ShapleyVIC: Shapley Variable Importance Cloud for Interpretable Machine Learning

Yilin Ning, Chenglin Niu, Mingxuan Liu, Siqi Li, Nan Liu

2024-01-07

Table of contents

ShapleyVIC Introduction

Variable importance assessment is important for interpreting machine learning models. Current practice in interpretable machine learning applications focuses on explaining the final models that optimize predictive performance. However, this does not fully address practical needs, where researchers are willing to consider models that are "good enough" but are easier to understand or implement. Shapley variable importance cloud (ShapleyVIC) fills this gap by extending current method to a set of "good models" for comprehensive and robust assessments. Building on a common theoretical basis (i.e., Shapley values for variable importance), ShapleyVIC seamlessly complements the widely adopted SHAP assessments of a single final model to avoid biased inference. Please visit GitHub page for source code.

Usage

As detailed in Chapter 3 ShapleyVIC analysis of variable importance consists of 3 general steps:

1. Training an optimal prediction model (e.g., a logistic regression model).
2. Generating a reasonable number of (e.g., 350) nearly optimal models of the same model class (e.g., logistic regression).
3. Evaluate Shapley-based variable importance from each nearly optimal model and pool information for inference.

Chapter 3 demonstrates ShapleyVIC application for binary outcomes, and Chapter 6 and Chapter 7 provide additional examples for applications for ordinal and continuous outcomes, respectively.

ShapleyVIC does not require variable centering or standardization, but requires some data checking and pre-processing for stable and smooth processing, which we summarize in Chapter 2.

The ShapleyVIC-based variable ranking can also be used with the AutoScore framework to develop clinical risk scores for interpretable risk prediction, which we demonstrate in Chapter 4 and Chapter 5.

Installation

The ShapleyVIC framework is now implemented using a **Python library** that trains the optimal model, generates nearly optimal models and evaluate Shapley-based variable importance from such models, and an **R package** that pools information across models to generate summary statistics and visualizations for inference.

Python library

- **Required:** Python version 3.6 or higher.
  - ***Recommended:*** latest stable release of Python 3.9 or 3.10.
- **Required:** latest version of git.

Execute the following command in Terminal/Command Prompt to install the Python library from GitHub:

- Linux/macOS:

```
pip install git+"https://github.com/nliulab/ShapleyVIC#egg=ShapleyVIC&subdirectory=python"
```

- Windows:

```
python.exe -m pip install git+"https://github.com/nliulab/ShapleyVIC#egg=ShapleyVIC&subdir
```

> **Note**
>
> - *ShapleyVIC uses a modified version of the SAGE library (version 0.0.4b1), which avoids occasional stack overflow problems on Windows but does not affect variable importance evaluation.*

R package

- **Required:** R version 3.5.0 or higher.
  - ***Recommended:*** use latest version of R with RStudio.

Execute the following command in R/RStudio to install the R package from GitHub:

```
if (!require("devtools", quietly = TRUE)) install.packages("devtools")
devtools::install_github("nliulab/ShapleyVIC/r")
```

Citation

Core paper

- Ning Y, Ong ME, Chakraborty B, Goldstein BA, Ting DS, Vaughan R, Liu N. Shapley variable importance cloud for interpretable machine learning. *Patterns* 2022; 3: 100452.

Method extension

- Ning Y, Li S, Ong ME, Xie F, Chakraborty B, Ting DS, Liu N. A novel interpretable machine learning system to generate clinical risk scores: An application for predicting early mortality or unplanned readmission in a retrospective cohort study. *PLOS Digit Health* 2022; 1(6): e0000062.

Clinical applications

- Deng X, Ning Y, Saffari SE, Xiao B, Niu C, Ng SYE, Chia N, Choi X, Heng DL, Tan YJ, Ng E, Xu Z, Tay KY, Au WL, Ng A, Tan EK, Liu N, and Tan LCS (2023). Identifying clinical features and blood biomarkers associated with mild cognitive impairment in Parkinson's Disease using machine learning. *European Journal of Neurology*, 00:1–9.

Contact

- Yilin Ning (Email: yilin.ning@duke-nus.edu.sg)
- Nan Liu (Email: liu.nan@duke-nus.edu.sg)

1 Data requirements

1.1 General requirements

- Currently ShapleyVIC applies to binary, ordinal and continuous outcomes.
- Code binary outcomes as 0/1, and ordinal outcomes as integers starting from 0.
- No space or special characters (e.g., `[`, `]`, `(`, `)`, `,`) in variable names. Replace them using `_`.
- Variable centering/standardization is *not required*.

1.2 Missing values and sparsity

- Handle missing entries appropriately before applying ShapleyVIC. Missing entry is not supported
- Check data distribution and handle data sparsity before applying ShapleyVIC. **Data sparsity may increase run time and lead to unstable results.**

1.3 Additional pre-processing for high-dimensional data

- Although theoretically permissible, it is not advisable to apply ShapleyVIC to data with a large number of variables.
- Screen out variables with low importance (e.g., based on univariable or multivariable analysis p-values) to reduce dimension (e.g., to $<$50 variables) before applying ShapleyVIC.

1.4 General suggestions on the size of explanation set

- Larger number of variables generally requires larger explanation set for stable results.
- Increase in the size of explanation set and/or number of variables increases time required to compute ShapleyVIC values.
- Use of $>$3500 samples in explanation set leads to long run time and is generally not recommended.

2 ShapleyVIC for Variable Importance Assessment

ShapleyVIC is model agnostic, and its benefits of has been demonstrated in empirical experiments in applications for multiple domains. This tutorial illustrates ShapleyVIC implementation using the Python library and R package, in a study with a binary outcome that predicts 2-year recidivism using a logistic regression of 6 binary variables. Chapter 5 and Chapter 6 provides reproducible examples for ShapleyVIC analysis of ordinal and continuous outcomes, respectively.

Cite the following papers for ShapleyVIC:

- Ning Y, Ong ME, Chakraborty B, Goldstein BA, Ting DS, Vaughan R, Liu N. Shapley variable importance cloud for interpretable machine learning. *Patterns* 2022; 3: 100452.

2.1 [Python] ShapleyVIC calculation

This part of the ShapleyVIC workflow is implemented in Python.

In this part of the workflow, we load and prepare data, train optimal logistic regression model, generate nearly optimal models, and compute Shapley-based variable importance for each model.

2.1.1 Load data

- Read data from CSV or Excel files.
- For this demo, use the integrated data in the library that contains 7214 samples analyzed in Experiment 1 (i.e., the recidivism prediction study) of the paper.

```
from ShapleyVIC import df_compas

compas = df_compas.load_data()
# See data description using the following command:
# help(df_compas.load_data)
compas.loc[:5]
```

y	age	race	prior	gender	juvenilecrime	currentcharge	train_test
0	0	0	1	1	1	0	train
1	0	1	1	1	1	0	train
1	0	1	0	1	0	0	train
0	0	1	0	1	0	0	train
0	0	0	0	1	1	0	train
0	0	0	1	1	1	1	train

- y: 2-year recidivism (the binary outcome, 1=event and 0=non-event).
- age, race, prior, gender, juvenilecrime, currentcharge: binary predictors.
- train_test: training/explanation set membership indicator (`"train"` for training and `"test"` for explanation). *Not to include in models.*

2.1.2 Prepare training and explanation sets

- When there is sufficient data, users can split the full dataset into a training set to train optimal and nearly optimal models, and an explanation set to compute ShapleyVIC values.
- Otherwise, users may use the full dataset to train models and compute ShapleyVIC values.

> **!** Important
>
> - *As detailed in Chapter 1, check for and handle data issues before applying ShapleyVIC.*
> - *This demo will show impact of data sparsity on ShapleyVIC results.*

In the experiment, we used 10% of the full dataset as explanation set. See Chapter 1 for general suggestions on the size of explanation set.

2.1.2.1 Use the `train_test` indicator available

```
dat_train = compas.loc[compas['train_test']=='train']
# Drop the indicator column after using it to split data:
dat_train = dat_train.drop(columns=['train_test'])
dat_train.reset_index(drop=True, inplace=True)

dat_expl = compas.loc[compas['train_test']=='test']
dat_expl = dat_expl.drop(columns=['train_test'])
dat_expl.reset_index(drop=True, inplace=True)
```

2.1.2.2 Random split for general cases

```
# Drop the column 'train_test' that indicates set membership in example data:
compas = compas.drop(columns=['train_test'])
# Generate row indices for training and explanation sets:
from sklearn.model_selection import train_test_split
i_train, i_expl = train_test_split(list(range(compas.shape[0])),
    test_size=int(0.1 * compas.shape[0]), random_state=0)

dat_train = compas.iloc[i_train, :]
dat_train.reset_index(drop=True, inplace=True)

dat_expl = compas.iloc[i_expl, :]
dat_expl.reset_index(drop=True, inplace=True)
```

2.1.3 Train optimal model

- Specify training data to initialize the model object and train the optimal model.
- `x`, `y`: predictors (as a data frame) and outcome from the training set.
- `outcome_type`: type of the outcome. Default is "binary" that is most common in clinical applications.
  - See Chapter 5 for "ordinal" and Chapter 6 for "continuous".
- `x_names_cat`: names of categorical predictors. Optional for binary predictors encoded as 0/1.
- `output_dir`: the directory to save key outputs to. Will be used as input in the subsequent R workflow.
- `save_data`: whether to save `x` and `y` to `output_dir` (default is to save). If not, `x` and `y` must be supplied separately in subsequent R analysis.

In this step, users also need to configure the `criterion` for defining nearly optimal models (see following sections for detail).

- Default is `criterion="loss"` that applies to all outcome types, where nearly optimal models exceed minimum loss by no more than `epsilon` (default is 0.05, i.e., 5%).
- `criterion="auc"` or `criterion="prauc"` is also supported for binary outcomes (but not other outcome types), where nearly optimal models have AUC or PRAUC within the 95% of that for the optimal model.

2.1.3.1 Loss criterion

Define nearly optimal based on loss, and `epsilon = 0.05` (the default) for the range of permissible loss.

Default option is to save input data `x` and `y`, so that they are not needed as input in the subsequent R workflow.

```
# Specify the name of outcome, which is 'y' in this example:
y_name = 'y'
from ShapleyVIC import model
model_object = model.models(
    x=dat_train.drop(columns=[y_name]), y=dat_train[y_name],
    outcome_type="binary",
    x_names_cat=['age','race','prior','gender','juvenilecrime','currentcharge'],
    output_dir="compas_output"
)
# To display the optimal logistic regression trained:
model_object.model_optim.summary().tables[1]
```

2.1.3.2 AUC criterion

Define nearly optimal based on AUC, in this case `epsilon` is not used. This specification does not affect the optimal model.

Users are advised to allocate different output folders when different `criterion` is selected to avoid confusion. In this example, results are saved to output folder `compas_auc_output`.

```
# Specify the name of outcome, which is 'y' in this example:
y_name = 'y'
from ShapleyVIC import model
model_object_auc = model.models(
    x=dat_train.drop(columns=[y_name]), y=dat_train[y_name],
    outcome_type="binary", criterion="auc",
    x_names_cat=['age','race','prior','gender','juvenilecrime','currentcharge'],
    output_dir="compas_auc_output"
)
# To display the optimal logistic regression trained:
model_object_auc.model_optim.summary().tables[1]
```

2.1.3.3 PRAUC criterion

Define nearly optimal based on PRAUC, in this case `epsilon` is not used. This specification does not affect the optimal model.

Users are advised to allocate different output folders when different `criterion` is selected to avoid confusion. In this example, results are saved to output folder `compas_prauc_output`.

```
# Specify the name of outcome, which is 'y' in this example:
y_name = 'y'
from ShapleyVIC import model
model_object_prauc = model.models(
    x=dat_train.drop(columns=[y_name]), y=dat_train[y_name],
    outcome_type="binary", criterion="prauc",
    x_names_cat=['age','race','prior','gender','juvenilecrime','currentcharge'],
    output_dir="compas_prauc_output"
)
# To display the optimal logistic regression trained:
model_object_prauc.model_optim.summary().tables[1]
```

2.1.3.4 Loss criterion, do not save data

Specify `save_data=False` to avoid saving `x` and `y` to output folder. See Chapter 3 for another example.

```
# Specify the name of outcome, which is 'y' in this example:
y_name = 'y'
from ShapleyVIC import model
model_object = model.models(
    x=dat_train.drop(columns=[y_name]), y=dat_train[y_name],
    outcome_type="binary",
    x_names_cat=['age','race','prior','gender','juvenilecrime','currentcharge'],
    output_dir="compas_output", save_data=False
)
# To display the optimal logistic regression trained:
model_object.model_optim.summary().tables[1]
```

	coef	std err	z	P	[0.025	0.975]
const	0.4455	0.107	4.160	0.000	0.236	0.655
age	1.5001	0.187	8.011	0.000	1.133	1.867
race	0.4164	0.053	7.858	0.000	0.313	0.520

	coef	std err	z	P	[0.025	0.975]
prior	-0.8543	0.061	-13.984	0.000	-0.974	-0.735
gender	0.3835	0.068	5.651	0.000	0.251	0.517
juvenilecrime	-0.8646	0.084	-10.238	0.000	-1.030	-0.699
currentcharge	-0.2544	0.056	-4.562	0.000	-0.364	-0.145

2.1.4 Generate nearly optimal models

As mentioned above, by default (i.e., `criterion="loss", epsilon = 0.05`) nearly optimal logistic regression models are defined as models with logistic loss less than $(1 + \varepsilon)$ times the minimum loss (i.e., logistic loss of the optimal model). Default value for ε is 5%.

When `criterion="auc"` or `criterion="prauc"` is specified, nearly optimal logistic regression models are defined as models with AUC (or PRAUC) within the 95% CI of that of the optimal logistic regression model.

The `criterion` (and `epsilon`) specified when initializing the model object is used to configure and sample nearly optimal models in the following steps.

- `u1` and `u2` are key hyper-parameters for generating nearly optimal models, which control the sampling range of initial models to fully explore the model space.
- Use the following command to generate a set of reasonable values for `u1` and `u2` (using `m=200` initial models), such that approximately 70%-80% of initial models are eligible.
  - The same command applies for all `criterion` but output varies.

2.1.4.1 Output for loss criterion (default)

```
u1, u2 = model_object.init_hyper_params(m=200)
(u1, u2)
```

```
Nearly optimal defined based on loss with epsilon=0.05.
```

```
(0.5, 80.3125)
```

2.1.4.2 Output for AUC criterion

```
u1, u2 = model_object_auc.init_hyper_params(m=200)
(u1, u2)
```

```
Nearly optimal defined based on auc.
```

```
(0.5, 38.125)
```

2.1.4.3 Output for PRAUC criterion

```
u1, u2 = model_object_prauc.init_hyper_params(m=200)
(u1, u2)
```

```
Nearly optimal defined based on prauc.
```

```
(0.5, 32.5)
```

- Use the following command to generate a final set of nearly optimal models (e.g., `n_final=250`) from 500 initial samples (`m=500`).
  - The same command applies for all `criterion` but outputs vary.

