DAV 5400 프로젝트 4(13주차) (100점)

scikit-learn을 사용한 예측 분석

****이 프로젝트는 3명 이하의 소규모 그룹으로 작업할 수 있습니다. ****

데이터 분석가로서 우리는 종종 데이터를 한 가지 형태로 가져와서 더 쉽게 다운스트림 분석을 할 수 있도록 변환하는 작업을 해야 합니다. 이 프로젝트에서는 교육 과정에서 배운 내용을 사용하여 예측 분석을 위한 데이터를 준비한 다음, `scikit-learn` 라이브러리에서 제공되는 도구를 사용하여 예측 모델을 구축합니다.

많은 업계에서 기존 고객과의 관계를 확대하는 데 관심이 있습니다. 이를 위해 기업은 종종 기존 고객에게 추가 상품을 판매하려고 시도합니다. 예를 들어, 특정 은행에 당좌 예금 계좌를 가지고 있는 고객에게 자동차 대출이나 모기지 등을 판매하려고 시도할 가능성이 높습니다. 한 대기업으로부터 특정 기존 고객이 회사에서 추가 상품을 구매할 가능성이 있는지 여부를 예측할 수 있는 모델 개발을 의뢰받았다고 가정해 보겠습니다. 이 회사는 고객 유지 및 영업 관행을 개선하기 위해 이러한 모델의 결과를 사용할 계획입니다.

작업할 데이터 집합은 응답/종속 변수(신제품 구매 여부를 나타내는) 1개와 설명/독립 변수 11개에 대한 14,000개 이상의 관찰로 구성됩니다. 이 회사는 신제품을 제공한 고객에 대한 데이터를 수집했습니다. 신제품에 가입했는지 여부에 대한 정보와 함께 일부 고객 정보 및 다른 두 제품(제품 A 및 제품 B)의 구매 행동에 대한 정보가 주어집니다. 데이터 세트에 대한 데이터 사전은 아래에 제공됩니다.

속성	설명
ID	고유 고객 식별자
대상	고객이 새 제품을 구매했는지 여부(N = 아니요, Y = 예)
로열티	고객 충성도 수준, 낮음부터 높음까지(0~3), 99 = 분류되지 않음
나이	고객 연령(연도)
도시	도시별 고유 코드(고객 거주지)
Age_p	고객 파트너의 연령(연도)
LOR_m	고객과 회사와의 관계 기간(개월)
Prod_A	고객이 이전에 제품 A를 구매했습니다(0=아니요, 1=예).
Type_A	제품 유형 A
Turnover_A	고객이 제품 A에 지출한 금액
Prod_B	고객이 이전에 제품 B를 구매했습니다(0=아니요, 1=예).
Type_B	제품 유형 B
Turnover_B	고객이 제품 B에 지출한 금액

이 프로젝트에서는 scikit-learn을 사용하여 질문에 답하게 됩니다:

"***고객이 추가 제품을 구매할지 여부를 가장 잘 예측할 수 있는 속성은 무엇인가요?***"

이 프로젝트에 필요한 작업은 두 가지 단계로 구분할 수 있습니다. **1단계에** 필요한 모든 작업은 이 수업 10주차까지 학습한 Python 및 Pandas 기술을 사용하여 완료할 수 있으며(즉, 사전 **`scikit-learn`** 지식이 없어도), 2단계에서는 데이터 프레임에 포함된 속성의 예측 품질을 평가하는 데 **`scikit-learn`을** 사용해야 합니다.

1단계: **데이터 수집, 데이터 준비 및 탐색적 데이터 분석(60점)**

- 작업의 목적과 목표를 설명하는 전문가 수준의 소개 내러티브를 작성하세요.
- 제공된 데이터 세트를 Github 리포지토리에 로드하고 Github 리포지토리에서 Jupyter Notebook으로 읽어들입니다.
- 데이터 세트와 데이터에 대한 관련 설명(예: "데이터 사전")을 공부합니다.
- **데이터 집합의 열 하위 집합을 사용하여** 판다 데이터 프레임을 만듭니다. **고객이 추가 제품을 구매했는지 여부를** 나타내는 **TARGET** 열, **Age, Type_A, Type_B, Ior_M** 열 및 선택한 **다른 두 개 이상의 열을** 포함해야 합니다.
- 예를 들어, 첫 번째 열의 **목표** 표시기를 숫자로 변환하면 'N'은 0이 되고 'Y'는 1이 될 수 있습니다.
- 탐색적 데이터 분석 수행: 선택한 각 열(**Age, Type_A, Type_B, Ior_M** 열 포함)에 대한 데이터 분포를 표시하고, 선택한 다른 열뿐만 아니라 TARGET 대 연령, TARGET 대 Type_A, TARGET 대 Type_B, TARGET 대 Ior_M에 대한 플롯을 표시합니다. 이러한 작업에 사용할 플롯 유형을 결정하는 것은 사용자의 몫입니다. EDA 결과를 설명하는 텍스트를 포함합니다.
- 하위 집합에 포함시킨 다른 열*(예: TARGET 지표 제외)*이 특정 고객이 추가 제품을 구매할 가능성이 있는지 예측하는 데 도움이 될 수 있는지에 대한 예비 결론을 설명하는 텍스트를 포함하세요.
- **Type_A** 및 **Type_B** 열 모두에 대해 더미 변수 집합을 만듭니다. 이는 프로젝트 4에서 scikit-learn을 사용하는 다운스트림 처리에서 범주형 데이터 값을 이진 표시기 변수로 변환해야 하기 때문에 필요합니다. 이를 위한 한 가지 가능한 접근 방식은 pandas **get_dummies()** 메서드를 참조하세요.

2단계: **예측 모델 구축(40점)**

- 1단계에서 구성한 판다 데이터 프레임의 데이터(더미 변수 포함)로 시작하세요.
- scikit-learn을 사용하여 적절한 회귀 모델을 구성하고 해당 모델에서 제공하는 정보를 사용하여 선택한 예측자 열(**나이, 유형_A, 유형_B, Ior_M** 포함) 중 어떤 것이 고객이 추가 제품을 구매할 가능성이 가장 정확하게 예측되는지 결정합니다. 이 작업을 **스키킷 학습으로 수행하는** 방법은 데이터 분석 실무자에게 달려 있습니다.
- 추가 분석을 위한 권장 사항과 함께 결론을 명확하게 기술하세요.

****힌트**** : 데이터스쿨의 머신러닝 동영상에서 4개의 예측 변수로부터 홍채 종을 예측하는 데 사용된 프로세스를 이해했다면, 이 프로젝트를 완료하는 데 배운 내용을 적용할 수 있을 것입니다.

Jupyter 노트북의 결과물은 출판물 수준의 전문가용 문서와 유사해야 하며, 명확하게 표시된 그래픽, 고품질 서식, 명확하게 정의된 섹션 및 하위 섹션 헤더가 포함되어야 하고 맞춤법 및 문법 오류가 없어야 합니다. 또한 Python 코드에는 간결한

설명 주석이 포함되어야 합니다.

이 프로젝트의 모든 작업을 **하나의 Jupyter 노트북에** 저장하고 Canvas 내 프로젝트 4 페이지를 통해 제출하세요. 반드시 **이름_성_프로젝트4**"(예: J_Smith_Project4)와 같은 명명법을 사용해 노트북을 저장하세요.). ***소규모 그룹은 주피터 노트북을 시작할 때 모든 그룹 구성원의 신원을 확인해야 하며, 각 팀원은 캔버스 내에서 팀 작업의 사본을 제출해야 합니다.***