대출 위험도 예측 모델

프로세스

- 데이터 탐색 (EDA)
  - 가지고 있는 데이터로부터 통찰력(Insight)을 얻는다.
- 라벨(답)의 존재 유무
  - 지도학습/비지도학습 결정
- 라벨의 형태
  - 분류/회귀 모델 결정
- Base-Line 모델로 가장 간단한 머신러닝 모형 구현
  - 아무런 설정없이 생성한 모델
  - EDA/데이터전처리 결과 확인을 위한 모델
  - 모델 선택, 튜닝의 기준이 될 모델
  - Base-Line 모델의 문제점을 파악하여 그것을 개선하는 방향으로 튜닝해 나간다.

프로세스

- 문제에 대한 이해
- 데이터에 대한 이해
- Base-Line 모델을 위한 특성공학(Feature Engineering)
  - 결측치 처리,
  - 이상치 처리
  - 문자열를 실수로 변환
  - Scaling
  - Feature Selection
- Base-Line 모델 선택 및 훈련
- 평가 지표에 맞춰 Base-Line 테스트 및 검증
- 검증 결과를 통해 문제점 발견
  - 데이터 전처리 반복
- Base-Line 모델 최적화
  - 하이퍼파라미터 튜닝
  - 검증 결과에 따라 튜닝 반복

문제에 대한 이해

- 은행에서 대출 대상자 데이터를 기반으로 2년내에 대출금 연체할 가능성이 있는지 여부를 예측하는 알고리즘 개발 의뢰

- 요청 세부사항
  - 대출 요청 고객이 2년내 대출금을 연체할지 여부를 예측하는 모델개발.
  - 현재 수입, 지출 등의 데이터에 대해 은행 자체의 분석을 진행하여 대출자가 미래에 돈을 갚을 수 있는지 확인
  - 수동적이고 시간이 소요되는 이러한 분석을 자동화

- 알고리즘 결과
  - 미래 일정 기간(2년) 내에 채무 불이행 할지 여부

- 평가 지표
  - **roc_auc 점수**

데이터 속성에 대한 이해

- SeriousDlqin2yrs
  - 목표 변수
  - 최근 2년 동안 90일 이상 연체한 적이 있는지 여부
  - 값: 1 (연체한 적 있음), 0 (연체한 적 없음)

- RevolvingUtilizationOfUnsecuredLines
  - 부동산과 할부 부채(installment debit)를 제외한 보유 자산 및 신용 대비 현재 운용할 수 있는 돈의 비율
    - 전체 운용가능한 돈 대비 현재 운용가능한 돈의 비율
  - float
  - ex) 신용카드 총한도가 100만원, 통장 잔액이 200만원인 상황에서 남은 신용카드 한도가 40만원인 경우
    (200만원+40만원)/(200+100) = 220 만원 /300 만원 = 0.8

데이터 속성에 대한 이해

- Age
  - 대출자의 나이
  - Integer
- NumberOfTime30-59DaysPastDueNotWorse
  - 최근 2년 동안 30일 ~ 59일 연체한 횟수
  - Integer
- DebtRatio
  - 전체수입 대비 월 부채 상환과 월 지출 합계의 비율
  - float
  - ex) 수입이 1000만원인 사람이 한달에 300만원 씩 부채를 갚고 있고
    그 외 지출(생활비등)하는 비용이 500만원인 경우
    (300만원 + 500만원)/1000만원 = 800만원/1000만원 = 0.8

데이터 속성에 대한 이해

- MonthlyIncome
  - 월 수입
  - Integer
- NumberOfOpenCreditLinesAndLoans
  - 대출자가 보유중인 담보 대출 및 신용 대출 건수
  - Integer
- NumberOfTimes90DaysLate
  - 과거 90일 이상 연체한 횟수
  - Integer
- NumberRealEstateLoansOrLines
  - 주택 담보 대출을 포함한 부동산 담보 대출 건수
  - Integer

데이터 속성에 대한 이해

- NumberOfTime60-89DaysPastDueNotWorse
  - 최근 2년간 60 ~ 89일 연체한 횟수
  - Integer
- NumberOfDependents
  - 대출자를 제외한 부양가족 수
  - Integer

데이터 분석 절차

- 데이터 읽기
- 데이터셋 크기 파악
- 앞/뒤 일부 데이터를 읽어 데이터 확인
  - 필요 없는 열의 존재 (발견하며 제거)
  - 열이름 변경 필요 여부(필요하면 변경)
  - Index로 사용할 컬럼 존재 (있으면 Index로 변경)

데이터 분석 절차

- EDA (Exploratory Data Analysis – 탐색적 데이터 분석)
  - 데이터가 실제(fact)를 어떻게 표현하고 있는지 이해
- 일반적인 EDA 내용
  - 통계치 확인
  - 결측치 확인 및 처리
    - 결측치 대체 또는 제거
  - 이상치(Outlier) 탐지
    - 이상치 처리
  - 변수간 상관관계 확인

Base-line 모델 선택, 훈련 및 평가지표 확인

- Base-line 모델 목적
  - EDA 결과를 확인
  - 모델 선택과 모델 튜닝의 기준점이 되는 모델
- 모델 훈련 프로세스
  - 패키지 로딩
  - 입력변수, 출력변수 분리
  - 학습/검증용 데이터 분리
  - 모델 정의
  - 모델 학습
  - 평가지표
- 모델 개선
  - 하이퍼파라미터 튜닝