2.1.4.4 Loss criterion (default)

```
model_object.draw_models(u1=u1, u2=u2, m=500, n_final=250, random_state=1234)
model_object.models_plot
```

```
model_object.models_near_optim.iloc[:5]
```

const	age_1	race_1	prior_1	gender_1	juvenilecrime_1	currentcharge_1	perf_metric
-0.2307	3.1195	0.5047	-1.1409	0.2644	-0.1170	0.2664	1.0280
0.5503	0.7759	0.8971	-1.1164	-0.3083	-0.6398	-0.1481	1.0285
0.1068	0.8697	-0.0176	-0.6963	0.6987	-0.5041	-0.1812	1.0187
0.9715	0.8669	-0.1101	-1.0772	0.6450	-1.3590	-0.3310	1.0212

const	age_1	race_1	prior_1	gender_1	juvenilecrime_1	currentcharge_1	perf_metric
-1.0476	2.0026	0.6911	-0.3203	1.4661	-0.6633	-0.0397	1.0438
0.4006	1.6629	0.1719	-0.5450	0.3218	-0.9498	0.6260	1.0445

2.1.4.5 AUC criterion

```
model_object_auc.draw_models(u1=u1, u2=u2, m=500, n_final=250, random_state=1234)
model_object_auc.models_plot
```

```
model_object_auc.models_near_optim.iloc[:5]
```

const	age_1	race_1	prior_1	gender_1	juvenilecrime_1	currentcharge_1	perf_metric
0.1989	2.5653	0.5665	-1.6820	0.7333	-0.4778	-0.5296	0.6742
0.5184	0.9964	0.7507	-1.0366	-0.0977	-0.7082	-0.1805	0.6671
0.8325	1.7529	0.5860	-1.0563	0.0572	-0.8658	-0.2591	0.6785
0.2104	1.0626	0.1152	-0.7446	0.6022	-0.6144	-0.2036	0.6758
0.8113	1.0597	0.0502	-1.0093	0.5654	-1.2084	-0.3077	0.6751
0.4500	1.7156	0.4195	-0.9999	0.4521	-0.7532	-0.5512	0.6765

2.1.4.6 PRAUC criterion

```
model_object_prauc.draw_models(u1=u1, u2=u2, m=500, n_final=250, random_state=1234)
model_object_prauc.models_plot
```

```
model_object_prauc.models_near_optim.iloc[:5]
```

const	age_1	race_1	prior_1	gender_1	juvenilecrime_1	currentcharge_1	perf_metric
0.2178	2.4837	0.5550	-1.6186	0.7065	-0.5075	-0.5085	0.6557
0.5161	1.4747	0.5260	-0.7212	0.3116	-0.9734	-0.3726	0.6636
0.7843	1.0923	0.0772	-0.9979	0.5519	-1.1830	-0.3038	0.6605
0.4497	1.6994	0.4193	-0.9889	0.4469	-0.7616	-0.5288	0.6631
-0.3412	0.7137	0.3837	-0.9295	1.2320	-0.6347	-0.2812	0.6622
-0.5807	0.5378	0.9135	-0.9719	0.6827	-0.4119	-0.4054	0.6583

2.1.5 Assess variable importance

This step assesses variable importance for each nearly optimal model generated in the previous step using the SAGE method, and write the results to the output folder for further processing in the subsequent R workflow. Parallel processing is used to reduce run time.

- `model_object`: the model object created above.
- `x_expl`, `y_expl`: predictors (as a data frame) and outcome from the explanation set.
- `n_cores`: number of CPU cores to use in parallel processing.
  - For a computer with `n` cores, do not use more than `n-1` cores.
- `threshold`: threshold parameter used in SAGE algorithm for convergence criterion. A reasonable value is 0.05 (default).
  - Smaller `threshold` value may improve accuracy of uncertainty measure but notably increases run time.
- The same command applies for all `criterion`.

2.1.5.1 Loss criterion

```
from ShapleyVIC import compute
m_svic = compute.compute_shapley_vic(
    model_obj=model_object,
    x_expl=dat_expl.drop(columns=[y_name]), y_expl=dat_expl[y_name],
    n_cores=7, # running on a MacBook Air with 8 cores
    threshold=0.05
)
```

2.1.5.2 AUC criterion

```
from ShapleyVIC import compute
m_svic = compute.compute_shapley_vic(
    model_obj=model_object_auc,
    x_expl=dat_expl.drop(columns=[y_name]), y_expl=dat_expl[y_name],
    n_cores=7, # running on a MacBook Air with 8 cores
    threshold=0.05
)
```

2.1.5.3 PRAUC criterion

```
from ShapleyVIC import compute
m_svic = compute.compute_shapley_vic(
    model_obj=model_object_prauc,
    x_expl=dat_expl.drop(columns=[y_name]), y_expl=dat_expl[y_name],
    n_cores=7, # running on a MacBook Air with 8 cores
    threshold=0.05
)
```

Note

- *Use built-in software (e.g., Activity Monitor/Task Manager) to monitor CPU and Memory usage. Avoid taking up 100% CPU, which can slow down computation.*
- *This step can be time consuming with larger number of variables and/or larger explanation data.*
- *For users' reference, the commands above took approximately 10-20 minutes on a 2022 MacBook Air (Apple M2 chip with 8-core CPU, 8-core GPU; 16GB unified memory; 256GB SSD storage).*

2.2 [R] ShapleyVIC summary and visualizations

This part of the ShapleyVIC workflow is implemented in R.

This part of the workflow works on output from Python (all saved in `output_dir`), pooling information across models to compute (and visualize) overall variable importance and derive ensemble variable rankings.

2.2.1 Compute overall importance

As detailed in the paper, raw Shapley-based variable importance needs to be adjusted based on variable colinearity to derive final ShapleyVIC values.

- `output_dir`: output folder generated from the Python workflow.
- `outcome_type`: type of outcome, as specified in the Python workflow.
- `criterion`: criterion to define nearly optimal models, as used in the Python workflow.
  - Loss criterion is assumed by default.
- `x` and `y`: training data specified in the Python workflow, required if `save_data=False` was specified when setting up `model.models(...)` in Python.

- `x_names_cat`: names of categorical variables, as specified in the Python workflow. Used when assessing variable colinearity from the training set. Optional for binary variables coded as 0/1.
- `x_names_display`: variable names to use in summary statistics and visualizations. If not provided, column names in the training set will be used.

2.2.1.1 Loss criterion

```
library(ShapleyVIC)
model_object <- compile_shapley_vic(
  output_dir = "compas_output", outcome_type = "binary",
  x_names_cat = c('age','race','prior','gender','juvenilecrime','currentcharge'),
  x_names = c("Age", "Race", "Prior criminal history", "Gender",
              "Juvenile criminal history", "Current charge")
)
```

```
Compiling results for binary outcome using loss criterion to define neaerly optimal models.
```

2.2.1.2 AUC criterion

```
library(ShapleyVIC)
model_object_auc <- compile_shapley_vic(
  output_dir = "compas_auc_output", outcome_type = "binary", criterion = "auc",
  x_names_cat = c('age','race','prior','gender','juvenilecrime','currentcharge'),
  x_names = c("Age", "Race", "Prior criminal history", "Gender",
              "Juvenile criminal history", "Current charge")
)
```

```
Compiling results for binary outcome using auc criterion to define neaerly optimal models.
```

2.2.1.3 PRAUC criterion

```
library(ShapleyVIC)
model_object_prauc <- compile_shapley_vic(
  output_dir = "compas_prauc_output", outcome_type = "binary", criterion = "prauc",
  x_names_cat = c('age','race','prior','gender','juvenilecrime','currentcharge'),
  x_names = c("Age", "Race", "Prior criminal history", "Gender",
              "Juvenile criminal history", "Current charge")
```

```
)
```

```
Compiling results for binary outcome using prauc criterion to define neaerly optimal models.
```

2.2.1.4 Loss criterion, data not saved in Python workflow

When training data was not saved in the Python workflow, they must be supplied as input (`x` and `y`) in this step. See Chapter 3 for another example.

```
# Prepare training data the same way as in Python workflow:
library(ShapleyVIC)
library(dplyr)
data("df_compas")
# Use the `train_test` indicator to filter out training data used in Python workflow:
df_train <- df_compas %>% filter(train_test == "train") %>%
  select(-train_test) %>% as.data.frame()
y_name <- "y"
x_names <- setdiff(names(df_train), y_name)
# Supply x and y when compiling results from Python workflow:
model_object <- compile_shapley_vic(
  output_dir = "compas_output", outcome_type = "binary",
  x = df_train[, x_names], y = df_train[, y_name],
  x_names_cat = c('age','race','prior','gender','juvenilecrime','currentcharge'),
  x_names = c("Age", "Race", "Prior criminal history", "Gender",
              "Juvenile criminal history", "Current charge")
)
```

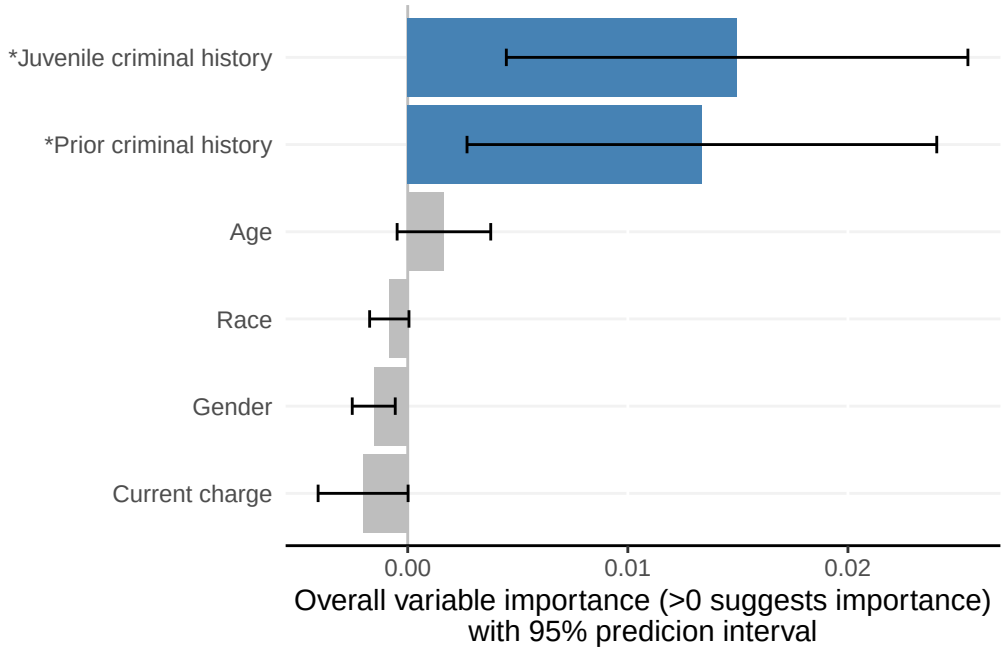
2.2.2 Visualize overall variable importance

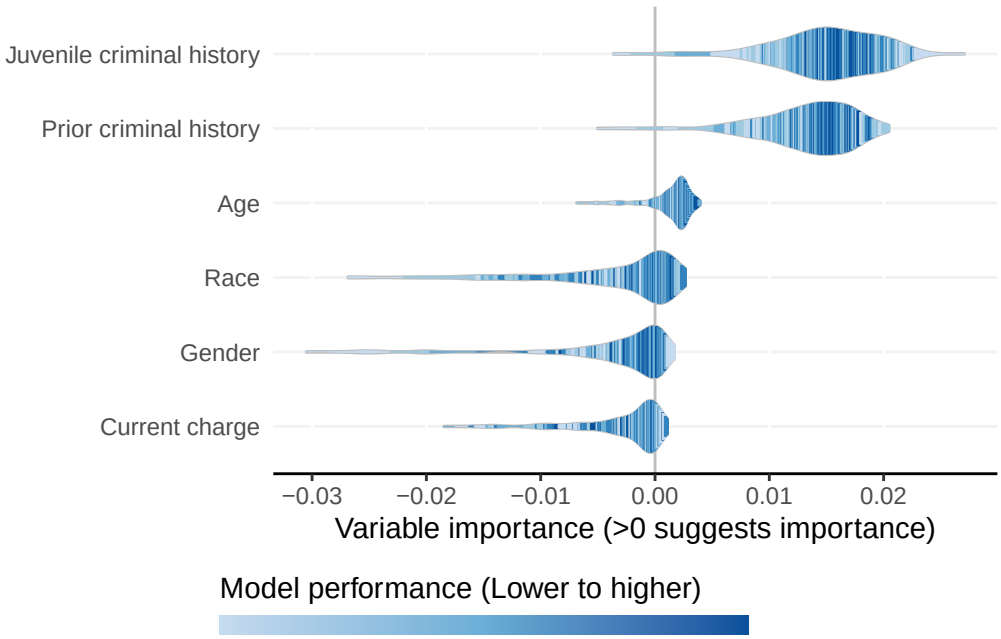
Each ShapleyVIC value (`shapley_vic_val`) is reported with a standard deviation (`sage_sd`). We pool information across models to compute overall variable importance and uncertainty interval, visualized using bar plot. The relationship between variable importance and model performance is visualized using violin plot.

- The same command applies for all `criterion` and generates similar results.
- For clarity, in the bar plot variables with significant overall importance are indicated by blue color and "*" next to variable names.

2.2.2.1 Results from loss criterion

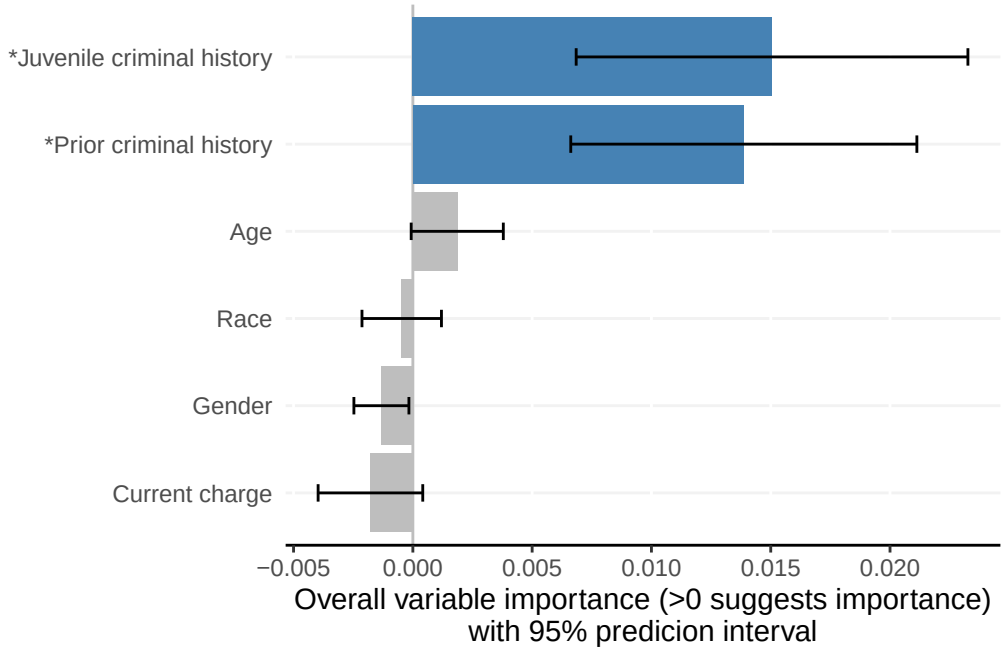
```
model_plots <- plot(model_object)
```

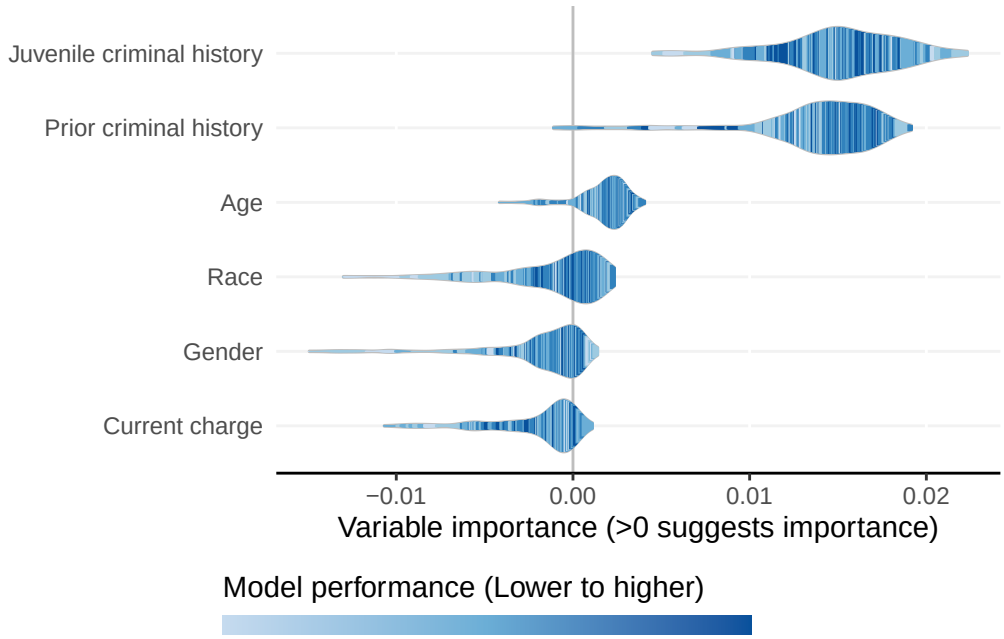




2.2.2.2 Results from AUC criterion

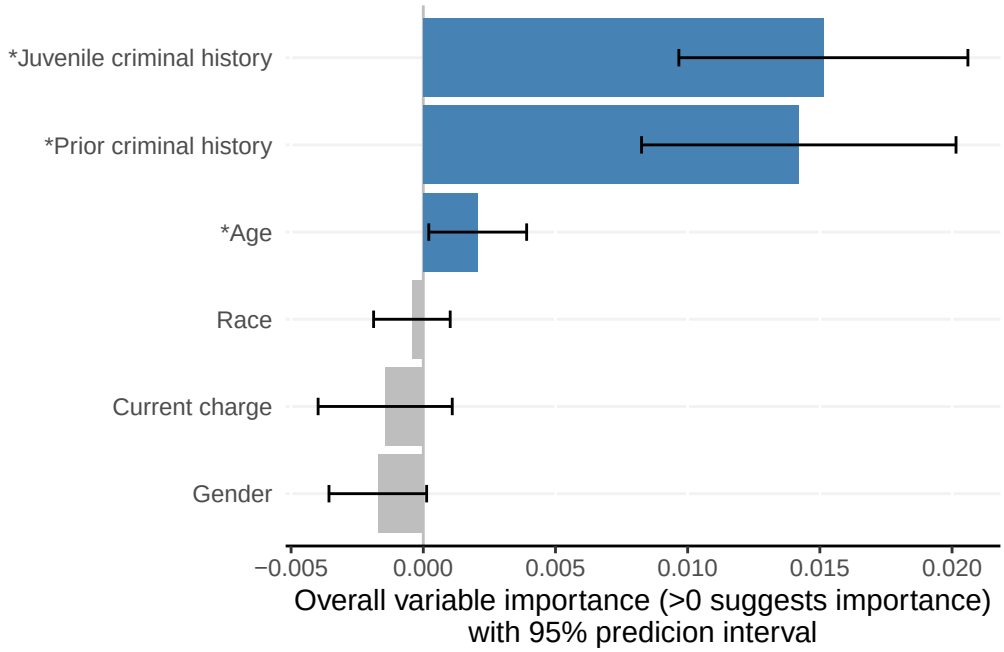
```
model_auc_plots <- plot(model_object_auc)
```

