

Package ‘NBvarsel’

August 30, 2026

Title Variable Selection via Cross-Validated Net Benefit

Version 0.1.1

Description Performs exhaustive or groupwise (backward elimination) variable selection for binary outcome prediction models using cross-validated Net Benefit as the optimization criterion. It supports predictor costs, restricted cubic splines, interaction terms, permutation importance, and parallel computation. It includes visualizations for model comparison and variable importance. References include Vickers AJ & Elkin EB (2006) [<doi:10.1177/0272989X06295361>](https://doi.org/10.1177/0272989X06295361), Van Calster B et al. (2018) [<doi:10.1016/j.eururo.2018.08.038>](https://doi.org/10.1016/j.eururo.2018.08.038), Vickers AJ et al. (2019) [<doi:10.1186/s41512-019-0064-7>](https://doi.org/10.1186/s41512-019-0064-7), and Baker SG et al. (2012) [<doi:10.1515/1557-4679.1395>](https://doi.org/10.1515/1557-4679.1395).

License GPL (≥ 3)

URL https://lasaibarrenada.github.io/NB_varsel/,
https://github.com/LasaiBarrenada/NB_varsel

BugReports https://github.com/LasaiBarrenada/NB_varsel/issues

Depends R ($\geq 4.1.0$)

Encoding UTF-8

LazyData true

RoxygenNote 8.0.0

Imports caret, DescTools, doParallel, dplyr, foreach, ggplot2, ggtext,
parallel, patchwork, pROC, rlang, rms, scales, stats, stringr,
tidyr

Suggests glmnet, ggsci, gt, knitr, quarto, testthat ($\geq 3.0.0$)

VignetteBuilder quarto

Config/testthat/edition 3

NeedsCompilation no

Author Lasai Barreñada [aut, cre] (ORCID:
 [<https://orcid.org/0000-0001-8020-0210>](https://orcid.org/0000-0001-8020-0210), Department of Development
and Regeneration ; Leuven Unit for Health Technology Assessment

Research (LUHTAR), KU Leuven, Belgium),
Laure Wynants [aut] (ORCID: <<https://orcid.org/0000-0002-3037-122X>>,
Department of Development and Regeneration, KU Leuven, Belgium;
Leuven Unit for Health Technology Assessment Research (LUHTAR), KU
Leuven, Belgium; Department of Epidemiology, CAPHRI Care and Public
Health Research Institute, Maastricht University, Maastricht,
Netherlands),
Ewout W. Steyerberg [aut] (ORCID:
<<https://orcid.org/0000-0002-7787-0122>>, Julius Centre for Health
Sciences and Primary Care, University Medical Centre Utrecht,
Utrecht, Netherlands),
Andrew Vickers [aut] (ORCID: <<https://orcid.org/0000-0003-1525-6503>>,
Department of Epidemiology and Biostatistics, Memorial Sloan
Kettering Cancer Center, New York, NY, USA),
Ben Van Calster [aut] (ORCID: <<https://orcid.org/0000-0003-1613-7450>>,
Department of Development and Regeneration, KU Leuven, Belgium;
Leuven Unit for Health Technology Assessment Research (LUHTAR), KU
Leuven, Belgium; Julius Centre for Health Sciences and Primary
Care, University Medical Centre Utrecht, Utrecht, Netherlands)

Maintainer Lasai Barreñada <lasai.barrenadataleb@kuleuven.be>

Repository CRAN

Date/Publication 2026-08-30 09:10:33 UTC

Contents

adnex_results	2
all_subset_plot	4
nb_varsel	5
VIF_plot	7
Index	8

adnex_results	<i>ADNEX Exhaustive Variable Selection Results</i>
---------------	--

Description

Pre-computed results from an exhaustive Net Benefit-based variable selection applied to ovarian tumour data from the International Ovarian Tumour Analysis (IOTA) consortium (phases 1–3). The dataset contains model-level summary statistics for the top 20 models per predictor count, drawn from a full exhaustive search over all 65,535 (2^16 - 1) predictor combinations.

Usage

adnex_results

Format

A data frame with 293 rows and 23 columns:

Model Comma-separated list of predictor names included in the model.

n_Preds Number of predictors in the model.

AUC Cross-validated area under the ROC curve.

Brier Cross-validated Brier score.

Total_Cost Sum of predictor group costs for the model.

Avg_Adj_Net_Benefit Average cost-adjusted Net Benefit across thresholds and CV folds.

