WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

MASTER THESIS

Variable importance measures in regression and classification methods

Institute for

Statistics and Mathematics

Vienna University of Economics and Business

under the supervision of

Univ.Prof. Dipl.-Ing. Dr.techn. Kurt Hornik

submitted by

Dipl.-Ing. Jakob Weissteiner

Jacquingasse 16
1030 Wien

Vienna, September 3, 2018

Contents

Abstract

This master thesis deals with the problem of determining variable importance for different kinds of regression and classification methods. The first chapter introduces relative importance metrics for multiple linear regression, which are based on a decomposition of the coefficient of determination. Chapter 2 serves as an introduction to a general variable importance measure motivated from causal inference, that can in principle be applied to a very large class of models. In Chapter 3 we discuss in detail different importance measures for random forests. In the course of that, we also review the main principles behind random forests by discussing the famous *CART* algorithm. At the end of chapter 3 we extend the unconditional permutation importance, introduced in the context of random forests, to linear and logistic regression. Chapter 4 deals with a heuristic approach to measure relative importance in a logistic regression setting, that is motivated by the *relative weights* method from linear regression. Here, the presented importance measure is as in the first chapter based on the amount of explained variance in the response variable i.e. dispersion importance. Chapter 5 deals with the application of the permutation importance measure on a credit scoring dataset in order to determine the most important predictor variables for an event of default. Simulation studies, which highlight the advantages and disadvantages of each method are presented at the end of each chapter.

Research Question

When building models for e.g. a binary response variable using different kinds of learners like a logit/probit model (possibly with regularization) or random forests, it is often of interest not only to compare these models w.r.t their performance on a test set, but also to compare the models from a structural point of view e.g. the "importance" of single predictors.

We are interested to know if there is a conceptual framework that unifies all or at least some methods for quantifying variable importance in a given regression or classification setting. A literature review on different techniques for measuring variable importance is conducted . Furthermore we want to outline and discuss the difference and similarities of various techniques as far as possible and investigate the already implemented packages in R. For this purpose we will analyze the R function *varImp()*, which already implemented a variable importance measure for different classes of regression and classification techniques. After that we additionally want to conduct an empirical study, where we investigate the importance of predictors in a credit scoring data w.r.t. a default indicator.

Chapter 1

Relative importance for linear regression

We will present in this chapter a short summary of metrics for measuring relative importance of single regressors in a multidimensional linear setting. All these methods are implemented in the R package **relaimpo**, which is documented in Grömping (2006). This chapter is mainly along the lines of Grömping (2006).

As stated in Grömping (2006) relative importance refers to the quantification of an individual regressor's contribution to a multiple regression model. Furthermore one often distinguishes between the following three types of importance Achen (1982):

- **dispersion importance:** importance relating to the amount of explained variance.
- **level importance:** importance with respect to the mean of the response.
- **theoretical importance:** change of the response variable for a given change in the explanatory variable.

The focus in this section will be entirely on dispersion importance. Another definition of relative importance, in the context of dispersion importance, was given from *Johnson and Lebreton* in Johnson and LeBreton (2004) as follows: *relative importance is the contribution each predictor makes to the coefficient of determination, considering both its* ***direct effect*** *(i.e. correlation with the response variable) and its* ***indirect or total effect*** *when combined with other explanatory variables.*

In the sequel we will list all importance metrics that are available in the package **relaimpo**.

1.1 The linear model and relative importance metrics

A simple linear multiple regression model can be formulated as

$$Y = X\beta + \epsilon, \quad Y \in \mathbb{R}^n,\, \beta \in \mathbb{R}^p,\, X \in \mathbb{R}^{n \times p}, \tag{1.1}$$

which reads component wise for $i \in \{1, \ldots, n\}$ as

$$y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i,$$

where y_i is the i-th observation of the response variable Y, β_i denotes the i-th regression coefficient, x_{ik} is the i-th observation of the k-th explanatory variable/regressor $X_k := (X)_{\cdot,k}$ and ϵ_i is defined as the i-th residual or unexplained part. Note that throughout this section, the first column of the design matrix is assumed to be constant. The key feature for a linear model is, as the name already suggests, that we assume a linear relationship between the response and the explanatory variables i.e. $Y = f(X) + \epsilon$, where $f : \mathbb{R}^{n\times p} \to \mathbb{R}^n$ is a linear mapping. The coefficients β are usually estimated by minimizing the sum of squared residuals (RSS) which is defined as

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{for } \hat{y}_i := \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip},$$

where we denoted the estimated coefficients and the fitted values by $(\hat{\beta}_i)_{i\in\{1,\ldots,p\}}$ respectively $(\hat{y}_i)_{i\in\{1,\ldots,n\}}$. Under the usual full rank assumption for the matrix X one has the following famous formula for the estimated coefficients:

$$\hat{\beta} = (X'X)^{-1}XY.$$

Some of the subsequent metrics for individual relative importance are based on the coefficient of determination.

Definition 1.1 (Coefficient of determination) *The coefficient of determination for the linear model defined in* (1.1) *is defined as*

$$R^2 := 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \in [0,1], \tag{1.2}$$

where we will use the following abbreviations: $TSS := \sum_{i=1}^{n}(y_i - \overline{y})^2$, $ESS := \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$. [1]

The second equality in definition 1.1 follows from the fact that

$$\underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{RSS} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}_{ESS} = \underbrace{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}_{TSS}.$$

From (1.2) one can see that the coefficient of determination measures the proportion of variance in the response variable Y, that is explained by the estimated model. It provides a measure of how well observed outcomes are captured by the model, based on the proportion of total variation of Y explained with the model. A value of 1 for R^2 indicates that we can perfectly explain the observed data with our model.

It can be shown, that in the case of a linear model the R^2 is equal to the square of the coefficient of multiple correlation.

Definition 1.2 (Coefficient of multiple correlation) *The coefficient of multiple correlation with respect to the model defined in* (1.1) *is defined as*

$$\mathcal{R} := (c' \cdot R_{XX}^{-1} \cdot c)^{1/2}, \quad c := \left(r_{X_1,Y}, \ldots, r_{X_p,Y}\right), \tag{1.3}$$

where $r_{X_i,Y}$ is the empirical pearson correlation coefficient between the i-th explanatory variable

[1] TSS stands for **Total Sum of Squares** whereas ESS stands for **Explained or Model Sum of Squares**.

and the response variable Y and

$$R_{XX} := \begin{pmatrix} r_{X_1,X_1} & \dots & r_{X_1,X_p} \\ \vdots & \ddots & \vdots \\ r_{X_1,X_p} & \dots & r_{X_p,X_p} \end{pmatrix}$$

is the correlation matrix of the explanatory variables X.

One can now show that the coefficient of determination equals the square of the coefficient of multiple correlation i.e. $R^2 = \mathcal{R}^2$ (see Appendix B.3). Thus in the setting of uncorrelated explanatory variables i.e. $R_{XX} = I_p$, one can conclude that the coefficient of determination is just the sum of the squared marginal correlation coefficients i.e. $R^2 = \sum_{i=1}^{p} r^2_{X_i,Y}$. Since in the univariate linear setting (including an intercept) the squared pearson correlation coefficient equals the coefficient of determination, we see that each regressor's contribution to the total R^2 in an orthogonal setting is just the R^2 from the univariate regression, and all univariate R^2-values add up to the total R^2. Thus one can perfectly measure the relative importance of a single regressor by means of its univariate coefficient of determination. Of course this breaks down if the regressors are correlated, which is often the case. Nevertheless one could consider the following metrics for measuring relative importance in a non orthogonal setting:

1.1.1 Simple relative importance metrics

(i) The metric *first*:
In this case one compares the univariate R^2-values of each regressor i.e. one measures how well can this individual regressor explain the model. This is motivated by the above discussed fact, that if we consider orthogonal regressors one can decompose the total R^2 into the sum of the individual R^2-values. In a general situation multicollinearity is present and one does not obtain a decomposition of the models R^2 by using this technique. Also this approach does not comply with the definition from Johnson and LeBreton (2004), which was stated at the beginning of Chapter 1, since it only captures direct effects.

(ii) The metric *last*:
A similar way is to compare what each regressor is able to explain in presence of all the other regressors. The metric *last* measures the increase in the total R^2 when including this regressor as the last one. If multicollinearity is present then this contributions again do not add up the the total R^2 of the model. Since one does not consider direct effects this metric does again not comply with the definition of relative dispersion importance given in Johnson and LeBreton (2004).

(iii) The metric *betasq*:
This approach measures relative importance by comparing the squared standardized estimated regression coefficients. They are calculated as follows:

$$\hat{\beta}^2_{k,standardized} := \left(\hat{\beta}_k \cdot \frac{s_{X_k}}{s_Y} \right)^2,$$

where s_{X_k} and s_Y denotes the empirical standard deviation of the variables X_k and Y. When comparing coefficients within models for the same response variable Y the denominator in the scaling factor is irrelevant. Again this metric does not provide a natural decomposition of the total R^2 (except the squared values in the case of orthogonal explanatory variables) and considers only indirect effects.

(iv) The metric *pratt*:
The *pratt* measure for relative importance is defined as the product of the previously defined standardized estimated coefficients and the marginal correlation coefficient i.e.

$$p_k := \hat{\beta}_{k,standardized} \cdot r_{X_k,Y}.$$

It can be shown that this definition yields an additive decomposition of the total R^2 i.e. $R^2 = \sum_{i=1}^{p} p_i$ (see Appendix B.3). Furthermore, it connects both direct (marginal) and indirect (conditional) effects of each regressor. Nevertheless a major disadvantage is, that this metric can yield negative values in which case it is not interpretable and one should not draw any conclusion from it.

1.1.2 Computer intensive relative importance metrics

The following two metrics require more computational effort compared to the simple metrics discussed above. Both of them yield a nonnegative decomposition of the total R^2. When decomposing the R^2 for regression models sequentially in the presence of correlated regressors, it turns out that each order of regressors yields a different decomposition of the model sum of squares (ESS). Division of the sequential ESS by TSS yields sequential R^2 contributions. The approach of the following two variable importance measures is based on sequential R^2 contributions, but takes care of the dependence with respect to the ordering by taking an unweighted and weighted average over all possible orders.

First we define the following notions:

The R^2 corresponding to a model using only $S \subset \{X_1, \ldots, X_p\}$ regressors is defined by $R^2(S)$.
The additional R^2 by adding regressors of the set $M \subset \{X_1, \ldots, X_n\}$ to an existing model with regressors $S \subset \{X_1, \ldots, X_p\}$ is for $S \cap M = \emptyset$ defined as:

$$R^2(M|S) := R^2(M \cup S) - R^2(S).$$

Furthermore we define by $S_k(r)$ the set of regressors entered into the model before the regressor X_k corresponding to the order $r := (r_1, \ldots, r_p)$.

(i) The metric *lmg*:
This metric simply corresponds to the empirical mean of all the sequential R^2-values. The relative importance measure of the k-th explanatory variable is therefore given as:

$$Lmg(X_k) := \frac{1}{p!} \sum_{r \in \mathcal{P}} R^2(\{X_k\}|S_k(r)) = \frac{1}{p!} \sum_{S \subset \{X_1,\ldots,X_p\} \backslash \{X_k\}} |S|! \cdot (p-1-|S|)! \cdot R^2(\{X_k\}|S),$$

where $\mathcal{P}$ denotes the set of all permutations of $\{r_1, \ldots, r_p\}$. From the definition one can see that this variable importance concept uses both direct effects (orders where X_k enters first in the model) and effects adjusted to other regressors (X_k enters last). A disadvantage of this concept is that a contribution of a regressor with estimated coefficient $\hat{\beta}_k$ equal to zero can be positive if this regressor is correlated with others (see Feldman (2005)). However, it was argued in Grömping (2007) that a zero coefficient does by no means indicate an unimportant variable in the case of correlated regressors from a causal/theoretical point of view. Nevertheless this "disadvantage" leads to the definition of the following metric.

(ii) The metric *pmvd*:
Per construction this metric guarantees that a regressor with estimated coefficient equal to zero does yield zero contribution in this relative importance measure. The formula is given as

$$Pmvd(X_k) := \sum_{r \in \mathcal{P}} w(r) R^2(\{X_k\}|S_k(r)),$$

where $w(r) := \frac{\prod_{i=1}^{p-1} \left(R^2(\{X_{r_{i+1}}, \ldots, X_{r_p}\}|\{X_{r_1}, \ldots, X_{r_i}\})\right)^{-1}}{\sum_{r \in \mathcal{P}} \prod_{i=1}^{p-1} \left(R^2(\{X_{r_{i+1}}, \ldots, X_{r_p}\}|\{X_{r_1}, \ldots, X_{r_i}\})\right)^{-1}}$ are weights derived from a set of axioms.

It was shown in Feldman (2005) that the weights are only positive for orders with all 0 coefficients regressors last, which leads to a share of 0 for these regressors.

We will discuss two more variable importance measures in the case of a linear regression setting. The *permutation importance* will be discussed in detail in chapter 3. In section 3.2 we will test this method on linear as well as logistic models. The *relative weights method* is introduced in 4.2.1.

1.1.3 Simulation study: relative importance measures

All of the six presented relative importance measures for linear regression are implemented in the R package **relaimpo**. Using this package they can be calculated for an object of class *lm* as follows:

```
calc.relimp(lm,type=c("lmg","pmvd","first","last","betasq","pratt")).
```

We will test now the different metrics on a simulated data set. The data set was generated according to the following linear model:

$$(X_1, \ldots, X_{12})' \sim \mathcal{N}(\mu, \Sigma) \qquad (\Sigma)_{i,j} := \begin{cases} 1 & , j = i \\ 0.9 & , i \neq j \leq 4 \\ -0.9 & , i \neq j \geq 9 \\ 0 & , \text{else} \end{cases}$$

$$\mu := (0, 0, 5, 10, 0, 0, 5, 10, 0, 0, 0, 50)'$$

$$(\beta_1, \ldots, \beta_{12})' = (5, 5, 2, 0, -5, -5, -2, 0, 0, 0, 10, 0)'$$

1. Linear model:

$$y_i = \beta_1 x_{1,i} + \ldots + \beta_{12} x_{12,i} + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, 100), \, i \leq N$$

Of the twelve features seven were influential. We constructed one (4 × 4) - block of positively correlated predictors, one (4 × 4) - block of independent predictors, which both have regression coefficients with the same absolute value and a (2×2) - block of negatively correlated predictors, where the variable X_{11} has in absolute value the largest coefficient. We chose the number of simulated data points to be $N = 1000$. For fitting a linear model we used the **lm()** function, which is contained in the R package **stats**.

Results of the raw values and ranks are presented in table 1.1 and table 1.2.

	first	last	betasq	pratt	lmg	pmvd
X_1	0.31677	0.00743	0.05419	0.13102	0.08978	0.16073
X_2	0.31596	0.00706	0.04995	0.12563	0.08854	0.14157
X_3	0.30625	0.00139	0.01007	0.05554	0.08110	0.03249
X_4	0.31331	0.00042	0.00311	0.03123	0.08156	0.01055
X_5	0.06439	0.06529	0.06561	0.06499	0.06483	0.06549
X_6	0.05450	0.05950	0.05994	0.05715	0.05802	0.05954
X_7	0.02846	0.01591	0.01609	0.02139	0.01941	0.01732
X_8	0.00087	0.00005	0.00006	0.00022	0.00024	0.00006
X_9	0.00002	0.00002	0.00002	-0.00002	0.00008	0.00002
X_10	0.00017	0.00003	0.00003	0.00007	0.00007	0.00003
X_11	0.24720	0.04142	0.22092	0.23369	0.14829	0.25296
X_12	0.23013	0.00043	0.00230	0.02303	0.11204	0.00319

Table 1.1: Relative importance metrics in a linear setting.

	first	last	betasq	pratt	lmg	pmvd
X_1	1	5	4	2	3	2
X_2	2	6	5	3	4	3
X_3	4	7	7	6	6	6
X_4	3	9	8	7	5	8
X_5	7	1	2	4	7	4
X_6	8	2	3	5	8	5
X_7	9	4	6	9	9	7
X_8	10	10	10	10	10	10
X_9	12	12	12	12	11	12
X_10	11	11	11	11	12	11
X_11	5	3	1	1	1	1
X_12	6	8	9	8	2	9

Table 1.2: Ranks of relative importance metrics in a linear setting.

We comment now on the obtained results for the different metrics.

1. **first**: This simple metric fails to identify some of the most influential predictors like X_5, X_6 and X_{11}. It also shows a strong preference for correlated predictors with small or even zero influence like X_3, X_4, X_{12}. This is due to the fact that is only displays the direct effect as discussed above.

2. **last**: This metric shows only the effect of a variable on the response when combined with all the other variables and thus no direct effects. This is the reason why the correlated influential predictors X_1 and X_2 are not ranked appropriately. Nevertheless it was able to figure out the relevance of the variable X_{11}.

3. **betasq**: The squared standardized coefficient is able to detect the most influential variables even though multicollinearity is present.

4. **pratt**: This natural decomposition of the R^2 yields basically the same result as the *betasq* metric. Only the rank of the correlated influential predictors X_1 and X_2 and the uncorrelated ones X_5 and X_6 are interchanged.

5. **lmg**: Shows a rather strong preference for correlated predictors with little or no influence like X_3, X_4 and X_{12}. It is also not ensured that variables with zero coefficients $X_4, X_8, X_9, X_{10}, X_{12}$ do have an importance score of zero.

6. **pmvd**: This metric ensures that non influential variables i.e. variables with zero coefficients have an importance score of (theoretically) zero. Furthermore it was simultaneously able to detect the most important variables and to yield a positive decomposition of the R^2.

Chapter 2

Variable importance in a general regression setting

This chapter gives a short introduction to the concepts developed in Van der Laan (2006) and summarizes the main results of the first part of said work.

In many current practical problems the number of explanatory variables can be very large. Assuming a fully parametrized model, such as a multiple linear regression, and minimizing the empirical mean of a loss function (e.g. RSS) is likely to yield poor estimators (overfitting) and therefore many applications demand a nonparametric regression model. The approach in prediction is often that one learns the optimal predictor from data and derives, for each of the input variables, a variable importance by considering the obtained fit. In Van der Laan (2006) the authors propose estimators of variable importance which are directly targeted at this parameters. Therefore this approach results in a separate estimation procedure for each variable of interest. We first will formulate the problem of estimating variable importance.

We are given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and n i.i.d observations of a random vector $O = (W^*, Y) \sim P_0$, where P_0 denotes the true underlying data generating distribution, $Y : \Omega \to \mathbb{R}$ is the outcome and $W^* : \Omega \to \mathbb{R}^n$ denotes the random vector of input variables which can be used to predict the outcome. Furthermore we define by $A := A(W^*)$ a function of the input variables for which we want to estimate the variable effect of $A = a$ relative to $A = 0$ e.g. A could be a simple projection on the k-th coordinate of W^* i.e. $A(W^*(\omega)) = W_k^*(\omega)$. Furthermore we define W such that $(W, A) \overset{(d)}{=} W^*$.

2.1 Variable importance measures

We will list now three related concepts of variable importance as presented in Van der Laan (2006). In order to obtain a well defined parameter of variable importance we will assume that $P[A = a|W] > 0$ and $P[A = 0|W] > 0,\ P_W\ a.s.$.

(i) The first proposed real valued parameter of variable importance of the predictor $\mathbb{E}_{P_0}[Y|A, W]$

on a model for P_0 is defined as the image of the following function

$$P \to \Psi(P)(a) := \mathbb{E}_P[\,(\mathbb{E}_P[Y|A=a,W] - \mathbb{E}_P[Y|A=0,W])\,].$$

The parameter $\Psi(P)(a)$ and the whole curve $\Psi(P) := \{\Psi(P)(a) : a\}$ are called the *a-specific marginal variable importance* and the *marginal variable importance of the variable A*, respectively.

(ii) The *a-specific W adjusted variable importance* is defined as the image of

$$P \to \Psi(P)(a,w) := \mathbb{E}_P[Y|A=a,W=w] - \mathbb{E}_P[Y|A=0,W=w],$$

where $w \in \{w : P(A=a|W=w)\cdot P(A=0|W=w) > 0\}$. From this definition one can see that $\Psi(P)(a) = \mathbb{E}_P[\Psi(P)(a,W)]$.

