3주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 1-3주차 패키지 문제의 조건 및 힌트는 R을 기준으로 하지만, Python을 사용해도 무방합니다.
- 제출형식은 R markdown으로 HTML, PDF 모두 가능합니다. **.R이나 .ipynb 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 5시 00분에 발표됩니다.
- 제출기한은 **목요일 자정까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요.

Chapter 1 모델링을 위한 전처리

3주차 패키지에서는 로지스틱 회귀 모델을 통해 분류 예측을 해보도록 하겠습니다. 분류 모델은 Y값이 이산형인 데이터일 때 어느 그룹 또는 레이블에 속하는지 찾아내는 모델입니다. 이번주에는 신용 등급(credit 변수)을 1또는 0으로 예측하는 분류 예측을 진행하겠습니다. 이번 챕터에서는 먼저 모델링을 위한 간단한 전처리를 통해 데이터를 정제해보겠습니다.

조건: tidyverse, data.table, magrittr, caret 패키지 사용 (이외의 패키지 사용 금지), %>% 연산자를 최대한 사용하여 한 줄의 코드로 표현

문제0. (기본 세팅) 패키지를 불러오고 디렉토리를 설정하세요.

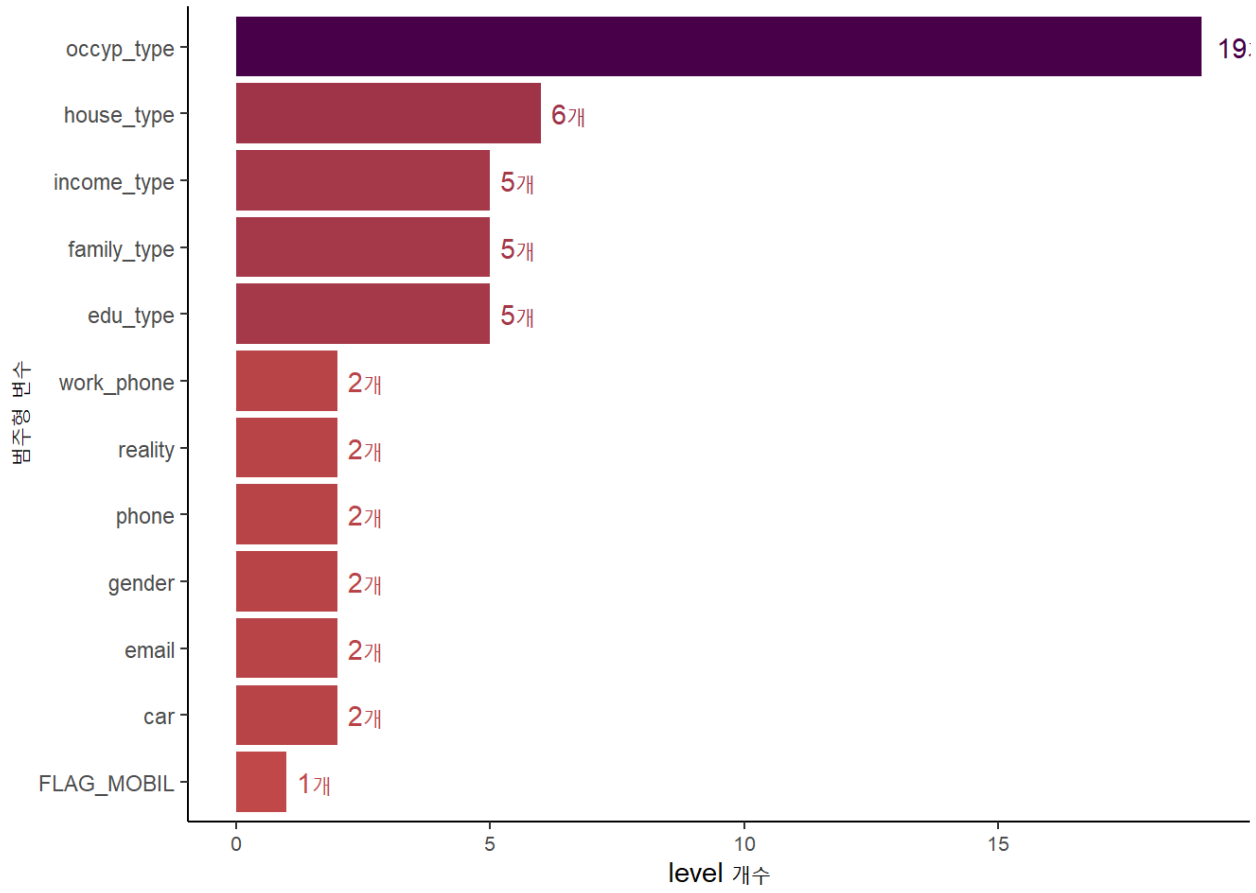
문제1. train.csv와 test.csv 데이터의 기본 구조를 파악하고 변수와 데이터 개수, 결측치 여부를 확인하세요.

