

Deductive Data Mining

Maxwell Hong, Ross Jacobucci, and Gitta Lubke
University of Notre Dame

Abstract

Data mining methods offer a powerful tool for psychologists to capture complex relations such as interaction and nonlinear effects without prior specification. However, interpreting and integrating information from data mining models can be challenging. The current research proposes a strategy to identify nonlinear and interaction effects by using a deductive data mining approach that in essence consists of comparing increasingly complex data mining models. The proposed approach is applied to 3 empirical data sets with details on how to interpret each step and model comparison, along with simulations providing a proof of concept. Annotated example code is also provided. Ultimately, the proposed deductive data mining approach provides a novel perspective on exploring interactions and nonlinear effects with the goal of model explanation and confirmation. Limitations of the current approach and future directions are also considered.

Translational Abstract

Increasingly large data sets provide the opportunity to search and find complex associations between predictors and an outcome. Data mining methods can serve to investigate even extremely large numbers of predictors simultaneously, however, the results of these methods are not necessarily easy to interpret. This is due to the fact that most data mining procedures are designed to predict future data (e.g., predicting e-mail spam or credit scores) rather than to clarify which predictors interact or have linear or nonlinear effects on an outcome. We provide a general modeling framework, called deductive data mining, for researchers to organize and interpret the results of data mining algorithms. Deductive data mining is shown to be an effective modeling guide in order to extract and select predicts and their specific effects. Empirical examples, analyses, and Monte Carlo simulations are provided to illustrate deductive data mining as well as annotated R code.

Keywords: deductive data mining, data mining, decision trees, model comparison

Statistical techniques for psychologists have undergone rapid development due to the ubiquity of larger and more diverse data sets, also known as “big data” (Harlow & Oswald, 2016). Traditional inference-based statistical methods such as multiple regression are limited when the number of predictor variables exceeds the sample size (Adjerid & Kelley, 2018). This limitation is only exacerbated in case the types of effects (linear, nonlinear, interactions) are not known. For instance, a single variable can be included in a regression model as a linear main effect. That variable may possibly require a transformation, such as taking the log of the variable, or require additional polynomial terms, such as quadratic or cubic. The effect of the same variable may also vary conditional on different levels of another predictor, and these interactions may have a nonlinear effect on the outcome. Even with as few as five predictors, the number of potential interactions


and nonlinear effects is often too large to test in standard multiple regression model. Issues that go along with analyzing big data will only increase in importance going forward because subfields in psychology, such as neuroscience, educational, developmental, and clinical research, are leveraging big data to answer research questions and enrich psychological theory (e.g., Ammerman, Jacobucci, Kleiman, Uyeji, & McCloskey, 2018; Miller, Lubke, McArtor, & Bergeman, 2016; Sinharay, 2016).

Data mining methods offer a viable alternative to traditional methods when the goal is to analyze the effects of many predictors simultaneously. However, it can be challenging to interpret results of data-driven algorithms because they are usually designed with an eye on prediction quality rather than an understanding of structural relations between variables. The overarching goal of this study is to provide practitioners a way to integrate data mining findings and translate results into meaningful interpretations. The deductive data mining (DDM) approach introduced in this article addresses the challenge to carry out a comprehensive exploratory analysis and translate the findings into a more parsimonious and interpretable model.

Confirmatory and Exploratory Analyses

Within the behavioral sciences, as well as within the field of data-mining, exploratory analyses are distinguished from confirmatory modeling strategies. Confirmatory models are sometimes

This article was published Online First January 9, 2020.

Maxwell Hong,  Ross Jacobucci, and Gitta Lubke, Department of Psychology, University of Notre Dame.

Ideas from this work were presented at the 2018 Society of Multivariate Experimental Psychology meeting and shared as a preprint on the Open Science Framework website, <http://dx.doi.org/10.31219/osf.io/bmwah>.

Correspondence concerning this article should be addressed to Gitta Lubke, Department of Psychology, University of Notre Dame, 390 Corbett Family Hall, Notre Dame, IN 46556. E-mail: glubke@nd.edu

also referred to as “explanatory models” in the data mining literature (Shmueli, 2010; Yarkoni & Westfall, 2017). Exploratory models are data-driven approaches that feature few, if any, prespecified effects. Confirmatory models serve to estimate effects that are specified based on theoretical considerations (Hastie, Tibshirani, & Friedman, 2009; Yarkoni & Westfall, 2017). Data mining models are essentially data-driven, and can have two distinct goals, the prediction of future outcomes and the selection of important predictors. In that latter sense, prediction models in data mining are similar to exploratory models in the behavioral sciences such as exploratory factor analysis where one of the goals is to identify which questionnaire items load on which factor. Although data mining models can serve to select important predictors and can accurately predict new data (e.g., random forests, Breiman, 2001) these models usually lack a clear interpretation of the interrelations between predictors.

The main emphasis in the behavioral sciences has been on building hypothesis or theory-driven confirmatory models, because a confirmatory model is usually straightforward to understand and interpret (AERA, APA, & NCME, 2014). The focus is on the interrelations of variables and model interpretability rather than on predicting new outcomes. Interpretability and optimizing prediction are not necessarily compatible. Consider for example single decision trees and ensembles of trees such as random forests or boosted trees. Decision trees are a powerful tool because they iteratively partition the data to capture interactions and nonlinear effects in a data-driven way without the need to prespecify any effects. Single trees are easy to interpret in principle, but it is known that the model structure can vary considerably from sample to sample, thus undermining interpretation as well as prediction quality. In order to assuage variability, several trees can be combined through so-called ensemble methods such as bagging, random forests, and boosted trees (Breiman, 1996; Friedman, 2001; Strobl, Malley, & Tutz, 2009). However, ensemble methods lack a clear interpretation compared with a single tree.

Relying solely on confirmatory statistical tools limits a psychological researcher’s ability to detect novel effects, such as nonlinear effects and interactions between variables, because researchers have to specify each effect before estimating the model. Interactions and nonlinear effects are crucial research findings and have received much attention in the social sciences (Aiken & West, 1991). The DDM approach introduced in this article addresses the challenge to conduct a comprehensive exploratory analysis and translate the findings into a more parsimonious and interpretable model.

The Current Study

DDM utilizes the strategy of model comparisons. Model comparisons within the context of ANOVA or structural equation modeling (SEM) are a common statistical tools for social scientists to estimate important effects (Bollen, 1989; Maxwell, Delaney, & Kelly, 2018). Model comparisons in the data mining community have mostly aimed at optimizing prediction (Hastie et al., 2009). DDM aims to provide a bridge between confirmatory modeling and pure exploratory prediction modeling. On a conceptual level, DDM parallels traditional model building by comparing different models that feature increasingly complex types of effects. Some data mining models such as the lasso only permit linear effects,

whereas others such as boosted trees permit linear as well as nonlinear main and interaction effects. Using prediction error as a diagnostic tool, the comparison of these increasingly complex data mining models permits to identify which effects are important in predicting a univariate outcome.

It is important to note however that while on a conceptual level there is certainly a parallel to testing nested models in an ANOVA or SEM framework, DDM is not an inference based approach that would permit model comparisons of nested models. Nested models are defined in terms of one model’s parameters being a subset of the other model’s parameters. The data mining methods that are compared in DDM do not meet this requirement. Equally important, the data mining community emphasizes that model comparisons and model validation need to be conducted with different partitions of the data to avoid capitalization on chance and sample idiosyncrasies that would not generalize to new data (Efron, 2014; Hurvich & Tsai, 1990; Lubke et al., 2017). DDM is a three stage process, and the stages are carried out in different partitions of the data.

In the following sections, we first provide a detailed description of model comparisons in the context of data mining algorithms. The goals of each stage of DDM, and models used to achieve said goals, are discussed in detail in the next section. DDM is illustrated with a toy example and two simulations. Furthermore, we provide three empirical analyses. Finally, code for researchers that builds on the R packages, *gbm* (Ridgeway, 2007), and *caret* (Kuhn, 2008) are provided. We conclude with a discussion of the implications and future research directions.

Deductive Data Mining

DDM is a model comparison approach consisting of three stages: exploration, interpretation, and confirmation. In the exploration stage, several models are compared: a (regularized) regression model with linear main effects, simple tree stumps which permit linear and nonlinear main effects, and more complex trees which permit linear and nonlinear main and interaction effects. The models, the rationale for their inclusion in the comparison, and the metrics used to compare their performance are described in detail below. The second stage serves to select effects and arrive at an interpretable model. The third stage consists of fitting a confirmatory regression model that includes the selected effects. Note that we advocate to partition the available sample data into two partitions. Partition 1 is used in Stages I and II, whereas Partition 2 (a “hold-out set”) is used only in the confirmation stage. This partition is necessary to avoid capitalization on chance.

Descriptions and summaries of each stage of DDM are presented in Figure 1. We will use a simulated example based on the findings of Grube and Agostinelli (1999) as a toy example throughout the next section. The original study aimed to predict average frequency of drinking with expected consequences of drinking as well as demographic variables.

