**[아홉 번째 과제]**

**IT응용시스템공학과**

**1494053**

**김희택**

**7.2 개인 대출 수락: 유니버설 은행은 전체 고객 확보에 관련하여 상대적으로 빠르게 성장하고 있는 신생 은행이다. 은행 고객들 중 대다수는 다양한 규모로 은행과 관계를 맺고 있는 채무 고객(예금자)들이다. 고객들 중 자산이 많은 고객의 수는 많지 않으며, 은행은 대출 사업을 확장하는 데 관심을 갖고 있다. 특히 은행은 부채를 가지고 있는 고객들을 개인 대출 고객으로 전환하기 위하여 다방면으로 노력하고 있다.**

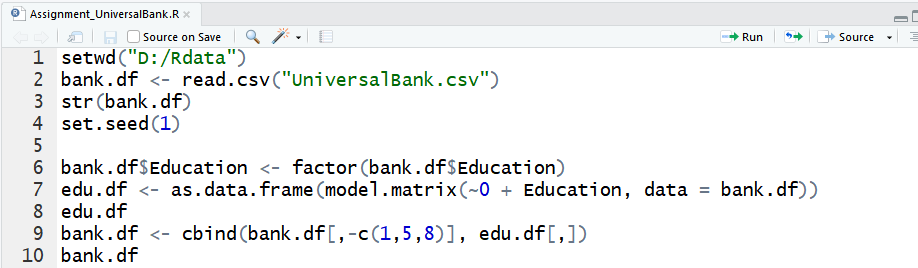
**작년에 은행이 채무 고객들을 대상으로 실시간 캠페인에서는 9% 이상의 성공적인 전환율을 보여주었다. 개인 마케팅 부서는 지난해의 실적에 힘입어 목표 마케팅을 활용한 보다 효과적인 캠페인을 고안하였다. 이 사례에 대한 분석의 목표는 이전 캠페인 데이터를 바탕으로 고객의 행동을 모형화하는 것으로서, 어떠한 요인들의 조합이 고객들로 하여금 개인대출 제의를 더 잘 수락할 수 있게 하는지를 분석하는 것이다. 이는 새로운 캠페인의 디자인을 위한 기초로서 제공될 것이다.**

**제공된 UniversalBank.csv 파일은 5,000명의 고객들에 대한 데이터를 포함하고 있다. 이 데이터는 인구통계학적 정보(나이, 소득 등), 고객과 은행의 관계 정보(담보부채권, 증권계좌 등), 그리고 과거 개인 대출 캠페인에 대한 고객들의 반응(개인 대출로 표기됨) 정보를 포함하고 있다. 5,000명의 고객들 중에 480명 (=9.6%) 만이 이전 캠페인에서 고객에게 제안된 개인대출을 받아들였다.**

