Buskirk, Trent D., Antje Kirchner, Adam Eck, and Curtis S. Signorino. 2018. "An Introduction to Machine Learning Methods for Survey Researchers." *Survey Practice* 11 (1). https://doi.org/10.29115/SP-2018-0004.

ARTICLES

An Introduction to Machine Learning Methods for Survey Researchers

Trent D. Buskirk[1], Antje Kirchner[2], Adam Eck[3], Curtis S. Signorino[4]

[1] Center for Survey Research and Department of Management Science and Information Systems, UMass Boston, [2] Survey Research Division, RTI International, [3] Department of Computer Science, Oberlin College, [4] Department of Political Science, University of Rochester

Keywords: machine learning, adaptive lasso, neural networks, classification and regression trees, random forests

https://doi.org/10.29115/SP-2018-0004

Survey Practice

Vol. 11, Issue 1, 2018

Machine learning techniques comprise an array of computer-intensive methods that aim at discovering patterns in data using flexible, often nonparametric, methods for modeling and variable selection. These methods offer an expansion to the more traditional methods, such as OLS or logistic regression, which have been used by survey researchers and social scientists. Many of the machine learning methods do not require the distributional assumptions of the more traditional methods, and many do not require explicit model specification prior to estimation.

Machine learning methods are beginning to be used for various aspects of survey research including responsive/adaptive designs, data processing and nonresponse adjustments and weighting. This special issue aims to familiarize survey researchers and social scientists with the basic concepts in machine learning and highlights five common methods. Specifically, articles in this issue will offer an accessible introduction to: LASSO models, support vector machines, neural networks, and classification and regression trees and random forests. In addition to a detailed description, each article will highlight how the respective method is being used in survey research along with an application of the method to a common example.

The introductory article will provide an accessible introduction to some commonly used concepts and terms associated with machine learning modeling and evaluation. The introduction also provides a description of the data set that was used as the common application example for each of the five machine learning methods.

What are Machine Learning Methods?

Machine learning methods are generally flexible, nonparametric methods for making predictions or classifications from data. These methods are typically described by the algorithm that details how the predictions are made using the raw data and can allow for a larger number of predictors, referred to as high-dimensional data. These methods can often automatically detect nonlinearities in the relationships between independent and dependent variables and can identify interactions automatically. These methods can be applied to predict continuous outcomes, generally referred to as regression type problems, or to predict levels of a categorical variable, generally referred to as classification problems. Machine learning methods can also be used to group cases based on a collection of variables known for all the cases.

Types of Machine Learning Algorithms

Generally, machine learning techniques can be divided into two broad categories, supervised and unsupervised. The goal of supervised learning is to optimally predict a dependent variable (also referred to as "output," "target," "class," or "label"), as a function of a range of independent variables (also referred to as “inputs,” “features,” or “attributes.”). A classical example of supervised machine learning with which survey and social scientists are familiar is ordinary least squares regression. Such a technique relies on a single (continuous) dependent variable and seeks to determine the best linear fit between this outcome and multiple independent variables. Unsupervised learning, on the other hand, is more complex, in that there is no prespecified dependent variable, and these methods focus on detecting patterns among all the variables of interest in a dataset. One of the most common unsupervised methods with which social scientists and market researchers might have some familiarity is hierarchical cluster analysis – also known as segmentation. In this case, the main interest is not on modeling an outcome based on multiple independent variables, as in regression, but rather on understanding if there are combinations of variables (e.g., demographics) that can segment or group sets of customers, respondents or members of a group, class, or city. The final output of this approach is the actual grouping of the cases within a data set, where the grouping is determined by the collection of variables available for the analysis.

Tuning Parameters for Machine Learning Methods

Unlike many traditional modeling techniques such as ordinary least squares regression, machine learning methods require a specification of hyperparameters, or tuning parameters before a final model and predictions can be obtained. These parameters are often estimated from the data prior to estimating the final model. It could be useful to think of these as settings or “knobs” on the “machine” prior to hitting the “start button” to generate the predictions. One of the simplest examples of a tuning parameter comes from K-means clustering. Prior to running a K-means clustering algorithm, the machine learning algorithm needs to know how many clusters it should produce in the end (i.e., K). The main point is that these tuning parameters are needed prior to computing final models and predictions. Many machine learning algorithms have only one such hyperparameter (e.g., K-means clustering, LASSO, tree-based models) and others require more than one (e.g., random forests, neural networks).