(iii) Both the above presented measures are special cases of the *a-specific V adjusted variable importance*, which is defined as

$$P \to \Psi(P)(a,v) := \mathbb{E}_P[\,(\mathbb{E}_P[Y|A=a,W] - \mathbb{E}_P[Y|A=0,W])\,|V=v\,].$$

This parameter is only well defined if for all w in the support of the conditional distribution $P_{W|V=v}$ it holds that $P(A=a|W=w)\cdot P(A=0|W=w) > 0$. Moreover if $V=W$ then the *a-specific V adjusted variable importance* is equal to the *a-specific W adjusted variable importance*. Furthermore if W is independent of V then the *a-specific V adjusted variable importance* equals the *a-specific marginal variable importance.*

In the context of a linear regression this model free variable importance parameters can be illustrated as follows:

- If $\mathbb{E}_P[Y|A,W] = \beta_0 + \beta_1 A + \beta_2 W$ then

$$\Psi(P)(a) = \Psi(P)(a,W) = \beta_1 a$$

- If $\mathbb{E}_P[Y|A,W] = \beta_0 + \beta_1 A + \beta_2 AW + \beta_3 W$ then

$$\begin{aligned}\Psi(P)(a,W) &= \beta_1 a + \beta_2 aW\\ \Psi(P)(a) &= \mathbb{E}_P[\Psi(P)(a,W)] = (\beta_1 + \beta_2\mathbb{E}_P[W])a\end{aligned}$$

- If $\mathbb{E}_P[Y|A,W] = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_1 A_2 + \beta_4 W$ then

$$\Psi(P)(a_1,a_2) = \Psi(P)(a_1,a_2,W) = \beta_1 a_1 + \beta_2 a_2 + \beta_3 a_1 a_2$$

This means if $\mathbb{E}_P[Y|A,W]$ is a linear combination of multi-way interactions, then the *a-specific W adjusted variable importance* is the linear combination of all interactions including a, obtained by deleting all interactions which do not depend on a.

In the case of a logit or probit model for the conditional mean e.g

$$\mathbb{E}_P[Y|A,W] = f(\beta_0 + \beta_1 A + \beta_2 W),$$

with f denoting the corresponding response function i.e.

- Logit: $f(x) := \dfrac{1}{1+\exp(-x)}$
- Probit: $f(x) := \Phi(x)$,

one is not able anymore to delete all terms which are independent of a as in the linear case. Nevertheless the variable importance measures can be calculated as

- Logit:

$$\Psi(P)(a) = e^{-\beta_0}\left(1-e^{-a\beta_1}\right)\mathbb{E}_P\left[\frac{e^{-\beta_2 W}}{\left(1+e^{-(\beta_0+\beta_1 a+\beta_2 W)}\right)\left(1+e^{-(\beta_0+\beta_2 W)}\right)}\right]$$

$$\Psi(P)(a,W) = \frac{e^{-\beta_0}\left(1-e^{-a\beta_1}\right)e^{-\beta_2 W}}{\left(1+e^{-(\beta_0+\beta_1 a+\beta_2 W)}\right)\left(1+e^{-(\beta_0+\beta_2 W)}\right)}$$

- Probit:

$$\Psi(P)(a) = \mathbb{E}_P\left[\Phi(\beta_0+\beta_1 a+\beta_2 W) - \Phi(\beta_0+\beta_2 W)\right]$$
$$\Psi(P)(a,W) = \Phi(\beta_0+\beta_1 a+\beta_2 W) - \Phi(\beta_0+\beta_2 W)$$

2.1.1 Interpretation of variable importance parameters

These measures of variable importance are motivated from their analogues in causal inference. For a binary treatment variable A, where we define the target group by $A=0$ the parameter of interest is given as:

$$\Psi(P_0) = \mathbb{E}_{P_0}[\,(\mathbb{E}_{P_0}[Y|A=0,W] - \mathbb{E}_{P_0}[Y|W])\,] = \mathbb{E}_{P_0}[\,\mathbb{E}_{P_0}[Y|A=0,W]\,] - \mathbb{E}_{P_0}[Y],$$

and thus compares the overall mean of the outcome to the mean of the target group, averaged over W. If we assume that the hypothetical full data is given by $X=(Y_0,Y,W)\sim P_X$, where Y_0 is the counterfactual outcome for $A=0$, i.e. the outcome as it would have happened under universal application of treatment A=0. Then the causal analogue to $\Psi(P_0)$ is:

$$\Psi(P_X) = \mathbb{E}_{P_X}[\,(\mathbb{E}_{P_X}[Y_0|W] - \mathbb{E}_{P_X}[Y|W])\,] = \mathbb{E}_{P_X}[Y_0 - Y].$$

Under the following assumptions (see Ritter et al. (2014)) it holds that $\Psi(P_0) = \Psi(P_X)$:

(i) The observed data structure (W,A,Y) is chronologically ordered i.e. W precedes A and A precedes Y.

(ii) The observed data structure equals a missing data structure $(W,A,Y) = (W,A,Y_0)$.

(iii) A is conditionally independent of Y_0 given W.

This can also be extended to a non binary treatment variable A and the event of interest given by $A = a$. Under this assumptions one can express the presented variable importance measures from section 2.1 in the following way:

$$\begin{aligned}\Psi(P)(a) &= \mathbb{E}_P[Y_a - Y_0],\\ \Psi(P)(a, W) &= \mathbb{E}_P[Y_a - Y_0|W],\\ \Psi(P)(a, V) &= \mathbb{E}_P[Y_a - Y_0|V].\end{aligned}$$

Thus the presented measures of variable importance can be interpreted as marginal causal effects. In the final section we will present the methodology for estimation of the *a-specific W adjusted marginal variable importance* for a discrete random variable A in a nonparametric model, which is presented in (Van der Laan, 2006, Section 2.1).

2.2 Discrete treatment variable, nonparametric model, a-specific variable importance

The first theorem establishes path-wise differentiability and a closed form expression of the efficient influence curve/canonical gradient of the marginal variable importance parameter.

Theorem 2.1 (Theorem 1, Ritter et al. (2014)) *Suppose that $O = (A, W, Y) \sim P_0$, where A is a discrete random variable with finite support. Assume that $P[A = a|W] > 0$ and $P[A = 0|W] > 0$, $P_0 - a.s.$. Consider a nonparametric model for P_0, and let $\Psi(P)(a)$ be the parameter of interest. Let $\psi_0 := \Psi(P_0)$. The efficient influence curve at P_0 of this parameter is given by:*

$$\begin{aligned}IC^*(O|P_0) =&(\theta_0(a, W) - \theta_0(0, W)) - \psi_0(a)\\ &+ \left[\frac{\mathbb{1}_{\{A=a\}}}{\Pi_0(a|W)}(Y - \theta_0(a, W)) - \frac{\mathbb{1}_{\{A=0\}}}{\Pi_0(0|W)}(Y - \theta_0(0, W))\right],\end{aligned}$$

where $\theta_0(a, W) := \mathbb{E}_{P_0}[Y|A = a, W]$ and $\Pi_0(a|W) := P_0(A = a|W)$ are the unknown nuisance parameters being the regression and the conditional distribution.

For notational convenience we will define $\psi_0 := \psi_0(a)$. The estimating function for ψ_0 based on $IC^*(O|P_0)$ is given by:

$$\begin{aligned}(O, \psi, \theta, \Pi) \to D(O|\psi, \theta, \Pi) :=&(\theta(a, W) - \theta(0, W)) - \psi(a)\\ &+ \left[\frac{\mathbb{1}_{\{A=a\}}}{\Pi(a|W)}(Y - \theta(a, W)) - \frac{\mathbb{1}_{\{A=0\}}}{\Pi(0|W)}(Y - \theta(0, W))\right],\end{aligned}$$

where $\theta(a, W) := \mathbb{E}_P[Y|A = a, W]$ and $\Pi(a|W) := P(A = a|W)$.

The following lemma yields an estimating equation for ψ_0.

Lemma 2.2 (Result 1, Van der Laan (2006)) *Assume $P(A = a|W)P(A = a|W) > 0$ $P_0 - a.s.$. Then it holds that*

$$\mathbb{E}_{P_0}[D(O|\psi_0, \theta, \Pi)] = 0, \text{ if either } \theta = \theta_0, \text{ or } \Pi = \Pi_0.$$

Proof.

$$\begin{aligned}\mathbb{E}_{P_0}[D(O|\psi_0,\theta,\Pi)] =&\mathbb{E}_{P_0}[(\theta(a,W)-\theta(0,W))]-\mathbb{E}_{P_0}[\psi_0(a)]\\ &+\int_\Omega\left[\frac{\Pi_0(a|W)}{\Pi(a|W)}(\theta_0(a,W)-\theta(a,W))-\frac{\Pi_0(0|W)}{\Pi(0|W)}(\theta_0(0,W)-\theta(0,W))\right]dP_0.\end{aligned}$$

Consider now the case where $\theta=\theta_0$ then the integral vanishes and also the first term by definition of $\psi_0(a)$. In the case of $\Pi=\Pi_0$ the integral is the negative of the first 2 terms. □

A double robust (i.e. consistent if either θ_0 or Π_0 is estimated consistently) locally efficient estimator can be constructed by solving the above defined estimating equation i.e. given estimators Π_n and θ_n of Π_0 and θ_0, one can estimate ψ_0 with:

$$\psi_n := P_n D(O|\theta_n,\Pi_n),$$

where we use the notation $Pf := \int f dP$ for the expectation operator and

$$\begin{aligned}D(O,\theta_n,\Pi_n) :=&(\theta_n(a,W)-\theta_n(0,W))\\ &+\left[\frac{\mathbb{1}_{\{A=a\}}}{\Pi_n(a|W)}(Y-\theta_n(a,W))-\frac{\mathbb{1}_{\{A=0\}}}{\Pi_n(0|W)}(Y-\theta_n(0,W))\right].\end{aligned}$$

Thus given n observations the estimator ψ_n can be written as:

$$\begin{aligned}\psi_n =&\frac{1}{n}\sum_{i=1}^n Y_i\left(\frac{\mathbb{1}_{\{A_i=a\}}}{\Pi_n(a|W_i)}-\frac{\mathbb{1}_{\{A_i=0\}}}{\Pi_n(0|W_i)}\right)-\sum_{i=1}^n\theta_n(a,W_i)\left(\frac{\mathbb{1}_{\{A_i=a\}}}{\Pi_n(a|W_i)}-\frac{\mathbb{1}_{\{A_i=0\}}}{\Pi_n(0|W_i)}\right)\\ &+\frac{1}{n}\sum_{i=1}^n\theta_n(a,W_i)-\theta_n(0,W_i).\end{aligned}\tag{2.1}$$

If one assumes a correctly specified model for Π_0, then we can set $\theta_n=0$, which results in:

$$\psi_n=\frac{1}{n}\left[\sum_{i=1}^n Y_i\left(\frac{\mathbb{1}_{\{A_i=a\}}}{\Pi_n(a|W_i)}-\frac{\mathbb{1}_{\{A_i=0\}}}{\Pi_n(0|W_i)}\right)\right].\tag{2.2}$$

In the case of a binary treatment or exposure variable $A\in\{0,1\}$ the estimators from formula (2.1) and formula (2.2) are implemented in the R package **multiPIM** (Ritter et al., 2014, Section 2.2). Nevertheless, in the remaining part of this theses we will no longer focus on these variable importance measures in a general regression context and rather move to importance measures in the context of random forests, which can deal with high dimensionality and are easily applicable.

Chapter 3

Variable importance in the context of random forests

In this chapter we introduce several ways to assess variable importance, when dealing with random forests. In applications random forests are a widely used tool for non-parametric regression or classification problems. They can be applied to "large-p small-N" problems, can deal with highly correlated explanatory variables as well as complex interactions between them. Furthermore they provide different variable importance measures that can be used to identify the most important features in a given setting. After a short overview on random forests, we will present the most commonly used variable importance measures. Finally we will compare them on a simulated data set. The first part of this chapter is based on Breiman (2001) as well as (Friedman et al., 2001, section 9.2 and chapter 15).

We will use the following notation throughout this chapter:

Let $X := (X_1, ..., X_K) \in \mathbb{R}^{N \times K}$ denote the input matrix , where N and K are the number of observations and features (explanatory variables, regressors, independent variables or predictors) respectively. Furthermore we will denote by X_k for $k \leq K$ the k-th column of X which represents a single feature and by x_n for $n \leq N$ the n-th row of X representing a single observation. The target or response variable will be denoted by $Y \in \mathbb{R}^N$ whereas y_n for $n \leq N$ denotes a single observation. We will identify by $T := (X, Y) \in \mathbb{R}^{N \times (K+1)}$ the training sample, upon we will build the model. For the sake of readability we will hereby use the notation of capital letters both for real valued column vectors representing realizations of a feature as well as for real valued random variables and refer to the meaning of it from the context.

A random forest is an ensemble of multiple decision trees. There are various methods available which randomly select first the training sets T_b, by selecting a subset of the rows of T, for the b-th individual decision tree and secondly at each node the used features $\{X_{i_1}, \ldots, X_{i_m}\}$ with $m \leq K$ and $i_m \in \{1, \ldots, K\}$ upon the split is made. This method is called *feature bagging*. In order to choose the feature $X^* \in \{X_{i_1}, \ldots, X_{i_m}\}$ that "best" binary splits at a certain node one solves an optimization problem with respect to a certain metric, which often measures the homogeneity of the target variable Y in the resulting subset. We will focus in this introduction on the famous CART algorithm introduced by Breiman et al. (1984) which outlines the main idea in recursive binary partitioning. The most popular metrics used for measuring the "best" split are:

- *Regression trees:*
  - *Minimum sum of squared errors:*
    At each node in a single tree the feature X^* and splitting point s^* is selected based on the N_{node} observations in this node as the solution to the following minimization problem:

    $$\min_{j,s}\left\{\sum_{y_i\in R_1(j,s)}(y_i-\overline{y_1})^2+\sum_{y_i\in R_2(j,s)}(y_i-\overline{y_2})^2\right\},$$

    where $R_1(j,s):=\{(X,Y)|X_j\leq s\}\subset N_{node}$ and $R_2(j,s):=\{(X,Y)|X_j>s\}\subset N_{node}$ are the two half spaces (i.e. rows of (X,Y)) representing a binary split with respect to the j-th feature and $\overline{y_{1,2}}:=\frac{1}{|R_{1,2}(j,s)|}\sum_{y_i\in R_{1,2}(j,s)}y_i$ denotes the mean within those subsamples.

- *Classification trees:*
  If the target variable Y is a factor with L levels then we define for $l\leq L$

  $$p_{1l}:=\frac{1}{|R_1(j,s)|}\sum_{y_i\in R_1(j,s)}\mathbb{1}_{\{y_i=l\}}$$

  as the proportion of level l in $R_1(j,s)$ resulting due to the resulting binary split. Analogously one can define p_{2l}. In each resulting node the observations are classified according to the majority vote i.e. in the left child node all the observations are classified according to the level $l^*:=\arg\max_l p_{1l}$. Instead of minimizing the mean squared error as above one seeks to minimize one of the following homogeneity measures:

  - *Gini impurity:* measures how often a randomly chosen element would be incorrectly labeled if it was randomly labeled according to the frequency of the levels in the subset. Formally defined as

    $$GI_1(j,s):=\sum_{l=1}^{L}p_{1l}\cdot(1-p_{1l}).$$

  - *Cross-Entropy:*

    $$E_1(j,s):=-\sum_{l=1}^{L}p_{1l}\cdot\log(p_{1l})$$

  - *Misclassification Error:*

    $$MCE_1(j,s):=\frac{1}{|R_1(j,s)|}\sum_{y_i\in R_1(j,s)}\mathbb{1}_{\{y_i\neq l^*\}}=1-p_{1l^*}$$

  In the special case of a binary target variable the measures result in

  $$\begin{aligned}&GI_1(j,s)=2p\cdot(1-p)\\&E_1(j,s)=-(p\log(p)+(1-p)l\log(1-p))\\&MCE_1(j,s)=1-\max(p,1-p),\end{aligned}$$

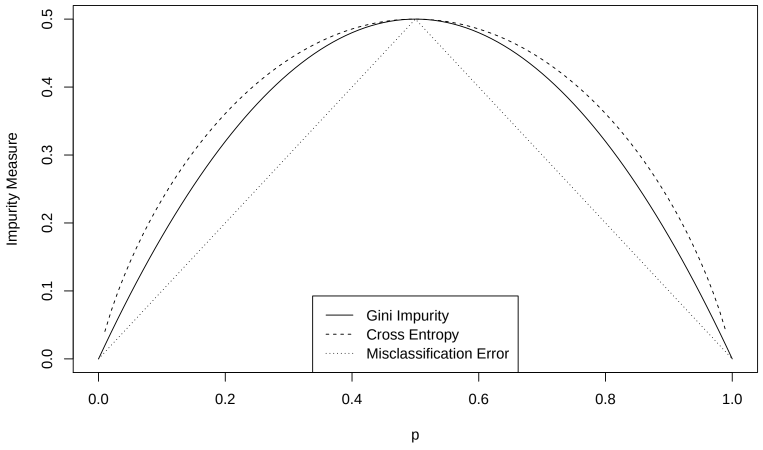


Figure 3.1: Node impurity measures for a binary target variable as a function of the proportion of the second class p. $E_1(p)$ was scaled to go through the point $(0.5, 0.5)$.

where p denotes the proportion of the second class. They are presented in figure 3.1. To decide upon a splitting feature and point the minimization is done by weighting the resulting measures in the two child nodes and adding them up.

e.g. for the Gini impurity at node k:

$$G_k := \min_{j,s} \left\{ GI_1(j,s) \cdot \frac{|R_1(j,s)|}{|R_1(j,s)| + |R_2(j,s)|} + GI_2(j,s) \cdot \frac{|R_2(j,s)|}{|R_1(j,s)| + |R_2(j,s)|} \right\}, \tag{3.1}$$

where the right hand side depends on k trough the observations considered.

One example for selecting training sets is *bagging* (bootstrap aggregation), where an individual training set T_k for the k-th decision tree is generated by a random selection of rows with replacement from the original training set T. By taking the majority vote, in the case of classification, or the mean of each terminal leaf one obtains predictions for each individual grown tree. The final prediction of the random forest is then again obtained by majority vote or averaging over all grown trees. By reaching a certain criterion such as the minimum number of observations in one node the growth of a tree is stopped.

The common element of all these procedures is that for the b-th tree a random vector Θ_b is generated. In the case of (feature) bagging with replacement Θ_b would be a vector of N i.i.d uniformly distributed random variables $U \sim \{1, \ldots, N\}$. Furthermore the sequence of random vectors $(\Theta_1, \Theta_2, \ldots, \Theta_B)$ is assumed to independent and identically distributed. An individual tree is grown using the training set T, Θ_b and the input X and will be denoted by $h(T, X, \Theta_b)$. This yields to a more formal definition of a random forest.

Definition 3.1 (Random Forest) *A random forest is a collection of regression or classification trees $\{h(T, X, \Theta_b), b \in B \subseteq \mathbb{N}\}$, where $\{\Theta_b\}_{b \in B \subseteq \mathbb{N}}$ is a sequence of independent and identically distributed random vectors. For an input X the output is obtained*

- *Classification: each tree casts a unit vote or for the most popular class for the input X. Upon* *this votes the classification is determined by the majority.*

- *Regression: each tree outputs the mean of the terminal leaf, where the considered input X is assigned to. Taking again the mean over all trees yields the final output.*

In the remaining part of the thesis we will for the sake of readability define a single tree by $h(T_b) := h(T, X, \Theta_b)$. Below a pseudo code for the implementation of a random forest is presented.

Algorithm 1: Pseudo code for implementing a random forest

```
1. for $b = 1$ to $B$ do
        (a) Draw a bootstrap sample $\Theta_b$ from the total number of rows $N$ of the training sample $T$ and construct $T_b$ .
        (b) Fit a single decision tree $h(T_b)$ by recursively repeating the following steps for each node until a stopping criterion (e.g. minimum number of observations) is met:
              i. Select randomly $m \leq p$ features: $\{X_{i_1}, \dots, X_{i_m}\}$.
             ii. Determine the feature $X^* \in \{X_{i_1}, \dots, X_{i_m}\}$ and a splitting point $s^*$ that best splits the data according to some impurity measure.
            iii. Conduct a binary split into two child nodes.
        (c) Output the random forest $\{h(T_b) : b \leq B\}$.
   end
2. For an input $x_j$, $j \leq N$ predict the response as following:
        (a) Regression: $\hat{f}^B_{rf}(x_j) := \frac{1}{B} \sum_{b=1}^{B} h(T_b)(x_j)$ where $h(T_b)(x_j)$ is the prediction of a single tree based upon the mean in the terminal leaf where $x_j$ falls into.
        (b) Classification: $\hat{f}^B_{rf}(x_j) :=$ majority vote of $\{h(T_b)(x_j) : b \leq B\}$ where $h(T_b)(x_j)$ is the prediction of the single tree based on a majority vote in the terminal leaf where $x_j$ falls into.
```

3.1 Variable importance measures

We will now discuss several popular methods for measuring variable importance in the context of random forests. The random forests will be based on the original CART implementation as outlined above and on a newer approach, where splits are conducted based on a conditional inference framework instead of the e.g. *Gini impurity* used in the CART algorithm. Finally we will compare the different methods using simulated data.

3.1.1 Gini importance

The basic idea is to assess variable importance for a feature X_j by accumulating over each tree the improvement in the splitting criterion metric in each split, that is conducted by X_j. In the case for a regression tree the splitting criterion metric would be simply the squared error. For classification we could theoretically use any of the above discussed impurity measures, while the

most common approach is to use the Gini impurity measure which leads us to the definition of the *Gini importance.*
Let M_b be the number of nodes in the b-th tree of the random forest $\{h(T_b)\}_{b\in B\subseteq\mathbb{N}}$ (not including terminal nodes i.e. leafs). Then the *Gini importance* for the feature X_j is defined as

$$I_{gini}(j) := \sum_{b=1}^{B}\left\{\sum_{m=1}^{M_b}(GI_m^{parent} - G_m)\mathbb{1}_{\{\text{split is made upon } X_j\}}\right\}, \tag{3.2}$$

where GI_m^{parent} is defined as the Gini impurity in the node m i.e. parent node w.r.t to the split and G_m is defined in equation (3.1) as the weighted Gini impurity resulting from the split.

