클린업 1주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 클린업 1-3주차 패키지 문제의 조건 및 힌트는 R을 기준으로 하지만, Python을 사용해도 무방합니다.
- 제출형식은 HTML, PDF 모두 가능합니다. **.ipynb 이나 .R 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 16시 00분에 발표됩니다.
- 제출기한은 **목요일 자정까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요. 또한 불성실한 제출로 인한 경고를 2회 받으실시도 퇴출이니 유의해주세요.

Chapter 1 : Data Preprocessing

이번 클린업 1주차에서는 주어진 데이터를 바탕으로 tidyverse 패키지를 이용해서 데이터를 전처리하고 시각화를 해보겠습니다. Tidyverse를 파이프(%>%) 연산자를 이용해서 함수를 직관적으로 표현할 수 있는 것이 특징이며,

TIdyverse 안에는 dplyr, magrittr, ggplot2 등 다양한 패키지가 포함되어 있습니다. 데이터 전처리부터 시각화까지 이번 주에는 TIdyverse를 이용해 봅시다!

문제1. Train, test 데이터를 불러온 뒤 데이터의 구조를 파악하세요.

(HINT) str, head, describe(rms package) 등을 활용하면 결측치와 데이터의 타입들을 쉽게 확인할 수 있습니다.

문제2. Train과 Test 데이터에서 Id, V1 변수는 필요하지 않으니 해당 열을 삭제해주세요.

(HINT) dplyr의 select와 magrittr의 %<>%를 사용하면 쉽게 열을 제가할 수 있습니다.

문제 3. 현재 변수 이름이 X.(변수이름). 구조로 읽기 어렵게 되어있으므로 (변수이름) 형태로 변환해주세요.

(Example) X.checking_status. -> checking_status

(HINT) strsplit 함수를 이용하면 원하는 split 단위를 선택할 수 있습니다.

문제4. personal_status는 (성별) (결혼상태) 구조로 되어있습니다. 이를 sex와 marital_status로 파생변수를 만들고 personal_status는 제거해주세요.

(HINT) 위에서 사용한 strsplit 함수와 dplyr의 mutate 함수를 이용하면 쉽게 새로운 파생변수를 만들 수 있습니다.

문제5. checking_status 열에서 no_checking, savings_status 열에서 no known savings을 결측값으로 바꿔주세요

문제6. credit_history에서 no credits/all paid을 no credits으로 marital_status에서 div/dep/mar, div/sep, mar/wid를 div/dep/mar로 만들어주세요.

문제7. 구간에 따라 나눠져 있는 변수들을 순서에 맞게 1, 2, 3, … 으로 바꿔주세요.

- 구간에 따라 나눠져 있는 변수는 checking_status, savings_status, empolyment입니다.

(+ 보너스 문제 1) 주어진 변수를 순서형 변수로 바꾸지 않고, as.factor를 이용해서 명목형 변수로 바꾸었을 때 문제점에 대해서 간단히 서술해주세요.

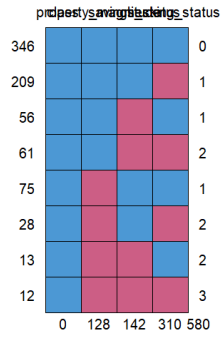
(+ 보너스 문제 2) employment에서 unemployed 변수는 결측치 처리를 하지 않았는데, 그 이유를 설명해주세요

문제8. 문자형 변수는 factor로, 정수형 변수는 numeric으로 바뀌주세요.

문제9. 앞선 했던 작업들을 test 데이터에서도 진행해주세요.

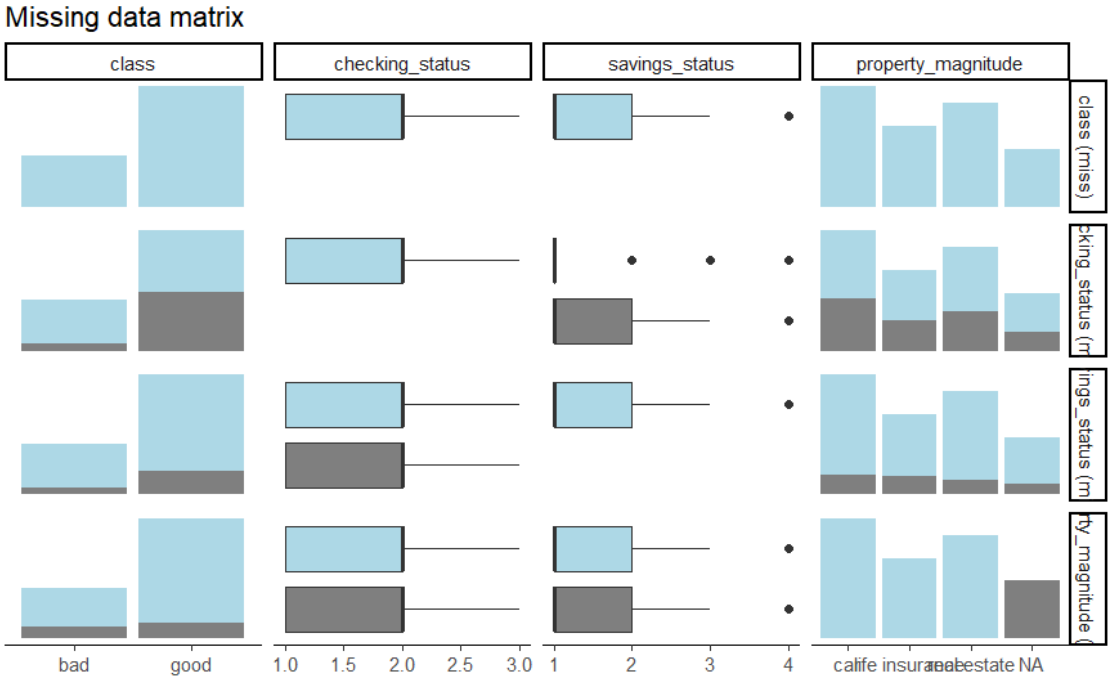
문제10. 아래와 같이 결측치의 패턴을 확인할 수 있는 plot을 그려보고 결측치 발생 여부에 대해 패턴이 존재하는지 확인해주세요. 변수 순서는 class, checking_status, savings_status, property_. 순입니다.