The Context for Machine Learning Methods: Explanation versus Prediction

Machine learning methods are algorithmic and focus on using data at hand to describe the data generating mechanism. In applying these more empirical methods in survey research, it is important to understand the distinction between models created and used for explanation versus prediction. Breiman (2001) refers to these two end goals as the two statistical modeling cultures, and Shmueli (2010) refers to them as two modeling paths. The first of these modeling paths consist of traditional methods or explanatory models that focus on explanation, while the second one consists of predictive models that focus on prediction of continuous outcomes or classification for categorical outcomes. While machine learning or algorithmic methods can be used to refine explanatory models, their most common application lies in the development of prediction or classification models. The goals and methods for constructing and evaluating models from each of these two paths overlap to some degree, but in many applications, there can be specific differences that should be understood to maximize their utility in both research and practice. We turn now to a brief overview of explanatory models and predictive models in an effort to elucidate some of the key distinctions in these approaches that are needed in order to understand how predictive models developed using machine learning methods are evaluated in practice.

A Recap of Explanatory Models

In many social sciences applications, a relevant underlying theoretical model posits a functional relationship between constructs and an outcome of interest. These constructs are then operationalized into variables that are then used in the explanatory model for exploration and hypotheses testing. For example, researchers who are looking to understand the adoption of new technologies might posit a path model that is informed by the underlying theoretical technology adoption lifecycle model (Rogers 1962). Taking one step beyond explanation, these models can also be used to make causal inferences about the nature of the relationships between the observed variables and the outcome of interest. Another common interest among survey researchers is understanding correlates of nonresponse as well as possible causal pathways of it. In fact, survey researchers have a long history of conducting nonresponse follow-up surveys to gather additional information thought to be related to survey participation, or in the causal pathway, that go beyond known auxiliary variables. An explanatory model can be constructed using all of the available information and then used to test various hypotheses about how the variables, or relationships among variables, impact survey participation. But this type of model may have very limited utility for predicting nonresponse as it contains variables not likely to be available from all sampled units prior to the survey.

Explanatory models are commonly used in research and practice to facilitate statistical inferences rather than to make predictions, per se. The underlying shape (e.g., linear, polynomial terms, nonlinear terms) and content of these models is often informed by the underlying theory, experience of the researcher, or prior literature. Well-constructed explanatory models are then used to investigate hypotheses related to the underlying theory as well as to explore relationships among the predictor variables and the outcome of interest. These models are constructed to maximize explanatory power (e.g., percentage of observed variance explained) and proper specification to minimize bias while also being attentive to parsimony. Hence, evaluation of these models focuses on goodness of fit; simplifications of the models are driven by evaluating the significance of the predictors and overall goodness of fit indices. The inclusion of important predictors in the final model is often quantified using effect size measures, confidence intervals, or p-values for estimated coefficients.

The Basics of Predictive Modeling

In contrast to explanatory models that explore relationships among observed variables or confer hypotheses, prediction or classification models are constructed with the primary purpose of predicting or classifying continuous or categorical outcomes, respectively, for new cases not yet observed. Prediction for continuous numeric variables, also referred to as quantitative variables, is usually referred to as a *regression problem*, whereas prediction for categorical, qualitative variables is referred to as a *classification problem*. For example, in responsive survey designs, it is often useful to have an accurate classification of which sampled units are likely to respond to a survey and which are not. Within an online survey panel context, it might also be useful to know which respondents are likely to leave an item missing on a questionnaire and which respondents are not. Armed with these predicted classifications, researchers and practitioners can tailor the survey experience in an attempt to mitigate the negative consequences of nonresponse or item missingness.