문제2. character 형태로 되어있는 변수들을 factor 형태로 바꾸세요.

문제3. 변수 의미 txt파일을 참고하여 남은 범주형 변수들도 factor 형태로 바꾼 뒤 str을 확인해주세요.

문제4. Factor형 형태의 변수들의 각 level이 몇 개인지 개수를 확인하고 다음과 같이 그래프를 그려주세요.

(n_distinct 함수 사용, hjust=-0.3, 그라데이션 색상 : high:#480048", low:"#C04848")



문제4-1. 그래프를 통해 필요 없는 변수를 확인하고 삭제하세요.

문제5. days_birth는 데이터 생성일로부터 몇일 전 태어났는지를 역으로 세는 변수입니다. (-1은 데이터 수집일 하루 전에 태어났음을 의미) 이 변수를 사용하여 나이(AGE) 파생변수를 생성하고 기존 변수는 삭제하세요. (반올림 사용 가능)

문제6. days_employed는 데이터 생성일로부터 몇일 전 업무를 시작했는지를 나타내는 변수입니다. (-1은 데이터 수집일 하루 전부터 일을 시작했음을 의미, 다만 양수 값은 고용되지 않은 상태를 뜻함) 이 변수를 사용하여 업무 년차(YEARS_EMPLOYED) 파생변수를 생성하고 기존 변수는 삭제하세요.

문제7. Test 데이터셋도 같은 방식으로 전처리 해주세요.

문제8. train 데이터를 학습용 데이터와 검증용 데이터로 분리하세요. (p=0.8, seed:123)

Chapter 2 분류모델 : 로지스틱 회귀

이번 챕터에서는 로지스틱 회귀 모델을 통해 'credit'변수에 대한 이진분류 성능을 평가해보겠습니다. 로지스틱 회귀는 가장 기본적인 분류 모형으로 회귀를 사용하여 어떤 범주에 속할 확률을 예측하고 그 확률에 따라 가능성이 더 높은 범주에 속하는 것으로 분류하는 알고리즘입니다.

또한 로지스틱 회귀는 일반선형회귀(OLS)와 마찬가지로 Ridge나 Lasso penalty를 적용할 수 있습니다. 두가지 방법을 모두 사용하여 모델링을 진행하고 결과를 비교해보도록 하겠습니다.

조건: glmnet, Epi, MLmetrics 패키지 사용

[로지스틱 회귀]

문제1-1. 전체 변수들을 가지고 로지스틱 회귀 모델을 만들고 결과를 보여주세요.

문제1-2. 변수선택법을 적용해보세요. 결과를 보여주고 문제1의 모델과 비교해주세요. *(어떤 알고리즘을 사용하든 정답은 없습니다. 사용한 방식이 무엇인지, 왜 이 방법을 사용했는지 설명해주세요.)*

문제1-3. 모델의 회귀계수의 신뢰구간을 구해보세요.

문제1-4. 오즈비와 회귀계수의 관계를 이용하여 회귀계수를 해석해보세요. *(변수들이 많기 때문에 모든 변수들을 일일이 말로 해석할 필요는 없습니다. 코드로 어떻게 구현하고 이를 어떻게 해석하면 되는지 예시로 하나정도만 설명해주시면 됩니다.)*

문제1-5. 0.5를 임계값으로 모델의 예측값(train error)을 구하고 confusion matrix를 만들어보세요.

문제1-6. Validation data를 통해 확률값이 나오도록 예측값을 구하고 이를 사용하여 ROC curve를 그리고 해석해보세요. (Epi 패키지 사용)

문제1-7. 위의 ROC curve에서 구한 최적의 임계값을 기준으로 Accuracy와 F1-score를 구하고 값을 저장해두세요. *(이후 세 모델의 비교를 위한 시각화에 사용될 것입니다. 두 metric은 직접 계산해도 되고 패키지를 사용해도 됩니다.)*

문제1-8. 같은 조건으로 전체 데이터를 다시 로지스틱 회귀 모형을 적합시키고 test 데이터셋에 대해 예측하세요.