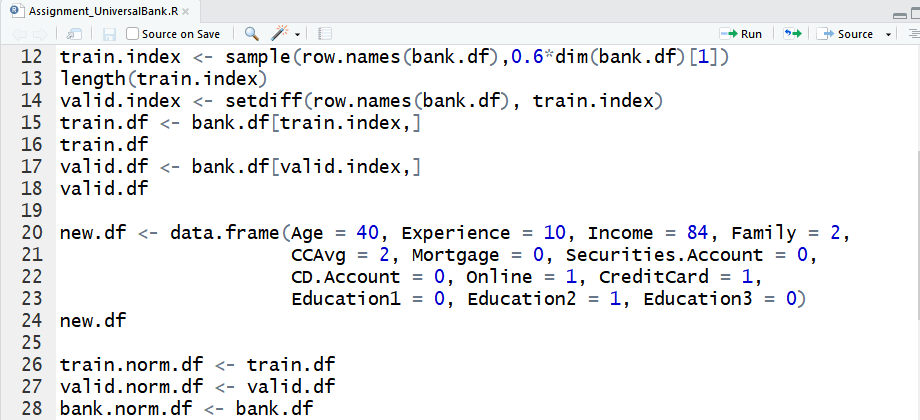
**데이터를 학습 세트 60%와 검증 세트 40%로 분할하시오.**

**a. Consider the following customer**

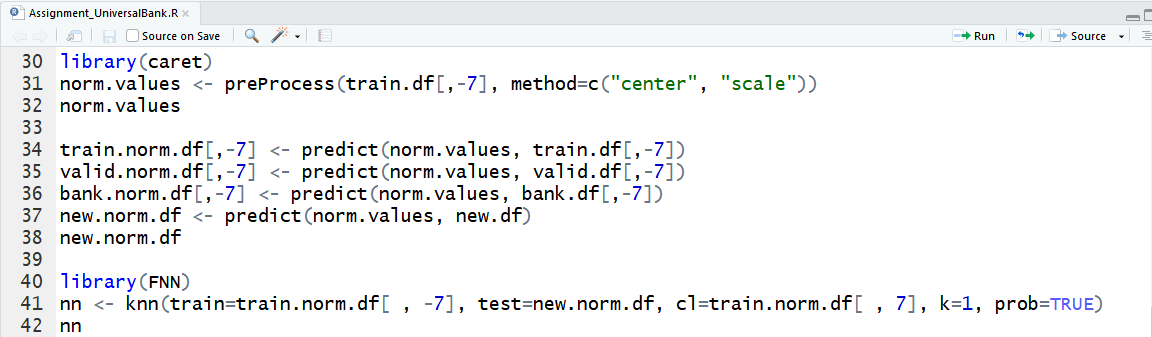
Age = 40, Experience = 10, Income = 84, Family = 2, CCAvg =2, Education\_1 = 0, Education\_2 = 1, Education\_3 = 0, Mortgage = 0, Securities Account = 0, CD Account = 0, Online = 1, and Credit card = 1. Perform a k-NN classification with all predictors except ID and ZIP code using k = 1. Remember to transform categorical predictors with more than two categories into dummy variables first. Specify the success class as 1 (loan acceptance), and use the default cutoff value of 0.5. How would this customer be classified?



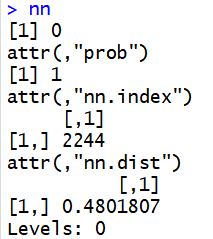
* 먼저 ID와 ZIP code를 제외하고 범주형 변수인 Education에 대해 가변수를 만들었습니다.



* 데이터를 학습 세트 60%와 검증 세트 40%로 분리하였습니다. 또한 새로운 데이터를 추가한 뒤, 정규화 작업을 진행하였습니다.

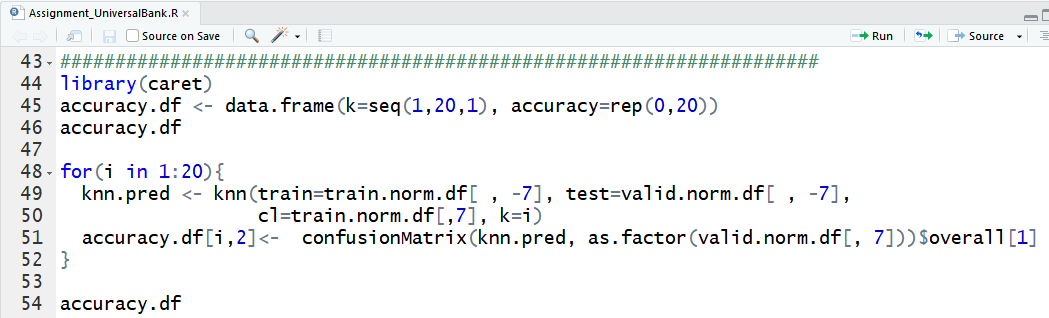


* 개인 대출 여부를 결정하는 Personal Loan이 목표변수입니다. 7번째 열인 Personal Loan 을 제외하고 predict 함수를 사용하여 데이터를 정규화 하였습니다. 또한, k를 1로 설정하여 k-최근접이웃을 실행하였습니다.

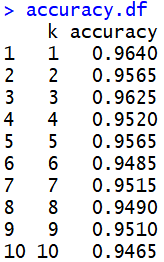
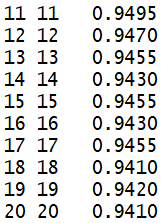


* 결과를 확인해보면, 대출수락을 하지 않는 것으로 볼 수 있습니다.