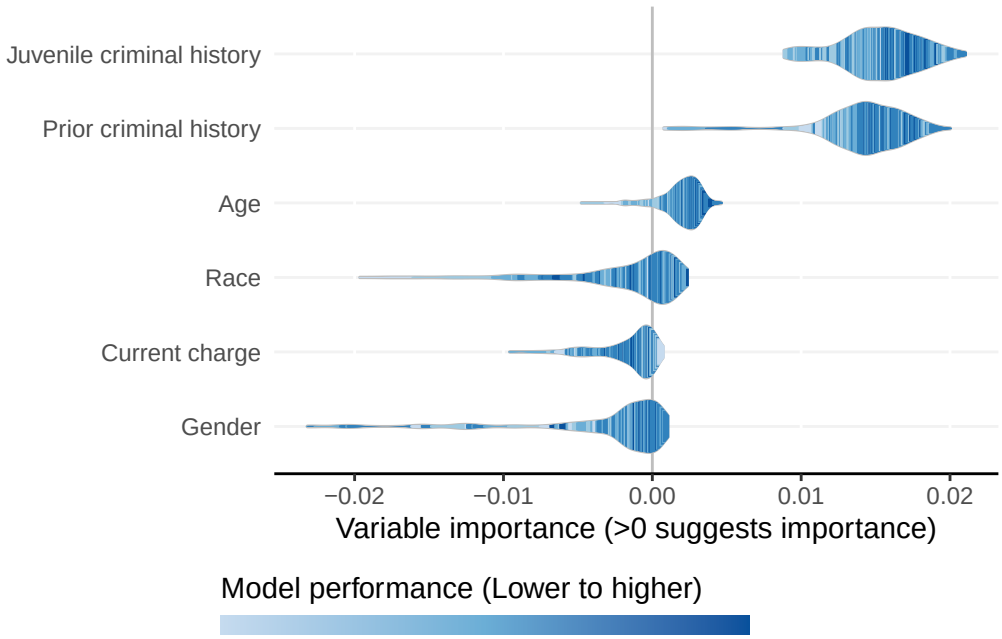

*Juvenile criminal history
*Prior criminal history
Age
Race
Gender
Current charge
−0.005
0.000
0.005
0.010
0.015
0.020
Overall variable importance (>0 suggests importance)
with 95% predicion interval


Juvenile criminal history
Prior criminal history
Age
Race
Gender
Current charge
−0.01
0.00
0.01
0.02
Variable importance (>0 suggests importance)
Model performance (Lower to higher)

2.2.2.3 Results from PRAUC criterion

```
model_prauc_plots <- plot(model_object_prauc)
```





> **Note**
>
> - *Plots above reproduce key findings reported in the paper: race had non-significant overall importance, and prior criminal history and juvenile criminal history had higher overall importance than other variables.*
> - *Overall importance of age now becomes non-significant when loss or AUC crterion were used, showing that data sparsity (only 20 [2.8%] of 721 subjects had age=1 in explanation data) leads to less stable results.*

The bar plot can be further edited using ggplot functions, e.g., edit text font size using `theme()` or add plot title using `labs()`:

```
library(ggplot2)
model_plots$bar + theme(text = element_text(size = 14)) + labs(title = "Bar plot")
```

To apply similar formatting to the violin plot, use the following function:

```
library(ggplot2)
plot_violin(x = model_object, title = "Violin plot",
            plot_theme = theme(text = element_text(size = 14)))
```

2.2.3 Ensemble variable ranking

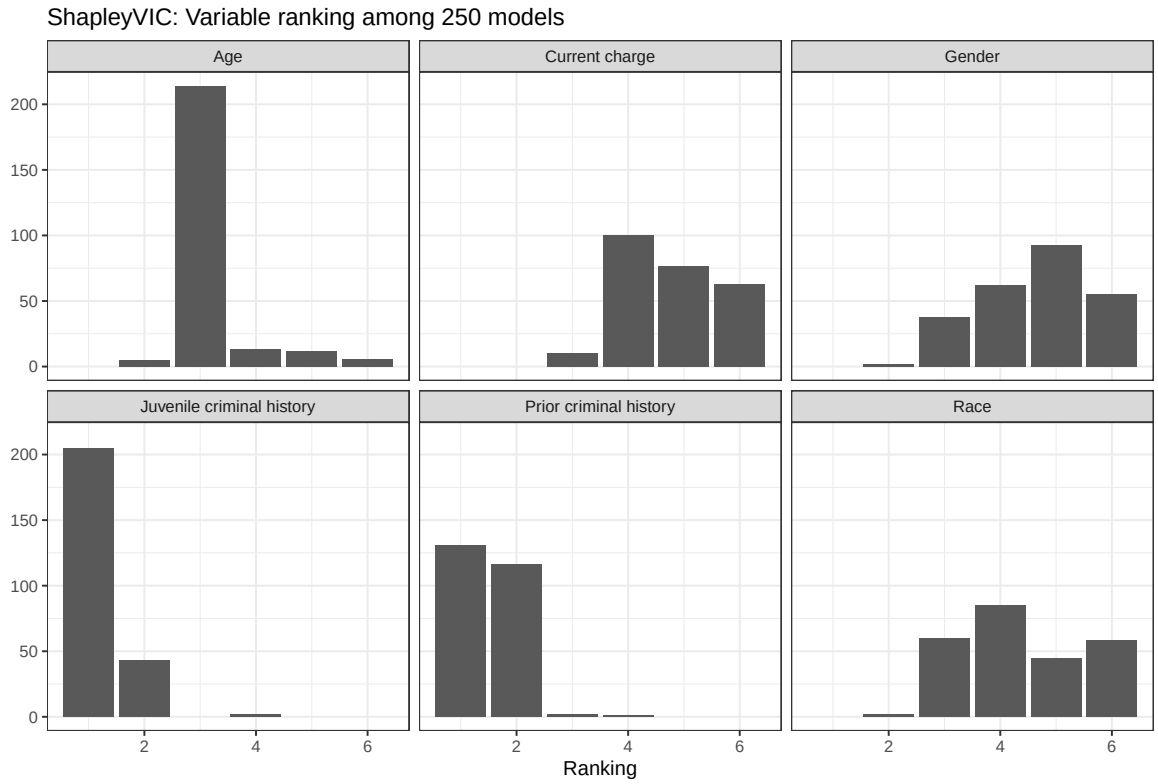
ShapleyVIC values can also be used to rank variables by their importance to each model. The bar plot of ranks may help identify models with increased reliance on specific variable of interest for further investigation.

The same command applies for all `criterion`. The example below illustrates for the model object using loss criterion.

```
val_ranks <- rank_variables(model_object)
head(val_ranks, 6)
```

```
  model_id                  Variable rank
1        0                       Age    5
2        0                      Race    3
3        0    Prior criminal history    1
4        0                    Gender    3
5        0 Juvenile criminal history    2
6        0            Current charge    5
```

```
library(ggplot2)
ggplot(val_ranks, aes(x = rank, group = Variable)) +
  geom_bar() +
  facet_wrap(~ Variable, nrow = 2) +
  theme_bw() +
  labs(x = "Ranking", y = "",
       title = "ShapleyVIC: Variable ranking among 250 models")
```



The ensemble ranking averages the ranks across models, and can be used to guide downstream model building, e.g., using AutoScore. See the next chapter for detailed demonstration.

```
rank_variables(model_object, summarise = TRUE)
```

```
                  Variable mean_rank
1 Juvenile criminal history     1.196
2    Prior criminal history     1.492
```

```
# To return variable ranking as named vector for convenient integration with
# AutoScore:
rank_variables(model_object, summarise = TRUE, as_vector = TRUE)
```

```
Juvenile criminal history    Prior criminal history
                    1.196                     1.492
```

3 AutoScore-ShapleyVIC for Interpretable Risk Score Development

Risk scores are widely used for clinical decision making and commonly generated from logistic regression models. Machine-learning-based methods may work well for identifying important predictors to create parsimonious scores, but such 'black box' variable selection limits interpretability, and variable importance evaluated from a single model can be biased. We propose a robust and interpretable variable selection approach using ShapleyVIC, and integrate it with the AutoScore framework for convenient development of risk scoring models.

In this chapter, we describe the application of the AutoScore-ShapleyVIC workflow using an empirical example in our paper, and provide code for generating a risk score (i.e., Model 2 in the paper) to predict the risk of 30-day readmission or death from 41 candidate variables.

In the next chapter, we provide a fully reproducible example to demonstrate the use of the AutoScore-ShapleyVIC workflow using a simulated data that is publicly available.

Cite the following papers for AutoScore-ShapleyVIC:

- Ning Y, Ong ME, Chakraborty B, Goldstein BA, Ting DS, Vaughan R, Liu N. Shapley variable importance cloud for interpretable machine learning. *Patterns* 2022
- Ning Y, Li S, Ong ME, Xie F, Chakraborty B, Ting DS, Liu N. A novel interpretable machine learning system to generate clinical risk scores: An application for predicting early mortality or unplanned readmission in a retrospective cohort study. *PLOS Digit Health* 1(6): e0000062.
- Xie F, Chakraborty B, Ong MEH, Goldstein BA, Liu N. AutoScore: A machine learning-based automatic clinical score generator and its application to mortality prediction using electronic health records. *JMIR Medical Informatics* 2020; 8(10): e21798.

3.1 [R] Prepare data

This part of the workflow is implemented in R.

3.1.1 Load R packages and data

```
if (!require(AutoScore, quietly = TRUE)) install.packages("AutoScore")
library(AutoScore)
library(tidyverse) # For convenient data manipulation and visualization

# Read the final clean data with 41 candidate variables and the binary outcome
# (`label`):
dat <- readRDS("dat_readmit_or_death.RDS")
```

3.1.2 Prepare training, validation and test datasets

- Use the `split_data()` function of the AutoScore package to split data into training (70%), validation (10%) and test (20%) sets for risk score development.
- Perform median imputation for vital signs and lab tests based on training set.

> **!** Important
>
> - *As detailed in Chapter 1, handle missingness (and any other potential data issue) before applying ShapleyVIC.*

```
set.seed(1234)
Out_split <- split_data(data = dat, ratio = c(7, 1, 2))
# Median imputation for vital signs and lab tests based on training set:
train_lab_test <- Out_split$train_set %>% select(Pulse:SODIUM)
train_lab_test_median <- apply(train_lab_test, 2, function(x) median(x, na.rm = TRUE))
Out_split <- lapply(Out_split, function(dat) {
  for (nm in names(train_lab_test)) {
    dat[, nm] <- ifelse(is.na(dat[, nm]), train_lab_test_median[nm], dat[, nm])
  }
  dat
})

train_set <- Out_split$train_set
validation_set <- Out_split$validation_set
test_set <- Out_split$test_set
```

- Prepare `output_dir` for ShapleyVIC, using `train_set` as training set and the first 3500 observations in `validation_set` as the explanation data.

```r
output_dir <- "score_output"
if (!dir.exists(output_dir)) dir.create(output_dir)
write.csv(train_set, file = file.path(output_dir, "train_set.csv"),
          row.names = FALSE)
write.csv(validation_set[1:3500, ],
          file = file.path(output_dir, "validation_set.csv"),
          row.names = FALSE)
```

3.2 [Python] Compute ShapleyVIC values

This part of the workflow is implemented in Python.

- Load data and set up input information.
- Data used in this analysis is sensitive, therefore we do not save training data to the output folder to avoid any potential data security issue.

```python
import os
import pandas as pd
output_dir = "score_output"
dat_train = pd.read_csv(os.path.join(output_dir, 'train_set.csv'))
dat_expl = pd.read_csv(os.path.join(output_dir, 'validation_set.csv'))

y_name = 'label'
x_names_cat = ['Gender','Race','Triage_Class_Code','DayofWeek','MI','CHF','PVD',
    'Stroke','Dementia','Pulmonary','Rheumatic','PUD','LiverMild','Diabetes',
    'DMcx','Paralysis','Renal','Cancer','LiverSevere','Mets','admit_cat',
    'resuscitation','VENTILATION']
from ShapleyVIC import model
model_object = model.models(
    x=dat_train.drop(columns=[y_name]), y=dat_train[y_name],
    x_names_cat=x_names_cat, outcome_type="binary", output_dir=output_dir,
    save_data=False
)
```

- Draw 350 nearly optimal models.

```python
model_object.draw_models(u1=0.2, u2=300, m=800, n_final=350)
```

- Compute ShapleyVIC values.

```
from ShapleyVIC import compute
m_svic = compute.compute_shapley_vic(
    model_obj=model_object,
    x_expl=dat_expl.drop(columns=[y_name]), y_expl=dat_expl[y_name],
    n_cores=10, # running on a PC with 40 logical processors
    threshold=0.025
)
```

3.3 [R] Develop risk score

This part of the workflow is implemented in R.

3.3.1 Rank variables using ShapleyVIC

- Compile ShapleyVIC output.
- Since data was not saved in the Python workflow, we explicitly specify it in the R analysis.
- Explicitly specify names of categorical variables, identical to those specified in the Python workflow.

