Wesleyan Economic Working Papers

http://repec.wesleyan.edu/
N°: 2014-003

Comparing Standard Regression Modeling to Ensemble Modeling: How Data Mining Software Can Improve Economists' Predictions

Joyce P. Jacobsen, Laurence M. Levin and Zachary Tausanovitch

December, 2014

WESLEYAN
UNIVERSITY

Department of Economics
Public Affairs Center
238 Church Street
Middletown, CT 06459-007

Tel: (860) 685-2340
Fax: (860) 685-2301
http://www.wesleyan.edu/econ

Comparing Standard Regression Modeling to Ensemble Modeling: How Data Mining Software Can Improve Economists' Predictions

Joyce P. Jacobsen, Wesleyan University
Laurence M. Levin, VISA
Zachary Tausanovitch, Network for Teaching Entrepreneurship

December 14, 2014

Abstract:

Economists' wariness of data mining may be misplaced, even in cases where economic theory provides a well-specified model for estimation. We discuss how new data mining/ensemble modeling software, for example the program TreeNet, can be used to create predictive models. We then show how for a standard labor economics problem, the estimation of wage equations, TreeNet outperforms standard OLS regression in terms of lower prediction error. Ensemble modeling also resists the tendency to overfit data. We conclude by considering additional types of economic problems that are well-suited to use of data mining techniques.

Corresponding author: Jacobsen; jjacobsen@wesleyan.edu; 860-685-2357

RRH: Comparing Standard Regression Models to Ensemble Models

JEL codes: C14; C51; J31

Keywords: data mining; ensemble modeling

Comparing Standard Regression Modeling to Ensemble Modeling: How Data Mining Software Can Improve Economists' Predictions

INTRODUCTION

Two currently popular buzzwords are "Big Data" and "data mining." The idea that data sets are too large to be subjected to traditional analysis methods, and that patterns are to be discerned in data utilizing new computational techniques has become engrained in the popular culture and media, even as few people actually understand what these methods entail.

Economists have traditionally been wary of the idea of data mining, even as data mining has become common in other areas such as market analyses. The general portrayal is that data mining is akin to a "fishing expedition," where one looks inductively for correlative patterns in data without being guided by the hand of theory. In other words, the model is derived inductively from the data, although it can then be formalized and verified by using another data set or a reserved subset of the data. As economists are generally taught to have a theory firmly in hand before approaching data, they are naturally wary of data mining techniques.

One of the main ways to estimate these models is a technique known as ensemble modeling. In ensemble modeling, the model is estimated many times on different subsamples, with the goal of taking the best qualities of each model and combining them into a single (ensemble) model, similar to a neural net (but the estimation technique is different, so this is not the same thing as a neural network). The actual estimation method can vary; the particular method we employ herein is tree-building using a particular proprietary algorithm implementation, TreeNet.

A scan of the economics literature does not turn up any articles utilizing ensemble method techniques (whether tree-based or neural net) prior to Varian's recent [2014] primer on big data, wherein he presents several short examples of how ensemble methods can be used. Currently, in the ensemble modelling literature, the types of empirical problems that economists generally deal with (often broken down in econometrics courses into cross-sectional, panel, and time-series methods) are really not mentioned. Only one article [Wichard and Ogorzalek 2007] addresses how one could do time series prediction using ensemble modeling, and the authors hail from an electrical engineering background. Indeed, there is an active research program to improve time series forecasting (cf. the International Competition of Time Series Forecasting held in 2011)—with little if any input from economists. Thus the economics profession's continued focus on regression-based estimation techniques seems somewhat narrow given the possibility of alternative nonlinear (including discrete) techniques that may be able to outperform regression in predictive power.

This misguided focus may be partly a result of the economist's tendency to evaluate models by goodness-of-fit measures (such as adjusted R-squared) rather than by prediction error. This is a natural tendency in cases where the available data set is small and nonreplicable. However, bootstrapping techniques, and the closely-related technique of boosting (which is what TreeNet, the program we discuss below, uses), are familiar to most economists and are a way of moving away from treating samples as *sui generis*. In boosting, the model is built upon a subset of the data, i.e., fitted to a subset of data—and even then each model in the ensemble is built on a separate random sample of the data subset, in a process called "training." Once the ensemble model is developed from these repeated random samples, the model can then be evaluated on the

remaining data which had not been used in the training process, as well as utilized on additional data sets to see how well it predicts in the out-of-sample test.

