DEGREE PROJECT IN THE FIELD OF TECHNOLOGY
ENGINEERING PHYSICS AND THE MAIN FIELD OF STUDY
COMPUTER SCIENCE AND ENGINEERING, SECOND CYCLE, 30 CREDITS
STOCKHOLM, *SWEDEN 2016*

Computing Random Forests Variable Importance Measures (VIM) on Mixed Continuous and Categorical Data

ADAM HJERPE

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF COMPUTER SCIENCE AND COMMUNICATION

KTH Computer Science
and Communication

Computing Random Forests Variable Importance Measures (VIM) on Mixed Numerical and Categorical Data

Beräkning av Random Forests variable importance measures (VIM) på kategoriska och numeriska prediktorvariabler

ADAM HJERPE
hjerpe.adam@gmail.com

Master's Thesis at CSC
Project performed at Handelsbanken Capital markets
Supervisor at Handelsbanken: Fredrik Bohlin
Supervisor at CSC: Hedvig Kjellström
Examiner: Danica Kragic

Abstract

The Random Forest model is commonly used as a predictor function and the model have been proven useful in a variety of applications. Their popularity stems from the combination of providing high prediction accuracy, their ability to model high dimensional complex data, and their applicability under predictor correlations. This report investigates the random forest variable importance measure (VIM) as a means to find a ranking of important variables. The robustness of the VIM under imputation of categorical noise, and the capability to differentiate informative predictors from non-informative variables is investigated. The selection of variables may improve robustness of the predictor, improve the prediction accuracy, reduce computational time, and may serve as a exploratory data analysis tool. In addition the partial dependency plot obtained from the random forest model is examined as a means to find underlying relations in a non-linear simulation study.

Random Forest (RF) är en populär prediktormodell som visat goda resultat vid en stor uppsättning applikationsstudier. Modellen ger hög prediktionsprecision, har förmåga att modellera komplex högdimensionell data och modellen har vidare visat goda resultat vid interkorrelerade prediktorvariabler. Detta projekt undersöker ett mått, variabel importance measure (VIM) erhållna från RF modellen, för att beräkna graden av association mellan prediktorvariabler och målvariabeln. Projektet undersöker känsligheten hos VIM vid kvalitativt prediktorbrus och undersöker VIMs förmåga att differentiera prediktiva variabler från variabler som endast, med aveende på målvariableln, beskriver brus. Att differentiera prediktiva variabler vid övervakad inlärning kan användas till att öka robustheten hos klassificerare, öka prediktionsprecisionen, reducera data dimensionalitet och VIM kan användas som ett verktyg för att utforska relationer mellan prediktorvariabler och målvariablel.

Contents

Chapter 1

Introduction

This thesis presents the random forest model as a means to identify informative variables in a supervised learning setting. That is we have data $\mathcal{D} = \{X^i, Y^i\}_{i=1}^n$ and we believe that there is a mapping f which maps the input-or predictor variables X^i onto the target variable Y^i with reasonable overall accuracy, and we want to identify a subset of variables $\{X_k\}_{k=1}^p$ which are truly informative to the target variable Y.

A variable importance measure may be used to find a ranking of predicting variables indicating operational risk exposure[1]. Let $\mathcal{D}$ be labeled data where $Y^i = 1$ denotes a large deviation between the market and settled prices for a trade, and $Y^i = 0$ denotes that the price difference lies within a predefined threshold. Examples of predicting variables $X^i = (X_1, .., X_p)$ are counter-party, instrument-type, changes-to-old-trades, and expert defined indicators which are computed over historical data. More generally the setting is such that the target variable $Y^i = 1$ denotes exposure to operational risk and $Y^i = 0$ denotes normal exposure, and where the predicting variables $X^i = (X_1, ..., X_p)$ are predefined indicators. Assessing an adequate measure of importance over the indicators may facilitate to identify a subset of indicators that seems likely to relate to risk exposure. This identified subset of variables is useful since it may reduce the amount of monitored data to only informative variables, and could be used to direct effort onto specific parts of the overall business process e.g. increasing the quality, resolution, and the amount of stored data that is informative with respect to risk exposure, or to improve operation procedures.

When choosing an adequate predictor model there is a trade off between flexibility and interpretability. Highly interpretable models such as the logistic regression model provides clear relations between the learned coefficients and the output of the model. However for some data the model is not sufficiently flexible and may

[1] Operational risk is defined as the risk of change in value caused by the fact that actual losses, incurred for inadequate or failed internal processes in a business.

not be used to distinguish important from non-important variables.

By increasing the flexibility of the chosen predictor model there is often a sacrifice in intepretability. The Random forest model is a highly flexible model which often provides high prediction accuracy, and the model is capable of modeling non-linear and complex interactions. Most importantly the model provides an importance measure of variables which may be used to identify informative variables. Further since the random forest model to a high extent adapts with respect to training data the corresponding importance measures are less biased with respect to a specified underlying model $y \sim f(x)$, e.g. compared with the linear-or logistic regression model which may simply learn a best fit under an occasionally too strict model. In addition random forests are applicable in both classification and regression settings.

1.1 Problem statement

The objective of this thesis is to find a variable importance measure over the predictor variables in a supervised learning setting, where the input data consists of both categorical and numerical predictor variables. An application for the importance measure is to differentiate informative from non-informative predictor variables in an internal audit setting. The investigated data, data which is yet not available, will consist of price deviations between settled and associated market prices, $Y_{\texttt{settled}} - Y_{\texttt{market}}$. The associated predictor variables are e.g. counter-party, instrument-type, changes-to-old-trades, valuation-method, trader, portfolio, trade-frequency and currency. The objective is to differentiate the predictor variables that correspond to noise from those variables that seems informative with respect to the price differences. The obtained measure will be used to identify what variable or variables, supposedly caused the price deviations. E.g. are the deviations caused by an internal error such as fraudulent behavior, or caused by an interaction of variables such as trade-frequency, trader and occurring in specific portfolios. Another idea where the variable importance measures can be proven useful is within anti-money laundering (AML) [54], where the laundering-examples used to train the model could e.g. be created by domain experts. An approach to tackle AML is by considering it a subfield of anomaly detection, and where the objective is to identify patterns which distinguish from normal behavior [63, 42].
The broad idea of the thesis project is to find a method that is capable of identifying informative predicting variables with respect to an observed outcome, and where the obtained variable importance measures will function as a screening tool used to explore relations in the data. Further the data considered for screening may to a great extent vary with respect to the nature of the variables as well as with respect to numerical-and categorical ranges. To tackle the lack of business data this thesis investigates the validity of the random forest variable importance measures on simulated data where the ground truth is known together with an experiment on real data.

1.2 Contributions

- This report investigates the random forest variable importance measure as a means to identify informative variables. The measures are applied to both regression-and classification settings. The importance measures are proven capable identifying informative variables in simulated data and the importance measures are shown useful in a live data setting.
- The project provides a survey of various variable selection techniques.
- A thorough background regarding tree based models, random forest techniques, and the random forest variable importance is presented.
- A largely automated pipeline of programs which facilitate analyzing parameter settings for the random forest model and selection of informative variables, implemented for the R system [46].

1.3 Outline

The project work is structured as follows,

- Chapter 2 presents a background of feature selection techniques and background regarding supervised learning models as a means to investigate relations between the target and predictor variables. Chapter 2 also presents a background describing tree based models and the random forest as a general technique to differentiate informative predictors from non-informative predictors with respect to the target variable.
- Chapter 3 presents applications of the random forest model and research regarding properties of the random forest variable importance measures.
- Chapter 4 presents the core of the project, a variable selection procedure used to differentiate informative variables from non-informative variables with respect to the target variable.
- In Chapter 5 the variable selection procedure is evaluated on data where the truth underlying the data is known. The robustness with respect to categorical noise is investigated, together with an application to real cencus data. The main results are discussed with respect to finding truly informative variables and with respect to prediction accuracy for the real census data.
- Chapter 6 outline ideas for further research, both by performing additional experiments and by a modification proposal to the investigated variable selection procedure.
- Appendix A provides an introduction to terminology and the field of machine- and supervised learning, together with an example illustrating the idea of variable selection.

1.4 Ethical considerations

In every computer based decision system there is a risk that something goes wrong, when a computer based system is installed questions regarding potential errors and liabilities should be carefully analyzed. To simply trust variable importance measures as a means to learn underlying truths regarding data is not reliable. The investigated variable importance measure must be carefully analyzed by domain experts before any conclusion can be made. Furthermore measures of associations does not imply causation, when inferring causal relationships from data a range of statistical tests must be applied together with a thorough statistical design. However importance measures may serve useful as a means to explore and search for informative variables. The variables may in hand be used to build a more robust predictor, to reduce the computational time, and to serve as a guide for exploratory data analysis techniques. Exploratory analysis techniques which relates observations to predictors may in hand be used to learn observational patterns in a variety of fields such as medicine, physics, psychology, and economy. E.g. learning what variables and to which strength that related to inflation, learning variables that relates to a particular disease, or learning what predictors that relates to the likelihood of observing new planets.

1.5 Relation to my education

This report mainly intersects the three fields Statistics, Computer Science and Mathematics. Course work including Times series analysis, Probability theory, Computer intense methods in mathematical statistics, Algorithms and Complexity, Artificial intelligence, Advanced Machine Learning, and Image recognition and classification provides a solid foundation handling Machine Learning techniques such as the Random Forest model. Further solid programming experience facilitates writing, to a large extent, automated pipe-line of programs which returns variable importance measures and graphs as a means to explore algorithmic convergence and data relations.

Chapter 2

Background

This chapter provides a summary of general variable selection techniques such as wrapper, embedded, and filter methods. Furthermore supervised learning models as a means to identify informative predictors with respect to the target variables are outlined. Lastly related concepts and background concerning the random forest model are presented, such as classification and regression trees, adaptive and non adaptive averaging procedures e.g. bagging and boosting, together with the random forest variable importance measures.

2.1 Feature selection

The field of feature selection studies the problem of finding a small set of predictors in supervised learning problems. Generally there are three main objectives governing variable selection, to improve prediction accuracy, to reduce the time needed for training, and to enhance interpretation of the learned predictor model [21]. Feature selection techniques are divided into three categories called *wrapper*, *embedded*, and *filter* methods [5].

Wrapper methods utilizes the predictor as a black box to score subsets of variables according to their predictive power, examples are forward selection and backward elimination which commonly are used in regression settings [32]. A feature is greedily included or excluded from the model based on the R^2 or adjusted R^2 measure. Forward and backward elimination strategies can be non-robust, meaning that a small change in the input data can result in very different models, the scapegoat could be that each predictor is discretely adjoined or excluded from the model.

Embedded methods perform variable selection in the process of training and are specific to the given classifier. An example facilitating the above non-robustness in linear regression settings is the *least absolute shrinkage and selection operator* (LASSO) [52] which combines continuously shrinking some of the coefficients towards 0 and setting others coefficients to exactly 0. Another example of an em-

bedded method is the random forest model. During the training process the model partitions the prediction variables into similar regions with respect to the target variable, and the number of times each predictor variable is selected as a partitioning variable can be used as a measure of variable importance.

Filter methods selects variables as a pre-processing step and are independent of the specific model. An example of a univariate filter method is to threshold each predictor variable based on the correlation between the predictor and the target variable e.g. using a mutual information measure [5]. An example of a multivariate filter method is the *Correlation based feature selector* [22] which selects subsets of features that are highly correlated with the target variable, such that the features are uncorrelated with each other.

2.2 Choice of model

There are many methods which provide good prediction results. Research in Neural nets is an ever more active field and the nets provide state-of-the-art prediction accuracy. Despite the good prediction results, the relation between the target and the predicting variables is not easily understood. The variables are transformed by the nets at each layer which hampers interpretation of how the prediction variables relates to the model output.

Another model commonly used in classification settings is the logistic regression model specified by (A.2). There are a few caveats applying the logistic regression model to our problem setting. Firstly the logistic regression model is a parametric model and could have difficulties providing good prediction results when faced with non-linear decision boundaries, an example is illustrated in figure A.2. The non-linearities may be resolved by *feature engineering* i.e. to adjoin the data by various transformation of the predictors. Another approach handling more complex decision boundaries is the non-parametric linear regression model [27]. The model is specified by

$$\text{logit}(p) = \log \frac{p}{1-p} = \sum_i \phi_i(x_i)$$

where ϕ_i are smooth functions of the predictor variables.

Secondly logistic regression has difficulties providing reliable parameter estimates in settings, called small n large p problems, where n denotes the number of training cases and p denotes the number of parameters to be estimated. Empirical research examining this property has indicated that when the number of *events per variable* (EPV) that is if the number of training cases over the number of predicting variables is less than 10, then the estimated coefficients where found to be biased in both negative and positive direction [43]. More recent studies have found that even

when the EPV is greater than 10 the logistic regression model may have insufficient power, which may lead to discarding predicting variable due to lack of significance [11].

