# Package 'MSML'

January 20, 2025

**Title** Model Selection Based on Machine Learning (ML)

**Version** 1.0.0.1

**Description** Model evaluation based on a modified version of the recursive feature elimination algorithm. This package is designed to determine the optimal model(s) by leveraging all available features.

**License** GPL (>= 3)

**URL** `https://github.com/mommy003/MSML`

**Encoding** UTF-8

**RoxygenNote** 7.3.1

**Depends** R (>= 2.10)

**Imports** r2redux, R2ROC

**LazyData** true

**NeedsCompilation** no

**Author** Hong Lee [aut, cph],
Moksedul Momin [aut, cre, cph]

**Maintainer** Moksedul Momin <`cvasu.momin@gmail.com`>

**Repository** CRAN

**Date/Publication** 2024-03-04 05:20:02 UTC

## Contents

cov_train . . . . . . . . . . 2
cov_valid . . . . . . . . . . 2
data_test . . . . . . . . . . 3
data_train . . . . . . . . . . 3
data_valid . . . . . . . . . . 4
model_configuration . . . . . . . . . . 5
model_configuration2 . . . . . . . . . . 6
model_evaluation . . . . . . . . . . 7
predict_validation . . . . . . . . . . 8

**Index** **12**

## cov_train *3 sets of covariates for training data set*

### Description

A dataset containing N sets of covariates (N=3 as an example here) intended for constant use across all model configurations (refer to the 'model_configuration2' function) when using a training dataset. Please note that if constant covariates are not required, this file is unnecessary (refer to the 'model_configuration' function).

### Usage

```
cov_train
```

### Format

A data frame for training dataset:

**V1** covariate 1

**V2** covariate 2

**V3** covariate 3

## cov_valid *3 sets of covariates for validation data set*

### Description

A dataset containing N sets of covariates (N=3 as an example here) intended for constant use across all model configurations (refer to the 'model_configuration2' function) when using a validation dataset. Please note that if constant covariates are not required, this file is unnecessary (refer to the 'model_configuration' function).

### Usage

```
cov_valid
```

### Format

A data frame for validation dataset:

**V1** covariate 1

**V2** covariate 2

**V3** covariate 3

## `data_test` *7 sets of PRSs for test dataset and target phenotype*

### Description

A dataset containing 7 sets of PRSs for test dataset and target phenotype

### Usage

```
data_test
```

### Format

A data frame for test dataset:

**V1** Feature 1 (or PRS1 constructed using the first subset of SNPs from GWAS summary statistics)

**V2** Feature 2 (or PRS2 constructed using the second subset of SNPs from GWAS summary statistics)

**V3** Feature 3 (or PRS3 constructed using the third subset of SNPs from GWAS summary statistics)

**V4** Feature 4 (or PRS4 constructed using the fourth subset of SNPs from GWAS summary statistics)

**V5** Feature 5 (or PRS5 constructed using the fifth subset of SNPs from GWAS summary statistics)

**V6** Feature 6 (or PRS6 constructed using the sixth subset of SNPs from GWAS summary statistics)

**V7** Feature 7 (or PRS7 constructed using the seventh subset of SNPs from GWAS summary statistics)

**phenotype** Phenotypic values

## `data_train` *7 sets of PRSs for training data set and target phenotype*

### Description

A dataset containing 7 sets of PRSs for training data set and target phenotype

### Usage

```
data_train
```

### Format

A data frame for training dataset:

**V1** Feature 1 (or PRS1 constructed using the first subset of SNPs from GWAS summary statistics)

**V2** Feature 2 (or PRS2 constructed using the second subset of SNPs from GWAS summary statistics)

**V3** Feature 3 (or PRS3 constructed using the third subset of SNPs from GWAS summary statistics)

**V4** Feature 4 (or PRS4 constructed using the fourth subset of SNPs from GWAS summary statistics)

**V5** Feature 5 (or PRS5 constructed using the fifth subset of SNPs from GWAS summary statistics)

**V6** Feature 6 (or PRS6 constructed using the sixth subset of SNPs from GWAS summary statistics)

**V7** Feature 7 (or PRS7 constructed using the seventh subset of SNPs from GWAS summary statistics)

**phenotype** Phenotypic values

## `data_valid` *7 sets of PRSs for validation dataset and target phenotype*

### Description

A dataset containing 7 sets of PRSs for validation dataset and target phenotype

### Usage

```
data_valid
```

### Format

A data frame for validation dataset:

**V1** Feature 1 (or PRS1 constructed using the first subset of SNPs from GWAS summary statistics)

**V2** Feature 2 (or PRS2 constructed using the second subset of SNPs from GWAS summary statistics)

**V3** Feature 3 (or PRS3 constructed using the third subset of SNPs from GWAS summary statistics)

**V4** Feature 4 (or PRS4 constructed using the fourth subset of SNPs from GWAS summary statistics)

**V5** Feature 5 (or PRS5 constructed using the fifth subset of SNPs from GWAS summary statistics)

**V6** Feature 6 (or PRS6 constructed using the sixth subset of SNPs from GWAS summary statistics)

**V7** Feature 7 (or PRS7 constructed using the seventh subset of SNPs from GWAS summary statistics)

**phenotype** Phenotypic values

## model_configuration *model_configuration function*

### Description

This function generates predicted values for the validation dataset by applying optimal weights to features, which were estimated in the training dataset for each model configuration. The total number of model configurations is determined by summing the combinations for each possible number of features, ranging from 1 to 'n' (C(n, k)), where 'n choose k' (C(n, k)) represents the binomial coefficient. Here, 'n' denotes the total number of features, and 'k' indicates the number of features included in each model. For example, with n=7, the total number of model configurations is 127.

### Usage

```
model_configuration(data_train, data_valid, mv, model = "lm")
```

### Arguments

| | |
|---|---|
| `data_train` | This includes the dataframe of the training dataset in a matrix format |
| `data_valid` | This includes the dataframe of the validation dataset in a matrix format |
| `mv` | The total number of columns in data_train/data_valid |
| `model` | This is the type of model (e.g. lm (default) or glm) |

### Value

This function will generate all possible model outcomes for validation and test dataset

### Examples

```
data_train <- data_train
data_valid  <- data_valid
mv=8
out=model_configuration(data_train,data_valid,mv,model = "lm")
#This process will produce predicted values for the validation datasets,
#corresponding to each model configuration trained on the training dataset.
#The outcome of this function will yield variables named 'predict_validation'
#and 'total_model_configurations.
#To print the outcomes run out$predict_validation and out$total_model_configurations.
#For details (see https://github.com/mommy003/MSML).
```

## `model_configuration2` *model_configuration2 function*

### Description

This function is similar to the model_configuration function, with the added capability to maintain constant variables across models during training and prediction (see cov_train and cov_valid in page 2). Additionally, users have the option to select between linear or logistic regression models.

### Usage

```
model_configuration2(
  data_train,
  data_valid,
  mv,
  cov_train,
  cov_valid,
  model = ”lm”
)
```

### Arguments

`data_train` This includes the dataframe of the training dataset in a matrix format

`data_valid` This includes the dataframe of the validation dataset in a matrix format

`mv` The total number of columns in data_train/data_valid

`cov_train` This includes dataframe of covariates for training dataset in a matrix format

`cov_valid` This includes dataframe of covariates for validation dataset in a matrix format

`model` This is the type of model (e.g. lm (default) or glm (logistic regression))

### Value

This function will generate all possible model outcomes for validation and test dataset

### Examples

```
data_train <- data_train
data_valid  <- data_valid
mv=8
cov_train <- cov_train
cov_valid  <- cov_valid
out=model_configuration2(data_train,data_valid,mv,cov_train, cov_valid, model = ”lm”)
#This process will produce predicted values for the validation datasets,
#corresponding to each model configuration trained on the training dataset.
#The outcome of this function will yield variables named 'predict_validation'
#and 'total_model_configurations.
#To print the outcomes run out$predict_validation and out$total_model_configurations.
#For details (see https://github.com/mommy003/MSML).
```

```
#If a user intends to employ logistic regression without constant covariates,
#we advise preparing a covariate file where all values are set to 1.
```

## model_evaluation *model_evaluation function*

### Description

This function will identify the best model in the validation and test dataset.