However it was shown in Strobl et al. (2007) that the Gini importance measure used in combination with the CART algorithm does not yield reliable results. The authors of Strobl et al. (2007) conducted several simulation studies where they showed that the Gini importance has a strong preference for continuous variables and variables with many categories. It sometimes completely fails to identify the relevant predictors. The reason for the bias induced by the Gini importance measure is due to preference of continuous variables or variables with many categories in a CART-like tree building process. Since the Gini importance measure is directly calculated as the improvement in the *Gini impurity* resulting from a split, it is strongly affected by this selection bias and does not yield reliable results, especially in a setting where the predictors vary in their scale or have a different number of categories. Thus we won't focus on this particular variable importance measure in the remaining part of this chapter. However it is accessible in the following R package:

The *Gini importance* measure is implemented by the R function **importance(**...**,type=2)** which is part of the package **randomForest** Liaw et al. (2015).

3.1.2 Permutation importance

Following Breiman (2001), we now focus on the "feature bagging" algorithm for growing a tree. This means that first a new training set T_b is drawn from the original training set T with replacement. Then a single tree is grown on the new training set using random feature selection at each node.

In the following we will need the definition of an out of bag (OOB) sample for a single tree. This is defined for the b-th tree as $T \setminus T_b$ i.e. the observations which were not used in the fitting of this single tree. After the whole forest has been trained the *permutation importance* of variable X_j is measured by comparing OOB prediction accuracy of a single tree i.e. classification rate (classification trees) or mean squared error (regression trees) before and after permuting the feature X_j. The idea behind that is if this feature was relevant for the prediction or had an influence on the target the accuracy should decrease. Finally averaging each decrease over all trees yields the *permutation importance.* This can be formalized in the case of a classification random forest as follows:

Let $O_b := T \setminus T_b$ be the out of bag sample for the b-th tree with $b \in \{1, \dots, B\}$. Then the *permutation importance* of the j-th feature $I_{permute}(j)$ is defined as:

$$I_{permute}(j) := \frac{1}{B}\sum_{b=1}^{B} \underbrace{\left\{ \frac{\sum_{i \in O_b} \mathbb{1}_{\{y_i = \hat{y}_i^b\}}}{|O_b|} - \frac{\sum_{i \in O_b} \mathbb{1}_{\{y_i = \hat{y}_{i,\pi_j}^b\}}}{|O_b|} \right\}}_{:= I_{permute}^b(j)}, \tag{3.3}$$

where $\hat{y}_i^b := h(T_b)(x_i)$ and $\hat{y}_{i,\pi_j}^b := h(T_b)(x_{i,\pi_j})$ is the predicted class of the b-th tree for the input x_i respectively for the permuted input $x_{i,\pi_j} := (x_{i,1}, \dots, x_{i,j-1}, x_{\pi_j(i),j}, x_{i,j+1}, \dots, x_{i,K})$.

This approach can be naturally extended to regression forests by substituting the classification rate $\sum_{i \in O_b} \mathbb{1}_{\{y_i = \hat{y}_i^b\}}$ in equation (3.3) by the mean squared error $\sum_{i \in O_b}(y_i - \hat{y}_i^b)^2$ and considering the increase in the MSE. A pseudo code for measuring the *permutation importance* is presented below.

Algorithm 2: Pseudo code for calculating permutation importance for a single feature X_j

```
1. Fit a random forest $\{h(T_b) : b \leq B\}$ on the training set $T$ using algorithm 1 presented above.
2. for $b = 1$ to $B$ do
       (a) Compute the OOB prediction accuracy of the $b$-th tree $h(T_b)$.
       (b) Permute randomly the observations of the feature $X_j$ in the OOB sample $O_b$ once.[1]
       (c) Recompute the OOB prediction accuracy of the $b$-th tree $h(T_b)$ using the permuted input.
       (d) Compute $I_{permute}^b(j)$.
   end
3. Compute the average decrease of prediction accuracy over all trees i.e. $I_{permute}(j)$.
```

In the following we will use the term *unconditional permutation importance* for the above discussed *permutation importance*. Again it was shown in Strobl et al. (2007) that the *unconditional permutation importance*, when using it in combination with the CART algorithm, does not yield good results. As presented in Strobl et al. (2007), using the CART algorithm in combination with the Gini impurity split criteria induces a bias towards continuous predictor variables or variables with many categories. This bias of course affects the permutation procedure. Variables that appear more often in trees and are situated closer to the root of each tree can affect the prediction accuracy of a larger set of observations when permuted.

It was also outlined in Strobl et al. (2007) that the sampling scheme for the training set T_k of the k-th tree does have a not negligible effect on the so far discussed variable importance measures. A training set T_k obtained via bootstrapping i.e. sampling N observations from T **with replacement** also induces a bias for continuous variables or variables with many categories. This bias is independent of the used algorithm and is also present when building an unbiased random forest

[1] Let S_n be the symmetric group of all permutations of $\{1, \dots, n\}$. A random permutation of S_n is a uniformly distributed random variable $\Pi : \Omega \mapsto S_n$ i.e. $\mathbb{P}[\Pi = \pi] = \frac{1}{n!}, \quad \forall \pi \in S_n$.

as in Hothorn et al. (2006), where the splitting criteria is based on a permutation test framework. Applying the method of Hothorn et al. (2006), predictors which attain statistical significance are candidates for the node split. Among those the split is made upon the predictor with the smallest p-value. This approach guarantees unbiased variable selection in the sense that continuous predictor variables or features with many levels are no longer favored when conducting a split.
All in all sampling for the training set T_k should be carried out independently of the algorithm used **without replacement**.

There are two different versions of the *unconditional permutation importance* in R available:

1. CART algorithm: biased
   - package: **randomForest** Liaw et al. (2015)
   - function: **importance(...,type=1)**
2. Conditional inference forests Hothorn et al. (2006): unbiased
   - packages: **party** Hothorn et al. (2017-12-12) and **partykit** Hothorn et al. (2017-12-13).
   - function: **varimp()**

In the next two sections we will discuss two more extension of the *unconditional permutation importance*, which can deal better with correlated predictor variables and missing data.
Thus up to now the most promising method is fitting random forests using the function **cforest()** of the packages **party** or **partykit** in combination with sampling with replacement (which is the default setting) and measure the importance via the function **varimp()**.

3.1.3 Conditional permutation importance

In a setting with highly correlated features the distinction between the marginal effect and the conditional effect of a single predictor variable on the target is crucial. Consider for example a group of several pupils between the age of 7 and 18 doing a basic knowledge test. The two predictor variables age A and the size of shoes S are used to predict the performance on the test Y. Since it is likely that the correlation of A and S is large, we will outline in the following why both variables will have a rather large importance when using the *unconditional permutation importance* discussed above. Nevertheless, when conditioning on the age A i.e. comparing only students with the same age it is then clear that the size of the shoes S is no longer associated with the performance Y. This is an example where a predictor may appear to be marginally influential but might actually be independent of the response variable when conditioned on another.
We will therefore discuss in this section a conditional variable importance measure based on the same idea of the *unconditional permutation importance*, which reflects the true impact of a predictor variable more reliable. It will be based on the partition of the feature space obtained by the fitted random forest. This section will be mainly along the lines of Strobl et al. (2008).

First we will outline why the *unconditional permutation importance* from section 3.1.2 favors correlated predictor variables. This is caused by the following two reasons:

1. Preference of correlated predictor variables in (early) splits when fitting the random forest.

   We will illustrate this effect by a simulation study similar as in (Strobl et al., 2008, section

2.1) by fitting a regression and a classification forests using the R function **cforest()** from the **partykit** package Hothorn et al. (2017-12-13).

Data sets were generated according to a linear and a binary response model as following:

$$(X_1, \ldots, X_{12})' \sim \mathcal{N}(0, \Sigma) \qquad (\Sigma)_{i,j} := \begin{cases} 1 & , j = i \\ 0.9 & , i \neq j \leq 4 \\ -0.9 & , i \neq j \geq 9 \\ 0 & , \text{else} \end{cases}$$

$$(\beta_1, \ldots, \beta_{12})' = (5, 5, 2, 0, -5, -5, -2, 0, 0, 0, 10, 0)'$$

(a) Linear model:

$$y_i = \beta_1 x_{1,i} + \ldots + \beta_{12} x_{12,i} + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, 0.25),\ i \leq N$$

(b) Binary response model:

$$y_i = \mathbb{1}_{\left\{\frac{\exp(\beta' \cdot x_i + \epsilon_i)}{1+\exp(\beta' \cdot x_i + \epsilon_i)} \geq 0.8\right\}} \qquad \epsilon_i \sim \mathcal{N}(0, 1),\ i \leq N$$

Of the twelve features seven were influential. We constructed in every case one (4×4) - block of positively correlated predictors, one (4×4) - block of independent predictors, which both have regression coefficients with the same absolute value and a (2×2) - block of negatively correlated predictors, where the variable X_{11} has in absolute value the largest overall coefficient. We want to compare the selection rate between those features w.r.t to the primary split and all splits in each tree.

For the simulation study we considered different values of the input parameter **mtry** for the **cforest()** function. This parameter determines the amount of randomly selected features at each node in each tree upon the split is made. Default values were used for the other parameters i.e. number of trees fitted were set to be 500. Figure 3.2 and 3.3 present the results for the linear model.

In figure 3.2 one can approximately see a uniform distribution of all features for $mtry = 1$, which is of course due to the random selection of the features i.e. there are no competitor features. Increasing the parameter *mtry* yields a different picture. For $mtry = 3$ the predictors X_3, X_4, X_{12} are clearly more often selected than X_5, X_6 even though they have coefficients equal to zero or in absolute value smaller than those of X_5, X_6 and therefore no or little impact on the response. If $mtry \in \{8, 12\}$ it seems that the "best" first split is made using X_1, nevertheless if $mtry = 8$ the features X_2, X_3, X_4 are, due to their high correlation, also selected quite often in the tree building process. Interestingly the variable with the strongest influence on the response X_{11} is never selected in this case and thus even less often than the variables X_3 and X_4, which have little or no influence.

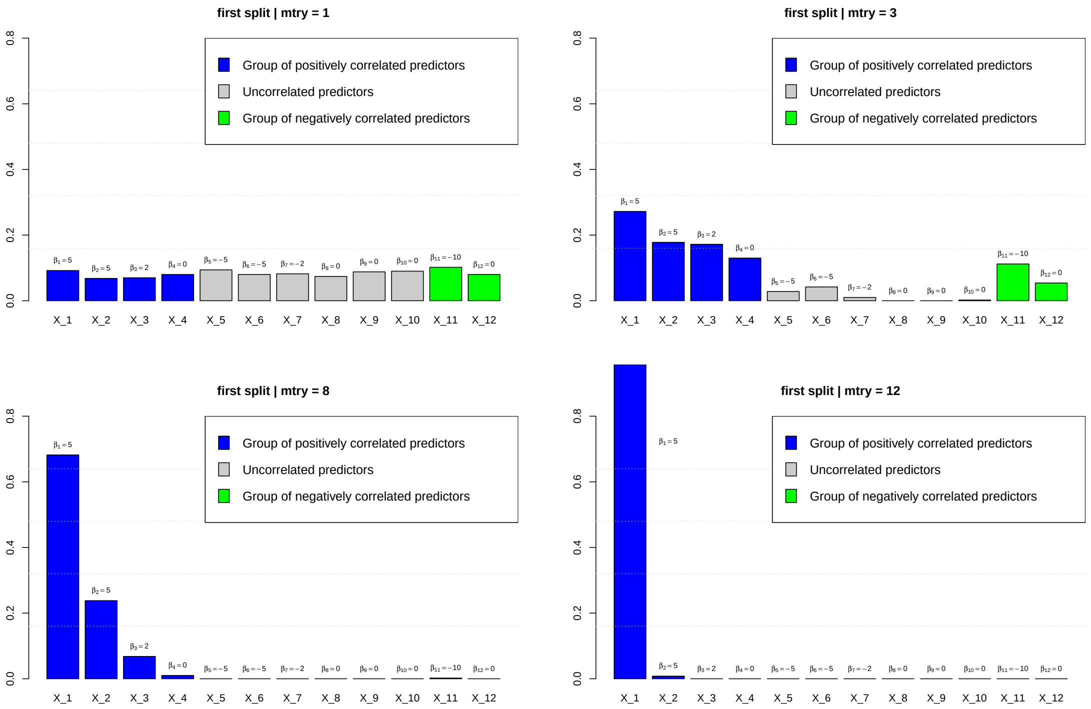


Figure 3.2: Selection rates in the linear model in the first split.

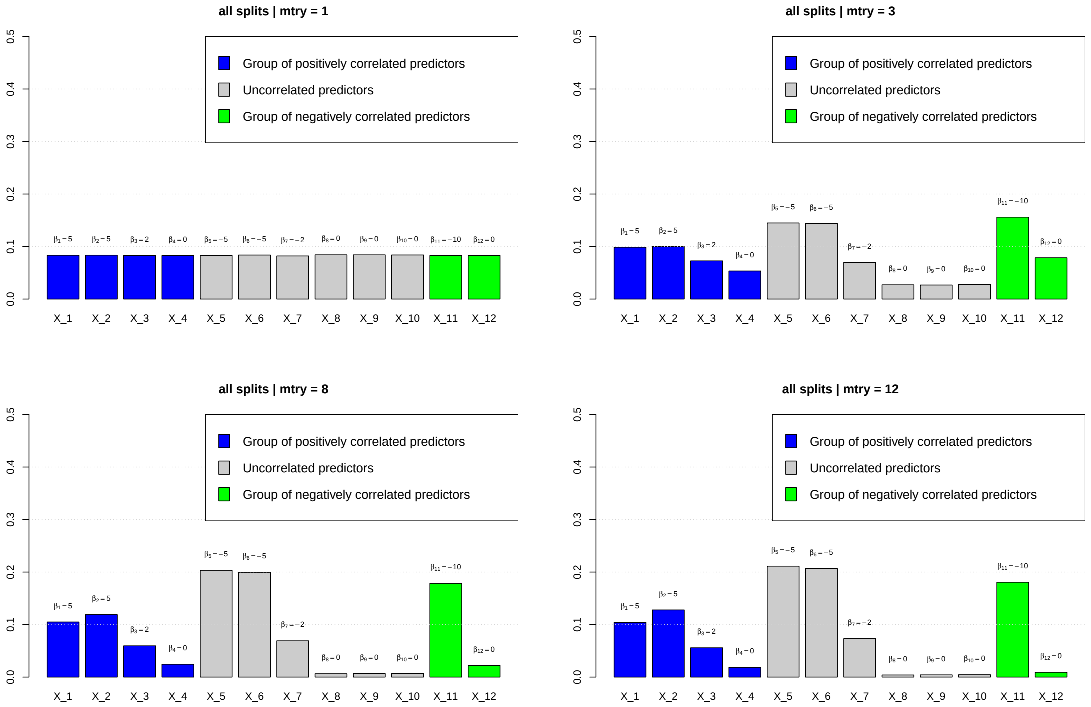


Figure 3.3: Selection rates in the linear model in all splits.

In figure 3.3 one can observe the same effect for $mtry = 1$ as in the case where only the first split was considered. For increasing *mtry* we get different results. Because variable selection is now conditionally on the previously selected variables i.e. the parent nodes, the correlated predictors which have little or no influence X_3, X_4, X_{12} are less often selected. However X_4 and X_{12} are still sometimes selected and thus more often than the other features without any influence X_8, X_9, X_{10}.

Similar results are obtained for the classification model. They are presented in figure 3.4 and 3.5.

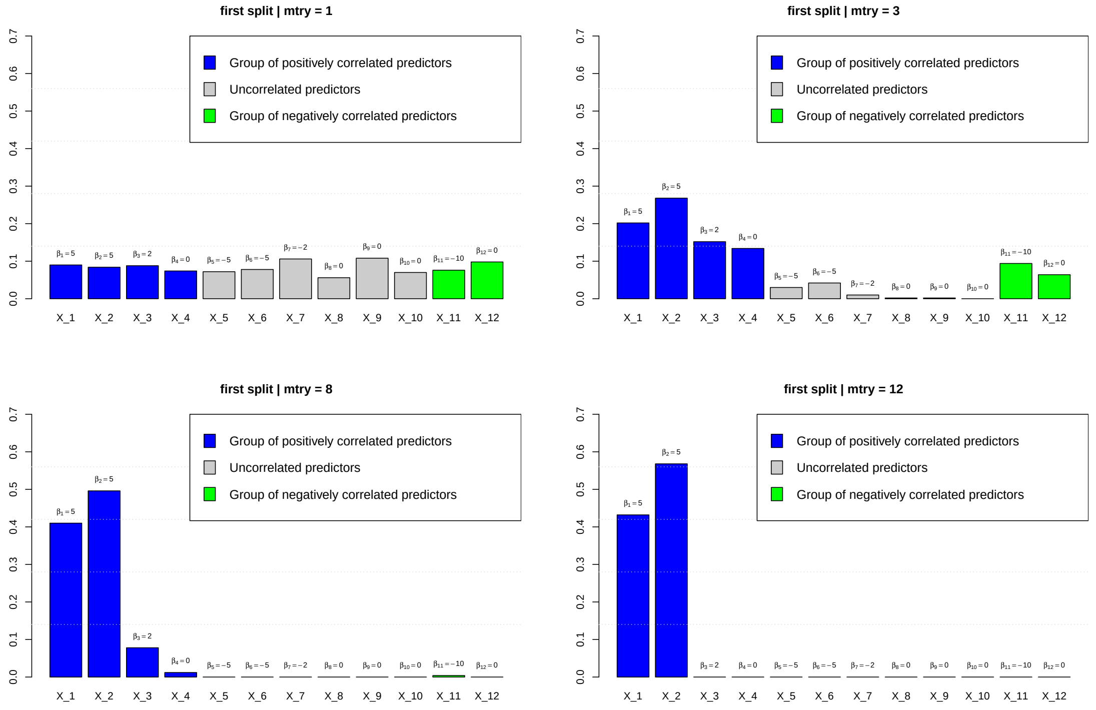


Figure 3.4: Selection rates in the binary response model in the first split.

2. The inherent null hypothesis for the *unconditional permutation importance.*

In the context of permutation tests one usually considers a null hypothesis which implies the independence of a certain predictor variable to the response **and** the remaining predictors. Under this global null hypothesis a permutation does not affect the joint distribution of them. Therefore a change in the joint distribution or a test statistic computed from it implies that the null hypotheses does not hold. The *unconditional permutation importance*, where one feature X_j is permuted against the target Y and the remaining features $Z := (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_K)$, corresponds to a null hypothesis that X_j is independent of both Y and Z:

$$\mathcal{H}_0 : X_j \perp Y \text{ and } X_j \perp Z$$

The "test statistic" used in this setting is the decrease in the prediction accuracy. A positive value of the *unconditional permutation importance* corresponds to a violation of this null hypothesis. This can be caused either if X_j and Y are not independent or the rather uninteresting case if X_j and Z are not independent. This implies naturally an advantage of

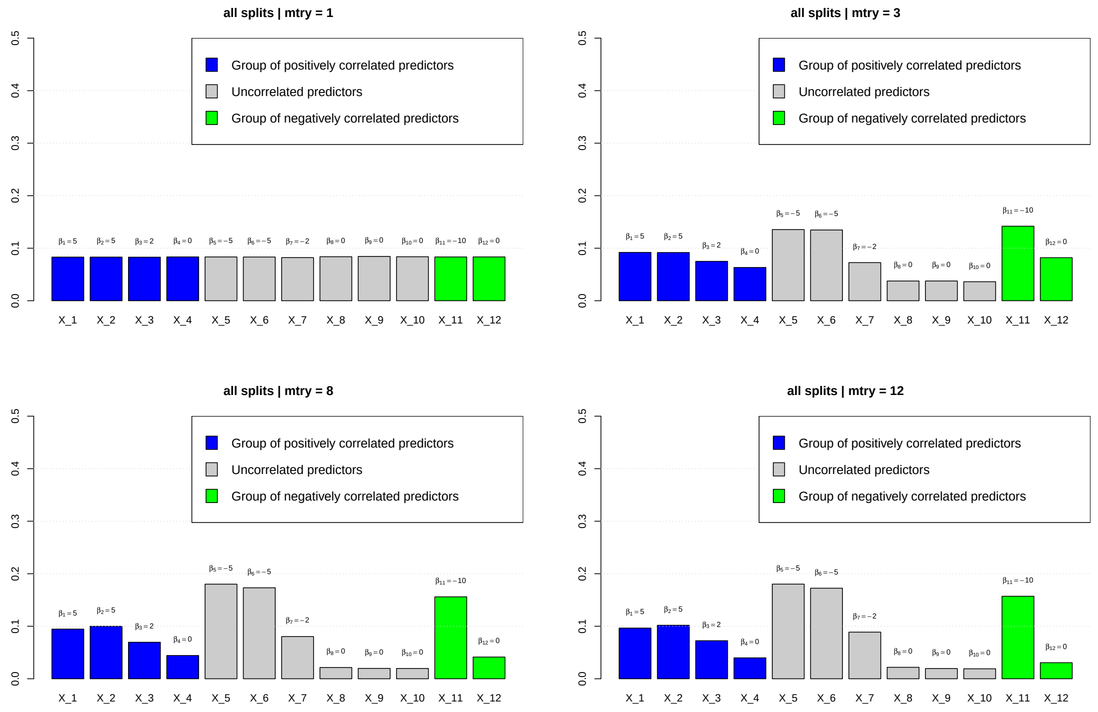


Figure 3.5: Selection rates in the binary response model in all splits.