A more viable approach to our setting is the *Random Forest* model [7]. Forests requires little parameter tuning and are applicable when the data consists of both numerical and categorical predicting variables. Forests are also applicable modeling non-linear decision without need of feature engineering [26], as indicated by the noisy circle data illustrated in figure A.5. Most importantly Forests have an embedded feature ranking technique called *variable importance measure* (VIM) which can be used as a tool guiding the selection of predictors for the final model. The VI measures computed from the noisy circle data is illustrated in figure A.6, note that the two variables X_1 and X_2 are the only informative predictors with respect to the target variable.
There is yet no uniform framework of what constitutes an important variable and the notion varies with respect to application. When two or more predictor variables interact work by Kohavi and John [35] have shown that subtle changes to the definition of what constitutes an informative variable implies surprisingly large changes concerning which variables are defined as relevant. The authors concludes that in practice one should look for features with respect to the specific learning algorithm and training data. Hence the fact that random forests are capable to model complex non-linear relations, as illustrated in figure A.1, implies that the associated importance measure can capture a broad set of relations.

Criticism concerning the superiority of Forests techniques used for variable selection is presented in the survey by Verikas et al. [57]. The paper shows that the non-parametric model k-Nerest-Neighboors (k-NN) provides-comparable and sometimes better results with respect to prediction accuracy by the reduced model. The k-NN model can easily model non-linear problems and there is only one tuning parameter. However the algorithm may have difficulties using categorical predictor variables since there is no general method of how to assign distances between different categories.

Lastly an interesting approach is the *Multivariate adaptive regression splines* (MARS) model. The MARS model handles data consisting of both numerical and categorical predictor variables in a natural way, the model is computationally scalable, and the model is also interpretable [26]. MARS models are however not commonly used for variable selection.

2.3 Ensemble methods and Random Forests

Trees are invariant under strictly monotone transformation of the individual predictors, they are robust to predictor outliers, and trees often obtain good prediction

results without the need of extensive parameter tuning. When faced with new data a reasonable initial step is to explore if there is a partitioning over the predictor variables such that the values of the associated target variable are as similar as possible. Decision trees works by searching for precisely such groups, the algorithm aims to partition the predictor space into high dimensional rectangles with the objective that the values of the target variable are as similar as possible.
The most common algorithms for creating decision trees are the CART (classification and regression trees) and the C4.5 algorithm developed by Breiman et al. [8] and independently introduced by Quinlan [45]. CART models are freely accessed from both the Python and the R programming languages. CART are built by the principle of recursive partitioning, more concretely in a regression setting, the trees are built by

1. The predictor space $X_1, .., X_p$ is divided into J distinct non-overlapping regions $R_1, .., R_J$

2. For every observations that falls into region R_j the associated response is predicted as the mean response over all training observations contained in region R_j

The regions R_j are chosen as high dimensional rectangles for computational simplicity and to facilitate interpretation. The goal is to find rectangles R_j that minimizes the residual sum of squares (RSS)

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{2.1}$$

where $\hat{y}_{R_j}$ is the mean response over all training observations within the jth hyperrectangle. Finding an optimal partitioning is however generally computationally intractable [31] and the splitting selection is instead chosen by a greedy heuristic called *recursive binary splitting*. For each variable X_j the feature space is binary partitioned into two regions

$$R_1(j, s) = \{X | X_j < s\} \quad \text{and} \quad R_2(j, s) = \{X | X_j \geq s\} \tag{2.2}$$

where the pair (j, s) is chosen such that the RSS error is minimized over the two regions. This selection process is then recursively repeated over all resulting regions until a stopping criteria is met, e.g. the resulting terminal-node contains fewer than n observations. If the tree is grown deep the predictor could easily overfit, and on the contrary if the tree is grown too shallow the predictor may fail to capture important relations in the data. A common strategy to tackle this trade off, more thoroughly explained in [26], works as follows. First a large tree T_0 is grown until a minimum node size is met. This large tree is then pruned using *cost complexity pruning*. We define a subtree $T \subset T_0$ as any tree that can be obtained by collapsing

non-terminal nodes of T_0. Let all regions R_m containing terminal nodes be indexed by m and let $|T|$ denote the number of terminal nodes in T. Further we define

$$\begin{aligned} N_m &= |\{x_i \in R_m\}|, \\ \hat{y}_m &= \frac{1}{N_m} \sum_{x_i \in R_m} y_i, \\ Q_m(T) &= \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{y}_m)^2, \end{aligned} \tag{2.3}$$

and the cost complexity criterion is defined as

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|. \tag{2.4}$$

For each α one searches for the subtree $T_\alpha \subseteq T_0$ that minimizes the objective balancing the tree complexity and the accuracy of the predictions, i.e. the subtree T_α that minimizes $C_\alpha(T)$. It can be shown that for every α there exists a unique smallest subtree T_α minimizing $C_\alpha(T)$, and the optimal trees T_α are found using *weakest link pruning*. The internal nodes that provides the smallest per-node increase in $\sum_m N_m Q_m(T)$ are sequentially collapsed until a single-node tree is obtained. The papers by Breiman et al. [8] and Ripley [47] show that the latter sequence must contain T_α. The complexity trade-off parameter α is chosen by five- or tenfold cross-validation and $\hat{\alpha}$ is chosen to the value that minimizes the cross-validated sum of squares.

The *impurity measure* $Q_m(T)$ is responsible to guide the partitioning of the predictor variables such that the associated target outcomes, which are similar, are grouped into shared hyper rectangles. In regression settings $Q_m(T)$ is chosen as the RME and must be modified when faced with categorical data. Common measures $Q_m(T)$ of node impurity used in classification settings are

$$\begin{array}{ll} \text{Misclassification error:} & \frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}. \\ \text{Gini index:} & \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}). \\ \text{Cross-entropy or deviance:} & -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \end{array} \tag{2.5}$$

where $\hat{p}_{mk} = \sum_{x_i \in R_m} I(y_i = k)$. The Gini index and the cross-entropy are both small if all of the $\hat{p}_{mk}$ are close to 0 or 1, which implies that the two measures prefer pure nodes. In general any impurity function Φ may be used. An impurity function is defined as a function over any K-tuple of non-negative numbers $(p_1, ..., p_K)$ satisfying $\sum_j p_j = 1$, with the following properties

1. Φ achieves maximum only for the uniform distribution.
2. Φ achieves minimum only at the points where all $p_j = 1,\ j = 1, .., K.$ (2.6)
3. Φ is symmetric with respect to $p_1, .., p_k$.

Furthermore impurity measures $i(\cdot)$ are related to an impurity function Φ by the following relation

$$i(t) = \Phi(p(1|t), .., p(K|t)) \tag{2.7}$$

where $p(j|t)$ is the estimated probability of class j in node t. Breiman et al. [8] showed that the Gini index prefers splits that puts the largest class into a pure node and all others into the other node, whereas the entropy criterion put their emphasis on balancing the sizes of the two children nodes. However in problems with a small number of classes both criteria should produce similar results.

Decision trees have a great advantage in interpretability. The logic of how the predictor space is partitioned and the associated purity of all terminal nodes is easily assessed and examined. However a major problem with decision trees are their high variance. A small change to the input data could alter the series of splits in the building process which in turn changes the interpretation of the overall tree, and which could decrease the prediction accuracy on unseen data.

A general variance reduction technique called *bagging* (bootstrap aggregating) introduced by Breiman [6] and independently found by Ho [28], can help mitigate the high variance for the decision tree predictor. The technique often also substantially increases the prediction accuracy when the base learners are chosen as high variance classifiers. The idea is, when a decision tree is chosen as the base classifier, to bootstrap B training sets from the original training data. For each bootstrap sample a decision tree is grown without pruning and the B classifiers are then aggregated into a single classifier,

$$\hat{f}_{\texttt{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x). \tag{2.8}$$

The rational for growing the trees without pruning is to learn as much structure of the data as possible, the increase in variance is then mitigated by averaging. Since all decision trees are trained on bootstrap samples the means and covariances are equal for all individual trees, and the total variance for the bagged predictor (2.8) is given by

$$\begin{aligned}
\texttt{var}(\hat{f}_{\texttt{bag}}) =& \frac{1}{B^2} E\left[(\sum X_i)^2\right] - E^2\left[\frac{1}{B}\sum X_i\right] = \frac{1}{B^2}\left[\sum_{i,j} X_i X_j - B\mu^2\right] \\
=& \frac{1}{B^2}\left[B(\sigma^2+\mu^2) + B(B-1)\rho - B\mu^2\right] \\
=& \frac{1}{B}\left[\sigma^2 + (B-1)\rho\right]
\end{aligned} \tag{2.9}$$

where $\sigma^2 = \texttt{var}X_i$ and $\rho = \texttt{cov}(X_i, X_j)$. Thus the variance decreases as the number of included classifiers B increase. Another treat provided by the bagged model

(2.8) is a test error estimate which is obtained without the need to use either cross-validation, or to use a validation set. On average the number of observation not contained in a bootstrap sample is approximately one third, referred to as the *out-of-bag* (OOB) samples. For each sample x_i there are approximately $B/3$ trees where this samples is OOB, and the estimate for the test error is obtained by averaging all predictions over those trees and over all samples. This estimate is however biased and can sometimes underestimate the true error by 10% [26]. The average size for the OOB samples over a n-set is explained by noting that the probability that an observation is included in a bootstrap sample equals $1 - (1 - 1/n)^n$. The latter expression quite rapidly converges to $1 - 1/e \approx 0.632$, with $n = 100$ the probability approximately equals 0.63.

The resulting bagged model (2.8) is clearly not as interpretable as a single classification tree due to the averaging effect. However for complex data sets the interpretation of single decision trees should be taken with care due to their inherent instability [6]. Most importantly the bagged decision tree model (2.8) provides a *variable importance measure* (VIM), which also mitigate some of the problems concerning the VIMs obtained from a single decision tree. For a single tree the importance measures could easily be *masked*. E.g. suppose that two variables X_1 and X_2 both provide equal prediction accuracy when either X_1 or X_2 is included in the decision tree. Once one of the variables is included, inclusion of the other variable does not improve prediction accuracy. Suppose further that the gini impurity criterion favors inclusion of the first variable over the inclusion of the second variable. Then by only measuring improvements over variables which partition the data, the equally important variable X_2 would obtain a variable importance measure of 0, albeit either variable could be used with respect to prediction accuracy. The bagged model (2.8) smooths out the instability inherent by the individual decision trees, and the variable importance measure are more reliable compared to the measures obtained from single trees [26].

The overall variance for the bagged model (2.9), as the number of included trees B increases, is bounded by the covariance between individual trees. The Random Forest model proposed by Breiman [7] further decrease the correlation between individual predictors by an additional randomization procedure. The individual classifiers are trained using bootstrap samples, and the additional randomization is imputed by attribute sampling. During training the trees are presented with only a uniformly drawn subset of features permissible as splitting variables. This modification further decreases the correlation between individual trees. The idea to select features at random was independently introduced by Ho [28] and by Amit and Geman [2].

Another *meta-algorithm*, which by [16] also been shown outperform the bagging procedure, is called *boosting*. The procedure sometimes provide higher prediction accuracy than random forests, especially in little or no noise data settings. The

idea is to adaptively grow a sequence of weak learners by a weighting procedure. Examples at the current sequence length that are proven difficult to predict are attached larger weights, and examples where the predictions are more accurate are attached smaller weights. Breiman [7] showed that the error rates obtained by random forests are comparable with those obtained by boosting, and further reports that the Random Forest model is more robust with respect to input noise than the boosting procedure. In classification settings the forests also provides smoother prediction boundaries by aggregating the class predictions provided by the individual trees into a single probability, which in hand may be used as an uncertainty measure [14].

2.4 Variable Importance

For a single CART model Breiman et al. [8] proposed a variable importance measure by using surrogate splits intended to mitigate the risk that the variable importance of a single variable is masked. For an aggregated model such as bagging or boosting the variable importance measure is not as limited by the overall size of the tree, and the number of splitting opportunities is vastly increased which implies that masking is less of a problem. Despite the averaging effect there is still a possibility that a single variable X_2 is not included in the model due to a slightly higher performance obtained when the data instead is partitioned by variable X_1 (say that X_2 is highly correlated with X_1). This in hand implies a tiny importance measure for variable X_2. The random forest model further reduces the likelihood of masking caused by the latter scenario by only allowing a random subset of features available at each split. Thus for all feature sets not including X_1 the likelihood that the correlated variable X_2 is used as a partitioning variable is increased. Since the individual trees in contrary to bagging are grown deep, the contribution in variable importance due to interaction effects is increased. The boosting procedure may on the other hand choose to ignore some variables completely.

