해당 모델의 성능을 다른 모델의 성능과 비교/대조합니다. 특정 모델을 선호하는 모델로 선택한 이유를 토론합니다. 선호하는 모델을 테스트 하위 집합에 적용하고 그 결과를 토론합니다. 선호하는 모델이 예상대로 잘 작동했나요? 모델 선택 작업의 일부로 사용된 Python 코드를 포함시키고, 모델에서 도출한 분류 성능 메트릭의 맥락에서 토론의 틀을 잡아야 합니다.

7) **결론(10점)**

주피터 노트북의 결과물은 출판물 수준의 전문 문서와 유사해야 하며, 명확하게 표시된 그래픽, 고품질 서식, 명확하게 정의된 섹션 및 하위 섹션 헤더, 맞춤법 및 문법 오류가 없어야 합니다. 또한 Pythion 코드에는 간결한 설명 주석이 포함되어야 합니다.

제공된 M11 과제 캔버스 제출 포털에서 Jupyter 노트북을 업로드합니다. 노트북을 저장할 때는 **이름_성_M11_assn**"(예: J_Smith_M11_assn)과 같은 명명법을 사용하여 저장해야 합니다. ***소규모 그룹은 Jupyter 노트북을 시작할 때 모든 그룹 구성원의 신원을 확인하고 각 팀원은 Canvas 내에서 팀 작업의 사본을 제출해야 합니다.***