Stage I: Exploration

The first stage in DDM is the exploration stage. The goal is to identify which predictors in the data have which type of effect in predicting the outcome. The focus is on linear and nonlinear main

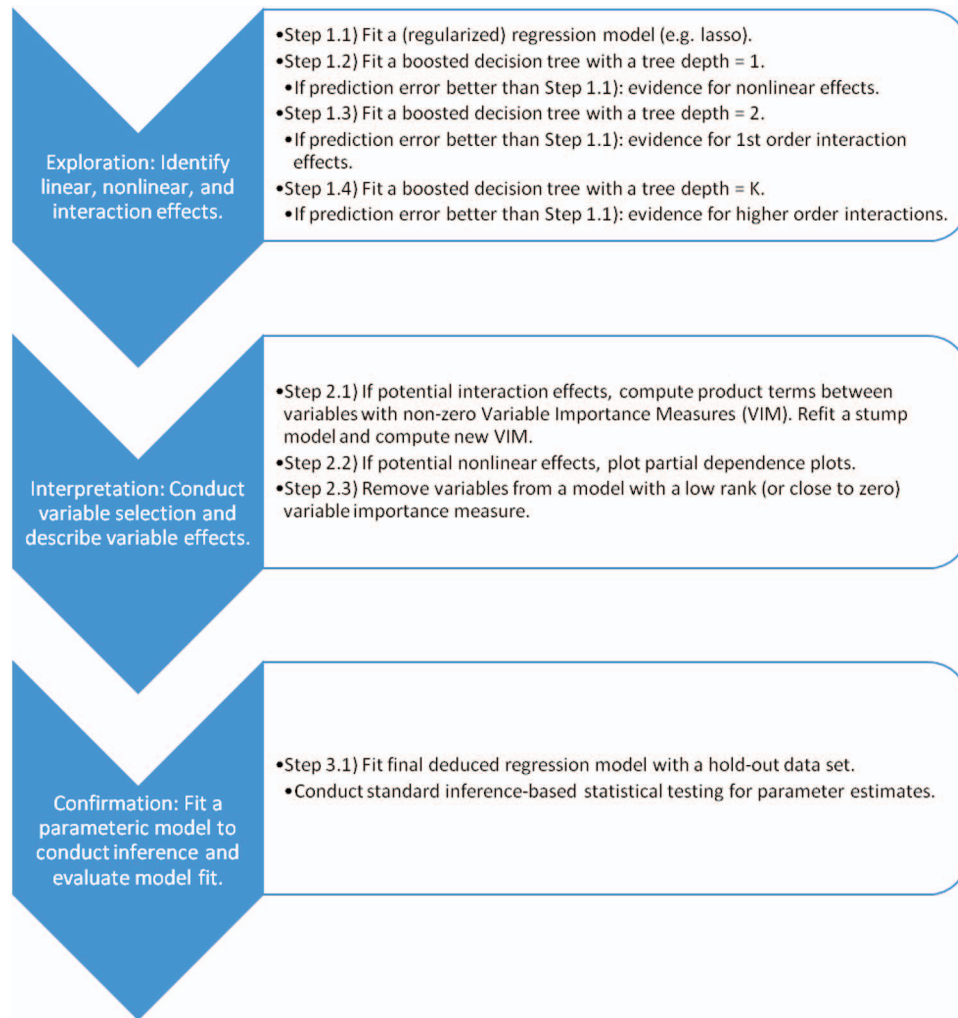


Figure 1. A flowchart of the deductive data mining (DDM) process. See the online article for the color version of this figure.

and interaction effects, and their importance in overall model prediction. DDM parallels traditional model building by starting with the most simple model (Maxwell et al., 2018). The increasingly complex models are compared using prediction error assessed with new data. For the current toy example, we present the prediction errors in Table 1.

A common metric used is the root mean square error (RMSE) or the proportion of out-of-sample explained variance R^2

Table 1
Stage I Prediction Error for Toy Example

Model	RMSE	R^2
Model 1: Main effect only regression model	1.029 (.049)	.081 (.028)
Model 2: Stump model	.989 (.051)	.147 (.039)
Model 3: Boosted decision tree (tree depth = 2)	.973 (.035)	.178 (.054)
Model 4: Boosted decision tree (tree depth = 5)	.975 (.051)	.175 (.065)

Note. RMSE = root mean square error. The standard deviation of over the cross-validated samples are in parenthesis.

(Kvalseth, 1985). RMSE can be thought of as an indicator of the discrepancy of an individual prediction on the raw scale and R^2 can be thought of as a more interpretable statistic for model accuracy. The prediction error for the exploratory phase is done with 10-fold cross-validation for all models using Partition 1 of the sample data.

Model 1. The first model fit to the data is a linear regression model that includes all available predictors. It is important to note if there are a larger or equal number of predictors relative to the sample size, then the regression has to be regularized using either a ridge regression approach or the least absolute shrinkage and selection operator (lasso) regression (Hoerl & Kennard, 1970; Tibshirani, 1991). The (regularized) simple regression model is used as the baseline model. Although this model only includes main effects of all available predictor variables, if a researcher has any theory, such as a specific nonlinear or interaction effects, they may include them during this stage in the original specified regression model. The inclusion of known effects can improve the exploratory search within DDM.

Using data simulated based on Grube and Agostinelli (1999), we fit a linear regression model with only linear main effects to predict frequency of alcohol units consumed. There were six predictors: age, gender, ethnicity (0 = White, 1 = non-White students), perceived negative consequences (e.g., getting a hangover), perceived social interaction (e.g., easier to talk to individuals), and affective behavior (e.g., feeling happy). All continuous variables were standardized prior to analysis. Using 10-fold cross-validation, the regression model achieved a prediction error with an RMSE of 1.029 ($SD = 0.049$) and R^2 of 0.081 ($SD = 0.028$).

Models 2 and 3. The second and third models are tree models. Decision trees are built iteratively by partitioning the data such that each partition is more homogenous with respect to the outcome. Similar to stepwise regression, the best predictor is selected using a criterion such as least squares. Instead of estimating a regression coefficient, in a tree model the best cutpoint is selected to partition the data into two mutually exclusive subgroups. In the alcohol example, suppose “perceived social interaction” is most strongly related to frequency of drinking, then the sample would be split on this variable into groups that score either high or low, and these groups would as a result be more homogenous with respect to frequency of drinking than the entire sample before the split. Note that a single split represents an approximation to a main effect, albeit crude. Trees with a single split are called stumps. The splitting can be repeated iteratively, resulting in a tree structure (see Figure 2). A subsequent split (e.g., “negative expectations” in Figure 2) is a conditional effect, and can approximate an interaction. In the alcohol example depicted in Figure 2, a person on average drinks 1.9 units of alcohol at one occasion. Increased affective state is the first important predictor of increased alcohol intake. If a participant scored higher than the cutoff of 0.56 (on a standardized scale) and, in addition does not perceive negative outcomes (hangover) as important, then the number of drinks is increased on average to 2.6 units.

The structure of a single tree can be highly variable due to sampling fluctuation. A different sample might result in a different first splitting variable, or different cutpoint, which alters the tree structure. To overcome this limitation, *ensemble methods* can

provide a more stable prediction. Examples are random forests or boosted trees. The latter are used here. The boosted tree model is an additive model of trees. For continuous outcomes the boosting process can be understood as follows. A tree is first fitted to the outcomes, and the residuals for each observation are computed. The next tree is fitted to these residuals, and added to the first tree. Again, the residuals are computed, and serve as the outcome to fit the third tree. The final predictions are created by averaging over all fitted trees. The difference in the predictions from a single decision tree or an ensemble are presented in Figure 3. The complexity of boosted decision trees mainly depends on three parameters: (a) the number of splits in a tree, d ; (b) the number of trees M ; and (c) the step-size or learning rate. The learning rate and number of trees are optimized for each boosted decision tree ensemble. One can tune across a large numbers of trees, such as from one to 5,000, and step-size rate, such as from 0.001 to 0.1. Unless specified otherwise, the number of trees and step-size in this study were evaluated across these values.

In the context of DDM, the maximum number of consecutive splits, d , is an important parameter. It is sometimes referred to as *interaction depth* because it determines the maximum order of interaction that can be approximated. As mentioned above, a tree with a depth of one is called a *stump* model and can only detect main effects as it does not permit any conditional splits. A stump can approximate nonlinear main effects because multiple trees are added to the model which may feature the same splitting variable but different split points. With the alcohol example using a tree depth of one, the prediction quality of the model has an R^2 0.147 ($SD = 0.039$) and RMSE 0.989 ($SD = 0.051$). The prediction quality of the stump model is better than the linear regression model, which suggests that there may be possible nonlinear main effects. Note that we do not yet know which of the predictor variables have a nonlinear effect.

Increasing d to two can maximally detect a first order interaction effect. Comparing trees with different interaction depths provides a rich picture of different nonlinear effects and interactions in data. The second model in DDM is a boosted decision tree model with a tree depth fixed to two. Compared with the stump, this model permits approximations to all possible linear and nonlinear main and first order interaction effects. The prediction quality of the model has an R^2 of 0.178 ($SD = 0.054$) and RMSE of 0.973 ($SD = 0.035$). Because the prediction quality has improved over both the stump model and the simple linear regression model, there is evidence that there is an interaction in the data between two variables. However, similar to the stump model, we do not know which variables interact with each other.

The fourth model is a boosted decision tree where the tree depth is larger than two, thereby including possible higher order interactions. Note that a tree depth of five means that the tree is allowed to have a total of five splits. Theoretically, such a tree would be able to capture fifth order interaction effect in the data. In practice, the order of interaction that is captured by a tree with a total of five splits is lower than five because other main effects will contribute to the tree structure. In our experience, a tree depth of five may be necessary to reliably capture two-way interactions, but this depends on the size of the main and interaction effects in the data. Given the overall goal of interpretability we focus on lower order interactions which and also be visualized in plots. A boosted decision tree depth of five fitted to the alcohol data has an R^2 0.175

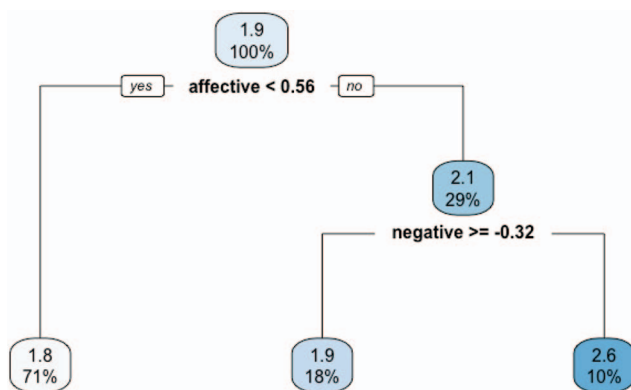


Figure 2. An example of a single decision tree when predicting frequency of drinking. The average number of drinks and sample size percentage are within each node. A split is described within each branch based on a single predictor (i.e., affective or negative expectations). See the online article for the color version of this figure.

