

goophi

2022.08.22.

Table of contents

1	Introduction	1
2	Import sample data	1
3	Data Setup Tab	2
4	Modeling with CV Tab	3

```
library(goophi)
```

1 Introduction

1) 본 문서는 goophi 패키지를 Shiny app에서 사용하는 것을 상정해 작성했습니다.

2) 본 문서의 케이스 스타일은 Camel case와 Snake case가 혼용되어 있습니다.

- Camel case : goophi의 함수명 및 파라미터명
- Snake case: 유저로부터 받는 입력, Shiny app의 server에서 사용(될 것이라고 예상)하는 Object명, snake case로 작성된 dependencies의 함수명 등

2 Import sample data

전처리가 완료된 샘플데이터를 불러옵니다. - NA가 없어야 함 - string value가 있는 열은 factor로 변환 - 한 열이 모두 같은 값으로 채워져 있을 경우 제외해야 함 - Date type column이 없어야 함

```
cleaned_data <- read.csv(file = "~/git/goophi/data/boston_c.csv", stringsAsFactors = TRUE)
str(cleaned_data)
```

```
'data.frame': 506 obs. of 14 variables:
 $ zn      : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus   : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas    : Factor w/ 2 levels "otherwise","Tract bounds river": 1 1 1 1 1 1 1 1 1 1 ...
 $ nox     : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm      : num  6.58 6.42 7.18 7 7.15 ...
 $ age     : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis     : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad     : int   1 2 2 3 3 3 5 5 5 5 ...
 $ tax     : int  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio : num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black   : num  397 397 393 395 397 ...
 $ lstat   : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv    : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
 $ Pcrime  : int   0 0 0 0 0 0 0 0 0 0 ...
```

3 Data Setup Tab

User Input	description
targetVar	목적 변수(target, label)
trainSetRatio	전체 데이터 중 train set의 비율 (range: 0 - 1)

1) User input을 다음과 같이 받습니다

```
target_var <- "Pcrime"
train_set_ratio <- "0.7"
seed <- "1234"
formula <- paste0(target_var, " ~ .") # user x (1 user targetVar )
```

2) Train-test split 작업이 완료된 Object를 저장하고, Train set을 보여줍니다.

```
split_tmp <- goophi::trainTestSplit(data = cleaned_data,
                                     target = target_var,
                                     prop = train_set_ratio,
                                     seed = seed)
```

)

```
data_train <- split_tmp[[1]] # train data
data_test <- split_tmp[[2]] # test data
data_split <- split_tmp[[3]] # whole data with split information
```

3) train set에 적용할 전처리 정보를 담은 recipe를 생성합니다

```
rec <- goopht::prepForCV(data = data_train,
                        formula = formula,
                        seed = seed
                      )
```

4 Modeling with CV Tab

grid search, cross validation을 통해 사용자가 선택한 모델을 fitting합니다.

User Input	description
algo	ML 모델 선택
engine	engine 선택
mode	mode 선택
metric	Best performance에 대한 평가지표 선택
v	Cross validation시 train set을 몇 번 분할할 것인지 입력