Separating the Wheat from the Chaff: On Feature Selection and Feature Importance in Regression Random Forests and Symbolic Regression

Sean Stijven
Department of Mathematics
and Computer Science
University of Antwerp
Antwerp, Belgium
seanstyven@gmail.com

Wouter Minnebo
Department of Mathematics
and Computer Science
University of Antwerp
Antwerp, Belgium
wouter.minnebo@gmail.com

Katya Vladislavleva
Department of Mathematics
and Computer Science
University of Antwerp
Antwerp, Belgium
katya@evolved-analytics.com

ABSTRACT

Feature selection in high-dimensional data sets is an open problem with no universal satisfactory method available. In this paper we discuss the requirements for such a method with respect to the various aspects of feature importance and explore them using regression random forests and symbolic regression. We study 'conventional' feature selection with both methods on several test problems and a case study, compare the results, and identify the conceptual differences in generated feature importances.

We demonstrate that random forests might overlook important variables (significantly related to the response) for various reasons, while symbolic regression identifies all important variables if models of sufficient quality are found. We explain the results by the fact that variable importances obtained by these methods have different semantics.

Categories and Subject Descriptors

I.6.4 [**Simulation and Modeling**]: Model Validation and Analysis; I.6.5 [**Simulation and Modeling**]: Model Development; I.1.2 [**Symbolic and Algebraic Manipulation**]: Algorithms

General Terms

Algorithms, Experimentation

Keywords

Feature Selection, Variable Selection, Variable Importance, Random Forests, Symbolic Regression, Genetic Programming

GECCO'11, July 12–16, 2011, Dublin, Ireland.

1. INTRODUCTION

Input-response datasets with a large number of variables are now more than ever a great challenge to analyze and model. Variable or feature selection algorithms are designed to reduce the number of variables in a meaningful way (see [8] for a general discussion). Variable selection is necessary since irrelevant or noisy variables can be detrimental for model accuracy, comprehensibility, and robustness, and hence can jeopardize project success and mislead system understanding. Feature selection has been studied across disciplines using different techniques [17, 9, 20, 16, 25, 1] and efforts are made to consolidate this knowledge [24].

While many different methods are available, few are accessible. In addition it is not straightforward to compare results obtained with different methods because they express different concepts of importance depending on the assumptions made by the method. Different interpretations of importance hamper the adoption of new feature selection methods since users need to fully understand a method to use it with confidence. The absence of a conventional definition of relevance and importance motivated us to explore and understand the necessary characteristics of a good method of assigning importances to data variables.

In our definition, something is important if its presence or absence matters. So, an input variable is important if its presence of absence matters in system understanding, i.e. in understanding of the behavior of the response variables. Importance of a variable can only be accessed from a plausible, accurate and insightful model - the functional relationship between the inputs and the response. An input variable is not important if it does not cause a change in the response, i.e. the partial derivative of the model over this variable (if it exists) is zero across the entire input space.

The problem is that unless input-response models are given analytically as continuous differentiable functions, the sensitivity analysis must be performed by manual exploration of the response's behavior with little if any quantification of importance.

In search for a good general-purpose method of measuring the importance of variables we want such a method to have the following properties:

- **Interpretability**–Obtained importances should reflect the importances of the true input variables, without transformation.

- **Strictness**–Only variables relevant to describing the response should be allocated importance, so spurious variables should not appear important.
- **Conservativeness**–At intermediate stages of importance analysis all potentially interesting variables should receive importance.
- **Reproducibility**–A result can only be considered correct if it is reproducible.
- **Universality**–It is desirable for importances to be mutually comparable, and in addition to be problem independent.

In general variable selection algorithms come in three varieties: filter, wrapper and embedded methods [8]. A filter method will not build any models, but only use the characteristics of the data, while a wrapper method will estimate variable importance based on model evaluation. The embedded method incorporates the feature selection in the model building process, such that variable importances are guiding the modeling process. Filter methods can be used as preprocessing to remove some spurious variables, but we argue that variable importances can not be assessed since no models are built. Hence a filter method can not estimate the impact of a variable's absence. Wrapper methods are by definition constrained by the model they operate on, because such methods handle the model as a black box simulation. Embedded methods are beneficial since models are built with partial knowledge of variable importance.

In this paper we examine two embedded techniques, regression random forests and symbolic regression via genetic programming, which are two commonly used methods for variable selection in high-dimensional real-life data. The variable selection strategy in both methods is based on the *quantifiable* variable importances induced by the input-response regression models. These importance values are scrutinized for both methods with respect to properties described above and we conclude that they are conceptually different and should not be used in a one-to-one comparison of an importance (relevance) of any given feature.

The rest of the paper is organized as follows: First we present the modeling methods from an algorithmic standpoint. Next we elaborate on the interpretations of variable importance obtained by those methods. This is illustrated the differences by several experiments and a case study. Lastly the conclusions are summarized and future work is outlined.