(HINT) mice 라이브러리의 md.pattern을 이용하면 아래의 plot을 손쉽게 그릴 수 있습니다.



문제11. 아래와 같이 결측치 여부에 따라 변수 간의 간략한 분포를 얻을 수 있는 plot을 그리고 이를 해석해주세요.

(HINT) mice 라이브러리의 missing_paris을 이용하면 아래의 plot을 손쉽게 그릴 수 있습니다.



문제9. 결측치 보간을 위한 다양한 방법이 있지만 여기서는 checking_status와 savings_status의 조합 중 최빈값으로 결측값 대체를 해주세요. Property_magenitude는 변수 그 자체의 최빈값을 대체해주세요.

(HINT) table 함수를 이용하면 각 변수가 같이 발생한 빈도를 구할 수 있습니다.

Chapter 2 : EDA

좋은 모델링을 하기 위해서는 변수의 특성을 파악하는 것이 중요합니다. 이번 EDA 파트에서는 변수 분포의 시각화를 중심으로 다루어 보겠습니다.

문제1. 각 변수의 특징을 알기 위해서 아래와 같이 변수의 분포를 시각화해주세요.

- 범주형 자료는 barplot을 이용해서, 연속형 자료는 히스토그램과 밀도함수 둘 다 이용해서 그려주세요.
- 범주가 3개 이하인 변수에서는 "#00AFBB", "#E7B800", "#999999" 색을 이용해주세요.
- 범주가 4개 이상인 변수에서는 palette “Set 3”를 이용해주세요.
- 연속형 변수에서는 #00AFBB 색을 이용해주세요.
- 테두리가 있는 경우에 테두리 color는 모두 ‘black’ 입니다.
- 연속형 변수에서 히스토그램의 alpha = 0.5, 밀도 함수의 alpha = 0.2입니다.

(HINT1) 색은 scale_fill_manual, scale_color_manual, scale_fill_discrete_qualitative 혹은 ggplot에 직접 색을 입력하는 할 수 있습니다.

(HINT2) grid extra 패키지의 grid.arrange를 이용하면 쉽게 그린 plot을 합칠 수 있습니다.

(+ 보너스 문제 3) 위의 시각화는 P(X)를 나타낸 것입니다. 클래스에 따른 변수의 분포를 보기 위해 P(X|Y)를 자유롭게 시각화해주세요.

Chapter3 모델링

이번에 사용할 모델은 랜덤포레스트와 로지스틱 회귀입니다. 랜덤포레스트는 여러 개의 트리모델을 앙상블한 모델 중의 하나입니다. 로지스틱은 일반화 선형 모델의 일종으로 분류 문제에 주로 사용되는 모델입니다. 두 모델에 대해 간략하게 알아보고 test를 예측해 보겠습니다.

문제1. 랜덤포레스트 패키지를 불러와주세요.

문제2. 랜덤포레스트 모델에는 여러가지 하이퍼 파라미터가 존재합니다. 각 하이퍼 파라미터에 대해서 설명해 주세요.

문제3. Ntree를 100으로 설정한 다음에 train 데이터로 학습을 한 후 변수별 중요도를 Importance plot으로 시각화하고 어떤 변수가 대출 승인에 영향을 미치는 지 설명해주세요.

(HINT) VarImpPlot을 이용하면 쉽게 random forest의 variable importance를 구할 수 있습니다.

문제4. 로지스틱 함수로 모든 변수를 다 넣고 돌린 다음에 summary 함수를 이용해서 중요하지 않은 변수를 선택해주세요.

문제5. CORElearn 패키지를 불러오고 릴리프(Relief) 알고리즘을 이용해 변수 별 중요도를 시각화해주세요.

(+ 보너스 문제 4) 릴리프 알고리즘에 대해 간단하게 설명해주세요.

문제6. Random forest의 variable importance와 로지스틱 함수의 summary 결과, 릴리프 알고리즘의 변수 별 중요도를 이용해 자유롭게 변수를 선택하고 Test Data를 예측해주세요.

문제7. 예측한 Test Data와 실제 Test Data 값의 RMSE(Root Mean Square Error)를 구해주세요.