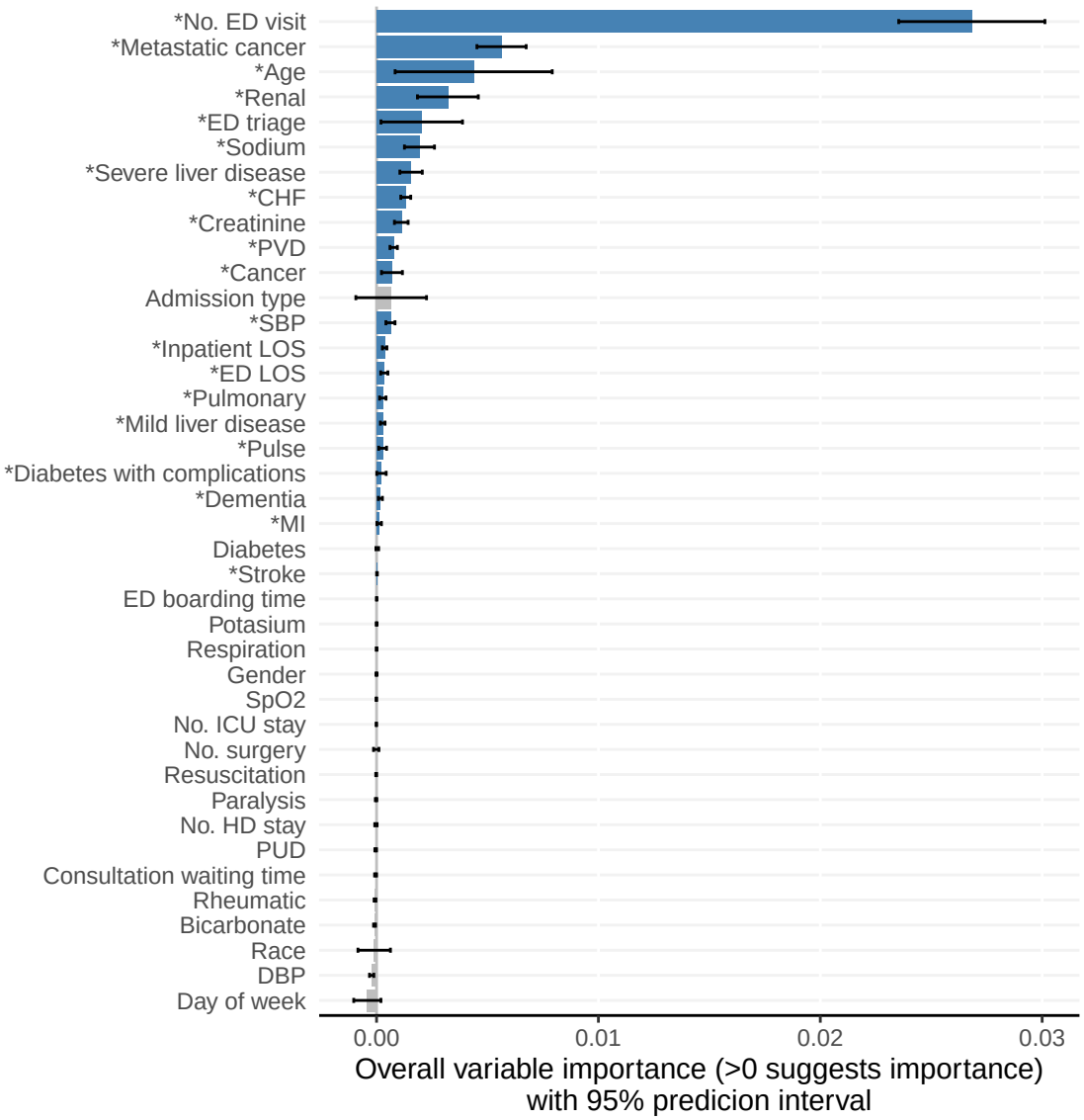
```
output_dir <- "score_output"
x_names_display <- c(
  "Age", "Gender", "Race", "ED LOS", "ED triage",
  "ED boarding time", "Consultation waiting time", "No. ED visit",
  "Day of week", "Inpatient LOS", "Ventilation", "Resuscitation",
  "No. surgery", "No. ICU stay",
  "No. HD stay", "Pulse", "Respiration", "SpO2",
  "DBP", "SBP", "Bicarbonate", "Creatinine",
  "Potasium", "Sodium", "MI", "CHF", "PVD", "Stroke",
  "Dementia", "Pulmonary", "Rheumatic", "PUD", "Mild liver disease",
  "Diabetes", "Diabetes with complications", "Paralysis", "Renal", "Cancer",
  "Severe liver disease", "Metastatic cancer", "Admission type"
)
y_name <- "label"
x_names <- setdiff(names(train_set), y_name)
library(ShapleyVIC)
model_object <- compile_shapley_vic(
  output_dir = output_dir, outcome_type = "binary",
  x = train_set[, x_names], y = train_set$label,
```

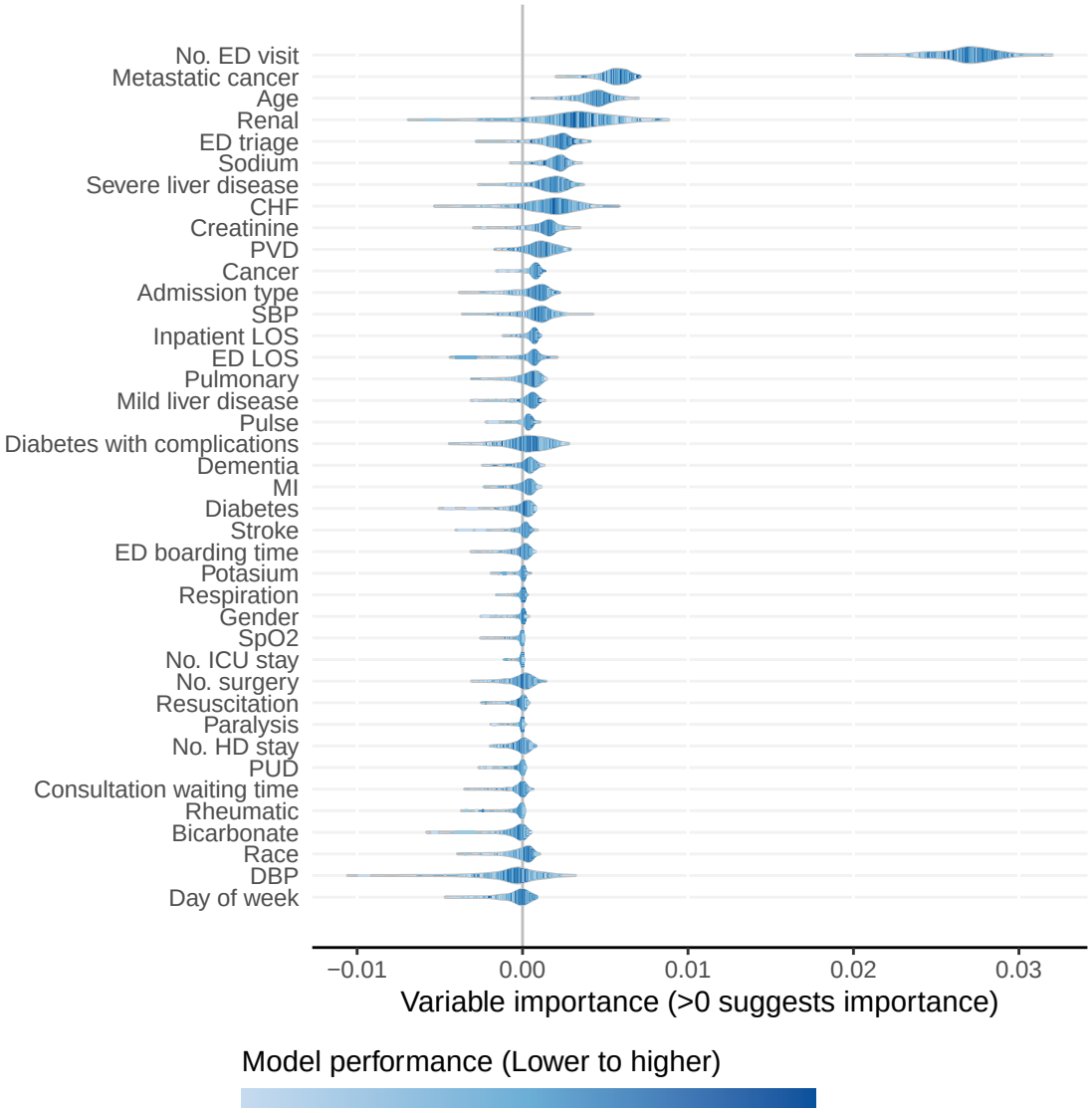
```
  x_names_cat = c(
    'Gender','Race','Triage_Class_Code','DayofWeek','MI','CHF','PVD',
    'Stroke','Dementia','Pulmonary','Rheumatic','PUD','LiverMild','Diabetes',
    'DMcx','Paralysis','Renal','Cancer','LiverSevere','Mets','admit_cat',
    'resuscitation','VENTILATION'
  ),
  x_names = x_names_display
)
```

- Visualize ShapleyVIC values for overall variable importance.

```
model_plots <- plot(model_object)
```

```
The following variables are excluded due to zero importance in all models analysed:

     Ventilation
```


*No. ED visit
*Metastatic cancer
*Age
*Renal
*ED triage
*Sodium
*Severe liver disease
*CHF
*Creatinine
*PVD
*Cancer
Admission type
*SBP
*Inpatient LOS
*ED LOS
*Pulmonary
*Mild liver disease
*Pulse
*Diabetes with complications
*Dementia
*MI
Diabetes
*Stroke
ED boarding time
Potasium
Respiration
Gender
SpO2
No. ICU stay
No. surgery
Resuscitation
Paralysis
No. HD stay
PUD
Consultation waiting time
Rheumatic
Bicarbonate
Race
DBP
Day of week
0.00
0.01
0.02
0.03
Overall variable importance (>0 suggests importance)
with 95% predicion interval



- Derive ShapleyVIC-based ensemble variable ranking.

```
ranking <- rank_variables(model_object, summarise = TRUE, as_vector = TRUE)
ranking
```

```
      No. ED visit  Metastatic cancer
          1.000000           2.182857
              Age             Sodium
          3.057143           6.931429
```

```
                Renal                   ED triage
             6.945714                    7.428571
Severe liver disease                         CHF
             9.071429                   10.365714
          Creatinine                         PVD
            10.642857                   12.074286
                 SBP                      Cancer
            14.120000                   14.597143
       Inpatient LOS                      ED LOS
            16.105714                   16.908571
  Mild liver disease                    Pulmonary
            17.785714                   18.040000
             Dementia Diabetes with complications
            19.697143                   20.291429
                Pulse                          MI
            20.691429                   21.214286
               Stroke
            24.185714
```

3.3.2 Develop risk score using AutoScore workflow

- Modify variable names in training, validation and test sets for publication-ready figures and printed output.

```
# Current raw variable names:
names(train_set)
```

```
 [1] "label"                    "Age"
 [3] "Gender"                   "Race"
 [5] "ED_LOS"                   "Triage_Class_Code"
 [7] "EDBoardingTime"           "ConsultationWaitingTime"
 [9] "n_ed_6mth"                "DayofWeek"
[11] "LOS_inp"                  "VENTILATION"
[13] "resuscitation"            "Total_Num_Surgery_last1yr"
[15] "Total_icu_count_last1yr"  "Total_hd_count_last1yr"
[17] "Pulse"                    "Respiration"
[19] "SPO2"                     "BP_Diastolic"
[21] "BP_Systolic"              "BICARBONATE"
[23] "CREATININE"               "POTASSIUM"
[25] "SODIUM"                   "MI"
[27] "CHF"                      "PVD"
```

```
[29] "Stroke"                    "Dementia"
[31] "Pulmonary"                 "Rheumatic"
[33] "PUD"                       "LiverMild"
[35] "Diabetes"                  "DMcx"
[37] "Paralysis"                 "Renal"
[39] "Cancer"                    "LiverSevere"
[41] "Mets"                      "admit_cat"
```

```r
# Modified variable names:
names(train_set)[-1] <- x_names_display
names(train_set)
```

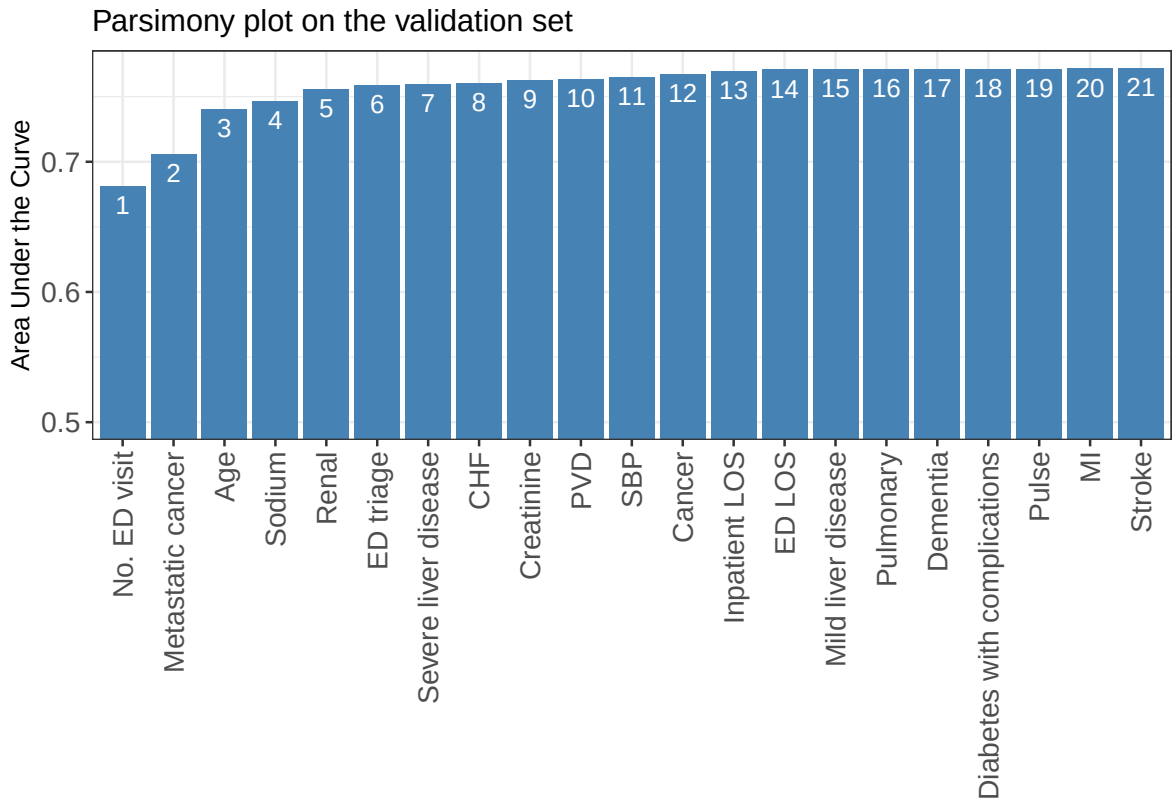
```
 [1] "label"                      "Age"
 [3] "Gender"                     "Race"
 [5] "ED LOS"                     "ED triage"
 [7] "ED boarding time"           "Consultation waiting time"
 [9] "No. ED visit"               "Day of week"
[11] "Inpatient LOS"              "Ventilation"
[13] "Resuscitation"              "No. surgery"
[15] "No. ICU stay"               "No. HD stay"
[17] "Pulse"                      "Respiration"
[19] "SpO2"                       "DBP"
[21] "SBP"                        "Bicarbonate"
[23] "Creatinine"                 "Potasium"
[25] "Sodium"                     "MI"
[27] "CHF"                        "PVD"
[29] "Stroke"                     "Dementia"
[31] "Pulmonary"                  "Rheumatic"
[33] "PUD"                        "Mild liver disease"
[35] "Diabetes"                   "Diabetes with complications"
[37] "Paralysis"                  "Renal"
[39] "Cancer"                     "Severe liver disease"
[41] "Metastatic cancer"          "Admission type"
```

```r
names(validation_set)[-1] <- x_names_display
names(test_set)[-1] <- x_names_display
```

- Based on the ensemble variable ranking, apply AutoScore STEP(ii) to select the best model with parsimony plot.

```r
AUC <- AutoScore_parsimony(
  train_set = train_set, validation_set = validation_set,
  rank = ranking, max_score = 100, n_min = 1, n_max = length(ranking)
)
```

```
Select 1 Variable(s):  Area under the curve: 0.6811
Select 2 Variable(s):  Area under the curve: 0.706
Select 3 Variable(s):  Area under the curve: 0.7406
Select 4 Variable(s):  Area under the curve: 0.7467
Select 5 Variable(s):  Area under the curve: 0.7555
Select 6 Variable(s):  Area under the curve: 0.7589
Select 7 Variable(s):  Area under the curve: 0.7595
Select 8 Variable(s):  Area under the curve: 0.7605
Select 9 Variable(s):  Area under the curve: 0.7624
Select 10 Variable(s):  Area under the curve: 0.7637
Select 11 Variable(s):  Area under the curve: 0.765
Select 12 Variable(s):  Area under the curve: 0.7674
Select 13 Variable(s):  Area under the curve: 0.7696
Select 14 Variable(s):  Area under the curve: 0.7708
Select 15 Variable(s):  Area under the curve: 0.7708
Select 16 Variable(s):  Area under the curve: 0.7713
Select 17 Variable(s):  Area under the curve: 0.7713
Select 18 Variable(s):  Area under the curve: 0.7712
Select 19 Variable(s):  Area under the curve: 0.7713
Select 20 Variable(s):  Area under the curve: 0.7715
Select 21 Variable(s):  Area under the curve: 0.7718
```

