DAV 6150 모듈 8 할당

KNN 및 SVM을 통한 분류

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

모듈 7 과제는 보험 회사 고객과 관련된 데이터가 포함된 Kaggle 기여 (https://www.kaggle.com/rluyck/insurance-company?select=Customer_data.csv)에서 제공된 데이터 세트를 사용했습니다. 이 데이터 집합은 하나의 응답/종속 변수(새로운 보험 상품 구매 여부를 나타내는)와 14개의 설명/독립 변수에 대한 14,000개 이상의 관측치로 구성되었습니다.

기억하시겠지만, 이 과제에서 여러분은 보험 회사로부터 특정 보험 회사 고객이 추가 보험 상품을 구매할 가능성이 있는지 여부를 예측할 수 있는 이항 로지스틱 회귀 모델을 개발하라는 과제를 받았습니다. 이 보험 회사는 고객 유지 및 영업 관행을 개선하기 위해 이러한 모델의 결과를 사용할 계획입니다.

이 과제의 경우 보험 회사에서 작업을 다시 검토하고 이항 로지스틱 회귀 모델 이외의 다른 분류 모델이 지정된 작업에 더 효과적인지 확인하기 위해 추가 분류 모델을 개발해 달라고 요청했습니다.

모듈 8 과제의 과제는 주어진 보험 회사 고객이 추가 보험 상품을 구매할 가능성이 있는지 여부를 예측하는 일련의 **K-최근 이웃 및 서포트 벡터 머신 모델을** 구성하고 비교/대조하는 것입니다(필요한 EDA 및 데이터 준비 작업을 완료한 후). 모델링할 응답 변수는 특정 보험 회사 고객이 추가 보험 상품을 구매했는지 여부를 나타내는 데이터 세트의 "**TARGET**" 속성입니다. 이러한 모델에 어떤 기능을 포함할지 결정하는 것은 데이터 과학 실무자의 몫입니다. 작업에는 EDA, 데이터 준비(필요에 따라 변환 포함), 기능 선택, 모델 성능 메트릭에 대한 철저한 평가가 포함되어야 합니다. 다음과 같이 과제를 시작하세요:

1) 모듈 7 과제와 함께 제공된 **M7_Data.csv** 파일이 DAV 6150 Github 리포지토리에 로드되었는지 확인합니다.
2) 그런 다음, 주피터 노트북을 사용하여 Github 리포지토리에서 데이터 세트를 읽고 Pandas 데이터 프레임에 로드합니다.
3) 필요에 따라 EDA 작업을 수행합니다. (**참고**: 모듈 7 과제에서 이미 완성한 고품질 EDA가 있는 경우 여기에 통합할 수 있습니다. M7 과제 EDA에 결함이 있는 경우, EDA 작업을 반복하여 M7 과제 EDA에서 확인된 모든 부족한 부분을 해결해야 합니다. 수정되지 않은 결함은 이 과제에 상응하는 감점을 받게

됩니다.)

4) 작업에 필요하다고 판단되는 기능 엔지니어링 또는 표준화 조정을 포함하여 필요한 데이터 준비 작업을 수행합니다.

5) 특징 선택 및/또는 차원 축소 기술에 대한 지식을 적용하여 모델에 포함할 **설명 변수를 최소 4개 이상** 식별합니다. 도메인 지식을 적용하여 수동으로 기능을 선택하거나, 순방향 또는 역방향 선택을 사용하거나, 다른 기능 선택 방법(예: 의사 결정 트리 등)을 사용할 수 있습니다. 데이터 과학 실무자가 데이터 집합에 사용할 가장 적절한 특징 선택 및/또는 차원 축소 기법을 결정하는 것은 사용자의 몫입니다.

6) 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 KNN** 모델과 **최소 두 개의 서로 다른 SVM 모델을** 구성합니다.

7) 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) **KNN + SVM 모델링(40점)**: 기능 선택/차원 축소 결정, KNN 모델의 "K" 값 선택 과정, SVM 모델의 하이퍼파라미터 선택 과정, SVM 모델 내에서 사용된 커널 함수(해당하는 경우)를 포함하여 KNN 및 SVM 모델링 작업을 설명하고 제시하세요. 이 섹션에는 기능 선택, 차원 축소 및 모델 구축에 사용된 모든 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? 선호하는 모듈 8 과제 모델의 성능은 모듈 7 과제에서 선호하는 이항 로지스틱 회귀 모델의 성능과 어떻게 비교되었는가? 모델 선택 작업의 일부로 사용된 Python 코드를 포함하고 모델에서 도출한 분류 성능 메

트릭의 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 M8 과제 캔버스 제출 포털에서 Jupyter 노트북을 업로드합니다. 노트북을 저장할 때는 **이름_이름_성_M8_assn**"(예: J_Smith_M8_assn_)과 같은 명명법을 사용해야 합니다.

소규모 그룹은 주피터 노트북을 시작할 때 모든 그룹 구성원의 신원을 확인하고 각 팀원은 캔버스 내에서 팀 작업의 사본을 제출해야 합니다.