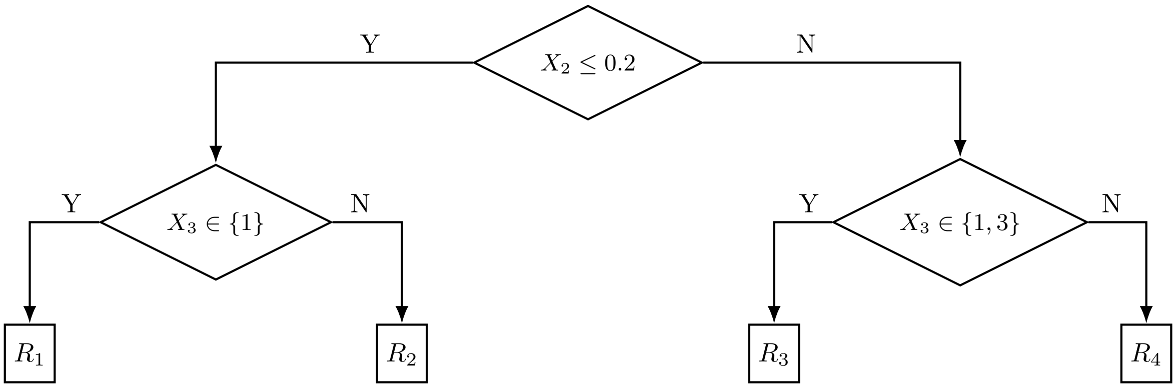
correlated features. We remark here that the selection frequency can be also used as a naive variable importance measure.

We will now focus on a more reliable importance measure which can deal better with correlated predictor variables. In order to measure only the impact of X_j to the response Y the *conditional permutation importance* measure is discussed. The null hypothesis can be formulated as follows:

$$\mathcal{H}_0 : (X_j \perp Y)|Z$$

If we are in a orthogonal setting i.e. $\mathbb{P}\left[(X_1, \ldots, X_K) \in A\right] = \prod_{j=1}^{K} \mathbb{P}\left[X_j \in A\right]$ both measures will theoretically yield the the same results. However if correlation is present the *unconditional permutation importance* will tend to overestimate the importance of non influential correlated predictors as outlined above. If Z consist only of categorical variables the grid to be conditioned on can be simply defined by the number of categories. However permuting the observed values of X_j conditioned on $Z = z$ is problematic if Z contains continuous variables. Even in the case where Z is categorical and having many levels this approach can sometimes be computationally infeasible. Thus the authors of Strobl et al. (2008) propose the grid, within the values of X_j are permuted, as the resulting partition on the space of the predictors, which is induced by the fitted tree. This can be also applied to continuous features. However for ease of computation they suggest to use all splitting criteria as cut points of the predictor space. The difference of both approaches is best explained using a simple example:

Consider three explanatory variables $X := (X_1, X_2, X_3) \in \mathbb{R}^2 \times \{1, 2, 3\}$. We want to measure the conditional importance of X_1 on the target Y conditioned on X_2, X_3. Suppose the following tree has been fitted



which leads to a partition of the feature space into:

$$R_1 = \{X | X_2 \leq 0.2, X_3 = 1\}, \qquad R_2 = \{X | X_2 \leq 0.2, X_3 \in \{2, 3\}\}$$
$$R_3 = \{X | X_2 > 0.2, X_3 \in \{1, 3\}\}, \qquad R_4 = \{X | X_2 > 0.2, X_3 = 2\}.$$

Following the second approach, where each split is used as a bisection of the entire feature space and not only of the current node, one would obtain the following grid:

$$\tilde{R}_1 = \{X | X_2 \leq 0.2, X_3 = 1\}, \quad \tilde{R}_2 = \{X | X_2 \leq 0.2, X_3 = 2\}, \quad \tilde{R}_3 = \{X | X_2 \leq 0.2, X_3 = 3\}$$
$$\tilde{R}_4 = \{X | X_2 > 0.2, X_3 = 1\}, \quad \tilde{R}_5 = \{X | X_2 > 0.2, X_3 = 2\}, \quad \tilde{R}_6 = \{X | X_2 > 0.2, X_3 = 3\},$$

which is more fine graded.

Nevertheless conditioning too strictly does not have a negative effect from a theoretical point of view as opposed to the other way around. Several ways were proposed to select the variables Z to be conditioned on. The most conservative is to select all remaining variables. One can also define for (continuous) predictors a threshold for some dependency measure such as the Pearson Correlation Coefficient, Kendall's Tau or Spearman's Rho and select only those variables for Z that attain a value greater than this threshold. The framework of Hothorn et al. (2006) provides also p-values, which measure the association between the predictor and the target variable. These can also be used for constructing Z. Furthermore this method can be applied to all types of variables on different scales including categorical.

A pseudo code for the *conditional permutation importance* is given below in algorithm 3.

The *conditional permutation variable importance* is accessible in the R packages **partykit** Hothorn et al. (2017-12-13) and **party** Hothorn et al. (2017-12-12) via the function **varimp(. . ., conditional=TRUE, threshold =.2)**. The variables to be conditioned on Z are determined as:

> [. . .] If conditional = TRUE, the importance of each variable is computed by permuting within a grid defined by the covariates that are associated (with $(1 - p)$ - value greater than threshold) to the variable of interest.

Using the **partykit** package the risk evaluated for the mean decrease in accuracy in the case of a regression forest is given by the log-likelihood instead of the mean squared error, which is used in the packages **party** and **randomForest**.

Algorithm 3: Pseudo code for calculating conditional permutation importance for a single feature X_j

```
1. Fit a random forest $\{h(T_b) : b \leq B\}$ on the training set $T$ using algorithm 1 (where the
   splitting rule can be also made using a permutation test framework as presented in Hothorn
   et al. (2006))

2. for $b = 1$ to $B$ do
   |   (a) Compute the OOB prediction accuracy of the $b$-th tree $h(T_b)$ (see section (3.1.2)).
   |   (b) Determine the variables $Z$ to be conditioned on, extract all cutpoints for each variable
   |       in the current tree and construct the grid by bisecting the feature space in each
   |       cutpoint.
   |   (c) Within this grid permute the observations of the feature $X_j$ in the OOB sample $O_b$.
   |   (d) Recompute the OOB prediction accuracy of the $b$-th tree $h(T_b)$ using the permuted
   |       input i.e. (for classification) compute
   |
   |           $\frac{\sum_{i \in O_b} \mathbb{1}_{\{y_i = \hat{y}^b_{i,\pi_j|Z}\}}}{|O_b|}$,
   |
   |       where $\hat{y}^b_{i,\pi_j|Z} = h(T_b)(x_{i,\pi_j|Z})$ is the prediction of the $b$-th tree after permuting the
   |       observations of $X_j$ within the grid defined by $Z$.
   |   (e) Compute $I^b_{permute,conditional}(j)$ using (a) and (d) analogously as for $I^b_{permute}(j)$.
   end

3. Compute the average decrease of prediction accuracy over all trees which we will denote by
   $I_{permute,conditional}(j)$ analogously using $I^b_{permute,conditional}(j)$ as for $I_{permute}(j)$.
```

3.1.4 Importance measure with missing data

In this section we will discuss a permutation importance measure similar to the *unconditional permutation importance* which can be applied in the case of missing data. This section is along the lines of Hapfelmeier et al. (2014).

There are several methods for handling missing values in a tree building process. One possibility is to stop the throughput of an observation x_j at the node where the splitting variable is missing in this observation and to use the mean or the majority vote of the response at this node for prediction. Another approach is to set the missing values as the majority of the observed values of all observations. Nevertheless the most popular method is to use *surrogate splits* that are based on additional predictors. Surrogate splits are constructed such that the resulting binary partition “mimics” the primary split that could not be made because of missing values of the splitting variable. Usually several surrogate splits are calculated at each node and ranked according to their ability of resembling the original split. If an observation contains further missing values in the surrogate splits they are selected according to their rank.

If surrogate splits are used the *unconditional permutation importance* discussed in section 3.1.2 is no longer applicable. This is due to the fact that surrogate splits are not directly affected by the permutation of the variable of interest. The authors of Hapfelmeier et al. (2014) therefore proposed a modification of the *unconditional permutation importance* in the case of missing values which we will discuss below.

The substantial difference to the previously discussed permutation measures is the following. Instead of permuting a predictor variable X_j, which could theoretically contain missing values, one randomly allocates observations at each node, where the split is conducted via X_j, to the two child nodes. The random assignment is done using the relative frequency in each of the two child nodes as a distribution. This can be formalized as follows:

Let the random variable $D_k \sim Ber(p_k)$ represent the decision if an observation at node k is assigned to the left child node $\mathbb{P}\left[x_m \text{ to left}\right] = \mathbb{P}\left[D_k = 1\right]$ or respectively to the right child node $\mathbb{P}\left[x_m \text{ to right}\right] = \mathbb{P}\left[D_k = 0\right]$. Furthermore let $p_k := \frac{n_{left}}{N_{node(k)}}$ where $N_{node(k)}$ is defined to be the number of observations at the k-th node and n_{left} represents the number of observations which were assigned to the left child node of node k.

The null hypothesis i.e. independence of X_j from Y and $Z := (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_K)$ is given as:

$$\mathcal{H}_0 : \mathcal{L}(D_k | X_j) = \mathcal{L}(D_k),$$

i.e. that the random assignment of an observation does not depend on the particular predictor variable X_j. All in all the procedure can be summarized as follows.

Algorithm 4: Pseudo code of the permutation importance with missing data for a single feature X_j

```
1. Fit a random forest $\{h(T_b) : b \leq B\}$ on the training set $T$ using algorithm 1 (where the
   splitting rule can be also made using the permutation test framework as presented in
   Hothorn et al. (2006)).
2. for $b = 1$ to $B$ do
       (a) Compute the OOB prediction accuracy of the $b$-th tree $h(T_b)$.
       (b) Randomly assign at each node $k$, where the splitting rule is defined via $X_j$ all
           observations to the two child nodes using the random variable $D_k$.
       (c) Recompute the OOB prediction accuracy of the $b$-th tree $h(T_b)$.
       (d) Compute $I^b_{permute,missing}(j)$ as the difference between the original and the recomputed
           OOB accuracy.
   end
3. Compute the average decrease of prediction accuracy over all trees $I_{permute,missing}(j)$.
```

This algorithm is accessible via the function **varimp(...,pre1.0_0 = FALSE)** in the R package **party** Hothorn et al. (2017-12-12) or as standard implementation of the function **varimp()** in the **partykit** package if missing values are present in the input X.

3.1.5 Simulation study: permutation importance measures for random forests

In this section we will compare the three different permutation variable importance measures presented in the previous sections. We will concentrate on unbiased random forests from the R package **party** since importance measures based on them are more reliable as shown in Strobl et al. (2007). The simulation will be based again on the linear model from section 3.1.3.

We fitted a random forest using the R function **cforest()** from the **party** package to $N = 1000$ simulated data points according to the linear model presented above:

predictor	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
β_i	5	5	2	0	-5	-5	-2	0	0	0	10	0

where the green and the blue variables were selected to be positively correlated with $r = 0.9$ respectively negatively correlated with $r = -0.9$.
As hyper parameters we used the default settings for different $mtry$ parameters i.e.

cforest(... ,control=cforest_unbiased(ntree=500, mtry)).

The permutation importance measures were calculated using the **varimp()** function as following:

- unconditional: **varimp(..., nperm=** 50**, pre1.0_0 = TRUE)**
- conditional: **varimp(..., conditional=TRUE, threshold=** 10^{-16}**, nperm=** 50)

- unconditional (missing data): **varimp(**. . ., **nperm**= 50, **pre1.0_0 = FALSE)**

The results of the mean importance scores w.r.t. the number of permutations are presented in figure 3.6 and 3.7.

One can clearly observe in figure 3.6 that the conditional variable importance measure yields much smaller importance scores as the unconditional ones. Furthermore the unconditional permutation importance measure and the permutation importance measure with missing data do basically coincide. This is not a surprising fact since we did not artificially create missing values in our data and thus they should coincide as has been shown by Hapfelmeier et al. (2014). The results also do show that the permutation importance scores are naturally strongly influenced by the parameter *mtry*. For $mtry \in \{1, 3\}$ the scores do not reflect the true underlying model, since variables with no or little influence X_3, X_4, X_{12} does have a large importance score. These are the correlated variables, but since the unconditional importance measures do not correct for multicollinearity, large importance scores are assigned to them. Nevertheless with increasing $mtry \in \{8, 12\}$ the correlated predictors with little or less influence are replaced by their correlated competitors which do have more influence X_1, X_2. Even though variable X_{12} does also show for $mtry = 8$ a higher importance than X_6. Finally when using all variables i.e. $mtry = 12$ (bagging) the true model is reflected the best. However using all variables in every split is rather unrealistic if the set of predictors gets too large and is also likely to induce correlated trees within a random forest.

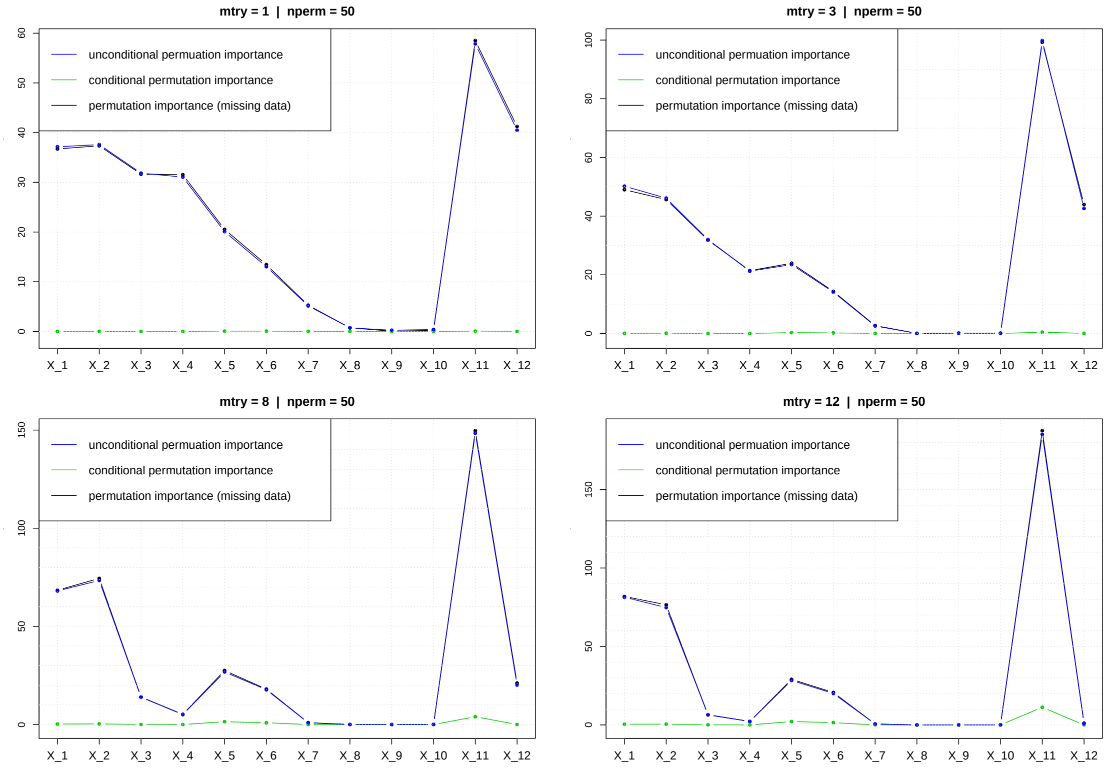


Figure 3.6: Comparison of permutation importance measures of random forests under multicollinearity.

Having a closer look at the conditional variable importance measure in figure 3.7 one can see that the true model is better reflected in all cases. Nevertheless correlated predictors with influence such as X_1, X_2 are now kind of downgraded for large values of the $mtry$ parameter.

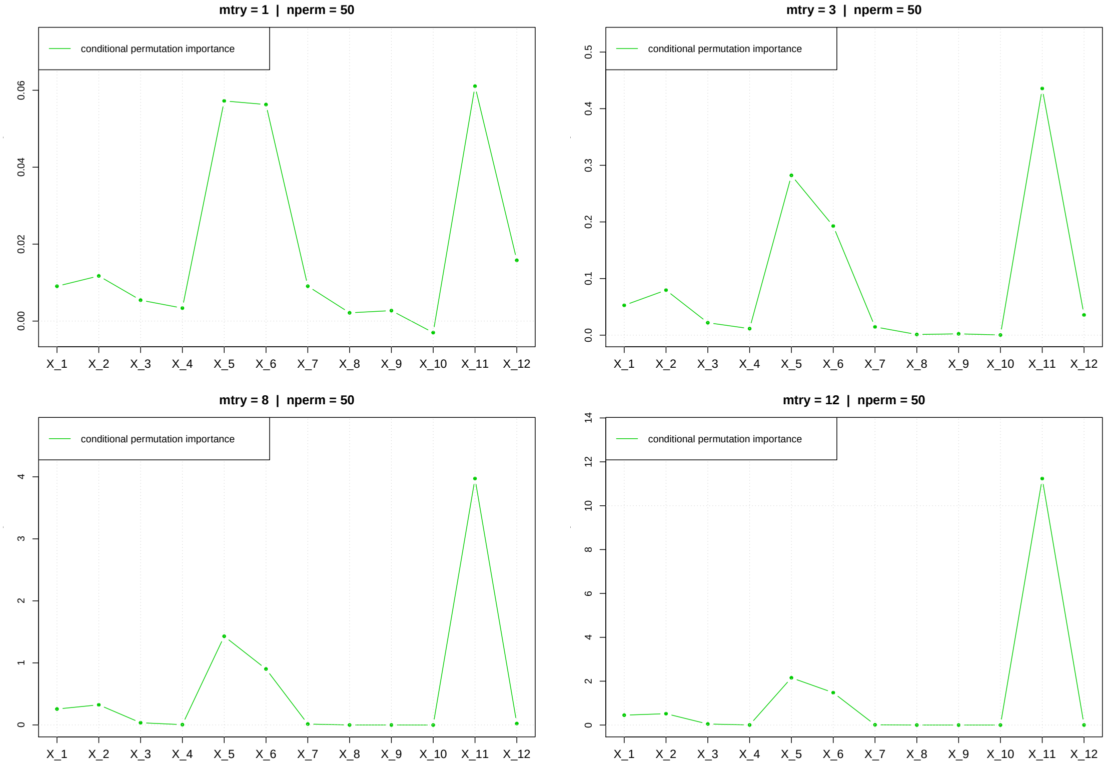


Figure 3.7: Conditional permutation importance measure for random forests under multicollinearity.

The ranks of the calculated importance scores are presented in table 3.1.

mtry		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_10	X_11	X_12
	missing data method	4	3	5	6	7	8	9	10	12	11	1	2
1	conditional	7	5	8	9	2	3	6	11	10	12	1	4
	unconditional	4	3	5	6	7	8	9	10	12	11	1	2
	missing data method	2	3	5	7	6	8	9	12	10	11	1	4
3	conditional	5	4	7	9	2	3	8	11	10	12	1	6
	unconditional	2	3	5	7	6	8	9	12	10	11	1	4
	missing data method	3	2	7	8	4	6	9	11	12	10	1	5
8	conditional	5	4	6	9	2	3	8	11	10	12	1	7
	unconditional	3	2	7	8	4	6	9	11	12	10	1	5
	missing data method	2	3	6	7	4	5	9	11	12	10	1	8
12	conditional	5	4	6	8	2	3	7	10	12	11	1	9
	unconditional	2	3	6	7	4	5	9	11	12	10	1	8

Table 3.1: Ranks of permutation importance measures for random forest under multicollinearity.

All in all it is maybe good to use both an unconditional and a conditional permutation importance measure for correct inference about variable importance.

3.2 Permutation importance for linear and logistic models

This section is dedicated to the application of the *unconditional permutation importance*, which has been presented in section 3.1.2, to other models than random forests. Specifically we will test this method for a linear and a logistic regression setting.

The *unconditional permutation importance* is based on a very simple idea, which does not require the special structure of a random forest model. Thus it can in principle be applied to all kinds of classification or regression models. We will briefly review the main steps when calculating the *unconditional permutation importance* for a unspecified model. It is worth noting that for a random forest model the following algorithm is applied to each tree in the forest and finally averaged, where the test set is the OOB sample.

