P-SAT

2주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 1-3주차 패키지 문제의 조건 및 힌트는 R을 기준으로 하지만, Python을 사용해도 무방합니다.
- 제출형식은 피피티(템플릿은 P-SAT기본 템플릿), 마크다운, HTML, PDF 모두 됩니다. .R이나 .ipynb 등의 소스코드 파일은 안됩니다. 완료 후 psat2009@naver.com로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는시간 이후에 하게 되며, 역시 랜덤으로 5시00분에 발표됩니다.
- 제출기한은 목요일 자정까지이며 지각시 벌금 5000원이 있습니다. 미제출 시 만원입니다. 패키지 2회 무단 미제출 시 퇴출이니 유의해 주세요.

Chapter 1. 모델링을 위한 전처리

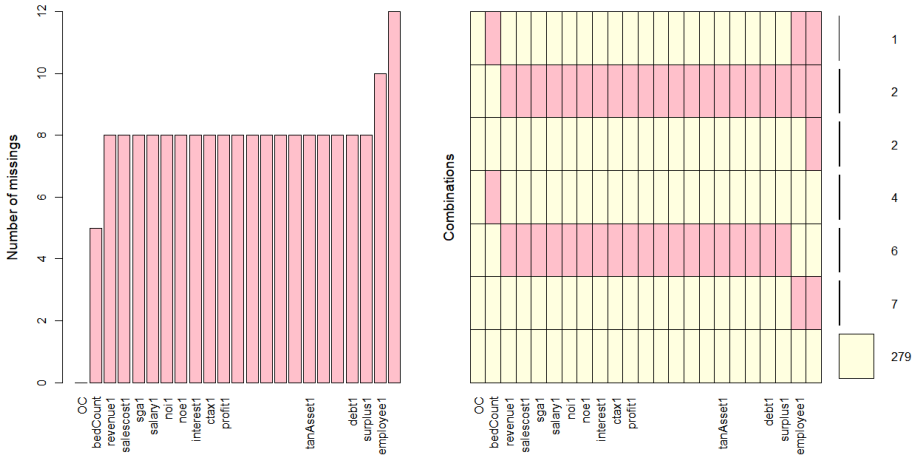
이번주는 모델링의 기본 프로세스를 익혀보겠습니다. 지도학습인 분류/회귀 모델을 hold-out 및 cv를 이용하여 평가하는 것이 목표입니다. chapter1에서는 저번주에 배운 tidyverse를 통한 전처리 방법을 복습해보며 분류 모델링에 쓰일 데이터를 정제해보겠습니다.

[조건: tidyverse, data.table, VIM 패키지 모두 사용(이외 패키지 금지), %>% 최대 활용]

문제 0 기본 세팅. tidyverse, data.table, VIM 패키지를 부른 후, setwd로 'data.csv'가 있는 폴더로 경로를 설정하고, fread로 'data.csv'를 불러오세요.

문제 1. '2'로 끝나는 변수를 모두 제거하세요. (단순히 변수의 이름 또는 인덱스 열거를 통한 제거가 아닌 적절한 함수를 사용할 것) ('2'로 끝나는 변수는 2016년 기준 값이고, '1'은 2017년 기준입니다. 정확한 모델링을 위해서는 이를 모두 고려해야 하지만, 전처리 방법 학습 및 간단한 모델링을 위해 2017년 기준 변수만 사용하도록 하겠습니다.)

문제 2. 'VIM' 패키지를 이용하여 다음과 같이 시각화 한 후 간단히 해석해 보세요. (아래처럼 변수가 모두 출력이 안될 수 있습니다. 사용 색: pink, lightyellow)



문제 3-1 NA imputation. 숫자 데이터의 NA값을 mean imputation을 통해 채우세요. (lapply 이용할 것)

문제 3-2 NA imputation. 범주 데이터의 NA값을 mode imputation을 통해 채우세요.

문제 4. 변수 'OC(병원 개/폐업 여부)'를 타겟 변수로 사용하기 위해 "open"을 1, "close"를 0으로 바꾸세요.

문제 5. 숫자 데이터 중 integer 자료형인 경우 num(numeric) 자료형으로 바꾸세요. (lapply 이용할 것)

Chapter 2. 분류모델

공모전 데이터 등 대부분의 경우 test set의 타겟 값이 주어지지 않기 때문에, validation 데이터셋을 통하여 test set의 모델의 성능을 간접적으로 평가하여 선택하는 것이 필요합니다. 이번 챕터에서는 **Hold-out과 CV 방식**을 통해 **로지스틱 회귀**와 트리기반 모델인 **랜덤포레스트**의 'OC'변수에 대한 이진분류 성능을 간접적으로 평가해 보겠습니다. (각 모델은 이번주 범주형자료분석 및 데이터마이닝 클린업에서 다룰 예정입니다.)

랜덤포레스트와 같이 모델의 파라미터가 많은 경우 최적의 파라미터 조합을 찾는 것이 필요한데, 이를 그리드 서치라고 합니다. 각 파라미터 조합 별로 CV를 진행하여 가장 좋은 성능을 보이는 파라미터 조합을 선택하는 것이 일반적입니다. 이번 챕터에서는 랜덤포레스트에 대한 **그리드서치 5-fold CV**를 해보겠습니다.