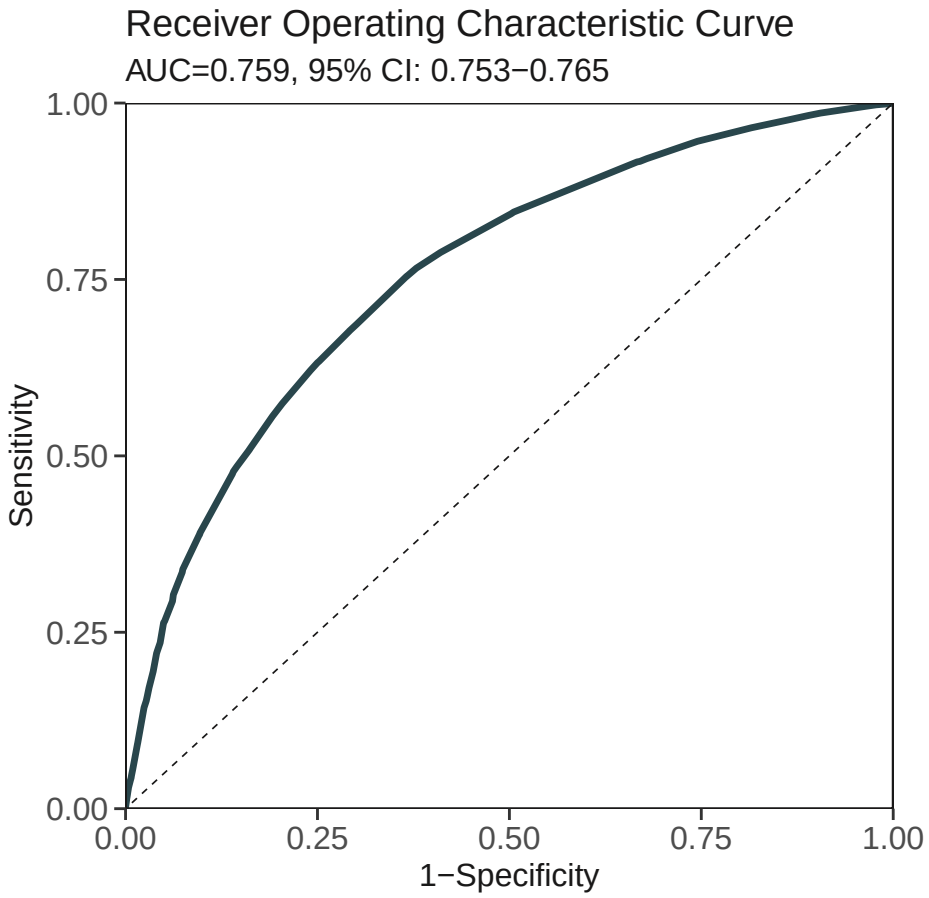


- This parsimony is somewhat smoother than that from random forest-based variable ranking used in AutoScore.
- A feasible choice is to select the top 6 variables, as adding additional variables does not substantially improve model performance.
- Apply AutoScore STEP(iii) to build initial scores from the top 6 variables.

```
cut_vec <- AutoScore_weighting(
  train_set = train_set, validation_set = validation_set,
  final_variables = names(ranking)[1:6], max_score = 100
)
```

```
****Included Variables:
      variable_name
1      No. ED visit
2 Metastatic cancer
3               Age
4            Sodium
```

```
5             Renal
6         ED triage
****Initial Scores:


=================  =========  =====
variable           interval   point
=================  =========  =====
No. ED visit       <1           0
                   [1,3)       14
                   >=3         32

Metastatic cancer  0            0
                   1           22

Age                <28          0
                   [28,46)      5
                   [46,78)     11
                   [78,87)     14
                   >=87        19

Sodium             <126        11
                   [126,132)    8
                   [132,138)    3
                   [138,141)    0
                   >=141        3

Renal              0            0
                   1            8

ED triage          P1           8
                   P2           5
                   P3 and P4    0
=================  =========  =====
```




```
***Performance (based on validation set):
AUC:  0.7589   95% CI: 0.7525-0.7654 (DeLong)
Best score threshold: >= 27
Other performance indicators based on this score threshold:
Sensitivity: 0.7546
Specificity: 0.6338
PPV:         0.2888
NPV:         0.9291
***The cutoffs of each variable generated by the AutoScore are saved in cut_vec. You can dec
```

- Users can apply additional AutoScore STEPs for subsequent model fine-tuning and evaluation.

4 AutoScore-ShapleyVIC Reproducible Example

This chapter provides a fully reproducible example to demonstrate in detail the use of the AutoScore-ShapleyVIC workflow, using a simulated dataset with binary outcome available from the AutoScore package. The data is described in detail in the AutoScore Guidebook.

4.1 [R] Prepare data

This part of the workflow is implemented in R.

4.1.1 Load data

- Load `sample_data` from the AutoScore package.
- As required by AutoScore, change the name of outcome variable to `label`.
- Read AutoScore Guidebook for detailed data requirement.

```
library(AutoScore)
data("sample_data")
names(sample_data)[names(sample_data) == "Mortality_inpatient"] <- "label"
check_data(sample_data)
```

```
Data type check passed.

No NA in data.
```

```
dim(sample_data)
```

```
[1] 20000    22
```

- As required by ShapleyVIC, code the binary outcome as 0/1.
- All variables are continuous.

```r
sample_data$label <- as.numeric(sample_data$label == "TRUE")
head(sample_data)
```

```
  Vital_A Vital_B Vital_C Vital_D Vital_E Vital_F Vital_G Lab_A Lab_B Lab_C
1      87     143      78     101      13    35.7      99   160  13.0    23
2      43     133      64      83      20    36.1      95   116  15.3    24
3      80     115      48      72      23    37.4      99   133   8.0    27
4     106     121      68      84      16    37.6      99   206  12.1    25
5      86     135      70      83      24    37.2      96   100  18.1    26
6      69     123      72      88      16    36.5      95   204  19.9    20
  Lab_D Lab_E Lab_F Lab_G Lab_H Lab_I Lab_J Lab_K Lab_L Lab_M Age label
1   0.0   105    34    12   0.8    98   4.4     0   136    16  66     0
2   0.8   108    36    12   0.6   322   4.3    55   141    17  79     0
3   1.3   111    30    11   2.9     0   4.4    40   142     0  86     0
4   0.0   102    39    14   3.0   214   4.4     0   134     6  69     0
5   2.3    96    36    13   2.7   326   3.8    20   134    26  65     0
6   2.5   101    31    10   0.8   103   4.2    38   138    14  68     0
```

4.1.2 Prepare training, validation, and test datasets

- Given large sample size (n=20000), split the data into training (70%), validation (10%) and test (20%) sets for risk score development.

```r
set.seed(4)
out_split <- split_data(data = sample_data, ratio = c(0.7, 0.1, 0.2))
train_set <- out_split$train_set
dim(train_set)
```

```
[1] 14000    22
```

```r
validation_set <- out_split$validation_set
dim(validation_set)
```

```
[1] 2000   22
```

```r
test_set <- out_split$test_set
dim(test_set)
```

```
[1] 4000   22
```

- Prepare `output_dir` for ShapleyVIC, using `train_set` as training set and `validation_set` as the explanation data.

> **!** Important
>
> - *As detailed in Chapter 1, check for and handle data issues before applying ShapleyVIC.*
> - *This demo uses data as-is because it is simulated clean data.*

```r
output_dir <- "mort_output"
if (!dir.exists(output_dir)) dir.create(output_dir)
write.csv(train_set, file = file.path(output_dir, "train_set.csv"),
          row.names = FALSE)
write.csv(validation_set,
          file = file.path(output_dir, "validation_set.csv"),
          row.names = FALSE)
```

4.2 [Python] Compute ShapleyVIC values

This part of the workflow is implemented in Python.

- Load data and set up input information.

```python
import os
import pandas as pd
output_dir = "mort_output"
dat_train = pd.read_csv(os.path.join(output_dir, 'train_set.csv'))
dat_expl = pd.read_csv(os.path.join(output_dir, 'validation_set.csv'))

y_name = 'label'
from ShapleyVIC import model
model_object = model.models(
    x=dat_train.drop(columns=[y_name]), y=dat_train[y_name],
    # No need to specify x_names_cat because all variables are continuous
    outcome_type="binary", output_dir=output_dir
)
```

- Set values for hyper-parameters `u1` and `u2`.

```
u1, u2 = model_object.init_hyper_params(m=200)
(u1, u2)
```

```
(0.5, 15.625)
```

- Draw 250 nearly optimal models from 500 initial samples.

```
model_object.draw_models(u1=u1, u2=u2, m=500, n_final=250, random_state=1234)
model_object.models_plot
```

- Compute ShapleyVIC values.

```
from ShapleyVIC import compute
m_svic = compute.compute_shapley_vic(
    model_obj=model_object,
    x_expl=dat_expl.drop(columns=[y_name]), y_expl=dat_expl[y_name],
    n_cores=25, # running on a PC with 40 logical processors
    threshold=0.05
)
```

> **Note**
>
> - *For users' reference, the command above took approximately 17 hours on a PC (Windows 10 Education; Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz 2.19GHz (2 processors); 128GB RAM).*

4.3 [R] Develop risk score

This part of the workflow is implemented in R.

4.3.1 Rank variables using ShapleyVIC

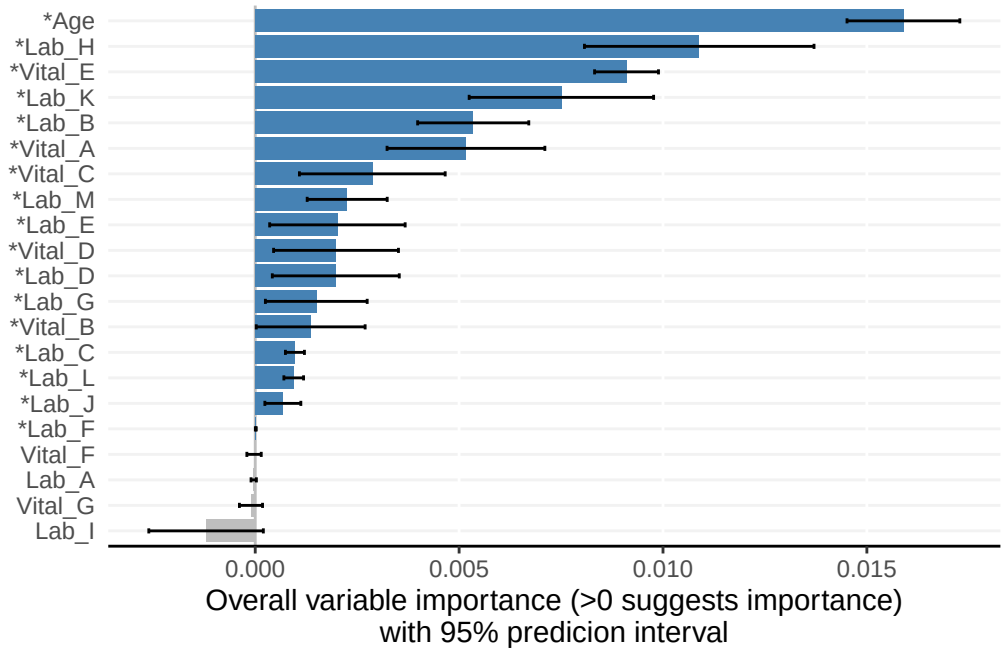
- Compile ShapleyVIC output.

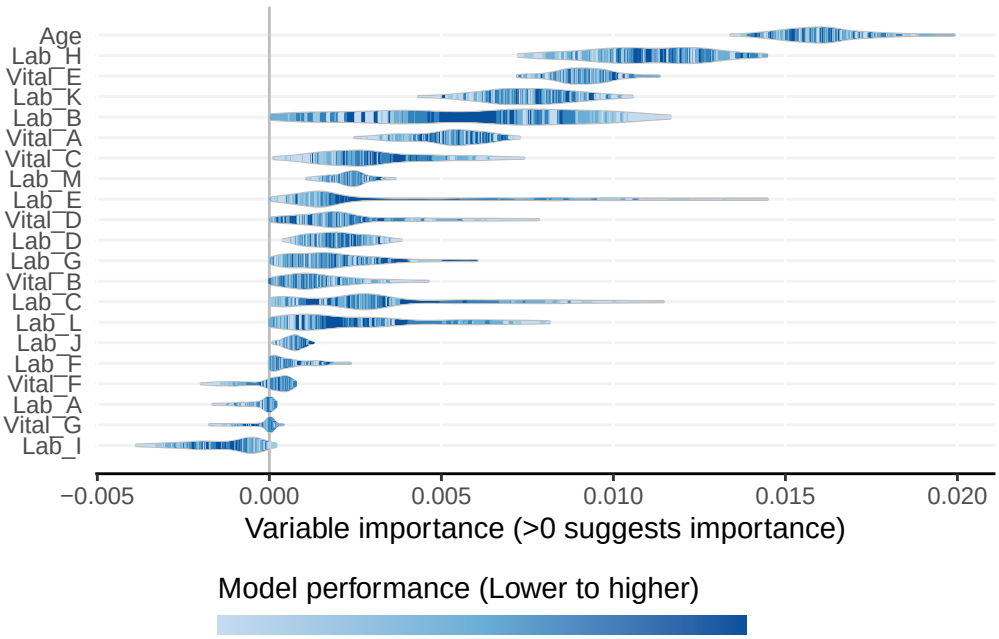
```
library(ShapleyVIC)
model_object <- compile_shapley_vic(
  output_dir = output_dir, outcome_type = "binary"
)
```

```
Compiling results for binary outcome using loss criterion to define neaerly optimal models.
```

- Visualize ShapleyVIC values for overall variable importance.

```
model_plots <- plot(model_object)
```





- Derive ShapleyVIC-based ensemble variable ranking.

```
ranking <- rank_variables(model_object, summarise = TRUE, as_vector = TRUE)
ranking
```

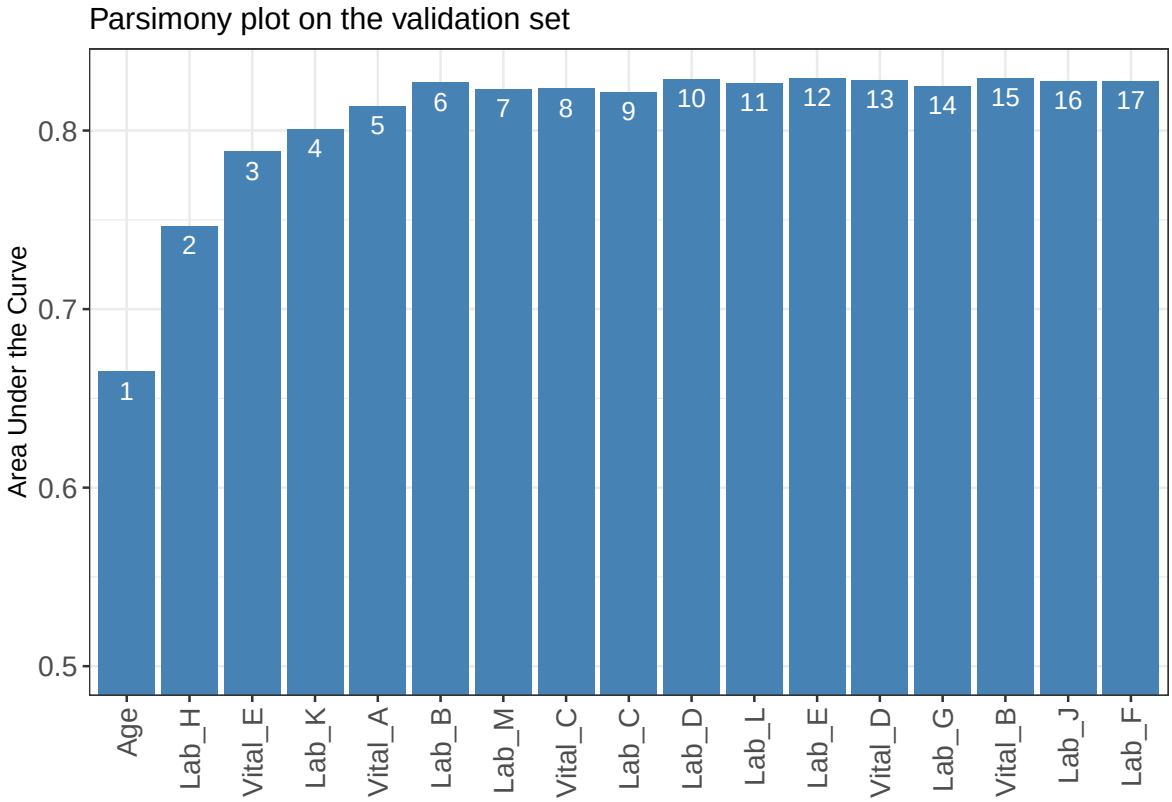
```
   Age   Lab_H Vital_E   Lab_K Vital_A   Lab_B   Lab_M Vital_C   Lab_C   Lab_D
 1.000   2.116   3.028   4.104   5.588   5.744   8.904   8.940   9.084   9.704
 Lab_L   Lab_E Vital_D   Lab_G Vital_B   Lab_J   Lab_F
10.120  10.268  11.628  12.952  13.220  13.328  16.052
```