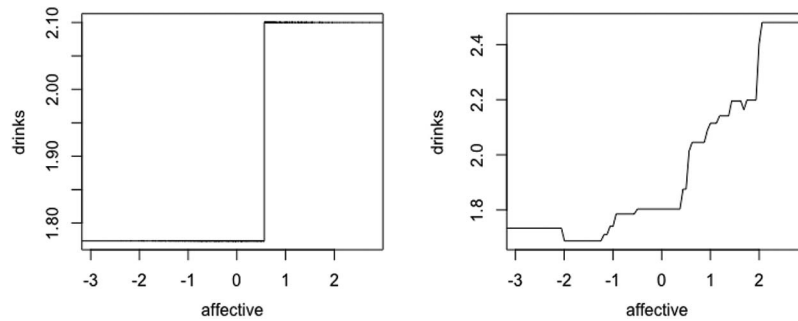


Figure 3. An example of predictions based on a single decision tree or a set of boosted decision trees when predicting frequency of drinking. The left panel shows a single decision tree that approximates a function using a step function. The right panel shows how boosted decision trees can be used to approximate a function by generating a smooth curve.

($SD = 0.065$) and RMSE 0.975 ($SD = 0.051$), which is about same prediction quality when the tree depth is two.

In sum, the exploration stage identified important nonlinear and interaction effects. However, the exploration stage does not provide any context on how the variables interact or what transformations of the variables are necessary to capture nonlinearities. The next stage serves to identify specific predictor effects and achieve interpretability.

Stage II: Interpretation

This stage aims to identify which predictors and their effects contributed to the model that had the best prediction performance in Stage I. There has been development on probing nonlinear effects and interaction terms within the data mining community (Elith, Leathwick, & Hastie, 2008; Friedman, 2001; Friedman & Popescu, 2008), and interpretable data mining is an ongoing field of interest (Caruana, Herlands, Simard, Wilson, & Yosinski, 2017). In DDM, we use variable importance measures (VIM) to identify which predictors contributed most to the performance of the best model, and an additional model comparison step to identify specific interactions.

VIM aim to quantify the contribution of each predictor to the overall prediction of the outcome. At any given split, the explained variance of the outcome can be quantified using a measure such as *MSE*. VIMs aggregate these quantifications over all splits for each predictor and trees. The result is a ranking of predictors according to their contribution to the overall explained variance.

The VIM of the six predictors in the alcohol example are shown in Table 2. The most important predictors were expected outcomes from drinking (affective, negative, and social) and age. Ethnicity and gender do not seem to play a major role in these data. Because VIMs aggregate marginal and interaction effects they do not reveal the type of effect (Auret & Aldrich, 2011; Kim & Kim, 2007; Miller et al., 2016). We know from the first stage of DDM that a model that permits nonlinear main effects and low order interactions has a better predictive performance. In the second stage of DDM, after identifying important predictors using VIM, an additional model is fitted to identify specific interactions. This is a stump model but fitted to an augmented data set in which centered cross-product variables are added of all predictors with VIMs larger than 0. Importantly, the new model will be able to disen-

tangle between main and interaction effects as it provides individual VIMs for each of the specified interaction terms.

In the alcohol example, we augmented the data with product terms of all variables. In the current example, we find that both negative and affective expectations interact and so does social and affective expectations. In cases with only a few important predictors, one may consider visual representations of nonlinear and interaction effects such as partial dependence plots (Friedman, 2001). Partial dependence plots depict the partial effect of a subset of variables given fixed values of all other variables. Such visualizations are usually limited to low-dimensional arguments, where one or two predictors are plotted in one graph. Although a set of partial dependence plots seldom capture a comprehensive depiction of the approximated function, they remain useful clues for low-order interactions and nonlinear effects (Friedman, 2001). We can plot the interaction and nonlinear effects with partial dependence plots in Figure 4. The plots shows how people drink more when there were few negative consequences and when people expect to be happy. Moreover, people also drank more when they assumed they would be more social and happy. There also appears to be a curvilinear relation between negative expectations and number of alcohol units consumed. This finding suggests that low to medium negative expectations, such as a hangover, are associated with an increased number of units alcohol consumed.

In sum, Stage II of DDM aims to provide guidelines to which nonlinear and interaction effects should be considered in the confirmatory stage. The example suggests that the variables gender

Table 2
Stage II Variable Importance Measures (VIM) for the Toy Example

VIM rank	Stump model	Tree depth = 2	Augmented tree
1	Negative (1)	Negative (1)	Negative (1)
2	Affective (.374)	Affective (.525)	Negative \times Affective (.570)
3	Age (.317)	Age (.374)	Age (.365)
4	Social (.262)	Social (.340)	Affective (.312)
5	Gender (.012)	Gender (.018)	Social (.300)
6	Ethnicity (.000)	Ethnicity (.000)	Social \times Affective (.287)

Note. The VIMs for the corresponding variable are in parenthesis.

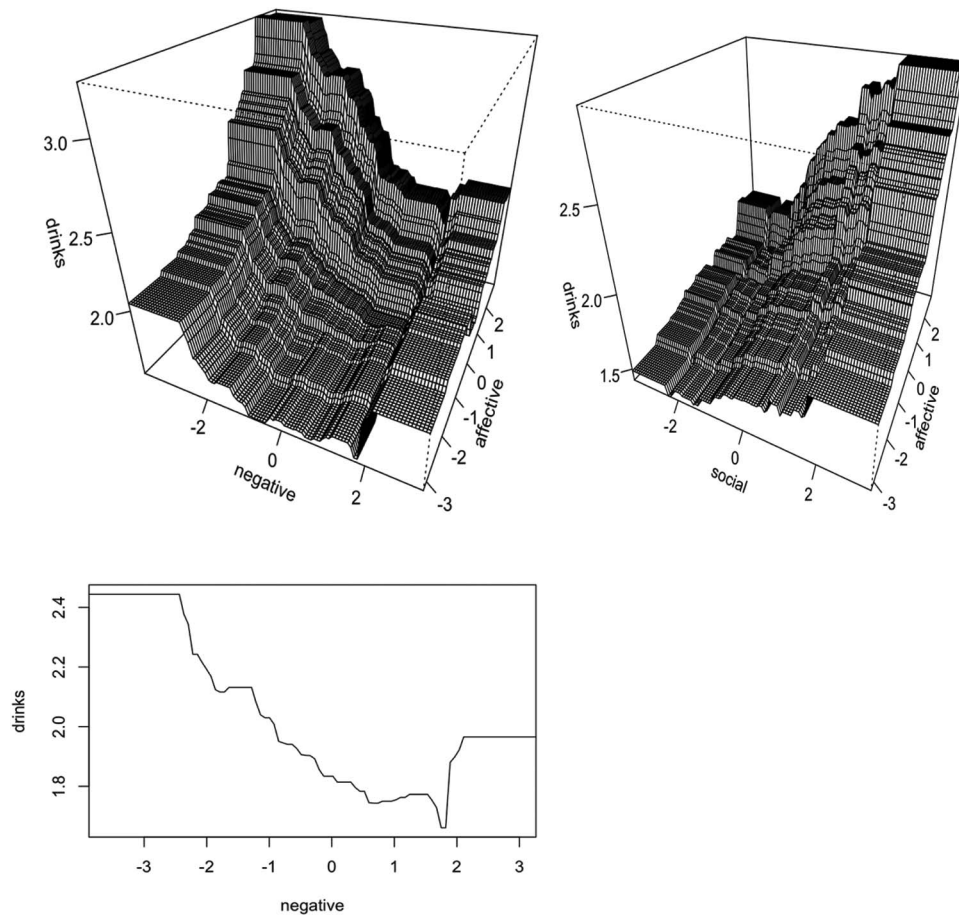


Figure 4. Example partial dependence plots of interactions and nonlinear effects. In the top two panels, there is evidence that there is an interaction between negative and affective expectations and interaction between social and affective expectations. In the bottom panel, there is a curvilinear relation between negative expectations and number of drinks one consumes.

and ethnicity are not playing an important role in predicting alcohol consumption. Furthermore, a nonlinear effect was identified between negative experiences and the outcome as well as two interaction effects.

It is important to note that in Stages I and II of DDM, multiple models have been fit to the data. Model comparisons and fitting a confirmatory model should not be conducted on the same data set in order to avoid capitalization on chance of finding an effect (Efron, 2014; Hurvich & Tsai, 1990; Lubke et al., 2017). Therefore, we propose to use a holdout set for the last stage of DDM, which focuses on confirmation.

Stage III: Confirmation

The third and final stage of DDM is to provide standard measures of model fit and parameter inference. To this end a final regression model is fitted with predictors and their effects identified in the final step of Stage II. This is done using a holdout data set (i.e., data not used in Stages I and II). Table 3 presents the final identified model which can be easily interpreted by any social scientist.

It is important to note there is a trade-off for how much data is allocated to the model confirmation stage when conducting DDM (e.g., Mosier, 1951). Most data mining algorithms can approximate any complex effect, but may require a large amount of data in order to achieve sufficient resolution (van der Ploeg, Austin, & Steyerberg, 2014). We expect sample size requirements for data

Table 3
Stage III Final Regression Model for Toy Example

Variable	B (SE)	p-value
(Intercept)	-.945 (.386)	.015
Age	.153 (.021)	<.001
Negative	-.111 (.021)	<.001
Social	.137 (.021)	<.001
Affective	.171 (.021)	<.001
Negative ²	.200 (.014)	<.001
Negative × Affective	-.138 (.021)	<.001
Social × Affective	.131 (.021)	<.001

Note. Negative² = squared term of negative.

mining algorithms likely to exceed sample sizes needed for complex parametric models. A recent study applied DDM to a sample of 62,227 parent ratings of childhood aggression, which was clearly sufficient to detect rather small interaction effects (e.g., standardized regression coefficients in the confirmatory four-group model around 0.02; Hendriks, Lunningham, Jacobucci, Hong, & Lubke, 2019). Future larger simulations are necessary to provide clear guidelines with respect to sample size requirements.