The *permutation importance* measure, introduced by Breiman [7], is one of the two most common variable importance measures. To measure the importance for variable X_i the idea is to permute all values of this variable, and the variable importance measure is defined as the difference in prediction accuracy caused by the permutation. If the variable consists of purely random noise the prediction accuracy will likely not be affected by permuting the values of this variable. Formally the variable importance is computed as follows. Let $\mathcal{B}^t$ denote the OOB samples for a tree t and let $L(T_t(\mathbf{x}_i), y_i)$ denote the prediction accuracy at the ith training example. The importance for variable X_j in tree t is defined as

$$VI^{(t)}(X_j) = \sum_{i \in \mathcal{B}^t} L(T_t(\mathbf{x}_i), y_i) - L(T_t(\mathbf{x}_{i,\pi_j}), y_i) \tag{2.10}$$

where $\mathbf{x}_{i,\pi_j} = (\mathbf{x}_{i,1}, .., \mathbf{x}_{\pi_j(i),j}, \mathbf{x}_{i,j+1}.., \mathbf{x}_{i,p})$, and where π_j is a random permutation of n integers. In classification settings the prediction accuracy $L(T_t(\mathbf{x}_i), y_i)$ is defined

as

$$L(T_t(\mathbf{x}_i), y_i) = \frac{\sum_{i \in \mathcal{B}^t} I(\hat{y}_i^t = y_i)}{|\mathcal{B}^t|} \tag{2.11}$$

where $\hat{y}_i^t = T_t(\mathbf{x}_i)$ denotes the prediction at point $\mathbf{x}_i$ by tree t, and $I(\cdot)$ denotes the indicator function. Whereas in regression settings the prediction accuracy $L(\hat{y}, y)$ is defined as the RMS error. The variable importance measure for variable X_j is computed as the sum of the importances over all trees in the forest,

$$VI(X_j) = \frac{\sum_{t \in B} VI^{(t)}(X_j)}{\texttt{ntree}}. \tag{2.12}$$

Another commonly used importance measure is the *Mean Decrease Impurity* (MDI). The importance measure for variable X_j is computed by the sum of all decreases in node impurities where variable X_j is used to partition the data. If we let t_L and t_R denote the two resulting children nodes when partitioning the data at node t, and we let N_t denote the number of examples reaching node t, the decrease in impurity is defined as

$$\Delta i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{2.13}$$

where $i(\cdot)$ is some impurity measure defined by (2.7), and $p_L = N_{t_L}/N_t$ and $p_R = N_{t_R}/N_R$. The resulting children nodes are obtained when the data is partitioned by the parent node at $s = (X_m < c)$. Lastly the MDI measure is defined by averaging over all trees T and all nodes t,

$$VI(X_m) = \frac{1}{\texttt{ntree}} \sum_T \sum_{t \in T : v(s_t) = X_m} p(t) \Delta i(s_t, t) \tag{2.14}$$

where $p(t)$ denotes the proportion N_t/N of samples reaching node t and $v(s_t)$ denotes the variable used to split node t. A common choice of node impurity measure is the gini-coefficient, a combination which commonly is denoted as the *Gini importance*. Whereas in regression settings a common impurity measure is the MSE. The permutation and the gini importances both captures non-linear relationships, as indicated by the noisy-circle data illustrated in figure A.6 for the permutation importance. Furthermore the two measures capture importances for variables which are correlated with informative predictors [51]. Both the permutation importance and MDI measure with the RME and gini impurity measures are freely available in the package `randomForest` by Liaw and Wiener [39] for the R system for statistical computing. The package provides an R interface of the Fortran program originally developed by Breiman and Cutler freely available at `http://www.stat.berkely.edu/users/breiman/`.

In settings with correlated predictors the permutation variable importance measure have been observed to put an increased importance onto variables which are correlated, likely due to the fact that correlated variables are preferred as splitting candidates earlier in the tree Strobl et al. [51]. The authors suggests a modified

variable importance measure following the logic of permutation tests [19]. The idea is to form a null hypothesis designed to investigate whether predictor variables are informative. If the null hypothesis H_0 is specified as a global null hypothesis, i.e all predictor variables are independent of the target variable ($Y \perp X_1, .., X_p$). The null hypothesis implies that the joint distribution then factorizes as

$$P(Y, X_1, .., X_p) = P(Y) \cdot P(X_1, .., X_p). \tag{2.15}$$

If the data is truly generated under the null hypothesis, a permutation of the target variable will not affect the joint distribution (2.15) due to the factorization. On the contrary if the null is false and the target variable is permuted, an observed deviation of the joint distribution or some reasonable test statistic computed from it is to be expected.

Under the global null hypothesis it is expected that the permutation importance measures are distributed as a zero mean random variable. A deviation from the null hypothesis is expected to imply a change in prediction accuracy which in hand implies a deviation to the permutation importances, the deviation in importance measures is chosen as test statistic to indicate deviations to the independence assumption. The null hypothesis which corresponds to the original permutation importance assumes that variable X_j is independent to both Y and to $Z = X_1, ..., X_{j-1}, X_{j+1}, .., X_p$. To mitigate importance deviations arising due to dependency between variables X and Z Strobl et al. suggests a modified permutation scheme where variable X_j is permuted only within groups of observations with $Z = z$.
The conditional permutation importance, given by [51], for variable X_j is computed by the 4 steps,

1. First the OOB-prediction accuracy is computed as in equation (2.11).

$$\frac{\sum_{i \in \mathcal{B}^t} I(\hat{y}_j^t = y_j)}{|\mathcal{B}^t|}. \tag{2.16}$$

2. For each variable Z_i used for conditioning. The cut-points which partitions this variable are bisected into a grid to form a permutation grid.

3. Within the permutation grid the values of X_j are permuted and the associated permutation accuracy is computed

$$\frac{\sum_{i \in \mathcal{B}^t} I(\hat{y}_{i,\pi_j}^t = y_i)}{|\mathcal{B}^t|} \tag{2.17}$$

   where $\hat{y}_{i,\pi_j}^t$ denotes the predictions of the t:th tree $T^t(\mathtt{x}_{i,1}, .., \mathtt{x}_{\pi_j(i),j}, \mathtt{x}_{i,j+1}.., \mathtt{x}_{i,p})$, and where the values of the input variable X_j are permuted with respect to the permutation grid.

4. The contribution to the permutation importance measure for variable X_j by tree t is computed as the difference between the non-permuted prediction accuracy (2.16) and the permuted prediction accuracy (2.17). The permutation importance measure for variable X_j is obtained by summing all importance contributions over all trees, exactly as for the original permutation importance measure (2.12). The conditional and the unconditional permutation importance measures simply differs with with respect to permutation scheme. For the conditional measure the ranges of values which are permuted for each variable are reduced, the size of the range depends on the number of bisection values for the conditioning variables Z_j.

Strobl et al. suggests a strategy used to guide what variables to use as conditioning variables for variable X_i. The conditioning variables Z_j to include are selected as those with empirical correlation, with variable X_i, that exceeds some threshold. A more general approach suggested by Hothorn et al. [30] is to compute p-values of conditional inference test as a measure of association. The latter strategy have the advantage that the same guiding procedure is applied whether the predictors are either categorically or numerical.

Chapter 3

Related Work

The main objective for the thesis project is to provide a technique used to identify informative variables in supervised learning settings. The random forest variable importance measure have been proven useful identifying informative variables in a variety of applications due to its high flexibility. This fact is indeed useful, firstly since the importance measures will be applied to a variety of data and secondly since the relation between the price differences and the predictor variables, specified in the problem statement 1.1, should to a large extent be specified by data. Furthermore the variable importance measure have been proven reliable in settings where the number of examples is less than the number of variables (small n large p), a property which may decrease the required amount of labeled data.

This chapter presents applications by the random forest model, where the presented applications are focused with respect to the nature of the random forest model and with respect to the random forest variable importance measures, together with studies where the relation between the target and predictor variables is known. The latter fact facilitates evaluation of the importance measures.

3.1 Random forest variable importances

The variable importance measure obtained by the Random Forest model is a frequently used measure for feature selection in a variety of fields. Work by Díaz-Uriarte and De Andres [15] investigated random forests used to select a set of informative genes. With respect to prediction accuracy the work showed that the random forest model provide similar performance as the k-NN, SVM and the Diagonal Linear Discriminant Analysis (DLDA). More importantly with respect to finding informative variables the authors showed that the random forest variable importance measures could be used to identify a small number of genes while preserving predictive accuracy. The parameter settings governing forests where shown, by both simulation and by real data, to be quite robust with respect to the parameters `mtry`, `nodesize` and `ntree`. The parameters controls the number of features

available at each partitioning, minimum number of examples reaching a node used as a stopping criteria, and the number of trees used in the forest. The work showed that a large value of `ntree` slightly increased the stability of the variable importance measures. In cases where the ratio, number of informative over total number of variables, is small an increase in `mtry` implied a slight increase in prediction accuracy. Lastly the parameter `nodesize` controlling the minimal size of the terminal nodes was observed to have negligible effects with respect to prediction accuracy. The work further investigates a problem concerning identification of predictors called the *multiplicity* problem, namely that there could exists two or more subsets of predictor variables, and where all subsets provide equal prediction accuracy. Multiplicity is a commonly observed problem in settings where the number of variables p is large compared to the number of observed cases n, and the issue is investigated in the statistical literature [25, 9]. The author states that in small n large p settings the multiplicity problems may be hard to solve. A proposal to tackle this issue may be to use a variety of techniques, and examining if there is a subset of variables which are selected by most of the models.

A simulation study conducted by Archer and Kimes [3] investigates the gini variable importance measure. The predictors were simulated following a multivariate normal distribution with covariance matrix consisting of 20 blocks, where the jth block of variables are correlated with $\rho_j = 0.05j - 0.05$. Within each block there is only one informative covariate and the association with the target variable is specified by

$$y_i = \begin{cases} 1 & \texttt{if} \quad \pi_{i,j} < u_{i,j} \\ 0 & \texttt{else} \end{cases}$$

where $u_{i,j}$ is a uniformly drawn random number and where π is given by

$$\pi_{i,j} = \frac{e^{\beta_l x_{i,j}}}{1 + e^{\beta_l x_{i,j}}} \quad j = 1 + 40k, \; k = 0, 1, .., 19$$

with the parameter β ranging over 7 levels, $\beta \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75\}$. The study reports the number of times the truly informative variable obtained the largest importance measure among all covariates over a total 400 number of simulation trials. This ratio is presented as a function of the 20 levels of correlation and the 7 levels of association between the predictors and the target variable. The random forest model successfully identified the true predictor more than 80% out of the 400 trails under moderate correlations $\rho < 0.5$ and with $\beta > 1$.

Strobl et al. [50] showed by a simulation study that the gini variable importance measure is biased onto categorical variables and observed that the importance measure increases as the number of factors increases. The mechanics underlying this bias likely derives from the gini impurity which have been shown to favors variables with e.g. more categories Kim and Loh [34], White and Liu [58]. Furthermore the gini index have been shown to prefer variables with a large amount of missing values

[49]. Due to these observations Strobl et al. propose a new random forest algorithm that utilizes conditional inference trees as base learners [30] instead of CART trees. The measures obtained by conditional tree based forest were shown more robust than the gini importance measure for categorical non-informative predictors [50].

Later Strobl et al. [51] showed that the permutation VIM is biased with respect to correlated predictor variables. It was shown by a simulation study where the target variable y is generated by 12 predictor variables as follows,

$$y_i = 5x_{i,1} + 5x_{i,2} + 2x_{i,3} - 5x_{i,5} - 5x_{i,6} - 2x_{i,7} + \epsilon_i$$

where ϵ_i is i.i.d. $N(0, 1/2)$ distributed noise and where $X_{i,1:12} \sim N(0, \Sigma)$. The correlation matrix Σ is specified such that all variables have unit variance $\sigma_{i,i} = 1$, the first 4 variables are correlated $\sigma_{i,j} = 0.9$ for $i \neq j \leq 4$, and where all other variable are uncorrelated $\sigma_{i,j} = 0$.
Notably it was shown that the permutation variable importance measure was larger for variable X_3 than the measures for the two variables X_5 and X_6, albeit that their absolute coefficient values are equal to the coefficient for the two most informative variables X_1 and X_2. Due to these observations a new variable importance measure called *conditional variable importance* is proposed. The proposed measure suppress the importances for the correlated variables into the order $v_1 > v_2 > v_5 > v_6 > v_3$, where v_i denotes the importance measure for variable X_i. Albeit the distinction between the importance measures v_3, v_5 and v_6 are quite small and the two importance measures v_1 and v_2 are always larger irrespective of the `mtry` parameter. The variable importance measure is freely available at the R programming package named **party** [29, 50, 51].

Genuer et al. [18] also explore the random forest variables importance measure when the predictors are correlated. The setting is such that 6 out of 200 variables are informative, and the data contains 100 cases. The variable importance measure is studied as the data is adjoined with $1, 10$ and 20 uninformative variables having a correlation of 0.9 with the most informative variable X_3. The study showed that the importance measure for variable X_3 decreases as more correlated variables are adjoined to the data. However the importance for variable 3 is never confused with noise, this also holds for the adjoined correlated predictors.

An exhaustive survey regarding forests in applications is provided in the paper by Verikas et al. [57]. The work reports summaries over multiple research papers ranging over a broad field of applications, categorized as, 17 studies in which the random forest outperformed or performed on the same level as other techniques e.g. customer churn prediction [60, 61, 13, 12], credit card fraud detection [59], Bacterial species identification Slabbinck et al. and automatic e-mail filing into folders [36]. Eleven small scale studies where the random forests is outperformed by other techniques such as spam email detection [1], customer loyalty prediction [10] and network intrusion detection [33, 62]. A total of 14 studies in which the random forest

variable importance measure is exploited e.g. identification of a small number of risk-associated single nucleotide polymorphisms (SNPs) among a large number of unassociated SNPs [41], customer retention and profitability [37], and categorizing cancer cases [53].