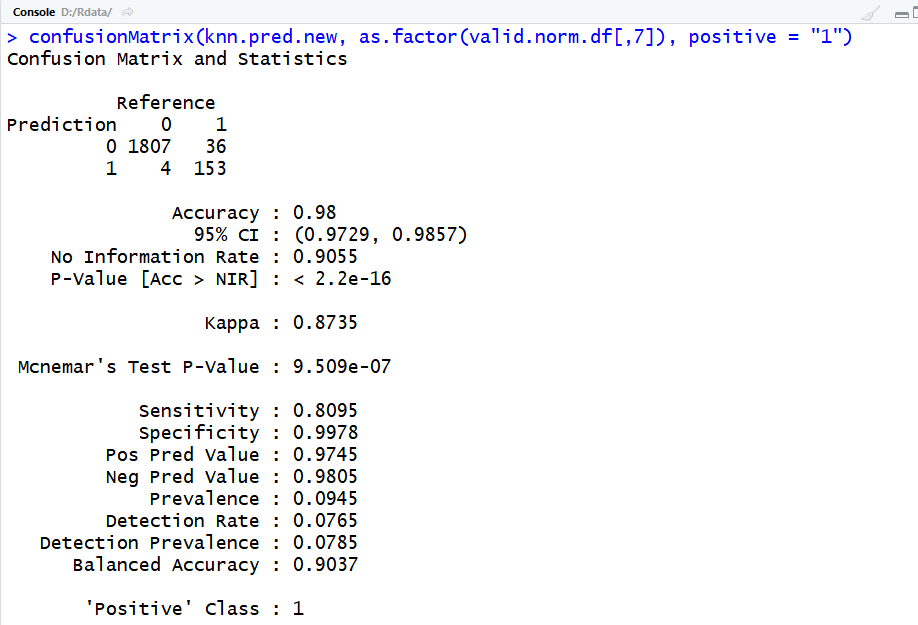
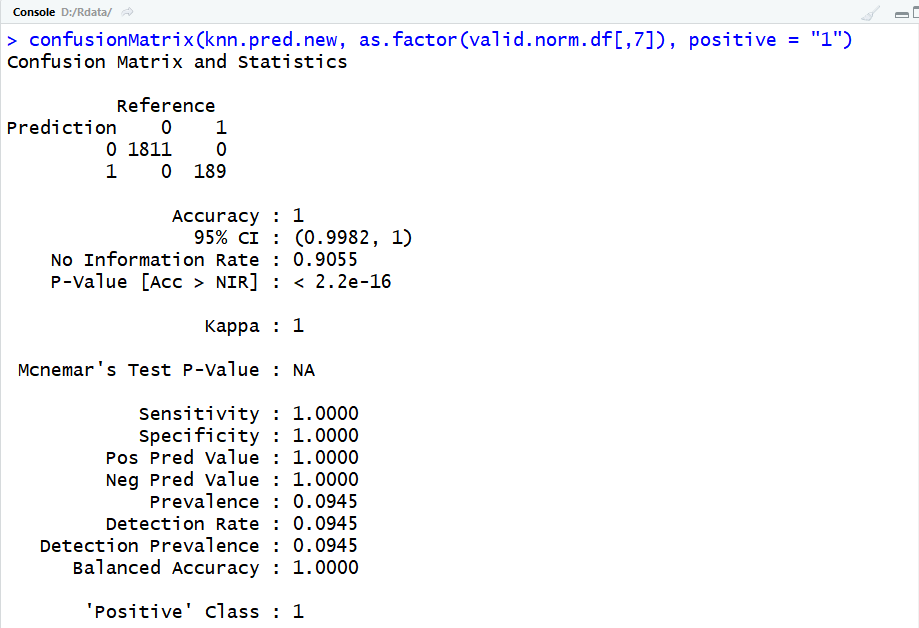
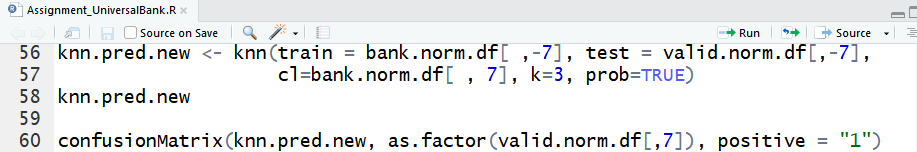
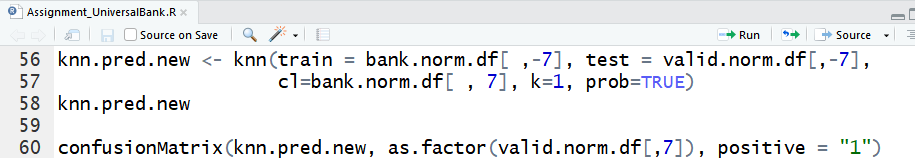
**c. 최적의 k를 사용하여 검증 세트에 대한 정오행렬표를 만드시오.**



