Assignment #2

Due date: 4월 14일 (화)

CommomBank.csv 파일은 *Common Bank*의 예금계좌를 소유하고 있는 5,000명의 고객 정보 데이터로 다음과 같은 정보를 포함하고 있다.

- **ID** : 고객 ID
- **Age** : 나이
- **Experience** : 경력
- **Income** : 연간 수입 ($1,000)
- **ZIP.code** : 우편번호
- **Family** : 가족 구성원 수
- **CCAvg** : 월간 평균 신용카드 사용액 ($1,000)
- **Education** : 1 (Undergraduate), 2 (Graduate), 3 (Advanced)
- **Mortgage** : 주택 담보액
- **PersonalLoan** : 개인 대출 상품 가입 여부(Reject or Accept) (0/1)
- **SecuritiesAccount** : Securities account 소유 여부 (0/1)
- **CDAccount**: CD account 소유 여부 (0/1)
- **Online** : 온라인 뱅킹 사용 여부 (0/1)
- **CreditCard** : Common Bank의 신용카드 소유 여부 (0/1)

*Common Bank*는 현재 예금계좌 소유 고객을 대상으로 개인 대출 상품의 가입을 홍보하기 위한 마케팅을 준비 중이다. 마케팅 부서에서는 어떤 그룹의 고객들을 타겟팅하여 집중적으로 마케팅 예산을 투입할 지를 고민 중이다. 따라서 *k*-NN을 활용하여 새로운 고객의 정보가 주어졌을 때 이 고객이 개인 대출 상품을 가입할 지를 예측해 보고자 한다.

1. 첫 4000명의 데이터를 training set으로, 나머지 1000명의 데이터를 test set으로 사용하자. Training set과 test set에서의 PersonalLoan 값의 분포를 비교해 보자.

2. 7-NN을 적용해 보자. 이때 ID와 ZIP.code는 feature에서 제외한다.

3. 다양한 *k* 값에 대해 *k*-NN을 적용해 보고 test set에 대한 예측 성능을 비교해 보자. *k*가 어떤 값을 가질 때 모델의 성능이 가장 우수한가?

4. Training set에 대해 5-fold cross validation을 5회 반복하여 best *k* 값을 찾아보자. best *k* 값으로 만들어지는 최종 model에 test set을 적용하여 model의 성능을 report하자.

5. 3번과 4번 training 방식의 장단점을 비교해보자.

아래의 파일을 제출해야 합니다.

- R script 파일 (comment 포함)
- 결과 리포트