[Lasso 로지스틱 회귀]

문제2-1. 범주형 변수들이 더미화된 디자인 행렬을 만드세요. *(model.matrix() 사용)*

문제2-2. CV로 최적의 람다를 찾고 찾은 최적의 람다로 Lasso 로지스틱 회귀 모델을 적합하세요. (seed:123)

문제2-3. 모델의 회귀계수를 확인하고 회귀계수가 없는 변수들이 있는 이유를 설명해주세요.

문제2-4. Validation 데이터를 통해 확률값이 나오도록 예측값을 구하고 이를 사용하여 ROC curve를 그리고 해석해보세요.

문제2-5. 위의 ROC curve에서 구한 최적의 임계값을 기준으로 Accuracy와 F1-score를 구하고 값을 저장해두세요. *(이후 세 모델의 비교를 위한 시각화에 사용될 것입니다. 두 metric은 직접 계산해도 되고 패키지를 사용해도 됩니다.)* .

문제2-6. 같은 조건으로 전체 데이터를 다시 Lasso 로지스틱 회귀 모형을 적합시키고 test 데이터셋에 대해 예측하세요.

[Ridge 로지스틱 회귀]

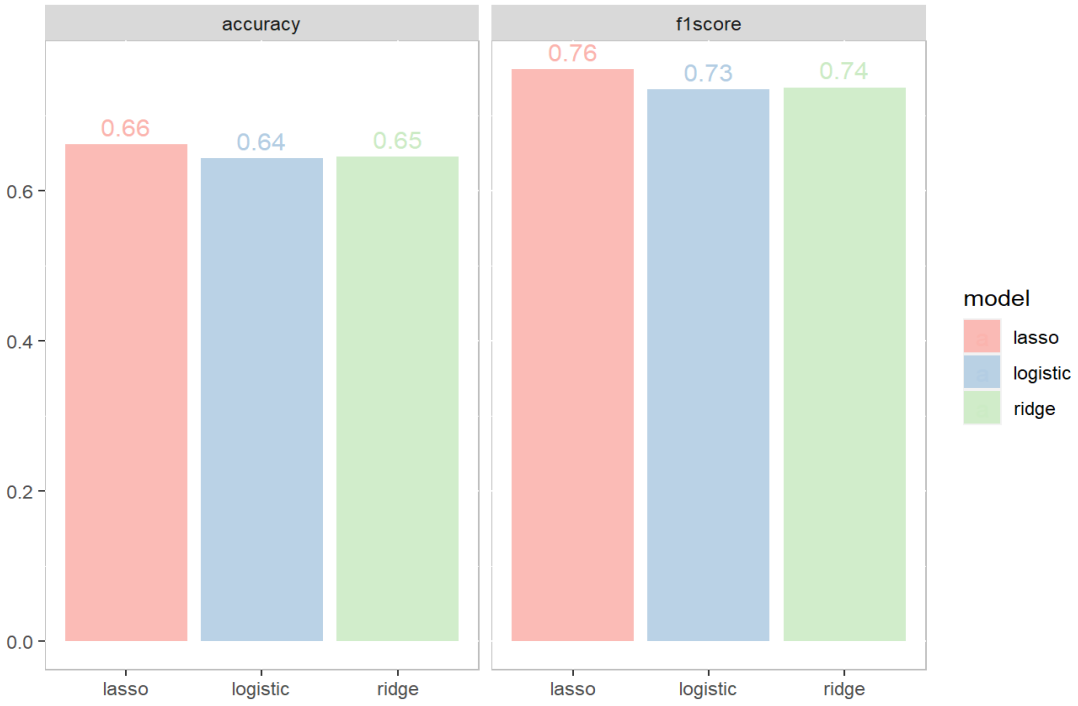
문제3-1. Lasso 로지스틱 회귀에 사용한 동일한 데이터를 사용하여 CV로 최적의 람다를 찾고 찾은 최적의 람다로 Ridge 로지스틱 회귀 모델을 적합하고 모델의 회귀계수들을 확인하세요. (seed:123)

문제3-2. Validation 데이터를 통해 확률값이 나오도록 예측값을 구하고 이를 사용하여 ROC curve를 그리고 해석해보세요.