* 먼저 최적의 k를 구하기 위해 accuracy.df를 구하였습니다.

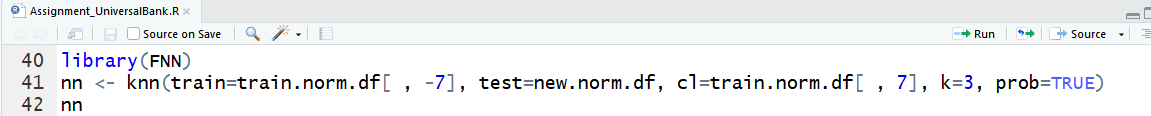
* 즉 k가 1, 3일 때, 최적인 것을 알 수 있습니다.



* k가 1, 3일 때, k-최근접이웃을 적용하여 confusionMatrix로 검증 세트에 대한 정오행렬 표 결과입니다.
* k가 작으면 노이즈(측정오차)가 발생할 가능성이 크므로, **가장 최적의 k는 3입니다.**

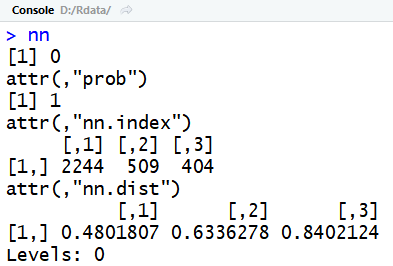
**d. Consider the following customer**

Age = 40, Experience = 10, Income = 84, Family = 2, CCAvg = 2, Education\_1 = 0, Education\_2 = 1, Education\_3 = 0, Mortgage = 0, Securities Account = 0, CD Account = 0, Online = 1 and Credit Card = 1. Classify the customer using the best k.