Furthermore the work examines the consistency and generality of the random forest variable importance measures, and are compared by selecting features using the k-NN classifier. Both the gini Δ_j and the permutation $\bar{D_j}$ variable importance measure are examined. Features are included into the final model by adding features corresponding to the highest values of the $\bar{D_j}$ measure, and by recursive feature elimination one-by-one based on $\bar{D_j}$. The variable selection by the k-NN classifier are based on forward selection (backward selection provided similar results). The study reports that the features selected by the k-NN model provide a higher prediction accuracy by both the SVM and k-NN classifier.
The paper furthermore reports that the random forest variable importance measure is capable to accurately identify almost all 21 informative variables in the 40-dimensional artificial Waveform data, albeit the measures over 4 of the informative variables are very hard to differentiate from noise. Previous work have shown that correct identification of all noise variables in the Waveform data is a difficult task [56, 55].

Chapter 4

Variable selection procedure

4.1 Variable selection

When searching for important variables one may assume e.g. an additive model and iteratively performing various feature transformations until reasonable prediction accuracy is obtained, or that the initial assumption of additive is deemed unreasonable. The process is illustrated by the prediction accuracy provided by the logistic regression model before and after the data is adjoined by the two features $X_1^{'} = X_1^2$ and $X_2^{'} = X_2^2$ shown in figures A.3 and A.4. When a sound result is obtained by the latter strategy it corresponds to a reasonable fit given the underlying model, e.g. a linear additive model. Another or an accompanied strategy may be to try a less biased model that provides a variable importance measure. The idea is that the variable importance measures obtained by a more flexible model may capture e.g. both linear and non-linear relations in the data. The random forest model is a non-parametric and highly flexible model. A reasonable fit and associated variable importance measures, when the random forest model is trained on the non-linear noisy circle data, is illustrated in figures A.5 and A.6. Variable selection using the random forest model can, following the pseudocode presented in Hapfelmeier and Ulm [24], be schematic summarized as

1. Assess (a) the OOB-error or (b) a cross-validated error of the forest.
2. Compute the importance measures of variables.
3. Reject a fraction of least important variables and refit the forest.
4. Assess (a) the OOB-error or (b) a cross-validated error of the forest.
5. Return to (a) step 2 or (b) step 3 until no further variables can be rejected.
6. Choose the model with (a) the lowest error or (b) the sparsest model with the error within a specified number of standard deviations to the lowest error (e.g. according to the 1 s.e. rule).

Often the preceding steps are based on averaged finding to achieve higher stability. Therefore, steps 1-5 can optionally be repeated separately, in conjunction and

within cross-validated runs.

The chosen method used to differentiate relevant from non-relevant variables, which fits the above general method, follows the strategy presented in the work by Genuer et al. [18]. The procedure consists of the following steps,

1. Compute the RF variable importance measures, excluding the variables with the smallest importances.
2. Order the m remaining variables in decreasing order of importance.
3. Construct, on separate data, a sequence of random forests greedily including the k first variables, for $k = 1, .., m$. The final model is chosen as the most sparse model with smallest OOB error.

The variable importance measures, used in the experiments are chosen as the two most commonly used importance measures. Namely the permutation importance measure specified in equation (2.12) and the mean decrease impurity measure. The impurity function governing the MDI is chosen as the gini impurity specified by (2.5) in classification settings, and chosen as the MSE in regression settings. Both measures are freely obtained by the `randomForest` package [39] in the R system programming language.

To further illustrate the steps governing the variable selection method the procedure when applied to the noisy circle data (A.4) is now outlined. First the permutation VIMs are computed, and the ordered variable importance measures are visually analyzed, illustrated in figure A.6. The 3 variables with the largest importance measures are selected adjoining the third step. Lastly the final model is chosen by levering the number of included variables and the associated OOB errors. The final model is chosen by inclusion of the 3 variables $V1, V2$ and $V18$. The greedy sequence of OOB errors as a function of included variables is shown in figure A.7. Note that the OOB error sequence is illustrated for $k = 1, .., 20$ number of included variables, but there are only 3 variables with non-zero importance measures. The OOB error minimum is obtained with 2 included variables $V1$ and $V2$, which also are the two truly informative variables.

In general any impurity function (2.6) is valid to specify the MDI measure, and the procedure also allows modification to the permutation importance measure by changing permutation schemes e.g. to the conditional permutation scheme (2.16). There are a few notable strengths supporting the above method. Firstly the work by Verikas et al. [57] showed good results identifying more or less all important variables in the 40 dimensional Waveform data simply by inspection of the variable importance measures, albeit there are at least 4 informative variables which cannot be differentiated from the noise variables. Identifying all informative variables in the latter data is considered a difficult task [55, 56]. The results by Verikas et al. indicates that almost all informative variables will sequentially be included-

and-ordered by decreasing importance measure by the greedy strategy at step 3. More importantly it is desired that the OOB error obtained from the random forest is proven capable to differentiate the 4 informative variables which were shown especially hard to separate from noise variables. Furthermore the method is computationally efficient since the importances measure are only calculated once at step 1, and thereafter the OOB errors are used to differentiate informative from non-informative variables. The fact that the OOB error rates provide an estimate for the test error further decreases the computational burden since there is no need to use separate validation data or use of cross-validation.

Chapter 5

Experiments

5.1 Investigated data

Friedman1 data
The data is obtained by the R system package `ml-bench` [38] originating from Friedman [17] and consists of 1000 number of cases. The underlying model is a 10 dimensional regression model specified by

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon \tag{5.1}$$

where $X_1, .., X_{10} \sim U(0, 1)$ and $\epsilon \sim N(0, 1)$. This data is used for two purposes, first we want to examine if variable selection procedure 4.1 is capable to differentiate the truly informative variables from the noise variables in a non-linear setting, and secondly to present the random forest partial dependency plot as a means exploring relationships between the target and predictor variables.

Waveform data
The data is obtained from the UCI machine learning repository [40] and consists of a simulated 40 dimensional classification data set with 5000 examples. Each class is generated from a combination of 2 or 3 base waves. The latter 19 variables are all noise with mean 0 and variance 1. More or less all informative variables have previously been identified using the random forest VIM by Verikas et al. [57], by visual inspection of the importances. However the authors states that there are 4 informative predictors with variable importance measures that are hard to distinguish from importances obtained from noise variables. The goal is to identify all truly informative predictors using the variable selection procedure presented in section 4.1. Furthermore the robustness of the variable importance is investigated by adjoining the Waveform data with categorical noise variables.

Waveform noise data
This data consists of the original waveform data described above adjoined with 7 categorical noise variables. The 7 variables contains $3, 5, 8, 9, 10, 12$ and 15 number

of categories, where each category equally likely. This data is used to assess the robustness of the variable selection procedure 4.1 when the data contains categorical noise.

Adult data
The data is collected by Barry Becker from the 1994 Census data base and is obtained from the UCI machine learning repository [4]. The data consists of 45222 number of cases and is downloaded as a training and test set. The number of training cases equals 30162 and the number of test cases is 15060. The predicting variables consists of (1) age, (2) workclass, (3) fnlwgt, (4) education, (5) education-num, (6) marital-status, (7) occupation, (8) relationship, (9) race, (10) sex, (11) capital-gain, (12) capital-loss, (13) hours-per-week, and (14) native-country.
The purpose of the data is to predict whether income exceeds \$50K/y based on census data.

5.2 Selection of variables

All experiments are conducted using the procedure specified in section 4.1. The first step from the procedure consists of optimizing the two parameters `mtry` and `ntree` over a randomly selected training set $\mathcal{D}_1$ with respect to the OOB error, the calculations are averaged over 40 iterations. Using the optimized parameters together with the associated variable importance measures, the second step consists of greedily including predictors into the final model where the predictors are ordered by decreasing variable importance measure. The final model is chosen as the least complex model, and where the model also obtains the smallest OOB error over randomly chosen training data $\mathcal{D}_2$ (disjoint from $\mathcal{D}_1$). Lastly we evaluate the final model and compare it to a random forest model trained using all predictors. Both models are trained on $\mathcal{D}_1$ and $\mathcal{D}_2$, and evaluated on separate testing data $\mathcal{D}_{\texttt{test}}$.
For all experiments the sizes for the 3 randomly chosen partitions $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_{\texttt{test}}$ are $0.8aN$, $0.8(1-a)N$ and $0.2N$ with $a = 0.6$ and N denoting the number of observations for each data set.

Freidman1 data
The parameter `mtree` and `ntree` are chosen by minimizing the OOB errors over 40 RF models illustrated in figure 5.1. The error rates are quite stable for `ntree` $\in \{500, 1500, 2500\}$, with a slightly smaller error range for `ntree` ≥ 1500. The parameter controlling the number of variables available at each data partitioning `mtry` provides a reasonable error rate at its default value $\lfloor\sqrt{10}\rfloor = 3$, the error decreases by increasing `mtry` > 3, and starts to slightly increase for `mtry` > 6.

By the permutation VIM shown in figure 5.2 it is noted that the VIM over the important variables $V1$ up to $V5$ is clearly distinguished from the VIM over the non-relevant variables $V6$ up to $V10$. The variable importance measures showed

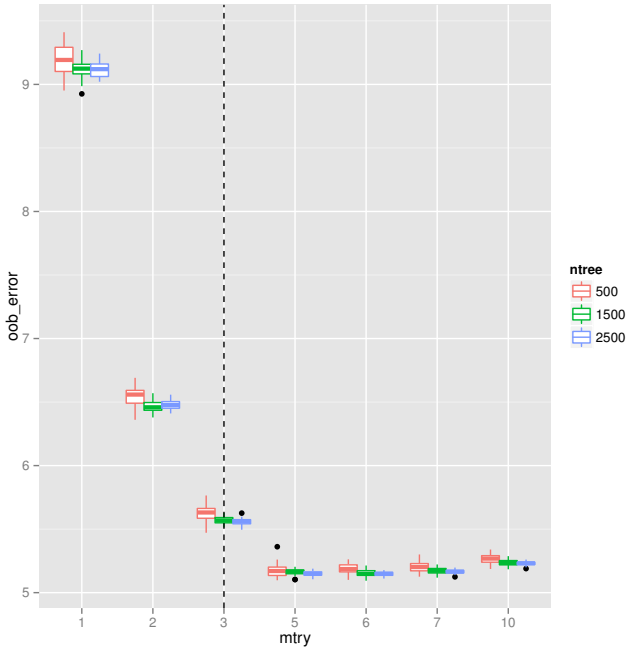

Figure 5.1: *OOB errors computed by* 40 *RFs for the Freidman1 data. The default value for* `mtry` *is marked by a vertical line.*

similar results for all values of `mtry`, results which are not illustrated in the this report. The gini importance measure showed similar results with respect to the relative order for the VIMs, however the measure did not approach zero for the non-important variables, but instead stagnated at a value around 240 compared to a VIM of 490 for variable $V3$, which was the smallest importance measure obtained over the truly relevant predictors.

The important variables are clearly distinguished from the non-important variables by the variable importance measures. However this feature is not always observed e.g. the Waveform data examined in the work by Verikas et al. [57]. To further facilitate the differentiation of informative from non-informative predictors the greedy OBB error sequence is investigated. The error sequence by inclusion of the k most important variables for $k = 1, 2, .., 10$ is illustrated in figure 5.3. The smallest OOB error is obtained by including the 5 most important variables.

The top included predictors $V1$ up to $V5$ suggested by the VIMs in combination by sequence of OOB errors are precisely those predictors that are related to the target variable by the generation process (5.1).

The random forest partial dependency plot is useful to further explore relationships between the target and the predictor variables. The partial dependency plot

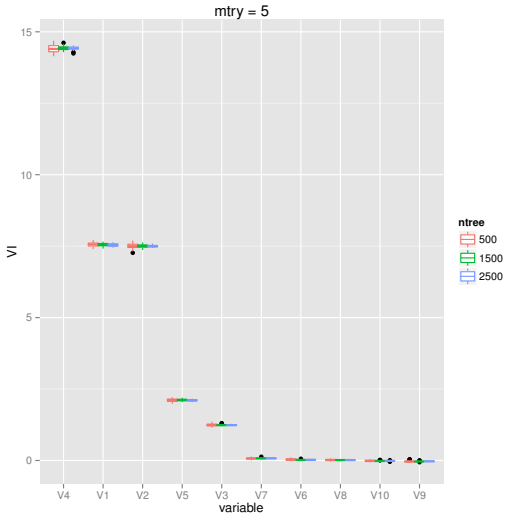

Figure 5.2: *Permutation VIM computed by* 40 *RFs for the Friedman1 data.*

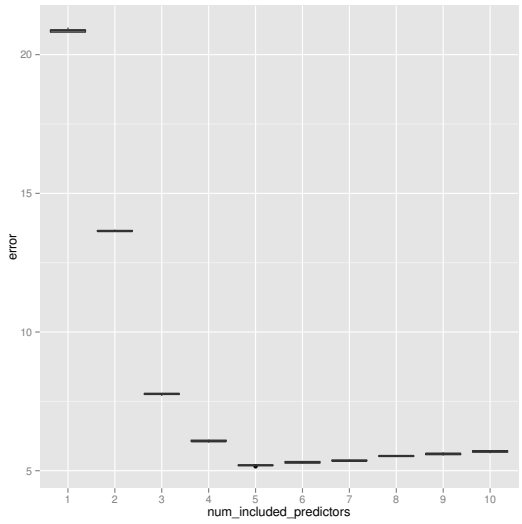

Figure 5.3: *Sequence of OOB errors for the Friedman1 data computed by* 10 *RFs with inclusion of* $1-10$ *predicting variables.*

for the selected variables $V1$ up to $V5$ is illustrated in figure 5.4. From the figure the linearity governing two predictors $V4$ and $V5$ is evident, as well as the non-linearity of variable $V3$, with a minimum at 0.5.