This final phase, similar to seeing how well a time-series model forecasts the next period, is, in our view, the acid test of a model. Indeed, as we show below, there is a tendency in using a regression model to overfit to the specific dataset, and thus regression models tend to underperform ensemble models when evaluated by mean squared prediction error (i.e., they have larger mean squared prediction error). We show this is the case both when the model is specified *a priori* and when modified through feedback from the sample (e.g., by using either a stepwise regression technique or by a "researcher feedback" approach), and also true both when the model is specified sparsely and when it is specified using the "kitchen sink" approach. Note the topic of how to create "appropriately" or "approximately" sparse regression models is also a newly active area of applied econometric research [Belloni, Chernozhukov, and Hansen 2014] now that larger data sets, defined as large by virtue of having both high observation counts and many variables per observation, become increasingly available to economists.

HOW TREENET WORKS

TreeNet is an implementation of the gradient boosting method invented by Stanford statistician Jerome Friedman in the late 1990s [Friedman 1999a,;1999b], also referred to as MART, or "Multiple Additive Regression Trees." It is sold by the company Salford Systems, which can provide full documentation on the methodology [Salford Systems 2001-2005].

TreeNet is a program that models data using a series of decision trees. A simple example: consider a binary model that asks the question: what determines whether or not you are accepted at a particular university? In the first pass, consider SAT composite score as the only

independent variable. While a probit or logit regression model would give you a coefficient and thus a probability of acceptance conditional on your score, a decision tree will instead return nodes; for instance, that SAT scores higher than 1800 lead to acceptance and lower scores lead to rejection. This is a one-decision-node, two-terminal-node tree, which could be represented by a horizontal line at SAT = 1800, which would act as a boundary between acceptance and rejection. Upon adding additional decision factors, we could then construct this tree with more nodes and reuse independent variables at separate nodes as many times as necessary in order to fit the model. For instance, if we also add GPA as an independent variable, the model could possibly state: if SAT > 1800 and GPA > 3.0 or if GPA < 3.0 and SAT > 2100 or if GPA > 3.5 and SAT < 1800 then accept; otherwise reject.

TreeNet then uses an optimizing technique called boosting in order to come up with the ensemble model. For instance, twenty percent of the data could be randomly selected for each model and a tree structure built on that subsample. Then TreeNet builds another decision tree on another random sample of the data, and moves points along the first tree closer to the second tree by adding additional nodes (or, if you prefer, adjusting the weights on the nodes). The researcher can determine how much the initial tree changes by setting the "learn rate." By repeating this process with possibly hundreds (or even thousands) more trees, the original few-node model starts to look more like a continuous curve.

A decision tree could of course be too specific to the data on which it is built. Imagine a tree with a node for every data point; this model would be irrelevant for any other data set. To mitigate that possibility, TreeNet builds the model with only a portion of the full data set and tests it on the remaining data. Through the use of a loss function, it determines how divergent the estimates become, and when additional trees cause more divergence than reduced error, TreeNet

determines that the optimal number of trees has been reached. Also, by setting a minimum number of observations for each terminal node (implying that a split can only be made if the categories have enough people in each one), the one-to-one correspondence will be avoided and this will aid the model in avoiding overfitting.

TreeNet provides visual diagnostics so that the researcher can see how the model is performing. Figure 1 gives an example of such a diagnostic diagram. This figure shows the divergence of error between the training and the testing samples. The less error for both models the better; also the less they diverge the more externally valid the model is (i.e., the better it works on the testing group). If the training error is fluctuating too much this can also be an indicator of low external validity (i.e., oversensitivity to number of trees). The green line indicates the number of trees at which the model has a balance between low error and low divergence test results; this is reported as the optimal number of trees.

<<Figure 1 here>>

Another tool for evaluating model fit and results is a gains chart, which compares how well the model predicts dependent variable values (or outcomes) relative to no model (i.e., predicting the mean value in the sample as the outcome for each person). Figure 2 gives an example of this diagnostic diagram. On the y-axis is the cumulative predicted sum of the dependent values for the observations, sorted along the x-axis by closeness to the actual value. The more area between this line and the 45-degree line, the better the prediction (so if the line is exactly 45 degrees, the model does not perform better than simply guessing the mean value for each person). For instance, the second green point on the graph shows that the best eleven percent of the predictions account for sixteen percent of the real total of the dependent variable.

<<Figure 2 here>>

In order to also calculate prediction error (as we do below), a sample can be divided into three parts, a training section, a testing section, and then (once the model has been set through the training and testing phases) a final section in which the model is used to see how well it predicts out of sample. This is the method that we use in evaluating our models as described below.