문제3-3. 위의 ROC curve에서 구한 최적의 임계값을 기준으로 Accuracy와 F1-score를 구하고 값을 저장해두세요. *(이후 세 모델의 비교를 위한 시각화에 사용될 것입니다. 두 metric은 직접 계산해도 되고 패키지를 사용해도 됩니다.)*

문제3-4. 같은 조건으로 전체 데이터를 다시 Ridge로지스틱 회귀 모형을 적합시키고 test 데이터셋에 대해 예측하세요.

문제3-5. 각각 세 모델의 Accuracy값과 F1score 값을 다음과 같이 시각화하고 결과를 해석해보세요. *(gather 사용 시 편리, alpha=0.9, palette='Pastel1')*



Chapter3 클러스터링

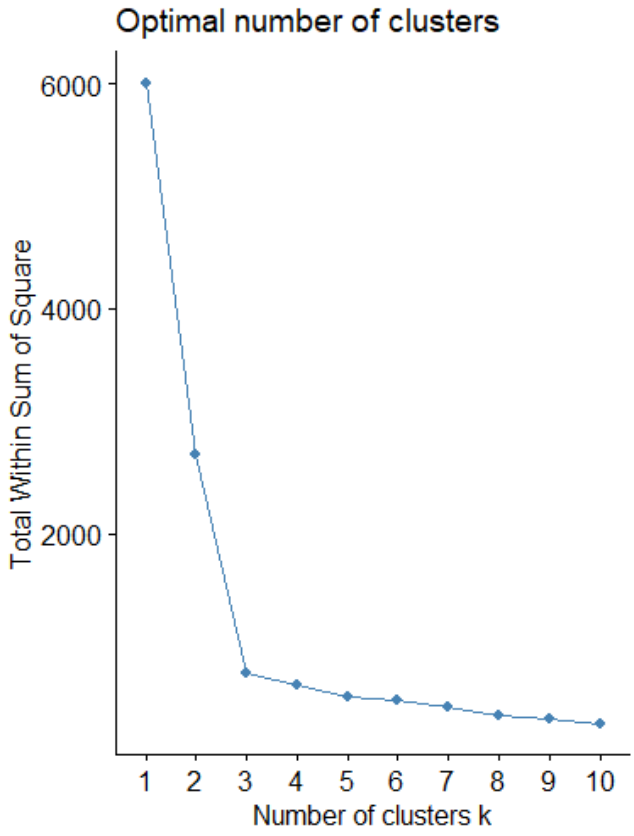
지금까지는 y값을 예측하는 지도학습에 대해 알아보았습니다. 이번 챕터에서는 비지도학습의 대표적인 모델인 K-means Clustering을 해보겠습니다. 거리를 기반으로 특성이 비슷한 데이터들을 묶어주는 군집화 방법으로, 변수들을 해석하기 위해 자주 사용합니다. 데이터는 cluseter 패키지에 내장되어 있는 xclarara 데이터를 사용하도록 하겠습니다.

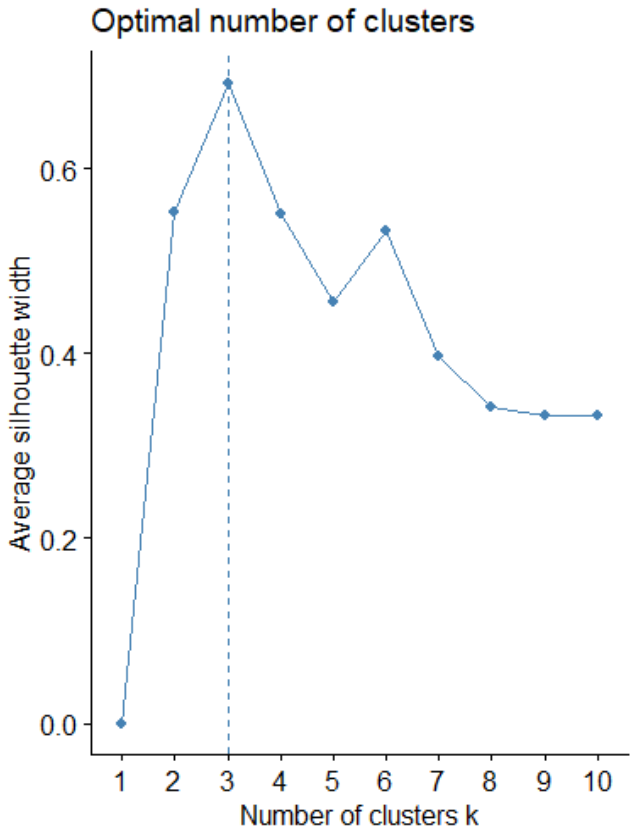
조건: caret, corrplot, cluster, factoextra, gridExtra 패키지 사용

문제1. 환경 내 저장된 데이터를 전부 삭제하고 cluster 패키지의 xclara 데이터를 불러오세요.