Under the assumption that the data is generated by an additive model we can further explore the relation between the target and predictor variables by illustrating $y - (\hat{k}_4 V4 + \hat{k}_5 V5)$ against variables $V1, V3$ and $V5$. The slopes $\hat{k}_4$ and $\hat{k}_5$ are estimated to $\hat{k}_4 \approx 9.61$ and $\hat{k}_5 \approx 3.67$ by the partial dependency plot 5.4, where the latter figure shows a clear squared relationship with respect to variable $V3$.

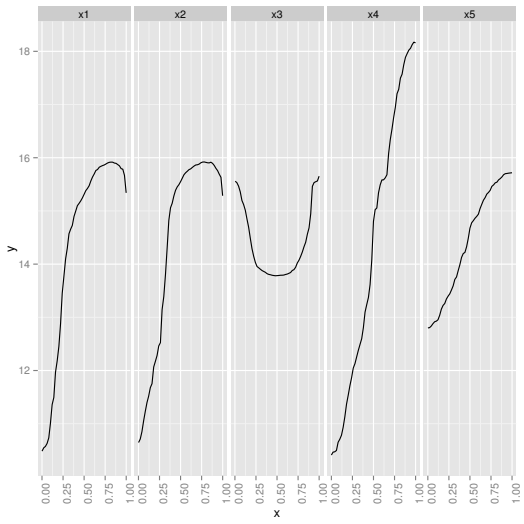

Figure 5.4: *Random forest partial dependency plot over the Friedman1 data.*

An illustration of y and of $(y - \hat{k}_4 V4 - \hat{k}_5 V5)$ versus the variables $V1, V3$ and $V4$ are shown in figures 5.5 and 5.6.

Waveform data

The parameter `mtree` and `ntree` are chosen by minimizing the OOB errors over 40 RF models illustrated in figure 5.7. The error rates are, as in the freidman1 data, quite stable for `ntree` $\in \{500, 1500, 2500\}$ and with a slightly smaller error range for `ntree` ≥ 1500. The parameter controlling the number of variables available at each data partitioning `mtry` provides a shared minimum error rate at its default value $\lfloor\sqrt{40}\rfloor = 6$ together with the value `mtry` $= 3$.
The associated permutation variable importance measures are shown in figure 5.8. From the figure it is observed that all truly important variables are included except variable $V1$ and $V21$ by the first 20 largest importances. The work by Verikas et al. [57] observed that the VIMs for variables $< 1, 2, 20, 21 >$ had almost equivalent values as the VIMs over the noise variables.

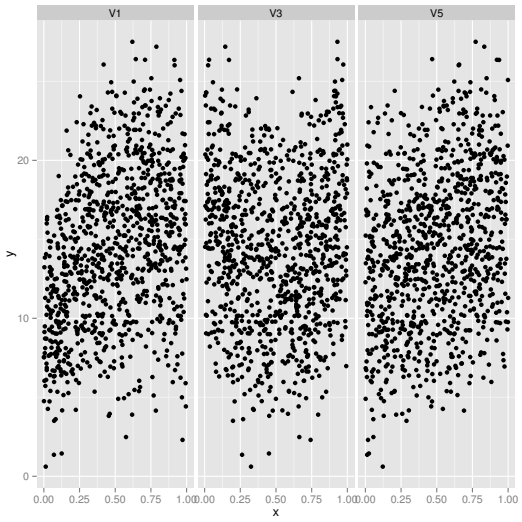

Figure 5.5: *y as a function over* 3 *predictors for the Freidman1 data.*

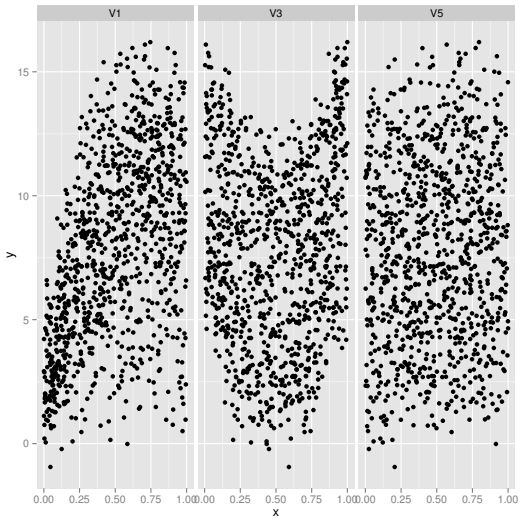

Figure 5.6: $y - 9.61V4 - 3.67V5$ *as a function over* 3 *predictors for the Freidman1 data. The slopes are estimated by random forest partial dependency plots.*

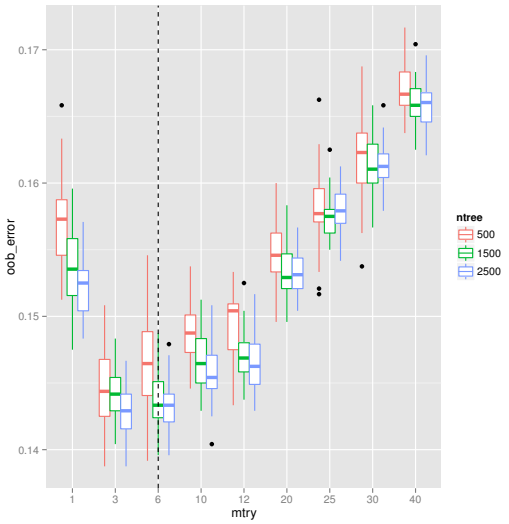

Figure 5.7: *OOB errors computed by* 40 *RFs for the Waveform data. The default value for* `mtry` *is marked by a vertical line.*

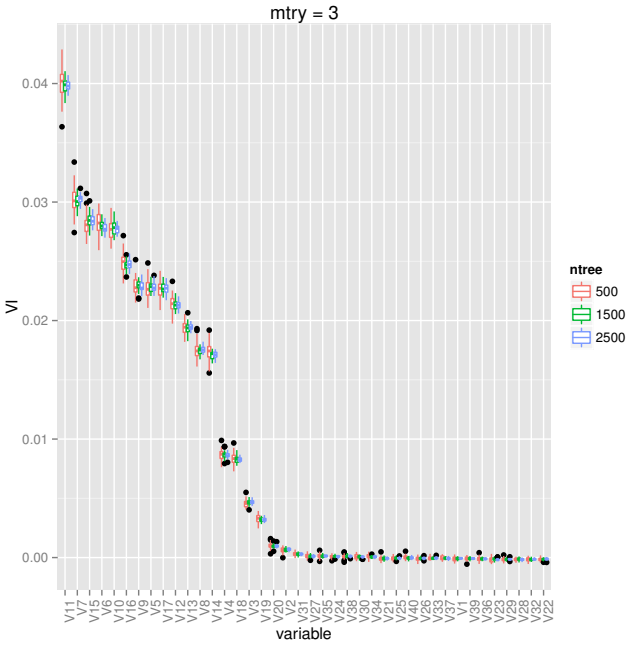

Figure 5.8: *Permutation VIM computed by* 40 *RFs for the Waveform data.*

The zero approaching VIMs is more easily observed illustrated in figure 5.9. It is observed that the ordered sequence of VIMs are deemed zero for variables ordered after the first 26 variables, and that the 27th variable is a truly important variable, albeit with zero importance measure. Similar VIMs were observed by the gini importance measure. However the gini measure using `mtry` $= 3$ did provide a greater distinguish for variables $< 2, 20 >$, but the measures are generally harder to distinguish from noise.

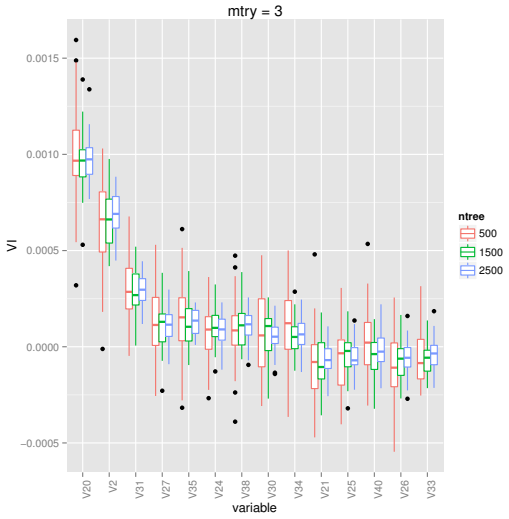

Figure 5.9: *Zoomed subset of* 14 *zero approaching permutation VIM from figure 5.8.*

The sequence of OOB errors used for variable selection are shown in figure 5.10. Recall that the variable importance measures previously shown in figure 5.9 do only allow inclusion of a sequence with a maximal length of 26. From the figure illustrating the OOB errors we observe that the minimal OOB error is obtained by $k = 27$ number of included variables. Notably that is exactly when the truly important variable $V21$ is included in the model. A more reliable procedure is to terminate the OOB error sequence in a mean sense. The minimal OOB error is obtained for $k = 25$ number of included variables. The final model obtained by the variable selection procedure 4.1 includes all informative variables except variable $V1$ and $V21$ and there are 6 included predictors which actually are random noise variables.

Noisy Waveform data

This data investigates how the VIMs alters when the Waveform data is adjoined with categorical noise variables. The optimal values for the parameter *mtry* shows a similar relationship as observed for the Waveform data. The OOB errors are il-

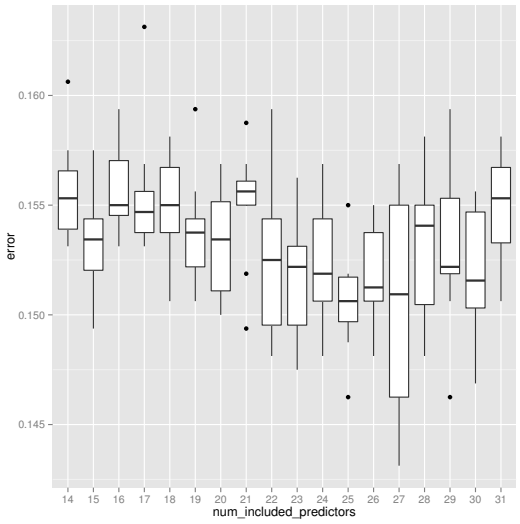

Figure 5.10: *Sequence of OOB errors for the Waveform data computed by* 10 *RFs with inclusion of* $14 - 31$ *predicting variables.*

lustrated in figure 5.11, with minimal OOB errors for $mtry = 3$ and for $mtry = 6$.

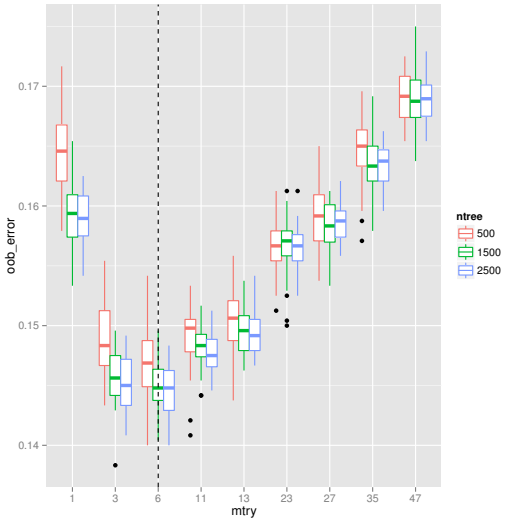

Figure 5.11: OOB errors computed by 40 RFs for the Waveform noise data. The default value for `mtry` is marked by a vertical line.

From the permutation VIMs shown in figure 5.12 we note that there are 5 categorical noise variables with a higher variable importance measure than the VIM obtained for the truly relevant variable $V21$.