4.3.2 Develop risk score using AutoScore workflow

- Based on the ensemble variable ranking, apply AutoScore STEP(ii) to select the best model with parsimony plot.

```
AUC <- AutoScore_parsimony(
  train_set = train_set, validation_set = validation_set,
  rank = ranking, max_score = 100, n_min = 1, n_max = length(ranking)
)
```

```
Select 1 Variable(s):  Area under the curve: 0.6649
Select 2 Variable(s):  Area under the curve: 0.7466
Select 3 Variable(s):  Area under the curve: 0.7881
Select 4 Variable(s):  Area under the curve: 0.8009
Select 5 Variable(s):  Area under the curve: 0.8137
Select 6 Variable(s):  Area under the curve: 0.8268
Select 7 Variable(s):  Area under the curve: 0.8232
Select 8 Variable(s):  Area under the curve: 0.8234
Select 9 Variable(s):  Area under the curve: 0.8215
Select 10 Variable(s):  Area under the curve: 0.8286
Select 11 Variable(s):  Area under the curve: 0.8267
Select 12 Variable(s):  Area under the curve: 0.8294
Select 13 Variable(s):  Area under the curve: 0.8281
Select 14 Variable(s):  Area under the curve: 0.8246
Select 15 Variable(s):  Area under the curve: 0.8293
Select 16 Variable(s):  Area under the curve: 0.8275
Select 17 Variable(s):  Area under the curve: 0.8278
```

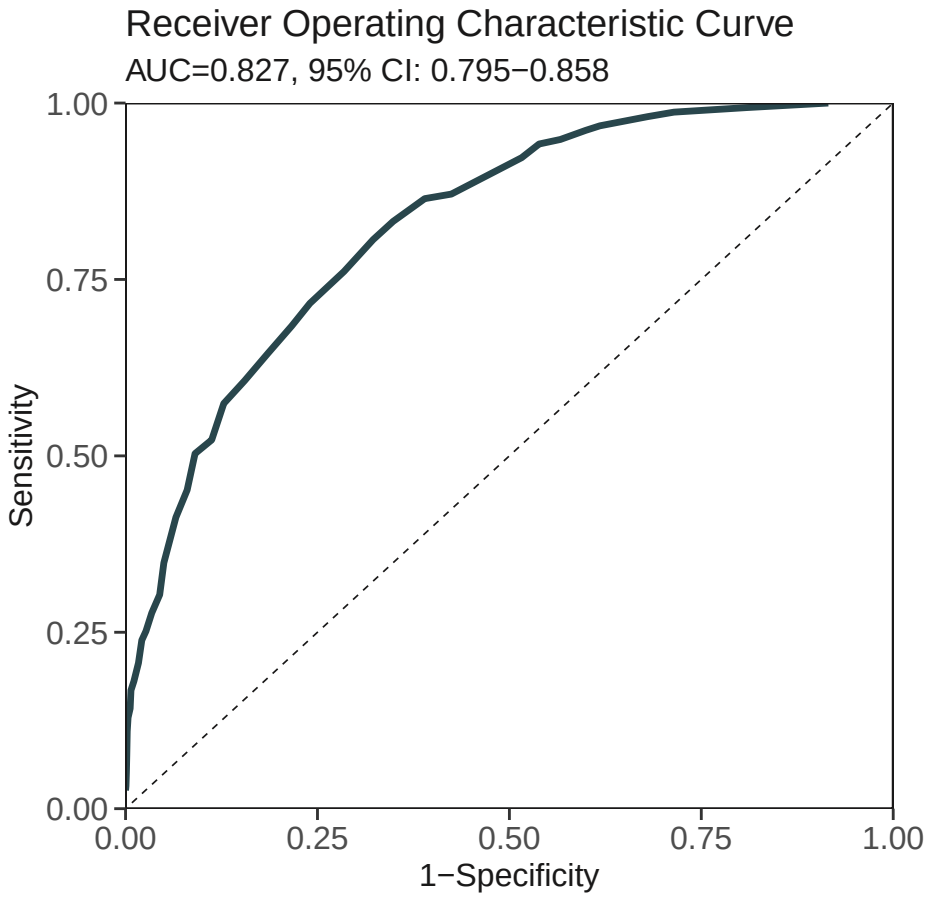


- This parsimony is somewhat smoother than that from random forest-based variable ranking used in AutoScore.
- A feasible choice is to select the top 6 variables, as adding additional variables does not substantially improve model performance.
- Apply AutoScore STEP(iii) to build initial scores from the top 6 variables.

```
cut_vec <- AutoScore_weighting(
  train_set = train_set, validation_set = validation_set,
  final_variables = names(ranking)[1:6], max_score = 100
)
```

```
****Included Variables:
  variable_name
1           Age
2         Lab_H
3       Vital_E
4         Lab_K
5       Vital_A
6         Lab_B
****Initial Scores:


========  ==========  =====
variable  interval    point
========  ==========  =====
Age       <35           0
          [35,49)       7
          [49,76)      17
          [76,89)      23
          >=89         27

Lab_H     <0.2          0
          [0.2,1.1)     4
          [1.1,3.1)     9
          [3.1,4)      15
          >=4          18

Vital_E   <12           0
          [12,15)       2
          [15,22)       7
          [22,25)      12
          >=25         15
```

Lab_K	<8	0
	[8,42)	6
	[42,58)	11
	>=58	14
Vital_A	<60	0
	[60,73)	1
	[73,98)	6
	[98,111)	10
	>=111	13
Lab_B	<8.5	0
	[8.5,11.2)	4
	[11.2,17)	7
	[17,19.8)	10
	>=19.8	12




```
***Performance (based on validation set):
AUC:  0.8268   95% CI: 0.7953-0.8583 (DeLong)
Best score threshold: >= 57
Other performance indicators based on this score threshold:
Sensitivity: 0.8065
Specificity: 0.6775
PPV:         0.1736
NPV:         0.9766
***The cutoffs of each variable generated by the AutoScore are saved in cut_vec. You can dec
```

- Users can apply additional AutoScore STEPs for subsequent model fine-tuning and evaluation.

5 ShapleyVIC for Ordinal Outcomes

As introduced in the ShapleyVIC paper, this method can be applied to regression models beyond the logistic regression. This chapter provides a reproducible example to demonstrate its application for ordinal outcomes using a simulated dataset with ordinal outcome from the AutoScore package. The data is described in detail in the AutoScore Guidebook.

Specifically, as demonstrated in a recent clinical application, we use ShapleyVIC to analyse the importance of all candidate variables in the simulated dataset, exclude variables that have non-significant contribution to prediction, and apply the stepwise variable selection (starting with all significant contributors) to build sparse regression models for prediction.

5.1 [R] Prepare data

This part of the workflow is implemented in R.

5.1.1 Load data

- Load `sample_data_ordinal` from the AutoScore package.
- Variable `label` is a simulated outcome label with 3 ordered categories.
- Among the 20 predictor variables, `Gender`, `Util_A` and the 5 comorbidity variables (`Comorb_A` to `Comorb_E`) are categorical, and the rest are continuous.

```
library(AutoScore)
library(dplyr)
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(purrr)
data("sample_data_ordinal")
head(sample_data_ordinal)
```

```
  label Age Gender Util_A Util_B Util_C    Util_D Comorb_A Comorb_B Comorb_C
1     1  63 FEMALE     P2      0   0.00 3.5933333        0        0        0
2     1  41 FEMALE     P2      0   0.96 3.6288889        0        0        0
3     1  86   MALE     P1      0   0.00 2.6502778        0        0        0
4     1  51   MALE     P2      0   0.00 4.9711111        0        0        0
5     1  23 FEMALE     P1      0   0.00 0.5352778        0        0        0
6     1  32 FEMALE     P2      0   4.13 4.4008333        0        0        0
  Comorb_D Comorb_E Lab_A Lab_B Lab_C Vital_A Vital_B Vital_C Vital_D Vital_E
1        0        0   117   3.9   136      91      19     100      70     152
2        1        0   500   3.6   114      91      16     100      70     147
3        0        0    72   4.1   136     100      18      99      65     126
4        0        0    67   5.0   122      73      17      97      46     100
5        0        0  1036   4.1   138      74      18      98      89     114
6        0        0   806   4.1   136      77      18      98      74     157
  Vital_F
1    25.7
2    22.6
3    25.7
4    24.9
5    25.7
6    25.3
```

```r
dim(sample_data_ordinal)
```

```
[1] 20000    21
```

```r
summary(sample_data_ordinal)
```

```
 label          Age            Gender            Util_A          Util_B
 1:16360   Min.   : 18.00   FEMALE:10137   P1       : 3750   Min.   : 0.0000
```

```
 2: 2449   1st Qu.: 50.00   MALE  : 9863   P2       :11307   1st Qu.: 0.0000
 3: 1191   Median : 64.00                  P3 and P4: 4943   Median : 0.0000
           Mean   : 61.68                                    Mean   : 0.9267
           3rd Qu.: 76.00                                    3rd Qu.: 1.0000
           Max.   :109.00                                    Max.   :42.0000
    Util_C            Util_D         Comorb_A  Comorb_B  Comorb_C  Comorb_D
 Min.   :  0.000   Min.   : 0.09806   0:18445   0:17401   0:19474   0:18113
 1st Qu.:  0.000   1st Qu.: 1.52819   1: 1555   1: 2599   1:  526   1: 1887
 Median :  0.600   Median : 2.46306
 Mean   :  3.535   Mean   : 2.76030
 3rd Qu.:  3.970   3rd Qu.: 3.61472
 Max.   :293.680   Max.   :23.39056
 Comorb_E      Lab_A            Lab_B           Lab_C          Vital_A
 0:19690   Min.   :  16.0   Min.   :1.500   Min.   :102.0   Min.   :  0.00
 1:  310   1st Qu.:  66.0   1st Qu.:3.800   1st Qu.:133.0   1st Qu.: 70.00
           Median :  83.0   Median :4.100   Median :136.0   Median : 81.00
           Mean   : 146.9   Mean   :4.155   Mean   :135.2   Mean   : 82.67
           3rd Qu.: 115.0   3rd Qu.:4.400   3rd Qu.:138.0   3rd Qu.: 93.00
           Max.   :3534.0   Max.   :8.800   Max.   :170.0   Max.   :197.00
   Vital_B         Vital_C          Vital_D          Vital_E
 Min.   : 1.00   Min.   :  0.00   Min.   :  5.00   Min.   :  0.0
 1st Qu.:17.00   1st Qu.: 97.00   1st Qu.: 62.00   1st Qu.:116.0
 Median :18.00   Median : 98.00   Median : 70.00   Median :131.0
 Mean   :17.86   Mean   : 97.96   Mean   : 71.23   Mean   :133.5
 3rd Qu.:18.00   3rd Qu.: 99.00   3rd Qu.: 79.00   3rd Qu.:148.0
 Max.   :48.00   Max.   :100.00   Max.   :180.00   Max.   :262.0
   Vital_F
 Min.   : 2.30
 1st Qu.:21.10
 Median :23.00
 Mean   :22.82
 3rd Qu.:24.80
 Max.   :44.30
```

- Recode the outcome labels to start from 0, which is required by ShapleyVIC.

```r
sample_data_ordinal$label <- as.ordered(as.numeric(sample_data_ordinal$label) - 1)
table(sample_data_ordinal$label)
```

```
    0     1     2
16360  2449  1191
```

5.1.2 Prepare training, validation, and test datasets

- Given large sample size (n=20000), split the data into training (70%), validation (10%) and test (20%) sets for regression model development.
- Stratify by the outcome variable (`label`) when splitting data.

```
set.seed(4)
out_split <- split_data(data = sample_data_ordinal, ratio = c(0.7, 0.1, 0.2),
                        strat_by_label = TRUE)
train_set <- out_split$train_set
dim(train_set)
```

```
[1] 14000    21
```

```
validation_set <- out_split$validation_set
dim(validation_set)
```

```
[1] 2000   21
```

```
test_set <- out_split$test_set
dim(test_set)
```

```
[1] 4000   21
```

- Prepare `ord_output` for ShapleyVIC, using `train_set` as training set and `validation_set` as the explanation data.

> **!** Important
>
> - *As detailed in Chapter 1, check for and handle data issues before applying ShapleyVIC. This demo uses data as-is because it is simulated clean data.*
> - *In this example the* ***`validation_set`*** *has 2000 samples, which is a reasonable sample size to be used as the explanation data. In cases with larger sample sizes, users should use a smaller subset as the explanation data (see Chapter 1 for detail).*

```
output_dir <- "ord_output"
if (!dir.exists(output_dir)) dir.create(output_dir)
write.csv(train_set, file = file.path(output_dir, "train_set.csv"),
          row.names = FALSE)
```

```r
write.csv(validation_set,
          file = file.path(output_dir, "validation_set.csv"),
          row.names = FALSE)
```

5.2 [Python] Compute ShapleyVIC values

This part of the workflow is implemented in Python.

- Load data and set up input information.
- For ordinal outcome, specify `outcome_type='ordinal'`.

```python
import os
import pandas as pd
output_dir = "ord_output"
dat_train = pd.read_csv(os.path.join(output_dir, 'train_set.csv'))
dat_expl = pd.read_csv(os.path.join(output_dir, 'validation_set.csv'))

y_name = 'label'
from ShapleyVIC import model
model_object = model.models(
    x=dat_train.drop(columns=[y_name]), y=dat_train[y_name],
    x_names_cat=['Gender', 'Util_A', 'Comorb_A', 'Comorb_B', 'Comorb_C', 'Comorb_D', 'Como
    outcome_type='ordinal', output_dir=output_dir
)
```

- Set values for hyper-parameters `u1` and `u2`.

```python
u1, u2 = model_object.init_hyper_params()
(u1, u2)
```

```
(0.5, 43.75)
```

- Draw 250 nearly optimal models from 500 initial samples.

```python
model_object.draw_models(u1=u1, u2=u2, m=500, n_final=250, random_state=1234)
model_object.models_plot
```

- Compute ShapleyVIC values.

```
from ShapleyVIC import compute
m_svic = compute.compute_shapley_vic(
    model_obj=model_object,
    x_expl=dat_expl.drop(columns=[y_name]), y_expl=dat_expl[y_name],
    n_cores=20, # running on a PC with 40 logical processors
    threshold=0.05
)
```

> **Note**
>
> - *For users' reference, the command above took approximately 18 hours on a PC (Windows 10 Education; Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz 2.19GHz (2 processors); 128GB RAM).*

5.3 [R] Develop prediction model

This part of the workflow is implemented in R.

5.3.1 Overall variable importance from ShapleyVIC

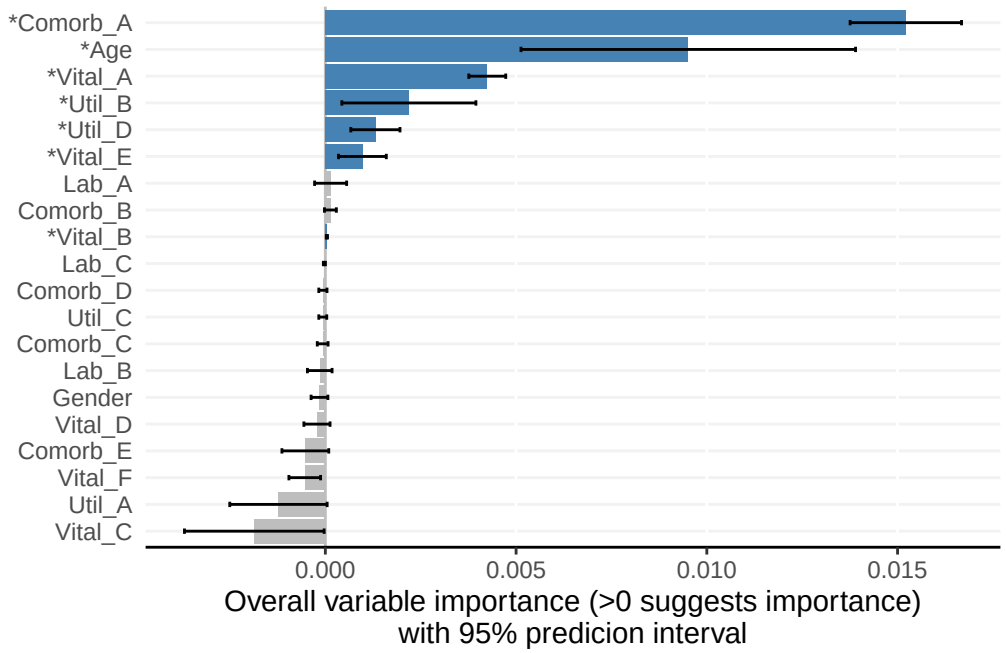
- Compile ShapleyVIC output.