AN APPLICATION EXAMPLE TAKEN FROM LABOR ECONOMICS

One of the most standard empirical applications in economics is to estimate an earnings equation based on human capital-related variables and other variables (such as demographic factors) in order to explain earnings differences among people. Such an equation is based on standard neoclassical theory (namely that greater amounts of human capital such as education and work experience will lead to higher hourly earnings as a return on those human capital investments), and in extension has even led to general agreement about the appropriate functional form, namely that the returns to human capital will tend to rise and then fall over the lifecycle as the person first invests in more human capital early in life and then will experience depreciation and thus reduced returns later in life. Thus it is common to fit a quadratic in work experience (or to tenure in a particular firm or position) to represent this pattern; indeed, quadratics in experience, age, or tenure generally outperform strictly linear models. Over time there has also been general consensus that the best model is to use a natural log of earnings as the dependent variable, thus estimating a semi-log specification (where the coefficients thus can be read as showing percentage effects of the independent variables on earnings).[1] These earnings models have been estimated thousands of times for different years, countries, and demographic

groups and thus represent one of the most accepted theoretical/empirical models in the economics literature.[2]

This type of model thus provides an ideal test to demonstrate how ensemble modeling can nonetheless outperform a well-established regression model that is built on a firm economic theory basis. Outperformance is here measured by mean squared prediction error rather than by R-squared or adjusted R-squared; TreeNet provides R-squareds but not adjusted R-squareds (since the node method does not adapt well to making the degrees-of-freedom adjustment calculation) so that we can also compare those measures, and as you will see TreeNet actually performs well in terms of R-squared as well while apparently avoiding the peril of overfitting more effectively than OLS regression.

We utilize two standard U.S. data samples for estimating our earnings regressions, the 2009 and 2010 Annual Demographic Samples, the March versions of the Current Population Survey. Along with self-reported earnings (where hourly earnings are calculated by taking annual earnings and divided by weeks worked times normal weekly hours worked), these data sets provide many standard variables that are viewed by consensus (and backed up by theoretical reasoning) as having an effect on earnings: education, race, gender, marital status, country of origin, whether or not one is a native English speaker; other income sources, and current location. While total work experience is not asked, it is common either to use age (and age-squared) as a proxy for this variable, or to construct potential experience as age minus education minus six. Note that we could potentially put in every reported variable and let TreeNet sort through them (real data mining); we do an initial cull since this paper is meant to be expository rather than factfinding.

Rather than looking at tables of coefficients and standard errors or t-statistics, since TreeNet doesn't produce any coefficients, we can look at variable importance plots and interaction plots. For example, Figure 3 shows a variable importance plot from a model that predicts earnings. The results are normalized to the most important variable (in TreeNet's view), which is the variable that has the largest effect on the tree results; in this case that variable is education. Notice that financial assistance income is relatively unimportant; indeed, dropping low importance variables may improve the model. Because TreeNet already considers nonlinear relationships in fitting the model, age-squared is already accounted for and need not be included specifically in the variable list; thus it has no additional effect when explicitly included.

<<Figure 3 here>>

Figure 4 (a and b) shows one-predictor and two-predictor plots of the effects of independent variables on the final result. Figure 4 (a) shows how age can affect the natural log of hourly earnings, indicating the nonlinear effect and how age has an increasing positive effect relative to the sample average with the crossover occurring in one's early thirties (so that after that age people make more than the mean earnings for the sample). Figure 4 (b) shows how age and educational attainment can jointly affect the natural log of hourly earnings, with rising earnings for older ages and more years of education (but more marginal effect from higher levels of education than from age); the marginal (one-predictor) graphs can be seen from this graph as well (compare the marginal view of age in Figure 4 (b) to Figure 4 (a)). These graphs are consistent with standard expectations regarding the relationships between age and education to earnings.

<<Figure 4 here>>

More importantly, we can evaluate goodness-of-fit for TreeNet vs. other models by utilizing both R-squared measures for in-sample fit, and prediction error measures for out-of-sample fit (i.e., predictive ability). We compare three different OLS specifications to the results obtained from TreeNet. The full set of variables that we used included education, gender, marital status (married or not), age, potential years of experience (constructed as age minus education years minus six), metropolitan status, whether one is white non-Hispanic or not, whether one was born in the US or not, region, veteran status, number of people in the household, nonlabor income, and weeks of work missed. TreeNet was given all these variables to use, except potential years of experience was not constructed; rather it was given age and education, the underlying components. OLS Model 1 ("Sparse") includes the subset of the variables that are widely believed from previous work to be of most importance in explaining earnings (education, potential experience, sex, race, marital status). OLS Model 2 ("All Variables") includes the full set of variables but no polynomial or interaction terms. OLS Model 3 ("Complex") includes the full set of variables along with polynomial and interaction terms.[3] All the variables used are listed in the Appendix. For this third model, we purposefully experimented so as to maximize adjusted R-squared, in other words to try to fit the model as closely as possible (by that criterion) to the combined testing-and-training subsample. Each model was built using two-thirds of the relevant data from the Current Population Survey, and then used on the remaining third of the data for predictive accuracy (i.e. to calculate mean squared prediction error for this subsample).