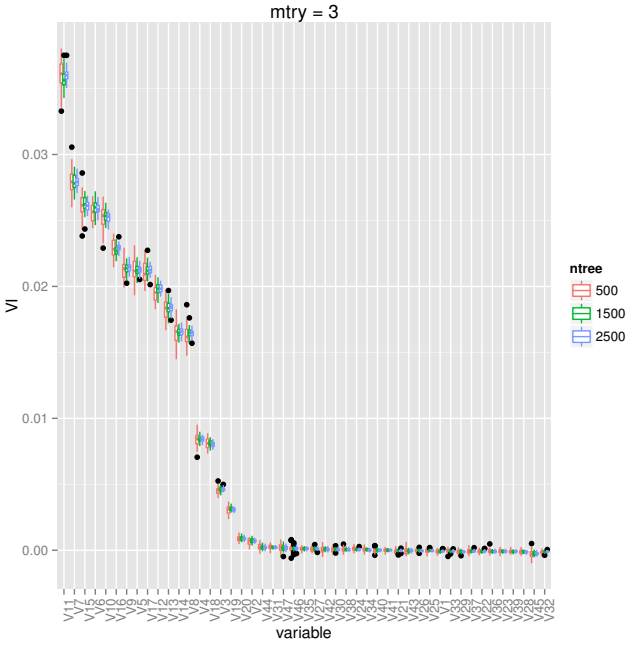

Figure 5.12: *Permutation VIM computed by* 40 *RFs for the Waveform noise data.*

The zero approaching measure are shown in figure 5.13 and there are a total of 30 non-zero important variables available for inclusion. Notably the top 19 predictors ordered by decreasing variables importance measure are equal and equally ordered for both the noise free and the noise Wave form data. However the two variables $V8$ and $V14$ are interchanged with respect to their variable importance measure.

The greedy sequence of OOB errors is shown in figure 5.14. In a mean sense the optimal number of predictors to include in the final model are $k = 26$. The final model includes all important variables except variable $V1$ and $V21$. The model contains 7 noise variables and where 3 noise variables are due to the adjoined categorical noise. The gini VIMs, not illustrated in this report, are highly non-robust with respect to the imputed noise. The random forest variable importance assigned a higher importance measure to 5 of the categorical noise variables than to the two truly important variables $V20$ and $V2$.

Adult data

This data consists of real census data obtained from the UCI machine learning repository [40]. The parameter `mtree` and `ntree` which minimizes the OOB errors are illustrated in figure 5.15. The error rates are stable for `ntree` $\in \{500, 1500, 2500\}$ and with a smaller error range for `ntree` ≥ 1500, especially for small values for

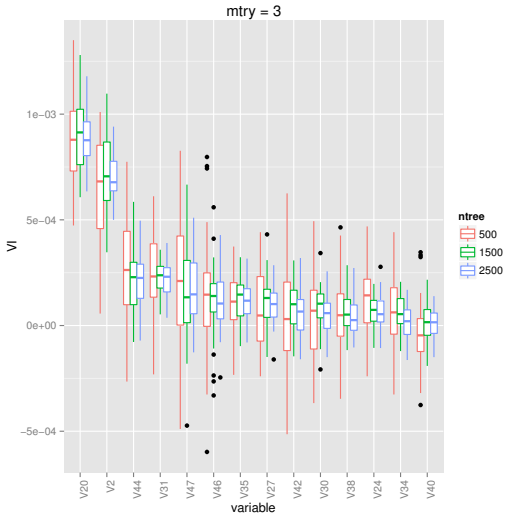

Figure 5.13: *Zoomed subset of* 14 *zero approaching permutation VIM from figure 5.12.*

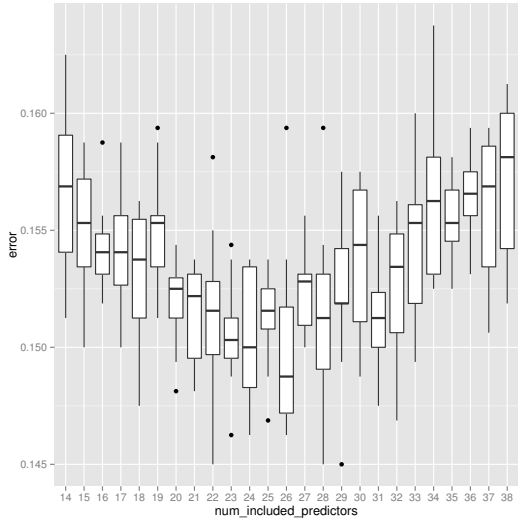

Figure 5.14: *Sequence of OOB errors for the Waveform noise data computed by* 10 *RFs with inclusion of* $14 - 38$ *predicting variables.*

`mtry`. The parameter `mtry` provides a distinct minimum at the default value `mtree` $= \lfloor\sqrt{14}\rfloor = 3$.

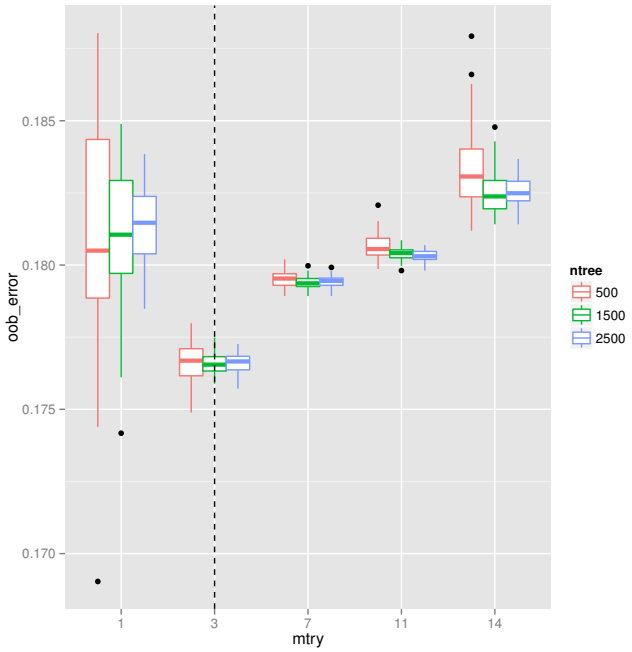

Figure 5.15: *OOB errors computed by* 40 *RFs for the Adult data. The default value for* `mtry` *is marked by a vertical line.*

The associated permutation VIMs illustrated in figure 5.16 implies a potential exclusion of the two variables $V14$ and $V3$, which corresponds to variables native-country and final-weight. Simply put, the two variables are weights used to produce accurate population estimates for the various items covered in the regular monthly Current Population Survey.

The sequence of OOB for the Adult data are shown in figure 5.17 with a minimum for $k = 9$ number of included variables, and where 10 included variables also provides a reasonably small OOB error. The 5 least important variables, ordered by decreasing importance, corresponds to `workclass`, sex, race, fnlwgy and native-country. By including variables based on the computed VIMs either 2 or 3 variables may be excluded from the final model. Thus by only basing exclusion of variables by the variable importance measure, the minimum obtained by including 9 variables would thus be missed. The confusion matrix on test data using default parameter settings and trained using all predictor variables is shown in table 5.2 and the confusion matrix using optimized parameters trained with the 9 most important variables is shown in table 5.1. The mean miss-classification errors for both models are shown in table 5.3.

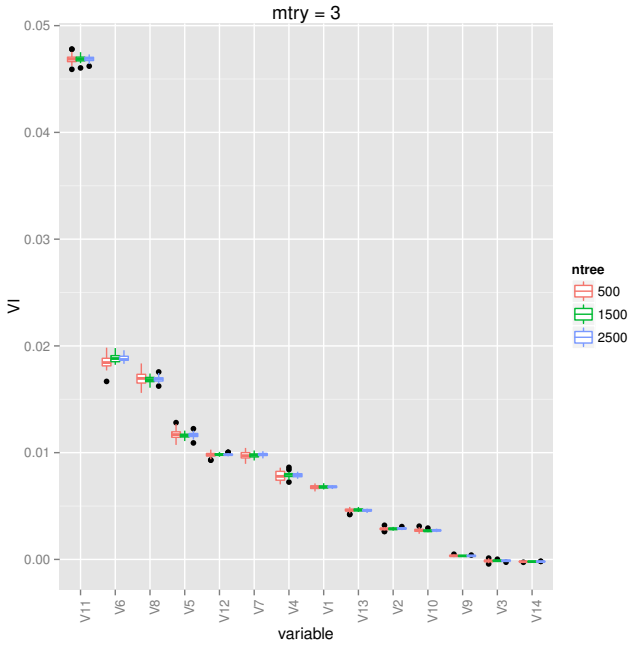

Figure 5.16: *Permutation VIM computed by* 40 *RFs for the Adult data.*

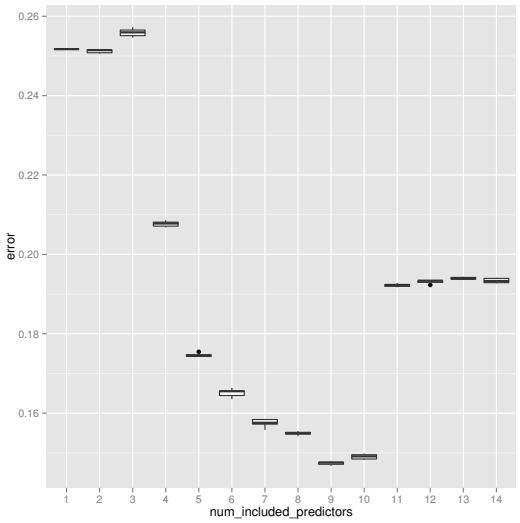

Figure 5.17: *Sequence of OOB errors for the Adult data computed by* 10 *RFs with inclusion of* $1 - 14$ *predicting variables.*

	target outcome	
predictions	≤ 50K	> 50K
≤ 50K	10535	1383
> 50K	825	2317

Table 5.1: *RF predicted values with optimized parameter settings and trained using the* 9 *most informative variables.*

	target outcome	
predictions	≤ 50K	> 50K
≤ 50K	11330	2707
> 50K	30	993

Table 5.2: *RF predicted values with default parameter settings for the Adult test data.*

	Default parameters	Optimized parameters
prediction error	0.1821	0.1456

Table 5.3: *Mean RF classification errors over* 5 *iterations on the Adult test data.*

5.3 Discussion

This work further strengthens the capacity of the random forest variable importance measures as a means to identify informative variables in both supervised classification-and regression settings. The permutation importance measure is shown capable differentiating informative variables from noise variables in a non-linear settings and where shown successful identifying all informative variables in the 10 dimensional non-linear Friedman1 data. The greedy sequence of OOB errors further provides a distinct minimum exactly when the 5 truly informative variables are included. The partial dependency plot used in combination with residual plots are shown useful investigating the underlying structure in the non-linear Friedman1 data, specifically both linear and squared contributions are identified and estimated by the partial dependency plot. Further the order of the importance measures obtained from the data are sensible with respect to the generated data.

Notably the importance measures used in combination with greedy inclusion of the top most important variables are shown capable identifying 19 out of all 21 informative variables in the 40 dimensional Waveform data, a task which has previously proven difficult [55, 56]. Choosing variables for inclusion merely based on the variable importance measures did not provide a clear distinction between the informative-and non-informative variables. The importance measure guided sequence of OOB errors provided a minimum, in a mean sense, when the top 25 important predictors are included in the model. However the permutation importances, shown in figure 5.9, are zero after the 27:th most important variable. Thus by combining the importance measures and the greedy OOB error sequence two

more non-informative variables are excluded from the final model. Further the permutation importance measures are shown robust when presented with categorical noise variables and the importance measure respects the relative order over the informative variables to a large extent, there are however two variables whose order where interchanged, this property is albeit in need of further validation. The gini importance measure are however observed to inaccurately order categorical noise variables as more informative than truly informative predictors, an observation not presented in this report. The final model includes a total of 25 variables, where 21 variables are truly informative and 4 variables are noise with respect to the target variable. The number of included variables in the final model are reduced by 62.5%.

The permutation importance measure applied are also shown useful when applied to the real census data, at least with respect to prediction accuracy. The measures computed over the census Adult data deems 2 or 3 variables as non-informative. The OOB error sequence however suggests further exclusion of variables, a distinct minimum is obtained by excluding the 6 least important variables. The mean missclassification error for the optimized random forest model equals 0.1456 and is equal to 0.1821 for a random forest model using default parameter settings. When investigating real data it is hard to assess the validity of the importance measure with respect to identifying truly informative predictors. From the 3 simulation trials the information obtained by the importance measures and the information obtained by the greedy sequence of OOB errors are used in a combination for assessing informative variables, the included predictor variables are further validated as truly informative with respect to the target variable. However with respect to prediction accuracy the total error when excluding the 6 least important variables is decreased by 25.01%.

Lastly the greedy OOB sequence of errors computed from the noisy circle data, illustrated in figure A.7, showed that the errors further decreased when two non-informative variables where included in the model. Thus it is only when the importance measures are combined with the greedy sequence of errors that only truly informative variables are selected by the variable selection procedure 4.1. However the reliability of the variable selection procedure needs to be validated on additional data.

Chapter 6

Future Work

The random forest model is not yet fully understood from a theoretical perspective and further work is needed before the mechanics underlying the variable importance measures are fully understood. Yet the variable importance measures is proven useful analyzing a variety of data. Previous work presented in the background section 2 investigates and describes certain properties regarding the variable importance measures, such as correlated settings [18, 3, 15, 51, 20], and data containing missing values [23]. These simulation studies could be extended by additional simulation trials designed to analyze certain relations between the target and predictor variables. E.g. additional non-linear settings could be investigated, data with strong predictor interactions and examining data with categorical informative variables, perhaps by changing the overall distributional setting such as in mixture models. Further it would be interesting to analyze boundary situations which balances cases when the importance measures are proven reliable from cases where the importance measure displays undesired properties. This information is useful not only to learn the nature of the random forest variable importance measure but also for identifying cases when there is a need of additional variable selection techniques used to differentiate informative from non-informative variables.
It would be useful to form benchmark data designed by properties regarding the predictor variables such as, correlations, missing-values, and skew ranges of values, together with various degrees of non-linearity between the target and predictor variables, and varying importance relations. Onto this benchmark data a variety of variable selection techniques could be tested. Comparing the performance of each technique to the different simulations may facilitate understanding of the *whens* and *whys* governing the variable selection techniques. This may further improve the communication between various variable selection techniques, especially those techniques where the measures are yet not fully understood from an analytic perspective.