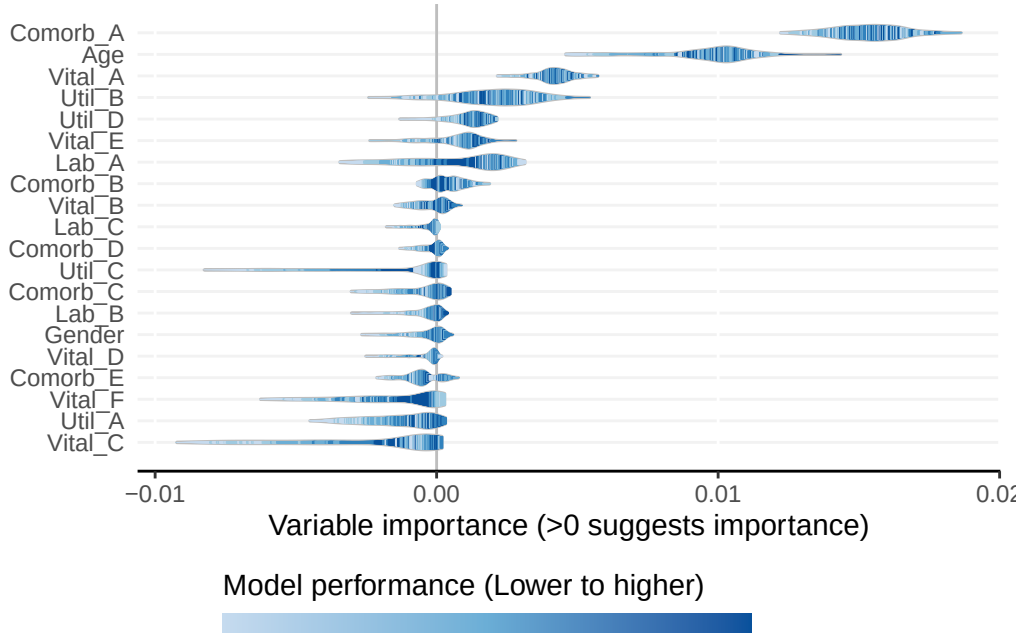
```
output_dir <- "ord_output"
library(ShapleyVIC)
model_object <- compile_shapley_vic(
  output_dir = output_dir, outcome_type = "ordinal",
  x_names_cat = c('Gender', 'Util_A', 'Comorb_A', 'Comorb_B', 'Comorb_C',
                  'Comorb_D', 'Comorb_E')
)
```

```
Compiling results for ordinal outcome using loss criterion to define neaerly optimal models.
```

- Visualize ShapleyVIC values for overall variable importance.

```
model_plots <- plot(model_object)
```


*Comorb_A
*Age
*Vital_A
*Util_B
*Util_D
*Vital_E
Lab_A
Comorb_B
*Vital_B
Lab_C
Comorb_D
Util_C
Comorb_C
Lab_B
Gender
Vital_D
Comorb_E
Vital_F
Util_A
Vital_C
0.000
0.005
0.010
0.015
Overall variable importance (>0 suggests importance)
with 95% predicion interval


Comorb_A
Age
Vital_A
Util_B
Util_D
Vital_E
Lab_A
Comorb_B
Vital_B
Lab_C
Comorb_D
Util_C
Comorb_C
Lab_B
Gender
Vital_D
Comorb_E
Vital_F
Util_A
Vital_C
−0.01
0.00
0.01
Variable importance (>0 suggests importance)
Model performance (Lower to higher)

5.3.2 ShapleyVIC-assisted backward selection

- Identify variables with significant overall importance.

```
vars_svic <- rank_variables(model_object, summarise = TRUE, as_vector = TRUE) %>%
  names()
vars_svic
```

```
[1] "Comorb_A" "Age"      "Vital_A"  "Util_D"   "Util_B"   "Vital_E"  "Vital_B"
```

- Starting with a model that include all variables above, develop a sparse regression model using AIC-based stepwise selection (implemented by the MASS package).
- Using the ordinal package to develop ordinal regression models, more specifically cumulative link model (CLM) with the logit link. See the ordinal package for detailed usage.

```
# Model with all ShapleyVIC-selected variables:
library(ordinal)
```

```
Attaching package: 'ordinal'
```

```
The following object is masked from 'package:dplyr':

    slice
```

```
m_svic_all <- clm(label ~ ., data = train_set[, c("label", vars_svic)])
summary(m_svic_all)
```

```
formula: label ~ Comorb_A + Age + Vital_A + Util_D + Util_B + Vital_E + Vital_B
data:    train_set[, c("label", vars_svic)]

 link  threshold nobs  logLik   AIC      niter max.grad cond.H
 logit flexible  14000 -7726.66 15471.32 5(0)  7.24e-07 1.0e+07

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
Comorb_A1  1.5184021  0.0659862  23.011  < 2e-16 ***
Age        0.0195374  0.0013333  14.654  < 2e-16 ***
```

```
Vital_A    0.0105223  0.0012776   8.236  < 2e-16 ***
Util_D    -0.0768266  0.0140619  -5.463 4.67e-08 ***
Util_B     0.1349605  0.0087143  15.487  < 2e-16 ***
Vital_E   -0.0062609  0.0009181  -6.819 9.14e-12 ***
Vital_B   -0.0039738  0.0127336  -0.312    0.755
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
0|1   2.8254     0.2961   9.543
1|2   4.1694     0.2980  13.993
```

```r
# Backward selection:
library(MASS)
```

```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```r
m_svic <- stepAIC(object = m_svic_all, scope = list(upper = ~ ., lower = ~ 1),
                  trace = FALSE)
summary(m_svic)
```

```
formula: label ~ Comorb_A + Age + Vital_A + Util_D + Util_B + Vital_E
data:    train_set[, c("label", vars_svic)]

 link  threshold nobs  logLik   AIC      niter max.grad cond.H
 logit flexible  14000 -7726.71 15469.42 5(0)  7.24e-07 4.1e+06

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
Comorb_A1  1.5186641  0.0659814  23.017  < 2e-16 ***
Age        0.0195405  0.0013332  14.657  < 2e-16 ***
Vital_A    0.0105274  0.0012775   8.240  < 2e-16 ***
Util_D    -0.0768318  0.0140616  -5.464 4.66e-08 ***
Util_B     0.1349614  0.0087136  15.489  < 2e-16 ***
```

```
Vital_E   -0.0062629  0.0009181  -6.822 8.99e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
0|1   2.8967     0.1885   15.37
1|2   4.2407     0.1914   22.15
```

```r
# Variables selected:
(x_svic <- setdiff(names(m_svic$model), "label"))
```

```
[1] "Comorb_A" "Age"      "Vital_A"  "Util_D"   "Util_B"   "Vital_E"
```

5.3.3 Compare with conventional backward selection

- Backward selection from all candidate variables.

```r
# Model with all variables:
m_all <- clm(formula = label ~ ., data = train_set)
summary(m_all)
```

```
formula:
label ~ Age + Gender + Util_A + Util_B + Util_C + Util_D + Comorb_A + Comorb_B + Comorb_C + 
data:    train_set

 link  threshold nobs  logLik   AIC      niter max.grad cond.H
 logit flexible  14000 -7706.19 15458.38 5(0)  8.70e-07 4.2e+08

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
Age              0.0196646  0.0013361  14.718  < 2e-16 ***
GenderMALE      -0.0890106  0.0452970  -1.965 0.049409 *
Util_AP2         0.0036111  0.0601693   0.060 0.952144
Util_AP3 and P4 -0.0231994  0.0696219  -0.333 0.738969
Util_B           0.1360881  0.0087351  15.579  < 2e-16 ***
Util_C          -0.0003512  0.0024950  -0.141 0.888066
Util_D          -0.0769431  0.0140759  -5.466 4.60e-08 ***
Comorb_A1        1.5267179  0.0661465  23.081  < 2e-16 ***
Comorb_B1        0.0322990  0.0664355   0.486 0.626846
```

```
Comorb_C1        0.2021264  0.1351520   1.496 0.134771
Comorb_D1        0.0261648  0.0771122   0.339 0.734377
Comorb_E1       -0.1684798  0.1917073  -0.879 0.379489
Lab_A            0.0004328  0.0001057   4.092 4.27e-05 ***
Lab_B           -0.0069865  0.0332327  -0.210 0.833489
Lab_C           -0.0067104  0.0046660  -1.438 0.150389
Vital_A          0.0106528  0.0012804   8.320  < 2e-16 ***
Vital_B         -0.0040444  0.0127466  -0.317 0.751019
Vital_C         -0.0013990  0.0069806  -0.200 0.841155
Vital_D          0.0006639  0.0016693   0.398 0.690843
Vital_E         -0.0062020  0.0009198  -6.743 1.55e-11 ***
Vital_F         -0.0243927  0.0063573  -3.837 0.000125 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
0|1    1.299      1.010   1.287
1|2    2.647      1.010   2.620
```

```r
# Backward selection:
m_back <- stepAIC(object = m_all, scope = list(upper = ~ ., lower = ~ 1),
                  trace = FALSE)
summary(m_back)
```

```
formula:
label ~ Age + Gender + Util_B + Util_D + Comorb_A + Comorb_C + Lab_A + Vital_A + Vital_E + Vi
data:    train_set

 link  threshold nobs  logLik   AIC      niter max.grad cond.H
 logit flexible  14000 -7708.06 15440.13 5(0)  8.70e-07 1.8e+07

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
Age         0.0196518  0.0013353  14.717  < 2e-16 ***
GenderMALE -0.0906919  0.0452646  -2.004  0.04511 *
Util_B      0.1358021  0.0087232  15.568  < 2e-16 ***
Util_D     -0.0771132  0.0140754  -5.479 4.29e-08 ***
Comorb_A1   1.5262185  0.0660905  23.093  < 2e-16 ***
Comorb_C1   0.2026533  0.1350978   1.500  0.13360
Lab_A       0.0004345  0.0001057   4.111 3.94e-05 ***
```

```
Vital_A     0.0106409  0.0012794   8.317  < 2e-16 ***
Vital_E    -0.0062111  0.0009193  -6.756 1.42e-11 ***
Vital_F    -0.0243158  0.0063537  -3.827  0.00013 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
0|1   2.3946     0.2376   10.08
1|2   3.7417     0.2397   15.61
```

```r
# Variables selected:
(x_back <- setdiff(names(m_back$model), "label"))
```

```
 [1] "Age"      "Gender"   "Util_B"   "Util_D"   "Comorb_A" "Comorb_C"
 [7] "Lab_A"    "Vital_A"  "Vital_E"  "Vital_F"
```

- ShapleyVIC-assisted backward selection developed a more parsimonious model (with 6 variables) than conventional backward selection (with 10 variables) without significantly impairing performance.

```r
# Performance of model from ShapleyVIC-assisted backward selection on test set:
fx_svic <- model.matrix(~ ., data = test_set[, x_svic])[, -1] %*% m_svic$beta
print_performance_ordinal(label = test_set$label, score = as.numeric(fx_svic),
                          n_boot = 100, report_cindex = TRUE)
```

```
mAUC: 0.7291     95% CI: 0.7062-0.7471 (from 100 bootstrap samples)
Generalised c-index: 0.6990      95% CI: 0.6803-0.7146 (from 100 bootstrap samples)
```

```r
# Performance of model from conventional backward selection on test set:
fx_back <- model.matrix(~ ., data = test_set[, x_back])[, -1] %*% m_back$beta
print_performance_ordinal(label = test_set$label, score = as.numeric(fx_back),
                          n_boot = 100, report_cindex = TRUE)
```

```
mAUC: 0.7351     95% CI: 0.7165-0.7559 (from 100 bootstrap samples)
Generalised c-index: 0.7033      95% CI: 0.6881-0.7197 (from 100 bootstrap samples)
```

6 ShapleyVIC for Continuous Outcomes

As introduced in the ShapleyVIC paper, this method can be applied to regression models beyond the logistic regression. This chapter provides a reproducible example to demonstrate its application for continuous outcomes using a simulated dataset from the AutoScore package, which is described in detail in the AutoScore Guidebook.

Specifically, as demonstrated in a recent clinical application, we use ShapleyVIC to analyse the importance of all candidate variables in the simulated dataset, exclude variables that have non-significant contribution to prediction, and apply the stepwise variable selection (starting with all significant contributors) to build sparse regression models for prediction.

> **!** Important
>
> - Although the outcome in this dataset is ordinal with 3 ordered categories, for demonstrative purpose we consider it as a continuous outcome in this chapter.
> - The previous chapter demonstrates how to analyse the same outcome as an ordinal variable.
> - In real-life clinical applications, users should take into consideration the nature of the outcome when choosing an analysis approach.

6.1 [R] Prepare data

This part of the workflow is implemented in R.

6.1.1 Load data

- Load `sample_data_ordinal` from the AutoScore package.
- Variable `label` is a simulated outcome label with 3 ordered categories.
- Among the 20 predictor variables, `Gender`, `Util_A` and the 5 comorbidity variables (`Comorb_A` to `Comorb_E`) are categorical, and the rest are continuous.

```
library(AutoScore)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(purrr)
data("sample_data_ordinal")

head(sample_data_ordinal)
```

```
  label Age Gender Util_A Util_B Util_C    Util_D Comorb_A Comorb_B Comorb_C
1     1  63 FEMALE     P2      0   0.00 3.5933333        0        0        0
2     1  41 FEMALE     P2      0   0.96 3.6288889        0        0        0
3     1  86   MALE     P1      0   0.00 2.6502778        0        0        0
4     1  51   MALE     P2      0   0.00 4.9711111        0        0        0
5     1  23 FEMALE     P1      0   0.00 0.5352778        0        0        0
6     1  32 FEMALE     P2      0   4.13 4.4008333        0        0        0
  Comorb_D Comorb_E Lab_A Lab_B Lab_C Vital_A Vital_B Vital_C Vital_D Vital_E
1        0        0   117   3.9   136      91      19     100      70     152
2        1        0   500   3.6   114      91      16     100      70     147
3        0        0    72   4.1   136     100      18      99      65     126
4        0        0    67   5.0   122      73      17      97      46     100
5        0        0  1036   4.1   138      74      18      98      89     114
6        0        0   806   4.1   136      77      18      98      74     157
  Vital_F
1    25.7
2    22.6
3    25.7
4    24.9
5    25.7
6    25.3
```

```r
dim(sample_data_ordinal)
```

```
[1] 20000    21
```

```
summary(sample_data_ordinal)
```

```
label          Age             Gender            Util_A          Util_B
1:16360   Min.   : 18.00   FEMALE:10137   P1       : 3750   Min.   : 0.0000
2: 2449   1st Qu.: 50.00   MALE  : 9863   P2       :11307   1st Qu.: 0.0000
3: 1191   Median : 64.00                  P3 and P4: 4943   Median : 0.0000
          Mean   : 61.68                                    Mean   : 0.9267
          3rd Qu.: 76.00                                    3rd Qu.: 1.0000
          Max.   :109.00                                    Max.   :42.0000
    Util_C            Util_D          Comorb_A  Comorb_B  Comorb_C  Comorb_D
Min.   :  0.000   Min.   : 0.09806   0:18445   0:17401   0:19474   0:18113
1st Qu.:  0.000   1st Qu.: 1.52819   1: 1555   1: 2599   1:  526   1: 1887
Median :  0.600   Median : 2.46306
Mean   :  3.535   Mean   : 2.76030
3rd Qu.:  3.970   3rd Qu.: 3.61472
Max.   :293.680   Max.   :23.39056
Comorb_E      Lab_A            Lab_B           Lab_C           Vital_A
0:19690   Min.   :  16.0   Min.   :1.500   Min.   :102.0   Min.   :  0.00
1:  310   1st Qu.:  66.0   1st Qu.:3.800   1st Qu.:133.0   1st Qu.: 70.00
          Median :  83.0   Median :4.100   Median :136.0   Median : 81.00
          Mean   : 146.9   Mean   :4.155   Mean   :135.2   Mean   : 82.67
          3rd Qu.: 115.0   3rd Qu.:4.400   3rd Qu.:138.0   3rd Qu.: 93.00
          Max.   :3534.0   Max.   :8.800   Max.   :170.0   Max.   :197.00
   Vital_B         Vital_C          Vital_D          Vital_E
Min.   : 1.00   Min.   :  0.00   Min.   :  5.00   Min.   :  0.0
1st Qu.:17.00   1st Qu.: 97.00   1st Qu.: 62.00   1st Qu.:116.0
Median :18.00   Median : 98.00   Median : 70.00   Median :131.0
Mean   :17.86   Mean   : 97.96   Mean   : 71.23   Mean   :133.5
3rd Qu.:18.00   3rd Qu.: 99.00   3rd Qu.: 79.00   3rd Qu.:148.0
Max.   :48.00   Max.   :100.00   Max.   :180.00   Max.   :262.0
   Vital_F
Min.   : 2.30
1st Qu.:21.10
Median :23.00
Mean   :22.82
3rd Qu.:24.80
Max.   :44.30
```

6.1.2 Prepare training, validation, and test datasets

- Given large sample size (n=20000), split the data into training (70%), validation (10%) and test (20%) sets for regression model development.
- For convenience, we reuse the stratified data split step from the previous chapter, but convert the outcome ("label") to continuous.

```
set.seed(4)
out_split <- split_data(data = sample_data_ordinal, ratio = c(0.7, 0.1, 0.2),
                        strat_by_label = TRUE)
train_set <- out_split$train_set
# For this demo, convert the outcome ("label") to continuous
train_set$label <- as.numeric(train_set$label)
dim(train_set)
```

```
[1] 14000    21
```

```
validation_set <- out_split$validation_set
# For this demo, convert the outcome ("label") to continuous
validation_set$label <- as.numeric(validation_set$label)
dim(validation_set)
```

```
[1] 2000   21
```

```
test_set <- out_split$test_set
# For this demo, convert the outcome ("label") to continuous
test_set$label <- as.numeric(test_set$label)
dim(test_set)
```

```
[1] 4000   21
```

- Prepare `cont_output` for ShapleyVIC, using `train_set` as training set and `validation_set` as the explanation data.