[조건: caret, MLmetrics, randomForest 패키지 모두 사용]

[모델1 로지스틱 회귀]

문제 1. 앞서 전처리한 데이터를 createDataPartition으로 7:3으로 train과 validation set으로 나누세요. (seed :1234, p: 0.3)

문제 2 Hold-out. train 데이터의 모든 변수를 이용하여 'OC'를 타겟으로 하는 로지스틱 회귀를 만들고 validation set의 Accuracy값을 구하세요.

문제 3 Feature selection & Hold-out. 변수선택법 중 ~~다중선택법~~을 이용하여 변수를 선택하고, 선택된 변수들로 로지스틱 회귀를 만들어 validation set의 Accuracy값을 구하세요.

[모델2 랜덤포레스트]

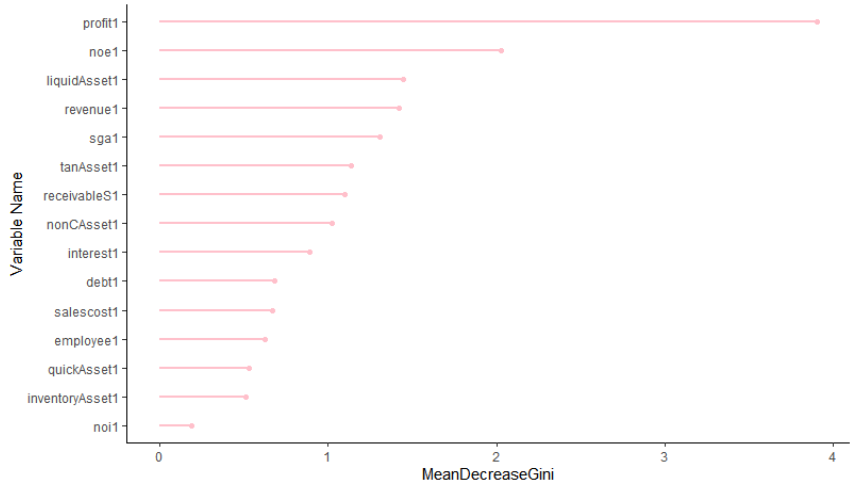
문제 4. mtry에 대한 그리드서치를 위해 expand.grid를 이용하여 다음과 같은 데이터 프레임을 만드세요. (데이터 프레임 명: acc_rf)

```
> acc_rf
  mtry acc
1    3  NA
2    4  NA
3    5  NA
```

문제 5. 로지스틱회귀에서 선택된 변수들로 랜덤포레스트에 대한 5-fold CV 그리드서치를 진행하여 acc_rf의 acc 변수에 해당 Accuracy값을 넣으세요. (ntree 파라미터를 10으로 설정하고, 이중 for문을 이용하여 직접 코드를 짤 것)

문제 6. acc_rf에서 가장 높은 Accuracy값의 행을 출력하세요.

문제 7. 가장 좋은 파라미터 조합으로 랜덤포레스트 모델을 학습시킨 후, varImpPlot과 ggplot을 이용해 다음과 같이 시각화 하여 이를 기반으로 모델을 해석해주세요. (사용 색: pink)



Chapter 3. 회귀모델

저번주에는 **'Boston' 데이터의 medv**에 대한 회귀분석을 하여 test set에 대한 RMSE값을 구했습니다. 이번 챕터에서는 같은 데이터로 **랜덤포레스트**를 이용하여 test set에 대한 RMSE를 구해봅시다.

분류를 위한 랜덤포레스트 모델링과 마찬가지로 **그리드서치 5-fold CV**를 통해 파라미터 조합을 선택한 후, 최종적으로 test set에 대한 RMSE를 구하는 것이 목표입니다.

[조건: MASS, caret, MLmetrics, randomForest 패키지 모두 사용]

문제 1. Boston 데이터를 8:2로 train과 test set으로 나누세요. (p: 0.2)

문제 2. expand.grid를 이용하여 다음과 같은 데이터 프레임을 만드세요. (데이터 프레임 명: RMSE_rf)

	mtry	ntree	RMSE
1	3	10	NA
2	4	10	NA
3	5	10	NA
4	3	100	NA
5	4	100	NA
6	5	100	NA
7	3	200	NA
8	4	200	NA
9	5	200	NA

문제 3. medv를 종속변수로 하는 랜덤포레스트에 대한 5-fold CV 그리드서치를 진행하여 RMSE_rf의 RMSE 변수에 해당 RMSE값을 넣으세요. (이중 for문을 이용하여 직접 코드를 짤 것)

문제 4. RMSE_rf에서 가장 낮은 RMSE값을 가진 행을 출력하세요.

문제 5. train set으로 그리드 서치로 나온 가장 좋은 조합의 파라미터의 랜덤포레스트를 학습시킨 후, test set의 RMSE를 구하세요.