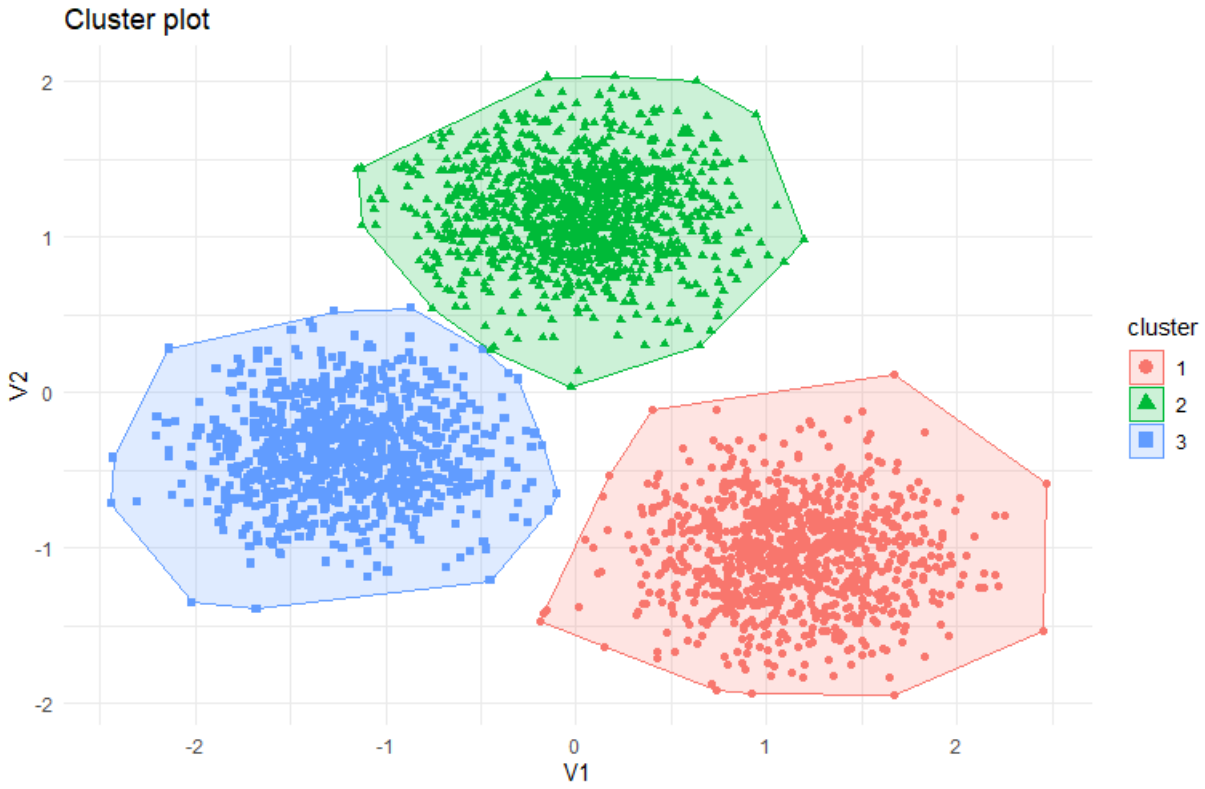
문제2. 데이터의 상관관계를 확인하고 스케일링을 해주세요. 또한 클러스터링 전에 데이터를 스케일링 해주어야 하는 이유를 적어주세요.

문제3. Fviz_nbclust 함수로 다음과 같이 시각화 한뒤 적절한 k 값을 선택하고 그 이유를 설명해주세요. (seed:123)





문제4. K-means clustering을 진행하고 다음과 같이 시각화하세요. (nstart = 1, iter.max = 100)



문제5. 사용된 변수 V1과 V2에 대해 다음과 같이 클러스터별로 박스 플랏을 시각화하여 비교하세요.