Avg_Net_Benefit Average Net Benefit (without cost adjustment) across thresholds and CV folds.

VIF_age, VIF_ca125, VIF_family_history, VIF_locules_gt_10, VIF_oncology_center, VIF_max_diam_lesion, VIF_pa
Permutation importance (delta Net Benefit) for each predictor. NA when the predictor is not included in the model.

Details

The original patient-level data are not publicly available. This dataset contains only aggregated model performance metrics and can be used to demonstrate the [all_subset_plot\(\)](#) and [VIF_plot\(\)](#) visualisation functions.

The analysis used 16 candidate predictors for a binary outcome (malignant vs. benign ovarian tumour) with grouped predictor costs reflecting clinical history (free), ultrasound examination (moderate cost), and blood biomarker (higher cost). Five-fold cross-validation was used with restricted cubic splines (3 knots) for continuous predictors.

The 16 candidate predictors are:

- **Clinical history** (no cost): patient age, family history of ovarian cancer, oncology centre, pain.
- **Ultrasound** (moderate cost): maximum lesion diameter, proportion solid, number of locules > 10, papillary count, papillary presence, acoustic shadows, ascites, irregular walls, bilateral, colour score, maximum solid diameter.
- **Blood biomarker** (higher cost): CA-125.

This is a filtered subset (top 20 models per number of predictors, ranked by Avg_Adj_Net_Benefit) of the full 65,535-model exhaustive search. The attribute "best_model_stats" contains the overall best model, and "n_total_models" records the total number of evaluated models.

Source

IOTA consortium data (phases 1–3). See <https://iotaplus.org/en> for information about the IOTA studies.

Examples

```
data(adnex_results)
head(adnex_results)

# Best model from the full search
attr(adnex_results, "best_model_stats")

# Visualise
all_subset_plot(adnex_results, filter = 7, size_dot = 1)
VIF_plot(adnex_results)$plot
```

all_subset_plot	<i>All-Subset Model Comparison Plot</i>
-----------------	---

Description

Creates a two-panel figure comparing model performance across all evaluated subsets. The upper panel shows the selected metric for each model, and the lower panel shows which predictors are included in each model.

Usage

```
all_subset_plot(
  all_models,
  metric = "Avg_Net_Benefit",
  y_axis = "Average Net Benefit",
  filter = 5,
  size_dot = 3,
  highlight_color = "red",
  tile_color = "#2A6EBB",
  p1_theme = NULL,
  p2_theme = NULL,
  data_dict = NULL
)
```

Arguments

all_models	Data frame returned in the all_models element of <code>nb_varsel()</code> . This should contain the model summary output used for plotting.
metric	Character string. Name of the column in all_models to plot on the y-axis. Defaults to "Avg_Net_Benefit".
y_axis	Character string. Label for the y-axis in the top panel. Defaults to "Average Net Benefit".
filter	Integer. Maximum number of models to display per number of predictors. Defaults to 5.
size_dot	Numeric. Point size used in the metric plot. Defaults to 3.

highlight_color	Character string. Color for the dashed reference lines marking the best model. Defaults to "red".
tile_color	Character string. Fill color used for included predictors in the heatmap. Defaults to "#2A6EBB".
p1_theme	A <code>ggplot2::theme()</code> object applied to the top panel, or NULL.
p2_theme	A <code>ggplot2::theme()</code> object applied to the bottom panel, or NULL.
data_dict	A named character vector used to relabel predictors. Format: <code>c("raw_name" = "Display Label")</code> . Defaults to NULL.

Value

A [patchwork](#) object containing two stacked ggplot panels.

nb_varsel	<i>Variable Selection via Cross-Validated Net Benefit</i>
-----------	---

Description

Performs exhaustive or groupwise variable selection for binary-outcome prediction models. Models are evaluated using cross-validated net benefit, with optional adjustment for predictor costs.

Usage

```
nb_varsel(
  data,
  outcome_var,
  costs = NULL,
  thresholds = seq(0.05, 0.3, by = 0.01),
  include_interactions = FALSE,
  mode = c("exhaustive", "groupwise"),
  group_size = 2,
  cv_folds = 5,
  seed = 123,
  verbose = TRUE,
  allow_parallel = TRUE,
  permutation = FALSE,
  splines = TRUE,
  n_knots = 3
)
```

Arguments

data	A data frame containing the predictor variables and the binary outcome.
outcome_var	Character string naming the binary outcome column (coded as 0/1 or as a two-level factor).