Predictive models are constructed from data and leverage associations between predictor variables and the outcome of interest. These models are constructed by minimizing both estimation variance and bias, and because of this, balance predictive models, in the end, may trade off some accuracy for improved empirical precision (Shmueli 2010). In contrast to many explanatory models, the actual functional form of the predictive model is often not specified in advance as these models place much less emphasis on the value of individual predictor variables and much more emphasis on the overall prediction accuracy. In fact, most predictive models that are constructed using various machine learning methods produce no table of coefficient estimates or specific statistics that evaluate the significance of a given predictor variable. And because the focus of these models is on prediction, they must use variables that are available prior to observing the outcome of interest. Such variables are said to have ex-ante availability. In the case of responsive designs, where a prediction of nonresponse is desired in real time throughout the field period, the types of ex-ante variables may include auxiliary variables known for all sampling units or paradata that are collected on all sampled units during an initial field period. Certainly, these variables should be associated with survey response, but they may not provide a complete picture of why sampled persons or households participate in the survey or answer a given item. But the purpose and use of these models has less to do with fully explaining or confirming the causal mechanisms of nonresponse and more to do with correctly classifying sampled units as respondents or nonrespondents, and using this classification as the basis of tailoring or adjustment.

Evaluating Predictive Models Created Using Machine Learning Methods

Compared to traditional statistical methods, machine learning techniques are more prone to overfitting the data, that is, to detecting patterns that might not generalize to other data. Model development in machine learning hence usually

Table 1 A typical confusion matrix for a binary classification problem displaying cell counts.

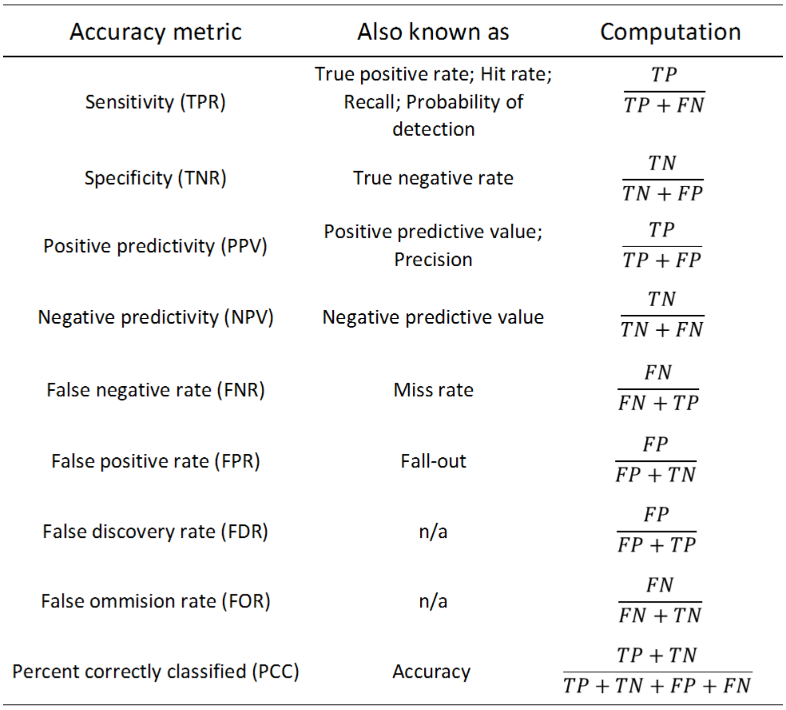
Actual class	Predicted class	
	Yes (1)	No (0)
Yes (1)	TP	FN
No (0)	FP	TN

TP = True positive; FN = False negative

FP = False positive; TN = True negative

relies on so-called *cross-validation* as one method to curb the risk of overfitting. Cross-validation can be implemented in different ways but the general idea is to use a subsample of the data, referred to as a *training* or *estimation sample*, to develop a predictive model. The remaining sample, not included in the training subsample, is referred to as a *test* or *holdout sample* and is used to evaluate the accuracy of the predictive model developed using the training sample. Some machine learning techniques use a third subsample for tuning purposes, that is, the *validation sample*, to find those tuning parameters that yield the most optimal prediction. In these cases, once a model has been constructed using the training sample and refined using the validation sample, its overall performance is then evaluated using the test sample. For supervised learners, these three samples contain both the predictor variables (or features) and the outcome (or target) of interest.