In the alcohol example, we find that the linear main effects identified by DDM (age, negative, social, and affective expectations), are statistically significant at the .05 level. The final deduced regression model was overall significant, $F(7, 1992) = 69.91$, $p < .001$ and R^2 of 0.197. Moreover, the regression coefficients that correspond to deduced nonlinear and two interaction effects were statistically significant. The final deduced model uncovered three complex effects and extended the original linear main effects only model. For instance, DDM uncovered how excessive negative expectations was found to not reduce drinking behavior. Moreover, the effect of both negative and social expectations depended on expected affective state.

In summary, DDM provides a guideline to determine what effects and which predictors are important to consider in a confirmatory model that is fitted to a hold-out set. The next section provides a set of small simulations, which is followed by three empirical examples that further illustrate how a model comparison approach can be implemented in practice.

Simulations

The alcohol example provided an illustration how comparing different data mining models in DDM can result in the specification of an interpretable regression model. One advantage of DDM is that it is able to identify specific nonlinear and interactions effects that would be difficult to identify in a confirmatory model due to the sheer number of potential effects. With simulated data, we aim to provide further evidence for this claim. In particular, we demonstrate how DDM can be used to identify different nonlinear effects, such as sinusoidal patterns or polynomials of higher terms. Furthermore, we also demonstrate that DDM can be used to identify higher order interaction effects, such as a three-way interaction term.

Simulation 1: Nonlinear Main Effects

The first simulation illustrates how DDM can identify different nonlinear main effects. The data generated was from a linear model:

$$y_i = \beta_1 F(x_{i1}) + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i$$

for $i = 1, \dots, n$ individuals. $F(\cdot)$ denotes a transformation of a predictor, such as a quadratic effect, that maps the predictor to the outcome variable and then standardized to have comparable effect sizes. The predictor variables were simulated from a standard normal distribution and the error term was manipulated to maintain a signal to noise ratio (SNR) of 2:1, (i.e., $SNR = \frac{\text{var}(\beta_0 + \beta_1 F(x_{i1}) + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5})}{\text{var}(\epsilon_i)}$). The weights for each predictor variable were generated such that that $\beta_1 = \beta_2 = \beta_3 = 0.3$ and β_4 and β_5 were noise variables. We manipulated the effect of the first variable to be either a linear, quadratic, cubic,

exponential, or sinusoidal function on the outcome variable. In sum, we had five data generation schemes. We generated 100 data sets for each condition and fit two models to the data, a boosted regression tree model or a linear regression model with no transformed predictors. It is important to note it does not make a difference whether one uses a stump model or varies tree-depth for the purposes of probing nonlinear effects. A stump model is able to capture nonlinear effects as long as there are enough iterations in the boosted tree algorithm. For completeness, we did both approaches and found no difference in prediction quality.

From each model, we computed the cross-validated RMSE and averaged across simulations and present them in Table 4. From these results, there are obvious differences to whether a linear approximation can be used to capture a nonlinear effect. If the true effect is linear, then the regression and boosted decision tree model have the same prediction error. Depending on the nonlinear effect, the regression tree may approximate the nonlinear effect with varying degrees of success. For instance, a quadratic effect would not be able to be approximated by a linear regression model unless specified the regression model. However, other effects such as an exponential effect may be better approximated. This idea is reaffirmed if we plot the partial dependence of x_1 on y in Figure 5.

The results show how the prediction error of a linear model has worse prediction error compared with a boosted decision tree. This finding confirms the proposed DDM approach to compare prediction error across models. Moreover, the discrepancy in prediction error depended on the true nonlinear effect. For instance, a quadratic effect would rarely be approximated using a linear model whereas an exponential effect can be. Moreover, identifying nonlinear effects using partial dependence plots is shown to be a powerful method to detect varying nonlinear effects in DDM.

Simulation 2: Higher Order Interactions

The second simulation illustrates how boosted regression trees coupled with a model comparison approach can handle higher order interactions. Data for this simulation were generated using a linear model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i1} x_{i2} x_{i3} + \epsilon_i$$

The predictors were simulated from a standard normal distribution and the signal to noise ratio was maintained at 2:1. However, the weights for the predictors were changed so that $\beta_1 = \beta_2 = 0.3$ and $\beta_4 = 0.5$, with other predictor main effects fixed to 0. The interaction effect, β_6 , was set to 0.0 or 0.3. In total, we have two data generation schemes to probe how DDM can approximate higher order interactions. We generated 100 data sets for each condition and fit the following models: Model 1—a regression model, Model 2—a stump model, Model 3—a boosted decision tree with tree depth fixed to two, Model 4—a boosted decision tree

Table 4
Average RMSE for Simulation 1

Model	Linear	Quadratic	Cubic	Exponential	Sinusoidal
Regression	.26	.61	.43	.43	.33
Boosted tree	.25	.26	.28	.28	.25

Note. RMSE = root mean square error.

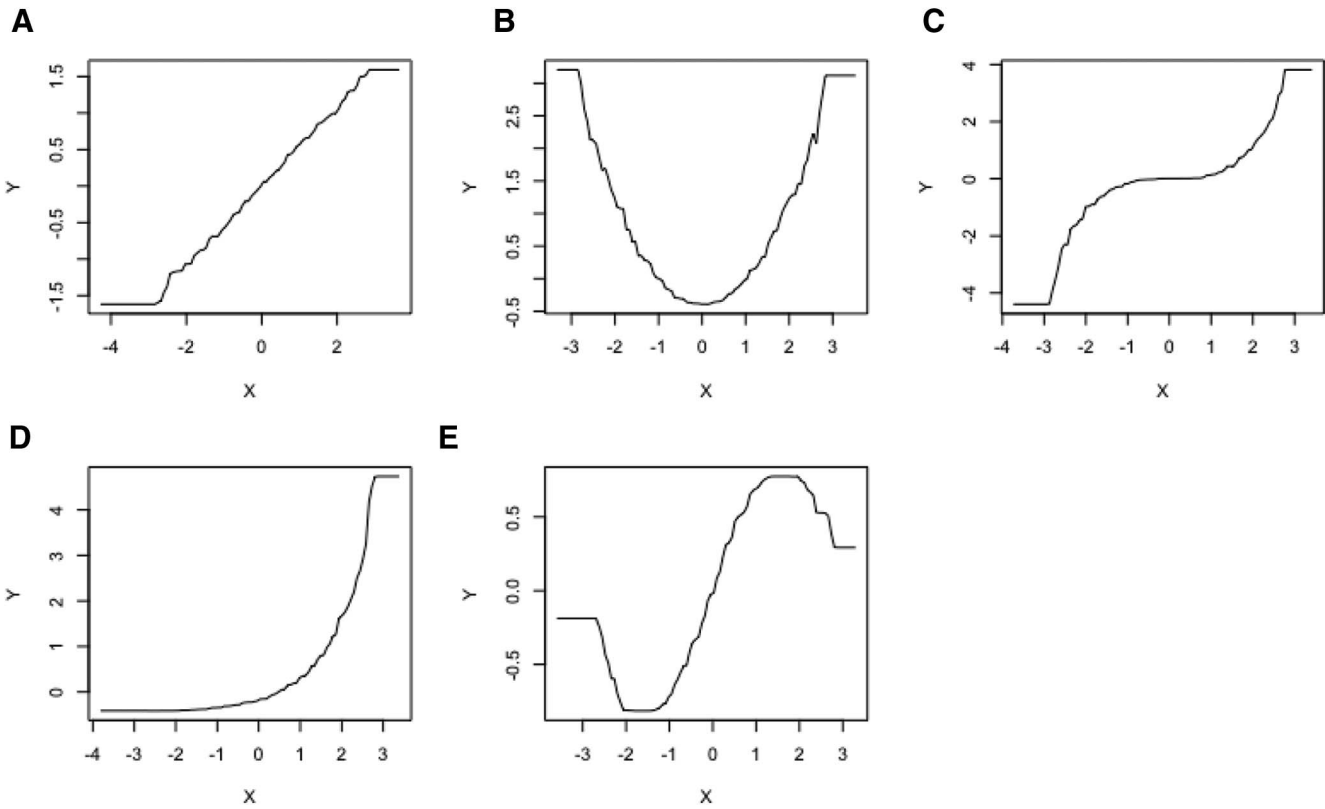


Figure 5. Partial dependence plot for a linear and a variety of nonlinear main effects of the variable, x_1 , against the outcome variable, y .

with tree depth fixed to 3, Model 5—a boosted decision tree model where the algorithm decides the optimal tree depth, Model 6—an augmented boosted decision tree using all possible three-way interactions.

We calculated the average RMSE for each simulation condition and present it in Table 5. When there is no interaction effect in the model, we can see that the prediction error is approximately the same across different models. However, if there is a three-way interaction, then Models 1–3 have worse prediction error because they were unable to capture the three-way interaction effect. Models 4–6 have similar prediction error and give evidence that there is a three-way interaction effect. We also compute the VIMs for each model and present them in Table 6. If one compares the VIM from Model 2–5, it is clear that there is some interaction in the data due to the change in the VIM ranking of x_3 . However, it is not clear which variables it interacts with. Model 6, or the augmented stump model, is able to provide estimates that disentangle between the

marginal effects of x_1 and x_2 and the three-way interaction between x_1 , x_2 and x_3 .

