Variable Importance Measures for Variable Selection in Multivariate Random Forests Supplementary Material

Sharmistha Sikdar, Giles Hooker and Vrinda Kadiyali

A.1 Pseudo-codes

A.1.1 MVRF using Sub-bagging Procedure

Algorithm A.1 MVRF using Sub-bagging Procedure

```
**Inputs**: training and testing sets, $x$ and $x^*$, size of subsample $l_N$, number of subsamples $r_N$
**for** $b$ in 1 to $r_N$ **do**
    Select subsample of size $l_N$ from training set $x$
    Build tree on subsample $b$
    Use tree at testing set $x^*$ to get prediction vector $(Y_{N,l_N,r_N})^b$
**end for**
Average the $r_N$ predictions to obtain $(\hat{Y}_{N,l_N,r_N})$
```

A.1.2 SI based VIMs for Significant Splits

Algorithm A.2 Computing SI based variable importance measures for significant splits

```
**Inputs**: training and testing sets, $x$ and $x^*$, subsample size $l_N$, number of subsamples $r_N$
**for** $b$ in 1 to $r_N$ **do**
    Select subsample of size $l_N$ from training set $x$
    Build tree on subsample $b$ with number of splitting nodes $Q_b$
    Use tree to predict on testing set $x^*$
    Initialize VIM vector of dimension $P \times 1$ for tree $b$ as $VIM_0^b = 0$
    **for** $j$ in 1 to $Q_b$ **do**
        Calculate magnitude of SI for split $j$ in tree $b$ as $SI_{bj}$
        Perform F test for $H_0$
        **for** $m$ in 1 to $P$ **do**
            **if** feature $m$ is used for split $j$ in tree $b$ **then**
                **if** $H_0$ is rejected **then**
                    $VIM_{0m}^b = VIM_{0m}^b + SI_{bj}$
                **end if**
            **end if**
            else $VIM_{0m}^b = VIM_{0m}^b$
        **end for**
    **end for**
**end for**
Average the $r_N$ predictions to obtain final estimate $(\hat{Y}_{N,l_N,r_N})$
Average the $r_N$ calculations of VIM vector $VIM_{N,l_N,r_N}^b$ to get $VIM_{N,l_N,r_N}^*$
```

A.1.3 Recursive Feature Elimination Strategy

Algorithm A.3 Proposed Recursive Feature Elimination Strategy

```
Inputs: training and testing sets, $x$ and $x^*$, number of bootstrap samples $B$, maximum number of iterations $maxiter$
Introduce a Gaussian random noise pseudo-variable $r$ to both training and testing sets
for $iter$ in 1 to $maxiter$ do
    for $b$ in 1 to $B$ do
        Build tree on subsample $b$
        Use tree to predict on testing set $x^*$
    end for
    Average across the $B$ trees to compute the average prediction error
    Compute VIM for each feature including pseudo-variable $r$.
    Remove features with VIM lower than that of $r$.
end for
```

A.2 Simulation Studies

Table A.1: Simulation Design

Variables	Non-sparse data setting	Sparse data setting
Explanatory		
X_1, X_5	$Uniform[0,1]$	$Uniform[0,1]$
X_2	$Binomial(1,0.5)$	$Binomial(1,0.5)$
X_3	$Poisson(50)$	$Poisson(50)$
X_4	$Binomial(1,0.25)$	$Binomial(1,0.5)$
Spurious		
X_6, X_8	$Binomial(1,0.2)$	$Binomial(1,0.9)$
X_7, X_{11}	$Uniform[0,1]$	$Uniform[0,1]$
X_9, X_{15}	$Uniform[0,0.5]$	$Uniform[0,0.5]$
X_{10}, X_{12}	$Binomial(1,0.15)$	$Binomial(1,0.9)$
X_{13}	$Uniform[0,0.25]$	$Uniform[0,0.25]$
X_{14}	$Binomial(1,0.125)$	$Binomial(1,1)$

Next, in Tables A.2 and A.3, we show the results of the remaining two simulation scenarios 3 (non-sparse data with correlated errors) and 4 (sparse data with correlated errors).

A.2.1 Scenario 3: Linear Model with non sparse data and correlated errors

Table A.2: Variable Ranking by naive and proposed VIMs under Scenario 3

				Mean Struc.		.. w/ F-test		Outcm. Diff.		.. w/ F-test	
Var.	True rank	Freq.	Incid.	Train	OOB	Train	OOB	Train	OOB	Train	OOB
X_1	1	2	1	2	3	2	3	1	1	1	1
X_2	2	3	4	3	4	3	4	2	3	2	2
X_3	4	1	1	1	2	1	2	3	2	3	3
X4	3	5	5	5	5	5	5	4	5	4	5
X5	5	4	1	4	1	4	1	5	4	5	4
TPR		100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FPR		0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Abbreviations: Freq. = Frequency based VIM, Incid. = Incidence based VIM, Mean Struc. VIM = Mean Structure based VIM, Outcome Diff. = Outcome Difference based VIM.

A.2.2 Scenario 4: Non-Linear Model with sparse data and correlated errors

Table A.3: Variable Ranking by naive and proposed VIMs under Scenario 4

				Mean Struc.		.. w/ F-test		Outcome Diff.		.. w/ F-test	
Var.	True rank	Freq.	Incid.	Train	OOB	Train	OOB	Train	OOB	Train	OOB
X_1	1	1	1	1	1	1	1	1	1	1	1
X_2	2	2	2	2	**9**	2	**7**	2	**9**	2	**8**
X_3	4	3	3	3	2	3	2	4	3	4	2
X_4	3	**6**	4	**7**	**8**	**7**	**9**	3	**8**	3	**9**
X_5	5	4	5	4	3	4	3	**6**	4	**6**	3
TPR		80%	100%	80%	60%	80%	60%	80%	60%	80%	60%
FPR		10%	0%	10%	20%	10%	20%	10%	20%	10%	20%

Bolded numbers indicate ranks that are lower than those of the spurious covariates.