> **!** Important
>
> - *As detailed in Chapter 1, check for and handle data issues before applying ShapleyVIC. This demo uses data as-is because it is simulated clean data.*
> - *In this example the* `validation_set` *has 2000 samples, which is a reasonable sample*

size to be used as the explanation data. In cases with larger sample sizes, users should use a smaller subset as the explanation data (see Chapter 1 for detail).

```
output_dir <- "cont_output"
if (!dir.exists(output_dir)) dir.create(output_dir)
write.csv(train_set, file = file.path(output_dir, "train_set.csv"),
          row.names = FALSE)
write.csv(validation_set,
          file = file.path(output_dir, "validation_set.csv"),
          row.names = FALSE)
```

6.2 [Python] Compute ShapleyVIC values

This part of the workflow is implemented in Python.

- Load data and set up input information.
- For continuous outcome, specify `outcome_type='continuous'`.

```
import os
import pandas as pd
output_dir = "cont_output"
dat_train = pd.read_csv(os.path.join(output_dir, 'train_set.csv'))
dat_expl = pd.read_csv(os.path.join(output_dir, 'validation_set.csv'))

y_name = 'label'
from ShapleyVIC import model
model_object = model.models(
    x=dat_train.drop(columns=[y_name]), y=dat_train[y_name],
    x_names_cat=['Gender', 'Util_A', 'Comorb_A', 'Comorb_B', 'Comorb_C', 'Comorb_D', 'Como
    outcome_type='continuous', output_dir=output_dir
)
```

- Set values for hyper-parameters `u1` and `u2`.

```
u1, u2 = model_object.init_hyper_params()
(u1, u2)
```

```
(0.5, 77.5)
```

- Draw 250 nearly optimal models from 500 initial samples.

```
model_object.draw_models(u1=u1, u2=u2, m=500, n_final=250, random_state=1234)
model_object.models_plot
```

- Compute ShapleyVIC values.

```
from ShapleyVIC import compute
m_svic = compute.compute_shapley_vic(
    model_obj=model_object,
    x_expl=dat_expl.drop(columns=[y_name]), y_expl=dat_expl[y_name],
    n_cores=20, # running on a PC with 40 logical processors
    threshold=0.05
)
```

> **Note**
>
> - *For users' reference, the command above took approximately 24 hours on a PC (Windows 10 Education; Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz 2.19GHz (2 processors); 128GB RAM).*

6.3 [R] Develop prediction model

This part of the workflow is implemented in R.

6.3.1 Overall variable importance from ShapleyVIC

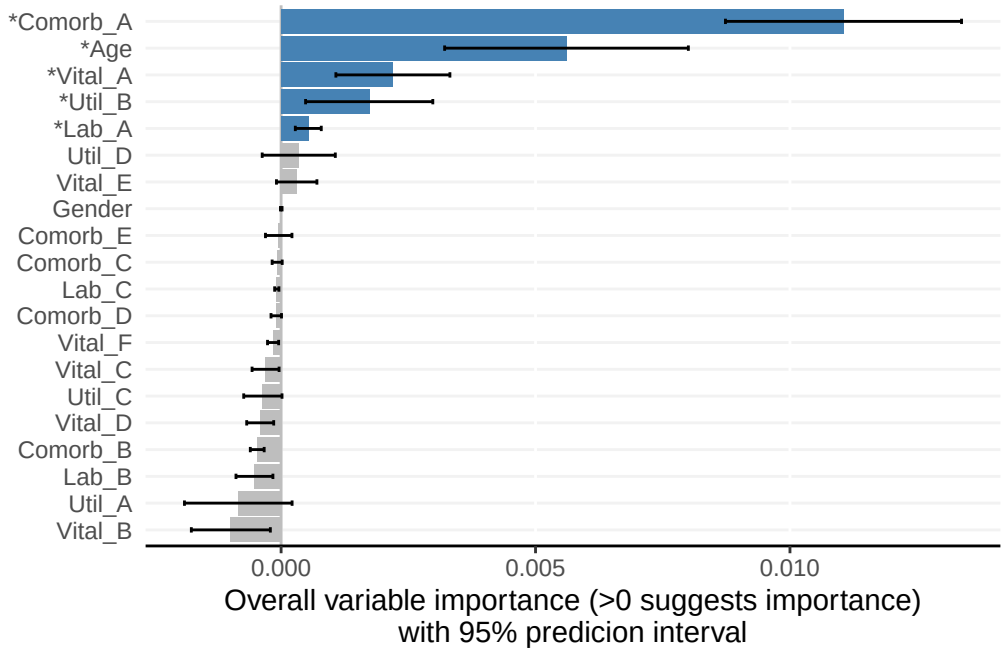
- Compile ShapleyVIC output.

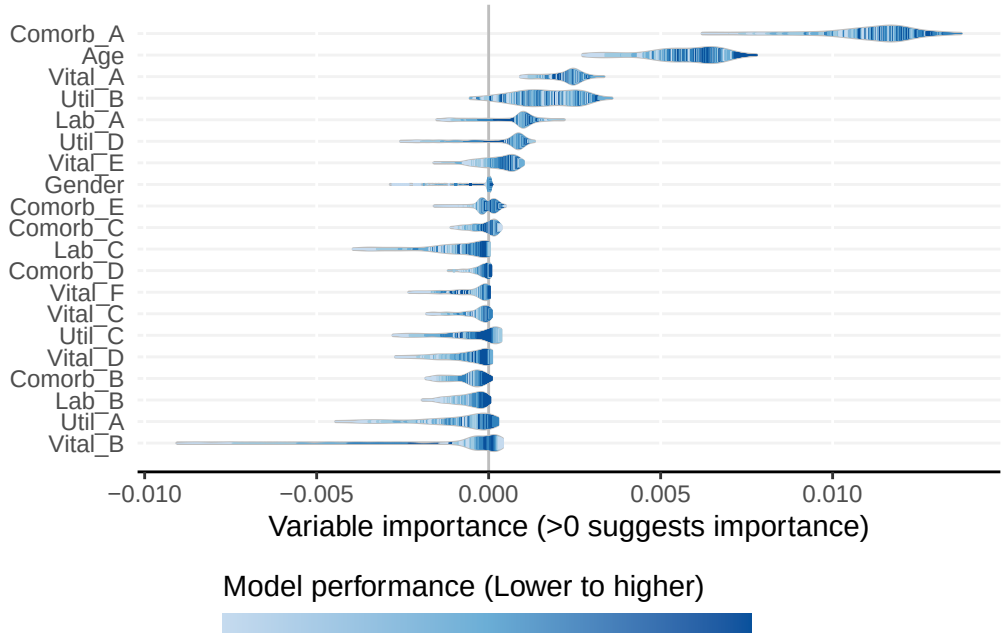
```
output_dir <- "cont_output"
library(ShapleyVIC)
model_object <- compile_shapley_vic(
  output_dir = output_dir, outcome_type = "continuous",
  x_names_cat = c('Gender', 'Util_A', 'Comorb_A', 'Comorb_B', 'Comorb_C',
                  'Comorb_D', 'Comorb_E')
)
```

```
Compiling results for continuous outcome using loss criterion to define neaerly optimal model
```

- Visualize ShapleyVIC values for overall variable importance.

```
model_plots <- plot(model_object)
```





6.3.2 ShapleyVIC-assisted backward selection

- Identify variables with significant overall importance.

```
vars_svic <- rank_variables(model_object, summarise = TRUE, as_vector = TRUE) %>%
  names()
vars_svic
```

```
[1] "Comorb_A" "Age"      "Vital_A"  "Util_B"   "Lab_A"
```

- Starting with a model that include all variables above, develop a sparse regression model using AIC-based stepwise selection (implemented by the MASS package).

```
# Model with all ShapleyVIC-selected variables:
m_svic_all <- lm(label ~ ., data = train_set[, c("label", vars_svic)])
summary(m_svic_all)
```

```
Call:
lm(formula = label ~ ., data = train_set[, c("label", vars_svic)])

Residuals:
     Min       1Q   Median       3Q      Max
-1.59070 -0.24560 -0.17416 -0.06995  1.98002

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.7564288  0.0268424  28.180  < 2e-16 ***
Comorb_A1   0.4327596  0.0165608  26.131  < 2e-16 ***
Age         0.0036084  0.0002453  14.707  < 2e-16 ***
Vital_A     0.0021720  0.0002606   8.335  < 2e-16 ***
Util_B      0.0370323  0.0020596  17.980  < 2e-16 ***
Lab_A       0.0001014  0.0000225   4.507 6.63e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5258 on 13994 degrees of freedom
Multiple R-squared:  0.0861,    Adjusted R-squared:  0.08577
F-statistic: 263.7 on 5 and 13994 DF,  p-value: < 2.2e-16
```

```r
# Backward selection:
library(MASS)
```

```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```r
m_svic <- stepAIC(object = m_svic_all, scope = list(upper = ~ ., lower = ~ 1),
                  trace = FALSE)
summary(m_svic)
```

```
Call:
lm(formula = label ~ Comorb_A + Age + Vital_A + Util_B + Lab_A,
    data = train_set[, c("label", vars_svic)])

Residuals:
     Min       1Q   Median       3Q      Max
-1.59070 -0.24560 -0.17416 -0.06995  1.98002

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.7564288  0.0268424  28.180  < 2e-16 ***
Comorb_A1   0.4327596  0.0165608  26.131  < 2e-16 ***
Age         0.0036084  0.0002453  14.707  < 2e-16 ***
Vital_A     0.0021720  0.0002606   8.335  < 2e-16 ***
Util_B      0.0370323  0.0020596  17.980  < 2e-16 ***
Lab_A       0.0001014  0.0000225   4.507 6.63e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5258 on 13994 degrees of freedom
Multiple R-squared:  0.0861,    Adjusted R-squared:  0.08577
F-statistic: 263.7 on 5 and 13994 DF,  p-value: < 2.2e-16
```

```r
# Variables selected:
(x_svic <- setdiff(names(m_svic$model), "label"))
```

```
[1] "Comorb_A" "Age"      "Vital_A"  "Util_B"   "Lab_A"
```

6.3.3 Compare with conventional backward selection

- Backward selection from all candidate variables.

```r
# Model with all variables:
m_all <- lm(label ~ ., data = train_set)
summary(m_all)
```

```
Call:
lm(formula = label ~ ., data = train_set)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55915 -0.25176 -0.17028 -0.05685  2.12399

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.287e+00  1.977e-01   6.511 7.75e-11 ***
Age              3.596e-03  2.447e-04  14.695  < 2e-16 ***
GenderMALE      -1.563e-02  8.864e-03  -1.763   0.0779 .
Util_AP2        -1.577e-05  1.178e-02  -0.001   0.9989
Util_AP3 and P4 -6.534e-04  1.360e-02  -0.048   0.9617
Util_B           3.718e-02  2.055e-03  18.094  < 2e-16 ***
Util_C           5.271e-05  5.093e-04   0.104   0.9176
Util_D          -1.450e-02  2.579e-03  -5.623 1.91e-08 ***
Comorb_A1        4.301e-01  1.652e-02  26.037  < 2e-16 ***
Comorb_B1        8.332e-03  1.323e-02   0.630   0.5288
Comorb_C1        3.536e-02  2.820e-02   1.254   0.2098
Comorb_D1        3.266e-03  1.521e-02   0.215   0.8300
Comorb_E1       -3.646e-02  3.598e-02  -1.013   0.3110
Lab_A            9.818e-05  2.245e-05   4.374 1.23e-05 ***
Lab_B           -3.988e-03  6.498e-03  -0.614   0.5394
Lab_C           -1.216e-03  9.222e-04  -1.318   0.1874
Vital_A          2.174e-03  2.599e-04   8.362  < 2e-16 ***
```

```
Vital_B          1.323e-04  2.445e-03   0.054   0.9568
Vital_C         -3.688e-04  1.364e-03  -0.270   0.7868
Vital_D          7.918e-05  3.276e-04   0.242   0.8090
Vital_E         -1.130e-03  1.747e-04  -6.467 1.04e-10 ***
Vital_F         -5.429e-03  1.255e-03  -4.327 1.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5242 on 13978 degrees of freedom
Multiple R-squared:  0.0927,    Adjusted R-squared:  0.09134
F-statistic: 68.01 on 21 and 13978 DF,  p-value: < 2.2e-16
```

```r
# Backward selection:
m_back <- stepAIC(object = m_all, scope = list(upper = ~ ., lower = ~ 1),
                  trace = FALSE)
summary(m_back)
```

```
Call:
lm(formula = label ~ Age + Gender + Util_B + Util_D + Comorb_A +
    Lab_A + Vital_A + Vital_E + Vital_F, data = train_set)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55557 -0.25120 -0.17067 -0.05771  2.13009

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.080e+00  4.611e-02  23.419  < 2e-16 ***
Age          3.597e-03  2.446e-04  14.710  < 2e-16 ***
GenderMALE  -1.583e-02  8.859e-03  -1.787   0.0739 .
Util_B       3.715e-02  2.053e-03  18.097  < 2e-16 ***
Util_D      -1.455e-02  2.578e-03  -5.644 1.69e-08 ***
Comorb_A1    4.302e-01  1.651e-02  26.062  < 2e-16 ***
Lab_A        9.902e-05  2.243e-05   4.415 1.02e-05 ***
Vital_A      2.180e-03  2.597e-04   8.392  < 2e-16 ***
Vital_E     -1.132e-03  1.746e-04  -6.482 9.37e-11 ***
Vital_F     -5.443e-03  1.254e-03  -4.341 1.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.524 on 13990 degrees of freedom
Multiple R-squared:  0.09236,   Adjusted R-squared:  0.09178
F-statistic: 158.2 on 9 and 13990 DF,  p-value: < 2.2e-16
```

```r
# Variables selected:
(x_back <- setdiff(names(m_back$model), "label"))
```

```
[1] "Age"      "Gender"   "Util_B"   "Util_D"   "Comorb_A" "Lab_A"    "Vital_A"
[8] "Vital_E"  "Vital_F"
```

- ShapleyVIC-assisted backward selection developed a more parsimonious model (with 5 variables) than conventional backward selection (with 9 variables) without significantly impairing performance.

```r
compute_mse <- function(y, y_pred) {
  mean((y - y_pred) ^ 2)
}
# Mean squared error (MSE) of the two models on test set:
c(ShapleyVIC_assisted_backward_selection = compute_mse(
  y = test_set$label, y_pred = predict(m_svic, newdata = test_set)),
  Conventional_backward_selection = compute_mse(
  y = test_set$label, y_pred = predict(m_back, newdata = test_set))
)
```

```
ShapleyVIC_assisted_backward_selection        Conventional_backward_selection
                             0.2743280                              0.2722482
```