The predictive accuracy for machine learning algorithms applied to continuous outcomes (e.g., regression problems) are usually quantified using a root mean squared error statistic that compares the observed value of the outcome to a predicted value. In classification problems, the predictive accuracy can be estimated using a host of statistics including: sensitivity, specificity, and overall accuracy. Generally, the computation of these and related measures of accuracy are based on a *confusion matrix*, which is simply a cross-tabulated table with the rows denoting the actual value of the target variable for every sample or case in the test set and the columns representing the values of the predicted level of the target variable for every sample or case in the test set. An example confusion matrix applied to a binary classification problem displaying the counts of cases in each of its four cells is displayed in Table 1. The abbreviations in Table 1 represent: the number of true positives – that is the number of cases that were predicted to be a "Yes" for the binary target variable that actually had that value; the number of false negatives – that is the number of cases that had an actual value of "Yes" for the target variable but which were predicted to be a "No"; the number of false positives – that is the number of cases that had an actual value of "No" but which were predicted to be a "Yes" and finally, the number of true negatives – that is the number of cases that had an actual value of "No" that were predicted to be as such.



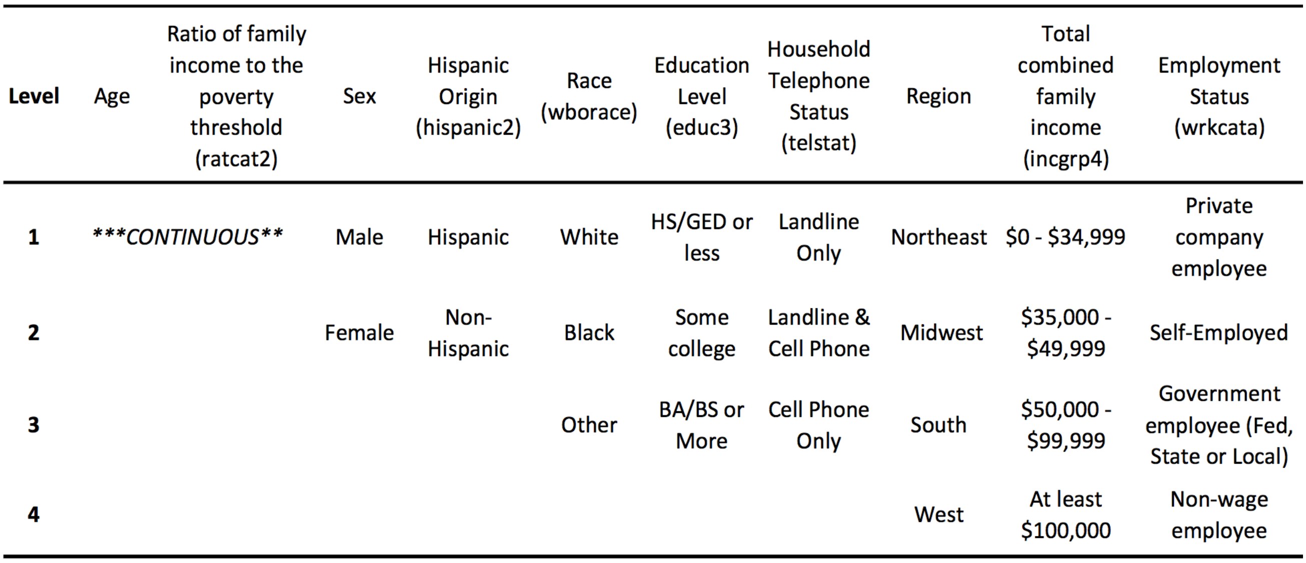
Accuracy metric	Also known as	Computation
Sensitivity (TPR)	True positive rate; Hit rate; Recall; Probability of detection	$\frac{TP}{TP+FN}$
Specificity (TNR)	True negative rate	$\frac{TN}{TN+FP}$
Positive predictivity (PPV)	Positive predictive value; Precision	$\frac{TP}{TP+FP}$
Negative predictivity (NPV)	Negative predictive value	$\frac{TN}{TN+FN}$
False negative rate (FNR)	Miss rate	$\frac{FN}{FN+TP}$
False positive rate (FPR)	Fall-out	$\frac{FP}{FP+TN}$
False discovery rate (FDR)	n/a	$\frac{FP}{FP+TP}$
False ommision rate (FOR)	n/a	$\frac{FN}{FN+TN}$
Percent correctly classified (PCC)	Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$

Table 2 A battery of accuracy metrics for binary classification problems defined in terms of the cells of the confusion matrix displayed in Table 1.