A.3 Box Plots and Confidence Interval of Top 5 Features

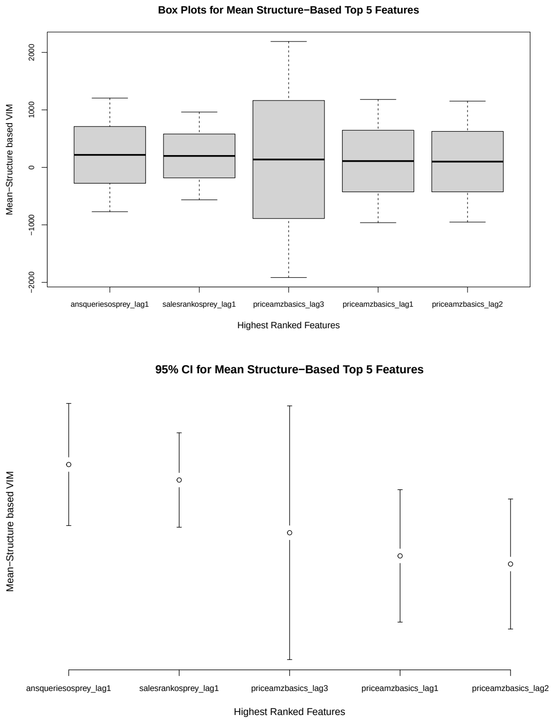

Figure A.1: Mean Structure-based VIM: Top Five Features.

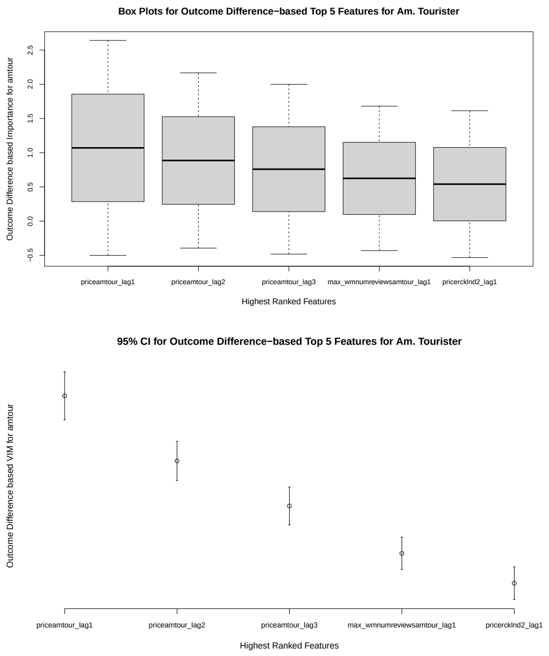

Figure A.2: Outcome Difference VIM-based Top Five Features (Brand: American Tourister).

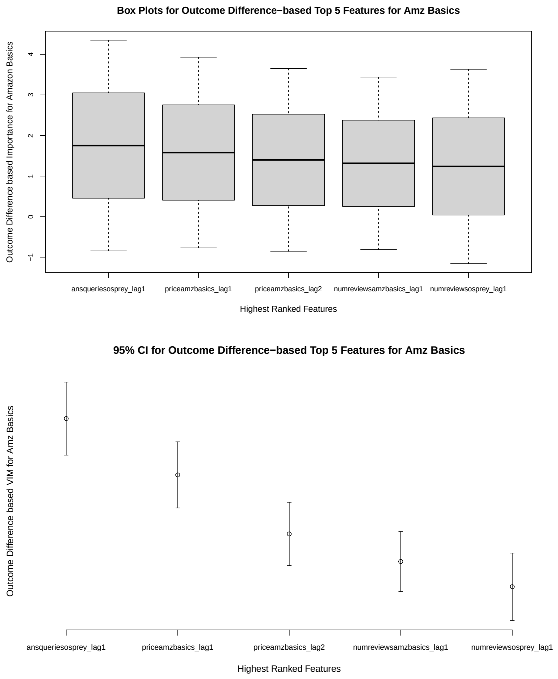

Figure A.3: Outcome Difference VIM-based Top Five Features (Brand: Amazon Basics).

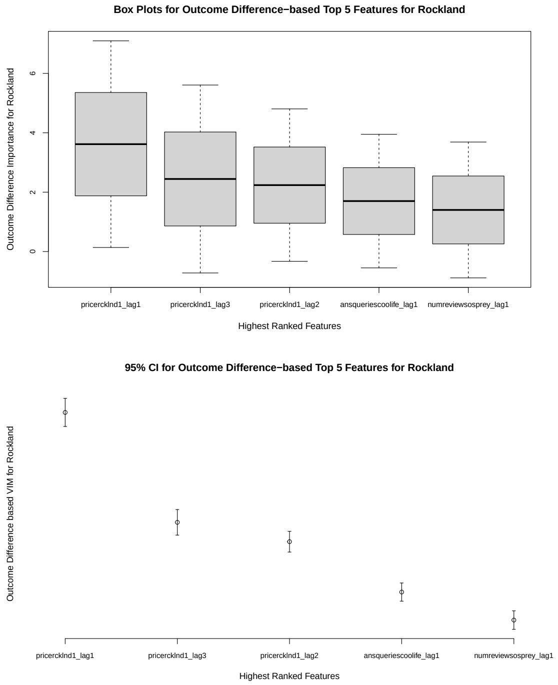

Figure A.4: Outcome Difference VIM-based Top Five Features (Brand: Rockland).