When the main concern is maximizing prediction accuracy a suggestion for further improvements, more formally specified in [18], is by forward selection or backward

elimination applied after the variable selection procedure. Inclusion or exclusion of variables are executed when the decrease in OOB error exceed a user defined threshold, or preferably by visual inspection of how the OOB error relates to inclusion or exclusion of variables. The number of included variables after the variable selection procedure are approximately reduced by $50, 63, 55$ and 64 percent. The first variable selection procedure can be seen as a filter technique which reduces the number of predictor variables before applying a wrapper method such as forward selection or backward elimination.
Note that the computational effort is greatly reduced compared to applying a wrapper method to the full data since the variable importance measures are computed only once at the filtering step. Another heuristic designed to guide removal of variables could be formed by computing the variable importance measures repeatedly after each variable is removed from the model. E.g. the importance measures could be computed on the model, with k variables included, resulting from the variable selection procedure 4.1. The variable with smallest variable importance is then repeatedly excluded from the model. This modification reduces the computational burden compared to backward elimination since at e.g. the first iteration there is only 1 candidate compared to k candidate variables considered for removal. It would be interesting to apply these variable selection techniques as a pre-processing step before training another classifier e.g. the SVM or the k-NN classifier on the reduced data.

Appendix A

Introduction to terminology and search of important variables

A.1 Supervised learning

Let $\mathcal{D} = \{(X_i, Y_i)\}_i$ be a data set consisting of input variables X_i commonly called *features* or *predictor* variables. For each tuple of predictor variables X_i there is an associated outcome Y_i called *target* variable. The field of *Supervised learning* works on the problem of learning a predictor function f that *generalizes*, that is given a not previously observed input variable the model should return a reasonable outcome.

The nature of the predictor function f can grouped into two different settings called *regression* and *classification* which are specified by the nature of the target variable Y. In a regression framework the associated outcomes spans a range of continuous values and so should the outcomes of the learned predictor function. Whereas in a classification setting the target variable ranges over a set of categories and there is no natural distance between different categories.

Further the model governing the predictor function can generally be grouped as a *parametric*, *semi-parametric*, or a *non-parametric* model. An example of a parametric model, with p predictor variables, is the linear regression model

$$\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_p X_p. \tag{A.1}$$

Parametric models generally have the advantage that they are easy to interpret. The relationship between the target variable Y and the predicting variables X is easily understood from the parameter estimates $\hat{\beta}$ and the relation between Y and X is specified as a closed form expression. However parametric models may be *misspecified* meaning that there is a mismatch between the specified model and the nature of the data. Which in turn may lead to poor prediction performance and unreliable parameter estimates. The latter error is a so called *bias* error, the learned

predictor is forced e.g. to use only linear combinations of the predictor variables, and in some cases this bias will be a too strict assumption.

Non-parametric models are generally more flexible and can model non-linear relationships more accurately, the models may on the other hand lack a clear interpretation between the predictor variables and the target variable. The latter type of models where there is no clear interpretation between the predictors X and the target variable Y are called *black box* models.
An example of a non-parametric model which provides a measure of importance between the predicting variables and the target variable is the *Random Forest* model. An illustration of the random forest model trained on two dimensional data, where the first predictor X_1 is continuous and the second X_2 is categorical with two categories, is shown in figure A.1.

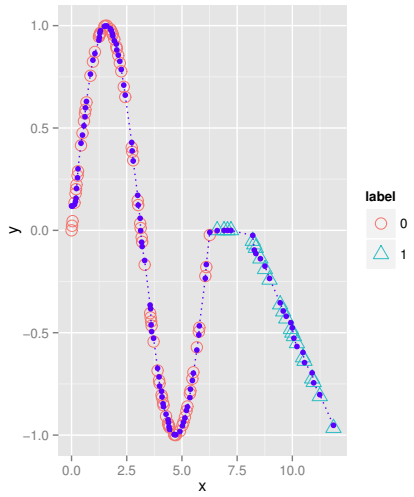

Figure A.1: *Random forest predictions trained on data generated from a non-linear function. The circle and triangles marks the test points and the solid dots marks the corresponding predictions.*

A.2 Relation between the target and predictor variables

In many supervised learning problems the objective is not merely to learn a predictor with high accuracy but one would also want to know which predictor variables that relates to the target variable. This information can be used to remove non-informative variables which in turn may reduce the risk of *overfitting* i.e. reduce the risk that the learned model describes random error instead of the underlying model.
The parametric linear regression model specified in section A.1 provides a clear relationship between the predictor and target variable and is commonly used as a

starting model. In settings when the target variable is categorical a commonly used parametric model is the *logistic regression model*. If the target variable Y ranges over two labels $\{0, 1\}$ the model is defined by

$$P(Y_i = 1|X_i, \beta) = \frac{\exp \sum_{i=0} \beta_i X_i}{1 + \exp \sum_{i=0} \beta_i X_i}, \quad X_0 \coloneqq 1. \tag{A.2}$$

The parameters β are learned from the data by minimizing the *log-likelihood*, that is finding values for β which minimizes the probability $\log p(Y_1, .., Y_n|X, \beta)$. The parameters that solves the latter equation are those that maximizes the probability of observing the data at hand $\{(Y_i, X_i)\}_i$ under the assumption that the data is generated from the logistic regression model.
Further the logistic regression model have a clear interpretation between the predictors X and the target variable Y given by the *log-odds*

$$\text{logit}(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{A.3}$$

when the variable X_i increases by one unit the predicted odds increases by e^{β_i}, all other variables kept fixed.
An example where the logistic regression model is learned on a 20 dimensional data set, describing a *noisy circle* in the two variables X_1 and X_2 generated by,

$$Y = \begin{cases} 1 & \text{if } \sqrt{X_1^2 + X_2^2} + \epsilon \leq 1 \\ 0 & \text{else} \end{cases}. \tag{A.4}$$

Where $\{X_i\}_{i=3}^{10} \sim \text{N}(0, 1)$ and $\{X_{i+10}\}_{i=1}^{10} \sim \text{Mu}(i)$[1]. The generated data together with the predictions from the trained logistic regression model is shown in figure A.2. The estimated coefficients for the predictors X_6, X_8 and X_{19} are all significant at levels of $0.001, 0.01$ and 0.01.
As seen in the figure the predictions are not reasonable and this example illustrates a model that is miss-specified. The estimated parameter coefficients together with the distribution of the relevant predictor X_1 and the non-relevant predictors X_6 and X_{19} are shown in figure A.3. When performing *multiple comparisons* it is often recommended to make some sort of adjustment to the p-values [44], after applying the *Bonferroni adjustment*[2] all parameter estimates are non-significant.
The logistic regression model specified by (A.2) classifies points X as label 1 if $\hat{p} \geq 1/2$ and otherwise as label 0. Where $\hat{p}$ is given by (A.2)

$$\hat{p} = p(Y = 1|X, \hat{\beta}) = \frac{\exp \hat{\beta} \cdot X}{1 + \exp \hat{\beta} \cdot X}. \tag{A.5}$$

[1] Mu(i) denotes the multinomial distribution e.g. p(X=k) = 1/(i+1), k < i.

[2] The Bonferroni adjustment controls the *familywise error rate, that is the probability of rejecting at least one true null hypothesis.*

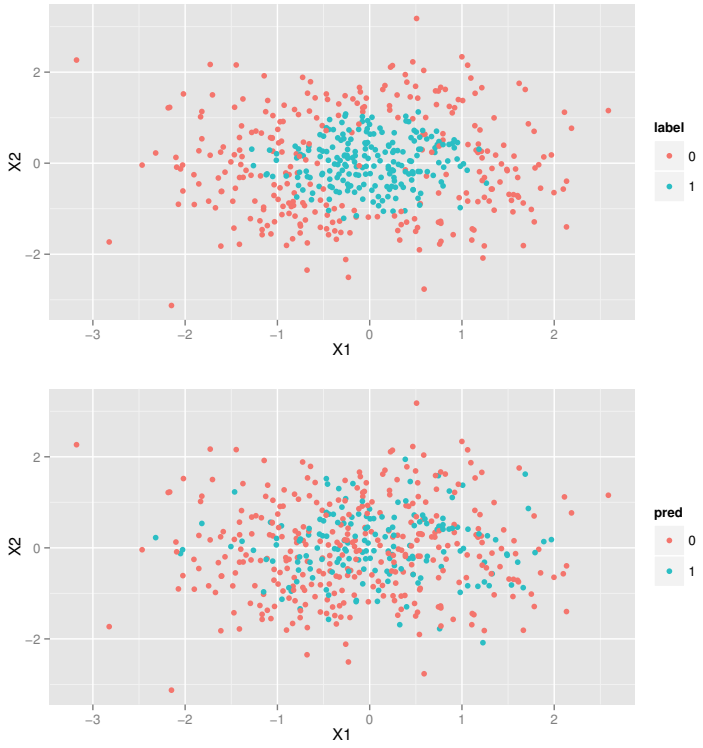

Figure A.2: *First row:* 20 *dimensional training data generated by* (A.4). *Second row: Logistic regression predictions on the above training data.*

By equation (A.5) the decision boundary which separates the two categories is specified by the hyperplane $\{X : \hat{\beta} \cdot X = 0\}$. Thus the poor performance given by the logistic regression model fitted on the noisy circle data (A.2) is expected due to the linear decision boundary. A solution handling this miss-specification problem is to adjoin the data by the two features $X_{21} = X_1^2$ and $X_{22}^2 = X_2^2$ which will improve the *separability* of the data i.e. the two adjoined predictors facilitates separating the points with label 1 from those with label 2. In general when accounting for all squared terms in a n dimensional learning problem the dimensionality is increased by n. Accounting for all interaction terms $X' = X_i X_k, (i \neq k)$ the dimensionality will further increase by $\binom{n}{2} = n(n-1)/2$. In total when accounting for all squared and all interactions terms in a 20 dimensional learning problem the dimensionality will increase by 210.

Training a logistic regression model on the noisy circle data generated by (A.4) adjoined by the two squared features provides two significant parameter estimates β_{21} and β_{22} which are the coefficients associated with X_1^2 and X_2^2, when using the Bonferroni adjustment. With levels of 0.001 and all other parameter estimates are non-significant. The predictions and the training data for the logistic regression

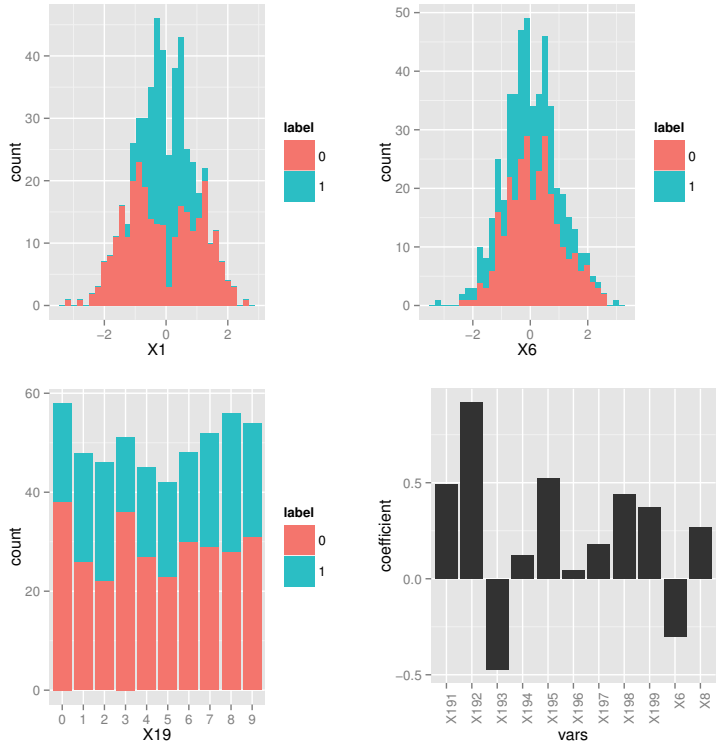

Figure A.3: *First* 3 *figures, top left to bottom right, illustrates the distribution of* 3 *out of* 10 *features generated by* (A.4)
Bottom left figure shows the estimated coefficients by the logistic regression model. The estimated coefficients $\beta_6, \beta_8, \beta_{92}$ *are significant with levels of* $0.001, 0.01, 0.01$. *Adjusting for multiple comparisons by the Bonferroni correction all estimates are non-significant.*

model are shown in figure A.4.