The first two sections of Table 1 show the mean squared prediction error (MSPE), R-squared, and adjusted R-squared (for the regression models only) for the four models (3 OLS, 1 TreeNet) on both the 2009 and 2010 CPS samples. For both the 2009 and 2010 CPS samples, TreeNet actually has higher R-squared than both the sparse and all-variable OLS specifications.

Note that adjusted R-squared differs little for those specifications relative to regular R-squared, as the sample sizes are quite large for the CPS (thousands of observations). Only the complex model, where the model has been specifically overfitted to the samples, outperform the TreeNet models in terms of goodness of fit.

<<Table 1 here>>

However, TreeNet outperforms all three OLS models in terms of lowest MSPE. In particular, TreeNet substantially outperforms the overfitted model on the 2009 data, which had the highest adjusted R-squared of all the OLS models.

As a final test of TreeNet's robustness, we used the 2009 TreeNet model to predict on the 2010 CPS data and vice versa. The results are shown in the last row of Table 1. In both cases, TreeNet still has lower or at least as low MSPE as the best-predicting OLS model.

One concern that researchers may have about moving away from standard regression methodologies is the lack of familiar output. While TreeNet (and related methods) does not provide the researcher with the standard type of output, namely a set of regression coefficients and accompanying standard errors, it does still provide a tangible output in the form of the actual decision trees (in the case of TreeNet at least—this would not be the case in the more black box-neural network method of predicting outcomes). These are provided as a SAS program that basically consists of a series of if-then-else statements. Essentially this program can be run on any new data set to generate not only predicted values for the dependent variable given the independent variables for any observation, but also to generate hypothetical cases.

Thus, in the labor economics example of predicting wages, one can easily perform a standard wage decomposition or a hypothetical such as asking what a woman would earn if she were subject to the male system on her independent variables. This method, known as the

Oaxaca-Blinder decomposition (as shown in the two important 1973 articles in the wage differential literature: Blinder [1973] and Oaxaca [1973]), is the standard way of creating adjusted wage ratios where adjustments can be made to account for differences in the average woman's experience and training relative to the average man's, and thus consider how much of the gender wage gap is attributable to differences in treatment as opposed to differences in the means by gender of the independent variables.

Table 2 shows the results using TreeNet as well as Models 1 and 3 to create adjusted gender earnings ratios. In each case women's earnings are adjusted to ask how much they would make if they had the same earning-related characteristics as do men (so the numerator is calculated by multiplying the women's equation coefficients times the men's mean values for the variables). Here the OLS models are the same as in Table 1, except the samples are split by gender and all variables and interaction terms involving gender are dropped from the equations. We drop using Model 2 from this table as it does not yield notably different results from the complex model.

<<Table 2 here>>

Interestingly, while in older samples this adjustment method tended to make the women/men hourly earnings ratio move closer to one than in the raw data, for both years and for both the sparse and complex models the ratio actually varies little from the actual mean values in the sample, and even makes the ratio worse in some cases. This may be attributable to the fact that these are recession years in the US and in this recession, many men lost their jobs and thus would not be in our sample of earners. In addition, women's educational attainment has recently overtaken men's attainment, making it less likely that this adjustment would make as much of a difference on earnings as in the recent past. However, the TreeNet adjustment moves the ratio

much more in the direction towards equality, implying that there may be something related to nonlinearities in the recent data that are not captured by the OLS models. Thus there is a substantive difference in this public policy-related calculation between the two model estimation methods that calls for additional investigation.

Of course, economists and other researchers who place primary emphasis on inference based on parameter values will still not be satisfied by this estimation approach, because there are still no coefficients to recover. Thus, as Stock [2010] points out that much of the profession has moved more towards trying to recover more credible estimates of a smaller number of key parameters, data mining techniques will be less acceptable for these situations when it is important to be able to measure a specific marginal effect (for instance, policy evaluation studies). In these cases, ensemble modelling may be of more use in early phases of model testing, for instance to reduce the set of covariates that might be used as controls in the final regression model. And in cases where carefully controlled experiments are carried out in order to measure the effect of one key element, then there will be little or no need for data mining techniques to be used.

CONCLUSIONS AND EXTENSIONS

In the example delineated in the preceding section, TreeNet has produced similar results to those that we can produce using OLS, and the results may be preferable in terms of lower MSPE. In addition, there can still be substantively different results from the two modeling methods that may have significant policy implications, as demonstrated by the adjusted earnings ratio calculations. This shows that use of ensemble models can complement or even replace the work that has traditionally been done using standard regression models. In cases where there is

less theoretical guidance regarding the functional form of the model, this methodology may be preferable because it automatically considers variable interaction and nonlinear forms. In addition, in cases where the variables to be included are not clearly delineated by theory, TreeNet provides an alternative modeling choice to *a priori* exclusion of variables.