2. MODELING METHODS

2.1 Random Forests

The Random Forest (RF) technique was introduced by Breiman [3] as an ensemble of binary decision trees. An accessible treatment is also provided in [18].

A forest is an ensemble of weak learners constructed such that every tree will independently divide the dataset into a training and test set, a technique also known as bagging (see [2]). The training set is referred to as in-bag, and the test set as out-of-bag (OOB). The training set can be sampled with replacement from the full dataset so the same point occur multiple times in a training set. A tree built from such data will be biased in the region of points with multiple occurrences. Even though it degrades the individual predictions, the prediction of the forest will benefit [2].

The training set is then recursively partitioned according to an information gain criterion, until a partition is sufficiently small or no information is gained by partitioning further. The last partition in such a series is called a leaf or terminal node.

Algorithm 1 Constructing a Random Forest

```
for Number of Trees in Forest do
  Divide dataset into TrainingSet and TestSet
  PartitionList $\leftarrow$ TrainingSet
  while PartitionList is not empty do
    PartitionList $\rightarrow$ Current
    if $|$Current$| >$ LeafSize then
      Left, Right $\leftarrow$ BestSplit( Current )
      PartitionList $\leftarrow$ Left, Right
```

At each step of the construction of a regression tree a partition is split in two smaller partitions by introducing a decision threshold on a variable, represented by a node. Data points with a value higher than the decision value for that variable will be grouped together, as are the data points with a value less than the decision variable. Candidate variables on which to split are determined by choosing a random subset out of all variables. Exhaustive search for the variable providing the highest information gain is performed on all these candidates and that variable is chosen to be the decision variable.

The prediction of a tree is determined by checking the decision value of the tree starting in the root node. Every next node to consider is the node to which the input point would be assigned during partitioning. This effectively creates a path from the root node to a leaf node. The prediction of a tree for any input point is the average of the responses of the training points in the leaf node reached by following such path. The prediction of a forest is the average of all tree predictions. Note that the predicted response of a tree is a multidimensional step function, see Figure 1.

2.2 Symbolic Regression

Symbolic Regression (SR) aims to capture the input-response behavior of the data with an algebraic expression. No assumptions are made about the model structure. Clearly, there exists an infinite number of possible expressions with all input parameters thus an exhaustive search is not a realistic approach. Instead SR uses genetic programming (GP) optionally complemented by other methods [15] to search the model space efficiently. In addition each iteration of symbolic regression explores a large set of models, frequently called a population. Each model is described as an expression tree. Ensemble-based SR uses an ensemble of models as a final solution. Note, that SR ensembles are collections of individually strong predictors in contrast to RF.

At every iteration the fitness of all individuals is assessed using criteria related to the prediction error. A new population is then generated by recombining and modifying these individuals through crossover and mutation. Crossover is the process of combining two parent individuals into two new child individuals, by using subtrees of both parents. Mutating an individual introduces random alterations in its

Algorithm 2 A simple Symbolic Regression algorithm.

```
Initialize CurrentPopulation
for Number of iterations do
    Determine fitness of all individuals
    for Size( CurrentPopulation ) do
        NewPopulation ← Empty
        Parent1 ← SelectParent( CurrentPopulation )
        if rand() < CrossRate then
            Parent2 ← SelectParent( CurrentPopulation )
            Child ← CrossOver( Parent1, Parent2 )
        else
            Child ← Mutate( Parent1 )
        NewPopulation ← Child
    CurrentPopulation ← NewPopulation
```

expression tree. The rate at which crossover and mutation occurs is instrumental in the convergence rate to good solutions. Using only crossover prevents the population from finding new good combinations since diversity is lost and some crucial operator or variable might be absent. On the other hand with too much mutation the search process is a more random exploration and the convergence rate will suffer as well. Satisfactory results were obtained using 90% crossover, 10% mutation rate.

Several enhancements to the described algorithm exist. We used Pareto-aware Symbolic Regression, or SR via Pareto genetic programming [22, 19]. This method uses additional objectives for model selection.

A complexity measure favoring simple models is assigned to each individual and the model building process is then redefined as the optimization of two (or more) objectives: the predictive power of the individuals and their simplicity. Note that there are multiple solutions since neither objective is strictly preferred, and all Pareto optimal solutions are considered. A solution is Pareto optimal with respect to other solutions if none of those alternative solutions improve at least one objective while not reducing any other objective's performance.

Pareto-aware Symbolic Regression is implemented in [4] which incorporates several additional enhancements, most notable elitism and niching. If all individuals are recombined into a new population, good solutions might be lost which is clearly undesirable. Elitism counters this by maintaining an archive, which is a set of current best models according to the Pareto optimality. An alternative niching strategy in a space of selected objectives can also be used. Quality preservation (elitism) is instrumental in the parent selection since it can be expected that the probability to create a child with good fitness is higher with an elite solution as parent. Niching is instrumental to preserve the population diversity.