This simulation demonstrates how one can deduce that a three-way interaction exists using DDM through prediction error and VIM. Furthermore, the simulation demonstrated how an augmented decision tree can disentangle the three-way interaction and other main effects. In sum, the simulations provide further evidence on how DDM can be used to explore more complex effects that would not be feasible to identify in stepwise regression due to the vast number of complex effects. DDM offers a general framework where new effects can be discovered which were not originally anticipated in the data. The next section provides three empirical examples that illustrate how a model comparison approach can be implemented in practice.

Table 5
Average RMSE for Simulation 2

Effect	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$\beta_6 = 0$.31	.31	.31	.31	.31	.31
$\beta_6 = 0.3$.51	.52	.52	.43	.40	.37

Note. RMSE = root mean square error.

Table 6
Variable Importance Measures (VIMs) From Simulation 2

Variable	Model 2	Model 3	Model 4	Model 5	Model 6
x_1	.38/.36	.38/.35	.63/.35	.50/.36	.35/.36
x_2	.38/.35	.38/.34	.60/.34	.50/.35	.36/.34
x_3	.02/.00	.02/.00	.34/.00	.25/.00	.00/.00
x_4	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
x_5	.00/.00	.00/.00	.11/.00	.05/.00	.00/.00
$x_1x_2x_3$	—	—	—	—	.40/.00

Note. The VIMs in each cell are placed where $\beta_6 = 0/\beta_6 = 0.3$.

Empirical Examples

In the following section, three empirical examples are presented to showcase how DDM can be used with empirical data. The first two examples focus on nonlinear effects and interaction terms, respectively. The third example demonstrates how to probe nonlinear effects and interaction terms in one sample and conduct parameter inference with new data. Example code that mirrors the empirical examples are provided in the [Appendix](#). The data sets used in the applied analysis are open source. The first and second examples are a reanalysis from the textbook by [Darlington and Hayes \(2016\)](#) and can be downloaded from their companion website (<http://afhayes.com/>). The data for the third example can be downloaded from the ICPSR database (www.icpsr.umich.edu/).

Empirical Example 1: Nonlinear Effects

The following example illustrates how to probe nonlinear effects using DDM. The data comes from a nationally representative survey of 340 people living in the United States. In this analysis, exploration and interpretation are conducted. Confirmation was not done due to the small sample size. The goal of the analysis is to show how the predictor, frequency of traditional news (news), is related to the outcome variable, political knowledge. Intuitively, the original idea was that the more news one consumes, the more knowledgeable one is about politics ([Darlington & Hayes, 2016](#)). Both the outcome and variable of interest are continuous variables. Two other covariates were also included in the model: age and gender. All continuous predictors are standardized before analyzing the data. The outcome variable was not standardized because the original metrics were considered meaningful by the original analysis. All models' prediction error in terms of RMSE and R^2 was assessed using fivefold cross-validation and are presented in [Table 7](#).

Exploration. The first model fit to the data is a main effects only regression model. The RMSE of this model is 4.150 and R^2 of 0.113. The simplest model is able to explain about 11% of the variance in the outcome variable. The second model fit to the data is a stump model. This model did have smaller RMSE (4.069) and more variance explained (0.136) compared with the main effect only regression model. This finding suggests that there may be an important nonlinear marginal effect in the data. For completeness, we varied the tree depth from one to 10. The optimal tree depth was four. The RMSE for the third model (4.040) was indeed smaller than both the stump and main effects regression model. Moreover, the proportion of variance explained was also larger (0.150). This finding suggests that there may be an interaction effect in the data up to the fourth order. In summary, the exploratory stage suggests that there may be an interaction effect up to

the fourth order that is important in the data. Moreover, there is also evidence that there are nonlinear effects.

Interpretation. Because the purpose of this analysis is to probe nonlinear effects, we revisit the stump model and plot the partial dependence of traditional news against political knowledge. From [Figure 6](#), we can infer that there may be a nonlinear main effect of news on political knowledge, such as quadratic effect. It is clear that there is most likely a nonlinear effect given the points of the data. The stump model is only approximating the nonlinear relation between the variable news and political knowledge. To obtain a better approximation, one would need to collect more data for the exploration stage.

A separate analysis would need to be done on a holdout for confirmation. Due to the small sample size, parameter inference was not conducted in order to provide more information during the exploration stage. New data would need to be collected to verify the quadratic effect in the model to see if the nonlinear transformation of the news variable is meaningful and replicable.

These findings demonstrate how DDM can be leveraged as an exploratory tool. In this case, political knowledge is originally assumed to have only a linear relation with news consumption using a simple regression model. However, DDM was able to uncover both that there is a nonlinear effect and a way to both visualize the effect and make it interpretable. In this case, excessive news consumption was shown to not be related to more knowledge about politics.

Empirical Example 2: Interaction Effects

The following empirical example demonstrates how DDM can probe interaction effects. The motivation for this analysis is to better understand how work exhaustion is related to work injuries for health care professionals ([Halbesleben, 2010](#)). Due to the small sample size, parameter inference was not conducted. In the study, 300 health care workers are asked about their physical and mental work-related exhaustion, where larger scores reflect more exhaustion. After a 6-month interval, participants are asked to rate their safety protocols during work-related activities. Other covariates were also measured, such as job tenure (i.e., experience as a health care worker), gender, and age. One theory was that safety must be related to how long a worker has been at the hospital and how tired they are during the job ([Darlington & Hayes, 2016](#)). However, no interaction was prespecified. All continuous predictors are standardized before analyzing the data. The outcome variable is left unstandardized because the original metric was again considered to be meaningful. The goal of this analysis is to investigate possible interactions pertaining to the variable work exhaustion. All models' prediction errors are presented in [Table 8](#).

Exploration. The first model fit to the data is a main effects regression model. The RMSE of this model was 0.955 and an R^2 of 0.206. The second model fit to the data is a stump model. The stump model had an RMSE of 0.982, which is larger than the main effect only regression model. Furthermore, the R^2 value was also smaller, at 0.181. This indicates that there is no evidence for nonlinear marginal effects. Furthermore, there may be several main effects that can be constrained to be linear that are poorly approximated by the stump model.

The third and fourth models fit to the data are boosted decision trees where the tree-depth was allowed to increase to 10. Given the

Table 7
Stage 1 Prediction Error for Empirical Example 1

Model	RMSE	R^2
Model 1: Main effect only regression model	4.150 (.110)	.113 (.071)
Model 2: Stump model	4.069 (.220)	.136 (.084)
Model 3: Boosted decision tree (tree depth = 4)	4.040 (.150)	.150 (.062)

Note. RMSE = root mean square error. The standard deviation of over the cross-validated samples are in parenthesis.

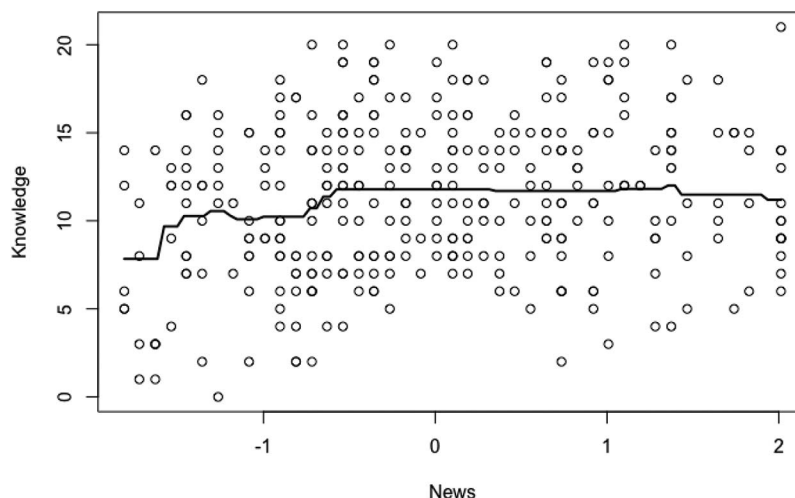


Figure 6. Partial dependence plot of news consumption on general political knowledge based on Empirical Example 1. News consumption is standardized so the axes correspond to standard deviation changes in the predictor variable. The predicted values of news consumption show a nonlinear effect.

difficulty of interpreting such an effect, one can also look at the prediction quality of models where the tree depth was constrained to be two. The RMSE for this model had some improvement compared with the main effects only regression model, but the R^2 value increases to 0.220 when the tree depth is two. The approximated interaction effects during this stage is not as large as the optimal boosted decision tree with a tree depth of six (R^2 value of 0.235).

In summary, the exploration stage suggests that there are no important nonlinear effects for the predictors. However, there is evidence for interaction effects.

Interpretation. In order to investigate which variables interact in the data, we first examine the VIMs in Table 9. The data are augmented to include all possible two-way interactions between the predictor variables that have VIM greater than 0: tenure, exhaust, and age. A stump model is fit to the augmented data set which is able to disentangle between main and interaction effects. The VIM suggest that the variables exhaust and tenure interact ($VIM = 0.593$). A partial dependence plot is used to provide visual guidance on what kind of interaction effect is in the data based on the model when the tree depth is fixed to two. Figure 7 shows the partial dependent plot of exhaust and tenure on safety. Clearly, there appears to be an interaction effect where lower work exhaust and more work tenure predicts improved work safety. This empirical example shows that DDM was able to uncover an interaction

effect between the exhaustion and tenure, with a large gain in work safety when workers are both less tired and have worked in their job for a longer period of time (Halbesleben, 2010).

Similar to the first empirical example, the sample size in this second example was not sufficient to portion the data and test the confirmatory model in a hold-out set. The third empirical example has a larger sample size, and we show how DDM can be used to probe nonlinear and interaction effects and test a final confirmatory model.