What other programs provide ensemble modeling routines? Three commonly-used statistical packages by economists--namely SAS, Stata, and the open-source package R--all provide options for ensemble modeling. SAS's data mining product, Enterprise Miner, was first issued in 1999. In Stata, the plugin/command boost is available (see Schonlau [2005] for a description of how it works). In R, there are currently at least seven ensemble packages available written by different people (see Bowles [2014] for a list and description).

What other types of economic modeling problems could be tackled using ensemble methods? Certainly projects that involve large quantities of data with multiple possible covariates: one area that Cook [2014] suggests as ripe for additional "big data" methods work is educational policy (as well as social policy more generally), where multiple variables available at the state, school district, school, and classroom level can be merged in and their interactions considered. For instance, Varian [2014] revisits the classic mortgage lending discrimination study done by Munnell, Tootell, Browne, and McEneaney [1996] to show how it could be redone using decision tree estimation. In this case, Varian notes that fitting a tree model that omitted race as an explanatory variable fit the data as well as a tree model that included it (where the regression model fitted in the study showed a significant negative effect of being black on the probability of receiving a mortgage).

In addition, many institutional research problems come to mind, such as trying to determine which factors lead some donors to give money to a charity out of a long list of

possible covariates, or trying to determine which individual characteristics lead a person to success or failure in some competitive venue such as getting into a competitive university or getting a loan. One could thus model institutional processes that happen in closed rooms without actually needing to know ahead of time how a variety of factors available to the decisionmakers are weighted in the decisionmaking process. Indeed, many empirical projects in economics receive much less guidance from economic theory, other than a list of potential factors that could have an impact on the outcome (think of consumer theory), than the earnings equation problem considered above. Thus economists may find themselves increasingly pressed to compete with other disciplines who are less wedded to a priori specifications in terms of who is able to make the more effective predictions of future outcomes. Perversely, the built-in tendency of ensemble methods to avoid overfitting may make these models more robust to specification error and thus more likely to avoid the general critique of economists' models that they are better at explaining the past than at predicting the future.

Acknowledgements

The authors wish to thank session participants at the May 2013 Eastern Economics Association meetings and an anonymous referee for helpful suggestions, and Wesleyan University for research support.

Notes

1. We do not consider whether the dependent variable should be transformed but rather stay with this standard method; similar to regression modeling, TreeNet does not also automatically test different transformations of the dependent variable.

2. There are literally so many such studies that no single survey of the wage regression literature exists. See Willis [1986] for a classic survey of human capital earnings functions; Jacobsen [2007], Chapter 7, for discussion and examples of many of these articles in the gender wage differential context.

3. A fourth model was run where experience and education were included as categorical values (i.e., as many dummies as the number of values minus one); this model did not outperform either model or 2 or model 3 by either the adjusted R-squared or the prediction error metrics, so we do not include results from it.

References

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2): 29-50.

Blinder, Alan. 1973. Wage Discrimination: Reduced Form and Structural Estimates. *Journal of Human Resources*, 8(4): 436-455.

Bowles, Mike. 2014. Ensemble Packages in R. Revolutions Blog (April 8), Revolution Analytics. http://blog.revolutionanalytics.com/2014/04/ensemble-packages-in-r.html.

Cook, Thomas D. 2014. 'Big Data' in Research on Social Policy. *Journal of Policy Analysis and Management*, 33(2): 544-547.

Friedman, Jerome H. 1999a. Stochastic Gradient Boosting. Technical Report, Dept. of Statistics, Stanford University.

-----. 1999b. Greedy Function Approximation: A Gradient Boosting Machine. Technical Report, Dept. of Statistics, Stanford University.

Jacobsen, Joyce P. 2007. *The Economics of Gender, Third Edition*. Malden, Mass.: Blackwell.

Munnell, Alicia H., Geoffrey M. B. Tootell, Lynne E. Browne, and James McEneaney. 1996. Mortgage Lending in Boston: Interpreting HMDA Data. *American Economic Review*, 86(1): 25-53.

Oaxaca, Ronald. 1973. Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, 14(3): 693-709.

Salford Systems. 2001-05. TreeNet: An Exclusive Implementation of Jerome Friedman's MART Methodology: Robust Multi-Tree Technology for Data Mining, Predictive Modeling and Data Processing. http://www.salford-systems.com/.

Schonlau, Matthias. 2005. Boosted regression (boosting): An introductory tutorial and a Stata plugin. *The Stata Journal*, 5(3): 330-354.

Stock, James H. 2010. The Other Transformation in Econometric Practice: Robust Tools for Inference. *Journal of Economic Perspectives*, 24(2): 83-94.

Varian, Hal R. 2014. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2): 3-28.

Wichard, Jorg D. and Maciej Ogorzalek. 2007. Time Series Prediction with Ensemble Models Applied to the CATS Benchmark. *Neurocomputing* 70(13-15): 2371-78.