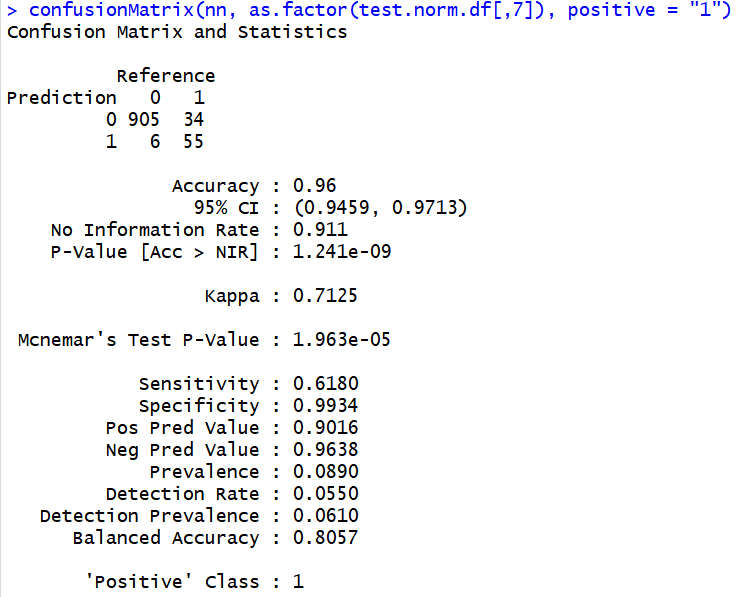
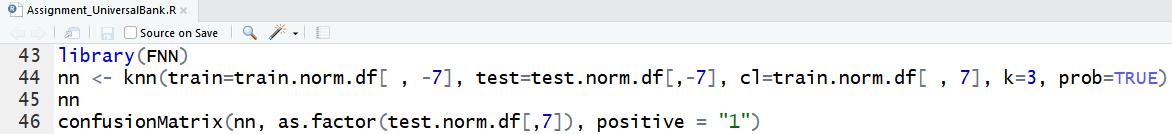
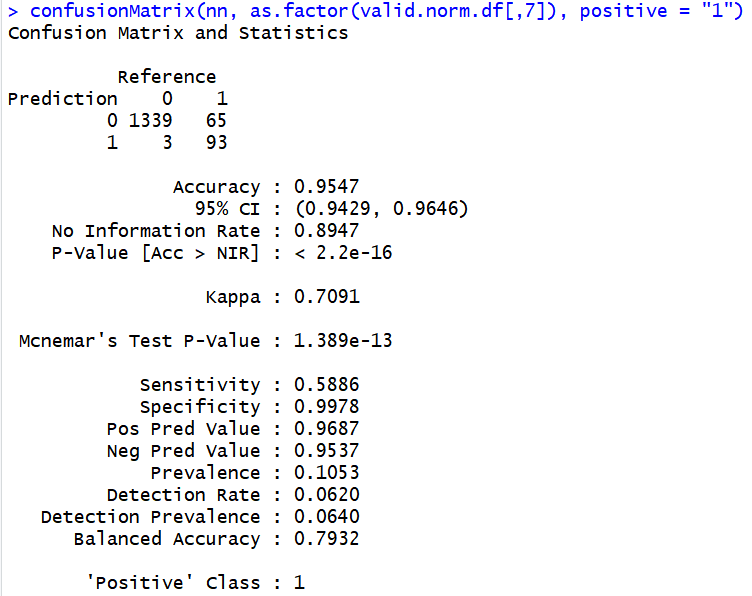
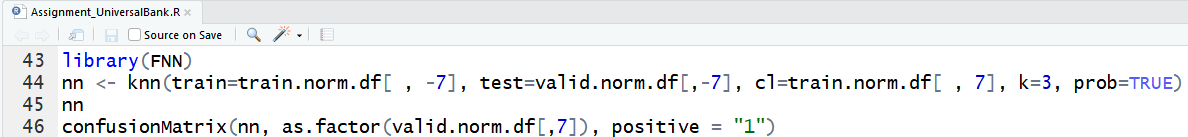
**베스트 k를 찾았으면, train.norm.df 가 아닌 bank.norm.df와 비교를 해야 한다.**

**knn(train=bank.nomr.df[,-7] …\_**



* k가 3일 때, 가장 가까운 데이터는 2244번이며, 그 거리는 0.4801807 이다. 새로운 데이터는 대출 수락을 하지 않는 것으로 확인할 수 있습니다.
* 클래스는 ‘0’에 속한다. 3/3 = 1 즉, 확률은 1이다. 이로써 대출을 받지 않는 것을 알 수 있다.

**e. 이번에는 데이터를 학습용, 검증용, 그리고 평가용 세트로 다시 분할하시오. (50%: 30%: 20%) 위에서 선택된 k를 사용하여 k-최근접이웃을 적용하시오. 평가 세트에 대한 분류행렬을 학습 세트 및 검증 세트의 정오행렬표와 비교하시오. 차이점을 찾아내고 그 이유에 대하여 설명하시오**



* 데이터를 학습용, 검증용, 평가용 각 50%, 30%, 20%으로 분할하였습니다. 위에서 선택된 k인 3을 이용하여 k-최근접이웃을 사용하였습니다. 위 결과는 평가 세트에 대한 분류행렬을 학습 세트와 검증 세트의 정오행렬표와 비교한 결과입니다. Accuracy가 학습 세트일 경우 더 높은 것을 확인할 수 있습니다. 즉 좋은 모델인 것을 알 수 있습니다.
* 0.9547, 0.96 두 개의 결과는 매우 비슷하다. 현재 만들어놓은 분류기는 크게 영향을 받지 않는 것을 알 수 있다.