<code>costs</code>	Predictor costs. Can be: • NULL (default): no costs. • A named numeric vector: per-predictor costs, e.g. <code>c(X1 = 0.1, X2 = 0.05)</code>. • An unnamed scalar: the same cost applied to every predictor. • A list of groups, each with elements <code>vars</code> (character vector) and <code>cost</code> (numeric). The group cost is added once if any variable in the group is selected.
<code>thresholds</code>	Numeric vector of decision thresholds at which net benefit is evaluated. Defaults to <code>seq(0.05, 0.3, by = 0.01)</code> .
<code>include_interactions</code>	Logical. If TRUE, interaction terms $((X1 + X2 + \dots)^2)$ are also evaluated in exhaustive mode. Defaults to FALSE.
<code>mode</code>	Character string specifying the selection strategy: "exhaustive" evaluates all predictor subsets, while "groupwise" performs backward elimination in groups of size <code>group_size</code> . Defaults to "exhaustive".
<code>group_size</code>	Integer. Number of variables removed at each step in groupwise mode. Defaults to 2.
<code>cv_folds</code>	Integer. Number of cross-validation folds. Defaults to 5.
<code>seed</code>	Integer. Random seed for reproducibility. Defaults to 123.
<code>verbose</code>	Logical. If TRUE, progress messages are printed. Defaults to TRUE.
<code>allow_parallel</code>	Logical. If TRUE, use parallel computation through <code>foreach</code> and <code>doParallel</code> . Defaults to TRUE.
<code>permutation</code>	Logical. If TRUE, compute permutation importance scores for each predictor. Defaults to FALSE.
<code>splines</code>	Logical. If TRUE, apply restricted cubic splines (via <code>rms::rcs()</code>) to continuous predictors. Defaults to TRUE.
<code>n_knots</code>	Integer. Number of knots used for restricted cubic splines. Defaults to 3.

Value

A list with the following elements:

best_model_stats A one-row data frame describing the top-performing model.

all_models A data frame with one row per evaluated model, containing columns `Model`, `n_Preds`, `AUC`, `Brier`, `Total_Cost`, `Avg_Adj_Net_Benefit`, and `Avg_Net_Benefit`. If `permutation = TRUE`, additional `VIF_` columns containing permutation importance scores are included.

Examples

```
set.seed(42)
n <- 500
df <- data.frame(
  X1 = rnorm(n), X2 = rbinom(n, 1, 0.7),
  X3 = rnorm(n), X4 = rbinom(n, 1, 0.5)
)
df$Y <- rbinom(n, 1, plogis(2 * df$X1 + 1.5 * df$X2))
```

```

harms <- c(X1 = 0.1, X2 = 0.05, X3 = 0.1, X4 = 0.0001)

result <- nb_varsel(
  data = df, outcome_var = "Y", costs = harms,
  mode = "exhaustive", splines = FALSE,
  allow_parallel = FALSE
)
result$best_model_stats

```

VIF_plot

*Permutation Importance Bar Plot***Description**

Creates a horizontal bar chart showing the average permutation importance of each predictor, measured as the change in net benefit relative to the full set of evaluated models.

Usage

```
VIF_plot(all_models, filter = NULL, color = "#2A6EBB", data_dict = NULL)
```

Arguments

<code>all_models</code>	Data frame returned in the <code>all_models</code> element of <code>nb_varsel()</code> . Must include VIF_ columns generated when <code>permutation = TRUE</code> .
<code>filter</code>	Integer or NULL. If specified, only the top filter models by Avg_Net_Benefit are used to compute average importance. Defaults to NULL, which uses all models.
<code>color</code>	Character string. Fill color for the bars. Defaults to "#2A6EBB".
<code>data_dict</code>	A named character vector used to relabel predictors. Format: <code>c("raw_name" = "Display Label")</code> . Defaults to NULL.

Value

A list with two elements:

plot A `ggplot2::ggplot` object showing the bar chart.

data A data frame containing the Variable, Average_Delta_NB, and mapped Label values.

Examples

```

data(adnex_results)
vif_results <- VIF_plot(adnex_results)
vif_results$plot
head(vif_results$data)

```

Index

* **datasets**

adnex_results, [2](#)

adnex_results, [2](#)

all_subset_plot, [4](#)

all_subset_plot(), [3](#)

ggplot2::ggplot, [7](#)

ggplot2::theme(), [5](#)

nb_varsel, [5](#)

nb_varsel(), [4](#), [7](#)

patchwork, [5](#)

VIF_plot, [7](#)

VIF_plot(), [3](#)