2.3 Prediction

It is clear that RF and SR have a different model structure, affecting the prediction of created models.

Figure 1 shows that the prediction from a decision tree results in a stepwise function, yet a forest can approximate smooth functions by combining trees. More trees yield smoother results in general, however the diversity within the ensemble is important as well for the predictive capabilities of the forest [5]. It is also noted that data points grouped in a leaf can be considered 'close' together, and the prediction resembles nearest neighbors, a notion formalized in [13].

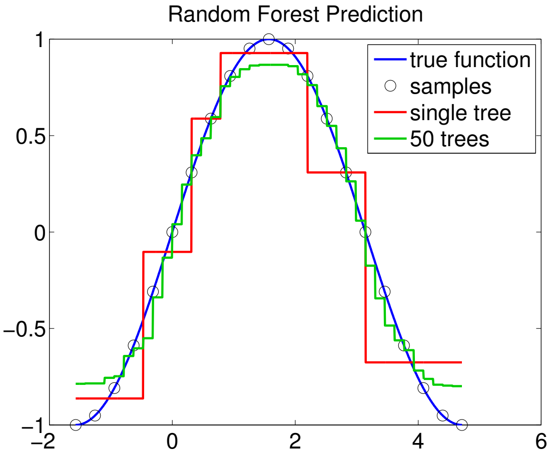

Figure 1: The blue function is sampled at the dots. The red line is a single tree's prediction, while the green line is the prediction of a forest of 50 trees.

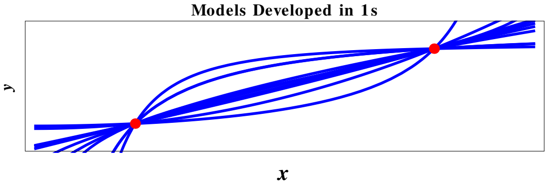

Figure 2: Response curves of models found by SR fitting the two points.

Symbolic regression provides multiple models. The use of model ensembles as the final solution is advisable to improve robustness. The model disagreement can be used to estimate a confidence in the ensemble prediction (see Figure 2).

3. VARIABLE IMPORTANCE

3.1 Random Forests

In RF the variable importance is determined by the change in prediction error when records are permuted. Each out-of-bag (OOB) datapoint is evaluated using the value of that variable from another OOB point in the dataset, the resulting point is called a perturbed datapoint. The reasoning is that important variables are split more often, influencing the prediction more than irrelevant variables. So a variable contributing to the prediction, will affect the new path through the tree of a perturbed datapoint such that it is more probable to end up in a leaf further away from its original leaf as the importance of the variable increases. The importance of a variable can then be defined as the average percentage of change in prediction error per tree. Note this definition is *interpretable* as defined in the Introduction. Note that other yet similar techniques are proposed as well in [12]. In classification problems it has been shown that the variable importance is biased towards variables with many categories [21], so caution is advised when mixing categorical and continuous variables. In [6] different strategies are proposed for selecting variables either for explanation or prediction. A comprehensive comparison of linear regression and random forests is presented in [7].

It could be argued that variable importance according to RF is a pure wrapper method since it relies on model evaluation. However, the forest prediction is not used, instead each tree contributes individually. This could be weighted by the individual tree prediction accuracy, but this is not a default strategy. Also note that the variable importance will closely reflect the choices made at the splitting phase of the algorithm, so labeling this method as embedded seems more appropriate.

3.2 Symbolic Regression

In SR the variable presence indicates whether a variable is potentially important. Irrelevant variables will account for extra complexity but do not provide additional predictive power. Clearly individuals incorporating them will perform worse by Pareto optimality than individuals using only relevant variables. This will in turn lower the chance of being chosen to produce children, so the presence of irrelevant variables is discouraged. Therefore the presence of a variable in a sufficiently evolved population will provide an indication on whether that variable is relevant for describing the response. This is also an *interpretable* measure as defined in the Introduction.

Presence-weighted and fitness-weighted variable importance are described as defined in [23].

The presence-weighted variable importance for variable x_i, $i = 1, ..., d$ in a set of models $\mathcal{M} = \{M_j, j = 1, ..., m\}$ is computed as a fraction of the models containing x_i. Such definition provides a robust estimation of relevance if set $\mathcal{M}$ is of high quality and sufficiently diverse, i.e. obtained using several independent runs.