Empirical Example 3: Interactions, Nonlinear Effects, and Variable Selection

Data for the third empirical example comes from the Health and Retirement Study (HRS) 2010 year. Only complete cases were used in the dataset, which left 7,839 participants. Due to the large sample size, this dataset is viable to conduct both exploration, interpretation, and confirmation. Therefore, 5,000 subjects were used to train and compare models using 10-fold cross-validation during both the exploration and interpretation stage. A hold out sample of the remainder 2,839 were used for confirmation. The outcome variable for this analysis is depression scores measured by the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977). The 17 predictor variables for this analysis were chosen based on sets of previously grouped variables: socioeconomic status, health, family life, cognitive ability, and functional limitations (HRS, 2016). The goal of this analysis is to consider all predictors simultaneously, and arrive at a final confirmatory model that provides correct effect sizes. All subsequent analyses are based on standardized predictors. The outcome variable was left on its original scale. The prediction error for the third empirical analysis are presented in Table 10.

Exploration. The first model fit to the data is a main effect regression model using all 17 predictors. The RMSE for the linear regression model was 1.539 and R^2 of 0.270. The stump model had approximately the same prediction error as the main effect regression model, with an $RMSE = 1.540$, and $R^2 = 0.276$. This

Table 8
Stage I Prediction Error for Empirical Example 2

Model	RMSE	R^2
Model 1: Main effect only regression model	.955 (.030)	.206 (.071)
Model 2: Stump model	.982 (.030)	.181 (.045)
Model 3: Boosted decision tree (tree depth = 2)	.955 (.030)	.220 (.078)
Model 4: Boosted decision tree (tree depth = 6)	.949 (.060)	.235 (.067)

Note. RMSE = root mean square error. The standard deviation of over the cross-validated samples are in parenthesis.

Table 9
Stage II Variable Importance Measures (VIMs) for Empirical Example 2

VIM rank	Stump model	Tree depth = 2	Tree depth = 3	Tree depth = 6	Augmented tree
1	Tenure (1.00)	Exhaust (1.00)	Exhaust (1.00)	Exhaust (1.000)	Tenure (1.00)
2	Exhaust (.984)	Tenure (.884)	Tenure (.823)	Tenure (.823)	Exhaust (.714)
3	Age (.536)	Age (.524)	Age (.421)	Age (.617)	Exhaust \times Tenure (.593)
4	Sex (.000)	Sex (.000)	Sex (.000)	Sex (.025)	Age (.348)
5					Tenure \times Age (.137)
6					Exhaust \times Age (.050)
7					Sex (.00)

Note. The VIMs for the corresponding variable are in parenthesis.

provides some evidence for important nonlinear effects in the data. The next model fit to the data is a boosted decision tree with the tree depth fixed to two. There was a slight improvement in RMSE, 1.529, and R^2 at 0.279, which is larger than the original regression model. If we compare the boosted decision tree model with varying tree depth, we see little difference in prediction quality. Taken together, one can posit that there may be nonlinear effects or interactions effects in the data. However, they are most likely weak compared to the linear main effects in the model.

Interpretation. The first set of analysis aims to identify important variables. Table 11 presents the VIMs based of each fitted boosted decision tree model along with the augmented decision tree. In order to probe which variable age interacts with, an augmented data matrix with all linear product terms between variables that have VIMs greater than 0 in the previous models. The prediction error with the augmented data matrix had an RMSE of 1.544 and R^2 of 0.267. Furthermore, the VIM from the augmented data suggest that overall health interacts with motor skills and instrumental activities of daily living (IADLS). In addition, age was also found to interact with motor skills and IADLS.

In order to further probe possible interaction effects, joint partial dependence plots of the posited interaction effects are presented in Figure 8. Younger people and people with worse overall health and

have more difficulty completing IADLS activities (e.g., preparing a meal) have larger self-reported CES-D scores. Similarly, younger people and people with worse health who have more limitations with motor skills activities (e.g., walking across the room) have larger self-reported CES-D scores. Other interactions can be similarly interpreted. The predictors are standardized so the axes correspond to standard deviation changes of the predictor variable.

In order to perform variable selection, we removed variables with consistently low VIMs (close to 0). Four variables (the number of siblings one has, education level, mother's education level, and number of children) were found to be unimportant. We also visualized each linear main effect using partial dependence plots. Most effects were linear and suggested that no transformations were necessary.

In summary, the second stage revealed that nonlinear transformations of the predictor variables may be unnecessary. Moreover, there is evidence supporting four interaction effects, namely between the variables age and motor skills, age and IADLS, health and motor skills, and health and IADLS. Clearly the variables contributing to the model are a subset of the linear main effects. However, there are small contributions based on the identified interaction effects.

Confirmation. Next we evaluated the regression model using the hold-out set to obtain effect size measures for each parameter estimate in a confirmatory fashion. Standardized regression weights, their standard errors, and p values are presented in Table 12. The global F test was found to be statistically significant, $F(18, 2820) = 57.36$, $p < .001$. Furthermore, we can see that only two of the interaction effects involving age appear to be relevant using a .05 significance level. Furthermore, some main effect variables, such as length of marriage, also have smaller effects that are not statistically significant. Because the final model is fitted in a hold-out set, parameter estimates and measures of goodness of fit are not confounded by capitalization on chance.

Table 10
Stage I Prediction Error for Empirical Example 3

Model	RMSE	R^2
Model 1: Regression model	1.539 (.07)	.270 (.029)
Model 2: Stump model	1.540 (.05)	.276 (.055)
Model 3: Boosted decision tree (tree depth = 2)	1.529 (.06)	.280 (.043)
Model 4: Boosted decision tree (tree depth = 4)	1.529 (.06)	.279 (.032)

Note. RMSE = root mean square error. The standard deviation of over the cross-validated samples are in parenthesis.

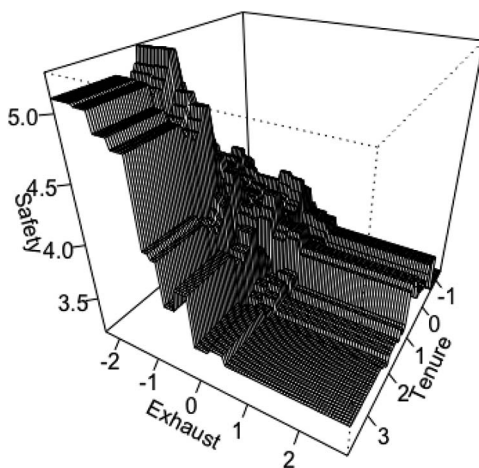


Figure 7. Partial dependence plot of work exhaustion and job tenure on job safety based on Empirical Example 2. Work exhaustion and job tenure are standardized so the axes correspond to standard deviation changes in the predictor variable. Less tired and more experienced workers have greater job safety.

Table 11
Stage II Variable Importance Measures (VIMs) for Empirical Example 3

VIM rank	Stump model	Tree depth = 2	Tree depth = 4	Augmented tree
1	Health (.1)	Health (.1)	Health (.1)	Health (.1)
2	Motor skills (.360)	IADLS (.374)	IADLS (.314)	Health × Motor Skills (.330)
3	IADLS (.349)	Motor skills (.373)	Motor skills (.305)	Age × Motor Skills (.239)
4	Employment (.152)	Employment (.179)	Age (.167)	Health × IADLS (.150)
5	Length of marriage (.11)	Age (.170)	Employment (.165)	Employment (.124)
6	Memory (.104)	Length of marriage (.144)	Length of marriage (.157)	Age × IADLS (.084)
7	Age (.096)	Memory (.119)	Chronic health conditions (.118)	Motor (.081)
8	Cognition (.085)	Chronic health conditions (.096)	Cognition (.100)	Memory (.709)
9	Mobility (.082)	Cognition (.095)	Memory (.097)	Mobility (.070)
10	Chronic health conditions (.067)	Mobility (.094)	Mobility (.087)	IADLS (.071)

Note. IADLS = instrumental activities of daily living. The VIMs for the corresponding variable are in parenthesis.

The third empirical example demonstrates how DDM leverages exploration and integrates results in a standard confirmatory setting. Previous analyses of the same data only considered main effects, and was limited to a subset of variables in a regression model. It is well-known that effect sizes are biased if not all relevant predictors are included in a model. DDM was able to reveal two interaction effects and eliminate four predictors from the initial set. The final model showed that age interacts with motor ability and daily activities.

Discussion

Identifying important complex effects, such as interactions and nonlinear effects, are vital for psychological theory. However, identifying complex effects can be daunting due to the sheer number of potential nonlinear and interaction effects even in a limited set of predictors. When there is little or no substantive theory that would suggest a complex effect, exploratory analyses need to be used to detect important effects in order to build substantive theory. Probing and interpreting larger data sources requires new methodology that is flexible and easy to use for social and behavioral scientists. Employing a DDM strategy provides a powerful approach to identify a suitable model for interpretation purposes.

In this article, we compared different data mining models and leveraged the flexibility of boosted decision tree models to approximate nonlinear effects and interactions. The proposed approach does not require any prior specification of effects. This flexibility is the key strength of data mining models. Unfortunately, boosted decision trees are notorious for being difficult to interpret and have been viewed as a “black-box” approach (Hastie et al., 2009). The difficulty of interpreting boosted decision trees makes the methodology hard to apply as a stand-alone method in behavioral science, especially when the goal is theory-building. The proposed DDM strategy consists of comparing increasingly complex models, thus permitting deduction of the type of effects as well as identification of the involved predictors.

Advantages Over Alternative Approaches

DDM has several advantages over alternative modeling strategies, such as stepwise regression. In previous research, VIMs based on tree ensembles were able to detect more predictors and provide stable rankings of predictors compared to stepwise

regression (Rossi, Amaddeo, Sandri, & Tansella, 2005). Strobl et al. (2009) also noted that stepwise procedures including interaction effects are limited normally to second order interactions. As demonstrated in our simulations, DDM is not restricted by this limitation. Moreover, it is rare in practice to evaluate multiple different types of nonlinear effects using stepwise regression. The process for testing all possible combinations of effects with each variable is simply too large given that parametric models commonly employ significance tests for variable selection. DDM overcomes this limitation by emphasizing prediction accuracy which is either done using k-fold cross-validation or is done with a hold out set. In that sense, DDM emphasizes the generalizability of results.