As mentioned earlier, there are a host of statistics that can be computed to estimate the accuracy of machine learning models applied to binary classification problems. Many of these statistics can be extended to the case of more than two levels in the target variable of interest. Since many of the survey related outcomes like survey response can be posed as a binary classification problem, we will illustrate these accuracy metrics using the confusion matrix that is given in Table 1. In Table 2, we define several common accuracy metrics for binary classification problems explicitly in terms of the cell counts displayed in Table 1. One additional metric that is not simply defined in terms of the cells of the confusion matrix is the area under the curve (AUC) and receiver operating characteristic (ROC) curve. This curve plots the true positive rate (sensitivity) versus the false positive rate (1-specificity) for various object values of a cutoff used for creating the binary classifications. Values of the AUC statistic that are close to 0.5 indicate very poor fitting classification models, while values that are higher and closer to 1 indicate more accurate classification models. The technical interpretation of the AUC and ROC curve statistic is the probability that the classification model will rank a randomly chosen "Yes" case higher than a randomly chosen "No" case.

Common Example Description

Within each of the four papers, we will apply the respective machine learning method to predict a simulated binary response outcome using several predictors using data from the 2012 US National Health Interview Survey (NHIS). Specifically, the demo data set (henceforth referred to as the DDS) consists of complete records from 26,785 adults aged 18+ that were extracted



Level	Age	Ratio of family income to the poverty threshold (ratcat2)	Sex	Hispanic Origin (hispanic2)	Race (wborace)	Education Level (educ3)	Household Telephone Status (telstat)	Region	Total combined family income (incgrp4)	Employment Status (wrkcata)
1	***CONTINUOUS**		Male	Hispanic	White	HS/GED or less	Landline Only	Northeast	$0 - $34,999	Private company employee
2			Female	Non-Hispanic	Black	Some college	Landline & Cell Phone	Midwest	$35,000 - $49,999	Self-Employed
3					Other	BA/BS or More	Cell Phone Only	South	$50,000 - $99,999	Government employee (Fed, State or Local)
4								West	At least $100,000	Non-wage employee

Table 3 Predictor variables used for generating the survey response outcome and modeling it using various machine learning methods.

from the 2012 public use data file. More complete details about this specific data set have been described elsewhere (Buskirk and Kolenikov 2015), and a complete description of both the NHIS study and the entire corpus of survey data is available at: http://www.cdc.gov/nchs/nhis.htm

The primary application of each of the methods we discuss in the papers in this special edition will be to predict a *binary survey response variable* using a battery of demographic variables available in the DDS including: region, age, sex, education, race, income level, Hispanicity, employment status, ratio of family income to the poverty threshold and telephone status. The exact levels of these predictor variables are provided in Table 3. The binary survey response variable was randomly generated from a simulated probit model that was primarily a nonlinear function of these demographic variables. More specific information about the exact form of the simulated probit models and how the binary survey response was randomly generated for each adult in the DDS are provided in the online technical appendix.

To evaluate model performance, we used a split sample cross-validation approach that created a single training data set (trainDDS) consisting of a random subset of approximately 85% of the cases in DDS along with a test data set (testDDS) consisting of the remaining cases. Each of the methods described in this special issue was applied to predict the simulated survey binary response variable using the core set of aforementioned demographic variables. Specifically, models were estimated using data from all cases in the trainDDS. In turn, these estimated models were then applied to the testDDS. The performance of each of the methods was measured by how well the estimated models predicted survey response status for cases in testDDS using the following accuracy metrics: percent correctly classified, sensitivity, specificity, balanced accuracy (average of sensitivity and specificity), and the AUC.

REFERENCES

Breiman, L. 2001. “Statistical Modeling: The Two Cultures.” *Statistical Science* 16 (3): 199–231.

Buskirk, T., and S. Kolenikov. 2015. “Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification.” *Survey Insights: Methods from the Field Weighting: Practical Issues and “How to” Approach*. http://surveyinsights.org/?p=5108.

Rogers, E.M. 1962. *Diffusion of Innovations*. New York, NY: Free Press of Glencoe.

Shmueli, G. 2010. “To Explain or to Predict?” *Statistical Science* 25 (3): 289–310.

SUPPLEMENTARY MATERIALS

<strong>Supplemental</strong> Datasets for all papers

Download: https://www.surveypractice.org/article/2718-an-introduction-to-machine-learning-methods-for-survey-researchers/attachment/9403.zip