The fitness-weighted variable importance eliminates the need for high quality in $\mathcal{M}$ (but does not eliminate the need for diversity) by weighting the presence indicator with the fitness of each model:

$$\mathcal{I}_i^{(FW)}(\mathcal{M}) = \sum_{j=1}^{m} \frac{fitness(M_j)}{\sum_{i=1}^{d} \delta(x_i, M_j)} \delta(x_i, M_j) \qquad (1)$$

Note that in the extreme case of a population of perfect models, the fitness-weighted variable importance is equivalent with the presence-weighted variable importance. In this paper we use the normalized fitness-weighted variable importances as defined by:

$$\mathcal{I}_i^{(NFW)}(\tilde{\mathcal{M}}) = \frac{\mathcal{I}_i^{(FW)}(\tilde{\mathcal{M}})}{\sum_{i=1}^{d} \mathcal{I}_i^{(FW)}(\tilde{\mathcal{M}})} * 100\% \qquad (2)$$

3.3 Experimental Comparison

Applying both RF and SR to a few test datasets highlight some potential problems with respect to variable importance. For both RF and SR several independent runs were conducted to ensure results are not significantly influenced by the random seed of either algorithm. The same dataset is used for both algorithms, and variables are uniformly sampled between zero and one unless stated otherwise.

In all RF runs a forest of 1000 trees was built, considering variable subsets of the default $|variables|/3$ size to determine the best split. For RF the median importance is plotted without confidence intervals for visual presentation.

For SR the variable importance is computed on all models obtained from the independent runs, and models were of sufficient quality for the variable importance to be representative. Accuracy was defined as $1 - R^2$ where R^2 is the square of the scaled correlation between prediction and response. The number of subtrees in an expression was used as complexity measure. Model age was used as the secondary complexity measure.

The colors used on bar plots are for visual presentation only, and convey no further information.

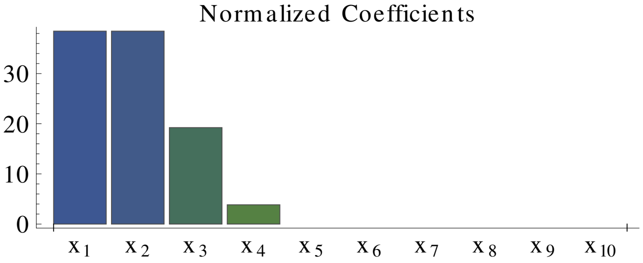

Figure 3: The normalized coefficients of the generating function from Eq. 3.

3.3.1 Linear Model with Spurious Variables

When presented with a large dataset, is is common that most of the variables are redundant and one typically wishes to remove them. Here the methods are tested for *strictness*, as defined in the Introduction. The generating function is given by Eq. 3.

$$y = 10x_1 + 10x_2 + 5x_3 + 1x_4 + 0x_5 + ... + 0x_{10} \qquad (3)$$

Variable importances are expected to correspond to the normalized coefficients, as illustrated in Figure 3. The method by which RF computes the variable importance could capture the same relation, and results are shown in Figure 4. Remarkable is that the x_4 variable's importance is underestimated. This suggests that RF is *strict* but not *conservative*.

Because the discussed variable importance for SR only reflects a variable's presence, even though fitness-weighted, it is expected that the relative importances will not be well preserved. As can be seen from the results in Figure 7 the variable importance of both x_3 and x_4 is overestimated. This more *conservative* behavior safeguards against removing relevant variables, but might cloud the relative importances.

We observed that both methods are not sensitive to the addition of more variables with zero coefficient.

3.3.2 Unbalanced Data

In real problems the available data might not be well distributed over the input space. This situation is mimicked by oversampling a variable such that its distribution is no longer uniform, here variable x_1 is sampled much denser in the interval $[0, .1]$. The generating function is given by Eq.3.

The result obtained with RF is shown in Figure 5. Observe that the importance of variable x_1 is underestimated by a large margin. In this case a perturbed datapoint will most likely be assigned a value from the dense area, were little change can be observed. These findings agree that RF is not *conservative*.

This is in contrast with the findings from SR, presented in Figure 8. Observe that in this example the unbalanced dataset does not change the variable importances of SR significantly. Since SR builds global models and the underlying model is unchanged, the local difference in density will only slightly influence the variable presence.

3.3.3 Correlated Variables

Frequently not all variables in a dataset are independent, and input variables are correlated. Such dataset was created by adding new variables related to a true variable. In this experiment variables x_i for $i = 1, ..., 10$ satisfy Eq. 3, while variables x_j for $j = 11, ..., 20$ are constructed by

$$x_j = 0.9 * x_1 + 0.1 * Noise, \tag{4}$$

where $Noise$ is uniformly distributed in [0,1].

In RF the variable importances change dramatically as illustrated in Figure 6. Because correlated variables provide splits of comparable quality, the true variable effectively loses splits to a variable it is correlated with, even though this correlated variable might have coefficient zero in the generating function. This causes the true variable to lose influence over the prediction path through the tree, and consequently its variable importance will be lower. These findings support the notion that RF is not *conservative.* Other research found that relevant correlated variables are overestimated and propose a conditional variable importance which greatly improves this behavior [20].