1. Divide the given data D in a training set $T \subset D$ and a test set $T^c \subset D$.
2. Fit a model $\mathcal{M}$ to the training set T.
3. Use the fitted model $\mathcal{M}$ to make predictions on the test set T^c.
4. Calculate the accuracy of the predictions using e.g. MSE, classification rate[2], the Lift measure or the Gini index.
5. Randomly permute the predictor of interest ceteris paribus in the test set T^c and recalculate the accuracy using the permuted data.
6. Calculate the decrease in accuracy.

In the following we test the *unconditional permutation importance* for a linear as well as a logistic regression model.

3.2.1 Permutation importance - linear regression setting

We will apply the *unconditional permutation importance* on two simulated data sets in the case of a linear regression setting. These were generated as:

$$y_i = \beta_1 x_{1,i} + \ldots + \beta_{12} x_{12,i} + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, 25),\ i \leq N := 1000$$

where

$$(X_1, \ldots, X_{12})' \sim \mathcal{N}(\mu, \Sigma) \qquad (\Sigma)_{i,j} := \begin{cases} 1 & ,j = i \\ 0.9 & ,i \neq j \leq 4 \\ -0.9 & ,i \neq j \geq 9 \\ 0 & ,\text{else} \end{cases}$$

$$\mu := (0,0,0,0,0,0,0,0,0,0,0,0)'$$

$$(\beta_1, \ldots, \beta_{12})' = \begin{cases} (5,5,2,0,-5,-5,-2,0,0,0,10,0)', & \text{Dataset 1} \\ (10,5,2,0,-10,-5,-2,0,0,0,10,0)', & \text{Dataset 2} \end{cases}$$

[2] By classification rate we mean the percentage of correctly classified labels i.e. $\hat{Y}_i = Y_i$

The training set and the test set were constructed according to a 70 : 30 rule. For measuring the accuracy we used the MSE. The linear model was fitted using the R function **lm** from the **stats** package. The results are presented in figure 3.8 and 3.9. Figure 3.8 displays the distribution of 500 repetitions. Ranks of the permutation importance method as well as of the absolute value of the estimated regression coefficients[3] are presented at the top of each bar in figure 3.9.

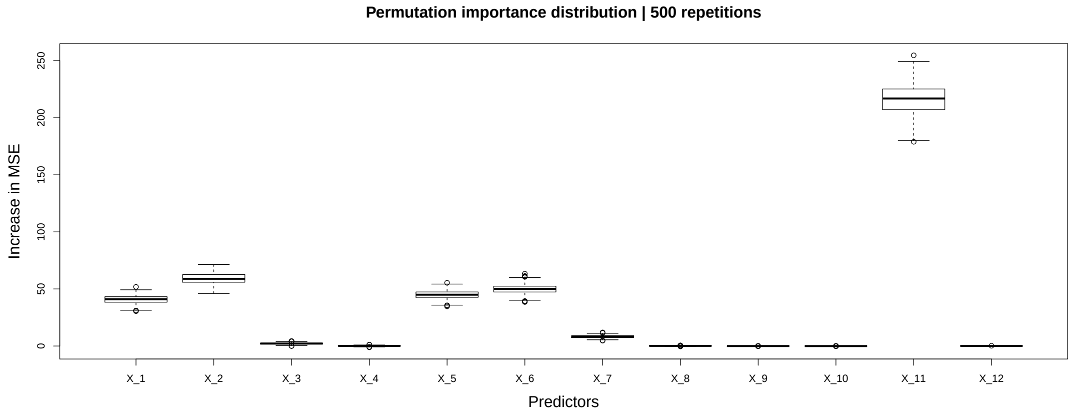


Figure 3.8: Permutation importance linear regression: Dataset 1, distribution.

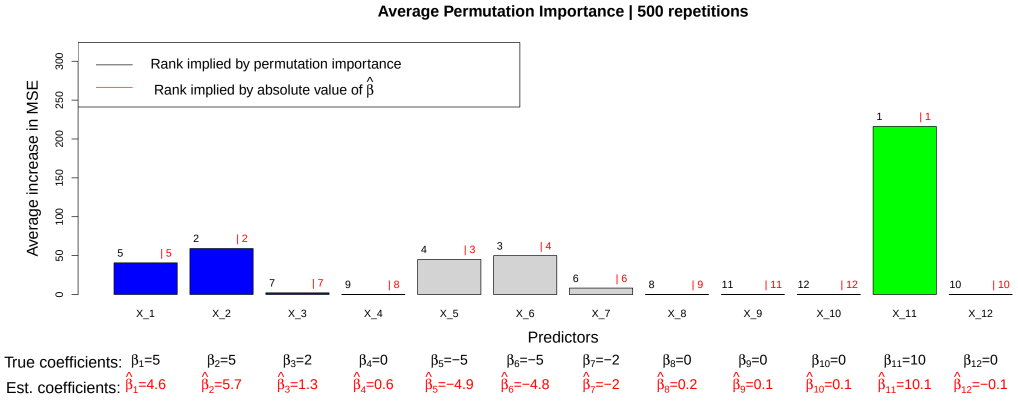


Figure 3.9: Permutation importance linear regression: Dataset 1.

From figure 3.9 we can see that this method ranks almost equally according to the absolute value of the estimated coefficients as well as the true coefficients, which is intuitively, due the definition of the unconditional permutation importance, not a surprising fact. One can also observe from figure 3.8 that the larger the estimated coefficient the larger the variance of the calculated permutation importance.

Results for the second dataset are presented in figure 3.10 and 3.11, where we can observe a similar pattern. All in all analyzing variable importance using the unconditional permutation importance yields similar results as one would obtain via the absolute value of the estimated coefficients. Furthermore one cannot observe a preference for correlated predictors when applying this method.

[3] Since the data X was chosen to be standardized, these are equal to the standardized regression coefficients, $betasq$, from section 1.1.1 when neglecting the division by s_Y.

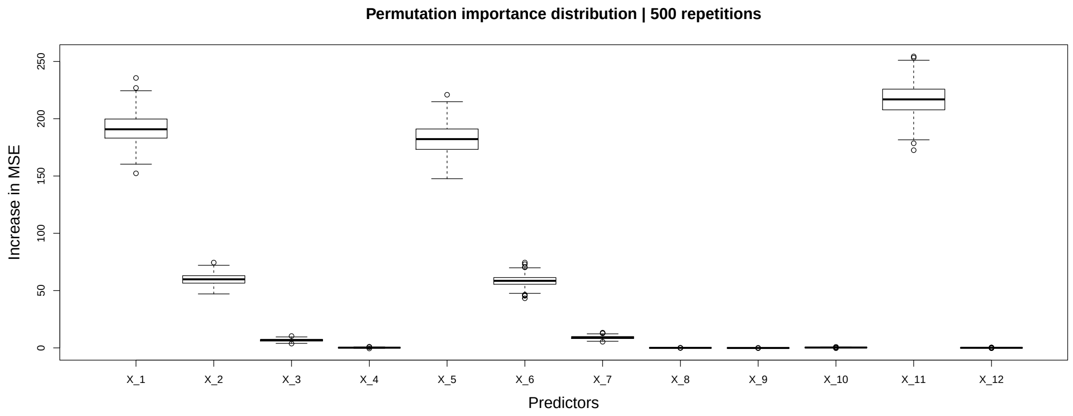


Figure 3.10: Permutation importance linear regression: Dataset 2, distribution.

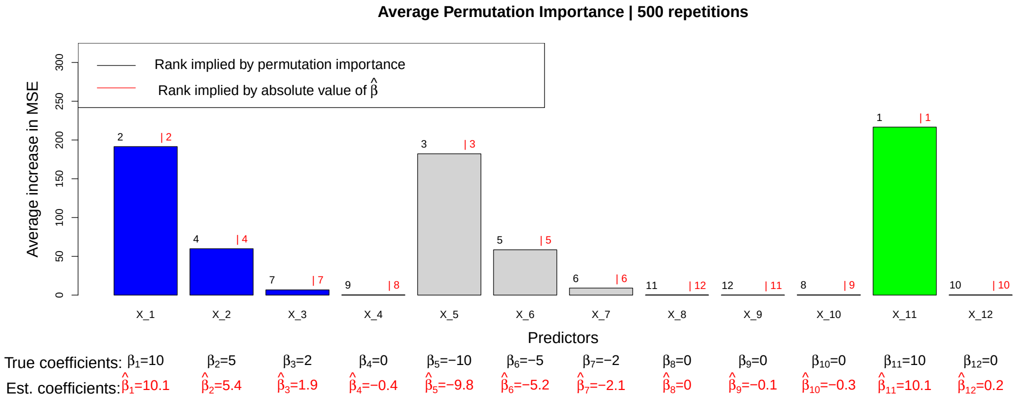


Figure 3.11: Permutation importance linear regression: Dataset 2.

3.2.2 Permutation importance - logistic regression setting

The same analysis as in the previous section is conducted in the case of a logistic regression model defined by[4]:

$$(X_1,\dots,X_{12})' \sim \mathcal{N}(0,\Sigma) \qquad (\Sigma)_{i,j} := \begin{cases} 1 & ,j=i \\ 0.9 & ,i\neq j\leq 4 \\ -0.9 & ,i\neq j\geq 9 \\ 0 & ,\text{else} \end{cases}$$

$$(\beta_1,\dots,\beta_{12})' = \begin{cases} (5,5,2,0,-5,-5,-2,0,0,0,10,0)', & \text{Dataset 1} \\ (10,10,2,0,-5,-5,-2,0,0,0,10,0)', & \text{Dataset 2} \end{cases}$$

$$y_i \text{ drawn from } Y_i \sim \text{Ber}\left(\left\{\frac{\exp\left(\beta'\cdot x_i\right)}{1+\exp\left(\beta'\cdot x_i\right)}\right\}\right), \qquad i\leq N := 1000.$$

[4] Ber(p) is defined to represent the Bernoulli distribution on $\{0,1\}$ with success probability $p\in[0,1]$.

For measuring the accuracy i.e. the classification rate, we used the empirical mean of the loss function l that is defined as: $l(\hat{y}_i, y_i) := |\mathbb{1}_{\{\hat{y}_i \geq 0.5\}} - y_i|$, where $\hat{y}_i$ denotes the estimated probability that $y_i = 1$ and $y_i \in \{0, 1\}$ is the observed data. Results are presented in figure 3.12 and 3.13. A similar pattern as in the linear case can also be observed in the logistic regression setting i.e. ranks of the *unconditional permutation importance* are almost equal to the absolute coefficient of the estimated regression coefficients, no preference of correlated predictors and larger variance for regressors that have in absolute value a large estimated regression coefficient.

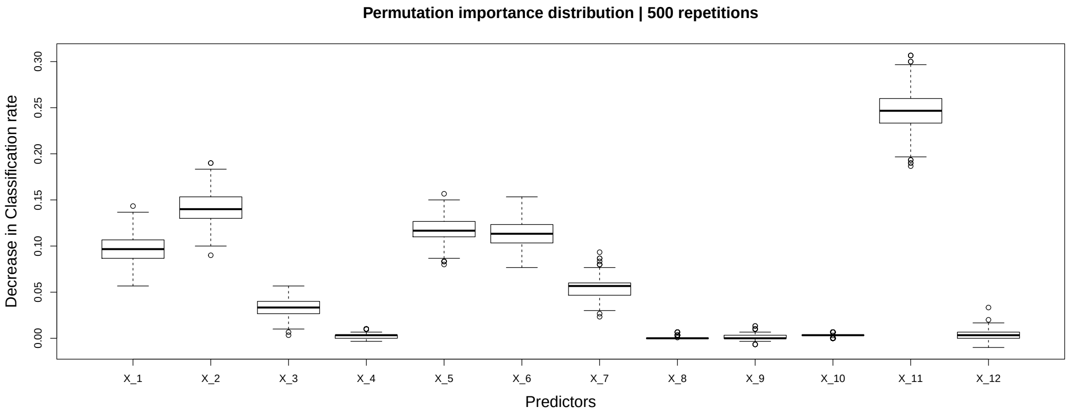


Figure 3.12: Permutation importance logistic regression: Dataset 1, distribution.

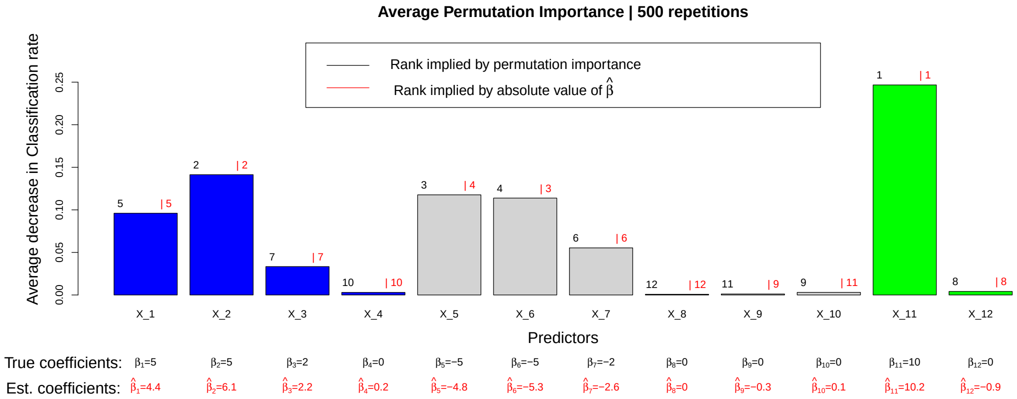


Figure 3.13: Permutation importance logistic regression: Dataset 1.

Additionally we compared the obtain ranks through the *unconditional permutation importance* with the **varImp()** function of the R **caret** package (Kuhn et al. (2018-03-29)), which computes the absolute value of the *z-statistic* of the MLE coefficients.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
varImp() function	6	5	7	10	2	3	4	12	9	11	1	8
Permutation Importance	5	2	7	10	3	4	6	12	11	9	1	8

Table 3.2: Ranks comparison varImp() function vs. permutation importance in a logistic setting.

Again these two measures almost coincide by being able to identify the most influential predictors $\{X_{11}, X_1, X_2, X_5, X_6\}$, whereas the varImp() function seems to down weight correlated variables e.g. $\{X_1, X_2\}$ in comparison to their uncorrelated counterparts $\{X_5, X_6\}$ and ranks X_7 above $\{X_5, X_6\}$. This effect can even be more dramatically observed for the second dataset in figure 3.15, where $\{X_5, X_6, X_7\}$ does almost have the same importance as $\{X_1, X_2\}$, although we increased the true coefficients in the data generating process from 5 to 10 for $\{X_1, X_2\}$

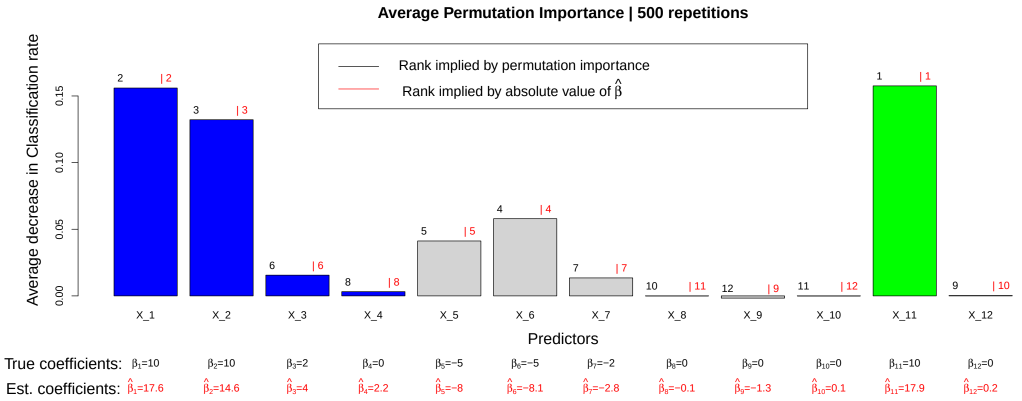


Figure 3.14: Permutation importance logistic regression: Dataset 2.

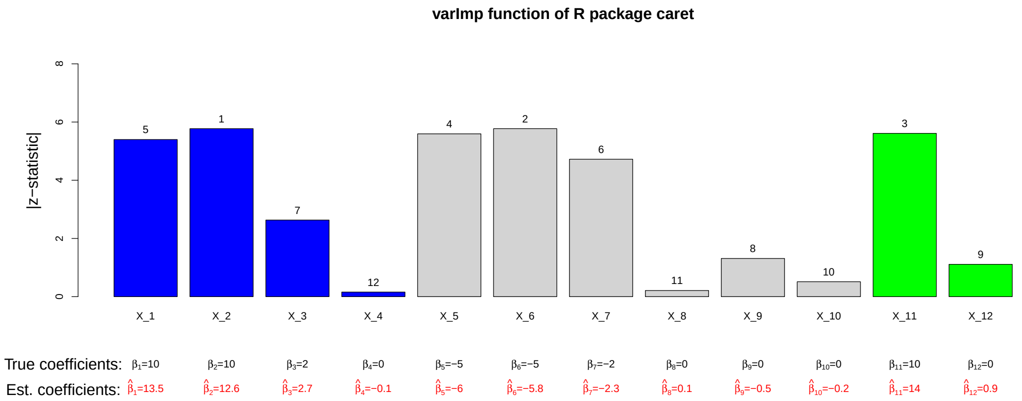


Figure 3.15: varImp results: Dataset 2.

The reason why the estimated coefficients from figure 3.14 and 3.15 differ is that in figure 3.14 they result from fitting a model on the training set $T \subset D$, while in figure 3.15 they are obtained by fitting a model on the whole dataset D.

Chapter 4

Variable importance in logistic regression

In this chapter we present one possible method to measure relative variable importance of regressors in a logistic regression setting.

4.1 Setting

This section briefly reviews the principles of logistic regression.

We are given a matrix of explanatory variables denoted by $X = (X_1, \ldots, X_p) \in \mathbb{R}^{n \times p}$ where $X_i \in \mathbb{R}^n$ denotes the i-th regressor variable. In the following we will not distinguish between the random Variable $X_i : \Omega \mapsto \mathbb{R}$ and the corresponding column of X i.e. the n observed values of $(X_i(\omega_1), \ldots, X_i(\omega_n))$ and denote both objects for the sake of readability by X_i. Thus X will also denote a p-dimensional random vector. From the context it should be clear what X_i or X represents. The i-th row of the matrix X will be denoted by x_i.

The response $Y = (Y_1(\omega_1), \ldots, Y_n(\omega_n)) \in \mathbb{R}^n$ is given as n realizations of i.i.d binary random variables $Y_i : \Omega \mapsto \{0, 1\}$. We furthermore define $\pi_i := \mathbb{P}[Y_i = 1|X]$ as the conditional probability that the dependent variable equals 1. In a logistic regression one tries to model π_i using X.

Nevertheless the difference of logistic regression in contrast to multiple linear regression is that the response and the explanatory variables are linked in a nonlinear way. The link function f (log odds) and the response function f^{-1} are given as:

$$\begin{aligned} f(\pi_i) &:= \log\left(\frac{\pi_i}{1-\pi_i}\right) \quad \in \overline{\mathbb{R}} \\ f^{-1}(x_i'\beta) &:= \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \quad \in [0, 1]. \end{aligned}$$

As in the standard linear regression setting one tries to model the conditional expectation of Y_i

given X in the following way:

$$\pi_i = \mathbb{P}[Y_i = 1|X] = \mathbb{E}[Y_i|X] \overset{!}{=} f^{-1}(X\beta) = \frac{e^{X'\beta}}{1+e^{X'\beta}}$$

which can be equivalently written in terms of the response function f as:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = f(\pi_i) = X\beta.$$

The goal is to estimate the vector of coefficients β. This is done by maximum likelihood estimation. Given the observed values $Y = (y_1, \ldots, y_n) \in \{0,1\}^n$ and $(x_1, \ldots, x_n)' \in \mathbb{R}^{n\times p}$ the likelihood function L as well as the log likelihood function $\ell := \log(L)$ can be written as:

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \mathbb{P}[Y_i = y_i|x_i](\beta) = \prod_{i=1}^{n} \pi_i(\beta, x_i)^{y_i} \cdot (1-\pi_i(\beta, x_i))^{1-y_i} \\ \ell(\beta) &= \sum_{i=1}^{n} y_i \log(\pi_i(\beta, x_i)) + (1-y_i)\log(\pi_i(\beta, x_i)), \end{aligned}$$

where $\pi_i(\beta, x_i) := \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$. The estimator of β denoted by $\hat{\beta}$ is now given by the MLE:

$$\hat{\beta} := \underset{\beta}{\operatorname{argmax}}\, \ell(\beta). \tag{4.1}$$

It is worth noting that there is no closed form solution of problem (4.1) and it must be computed numerically, which is usually done by using the *Fisher scoring algorithm* that results in an *iterative weighted least squares (IWLS) algorithm.*

4.2 Extension of relative weights to logistic regression - a heuristic approach

In this section we will discuss one possible way to assess relative variable importance in the context of logistic regression. At the end we conduct an empirical study on a simulated data set to test the method in greater detail. This section is based on Tonidandel and LeBreton (2010) and Johnson (2001).