Limitations

DDM has several limitations, which point to potential future directions. First of all, it is important to note that the exploration stage requires cut-offs for what is and what is not a meaningful difference in prediction error. Although every statistical analysis has some degree of inherent subjectivity, a straightforward way to address the current limitation is to add another layer of resampling during the exploration phase by carrying out DDM on multiple bootstrap samples. This provides the opportunity to construct confidence intervals for all prediction errors and model descriptive statistics (Efron & Tibshirani, 1994). In case of missing data, one can use MICE for imputation, and carry out DDM in each of the imputed data sets (Van Buuren & Groothuis-Oudshoorn, 2011). This approach was recently used in a large two-cohort study investigating predictors of childhood aggression (Hendriks et al., 2019).

Another limitation of DDM is the sample size requirement that should be used for each step. A well-known issue within psychology is that many studies are underpowered (Maxwell et al., 2018). If a study does not have an adequate sample size, it may be difficult to reliably detect any effect. This problem is exacerbated in case of nonlinear or interaction effects. Sample size requirements for data mining methods with behavioral data have not yet been investigated, and findings from DDM depend on stable and consistent results. Large scale simulation studies would be able to address this shortcoming of the current article.

Furthermore, more research needs to be done on the effects of multicollinearity in data mining methods, specifically with boosted regression trees. As pointed out in Friedman and Popescu (2008),

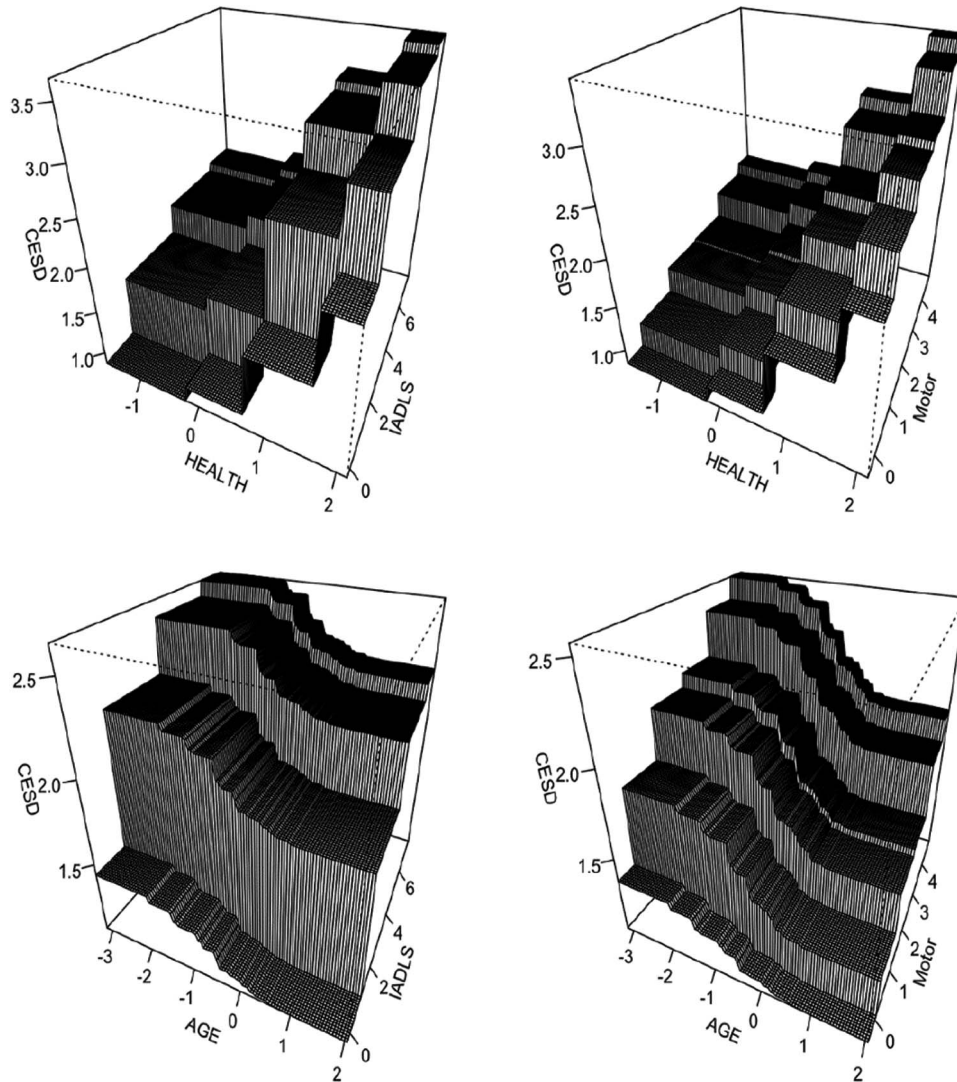


Figure 8. Partial dependence plot of instrumental activities of daily living (IADLS), motor skills, and age on Center for Epidemiologic Studies Depression Scale (CES-D) based on Empirical Example 3. Younger people who have more difficulty completing IADLS activities (e.g., preparing a meal) have larger self-reported CES-D scores. Younger people who have more limitations with gross motor skills activities (e.g., walking across the room) have larger self-reported CES-D scores. Other interactions can be similarly interpreted. The predictors are standardized so the axes correspond to standard deviation changes in the predictor variable.

it is possible that a large degree of collinearity among predictor variables would impact the interpretation of a single tree by biasing VIMs or when evaluating partial dependence. If the data has a large number of variables that are collinear, it is not possible to easily distinguish between interactions and nonlinear effects among a subset of variables. A few ways to overcome this issue is to modify the splitting rule by discouraging the entry of superfluous variables during each split (Friedman & Popescu, 2008). However, this method has not been rigorously tested nor implemented in current commercial software. Other extensions exist such as using a conditional VIM measure has been found to reduce bias in tree ensembles such as random forests (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). However, conditional VIMs have not been extended to a boosted decision tree framework.

Moreover, a subtle difference between random forests and boosted regression trees is that one typically tunes across a larger tree depth for random forests than boosted decision trees. Random forests also have a different set of parameters, such as the number of randomly selected variables considered at each split. If one tunes with a sufficiently large number of iterations in the boosted tree algorithm and considers a small range of the shrinkage parameter, it should reduce the problems of multicollinearity (Friedman, 2001). Another way to handle multicollinearity is to do some sort of dimension reduction prior to data analysis. For the empirical studies in the study, the authors computed a Pearson correlation matrix of the predictors and flagged variables that were highly collinear. Potential multicollinearity effects were eliminated by checking the face validity of each question to see if it makes sense

Table 12
Stage III Regression Model for Empirical Example 3

Variable	<i>B</i> (<i>SE</i>)	<i>p</i> -value
(Intercept)	1.164 (.032)	<.001
Employment	-.12 (.035)	.001
Income	-.042 (.032)	.184
Age	-.352 (.05)	<.001
Health	.363 (.037)	<.001
IADLS	.162 (.042)	<.001
Chronic health conditions	.063 (.037)	.084
Length of current marriage	-.02 (.037)	.594
Self-reported memory	.121 (.032)	<.001
Cognition score	-.092 (.032)	.003
Body mass index	-.126 (.031)	<.001
Number of people living in house	.056 (.031)	.074
Mobility index	.194 (.059)	.001
Gross motor skills	.192 (.071)	.007
Self-report belief to live 85+	-.074 (.031)	.016
IADLS × Age	-.061 (.03)	.042
Gross Motor Skills × Age	-.076 (.027)	.004
IADLS × Health	.025 (.035)	.476
Gross Motor Skills × Health	.039 (.033)	.236
<i>R</i> ²	.268	

Note. IADLS = instrumental activities of daily living.

that the two variables would be correlated and remove redundant variables with similar information.

A well-known general caveat of detecting interactions using any decision tree algorithm is the X-OR problem, where neither variables have a main effect (Strobl et al., 2009). The X-OR problem is caused by the fact that the boosted decision tree algorithm would never select one of the participating variables and the interaction effect would remain undetected. This is a larger issue in the decision tree literature, and our goal is not to propose a solution to this problem. One way to overcome this issue is to specify all possible linear interaction effects in the data a priori. However, preliminary work showed that this approach would be intractable due to computation reasons. When the number of predictors becomes really large the number of linear interaction combinations (i.e., $p(I-p)$) becomes so large that the computation time becomes unreasonable in practice to achieve a good prediction error. The problem is further exacerbated when one considers higher order interactions.

Future Directions

Given the general framework of DDM, one can consider including alternative data mining methods, such as the lasso with interactions and quadratic effects (implemented in the R package *hierNet*; Bien, Taylor, & Tibshirani, 2013), to overcome pitfalls due to other methods (Hoerl & Kennard, 1970; Tibshirani, 1991). We see this as an addition to the use of boosted trees with a depth of one (stumps) or two, because the hierarchical lasso limits interaction effects to be linear. Integrating the hierarchical lasso would then provide the opportunity to rule out of include nonlinear interaction effects. However, this should first be evaluated as part of a simulation. Finally, the simulations in the current study can be expanded to other conditions. The current simulations aimed at providing proof of concept, and further work is needed to evaluate the strengths and weaknesses of DDM in different data settings.

Future directions and extensions include the development of new statistics to probe nonlinear interaction effects. Although trees can naturally accommodate these kinds of effects, in practice they would be quite difficult to interpret. Current tools for identifying nonlinear interactions could be done by identifying all potential general interaction effects by using Friedman's H statistic and plotting the joint partial dependence for different sets of variables (Friedman & Popescu, 2008). However, this would be computationally demanding and interpreting the joint effect is nontrivial. Computationally efficient and descriptive tools that can detect and probe nonlinear interactions would be a natural next step when interpreting tree ensembles.