The best models obtained with SR do not contain the correlated noise variables. But since those correlated variables can still serve as a surrogate for the true variable to build rough approximations, they will remain active in the population. Consequently SR will always allocate some importance to the noise variables, as can be observed in Figure 9.

3.3.4 Relatively Weak Variables

Here we examine how both methods deal with variables which are considerably less important than others, yet not redundant. This test verifies whether the methods are *conservative*, as defined in the Introduction. The generating function is given by Eq. 5.

$$y = 10x_1 + 10x_2 ... 10x_8 + 1x_9 + 0x_{10} \tag{5}$$

Random forests produces the variable importance shown in Figure 11. It is observed that the variable with relatively low coefficient (x_9) has an importance close to zero, supporting the observation that RF is not *conservative.* In addition the relative importance of the variables x_1 through x_8 vary considerably, implying that RF is not *universal.* Since most variables are of equal importance neither will consistently provide a better split than its peers, but still always better than the variable with a low coefficient. The potential splits on the equally important variables differ only slightly according to the splitting criterion. They will reflect the distribution of the input variables which might exhibit slight non-uniformities due to the limited sample size. The splitting criterion amplifies these slight variations in density.

The same situation is better handled in SR, of which results are shown in Figure 12. While the importances are not exactly the same, this is expected since the population can still yield suboptimal models causing a small variation in the importances, even though the variables have equal coefficients.

Observe that the importance of a variable with lower coefficient is overestimated. In the specific case that many significant variables' contribution is approximately equal, Latent Variable Symbolic Regression (LVSR) performs better than SR with respect to *strictness* as defined in the introduction (see [14]).

4. CASE STUDY

Human development is an important and interesting subject, more so because it concerns every human on the planet. Many social, economical and personal factors are combined in a single number known as the Human Development Index (HDI) [10]. This index is an indicator expressing the general quality of life of the human population but not limited to material well-being. It uses three dimensions of development: health, knowledge, income. Health is measured by life expectancy at birth. Knowledge is measured by combining the expected years of schooling for a school-age child in a country today with the mean years of prior schooling for adults aged 25 and older. Income is measured in purchasing power adjusted per capita gross national income. These three dimensions are then combined using the geometric mean.

The HDI is not perfect, but the data collected in the process surely holds information and is subject of active research [11].

In this case study we compare the variable importances obtained by RF and SR, when estimating several parameters from the HDI dataset. The initial dataset is available at [10] and contains many missing values.

While RF can produce a proximity matrix by which missing values could be filled in, it seemed prudent to pre-process the dataset independent of the methods to be compared. The dataset was pre-processed as follows: If a value was missing for a specific variable in more than 50 countries, the variable was removed. If a country had more than 30 missing values on the remaining variables, the country was removed. The remaining missing values were replaced by the average value over countries within the same development group (low, medium, high, very high developed country). In addition the variable 'Antenatal Care' was removed because no data was available for any country in the very high developed group.

The variables 'Total Satisfaction Freedom of Choice', 'Total Purposeful Life' are modeled because any relationship between the development of a country and perceived freedom and purpose would be interesting. For both these variables data was collected through questionnaires were participants were asked whether they were satisfied with their freedom of choice, or whether they find their lives purposeful, respectively. The percentage of the population who answers 'yes' to either question is the response variable. Note that the dataset provides both a total percentage and a percentage for the female population to make comparison across genders possible. Only the total percentage is modeled and the female percentage is removed a priori, since this variable holds exactly the same information defeating the purpose of modeling.

The GDP per Capita was also modeled to compare both techniques on a non-linear problem. The Gross Domestic Product (GDP) is the value of all products and services produced within a country in a year. This is typically divided by the population to make international comparison possible, considering that more people are able to produce more value. While the explicit formula is known in this case and the true variables GDP and population are present in

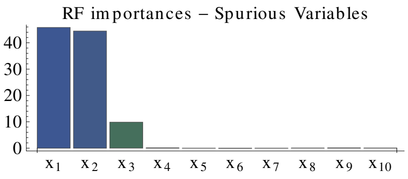

Figure 4: Variable importances according to RF, using a uniformly sampled dataset from the generating function given by Eq. 3.
Note the underestimated importances of variables x_3 and x_4.

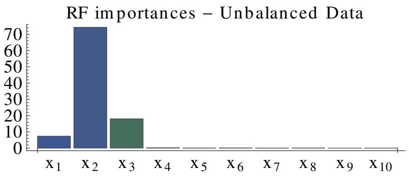

Figure 5: Variable importances according to RF, where the dataset sampled from the generating function given by Eq. 3 is unbalanced in x_1. Note the gross underestimation of variable x_1.