Willis, Robert J. 1986. Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions. *Handbook of Labor Economics*, 1: 525-602.

Figure 1
TreeNet diagnostic for training-testing divergence and optimal number of trees

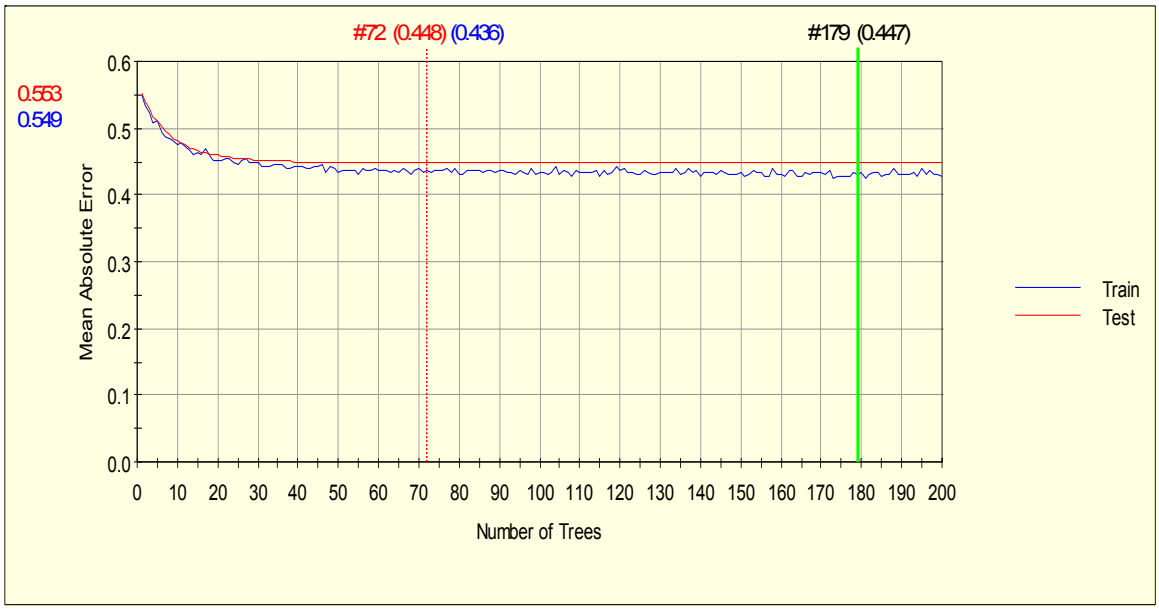


Figure 2
An Example of a TreeNet Gains Chart

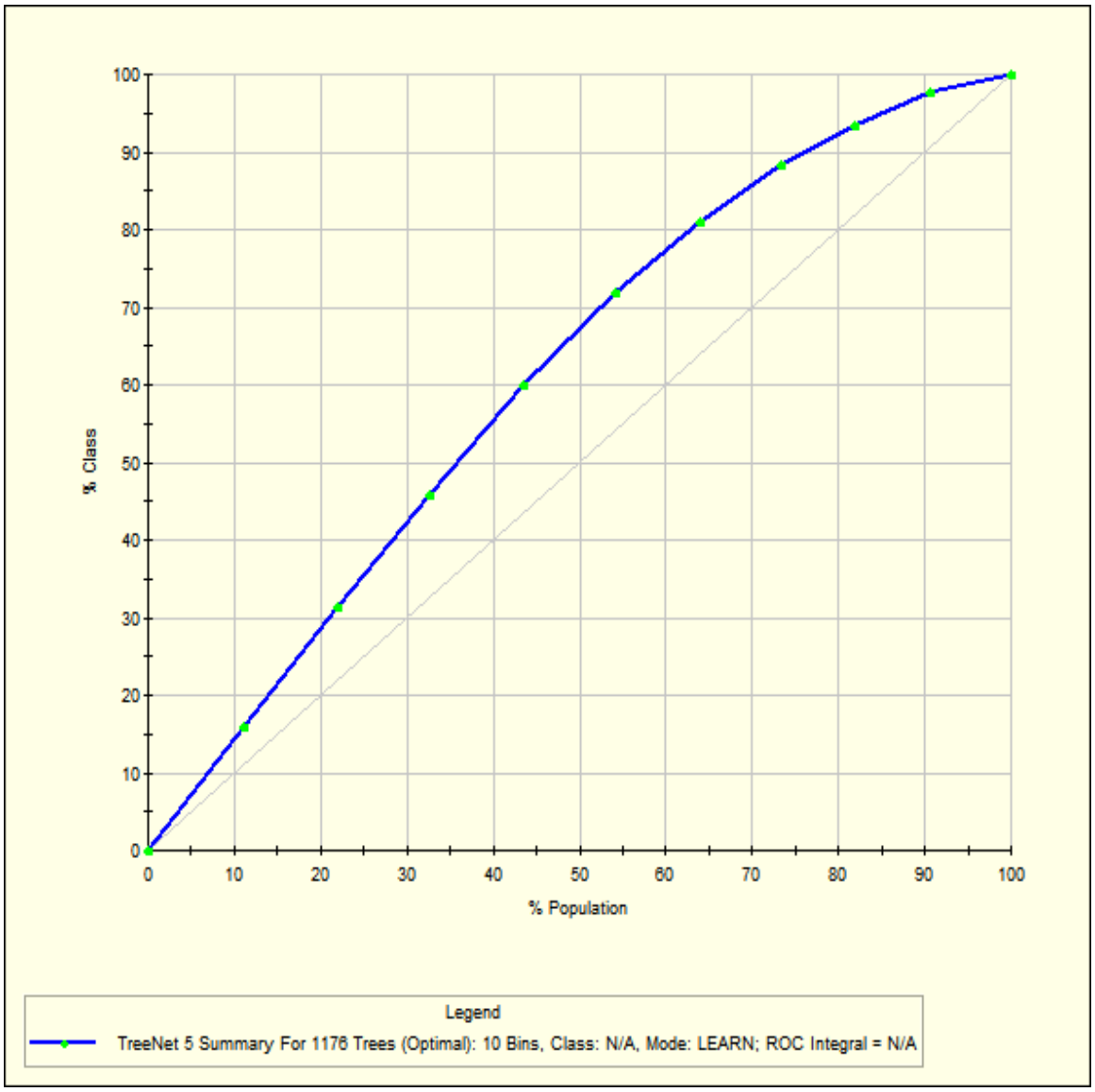
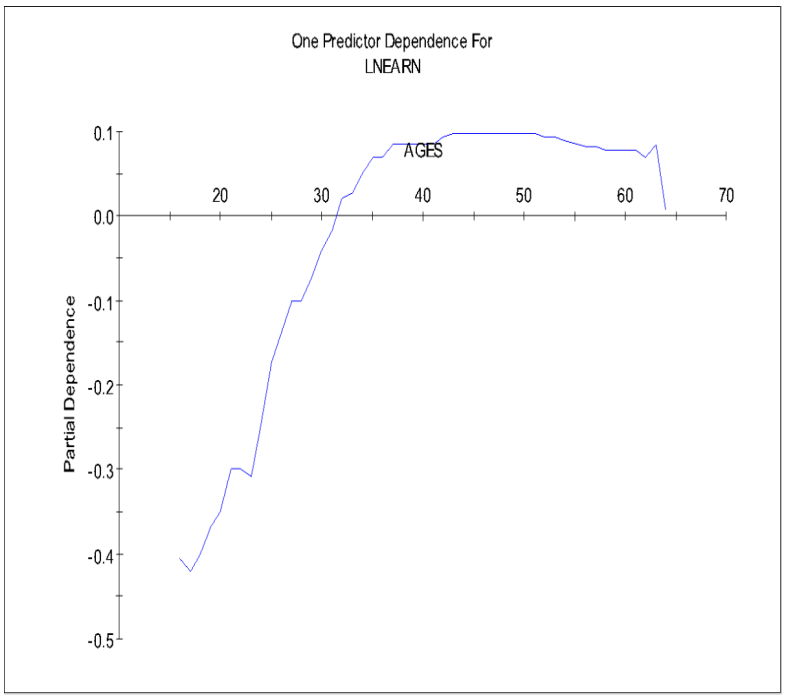


Figure 3
An Example of a TreeNet Variable Importance Plot

Variable	Score	
Education	100.00	|||||||||||||||||||||||||||||||||||||||||
Age	63.83	||||||||||||||||||||||||
Child support $ per year	56.82	||||||||||||||||||||||
Social security $ per year	41.29	||||||||||||||||
Gender	33.25	||||||||||||
Marriage Status	29.76	||||||||||
Country of Birth	18.08	|||||||
Race (white dummy)	16.29	||||||
Experience	16.11	||||||
Born in English country	8.82	|||
From the South	8.37	|||
Other income	4.63		
Financial assistance income	2.91	
Age Squared	0.00	

Figure 4
Examples of TreeNet Interactions Plots

(a) one predictor (age for earnings)



(b) two predictors (age and education for earnings)