Conclusion

In summary, DDM provides researchers with a new framework to probe nonlinear and interaction effects for many predictors simultaneously, and arrive at a final confirmatory model for social and behavioral data. As a principled approach, DDM makes the process of theory building explicit, no longer requiring researchers to perform this step clandestinely (McArdle, 2013), with suboptimal procedures. We encourage researchers to detail and document each aspect of their analysis, taking full advantage of new statistical tools to perform an exploratory analysis. Instead of viewing DDM as a replacement for traditional tools, we see it as supplemental, stepwise process aimed at situations where researchers may lack a priori theory, and/or suspect the existence of nonlinear and interaction effects. The general framework makes the deductive approach a useful tool to discover and investigate more nuanced relationships that would be difficult or impossible to detect by using model selection with parametric models.

References

- Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, 73, 899–917. <http://dx.doi.org/10.1037/amp0000190>
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.
- Ammerman, B. A., Jacobucci, R., Kleiman, E. M., Uyeji, L. L., & McCloskey, M. S. (2018). The relationship between nonsuicidal self-injury age of onset and severity of self-harm. *Suicide & Life-Threatening Behavior*, 48, 31–37. <http://dx.doi.org/10.1111/sltb.12330>
- Auret, L., & Aldrich, C. (2011). Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems*, 105, 157–170. <http://dx.doi.org/10.1016/j.chemolab.2010.12.004>
- Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics*, 41, 1111–1141. <http://dx.doi.org/10.1214/13-AOS1096>
- Bollen, K. (1989). *Structural equations with latent variables*. New York, NY: Wiley. <http://dx.doi.org/10.1002/9781118619179>
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123–140. <http://dx.doi.org/10.1007/BF00058655>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <http://dx.doi.org/10.1023/A:1010933404324>

- Caruana, R., Herlands, W., Simard, P., Wilson, A. G., & Yosinski, J. (2017). Proceedings of NIPS 2017 Symposium on interpretable machine learning. *ArXiv Preprint ArXiv:1711.09889*.
- Darlington, R. B., & Hayes, A. F. (2016). *Regression analysis and linear models: Concepts, application and implementation*. New York, NY: Guilford Press.
- Efron, B. (2014). Estimation and Accuracy after Model Selection. *Journal of the American Statistical Association*, 109, 991–1007. <http://dx.doi.org/10.1080/01621459.2013.823775>
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.
- Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77, 802–813. <http://dx.doi.org/10.1111/j.1365-2656.2008.01390.x>
- Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29, 1189–1232.
- Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2, 916–954. <http://dx.doi.org/10.1214/07-AOAS148>
- Grube, J. W., & Agostinelli, G. E. (1999). Perceived consequences and adolescent drinking: Nonlinear and interactive models of alcohol expectancies. *Psychology of Addictive Behaviors*, 13, 303–312. <http://dx.doi.org/10.1037/0893-164X.13.4.303>
- Halbesleben, J. R. B. (2010). A meta-analysis of work engagement: Relationships with burnout, demands, resources and consequences. In A. Bakker & M. P. Leiter (Eds.), *Work engagement: Recent developments in theory and research*. London, UK: Routledge.
- Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, 21, 447–457. <http://dx.doi.org/10.1037/1037-00000120>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. *Elements*. Advance online publication. <http://dx.doi.org/10.1007/978-0-387-84858-7>
- Health and Retirement Study, (RAND HRS Longitudinal File) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (2016).
- Hendriks, A., Lunningham, J., Jacobucci, R., Hong, M., & Lubke, G. (2019). *Predicting childhood aggression: Mining large data followed by confirmatory models*. Manuscript submitted for publication.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67. <http://dx.doi.org/10.1080/00401706.1970.10488634>
- Hurvich, C. M., & Tsai, C.-L. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44, 214–217.
- Kim, Y., & Kim, H. (2007). Application of random forests to association studies using mitochondrial single nucleotide. *Genomics & Informatics*, 5, 168–173.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28, 1–26. <http://dx.doi.org/10.18637/jss.v028.i05>
- Kvalseth, T. O. (1985). Cautionary note about r^2 . *The American Statistician*, 39, 279.
- Lubke, G. H., Campbell, I., McArtor, D., Miller, P., Luningham, J., & van den Berg, S. M. (2017). Assessing model selection uncertainty using a bootstrap approach: An update. *Structural Equation Modeling*, 24, 230–245. <http://dx.doi.org/10.1080/10705511.2016.1252265>
- Maxwell, S., Delaney, H., & Kelly, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). New York, NY: Routledge.
- McArdle, J. J. (2013). Exploratory data mining using decision trees in the behavioral sciences. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 25–69). New York, NY: Routledge. <http://dx.doi.org/10.4324/9780203403020>
- Miller, P. J., Lubke, G. H., McArtor, D. B., & Bergeman, C. S. (2016). Finding structure in data using multivariate tree boosting. *Psychological Methods*, 21, 583–602. <http://dx.doi.org/10.1037/met0000087>
- Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement*, 11, 5–11. <http://dx.doi.org/10.1177/001316445101100101>
- Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1, 385–401. <http://dx.doi.org/10.1177/014662167700100306>
- Ridgeway, G. (2007). Generalized boosted models: A guide to the gbm package. *Compute*, 1, 1–12.
- Rossi, A., Amaddeo, F., Sandri, M., & Tansella, M. (2005). Determinants of once-only contact in a community-based psychiatric service. *Social Psychiatry and Psychiatric Epidemiology*, 40, 50–56.
- Shmueli, G. (2010). To explain or predict? *Statistical Science*, 25, 289–310. <http://dx.doi.org/10.1214/10-STS330>
- Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues and Practice*, 35, 38–54. <http://dx.doi.org/10.1111/emip.12115>
- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9, 307. <http://dx.doi.org/10.1186/1471-2105-9-307>
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14, 323–348. <http://dx.doi.org/10.1037/a0016973>
- Tibshirani, R. (1991). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, Methodological*, 58, 267–288.
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Multivariate imputation by chained equations. *Journal of Statistical Software*. Advance online publication. <http://dx.doi.org/10.18637/jss.v045.i03>
- van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, 14, 137. <http://dx.doi.org/10.1186/1471-2288-14-137>
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons From machine learning. *Perspectives on Psychological Science*, 12, 1100–1122. <http://dx.doi.org/10.1177/1745691617693393>

(Appendix follows)

Appendix

Example Code

```
##### Example Code #####

#### Libraries and other preliminaries
set.seed(12345)
require(caret)
require(gbm)

#### Training Data
train.data

#### Test Data
test.data

#### Training Error Specification
control <- trainControl(method="cv", number=10)

#### Model 1: Regression
mod1 <- train(y~., data = train.data, method='glm', trControl = control)

#### Model 2: Stump Model
gbmGrid <- expand.grid(interaction.depth = c(1),
  n.trees = seq(10,5000,5),
  shrinkage = c(0.1, .01, .05, .001, .005),
  n.minobsinnode = 10)
mod2 <- train(y~., data = train.data, method='gbm',
  tuneGrid = gbmGrid, trControl = control)

#### Model 3: kth order Boosted Regression Tree
k <- 2
gbmGrid <- expand.grid(interaction.depth = k,
  n.trees = seq(10,5000,5),
  shrinkage = c(0.1, .01, .05, .001, .005),
  n.minobsinnode = 10)
mod.3 <- train(y~., data = train.data, method='gbm',
  tuneGrid = gbmGrid, trControl = control)

#### Model 4: Fully Tuned Boosted Regression Tree
gbmGrid <- expand.grid(interaction.depth = 1:10,
  n.trees = seq(10,5000,5),
  shrinkage = c(0.1, .01, .05, .001, .005),
  n.minobsinnode = 10)
mod4 <- train(y~., data = train.data, method='gbm',
  tuneGrid = gbmGrid, trControl = control)
```

(Appendix continues)


```
#### Model 5: Informed Grid and Stump Model
gbmGrid <- expand.grid(interaction.depth = c(1),
  n.trees = seq(10,5000,5),
  shrinkage = c(0.1, .01, .05, .001, .005),
  n.minobsinnode = 10)

#### Including an interaction term
train.inf <- train.data
int <- train.inf[,1]*train.inf[,2]
train.inf <- cbind(int = int,train.inf)

mod.5 <- train(y~., data = train.inf, method='gbm',
  tuneGrid = gbmGrid, trControl = control)

#### Model 6: Regression Model With Specified Effects
#### Including Interaction Effect
mod.6 <- train(y~., data = train.inf, method='glm', trControl = control)

#### Model 7: Regression Model on Final Data Set
#### Specify Interaction
int <- test.data[,1]*test.data[,2]
test.data <- cbind(int = int,test.data)

mod.7 <- lm(y~., data = test.data)
summary(mod.7)

#### Model Comparison of Prediction Error
compare <- resamples(list(mod.1 = mod.1, mod.2 = mod.2, mod.3 = mod.3,
  mod.4 = mod.4, mod.5 = mod.5, mod.6 = mod.6, mod.7 = mod.7))

summary(compare)

#### Variable Importance Example
mod1.f <- mod.1$finalModel
sort(relative.influence(mod1.f, scale = T),T)

#### Partial Dependence Plot
#### Marginal Plot
plot.out <- plot(mod1.f, i.var = "X1", return.grid = TRUE)
plot(plot.out, type = 'l', lwd = 1, xlab = 'X1', ylab = 'Y')

#### Joint Plot
grid <- gbm::plot.gbm(mod1.f, i.var = c("X1","X2"),perspective = TRUE, return.grid = TRUE)
x <- as.numeric(unique(grid[, 1]))
y <- as.numeric(unique(grid[, 2]))
z <- matrix(grid[, 3], length(unique(x)), length(unique(y)))

persp(x = x, y = y, z = z, theta = 25, phi = 30,
  ticktype = "detailed",
  zlab = 'Y', xlab = 'X1', ylab = 'X2')
```

Received February 4, 2019
Revision received November 7, 2019
Accepted November 18, 2019 ■