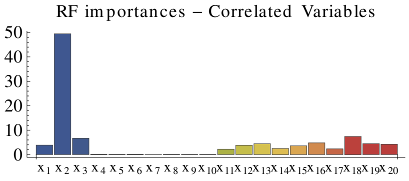

Figure 6: Variable importances according to RF, using a dataset from Section 3.3.3. The true variable x_1 is grossly underestimated and furthermore indistinguishable from variables it is correlated with.

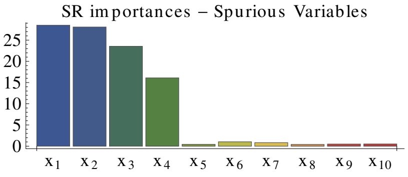

Figure 7: Variable importances according to SR, using a uniformly sampled dataset from the generating function given by Eq. 3.
Note the overestimated importances of variables x_3 and x_4.

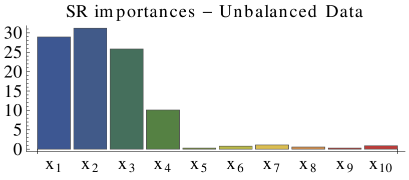

Figure 8: Variable importances according to SR, where the dataset sampled from the generating function given by Eq. 3 is unbalanced in x_1. The importances are still comparable with the unbalanced dataset (Fig. 7).

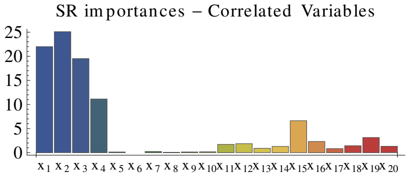

Figure 9: Variable importances according to SR, using a dataset from Section 3.3.3. Importances of the true variables remain comparable to those in Fig. 7, but extra variables have a non-neglectable importance.

the dataset, the difficulty lies in extracting the two relevant variables from the many irrelevant but correlated variables. The variable ‘GNI per Capita’ is removed a priori, this variable by itself approximates GDP per Capita.

4.1 Purposeful Life

The relation between country development and Purposeful Life is illustrated in Figure 13. Note that countries with a higher HDI exhibit a larger variance in the Purposeful Life.

Importances obtained by SR displayed in Figure 19, models of good quality showed a linear relation between variables. The SR variable importances suggest that the most important variables determining whether life is perceived as purposeful are: Hospital Beds, Fertility, Age, Maternity Leave, Undernourishment. We roughly relate these variables to health, children and food. We find these factors not unreasonable in order to enjoy a purposeful life.

Importances obtained by RF are shown in Figure 16. Observe that only the most important variables are similar to those identified by SR, and variables with lower importances are incomparable. Since the underlying model was suggested to be linear by SR the observations made in the experiments from the previous section are applicable.

4.2 Freedom of Choice

The relation between Freedom of Choice Satisfaction and country development is presented in Figure 14. While the higher developed countries score higher, a large variance is observed.

Importances obtained by SR are shown in Figure 20, models of good quality showed a linear relation between variables. The SR variable importances suggest that the most important variables determining the satisfaction with ones freedom of choice are: Standard of Living Satisfaction, Healthcare Satisfaction, Respect, Employment, Air and Water Quality. So this one satisfaction variable aggregates information of different aspects in life. We find these variable importances intuitively plausible.

Importances obtained by RF are visualized in Figure 17. Again the most important variables agree with those found by SR with variations in relative importance, but variables with lower importances are different. For example the Income Gini Coefficient has a relatively high importance score, while this variable is not present at all in the top 20 important variables obtained with SR. In the absence of domain expertise this can have a significant impact on a decision making process. This illustrates that caution is warranted when performing variable selection, and consulting multiple methods is advised.

4.3 GDP per Capita

The relation between GDP per Capita and country development is shown in Figure 15.

Importances obtained by SR are presented in Figure 21. While neither GDP nor Population scores high in importance, it is noted that the explicit formula $GDP/Population$ was retrieved in a few models. But since the majority of the population achieved a much lower level of quality, they will obscure these variables even though the importance is fitness weighted. This is an example of why models of sufficient quality should be used to estimate importances, and highlights the consequences of not doing so.

Importances obtained by RF are displayed in Figure 18. Neither the GDP nor Population variable is present in the top 20 most important variables, as was the case with SR. However it is observed that correlated variables dominate other variables. Many variables with ‘HDI’ in their name are correlated, and are also correlated with GDP per Capita. This agrees with findings in other research regarding RF variable importance with correlated variables [20].

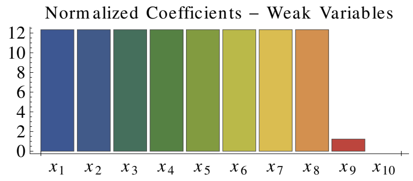

Figure 10: The normalized coefficients of the generating function as defined by Eq. 5. By computing the variable importances for this linear problem one can verify whether the used method is *conservative*.