The method of *relative weights* has been originally proposed for multiple linear regression in order to evaluate more accurately the importance of predictors under multicollinearity. The concept of relative importance of a predictor is in this section defined as the contribution each regressor makes to the variance of the predicted response, when considered by itself and in combination with other predictors. Thus similar as in chapter 1 (decomposition of the coefficient of determination), the importance of a predictor is solely measured as the proportionate contribution each predictor makes to the coefficient of determination.

We will first outline the concept of relative weights in a multiple linear regression setting. After that we will present an extension of this method to logistic regression.

4.2.1 Relative weights in multiple linear regression

Let $Y \in \mathbb{R}^n$ be the response and $X \in \mathbb{R}^{n\times p}$ the explanatory variables in the multiple linear regression setting, where we assume that X and Y are in standard score form i.e. exemplary for X:

$$||X_k||_2^2 = \sum_{i=1}^{n} x_{ik}^2 = 1 \quad \forall k \in \{1,\dots,p\} \tag{4.2}$$

$$\sum_{i=1}^{n} x_{ik} = 0 \quad \forall k \in \{1,\dots,p\} \tag{4.3}$$

From that it follows immediately that the standard deviation of each column of X and Y is given by $s_Y = s_{X_k} = (n-1)^{-\frac{1}{2}}$. The basic idea is to find orthogonal regressors Z which "best" approximate the original explanatory variables X.

Let the singular value decomposition (SVD) of X be given as:

$$X = P\Delta Q',$$

where $P \in \mathbb{R}^{n\times n}$, $\Delta \in \mathbb{R}^{n\times p}$ and $Q \in \mathbb{R}^{p\times p}$. Under the assumption that X has full rank p the diagonal elements of Δ, the squared singular values, are not equal to zero. The SVD decomposition can be equivalently written in reduced form as:

$$X = \tilde{P}\tilde{\Delta}Q',$$

where $\tilde{P} \in \mathbb{R}^{n\times p}$ and $\tilde{\Delta} \in \mathbb{R}^{p\times p}$.

It was shown in Johnson (1966) that the desired minimal (in the least square sense) orthogonal linear transformation of X, denoted by Z is given by:

$$Z = \tilde{P}Q'.$$

Formally $Z = \tilde{P}Q'$ and $T = Q\tilde{\Delta}^{-1}Q'$ is the solution to the following optimization problem:

$$\begin{aligned}
&XT = Z\\
&ZZ' = I\\
&\min_{Z} tr\left\{(X-Z)'(X-Z)\right\}.
\end{aligned}$$

Since X was assumed to be in standard score form we get that[1]:

$$X'\mathbf{1}_n = Q\tilde{\Delta}\tilde{P}'\mathbf{1}_n = \mathbf{0}_n \iff \tilde{P}'\mathbf{1}_n = \Delta^{-1}Q'\mathbf{0}_n = \mathbf{0}_n \implies Z'\mathbf{1}_n = \mathbf{0}_n.$$

Thus we can conclude that Z is also column wise centered. Exploiting the orthogonality of Z i.e. $ZZ' = I_n$ we can infer that the column wise standard deviations are all equal and given by $s_{Z_j} = (n-1)^{-\frac{1}{2}}\ \forall j \in \{1,\dots,p\}$.

[1] We denote by $\mathbf{1}_n$ the n-dimensional vector where each component equals 1. Similarly $\mathbf{0}_n$ is the n-dimensional null vector.

The relative weights $\epsilon \in \mathbb{R}^p$, as a measure of variable importance, can be calculated as follows:

1. Regress Y on Z and calculate the **standardized** regression coefficients $\hat{\beta}$ (beta weights) which are the same as the unstandardized coefficients $\hat{b}$ after transforming X and Y in standard score form:

   $$\hat{\beta}_i = \hat{b}_i \cdot \frac{s_{Z_i}}{s_Y} = \hat{b}_i \cdot 1 \quad \text{where}$$
   $$\hat{b} = (Z'Z)^{-1}Z'Y = Z'Y = Q\tilde{P}'Y$$

   Since Z is orthogonal it holds that $(\hat{\beta}_i)^2 = (r_{Z_i,Y})^2$ and therefore we can decompose the coefficient of determination in the following way [2] :

   $$R^2_{Y|Z} = \mathcal{R}^2_{Y|Z} = \sum_{i=1}^{p} r^2_{Z_i,Y} = \sum_{i=1}^{p} (\hat{\beta}_i)^2.$$

   Thus the squared beta coefficients represents the relative proportion of explained variance in Y for each regressor and does coincide with our original definition of variable importance in this section. Nevertheless since the orthogonal regressors Z are only approximations of the original predictors X, they have to be "linked" back to X.

2. Relate the orthogonal variables Z back to the original predictors X:

   This can be done by regressing each column of X, i.e. each predictor X_j, on Z. The resulting **standardized** regression coefficients $\hat{\Lambda}^* = (\hat{\Lambda}^*_1, \ldots, \hat{\Lambda}^*_p) \in \mathbb{R}^{p \times p}$ are given for $k, j \in \{1, \ldots, p\}$ by:

   $$\hat{\Lambda}^*_{kj} = \hat{\Lambda}_{kj} \cdot \frac{s_{Z_k}}{s_{X_j}} = \hat{\Lambda}_{kj} \cdot 1 \quad \text{where}$$
   $$\hat{\Lambda} = (Z'Z)^{-1}Z'X = IZ'X = Q\tilde{\Delta}Q'$$

   Since Z is orthogonal $(\hat{\Lambda}^*_{kj})^2 = (r_{Z_k,X_j})^2$ represents the proportion in variance of X_j accounted for by Z_k. By the symmetry of $\hat{\Lambda}^*$ it follows that $(\hat{\Lambda}^*_{kj})^2$ equivalently represents the proportion in variance of X_j accounted for by Z_k. Furthermore since Z is a linear transformation of X we get that $R^2_{X_j|Z} = \sum_{k=1}^{p}(r_{Z_k,X_j})^2 = \sum_{k=1}^{p}(\hat{\Lambda}^*_{kj})^2 = 1$.

3. In order to estimate the variable importance of X_j with respect to Y one can now multiply the proportion in variance of X_j accounted fo by each Z_k : $((\hat{\Lambda}^*_{1j})^2, \ldots, (\hat{\Lambda}^*_{pj})^2) = ((\hat{\Lambda}^*_{j1})^2, \ldots, (\hat{\Lambda}^*_{jp})^2)$ with the proportion in variance of Y accounted for by each Z_k i.e. for $j \in \{1, \ldots, p\}$:

   $$Varimp(X_j, Y) \approx \epsilon_j := \sum_{k=1}^{p} (\hat{\Lambda}^*_{jk})^2 \cdot (\hat{\beta}_k)^2$$

   Finally the relative weights with respect to X and Y are given by $\epsilon \in \mathbb{R}^p$.

It is worth mentioning that ϵ sums up to the coefficient of determination of the initially

[2] By $R^2_{Y|Z}$ we denote the coefficient of determination resulting when regressing Y on Z

defined regression problem i.e. $R^2_{Y|X}$. This holds since

$$\begin{aligned} R^2_{Y|X} &= \mathcal{R}^2 = R'_{XY}(R_{XX})^{-1}R_{XY} = Y'X(Q\tilde{\Delta}\tilde{\Delta}Q')^{-1}X'Y = \\ &= Y'X(Q\tilde{\Delta}Q'Q\tilde{\Delta}Q')^{-1}X'Y = (Y'XQ\tilde{\Delta}^{-1}Q')\underbrace{(Q\tilde{\Delta}^{-1}Q'X'Y)}_{:=w} = w'w = \sum_{i=1}^{p} w_i^2 \end{aligned}$$

and the fact that w equals the standardized regression coefficients $\hat{\beta}$

$$w = Q\tilde{\Delta}^{-1}Q'X'Y = Q\tilde{\Delta}^{-1}Q'Q\tilde{\Delta}\tilde{P}'Y = Q\tilde{P}'Y = \hat{\beta},$$

by using the fact that the column sums of $\hat{\Lambda}^*$ are equal to 1:

$$\sum_{j=1}^{p} \epsilon_j = \sum_{k=1}^{p} (\hat{\beta}_k)^2 \underbrace{\sum_{j=1}^{p} (\hat{\Lambda}^*_{jk})^2}_{=1} = R^2_{Y|X}.$$

Thus we can express each relative weight ϵ_j as percentage of predictable variance accounted for by each X_j when dividing by their sum.

4.2.2 Relative weights in logistic regression - a heuristic approach

Applying the above described *relative weights* method in logistic regression requires modifications of the $\hat{\beta}$ coefficients. Instead of using standardized linear regression coefficients $\hat{\beta}$, one requires standardized **logistic** regression coefficients denoted by $\hat{\beta}_{log}$.

Let $Y \in \{0,1\}^n$ denote the binary response variable of interest and $\hat{b}_{log} \in \mathbb{R}^p$ denote the MLE when modeling $\mathbb{E}[Y|Z]$ as $f^{-1}(Z \cdot b_{log})$ where f denotes the in section 4.1 defined (logit) link function. The model can be equivalently written for $i \in \{1,\dots,n\}$ as

$$\text{logit}(Y_i) := f(\mathbb{E}[Y_i|z_i]) = z_i' b_{log} \tag{4.4}$$

In order to standardize $\hat{b}_{log}$ one basically follows the same approach as in the linear regression when considering equation (4.4). Nevertheless since the empirical standard deviation of $\text{logit}(Y_i)$ cannot be computed from data, because $\text{logit}(Y_i) \in \{-\infty, +\infty\}$, we have to estimate $s_{\text{logit}(Y)}$ differently. To obtain a standardized logistic regression coefficient the following suggestion was made by Menard (2004):

Let $R^2_{Y,\hat{Y}}$ be the coefficient of determination obtained when performing standard linear regression of Y onto the predicted values $\hat{Y}$. Then one can estimate $s_{\text{logit}(Y)}$ by $\dfrac{s_{\text{logit}(\hat{Y})}}{R_{Y,\hat{Y}}}$. This is motivated by the fact that in linear regression it holds that $R_{Y,\hat{Y}} = \dfrac{s_{\hat{Y}}}{s_Y}$.

Finally Menard (2004) proposed the following expression for standardized logistic regression coefficients:

$$\hat{\beta}_{log}(i) = \frac{\hat{b}_{log}(i) \cdot s_{Z_i} \cdot R_{Y,\hat{Y}}}{s_{\text{logit}(\hat{Y})}} \quad i \in \{1,\dots,p\} \tag{4.5}$$

Thus one can calculate relative weights, similar to those in the linear regression setting, by performing the following algorithm:

1. Calculate the orthogonal predictor matrix Z, regress X on Z and obtain $\hat{\Lambda}^*$ (see section 4.2.1).

2. Perform a logistic regression of Y on Z and obtain standardized logistic regression coefficients $\hat{\beta}_{log}(i)$ from equation (4.5).

3. Calculate $\epsilon_j := \sum_{k=1}^{p} (\hat{\Lambda}^*_{jk})^2 \cdot (\hat{\beta}_{log}(k))^2 \quad j \in \{1, \dots, p\}$.

The resulting ϵ represents the contribution of each of the original X_j predictors in terms of predicting the categorical criterion Y.

4.3 Simulation study

In this section we will test the *relative weights* method for logistic regression on a simulated data set and compare it to the **varImp()** function of the R **caret** package (Kuhn et al. (2018-03-29)), which computes the absolute value of the *z-statistic* of the MLE coefficients. Sample code for calculating the relative weights in R is provided at the end of this section.

We used two simulated data sets each consisting of $N = 1000$ data points:

1. The first data set was generated in the following way:

$$(X_1, \dots, X_{12})' \sim \mathcal{N}(0, \Sigma) \qquad (\Sigma)_{i,j} := \begin{cases} 1 & , j = i \\ 0.9 & , i \neq j \leq 4 \\ -0.9 & , i \neq j \geq 9 \\ 0 & , \text{else} \end{cases}$$

$$(\beta_1, \dots, \beta_{12})' = (5, 5, 2, 0, -5, -5, -2, 0, 0, 0, 10, 0)'$$

$$y_i \text{ drawn from } Y_i \sim \text{Ber}\left(\left\{\frac{\exp(\beta' \cdot x_i)}{1 + \exp(\beta' \cdot x_i)}\right\}\right), \qquad i \leq N.$$

In summary we constructed

predictor	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
β_i	5	5	2	0	-5	-5	-2	0	0	0	10	0

where the green and the blue variables were selected to be positively correlated with $r = 0.9$ respectively negatively correlated with $r = -0.9$.

2. The second data set was generated similar as the one above, except that we introduced a pair of correlated regression without influence on the response variable i.e.

predictor	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
β_i	5	5	2	0	-5	-5	-2	0	0	0	10	0

where the green and the blue variables were selected to be positively correlated with $r = 0.9$ respectively negatively correlated with $r = -0.9$ and the red variables were selected to be positively correlated with $r = 0.99$.

4.3.1 Results - simulated data 1

The results of our empirical study for the first simulated data set are presented in figure 4.1. Ranks are shown at the top of each bar. We can see that the *relative weights* method strongly favors correlated variables that have zero or less influence on the target variable: $\{X_3, X_4, X_{12}\}$. Especially X_{12} is ranked second of all the variables. This is due to the high correlation with the most influential variable X_{11} and the definition of variable importance, as the proportion in explained variance, in this section. Since by the *Cauchy Schwarz inequality* and the fact that X is standardized we can conclude that

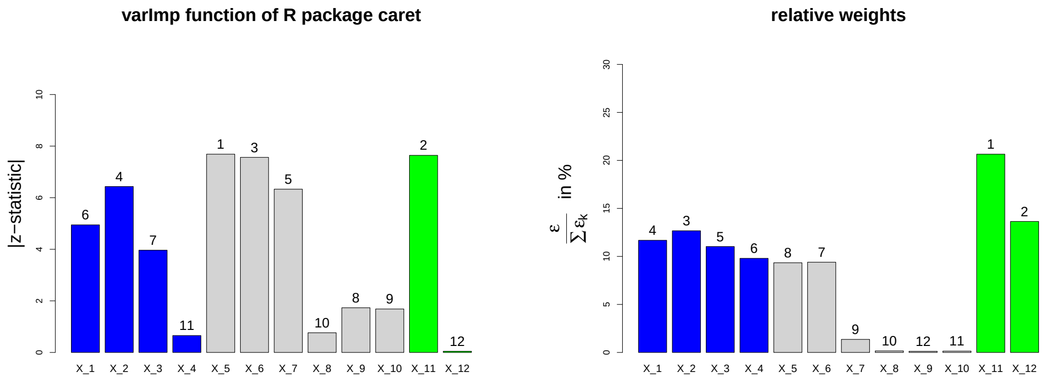


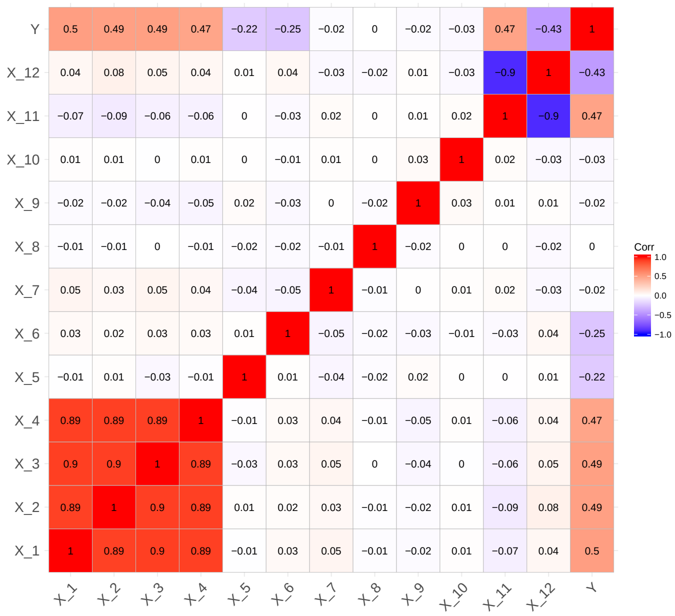
Figure 4.1: Comparison of relative weights to varImp function. Simulated data 1.

$$|Cov(X_{11}, Y) - Cov(X_{12}, Y)| = |Cov(X_{12}, Y)| = |\mathbb{E}[(X_{11} - X_{12})Y]| \leq \sqrt{\mathbb{E}[(X_{11} - X_{12})^2]\mathbb{E}[Y^2]} =$$
$$= \sqrt{2(1 - Cor(X_{11}, X_{12}))} \cdot \sqrt{\mathbb{E}[Y^2]}$$

If $Cor(X_{11}, X_{12})) \approx 1$ it follows that $Cov(X_{11}, Y) \approx Cov(X_{12}, Y)$ and subsequently $Cor(X_{11}, Y) \approx Cor(X_{12}, Y)$. The definition of variable importance and the fact that ϵ_{11} is large implies that it is also very likely to obtain a large ϵ_{12}, although it has zero influence in the true underlying model. Two variables which are highly correlated with each other and with the response variable may have very different regression coefficients or *z-statistics*. Nevertheless as we defined *relative weights* here, variables such as these should have very similar relative weights, because they are very similar to each other and predict the dependent variable about equally.

The correlation structure of the simulated data set 1 is shown in figure 4.2. The first row represents what is often called the “direct effect” of the explanatory variables with the response i.e. the (zero order) correlation $r_{X_j,Y}$ $\forall j \in \{1, \dots, 12\}$.

The left plot of figure 4.1 shows the obtained values of the **varImp** function. Here we can see that the absolute value of the *z-statistic* does better reflect the true model w.r.t to each individual



	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_10	X_11	X_12	Y
Y	0.5	0.49	0.49	0.47	−0.22	−0.25	−0.02	0	−0.02	−0.03	0.47	−0.43	1
X_12	0.04	0.08	0.05	0.04	0.01	0.04	−0.03	−0.02	0.01	−0.03	−0.9	1	−0.43
X_11	−0.07	−0.09	−0.06	−0.06	0	−0.03	0.02	0	0.01	0.02	1	−0.9	0.47
X_10	0.01	0.01	0	0.01	0	−0.01	0.01	0	0.03	1	0.02	−0.03	−0.03
X_9	−0.02	−0.02	−0.04	−0.05	0.02	−0.03	0	−0.02	1	0.03	0.01	0.01	−0.02
X_8	−0.01	−0.01	0	−0.01	−0.02	−0.02	−0.01	1	−0.02	0	0	−0.02	0
X_7	0.05	0.03	0.05	0.04	−0.04	−0.05	1	−0.01	0	0.01	0.02	−0.03	−0.02
X_6	0.03	0.02	0.03	0.03	0.01	1	−0.05	−0.02	−0.03	−0.01	−0.03	0.04	−0.25
X_5	−0.01	0.01	−0.03	−0.01	1	0.01	−0.04	−0.02	0.02	0	0	0.01	−0.22
X_4	0.89	0.89	0.89	1	−0.01	0.03	0.04	−0.01	−0.05	0.01	−0.06	0.04	0.47
X_3	0.9	0.9	1	0.89	−0.03	0.03	0.05	0	−0.04	0	−0.06	0.05	0.49
X_2	0.89	1	0.9	0.89	0.01	0.02	0.03	−0.01	−0.02	0.01	−0.09	0.08	0.49
X_1	1	0.89	0.9	0.89	−0.01	0.03	0.05	−0.01	−0.02	0.01	−0.07	0.04	0.5

Figure 4.2: Correlation structure of simulated data set 1.

influence on Y.

4.3.2 Results - simulated data 2

The results of our empirical study for the second simulated data set are shown in figure 4.3. Ranks are presented at the top of each bar. From 4.3 we can confirm that correlation between regressors, where none of them has an influence on the response variable, does not lead to large relative weights e.g. $\{X_9, X_{10}\}$. The results obtained from the **varImp** function are similar for both data sets.

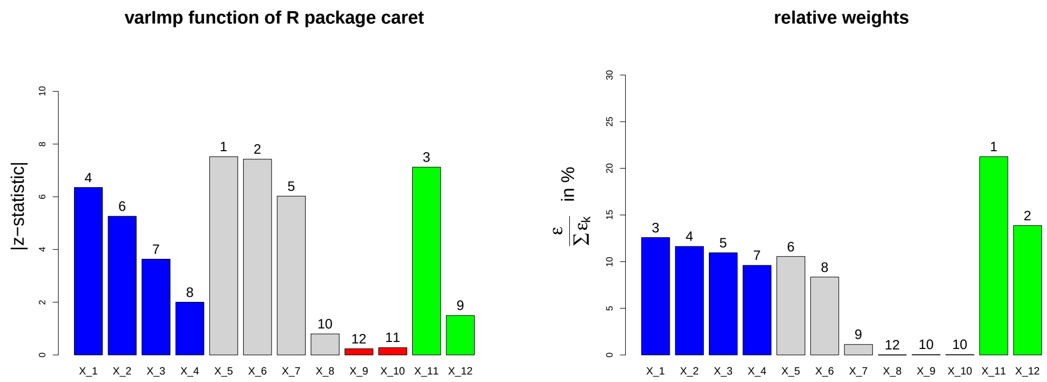


Figure 4.3: Comparison of relative weights to varImp function. Simulated data 2.