Mentioned in section A.1 parametric models generally are easier to interpret compared to non-parametric models but may on the other hand be miss-specified as illustrated in figure A.2. The non-parametric random forest model previously learned on the non-linear data illustrated in figure A.1 is also applicable in classification settings. Predictions by the random forest model learned on the noisy circle data, not adjoined by squared features, is illustrated in figure A.5. The random forest model also provides a measure of importance which measures the importance of the predictors modeling the target variable under the trained model. The measure of importances are shown in figure A.6.

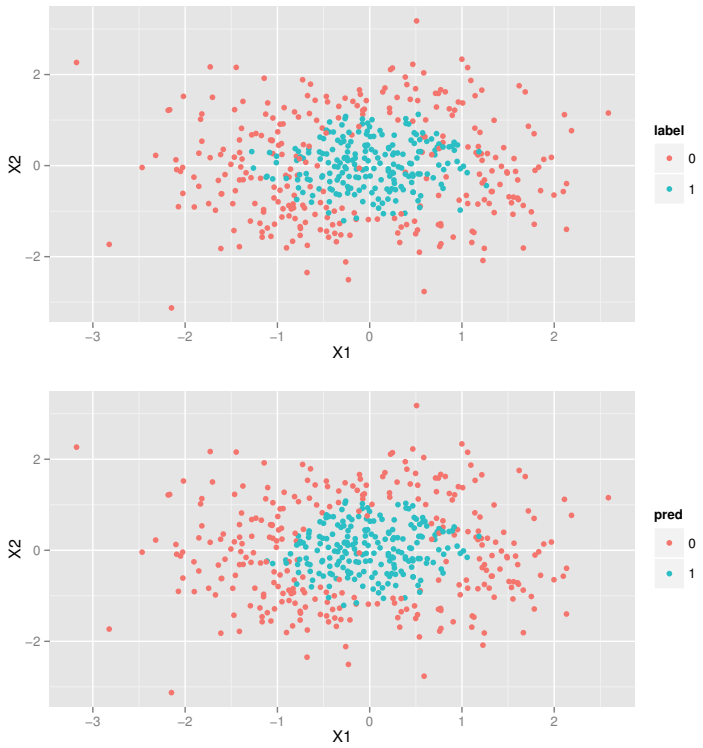

Figure A.4: *First row: Training data generated by* (A.4)
Second row: Associated predictions using a logistic regression model when adjoining the data by two additional features $X_{21} = X_1^2$ *and* $X_{22} = X_2^2$.

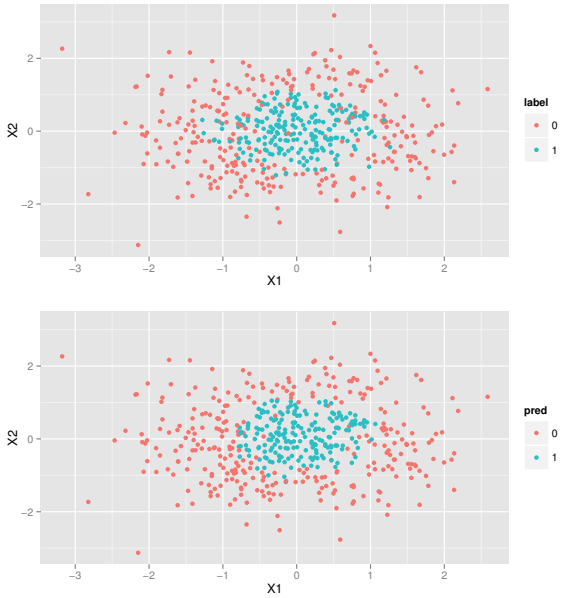

Figure A.5: *First row:* 20 *dimensional training data generated by* (A.4). *Second row: Random forest predictions on the above training data.*

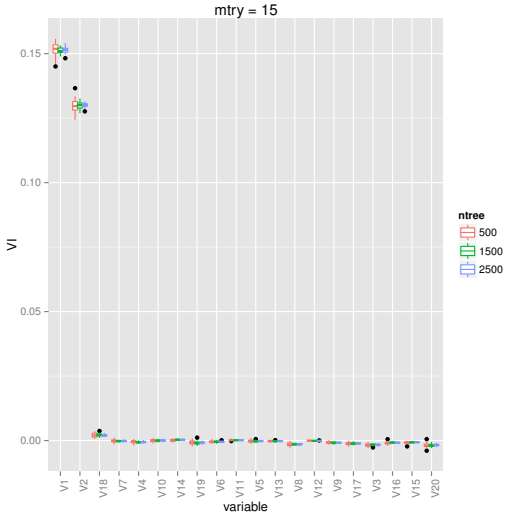

Figure A.6: *Permutation importance measures computed by* 10 *RFs learned on the data illustrated in figure A.5. Only the two predictors* X_1 *and* X_2 *are informative modeling the target variable.*

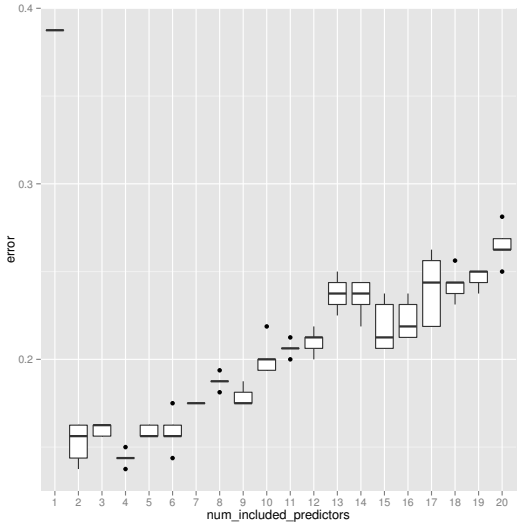

Figure A.7: *Sequence of OOB errors computed by including k variables for $k = 1, .., 20$ ordered by decreasing permutation variable importance measure. This sequence guides the variable selection procedure specified in chapter 4.*

Bibliography

[1] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair. Bayesian additive regression trees-based spam detection for enhanced email privacy. In *Availability, Reliability and Security, 2008. ARES 08. Third International Conference on*, pages 1044–1051. IEEE, 2008.

[2] Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588, 1997.

[3] Kellie J Archer and Ryan V Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52 (4):2249–2260, 2008.

[4] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. URL `http://www.ics.uci.edu/$\sim$mlearn/{MLR}epository.html`.

[5] Avrim L Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271, 1997.

[6] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, August 1996. ISSN 0885-6125. doi: 10.1023/A:1018054314350. URL `http://dx.doi.org/10.1023/A:1018054314350`.

[7] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL `http://dx.doi.org/10.1023/A:1010933404324`.

[8] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A., 1984.

[9] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

[10] Wouter Buckinx, Geert Verstraeten, and Dirk Van den Poel. Predicting customer loyalty using the internal transactional database. *Expert Systems with Applications*, 32(1):125 – 134, 2007. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2005.11.004. URL `http://www.sciencedirect.com/science/article/pii/S0957417405003143`.

[11] Delphine S Courvoisier, Christophe Combescure, Thomas Agoritsas, Angèle Gayet-Ageron, and Thomas V Perneger. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *Journal of clinical epidemiology*, 64(9):993–1000, 2011.

[12] Kristof Coussement and Dirk Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, 34(1): 313 – 327, 2008. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2006.09.038. URL `http://www.sciencedirect.com/science/article/pii/S0957417406002806`.

[13] Kristof Coussement and Dirk Van den Poel. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications*, 36(3):6127–6134, 2009.

[14] A Criminisi, J Shotton, and E Konukoglu. Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114*, 5(6):12, 2011.

[15] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):1, 2006.

[16] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000.

[17] Jerome H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991.

[18] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

[19] Phillip I. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2004. ISBN 038720279X.

[20] Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Correlation and variable importance in random forests. *arXiv preprint arXiv:1310.5726*, 2013.

[21] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[22] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[23] A Hapfelmeier and K Ulm. Variable selection by random forests using data with missing values. *Computational Statistics & Data Analysis*, 80:129–139, 2014.

[24] Alexander Hapfelmeier and Kurt Ulm. A new variable selection approach using random forests. *Computational Statistics & Data Analysis*, 60:50–69, 2013.

[25] Frank Harrell. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Institute of Mathematical Statistics, 2015.

[26] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[27] Trevor J Hastie. Nonparametric logistic regression. Technical report, 1983.

[28] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference*. IEEE, 1995.

[29] Torsten Hothorn, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro, and Mark Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

[30] 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.

[31] Laurent Hyafil and Ronald L Rivest. Constructing optimal binary decision trees is np-complete. *Information Processing Letters*, 5(1):15–17, 1976.

[32] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370.

[33] Jyh-Shing Roger Jang. Anfis: adaptive-network-based fuzzy inference system. *Systems, Man and Cybernetics, IEEE Transactions on*, 23(3):665–685, 1993.

[34] Hyunjoong Kim and Wei-Yin Loh. Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, 2011.

[35] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[36] Irena Koprinska, Josiah Poon, James Clark, and Jason Chan. Learning to classify e-mail. *Inf. Sci.*, 177(10):2167–2187, May 2007. ISSN 0020-0255. doi: 10.1016/j.ins.2006.12.005. URL `http://dx.doi.org/10.1016/j.ins.2006.12.005`.

[37] Bart Larivière and Dirk Van den Poel. Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 29(2):472–484, 2005.

[38] Friedrich Leisch and Evgenia Dimitriadou. *mlbench: Machine Learning Benchmark Problems*, 2010. R package version 2.1-1.

[39] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[40] M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

[41] Kathryn L Lunetta, LBrooke Hayward, Jonathan Segal, and Paul Van Eerdewegh. Screening large-scale association study data: exploiting interactions using random forests. *BMC genetics*, 5(1):1, 2004.

[42] Animesh Patcha and Jung-Min Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer networks*, 51(12):3448–3470, 2007.

[43] Peter Peduzzi, John Concato, Elizabeth Kemper, Theodore R Holford, and Alvan R Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379, 1996.

[44] Daniel J Mundfrom Jamis J Perrett, Jay Schaffer, Adam Piccone, and Michelle Roozeboom. Bonferroni adjustments in tests for regression coefficients. *Multiple Linear Regression Viewpoints*, 32:1–6, 2006.

[45] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[46] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

[47] Brian D Ripley. *Pattern recognition and neural networks*. Cambridge University Press, 2007.

[48] Bram Slabbinck, Bernard De Baets, Peter Dawyndt, and De Vos Pauls. Towards large-scale fame-based bacterial species identification using machine learning techniques. *Systematic and applied microbiology*, 32(3):163–176, 2009.

[49] Carolin Strobl, Anne-Laure Boulesteix, and Thomas Augustin. Unbiased split selection for classification trees based on the gini index. *Computational Statistics & Data Analysis*, 52(1):483–501, 2007.

[50] 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):1, 2007.

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

[52] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[53] Sandro Tognazzo, Bovo Emanuela, Fiore Anna Rita, Guzzinati Stefano, Monetti Daniele, Stocco Cramen Fiorella, and Zambon Paola. Probabilistic classifiers and automated cancer registration: An exploratory application. *Journal of biomedical informatics*, 42(1):1–10, 2009.

[54] Brigitte Unger and Frans van Waarden. How to dodge drowning in data? rule- and risk-based anti money laundering policies compared. *Review of Law & Economics*, 5(2):953–985, 2009.

[55] A. Verikas, M. Bacauskiene, D. Valincius, and A. Gelzinis. Predictor output sensitivity and feature similarity-based feature selection. *Fuzzy Sets and Systems*, 159(4):422 – 434, 2008. ISSN 0165-0114. doi: http://dx.doi.org/10.1016/j.fss.2007.05.020. URL `http://www.sciencedirect.com/science/article/pii/S0165011407002631`. Theme: Information Processing.

[56] Antanas Verikas and Marija Bacauskiene. Feature selection with neural networks. *Pattern Recognition Letters*, 23(11):1323–1335, 2002.

[57] Antanas Verikas, Adas Gelzinis, and Marija Bacauskiene. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2): 330–349, 2011.

[58] Allan P White and Wei Zhong Liu. Technical note: Bias in information-based measures in decision tree induction. *Machine Learning*, 15(3):321–329, 1994.

[59] C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, and N. M. Adams. Transaction aggregation as a strategy for credit card fraud detection. *Data Min. Knowl. Discov.*, 18(1):30–55, February 2009. ISSN 1384-5810. doi: 10.1007/s10618-008-0116-z. URL `http://dx.doi.org/10.1007/s10618-008-0116-z`.

[60] Yaya Xie, Xiu Li, EWT Ngai, and Weiyun Ying. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36 (3):5445–5449, 2009.

[61] Weiyun Ying, Xiu Li, Yaya Xie, and Ellis Johnson. Preventing customer churn by using random forests modeling. In *Information Reuse and Integration, 2008. IRI 2008. IEEE International Conference on*, pages 429–434. IEEE, 2008.

[62] Anazida Zainal, Mohd Aizaini Maarof, Siti Mariyam Shamsuddin, and Ajith Abraham. Ensemble of one-class classifiers for network intrusion detection system. In *Information Assurance and Security, 2008. ISIAS'08. Fourth International Conference on*, pages 180–185. IEEE, 2008.

[63] Jiong Zhang, Mohammad Zulkernine, and Anwar Haque. Random-forests-based network intrusion detection systems. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(5):649–659, 2008.