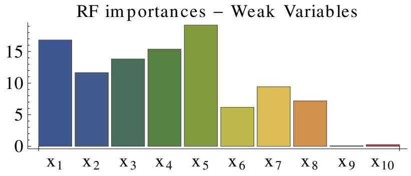

Figure 11: Variables with equal coefficient can still vary in importances obtained by RF. Note variable x_9 appears to be irrelevant, while it has a non zero coefficient in the generating function given by Eq. 5.

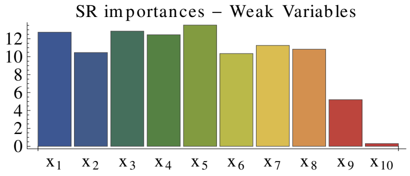

Figure 12: Variables with equal coefficient vary in importances obtained by SR, but do not differ as much as the importances obtained by RF in Figure 11. The generating function is given by Eq. 5.

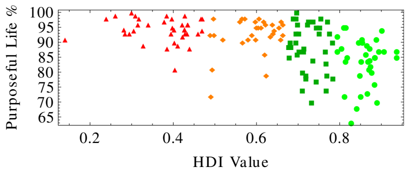

Figure 13: The HDI plotted versus Purposeful Life. Countries with a higher HDI score less on average due to higher variance in Purposeful Life.

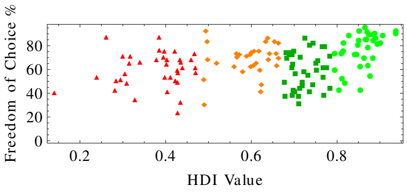

Figure 14: The HDI plotted versus Freedom of Choice Satisfaction. The higher developed countries score higher. The variance is large.

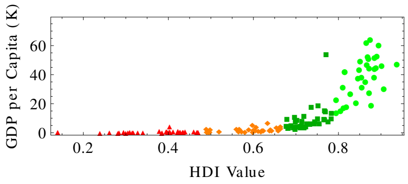

Figure 15: The HDI plotted versus the GDP per Capita, which is closely related to the IncomeIndex, one of the three dimensions of the HDI.

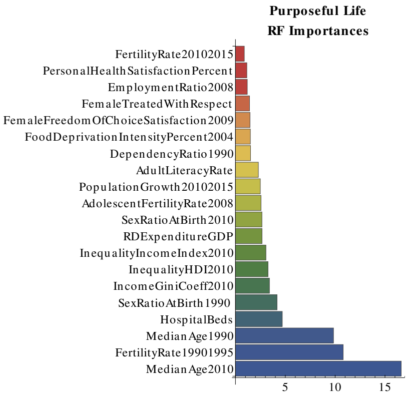

Figure 16: Importances obtained with RF modeling Purposeful Life.

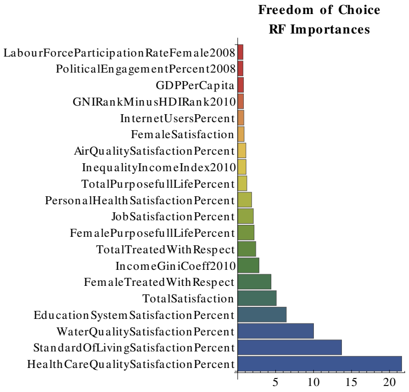

Figure 17: Importances obtained with RF modeling Freedom of Choice.

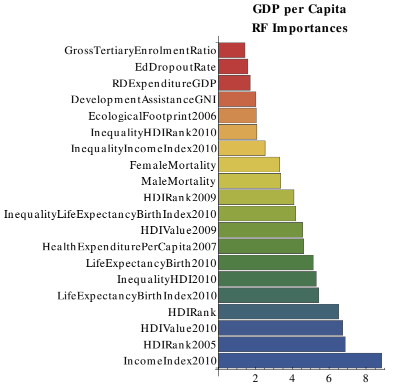

Figure 18: Importances obtained with RF modeling GDP per Capita.

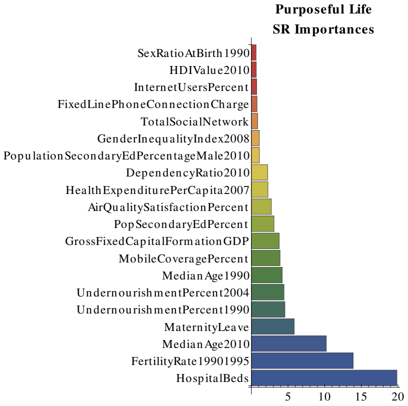

Figure 19: Importances obtained with SR modeling Total Purposeful Life.

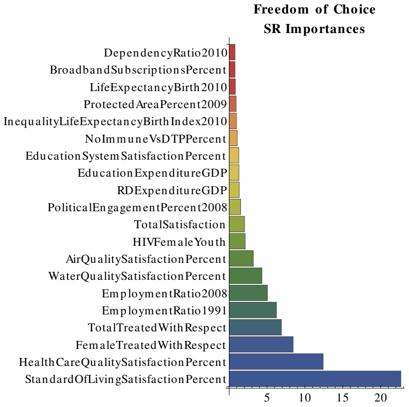

Figure 20: Importances obtained with SR modeling Freedom of Choice.