4.3.3 Source code

This section provides a possible implementation of the above discussed relative weights method in R.

```r
#relativeWeights() function
#------------------------------------------------------------------
#D ... Data frame, where the target variable column must be named ’Y’ and of type factor.

relativeWeights <- function(D){
  #response must be named Y
  Y <- D[,grepl(’Y’,names(D),fixed=TRUE)]
  X <- D[,!grepl(’Y’,names(D),fixed=TRUE)]
  X <- scale(X) #standardize data
  #REDUCED singular value decomposition
  X.svd <- svd(X)
  Q <- X.svd$v
  P <- X.svd$u
  #orthogonal regressors with minimal squared distance to X
  Z <- tcrossprod(P,Q)
  #regress X on Z
  Lambda <- crossprod(Z,X)
  sdZ <- apply(Z,2,sd) #standard deviation of columns of Z
  Lambda_stand <- Lambda*rep(sdZ,each=nrow(Lambda)) #standardized regression coefficients
  #logistic regression of Y on Z
  logrfit <- glm(Y˜0+Z,family=binomial(link=’logit’))
  b <- coef(logrfit)
  #standardize logistic regression coefficient (Mennard 2004)
  Yhat <- predict(logrfit,newdata=D,type="response")
  logit_Yhat <- log(Yhat/(1-Yhat))
  sqR <- cor(Yhat,as.numeric(Y)-1) #square root of R-squared
  beta <- b*sdZ*(sqR/sd(logit_Yhat))
  #calculate relative weights
  epsilon <- (Lambda_standˆ2)%*%(betaˆ2)
  PropWeights <- (epsilon/sum(epsilon))
  rank <- rank(-epsilon)
  return(data.frame(’variables’=names(D)[1:(ncol(D)-1)],’epsilon’=epsilon, ’prop_epsilon’=round(PropWeights*100,2),’rank’=rank))
}
```

Chapter 5

Application - credit scoring data

In this section we will discuss one possible application of variable importance measures on a real world data set. We will focus on the permutation variable importance measure for missing data (see section 3.1.4). This is due to the fact that it is the most efficient when comparing computational costs as well as that it can handle missing data points without the need of some imputation methods. The goal of this empirical study will be to determine the "most important" covariates that drive the default event.

5.1 Data description

We used a credit scoring data set, where the response variable $Y \in \{0, 1\}$ is given as a default indicator. As explanatory variables financial and market information on firm level are used together. Accounting statements, which are updated quarterly, are obtained from the S&P Capital IQ's Compustat North America database. Given the accounting statements 39 ratios, which measure interest coverage, liquidity, capital structure, profitability and the efficiency of the firm, were computed. For computing market variables, monthly and daily stock and index prices from the Center of Research in Security Prices (CPRS) were used. A detailed description of the variables used can be found in A.1.

Issuer credit ratings from the big three credit rating agencies S&P, Moody's and Fitch are used for the analysis. S&P ratings are collected from the S&P Capital IQ's Compustat North America Ratings file. The ratings from Moody's and Fitch are provided by the credit rating agencies themselves.

For the default and failure information a binary default indicator was constructed[1].

[1] A default is defined as any filing for bankruptcy under Chapter 7 (Liquidation) or Chapter 11 (Reorganization) of the United States Bankruptcy Code that occurred in the one year window following the rating observation. The default indicator was set to one if either from **Moody's Default & Recovery Database** or from the **UCLA-LoPucki Bankruptcy Research Database** a default was recored according to the above definition

Issuer credit ratings were augmented in the following way.

SPR& Fitch		Moodys	
AAA	AAA	Aaa	Aaa
AA+	AA	Aa1	Aa
AA		Aa2	
AA-		Aa3	
A+	A	A1	A
A		A2	
A-		A3	
BBB+	BBB	Baa1	Baa
BBB		Baa2	
BBB-		Baa3	
BB+	BB	Ba1	Ba
BB		Ba2	
BB-		Ba3	
B+	B	B1	B
B		B2	
B-		B3	
CCC+	CCC/C	Caa1	Caa
CCC		Caa2	
CCC-		Caa3	
CC			
C		Ca	Ca

Table 5.1: Aggregation scheme of ratings in the credit scoring data set.

5.2 Results

After data cleaning and preprocessing the final dimension of the data set has been reduced to 21 397 observations of 58 variables including the response default indicator (see A.1 for a full list of the variables used). Additionally we added three missing indicators for the different ratings in order to investigate if a missing rating does have a effect w.r.t predictability of a default event.

Block correlation structures of the data are presented in figures 5.1, 5.2 and 5.3, where the Spearman's Rho rank correlation coefficient was used for the ordered factors of the rating variables.

The ratings of the different CRAs are naturally highly correlated. Nevertheless due to many missing values of the variables **Fitch** and **Moodys** in comparison to the variable **SPR** we expect them to have a lower permutation importance, as was outlined by Hapfelmeier et al. (2014).

	Moodys	Fitch	SPR
% of NA's	36.8	78.4	4.7

Table 5.2: Percentage of missing values in the rating variables.

The pair (**RSIZE**,**MKTEQ**) exhibits the largest correlation of the market variables. This is due to the fact that **RSIZE** is just a monotone transformation of **MKTEQ** (see A.1). We can also identify some groups of variables of the accounting ratios that are highly correlated e.g. (**R2**, **R3**), (**R17M**, **R11M**), (**R11**, **R12**, **R17**, **R18**) or (**IAT**,**ISALE**).

In the following we will fit four different regression random forests using the R **party** package (Hothorn et al. (2017-12-12)) each consisting of $ntree = 100$ conditional inference trees. The random selection parameter *mtry* was chosen to be an element of $\{3, 7, 12, 20\}$. For all other parameters default settings were used.

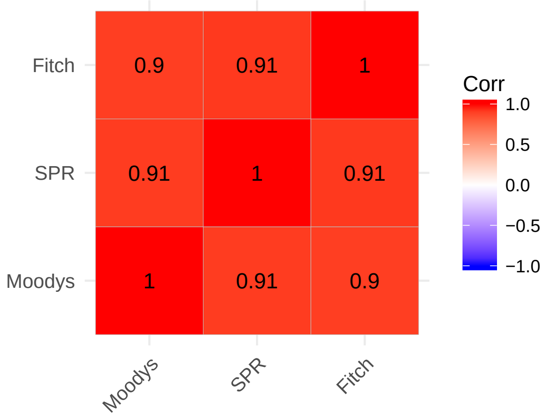


Figure 5.1: Block correlation structure of rating variables.

In figure 5.4 the OOB performance of the fitted forests is measured with the *Gini index* $\in [0, 1]$ and the *Lift measure*. Details of those performance measures can be found in Appendix B. These measures were calculated using the R package **ROCR** (Sing et al. (2005)). The straight red line corresponds to a random model and the green curve denotes the maximal possible value of the lift. The x-axis shows the percentage of positive predictions i.e. we sort the predicted probabilities of a default in decreasing order and predict for the top x-percent a default event. One can see from figure 5.4 that there is hardly any difference with respect to the parameter *mtry* when measuring the OOB performance.

In a next step we calculated the permutation variable importance with missing data, that has been discussed in detail in section 3.1.4. For that we used the **varimp()** function with the following parameters:

varimp(. . ., nperm=10, pre1.0_0 = FALSE).

To cater for possible dependence on random variation the permutation importance scores were calculated ten times and then averaged (*nperm=10*). In table 5.3 and figure 5.5 results of the calculated variable importance are shown (on the y-axis the increase in MSE is measured).

First of all one can observe that the absolute value of the scores (average increase in the MSE) obtained from the permutation importance are rather small. This is clear since default events are

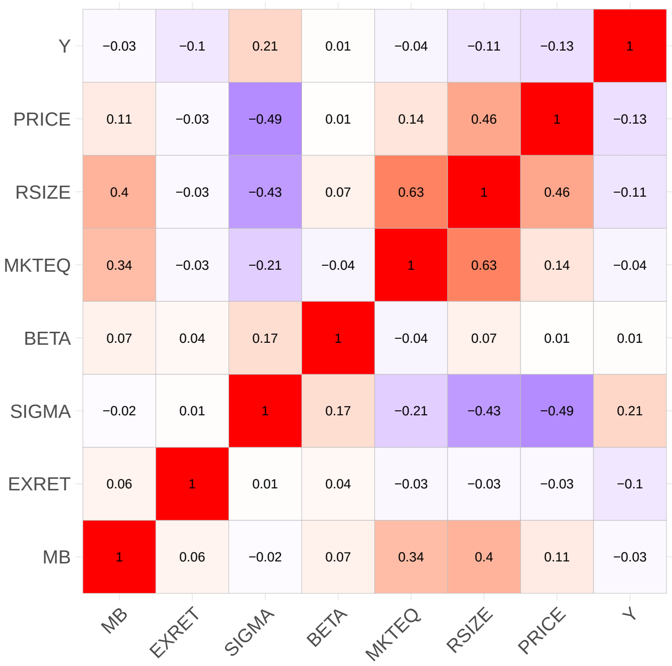


Figure 5.2: Block correlation structure of market variables.

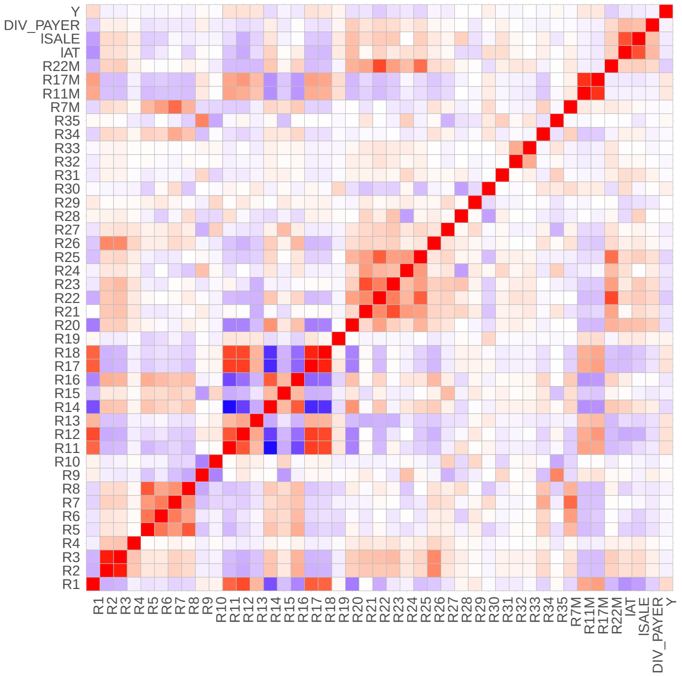


Figure 5.3: Block correlation structure of accounting ratios.

rare in the observed data and a regression random forest predicts the mean of a terminal leaf, which is then also rather small. Additionally we used the MSE as error function which results in small permutation importance values. Nevertheless we don't care what the values are per se rather than the relative predictive strengths of the features. Instead of interpreting the raw score one should better stick to the obtained ranks of the scores.

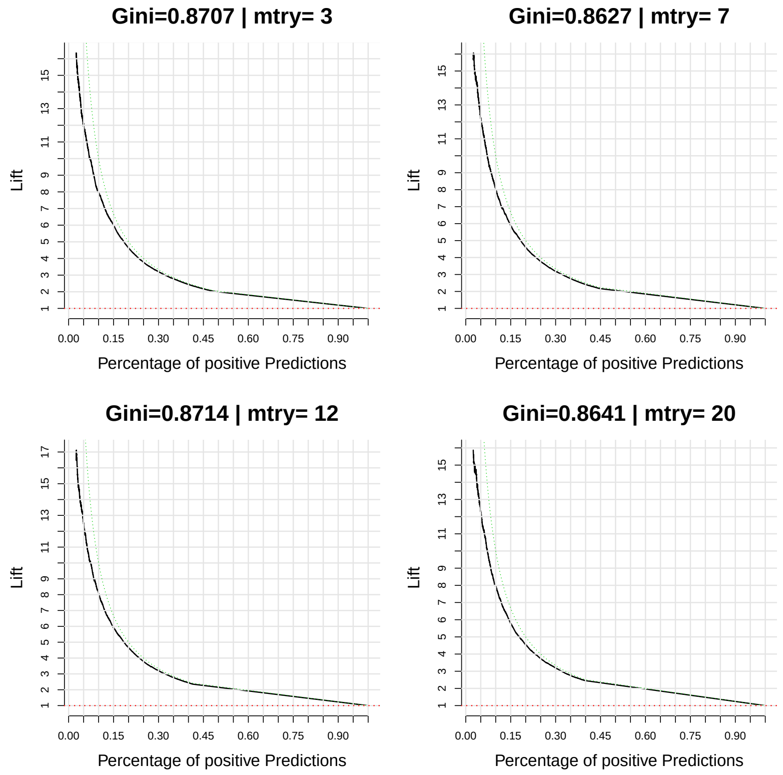


Figure 5.4: Performance measures of fitted random forests.

The scores for the rating variables are as expected. The more missing values they have the less important they are. Here we see that **SPR** is the most important variable for the models with mtry $\in \{7, 12, 20\}$. This is plausible since the rating naturally is a strong predictor of a default event (i.e. it is rather unlikely that a company with a top rating defaults one year after the rating assignment).

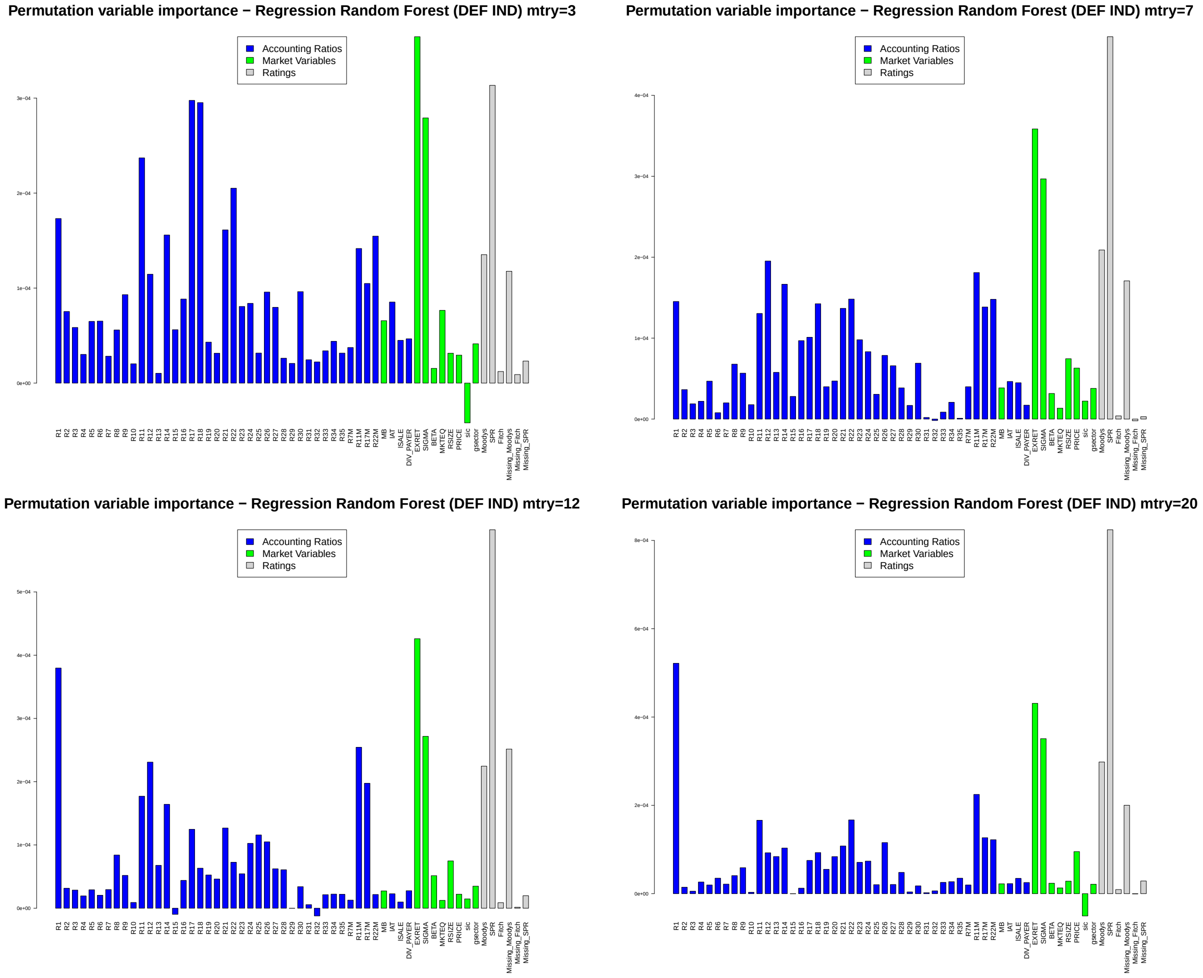
Furthermore one can see from table 3.1 that **EXRET** is ranked always under the top three variables. This variable denotes the average excess return of the company over the S&P 500 in the past three months. One possible interpretation is that companies close to financial distress do have a rather small return on equity. Also **SIGMA**, the standard deviation of the stock returns over the past three month plays an important role and is ranked under the top 5 variables in all models.

From the group of accounting ratios it turns out that with increasing *mtry* parameter the variable **R1** seems to be the most important. This variable denotes the interest paid on assets. Usually the higher the average cost of refinancing for a firm is the more risk is assessed to the company. This could be an explanation for the importance score of this variable.

Finally one can also see that the missing indicator of the rating from Moody's is three times ranked under the top 10 variables suggesting that a missing rating observation from the data does influence a default event.

	mtry= 3	mtry= 7	mtry= 12	mtry= 20
EXRET	1	2	2	3
Missing_Moodys	-	7	6	7
Moodys	-	4	8	5
R1	8	-	3	2
R11	6	-	10	9
R11M	-	6	5	6
R12	-	5	7	-
R14	10	8	-	-
R17	3	-	-	-
R17M	-	-	9	10
R18	4	-	-	-
R21	9	-	-	-
R22	7	9	-	8
R22M	-	10	-	-
SIGMA	5	3	4	4
SPR	2	1	1	1

Table 5.3: Ten most important covariates per model.



Permutation variable importance − Regression Random Forest (DEF IND) mtry=3

Permutation variable importance − Regression Random Forest (DEF IND) mtry=7

Permutation variable importance − Regression Random Forest (DEF IND) mtry=12

Permutation variable importance − Regression Random Forest (DEF IND) mtry=20

Figure 5.5: Results: permutation variable importance with missing data.

Appendix A

Credit scoring dataset

A.1 Data description

The data source as well as the following table has been provided by Laura Vana and Rainer Hirk, two PhD. students of Professor Kurt Hornik at the Vienna University of Economics and Business.

Table A.1: **Collection of accounting ratios.** The table contains information for the accounting ratios used in the context of credit risk. Ratios with codes in bold were found relevant for explaining credit risk in at least one of the studies listed under the Source column. Entry *other* in the Source column refers to expert opinions or usage in industry.