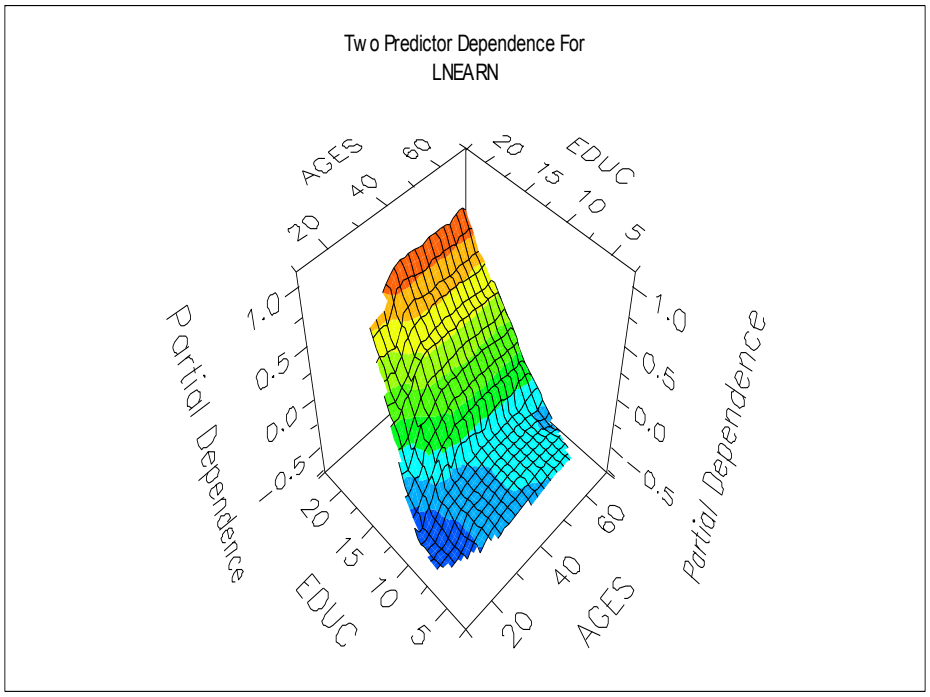


Table 1
Mean Squared Prediction Error and R-squareds, OLS and TreeNet Models, 2009 and 2010 CPS Samples

Year	2009			
Model	TreeNet	Sparse	All Variables	Complex
MSPE	0.373	0.378	0.375	0.510
R-Squared	0.334	0.315	0.322	0.359
Adj R-Squared	-	0.314	0.321	0.349

Year	2010			
Model	TreeNet	Sparse	All Variables	Complex
MSPE	0.394	0.399	0.397	0.399
R-Squared	0.314	0.303	0.306	0.321
Adj. R-Squared	-	0.303	0.305	0.313

Year	2009 data, 2010 model	2010 data, 2009 model
Model	TreeNet	TreeNet
MSPE	0.375	0.397
R-Squared	0.314	0.334

Table 2
Comparison of actual and adjusted female/male earnings ratios,
2009 and 2010 CPS Samples

Model	2009 data	2010 data
Actual values from sample	0.77	0.79
TreeNet	0.85	0.88
Sparse	0.76	0.79
Complex	0.79	0.76

Appendix

Variables Used in Regressions and TreeNet

Dependent Variable

Lnearn — The log real wage, calculated by taking the log of the person's earnings from the previous year, divided by hours usually worked times weeks worked last year.

Variables in "Sparse" Model

Education — A recoding from the level of schooling completed to the approximate years of schooling that would normally have been spent to achieve that level of attainment
Experience — Potential experience, calculated as age minus years of education minus six.
Gender — A dummy variable coded 1 for woman
Race — A dummy variable coded 1 for anyone whose ethnicity is more than 66 percent from a white nonshipanic background
Marriage — A dummy variable coded 1 for a person that is legally married
Serve — A dummy variable coded 1 for people who have ever served on active duty in the United States military
Rural — A dummy variable based on the metropolitan status provided by the CPS, coded 1 for people whose primary living space is non-metropolitan
South — A dummy variable coded 1 for a person living in the southern region of the United States

Additional Variables in "All Variable" Model

Age — An integer variable representing the age of the person
Non-English — A dummy variable coded 1 for people who were born in countries where English is not the primary language spoken
NotborninUS — A dummy variable coded 1 if the person was born outside the United States
h_numper — An integer variable for the number of people living in the person's household
Losewks — An integer variable that represents the number of self-reported weeks that the person failed to show up for scheduled work
ss_val — The total payment in dollars that the person receives in yearly social security payments
csp_val — The total payment in dollars that the person receives in yearly child support payments
vet_yn — A dummy variable for whether or not the person receives veterans payments
fin_yn — A dummy variable for whether or not the person receives financial assistance
oi_yn — A dummy variable for whether or not the observation receives any other income

Additional Interaction and Polynomial Terms in "Complex" Model

Agesq — The square of the age variable
Edsq — The square of the education variable
Exsq The square of the experience variable
Wed — An interaction term between being a female and education
Wedsq — The Wed variable squared
Wex — An interaction term between being a female and experience
Wexsq — The Wex variable squared
(Note: additional higher-order terms and interactions were tested and rejected.)