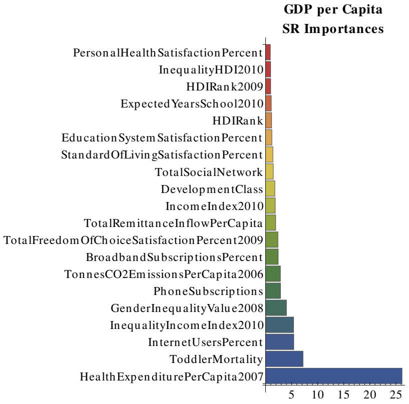

Figure 21: Importances obtained with SR modeling GDP per Capita.

5. CONCLUSION AND FUTURE WORK

Random forests (RF) can efficiently find important variables in the presence of many irrelevant variables. When many variables are equally important the variable importances vary randomly since such variables are not recognized as truly distinct. In addition RF might value a variable considerably less than expected, and correlation with spurious variables amplifies this behavior. Furthermore the variable importance is influenced by the data distribution, leaving them prone to misinterpretation. In general, caution is advised when using RF to decide which variables to retain, even if the dataset is known not to exhibit strong correlations or unevenly balanced data. Because data points in leaf nodes are similar in a nearest neighbor sense, variables selected by RF express proximity.

Symbolic regression (SR) performs well throughout all tests. The model building process is ideally a continuous evolution, and while convergence is assumed, there is little information available about the speed of convergence or the proximity to acceptable solutions. A population of insufficient quality will yield unreliable variable importances, so model quality must be verified when drawing conclusions. It remains important to look for robust algorithmic configurations that ensure the discovery of models of sufficient pre-defined quality. A more formal framework to establish the importance of variables given an expression tree is needed as well.

Acknowledgements

The authors would like to thank Mark Kotanchek and Evolved Analytics LLC for providing licenses of DataModeler, the anonymous reviewers for useful comments and the Department of Mathematics and Computer Science of the University of Antwerp for their help in funding this project.

6. REFERENCES

[1] H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64:115–123, Mar. 2008.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, August 1996.

[3] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 10.1023/A:1010933404324.

[4] Evolved Analytics LLC. *DataModeler Release 1.0.* Evolved Analytics LLC, 2010.

[5] M. Gashler, C. Giraud-Carrier, and T. Martinez. Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous. In *2008 Seventh International Conference on Machine Learning and Applications*, pages 900–905. IEEE, Dec. 2008.

[6] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recogn. Lett.*, 31:2225–2236, October 2010.

[7] U. Grömping. Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, 64:308–319, Nov. 2009.

[8] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[10] Human Development Reports. http://www.hdr.undp.org/.

[11] Human Development Research Papers. http://hdr.undp.org/en/reports/global/hdr2010/papers/.

[12] H. Ishwaran. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537, 2007.

[13] Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101:578–590, June 2006.

[14] T. McConaghy. *Latent Variable Symbolic Regression for High-Dimensional Inputs*, pages 103–118. Springer, 2010.

[15] R. K. McRee. Symbolic regression using nearest neighbor indexing. In *Proceedings of the 12th annual conference companion on Genetic and evolutionary computation*, GECCO '10, pages 1983–1990, New York, NY, USA, 2010. ACM.

[16] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, Aug. 2005.

[17] A. Rakotomamonjy. Variable selection using svm-based criteria. *Journal of Machine Learning Research*, 3:1357–1370, 2003.

[18] D. S. Siroky. Navigating random forests and related advances in algorithmic modeling. *Statistics Surveys*, 3:147–163, 2009.

[19] G. Smits and M. Kotanchek. Pareto-front exploitation in symbolic regression. In U.-M. O'Reilly, T. Yu, R. L. Riolo, and B. Worzel, editors, *Genetic Programming Theory and Practice II*, chapter 17, pages 283–299. Springer, Ann Arbor, 13-15 May 2004.

[20] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307+, July 2008.

[21] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25+, Jan. 2007.

[22] K. Vladislavleva. *Model-based Problem Solving through Symbolic Regression via Pareto Genetic Programming.* PhD thesis, Tilburg University, 2008.

[23] K. Vladislavleva, K. Veeramachaneni, M. Burland, J. Parcon, and U.-M. O'Reilly. Knowledge mining with genetic programming methods for variable selection in flavor design. In *GECCO*, pages 941–948, 2010.

[24] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu. Advancing feature selection research − asu feature selection repository. Technical report, Arizona State University, June 2010.

[25] H. Zou and T. Hastie. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society B*, 67:301–320, 2005.