Category	Code	Ratio	Formula	Source
interest coverage	**R1**	Interest rate paid on assets	XINT/AT	other
	R2	Interest coverage ratio (I)	EBITDA/XINT	Altman and Sabato (2007); Baghai et al (2014); Puccia et al (2013)
	R3	Interest coverage ratio (II)	(EBIT+XINT)/XINT	Alp (2013); Altman and Sabato (2007); Puccia et al (2013)
	R4	Free operating cash-flow coverage ratio	(OANCF − CAPX + XINT)/ XINT	Hunter et al (2014); Puccia et al (2013)
liquidity	**R5**	Current ratio	ACT/LCT	Beaver (1966); Ohlson (1980); Zmijewski (1984)
	R6	Cash to current liabilities	CH/LCT	Tian et al (2015)
	R7	Cash&equivalents to assets	CHE/AT	Tian et al (2015)
	R7M	Cash&equivalents to market assets	CHE/(MKTVAL + LT + MIB)	Tian et al (2015)
	R8	Working capital ratio	WCAP/AT	Altman (1968); Altman and Sabato (2007); Beaver (1966); Ohlson (1980)
	R9	Net property plant and equipment to assets	PPENT/AT	Alp (2013); Baghai et al (2014)
	R10	Intangibles to assets	INTAN/AT	Altman and Sabato (2007)
capital structure/ leverage	**R11**	Liabilities to assets (I)	LT/AT	Altman and Sabato (2007); Campbell et al (2008); Ohlson (1980)
	R11M	Liabilities to market assets	LT/(MKTVAL + LT + MIB)	Tian et al (2015)
	R12	Debt ratio (I)	(DLC + DLTT)/AT	Baghai et al (2014); Beaver (1966); Zmijewski (1984)

	R13	Debt to EBITDA	(DLC + DLTT)/EBITDA×(EBITDA> 0)	Puccia et al (2013)
	R14	Equity ratio	SEQ/AT	Min and Lee (2005)
	R15	Equity to net fixed assets	SEQ/PPENT	Min and Lee (2005)
	R16	Equity to liabilities	SEQ/LT	Altman and Sabato (2007)
	R17	Debt to capital (I)	(DLC + DLTT)/(SEQ + DLC + DLTT)	Hunter et al (2014); Puccia et al (2013); Tennant et al (2007)
	R17M	Debt to capital market	(DLC + DLTT)/(MKTEQ + DLC + DLTT)	Hunter et al (2014); Puccia et al (2013); Tennant et al (2007)
	R18	Long-term debt to long-term capital	DLTT/(DLTT + SEQ)	Puccia et al (2013)
	R19	Short term debt to common equity	DLC / (SEQ - PSTK)	Altman and Sabato (2007)
profitability	**R20**	Retained earnings to assets	RE/AT	Alp (2013); Altman (1968); Altman and Sabato (2007)
	R21	EBITDA to assets	EBITDA/AT	Altman and Sabato (2007)
	R22	Return on assets	NI/AT	Altman and Sabato (2007); Campbell et al (2008); Zmijewski (1984)
	R22M	Return on market assets	NI/(MKTEQ + LT + MIB)	Campbell et al (2008); Tian et al (2015)
	R23	Return on capital	EBIT/(SEQ + DLC + DLTT)	Puccia et al (2013), variant in Ohlson (1980)
	R24	EBIT margin	EBITDA/SALE	Altman and Sabato (2007); Baghai et al (2014); Puccia et al (2013)
	R25	Net profit margin	NI/SALE	Altman and Sabato (2007)
cash-flow	**R26**	Operating cash-flow to debt	OANCF/(DLC + DLTT)	Beaver (1966); Hunter et al (2014); Puccia et al (2013); Tennant et al (2007)
	R27	Capital expenditure ratio	OANCF/CAPX	Puccia et al (2013); Tennant et al (2007)
efficiency	**R28**	Asset turnover	SALE/AT	Altman (1968); Altman and Sabato (2007); Beaver (1966)
	R29	Accounts payable turnover	SALE/AP	Altman and Sabato (2007)
	R30	Current liabilities to sales	LCT/SALE	Tian et al (2015)
	R31	Employee productivity	SALE/EMP	other
growth	R32	Inventories growth	$(\text{INVT}_t - \text{INVT}_{t-1})/\text{INVT}_t$	Tian et al (2015)
	R33	Sales growth	$(\text{SALE}_t - \text{SALE}_{t-1})/\text{SALE}_t$	other
	R34	R&D	XRD/AT	ALP
	R35	CAPEX to assets	CAPX/AT	Alp
	lSALE	log sales	log(SALE)	Campbell et al (2008); Tian et al (2015)
	lAT	log assets	log(AT)	Campbell et al (2008); Tian et al (2015)
	DIV_PAYER	dividend payer or not	(DVT > 0)	Alp (2013)
market	**MKTEQ**	Market equity	PRC * SHROUT	Campbell et al (2008); Tian et al (2015)

	MB	Market to book ratio	MKTEQ/(SEQ + 0.1(MKTEQ-SEQ))	Campbell et al (2008); Tian et al (2015)
	SIGMA	volatility systematic risk	regression sd	Campbell et al (2008); Tian et al (2015)
	BETA	idiosyncratic risk	regression beta1	Campbell et al (2008); Tian et al (2015)
	RSIZE	size relative to total cap of an index	log(MKTEQ/TOTAL CAPITALIZATION)	Campbell et al (2008); Tian et al (2015)
	PRICE	average stock price during the year	log(min(PRC, 15))	Campbell et al (2008); Tian et al (2015)
	EXRET	average excess return over index		Campbell et al (2008); Tian et al (2015)
other	**SIC**	Standard Industrial Classification		
	GSECTOR	Global Industry Classification Standard		
	MOODYS	augmented rating		
	SPR	augmented rating		
	FITCH	augmented rating		

A.1.1 Details to ratio computation

- First compute the ratio as numerator/denominator.
- If the denominator is ≤ 0.001 (i.e., 1000$) set the ratio equal to zero.

References

Acharya V, Davydenko SA, Strebulaev IA (2012) Cash holdings and credit risk. Review of Financial Studies 25(12):3572–3604

Agarwal V, Taffler R (2008) Comparing the performance of market-based and accounting-based bankruptcy prediction models. Journal of Banking & Finance 32(8):1541–1551

Alp A (2013) Structural shifts in credit rating standards. The Journal of Finance 68(6):2435–2470

Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23(4):589–609

Altman EI, Sabato G (2007) Modelling credit risk for SMEs: Evidence from the US market. Abacus 43(3):332–357

Baghai RP, Servaes H, Tamayo A (2014) Have rating agencies become more conservative? implications for capital structure and debt pricing. The Journal of Finance 69(5):1961–2005

Balcaen S, Ooghe H (2006) 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. The British Accounting Review 38(1):63–93

Bauer J, Agarwal V (2014) Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test. Journal of Banking & Finance 40:432–442

Beaver WH (1966) Financial ratios as predictors of failure. Journal of Accounting Research 4:71–111

Blume ME, Lim F, MacKinlay AC (1998) The declining credit quality of us corporate debt: Myth or reality? Journal of Finance 53(4):1389–1413

Bongaerts D, Cremers KJM, Goetzmann WN (2012) Tiebreaker: Certification and multiple credit ratings. The Journal of Finance 67(1):113–152

Campbell JY, Hilscher J, Szilagyi J (2008) In search of distress risk. The Journal of Finance 63(6):2899–2939

Deakin EB (1972) A discriminant analysis of predictors of business failure. Journal of Accounting Research 10(1):167–179

Duffie D, Lando D (2001) Term structures of credit spreads with incomplete accounting information. Econometrica pp 633–664

Edmister R (1972) An empirical test of financial ratio analysis for small business failure prediction. Journal of Financial and Quantitative Analysis 7(2):1477–1493

Eklund J, Karlsson S (2007) Forecast combination and model averaging using predictive measures. Econometric Reviews 26(2–4):329–363

Fernandez C, Ley E, Steel MF (2001) Benchmark priors for Bayesian model averaging. Journal of Econometrics 100(2):381–427

González-Aguado C, Moral-Benito E (2013) Determinants of corporate default: a bma approach. Applied Economics Letters 20(6):511–514

Grün B, Hofmarcher P, Hornik K, Leitner C, Pichler S (2013) Deriving consensus ratings of the big three rating agencies. Journal of Credit Risk 9(1):75–98

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: A tutorial. Statistical Science 14(4):382–401

Hunter R, Dunning M, Simonton M, Kastholm D, Steel A (2014) Corporate rating methodology. including short-term ratings and parent and subsidiary linkage. Tech. rep., Fitch Ratings

Jackson CH, Sharples LD, Thompson SG (2010) Structural and parameter uncertainty in Bayesian cost-effectiveness models. Journal of the Royal Statistical Society C 59(2):233–253

Jarrow RA, Turnbull SM (1995) Pricing derivatives on financial securities subject to credit risk. Journal of Finance 50:53–53

Johnson SA (2003) Debt maturity and the effects of growth opportunities and liquidity risk on leverage. Review of Financial Studies 16(1):209–236, DOI 10.1093/rfs/16.1.0209, URL http://rfs.oxfordjournals.org/content/16/1/209.abstract

Johnstone D (2007) Discussion of Altman and Sabato. Abacus 43(3):358–362

Jones S, Hensher DA (2007) Modelling corporate failure: A multinomial nested logit analysis for unordered outcomes. The British Accounting Review 39(1):89–107

Ley E, Steel MF (2007) Jointness in Bayesian variable selection with applications to growth regression. Journal of Macroeconomics 29(3):476–493, Special Issue on the Empirics of Growth Nonlinearities

Ley E, Steel MF (2009) On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. Journal of Applied Econometrics 24(4):651–674

Löffler G (2013) Can rating agencies look through the cycle? Review of Quantitative Finance and Accounting 40(4):623–646

Maltritz D, Molchanov A (2013) Analyzing determinants of bond yield spreads with Bayesian model averaging. Journal of Banking & Finance 37(12):5275–5284

McNeil AJ, Wendin JP (2007) Bayesian inference for generalized linear mixed models of portfolio credit risk. Journal of Empirical Finance 14(2):131–149

Merton RC (1974) On the pricing of corporate debt: The risk structure of interest rates. The Journal of Finance 29(2):449–470

Min JH, Lee YC (2005) Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications 28(4):603–614

Ohlson JA (1980) Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 18(1):pp. 109–131

Pesaran MH, Schleicher C, Zaffaroni P (2009) Model averaging in risk management with an application to futures markets. Journal of Empirical Finance 16(2):280–305, DOI http://dx.doi.org/10.1016/j.jempfin.2008.08.001, URL http://www.sciencedirect.com/science/article/pii/S0927539808000583

Puccia M, Collett LA, Kernan P, Palmer AD, Mettrick MS, Deslondes G (2013) Request for comment: Corporate criteria. Tech. rep., Standard and Poor's Rating Services

Rubin DB (1981) The Bayesian bootstrap. The Annals of Statistics 9(1):130–134

Shumway T (2001) Forecasting bankruptcy more accurately: A simple hazard model. The Journal of Business 74(1):101–124

Tamari M (1966) Financial ratios as a means of forecasting bankruptcy. Management International Review 6(4):15–21

Tennant J, Metz A, Cantor R (2007) Moody's financial metrics: Key ratios by rating and industry for global non-financial corporations. Tech. rep., Moody's Investors Service

Tian S, Yu Y, Guo H (2015) Variable selection and corporate bankruptcy forecasts. Journal of Banking & Finance 52:89–100

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological) pp 267–288

Trustorff JH, Konrad PM, Leker J (2011) Credit risk prediction using support vector machines. Review of Quantitative Finance and Accounting 36(4):565–581, DOI 10.1007/s11156-010-0190-3, URL http://dx.doi.org/10.1007/s11156-010-0190-3

Vassalou M, Xing Y (2004) Default risk in equity returns. The Journal of Finance 59(2):831–868

Zellner A (1986) Bayesian estimation and prediction using asymmetric loss functions. Journal of the American Statistical Association 81(394):446–451

Zmijewski ME (1984) Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research 22:59–82

Appendix B

Performance and goodness of fit measures

B.1 Gini index

The Gini index $G(\mathcal{M}, \mathcal{H}) \in [-1, 1]$ is a measure to quantify the performance of a classification model $\mathcal{M}$, which is usually calculated on a hold out sample $\mathcal{H}$. We will restrict ourselves to a binary response variable $Y \in \{0, 1\}$. One possible way to define the Gini index is given as:

$$G(\mathcal{M}, \mathcal{H}) := 2AUC(\mathcal{M}, \mathcal{H}) - 1,$$

where $AUC(\mathcal{M}, \mathcal{H}) \in [0, 1]$ is the area under the ROC (receiver operating characteristics) curve. Thus the Gini index is just a re parametrization of the AUC measure, that ensures values in the interval $[-1, 1]$. The ROC curve is a graphical plot that illustrates the diagnostic ability of a classification model $\mathcal{M}(\theta)$ as its discrimination threshold $\theta \in [0, 1]$ is varied. Formally one can define the ROC curve as follows:

Definition B.1 (ROC curve) *Let $\theta \in [0, 1]$ be the cutoff for a positive prediction (i.e. prediction of class 1) and $\mathcal{M}(\theta)$ the corresponding binary classifier. Furthermore we denote by $TPR(\theta)$ and $FPR(\theta)$ the true respectively false positive rate when using the model $\mathcal{M}(\theta)$. Then the ROC curve is defined as*

$$ROC := \{(FPR(\theta), TPR(\theta)) : \theta \in [0, 1]\}.$$

	AUC	G
Best possible Model	1	1
Random Model	0.5	0
Worse than Random	< 0.5	< 0

Table B.1: AUC and Gini index.

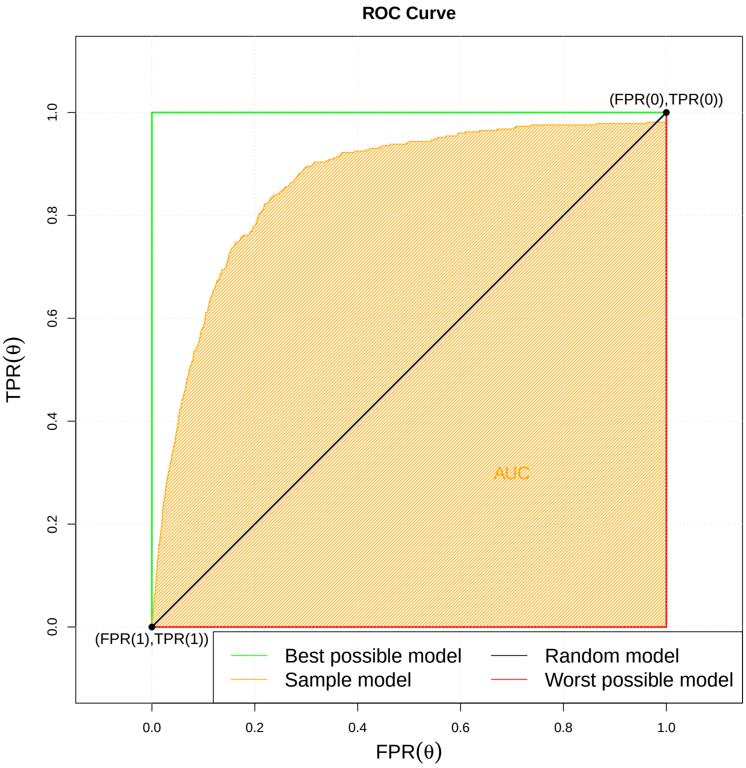


Figure B.1: ROC curve.

B.2 Lift

Another popular performance measure for classification models is the *Lift*. We again assume a binary response variable $Y \in \{0, 1\}$. Furthermore let $\mathcal{M}(\theta)$ be the model predicting the probability p_i of Y_i being of class 1. The *Lift* $L(\theta)$ measures the performance of $\mathcal{M}(\theta)$ by

1. sorting the true labels $(Y_{p_{(1)}}, \ldots, Y_{p_{(n)}})$ by the corresponding predictions $(p_{(1)}, \ldots, p_{(n)})$ in decreasing order.

2. For a given cutoff $\theta \in [0, 1]$ calculate the mean response of the predictions that are larger than θ i.e. calculate
$$m(\theta) := \frac{1}{|\{i \leq n,\, \theta \leq p_{(i)}\}|} \sum_{i \leq n, \theta \leq p_{(i)}} Y_i.$$

3. Finally the Lift is defined as the ratio of the mean response in the set of top predictions to the mean response in the whole dataset i.e.

$$L(\theta) := \frac{m(\theta)}{\frac{1}{n}\sum_{i=1}^{n} Y_i}.$$

Instead of a cutoff θ one could analogously calculate the *Lift* given a desired rate of positive predictions (*rpp*) i.e. how many predictions should be classified as the positive class in total. The definition of the *Lift* can then be also made based on the *rpp*, since there is a one to one relation of θ and *rpp*.

B.3 Coefficient of determination

In a multiple linear regression setting, the *coefficient of determination*, denoted by $R^2_{Y|X}$[1], is the proportion in the variance of the dependent variable Y that is predictable from the independent variables X. It was introduced in Definition 1.1. We will first show that the *coefficient of determination* equals the squared multiple coefficient of correlation $\mathcal{R}^2$, that was defined in Definition 1.2.

We will denote by $\tilde{X}$ and $\tilde{Y}$ the variables which were transformed from X and Y to the standard score form i.e. centered and column length equals 1 (see (4.2)).
Since $R^2_{Y|X}$ is invariant under centering and rescaling of variables we get

$$R^2_{Y|X} = R^2_{\tilde{Y}|\tilde{X}} = 1 - \frac{||\tilde{Y} - \hat{\tilde{Y}}||_2^2}{||\tilde{Y} - \overline{\tilde{Y}}||_2^2} = ||\tilde{Y} - \hat{\tilde{Y}}||_2^2 \overset{*}{=}$$

using furthermore the fact that

$$\tilde{Y} - \hat{\tilde{Y}} = \underbrace{(I - \tilde{X}(\tilde{X}'\tilde{X}))^{-1}\tilde{X}')}_{:=Q_{\tilde{X}}}\tilde{Y} = Q_{\tilde{X}}\tilde{Y},$$

we get that

$$\overset{*}{=} 1 - \tilde{Y}'Q'_{\tilde{X}}Q_{\tilde{X}}\tilde{Y} = 1 - Y'Q_{\tilde{X}}Q_{\tilde{X}}\tilde{Y} = 1 - \tilde{Y}'Q_{\tilde{X}}\tilde{Y} = \tilde{Y}'(I - Q_{\tilde{X}})\tilde{Y} = \tilde{Y}'(\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}')\tilde{Y} \overset{**}{=}$$

where we used that $Q_{\tilde{X}}$ is the orthogonal projection on the orthogonal complement of the column space of $\tilde{X}$ and therefore symmetric and idempotent. Finally using that the correlation coefficients can be calculated, when transforming to standard score form, as $R_{\tilde{X}\tilde{X}} = \tilde{X}'\tilde{X}$ and $R_{\tilde{Y}\tilde{X}} = \tilde{Y}'\tilde{X}$ we can conclude that

$$\overset{**}{=} R'_{\tilde{Y}\tilde{X}}(R_{\tilde{X}\tilde{X}})^{-1}R_{\tilde{Y}\tilde{X}} = R'_{YX}(R_{XX})^{-1}R_{YX} = \mathcal{R}^2.$$

Finally we will give a proof why the *Pratt measure* defined in 1.1.1 yields a additive decomposition of the *coefficient of determination*. For that we will use the fact that the standardized regression coefficient of the original regression of Y on X denoted by $\hat{\beta}_{standardized}$ and the non standardized coefficient of the regression of $\tilde{Y}$ on $\tilde{X}$ are equal when transforming the data to standard score form. Thus we can conclude that

$$R^2_{Y|X} = R^2_{\tilde{Y}|\tilde{X}} = \underbrace{(\tilde{Y}'\tilde{X}(\tilde{X}'\tilde{X})^{-1})}_{\hat{\beta}_{standardized}}(\tilde{X}'\tilde{Y}) = \hat{\beta}'_{standardized}R_{\tilde{X}\tilde{Y}} =$$

$$= \hat{\beta}'_{standardized}R_{XY} = \sum_{i=1}^{p}\hat{\beta}_{i,standardized} \cdot r_{X_i,Y}$$

[1] We will denote by $R^2_{Y|X}$ the coefficient of determination resulting from regressing Y on X.

Bibliography

Christopher H Achen. *Interpreting and using regression*, volume 29. Sage, 1982.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Barry E Feldman. Relative importance and value. *Manuscript version 1.1*, 2005. URL http://www.prismaanalytics.com/docs/RelativeImportance.pdf.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Ulrike Grömping. Relative importance for linear regression in r : the package relaimpo. *Journal of statistical software*, 17(1):1–27, 2006.

Ulrike Grömping. Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147, 2007.

Alexander Hapfelmeier, Torsten Hothorn, Kurt Ulm, and Carolin Strobl. A new variable importance measure for random forests with missing data. *Statistics and Computing*, 24(1):21–34, 2014.

T Hothorn, H Seibold, and A Zeileis. partykit: A toolkit for recursive partytioning. r package version 1.2-0. 2017-12-13. URL https://CRAN.R-project.org/package=partykit.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.

Torsten Hothorn, Kurt Hornik, Carolin Strobl, and Achim Zeileis. Party: A laboratory for recursive partytioning. r package version 1.2-4. 2017-12-12. URL https://CRAN.R-project.org/package=party.

Jeff W Johnson and James M LeBreton. History and use of relative importance indices in organizational research. *Organizational Research Methods*, 7(3):238–257, 2004.

Jefir W Johnson. Determining the relative importance of predictors in multiple regression: Practical applications of relative weights. *Child Development*, 59:969–992, 2001.

Richard M Johnson. The minimal transformation to orthonormality. *Psychometrika*, 31(1):61–66, 1966.

Max Kuhn et al. Caret package: R package version 6.0-79. 2018-03-29. URL https://CRAN.R-project.org/package=caret.

Andy Liaw, Matthew Wiener, et al. Breiman and cutler's random forests for classification and regression. r package version 4.6-12. 2015. URL https://www.stat.berkeley.edu/~breiman/RandomForests/.

Scott Menard. Six approaches to calculating standardized logistic regression coefficients. *The American Statistician*, 58(3):218–223, 2004.

Stephan J Ritter, Nicholas P Jewell, Alan E Hubbard, et al. R package multipim: a causal inference approach to variable importance analysis. *Journal of Statistical Software*, 57(1):1–29, 2014.

Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941, 2005.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007. URL http://rocr.bioinf.mpi-sb.mpg.de.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307, 2008.

Scott Tonidandel and James M LeBreton. Determining the relative importance of predictors in logistic regression: An extension of relative weight analysis. *Organizational Research Methods*, 13(4):767–781, 2010.

Mark J Van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

List of Figures

List of Tables