### Usage

```
model_evaluation(dat, mv, tn, prev, pthreshold = 0.05, method = "R2ROC")
```

### Arguments

| | |
|---|---|
| dat | This is the dataframe for all the combinations of the model in a matrix format |
| mv | The total number of columns in data_train/data_valid |
| tn | The total number of best models to be identified |
| prev | The prevalence of disease in the data |
| pthreshold | The significance p value threshold when comparing models (default 0.05) |
| method | The methods to be used to evaluate models (e.g. R2ROC (default) or r2redux) |

### Value

This function will generate all possible model outcomes for validation and test dataset

### Examples

```
dat <- predict_validation
mv=8
tn=15
prev=0.047
out=model_evaluation(dat,mv,tn,prev)
#This process will generate three output files.
#out$out_all, contains AUC, p values for AUC, R2, and p values for R2,
#respectively for all models.
#out$out_start, contains AUC, p values for AUC, R2, and p values for R2,
#respectively for top tn models.
#out$out_selected, contains AUC, p values for AUC, R2, and p values for R2,
#respectively for best models.  This also includes selected features for models
#For details (see https://github.com/mommy003/MSML).
```

`predict_validation` *target phenotype and 127 sets of model configurations based on validation dataset*

## Description

A dataset containing target phenotype and 127 sets of model configurations based on validation dataset

## Usage

```
predict_validation
```

## Format

A data frame for predicted values for target dataset from model configurations_test:

**V1** Phenotypic values in target dataset

**V2** predicted values for target dataset from model configuration1

**V3** predicted values for target dataset from model configuration2

**V4** predicted values for target dataset from model configuration3

**V5** predicted values for target dataset from model configuration4

**V6** predicted values for target dataset from model configuration5

**V7** predicted values for target dataset from model configuration6

**V8** predicted values for target dataset from model configuration7

**V9** predicted values for target dataset from model configuration8

**V10** predicted values for target dataset from model configuration9

**V11** predicted values for target dataset from model configuration10

**V12** predicted values for target dataset from model configuration11

**V13** predicted values for target dataset from model configuration12

**V14** predicted values for target dataset from model configuration13

**V15** predicted values for target dataset from model configuration14

**V16** predicted values for target dataset from model configuration15

**V17** predicted values for target dataset from model configuration16

**V18** predicted values for target dataset from model configuration17

**V19** predicted values for target dataset from model configuration18

**V20** predicted values for target dataset from model configuration19

**V21** predicted values for target dataset from model configuration10

**V22** predicted values for target dataset from model configuration21

**V23** predicted values for target dataset from model configuration22

**V24** predicted values for target dataset from model configuration23

**V25** predicted values for target dataset from model configuration24

**V26** predicted values for target dataset from model configuration25

**V27** predicted values for target dataset from model configuration26

**V28** predicted values for target dataset from model configuration27

**V29** predicted values for target dataset from model configuration28

**V30** predicted values for target dataset from model configuration29

**V31** predicted values for target dataset from model configuration30

**V32** predicted values for target dataset from model configuration31

**V33** predicted values for target dataset from model configuration32

**V34** predicted values for target dataset from model configuration33

**V35** predicted values for target dataset from model configuration34

**V36** predicted values for target dataset from model configuration35

**V37** predicted values for target dataset from model configuration36

**V38** predicted values for target dataset from model configuration37

**V39** predicted values for target dataset from model configuration38

**V40** predicted values for target dataset from model configuration39

**V41** predicted values for target dataset from model configuration40

**V42** predicted values for target dataset from model configuration41

**V43** predicted values for target dataset from model configuration42

**V44** predicted values for target dataset from model configuration43

**V45** predicted values for target dataset from model configuration44

**V46** predicted values for target dataset from model configuration45

**V47** predicted values for target dataset from model configuration46

**V48** predicted values for target dataset from model configuration47

**V49** predicted values for target dataset from model configuration48

**V50** predicted values for target dataset from model configuration49

**V51** predicted values for target dataset from model configuration50

**V52** predicted values for target dataset from model configuration51

**V53** predicted values for target dataset from model configuration52

**V54** predicted values for target dataset from model configuration53

**V55** predicted values for target dataset from model configuration54

**V56** predicted values for target dataset from model configuration55

**V57** predicted values for target dataset from model configuration56

**V58** predicted values for target dataset from model configuration57

**V59** predicted values for target dataset from model configuration58

**V60** predicted values for target dataset from model configuration59

**V61** predicted values for target dataset from model configuration60

**V62** predicted values for target dataset from model configuration61

**V63** predicted values for target dataset from model configuration62

**V64** predicted values for target dataset from model configuration63

**V65** predicted values for target dataset from model configuration64

**V66** predicted values for target dataset from model configuration65

**V67** predicted values for target dataset from model configuration66

**V68** predicted values for target dataset from model configuration67

**V69** predicted values for target dataset from model configuration68

**V70** predicted values for target dataset from model configuration69

**V71** predicted values for target dataset from model configuration70

**V72** predicted values for target dataset from model configuration71

**V73** predicted values for target dataset from model configuration72

**V74** predicted values for target dataset from model configuration73

**V75** predicted values for target dataset from model configuration74

**V76** predicted values for target dataset from model configuration75

**V77** predicted values for target dataset from model configuration76

**V78** predicted values for target dataset from model configuration77

**V79** predicted values for target dataset from model configuration78

**V80** predicted values for target dataset from model configuration79

**V81** predicted values for target dataset from model configuration80

**V82** predicted values for target dataset from model configuration81

**V83** predicted values for target dataset from model configuration82

**V84** predicted values for target dataset from model configuration83

**V85** predicted values for target dataset from model configuration84

**V86** predicted values for target dataset from model configuration85

**V87** predicted values for target dataset from model configuration86

**V88** predicted values for target dataset from model configuration87

**V89** predicted values for target dataset from model configuration88

**V90** predicted values for target dataset from model configuration89

**V91** predicted values for target dataset from model configuration90

**V92** predicted values for target dataset from model configuration91

**V93** predicted values for target dataset from model configuration92

**V94** predicted values for target dataset from model configuration93

**V95** predicted values for target dataset from model configuration94

**V96** predicted values for target dataset from model configuration95

**V97** predicted values for target dataset from model configuration96

**V98** predicted values for target dataset from model configuration97

**V99** predicted values for target dataset from model configuration98

**V100** predicted values for target dataset from model configuration99

**V101** predicted values for target dataset from model configuration100

**V102** predicted values for target dataset from model configuration101

**V103** predicted values for target dataset from model configuration102

**V104** predicted values for target dataset from model configuration103

**V105** predicted values for target dataset from model configuration104

**V106** predicted values for target dataset from model configuration105

**V107** predicted values for target dataset from model configuration106

**V108** predicted values for target dataset from model configuration107

**V109** predicted values for target dataset from model configuration108

**V110** predicted values for target dataset from model configuration109

**V111** predicted values for target dataset from model configuration110

**V112** predicted values for target dataset from model configuration111

**V113** predicted values for target dataset from model configuration112

**V114** predicted values for target dataset from model configuration113

**V115** predicted values for target dataset from model configuration114

**V116** predicted values for target dataset from model configuration115

**V117** predicted values for target dataset from model configuration116

**V118** predicted values for target dataset from model configuration117

**V119** predicted values for target dataset from model configuration118

**V120** predicted values for target dataset from model configuration119

**V121** predicted values for target dataset from model configuration120

**V122** predicted values for target dataset from model configuration121

**V123** predicted values for target dataset from model configuration122

**V124** predicted values for target dataset from model configuration123

**V125** predicted values for target dataset from model configuration124

**V126** predicted values for target dataset from model configuration125

**V127** predicted values for target dataset from model configuration126

**V128** predicted values for target dataset from model configuration127

# Index

∗ **All**
  model_configuration, 5
  model_configuration2, 6
∗ **Identify**
  model_evaluation, 7
∗ **best**
  model_evaluation, 7
∗ **combinations**
  model_configuration, 5
  model_configuration2, 6
∗ **datasets**
  cov_train, 2
  cov_valid, 2
  data_test, 3
  data_train, 3
  data_valid, 4
  predict_validation, 8
∗ **models**
  model_evaluation, 7
∗ **model**
  model_configuration, 5
  model_configuration2, 6
∗ **possible**
  model_configuration, 5
  model_configuration2, 6

cov_train, 2
cov_valid, 2

data_test, 3
data_train, 3
data_valid, 4

model_configuration, 5
model_configuration2, 6
model_evaluation, 7

predict_validation, 8